COMPARISON COUNT REGRESSION MODELS FOR OVERDISPERSED ALGA DATA

Esin Avcı[1*], Sibel Alturk[2,] & Elif Neyran Soylu[3]
[1] The University of Giresun, Science and Art Faculty, Department of Statistics, Gure Yerleskesi, 28000 Giresun, Turkey
[2,3] The University of Giresun, Science and Art Faculty, Department of Biology, Gure Yerleskesi, 28000 Giresun, Turkey

ABSTRACT
Count data become widely available in many diciplines. The most popular distribution for modeling count data is the Poisson distribution which assume equidispersion (Variance is equal to the mean). Since observed count data often exhibit over or under dispersion, Poisson models become less ideal for modeling. To deal with a wide range of dispersion levels, Quasi Poisson regresion, Negative Binomial regression and lately Conway-Maxwell-Poisson (COM-Poisson) regression used as an alternative regression models. We compare the COM-Poisson to all other regression models and illustrate its advantage and usefulness using over-dispersed alga data.

Keywords: *COM-Poisson regression, over-dispersed count data, Navicula cryptocephala Kutzing*.

1. INTRODUCTION

Count data arise in many fields, including biology, healthcare, psychology, marketing and more. When response variable is a count and the researcher is interested in how this count changes as the explanatory variable increase. Classical Poisson regression is the most well-known methods for modeling count data, but its underlying assumption of equi-dispersion (i.e., an equal mean and variance) limits its use in many real-world applications with over-or under dispersed (i.e., the variance is larger than the mean or smaller than the mean) data.This excess variation may accour incorrect inference about parameter estimates, standart errors, tests and confidence intervals. Overdispersion frequently arises for various reasons, including mechanisms that generate excessive zero counts or cencoring. As a result over-dispersed count data are common in many areas which in turn, has led to the development of statistical methodology for modeling over-dispersed data [1].For over-dispersed data, the Negative Binomial model is a popular choice [2]. Other over-dispersion models include Poisson mixtures [3], Quasi-Poisson regression and Conway-Maxwell-Poisson. Quasi-Poisson regression produces regression estimates equivalent to Poisson regression, but standart errors larger than classical Poisson regression [4].However, these models are not suitable for under-dispersion. A flexiable 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 a special ceses [5]. The COM-Poisson distribution has been used in a variety of count data application and has been extended methedologyically in various direction [6].

This article is organized as follows, Section 2 briefly describes the Poisson, Quasi-Poisson, Negative Binomial and COM-Poisson regression respectively. Section 3 presents dataset on Navicula cryptocephala Kutzing, where a sample of over-dispersed data. All mentioned regression models were applied and the results were compaired. Section 4 presents conclusion.

2. METHODS

2.1. Poisson Models

Poisson regression 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

$$f(y;\mu)=\frac{\exp(-\mu)\mu^{y}}{y!} \qquad (1)$$

The canonical link is $g(\mu)=\log(\mu)$ resulting in a log-linear relationship between mean and linear predictor. The variance in the Poisson model is identical to the mean, thus the dispersion is fixed at $\phi=1$ and the variance function is $V(\mu)=\mu$ [7]. The mean Poisson regression can be assumed to follow a log link, $E(Y_i)=\mu_i=exp(x_i{}'\beta)$, where x_i denotes the vector of explanatory variables and β the vector of regression parameters. The maximum likelihood estimates can be obtained by maximizing the log likelihood.

2.2. Quasi-Poisson Models

In order to relax the Poisson assumption of equidispersion, Quasi-Poisson methods represent a potential solution. By this way few assumptions about the dispersion for the dependent variable are required; only the relationship between the outcome mean and the explanatory variables, and between the mean and variance must be specified [8].This strategy leads to the same coefficient estimates as the standart Poisson models but inference is adjusted for over-dispersion [7].

2.3. Negative Binomial Models

A second way of modeling over-dispersed count data is to assume a Negative Binomial (NB) distribution for $(y_i|x_i)$ which can arise as a Gamma mixture of Poisson distributions. One parameterization of its probability density function is

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{(y+\theta)}} \qquad (2)$$

with mean μ and shape parameter θ; $\Gamma(.)$ is the Gamma function. For every fixed θ , this is another special case of the GLM framework. It also has $\phi = 1$ but with variance function $V(\mu) = \mu + \frac{\mu^2}{\theta}$[7]. The mean of NB regression can also be assumed to follow the log link, $E(Y_i) = \mu_i = exp(x_i{}'\beta)$, and the maximum likelihood estimates can be obtained by maximizing the log likelihood.

