American Journal of Mathematics and Statistics 2018, 8(6): 179-183
DOI: 10.5923/j.ajms.20180806.03

Modelling Count Data; A Generalized Linear Model Framework

Obubu Maxwell[1,*], Babalola A. Mayowa[2], Ikediuwa U. Chinedu[1], Amadi E. Peace[3]

[1]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria
[3]Department of Statistics, Abia State Polytechnic, Aba, Nigeria

Abstract Count Data Models allow for regression-type analyses when the dependent variable of interest is a numerical count. They can be used to estimate the effect of a policy intervention either on the average rate or on the probability of no event, a single event, or multiple events. The mostly used distribution for modeling count data is the Poisson distribution (Horim and Levy; 1981) which assume equidispersion (Variance is equal to the mean). Since observed count data often exhibit over or under dispersion, the Poisson model becomes less ideal for modeling. To deal with a wide range of dispersion levels, Negative Binomial Regression, Generalized Poisson Regression, Poisson Regression, and lately Conway-Maxwell-Poisson (COM-Poisson) Regression can be used as alternative regression models. We compared the Generalized Poisson regression to all other regression models and also stated their advantages and usefulness. Data were analyzed using these four methods, the results from the four methods are compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion with the Generalized Poisson Regression having the smallest AIC and BIC values. The Generalized Poisson Regression Model was considered a better model when analyzing road traffic crashes for the data set considered.

Keywords Over-dispersion, Count Data, Negative Binomial Regression, Generalized Poisson Regression, Conway-Maxwell Poisson, Akaike Information Criterion, Equidispersion

1. Introduction

Count data is a statistical data type, a type of data in which the observations can take only the non-negative integer values {0, 1, 2, 3, ...}, and where these integers arise from counting rather than ranking [1-3]. The statistical treatment of count data is distinct from that of binary data, in which the observations can take only two values, usually represented by 0 and 1, and from ordinal data, which may also consist of integers but where the individual values fall on an arbitrary scale and only the relative ranking is important. Count data models have a dependent variable that is counts (0, 1, 2, 3, and so on) [4]. Most of the data are concentrated on a few small discrete values. Examples include: the number of children a couple has, the number of doctor's visit per year a person makes, and the number of trips per month that a person takes. Count data arise in many fields which includes; biology, healthcare, psychology, marketing and many more. When response variable is a count and the researcher is interested in how this count changes as the explanatory variable increases. Classical Poisson regression is the most well-known methods for modeling count data, but its underlying assumption of equidispersion limits its use in many real-world applications with over-or under dispersed data [5-7]. This excess variation may result to incorrect inference about parameter estimates, standard errors, tests and confidence intervals. Over-dispersion mostly arises for various reasons including mechanisms that generate excessive zero counts or censoring [9-11]. As a result over-dispersed count data are common in many areas which in turn, have led to the development of statistical methodology for modeling over-dispersed data. For over-dispersed data, the Negative Binomial model is a popular choice. Other over-dispersion models include Poisson mixtures and Conway-Maxwell-Poisson. A flexible alternative that captures both over- and under-dispersion is the Conway-Maxwell-Poisson (COM-Poisson) distribution. The COM-Poisson is a two-parameter generalization of the Poisson distribution which also includes the Bernoulli and Geometric distributions as special cases [12]. The COM-Poisson distribution has been used in so many count data application and has been extended methodologically in various directions. Therefore in this work, because of the problem of model selection and the appropriate method to

* Corresponding author:
maxwellobubu@gmail.com (Obubu Maxwell)
Published online at http://journal.sapub.org/ajms

apply in the analysis of auto-crash data bearing in mind their underlying assumptions, we wish to find the model that is most adequate.

2. Materials and Method

In this section we shall review the models that most widely used in the analysis of count data which include: the Poisson models, Conway- Maxwell- Poisson models, Generalized Poisson Regression model, and the Negative Binomial Regression Model.

2.1. Poisson Models

This is a special case of Generalized Linear Models (GLM) framework. The simplest distribution used for modeling count data is the Poisson distribution with probability density function.

$$P_r(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \; for \; y = 0,1,2, \ldots$$

For $\mu > 0$. The mean and variance of this distribution can be shown to be $E(Y) = var(Y) = \mu$. Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data.

