Science Journal of Applied Mathematics and Statistics
2016; 4(6): 256-262
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20160406.12
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Statistical Models for Count Data

Alexander Kasyoki Muoka[1], Oscar Owino Ngesa[2], Anthony Gichuhi Waititu[3]

[1]Department of Basic and Applied Sciences, Jomo Kenyatta University of Agriculture and Technology-Westlands campus, Nairobi, Kenya

[2]Mathematics and Informatics department, Taita Taveta University College, Voi, Kenya

[3]Department of Basic and Applied Sciences, Jomo Kenyatta University of Agriculture and Technology-Westlands campus, Nairobi, Kenya

Email address:

alexanderkasyoki@gmail.com (A. K. Muoka), oscanges@ttuc.ac.ke (O. O. Ngesa), agwaititu@gmail.com (A. G. Waititu)

To cite this article:

Alexander Kasyoki Muoka, Oscar Owino Ngesa, Anthony Gichuhi Waititu. Statistical Models for Count Data. *Science Journal of Applied Mathematics and Statistics.* Vol. 4, No. 6, 2016, pp. 256-262. doi: 10.11648/j.sjams.20160406.12

Received: September 13, 2016; **Accepted**: September 23, 2016; **Published**: October 15, 2016

Abstract: Statistical analyses involving count data may take several forms depending on the context of use, that is; simple counts such as the number of plants in a particular field and categorical data in which counts represent the number of items falling in each of the several categories. The mostly adapted model for analyzing count data is the Poisson model. Other models that can be considered for modeling count data are the negative binomial and the hurdle models. It is of great importance that these models are systematically considered and compared before choosing one at the expense of others to handle count data. In real world situations count data sets may have zero counts which have an importance attached to them. In this work, statistical simulation technique was used to compare the performance of these count data models. Count data sets with different proportions of zero were simulated. Akaike Information Criterion (AIC) was used in the simulation study to compare how well several count data models fit the simulated datasets. From the results of the study it was concluded that negative binomial model fits better to over-dispersed data which has below 0.3 proportion of zeros and that hurdle model performs better in data with 0.3 and above proportion of zero.

Keywords: Count, Modeling, Simulation, AIC, Compare

1. Introduction

Count data is encountered on daily basis and dealings. More understanding of such data and extraction of important information about the data needs some statistical analysis or modeling. Different count data may possess different characteristics and therefore cannot be used with particular count data models. Poisson regression model provides a basis for the analysis of count data. Many practitioners choose to use Poisson model when faced with data analysis involving count data even without ensuring that all assumptions of this model are met. The systematic way for choosing a model for fitting a particular data is that one should test whether the model's assumptions are met rather than just going the naive way of fitting a model. Cases arise when these assumptions are violated and therefore a need to go for an alternative model.

Some of the alternative models that can be considered for modeling count data are negative binomial and hurdle models [1, 2] A hurdle model is mixed by a binary outcome of the count being below or above the hurdle (the selection variable), with a truncated model for outcomes above the hurdle. One of the common assumptions has been that all count data follows Poisson distribution and therefore the mean and the variance are equal. However, this is not the case as the data may show some deviation from this assumption in having greater variance than the mean. Another case is whereby particular count data models can handle data with a particular amount of zeros and therefore cannot handle data with excess zeros. In some cases these zeros cannot be ignored because they are of great importance as they are meaningful.

Oftenly, in dealing with count data the mean is not equal to variance as assumed by Poisson. Over-dispersion may be caused by occurrence of many zero counts than a Poisson model would predict. These zeros cannot be deemed meaningless as they frequently have special status. For instance, in counting the number of plants affected by a particular disease in the field, a plant may not have the symptoms of the disease because it is resistant to the disease

or simply because the disease causing micro-organism has not landed on it. The hurdle models are based on Poisson regression and negative binomial regression respectively but is used for modeling excess zeros.

