İstanbul Commerce University Journal of Science, 20(40), Fall 2021, 247-268.

İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

Istanbul Commerce University Journal of Science

http://dergipark.org.tr/ticaretfbd

Research Article / Araştırma Makalesi

COMPARISON OF SOME COUNT MODELS IN CASE OF EXCESSIVE ZEROS: AN APPLICATION

AŞIRI SIFIR DURUMUNDA BAZI SAYIM MODELLERİNİN KARŞILAŞTIRILMASI: BİR UYGULAMA

Öznur İŞÇİ GÜNERİ[1] **Burcu DURMUŞ[2]** **Aynur İNCEKIRIK[3]**

Corresponding Author / Sorumlu Yazar
oznur.isci@mu.edu.tr

Received / Geliş Tarihi
15.03.2021

Accepted / Kabul Tarihi
28.06.2021

Abstract

Different regression models have been developed in the literature for count data. Among these, the most well-known regression models are Poisson and negative binomial regression models. Poisson or negative binomial models are suitable if there are not many zero-valued terms. When there are excessive zeros in count data, zero-inflated Poisson models are the most preferred models in the case of equal dispersion, and zero-inflated negative binomial models are the most preferred models in case of overdispersion. Other models used in the case of too many zeros are the Poisson Hurdle and negative binomial Hurdle models. In this study, these models are compared for a sample data set. For this purpose, LL, AIC, BIC and Vuong test statistics were used.

Keywords: Count data, Hurdle model, negative binomial model, poisson model, zero-ınflated models.

Öz

Sayma verileri için literatürde farklı regresyon modelleri geliştirilmiştir. Bunlar arasında en bilinen regresyon modelleri Poisson ve negatif binomial regresyon modelleridir. Poisson ya da negatif binomial modeller eğer fazla sıfır değerli terimler yoksa uygun olur. Sayma verilerinde aşırı sıfır olduğunda eşit yayılım durumunda zero-inflated Poisson, aşırı yayılım durumunda zero-inflated negatif binom modelleri en çok tercih edilen modellerdir. Çok fazla sıfır olması durumunda kullanılan başka bir model de Poisson Hurdle ve negatif binomial Hurdle modelleridir. Bu çalışmada örnek bir veri seti için bu modeller karşılaştırılmıştır. Bu amaçla LL, AIC, BIC ve Vuong test istatistiği kullanılmıştır.

Anahtar Kelimeler: Hurdle model, negatif binomial model, poisson model, sayma verisi, sıfır şişirilmiş modeller.

[1]Mugla Sitki Kocman University, Faculty of Science, Department of Statistics, Kotekli Campus, Mugla, Turkey. oznur.isci@mu.edu.tr, Orcid.org/0000-0003-3677-7121.

[2]Mugla Sitki Kocman University, Rectorate Performance Analysis Unit, Kotekli Campus, Mugla, Turkey. burcudurmus@mu.edu.tr, Orcid.org/0000-0002-0298-0802.

[3]Manisa Celal Bayar University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Manisa, Turkey. aynur.incekirik@bayar.edu.tr, Orcid.org/0000-0002-1825-0097.

1. INTRODUCTION

Count models have a wide range of applications, especially in fields such as public health, epidemiology, psychology, social sciences, economics, demography, sociology, insurance and educational sciences. Poisson regression (PR), which is one of the widely used count models, uses the assumption that the conditional variance of the dependent variable is equal to the conditional mean, while negative binomial regression (NBR) is used in the case of overdispersion. Applying Poisson regression causes bias in parameter estimates and standard errors in case of overdispersion (Khoshgoftaar et al., 2005). In case of overdispersion, except for negative binomial distribution, generalized Poisson regression model, generalized negative binomial regression model, quasi model can be applied. Apart from these, Poisson-inverse Gaussian, Poisson-Lognormal are other methods used (Denuit et al., 2007).

Count data has zero values by nature and the classical least squares (OLS) method does not give good estimates because it does not show normal distribution. The presence of more than expected zero values in the data set is defined as zero-inflation (Martin et al., 2006; Cui & Yang, 2009). In data sets where most of the observations are zero, excluding the zero values from the analysis causes to obtain incorrect results. Zero-inflation count data may lack equality of mean and variance. In such a case, over-dispersion or under-dispersion must be taken into account. When there are excessive zeros in the data set, new models are needed for such data because when there are many zeros in the sample, Poisson and negative binomial distributions cannot predict well enough. Therefore, Lambert (1992) first proposed the zero-inflation Poisson (ZIP) model with an application of manufacturing defects. Later, Green conducted a study in 1994 on taking excessive zeros and sample selection into account in Poisson and negative binomial regression models.

Famoye and Singh (2006) proposed the zero-inflation generalized Poisson (ZIGP) model, which is an extension of the generalized Poisson distribution. Another widely used method is the negative binomial model, which can be preferred in cases where the Poisson mean has a gamma distribution. A natural extension of the negative binomial model is the zero-inflation negative binomial (ZINB) model when there are excess zeros in the data (Mwalili et al., 2008).

When you want to use the zero-inflation regression model, first consider whether a conventional negative binomial model is good enough. Just the presence of too many zeros in the dataset doesn't mean you need a zero-inflation model. There are two types of zeros in the zero-inflation model, namely "real zeros" and "excess zeros". Of course, there are situations where a zero-inflation model makes sense in terms of theory or common sense. For example, if the dependent variable is the number of children born in a sample of women aged 50, it is reasonable to assume that some women are biologically infertile. For these women, no change in predictive variables can change the expected number of children (Allison, 2012).

Another popular approach to modeling excess zeros in count data is to use truncated models. The Hurdle model is an example of truncated patterns for census data (Cragg, 1971). A failure to account for the correct type of over or under dispersion leads to very different estimates of the regression parameters and incorrect inferences about the model parameters (McCullagh & Nelder, 1989; Ver Hoef & Boveng, 2007). In the literature, hurdle and ZIP models are widely used for analyzing count responses with excessive zeros. However, hurdle and ZIP models do not allow for underdispersion with excessive zeros, these models apply only when there is overdispersion in the response variable (Lee at al., 2016).

Zero-inflated count models offer a way of modeling the excess zeros in addition to allowing for overdispersion in a standard parametric model. However, the hurdle model is flexible and can handle under-dispersion, overdispersion, and excess zeros problem (Workie & Gedef, 2021).

Zero-inflated models are used in many studies to model data which has high zero density. Ridout et al. (1998) reviewed some zero inflated models and hurdle models and gave an example on biological count data. Yip and Yau (2005) studied on zero-inflated distributions for claim frequency and they used the generalized Pearson $\chi 2$ statistic and information criteria. Greene (2005) has compared Zero Inflated and Hurdle Models. In this work, several extensions of the models are described and an application to health care demand data for comparison of the models is presented. Flynn (2009) compared traditional Poisson and Negative Binomial models with the Zero Inflated Models. Mouatassim and Ezzahid (2012) compared Poisson and zero-inflated Poisson model for health insurance and they used Vuong test for model comparison. A new zero-inflated regression model for zero-inflated count data and a new regression model so called Poisson quasi-Lindley regression model for over-dispersed count data are proposed by Altun (2018, 2019). Erdemir and Karadağ (2020) investigated models for count data with excessive zeros in non-life insurance.

