Journal of Data Science 511-530 , DOI: 10.6339/JDS.201807_16(3).0004

Using Count Regression Models to Determine the Factors which Effects the Hospitalization Number of People with Schizophrenia

Esin Avcı*

a Department of Statistics, University of Giresun, Giresun, 28000 TURKEY,

E-mail:esinavci@hotmail.com

*Abstract:*In medical literature, researchers suggested various statistical procedures to estimate the parameters in claim count or frequency model. In the recent years, the Poisson regression model has been widely used particularly. However, it is also recognized that the count or frequency data in medical practice often display over-dispersion, i.e., a situation where the variance of the response variable exceeds the mean. Inappropriate imposition of the Poisson may underestimate the standart errors and overstate the significance of the regression parameters, and consequently, giving misleading inference about the regression parameters. This article suggests the Negative Binomial (NB) and Conway-Maxwell-Poisson (COM-Poisson) regression models as an alternatives for handling overdispersion. All mentioned regression models are applied to simulation data and dataset of hospitalization number of people with schizophrenia, the results are compared.

*Keywords:*Poisson regression; Negative Binomial regression,COM-Poisson regression, hospitalization number of schizophrenia patients

1 Introduction

Schizophrenia, is one of the most disabling and emotionally devastating illnesses known to man. However, because they have been misunderstood for so long, they have received relatively little attention and have adversely affected victims. Schizophrenia is not a split personality, a rare and very different disorder. Like cancer and diabetes, schizophrenia has a biological basis; it is not caused by personal weakness or bad parenting. Although there is no known treatment for schizophrenia, it is a very treatable disease. Most schizophrenia patients respond to drug treatment and many can continue to be productive and satisfying (http://www.schizophrenia.com/family/schizintro.html).Schizophrenia is, in fact, a relatively common disease, there are at least 500,000 schizophrenia patients in Turkey, one of the country's leading medical associations has announced on World Mental Health Day.

Most people with schizophrenia (or another serious psychiatric disorder) will likely have to be hospitalized for at least a short time. The hospitalization may be voluntary (requested by the patient himself) or may be involuntary, meaning it is up to the discretion of the treating psychiatrist, emergency room staff, or a courtroom. At the point of hospitalization, a person may be in very bad patterns-the patient feels frightened, out of control and abandoned. The benefits of hospitalization could be summarized as follows; to keep the patients from harming themselves or others, monitored by trained professionals for symptoms and medication reactions, providing a safe place for patients to concentrate on concentrating and recovering themselves and making long-term treatment plans(http://www.schizophrenia.com/family-/schizintro.html).

Monitoring the trends in hospitalization numbers and conceive the factors that effect the hospitalization numbers are significant aspects of schizophrenia (or another serious psychi-atric disease) surveillance studies.

Data on the number of hospitalization is count, to determine the factors that affects Hospitalization number, regression analysis of counts is used. Count data usually analyzed with the Poisson regression. The characteristics of the Poisson regression mean and variance must be the same, whereas in fact the count data is often becoming variance greater than the mean, which is often reffered to over-dispersion. To deal with the problem over-dispersion, modelling can be done with Negative Binomial (NB) and a flexiable alternative that captures both over-and under-dispersion is the Conway-Maxwell-Poisson (COM-Poisson) regression because it does not require the mean value equal to the value of variance.

The main objective of this study is to assess the effects of demographic factors, habits, and relation with family and environment on the hospitalization number of people with schizophrenia by using the Poisson, NB and COM-Poisson regression models.

This article is organized as follows, Section 2 briefly describes the Poisson, Negative Binomial and COM-Poisson regression methods respectively. Section 3 presents simulation data and dataset on hospitalization number of people with schizophrenia, where a sample of over-dispersed data. All mentioned regression models are applied and the results are compared. Section 4 presents conclusion.