With these statistical models for handling count data, it is difficult to know which one to choose by just someone's intuitive feelings. Reference [3] proposed a comparative approach for handling count data by comparing five regression models on how they fitted their count data using the Akaike Information Criterion (AIC). They concluded that a specific model is not preferable to the other, but that one needs to choose model critically. With respect to how important it is to choose the best model and how many models and model specifications that exist one could presume that there are many comparisons of models in literature, but this is not the case [4]. Traditionally, the linear regression model, ordinary least squares (OLS) was used in modeling count data with the underlying assumption that the outcome of interest was normally distributed [5]. However, this assumption may not hold for some cases of count data even after applying some data transformation techniques.

Researchers and other practitioners may wish to study one or more sets of count data which have diverse characteristics to make important decisions or conclusions. In this case there could be difficulties in choosing one among the different models for handling count data. The main objective of this study was to investigate the performance of several count data models and inform their choice. Specifically, the study aimed at doing a review of models for count data and comparing them using simulated data of different characteristics.

2. Count Data Models Review

Counts are non-negative integers. They represent the number of occurrences of an event within a fixed period of time. In many economic and scientific contexts the dependent or the response variable of interest (y) is a count which we wish to analyze in terms of a set of covariates (x). Unlike in the case of a classical regression model, the response variable is a discrete with a distribution that places the probability mass at non-negative integer values only. Regression models for counts, like other limited or discrete variable models are nonlinear with many properties and special features intimately connected to discrete-ness and non-linearity [6]. Despite the fact that count data regression modeling techniques have rather recent origin, the statistical analysis of count data has a long history. Most of the early statistical count analyses concerned univariate independent and identically distributed random variables within the framework of discrete parametric distributions [7].

2.1. Basis for Count Data Models

The foundation for the development of count data models is the Poisson distribution. Most of the count data models belong to Generalized Linear Models.

Nearly all of count models have a basic structure as described by [8] in the equation:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

To isolate the predicted mean count on the left side of the above equation, we take exponential on both sides of the equation, giving

$$\mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}$$

Both of the above expressions are important in defining the terms in the count data models as it shall be discussed later. Reference [8] stated that an important feature of using the natural log link in the count data linear relationship model is that it guarantees that the predicted values will be positive, that is, $\mu > 0$. The several models for handling count data which are Poisson model, negative binomial model and hurdle models are as discussed in the following sections.

2.2. Poisson Model

The Poisson distribution is usually used as a standard model for count data and was derived as a limiting case of the binomial distribution by Poisson. It was the first model specifically used to model counts and it still stands at the base of many types of count models available to analysts. In Poisson modeling, it is assumed that the mean and the variance are equal. This makes it unsatisfactory to use Poisson model on real study data. A Poisson random variable has the probability distribution function (pdf)

$$f(y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad (1)$$

for $y_i = 0, 1, 2, \ldots$

The mean and the variance are

$$E(y_i) = \text{var}(y_i) = \mu_i$$

The expected value μ_i is a linear function of p predictors that take the values $\mathbf{X}_i' = (x_{i1,\ldots,} x_{ip})$ for the ith case so that

$$\mu_i = \mathbf{X_i'\beta}$$

Where $\boldsymbol{\beta}$ is a vector of parameters to be estimated.

In this case, we define the link function as $\eta_i = g(\mu_i)$ so that it is assumed that the transformed mean follows a linear model and we write

$$\eta_i = \mathbf{X_i'\beta}$$

Where η_i is known as the linear predictor.

From equation (1), it can be seen that the Poisson distribution belongs to the exponential family since it can be expressed in the probability density function form

$$f(y_i) = \exp\left(\frac{y_{i\theta_i} - b(\theta_i)}{a_i(\phi)} + c\,(y_i, \phi)\right) \quad (2)$$

Where θ_i and ϕ are location and scale parameters and $a_i(\phi)$, $b(\theta_i)$ and $c\ (y_i, \phi)$ are known functions. Moreover, if yi has a distribution in the exponential family then its variance and mean are

$$E(y_i) = \mu_i = b'(\theta_i) \tag{3}$$

$$\text{var } (y_i) = \sigma_i^2 = b''(\theta_i) / p_i \tag{4}$$

Where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$ respectively and pi is a known prior weight, usually 1. The relationship in equations (3) and (4) can be proved for Poisson distribution.