There are also hurdle models as an alternative to zero-inflated models. Boucher et al. (2008) used compound frequency models and they examined different risk classification models for count data by using Score and Haussmann tests. Yang et al. (2012) proposed new link functions for hurdle Poisson and hurdle negative binomial to deal with zero-inflation, overdispersion and missing observations in clinical trials. Sarul and Şahin (2015) compared Poisson models, zero-inflated models and hurdle models for claim frequency data. Baetschmann and Winkelmann (2017) introduced a new dynamic hurdle model for zero-inflated count data. Sakthivel and Rajitha (2017) compared methods with back propagation neural network for modeling the count data which has excessive number of zeros by using mean square error for model selection.

Although there are many publications on overdispersion in the literature, fewer publications are made because under-dispersion is a less common situation. Conway-Maxwell-Poisson (COM-Poisson) distribution can handle under dispersed count data. It is a flexible distribution that can account for under dispersion usually encountered in some types of count data (Shmueli et al. 2005; Sellers and Shmueli 2010). In Figure 1, frequently used models in count data are given.

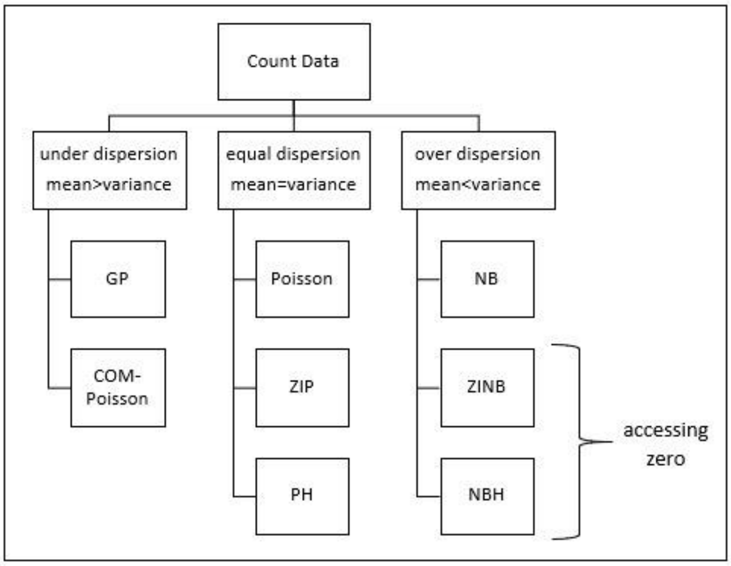

Figure 1. The Frequently Used Models in The Count Data Analysis

Hurdle models assume that there is only one process where zero can be generated, whereas zero-inflation models assume that there are two different processes that can produce zero. The first is the on-off part, which is a binary process. System π likely "off" and 1 - possibly "open" (where π is known as inflation probability). When the system is "off", only zero counting is possible. This part is the same for zero-inflation and Hurdle models. The second part is the counting part that occurs when the system is "on". This is where the Zero-inflation and Hurdle models differ. In zero-inflation models, the numbers can still be zero. In Hurdle models, they must be different from zero. For this section, zero-inflation models use a "normal" discrete probability distribution, while Hurdle models use a discrete probability distribution cut from zero.

To give an example to explain the Hurdle model; a car manufacturer wants to compare two quality control programs for its cars. It will compare them according to the number of warranty claims made. For each program, a randomly selected set of clients are monitored for one year and the number of warranty claims they file is counted. The "closed" state means "there is zero claim", "open" state means "at least one claim has been made". In the zero-inflation model, researchers discovered that some repairs on cars were fixed without filing a warranty claim. In this way zeros are a mixture of the absence of quality control problems as well as the presence of quality control problems involving no warranty claims. "closed" means "zero claims" while "open" means "at least one claim has been made or repairs have been done without a claim (James, 2014).

In count regression models, parameter estimates are commonly obtained using the Maximum Likelihood method (Karen & Kelvin, 2005). Information criteria such as Akaike information criterion (AIC) and Bayes information criterion (BIC) can be used to select the appropriate model. In addition, model comparisons with Vuong test statistics, which are widely used in zero-inflated models, are also made.

In the second section of this study, counting regression models are presented. In the third section, the model selection criteria used in the study are explained. In the 4th section, analyzes are made on a sample data set. In the 5th section, the results are evaluated.

2. MATERIALS AND METHODS

2.1. Poisson Regression (PR)

When the dependent variable consists of discrete and non-categorical counting data, the first method used is Poisson regression analysis. In Poisson regression analysis, it is assumed that the dependent variable y_i shows a Poisson distribution (Deniz, 2005). Probability density function for Poisson distribution with λ parameter (Sinharay, 2010) is as follows;

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

In this expression, y_i is the number of occurrences of events, and λ is the rate of repetition of events per unit of time. In other words, λ gives the mean of the distribution. Here the probability changes as a function of the λ value. The Poisson probability distribution is slanted to the right. However, as λ_i grows, the distribution approaches the normal distribution. The Poisson distribution is mostly used to model the number of rare events occurring. The most prominent feature of the Poisson regression model is that mean and variance are equal to each other;

$$E(y) = \lambda \text{ and Var (y)} = \lambda \tag{2}$$

Over or under-dispersion data sets cannot be modeled with the Poisson distribution because distortions are seen in the assumption that the conditional expected value is equal to the variance and the assumption is not satisfied. In practice, count variables show overdispersion, as they generally have greater variance than the average. The overdispersion of the data is caused by the number of observed zero values exceeding the zero values revealed by the Poisson model and unobserved heterogeneity (Kibar, 2008). The mean of the Poisson distribution, λ, is assumed to be a linear function of the arguments x_i. Poisson regression model can be given as follows;

$$log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + , \dots, + \beta_m x_m = x_i' \beta \tag{3}$$

In this equation, λ_i is an exponential function of independent variables. The value of λ_i can be written as follows;

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_{2+} +, \dots, + \beta_m x_m) = \exp(x_i' \beta) \tag{4}$$

Poisson regression is estimated by the maximum probability estimate. Log likelihood function of Poisson model (Shalabh, 2020);

$$LL_{Poisson} = \ln\mathrm{L}(\mathrm{y}, \lambda) = \sum_{i=1}^{n} y_i \ln(\lambda_i) - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \ln(y_i!) \tag{5}$$

After selecting the appropriate link function, the log likelihood function can be maximized for a given dataset using some numerical optimization techniques. The Poisson regression model usually requires a large sample.

