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

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

Julio Cezar Souza Vasconcelos

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

Piracicaba
2021

Julio Cezar Souza Vasconcelos
Degree in Mathematics

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2021

SUMMARY

RESUMO

O modelo de regressão paramétrico, semiparamétrico e de efeito aleatório baseado na extensão da distribuição Gaussiana inversa generalizada

Propomos um modelo de regressão baseado na distribuição de quatro parâmetros denominada odd log-logistic Gaussiana inversa generalizada (OLLGIG) com dois componentes sistemáticos adequados para dados unimodais e bimodais que estendam o modelo de regressão GIG heterocedástico. Os modelos de regressão aditivo, parcial ou semiparamétrico podem ser uma opção quando a variável resposta e a variável explicativa tem uma relação não linear, ou seja, não é mais levado em conta uma pressuposição fundamental de linearidade entre essas variáveis. Pensando nisso é proposto três modelos flexíveis denominados de modelos de regressão aditivo, parcial e semiparamétrico baseado na distribuição OLLGIG com uma estrutura sistemática, considerando três diferentes tipos de suavizações penalizadas gerados por splines. Muitos estudos nas áreas de saúde pública, economia, agronomia, medicina, biologia e ciências sociais, entre outros, envolvem observações repetidas de uma variável resposta. A expressão "medidas repetidas" é utilizada para designar medidas obtidas para a mesma variável ou na mesma unidade experimental em mais de uma ocasião. Vários projetos experimentais com medidas repetidas são comuns, como split-plot, crossover e longitudinal. Esses tipos de investigações são denominados estudos de dados correlacionados e desempenham um papel fundamental na análise dos resultados, onde é possível caracterizar alterações nas características de um indivíduo, associando essas variações a um conjunto de covariáveis. Devido à sua natureza, as medidas repetidas possuem uma estrutura de correlação que desempenha um papel importante na análise desses tipos de dados, além disso, a distribuição da variável resposta pode apresentar assimetria ou bimodalidade. Assim, é introduzida uma regressão com intercepto aleatório normal com base na distribuição OLLGIG. Na regressões linear e com efeito aleatório e adotado o método de máxima verossimilhança, já para os modelos: aditivo, parcial e semiparamétrico OLLGIG e utilizado o método de máxima verossimilhança penalizada para estimar os parâmetros do modelos propostos. Além disso, diversas simulações são realizadas para diferentes configurações de parâmetros e tamanhos de amostras para verificar a precisão dos estimadores de máxima verossimilhança e máxima verossimilhança penalizada. São realizadas análises de diagnósticos baseada em case-deletion e resíduos quantílicos. Para comprovar a potencialidade dos modelos de regressão propostos, são realizados ajustes com dados reais.

Palavras-chave: Gerador Odd Log-Logístico, Modelo Aditivo, Modelo Parcial, Modelo Semiparamétrico, Efeito Aleatório

ABSTRACT

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

We propose a regression model based on the four-parameter distribution called generalized inverse Gaussian odd log-logistic (OLLGIG) with two systematic components suitable for unimodal and bimodal data that extends the heteroscedastic GIG regression model. Additive, partial or semi-parametric regression models can be an option when the response variable and the explanatory variable have a nonlinear relationship, that is, the fundamental assumption of linearity between these variables does not hold. With this in mind, three flexible models are proposed, namely additive, partial and semiparametric regression models based on the OLLGIG distribution with a systematic structure, considering three different types of penalized smoothings generated by splines. Many studies in the areas of public health, economics, agronomics, medicine, biology and social sciences, among others, involve repeated observations of a response variable. The expression " repeated measures " is used to designate measurements obtained for the same variable or in the same experimental unit on more than one occasion. Various experimental designs with repeated measurements exist, such as split-plot, crossover and longitudinal. These types of investigations are called studies of correlated data and play a fundamental role in the analysis of results, where it is possible to characterize changes in the characteristics of an individual by associating these variations to a set of covariates. Due to their nature, the repeated measures have a correlation structure that plays an important role in the analysis of these types of data. In addition, the distribution of the response variable may present asymmetry or bimodality. Thus, a regression with a normal random intercept is introduced based on the OLLGIG distribution. In linear and random regressions, the maximum likelihood method is adopted for the models: additive, partial and semiparametric OLLGIG and the penalized maximum likelihood method are used to estimate the parameters of the proposed models. In addition, several simulations are performed for different parameter configurations and sample sizes to verify the accuracy of the maximum likelihood and penalized maximum likelihood estimators. Diagnostic analyses based on case-deletion and quantile residuals are performed. To prove the potential of the proposed regression models, adjustments are made with real data.

