Statistical Models for Analyzing Count Data

Felix Boakye Oppong, Edward ChongsiMbukam, Anthony Adomah Agyapong

Abstract – Countdata is encountered in different forms. This include count data without zeros, count data with excessive number of zeros, counts with large observations, and many others. Different statistical models are used in analyzing these forms of count data. This paper provides an overview of the model frameworks and possible selection criteria that are appropriate for analyzing the various forms of count data.

Keywords – Hurdle model, Negative binomial regression, Overdispersion, Poisson Regression, Quasi-Poisson model, zero-inflated models, zero-truncated models.

1 INTRODUCTION

Count data is encountered in almost all areas of research including, economics, medicine, management, industrial organizations and many more [1]. Examples of count data include the number of road accidents on a particular high way in a specified period of time, the number of insurance claims paid by an insurance company in a year, the total number of epileptic seizures in a week, number of defective chips in a batch of manufactured computer chips, etc. The most common approach used in modeling this type of data is the Poisson regression. However, due to overdispersion (extra variability) associated with the use of the Poisson model, practitioners routinely make use of the negative binomial model as an alternative [2].

Although Poisson and negative binomial models are the building blocks for count data, there are a number of extensions to these models that accommodate special features of the available data. Some of these extensions include hurdle effects, zero inflation, zero truncation, sample selection and severalothers [3], [4], [5]. In a similar fashion, generalized linear mixed models (GLMM) extend the Poisson and negative binomial models for the analysis of longitudinal count data [6]. Moreover, several proposals have been made to extend the Poisson and negative binomial models to accommodate multivariate data[7]. This paper explores the Poisson and negative binomial models, with extensions including the zero-inflated, hurdle and zero-truncated models. Different datasets are considered to illustrate the use of these model extensions in practice.

The organization of the paper is as follows. Section 2 introduces the Poisson and negative binomial models, as well as the quasi Poisson model. Extensions of the Poisson and negative binomial models are discussed with practical examples in Sections 3. Some remarks on software is presented in Section 4. Finally, Section 5 provides concluding remarks to the entire write-up.

2 POISSON AND NEGATIVE BINOMIAL MODEL

2.1 Poisson Regression

By definition, the number of events that occur in a given period of time follows a Poisson distribution [8]. A random variable Y is said to have a Poisson distribution with parameter μ, if it takes integer values y = 0, 1, 2, 3,.... That is, it is assume that the response variable y_i is a count, which can take values 0, 1, 2, ….. A probability model for the Poisson distribution can be written as:

$$f(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ... \quad \mu > 0,$$

The mean and the variance of the Poisson distribution is shown to be $E(\mathrm{Y}) = Var(\mathrm{Y}) = \mu$. The Poisson regression model is specified using the generalized linear model (GLM) notation as:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k = \boldsymbol{x}_k'\boldsymbol{\beta}$$

GLM extends linear regression model and allow for the modeling of data that follow other probability distributions aside the normal distribution [9]. In GLM, three components are distinguished namely, the random component, the systematic component and the link function. The random component specifies a probability distribution for the response variable (*Y*), the systematic component identifies the set of predictors used, and the link function species a function that maps E(Y) to the systematic component [10]. Here, g is the canonical link function and $g(\mu_i) = \eta_i$ is the mean response. The linear predictors can be related to the mean of the response as

$$\mu_i = g'(\eta_i) = g^{-1}(\boldsymbol{x}_k'\boldsymbol{\beta})$$

Although the identity link is sometimes used, for the Poisson regression, the log link is the most common link function. In using the log link, it guarantees that, all of the fitted values for the response variable will be positive. The method of maximum likelihood is used in the estimation of the parameters of the Poisson regression model. The specification

- *Felix Boakye OPPONG, graduate Biostatistician, Hasselt University, Belgium. E-mail: atomistic4u@gmail.com*
- *Edward Chongsi MBUKAM, graduate Biostatistician, Hasselt University, Belgium.*
- *Anthony Adomah AGYAPONG, masters of Biostatistics, Hasselt University, Belgium.*

of the (log) - likelihood and inference can be found in [10].

