IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 01-06

Statistical Modeling of Count Data using Negative Binomial - Generalized Lindley Distribution

K.M.Sakthivel[1], C.S.Rajitha[2], K.B.Alshad[3]
[1]*Department of Statistics, Bharathiar University, Coimbatore 641046 Tamilnadu, India*
[2]*Department of Statistics, Bharathiar University, Coimbatore 641046 Tamilnadu, India*
[3]*Department of Statistics, Bharathiar University, Coimbatore 641046 Tamilnadu, India*

Abstract: *For analyzing the count data, traditional probability distributions such as Poisson and negative binomial distributions are considered to be the most suitable models. But in some situations, count data shows large number of zeros that cause heavy tail which leads to over dispersion. In these situations, it is observed that these traditional statistical count models cannot be used efficiently. In order to overcome this problem, many mixed distributions have been introduced in the statistical literature. Among these distributions, Poisson and negative binomial were used as base line distribution for analyzing over dispersed count data. In this paper, we have proposed a mixture of negative binomial mixture distribution with generalized Lindley distribution and the resulting distribution is named as negative binomial-generalized Lindley (NB-GL) distribution. Further we have obtained some vital characteristics of the distribution such as mean, variance and factorial moments. Also we used maximum likelihood estimation method for estimation of parameters of proposed distribution.*

Keywords: *Mixture Distributions, Negative binomial distribution, Generalized Lindley distribution, Maximum Likelihood Estimation*

I. INTRODUCTION

Count data analysis plays a significant role in various fields such as insurance, public health and road accidents (Panjer, 2006). Usually Poisson distribution is used for count data analysis if the dispersion index (equi-disperson) gives the value one (Johnson e.al 2005). But in many applications, count data does not satisfy this assumption. According to Karlis and Xekalaki, 2005, Poisson distribution does not provide flexible analysis for count data sets, since most of the applications count data shows over dispersion. Hence many mixed distributions were introduced in literature for modeling over dispersed count data(Raghavachari etal., 1997; panjer, 2006; Karlis and Xekalaki, 2005). Among the mixed distributions, negative binomial(NB) distribution is one of the most commonly used distribution for modeling the over dispersed count data. Greenwood and Yule, 1920 used mean of the Poisson random variable is distributed as a gamma random variable. It is considered as a very good alternative for Poisson distribution in count modeling. However, in some applications NB distribution may not suitable due to the over dispersed count data.

One major cause of over dispersion is the existence of excess number of zero counts. And this phenomenon is called zero inflation. Further, recent studies shows that mixed negative binomial distributions perform better compare to the Poisson mixture distributions and traditional distributions for modeling the zero inflated count data. Some of them are negative binomial –inverse Gaussian distribution (Gomez-Deniz et.al , 2008), negative binomial –Lindley distribution (Zamani and Ismail, 2010), negative binomial-exponential distribution(Ranger and Willmot, 1981), negative binomial –generalized exponential distribution (Aryuyuen and Bodhisuwan, 2013), negative binomial –Sushila distribution (Yamrubboon et al, 2017) etc.

The aim of this paper is to introduce an alternative distribution named as negative binomial-generalized Lindley (NB-GL) distribution for modeling the count data with excess number of zero counts. It is obtained by mixing the NB distribution with generalized –Lindley distribution (Zakerzadeh and Dolati, 2009). The organization of the paper is as follows. In Section 2, the probability mass function and some possible shapes of the proposed NB-GL distribution are provided. Section 3 discusses some characteristics of the NB-GL distribution. Parameter estimation method is provided in section 4. Section 5 provides applications of the NB-GL to a real data set. And finally in section 6, the conclusion of the study is given.

Negative binomial-Generalized Lindley distribution

In this part a new negative binomial mixture distribution is provided which is obtained by mixing the negative binomial distribution with generalized Lindley distribution (Zakerzadeh and Dolati, 2009).

Definition: Let X be a random variable following $\text{NB-GL}(r,\alpha,\theta,\gamma)$ distribution where X has the NB distribution with parameter $r>0$ *and* $p=\exp(-\lambda)$ and λ follows GL distribution with parameters α,θ,γ. i.e.; $X/\lambda \sim \text{NB}(r,\ p=\exp(-\lambda))$ and $\lambda \sim GL(\alpha,\theta,\gamma)$ for $r,\theta,\alpha,\gamma>0$.