In a Poisson model, the mean variance equality can be tested with the dispersion test. The test simply tests this assumption as a null hypothesis against an alternative where Var(λ)= λ+c∗f(λ), the constant $c<0$ means underdispersion and $c>0$ means overdispersion. The function f(.) is some monoton function (often linear or quadratic; the former is the default). The resulting test is equivalent to testing H_0:c=0 vs. H_1:c≠0 and the test statistic used is a t statistic which is asymptotically standard normal under the null.

2.2. Negative-Binomial Regression (NBR)

Possible values of y_i in negative binomial regression are again non-negative integer values such as 0,1,2,… etc. as in Poisson regression. Although negative binomial regression is a special case of Poisson regression, it is used as an alternative method in cases where zero values show over-dispersion (or under-dispersion) in applications. Negative binomial regression is a generalization of Poisson regression where the variance is equal to the mean calculated by the Poisson model and which relaxes the restrictive assumption. The negative binomial distribution has one more parameter, different from the Poisson distribution. Therefore, the second parameter can be used to adjust the variance independently of the mean. This model is based on a Poisson-Gamma mixed distribution. The Poisson distribution can be generalized by including a gamma noise variable with mean *1* and scale parameter *v*. The negative binomial distribution of Poisson-Gamma mixture obtained with α spread parameter is as follows (NNCS, 2020),

$$P(y_i|\lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i}, i = 1, 2, \dots, n$$
$$\lambda_i = t_i \lambda\,, \;\; \alpha = \frac{1}{v} \tag{6}$$

The expected value of the NBR model is $E(y) = \lambda$ and variance $Var(y) = \lambda + \alpha\lambda^2$ a quadratic function of the mean for α>0, equal to the Poisson variance if α=0. NBR model with t_i exposure time and $\beta_1, \beta_2, \dots, \beta_k$ unknown parameters can be shown as follows;

$$\lambda_i = \exp(\ln(t_i)\, \beta_{1i} x_{1i} + \beta_{2i} x_{2i}, \dots, \beta_{ki} x_{ki}) \quad (7)$$

Regression coefficients are estimated using the maximum likelihood method (Cameron et al., 2013). The log likelihood function of the negative binomial model can be given as follows (Zwilling, 2013);

$$LL_{NB} = \ln L(\alpha, \beta) = \sum_{i=1}^{n} (y_i ln\alpha + y_i(\alpha_i \beta_i) - (y_i + \frac{1}{\alpha}) \ln\left(1 + \alpha e^{\alpha_i \beta_i}\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma(\frac{1}{\alpha})) \quad (8)$$

Parameters are obtained by iterative solution methods.

2.3. Zero-Inflated Poisson Regression (ZIP)

One of the alternative methods used to analyze over-dispersed data is the zero-value weighted Poisson regression (ZIP) model. Zero value weighted Poisson regression is also used in modeling the dependent variable when the data set contains more zero values than expected. In the ZIP model, it is assumed that its dependent variable consists of two different data groups. These are structural zeros that always come from the zero group and the values that always come from the non-zero group, that is, the group defined as sampling zero (Peng, 2013). ZIP regression can be written as follows to explain the excess zeros in the dependent variable y_i (Lambert, 1992),

$$\Pr(y_i/x_i) = \begin{cases} \pi_i + (1 - \pi_i)\exp(-\lambda_i), & y_i = 0, \\ (1 - \pi_i)\exp(-\lambda_i)\lambda_i / y_i!, & y_i > 0. \end{cases} \quad (9)$$

In this model, $0 \le \pi_i \le 1$ and $\lambda_i > 0$. The mean of the ZIP model is shown as $E(y) = (1 - \pi)\lambda$ and its variance as $Var(y) = (1 - \pi)\lambda(1 + \pi\lambda)$. If $\pi = 0$, the ZIP model turns into PR. If $\pi_i > 0$, it is an indicator of overdispersion. The ZIP model is a two-piece model. From these parts, the log function is used to model positive numbers from both structural zero and sampling zero, as well as positive numbers from Poisson and negative binomial distributions. The other part is the logit function. This part is used to model the zeros in the data set (Peng, 2013). The log likelihood function for y_i dependent variable can be written as follows (Yau & Lee, 2001),

$$LL_{ZIP} = \sum_{i=1}^{n} \left(I_{yi=0} \log\left(\pi_i + (1 - \pi_i) e^{-\lambda_i}\right) + I_{yi>0} \log\left((1 - \pi_i) \frac{\lambda_i^{y_i} e^{\lambda_i}}{y_i!} \right) \right)$$
$$= \sum_{i=1}^{n} I_{yi=0} \log\left(\pi_i + (1 - \pi_i) e^{-\lambda_i}\right) + I_{yi>0} \log\left((1 - \pi_i) + y_i log\lambda_i - \lambda_i - log y_i!\right)) \quad (10)$$

The I. expression given in equation 10 is the indicator function for the specified event. From here, the parameters λ_i and π_i can be obtained using the link functions.

$$\log(\lambda) = B\beta \quad (11)$$

and

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathrm{G}\gamma \tag{12}$$

In equations 11 and 12, B and G are covariant matrices and are unknown parameter vectors (Yau, 2002; Yeşilova et al., 2010). The parameters β and γ can be obtained using maximum likelihood estimates.

2.4. Zero-Inflated Negative Binomial Regression (ZINB)

Zero value weighted ZINB model is used as an alternative method in cases where there is zero weighted and overdispersion data sets. This model has been defined as an improved version of the NB model (Greene, 1994). As with the ZIP model, zero and non-zero observations are modeled separately. However, unlike ZIP regression, non-zero observations in ZINB are modeled by NB regression. An alternative regression method is ZINB in modeling the dependent variable y_i in the case of overdispersion with many zero values. The ZINB model equation is as follows (Ridout et al., 2001):

$$\Pr(y_i/x_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\lambda_i)^{-\alpha^{-1}}, & y_i = 0, \\ (1-\pi_i)\dfrac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\dfrac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{y_i+\frac{1}{\alpha}}}, & y_i > 0. \end{cases} \tag{13}$$

In equation 13, the parameters π_i and λ_i depend on covariates and α>0 is an overdispersion parameter. The expected value of the ZINB model is shown as $E(y) = (1-\pi)\lambda$ and its variance as $Var(y) = E(y)(1+\alpha\lambda+\pi\lambda)$. In case of α> 0 and $\pi > 0$, there is overdispersion. ZINB log likelihood function for y_i (Yau, 2002):

$$\begin{aligned} LL_{ZINB} &= L(\lambda,\alpha,\pi;y) = \sum_{i=1}^{n}\left(I_{yi=0}\log\left(\pi_i + (1-\alpha\lambda_i)^{-\alpha^{-1}}\right)\right) \\ &+ I_{yi>0}\log\left((1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\frac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{y_i+\frac{1}{\alpha}}}\right) \\ &= \sum_{i=1}^{n}\left(I_{yi=0}\log\left(\pi_i + (1-\pi_i)(1-\alpha\lambda_i)^{-\alpha^{-1}}\right)\right) + I_{yi>0} \\ &\log\left((1-\pi_i)\frac{1}{\alpha}\log(1+\alpha\lambda_i)y_i\log(1+\frac{1}{\alpha\lambda_i}) + log\Gamma\left(y_i+\frac{1}{k}\right) - log\Gamma\left(\frac{1}{\alpha}\right) - log y_i!\right) \end{aligned} \tag{14}$$