It is important to emphasize that, for the standard Poisson regression, the conditional mean of the outcome is equal to its conditional variance. However, in practice, there is often overdispersion in the sense that, the conditional variance exceeds the conditional mean. That is, overdispersion occurs when the observed counts have higher variability than that expected by the Poisson regression model [2]. Some of the consequences of overdispersion includes the underestimation of standard errors, which leads to wrongfully inflating the significance level.

2.2 Quasi-Poisson model

The issue of overdispersion can be addressed by using the so called quasi-Poisson model. With this approach, the standard Poisson regression model is adjusted to estimate an additional dispersion parameter. For the standard Poisson model, the mean variance relationship is specified as $var(Y) = \phi E(Y)$, and ϕ is constrained to be equal to one. However, in the presence of overdispersion, the value of ϕ is mostly greater than one. Although underdispersion is quite rare, it comes about when the variance is smaller than the mean. In that case, the value of ϕ is less than one [2]. With the quasi-Poisson model as opposed to the standard Poisson model, ϕ is not constrained to be equal to 1, but is estimated from the data. Consequently, the parameter estimates of the quasi-Poisson model is the same as that of the standard Poisson model. However, their standard errors are adjusted (increased), which in effect affects their significance level. This is achieved by multiplying the covariance matrix of the parameters by the value of the overdispersion parameter ϕ [10].

2.3 Negative Binomial Regression

Aside the use of the quasi-Poisson model in accounting for overdispersion, a much more formal alternative is to use a negative binomial model [11]. The negative binomial distribution is quite similar to the Poisson distribution, however, unlike the Poisson distribution, the variance of the negative binomial distribution exceeds its mean. As such, if the variance of the observed outcome is suspected to be larger than its mean, the negative binomial distribution is more suitable compared to the Poisson distribution. Using the notation in [12], the negative binomial distribution can be parameterized as a mixture of Poisson-gamma as follows:

$$f(y_i;\mu_i,\alpha) = \begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, \alpha, \mu_i > 0$$

With this parameterization, μ is the mean of Y, whereas α is the heterogeneity (overdispersion) parameter. Unlike the Poisson distribution whose variance is as well equal to μ, the variance of the negative binomial distribution is $\mu + \alpha\mu^2$. As such, the negative binomial distribution is over-dispersed compared to the Poisson distribution. The presence of overdispersion results in values of α greater than zero [2]. However, as $\alpha \to 0$, the negative binomial distribution converges to the Poisson distribution. This form of parameterization of the negative binomial distribution is called the NB2 (NB for negative binomial). Basically, two different forms of the negative binomial models namely, NB1 and NB2 are in use. However, the NB2 model is often the preferred choice since, it reduces to the Poisson distribution when α is 0 [13]. Hilbe [13] provides the detail specification of the (log)-likelihood as well as inference for the negative binomial model.

2.4 Illustration 1, Horseshoe crab data

The data from a study of nesting horseshoe crabs [2], is used as the first illustrative data. The study investigated factors that affects the number of male crabs (called satellites), residingnear a female crab's nest. Several variables were used as explanatory variable, however, we consider only width of the female crab as predictor of the number of satellites attached to her in her nest. First, a Poisson regression is fitted to the data, followed by the quasi Poisson and negative binomial models.

The parameter estimates of the fitted models together with relevant output are presented in Table 1. As expected, the parameter estimates for the Poisson and quasi Poisson are the same, however, their standard errors differ. In checking the goodness of fit of the Poisson model, overdispersion is found to be present in the horseshoe crab data. This can be seen in the underestimation of the standard errors of the Poisson model, as compared to that of the quasi Poisson and the negative binomial models. Aside comparing the standard errors, dispersion can also be measured by using the scaled deviance and/or the scaled Pearson chi-square. When the values of these statistics are much larger than one (1), the Poisson mean variance relationship may not be valid and the data is said to be overdispersed.

Although all three models lead to the same conclusion, that is, a strong positive width effect on the number of satellites, in some analyses, the results of the Poisson model might differ from that of the quasi Poisson and/or negative binomial models.

Table 1. Parameter estimates for the Poisson, quasi Poisson and the negative binomial models.

