Fitting of Generalized Poisson Regression and Negative Binomial Regression models for analyzing of count time series event

Parykhan Abdulla Omer [1], Talar Mohamad Hussian [2]

[1, 2] Salahaddin University, Erbil College of Administration and Economics - Statistics & Informatics Department

***Abstract*—The focus of this paper is fitting the appropriation of two regression models of discrete count data, Poisson and Negative binomial regression models. The question is which one of these two models is the best choice for predicting the number of (EIA pax aircraft movement from Erbil international Airport) during specific period of time from (2015-2021). To model count data, Poisson regression has been widely used. It is frequently criticized, nevertheless, for the strong requirement of equidispersion in Poisson regression, which entails equality between the variance and mean of the dependent variable. Count data frequently displays excess zeroes and over-dispersion in many applications. While Negative binomial regression can model over-dispersed count data. There are instances of both overdispersion and underdispersion. One technique that can deal with overdispersion and underdispersion is Generalized Poisson regression (GPR). The data set is fitted to the specific models by a method called Maximum Likelihood estimation. This means that the unknown coefficients are estimated such as the likelihood of getting the given data is as large as possible. The dependent variable for the real data was the number of EIA pax aircraft movement weekly with the 13 independent variables. Four criterions used to check over dispersion and goodness of fit like Pearson χ^2statistic, Deviance, AIC and BIC as test statistic; these are the common ways of comparing likelihoods between different models with respect to the number of estimated parameters. Empirical results supported the Negative Binomial Regression Model fitted data set very well depending on the values of these criterions, as their smaller values indicate the best model. modeled dataset by available statistical software like SPSS V25 and Stata V16 and Stratigraphic V15.**

***Keywords*— Methodology, Poisson Regression, Negative Binomial Regression, Model selection**

I. INTRODUCTION

The most popular technique to model the relationship between dependent and independent variables is regression modeling. Different regression models are applied in real-world situations depending on whether the dependent variable is continuous or discrete. The dependent variable is frequently count data made up of integers that cannot have a negative value. The number of accidents involving natural gas pipelines, the number of airline delays, the number of party switches among deputies during an election year, the number of strikes per year in a nation, the number of accidents involving motor vehicles or the workplace that occur in a given day are all examples of count data. When this occurs, using standard regression analysis will result in skewed projected coefficients (See King, 1988).

The Poisson regression model is the most widely used regression model for count data. The dependent variable count data, which is used in the Poisson regression model, is derived from the Poisson distribution. Poisson regression models work well for data with an equal spread. The expected value and variance of the dependent variable has to be equal in order for there to be equal dispersion. Rarely is this the case. In several disciplines, including marketing, public health, and biomedical science, count data is particularly prevalent. Count data are typically used to model the number of occurrences of an event over a fixed time period. Regarding regression models, the classic linear regression model is unsuitable for the analysis of count data, since it violates the assumption of normality. Thus, generalized linear models are used to analyze data when linearity and normality assumptions are no longer valid.

Nelder and Wedderburn's (1972) first description of the Generalized Linear (GLM) has been further expanded upon and clarified by (See McCullagh and Nelder, 1989). It provides for alternative models of the mean than the classic linear regression, which models the mean as a linear function of the covariance. All GLMs have three components: a random component that determines the output variable's distribution; a systematic component that describes the covariates in linear form; and a link function that links the random component and the systematic components. The classic OLS regression is appropriate if the output variable's distribution is normal. Other distributions, such as binomial distributions, Poisson distributions, Negative Binomial distributions, etc., can be used in addition to the normal distribution (See Jiang, 2018).

To assess count data, Poisson regression and Negative binomial regression are widely employed. It is suitable for studying rate data as well. In extended linear models, the Poisson regression model belongs to a class of models (GLM).

Polytechnic Journal of Humanities and Social Sciences
Volume 4 No. 2 (2023)
Received 30 july 2022; Accepted 4 September 2022; Regular research paper: Published 15 August 2023
Copyright © 2023 Parykhan,Talar. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0)

It uses natural log as the link function and models the expected value of dependent variable. The model's natural log makes sure that the dependent variable's predicted values can never be negative. In Poisson regression, it is assumed that the dependent variable will follow a Poisson distribution. The Poisson distribution calls for the mean to be equal to the variance. There is frequently over-dispersion in count data. Over-dispersion occurs when the variance is significantly larger than the mean. The data is referred regarded as being over-dispersed when this occurs. The over-dispersion must be accounted by the analysis methods appropriate to the data. Poisson regression is insufficient for the analysis of over-dispersed data. Negative binomial regression is hence more suitable for over-dispersed data to overcome over-dispersion. This is due to the fact that negative binomial regression inherently has a higher variance than mean, which allows for over-dispersion. (See Brännäs and Johansson, 1994.)

2. Methodology and Methods

2.1 Generalized Linear Models (GLM)

Flexible expansions of ordinary general linear regression, generalized linear models (GLM) allow for the inclusion of dependent variable with distribution other than the normal distribution. The flexibility of a GLM is desired because the distribution is originally unknown and the purchasing behavior does not always follow a normal distribution. A GLM allows a link function to be related to the dependent variable, and furthermore, the variance of each measurement can be a function of its predicted value (See Olsson, 2002). In 1972, GLMs were first made available. GLMs provide a combined method for analyzing these many regression models rather than requiring separate studies for each one. Poisson regression models, linear regression models, zero-inflated regression models, logistic regression models, negative binomial regression models, the Poisson Hurdle model, and many other models are included in the Generalized linear models. These models use a common method for estimating parameters, and is one of their unique characteristics (See Adeti, 2016).