The $I\cdot$ expression given in equation 14 is the indicator function for the specified event. λ_i and π_i parameters can be obtained by using link functions (Lambert, 1992).

$$\log(\lambda) = \mathrm{X}\beta \tag{15}$$

and

$$\log\left(\frac{\pi}{1-\pi}\right) = G_\gamma \tag{16}$$

Here, X and G are covariate matrices, β and γ are unknown parameter vectors of dimensions (p+1)x1 and (q+1)x1, respectively. The parameters β and γ can be obtained using maximum likelihood estimates. Zero-inflation negative binomial models are not recommended for small samples. What constitutes a small sample is not clearly defined in the literature (Mamun, 2014).

2.5. Hurdle Regression

Hurdle models were first proposed by a Canadian statistician Cragg (1971), and later developed further by Mullahy (1986). These models are used for data sets with many zero values. Hurdle models consist of two stages. First, binary responses showing positive counts (1) versus zero counts (0); the second is the process in which only positive counts occur (Yeşilova et al., 2010). Binary responses are modeled using the logit connection function. Positive counts are modeled using the zero-value truncated counting model, that is, the log link function (Rose et al., 2006). In Hurdle models, Poisson Hurdle (PH) model is used if the counting part shows Poisson distribution, and NB Hurdle (NBH) model is used in case of negative binomial distribution (Gerdtham, 1997). The hurdle model is flexible and can handle both under and overdispersion problem. Hurdle models are widely used especially in healthcare applications.

2.5.1 Poisson Hurdle model (PH)

Positive observations based on truncated count data $y_i > 0$ are called the PH model when modeled using the poisson distribution. The Hurdle model is defined in the Poisson case as follows (Dalrymple et al., 2003):

$$P(y_i = 0/x) = 1 - p(x),$$
$$P(y_i = q/x, z) = \frac{p(x)\exp\left(-\lambda(z)\right)\lambda(z)^q}{q!\,(1-\exp\left(-\lambda(z)\right))}\ ,\quad q = 1,2,\ldots \tag{17}$$

In equation 17, x and z are covariate matrices. In this equation, $p(x)$ ve $\lambda(z)$ are modeled using logit and log-linear functions respectively. The Hurdle model formulation is very similar to the ZIP model but the Hurdle model keeps the class zero from the non-zero by modeling the non-zero y_i's with a truncated Poisson distribution. It is expressed as $\lambda(z)$ and p_i,

$$log\left(1-\lambda(z)\right) = x_i'\beta, \tag{18}$$

and

$$logit(p_i) = z_i'\alpha \tag{19}$$

β and α given in equation 18 and equation 19 are unknown parameter vectors respectively. The mean of the PH model is shown as $E(y) = (1-\pi)E(YIY > 0) = (1-\pi)\frac{\lambda}{1-e^{-\lambda}}$ and its variance as $Var(y) = (1-\pi)var(YIY > 0) + \pi(1-\pi)[E(YIY > 0)]^2$. For PH, the log likelihood function is written as:

$$\begin{aligned} LL_{PH} &= \sum_{y_i>0} x_i\beta - \sum_{i=0}^{n}\log(1+\exp(x_i\beta)) \\ &+ \sum_{y_i>0}[y_i z_i\alpha - \exp(z_i\alpha) - \log(1-\exp(-\exp(-\exp(z_i\alpha))) - \log(y_i)\,!)] \\ &= LL(\beta) + LL(\alpha) \end{aligned} \tag{20}$$

The parameters β and α can be obtained using maximum likelihood estimates.

2.5.2. Negative binomial Hurdle model (NBH)

If there is additional zero-inflation in the NB model, NBH model is used among other alternative models. The probability function of the Negative Binomial Hurdle Model is as follows (Sarul & Şahin, 2015):

$$\Pr(y_i/x_i) = \begin{cases} \pi_0 & , y_i = 0 \\ (1-\pi_i), \dfrac{g}{1-(1+\alpha\lambda)-\alpha^{-1}} & , y_i > 0 \end{cases} \tag{21}$$

where $g = g(y;\lambda,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{(y+1)\Gamma(\alpha^{-1})}(1+\alpha\lambda)^{-\alpha^{-1}-y\alpha^{y}\lambda^{y}}$. The mean of the NBH model is shown as $E(y) = (1-\pi)\frac{\lambda}{1-(1+\alpha\lambda)^{-\frac{1}{\alpha}}}$ and its variance as $Var(y) = (1-\pi)var(YIY > 0) + \pi(1-\pi)[E(YIY > 0)]^2$. The log likelihood function of NBH:

$$LL_{NBH} = \ln\big(f(0)\big) + \{\ln[1-f(0)] + lnP(t)\} \tag{22}$$

In equation 22, $f(0)$ represents the probability of the binary part and *p(j)* the probability of a positive count. The probability of zero when using the logit model,

$$f(0) = P(y=0;x) = \frac{1}{1+\exp(xb1)} \tag{23}$$

and

$$1-f(0) = P(y=0;x) = \frac{\exp(xb1)}{1+\exp(xb1)} \tag{24}$$

For both parts of the NBH model, the log likelihood function can be written as (Yeşilova et al., 2010):

$$\begin{aligned} LL_{NBH} &= cond\{y=0, \ln(\frac{1}{1-\exp(xb1)}, \ln(\frac{\exp(xb1)}{1+\exp(xb1)}) \\ &+y * \ln\left(\frac{\exp(xb)}{1+\exp(xb)}\right) - \ln(\frac{1+\exp(xb)}{\alpha}) + \ln\Gamma\left(\mathrm{y}+\frac{1}{\alpha}\right) - \ln\Gamma\left(\frac{1}{\alpha}\right) \\ &-\ln(1-(1+\exp(xb))\left(-\frac{1}{\alpha}\right))\} \end{aligned} \tag{25}$$

3. MODEL SELECTION

Pearson statistics, deviance statistics (Deviance), Log-likelihood(LL), Akaike Information Criteria (AIC) and Bayes Information Criteria (BIC) are commonly used criteria in testing the goodness of fit of regression models. Since LL, AIC, BIC and Vuong statistics are used in this study, these statistics are explained below.

