EPH - International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 2 Issue 1 June 2016

DOI:https://doi.org/10.53555/eijms.v5i1.40

ANALYSIS FOR COUNT AND CATEGORICAL DATA USING REGRESSION MODELS

Taghreed Abdul-Razek Abdul-Motaleb Al-Said, Ph.D[1]*
*[1]*A lecturer of Statistics at AL AZHAR University, Faculty of Commerce, Department of Statistics, Cairo, Egypt.*
*[2]*Assistant Professor of Statistics at King Abdul-Aziz University, Faculty of Science, Department of Statistics*

Corresponding Author:-

Abstract:-
Counts are non-negative integers. It represents the number of occurrences of an event within a fixed period. Measurements scales of categorical variable consist of two main types of measurement scales, ordered scales that named ordinal variables and unordered scales that named nominal variables. Regression models are the most frequently used statistical models for analyzing count data such as Poisson and negative binomial regression models; logistic regression models are used with binary and categorical variables. The main goals of this research are considering these models along estimating the parameters of them, discuss the proper model of each type of data, and make a comparison between models using suitable statistical programs for analyzing the two data sets.

Keywords:-*oisson regression model; Negative Binomial regression models; Over/ under dispersion groups; logistic regression models; anemia diseases; situ-simple herniotomy operation and tension free-mesh inguinal repair*

1. INTRODUCTION

Poisson distribution has been verified to be the best distribution to describe count data. Poisson regression model is the earliest description model for describing the relation between a Poisson distributed dependent variable that is count, and one or more independent variables. It can be used in many applications such as medical, education, and many other applications. It suffers a potential problem, the assumption of the equality of the variance, and the mean which is violated; the over/under dispersion occurs, and the standard errors estimated will be biased which will then lead to incorrect test statistics. [Lawless (1987)]

There are many early works done to correct and to treat over/under dispersion problems such as the Poisson quasi likelihood method, or alternatively using the negative binomial regression models or logistic regression models. [Cameron and Trivedi (1986)]

Logistic regression models are the most popular models for dealing with binary data and categorical data. The parameters β in logistic models determine the rate of increase or decrease of the S-shaped curve for the probability of success $\pi(x)$. The sign of β indicates whether the curve ascends $(\beta > 0)$ or descends $(\beta < 0)$, and the rate of change increases as $|\beta|$ increases. When $\beta = 0$ the curve becomes a horizontal straight line. The binary response Y is then independent of X. For logistic regression parameter β, that line has slope equal to $\beta\pi(x)[1 - \pi(x)]$. The slope approaches 0 as the probability approaches 1.0 or 0. [Agresti (2007)]

This research introduces a suitable comparison between Poisson, negative binomial regression models, and logistic regression models to treat and analysis count and dispersion data, binary and categorical data respectively. Section (2) has details of the Poisson, negative binomial, and logistic regression models, some examples of these models and literature review of some studies that consider or use them. Section (3) has estimation methods for parameters along with many criterions for testing the accurate and significance of models and parameters. Section (4) has two suitable applications of the Poisson, negative binomial, and logistic models using suitable data sets. The conclusions and recommendation will be included in section (5), and finally references will be in section (6).

2. The Poisson and Negative Binomial and Logistic Regression Models

Cameron and Trivedi (1998) mentioned that count data regression models are useful in studying the occurrence rate per unit of time conditional on some covariates. Count response variable addresses non-negative integer responses need suitable regression models such as binary logistic, probit, grouped logistic, ordinal logistic, Poisson and negative binomial regression models.