Additionally, it enables the linear model of several variables to be connected to a dependent variable using any number of different link functions. According to (Zurr et al., 2009), there are three processes involved in creating a GLM: a) selecting a dependent variable's distribution (Y), b) defining covariates (X), and c) selecting a link function between the dependent variable's mean (E(Y)) and a linear combination of the covariates (βX). consider a general linear model describes the observation ($i = 1,2, \dots, n$) of Y (the dependent or response variable) is as a linear function of ($p-1$) independent variables ($x_1, x_2, \dots, x_{p-1}$) as follows:

$$y_i = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i \quad \dots (1)$$

can be written in matrix form in the following way:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad \dots (2)$$
$$E(\underline{Y}) = X\underline{\hat{\beta}}$$

Where (Y) is a vector that consists of the observations of the dependent variable

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

And X is a matrix with dimension ($n \times p$), where $p = k + 1$

$$, \underline{X} = \begin{pmatrix} 1 & x_{11} & \dots\dots & x_{1k} \\ 1 & x_{21} & \dots\dots & x_{2k} \\ : & : & & : \\ . & . & & . \\ 1 & x_{n1} & & x_{nk} \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ . \\ \beta_k \end{pmatrix} \quad , and \ \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ . \\ \varepsilon_n \end{pmatrix}$$

From $\underline{X}$ the first column corresponds to the intercepts and the other columns contain the values of the independent variables (See Zurr et al., 2009).

$\underline{\beta}$ is a vector $\beta = (\beta_o, \beta_1, \beta_2, .., \beta_p)'$ containing the parameters p which are to be estimated and finally, ε is the residual vector $\underline{\varepsilon} = (\varepsilon_o, \varepsilon_1, \varepsilon_2, .., \varepsilon_p)'$

2.2 Poisson Regression Model

One of the distributions that is most frequently used in statistical applications is the Poisson distribution. Because the distribution's mean and variance are equal, it does have its limitations. In many situations this assumption is not realistic. To deal with this problem, researchers have used several Poisson definitions. Usually, to adjust for over-dispersion or under-dispersion, a mixing distribution is included or the Poisson is estimated with additional parameters. The Poisson regression model attempts to at modeling a counting variable Y, which counts how frequently a specific event occurs over a specified time period.

Poisson regression is frequently applied to count data. Count data is defined as "the number of occurrences of a behavior in a specific period of time" by Coxe et al (2009). Integers must only be non-negative in count data (See Karazsia et al, 2008). Hence A generalized linear regression model with a logarithmic link function is called Poisson regression. The phrase "models expanding the ordinary regression model to encompass non-normal dependent distributions and modeling function of the mean" was used by Agresti (2013) to describe generalized linear models (GLM). Three main assumptions govern Poisson regression as a generalized linear model (See Durrant, 2016 and Pesonen, 2018).

Inside the generalized family of linear models, the Poisson regression method is found (Hoffman, 2004; Agresti, 1996). In two ways, these models broaden the use of ordinary linear regression. In order to ensure that the dependent variable has conditional distributions that are not normal, they first define

them as linear functions of the explanatory variables, which specify transformations of the conditional means rather than the mean itself. The dependent variable's distribution can be skewed under various circumstances. The frequencies peak their highest point at the lowest number and rapidly decrease as they ascend. The Poisson distribution from the discrete distribution family can be expressed to represent variables with such asymmetric right-slope distributions (Moksony and Hegedus, 2014). Let x_i and x_i represent observations from a set of data. The numbers x_i and y_i in this instance represent a vector of arguments and dependent variables, respectively. The dependent variable y_i is assumed to exhibit the Poisson distribution in a Poisson regression study. Let xi and yi be observations from a data set. Here, the numbers xi and yi are respectively a vector of independent and dependent variables. Poisson regression analysis assumes that the yi shows the Poisson distribution. The probability density function for the Poisson distribution with the parameter λi is given in the following formula;

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \quad , \quad y_i = 0, 1, 2, \dots \qquad \dots (3)$$

The number of events occurring is denoted by the symbol yi, and the ratio of events occurring per unit of time is denoted by the symbol λi. In other words, λi provide the distribution's average. The probability here changes as a function of λi. The Poisson probability distribution has a right-angled skew. However, when λi increase, the distribution gets closer to the normal distribution. The Poisson regression model's equal mean and variance is its most important feature. Because distortions are apparent in the assumption that the conditional expected value is equal to the variance and the assumption is not met, over- or under-dispersed data sets cannot be described by the Poisson distribution. In this case, updating the data set or starting the analysis with different methods may be a solution. The expected value and variance of yi are given in Equation (4).

$$\lambda_i = E(y_i|x_i) = Var(y_i|x_i) \qquad \dots (4)$$

The link function illustrating the relationship between the expected value and the independent variables must have the form specified in Equation (5) in order to ensure that the expected value of yi does not take negative values (See Cameron and Trivedi, 1998).

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots., \beta_m x_m \qquad \dots (5)$$

In this equation, λi is an exponential function of the arguments. λi is the same as given in Equation below:

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots., \beta_m x_m) = e^{x_i'\beta} \qquad \dots (6)$$

Where β_0, β_1,…,$_m$ represent the unknown parameters.

There are many methods to calculating β estimators in the Poisson regression analysis based on the distribution of the dependent variable yi. Maximum likelihood (MLE) method, artificial maximum likelihood (PMLE) method, and generalized linear models (GLM) are the most commonly applied and well-known of these techniques. The most used method for regression models is (Newton Raphson iteration) approach is typically employed in the likelihood method (MLE). The Poisson regression model's log likelihood function is as follows given an observation set:

$$L(\beta|(y,x)) = \sum_{i=1}^{n} P(y_i|\lambda_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \qquad \dots . (7)$$

When the logarithm of this function is taken, Equation below is obtained.

$$lnL(\beta) = \sum_{i=1}^{n} (y_i \ln(\lambda_i) - \lambda_i - lny_i!) \qquad \dots (8)$$

Accordingly, the Poisson MLE of ( ) value is calculated from the expression in Equation 9 (See Durmuş and Güneri, 2020).

