Journal of New Theory
ISSN: 2149-1402

35 (2021) 48-61
Journal of New Theory
https://dergipark.org.tr/en/pub/jnt
Open Access

Models for Overdispersion Count Data with Generalized Distribution: An Application to Parasites Intensity

Öznur İşçi Güneri[1], Burcu Durmuş[2]

Article History
Received: 23 Mar 2021
Accepted: 04 Jun 2021
Published: 30 Jun 2021
Research Article

Abstract – The Poisson regression model is widely used for count data. This model assumes equidispersion. In practice, equidispersion is seldom reflected in data. However, in real-life data, the variance usually exceeds the mean. This situation is known as overdispersion. Negative binomial distribution and other Poisson mix models are often used to model overdispersion count data. Another extension of the negative binomial distribution in another model for count data is the univariate generalized Waring. In addition, the model developed by Famoye can be used in the analysis of count data. When the count data contains a large number of zeros, it is necessary to use zero-inflated models. In this study, different generalized regression models are emphasized for the analysis of excessive zeros count data. For this purpose, a real data set was analysed with the generalized Poisson model, generalized negative binomial model, generalized negative binomial Famoye, generalized Waring model, and the foregoing zero-inflated models. Log-likelihood, Akaike information criterion, Bayes information criterion, Vuong statistics were used for model comparisons.

Keywords – *Count data, overdispersion, generalized distribution models, zero-inflated*

Mathematics Subject Classification (2020) – 62P10, 62J05

1. Introduction

In regression analysis, the relationship between a dependent variable and one or more independent variables is examined. When the dependent variable consists of count data, count regression models are used instead of classical regression analysis. Count data can be expressed as observations consisting of nonnegative integers that can take the value of zero or a value greater than zero and show a discrete distribution [1]. Count data are generally right skewed and do not show normal distribution [2]. Different count regression models are used according to the mean and variance in modelling this type of data. Poisson regression analysis is based on the assumption of equidispersion, in other words, equality of mean and variance(mean=variance) in cases where the dependent variable is count data. In practice, however, even the distribution of data is rare. In practice, however, the variance usually exceeds the mean. This occurrence of non-Poisson variation is known as overdispersion (mean>variance) [3]. In cases where the variance is smaller than the mean, the data is considered underdispersion (mean<variance). Modelling overdispersion or underdispersion count data with inappropriate models can lead to overestimated standard errors and misleading inferences [4]. In this case, regression models containing the dispersion parameter should be used in the dataset instead of Poisson

[1]oznur.isci@mu.edu.tr (Corresponding Author); [2]burcudurmus@mu.edu.tr
[1]Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla, Turkey
[2]Rectorate Performance Analysis Unit, Muğla Sıtkı Koçman University, Muğla, Turkey

regression. Two approaches explain the overdispersion that occurs in Poisson regression. One of these approaches is the semi-likelihood approach, and the other approach is the mixed Poisson model approach [5]. Besides several models for modelling overdispersion count data, such as negative binomial distributions, quasi-Poisson, and other Poisson mixes, there are several models for underdispersion count data [6,7]. Harris, Yang, and Hardin (2012) proposed a generalized Poisson (GP) regression model for underdispersion count data [8].

Count data can be expressed as observations made up of nonnegative integers and showing a discrete distribution. For this reason, count data are generally right-skewed and do not show normal distribution. Poisson regression or alternative models are used in modelling this type of data, depending on the state of mean and variance. Count data can be analysed using regression models based on the Poisson distribution in the case of equidispersion. However, other discrete regression models, such as the generalized negative binomial distribution (GNB), can be used in case of overdispersion [9,10]. Also, a model was investigated by Famoye (1995) to show its use for analysing grouped binomial data [11]. In case of overdispersion, GNB is recommended besides the negative binomial distribution. GP has been found to be useful in fitting overdispersion and under dispersion count data to a model [12]. GNB is a simplification based on the generalized negative binomial distribution.

The generalized Waring distribution is an extension of the negative binomial distribution. This distribution is known as the beta negative binomial distribution. Waring distribution was first proposed and used by Irwin (1968) to model accident number data [13]. One advantage of this model over the negative binomial model is that researchers can distinguish unobserved heterogeneity from internal factors of each individual's characteristics and covariates that can affect the variability of the data. These models obtain parameter estimates by including the effect from overdispersion into the model.

Generalized Famoye (GNB-F) and generalized Waring (GNB-W) models have different applications in the literature. Various applications of GNB-F have been demonstrated in physics, ecology, medicine, etc. [14-16]. Another issue to be considered in the analysis of count data is the zeros' density in the dependent variable. Count data have zero values by nature, and the classical ordinary least squares (OLS) method does not give good estimates because it does not show a normal distribution.

The presence of more than expected zero values in the data set is defined as zero inflation [17,18]. It is more appropriate to analyse such data sets with zero-inflated models that take into account zeros [19]. Zero-inflated models are used in different fields such as econometrics, demography, medicine, public health, biology, agriculture, etc. Failure to use appropriate methods to analyse zero-inflated data may result in biased parameter estimates, smaller standard errors, and inconsistent results [20]. Zero-inflated count data may lack equality of mean and variance. In such a case, overdispersion or underdispersion must be taken into account.

