Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 5359-5366

http://www.ripublication.com

Comparison Count Regression Models for the Number of Infected of Pneumonia

Mohammed Jasim Mohammed Hussein and Hanan Ali Hamodi

Department of Statistics, College of Administration and Economics, University of Baghdad, Iraq.

Abstract

In this paper Regression model for count data was construct to study the influencing factors on the number of patients infected with Pneumonia for Iraqi children under the age of five years. Three type of distributions was used to find the best fit of the model. The results of Log-Likelihood and AIC indicated the Zero inflated -Geometric is the best fit for this model.

1. Introduction

Knows regression analysis in general as a measure of the average athletic relationship between two variables or more in terms of units of measurement the explanatory variables in the relationship is often called relations of this type the regression models. Regression analysis divided into two parts non-linear regression and linear regression. In statistics, 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, 4, 5, ... }, and where these integers arise from counting rather than ranking. There are count data in different areas of life such as results of the negative health, and traffic accidents, are all treated as events are counting. In addition, depend on her observations forms dispersion as follows:-

- Over Dispersed
- Equi Dispersed
- Under Dispersed

2.1. Negative Binomial (I)

Under Poisson distribution, the mean λ supposedly be homogeneous within classes. Assuming that λ gamma distribution with mean $E(\lambda)=\mu$ and variance $Var=\mu^2\theta^{-1}$, and Y/λ distributed of Poisson with conditional mean $E(y/\lambda)=\lambda$ then the marginal distribution for (Y) followed by Negative Binomial distribution with probability density function:

$$P_r(Y=y)=\int P_r(Y=y\mid\lambda)f(\lambda)d\lambda$$
$$=\frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \quad (1)$$

The mean $E(Y)=\mu$ and variance

$$Var(Y)=\mu+\mu^2\theta^{-1}$$
$$=\mu+\frac{\mu^2}{\theta}$$

And can be generate different kinds of Negative Binomial distribution for example if $\theta=\alpha^{-1}$ then (Y) is Negative Binomial with $E(\lambda)=\mu$ and $V(Y)=\mu(1+\alpha\mu)$. Knowing that θ is the inverse of the dispersion parameter and $\theta=1/\alpha$ then the equation (1) can be written as follows:

$$p(y)=\frac{\Gamma(1/\alpha+y)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{1/\alpha}{1/\alpha+\mu}\right)^{1/\alpha}\left(1-\frac{1/\alpha}{1/\alpha+\mu}\right)^{y}$$
$$=\frac{\Gamma(1/\alpha+y)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \quad (2)$$

In addition, if α equals zero then $E(Y)=Var(Y)$ and produces Poisson distribution and if $\alpha>0$ then $Var(Y)>E(Y)$ thus the distribution allows to deal with the case over dispersion. Moreover, there is special case for Negative Binomial distribution (I) if $\theta=1$ will be the distribution is Geometric, used in this research, it is

the only possible case for the application of Geometric distribution. If we assume that, the mean or the corresponding value is

$$E(Y \mid x) = \mu = e \exp\left(x^T \beta\right) \tag{3}$$

The maximum likelihood function for regression model of (Negative Binomial I) is

$$\ell(\beta,\alpha) = \sum \left\{ \sum_{r=1}^{y-1} \log(1+\alpha r) \right\} - y \log(\alpha) - \log(y!) + y \log(\alpha\mu) - \left(y - \alpha^{-1}\right) \log(1+\alpha\mu) \tag{4}$$

Therefore estimation of maximum likelihood function of $(\hat{\beta}, \hat{\alpha})$ can be obtained by maximizing $\ell(\beta,\alpha)$ for (β) and(α) the equations are:

$$\frac{\partial \ell(\beta,\alpha)}{\partial \beta_j} = \sum \frac{(y-\mu)x}{1+\alpha\mu} = 0 \qquad , j = 1,2,3,\ldots, p \tag{5}$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha} = \sum \left\{ \sum_{r=1}^{y-1} \left(\frac{r}{1+\alpha r} \right) \right\} + \alpha^{-2} \log(1+\alpha\mu) - \frac{\left(y+\alpha^{-1}\right)\mu}{(1+\alpha\mu)} = 0 \tag{6}$$

The maximum likelihood estimates for $(\hat{\beta}, \hat{\alpha})$ can be solved at the same time. Includes conducted a sequential, in the first iteration by the initial value for $\alpha, \alpha_{(0)}$ the $\ell(\beta,\alpha)$ maximizing for (β) it produces (β_1). The equation (5) is equivalent to the weighted least squares, so with a simple modification of this process can be performed by use of a regression (IWLS) similar to those in the Poisson distribution, and in the second iteration placed (β) fixed at (β_1) the $\ell(\beta,\alpha)$ maximizing for (α) is produces (α_1). In equation (6) can be extracted by the use of repetition Newton-Raphson , by imposition β as fixed, and α as fixed can be obtained on the maximum likelihood estimated for $(\hat{\beta}, \hat{\alpha})$.