2 Count Regression Models

The number of occurrence of any event within a specified time, can be described as counting data. In case of dependent variable is a count and researcher is interested in how this count changes as the explanatory variable increases count data regression model is used. Modelling count data has been widely used in actuarial sciences, Aitkin et al. (1990) and Renshaw (1994) fitted Poisson regression to two different set of U.K. motor claim data, and in biostatistics and demography, Frome (1983) modelled the lung cancer death rates among British physicians who were regular cigarette smokers. In recent years this model has been used frequently in economy, political science and sociology, Lord (2006) modeled motor vehicle crashes by using Poisson-Gamma model and Riphahn et al. (2003) fitted the model to German Socioeconomic Panel (GSOEP) data. Famoye and Singh (2006) proposed

a zero-inflated generalized Poisson (ZIGP) regression model to model domestic violence data with too many zeros. Pararai et al. (2010) derived the Generalized Poisson-Poisson mixture regression (GPPMR) model to handle accurate, underreported and overreported counts. Famoye et al. (2004) applied generalized Poisson regression (GPR) model for identifying the relationship between the number of accidents and some covariates. Ozmen and Famoye (2007) applied the Poisson, NB, GP, ZIP and ZIGP to zoological data set where the count data may exhibit evidence of many zeros and over-dispersion.

Because of counts are all positive integers and for rare events the Poisson distribution (rather than the normal) is appropriate. However, it is suitable only for modeling equi-dispersed (i.e., an equal mean and variance) distribution. Many real data do not adhere to this assumption (over- or under-dispersed data) and inappropriate imposition of Poisson regression model may underestimate the standard errors and overstate the significiance of regression coefficients. For over-dispersed data, the Generalized Estimation Equations (GEEs) method with Negative Binomial distribution is a popular choice and increase efficiency of estimates (Lord 2006, Hilbe 2011). Other overdispersion models include Poisson mixtures (McLachlan, 1997) and quasi-Poisson model is characterized by the first two moments (mean and variance) (Wedderburn 1974). However, these models are not suitable for under-dispersed data. Under-dispersed data is less commonly observed. In cases the sample is small and the sample mean is very low and can be caused by the data generating process that is independent from the sample size or mean (Oh et al. 2006). A few models exist that allow for both over- and under-dispersed data. One example is the restricted generalized Poisson regression models of Famoye (1993). It is called "restricted" model, because it belongs to an exponential family under the condition that the distribution parameter is constatnt. The Conway-Maxwell-Poisson (COM-Poisson) regression model is an alternative model to fit data sets of varying dispersion. COM-Poisson distribution is a two parameter generalization of the Poisson which also includes the Bernoulli and geometric distribution that allows for over- and under-dispersion. The distribution was briefly introduced by Conway and Maxwell in 1962 for modeling queuing systems with state-dependent service rates. The statistical properties of the COM-Poisson distribution, as well as methods for estimating its parameters were established by Shamueli et al. 2005. The COM-Poisson distribution has been used in a variety of count data application. (Consul 1989, Consul and Famoye 1992, Famoye et al. 2004, Wang and Famoye 1997, Conway and Maxwell 1962, Shmueli et al. 2005, Lord et al. 2008, Lord et al. 2010, Khan and Khan 2010). Benson and Friel (2017) provided a new rejection sampler for the COM-Poisson distribution which significantly reduces the CPU time required to perform inference for COM-Poisson regression. Saghir and Lin (2013) proposed Shewhart-type multivariate control chart to monitor multivariate COM-Poisson (MCP) chart, based on the MCP distribution. Chanialidis et al. (2017) illustrated the method and the benefits of using a Bayesian COM-Poisson regression model, through a simulation and two real-world data sets with different levels of dispersion.

Count data are commonly modeled with the Poisson distribution, with both mean and variance being equal. Due to heterogeneity (difference between individuals) and contagion (dependence between the occurrence of events), the variance is usually broader than the average, which makes the Poisson assumption more restrictive. By placing a Gamma distribution prior a Negative Binomial (NB) can be generated. Therefore, the NB distribution is also known as the Gamma-Poisson distribution. Because of the variance larger than the mean, the NB usually favored over the Poisson distribution for modeling over-dispersed counts (Zhou et al. 2012).

The regression analysis of counts is commonly performed under the Poisson and NB likelihoods, whose parameters are usually estimated by finding the maximum of the nonlinear log likelihood (Long 1997, Cameron and Trivedi 1998, Agresti 2002, Winkelmann 2008).