Greene (2008) defined Poisson regression model as the basis model for describing the relation between Poisson distributed response variable Y, and one or more independent variables which are themselves random variables. The Poisson regression model is often used for modeling count data that has number of useful extensions for count models such as negative binomial models. The Poisson model has the conditional mean function $E[(y_i|x_i)] = \lambda_i$, and conditional variance function $var[(y_i|x_i)] = \lambda_i$, where the parameters are $\lambda_i = \exp(x'_i \beta)$

Hilbe (1994) stated negative binomial models for dealing with count data when the conditional variance exceeds the conditional mean in Poisson model. He mentioned that Log negative binominal regression is a valuable member of the family of generalized linear models for discrete data when there is independence between observations, and heterogeneity is not due to longitudinal effects. Yaacob, et al. (2010) mentioned that Poisson regression model is a good starting point of analyzing count data models such as the numbers of patients visit doctors, the number of patent awarded to a firm or bank, the number of road accident death, the number of dengue fever cases are restricted to a single digit or integer with quite low number of events.

Agresti (2007) pointed that over dispersion occurs when the model omits important predictors, when the link function between response and predictor variables is mistaken, or data include outliers. The Poisson model fails to be sufficient for these problems and also for interaction terms where predictors need to be transformed to another scale. Hilbe (2007) mentioned that this phenomenon is common when there is positive correlation between responses or there is an excess variation between response probabilities, or counts. It also arises when there are violations in the distributional assumptions of the data, such as clustered data, and thereby violates the likelihood independence of observation assumption.

Zeileis, et al. (2008) suggested to use negative binomial hurdle model to fit inflated zero observations realize in count data. Zou, et al. (2012) proposed mixture negative binomial regression models to address the unobserved heterogeneity problem in vehicle crash data. Piza (2012) used negative binomial model with crime data. He mentioned that most crimes incidents are distributed as rare event counts.

Finlayson (2010) assumed data as a mixture of two separate data generation processes, one generates only zeroes and the other process generated counts from negative binomial model. The result of a Bernoulli trial is used to determine which of the two processes generates observation. Molla and Muniswamy (2012) suggest using the power of score test for negative binomial regression model to deal with over dispersion problems. The proposed score test was compared with likelihood ratio and the Wald test via Monte Carlo simulation technique. The suggested forms of the test were also used with two real datasets such as using numerical illustration.

2.1 The Poisson regression model:

The Poisson regression is the most popular, and the standard technique for analyzing model of count data. It deals with rare events. It named as a log-linear model. It has the following form:

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0,1,2, \ldots \qquad (1)$$

Where the variance of y equals the expected value, $E(x) = Var(x)$. The Poisson can be considered as a negative binomial with a heterogeneity parameter value of zero. Choosing between Poisson and negative binomial models depends on the nature of the distribution of the dependent variable. Analysis commonly selects negative binomial regression purely because the assumptions of Poisson models are often not observed specially the social data. [Lawless (1987)]

2.2 The negative binomial regression models:

Negative binomial regression models deal with over dispersion count data. These models have mean μ_i, and variance function equals to $\mu_i + \alpha \mu_i^p$. There are two special forms of the negative binomial regression model, in addition the Poisson regression model when the dispersion parameter $\alpha = 0$. The negative binomial 2 is the standard formulation of the negative binomial models with P=2. It has the following from:

$$f(y\backslash\mu, \alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad (2)$$

The dispersion heterogeneity parameter is α , the mean and the variance functions are $E(y) = \mu$, $V(y) = \mu + \alpha \mu^2$ respectively. The negative binomial 1 has the variance function $(1+\alpha)\mu_i = \varphi\mu_i$, as in generalized linear models. It sets P=1 and can be defined as follows:

$$f(y) = \left(\prod_{j=0}^{y-1}(j + a)\right)\frac{1}{y_!}\left(\frac{a}{a+\mu}\right)^{a}\left(\frac{1}{a+\mu}\right)^{y}\mu^{y} \qquad (3)$$

Many other values of p have the same density except that α^{-1} is replaced by $\alpha^{-1}\mu^{2-p}$. [Cameron and Trivedi (1986)]

