Journal of
Applied & Computational Mathematics

Abdulkabir et al., J Appl Computat Math 2015, 4:5
DOI: 10.4172/2168-9679.1000253

Research Article Open Access

# An Empirical Study of Generalized Linear Model for Count Data

**Muritala Abdulkabir[1]*, Udokang Anietie Edem[2], Raji Surajudeen Tunde[2] and Bello Latifat Kemi[2]**

[1]*Statistics Department, University of Ilorin, Ilorin, Nigeria*
[2]*Mathematics and Statistics, Federal Polytechnic, Offa, Kwara State, Nigeria*

**Abstract**

This paper deals with an empirical study of generalized linear model (GLM) for count data. In particular, Poisson regression model which is also known as generalized linear model for Poisson error structure has been widely used in recent years; it is also used in modeling of count and frequency data. Quasi Poisson model was employ for handling over and under dispersion which the data was found to be over dispersed and another way of handling over dispersion is negative binomial regression model. In this study, the two regression model were compare using the Akaike information criterion (AIC), the model with minimum AIC shows the best which implies the Poisson regression model.

**Keywords**: Poisson regression model; Quasi Poisson model; Negative binomial regression model

## Introduction

Generalized linear models (GLMs) represent a class of regression models that allow us to generalize the linear regression approach to accommodate many types of response variables including count, binary, proportions and positive valued continuous distributions [1,2]. Because of its flexibility in addressing a variety of statistical problems and the availability of software to fit the models, it is considered a valuable statistical tool and is widely used. In fact, the generalized linear model has been referred to as the most significant advance in regression analysis in the past twenty years [2].

A generalized linear model (GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, *Yi* (for the *i*th of *n* independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of *Yi* was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.

2. A linear predictor—that is a linear function of regressors,

$$y_{ij} = \alpha_i + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function, g (.) which transforms the expectation of the response variable, $\mu_i = E(y_i)$, to the linear predictor:

$$g(\mu_i) y_{ij} = \alpha_i + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

### Assumptions and diagnostics

Similar to the linear model approach, there are key assumptions that must be met when computing a p-value using the GLM approach and violation of any of these assumptions may compromise the interpretation of model results by producing biased standard errors and thus unreliable p-values. However, disagreements in the literature on what constitutes key assumptions, decisions and checks for generalized linear modeling. Because the type I error (the p-value) on the improvement in fit with the GLM is calculated from the chi-square distribution which assumes homogenous, normal, and independent deviations centered on zero [3] it follows that these are considered key assumptions for GLMs. There is a general consensus that the assumptions of homogeneity and independence of residuals must be met [2-5] however, point out that the independence assumption can be relaxed to "at least uncorrelated". The importance of normality of residuals in GLMs, on the other hand, is debated. Some authors [2,3] suggest that normality of the residuals must be met to correctly interpret the results while others [6] note that normally distributed errors are not a condition of GLM quality but simply a description of model behavior. In addition to the assumptions of the chi-square distribution stated above, [4] Breslow also considers the correct specification of the variance function (ν), the over dispersion factor (θ) and the link function (*g*) to be critical assumptions underlying GLMs. The objective of this study are to determine appropriate generalized linear models (GLM) that are suitable for count data and investigate the presence of over dispersion in the model parameter. The data used for were simulated data from R statistical software with sample size 250.

## Components of a Generalized Linear Model

### Random component

The random component of GLMs defines the probability distribution of the response. We specify that independent observation $Y_1, Y_2 \ldots Y_n$ on the response have a probability distribution that depends on Ø. the probability density function belongs to the exponential family, which is written in two common forms.

One form of the exponential family is given by Dobson

$$f(y:\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

Where a (.), b (.), c (.) and d (.) are known functions. If a (y) = y, the distribution is in canonical form b(Ø) and is called the natural parameters, in addition to they are treated as nuisance parameters, whose values are assumed known.

