University of São Paulo
"Luiz de Queiroz" College of Agriculture

The parametric and semiparametric regression models based on the generalized odd log-logistic family

Fábio Prataviera

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2020

Fábio Prataviera
Degree in Statistics

The parametric and semiparametric regression models based on the generalized odd log-logistic family

Advisor:
Prof. Dr. **EDWIN MOISES MARCOS ORTEGA**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2020

CONTENTS

RESUMO

Modelos de regressão parametricos e semiparametricos baseados na família generalizada odd log-logística

Nesse trabalho foram realizadas diferentes análises via modelos de regressão considerando a família geradora de novas distribuições, denominada de *generalizada odd log-logística-G* (GOLL-G). As distribuições nesta família apresentam maior flexibilidade, como por exemplo, funções de densidades bimodais. Com base na família GOLL-G, foram propostos: modelos de regressão com diferentes estruturas de regressão; modelo semi-paramétrico inflacionado de zeros modelando os parâmetros via splines penalizados; Para todas as abordagens o recurso computacional para implementação dos modelos foi o *software* R, sendo apresentados trechos de comandos ao longo do documento assim como breve descrições dos códigos usados. Os resultados obtidos nas aplicações mostram que o modelo proposto pode ser uma alternativa interessante, principalmente quando os dados apresentam assimetria e bimodalidade.

Palavras-chave: Dados censurados, Fração de cura, Inflação de zeros, Spline cúbico, Simulação

ABSTRACT

The parametric and semiparametric regression models based on the generalized odd log-logistic family

In this work, several analyzes were performed through regression models considering the family of new distributions, called *generalized odd log-logistic-G* (GOLL-*G*), the distributions in this family have greater flexibility, such as functions of bimodal densities. Based on the GOLL-*G* family, we proposed: regression models with different regression structures; inflated semi-parametric model of zeros modeling of the parameters via penalized splines; For all the modeling approaches presented, the computational resource for the implementation of the models was *software* R, throughout the document as well as brief descriptions of the codes used. The results obtained in the applications show that the proposed model can be an interesting alternative, especially when the data present asymmetry and bimodality.

Keywords: Censored data, Cure rate, Cubic spline, Zero inflated, Simulation

1 INTRODUCTION

Regression analysis is a commonly used statistical technique applied in many scientific fields. The linear regression model with normal distribution is generally used to model data having symmetric distribution. However, various phenomena cannot always be modeled with the normal distribution, be it for the lack of symmetry, the existence of bimodality or the presence of atypical values.

In past decades, when the phenomenon of interest did not satisfy the assumption of normality of the response variable, some type of transformation was applied at least to obtain symmetric behavior of the data. However, recently it has become more attractive to propose new regression models to model different types of data.

In this work we use regression model to solve problems in different areas. For example, in survival analysis the study of the lifetime of patients with a particular disease and the study of the failure time of an electronic component. The study of times is called survival analysis in the medical area and reliability analysis in the industrial area.

There are also situations where continuous data can include a high percentage of zeros. In these situations, continuous distributions can not be used. The data that contain excessive zeros can be analyzed by a mixture of two distributions: a continuous distribution (with positive support) and a degenerate distribution at zero, i.e. a model whose mixed discrete-continuous probability and distribution functions.

Among the different proposed models and families of distribution, it is notable that only a small number take bimodal forms. In this work a new model based on the *generalized odd log-logistic - G* (Cordeiro *et al.* 2017) family is proposed. We consider the bade distribution the Maxwell distribution.

The Maxwell (or Maxwell-Boltzmann) distribution is an important model in physics, chemistry and statistical mechanics. It forms the basis of the kinetic energy of gases and explains several fundamental properties of gases including pressure and diffusion. In statistical mechanics, it is related to properties of molecules in thermal equilibrium from the microscopic perspective. The Maxwell distribution is also important in kinetic translational energies for molecules. For example, Prigogine and Xhrouet (1949) discussed this distribution for chemical reactions in gases and Brilliantov and Poschel (2000) studied its deviations in granular gases with constant coefficient of restitution.

In recent years, the Maxwell distribution has been used to model failure times in survival and reliability analysis and some of its extended forms have been investigated. Krishna *et al.* (2012) addressed reliability estimation in the Maxwell distribution with progressively type-II censored data, Kazmi *et al.* (2012) explored a heterogeneous population by means of two mixture components of Maxwell distributions, Tomer and Panwar *et al.* (2015) considered point and interval estimation procedures for the Maxwell distribution in the presence of type-I progressively hybrid censored data, Dey *et al.* (2016) presented its structural properties and different methods of estimation, Iriarte *et al.* (2016) defined the gamma-Maxwell distribution and, more recently, Venegas *et al.* (2017) proposed the transmuted exponentiated Maxwell distribution. However, none of these papers deal with bimodality to real data and do not even present regression models for the extensions of the Maxwell distribution. We aim to fill up this gap.