A flexiable alternative that captures both over-dispersion and under-dispersion is the 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 (Shmueli et al. 2005). The COM-Poisson distribution has been used in a variety of count data applications and has been extended methodologically in various directions (Sellers et al. 2012).

2.1. Poisson Regression Model

The most basic regression model for counts is the Poisson regression model (Long 1997, Cameron and Trivedi 1998, Agresti 2002, Winkelmann 2008), which can be expressed as

$$y_i \sim Pois(\lambda_i), \qquad \lambda_i = exp(x_i^T \beta) \tag{1}$$

where $x_i = [1, x_{i1}, \dots, x_{iP}]^T$ is the covariate vector for sample i. The Newton-Raphson method can be used to iterarively find the maximum log likelihood of β (Long, 1997). A serious constraint of the Poisson regression model is that it assumes equal-dispersion, i.e., $E[y_i|x_i] = Var[y_i|x_i] = exp(x_i^T \beta)$. In practice, however, count data are often overdispersed, due to heterogeneity and contagion (Winkelmann, 2008). To model overdispersed counts, the poisson regression model can be modified as

$$y_i \sim Pois(\lambda_i) \qquad \lambda_i = exp(x_i^T \beta)\epsilon_i \tag{2}$$

where ϵ_i is a nonnegative multiplicative random-effect term to model individual heterogenety (Winkelmann, 2008).

2.2. Negative Binomial Regression Model

The Negative Binomial model (NB) (Long 1997, Cameron and Trivedi 1998, Winkelmann 2008, Hilbe 2007) is constructed by placing a Gamma prior on ϵ_i as

$$\epsilon_i \sim Gamma\left(r, \frac{1}{r}\right) = \frac{r^r}{\Gamma(r)} \epsilon_i^{r-1} e^{-r\epsilon_i} \tag{3}$$

where $E[\epsilon_i] = 1$ and $Var[\epsilon_i] = r^{-1}$. Marginalizing out ϵ_i in (2), we have a NB distribution

parameterized by mean $\mu_i = exp(x_i^T\beta)$ and inverse dispersion parameter ϕ (the reciprocal of r) as $f_Y(y_i) = \frac{\Gamma(\phi^{-1}+y_i)}{y_i!\Gamma(\phi^{-1})}(\frac{\phi^{-1}}{\phi^{-1}+\mu_i})^{\phi^{-1}}(\frac{\mu_i}{\phi^{-1}+\mu_i})^{y_i}$, thus

$$E[y_i|x_i] = exp(x_i^T\beta) \quad (4)$$

$$Var[y_i|x_i] = E[y_i|x_i] = \phi E^2[y_i|x_i] \quad (5)$$

The maximum log likelihood of β and ϕ can be found numerically with the Newton-Raphson method (Lawless, 1987).

2.3. 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 (Shmueli et al. 2005, Guikema and Coffelt 2008, Lord et al. 2010, Zou and Lord 2012). The COM- Poisson distribution was firstintroduced in 1962 by Conway and Maxwell; only in 2008 it was evaluated in the context of a GLM by Guikema and Coffelt (2008), Lord et al. (2008) and Sellers and Shmueli (2010). 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 (Sellers et al. 2010); since its revival, it has been further developed in several directions and applied in multiple fields (Sellers et al. 2012).

$$P(y;\lambda,\upsilon) = \frac{\lambda^y}{(y!)^\upsilon Z(\lambda,\upsilon)} \quad (6)$$

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 $(\upsilon \to \infty\, with\, probability\, \frac{\lambda}{1+\lambda})$ (Shmueli et al. 2005).

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

$$\eta(E(Y) = log\,\lambda = X'\beta = \beta_0 + \sum_{j=1}^{p}\beta_j X_j \quad (7)$$

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 (Sellers et al. 2013).

2.5. Testing for Variable Dispersion

Sellers and Shmueli (2010) 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 (Sellers et al. 2013).

Since NB regression reduce to Poisson regression in the limit as $\theta \to 0$, 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 (Ismail and Zamani, 2013).