Poisson Parameter estimation

The parameters of Poisson model are estimated by maximum likelihood approach using an iteratively re-weighted least squares algorithm. From equation (2) the log-likelihood for the sample $y_1,\ldots y_n$ is

$$l = \sum_{i=1}^{n} \left(\frac{y_{i\theta_i} - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right) \tag{5}$$

The maximum likelihood estimates are then obtained by solving the score equations

$$s(\beta_j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\phi_i V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} \right) = 0 \tag{6}$$

For parameters β_j, where V (μ_i) is a variance function.

By assuming that

$$\phi_i = \frac{\phi}{a_i} \tag{7}$$

Where ϕ is a single dispersion parameter and a_i are known prior weights, the estimating equations can then be written as

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{a_i (y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} \right) = 0 \tag{8}$$

The score equation above is then solved by using Fisher's scoring iterative algorithm whereby in the r^{th} iteration, the new estimate $\beta^{(r+1)}$ is obtained from the previous estimate $\beta^{(r)}$ by using the equation

$$\beta\ (r+1) = \beta\ (r) + s\ (\beta\ (r))\ E(H(\beta^{(r)}))^{-1} \tag{9}$$

Where H is the Hessian matrix (matrix of second derivatives of the log-likelihood). equation (9) can be rewritten as:

$$\beta^{(r+1)} = (X^T W^{(r)} X)^{-1} X^T W^{(r)} z^{(r)} \tag{10}$$

Where $W^{(r)} = diag(w_i)$,

The working dependent variable $z_i^{(r)} = \eta_i^{(r)} + (y_i - \mu_i^{(r)}) g'(\mu_i^{(r)})$ and $w_i^{(r)} = \frac{a_i}{V(\mu_i^{(r)}) (g'(\mu_i^{(t)}))^2}$

The procedure is usually repeated until successive estimates converge (that is, change by less than a specified small amount)

For Poisson regression model we consider the link

$$\eta_i = \log(\mu_i) \tag{11}$$

The derivative of the link is

$$\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i} \tag{12}$$

The working dependent variable is therefore

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i} \tag{13}$$

By assuming that the prior weight a_i is 1 then the iterative weight is

$$w_i = \frac{1}{V(\mu_i)(g'(\mu_i)^2} = \frac{1}{\left[\mu_i \frac{1}{\mu_i^2}\right]}$$

which gives

$$w_i = \mu_i \tag{14}$$

The parameters are then estimated using equation (10).

2.3. Negative Binomial Model

Negative binomial distribution is used for modeling over-dispersed count data and is a standard generalization of the Poisson. Of the two types of negative binomial discussed in the literature, we shall consider NB2 because of the advantages associated with it.

Let $y_i (i = 1, 2, \ldots, n)$ be a non-negative integer valued random variable rep-resenting the i^{th} outcome and y_i be the associated outcome of interest. The unconditional negative binomial distribution of yi is expressed as:

$$p(y_i) = \frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha) y_i!} \left(\frac{\beta}{1+\beta} \right)^{y_i} \left(\frac{1}{1+\beta} \right)^{\alpha}, y_i = 0, 1, 2, \ldots \tag{15}$$

The above distribution has mean

$$E(y_i) = \alpha\beta \tag{16}$$

and variance

$$\text{var}(y_i) = \alpha\beta + \alpha\beta^2 \tag{17}$$

For building a regression model, the negative binomial distribution can be expressed in terms of parameters $\mu = \alpha\beta$ and $k = 1/\alpha$ so that E (yi) = μ and $\text{var}(y_i) = \mu + k\mu^2$

The model can be expressed in terms of log link as follows:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{18}$$

For p covariates and the regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are to be estimated.

We typically assume that $y_i \sim Negbin(\mu_i, k)$ and taking exponential on equation (18), then distribution (15) can be written as:

$$p(y_i) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})(y_i + 1)} \left(\frac{\alpha e^{x_i.\beta}}{1 + \alpha e^{x_i.\beta}} \right)^{y_i} \left(\frac{1}{1 + \alpha e^{x_i.\beta}} \right)^{1/\alpha} \quad (19)$$

Where, $\mu = \alpha\beta$, $k = 1/\alpha$, $\mu_i > 0$, for $i = 1, 2, ..., n$ and α is the negative binomial over-dispersion parameter.