### Systematic component

The systematic component of a generalized linear model (GLM)

***Corresponding author:** Abdulkabir M, Statistics Department, University of Ilorin, Ilorin, Nigeria, Tel: 08055027002; E-mail: kaybeedydx@gmail.com

**Received** April 16, 2015; **Accepted** August 17, 2015; **Published** August 26, 2015

**Citation:** Abdulkabir M, Edem UA, Tunde RS, Kemi BL (2015) An Empirical Study of Generalized Linear Model for Count Data. J Appl Computat Math 4: 253. doi:10.4172/2168-9679.1000253

specifies the explanatory variable. We define ή as a linear combination variable $x_j$, j=1,2,..,p. the predictor is expressed as a linear combination of unknown parameter $\beta_j$. For the $i^{th}$ observation $\eta_i = \sum x_j \beta_j \quad i = 1,2,\ldots,n$ where $x_{ij}$ is the value of the $j^{th}$ explanatory variable at the observation. In the matrix notation, we write $\eta = X\beta$ where ή is a vector of n linear predictor $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^T$, β is a vector p parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_2)^T$ and X is the (nxp) model matrix written as

$$X = \begin{bmatrix} x_1^T \\ . \\ . \\ . \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & . & . & x_{1p} \\ . & & & \\ . & & & \\ . & & & \\ x_{n1} & . & . & x_{np} \end{bmatrix}$$

Each row of the model matrix X refers to a different observation and each refer to a different covariate.

### Link function

The link function ή is so called because it relates, or links, the systematic and random component of GLM. It specifies how the expected value of the response relates to explanatory variable. Let μ=E($Y_i$), then ή=g (μ) and $g(\mu) = \sum_{j=1}^{p} x_{ij} \beta_j$ i=1,2,…,n

In matrix notation $\eta = g(\mu) = X\beta$ the link function is usually chosen based on the form of the distribution of the response [7]. The choice of a link function is similar to the choice of a transformation of the response; expect that the link function is a transformation of the mean, not of the individual observations (Table 1).

The canonical link occurs when $\emptyset = \eta$ where $\emptyset$ is the canonical parameter define above. For example, when the response variable is normal the canonical link is the identity function, which does not restrict the value $\eta$ and $\mu$ are allowed to take. When the response is Poisson, the canonical link is the logarithmic function, which is chosen to ensure $\mu > 0$. When the response is binomial, the canonical link is the logit function, which ensures $0 < \mu < 1$. McCullagh and Nelder [5] stated that the canonical link is favored as it results in mathematical appealing properties of the model.

## Materials and Methods

### Poisson regression model

According to McCullagh and Nelder [5] the simplest distribution used for modeling count data is the Poisson distribution, thus Poisson regression model is a special case of the generalized linear model (GLM) framework. The variance in the Poisson model is identical to mean, thus the dispersion is fixed at theta given to be 1 and the variance function is $V(\mu) = \mu$.

### Over dispersion in Poisson model

- If the conditional mean is greater than the conditional variance.
- Another common problem with Poisson regression model is excess zeros, when over dispersion is a problem we make use of negative binomial regression model, it will adjust β estimate and standard errors.

### Quasi-Poisson model

The quasi Poisson model is a way of dealing with over-dispersion that is, use the mean regression function and the variance function from the Poisson Generalized linear model (GLM) but to leave the dispersion parameter unrestricted. Thus, is not assumed to be fixed at 1 but is estimated from the data [7]. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion.

### Negative binomial regression model

Another way of modeling over dispersion count data is to assume a Negative binomial distribution for $y_i/x_i$ which arises as a gamma mixture of Poisson distribution [1,5].

### Akaike Information Criterions (AIC)

AIC is a statistical measure of the likelihood of a model parameter for the complexity of the model. It is useful when comparing two or models for data, which implies that all the data, must have the same independent variables. The smaller the AIC the better fitted models of the parameter estimate.

## Count Data

Count data are non-negative integers, they represent the number of occurrence of an event within a fixed period, e.g. number of trade in a time interval, number of given disaster, number of crime on campus per semester etc.

For count data $Y_1, Y_2, \ldots Y_n$ we will assume the model for $Y_i$ to be Poisson regression model with equal mean and variance (Appendix).