2.6. Akaike Information Criteria (AIC)

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

$$AIC = -2lnL + 2p \tag{8}$$

where lnL denotes the fitted log likelihood and p the number of parameters (Ismail and Zamani, 2013). A relatively small value of AIC is favorable for the fitted model.

3 Application

3.1. Simulated Data

To demonstrate the flexibility of the COM-Poisson distribution, 500 data are derived from Poisson, Negative Binom and COM-Poisson, respectively. The inversion method is particularly simple to sample an integer value from the COM-Poisson distribution. The COM-Poisson probabilities are summed up starting from $P(Y = 0)$, until this sum exceeds the value of a simulated Uniform(0,1) variable. Y is then an observation from the COM-Poisson distribution (Minka et al, 2003). Goodness of fit (associated p-values provided in parentheses) of each distribution and estimated parameters are given in Table 1. The analysis are performed in R program. glm() function from "stats" package and cmp() function from "COMPoissonReg" package are used, respectievely.

Table 1 illustrated that, while the Poisson distribution was meaningful only for identical distribution, the Negative Binom distribution was flexible for many distributions except under-dispersion. However, COM-Poisson distribution was flexible for all considered distributions. Furthermore the estimated parameters almost same for Poisson and Negative Binomial distributions.

Table 1: Goodness of fit on simulated data of size 500

Distribution	Estimated Parameter		
	Poisson	Negative Binom	COM-Poisson
Poisson ($\boldsymbol{\lambda = 4}$)	λ =4.002	λ=4.002, θ=29.262	λ =3.243, υ=0.862
	(0.105)	(0.231)	(0.561)
Negative Binom	λ =3.39	λ=3.39, θ=5.51	λ =1.775, υ=0.538
($\boldsymbol{\lambda}$ =4, θ=6)	(0.000)	(0.556)	(0.999)
COM-Poisson ($\boldsymbol{\lambda}$=12, $\boldsymbol{\upsilon}$=5)	λ =1.242	λ=1.242, θ=2.333	λ =13.916, υ=5.150
	(0.000)	(0.000)	(0.651)
COM-Poisson ($\boldsymbol{\lambda}$=3, $\boldsymbol{\upsilon}$=0.4)	λ =16.078	λ=16.078, θ=12.195	λ =2.634, υ=0.362
	(0.000)	(0.658)	(0.987)

To illustrate the flexibility of the COM-Poisson model, the considered distributions (Poisson, Negative Binom and COM-Poisson) are above regressed with arbitrarily chosen single explanatory variable X and determined coefficients (β_0=3, β_1=0.5). By fitting the considered distribution (Poisson, Negatif Binom and COM-Poisson) the estimated model parameters, Log likelihood and AIC values are given in Table 2.

Under the response variable had Poisson distribution, the Poisson regression model is given the smallest Log-likelihood and AIC values. Hence, the Poisson regression model was the best fit among the estimated models for statistically significant coefficients. The second best fit model for statistically significant coefficients was the COM-Poisson regression model. The dispersion parameter (υ) was estimate quite near to 1. For negative Binomial distributed responce (over-dispersed data), COM-Poisson and negative Binom are estimated the model parameters, Log likelihood and AIC values almost the same. The COM-Poisson regression model is estimated the dispersion parameter (υ) as a 0.524. Finally for COM-Poisson distributed responce, the Poisson and negative Binom regression models are obtained similar estimation results, the explanatory variable was not statistically significant. It is concluded from the Table 2, under different distributed responce variable the COM-Poisson regression model was the best model.