		Poisson, log link			quasi-Poisson			negative binomial		
Effect		estimate	s.e	*p*-value	estimate	s.e	*p*-value	estimate	s.e	*p*-value
intercept		-3.30	0.54	< 0.0001	-3.30	0.99	0.0008	-4.0525	1.3528	0.0027
width		0.16	0.02	< 0.0001	0.16	0.04	< 0.0001	0.1921	0.0510	0.0002

s.e: standard error

3 EXTENSIONS OF POISSON AND NEGATIVE BINOMIAL MODELS

3.1 Zero inflated Poisson/ negative binomial models

In some datasets, the amount of observations that have a value equal to zero is higher than it would be expected in a Poisson model. For such datasets, the zero-inflated Poisson (ZIP) model [4] is appropriate. ZIP assumes that, the data comes from two sub-populations namely, the excess/inflated zero population, and the population consistent with the Poisson distribution. With this, the response takes the value zero (0) with probability of p and has a Poisson distribution with probability of *1-p.* Given that y_i follows a ZIP distribution with parameters p and μ, its probability distribution can be specified as

$$p(y_i \mid \mu_i, p_i) = \begin{cases} p_i + (1-p_i)e^{-\mu_i} & \text{if } y_i = 0 \\ (1-p_i)\dfrac{\mu_i^{y_i}}{y_i!}e^{-\mu_i} & \text{if } y_i = 1,2,3,\ldots \end{cases}$$

such that $0 \le p_i \le 1$, $\mu_i > 0$. μ_i is the expected Poisson count for the *i*th individual, and p_i is the probability of extra zeros. The mean and variance of theZIP are $(1-p_i)\mu_i$ and $(1-p_i)(1+p_i\mu_i)\mu_i$ respectively. In the absence of extra zeros, in which case $p_i = 0$, the ZIP reduces to the Poisson model with both mean and variance equal to μ_i. In fitting the ZIP model, a logistic regression is used to model the probability of zero counts, followed by a standard Poisson model for the non-zero counts. The data on horseshoe crab is ones again considered for the illustration of the ZIP model. The histogram of the distribution of the number of satellites attached to a female crab is provided in Fig 1. Most of the crabs have zero satellites attached to them in their nest. That is, out of the 173 crabs, 62 (approximately 36%) had nosatellites around them in their nest.Therefore, a ZIP model is fitted, to study whether the width of a female crab has an effects on the number of satellites staying near her nest. Also, we will use the ZIP model to find out if having no satellite around a female crab's nest is dependent on the crab's width.

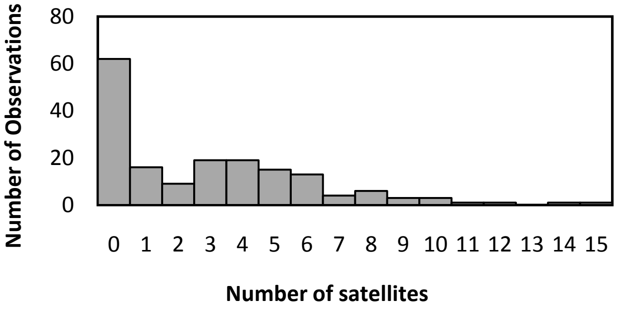

Fig 1. Number of satellites attached to a female crab

The parameter estimates of the resulting model together with other relevant output are presented in Table 2. Contrary to the results of the previous models namely, Poisson, quasi Poisson and negative binomial, the width of a female crab has no effect on the number of satellites attached to it. However, from the binomial model for the inflated zeros, the higher the width of the female crab, the lower the odds of having zero (0) satellite attach to it.In other words, if a crab was to increase its width by 1cm, the odds that it will have no crab attached to her in its nest would decrease by a factor of exp(-0.5010) = 0.606.

