Poisson Regression Models for Count Data: Use in the Number of Deaths in the Santo Angelo (Brazil)

Suzana Russo*, Diego Flender and Gabriel Francisco da Silva

Department of Statistic, Federal University of Sergipe, São Cristóvão, Brazil

Abstract: When speaking about data, presuppose its good quality otherwise the accuracy of information would be affected, which would lead to false interpretations. In Health Statistics data is obtained through surveys presented in its simplest expression, taking advantage of existing records; making an inquiry or by means of experiments. The rational organization of the data allows characterizing the priority issues and thus establishing health programs. To analyze the mortality data it is necessary to consider the mortality rate of certain age groups, so that we can find data which shows the prevalence of major groups of deaths. The analysis of data is followed by subsequent formulation of the Poisson regression models, where each group in question by age group is represented by a number of counting time. The Poisson regression model is a specific type of Generalized Linear Models (GLM) and non-linear. As [1], its main features are: a) to provide, in general, a satisfactory description of experimental data whose variance is proportional to the mean. b) It can be deduced theoretically from the first principles with a minimum of restrictions c) If events occur independently and randomly in time with constant average rate of occurrence, the model determines the number of time specified. At the end of this study, it could be seen through the analysis of the data that the age group from 70 to 79 years old sustains the highest incidence of deaths with 21.1%. Then comes the range of 60 to 69 years old with the morality rate of 20%. This was recorded for the time worked in January 2000 to December 2004. The death rate was 52.27and variance was equal to 102.43 in the city of Santo Angelo (Brazil). It was further found that the data analyzed over dispersion variance greater than average. AS a result it was necessary to remove the over dispersion to find the appropriate template. With the pattern found, some short-term forecasts were made.

Keywords: Deaths, Poisson regression models, Over dispersion.

1. INTRODUCTION

Conventionally the ‘Cause Mortality Statistics’ represents the basis or the various causes of death; therefore assigning each death one single cause is inaccurate. However the cases of violent deaths or acute infectious diseases can be associated to a single sourced death.

In order to study data on mortality, it is necessary to consider mortality in certain age groups, so that the data is showing the main groups in occurrence of deaths. For the data analysis, ‘Poisson Regression Models’ depict each group in question by age group next to a series of counts in time. Thus, we find the equations generated by the model, which allows us to make predictions in a short term.

In this study we sought to introduce the theoretical form on the structural modeling of time series by means of ‘Poisson’, representing the number of deaths by age group.

Considering the lack of data on the mortality of St. Angelo, this research will contribute to further information that would be of great importance to the local or public agencies.

At present, we lack an information system that is able to provide such knowledge with a dynamic update.

2. LITERATURE REVIEW

2.1. Poisson Regression Models

After the logistic regression model Poisson regression is the most widely used generalized linear models. Poisson regression models are applied when the response is a count, as the number of events in time.

The Poisson regression model is a specific type of GLM and non-linear. As [1], its main features are:

a) provides, in general, a satisfactory description of experimental data which is proportional to the average variance;

b) can be deduced theoretically from first principles with a minimum of restrictions;

c) If events occur independently and randomly in time with constant average rate of occurrence, the model determines the number of time.

In the last two decades have made experiments in various fields of health, such as the study of pollution vs deaths of engineering, such as in improving products and processes. Many experiments involve variables that do not have a normal distribution, the GLM are a useful alternative to traditional methods of data analysis that need changes. The GLM, introduced

*Address corresponding to this author at the Department of Statistic, Federal University of Sergipe, São Cristóvão, Brazil; Tel: +55-7921056362; E-mail: suzana.ufs@hotmail.com

ISSN: 1814-8085 / E-ISSN: 1927-5129/12

by [2] play an important role in statistics, since it generalizes the traditional normal linear regression, opening the range of options for the distribution of the response variable and giving greater flexibility to the connection between average and the systematic part of the model [3].