Table 2: Estimated model parameters and AIC for simulated data of size 500

Distribution	Estimated Parameter		
	Poisson	Negative Binom	COM-Poisson
Poisson ($\boldsymbol{\beta_0}$=3, $\boldsymbol{\beta_1}$=0.5)	$\hat{\beta}_0$=2.997(0.018)* $\hat{\beta}_1$=0.482(0.029)* Loglik=-1516.585 AIC=3037.2	$\hat{\beta}_0$=2.997(0.018)* $\hat{\beta}_1$=0.482(0.029)* Loglik=-1516.586 AIC=3039.2 $\hat{\theta}$=77302	$\hat{\beta}_0$=2.997(0.195)* $\hat{\beta}_1$=0.482(0.042)* Loglik=-1516.585 AIC=3039.171 υ=1.00
Negative Binom ($\boldsymbol{\beta_0}$=3, $\boldsymbol{\beta_1}$=0.5, θ=5)	$\hat{\beta}_0$=2.927(0.048)* $\hat{\beta}_1$=0.533(0.077)* Loglik=-1168.754 AIC=2341.5	$\hat{\beta}_0$=2.939(0.060)* $\hat{\beta}_1$=0.524(0.099)* Loglik=-1129.293 AIC=2264.6 $\hat{\theta}$=5.622	$\hat{\beta}_0$=1.590(0.073)* $\hat{\beta}_1$=0.505(0.063)* Loglik=-1131.30 AIC=2268.599 υ=0.524
COM-Poisson ($\boldsymbol{\beta_0}$=3, $\boldsymbol{\beta_1}$=0.5, $\boldsymbol{\upsilon}$=5)	$\hat{\beta}_0$=1.366(0.072)* $\hat{\beta}_1$=0.163(0.123) Loglik=-630.835 AIC=1265.7	$\hat{\beta}_0$=1.366(0.073)* $\hat{\beta}_1$=0.163(0.123) Loglik=-630.838 AIC=1267.7 $\hat{\theta}$=107816	$\hat{\beta}_0$=3.101(0.286)* $\hat{\beta}_1$=0.659(0.250)* Loglik=-452.243 AIC=910.486 υ=5.274

* 5% statistically significant

3.2. Hospitalization Number of People with Schizophrenia Data

In this article, the data which is based on 205 schizophrenia patients for a four-year period of 2011-2014, is obtained from Community Mental Health Center is located within Prof. Dr. A. Ilhan Ozdemir state hospital in Giresun. Besides age, Table 3 showed the rating factors and rating class for the number of hospitalization.

Table 3. Rating factors and rating classes for the number of hospitalization

Rating Factors	Rating Classes
Gender	Male
	Female
Education Status	Illiterate
	Primary School
	Secondary School
	High School
	College
	Undergraduate
	Graduate
Marital Status	Single
	Married
	Divorced
	Widow
Income Status	Unavailable
	Working
	Someone is looking
	State protection
	Retired
Urban Status	City
	Bent
Live Alone	Yes
	No

Rating Factors	Rating Classes
Family Disease	**No**
	Yes
Substance Abuse	No
	Yes
Relation with Family and Environment	Not good
	Not good not bad
	Good
Activity Status	Passive
	Active

The sample consisted of 65% men and 35% women. The patients ranged in age from 17 to over 78 years. Most of the were single (55%). While the most observed level of education was primary school (42%), the least observed was graduate level education (0.5%). In terms of income statu, the patients were 34% under state protection and 9% working. About 70% of the patients were lived in city. About 92% of the patients were not lived alone. In terms of family disease, 46% of the patients were had family disease history. About 56% of the patients were substance abused, 49% of the patients were had good relationship with the family and enviroment. Finally about 70% of patients were had active life.

The number of hospitalization ranged from 0 to 35 with a mean 2.73 (variance 11.60). The distribution of the number of hospitalization is presented in Figure 1 is skewed to the right. In other words, few patients are hospitalized 10 or more times. The modal number of hospitalization were zero and one hospitalization (21%) followed by two hospitalization (17%).

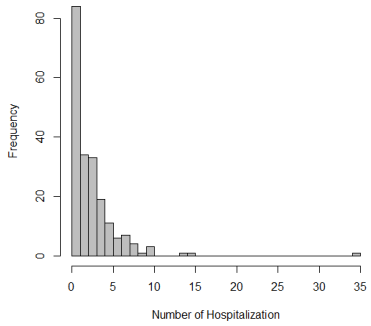


Figure 1: Distribution of Hospitalization number

All the models described above are applied to data set. At the end of the section, all fitted models are compared highlighting that the modelled mean function was similar but the fitted likelihood and AIC were different. The analysis are performed in R program. glm() function from "stats" package, glm.nb() function from "MASS" package and cmp() function from "COMPoissonReg" package are used, respectievely.