Suppose that we have a sample of n observations $y_1, y_2, \ldots, y_n$ which can be treated as realizations of independent Poisson random variables, with $Y_i \sim P(\mu_i)$ and suppose that we want to let the mean μ_i (and therefore the variance) depend on a vector of explanatory variables x_i [13-15].

We could entertain a simple linear model of the form

$$\mu_i = x_i\beta$$

but this model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative. A straightforward solution to this problem is to model instead the logarithm of the mean using a linear model. Thus, we take logs calculating

$$\eta_i = log(\mu_i)$$

and assume that the transformed mean follows a linear model

$$\eta_i = x_i\beta$$

Thus, we consider a generalized linear model with link log. Combining these two steps in one we can write the log-linear model as

$$log(\mu_i) = x_i\beta.$$

In this model the regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictor x_j . In other words increasing x_j by one unit is associated with an increase of β_j in the log of the mean. Exponentiating the above equation, we obtain a multiplicative model for the mean itself:

$$\mu_i = exp\{x_i\beta\}.$$

In this model, an exponentiated regression coefficient $exp\{\beta\}$ represents a multiplicative effect of the $j-th$ predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor $exp\{\beta_j\}$. A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive [16]. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to a much simpler model.

The Likelihood function for the Poisson model is;

$$L(\beta|y, X) = \prod_{i=1}^{N} Pr(y_i|\mu_i) = \prod_{i=1}^{N} \frac{exp(-\mu_i)\mu_i^{y_i}}{y!}$$

2.2. Conway-Maxwell-Poisson (COM-Poisson) Models

The Conway Maxwell Poisson (COM-Poisson) distribution with two parameters was originally developed as a solution to handling queueing systems with state-dependent arrival or service rates. This distribution generalizes the Poisson distribution by adding a parameter to model over-dispersion and under-dispersion and includes the geometric distribution as a special case and the Bernoulli distribution as a limiting case. The COM-Poisson distribution is a two parameter generalization of the Poisson distribution that is flexible enough to describe a wide range of counts data distributions, since its revival, it has been further developed in several directions and applied in multiple fields.

The COM-Poisson probability distribution function is given by the equation:

$$P(X = j) = \frac{1}{Z(\lambda,\nu)}\frac{\lambda^j}{(j!)^\nu} \; , \; j \in Z^+ = \{0,1,2, \ldots\}$$

Where $Z(\lambda, \nu)$ is a normalizing constant defined by

$$Z(\lambda, \nu) = \sum_{i=0}^{\infty} \frac{\lambda^i}{(i!)^\nu}$$

The domain of admissible parameters for which defines a probability distribution is $\lambda, \nu > 0, and \; 0 < \lambda < 1$, $\nu = 0$. The introduction of the second parameter ν allows for either sub or super-linear growth of the ratio $P(X = j - 1)/P(X = j)$, and allows X to have variance either less than or greater than it's mean. Of course, the mean of $X \sim CMP(\lambda, \nu)$ is not, in general, λ. Clearly, in the case where $\nu = 1, X \sim CMP(\lambda, 1)$ has the Poisson distribution $P_o(\lambda)$ and the normalizing constant $Z(\lambda, 1) = e^\lambda$. Note, other choices of ν also give rise to well-known distributions. For example, in the case where $\nu = 0$ and $0 < \lambda < 1$, X has a geometric distribution, with $Z(\lambda, 0) = (1 - \lambda)^{-1}$. In the limit $\nu \to \infty$, X converges in distribution to a Bernoulli random variable with mean $\lambda(1 + \lambda)^{-1}$ and $\lim \nu \to \infty \, Z(\lambda, \nu) = 1 + \lambda$. In general, of course, the normalizing constant $Z(\lambda, \nu)$ does not permit such a neat, closed-form expression. Asymptotic results are available, however. Gillispie and Green [17] prove that, for fixed ν,

$$Z(\lambda,v)\sim\frac{exp\left\{v\lambda^{1/v}\right\}}{\lambda^{(v-1/2v)}(2\pi)^{(v-1)/2}\sqrt{v}}\left(1+O\left(\lambda^{-1/v}\right)\right)$$