3.1. Log Likelihood (LL)

The log likelihood (LL) test is one of the most widely used tests for comparing different models. The LL test can be used to test for the presence of overdispersion. To test the Poisson model against 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;

$$LL = 2(lnL_1 - lnL_0) \tag{26}$$

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 (Wang & Famoye, 1997). When choosing the model over LL value, the model with the largest log-likelihood value is determined as the appropriate model.

3.2. Akaike Information Criteria (AIC)

This criterion, which is widely used to compare different models is;

$$AIC = -2log(\mathcal{L}) + 2k \tag{27}$$

In this equation, $\mathcal{L}$ is the maximum value of the log likelihood function; k represents the number of explanatory variables. Among the existing models, the model with the smallest 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. This value can be written as follows (Akaike, 1973; Sugiura, 1978; Hurvich & Tsai, 1989);

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} = \frac{2kn}{n-k-1} - 2ln(L) \tag{28}$$

3.3. Bayes Information Criterion (BIC)

Akaike proposed the BIC (Bayesian Information Criterion) model selection criterion for selected model problems in linear regression (McQuarrie & Tsai, 1998). AIC and BIC criteria are usually given together. Equality regarding the Bayesian measure of knowledge is as follows:

$$BIC = -2log(\mathcal{L}) + klog(n) \tag{29}$$

As with the AIC, the model with the lowest BIC value among the available models is selected as the appropriate model.

3.4. Vuong Statistic (V)

The Vuong test statistic is used to compare two models' fit to the same data with maximum probability. Specifically, it tests the null hypothesis arguing that the two models fit the data equally well. Vuong statistics is calculated as follows (Vuong, 1989);

$$V = \frac{\overline{m}\sqrt{n}}{sd(m)} \tag{30}$$

Where $\overline{m}$ is the mean of m_i, $sd(m)$ represents the standard deviation and n represents the sample size. m_i is expressed as follows:

$$m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right) \quad (31)$$

Vuong test statistics has a standard normal distribution. If the significance level is taken as α= 0.05 and if V> 1.96, it means that the first model is closer to the real model, yet, if V <-1.96, then it means that the second model is closer to the real model. If the calculated value is not between ± 1.96, then it means that that there is no difference between using the first or the second model.

4. EXPERIMENTAL RESULTS

The data set used in the current study is from Hemmingsen et al. (2005), who investigated the number of parasites in a study carried out for three years in four regions off the coast of Norway. The "intensity" variable, which indicates the number of parasites, was taken as the dependent variable. Independent variables are the variables of depth, weight, length, age, and area. Original observation values consist of 1254 data. But some observation values were excluded because they did not exist. As in the current study the dependent variable was count data and the analyses were made for PR, NBR, ZIP, ZINB, PH and NBH models. To evaluate the goodness of fit of the models, log likelihood, AIC, BIC and Vuong statistics values were calculated. Stata and RStudio were used for analysis. The histogram for the distribution of the number of parasites (intensity) is given in Figure 2. The distribution conforms to the Poisson distribution. Descriptive statistics are given in Table 1.

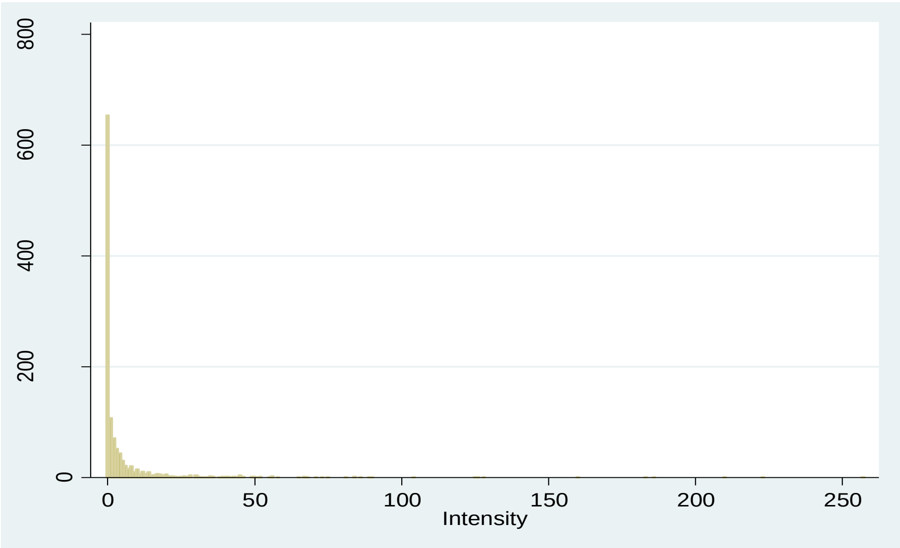

Figure 2. Frequency Distribution of Parasites Numbers (Intensity)

Table 1. Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Intensity	1191	6.209.068	1.964.186	0	257
Depth	1191	1.763.115	7.174.705	50	293
Weight	1191	1.717.688	1355.43	34	9990
Length	1191	5.353.065	1.418.831	17	101
Age	1191	4.118.388	190.539	0	10
Area	1191	256.843	1.078.504	1	4

4.1. Poisson Regression (PR)

Poisson regression is often used in modeling census data. However, in order for this model to be used, it must conform to the Poisson distribution and show an equal spread. That is, the mean of the dependent variable must be equal to its variance. The results obtained for the Poisson regression analysis are given in Table 2.

Table 2. Poisson Regression Analysis

Poisson regression				Number of obs = 1,191 LR chi2(5) = 5086.40 Prob > chi2 = 0.0000		
Log likelihood = -11316.054				Pseudo R2 = 0.1835		
Intensity	**Coef.**	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Depth	.0041254	.0001935	21.31	0.000	.003746	.0045047
Weight	-.0002165	.0000321	-6.74	0.000	-.0002795	-.0001535
Length	-.0272516	.0025701	-10.60	0.000	-.0322889	-.0222144
Age	.1241356	.0112843	11.00	0.000	.1020188	.1462523
Area	.5170952	.0140313	36.85	0.000	.4895943	.544596
_cons	.7040017	.0929926	7.57	0.000	.5217395	.8862639

When the Poisson regression model was examined, all variables were found to be statistically significant ($p \leq 0.05$). The presence of overdispersion was tested using the "dispersiontest" function of the AER package of the R software (z = 2.9025, p-value = 0.001851, dispersion(c)= 2.786516). Thus, in data set, overdispersion was detected. In some cases, a misspecified model may present a symptom such as an overdispersion problem. A common cause of overdispersion is excess zeros which are generated by an additional data generation process. In this case, the zero-inflation model should be considered.

The zero.test function of the “vcdExtra” package of the R software was used to test whether the Poisson distribution is suitable for processing zero frequencies in the data set (Chi-square = 178219.44184, df = 1, pvalue: < 2.22e-16).

4.2. Negative Binomial Regression (NBR)

The results obtained for the negative binomial regression analysis are given in Table 3.