$$\sum_{i=1}^{n} (y_i - \lambda_i) x_i = 0 \qquad \dots (9)$$

2.3 Generalized Poisson Regression model

the generalized Poisson distribution presented by Consul and Jain (1973), Famoye in (1993) derived the generalized Poisson regression (GPR) model. These distributions can handle under-distributed, over-distributed, and evenly-distributed count data. The Poisson regression model's equally scattered nature is its standout characteristic (See Famoye, 1993). However, in applications, the data sets typically have a variation that is above the average. therefore, they show overdispersion. The Poisson model's prediction of the number of zero values and the presence of unobserved heterogeneity together led to the over dispersion of data (See Kibar, 2008). The coefficient estimate is untouched by over dispersion in the model; however, the estimate is affected by the standard error. As a result, the model's dependability is diminished (See Al-Ghirbal and Al-Ghamdi, 2006). When there is over dispersion in the data set, the generalized Poisson distribution is as follows (Equation 10) (See Pamukçu et al., 2014);

$$P(y_i|\lambda_i, \alpha) = \frac{\lambda_i(\lambda_i + \alpha y_i)^{y_i - 1} e^{-\lambda_i - \alpha y_i}}{y_i!} \qquad y_i = 0,1,2, \dots \qquad \dots (10)$$

Where $\lambda_i > 0$ and $\max\left(-1, \frac{-\lambda_i}{4}\right) < \alpha < 1$. Also, the mean and variance of the generalized Poisson distribution are equations (11 and 12):

$$\mu_i = E(y_i) = \frac{\lambda_i}{1 - \alpha} \qquad \dots (11)$$

$$Var(y_i) = \frac{\lambda_i}{1 - \alpha^3} = \frac{\lambda_i}{1 - \alpha^2} E(y_i) = \emptyset E(y_i) \qquad \dots (12)$$

The term $\emptyset = \frac{\lambda_i}{1 - \alpha^2}$ in particular acts as a dispersion factor. It is obvious that the generalized Poisson distribution for $(\alpha = 0)$ is the general Poisson distribution with the parameter λi. Under dispersion occurs when $(\alpha < 0)$, whereas overdispersion occurs when $(\alpha > 0)$ (See Yang et al., 2009). The standard error will be below the estimate and the regression parameters will be interpreted incorrectly when there is over dispersion. With the use of a log-link function, like in Equation (13), the independent variables are combined in the regression model based on the GP distribution.

$$\frac{\lambda_i}{1-\alpha} = \mu_i = E(y_i|x_i) = e^{x_i'\beta} \qquad \dots(13)$$

2.4 Negative Binomial Regression Model

According to (Xia, et al., 2012), the Negative binomial (NB) regression model is the most well-known alternative to Poisson regression, handles overdispersion by explicitly modeling the associated events through a latent variable. They further stated that the NB expands the Poisson regression model by including the fact that the mean μ_i of Y_i is now controlled by a heterogeneity component ϵ_i that is independent of X_i in addition to X_i (See Adeti, 2016). Negative binomial distribution is another model for count data and is one of the most often distribution that used as an alternative to Poisson distribution (See Zuur et al., 2009). Assume that y is a random variable and its probability mass function (p.m.f) is as follows (See Hilbe, 2011).The density function for the Negative Binomial model is as follows if it is assumed that $\exp(\epsilon_i)$ has a Gamma distribution (θ, θ):

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\,\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y} \qquad y = 0,1,2,\dots \qquad \dots.(14)$$

Where $\theta = \frac{1}{\alpha}$ which is the dispersion parameter

$$E(Y_i = y|X_i) = \mu_i = E(\exp(X_i^T\beta + \epsilon_i)) = E(\exp(X_i^T\beta)\exp(\epsilon_i)) = \exp(X_i^T\beta)$$

And that $E(\exp(\epsilon_i)) = 1$ hence $E(\exp(X_i^T\beta + \epsilon_i)) = E(\exp(X_i^T\beta))$ thus, the expected value μ_i does not change even if we assume that a Negative Binomial distribution or a Poisson distribution. Since the dispersion parameter > 0, under the Negative Binomial distribution. $Var(Y_i = y|X_i) = \mu_i\left(1 + \frac{\mu_i}{\theta}\right) > \mu_i$, Implying that $Var(Y_i = y|X_i)/E(Y_i = y|X_i) = 1 + \frac{\mu_i}{\theta}$. This implies that, the variance of the NB is greater than its mean, hence addressing the issue of overdispersion.

The NB distribution's dispersion parameter, represented by the symbol θ, which indicates the degree of over dispersion. In the event that $\theta = 0$, the Negative Binomial regression model also transforms into a Poisson regression. In their investigations in a variety of domains, several researchers have thought about using both Poisson and Negative Binomial models (See, for example, Miaou, 1994; Kibria, 2006; and Chipeta, et al., 2014), which may be calculated from the probability density function as:

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\,\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y} = \frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma(r)} \qquad \dots(15)$$

The likelihood function for the negative binomial model is:

$$L = \prod_{i=1}^{N} P\,(Y_i = y|X_i) = \prod_{i=1}^{N}\left(\frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma(r)}\right) \qquad \dots(16)$$

And the log-likelihood is

$$\ell = \sum_{i=1}^{N} ln\left(\frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma(r)}\right) \qquad \dots(17)$$

$$= \sum_{i=1}^{N} ln\left(\Gamma(y+r)\right) + \sum_{i=1}^{N}\ln(p^r) + \sum_{i=1}^{N}\ln((1-p)^y) - \sum_{i=1}^{N}\ln\left(\Gamma(y+1)\right) - \sum_{i=1}^{NN}\ln(\Gamma(r)) \qquad \dots(18)$$

$$= \sum_{i=1}^{N} ln\left(\Gamma(y+r)\right) + Nrln(p) + yln(1-p) - \ln\left(\Gamma(y+1)\right) - Nln\left(\Gamma(r)\right) \qquad \dots(19)$$

To find the maximum likelihood estimates (MLE) of (r) and (p), we take derivatives of (ℓ) with respect to (r) and (p) and equate them to zero.

$$\frac{d\ell}{dp} = \frac{Nr}{p} + \sum_{i=1}^{N}\left(\frac{-y}{1-p}\right) = 0 \qquad \dots(20)$$

To get the maximum likelihood estimate of (p)

$$p = \frac{Nr}{Nr + \sum_{i=1}^{N} y} \qquad \dots(21)$$

Differentiating ℓ in relation to (r) and substituting the maximum likelihood estimate of (p) to eliminate (p) from the equation,

$$\frac{d\ell}{dr} = Nln(p) - N\psi(r) + \sum_{i=1}^{N}\psi(y+r) \qquad \dots(22)$$

$$\frac{d\ell}{dr} = Nln\left(\frac{Nr}{Nr + \sum_{i=1}^{N} y}\right) - N\psi(r) + \sum_{i=1}^{N}\psi(y+r) = 0 \qquad \dots(23)$$

The equation $\frac{d\ell}{dr}$ has no closed-form solution so the root has to be found with numerical methods.