The zero-inflated generalized Poisson (ZIGP) model is an extension of the generalized Poisson distribution [21]. Other widely used methods are the zero-inflated negative binomial (ZINB) model and Hurdle models in case of excess zeros in the data [22,23]. There are two types of zeros in the zero-inflated model: "real zeros" and "excess zeros". There are situations where a zero-inflation model makes sense in terms of theory or common sense. Altun (2018) proposed Poisson-Lindley distribution for overdispersion data. The Poisson-Lindley distribution arises when the parameter of the Poisson distribution has the Lindley distribution [24]. Unlike the Poisson distribution, the Poisson-Lindley distribution allows for overdispersion. Therefore, this model is a good option for modelling datasets that are overdispersion and zero-inflated.

In this study, some generalized models used for count data with overdispersion are discussed. These models are generalized Poisson (GP), generalized negative binomial Famoye (GNB-F), generalized negative binomial (GNB), generalized negative binomial Waring (GNB-W), zero-inflated negative binomial (ZINB), zero-inflated negative binomial Waring (ZINB-W) and zero-inflated negative binomial Famoye (ZINB-F) regression models.

2. The Generalized Models

2.1. Generalized Poisson Regression Model (GP)

The most widely used regression model for count data sets is the Poisson regression model with the log-link function. The most prominent feature of the Poisson model is its equidispersion. Still, in implementations, data sets often have a variance that exceeds the mean. When there is overdispersion in the data set, the generalized Poisson distribution is as follows [25];

$$f(y_i, \theta_i, k) = \frac{\theta_i(\theta_i + ky_i)^{y_i-1} e^{-\theta_i - ky_i}}{y_i!} \quad y_i = 0, 1, 2, \dots . \tag{1}$$

here $\theta_i > 0$ and $\max\left(-1, \frac{-\theta_i}{4}\right) < k < 1$. Also, the expected value and variance of the generalized Poisson distribution can be written as:

$$\mu_i = E(Y_i) = \frac{\theta_i}{1-k}$$
$$var(Y_i) = \frac{\theta_i}{(1-k)^3} = \frac{\theta_i}{(1-k)^2} \tag{2}$$
$$E(Y_i) = \emptyset E(Y_i)$$

In particular, the term $\emptyset = 1(1-k)^2$ plays the role of a dispersed factor. It is clear that the generalized Poisson distribution for $k = 0$ is the general Poisson distribution with the parameter of θ_i. When $k < 0$, under dispersion, occurs, while when $k > 0$, overdispersion occurs [26]. The presence of overdispersion will cause the standard error to be below estimate and misinterpretation of the regression parameters. As a result, a number of estimation methods have been proposed to model data in the occurrence of overdispersion. These models include the quasi-Poisson or quasi-binomial regression model and the negative binomial distribution. Parameter estimates of these models are similar to the simple Poisson approach, but confidence intervals are larger [27]. As a result, the models will give different results in terms of the significance of the coefficients.

2.2. Generalized Negative Binomial: Famoye (GNB-F)

The GNB-F model assumes that the value of θ is an unknown scalar parameter. So, the probability mass function of the distribution, mean, and variance are given as:

$$P(Y = y) = \frac{\theta}{\theta + \emptyset y}\binom{\theta + \emptyset y}{y}\mu^y(1-\mu)^{\theta - y - \emptyset y}$$
$$0 < \mu < 1, \qquad 1 \le \emptyset < \mu^{-1} \quad , \qquad \theta > 0 \text{ and } y_i \in (0,1,2, \dots) \tag{3}$$
$$E(Y) = \theta_i \mu (1 - \emptyset\mu)^{-1}$$
$$var(Y_i) = \theta_i \mu (1 - \emptyset\mu)^{-1}(1 - \emptyset\mu)^{-3}$$

Its main difference from the negative binomial model is that the θ parameter is unknown in Equation 2, but a known parameter in Equation 3 $\sigma = \emptyset > 1$. As the ∅ value approaches 1, the variance approaches the negative binomial. Thus, the parameter is generalized to have greater variance than is allowed in the GNB-F model. To compare the results of the Poisson and negative binomial distribution, the log link is as follows:

$$log(\mu) = x\beta \tag{4}$$

2.3. Generalized Negative Binomial Regression Model (GNB)

The negative binomial distribution is the first distribution to consider when the variance is greater than the mean. Negative binomial regression is used as an alternative to Poisson regression because these two methods fit the model by using the same connection log link function [28]. NB model is often used to model overdispersion count result variables. The assumption that the Poisson parameter changes proportionally to the chi-square leads to the negative binomial distribution.