2.3 The logistic regression model:

The Logistic regression model is the best fitting models for describing relationship between binary (dichotomous) or ordinal dependent variable and a set of independent variables. It is flexible, easily used and leads to a meaningful interpretation. [Pohar, et al. (2004)] The binary logistic regression model has the response Y that takes only one of two possible values which is denoted by 0 for failure and 1 for success. The model is defined as follows:

$$\text{logit}[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = a + \beta x \qquad (4)$$

The probability of success defined by $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$ where the regression parameter is β. The binary logistic regression formula implies the following formula for the probability $\pi(x)$ using the exponential form:

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \qquad (5)$$

The general form of the logistic regression model with multiple explanatory variables, Kpredictor variables $x_1, x_2, \ldots, x_k$ can be defined as follows:

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (6)$$

The parameters β_k refer to the effect of x_k on the log odds of (Y = 1), controlling the other explanatory variable. [Agresti (2007)] The component structural of the logistic regression model sets the logit link between the probability of success and a linear combination of the covariates as follows:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \qquad (7)$$

The model parameter vector is $\beta = (\beta_1, \ldots, \beta_k)^T$. It is a vector of k elements considering the model parameters which are to be estimated. The binary data formula can summarize in contingency tables. The response vector $y = (y_1, \ldots, y_n)^T$ contains the observed binary outcomes of n independent random variables $Y_1, \ldots, Y_n$ which has binomial distribution, $Y_i \sim$ binomial $(1, \pi_i)$. The joint probability of the sample $y_1, \ldots, y_n$:

$$P(Y_1, \ldots, Y_n = y_n; \pi) = \prod_{i=1}^{n}\pi_i^{y_i}(1 - \pi_i)^{(1-y_i)} \qquad (8)$$

Where the success probabilities $\pi(\pi_1, \ldots, \pi_n)^T$. [Agresti (2007)] and [Konis (2007)]

3. The Estimation Methods of Parameters and Goodness of Fit Criterions

There are many estimation methods of estimating the parameters of the Poisson, negative binomial, and logistic regression models. Also, there are many criterions for testing the significance of models and parameters. Section (3.1) has the estimation methods of parameters and section (3.2) has the criterions for goodness of fit to test significance of models and parameters.

3.1 The Estimation Methods of Parameters:

The maximum likelihood method for estimating parameters of negative binomial 1 solves the following associated first-order conditions:

$$\sum_{i=1}^{n}\{(\sum_{j=0}^{y_{i-1}}\frac{\alpha^{-1}\mu_i}{(j+\alpha^{-1}\mu_i)})x_i+\alpha^{-1}\mu_i x_i\}=0$$

$$\sum_{i=1}^{n}\frac{1}{\alpha^2}\{-(\sum_{j=0}^{y_{i-1}}\frac{\mu_i}{(j+\alpha^{-1})})-\alpha^{-2}\mu_i\ln(1+\alpha)-\frac{\alpha}{1+\alpha}+y_i\alpha\}=0 \quad (9)$$

Estimation based on the first two moments of negative binomial 1 yields the Poisson generalized model estimator [Cameron and Trivedi, (1998)] The Pseudo maximum likelihood estimator has not a closed form and is obtained by using correct specification of the mean in the framework of an exponential family for estimating parameters of the negative binomial model 2 where parameters are distributed as follows:

$$\hat{\beta}_{NB2}=N[\underline{\beta}, V_{NB2}(\hat{\underline{\beta}})=[\sum_{i=1}^{n}\hat{\lambda}_i \underline{x_i}x'_j]^{-1}] \quad (10)$$

The maximum likelihood for the logistic regression models is defined as follows:

$$L=\prod_{i=1}^{n}\left[\frac{e^{\alpha+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\alpha+\sum_{j=1}^{p}\beta_j x_j}}\right]^{Y_i} x \left[\frac{1}{1+e^{\alpha+\sum_{j=1}^{p}\beta_j x_j}}\right]^{1-Y_I} \quad (11)$$