Parameter estimation

We estimate α and β using Maximum likelihood estimation. The likelihood function is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{i=1}^{n} \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})(y_i + 1)} \left(\frac{\alpha e^{x_i.\beta}}{1 + \alpha e^{x_i.\beta}} \right)^{y_i} \left(\frac{1}{1 + \alpha e^{x_i.\beta}} \right)^{1/\alpha} \end{aligned} \quad (20)$$

and the log-likelihood function is

$$\ln L(\alpha, \beta) = \prod_{i=1}^{n} p(y_i) = \sum_{i=1}^{n} \begin{pmatrix} y_i \ln \alpha + y_i (x_i.\beta) - \left(y_i + \frac{1}{\alpha} \right) \\ \ln(1 + \alpha e^{x_i.\beta}) + \ln \Gamma\left(y_i + \frac{1}{\alpha} \right) \\ - \ln(y_i + 1) - \ln \Gamma\left(\frac{1}{\alpha} \right) \end{pmatrix} \quad (21)$$

The values of α and β that maximizes $\ln L(\alpha, \beta)$ will be the maximum likelihood estimates. The MLE's can be obtained using the fishers' scoring algorithm.

2.4. Hurdle Models

The idea of hurdle comes from considering the data as being generated by a process that commences generating positive counts only after crossing a zero barrier or hurdle and therefore until the hurdle is crossed, the process generates zeros. This implies that the hurdle is crossed if a count is greater than zero. However, for the binary component, values below the hurdle are given the value of 0 and above the hurdle point they are given the value of 1. Starting with the binomial process, suppose that is the probability value when the value for response variable is zero and that 1 is a probability value when the response variable is a positive integer. The probability mass function is given by:

$$\Pr(Y = y) = \begin{cases} \pi & y = 0, \\ 1 - \pi & y = 1, 2, ... \end{cases} \quad (22)$$

The zero-truncated poisson has the probability mass function

$$\Pr(Y = y \mid Y \neq 0) = \begin{cases} \dfrac{\lambda^y}{(e^y - 1)!} & y = 1, 2, ... \\ 0 & otherwise \end{cases} \quad (23)$$

Thus the unconditional probability mass function for Y is

$$\Pr(Y = y \mid Y \neq 0) = \begin{cases} \pi & y = 0 \\ (1 - \pi) \dfrac{\lambda^y}{(e^\lambda - 1) y!} & y = 1, 2, ... \end{cases} \quad (24)$$

and the log likelihood for the i[th] observation assuming the observations are independent and identically distributed is

$$\ln L(\pi_i, \lambda_i, y_i) = \begin{cases} \ln \pi_i & y = 0 \\ \ln(1 - \pi_i) \dfrac{\lambda_i^{y_i}}{(e^{\lambda_i} - 1) y_i !} & y = 1, 2, ... \end{cases} \quad (25)$$

So that if we model π_i using the complementary log-log link and λ_i using the log link, we have

$\pi_i = e^{-e^{x_i \beta_1}}$ and $\lambda_i = e^{x_i \beta_2}$

Thus the log likelihood can be written

$$\begin{aligned} \ln L &= \ln \left\{ \prod_{i \in \Omega_0} \left(e^{-e^{x_i \beta_1}} \right) \prod_{i \in \Omega_1} \left(1 - e^{-e^{x_i \beta_1}} \right) \prod_{i \in \Omega_1} \left(\frac{e^{y_i x_i \beta_2}}{\left(e^{e^{x_i \beta_2}} - 1 \right) y_i !} \right) \right\} \\ &= \left\{ \sum_{i \in \Omega_0} -e^{x_i \beta_1} + \sum_{i \in \Omega_1} \ln \left(1 - e^{-e^{x_i \beta_1}} \right) \right\} \\ &\quad + \left\{ \sum_{i \in \Omega_1} y_i x_i \beta_2 - \sum_{i \in \Omega_1} \ln \left(e^{e^{x_i \beta_2}} - 1 \right) - \sum_{i \in \Omega_1} \ln y_i ! \right\} \\ &= \ln L_1(\beta_1) + L_2(\beta_2) \end{aligned} \quad (26)$$

Where $\Omega_0 = \{ i \mid y_i = 0 \}, \Omega_1 = \{ i \mid y_i \neq 0 \}$ and $\Omega_0 \cup \Omega_1 = \{ 1, 2, ..., N \}$.