As $\lambda\to\infty$, confirming a conjecture made by Shmueli et al [18-19]. This asymptotic result may also be used to obtain asymptotic results for the probability generating function of $X\sim CMP(\lambda,v)$, since it may be easily seen that

$$Es^X=\frac{Z(s\lambda,\ v)}{Z(\lambda,\ v)}$$

2.3. The Generalized Poisson Regression Model

The advantage of using the generalized Poisson regression model is that it can be fitted for both over-dispersion, $Var(y_i)>E(y_i)$, as well as under-dispersion, $Var(y_i)>E(y_i)$. Suppose is a count response variable that follows a generalized Poisson distribution, the probability density function of $y_i, i=1,2,\dots,n$ is given as (Famoye (1993), Wang and Famoye (1997)) [20];

$$f_i(y_i,\mu_i,\alpha)=\left(\frac{\mu_i}{1+\alpha\mu_i}\right)\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\exp\left[\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right]$$

$$y_i=0,1,2,\dots,and\ \mu_i=\mu_i(x_i)=exp(x_i\beta),$$

Where x_i is a $(k-1)$ dimensional vector of covariates including demographic factors, driving habits and medication use, and β is a $k-$dimensional vector of regression parameters. For details on the generalized Poisson regression model, the reader is referred to Famoye (1993) [21]. The mean and variance of Y_i are, respectively, given by

$$E(Y_i|x+i)=\mu_i$$

and

$$V(Y_i|x_i)=\mu_i(1+\alpha\mu_i)^2$$

The generalized Poisson regression model above is a generalization of the standard Poisson regression (PR) model. When $\alpha=0$ the probability function model, the equality constraint is observed between the conditional mean $E(Y_i|x_i)$ and the conditional variance $V(Y_i|x_i)$ of the dependent variable for each observation. In practical applications and in "real" situations, this assumption is questionable since the variance can either be larger or smaller than the mean. If the variance is not equal to the mean, the estimates in PR model are still consistent but are inefficient, which leads to the invalidation of inference based on the estimated standard errors.

2.4. Negative Binomial Regression

Negative binomial regression is similar to regular multiple regression except that the dependent (Y) variable is an observed count that follows the negative binomial distribution. Thus, the possible values of Y are the nonnegative integers: 0, 1, 2, 3, and so on. Negative binomial regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This formulation is popular because it allows the modelling of Poisson heterogeneity using a gamma distribution [22].

The Poisson distribution may be generalized by including a gamma noise variable which has a mean of 1 and a scale parameter of ν. The Poisson-gamma mixture (negative binomial) distribution that results is

$$Pr(Y=y_i|\mu_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}$$

$$\mu_i=t_i\mu$$

$$\alpha=\frac{1}{v}$$

The parameter μ is the mean incidence rate of y per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t_i to represent the exposure for a particular observation. When no exposure given, it is assumed to be one. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t.

The results below make use of the following relationship derived from the definition of the gamma function

$$ln\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right)=\sum_{j=0}^{y_i-1}ln(j+\alpha^{-1})$$

In negative binomial regression, the mean of y is determined by the exposure time t and a set of k regressors variables (the x's). The expression relating these quantities is

$$\mu_i=exp(ln(t_i)+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki})$$

Often, $x_1\equiv1$, in which case β_1 is called the intercept. The regression coefficients $\beta_1,\beta_2,\dots,\beta_k$ are unknown parameters that are estimated from a set of data. Their estimates are symbolized as $b_1,b_2,\dots,b_k$. Using this notation, the fundamental negative binomial regression model for an observation i is written as

$$Pr(Y=y_i|\mu_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

2.5. Multicollinearity Test

One formal way of detecting Multicollinearity is by the use of the variance inflation factors (VIF) [23]. The VIF is used to test for the presence of Multicollinearity, and is given by

$$VIF=\frac{1}{1-R_j^2}$$

Where R_j^2 is the coefficient of determination of a regression of an explanatory variable j on all the other explanatory variables. A VIF value of 10 and above indicates a Multicollinearity problem.