2.2. Models of Excess Zeros

The count data may appear (over/under dispersion) or include many zeros. These properties suggested the use of models such as zero-inflated regression models and

Hurdle regression. These models has been used successfully in economy, medicine, biology, and epidemiology. The most important characteristics of these models is that they are suitable for mixed data of two society. One of them has zeros and other the numbers for discrete distributions. The models of zero-inflated and Hurdle can be summarized as comes:

$$p(Y = y / w) = w\delta_{\circ}(y) + (1-w) f(y) \qquad (7)$$

Y: variable count, W: increase the percentage of zeros

$$\begin{cases} \delta_{\circ}(y) = 1 & , \quad y = 0 \\ \delta_{\circ}(y) = 0 & , \quad o.w \end{cases}$$

$f(y)$: Density function the count

Moreover, if the-

$w \neq 0$, $f(0) = 0$ Produces Hurdle model in (7), $w \neq 0$, $f(0) \neq 0$ Produces zero-inflated model in (7)

If $w > 0$ in (7) produces zero-inflated or Hurdle, if $w < 0$ in (7) produces zero-inflated model. $f(y)$ Either be (Binomial or Geometric or Poisson or Negative Binomial or Generalized Poisson). In fact increase the number of zeros in conformity with impressions event none numbered for one reason or other, because it cannot take counting, the following explanation of all about model.

2.2.1.Zero-Inflated Regression Models

The zero-inflated density is a mixture of a point mass at $zero\, I_{\{0\}}(y)$ and a count distribution $f_{count}(y; x, \beta)$. The probability of observing a zero count is inflated with probability $\pi = f_{zero}(0; z, \gamma)$:

$$f_{zero\inf l}(y; x, z, \beta, \gamma) = f_{zero}(0; z, \gamma) \cdot I_{\{o\}}(y) + (1 - f_{zero}(0; z, \gamma)) \cdot f_{count}(y; x, \beta) \quad (8)$$

Where $I(\cdot)$ is the indicator function and the unobserved probability π of belonging to the point mass component is modelled by a Binomial GLM $\pi = g^{-1}(z^T\gamma)$ .The corresponding regression equation for the mean is:

$$\mu_i = \pi_i \cdot 0 + (1-\pi_i)\cdot \exp(x_i^T\beta) \quad (9)$$

And using the canonical log link. The vector of regressors in the zero-inflation model z_i and the regressors in the count component x_i need not to be distinct in the simplest case, $z_i = 1$ is just an intercept. The full set of parameters of β, γ and potentially the dispersion parameter θ (if a Negative Binomial count model is used) can be estimated by ML. Inference is typically performed for β, γ and, while θ is treated as a nuisance parameter even if a Negative Binomial model is used.

2.2.2 Hurdle Regression Models

The Hurdle model combines a count data model $f_{count}(y; x; \beta)$ (that is left-truncated at y = 1) and a zero Hurdle model $f_{zero}(y; z; \gamma)$(right-censored at y = 1) :

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; z, \gamma) & y = 0 \\ (1 - f_{zero}(0; z, \gamma)).\, f_{count}(y; x, \beta)/(1 - f_{count}(0; x, \beta)) & y > 0 \end{cases} \quad (10)$$

The parameters of models are β, γ . And one or two additional dispersion parameters θ (if f_{zero} , f_{count} or both are Negative Binomial densities) are estimated by ML, where the specification of the likelihood has the advantage that the count and the Hurdle component can be maximized separately. The corresponding mean regression relationship is given by:

$$\log(\mu_i) = x_i^T\beta + \log(1 - f_{zero}(0; z_i, \gamma)) - \log(1 - f_{count}(0; x_i, \beta)) \quad (11)$$

and using the canonical log link. For interpreting the zero model as a Hurdle , useful if the same regressors $x_i = z_i$ are used in the same count model in both components $f_{count} = f_{zero}$:

A test of the hypothesis $\beta = \gamma$ then tests whether the Hurdle is need or not.

3. APPLICATION

Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. Data are collected from the Multiple Indicator Cluster Survey (MICS4) in 2011. MICS produces a wide range of scientifically built and tested indicators to provide a realistic and detailed picture of the fulfillment of critical children and women rights across the world. 36,580 households were sampled in MICS4, In order to model the effect of variables on the number of children that has Pneumonia for all districts in all governorates of Iraq, the following variables has been considers for the model:-

Y: The Number of Infected of Pneumonia.

X_1: Number of family members.

X_2: Number of children under the age of five in each family.

X_3: Number of doses of DPT.