Table 2. Parameter estimates, standard errors and p-values for the zero inflated Poisson model

	Poisson with log link			binomial with logit link		
Effect	Estimate	s.e	*p* -value	Estimate	s.e	*p* – value
(Intercept)	0.574	0.593	0.333	12.423	2.693	< 0.0001
Width	0.035	0.022	0.114	-0.502	0.104	< 0.0001

s.e: standard error

Since the Poisson, quasi Poisson and negative binomial models fail to account for the inflated zeros, their results show that crabs with higher widths have more satellites around them. However, from the ZIP model, having higher width decreases the odds of having no satellites attached to the female crab. A straightforward question to ask is, which of the models namely, standard Poisson, quasi Poisson, negative binomial or ZIP model do we consider 'best'? We can use either AIC or BIC to answer this question. AIC and/or BIC are/is used for comparing non-nested models based on their maximum likelihood. Oppong [9] presents the mathematical formulation of AIC/BIC, together with their basic properties. Here, we presents the results from using AIC as the criterion for finding the "best" model. It should however be emphasized that, using BIC as the selection criterion leads to the selection of the same model. The AIC for the standard Poisson, negative Binomial and ZIP models are respectively 927.2, 757.3 and 737.6. The AIC of the quasi Poison model (QAIC - quasi AIC) is not provided since its validity in comparison to AIC has not been established [14]. QAICs have been developed only to be used within the quasi class of models and not between quasi models and models with distributional forms [15]. From the reported AICs, the ZIP model is the best fitting model. It captures the excess zeros present in the analyzed data.

When one suspects overdispersion together with excess zeros, the zero inflated negative binomial model is the preferred choice of model. For the analysis of the horseshoe crab data, in comparing the quasi Poisson and negative binomial models to the standard Poisson model, overdispersion was found to be present. Further, comparing the fit of the ZIP model (which accounts for the excess zeros) to the negative binomial model using their AIC's, we can conclude the existence of excess zeros in the horseshoe crab data. As such, we fit a zero-inflated negative binomial model to the data, to account for both overdispersion and excessive zeros. Although the parameter estimates and standard errors for the zero-inflated negative binomial model are not presented, they lead to the same conclusion as the zero-inflated Poisson model. However, the AIC of the zero-inflated negative binomial model is smaller than that of the ZIP model: 716.63 as against 737.64. Consequently, the fit of the zero-inflated negative binomial model surpasses that of the ZIP model.

3.2 Hurdle Poisson/ negative binomial model

Hurdle models, assuming either Poisson or negative binomial distributions are similar to, but different from their zero inflated counterparts. They are similar in the sense that, they have both been developed to account for zero-inflated outcome data with overdispersion (negative binomial) or without overdispersion (Poisson distribution). However, these two model frameworks differ in terms of their analysis and interpretation of the zero counts [8].

Zero-inflated Poisson/negative binomial splits the zero data into the genuine Poisson/ negative binomial zero data and the zeros which are due to the structure of the data. As an example, consider a data on the number of sexually transmitted diseases (STDs) contracted by a group of individuals living in an STD prone locality, in a period of 10 years. Genuinely, some of the participants may not contract STD within the study period. This represents the genuine zero data. On the other hand, some individuals may have no STD count within the 10 years period because they do not have sex. These zeros are observed as a results of the structure of the data. That is, all things been equal, these participants cannot have STDs.

On the contrary, hurdle model does not split the zero data. It rather considers all the zero data to emanate from one structure. That is, with hurdle models, it is assumed that zeros are generated by a different process other than that generated by the positive counts. In this sense, the zeros act as hurdles that one needs to cross before getting to the positive counts. To illustrate this, consider a data that records the number of hours that people use their computers in a month. Obviously, people who do not use computers will records zero (0) hours of computer use. Conversely, all things been equal, computer users cannot record zero computer use in one month.Ideally, a hurdle model combines a count data model, $f(y;x,\beta)$, for the non-zero counts and a zero hurdle model, $f(y;z,\tau)$, for the zero counts. For modelling the non-zero counts, the Poisson or negative binomial distribution can be used. On the other hand, a binomial model or a censored count distribution can be used for the zero hurdle model [11]. The model can be specified in terms of β and τ as:

$$f(\text{y};x,z,\beta,\tau)=\begin{cases} f(y=0;z,\tau) & \text{if } y=0 \\ (1-f(y=0;z,\tau))\dfrac{f(y>0;x,\beta)}{1-f(y=0;x,\beta)} & \text{if } y>0 \end{cases}$$