### Poisson regression model

McCullagh and Nelder [5] suggest that the Poisson distribution is the nominal distribution for count data in much the same way that the normal distribution is the benchmark for continuous data (Table 2) [8].

### Interpretation

The regressor is highly significant and the standard errors are appropriate. This will also be confirmed by the following models that deal with over-dispersion (excess zeros) that is the quasi Poisson regression model.

### Quasi-Poisson regression model

The quasi Poisson model is estimated when there is presence of over dispersion or excess zeros in Poisson model thus the regressor for both quasi and Poisson model and the same AIC in model which are highly significant.

From the Quasi Poisson model the estimated dispersion parameter were give as 1.195441 which greater than 1 indicating that over-dispersion is present in the data. The resulting from quasi Poisson regression tests of the coefficients are the same as to the results obtained from the Poisson regression with standard errors, leading to the same conclusions as before (Table 3).

### Negative binomial regression model

Another way of dealing with over-dispersion in a count data is to use a negative binomial regression model. Comparing the two models using the AIC, the Poisson regression has the minimum AIC with 579.15 with that of negative binomial regression which implies that the Poisson regression model is best (Table 4).

## Results and Conclusions

Based on the analysis the following are the resulting conclusions:

| Link | Function |
|---|---|
| Identity | μ |
| Logarithmic | Log(μ) |
| Logit | Log[μ/(1-μ)] |
| Probit | $\phi^{-1}(\mu)$ where $\phi^{-1}(.)$ is the normal cumulative distribution function |
| Complementary log-log | Log[-log(1-μ)] |
| Power | $\begin{cases}\mu^{\lambda} & \lambda \neq 0 \\ \log \mu & \lambda = 0\end{cases}$ |

**Table 1:** List some common link functions.

| Coefficients: | Estimate | Std Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.3464 | 0.08452 | -4.099 | 4.16e-05 |
| X | 1.09615 | 0.05619 | 19.507 | < 2e-16 |

(Dispersion parameter for poisson family taken to be 1)
AIC: 579.15
Number of Fisher Scoring iterations: 5

**Table 2:** Poisson Regression Model.

| Coefficients: | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.34640 | 0.09241 | -3.749 | 0.000221 |
| X | 1.09615 | 0.06144 | 17.841 | < 2e-16 |

(Dispersion parameter for quasipoisson family taken to be 1.195441)
AIC: NA
Number of Fisher Scoring iterations: 5

**Table 3:** Quasi-Poisson Regression Model.

| Coefficients: | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.34657 | 0.08457 | -4.098 | 4.16e-05 |
| X | 1.09637 | 0.05630 | 19.473 | < 2e-16 |

(Dispersion parameter for Negative Binomial (1237.558) family taken to be 1)
AIC: 581.15

**Table 4:** Negative Binomial Regression Model.

- The Poisson regression model was used to fit a model, the parameters of the fitted model were found to be significant.
- Quasi Poisson regression was used to test for over dispersion and it was found that there is over dispersion in Poisson regression model which lead to use of negative binomial regression model.
- The goodness-of-fit shows that the model is appropriate.
- Using the AIC the Poisson regression model give an appropriate model having the minimum AIC in the analysis.

## References

1. Nelder JA, Wedderburn RWM (1972) Generalized linear models. Journal of the Royal Statistical Society, Series A 135: 370-384.
2. Hoffmann JP (2004) Generalized linear models: An applied approach. Pearson: Boston.
3. Dobson AJ (2002) An Introduction to Generalized Linear Models. Chapman and Hall, London, UK.
4. Breslow N (1996) Generalized linear models: Checking assumptions and strengthening conclusions. Statistica applicata 8: 23-41.
5. McCullagh P, Nelder JA (1989) Generalized Linear Models (2ndEdn). London: Chapman & Hall
6. Gill J (2001) Generalized linear models: A unified approach. Sage University Paper London.
7. Myers RH, Montgomery DC, Vining GG (2002) Generalized Linear Models. John Wiley & Sons, New York.
8. (2012) Development Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.