X_4: Mother's education level

X_5: Regions of Iraq (Kurdistan, Central and Southern of Iraq)

X_6: Sex of the child

X_7: Type of breastfed

X_8: BCG immunization

By estimating the parameters of the model was conducted goodness of fit for (Y) variable which represents the number of Pneumonia patients, through the use of function (com.fit) in R language, the results of the values of parameters of the distribution is $\nu=0, \lambda=0.45$.

Since $\nu = 0$, $\lambda < 1$, this refers that the data are distributed Geometric distribution it is a special case of Com-Poisson distribution. To find the best molded which fit the data we used the following count regression model.

1. Geometric distribution.
2. Hurdle - Geometric distribution model.
3. Zero inflated -Geometric distribution model.

Table 1. Coefficient estimates, standard error, number of significant parameters, maximized log-likelihood and AIC value for models

Model / Coeff	Geometric		Hurdle-GE		zero-inflated -GE	
	Coeff.	St- Error	Coeff.	St- Error	Coeff.	St- Error
Intercept	3.94944	0.2395*	3.38337	0.37238*	3.77058	0.27299*
X_1	-0.27115	0.0310*	-0.31705	0.05366*	-0.35283	0.03969*
X_2	-0.12577	0.0274*	-0.07662	0.04596	-0.03903	0.03713
X_3	0.12581	0.0220*	0.12828	0.03396*	0.11402	0.02543*
X_4	-0.12096	0.0374*	-0.04395	0.06198	-0.09675	0.04618*
X_5	0.16178	0.0694*	0.24462	0.10055*	0.17165	0.08053*
X_6	-0.20945	0.0619*	-0.15316	0.09114	-0.18919	0.07042*
X_7	-1.29383	0.0802*	-1.18388	0.13975*	-1.17271	0.09335*
X_8	-1.19016	0.0927*	-1.31726	0.17601*	-1.18289	0.10818*
Intercept			5.67188	0.42730*	-19.6959	4.6500*
X_1			-0.30478	0.05326*	-5.6416	1.6373*
X_2			-0.20335	0.04621*	5.3130	1.2276*
X_3			0.14272	0.03703*	0.6405	0.3375
X_4			-0.24323	0.06450*	1.1916	0.5467*
X_5			0.07630	0.12179	-0.3995	1.0444
X_6			-0.31920	0.10390*	0.8919	0.7343
X_7			-1.65579	0.12507*	4.6682	1.2333*
X_8			-1.37705	0.13956*	1.2460	1.0566
no. parameters	9		15		13	
Log-likelihood	-2509.212		-2479		-2478	
AIC	5036		4993.844		4991.407	

* at the level of significant 0.05

4. CONCLUSIONS

This paper is related with the response variable of interest is a count, for count data, the most widely used regression model is Com-Poisson regression. When data display over-dispersion, the common solution is to Geometric regression. In order to model the effect of variables the number of Pneumonia patients, Geometric, Hurdle - Geometric and Zero inflated -Geometric regression models are fitted respectively. The results of Log likelihood and AIC indicated the Zero inflated –Geometric distribution is the best fit for this model.

REFERENCES

[1] Achim Zeileis, Christian Kleiber, Simon Jackman "Regression Models for Count Data in R". Journal of Statistical Software, Volume 27, Issue 8. 1-25,2008.

[2] B. Sutradhar and K. Das, "A higher order approximation to likelihood inference in the Poisson mixed model", *Statist, Prob, Lett*, Vol 52,PP 59-67, 2001.

[3] C. Dean and R. Balshaw, " Efficiency lost by analyzing counts rather than even times in Poisson and overdispersed Poisson regression models", *Journal of American Statistical Association*, Vol 92, 1387-1398, 1997.

[4] D. Lord , S. Washington and J. Ivan, " Poisson, Poisson (Gamma and Zero inflated regression models of motor vehicle crashes: Balancing . statistical fit and theory", *Accident analysis and Prevention*, Vol 37, 35-46, 2005.

[5] D . Moore " Asymptotic properties of moment estimators for overdispersed counts and proportions", *Biometrics* Vol 73, 583-588, 1986 .

[6] Joseph Ngatchou-Wandji1_ and Christophe Paris , On the Zero-Inated Count Models with Application to Modelling Annual Trends in Incidences of Some Occupational Allergic Diseases in France . Journal of Data Science 9, 639-659,2011

[7] J. Mullahy, "Speci_cation and testing of some modi_ed count data models" . Journal of Econometrics 33, 341-365,1986.

[8] J. Klein, " Semi parametric estimation of random effects using Cox model based on the EM algorithm", Biometrics , Vol 48, 795-806, 1992.

[9] Noriszura Ismail and Abdul Aziz Jemain ," Handling Overdispersion with Negative Binomial and Generalized Poisson regression Models", Casualty Actuarial Society Forum,Winter 2007 .

[10] N. Breslow and X. Lin , " Bias correction in generalized linear mixed models", *Biometrics* Vol 82, 81-91, 1995.