Table 3. Negative Binomial Regression Analysis

Negative binomial regression					Number of obs =	1,191
					LR chi2(5) =	129.74
Dispersion = mean					Prob > chi2 =	0.0000
Log likelihood = -2543.4871					Pseudo R2 =	0.0249
Intensity	**Coef.**	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Depth	.007029	.0012848	5.47	0.000	.0045109	.009547
Weight	-.0000879	.0001376	-0.64	0.523	-.0003576	.0001818
Length	-.0421083	.0147863	-2.85	0.004	-.0710888	-.0131277
Age	.2211496	.0553297	4.00	0.000	.1127053	.3295938
Area	.2487281	.0632454	3.93	0.000	.1247694	.3726868
_cons	1.117.684	.5818868	1.92	0.055	-.0227928	2.258.161
/lnalpha	1.654.779	.0541407			1.548.665	1.760.893
alpha	5.231.923	.2832601			4.705.185	5.817.629

Likelihood-ratio test of alpha=0: chibar2(01) = 1.8e+04 Prob>=chibar2 = 0.000

NBR can be used for overly distributed count data; that is, when conditional variance exceeds conditional mean. When the model was examined, the variables other than the "weight" variable were found to be significant. In this model, alpha (α=5,231) represents the dispersion parameter. The Poisson model is the model in which this α value is limited to zero. In other words, when the dispersion parameter is zero, the negative binomial distribution is equal to the Poisson distribution. Here it was found quite different from zero. A common cause of overdispersion is excessive zeros caused by an additional data generation. In this case, the zero-inflation model should be considered again.

4.3. Zero Inflated Poisson Regression (ZIP)

In the data set, 651 observations among 1191 observations consist of zeros. For this reason, the ZIB model was tried. The results obtained for the zero-inflated Poisson regression analysis are given in Table 4.

Table 4. Zero-inflated Poisson Regression Analysis

Zero-inflated Poisson regression				Number of obs =		1,191
				Nonzero obs =		540
				Zero obs =		651
Inflation model = logit				LR chi2(5) =		3255.91
Log likelihood = -7157.201				Prob > chi2 =		0.0000
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Intensity						
Depth	.0013499	.0002057	20.607	0.000	.0009467	.0017531
Weight	.0001842	.0000319	5.78	0.000	.0001217	.0002467
Length	-.058611	.0027065	-21.66	0.000	-.0639157	-.0533063
Age	.083538	.0113188	7.38	0.000	.0613535	.1057225
Area	.3155613	.01262	25.00	0.000	.2908266	.340296
_cons	3.761.493	.0976798	38.51	0.000	3.570.044	3.952.942
inflate						
Depth	-.0066406	.0009475	-7.01	0.000	-.0084977	-.0047836
Weight	.0004211	.0001239	3.40	0.001	.0001782	.0006639
Length	-.0287779	.0124624	-2.31	0.021	-.0532037	-.0043521
Age	-.1218325	.0504857	-2.41	0.016	-.2207826	-.0228824
Area	-.0576589	.0627729	-0.92	0.358	-.1806915	.0653738
_cons	2.834.424	.4768654	34.455	0.000	1.899.785	3.769.063

Vuong test of zip vs. standard Poisson: z =11.19 Pr>z = 0.0000

Zero-inflated Poisson regression model given in Table 4 is statistically significant (Prob> chi2 = 0.000). The first model gave similar results to Poisson regression analysis. However, in the second model, the variable "Area" was found to be insignificant.

Vuong testing compares the ZIP model with a classical Poisson regression model. Significance of the Z test indicates that the ZIP model is better (z = 11.19 Pr > z = 0.0000). This model has both a count model and a logit model. According to the ZIP model, all the "inflate" variables except for "Area" were found to be significant. The ZIP model can be applied both when the zero observation values are too high and when there is equal dispersion.

4.4. Zero Inflated Negative Binomial Regression (ZINB)

ZINB distribution was applied as there were both overdispersion and zero values in the data set. The results obtained for the zero-inflated negative binomial analysis are given in Table 5.

Table 5. Zero-inflated Negative Binomial Regression Analysis

Zero-inflated negative binomial regression Inflation model = logit Log likelihood = -2491.515			Number of obs = 1,191 Nonzero obs = 540 Zero obs = 651 LR chi2(5) = 92.13 Prob > chi2 = 0.0000			
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Intensity						
Depth	.001893	.001355	14.611	0.162	-.0007629	.0045488
Weight	-.0001165	.00014	-0.83	0.405	-.0003908	.0001578
Length	-.0397358	.015112	-2.63	0.009	-.0693549	-.0101168
Age	.211129	.0540591	3.91	0.000	.1051752	.3170828
Area	.3065592	.0655921	24.563	0.000	.1780011	.4351173
_cons	44.318	.6164726	3.33	0.001	.8417357	3.258.264
inflate						
Depth	-.1295182	.0338827	-3.82	0.000	-.195927	-.0631093
Weight	.0004932	.0005902	0.84	0.403	-.0006637	.0016501
Length	-.0207498	.0739934	-0.28	0.779	-.1657742	.1242746
Age	-.1258854	.2887396	-0.44	0.663	-.6918047	.4400338
Area	125.639	.4016089	41.334	0.002	.4692512	2.043.529
_cons	1.050.094	3.156.974	12.114	0.001	431.338	1.668.849
/lnalpha	1.449.799	.0608788	23.81	0.000	1.330.478	1.569.119
alpha	4.262.257	.2594811			3.782.853	4.802.416

Vuong test of zinb vs. standard negative binomial: z = 6.54 Pr>z = 0.0000

Again in this model, the significance of the coefficient values changed. The Vuong test compares the ZINB model with a classical NB model. Significance of the Z test (z = 6254 Pr> z = 0.0000) indicates that the ZINB model is better.

4.5. Hurdle Regression

4.5.1 Poisson logit Hurdle regression (PH)

One of the models used when there are too many zeros in the observation values is the PH regression model. Care should be taken in interpreting these models because λ is not the expected result, but the mean of a fundamental distribution containing zeros. The results obtained for the PH model are given in Table 6.

Table 6. Poisson Logit Hurdle Regression Analysis

Poisson-Logit Hurdle Regression Log likelihood = -7155.9229				Number of obs = 1,191 Wald chi2(5) = 83.50 Prob > chi2 = 0.0000		
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
logit						
Depth	-.0066536	.0009462	-7.03	0.000	-.0085081	-.0047992
Weight	.0004247	.0001237	3.43	0.001	.0001822	.0006672
Length	-.0286749	.0124465	-2.30	0.021	-.0530695	-.0042803
Age	-.1228605	.0503524	-2.44	0.015	-.2215493	-.0241716
Area	-.0598937	.0627175	-0.95	0.340	-.1828177	.0630303
_cons	2.838.815	.4763591	5.96	0.000	1.905.168	3.772.462
Poisson						
Depth	.0013485	.0002057	6.56	0.000	.0009454	.0017516
Weight	.0001815	.0000321	5.65	0.000	.0001186	.0002445
Length	-.0585348	.0027151	-21.56	0.000	-.0638564	-.0532132
Age	.0839956	.0113305	7.41	0.000	.0617882	.1062031
Area	.316694	.0126652	25.01	0.000	.2918707	.3415173
_cons	3.755.515	.0979291	38.35	0.000	3.563.577	3.947.452

4.5.2 Negative binomial logit Hurdle regression (NBH)

Since there was overdispersion in the observation values, the analysis was done with the negative binomial Hurdle model.