The log likelihood above describes the sum of a log log likelihood for the binary outcome model, $L_1(\beta_1)$, and a log likelihood for a truncated-at-zero Poisson model, $L_1(\beta_1)$. The same explanatory variables x_i are used in both cases but the fitted parameters β_1 and β_1 are separate and must not equal each other.

3. Methodology

This study utilized a simulation technique in R to generate data that was used for comparing the count data models.

The following pseudo code was used for simulation purpose:

1. Define the sample size for the data to be simulated.
2. Set the number of simulations. In this case the data sets were simulated 2000 times.
3. Generate random numbers with different proportion of

zeros in the sample for the purpose of defining characteristics of different datasets.

i. Sample command was used in order to generate count datasets
ii. The range for the count data was declared. In this case random integers in the range 0 to 20 were considered
iii. Specify the proportions of zeros using the probability command as an argument for the sample function. For the first dataset set the probability of zeros in the sample was 5%.
iv. Repeat the simulation of the datasets with 10%, 25%, 50%, 75%, and 90% zero proportions.

1. Simulate the covariates to be used in modeling. This was achieved by assuming that the covariates followed uniform distribution.
2. Retrieve each of the simulated datasets which represent data of different characteristics as defined by the pre-specified proportion of zeros.
3. Obtain the summary of the different datasets using each of the three count data models.
4. Note the average AIC's of the different count data models under each data set.
5. Compare the AIC's to determine which model fits better to each of the simulated data sets.

3.1. The Pre-Specified Zero Proportions

In this study, we simulated count data with particular proportions of zero so as to get sets of count data with diverse characteristics. For this study, the concern was goodness-of-fit for different count data models by comparing their average AIC's.

3.2. Random Number Generation

By definition, a random number is one in which there is no way possible to a priori determine its value. Most statistical analysis software packages include random number generators. However, these generated random numbers are not truly random. Usually, one specifies a seed; when replications are performed using the same seed, the generated numbers are identical to the first. Hence, the values are pseudo-random [9]. However, this limitation is actually an advantage in that the researcher can check for errors in model programming and run the analysis again with the same generated sample [10].

Another feature of Monte Carlo random sampling pertains to the desired distributions. Typically, the random numbers are drawn from a uniform distribution, which is then followed by a transformation to the desired distribution. This is because the U (0, 1) distribution with its $0 \le x \le 1$ range, can be used to simulate a set of random probabilities, which are used to generate other distribution functions through the inverse transformation and acceptance-rejection methods [10]. The random number generation was performed using R via a generic sample command.

3.3. Sample and Simulation Size

It is important to determine the appropriate sample size for each simulate. This is because too small sample size is not sufficient enough to assume that estimates are asymptotically normal. On the other hand, as the sample size increases to infinity the computer time also increases. The sample size was set at 100.

Determining the number of simulations was also an important concern since too few replications may result in inaccurate estimates and too many replications may unnecessarily overburden computer time and performance [9]. Reference [11] were able to sufficiently compare the goodness of fit for several competing models using 1,000 simulations. Likewise, [12] selected 1,000 simulations when researching misspecification in negative binomial ZIP models. Reference [13] set the number of simulations at 2,000 when researching the asymptotic properties of the ZIP model.

Most simulations published recently report upward from 1,000 trials, and simulations of 10,000 and 25,000 trials are common [10]. Given the previously noted problems with convergence for the negative binomial ZIP model, it seems prudent to minimize the number of simulations as much as possible. However, it is also important to simulate under conditions already found to produce asymptotic results. Hence, similar to [14] comparison study, the number of simulations was set at 2,000 for each condition.

3.4. Comparison of the Models Goodness-of- Fit

The goodness of fit of a statistical model describes how well it fits into a set of observations. Suppose we are interested in model selection for a set of N observations where we have say, q model parameters to be estimated. Furthermore, let L denote the log-likelihood, then the two commonly used goodness-of-fit statistics for model selection are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) calculated as follows:

$$AIC = -2L + 2q$$

$$BIC = -2L + q\ln(N) \tag{27}$$

In this study, each simulated dataset was analyzed with each of the three count data models to obtain the model's summary. AIC was used to compare how different models fit to different sets of the simulated count data. AIC is a linear transformation of the log-likelihood statistic with a result that is positive in sign and is interpreted in a lower-is-better fashion [14]. The advantage is that the AIC can be used to descriptively compare all models regardless of whether one is nested or not within another.