Keywords: Odd Log-Logistic Generator, Additive Model, Partial Model, Semiparametric Model, Random Effect

1 INTRODUTION

The inverse Gaussian (IG) distribution is used to model many phenomena in diverse areas, such as economics, engineering, business, social policy, real estate market, and natural events, among others. An extension of the IG distribution that has been used widely is the generalized inverse Gaussian (GIG), which has positive support. It was initially proposed by Good (1953) in a study of population frequencies. Many papers have examined the structural properties of the GIG distribution. Sichel (1975) used it to produce mixtures of Poisson distributions. The behavior of the GIG distribution and various of its statistical properties were addressed by Jørgensen (1982) and Atkinson (1982). Dagpunar (1989) proposed algorithms to simulate this distribution. Nguyen et al. (2003) pointed out that it has positive asymmetry. Madan et al. (2008) demonstrated that the Black-Scholes formula in finance can be expressed in terms of a function of the GIG distribution. More recently, Koudou and Ley (2014) published a list of its properties and Lemonte and Cordeiro (2011) described some mathematical properties of the exponentiated generalized inverse Gaussian distribution (EGIG).

In the majority of experiments, the response variable is influenced by explanatory variables that elucidate determined characteristics of individuals. Thus, the inclusion of covariables in a regression model is a way to represent the heterogeneity of a population. These covariables, in turn, should be considered in some way in the model to increase its predictive power. The statistical literature contains many types of regression models, such as the semiparametric generalized linear model proposed by Green and Yandell (1985), in which the authors added a nonparametric term in the linear predictor. Another extension of generalized linear models is the generalized additive model (GAM) proposed by Hastie and Tibshirani (1990), in which the term that is controlled in parametric form is altered by an arbitrary function and comes to be controlled in nonparametric form, estimated by smoothed curves (e.g., splines). Rigby and Stasinopoulos (2001) developed generalized additive models for location, scale and shape (GAMLSS), which are widely used in various areas of science. This type of modeling is very flexible, because it allows modeling the location, scale and shape parameters simultaneously.

Several papers have proposed regression models with random effects. Among these works are those of Muniz-Terrera et al. (2016), who developed random effect parametric and nonparametric regressions to analyze cognitive test data; Coupé (2018), who reported advances in statistical modeling in linguistics based in linear mixed-effects regressions; Ho et al. (2019), who presented an analysis of microbiome relative abundance data using a zero-inflated beta GAMLSS model and a meta-analysis of studies using random effects models; Hashimoto et al. (2019), who introduced the random effect log-Burr XII regression; and Dirmeier et al. (2020), who presented host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. In this sense, in this work we propose parametric, semiparametric and random effect regression models.

For this purpose, our first objective is to define a new four-parameter distribution called the odd log-logistic generalized inverse Gaussian (OLLGIG) to model data pertaining to areas such as the real estate market, engineering and natural phenomena, among others. This model is noteworthy because besides modeling unimodal data, it can also model data where the response variable is bimodal, making it an alternative that in many cases can be more effective than mixture models, in which different situations require two distributions to enable modeling data where the response variable has two modes, making the model more complex. In turn, with respect to semiparametric models, our second objective is to build a regression model based on the OLLGIG distribution that can model unimodal and bimodal data as well as data using extensions of linear regression models, such as the additive, partial and semiparametric cases, in which the different systematic penalized smoothing routines, consisting of splines, are considered in the systematic component. And finally, our third objective is to analyze the correlated data in the presence of bimodality and asymmetry, and based on the studies described, to perform regression with

normal random intercepts based on the OLLGIG for the purpose of considering the possible presence of heterogeneity among some cities in the state of São Paulo, Brazil.

