Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Esin and Emel, J Biom Biostat 2016, 7:1
DOI: 10.4172/2155-6180.1000278

Research Article Open Access

# Comparison of the Count Regression Models in Evaluation of the Effects of Hazelnut Harvest Season Variations on Pulmonary Aspergillus

**Esin A[1]* and Emel U[2]**

[1]*Department of Applied Statistics, Giresun University, Giresun, Turkey*
[2]*Department of Medical Microbiology, Giresun University, Giresun, Turkey*

## Abstract

Pulmonary aspergillosis has recently emerged as a worldwide health care problem especially in patients with underlying lung disease. The objective of this study was to compare the Poisson and COM-Poisson regression models and to find the best fitted model for determining the effect of hazelnut harvest season on pulmoner aspergillosis. The data obtained from the state hospital of our cityin the time period of two years, from September, 2012 to August, 2014. A retrospective study was conducted. Respiratory specimens which showed repeated isolation of Aspergillus were included in the study however only one of the samples was analysed. Cases were classified according to revised definitions given by European Organization for Research and Treatment of Cancer/Invasive Mycosis Study Consensus Group (EORT/MSG). Culture positive 36 patients were detected from 3457 patients. Poisson and Conway-Maxwell-Poisson (COM-Poisson) regression models were compared to determine the best fitted model for identifying the number of new pulmonary aspergillosis cases in hazelnut harvest season. To describe the best fitted model of count data, dispersion, deviance and Akaike Information Criteria (AIC) test statistics were used. Based on statistical test for dispersion, the under-dispersion was found non-significant. This results clearly indicate that Poisson regression model is more approtiate for pulmonary aspergillosis data when compared to COM-Poisson regression model. Deviance and AIC values also confirm this result. Poisson regression model and COM- Poisson regression model were compared with statistical tests. According to statistical tests Poisson regression model was found to be the best fit model for pulmonary aspergillosis data.

**Keywords:** AIC; Poisson regression; COM-Poisson regression; Pulmonary aspergillosis; Hazelnut harvest season.

## Introduction

Pulmonary aspergillosis has recently emerged as a worldwide health care problem especially in patients with underlying lung disease. Aspergillus spp. is a saprophytic and pathogenic fungus with a cosmopolitan distiribution. It is best known for its colonisation of tree nuts. Aspergillus infections can occur while hosts are still in the field, but often show no symptoms until postharvest storage and/or transport [1,2].

In Turkey on Black Sea Cost, hazelnuts are harvested annually in late August and September. During the harvest season spores of fungi disperse with the wind and many COPD and asthma patients apply to the hospital with exacerbation of the disease secondary to infection following the harvest. These exacerbations cause extensive use of broad spectrum antibiotics, immunosuppressive agents and increasing population of terminally ill patients.

In our study the data obtained from a state hospital in the time period of two years, from September, 2012 to August, 2014. A retrospective study was conducted. Demographic details, clinical and radiological findings, predisposing factors and treatments were noted down. Cases were classified as proven, probable and possible invasive Aspergillosis according to revised definitions given by European Organization for Research and Treatment of Cancer/Invasive Mycosis Study Consensus Group (EORT/MSG) [3]. According to the suggestion of this group, all culture positive patients and those fullfilling the following clinical criteria were included in the study; 1) Long term sterioid therapy 2) radiological features suggestive of Aspergillosis 3) Microbiological findings. To rule out the changes of contamination respiratory specimens which showed repeated isolation of Aspergillus were included in the study however only one of the sample taken for analysis. To demonstrate the relationship with hazelnutfarm harvest season (August and September), number of the patients were divided into two part according to their hospital visits whether in the nuts picking season or not. Culture positive 36 patients were detected.

All samples were identified as Aspergillus spp. morphologically. Patients were on steroid treatment at a minimum dose of 0.3 mg/kg/day of prednisone equivalent for more than three weeks. They had at least one radiological findings supporting the diagnosis of pulmonary aspergillosis. Despite the antifungal therapy 22.2% of the patients died. In clinical work the researcher often encounters situations where the outcome variable is numeric, but in the form of counts. Often it is a count of rare events for a certain period of time. To explore the relationship between dependent variable and some or all of the explanatory variables regression analysis are used [4]. If the dependent variable characterized by rare events count data and all the counts are positive integers Poisson distribution is the most common for fitting. The main assumption of Poisson distribution is that the mean and variance are equal (equidispersion). Unfortunately many real data do not adhere this assumption [5]. In case of violation of assumption, two-parameter generalized form of the Poisson distribution, called the Conway-Maxwell-Poisson (COM-Poisson) distribution allows for over (i.e., the variance is larger than the mean)-and under (i.e., the variance is smaller than the mean) dispersion [6].