4. Results and Discussion

Simulation Results and Discussion

It is important to note that the results of this work were limited to the assumptions that the count data has at least some zero count and that the zeros have an importance attached to them.

A sample of 100 count data points composed of different proportions of zeros was simulated. This was based on the covariates which had been simulated assuming a uniform distribution. The count data was simulated with 5%, 10%, 25%, 50%, 75% and 90% proportions of zero count respectively. At each level of zeros, the simulation was performed 1000 times. A regression was performed for each of the simulated data set with the same covariates. The average AIC based on each of the three models (Poisson, Negative binomial and Hurdle) was obtained. The mean and the variance of the simulated response variable were also noted.

From table 1, for 0.05 proportion of zeros in the simulated count data set, the average AIC for Poisson model (850.44) was the highest while that for negative binomial (678.63) was the least. AIC is usually interpreted in the lower is better fashion. The mean of the response variable was 11.39 while the variance was found to be 40.97. These results are interpreted to mean that if the response variable is made of about 5% zeros and it is also over-dispersed (variance > mean), then the best model to use is the negative binomial as opposed to the Poisson and the hurdle models.

When the count data had 0.10 proportion of zeros, the average AICs for Poisson, negative binomial and hurdle models were found to be 882.27, 670.72, 737.12 respectively. The mean and the variance of response variable was 10.21 and 41.46. The implication of these results is that in modeling over-dispersed count data, if the response variable is composed of approximately 10% zero count, then the negative binomial is the best model to use as it fits well to the data as explained by its average AIC value.

Poisson model fits badly to over-dispersed count data with about 25% proportion of zeros. Negative binomial and hurdle models are better compared to Poisson as depicted by their lower AICs of 627.88 and 684.52 respectively. Of all the models under consideration, negative binomial scores the best in terms of AIC values at this level (0.25) of zero proportions in the count data.

Table 1. *Summary of results of data simulation.*

	AICs			Response variable	
Zero proportion	**Poisson**	**Negative binomial**	**Hurdle**	**Mean**	**Variance**
0.05	850.44	678.63	794.57	11.39	40.97
0.10	882.27	670.72	737.12	10.21	41.46
0.25	1007.33	627.88	684.52	7.88	47.30
0.50	1055.86	1057.82	501.48	5.00	39.76
0.75	847.05	849.00	308.19	2.37	24.84
0.90	566.77	568.72	134.25	1.10	13.10

For the count data with about 50% zeros, the obtained average AICs for Poisson, negative binomial and hurdle models were 1055.86, 1057.82 and 501.48 respectively. The interpretation of these results were that when approximately half of the count data has a zero count and the data is over-dispersed then the best model to use is the hurdle model as opposed to the Poisson and negative binomial models which have relatively higher AICs at this level. For 0.75 and 0.90 proportion of zero in the count data, the AIC for hurdle model remains the least amongst the three count data models.

These results are further explained by graph for AICs versus zero proportion in figure 1. The graph shows that the negative binomial dominates the other models up to a point when the count data has about 30% zeros after which the hurdle model dominates and becomes better and better as the proportion of zeros increase. The implication of these is that the hurdle model can be used to handle over-dispersed count data with excess zeros.