Several models are fitted by including different rating factors; first the main effects only, then the main effects plus each of the paired interaction factors to assess the magnitude of the effect of several factors acting jointly over and above their effects are considered seperately. 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 was not significant).

Turning first to the main effects model, Table 4 showed the results of parameter estimates, standart error, AIC and Log likelihood.

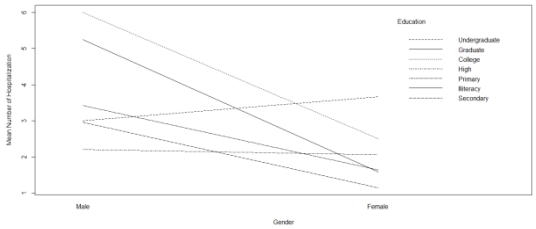
The formal test for significance of over-dispersion, the log-likelihood ratio, which is -2×(log-likelihood of Poisson regression-log-likelihood of Negative Binomial regression), is computed. The log-likelihood ratio is became 143.58, which exceeded critical value of chi-square, giving evidence of over-dispersion. The estimated dispersion parameter for COM-Poisson model was $\upsilon = 0.1289$, indicating over-dispersion. To test the significance of dispersion parameter a hypothesis test which established by Sellers and Shmueli (2010) is used. The p value found 0, indicating over-dispersion that requires a COM-Poisson regression instead of Poisson regression. Based on both tests, evidence of over-dispersion indicates inadequate fit of the Poisson model. The best model for the main effects models is chosen according to smallest log-Liklihood and AIC values. In terms of Log likelihood and AIC, the Negative Binomial model showed best fit for main effects model.

Table 4. Parameter estimates, standart error and AIC value for main effects models

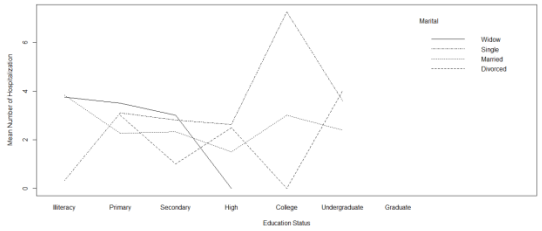
	Classic Poisson		Negative Binomial		COM-Poisson	
Factors	Estimated Coefficent	Standart Error	Estimated Coefficent	Standart Error	Estimated Coefficent	Standart Error
Intercept	0.0428	0.2349	-0.0893	0.3682	-0.4592	0.1467
Age	0.0221	0.0221*	0.0247	0.0065*	0.0076	0.0025*
Gender	-0.4779	0.1043*	-0.4881	0.1578*	-0.1794	0.0686*
Education Statu	0.0845	0.0317*	0.1082	0.0519*	0.0286	0.0187*
Substance Abuse	0.2760	0.0954*	0.2752	0.1501*	0.0981	0.0589*
Relation with Family and Environment	-0.1807	0.0505*	-0.2070	0.0827*	-0.0599	0.0304*
Dispersion parameter	-	-	1.821	0.319	0.1289	0.0549
Log-liklihood	-498.56		-426.77		-431.10	
AIC	1009.1		867.54		876,19	

* $p<0.05$ (significant covariate)

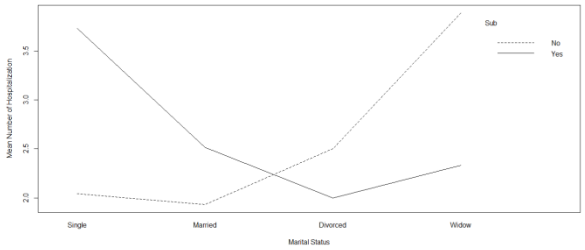
The significance of the interactions effect are looked at by adding them into the main effects model one at a time and retaining the significant interactions. Accordingly, all the three-way and higher-level interactions effects are obtained non-significant. From the two-way interactions only gender and education status, education status and marital status, and marital status and substance abuse were significant. The interaction plots are also used to assess the effect a pair of factors has on the response by plotting, for each value of one of the factors, a line between mean response at the low level of the other factor to the mean response at the high level. An interaction effect is indicated when the lines for different levels of the first factor have unequal slopes. The plots in Figure 2 confirmd the presence of interactions gender and education status, education status and marital status, and marital status and substance abuse. The first plot (a) indicated a decrease in the mean number of hospitalization from the males to the females almost all of the education status except Undergraduate. The second plot (b) revealed that for single category the mean number of hospitalization increased as their education status increased from illeteracy to college and decrased at undergraduate, for widow category the mean number of hospitalization decreased as their education status increased from illeteracy to high scholl. The Third plot (c) revealed that for non-substance abuse category the mean number of hospitalization decreased as marital status changed from the single to the divorced and increased at widow class.