Theorem 1: Let $X \sim NB\text{-}GL(r,\theta,\alpha,\gamma)$. Then the probability mass function of X is given by $p(X=x\mid\lambda)=\binom{r+x-1}{x}\sum_{j=0}^{x}\left[\binom{x}{j}(-1)^j\left(\frac{\theta}{(\theta+r+j)}\right)^{\alpha+1}\times\left[\frac{(\theta+r+j+\gamma)}{(\theta+\gamma)}\right]\right]$, where $x>0$ & $r,\theta,\alpha,\gamma>0$

Following figures (Fig 1 - 4) represents the PMF of NB-GL distribution for different values of parameters α,θ,r *and* γ

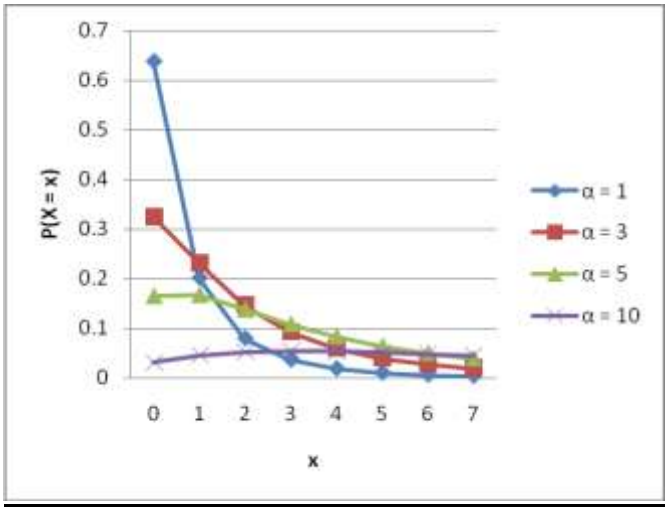

Figure 1: PMF of NB-GL distribution for the parameters α

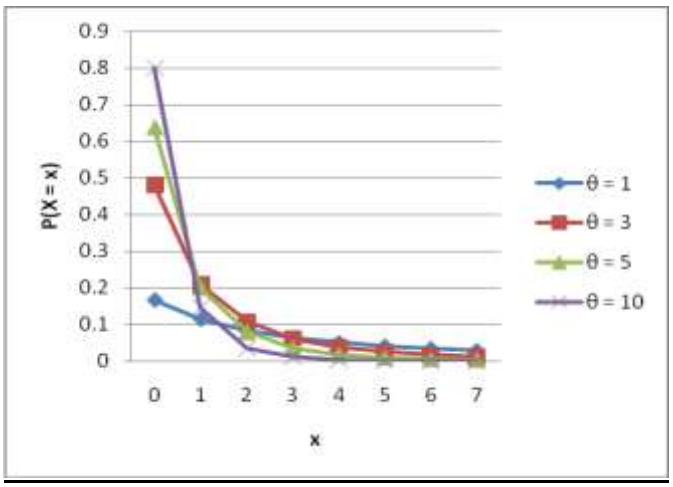

Figure 2: PMF of NB-GL distribution for the parameters θ

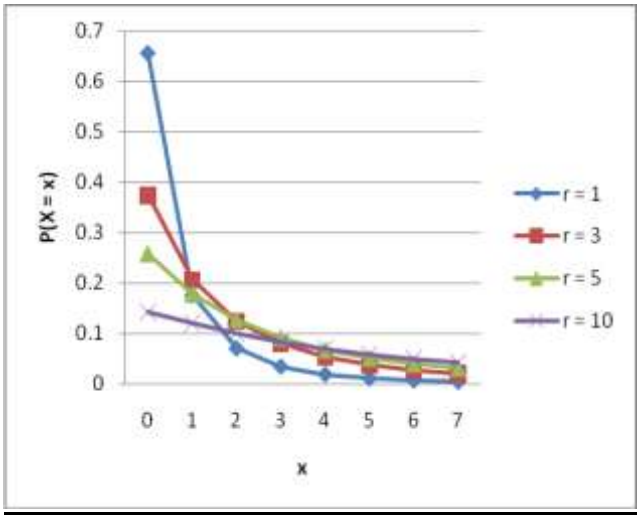

Figure 3: PMF of NB-GL distribution for the parameters r

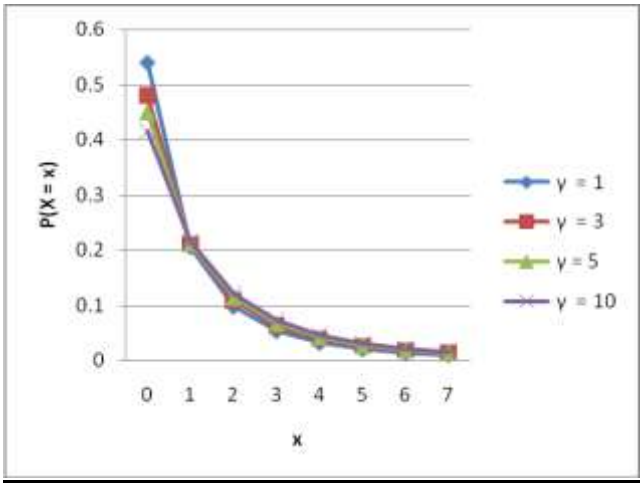

Figure 4: PMF of NB-GL distribution for the parameters γ

Characteristics of the NB-GL distribution

In this section, we provide some characteristics of the NB-GL distribution. Since the factorial moment is one of the most significant characteristics of the probability distribution, we defined the factorial moments as follows.