Table 7. Negative Binomial-Logit Hurdle Regression

Negative Binomial-Logit Hurdle Regression Log likelihood = -2513.7673				Number of obs = 1,191 Wald chi2(5) = 83.50 Prob > chi2 = 0.0000		
	Coef.	**Std. Err.**	**z**	**P>z**	**[95% Conf.**	**Interval]**
logit						
Depth	-.0066536	.0009462	-7.03	0.000	-.008508	-.0047992
Weight	.0004247	.0001237	3.43	0.001	.0001822	.0006672
Length	-.0286749	.0124465	-2.30	0.021	-.0530695	-.0042803
Age	-.1228604	.0503524	-2.44	0.015	-.2215493	-.0241716
Area	-.0598937	.0627175	-0.95	0.340	-.1828177	.0630303
_cons	2.838.814	.476359	5.96	0.000	1.905.168	3.772.461
neg binomial						
Depth	.0023168	.0014701	1.58	0.115	-.0005647	.0051982
Weight	.0001738	.0001714	1.01	0.311	-.0001621	.0005098
Length	-.0768143	.018983	-4.05	0.000	-.1140202	-.0396084
Age	.2018122	.0639428	42.430	0.002	.0764865	.3271378
Area	.2809381	.0733809	3.83	0.000	.1371142	.424762
_cons	3.402.976	.7766218	4.38	0.000	1.880.825	4.925.127
/lnalpha	1.670.232	.2715551	6.15	0.000	1.137.994	220.247

The variables "Area" in the logit part of the model, "Depth" and "Weight" in the negative binomial part are insignificant while the other variables are significant.

5. CONCLUSION AND SUGGESTIONS

Commonly used models in count data are PR and NB models. The applicability of PR to the data obtained based on the count depends on the fact that the mean and variances of the data set are equal. A greater than average variance indicates overdispersion in the data set. In this case, different count data models are used. Among these, NB is the most preferred model.

In the count data, the dependent variable also takes the value zero. In this case, analyzes can be made by determining inflate variables. Zero dispersion occurs when there are more than expected zero values in the data set. Count data with zero inflated and (or) overdispersion is common in a wide variety of disciplines. In case of zero inflated, it is appropriate to use ZIP, ZINB, PH, NBH or generalized models. In count models, the distribution parameter is used to see if there is overdispersion. In addition, the Vuong test is applied to compare non-nested models. In the model selection, according to the Chi-square (χ^2) distribution with one degree of freedom table value, the model with the largest LL and the smallest AIC and BIC values is determined as the best model. In the current study, 6 different models were tested on the sample data set and a comparison was made in terms of LL, AIC and BIC values. Among these models, the smallest AIC and BIC and the largest LL values were found for the ZINB model. Table 8 gives the results collectively.

Table 8. Information Criteria for Models

Count Models	LL	AIC	BIC
PR	-11316.054	22644.11	22674.6
NB	-2543.4871	5100.974	5136.552
ZIP	-7157.201	14338.4	14399.39
ZINB	**-2491.515**	**5009.029**	**5075.102**
PH	-7155.9229	14335.85	14396.84
NBH	-2513.7673	5053.535	5119.608

When PR and NB distributions are compared, NB distribution gives smaller AIC, BIC, which is an expected result. In this study, among 1191 observations, 540(45.34%) observations consist of positive values and 651(54.66%) observations consist of zeros. Therefore, analyzes were obtained for zero-inflated models. When Hurdle model and zero inflated models are compared, it is seen that better results are obtained for zero-inflated models.

As a result, while building a model, we must consider all other alternative methods including the simpler count models such as PR and NB models. In terms of the results obtained, the goodness of criteria, the Voung statistics, and LL tests were parallel to each other. We concluded that ZIP model is superior to the standard PR model and ZINB model is superior to NB in this study. Studies in the literature support this result. Results also showed that estimated regression coefficients and standard errors differed across different models. However, it is more reasonable to say that which model is the best for the data depends on the data structure. Also statistical software packages have recently developed a procedure to fit zero-inflated models.

Contribution of the Authors
The contributions of the authors to the article are equal. In this study, Öznur İŞÇİ GÜNERİ contributed to the creation of the idea for the article, conducting the necessary research and examination, analysis, interpretation, and writing the article. Burcu DURMUŞ contributed to the research, data collection, literature review, and the creation of graphics and figures. Aynur İNCEKIRIK contributed to writing the formulas, interpretation, development of references, and language of the article.

Thank
Assisting with the R program in the study, we would like to thank Assoc. Prof. Dr. Nevin GÜLER DİNÇER.

Conflict of Interest Statement
The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript, and there is no financial interest to report.

Statement of Research and Publication Ethics
Research and publication ethics have complied in this study.

REFERENCES

Akaike, H. (1973). *Information theory and an extension of the Maximum Likelihood Principle*. Second International Symposium on Information Theory. Budapest, Academiai Kiado, 267-281.

Allison, P. (2012). Do we really need zero-inflated models? Retrieved April 15, 2021 from https://statisticalhorizons.com/zero-inflated-models

Altun, E. (2018). A new zero-inflated regression model with application. *Journal of Statisticians: Statistics and Actuarial Sciences*. 11(2), 73-80.

Altun, E. (2019). A new model for over-dispersed count data: Poisson Quasi-Lindley regression model. *Mathematical Sciences.* 13(3). 241-247.

Baetschmann, G. & Winkelmann, R. (2017). "A dynamic Hurdle model for zero-inflated count data. *Communications in Statistics-Theory and Methods*. 46(14), 7174-7187.

Boucher, J.P. & Denuit, M. (2008). Credibility premiums for the zero-inflated Poisson model and new hunger for bonus interpretation. *Insurance: Mathematics and Economics*. 42, 727-735.

Cameron, A.C. & Trivedi, P.K. (2013). Regression analysis of count data, Econometric society monograph (2nd Edition). *Cambridge University Press*. England.

Cragg, J.G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*. 829-844.

Cui, Y. & Yang, W. (2009). Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of Theoretical Biology*. 256, 276-285.

Dalrymple, M.L., Hudson, I.L. & Ford, R.P.K. (2003). Finite mixture, zero-inflated Poisson and Hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41, 491-504.