(a)



(b)



(c)

Figure 2: Interaction plots

Table 5 showed the results of parameter estimates, standard error, AIC and Log likelihood for the main effects plus interaction effects model.

Table 5. Parameter estimates, standart error and AIC value for models

	Classic Poisson		Negative Binomial		COM-Poisson	
Factors	Estimated Coefficent	Standart Error	Estimated Coefficent	Standart Error	Estimated Coefficent	Standart Error
Intercept	0.0451	0.2144	0.0890	0.3200	-0.4294	0.1427*
Age	0.0188	0.0042*	0.0178	0.0066*	0.0071	0.0027*
Gender	-1.2280	0.1870*	-1.1929	0.2633*	-0.5057	0.1311*
Marital Statu	0.4631	0.0917*	0.4298	0.1456*	0.1825	0.0573*
Substance Abuse	0.5341	0.1148*	0.4953	0.1727*	0.2115	0.0765*
Gender*Education	0.3178	0.0603*	0.3239	0.0933*	0.1250	0.0388*
Education*Marital	-0.1480	0.0350*	-0.1366	0.0515*	-0.0573	0.0234*
Marital*Substance Abuse	-0.5039	0.1042*	-0.4550	0.1568*	-0.2006	0.0683*
Dispersion parameter	-	-	2.079	0.389	0.1733	0.0586
Log-liklihood	-479.39		-420.64		-423.32	
AIC	974.77		859.28		864.64	

* $p<0.05$ (significant covariate)

To control significance of over-dispersion, the log-likelihood ratio, which was -2×(log-likelihood of Poisson regression-log-likelihood of Negative Binomial regression), is computed. The log-likelihood ratio is became 117.5, which exceeded critical value of chi-square, giving evidence of over-dispersion. The estimated dispersion parameter for COM-Poisson model was $\upsilon = 0.1733$, indicating over-dispersion. To test the significance of dispersion parameter a hypothesis test which established by Sellers and Shmueli (2010) is used. The p value are found 0, indicating over-dispersion that requires a COM-Poisson regression instead of Poisson regression. Based on both test, evidence of over-dispersion indicated inadequate fit of the Poisson model. The best model for the main effects plus interaction effects models is chosen according to smallest log-Liklihood and AIC values. In terms of Log likelihood and AIC, the Negative Binomial model showed best fit for main effects model. Both Log likelihood and AIC values are decreased with added the interaction effects on the main effects models. The results from the Negative Binomial model analysis are presented in Table 5.

Acknowledgements

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

References

http://www.schizophrenia.com/family/schizintro.html

[1] Aitkin, M., Anderson, D., Francis, B., Hinde, J. (1990). Statistical modelling in GLIM. New York: Oxford University Press.

[2] Renshaw, A.E. (1994). Modelling the claims process in the presence of covariates. ASTIN Bulletin, 24(2): 265-285.

[3] Frome, E.L. (1983). The analysis of rates using Poisson regression models. Biometrics, 39: 665-674.

[4] Lord, D. (2006). Modelling motor vehicle crashes using Poisson-Gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis&Prevention, 38(4): 751-766.

[5] Riphahn, R., Wambach, A., Million, A. (2003). Incentive effect in the demand for health care: a bivariate panel count data estimation. Journal of Applied Econometrics, 18(4): 387-405.