Table 1 shows that all the variables have VIF values <10. Thus all the variables can be included in the subsequent analyses and modeling with the Poisson regression, Generalized Poisson regression, and Negative Binomial

Regression.

Table 1. Multicollinearity Test

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	NUMBER OF CRASHES	.609	1.643
	WEEK	.971	1.030
	NUMBER OF CAUSES	.621	1.611

2.6. Akaike Information Criterion (AIC)

When several models are available, one can compare the models performance based on several likelihood measures which have been proposed in statistical literatures. One of the most popularly used measures is AIC [24-25]. The AIC penalized a model with larger number of parameters, and is defined as

$$AIC = -2lnL + 2p = -2[lnL - p]$$

Where lnL denotes the fitted log likelihood and p the number of parameters. A relatively small value of AIC is favorable for the fitted model.

2.7. Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred [26]. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The BIC is formally defined as

$$BIC = ln(n)k - 2ln(\hat{L})$$

Where,

$\hat{L} =$ the maximized value of the likelihood function of the model.

$n =$ the number of data points in the observed data, the number of observations, or equivalently, the sample size.

$k =$ the number of parameters estimated by the model.

3. Results and Discussion

Auto crash data was collected from the Federal Road Safety Corps National Headquarters Abuja and data wase analyzed using R Software and the results obtained are given below. Before performing the analysis on the four methods used, the data were tested for Multicollinearity. The test results are shown on the table below.

Table 2. Parameter estimates, standard error and AIC value for the models

	POISSON REG.		NEGATIVE BINOMIAL REG.		GENERALIZED POISSON REG.		COM-POISSON REG	
	Estimated coefficient	Std Error	Estimated coefficient	Std Error	Estimated coefficient	Std Error	Estimated coefficient	Std Error
Intercept	2.908	0.053	2.908	0.228	2.255	0.125	16.946	6.213
Number of crashes	0.072	0.007	0.072	0.030	0.079	0.027	2.185	0.956
Season (Week of the year)	-0.006	0.001	-0.006	-0.005	-0.004	0.005	-0.160	0.151
Number of causes								
AIC	**2325.8**		**944.078**		**896.0278**		**951.01**	
BIC	**2322.7**		**935.011**		**891.0271**		**950.08**	

4. Conclusions and Recommendations

Poisson Regression Model, Generalized Poisson Regression Model, Negative Binomial Regression Model, and Conway-Maxwell Poisson regression model were compared to determine a better model used in modeling auto-crashes in Nigeria. The criterion for selection of the best model used is AIC and BIC values. Best model is the model that has the smallest AIC and BIC value.

Based on the values on Table 2 above, the model with the smallest AIC and BIC value is the Generalized Poisson Regression model. Thus, the best model for analyzing traffic crash data in Nigeria is the Generalized Poisson Regression model.

REFERENCES

[1] A. Zeileis, C. Kleiber, S. Jackman, Regression Models for Count Data in R, *Journal of Statistical Software.* 27, Issue 8, 1-25 (2008).

[2] Bozdogan, H. (2000), "Akaike's Information Criterion and Recent Developments in Information Complexity". Mathematical Psychology, 44, 62-91.

[3] Cameron, A.C and Trivedi, P. K (1998), "Regression analysis of count data." Cambridge University press Cambridge, UK.

[4] Consul P. C. and Famoye F. (1992), "Generalized Poisson regression model", communications in statistics (theory and methodology) vol. 2, no.1, 89-109.

[5] Obubu, M. Nwokolo P.C (2016), "Prevalence of Breast Cancer in Delta State, Nigeria." World Journal of Probability and Statistics. Vol. 2, No. 2, Pp 1-9.

[6] Osuji G.A., Obubu, M., Obiora-Ilouno H.O (2016), "Uterine Fibroid on Women's Fertility and Pregnancy Outcome in Delta State, Nigeria." Journal of Natural Sciences Research, Vol. 6, No 2, pp. 27-33.

[7] Osuji G.A., Obubu, M., Obiora-Ilouno H.O (2016), "An investigation on the causes of Low birth weight in Delta State, Nigeria" European Journal of Statistics and Probability Vol. 4, No 1, pp. 1-6.