3. Testing the Goodness of Fit of the Model (Model Selection)

The goodness of fit of the regression line modified to a data set in linear regression models refers to how well the regression line adapted with the data set. The distribution of the observations around the model's form should be assessed after the parameters have been estimated since the better the model fits the data, the closer the observations are to the projected model. In other words, it would be preferable to alter the explanatory variables to account for the change in (y_i) (Koutsoyiannis, 1989). the generally used criteria for testing the goodness of fit of any model include the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), the Pearson statistic χ^2, the Deviation statistic, and the pseudo R^2measurement (see Durmus and Guneri, 2020).

3.1 Pearson Statistics

One of the fundamental standards of goodness of fit is

Pearson's statistic, which is usually chosen to establish if the series is overspread. Equation following below provides Pearson statistics for a model with λi mean and ωi variance.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\widehat{\omega}_i} \quad \dots(24)$$

The value of this test is used to assess whether the dispersion of the series is over. It will be $\omega i = \lambda i$ as a natural extension of the Poisson distribution when Pearson statistic is used for Poisson regression, and the formula will take the form of Equation below:

$$\chi_p^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \quad \dots(25)$$

it indicates that the data are not suitable for the model and the presence of over dispersion status, If the ratio of calculated χ_p^2 to the degree of freedom is more than 1. The calculated χ_p^2 value will likewise be compared with the value $(n-k)$. If the series $\chi_p^2 > (n-k)$ is over dispersion, If the series $\chi_p^2 < (n_{\overline{4.}} k)$ is said to be under dispersion (See Deniz, 2005).

3.2 Deviation Statistics

deviance statistics is one of the techniques used to measure goodness of fit, also called 'G square statistic'. Deviation statistics are expressed by Equation 26.

$$G^2 = 2\sum_{i=1}^{n} y_i ln\left(\frac{y_i}{\lambda_i}\right) \quad \dots(26)$$

The convergence of this statistical value to zero indicates that the model fit has increased. If the statistical value is equal to (0), 'model fit is perfect (See Dumus and Guneri, 2020).

3.3 Akaike Information Criterion (AIC)

When comparing statistical models fitted by maximum likelihood (ML) to the same data, usually for non-nested models, to statistical models fitted by other methods, the Akaike Information Criterion (AIC) is used to measure the relative superiority of each model for the given data set. The statistic penalizes for the amount of predictors employed in the model and takes into account model parsimony, as a result:

$$AIC = -2L + 2k \quad \dots(27)$$

Clearly, the first term is a penalty for the number of parameters, while the second term is a deviation. The most effective statistical model is the one with the lowest AIC value.

The comparative superiority of statistical models for a given set of data is assessed using the Akaike information criterion (AIC). Hirotugu Akaike developed and published the AIC in 1973. As according Mazerolle (2004), the AIC gives an impartial technique for determining which of several competing models is the most frugal. AIC values for given data are meaningless, but when they are compared to the AIC values of competing models, they take on meaning. The model with the smallest value of AIC among competing models is the best (ideal) model for the given data set (Mazerolle, 2004). According to Mazerolle (2004), the AIC is used to choose a model that fits the data well but has a small number of parameters; as a result, the AIC penalizes the addition of parameters (See Adeti, 2016).

3.4 Bayesian Information Criterion (BIC)

Another estimator evaluating model fit for a given data among different types of non-nested model is the Bayesian information criterion (BIC), and its formula is as follows:

$$BIC = -2logL + klog\ n \quad \dots(28)$$

Where:

L: The model's maximum likelihood function.

k: Number of model parameters.

n: Number of observations (sample size).

The best model to fit the data is the one with the minimum value of BIC (See Cameron and Trivedi, 2013).

The Likelihood Ratio Test

A statistical technique called the likelihood ratio test (LR) is used to compare two "nested models." and determine which model fits the data better, its formula is given as (See Hilbe, 2011 and Zurr et al., 2009)

$$LR = -2log\left(\frac{L_1}{L_2}\right) \quad \dots(29)$$

L_1: The likelihood of the first model.

L_2: The likelihood of the second model.

5. Data Collection

The sample dataset of this study was limited on the Erbil International Airport and the observations were made of the number of EIA Pax aircraft Movement as a Y dependent (response) variable in the Airport transfer process in the specific period of time. The sample consisted of 336 weeks which have been collected during (7) years period; beginning from 1th January 2015 through to 31th December 2021 of all people that they had taken Erbil International Airport as a way to move to other countries. A set of variables were taken weekly, where (13) of them represents independent (explanatory) variables (Total EIA passenger, Male arrival, Male departures, Female arrival, Female departures, total Domestic Passengers, total Domestic EIA Movement, total international Passengers, total international EIA Movement, infant Passengers <2 (less than two years), adult Passengers > 2 (older than two years), business Class Passengers, economy Class Passengers} and the variable under study is dependent variable which is the number of EIA Pax aircraft Movement weekly over a period of seven years was measured.

6. Application, Results and Discussion

The number of occurrences is assumed to be independently identically distributed with a discrete probability distribution when evaluating count data variable. The Poisson and Negative Binomial distributions are the most typical probability distributions used to describe count data. The data

was also analyzed using generalized linear models using Poisson regression and Negative binomial regression model. Estimating log functions of a count variable is the objective of the Poisson regression model. The Poisson distribution (rather than the Normal distribution) is preferable since count variables are all positive integers and for rare events, as the Poisson mean > 0. Analysts typically look for alternatives to the Poisson model, such as the negative binomial model, because observed data will almost always display pronounced overdispersion. The equidispersion restriction of the Poisson model is relaxed through the use of a functional form called the negative binomial model.