This iterative solution procedure is available in popular statistical procedures such as the SPSS and SAS Packages for maximizing such equations. [Dayton (1992)]

The iterative reweighted least squares can also be used to estimate the parameters in nonlinear negative binomial model. For the binary logistic regression where (y=0 or y=1), the iterative reweighted least squares is equivalent maximizing the log-likelihood of the Bernoulli distributed process using Newton's method. [Zou, et al. (2012)]

3.2 Goodness of fit Criterion:

The most frequently used measures for goodness-of-fit in generalized linear models are the Pearson chi-squares, and the deviance. The Pearson chi-squares statistic can be defined as follows:

$$\sum\frac{(y_i-\mu_i)^2}{Var(y_i)} \quad (12)$$

The statistic has asymptotic chi-squares distribution with (n – p) degrees of freedom, where the number of rating classes is n, and the number of parameters is p. [Ismail and Jemain (2007)]

The deviance measure of negative binomial models is used to compare two models. It has approximately chi-squared distribution with k-degrees of freedom. [McCullough and Nelder (1989)]

The maximum likelihood- ratio test assesses the adequacy of the negative binomial models, and generalized Poisson model. The statistic has asymptotic distribution of probability mass of one-half at zero, and one-half-chi-squares distribution with one degree of freedom. The null hypothesis is rejected if the statistic greater $x^2_{(1-2\alpha,1)}$ [Cameron and Trivedi (1998)]

4. Applications of Binary and Count Data Models

4.1 Application 1:

The first data set is from: https://stats.idre.ucla.edu/stata/dae/negative-binomialregression/ to analysis the relation between the absent in the school and many variables. The concerned cases are students 314 selected from two high school juniors. The response variable is days absent. The variable math is the standardized math score for each student. The variable program has three-levels which indicate the type of instructional program in which the student is enrolled. The SPSS package and R program is used to describe the number of absent days and the math test. The results show the conditional mean of our outcome variable is much lower than its variance.

The average numbers of days absent by program type are seem to suggest that program type is a good candidate for predicting the number of absent days outcome variable. The mean value of the outcome appears much varied by program. The variances within each level of program are higher than the means within each level. These differences suggest that over-dispersion problem is present and negative binomial regression model would be appropriate for describing such data set. Table (1) shows the average and variances of days absent for each program level:

Table 1: Average days absent for each program level

Program	mean	Variance	N
1	10.65	67.259	40
2	6.93	55.447	167
3	2.67	13.939	107
Total	5.96	49.519	314

The estimated parameters of negative binomial model is in Table (2) with 95% Wald confidence interval for estimated parameters reflects the predictors are statistically significant for absent days:

Table 2: Estimated parameters of negative binomial model

			95% Wald confidence interval		Hypothesis test			Exp(B)	95% Wald CI Exp (B)	
parameter	B	Standard error	lower	upper	Wald chisquare	degrees of freedom	Sig.	Exp(B)	lower	upper
Intercept	2.615	0.1992	2.225	3.005	172.34	1	0	13.667	9.25	20.195
Program=3	-1.279	0.2048	-1.68	-0.877	38.988	1	0	0.278	0.186	0.416
Program=2	-0.441	0.1852	-0.804	-0.078	5.661	1	0.017	0.644	0.448	0.925
Program=1	0									
Math	-0.006	0.0025	-0.011	-0.001	5.55	1	0.018	0.994	0.989	0.999

The likelihood ratio chi-square value in Table (3) provides the significance of negative binomial model for fitting data set:

Table (3): Omnibus test

Likelihood ratio chi-square	Degrees of freedom	Sig.
67.184	3	0.000

The binary logistic regression can be applied with the data by assuming the dependent variable days absent with two categories 0 for students make one absent day and 1 for students have not any absent days regardless the number of absent days. The classification Table (4) after applying logistic regression model shows 81.5% overall percentage of correct classification of the cases as follows:

Table 4: Classification results

Observed		Predicted		Percentage Correct
		not absent	absent	
Step 1	not absent	2	55	3.5
	absent	3	254	98.8
	Overall Percentage			81.5
The cut value is 0.500				

The logistic regression model is as follows:

$$\text{logit}[P(Y = 1)] = 5.574 - 0.12\ \text{Math} - 1.262\ \text{Prog} - 0.78\ \text{Gender} \qquad (13)$$

4.2 Application 2:

The second application is about inguinal hernia that remains one of the most common surgical problems. It affects males by 25%, females by 2% and 4.4 % in children which 10 times more common in boys. [Decker, et al (2019)] and [Hammoud (2019)]

Surgical procedures fall into three categories: open repair without mesh, laparoscopic repair with a mesh and extra peritoneal repair. The goals of surgery include preventing the hernia recurrence, returning the patient to normal activities quickly and minimizing the postsurgical discomfort and adverse effects of surgery. [Baig, et al (2019)] The data set are selected from general surgery department in a University hospital in Egypt. It includes 137 patients aged between 12 and 35 years managed during a period of 5 years. Cases were randomly selected from two groups, situ-simple herniotomy operation (group A) and tension free-mesh inguinal repair (group B). The data are analyzed by the SPSS package; first classification Table (5) shows 52.6% correctly classified:

Table 5: First classification results

Observed		Predicted		
		Type		Percentage Correct
		Herniotomy	Herniorahy	
Type	Herniotomy	72	0	100.0
	Herniorahy	65	0	.0
Overall Percentage				52.6%

The classification results in Table (6) after using the logistic model shows 89.8% correctly classified which increased by 30.3%:

Table 6: Classification results of logistic model

Observed		Predicted		
		Type		Percentage Correct
		Herniotomy	Herniorap hy	
Type	Hernioto my	66	6	91.7
	Herniorap hy	8	57	87.7
Overall Percentage				89.8%

Table (7) has Nagelkerke R Square, -2 Log likelihood and Cox & Snell R Square equals which reflects the ability of logistic regression model in analysis the data set:

Table 7: Goodness of fit results of logistic model

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
65.695	.595	.794

Table (8) includes estimated values of parameters in logistic model which show variables: side and Age not significance whereas the follow up, time of operation and sex are significance predictor variables in logistic model:

Table 8: Variables in the logistic model

		B	S.E.	Wald	df	Sig.	Exp (B)
Step 1	Follow up	0.958	0.316	9.170	1	0.002	2.606
	Side			0.094	2	0.954	
	Side (1)	-17.193	60.203	0.082	1	0.775	0.000
	Side (2)	-0.076	0.671	0.013	1	0.910	0.927
	Time	0.503	0.094	28.512	1	0.000	1.654
	Sex (1)	-2.634	1.075	6.008	1	0.014	0.072
	Age	0.048	0.064	0.555	1	0.456	1.049
	Constant	-19.918	4.018	24.568	1	0.000	0.000

4. Conclusions and Recommendations

This research reviewed and applied many statistical regression models to analysis count and binary data respectively. Poisson regression assumption did not meet in application 1 and 2, so negative binomial model is used for analyzing application 1. Results show good fit of the variables and the used model in applying negative binomial in application 1 and logistic model in the two applications. The research supposes two values of the response variable for the first application and applies the logistic model which reflects better estimation of the parameters and the correct classification percentage. Also, the logistic model is applied with a medical real data set in the second application which lends a clinically meaningful interpretation. Some criteria for evaluation models and parameters are reviewed and applied such as Pearson Chi-Square statistic and Cox and Snell R-Square. Finally, if the dependent variable is count, binary or categorical the negative binomial and logistic models are the suitable use models to describe and analysis such data and on the bases of the goal of the research the dependent variable can used as count or category. It is suitable to try using other models such as Bayesian approach and operation research techniques for dealing with these types of data and applying alternative methods for goodness of fit criterion in solving experimental design problems.