[8] Nantulya, V.M and Reich M.R (2002), "The neglected epidemic: Road traffic injuries in developing countries". Br. Med. Journal, 324:1139-1141.

[9] N. Ismail, A.A. Jemain, Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models, *Casualty Actuarial Society Forum.* Winter:103-158 (2007).

[10] O'Neill B. and Mohan D. (2002), "Reducing motor vehicle crash deaths and injuries in newly motorizing countries". BMJ., 324:1142-1145.

[11] S.D. Guikema, J.P. Coffelt, A Flexible Count Data Regression Model for Risk Analysis, *Risk Analysis.* 28, Issue 1, 213-223 (2008).

[12] Osuji G.A., Obubu, M., Nwosu, C.A (2016), "Preconception sex selection using proper ovulation timing" World Journal of Probability and Statistics Research U.S.A, Vol. 2, No 1, pp. 1-12. http://wjmcr.com.

[13] Osuji, G.A., Okoro, C.N., Obubu, M., Obiora-Ilouno H.O. (2016) "Effect of Akaike Information Criterion on Model Selection in Analyzing Auto-Crash Variables." International Journal of Sciences: Basic and Applied Research (IJSBAR). Vol. 26, No 1, pp. 98-109.

[14] Osuji G.A., Obubu, M., Obiora-Ilouno H.O., Okoro, C.N (2015), "Post-Partum Hemorrhage in Delta State, Nigeria; A Logistic Approach." International Journal of Sciences: Basic and Applied Research (IJSBAR) Canada Vol. 24, No 6, pp. 45-53.

[15] D. Lord, S.R. Geedipally, S.D. Guikema, Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion, *Risk Analysis.* 30, Issue 8, 1268-1276 (2010).

[16] G.J. McLachlan, On the EM Algorithm for Overdispersed Count Data, *Statistical Methods in Medical Reseach.* 6, 76-98 (1997).

[17] Gillispie, S. B. and Green, C. G. (2015). Approximating the Conway-Maxwell-Poisson distribution normalizing constant. Statistics 49: 1062-1073.

[18] G. Shmueli, T.P. Minka, J.B. Kadane, S. Borle, P. Boatwright, A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution, *Journal of The Royal Statistical Society. Series C (Applied Statistics).* 54, Issue 1, 127-142 (2005).

[19] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the COM-Poisson. J. R. Stat Soc. Ser. C 54: 127–142.

[20] Famoye F, John T. W. and Karan P. S. (1997), "On the generalized Poisson regression model with an application to accident data". Journal of data science 2 (2004), 287-295

[21] Famoye, F. (1993). Restricted generalized Poisson regression model. Communications in Statistics – Theory and Methods 22, 1335-1354.

[22] K.F. Sellers, G. Shmueli G, Data Dispersion: Now you see it...Now you don't, *Communication in Statistics: Theory and Methods.* 42, Issue 17, 3134-47 (2013).

[23] Osuji G.A., Obubu, M., Nwosu, C.A (2016), "Stock Investment Decision in Nigeria; A PC Approach" World Journal of Multidisciplinary and Contemporary Research, Vol. 2, No 1, pp. 1-11. http://wjmcr.com.

[24] Osuji G.A., Obubu, M., Obiora-Ilouno H.O., Nwosu, D.F (2015), "Perinatal Mortality and Associated Obstetric Risk Factors in Urban Delta State, Nigeria; Rural-Urban Differences." International Journal of Mathematics and Statistics Studies Vol. 3, No. 5, PP 32-46.

[25] Osuji G.A., Obubu, M., Obiora-Ilouno H.O (2015), "An Investigation on Crime Rate in Southeastern Nigeria." European Journal of Statistics and Probability Vol. 3, No 4, pp. 1-9.

[26] Obubu, M., Okoye Valentine, Omoruyi Frederick, Ngonadi Lilian Oluebube (2017), "Infant Mortality; a continuing social problem in Northern Nigeria: Cox Regression Approach. American Journal of Innovative Research and Applied Sciences.2017; 5(5): 1-5.