Figure bellow shows number of flights distributed monthly for seven years ago in Erbil airport; different color lines plotted against time. Where the Y-axis represents EIA pax aircraft movement against the X- axis which is the time. Most of the lines of this study increased over time, some of trends somewhere have the similarity over time, this similarity indicates that many of the variables had the same behaviors or we can say semi similarity but generally have the same direction, this because of the similarity of correlation between independent variables look at the correlation matrix in index table (). There is only two lines have different behavior look at the (2017and 2020), this express that the air movement internationally demand for travelling did not show highly significant improvement as a new epicenter of COVID-19 emerged in several countries, leading to a re-imposition of travel restrictions in (2020). The struggle against ISIS, which began in 2017, has had a significant impact (affected) on Kurds' daily lives, particularly through its effects on the KRG economy, which leaded to stopped travelling on Erbil airport during period of time. as we saw in line of 2017.

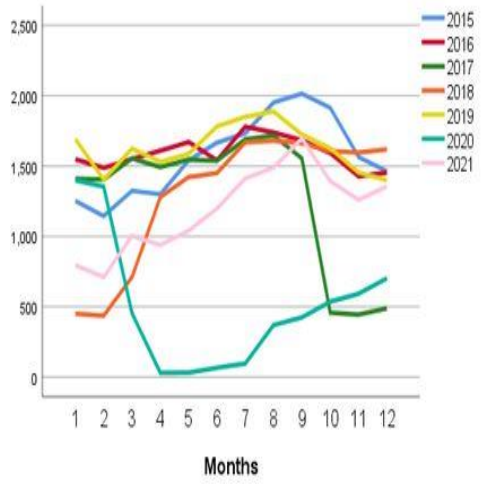

Figure 1: Graphical Representation of EIA pax aircraft movement monthly

Table 1: EIA pax aircraft movement Regionally

Y	Africa Movement	Europe Movement	Middle East Movement	Total Movement
2015	335	1953	16576	18864
2016	475	1619	16986	19080
2017	430	1402	13462	15294
2018	400	1720	13442	15562
2019	560	1988	17012	19560
2020	281	588	5184	6053
2021	472	1459	12359	14290
Total	2953	10729	95021	108703

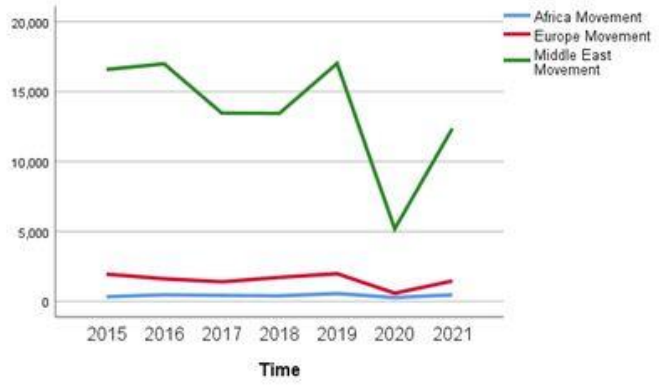

Figure 2: Graphical Representation of EIA pax aircraft movement Regionally

Figure (2) and table (1) show clearly the effect of COVID 19 on travelling regionally as we see in (2020) raghly decrease totally from the number of Air pax movement Compared to other years, because the Air movement internationally demand for travelling did not show highly significant improvement when this disease emerged in several countries.

Table 2: Passengers Demographic

	Africa	Europe	Middle East	Total
2015	21450	146961	1497290	1665701
2016	33508	143152	1637612	1814272
2017	32970	134776	1438785	1606531
2018	42102	152719	1339042	1533863
2019	54386	179956	1675443	1909785
2020	22742	46067	437437	506246
2021	31482	133653	1116597	1281732
Total	238640	937284	9142206	10318130

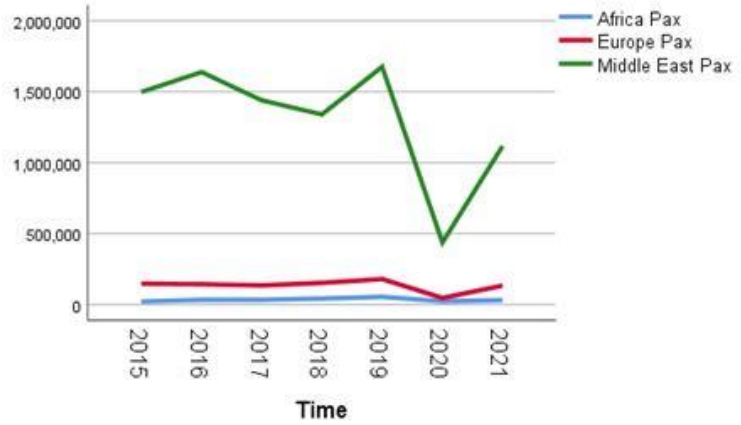

Figure 3: EIA Pax aircraft movement (Passengers Demographic)

Figure (3) and table (2) express the number of Passengers demographically. Covid 19 affected on the number of passengers demographically, there is a great convergence with the number of aircraft flights movement as we mentioned above especially the Middle East and Europe. While Africa did not have the problem of this disease (COVID 19). So, the year (2020) was more affected by this disease, which expanded in many countries, as we shown in Europe and the Middle East.

Table 3: Descriptive statistics of research variables

Variable	Definition	Mean	Minimum	Maximum	Variance	S.d
Y	EIA aircraft Movement	323.23	0	578	16045.880	126.672
X_1	Total EIA passenger	30702.41	30	55060	175048988.631	13230.608
X_2	Male arrival	11133.56	6341	21769	15794545.436	3974.235
X_3	Male departures	10076.13	2823	21858	17130554.311	4138.907
X_4	Female arrival	4931.72	1381	10595	3607088.475	1899.234
X_5	Female departures	4535.85	1324	10568	4044028.264	2010.977
X_6	Total Domestic Passengers	9107.47	0	19426	20572924.053	4535.739
X_7	Total Domestic Movement	101.79	0	216	1966.513	44.345
X_8	Total international Passengers	21618.85	0	48453	128134548.394	11319.653
X_9	Total international Movement	221.79	0	396	10416.032	102.059
X_{10}	Infant Passengers	505.12	0	1856	87395.687	295.628
X_{11}	Adult Passengers	30187.26	25135	58335	18585761.528	13632.856
X_{12}	Business Class Passengers	1129.45	0	2218	195702.702	442.383
X_{13}	Economy Class Passengers	29048.78	25133	55033	15772609.406	12558.965