6. References:

[1]. Agresti A. (2007). An Introduction to Categorical Data Analysis. New Jersey. John Wiley & Sons, Inc., Hoboken.
[2]. Cameron A. C. and Trivedi P. K. (1986). Econometric models based on count data: cComparisons and applications of some estimators and tests. Journal of Applied Econometrics, Vol. 1, pp. 29-54.
[3]. Cameron A. C. and Trivedi P. K. (1998). Regression Analysis of Count Data. First edition, New York. Cambridge University Press. United Kingdom .Published in the United States by Cambridge University Press.
[4]. Dayton C. M. (1992). Logistic regression analysis. Statistics & Evaluation.
[5]. Finlayson G. E. O. (2010). The additional cost of chronic disease in Manitoba. Winnipeg, MB: Manitoba Centre for Health Policy.
[6]. Greene W.H. (2008). Econometric Analysis. Fourth edition, New York University.
[7]. Hilbe J. M. (1994). Log-Negative Binomial Regression as a Generalized Linear Model. Arizona Arizona State University, Department of Sociology and Graduate College Committee on Statistic.
[8]. Ismail N. Abdul Aziz Jemain A.A. (2007). Handling Over dispersion with Negative Binomial and Generalized Poisson Regression Models. Actuarial Society *Forum,* Winter 2007
[9]. Lawless J. F. (1987). Negative binomial and mixed poission regression. The Canadian Journal of Statistics, Vol. 15, No. 3, PP. 209-225
[10]. McCullough P. and Nelder J. A. (1989). *Generalized Linear Models*, second edition. Chapman & Hall, London. Murphy
[11]. Molla D.T. and Muniswamy B. (2012). Power of Test for over dispersion Parameter in Negative Binomial Regression. Journal of Mathematics. Volume (1). No. 4, ISSN: 22785728, PP. 29-36.
[12]. Pohar M. Blas, M. & Turk S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodoloski zvezki, 143-161.
[13]. Zou Y. Zhang, Y. and Lord D. (2012). Application of finite Mixture of Negative Binomial Regression Models with Varying Weight Parameters for Vehicle Crash Data Analysis. Accident Analysis & Prevention, Vol. 50, pp. 1042-1051.

List of Sites

[1]. David C. (2003). Modeling survival data in medical research, Second Edition. Chapman & Hall http://www.ats.ucla.edu/stat/stata/output/stata_nbreg_output.htm
[2]. Decker E. Currie A. and Baig MK. (2019). Prolene hernia system versus Lichtenstein repair for inguinal hernia: a meta-analysis. https://www.ncbi.nlm.nih.gov/pubmed/30771031
[3]. Mohamad Hammoud; Jeffrey Gerken. Inguinal Hernia ttps://www.ncbi.nlm.nih.gov/books/NBK513332/
[4]. Negative Binomial and Generalized Poisson Regression Models. casualty actuarial society forum www.casact.org
[5]. Piza, E. (2012). Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts. Rutgers Center on Public Security: http://www.rutgerscps.org/docs/CountRegressionModels.pdf
[6]. Saffari, S. E., Adnan, R., and Greene, W. (2012). Hurdle Negative Binomial Regression Model With Right Censored Count Data. Stern School of Business, New York University. PP. 181-194. http://www.idescat.cat/sort/sort362/36.2.4.saffari-etal.pdf
[7]. Yaacob, W. F. W. Lazim, M. A. and Wah, Y. B. (2010). A Practical Approach in Modelling Count Data. Malaysia Institute of Statistics. Faculty of Computer and Mathematical Science P.P. 176-183. http://instatmy.org.my/downloads/RCSS'10/Proceedings/17P.pdf
[8]. Zeileis A. Kleiber C. and Jackman S. (2008). Regression Models for Count Data in R. Journal of Statistical Software, http://www.jstatsoft.org/ , Volume 27, Issue 8.