Based on the proposed model, we try to solve problems from different areas based on regression models. However, in many situations the relationship between the response variable and the explanatory (or covariate) variable has no linear relationship. This can often make it difficult to explain this relationship. In this work we propose the generalized odd log-logistic Maxwell parametric and semiparametric regression models to solve the problems above. In addition, to solve the issue of nonlinear behavior and in

order to obtain a more flexible model for the data we use cubic splines in this work. Thus, the inclusion of cubic splines in the model requires, in addition to descriptive and exploratory analyzes, diagnostic analyzes to assess the suitability of the model.

Another situation that occurs in many studies in several fields aim to determine how a set of explanatory variables influence other variables expressed as ratios or proportions, i.e., random experiments that produce results in the interval $(0, 1)$. Several researchers tried to model this type of data. For example, Ferrari and Cribari-Neto (2004) pioneered a regression in which the parameters are interpreted as mean and precision, Bayes *et al.* (2012) proposed a robust regression for proportions based on the beta rectangular distribution, Lemonte and Bazán (2016) defined a class of Johnson SB distributions and its associated regression for rates and proportions, Mazucheli *et al.* (2019) proposed a unit-Lindley distribution and its associated regression for proportional data. In these terms, our main aim is to propose a regression based on the generalized odd log-logistic beta ("GOLLBE" for short) distribution to model proportional data with bimodality.

All computational scripts of the new regression model were implemented in the `R` software using the `gamlss` package (Stasinopoulos and Rigby, 2007).

2 CONCLUSION

In this work we propose parametric and semiparametric regression models based on the family *generalized odd log-logistic-G*. Thus, two new models were proposed for the *generalized odd log-logistic Maxwell* (GOLLMax) distribution for for data on positive support and the *generalized odd log-logistic beta* (GOLLBE) distribution for data analysis in the unit interval. Various mathematical properties of the GOLLMax and GOLLBE distribution are investigated. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality. The former class of GOLLMax regression models is very suitable for modeling censored and uncensored lifetime data.

Based on the GOLLMax distribution, we propose a new distribution called *zero adjusted generalized odd log-logistic Maxwell* (ZAGOLLMax). For this model, we present a ZAGOLLMax semiparametric regression model to analyze soil microbiology data. We also propose a mixture model called *generalized odd log-logistic Maxwell mixture* (GOLLMaxM), with application to a prostate cancer dataset.

We use the `gamlss` script in the `R` package to obtain the maximum likelihood estimates and perform asymptotic tests for the model parameters based on the asymptotic distribution of the estimates.

References

Bayes, C.L., Bazán, J.L. and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, **4**, 841-866.

Brilliantov, N.V. and Poschel, T. (2000). *In Granular Gases*. Springer, Berlin.

Cordeiro, G.M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E.M.M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

Dey, S., Dey, T., Ali, S. and Mulekar, M.S. (2016). Two-parameter Maxwell distribution: Properties and different methods of estimation. *Journal of Statistical Theory and Practice*, **10**, 291-310.

Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Iriarte, Y.A., Astorga, J.M., Bolfarine, H. and Gómez, H.W. (2017). Gamma-Maxwell distribution. *Communications in Statistics-Theory and Methods*, **46**, 4264-4274.

Kazmi, S., Aslam, M. and Ali, S. (2012). On the Bayesian estimation for two component mixture of Maxwell distribution, assuming type I censored data. *SOURCE International Journal of Applied Science and Technology*, **2**, 197-218.

Krishna, H. and Malik, M. (2012). Reliability estimation in Maxwell distribution with progressively type-II censored data. *Journal of Statistical Computation and Simulation*, **82**, 623-641.

Lemonte, A.J. and Bazán, J.L. (2016). New class of Johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**, 727-746.

Mazucheli, J., Menezes, A.F.B. and Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, **46**, 700-714.

Prigogine, I., Xhrouet, E. (1949). On the perturbation of Maxwell distribution function by chemical reactions in gases. *Physica*, **15**, 913-932.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R, *J. Stat. Softw.* **23**, pp. 1-46.

Tomer, S.K. and Panwar, M. S. (2015). Estimation procedures for Maxwell distribution under type-I progressive hybrid censoring scheme. *Journal of Statistical Computation and Simulation*, **85**, 339-356.

Venegas, O., Iriarte, Y.A., Astorga, J.M., Borger, A., Bolfarine, H. and Gómez, H.W. (2017). A New Generalization of the Maxwell Distribution. *Applied Mathematics and Information Sciences*, **11**, 867-876.