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

The Conway-Maxwell-Poisson (COM-Poisson) distribution has been re-introduced bystatisticians to model count data characterized by either over- or under-dispersion [5,9,10,11]. The COM-Poisson distribution was firstintroduced in 1962 by Conway and Maxwell [12]; only in 2008 it was evaluated in the context of a GLM by Guikemaand Coffelt *et al*. (2008) [13], Lord *et al*. (2008) [10] and Sellers and Shmueli (2010) [14]. The COM-Poisson distribution is a twoparameter generalization of the Poisson distribution that is flexible enough to describe a wide range of count datadistributions [14]; since its revival, it has been further developed in several directions andapplied in multiple fields [5].

The COM-Poisson probability distribution function [5] is given by the equation:

$$P(y;\lambda,\upsilon) = \frac{\lambda^y}{(y!)^\upsilon Z(\lambda,\upsilon)} \qquad (3)$$

for a random variable Y, where $Z(\lambda,\upsilon) = \sum_{s=0}^{\infty}\frac{\lambda^s}{(s!)^\upsilon}$, and $\upsilon \geq 0$ is a normalizing constant; υ is considered the dispersion parameter such that $\upsilon > 1$ represents under-dispersion, and $\upsilon < 1$ over-dispersion. The COM-Poisson distribution includes three well-known distribution as special cases: Poisson ($\upsilon = 1$), Geometric ($\upsilon = 0, \lambda < 1$), and Bernoulli ($\upsilon \to \infty \; with \; probability \; \frac{\lambda}{1+\lambda}$)[5].

Taking a GLM approach, Sellers and Shmueli (2010) [14] proposed a COM-Poisson regression model using the link function,

$$\eta(E(Y) = \log\lambda = X'\beta = \beta_0 + \sum_{j=1}^{p}\beta_j X_j \qquad (4)$$

Accordingly, this function indirectly models the relationship between $E(Y)$ and $X'\beta$, and allows for estimating β and υ via associated normal equations. Because of the complexity of the normal equation, using $\beta^{(0)}$ and $\upsilon^{(0)} = 1$, as starting values. These equations can thus be solved via an appropriate iterative reweighted least squares procedure (or by maximizing the liklihood function directly using an optimization program) to determine the maximum likelihood estimates, $\hat{\beta}$ and $\hat{\upsilon}$. The associated standart errors of the estimated coefficients are derived using the Fisher Information matrix [1].

2.5. Testing for Variable Dispersion

Sellers and Shmueli (2010) [14] established a hypothesis testing procedure to determine if significant data dispersion exists, thus demonstarting the need for a COM-Poisson regression model over a simple Poisson regression model; in other words, they test whether ($\upsilon = 1$) or otherwise [1].

Since NB regression reduce to Poisson regression in the limit as $\theta \rightarrow 0$, The test of overdispersion in Poisson vs Negative Binom Poisson,is the test whether ($\theta = 0$) or otherwise, can be performed using Likelihood Ratio Test (LRT), $T = 2(lnL_1 - lnL_0)$, where lnL_1 and lnL_0 are the models log likelihhood under their respective hypothesis. To test the null hypothesis at significance level α, the critical value of chi-square distribution with significance level 2α is used. The null hypothesis is rejected when T value exceed critical value of chi-square [15].

2.6. Akaike Information Criteria (AIC)

When several models are available, one can compare the models performance based on several likekihood measure which have been proposed in statistical literatures. One of the most popular used measure is AIC. The AIC penalized a model with larger number of parameters, and is defined as

$$AIC = -2lnL + 2p \tag{5}$$

where lnL denotes the fitted log likelihood and p the number of parameters [15]. A relatively small value of AIC is favorable for the fitted model.

3. DATA ANALYSIS

Navicula Cryptocephala Kutzing is one of the Epilithic algae, occur in Freshwater (more oligotrophic) habitats, including slightly humic waters. Between June 2013 and May 2014 thenumber of Navicula Cryptocephala Kutzing data are collected from four different station that located on Batlama stream.

In order to model the effect of station and water temperature on the number of Navicula Cryptocephala Kutzingall the models described above applied to data set. At the end of the section, all fitted models are compared highlighting that the modelled mean function is similar but the fitted likelihood and AIC are different. The analysis are performed in R program. Respectievely, glm() function from "stats" package, glm.nb() function from "MASS" package and cmp() function from "COMPoissonReg" package are used.