The parameters β and τ can be estimated by maximum likelihood. Given μ_i as the expected number of events for the *i*th subject, the mean regression relationship is written using the canonical log link as

$$\log(\mu_i)=\boldsymbol{x}_i'\boldsymbol{\beta}+\log(1-f(y=0;z,\tau))-\log(1-f(y=0;x,\beta))$$

Depending on the nature of the data at hand, one can opt for either the zero-inflated model or the hurdle model. It should be emphasized that, these two models have different interpretations and can lead to different results. Evidently, considering the horseshoe crab data, the zero-inflated model is the best model choice, in the sense that, a female crab can possibly (and more likely) have zero satellites attached to it in its nest. That is, the process generating the non-zero outcomes is the same process generating the zero counts.

3.3 Zero truncated Poisson/ negative binomial models

If the distribution of*Y*is Poisson but cannot take zero as an outcome, the zero-truncated Poisson regression model is used [16]. To be more precise, the zero-truncated Poisson regression is used to model count data for which the values (response) cannot be zero. This kind of data is particularly common in studies where individuals become part of the sample only after the first count has been observed. An example is data that records the number of times patients in a malaria clinic contract malaria in a five year period. Obviously, all patients in a malaria clinic should have contacted malaria. Hence, the least possible outcome is one malaria case. With this kind of data, zero counts cannot be observed for any of the patients in the sample, hence the data is said to be truncated at zero. It is worth mentioning that, truncation can occur at any value. However, zero-truncation is a very common occurrence in practice, and require attention. Cameron and Trivedi [11] emphasized that suitable modification should be made to the likelihood function of the Poisson/ negative binomial likelihood, to prevent inconsistent parameter estimates associated with truncation. The density of the zero-truncated Poisson is specified as:

$$f(\text{y}_i\mid\mu_i,\text{y}_i>0)=\frac{f(\text{y}_i\mid\mu_i)}{1-F(0\mid\mu_i)}=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-\text{e}^{-\mu_i})},\; y_i=1,2,3,\ldots$$

Construction of (log)-likelihood and obtaining maximum likelihood estimates of the specified density is straightforward. Similar to the standard Poisson model, the zero-truncated Poisson model also suffers from the adverse effects of over-dispersion. In the presence of overdispersion, the zero-truncated Poisson model leads to biased and inefficient parameter estimates [17]. To address this, the zero-truncated negative binomial is considered.

With overdispersed count data for which zero cannot be a possible outcome, the zero-truncated negative binomial model is the best choice [16]. The density of the zero-truncated negative binomial distribution can be written as

$$f(\mathrm{y}_i \mid \mu_i, \alpha, \mathrm{y}_i > 0) = \frac{f(\mathrm{y}_i \mid \mu_i, \alpha)}{1 - F(0 \mid \mu_i, \alpha)}$$

$$= \begin{pmatrix} y_i + \dfrac{1}{\alpha} - 1 \\ \dfrac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i} \right)^{y_i} \frac{1}{1 - \left(\dfrac{1}{1+\alpha\mu_i} \right)^{\frac{1}{\alpha}}}$$

Long and Freese [17] provide details on model specification and inferences thereafter.

3.4 Illustration 2, Malaria data.

Data on the number of malaria cases recorded in a group of children treated in a malaria clinic from 2006 to 2010 is considered. The data records the number of times each of 174 children contracted malaria in the five year period. For illustration purpose, we restrict ours analysis to the 127 children with complete records at the end of the study period. The malaria clinic is meant mainly for the treatment of malaria cases. As such, patients who visit this clinic have recorded at least one malaria case. This is a perfect example of a zero truncated count data – it is not possible for any of the study participants to record a zero malaria case. Aside the response, which is the number of malaria cases recorded in five years, the data records the gender and whether or not a child lives in a residential area, as possible predictors.