The GNB model is based on the simplification of the generalized negative binomial distribution. If $\sigma = \emptyset$ and $\mu = \pi/(1 + \emptyset\pi)$ expressions are substituted for σ and μ expressions given in Equation 3, the parameter becomes a vector of observation-specific known constants. When the θ parameter is known, while $\emptyset > 1$, the σ parameter is not negative in the generalized negative binomial distribution. Thus, under these conditions, the probability mass function, mean, and variance are given by:

$$P(Y = y) = \frac{n}{n + \sigma y}\binom{n + \sigma y}{y}\left(\frac{\pi}{1 + \sigma\pi}\right)^{y}\left(1 - \frac{\pi}{1 + \sigma\pi}\right)^{n - y - \sigma y}$$

$$E(Y) = n\frac{\pi}{1 + \sigma\pi}\left(1 - \frac{\pi}{1 + \sigma\pi}\right)\sigma^{-1} \tag{5}$$

$$var(Y) = n\frac{\pi}{1 + \sigma\pi}\left(1 - \frac{\pi}{1 + \sigma\pi}\sigma\right)(1 + \sigma\pi)^{-3}$$

$$= n\pi(1 + \sigma\pi)(1 + \sigma\pi - \pi)$$

Therefore, the variance is equal to binomial variance, $= 0$. It is equal to negative binomial variance $if\ \sigma = 1$. Here, if $\sigma > 0$, GNB generalizes the binomial distribution in the regression model.

2.3.1. Generalized Waring Regression Model (GNB-W)

The negative binomial distribution is a limiting case of the generalized Waring distribution. This distribution provides a model for the distribution of accidents. Here the variance is divided into three components. The first of these is the usual random component in classical accident theory; the other two can often be described as separate variances due to "liability" and "proneness". The sum of the last two components is the only component defined by the variation in sensitivity in classical theory [13]. The generalized Waring distribution (the number of crashes A) depends on three parameters: $\pi, a,$ and k. However, the "three-component distribution" is not necessarily a generalized Waring distribution. The generalized Waring distribution must satisfy the following conditions:

i. $Y\ /\ x, \lambda_x, v \sim$ Poisson (μ_x)
ii. $\lambda_x, /v \sim$ Gamma (a_x, v)
iii. $v \sim$ Beta (ρ, k)

In Irwin's study on accident data, λ/v is specified as "accident liability" and v as "accident proneness". Thus, the mass density function is given by:

$$P(Y = y) = \frac{\Gamma(a_x + \rho)\Gamma(k + \rho)}{\Gamma(\rho)\Gamma(a_{x+k+\rho)}}\frac{(a_x)_y(k)_y}{(a_x + k + \rho)_y}\frac{1}{y!} \tag{6}$$

where $a_x\ , k\ , \rho > 0$; $a_x = \mu(\rho - 1)/k$ and $(a)_w$ is the Pochhammer notation $\Gamma(a + w)/\Gamma(w)$, if a>0.

$$E(Y) = \mu = \frac{a_x k}{\rho - 1} \tag{7}$$

$$var(Y) = \mu + \mu\left(\frac{k + 1}{\rho - 2}\right) + \mu^2\left\{\frac{k + \rho - 1}{k(\rho - 2)}\right\}$$

2.3.2. Zero-Inflation Model

Count models can also take zero value due to their nature. However, having more than the expected number of zero values in the data set is defined as zero inflation. In the datasets where most of the observations are zero, excluding the zero values from the analysis leads to incorrect results. Zero-inflated count data may lack equality of mean and variance. Therefore, when there are too many zeros, it may not be appropriate to use Poisson and other models. It is more appropriate to use zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson Hurdle (PH) or negative binomial Hurdle (NBH) regression methods in modelling dependent variables with more than the expected number of zero values [29].

Hardin and Hilbe (2012) describe two origins of the zero result [15]: Those who do not enter the counting process and those who enter the counting process and have a zero result. Therefore, the model should be divided into different parts, one being zero count $y = 0$ and the other being nonzero $y > 0$. The zero-inflated model can be given as:

$$P(Y = y) = \begin{cases} p + (1-p)f(y), & y = 0 \\ (1-p)f(y), & y > 0 \end{cases} \tag{8}$$

In the above equation, p is the probability that the binary process will result in zero results. Here $0 \leq p < 1$ and $f(y)$ is the probability function. Famoye and Singh (2006) proposed the zero-inflated generalized Poisson (ZIGP) model, an extension of the generalized Poisson distribution. In another widely used method, the negative binomial model may be preferred where the Poisson mean has a gamma distribution [21]. A natural extension of the negative binomial model, the zero-inflated negative binomial (ZINB) model, is used in case of excess zeros in the data [23].

For the Waring distribution and Famoye's proposed models, it is more appropriate to use zero-inflated versions if there are too many zeros in the data. In this context, ZINB-W and ZINB-F distributions have been proposed for models based on zero.

3. Model Selection

The fact that all p values in the model selection are less than 0.05 means that all explanatory variables are suitable for the model. However, the fact that all the explanatory variables are significant does not mean that the regression model applied will be suitable for the data. Various tests are used to determine which model is more suitable for count data. In this study, Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), log-likelihood (LL) value and Vuong statistics were used. The interpretation that a model is good can be made when the AIC and BIC value is the smallest, or the LL value is the largest. Vuong test is one of the tests used to compare non-nested models. Apart from nested model comparisons, possible binary models can also be compared with the Vuong test. It is a widely used test, especially in zero-inflated model comparisons. In this way, it can be determined which models are suitable for models with excessive zeros.