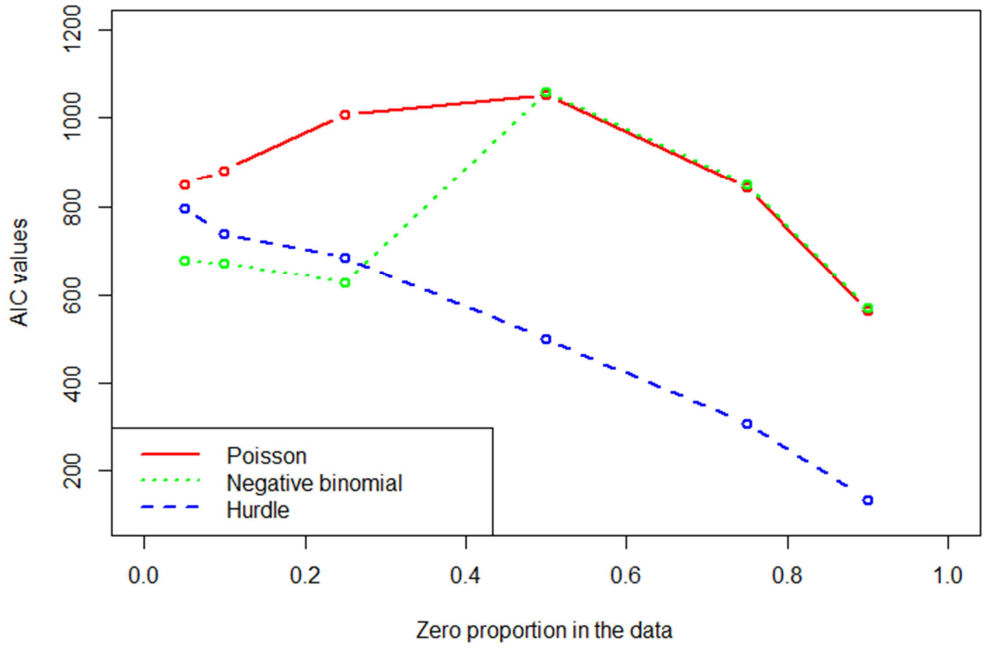

Figure 1. *A comparative graph of AICs for different count data models.*

5. Conclusions and Recommendations

From the results of simulation, it can be concluded that for over-dispersed count data in which the proportion of zero is about 5% and that of non-zero count is about 95%, then the best model to use is negative binomial. The model still performs the best for 0.10 and 0.25 proportions of zero in the count data. In short, negative binomial fits well to count data composed of up to below 30% proportions of zero. It was also concluded that for 0.50, 0.75 and 0.90 proportions of zero the hurdle model fits the best as compared to Poisson and negative binomial models. Hurdle model gets better and better as the proportion of zero increases from 0.30 to 0.90 and therefore it can be used when the count data has the over-dispersion property and excess zeros.

Based on the results of this work, we recommend that in modeling count data, apart from considering whether the model's assumptions are met, the researcher or any other practitioner should also consider the zero proportion in the dataset. If the proportion of zero in the count data is below 0.30 and the data is also over-dispersed, then negative binomial model should be used. Otherwise, if the zero proportion is about 0.30 or more then hurdle model could be a better choice.

References

[1] Dalrymple, M. L., Hudson, I., & Ford, R. P. K. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis*, 41 (3), 491-504.

[2] Gurmu, S., & Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics*, 14 (4), 469-477.

[3] Johansson, A. (2014). A comparison of regression models for count data in third party automobile insurance.

[4] Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention, 37 (1), 35-46.

[5] Frees, E. W. (2010). *Regression modeling with actuarial and financial applications.* Cambridge University Press.

[6] Cameron, A., & Trivedi, P. (1999). *Regression analysis of count data.* Cam-bridge University Press.

[7] Johnson, N. L., Kotz, S., & Kemp, A. (1992). *Univariate distributions*. New York, John Wiley.

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

[9] Bonate, P. L. (2001). A brief introduction to monte carlo simulation. *Clinical pharmacokinetics*, 40 (1), 15-22.

[10] Mooney, C. Z. (1997). Monte carlo simulation (quantitative applications in the social sciences).

[11] Min, Y., & Agresti, A. (2005). Random e ect models for repeated measures of zero-in ated count data. *Statistical Modelling,* 5 (1), 1-19.

[12] Civettini, A. J., & Hines, E. (2005). Misspeci cation e ects in zero-in ated negative binomial regression models: Common cases. In *Annual meeting of the southern political science association. new orleans, la.*

[13] Lambert, D. (1992). Zero-in ated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 (1), 1-14.

[14] Miller, J. M. (2007). *Comparing poisson, hurdle, and zip model fit under varying degrees of skew and zero-inflation.* University of Florida