We first fit a zero-truncated Poisson model to the data and then, compare its fit to a zero-truncated negative binomial model. Likewise, the data will be studied for the presence/absence of overdispersion. Although the identity link is sometimes adequate for the zero-truncated negative binomial model, just like with the Poisson GLM, the log link is used. Given μ_i as the mean malaria incidence for the *i*th child, the zero-truncated Poisson/negative binomial regression model can be specified in terms of gender and residency as

$$\boldsymbol{log(\mu_i) = \beta_0 + \beta_1 \mathbf{gender}_i + \beta_2 \mathbf{Residency}}$$

Both **gender** and **Residency** is coded as 0, 1. 1 corresponds to females and children living in non-residential area for gender and residency respectively. The results from fitting both models are presented in Table 3. As noted earlier, the negative binomialdistribution converges to the Poisson distribution as $\alpha \to 0$. The presence of overdispersion is observed by the significant effect of α, as presented in Table 3. This implies that, α, is significantly different from zero. Hence, the fit of the zero-truncated negative binomial model is preferred to that of the zero-truncated Poisson model, which does not account for overdispersion.

Table 3. Parameter estimates for the zero truncated Poisson and negative binomial models

	Zero-truncated Poisson			Zero-truncated negative binomial		
Effect	Estimate	Standard Error	*p* - value	Estimate	Standard Error	*p* - value
(Intercept)	0.6586	0.1734	0.00015	0.5507	0.2181	0.0116
Female	-0.1410	0.0969	0.1458	-0.1593	0.1378	0.2476
Non-residency	0.8447	0.1742	<0.0001	0.9169	0.2167	<0.0001
α	-	-	-	1.4561	0.3741	<0.0001

Is it necessary to worry about the fact that zeros are not observed as possible counts in the data? Why not fit just a standard Poisson or a standard negative binomial model? To answer these questions, we fit a standard Poisson and a negative binomial model to the data and compare their AICs to that of the zero-truncated Poisson and the zero-truncated negative binomial models. The AIC are 564.73, 553.76, 553.70 and 533.51 respectively for the standard Poisson, negative binomial, zero-truncated Poisson and the zero-truncated negative binomial models. Evidently, it is necessary to account for the truncated zeros in the analysis. The zero-truncated negative binomial model fits the data well, compared to the other models. From the reported AICs, the standard negative binomial model is comparable to the zero-truncated Poisson model. This might in part be due to the overdispersion that is not accounted for by the zero-truncated Poisson model. After accounting for overdispersion using the zero-truncated negative binomial model, the fit is drastically improved. As noted in [9], a difference in AIC of more than 10 is often considered a strong evidence in selecting the model with smaller AIC. Hence, among all the fitted model, the zero-truncated negative binomial model is considered the 'best' fitting model.

4 SOFTWARE

Several software packages are available for fitting models to count data. In this paper, we restrict attention to **SAS, R and Stata**. In fitting a standard Poisson regression model, the

GENMODprocedure in **SAS**, the **glm** function in **R**and the**poisson** command in**Stata**can be used. For the standard negative binomial model, the **SASGENMOD** procedure, the **glm nb**(negative binomial) function in the **MASS** library of **R**, and the **nbreg** command in **Stata** can be used. By specifying the distribution as 'zip', the **GENMOD** procedure in **SAS** can be used in fitting a zero-inflated Poisson model. Likewise, specifying the distribution as 'zinb' in the model statement of the **GENMOD** procedure in **SAS** fits a zero-inflated negative binomial model. The **pscl** package in **R** together with the **zeroinfl** function is required for fitting a zero-inflated Poisson model. Using the same **zeroinfl** function, and specifying the distribution as "negbin" fits the zero-inflated negative binomial model in **R**. The **'zip'** and **'zinb'** commands are used in **Stata** to fit the zero-inflated Poisson and zero-inflated negative binomial models respectively.

For fitting a hurdle Poisson/ negative binomial model, the **FMM** (finite mixture model) procedure in **SAS** can be utilized. In **R**, the **pscl** package together with the **hurdle** function is used in fitting a hurdle model. Currently, **Stata** does not have a specific command for fitting a hurdle model. On the other hand, several written commands are available to estimate the parameters of a hurdle model. McDowell [18] presents how to use a combination of existing commands to estimate the parameters of a hurdle model in**Stata.** Likewise,Hilbe [19] has written several commands in **Stata** for estimating the parameters of a hurdle model. These include the **hplogit** (hurdle poisson model) and **hnblogit** (hurdle negative binomial model) commands.