3.1. Log-Likelihood (LL)

The advantage of using the maximum likelihood method (ML) is that the log-likelihood (LL) test can be used for model comparisons. The LL test can be used to test for the presence of overdispersion. To test the Poisson model against the GP model, where α is the overdispersion parameter, the hypothesis is expressed as $H_0: \alpha = 0$ and $H_0: \alpha \neq 0$. Probability ratio statistics is calculated as;

$$LL = 2(lnL_1 - lnL_0) \tag{9}$$

Where L_1 and L_0 are the log-likelihood under the respective hypothesis. LL has an asymptotic chi-square distribution with one degree of freedom [30]. When choosing the model over the LL value, the model with the largest log-likelihood value is determined as the appropriate model.

3.2. Akaike Information Criterion (AIC)

This criterion, which is widely used to compare different models, can be expressed as follows [31];

$$AIC = 2k - 2log(L) \tag{10}$$

In this equation, L represents the maximum value of the log-likelihood function, and k represents the number of explanatory variables. Among the existing models, the model with the lowest AIC value is selected as the appropriate model. In cases where the number of parameters is larger than the sample size, the AICc proposed by Hurvich and Tsai should be used instead of AIC [32]. This value can be written as follows [31-33];

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} = \frac{2kn}{n-k-1} - 2\ln(L) \tag{11}$$

3.3. Bayes Information Criteria (BIC)

Akaike derived the BIC (Bayesian Information Criterion) model selection criteria for selected model problems in linear regression [34]. The equation regarding the Bayesian measure of knowledge is as follows:

$$BIC = -2log(\mathcal{L}) + klog(n) \tag{12}$$

As in the Akaike information criterion, the model with the smallest BIC value among the available models is selected as the appropriate model.

3.4. Vuong Test

The Vuong statistic is used to compare non-nested models such as ZIP, NB and ZINB. This test is a statistic used when there is no missing observation in the data set [35,36]. Equations used for the Vuong test are given in Equation 13 and Equation 14.

$$Vuong = \frac{\overline{m}\sqrt{n}}{\sum\sqrt{\frac{m-\overline{m}}{n-1}}} = \frac{\overline{m}\sqrt{n}}{s_m} \tag{13}$$

Here, m_i is a random variable, $\overline{m}$ is the mean of m_i, s_m is the standard deviation, and n is the sample size. Suppose we want to compare the probability density functions of the ZIP and ZINB models. The H_o and H_1 hypotheses are as follows:

H_0: ZIP and ZINB distribution functions are equal

H_1: ZIP and ZINB distribution functions are not equal

Probability density functions with f_1 and f_2 , the representation way of m_i is as follows;

$$m_i = \log\left(\frac{f_1(y_i/x_i)}{f_2(y_i/x_i)}\right) \tag{14}$$

Within the family of ZIP models, testing if a Poisson model is adequate corresponds to testing: $H_0 = \emptyset_i = 0$ vs. $H_0 = \emptyset_i > 0$. In the interpretation of the Vuong test value having a normal distribution (e.g. for $\alpha = 0.05$ significance level), if the Vuong value is greater than 1.96, the first model is interpreted as "closer" to the real model; if the Vuong value is less than -1.96, the second model can be interpreted as "closer" to the real model. If the calculated value is not between $(-1.96;\ 1.96)$, it is interpreted as "there is no difference between using the first or the second model" [37].

4. Experimental Results

In this study, the data from a three-year study conducted by Hemmingsen et al. (2005) in four regions off the Norwegian coast to count parasites [38] were used. The dependent variable is the number of parasites (Intensity), while the independent variables are depth, length of the fish, and area. In addition, missing observations in the original data were removed from the model. Since the data set contains a large number of zeros, it was tested in zero-inflated models and generalized models. Statistical analysis of the study was made using Stata 14 software program. The frequency distribution showing the parasite density is given in Figure 1.

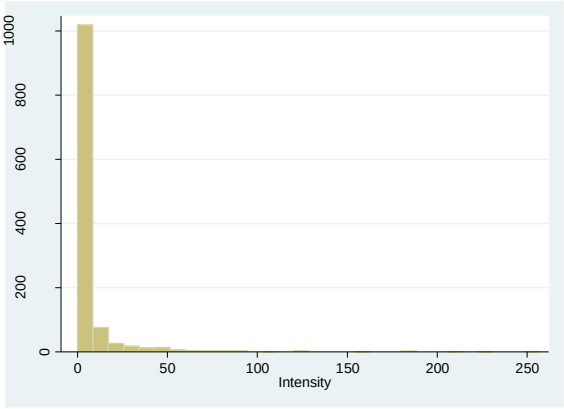

Fig 1. Frequency distribution of the number of parasites

4.1. Generalized Poisson Regression Model (GP)

The Poisson regression model, one of the most widely used generalized models, was first tried because overdispersion was detected in observation values. The results obtained are given in Table 1. IRR values show exp β values in Tables.