To obtain a first overview of the dependent variable, we are employed a histogram of the observed count frequencies. The histogram (Figure 1.) illustrates that the marjinal distribution exhibits substantial variation.

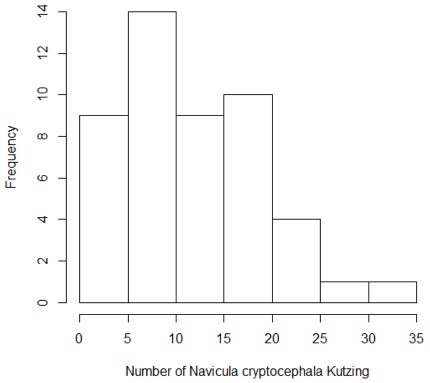

Figure 1. Frequency of Navicula Cryptocephala Kutzing

A second step in the exploratory analysis is to look at pairwise bivariate displays of the dependent variable against each of the regressors bringing out the partial relationships.

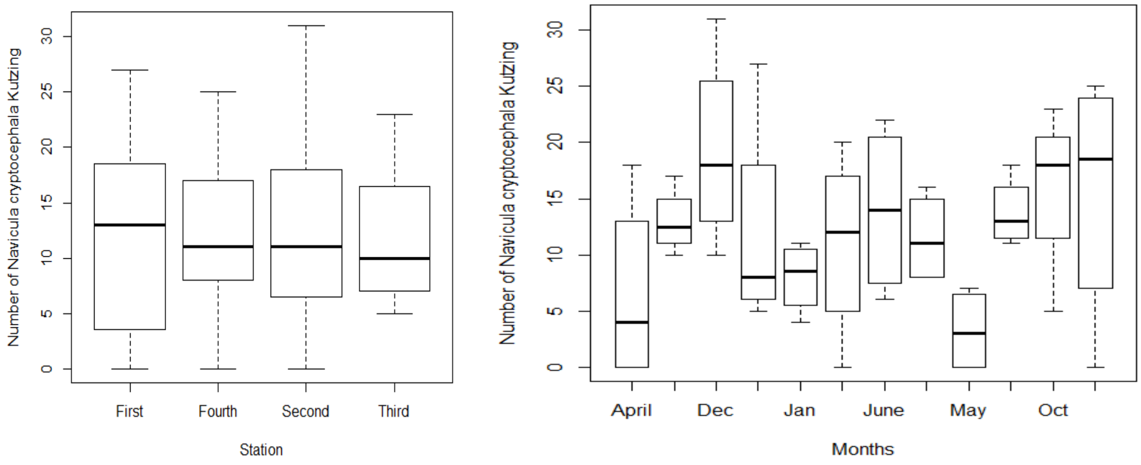

Figure 2. Box Plot of the dependent variable against station and Months

All displays (Figure 2.) show that the number of Navicula Cryptocephala Kutzing increases or decreases with the regressors. The number of Navicula Cryptocephala Kutzing is slightly lower for third station compared to first station. The number of Navicula Cryptocephala Kutzing is increases with September but decreases with May.

Regressing the number of Navicula Cryptocephala Kutzing on station and months (model-1) for all mentioned models. The best model is chosen using backward stepwise based on both AIC ($AIC_{model(2)} < AIC_{model(1)}$)and p values (drop a covariate if it is not significant). Based on two test, the effect of station is not found statistically significant, we drop the station covariate, and continous with only months effect (model-2). Results of this analysis are given in table 1. Table 1 shows the parameter estimates, standart error and AIC value to chosen best model.

Table 1. Parameter estimates, standart error and AIC value for models

		Classic Poisson		***Quasi-Poisson***		***Negative Binomial***		***COM-Poisson***	
		Estimated Coefficent	Standart Error	*Estimated Coefficent*	*Standart Error*	*Estimated Coefficent*	*Standart Error*	*Estimated Coefficent*	*Standart Error*
1	**Intercept**	2.8804	0.1239	2.8804	0.2705	3.0053	0.3354	0.3827	0.1481
	Station	-0.0104	0.0373	-0.0104	0.0814	-0.0274	0.0978	-0.0021	0.0165
	Months	-0.0604	0.0122*	-0.0604	0.0267*	-0.0739	0.0318*	-0.0120	0.0060*
	Dispersion parameter	-	-	-	-	2.0562	0.570	0.1417	0.0467
2	**Intercept**	2.8543	0.0822	2.8543	0.1778	2.9320	0.2300	0.3774	0.1412
	Months	-0.0604	0.0122*	-0.0604	0.0265*	-0.0731	0.0318*	-0.0120	0.0060*
	Dispersion parameter	-	-	-	-	2.0524	0.569	0.1416	0.0467
1	**Log-liklihood**	-219.19		-		-164.61		-160.97	
	AIC	444.37		-		337.21		333.94	
2	**Log-liklihood**	-219.23		-		-164.64		-160.98	
	AIC	442.45		-		335.29		331.95	