Theorem 2: If $X \sim NB\text{-}GL(r,\theta,\alpha,\gamma)$, then the factorial moment of order k of X is given by

$$\mu_k = \frac{\Gamma(r+k)}{\Gamma r} \sum_{j=0}^{k} \binom{k}{j} (-1)^j \left(\frac{\theta}{\theta-k+j}\right)^{\alpha+1} \left(\frac{\theta-k+j+\gamma}{\theta+\gamma}\right)$$

Where $k = 1,2,3,\ldots for\ r,\theta,\alpha,\gamma > 0$

From the factorial moments we can easily obtain the mean and variance of the NB-GL distribution as

$$\mu_1 = E(X) = r\left[\chi_1\delta_1 - 1\right]$$

$$\mu_2 = V(X) = r(r+1)\left[\chi_2\delta_2 - \chi_1\delta_1 + 1\right]$$

where

$$\left(\frac{\theta}{\theta-1}\right)^{\alpha+1}=\chi_1,\ \left(\frac{\theta}{\theta-2}\right)^{\alpha+1}=\chi_2,$$

$$\left(\frac{\theta-1+\gamma}{\theta+\gamma}\right)=\delta_1,\ \left(\frac{\theta-2+\gamma}{\theta+\gamma}\right)=\delta_2$$

Parameter estimation

The parameter of the NB-GL distribution is obtained by using the method of maximum likelihood estimation method.

The log likelihood function of the $\text{NB - GL}(r,\theta,\alpha,\gamma)$ is obtained as

$$\log L(r,\alpha,\theta,\gamma)=\sum_{x=0}^{k} n_x \log p_x$$

$$L(r,\alpha,\theta,\gamma)=\sum_{x=0}^{k} n_x \log\left[\begin{array}{l}\binom{r+x-1}{x}\\ \times\sum_{j=0}^{x}\binom{x}{j}(-1)^j\left(\frac{\theta}{(\theta+r+j)}\right)^{\alpha+1}\\ \times\left(\frac{(\theta+r+j+\gamma)}{(\theta+\gamma)}\right)\end{array}\right]$$

By equating the partial derivatives of the log likelihood to zero with respect to the parameters r,θ,α,γ we obtain the parameters of the NB-GL distribution. The score equations are given below.

$$\frac{\partial}{\partial r}L(r,\alpha,\theta,\gamma)=0$$

$$\Rightarrow\sum_{x=0}^{k} n_x\frac{1}{\delta}\left[\begin{array}{l}\frac{\Gamma(r+x)}{\Gamma r}k\left(\begin{array}{l}(\theta+r+j)\\ \frac{-(\theta+r+j+\gamma)(\alpha+1)}{(\theta+r+j)^{\alpha+2}}\end{array}\right)\\ +k\frac{(\theta+r+j+\gamma)}{(\theta+r+j)^{\alpha+1}}\left[\frac{\left\{\begin{array}{l}\Gamma r\Gamma(r+x)'\\ -\Gamma(r+x)\Gamma r'\end{array}\right\}}{(\Gamma r)^2}\right]\end{array}\right]=0,$$

$$\frac{\partial}{\partial\theta}L(r,\alpha,\theta,\gamma)=0$$

$$\Rightarrow \sum_{x=0}^{k} n_x \frac{-\delta(r+j)}{\delta}\left[\frac{1}{(\theta+r+j+\gamma)(\theta+\gamma)} + \frac{(\alpha+1)}{\theta(\theta+r+j)}\right]=0$$

$$\frac{\partial}{\partial\alpha}L(r,\alpha,\theta,\gamma)=0$$

$$\Rightarrow \sum_{x=0}^{k} n_x \frac{1}{\delta}\delta\ln\left(\frac{\theta}{(\theta+r+j)}\right)=0$$

$$\frac{\partial}{\partial\gamma}L(r,\alpha,\theta,\gamma)=0$$

$$\Rightarrow \sum_{x=0}^{k} n_x \frac{1}{\delta}\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\left(\frac{\theta}{(\theta+r+j)}\right)$$

Where

$$\delta=\sum_{j=0}^{x}\binom{x}{j}(-1)^j\left(\frac{\theta}{(\theta+r+j)}\right)^{\alpha+1}\left(\frac{(\theta+r+j+\gamma)}{(\theta+\gamma)}\right),$$

$$k=\frac{1}{\Gamma(x+1)}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta^{\alpha+1}}{(\theta+\gamma)}$$

Using these equations, we can obtain the MLE's of r,α,θ,γ respectively using R software.