***Corresponding author:** Esin A, Faculty of Art and Science, Department of Applied Statistics, Giresun University, Giresun, Turkey, Tel: +90 454 310 1000; E-mail: esinavci@hotmail.com

**Received** November 30, 2015; **Accepted** January 02, 2016; **Published** January 16, 2016

**Citation:** Esin A, Emel U (2016) Comparison of the Count Regression Models in Evaluation of the Effects of Hazelnut Harvest Season Variations on Pulmonary Aspergillus. J Biom Biostat 7: 278. doi:10.4172/2155-6180.1000278

In this study, we compared the Poisson to COM-Poisson regression model and illustrate its usefulness by using two years period of pulmonary aspergillosis data.

## Poisson regression models

Poisson regression is one of the members of the Generalized Linear Models (GLM) framework. The simplest distribution used for modeling count data is the Poisson distribution with probability density function

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

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 ($\mu$) = $\mu$. 7 The mean Poisson regression can be assumed to follow a log link, $E\ Yi = \mu i = (xi'\beta)$, where $xi$ 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.

## Conway-Maxwell-Poisson (Com-Poisson) models

the Conway-Maxwell-Poisson (COM-Poisson) distribution has been re- introduced by statisticians to model count data characterized by either over- or under- dispersion [6-10]. The COM-Poisson distribution was first introduced in 1962 by Conway and Maxwell; in 2008 Guikema et al. [8] evaluated it in the context of a GLM, in the same year Bayesian approach of COM-Poisson are exteneded by Lord et al. [9]. Sellers and Shmueli [11] enlarged the application area by developing COM Poisson Reg code in R package program. The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that is flexible enough to describe a wide range of count data distributions [11].

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

$$f(y;\lambda,\upsilon) = \frac{\lambda^{y}}{(y!)^{\upsilon}\, z(\lambda,\upsilon)} \qquad (2)$$

for a random variable Y, where Z($\lambda$,$\upsilon$) = and $\upsilon \geq 0$ is a normalizing constant; $\upsilon$ 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,< 1), and Bernoulli well-known distribution as special cases: Poisson ($\upsilon$ = 1), Geometric ($\upsilon$ = 0, λ < 1),

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

$$\eta(E(Y) \ =\ log\,\lambda =\ X'\beta = \beta_0 + \sum\nolimits_{j=1}^{p} \beta_j X_j \qquad (3)$$

Because of indirectly relationship between ($Y$) $'\beta$, estimating β and $\upsilon$ via associated normal equations become complex. Using $^{(0)}$ and $^{(0)}$ = 1, as starting values. These equations can thus be solved via an appropriate iterative reweighted least squares procedure (or by maximize

ing the likelihood function directly using anoptimization 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 [12].

## Testing for variable dispersion

In statistical analysis of count data, if the variance of the random variable is constrained to equal the mean which is called equidispersion, the Poisson regression model is usually adequate. Because of different factors can invalidate equidispersion hypothesis in the data its implicit restriction: consequently, data often exhibit overdispersion (i.e., the variance is larger than the mean) and, occasionally they exhibit underdispersion whith the mean exceeding the variance [13].

In GLM to detect over or under-dispersion simply, the researcher may look at the rule of thumb that the mean deviance, that is deviance/ degree of freedom should be close to unity. Deviance theoretically allows one to determine if the fitted GLM model is significantly worse than the saturated model [14].