Deniz, Ö. (2005). Poisson regresyon, *İstanbul Commerce University Journal of Science.* 4(7), 59-72.

Denuit, M., Marećhal, X., Pitrebois, S. & Walhin, J.F. (2007). Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems. *John Wiley & Sons.*

Erdemir, Ö.K. & Karadağ, Ö. (2020). On comparison of models for count data with excessive zeros in non-life insurance. *Sigma Journal of Engineering and Natural Sciences*, 38(3), 1543-1553.

Famoye, F. & Singh, K.P. (2006). Zero-inated generalized Poisson regression model with an application to domestic violence data", *Journal of Data Science*, 4(1), 117-130.

Flynn, M. (2009). More flexible GLMs zero-inflated models and hybrid models. Casualty Actuarial Society E-Forum, 148-224, Retrieved June 12, 2021 from https://www.casact.org/pubs/forum/09wforum/flynn_francis.pdf

Gerdtham, U.G. (1997). Equity in health care utilization: further tests based on Hurdle models and Swedish micro data. *Health Economics*, 6, 303-319.

Greene, W. (2005). Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*. 1(2), 113-218.

Greene, W.H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *NYU Working Paper*. No. EC-94-10.

Hemmingsen, W., Jansen, P.A. & MacKenzie, K. (2005). Crabs, leeches and trypanosomes: an unholy trinity? *Marine Pollution Bulletin*, 50(3), 336-339.

Hofstetter, H., Dusseldorp, E., Zeileis, A. & Schuller, A.A. (2016). Modeling caries experience: advantages of the use of the Hurdle model. *Caries Research*, 50(6), 517–26.

Hurvich, C.M. & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

James, D. (2014). What is the difference between zero-inflated and Hurdle models? Retrieved April 16, 2021 from https://stats.stackexchange.com/questions/81457/what-is-the-difference-between-zero-inflated-and-hurdle-models

Karen, C.H.Y. & Kelvin, K.W.Y. (2005). On modeling claim frequency data in general insurance with extra zeros. *Mathematics and Economics*. 36, 153-163.

Khoshgoftaar, T.M., Gao, K. & Szabo, R.M. (2005). Comparing software fault predictions of pure and zero-inflated Poisson regression models. *International Journal of Systems Science.* 36(11), 707-715.

Kibar, F.T. (2008). *Trafik kazaları ve Trabzon bölünmüş sahil yolu örneğinde kaza tahmin modelinin oluşturulması* [Master Thesis]. Karadeniz Technical University Graduate Institute of Natural and Applied Sciences. Trabzon.

Lambert, D. (1992). Zero-inated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 34(1), 1-14.

Lee, Y., Moudud, A., Noh, M., Rönnegård, L. & Skarin, A. (2016). Spatial modeling of data with excessive zeros applied to Reindeer Pellet-group counts. *Ecology and Evolution*. 6, 7047–7056

Mamun, A. (2014). *Zero-inflated regression models for count data: an application to under-5 deaths* [Master Thesis]. Ball State University Muncie. Indiana.

McCullagh, P. & Nelder, J.A. (1989). Generalized Linear Models (Second Edition). *Chapman and Hall*, New York, USA.

McQuarrie, A.D.R. & Tsai, C. (1998). Regression and time series model selection. *World Scientific Publishing Company.* Singapore.

Mouatassim, Y. & Ezzahid, E.H. (2012). Poisson regression and zero-inflated Poisson regression: Application to private health insurance data. *European Actuarial Journal*. 2(2), 187-204.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics,* 33(3), 341-365.

Mwalili, S.M., Lesaffre, E. & Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical Methods in Medical Research*, 17(2), 123-139.

NNCS Statistical Software (2020). Negative binomial regression, Chapter 326. Retrieved April 15, 2021 from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

Peng, J. (2013). *Count data models for injury data from the national health interview survey* [Master Thesis]. *The Ohio State University Graduate Program in Public Health*, Columbus.

Ridout, M., Demetrio, C.G.B. & Hinde, J. (1998). *Models for count data with many zeros.* International Biometric Conference, Cape Town, Retrieved June 12, 2021 from https://www.kent.ac.uk/smsas/personal/msr/webfiles/zip/ibc_fin.pdf.

Ridout, M., Hinde, J. & Demetrio, C.G.B. (2001). A score test for a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57, 219-233.

Rose, C.E, Martin, S.W., Wannemuehler, K.A. & Plikaytis, B.D. (2006). On the of zero-inflated and Hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, 16, 463-481.

Sakthivel, K.M. & Rajitha, C.S. (2017). A comparative study of zero-inflated, Hurdle models with artificial neural network in claim count modeling. *International Journal of Statistics and Systems.* 12(2), 265-276.

Sarul, L.S. & Şahin, S. (2015). An application of claim frequency data using zero inflated and Hurdle models in general insurance. *Journal of Business, Economics and Finance*, 4(4), 732-743.

Sellers, K.F. & Shmueli, G. (2010). A flexible regression model for count data. *Annals Applied Statistics*. 4(2), 943-961.

Shalabh, K. (2020). Poisson regression models, Chapter 15. Retrieved April 12, 2021 from http://home.iitk.ac.in/~shalab/regression/Chapter15-Regression-PoissonRegressionModels.pdf

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S. & Boatwright, P. (2005). A useful dis-tribution for fitting discrete data: revival of the conway-maxwell-Poisson distribution. *Applied Statistics*, 54, 127–142.

Sinharay, S. (2010). Discrete probability distributions, ETS. *Elsevier.* Princeton, NJ, USA.

Sugiura, N. (1978). Further analysts of the data by Akaike's Information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*. 7(1),13-26.

Ver Hoef, J.M. & Frost, K.J. (2003). A bayesian hierarchical model for monitoring Harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics*, 10, 201–219.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 57, 307-333.

Wang, W. & Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression, *Journal of Population Economics.*10, 273-283.

Workie, M.S. & Gedef, A.A. (2021). Bayesian zero-inflated regression model with application to under-five child mortality. *Journal of Big Data*. 8(4), 1-23.

Yang, J., Li, X. & Liu, G.F. (2012). Analysis of zero-inflated count data from clinical trials with potential dropouts", *Statistics in Biopharmaceutical Research*. 4(3), 273-283.

Yau, K.K.W. & Lee, A.H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, 20, 2907-2920.

Yau, Z. (2002). *Score tests for generalization and zore-inflation in count data modeling* [Unpublished Ph.D. Thesis]. University of South Caroline. Columbia.

Yeşilova, A., Kaydan, M.B. & Kaya, Y. (2010). Modeling insect-egg data with excess zeros using zero-inflated regression models. *Hacettepe Journal of Mathematics and Statistics*, 39(2), 273-282.

Yip, K.C. & Yau, K.K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2), 153-163.

Zwilling, M. L. (2013). Negative binomial regression. *The Mathematica Journal, Wolfram Media*. 15, 1-18.