[6] Famoye, F., Singh, K.P. (2006). Zero-Inflated Generalized Poisson regression model with an application to domestic violence data, Journal of Data Science , 4, 117-130.

[7] Pararai, M., Famoye, F., Lee, C. (2010). Generalized Poisson-Poisson Mixture model for misreported counts with an application to smoking data, Journal of Data Science, 8, 607-617.

[8] Famoye, F., Wulu, J.T., Singh, K.P. (2004). On the Generalized Poisson Regression model with an application to accident data, Journal of Data Science, 2, 287-295.

[9] Ozmen, I., Famoye, F. (2007). Count Regression Models with an application to zoological data containing structural zeros, Journal of Data Science, 5, 491-502.

[10] Hilbe, J.M. (2011). Negative Binomial regression, Second edition, Cambridge University Press.

[11] McLachlan, G.J. (1997). On the EM algorithm for overdispersed count data. Statistical methods in Medical Research, 6: 76-98.

[12] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, Biometrika, 61: 439–447.

[13] Oh, J., Washington, S.P., Nam, D. (2006). Accident prediction model for railway-highway interfaces, Accident Analysis&Prevention, 38(2): 346-356.

[14] Famoye, F. (1993). Restricted generalized Poisson regression models, Commun. Statist. Theor. Meth, 22:1335-1354.

[15] Conway, R.W., Maxwell, W.L. (1962). A queuing model with state dependent service rates, Journal of Industrial Engineering, 12:132-136.

[16] Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P. (2005). 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(1): 127-142.

[17] Consul, P.C. (1989). Generalized Poisson Distribution: Properties and Application. New York: Marcel Dekker.

[18] *Consul, P.C., Famoye, F. (1992). Generalized Poisson regression model, Communications in Statistics (Theory & Method), 2(1): 89-109.*

[19] Famoye, F., Wulu, J.T., Singh, K.P. (2004). On the generalized Poisson regression model with an application to accident data, Journal of Data Science, 2: 287-295.

[20] Wang, W., Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression, Journal of Population Economics, 10: 273-283.

[21] Lord, D., Guikema, S.D., Geedipally, S.R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes, Accident Analysis & Prevention, 40(3): 1123–1134.

[22] Lord, D., Geedipally, S.R., Guikema, S.D. (2010). Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion, Risk Analysis, 30(8): 1268-1276.

[23] Khan, N.M. and Khan, M.H. (2010). Model for analysing count with over-, equi-and under-dispersion in actuarial statistics, Journal of Mathematics and Statistics, 6(2):92-95.

[24] Benson, A., Friel, N. (2017). Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution, arXiv:1709.03471v1 [stat.CO] 11 Sep.

[25] Saghir, A., Lin, Z. (2013). Control chart for monitoring multivariate COMPoisson attributes, Journal of Applied Statistics, 41(1), 200–214.

[26] Chanialidis, C., Evers, L., Neocleous1, T., Nobile, A. (2017). Efficient Bayesian inference for COM-Poisson regression models, Statistics and computing, ISSN 0960-3174.

[27] Zhou, M., Li, L., Dunson, D., Carin, L. (2012). Lognormal and gamma mixed negative binomial regression, International Conference on Machine Learning (ICML2012), Edinburgh, Scotland.

[28] Long, S.J. (1997). Regression Models for Categorical and Limited Dependent Variables. SAGE.

[29] Cameron, A.C., Trivedi, P.K. (1998). Regression Analysis of Count Data. Cambridge, UK.

[30] Agresti, A. (2002). Categorical Data Analysis. WileyInterscience, 2nd edition.

[31] Winkelmann, R. (2008). Econometric Analysis of Count Data. Springer, Berlin, 5th edition.

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

[33] Hilbe, J.M. (2007). Negative Binomial Regression. Cambridge University Press.

[34] Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. Canadian Journal of Statistics, 15(3):209-225.

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

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

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

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

[39] 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, Spring 2013.

[40] Minka, T.P., Shmueli, G., Kadane, J.B., Borle, S., Boatwright, P.,"Computing with the COM-Poisson distribution". Technical Report Series, Carnegie Mellon University Department of Statistics, Pennsylvania, (2003).