Sellers et al. [11] established a hypothesis testing procedure to demonstarting the need for a COM-Poisson regression model over a simple Poisson regression model by determine if significant data dispersion exists or not, in other words, they test whether ($\upsilon$ = 1) or otherwise [9]. The test statistics,

$$C = -2\log\Lambda = -2\left[\log \mathrm{L}\left(\hat{\beta}^{(o)},\hat{\upsilon}=1\right) - \log \mathrm{L}\left(\hat{\beta},\hat{\upsilon}\right)\right] \qquad (4)$$

where Λ is the likelihood ratio test statistic, $\beta$ 0 are the maximum likelihood estimates obtained $\hat{\beta}^{(o)}$ under : $\upsilon$ = 1 (i.e., the Poisson estimates), and $\left(\hat{\beta},\hat{\upsilon}\right)$, $\upsilon$ are the maximum likelihood estimates under the general state space for the COM-Poisson distribution with 1 degree of freedom. For small samples, the test statistic distribution can be estimated via bootstarp [11].

## Akaike information criteria (AIC)

In statistical literatures, based on several likelihood measure, one can compare several models performance. One of the most regularly used measure is AIC. The AIC penalized a model with larger number of parameters, and is defined as

$$AIC = -2lnL + 2p \qquad (5)$$

where $lnL$ denotes the fitted log likelihood and $p$ the number of parameters [15]. A relatively small value of AIC is prefered for the fitted model. Analyses were performed using R program. Respectievely, glm () function from "stats" and cmp() function from "COMPoissonReg" package were used.

## Result and Conclusion

To obtain the first overview of the dependent variable, the histogram of the observed count frequencies were employed. Figure 1 illustrates the histogram, the marjinal distribution exhibits substantial variation (Figures 1 and 2).

Figure 2 presents line plot of total number of patients with positive culture. From the plot we could conclude that, especially in August and September the number of patients with positive culture were increased.

Poisson regression model was fitted to the data, regressing the number of patients with positive culture (Y) on hazelnutfarm harvest season (X) (1: Yes, 0: No); the estimated coefficients were given by $\hat{\beta}_o$ = −0.1054, $\beta$1 = 1.6094, then COM- Poisson regression model was fitted to determine whether equi-dispersion was a reasonable assumption. After dividing the COM-Poisson coefficients by $\upsilon$ dispersion parameter (-2.1822/1.5082 = 1.4469), the results in Table 1 indicate that the regression n parameters for two models had similar estimates in terms of the coefficient magnitudes.

While the estimated dispersion parameter for Poisson regression model is 0.84, for COM-Poisson model is $\upsilon$ = 1.51, indicating under-dispersion. To determine whether the dispersion parameter is

| | Classic Poisson | | | COM-Poisson | | |
|---|---|---|---|---|---|---|
| | **Estimated Coefficent** | **Standart Error** | **z value** | **Estimated Coefficent** | **Standart Error** | **z value** |
| Intercept Season<br>Dispersion parameter | - 0.1054 | 0.2357 | - 0.4472 | 0.1437 | 0.4007 | 0.3586 |
| | 1.6094 | 0.3333 | 4.8287* | 2.1822 | 0.8569 | 2.5466* |
| | 0.84 | | | 1.51 | [95%CI = (0.70;4.02)] | |
| Deviance AIC | | 18.478 | | | 26.988 | |
| | | 68.065 | | | 69.382 | |