One can fit a zero - truncated Poisson/ negative binomial model in **SAS** with **PROC NLMIXED** by specifying its log likelihood function. Similarly, by stating that **DIST=TRUNCPOISSON** or **TRUNCNEGBIN**, the **FMM** procedure can be used to fit a zero - truncated Poisson or negative binomial model in **SAS**. In **R**, the **vglm** function in the **VGAM** package can be used to fit a zero – truncated Poisson/ negative binomial model. This function fits a flexible class of models called the vector generalized linear models, to a wide range of distributions [20]. Specifying family = 'pospoisson' and 'posnegbinomial' via the **vglm** function fits the zero – truncated Poisson and zero – truncated negative binomial models respectively. In **Stata**, the **ztp** command fits a zero-truncated Poisson model, whereas the **ztnb** or **tnbreg** fits a zero – truncated negative binomial model

5 CONCLUSION

In the analysis of count data, the choice of model needs careful consideration. These considerations should be based on several factors including the features of the available data. These features include overdispersed count data, data with excess zeros, data that cannot take zero outcome, etc. Several extension to the commonly used standard Poisson/ negative binomial models are available to accommodate these special features. These model extensions include, but are not limited to the quasi Poisson, zero inflated Poisson/ negative binomial, hurdle Poisson/ negative binomial and the zero truncated Poisson/ negative binomial models. With advancement in software/ tools for statistical data analysis, these model extensions can be easily fitted in **SAS, R, Stata** and with many other statistical software packages.

REFERENCES

[1] Greene, W.(2007). Functional form and heterogeneity in models for count data. *Foundations and Trends® in Econometrics*, 1(2), 113-218.

[2] Agresti A. (2007).*An Introduction to Categorical Data Analysis*. (2nd Edition). John Wiley & Sons, Hoboken, New Jersey. 2007.

[3] Rose, C.E., Martin, S.W., Wannemuehler, K.A. and Plikaytis, B.D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4), 463-481.

[4] Lambert, D.(1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34 (1), 1 – 14.

[5] Terza, J. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics*, 84(1), 129–154.

[6] Molenberghs, G and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

[7] Riphahn, R.T., Wambach, A. and Million, A. (2003). Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of applied econometrics*, 18(4), 387-405.

[8] Hu M.C., Pavlicova M., Nunes E.V. (2011). Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *Am J Drug Alcohol Abuse*,**37**(5), 367 - 375.

[9] Oppong, F. B. (2016). Other Distributions for a Continuous Response Aside the Normal Distribution in a Linear Regression Model. *Mathematical Theory and Modeling*. **6**(5), 75-80.

[10] Nelder, J. and Wedderburn R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A.*, **135**: 370-384.

[11] Cameron, A.C, Trivedi, P.K. (1998).*Regression Analysis of Count Data*. Cambridge University Press.

[12] Zwilling, M.L. (2013). Negative Binomial Regression. *The Mathematica Journal*, **15**: 1-18.

[13] Hilbe, J.M. (2011). Modeling Count Data. International Encyclopedia of Statistical Science, Springer Berlin Heidelberg, 836-839.

[14] Ver Hoef, J.M. and Boveng, P.L. (2007). QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA? *Ecology*, 88(11), 2766-2772.

[15] Burnham, K.P. and Anderson, D.R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

[16] Ridout, M., Demetrio, C.G and Hinde, J. (1998). Models for count data with many zeros. Proceedings of the XIXth international biometric conference. 19: 179-192.

[17] Long, J. S. and Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata.* (3rd Edition).Stata Press.

[18] McDowell, A.(2003). From the help desk: hurdle models. *The Stata Journal*, 3(2), 178-184.

[19] Hilbe, J. M. (2014). *Modeling count data.* Cambridge University Press.

[20] Yee, T.W.(2010). The VGAM package for categorical data analysis. *Journal of StatisticalSoftware*, 32(10), 1-34.