Table 1. Generalized Poisson regression model (GP)

Generalized Poisson regression	Number of obs	=	1191
	LR chi2(3)	=	89.68
Dispersion = .8910582	Prob > chi2	=	0.0000
Log-likelihood = −2566.7445	Pseudo R2	=	0.0172

Intensity	**IRR**	**Std. Err.**	***z***	***P* > *z***	**[95% Conf.**	**Interval]**
Depth	1.004756	.0006329	7.50	0.000	1.00351	1.006003
Length	.9981571	.0028743	−0.64	0.521	.9925496	1.003796
Area	1.123788	.047966	2.43	0.015	1.022953	1.234562
_cons	2.024088	.2408844	2.93	0.003	1.262374	3.245417
/atanhdelta	1.427039	.0548193			1.319595	1.534483
delta	.8910582	.0112936			.8666833	.9111886

Likelihood-ratio test of delta=0: chi2(1) = 1.8*e* + 04 Prob>=chi2 = 0.0000

According to the results presented in Table 1, the length was found to be insignificant in terms of the number of parasites. The area and depth variables were found to be significant. According to the GP model, the length variable was found insignificant ($p > 0.05$). Area and depth variables are significant ($p < 0.05$). Accordingly, a one-unit increase in depth increases the parasite intensity approximately 1.005 times. When the area changes, the parasites density increases approximately 1.124 times. When the model is evaluated as a whole, it is significant according to the chi-square test.

4.2. Generalized Negative Binomial: Famoye (GNB-F)

Results for GNB-F are shown in Table 2. According to this model, the depth, length, and area variables are significant ($p < 0.05$). The model was again found to be statistically significant ($Prob > chi2 = 0.0000$).

Table 2. Generalized negative binomial: Famoye (GNB-F)

Generalized negative binomial-Fregression		Number of obs	=	1191		
		LR chi2(3)	=	120.96		
Log likelihood = −2550.873		Prob > chi2	=	0.0000		
Intensity	**IRR**	**Std. Err.**	***z***	***P > z***	**[95% Conf.**	**Interval]**
Depth	1.006702	.0013168	5.11	0.000	1.004.125	1.009286
Length	.9756556	.0050295	−4.78	0.000	.9658476	.9855632
Area	1.245722	.0802515	3.41	0.001	1.097957	1.413373
_cons	15.753	6.412282	6.77	0.000	7.093826	34.98212
/lnphim1	−6.154845	2.049813			−101.724	−2.137286
/lntheta	−1.64351	.0809544			−1.802178	−1.484842
phi	1.002123	.0043521			1.000038	1.117975
theta	.1933004	.0156485			.1649393	.2265381

4.3. Generalized Negative Binomial Regression Model (GNB)

The GNB distribution is one of the most widely used models in cases where the variance is greater than the mean; that is, in the case of overdispersion. The results obtained are given in Table 3. The fact that $\alpha = 5.34$ dispersion parameter is greater than zero indicates that it is overdispersion. According to the GNB model, the depth, length, and area variables are significant ($p < 0.05$). One unit increase in depth increases the parasite intensity nearly 1.006 times. When the area changes, the parasite intensity increases approximately 1.258 times.

Table 3. Generalized negative binomial regression model (GNB)

Generalized Poisson regression		Number of obs	=	1191		
		LR chi2(3)	=	114.65		
		Prob > chi2	=	0.0000		
Log likelihood = -2551.0291		Pseudo R2	=	0.0220		
Intensity	**IRR**	**Std. Err.**	***z***	***P > z***	**[95% Conf.**	**Interval]**
Depth	1.006555	.0012471	5.27	0.000	1.004114	1.009002
Length	.9753681	.004905	−4.96	0.000	.9658017	.9850293
Area	1.258504	.0772588	3.75	0.000	1.115834	1.419414
_cons	3.096207	1.192705	2.93	0.003	1.45524	6.587573
/lnalpha	1.675206	.0539381			1.56949	1.780923
alpha	5.339897	.288024			4.804195	5.935333
Likelihood-ratio test of delta= 0: chibar2(01) = 1.8e + 04 Prob>=chibar2 = 0.0000						

4.3.1. Generalized Waring Regression Model (GNB-W)

The results obtained for the GNB-W model are given in Table 4. The depth, length and area variables are significant according to the GNB-W model ($p < 0.05$).

Table 4. Generalized Waring regression model (GNB-W)

Generalized negative binomial-W regression	Number of obs	=	1191
	LR chi2(3)	=	106.77
Log likelihood = −2544.401	Prob > chi2	=	0.0000

Intensity	IRR	Std. Err.	*z*	*P* > *z*	[95% Conf.	Interval]
Depth	1.008.675	.0014077	6.19	0.000	1.005919	1.011437
Length	.9783038	.0058167	−3.69	0.000	.9669695	.9897709
Area	1.204.492	.0851747	2.63	0.009	1.048605	1.383554
_cons	2.089.578	.9044374	1.70	0.089	.8946043	4.880748
/lnrhom2	−.4083326	.4713641			−1.332189	.5155241
/lnk	−1.508367	.0772628			−1.659799	−1.356935
rho	2.664758	.3133429			2.263899	3.674516
k	.221271	.017096			.1901771	.2574488

One unit increase in depth increases the parasites intensity approximately 1.008 times. When the area changes, the parasite intensity increases about 1.204 times.