**Table 1:** Estimated regression models for patients with positive culture.

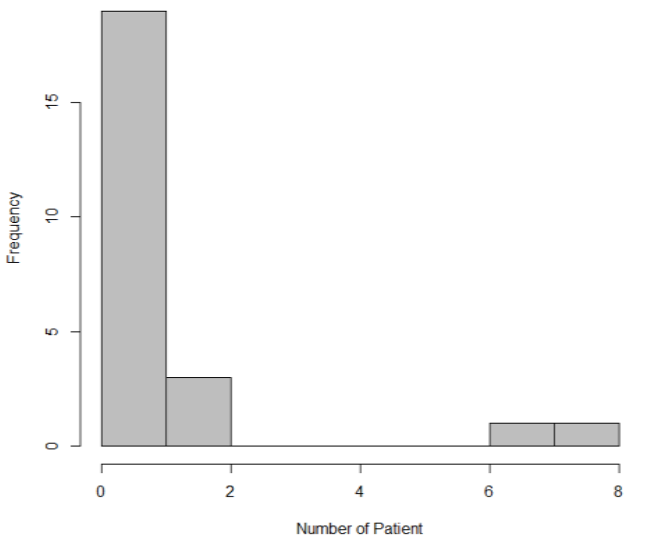

**Figure 1:** Frequency of patient with positive culture.

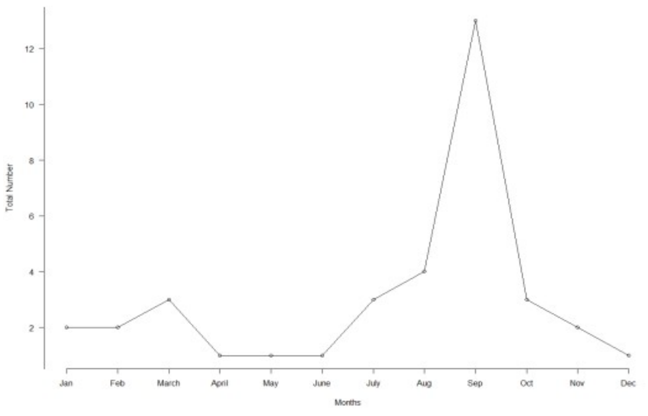

**Figure 2:** Line plot of patient with positive culture.

significant or not a hypothesis test which established by Sellers and Shmueli [11] was used. The p value was found 0.4085, and the 95% bootstrap confidence interval for $\upsilon$ include the value 1 (using 1000 samples). Indicating that, it is reasonable to assume a Poisson to model this relationship. Table 1 display estimated regression models for patients with positive culture.

According deviance and AIC, the Poisson regression model is better than the COM-Poisson regression model. In terms of model interpretation, the Poisson regression indicates that pulmonary aspergillosis is efected from hazelnut harvest season about five times more.

In this paper Poisson regression model was compared with COM-Poisson model for modeling pulmonary aspergillosis data. We revealedthat under-dispersion was not statistically significant and thus it was reasonable to assume a Poisson model to investigate the effects of hazelnut harvest season on pulmonary aspergillosis.

## Acknowledgement

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

## References

1. Zmeili OS, Soubani AO (2007) Pulmonary aspergillosis: A clinical update. QJM 100: 317-334.
2. Bayman P, Baker JL, Mahoney NE (2002) Aspergillus on tree nuts: incidence andassociations. Mycopathologia. 155: 161-169.
3. Pauw DB, Walsh TJ, Donnelly JP, Stevens DA, Edwards JE, et al. (2008) Revised definitions of invasive fungal disease from the European Organization for Research and Treatment of Cancer/Invasive Fungal Infections Cooperative Group. The National Institute of Allergy and Infectious Diseases Mycoses Study Group (EORTC/MSG) Consensus Group. Clin Infect Dis 46: 1813-1821.
4. Research methods-II: Multivariate Analysis. Poisson Regression Analysis pp: 136-143.
5. Sellers KF, Borle S, Shumeli G (2012) The COM-Poisson model for count data: a survey of methods and applications. Applied Stochastic Models in Business and Industry 28: 104-116.
6. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005) A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. Appl Statis 54: 127-142.
7. Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. Journal of Statistical Software 27: 1-25.
8. Guikema SD, Coffelt JP (2008) A flexible count data regression model for risk analysis, Risk Analysis 28: 213-223.
9. Lord D, Geedipally SR, Guikema SD (2010) Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting under-dispersion. Risk Analysis 30: 1268-1276.
10. Zou Y, Lord D, Geedipally SR (2012) Over-and under-dispersed crash data: comparing the conway-maxwell-poisson and double-poisson distributions. 91st TRB Annual Meeting pp: 22-26.
11. Sellers KF, Shmueli G (2010) A Flexible Regression Model for Count Data, The Annals of Applied Statistics 4: 943-961.
12. Sellers KF, Shmueli G (2013) Data dispersion: now you see it...now you don't. communication in statistics: Theory and Methods 42: 3134-3147.
13. Giuffre O, Grana A, Giuffre T, Marino R (2013) Accounting for dispersion and correlation in estimating safety performance functions. An overview statrting fom a case study. Modern Applied Science 7: 11-23.
14. Myers RH, Montgomery DC, Vining GG, Generalized Linear Models with Applications in Engineering and the Science, John Wiley & Sons.
15. Ismail N, Zamani H (2013) 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.