If the variance of dependent variable is higher than the mean (which is typically the case), this is called overdispersion and requires an additional dispersion parameter. Table (3) gives a brief summary measure of the dataset; each variable has 336 valid observations (from 2015 to 2021). The counts range of Y from a minimum value of (0) to a maximum value of (578) with mean (323.23). The results show that the variance of all variables is greater than the mean which means that there is over dispersion in the dataset. According to the standard errors of the estimates in the dataset, it was small and this shows that there is a high precision. The mean, Minimum, Maximum, Variance and stander deviation of the total EIA passenger was (30702.41, 30, 55060, 175048988.631 and 13230.608) by respectively. While, the mean of the (Male arrival, Male departures, Female arrival and Female departures) were (11133.56, 10076.13, 4931.72 and 4535.85) by respectively. The mean, Minimum, Maximum, Variance and stander deviation of the total Domestic Passengers were (9107.47, 0 , 19426, 20572924.053 and 4535.739) by respectively. The mean of the (Total Domestic Movement, Total international Passengers, Total international Movement, Infant Passengers, Adult Passengers, Business Class Passengers and Economy Class Passengers) were (101.79, 21618.85, 221.79, 505.12, 30187.26, 1129.45 and 29048.78) by respectively.

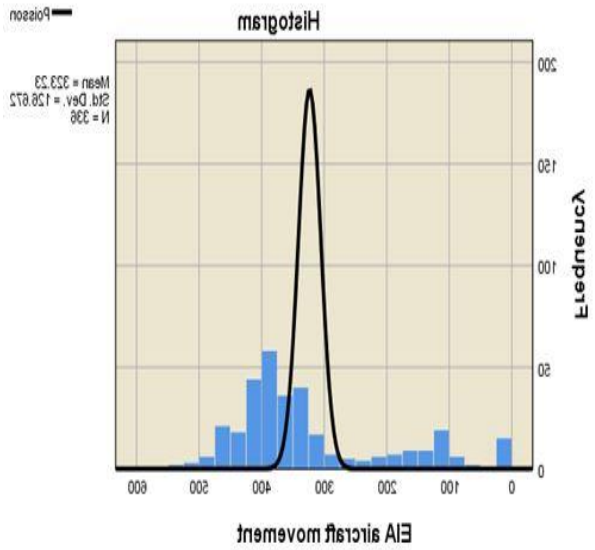

Figure 4: the Number of EIA pax aircraft movement during (2015-2021)

We got the output of response count variable that, this histogram does not distribute normally. Looks The figure (4) shows that data from (2015–2021) were used, and the dependent variable's Poisson distribution did not adequately cover the data due to the dataset's excessive dispersion and high variance. Therefore, the Poisson regression assumption that the number of cases has variance equal to the mean at each level of the covariates is neither violated or absent. Clearly shows underestimate the probability at 0 and overestimate the probability over 300. If we check it again, is it approximately negative binomial distribution? the answer is not, because the data set different from negative binomial and we know that a negative binomial distribution is not more appropriate for that sample dataset.

6.1 Interpretation of the results of Poisson regression model

This section's table (4) lists the regression coefficients for the Poisson regression model along with their standard error values and test statistics for the coefficients of related parameters (4). As can be observed, several parameters' maximum likelihood coefficients are statistically significant at 5%; statistically significant coefficients indicate that the relevant variable has a positive or negative impact on the number of EIA pax aircraft movements. we illustrate the output of Poisson regression model in table below accordingly of the software used, as the data were found to be over dispersed, the analysis of Poisson regression model will not be the best results. When the coefficients are examined by the Wald Chi square test, it is seen that (X_1 , X_2 , X_7 , X_8 , X_9 , X_{10} , X_{12}) of them are significant because their (p-values) are less than 5% level, with intercept too, while (X_3 , X_4 , X_5 , X_6 , X_{11} , X_{13}) of estimated coefficient are not significant because their p-values are greater than the 5% level.

Table 4: Results of Parameter Estimates (Fitting) Generalized Poisson Regression Model

Iteration 0: log likelihood	-2792.9198	Number of obs.	336	
Iteration 1: log likelihood	-2687.6075	Optimization	MLE	
Iteration 2: log likelihood	-2687.2785	Pseudo R2	0.7810	
Iteration 3: log likelihood	-2687.2785	Link function	Log	
Explanatory variables	Coefficient B	Std. Error	Hypothesis Test Wald Chi-Square	Sig,
Constant	4.244312	.0444109	95.57	0.000
Total EIA passenger - X_1	.0000134	3.48e-06	3.85	0.000
Male arrival - X_2	8.95e-06	3.40e-06	2.63	0.009
Male departures - X_3	-1.31e-06	3.97e-06	-0.33	0.742
Female arrival - X_4	7.86e-06	8.95e-06	0.88	0.380
Female departures - X_5	-.0000167	9.26e-06	-1.81	0.071
Total Domestic Passengers - X_6	-3.46e-06	4.76e-06	-0.73	0.467
Total Domestic Movement - X_7	.0023479	.0003975	5.91	0.000
Total international Passengers - X_8	-.0000205	3.89e-06	-5.28	0.000
Total international Movement - X_9	.0042403	.0002823	15.02	0.000
Infant Passengers - X_{10}	-.0001986	.0000579	-3.43	0.001
Adult Passengers - X_{11}	1.73e-06	1.88e-07	0.92	0.355
Business Class Passengers - X_{12}	.0001855	.0000467	3.97	0.000
Economy Class Passengers - X_{13}	3.31e-06	3.91e-06	0.85	0.397

Count data usually expressed the variance greater than mean. Hence, we shall check the value of variance and mean after fitting Poisson regression Count data 'model. Therefore, the over-dispersion must be check it as shown below:

$H_0: Mean = Variance \quad H_1: Mean < Variance$

Table 5: Omnibus test of Model Effects for Generalized Poisson Regression Model

Likelihood Ratio Chi-Square	DF	Sig.
19172.174	13	0.000

The Omnibus Test from table (5) shows the significance of

your overall model and is reported with Chi Square (χ^2), Under the hypothesis that the current model is acceptable for the combinations of independent variables, all measures have approximately Chi-Square distributions. It reflects the likelihood ratio test result for a model's overall fit equal to (19172.174) and ($p - value = < 0.05$), it means significant. The significance of χ^2statistics implies the existence of over-dispersion. Therefore, in the next section, we apply Negative Binomial model to handle the issue of over-dispersion.