4.3.2. Zero-Inflation

In the study, 651 observations within 1191 observations were found to contain zero values. In other words, approximately 55% of the parasite count data consists of zero observation. For this reason, analyses have also been made with zero-inflated models. The results of ZINB, ZINB-W, and ZINB-F models are given below.

4.3.2.1. Zero-Inflated Negative Binomial Regression (ZINB)

Zero-inflated models consist of two parts. ZINB model results are given in Table 5. The length variable is specified as the inflate variable. One unit increase in depth increases the parasites intensity approximately 1.006 times. When the area changes, the parasite intensity increases approximately 1.238 times. As the length decreases, the parasite density decreases (−0.223).

Table 5. Zero-inflated negative binomial regression (ZINB)

Zero-inflated negative binomial regression	Number of obs	=	1191
	Nonzero obs	=	540
	Zero obs	=	651
Inflation model = logit	LR chi2(3)	=	128.03
Log likelihood = −2539.562	Prob > chi2	=	0.0000

Intensity	IRR	Std. Err.	*z*	*P* > *z*	[95% Conf.	Interval]
intensity						
Depth	1.006497	.001221	5.34	0.000	1.004107	1.008893
Length	.9693981	.00494	−6.10	0.000	.9597642	.9791287
Area	1.238253	.0743365	3.56	0.000	1.100801	1.392868
_cons	4.755356	1.888854	3.93	0.000	2.183137	10.35822
inflate						
Length	−.2229618	.0809777	−2.75	0.006	−.3816751	−.0642484
_cons	5.721762	2.252244	2.54	0.011	1.307444	10.13608
/lnalpha	1.584263	.061479	25.77	0.000	1.463767	1.70476
alpha	4.875699	.2997532			4.322209	5.500066

Vuong test of zinb vs. standard negative binomial: $z = 2.58$ Pr> $z = 0.0049$

Using the ZINB distribution is more meaningful than the standard negative binomial distribution according to the Vuong test ($z = 2.58$ Pr $> z = 0.0049$). The variables are significant for both parts of the model.

4.3.2.2. Zero-Inflated Negative Binomial Regression-W (ZINB-W)

The ZINB-W model was compared with the standard Waring model. Table 6 shows the results. One unit increase in depth increases the parasites intensity approximately 1.008 times. When the area changes, the parasite intensity increases approximately 1.164 times. As the length decreases, the parasite density decreases (-0.175).

Table 6. Zero-inflated negative binomial regression-W (ZINB-W)

Zero-inflated gen neg binomial-W regression	Number of obs	=	1191
Regression link :	Nonzero obs	=	540
Inflation link: logit	Zero obs	=	651
	Wald chi2(3)	=	126.93
Log likelihood = -2530.653	Prob > chi2	=	0.0000

Intensity	**IRR**	**Std. Err.**	***z***	***P* > *z***	**[95% Conf.**	**Interval]**
intensity						
Depth	1.008.867	.0013606	6.55	0.000	1.006204	1.011537
Length	.9667968	.0063723	−5.12	0.000	.9543876	.9793674
Area	1.163945	.0820843	2.15	0.031	1.013686	1.336477
_cons	4.490452	2.178.058	3.10	0.002	1.735488	1.161873
inflate						
Lengt	−.1750797	.0808031	−2.17	0.030	−.3334509	−.0167085
_cons	4.615169	2.244926	2.6	0.040	.2151952	9.015143
/lnrhom2	−.9190866	.4479906			−1.797132	−.0410411
/lnk	−1.337637	.1120339			−1.55722	−1.118055
rho	2.398883	.1786959			2.165774	2.95979
k	.262465	.029405			.2107211	.3269151

Vuong test of zinbregw vs. gen neg binomial(W): $z = 29.81$ Pr $> z = 0.0000$

Bias-corrected (AIC) Vuong test: $z = 29.81$ Pr $> z = 0.0000$

Bias-corrected (BIC) Vuong test: $z = 29.79$ Pr $> z = 0.0000$

According to the Vuong test, this model is more significant than the generalized negative binomial distribution. The number of parasites increases as depth, length, and area change.

4.3.2.3. Zero-Inflated Negative Binomial Regression-F (ZINB-F)

The results found for ZINB-F are as in Table 7. When this model was compared with the results obtained with GNB-F, the Vuong test was found to be significant.

Table 7. Zero-inflated negative binomial regression-F (ZINB-F)

Zero-inflated gen neg binomial-F regression		Number of obs	=	1191		
Regression link :		Nonzero obs	=	540		
Inflation link: logit		Zero obs	=	651		
		Wald chi2(3)	=	139.87		
Log likelihood = −2539.562		Prob > chi2	=	0.0000		
Intensity_	**IRR**	**Std. Err.**	***z***	***P* > *z***	**[95% Conf.**	**Interval]**
Intensity_						
Depth	1.006497	.0012211	5.34	0.000	1.004107	1.008893
Length	.9693989	.0049403	−6.10	0.000	.9597644	.9791302
Area	1.238.237	.0743426	3.56	0.000	1.100775	1.392866
_cons	2.318.361	9.117.887	7.99	0.000	10.72536	50.11299
inflate						
Length	−.2229628	.0809865	−2.75	0.006	−.3816934	−.0642322
_cons	5.721764	2.252445	2.54	0.011	1.307053	10.13647
/lnphim1	−15.16193	695.4499			−1378.219	1347.895
/lntheta	−1.584.224	.061541			−1.704842	−1.463606
phi	1	.0001809			1	.
theta	.2051068	.0126225			.181801	.2314003
Vuong test of zinbregf vs. gen neg binomial(F): $z = 3.97$ Pr> $z = 0.0000$						
Bias-corrected (AIC) Vuong test: $z = 3.74$ Pr > $z = 0.0001$						
Bias-corrected (BIC) Vuong test: $z = 3.14$ Pr > $z = 0.0008$						