According to [4] the Poisson regression model is a specific type of generalized linear models (GLM) whose parameters can be estimated using the maximum likelihood method, with the likelihood function given by:

$$L = \prod_{i=1}^{n} \Pr(Z_i / \lambda_i\ ) = \prod_{i=1}^{n} \frac{e^{-\lambda} . \lambda_i^Z}{Z!} \qquad (2)$$

and the log-likelihood function equal to: $\log L = \sum (Z_i . \log(\lambda_i\ ) - \lambda_i\ ) - \sum \log(Z)!$

The constant function Z, given by $\sum \log(Z)!$, may be omitted, it does not involve λ [5, 6].

The systematic component admits the existence of a link function log(λ_i)between the means of observations and the linear structure of the model given by log(λ_i) $= \beta\ x_i^T$. Thus, the use of the connection log function ensures that the adjusted values of λ_i remain in the range [0,∞). The link function log relates the linear predictor $\beta\ x_i^T$ the expected value λ_i of vetor Z_i. The Poisson model with log link is sometimes called a log-linear model [6,7].

Transforming the link function log are obtained the following expression for the dependent variable:

$$\lambda = e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k} \qquad (3)$$

where $\beta\ x_i^T$ is the linear predictor, $\beta = (\beta_1, ..., \beta_k)^{T'}$ is the unknown parameters vector to be estimated and $x_i = (x_{i1}, ..., x_{ik})^T$ represent the values of *k* explanations variables [8].

The normal, binomial, Poisson, gamma, inverse normal, and negative binomial distributions are members of the exponential family.

Statistic of Interest

After convergence, which may be made through the Newton-Raphson algorithm, must examine the following statistics:

$(X^T W X)^{-1}$, the covariance matrix estimate for $\hat{\beta}$;

The log-likelihood function $l = l(Z, \lambda) = \sum_{i=1}^{N} \{Z_i \log \lambda_i - \lambda_i\}$

A plot of standardized residuals versus fitted values, no trend is an indication that the functional relationship variance/mean proposal for the data is satisfactory. Residual graphs versus covariates that are not in the model are quite useful. If no covariate is needed, then one should not find any trend in these plots. Data with gross errors can be detected with great residuals, or the fitted model should require more covariates, for example, higher-order interactions. The graphic inspection is a powerful means of inference in GLMs [9].

The degrees of freedom associated with the deviation is defined by $v = n - p$. To test a model, we compare the G^2 and your degrees of freedom, v, with a theoretical probability distribution. Generally, we adopt the chi-square distribution.

In practice, is content to test a model (loosely) by comparing the deviation from the critical value $\chi^2_{n-p}(\alpha)$ of chi-square distribution with a significance level equal to α. If this is greater than $\chi^2_{n-p}(\alpha)$, the model is rejected, and if less than $\chi^2_{n-p}(\alpha)$ or equal, we accepted the model [7].

It should be expected that a good fit to the model data has a deviation close to their degrees of freedom. If the deviance G^2 or the χ^2 exceed the value of its degrees of freedom, it is said that the model is inadequate can handle is a problem of overdispersion [10].

To evaluate the existence of overdispersion must employ a standard criterion. If it adopts the deviation G^2 exceeds the critical value $\chi^2_{n-p}(\alpha)$ up to 10%., whichever is greater, the model is rejected. The effect of this correction is the minimum in the punctual estimators. This method is based on the generalization of [11] for Poisson models.

3. ANALYSIS OF DATA THROUGH THE POISSON REGRESSION MODELS

The data relating to deaths in the city of Santo Angelo (Brazil), were collected in pre-prepared worksheets with the Civil Registry Office in this locality and refer to the period from January 2002 to December 2006.

Analysis of Mortality Data

The data analysis facilitates the identification of deaths by age group and the application of Poisson regression models, as can be seen in Table 1. Through Table 1 shows the incidence of deaths by age group in the city of Santo Angelo.

Table 1: Relationship of Age with Deaths

Age	cit.	Frequency
Less 1	22	4%
01 to 09	5	0,9%
10 to 19	10	1,8%
20 to 29	10	1,8%
30 to 39	17	3,1%
40 to 49	48	8,6%
50 to 59	71	12,8%
60 to 69	111	20,0%
70 to 79	117	21,1%
80 to 89	106	19,1%
More 90	38	6,8%
TOTAL	555	100%

Table **1** shows that the age group with the highest number of deaths is from 70 to 79 years old with 21.1% of them after coming to the 60 to 69 years old with 20%. Where you can see that there is a high life expectancy.