6.2 Interpretation of the results of Negative Binomial Regression Mode

The estimated parameters of Negative Binomial regression model are given in table (6). According of the results of the Wald Chi square test, it shows that (X_1 , X_2 , X_5 , X_7 , X_8 , X_9 , X_{10} , X_{12}) estimated coefficient are significant because their (p-values) are less than 5% level, with intercept too, while (X_3 , X_4 , X_6 , X_{11} , X_{13}) of estimated coefficient are not significant because their p-values are greater than the 5% level.

Table 6: The Results of Parameter Estimates (Fitting) of Negative Binomial Regression Model

Iteration 0: log likelihood	-1999.9838	Number of obs.	336
Iteration 2: log likelihood	-1945.1496	Optimization	MLE
Iteration 3: log likelihood	-1944.4893	Pseudo R2	0.1189
Iteration 4: log likelihood	-1944.4893		

Explanatory variables	Coefficient B	Std. Error	Hypothesis Test Wald Chi-Square	Sig,
Constant	3.837038	.0523234	73.33	0.000
Total EIA passenger - X_1	.0000163	5.83e-06	2.80	0.005
Male arrival - X_2	.0000179	5.57e-06	3.21	0.001
Male departures - X_3	-7.87e-06	6.56e-06	-1.20	0.230
Female arrival - X_4	8.31e-06	.0000142	0.58	0.559
Female departures - X_5	-.0000345	.0000148	-2.33	0.020
Total Domestic Passengers - X_6	1.82e-06	7.79e-06	0.23	0.815
Total Domestic Movement - X_7	.0033766	.0006435	5.25	0.000
Total international Passengers - X_8	-.0000303	6.35e-06	-4.77	0.000
Total international Movement - X_9	.0053215	.0004385	12.14	0.000
Infant Passengers - X_{10}	-.000506	.0000932	-5.43	0.000
Adult Passengers - X_{11}	1.82e-06	3.06e-06	0.59	0.553
Business Class Passengers - X_{12}	.0004378	.0000677	6.47	0.000
Economy Class Passengers - X_{13}	4.20e-06	6.50e-06	0.65	0.518

in table (7) demonstrates that the model fits the variables properly because the general significance is (0.000), which is within the 95% confidence interval. The highly high Chi-Square value of (1485.58) for the Likelihood Ratio indicates that the model has properly explained the variables it contains. From the aforementioned, it can be inferred that the model was able to show a substantial link between the dependent and independent variables. usually, count data expressed that the variance greater than mean. Here also, we will check the value of variance and mean after fitting Count data of negative binomial regression 'model. the over-dispersion must be check by hypothesis as shown below:

$H_0: Mean = Variance \quad H_1: Mean < Variance$

Table 7: Omnibus Test of Model Effects for Negative Binomial Regression Model

Likelihood Ratio Chi-Square	DF	Sig.
1485.58	13	0.000

Again, the Omnibus test from table (7) shows the significance of the model and is reported with Chi Square (χ^2). Under the hypothesis that the current model is acceptable for the combinations of independent variables, all measures have roughly Chi-Square distributions. It reflects the likelihood ratio test result for a model's overall fit equal to (1485.58) and ($p - value = 0.000 < 0.05$), it means significant. The significance of χ^2statistic implies the existence of over-dispersion.

7. Model Selection

In using Poisson regression model, equidispersion makes the assumption that the variance's mean value must be met. It appears that over dispersion is the case because this assumption is rarely true.

For detecting the over dispersion, it can be seen from the value of Null Deviance / DF=322 or Pearson χ^2/ DF=322. If the value of Null Deviance / DF, which equal to (2890.4281/322=8.9765) or Pearson χ^2/ DF, which equal to (2392.7301/322=7.4308) is greater than 1, when it is greater than 1, there is over dispersion; when it is less than 1, there is under dispersion. Negative binomial regression can be used as an additional method to handle over dispersion on the Poisson regression model in addition to the Generalized Poisson regression model.

The ideal (best) model is the one with the lowest AIC and BIC values. To assesses the fit of the two models In Table (8), with a significance level of 15%, the likelihood of a negative binomial regression model is shown, along with the smallest AIC and BIC values for each combination of variables ranging from eight combinations of predictor variables. The Poisson regression model of the dataset had the largest values of all criterions and indicating a poor fit to the data. Using the Negative Binomial regression as alternative of Poisson regression model, it had the smaller values of that criterions when we compared with Poisson regression model and indicating better goodness of fit with the data and is considered as the best models. where the Deviance statistic is distributed approximated a Chi-square distribution. The null Deviance equal (2890.4281) as Chi square distributed with the model degree of freedom (1) but the Residual Deviance equale (8.9764) as Chi-square distributed with the model degrees of freedom (322), also it shows the AIC value of the model equal to (16.0790) and the value of BIC equal to (1017.318). hence a model with 14 estimates parameters of Negative Binomial regression model with the smallest value of (AIC=11.6576) and (BIC=-1406.08) will be the optimal.

Table 8: Goodness of Fit for both regression model

Assessment parameter	Generalized Poisson regression model	Negative binomial model
Pearson χ^2 statistic	2392.7301	299.9627
Residual Pearson χ^2statistic	7.4308	0.9316
Null Deviance	2890.4281	267.0299
Null Deviance degree of freedom	335	335
Residual Deviance	8.9764	1.4504
Residual degree of freedom	322	322
AIC	16.0790	11.6576
BIC	1017.318	-1406.08

8. Conclusions

During analyzing the fitting of Generalized Poisson and Negative Binomial Regression models for output of count time series event as indicated from the practical part, the following conclusions have been drawn:

1. We can consider that the direction of the number of EIA pax aircraft movement of most of the past years will increase in the middle of every year and continue to increase and then decrease at the end of every year except the (2017 and 2020) there is a difference.