Application study

Here we considered a real data set which provides information on 9461 automobile insurance policies and fit the data set with the Poisson, NB, NB-L and the proposed NB-GL distribution. The data set was taken from Zamani and Ismail (2010). Table1 provides the fitting of different distributions to this data set. For comparing the expected and observed values of the data set for different distributions, we used the estimated log-likelihood and chi-square test statistic. Chi-square test is used for finding the goodness of fit of the data. Here we set the null hypothesis as data follow whatever distribution that is being tested, which includes Poisson, NB, NB-L, and NB-GL with given parameter estimates and the alternative hypothesis is set as data follow some other distributions. Based on the p- value and the log likelihood, the NB-GL distribution provides better fit to the given data.

Table 1. Observed and expected frequencies

No. of accidents	No. of policies	Poisson	NB	NB-L	NB-GL
0	7840	7638.3	7843.3	7853.6	7850.0
1	1317	1634.6	1290.2	1287.4	1310.3
2	239	174.9	257.7	247.6	237.8
3	42	12.5	54.5	54.2	41.7
4	14	0.7	11.8	13.2	14.2

5	4	0	2.6	3.5	3.5
6	4	0	0.6	1.0	2.6
7	1	0	0.2	0.3	0.6
8+	0	0	0.1	0.2	0.3
Parameter estimates		$\hat{\lambda} = 0.214$	$\hat{r} = 0.70$ $\hat{p} = 0.765$	$\hat{r} = 4.63$ $\hat{\theta} = 23.55$	$\hat{r} = 2.421$ $\hat{\theta} = 9.873$ $\hat{\alpha} = 0.855$ $\gamma = 29.27$
Chi-square		293.8	8.65	6.9976	4.7078
p-value		<0.01	0.07	0.32	0.6714
Log Likelihood		-5490.78	-5348.00	-5344.00	-5334.94

II. CONCLUSION

In this paper, we proposed a new mixed negative binomial distribution, named as negative binomial–generalized Lindley (NB-GL) distribution. Some characteristics of the negative binomial –generalized Lindley distribution such as factorial moments, mean, variance etc are determined. Using method of maximum likelihood estimation, we obtained the parameters of the proposed NB-GL distribution. And the NB-GL distribution is applied to a real data set and compared the performance with some other already existing distributions in terms of p-value and log likelihood. And the result shows that the proposed distribution is a flexible alternative to other distributions for modeling the count data with excess number of zero counts.

REFERENCE

[1]. Aryuyuen, S., & Bodhisuwan, W. "The negative binomial-generalized exponential (NB-GE) distribution". *Applied Mathematical Sciences*, vol 7(22), 1093- 1105, 2013.

[2]. Gomez-Deniz, E., Sarabia., J. M., & Calderin-Ojeda., E. "Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications". *Insurance Mathematics and Economics*, vol 42, 39-49, 2008.

[3]. Greenwood, M., & Yule, G. U. "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents".*Journal of the Royal StatisticalSociety*, vol 83(2), 255-279, 1920.

[4]. Johnson, N. L., Kemp, A. W., & Kotz, S. "Univariate Discrete Distributions, 3rd, Wiley Series in Probability and Statistics", John Wiley and Sons, Inc. Hoboken, New Jersey, U.S.A.2005

[5]. Karlis, D. & Xekalaki, E. "Mixed Poisson distributions",*International Statistical Review / Revue Internationalede Statistique,*vol 73(1), 35-58, 2005.

[6]. Panjer, H. H. "Mixed Poisson Distributions, In Encyclopedia of Actuarial Science", John Wiley and Sons, Ltd. Hoboken, New Jersey, U.S.A,2006.

[7]. Raghavachari, M., Srinivasan, A., & Sullo, P. "Poisson mixture yield models for integrated circuits: A critical review", *Microelectronics Reliability*, vol 37(4), 565-580,1997.

[8]. Ranger,H., & Willmot, H. "Finite sum evaluation of the negative binomial exponential model", *Astin Bulletin vol* 12, 133–137,1981.

[9]. Yamrubboon, D., Bodhisuwan, W., Pudprommarat, C.,& Saothayanun, L. "The Negative Binomial-Sushila Distribution with Application in Count Data Analysis".*Thailand Statistician*, vol 15(1), 69-77,2017.

[10]. Zamani, H., & Ismail, N. "Negative binomial-Lindley distribution and its application",*Journal of Mathematics and Statistics*, vol 6(1), 4-9,2010.

[11]. Zakerzadeh, H., &Dolati, A. "Generalized Lindley distribution",*Journal of Mathematical Extension, vol* 3(1), 1-1s7,2009.