Table **2** shows the parameters of the Poisson regression model.

Table 2: Summary of Model Parameters

	Coefficient	Standard Error(SE)	Coefficient/SE	p-value
X	3,96	0,018	221,56	0

In Table **2** are the model parameters. The Table **3** shows the evaluation criteria of the serial number of deaths in the city of Santo Angelo, by modeling the Poisson regression models.

Table 3: Criteria for Evaluation of the Model

Criteria	Degree of Freedom	Values	Values/DF
Scale of Deviance (G^2)	58	119,21	2,06
Scale of Pearson (X^2)	58	113,40	1,96
Likelihood	58	-232,60	

Table **3** shows that the scale of Deviance and Scale of Pearson that are not suitable for the model, because their degrees of freedom are not near to 1. Therefore one must analyze the data overdispersion. Table 4 shows the data after the overdispersion.

Table 4: Criteria for Evaluation of the Model

Criteria	Degree of Freedom	Values	Values/DF
Scale of Deviance (G^2)	58	58,00	1,00
Scale of Pearson (X^2)	58	55,17	0,95
Likelihood	58	-4936,17	

Table **4** shows that the degrees of freedom are close to 1 this implies that the model is appropriate. After confirmation of the model in question is held now, the short-term prediction, using values of the series that were considered outside the normal range, which follows in Table **5** the actual and predicted values of the model:

Table 5: Short Term Forecast for Total Number of Deaths

Real Value	Prediction
57	47,5
32	47,3

4. CONCLUSION

At the end of this study could be seen that through the analysis of the data could be found that the age group from 70 to 79 years is most prevalent, and also found the model for the serial number of deaths in the city of St. Angelo.

The statistics of the Pearson and Deviance divided by degrees of freedom were used to detect overdispersion or underdispersion had the series in question, taking into consideration that the mean and variance in a Poisson distribution are equal, this implies that the statistic Pearson divided by degrees of freedom should be approximately 1 (one). Values larger than one indicate overdispersion, ie the variance is greater than average, or, underdispersion when the variance is less than the average. Thus, we can expose the series that showed overdispersion or underdispersion.

REFERENCES

[1] Cordeiro GM. Modelos Lineares Generalizados. São Paulo, Campinas UNICAMP/UFPE: 1986.

[2] Nelder JA, Wedderburn R, W., M. Generalized linear models. J Royal Statist Soc v. 1972; 135: pp. 370-384. http://dx.doi.org/10.2307/2344614

[3] Dobson AJ. An introduction to generalized linear models. 2 ed. Chapman & Hall/CRC Press pp. 225.2002.

[4] Cordeiro GM. Introdução à Teoria de verossimilhança. Livro Texto do 10º Simpósio Nacional de Probabilidade e Estatística. UFRJ/ABE. Rio de Janeiro 1992.

[5] Ferrari SLP, David JSE, André PA, Pereira LAA. Use of overdispersed regression models in analyzing the association between air pollution and human health. Relatório Técnico, RTMAE-2002-10, IME-USP 2002.

[6] Rippon P, Rayner J. Assessing Poisson and Logistic Regression Models Using Smooth Tests. Research Online 2011: pp. 1-4.

[7] Schafer JL. Analyses of incomplete multivariate data. London Chapman & Hall 1997.

[8] Mccullagh P, Nelder JA. Generalized Linear Models. Third Edition. New York: Chapman and Hall/CRC. Reprint 1989.

[9] Piegorsch WW. An introduction to binary response regression and associated trend analyses. J Qualit Technol 1998; 30(3): 269-81.

[10] Wang P, Puterman ML, Cockburn LEN. Mixed Poisson Regression Models With Covariate Dependent Rates. Biometrics 1996; 52: pp. 381-400. http://dx.doi.org/10.2307/2532881

[11] Breslow NE. Extra-Poisson variation in log-linear models. Appl Statist 1984; 33: 38-44. http://dx.doi.org/10.2307/2347661

Received on 20-02-2012 Accepted on 11-03-2012 Published on 04-05-2012

http://dx.doi.org/10.6000/1927-5129.2012.08.02.01