* p<0.05 (significant covariate)

Comparing Poisson model with other three models, we find that the ratio $\widehat{\beta_1}/\hat{\sigma}\widehat{\beta_1}$ is 4.95, 2.28, 2.30 and 2 respectively. As expected Quasi-Poisson and NB are smaller then Poisson model. After dividing the COM-Poisson coefficients by υ dispersion parameter (-0.012/0.1416=0.0847), the results in table 1 indicate that the regression parameters for all models have similar estimates in terms of the coefficient magnitudes.

For testing overdispersion in Poisson versus NB regression, the likelihood ratio is 2[-164.64-(-219.23)]=109.18 indicating that the null hypothesis is rejected and the NB is more adequate. The estimated dispersion parameter for COM-Poisson model is $\upsilon = 0.1416$, indicating over-dispersion. To determine whether the dispersion parameter is significiant or nota hypothesis test which established by Sellers and Shmueli (2010) [14] is used. The p value are around zero found 0, indicating dispersion that requires a COM-Poisson regression instead of Poisson regression.

In terms of Log likelihood and AIC, the COM-Poisson shows best fit for model-2.

4. CONCLUSION

This study is related with the response variable of interest is a count, that is, takes on nonnegative integer values. For count data, the most widely used regression model is Poisson regression. Poisson regression is limiting in equidispersion assumption. When data display over-dispersion, the common solution is to use Quasi-Poisson and Negative Binomial regression. Lately COM-Poisson regression used to fit over- or under- dispersed data.

In order to model the effect of station and water temperatureon the number of Navicula Cryptocephala Kutzing, Poisson, Quasi-Poisson, NB and COM-Poisson regression models are fitted respectively. The results indicated that the regression parameters of all models had similar estimates and the ratio, for Quasi-Poisson and NB models were smaller then Poisson models. Both of test for overdispersion indicated that NB and COM-Poisson regression were more adequate than Poisson model. In terms of Log likelihood and AIC, the COM-Poisson shows best fit for model-2.

5. ACKNOWLEDGEMENTS

The authours thank Kimberly Sellers for her insightful comments and help to run COMPoissonReg code.

6. REFERENCES

[1] 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).

[2] J.M. Hilbe,"Negative Binomial Regression". 2nd edition. Cambridge University Press, London (2011).

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

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

[5] 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).

[6] K.F. Sellers, S. Borle, G. Shmueli, The CMP Model for Count Data: A Survey of Methods and Applications, *Applied Stochastic Models in Business and Industry*. **28,** Issue 2, 104- 116 (2012).

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

[8] P. McCullagh, J.A. Nelder, "Generalized Linear Models". 2nd edition. Chapman & Hall. London (1989).

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

[10] 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).

[11] Y. Zou, D. Lord , S.R. Geedipally, Over- and Under-Dispersed Crash Data: Comparing the Conway-Maxwell-Poisson and Double-Poisson Distributions, *91st TRB Annual Meeting*. *January 22-26*. (2012).

[12] R.W. Conway, W.L.A. Maxwell, Queuing Model with State Dependent Service Rates, *Journal of Industrial Engineering*. **12**, 132-136 (1962).

[13] D. Lord, S.D. Guikema, S. Geedipally, Application of the Conway-Maxwell- Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes, *Accident Analysis & Prevention*. **40**, Issue 3, 1123-1134 (2008).

[14] K.F. Sellers, G. Shmueli, A Flexible Regression Model for Count Data, *The Annals of Applied Statistics*. **4,** Issue 2, 943-961(2010).

[15] N. Ismail, H. Zamani, Estimation of Claim Count Data using Negative Binomial, Generalized Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Generalized Poisson Regression Models. *Casualty Actuarial Society E-Forum*. (Spring 2013).