5. Conclusion

Modelling discrete data is a special type of regression. As is known, linear regression analysis can be used in cases where the dependent variable is continuous. However, the data to be used in the analysis may not always be available continuously. In such cases, if the data are discontinuous, analyses using linear regression models will give ineffective, inconsistent, and contradictory results. Therefore, count data models should be used when the dependent variable consists of nonnegative discrete values. One of the most common models used in count data analysis is the Poisson regression model. The most important feature of the Poisson regression model is that the variance and mean are equal. Generally, this feature cannot be provided in practice. In this case, negative binomial regression analysis or generalized Poisson regression analysis is widely used. In addition, cases where count data contain too many zero values are encountered in many areas. In such cases, the zero-inflated Poisson, zero-inflated Negative Binomial, Poisson Hurdle and Negative Binomial Hurdle regression models can be preferred.

Table 8. Model Selection

Count Models	**LL**	**AIC**	**BIC**
GP	−2566.7445	5143.489	5168.902
GNB-F	−2550.873	5113.747	5144.242
GNB	−2551.0291	5112.058	5137.471
GNB-W	−2544.401	5100.802	5131.298
ZINB	−2539.562	5093.124	5128.702
ZINB-W	**−2530.653**	**5077.306**	**5117.967**
ZINB-F	−2539.562	5095.124	5135.785

This study aims to compare the generalized Famoye and Waring models with classical methods apart from the commonly used methods. Thus, GP, GNB-F, GNB and GNB-W models were examined by considering an overdispersion data set. Since approximately 55% of the data set consists of zero, the zero-inflated models of these models were also tested, and a model comparison was made. As a result, LL, AIC, and BIC values for the six count models are given in Table 8. Due to the large number of zeros in the data set we were using, zero-inflated models yielded better results. Among these, the highest LL value and the lowest AIC and BIC values were obtained for the ZINB-W model.

The study focused on generalized models, especially on count regression models. For these models, their performances can also be investigated by conducting a simulation study. In the case of different rates of zero values and outliers in the data set, the models' performances can be compared. Thus, the reliability of the obtained results can be increased by selecting the appropriate model for the data structure.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] D. Kılıç, H. Bayrak, *A Comparison on Count Data Models: Example of Problems That Occured in E-Commerce Over the Turkey,* Selçuk University Journal of Science Faculty, 46(2) (2020) 85–102.

[2] B. Pittman, E. Buta, S. Krishnan-Sarin, S. S. O'Malley, T. Liss, R. Gueorguieva, *Models for Analyzing Zero-Inflated and Overdispersed Count Data: An Application to Cigarette and Marijuana Use*, Nicotine and Tobacco Research 22(8) (2018) 1390–1398.

[3] C. B. Dean, *Testing for Overdispersion in Poisson and Binomial Regression Models*, JASA 87(418) (1992) 451–457.

[4] M. Harris, J. Marti, Y. Bhatti, H. Watt, J. Macinko, A. Darzi, *Explicit Bias towards High-Income Country Research: A Randomized, Blinded, Crossover Experiment in English Clinicians*, Health Affairs 36(11) (2017) 1–9.

[5] A. Yeşilova, R. Atlıhan, *Analysing the Effects of Different Temperatures on Egg Numbers of Scymnus subvillosus Using Mixture Poisson Regression*, Yüzüncü Yıl University Journal of Agricultural Sciences 17(2) (2007) 73–79.

[6] Z. Yang, J. W. Hardin, C. L. Addy, Q. H. Vuong, *Testing Approaches for Overdispersion in Poisson Regression Versus the Generalized Poisson Model*, Biometrical Journal 49 (2007) 565–584.

[7] J. M. Hilbe, *Modelling Count Data* (First Edition). New York, 2014, Cambridge University Press.

[8] T. Harris, J. M. Hilbe, J. W. Hardin, *Modeling Count Data with Generalized Distributions*, The Stata Journal 14(3) (2012) 562–579.

[9] G. C. Jain, P. C. Consul, *A Generalized Negative Binomial Distribution*, SIAM Journal of Applied Mathematics 21(4) (1971) 501–513.

[10] P. C. Consul, H. C. Gupta, *The Generalized Negative Binomial Distribution and Its Characterization by Zero Regression*, SIAM Journal on Applied Mathematics 39 (1980) 231–237.

[11] F. Famoye, *Generalized Binomial Regression Model*, Biometrical Journal 37 (1995) 581–594.