2. The behaviors of two different line (2020 and 2017) year, express that the air movement internationally demand for travelling did not show highly significant improvement as a new epicenter of (COVID-19) emerged in several countries, leading to a re-imposition of travel restrictions in (2020). In 2017 the struggle against ISIS has had a significant impact on Kurds' daily lives, particularly on the KRG economy, which leaded to stopped travelling on Erbil airport during period of time.

3. Covid 19 affected on the number of passengers demographically, there is a great convergence with the number of aircraft flights movement, especially the Middle East and Europe. While Africa did not have the problem of this disease (COVID 19).

4. the Standard error value for the dependent variable is a sign that the dependent variable may be accurately predicted. The models' high chi-square values further demonstrated how well they explained the variables they contained. The Generalized Poisson and Negative Binomial Regression Model's Omnibus Test of Model effects revealed that the models fit the data very well for the chosen variables even though their p-values are less than (5%) and they are over dispersed.

5. it shown that the same coefficients are significant for both models only the variable (Female departures-X_5) in the Generalized Poisson regression model is not significant, but enhanced statistically important variable in the Negative Binomial regression model by the value of the p-value of the Wald Chi Square test.

6. the condition of (over dispersion) in the Generalized Poisson regression cannot gives adequate results. Therefore, applied Negative Binomial regression to the same data. Find out that Negative Binomial is more capable than Generalized Poisson Regression model by using (Person χ^2, Deviance, AIC and BIC), it found to be that the Negative Binomial regression model had more performance and a superior model than Generalized Poisson regression model due to less than their values of that criterions.

References:

Adeti, F. (2016): "Modelling Count Outcomes from Dental Caries in Adults: A Comparison of Competing Statistical Models", A thesis Submitted to the department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi.

Agresti, A. (1996): "An Introduction to Categorical Data Analysis", John Wiley and Sons, New York.

Brännäs, K. and Johansson, P. (1994): "Time series count data regression. Communications in Statistics theory and Methods", 23(10), pp.2907-2925.

Al-Ghirbal A.S. and Al-Ghamdi A.S. (2006): "Predecting Severe Accidents Rates at Roundabouts Using Poisson Distribution", TRB Annual Meeting, 06-1684.

Cameron, A. C.; Trivedi, P. K. (1998): " Regression analysis of count data", Cambridge University Press. ISBN 978-0-521-63201-0.

Chipeta, M. G., Ngwira, B. M., Simoonga, C., & Kazembe, L. N. (2014): "Zero adjusted models with applications to analysing helminths count data", BMC research notes, 7(1), 856.

Deniz Ö. (2005): "Poisson Regresyon Analizi", İstanbul Ticaret Üniversitesi, Fen Bilimleri Dergisi, 4(7): 59-72.

Durrant, G. (2016): "Poisson Regression Modes for Count Data", Available from: https://www.slideshare.net/synchrony/poisson-regression-models-for-count-data-63688148 [Accessed on 14 July 2017].

Durmuş, B. and Güneri, Ö. İ. (2020): "An Application of the Generalized Poisson Model for Over Dispersion Data on The Number of Strikes Between 1984 and 2017", alphanumeric journal, the Journal of Operations Research, Statistics, Econometrics and Management Information Systems Volume 8, Issue 2.

Famoye, F. (1993): "Restricted Generalized Poisson Regression Model", Communications in Statistics Theory and Methods, 22(5), 1335-1354.

Hilbe, J. M. (2011): "Negative binomial regression". Cambridge University Press.

Hoffman, J. (2004): "Generalized Linear Models, Boston, Pearson Education Inc.

Jiang, Yuan (2018): "Analysis of Consumption Behaviors and Market Structure with Excess Zeros and Over-Dispersion", A dissertation to the Graduate School of the University of Florida in Partial fulfillment of the Requirements for the Degree of Doctor of Philosophy, University of Florida.

Karazsia, B. and Van Dulmen M. (2008): "Regression Models for Count Data: illustrations using longitudinal predictors of childhood injury", 33(10):1076-84 Available from: https://www.ncbi.nlm.nih.gov/pubmed/18522994. [Accessed on 14 July 2017].

King, G. (1988): "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model", American Journal of Political Science, 3(3), 838-863.

Kibar, F.T. (2008): "Trafik Kazaları ve Trabzon Bölünmüş Sahil Yolu Örneğinde Kaza Tahmin Modelinin Oluşturulması", Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

Kibria, B. G. (2006)" "Applications of some discrete regression models for count data:, Pakistan Journal of Statistics and Operation Research, 2(1), 1-16.

McCullagh, P. & Nelder, J. A. (1989): "Generalized Linear Models", 2nd Edition, Published by Chapman and Hall/CRC 532 Pages, New York, ISBN 9780412317606.

Miaou, S. P. (1994): "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions", Accident Analysis & Prevention, 26(4), 471-482.

Moksony, F. and Hegedus, R. (2014): "The Use of Poisson Regression in the Sociological Study of Suicide", Corvinus Journal of Sociology and Social Policy, 5(2), 97-114.

Olsson, U. (2002): "Generalized Linear Models: An Applied Approach" Student literature, ISBN-13: 978-9144041551.

Pamukcu, E., Colak, C. and Halisdemir, N. (2014): "Modeling of The Number of Divorce in Turkey Using the Generalized Poisson, Quasi-Poisson and Negative Binomial Regression", Turkish Journal of Science & Technology, 9(1), 89-96.

Pesonen, T. (2018): "Predicting Real Estate Sales Volume in Finland", Master of Science in Economics and Business Administration, Information and Service Management, Aalto University, P.O. BOX 11000, 00076 AALTO.

Yang, Z., Hardin, J.W. and Addy, C.L. (2009): "A Score test for Overdispersion in Poisson Regression Based on The Generalized Poisson-2 Model", Journal of Statistical Planning and Inference. 139, 1514-1521.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. and Smith, G. M., (2009): "Mixed effects models and extensions in ecology with R", Springer Science & Business Media.