[12] P. C. Consul, F. Famoye, *Generalized Poisson Regression Model*, Communications in Statistics-Theory and Methods 21 (1992) 89–109.

[13] J. O. Irwin, *The Generalized Waring Distribution Applied to Accident Theory*, Journal of the Royal Statistical Society Series A (General) 131(2) (1968) 205–225.

[14] K. Wang, K. W. Y. Kelvin, H. L. Andy, *A Zero-Inflated Poisson Mixed Model to Analyze Diagnosis Related Groups with Majority of Same-Day Hospital Stays,* Computer Methods and Programs in Biomedicine 68 (2002) 195–203.

[15] J. W. Hardin, J. M. Hilbe, *Generalized Linear Models and Extensions*, 3rd ed. College Station, TX: Stata Press, 2012.

[16] J. Rodriguez-Avi, A. Conde-Sanchez, A. J. Saez-Castillo, M. J. Olmo-Jimenez, A. M. Martinez-Rodriguez, *A Generalized Waring Regression Model for Count Data*, Computational Statistics and Data Analysis 53 (2009) 3717–3725.

[17] S. W. Martin, C. E. Rose, K. A. Wannemuehler, B. D. Plikaytis, *On of the Zero-inflated and Hurdle Models for Modelling Vaccine Adverse event Count Data*, Journal of Biopharmaceutical Statistics 16 (2006) 463–481.

[18] Y. Cui, W. Yang, *Zero-Inflated Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci Underlying Count Trait with Many Zeros*, Journal of Theoretical Biology 256 (2009) 276–285.

[19] M. Ridout, J. Hinde, C. G. B. Demetrio, *A Score Test for a Zero-Inflated Poisson Regression Model Against Zero- Inflated Negative Binomial Alternatives*, Biometrics 57 (2001) 219–233.

[20] J. M. Miller, *Comparing Poisson, Hurdle and Zip Model Fit Under Varying Degrees of Skew and Zero-Inflation*, Doctoral Thesis, University of Florida, 2007.

[21] F. Famoye, K. P. Singh, *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data*, Journal of Data Science 4(1) (2006) 117–130.

[22] C. Czado, V. Erhardt, A. Min, S. Wagner, *Zero-Inflated Generalized Poisson Models with Regression Effects on The Mean, Dispersion and Zero-Inflation Level Applied To Patent Outsourcing Rates*, Statistical Modelling 7 (2007) 125–153.

[23] S. M. Mwalili, E. Lesare, D. Declerck, *The Zero-Inflated Negative Binomial Regression Model with Correction for Misclassi_Cation: An Example in Caries Research*, Statistical Methods in Medical Research 17(2) (2008) 123–139.

[24] E. Altun, *A New Zero-Inflated Regression Model with Application*, Journal of Statisticians: Statistics and Actuarial Sciences 11(2) (2018) 73–80.

[25] E. Pamukçu, C. Colak, N. Halisdemir, *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) (2014) 89–96.

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

[27] M. Logan, *Biostatistical Design and Analysis Using R*, Willey-Blackwell, 2010.

[28] J. F. Lawless, *Negative Binomial and Mixed Poisson Regression*, The Canadian Journal of Statistics 15(3) (1987) 209–225.

[29] Y. Kaya, A. Yeşilova, *Investigation of E-Mail Traffic by Using Zero-Inflated Regression Models*, Anadolu University of Sciences & Technology-A: Applied Sciences & Engineering 13(1) (2012) 51–63.

[30] W. Wang, F. Famoye, *Modeling Household Fertiliq' Decisions with Generalized Poisson Regression*, Journal of Population Economics 10 (1997) 273–283.

[31] H. Akaike, *Information Theory and An Extension of The Maximum Likelihood Principle*, Second International Symposium on Information Theory (1973) 267–281, Academia Kiado, Budapest.

[32] C. M. Hurvich, C. L. Tsai, *Regression and Time Series Model Selection in Small Samples*, Biometrika 76 (1989) 297–307.

[33] N. Sugiura, *Further Analysts of the Data by Akaike's Information Criterion and the Finite Corrections*, Communications in Statistics - Theory and Methods 7(1) (1978) 13–26.

[34] A. D. R. Mcquarrie, C. Tsai, *Regression and Time Series Model Selection*, World Scientific Publishing Company, Singapore, 1998.

[35] Q. H. Vuong, *Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses*, Econometrica 57 (1989) 307–333.

[36] K. H. Pho, L. Sel, L. Sal, T. M. Lukusa, *Comparison Among Akaike Information Criterion, Bayesian Information Criterion and Vuong's Test in Model Selection: A Case Study of Violated Speed Regulation in Taiwan*, Journal of Advanced Engineering and Computation 3(1) (2019) 293–303.

[37] N. Ismail, H. Zamani, *Estimation of Claimcount Data Using Negative Binomial, Generalized Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Generalized Poisson Regression Models*, Casualty Actuarial Society E-Forum 41(20) (2013) 1–28.

[38] W. Hemmingsen, P. A. Jansen, K. Mackenzie, *Crabs, Leeches and Trypanosomes: An Unholy Trinity?*, Marine Pollution Bulletin 50(3) (2005) 336–339.