We also describe as special cases the generalized inverse Gaussian (GIG) and inverse Gaussian (IG) distributions, obtain some mathematical properties and discuss the maximum likelihood and penalized maximum likelihood estimation of the parameters. For these models, we present some ways to include global influence (case deletion), and also develop residual analyses based on quantile residuals. Several simulation studies are presented for different configurations of the parameters and sample sizes, and the empirical distribution of the residuals is shown and compared with the standard normal distribution. The results of these studies suggest that the empirical distribution of the quantile residuals for the OLLGIG regression model with two regression structures, along with the additive, partial and semiparametric models, as well as those with random effect on the intercept, have high concordance with the standard normal distribution. This qualifying material is organized as follows.

In Chapter 2, we define a new four-parameter model called the odd log-logistic generalized inverse Gaussian distribution, which extends the generalized inverse Gaussian distributions. We obtain some structural properties of the new distribution and construct an extended regression model based on this distribution with two systematic structures. We adopt the method of maximum likelihood to estimate the model parameters. In addition, various simulations are performed for different parameter settings and sample sizes to check the accuracy of the maximum likelihood estimators. We provide a diagnostics analysis based on case-deletion and quantile residuals. Finally, the potential of the new regression model to predict urban property values is illustrated by means of real data.

In Chapter 3 we propose three flexible regression models, called additive, partial and semiparametric, based on the odd log-logistic generalized inverse Gaussian distribution under three types of penalized smoothing. We adopt the penalized maximum likelihood method to estimate the parameters of the proposed regression models. Furthermore, we present several simulations carried out for different configurations of the parameters and sample sizes to verify the precision of the penalized maximum likelihood estimators. The regression is applied to ethanol data and air quality data.

In Chapter 4, a random effect regression is defined to model correlated data. The maximum likelihood is adopted to estimate the parameters and various simulations are performed for correlated data. Residuals are proposed for the new regression whose empirical distribution is close to normal. The usefulness of the regression is verified based on the average price per hectare of bare land in 10 cities in the state of São Paulo (Brazil).

References

Atkinson, A. (1982). The simulation of generalized inverse gaussian and hyperbolic random variables. *SIAM Journal on Scientific and Statistical Computing*, 3(4):502–515.

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, 9:513.

Dagpunar, J. (1989). An easily implemented generalised inverse gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710.

Dirmeier, S., Dächert, C., van Hemert, M., Tas, A., Ogando, N. S., van Kuppeveld, F., Bartenschlager, R., Kaderali, L., Binder, M., and Beerenwinkel, N. (2020). Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. *PLoS Computational Biology*, 16(2):e1007587.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models*, pages 44–55. Springer.

Hashimoto, E. M., Silva, G. O., Ortega, E. M., and Cordeiro, G. M. (2019). Log-burr xii gamma–weibull regression model with random effects and censored data. *Journal of Statistical Theory and Practice*, 13(2):1–21.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Ho, N. T., Li, F., Wang, S., and Kuhn, L. (2019). metamicrobiomer: an r package for analysis of microbiome relative abundance data using zero-inflated beta gamlss and meta-analysis across studies using random effects models. *BMC Bioinformatics*, 20(1):188.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer Science & Business Media, New York.

Koudou, A. E. and Ley, C. (2014). Efficiency combined with simplicity: new testing procedures for generalized inverse gaussian models. *Test*, 23(4):708–724.

Lemonte, A. J. and Cordeiro, G. M. (2011). The exponentiated generalized inverse gaussian distribution. *Statistics & Probability Letters*, 81(4):506–517.

Madan, D., Roynette, B., and Yor, M. (2008). Unifying black–scholes type formulae which involve brownian last passage times up to a finite horizon. *Asia-Pacific Financial Markets*, 15(2):97–115.

Muniz-Terrera, G., Hout, A. v. d., Rigby, R., and Stasinopoulos, D. (2016). Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical Methods in Medical Research*, 25(2):741–753.

Nguyen, T. T., Chen, J. T., Gupta, A. K., and Dinh, K. T. (2003). A proof of the conjecture on positive skewness of generalised inverse gaussian distributions. *Biometrika*, pages 245–250.

Rigby, R. and Stasinopoulos, D. (2001). The gamlss project: a flexible approach to statistical modelling. In *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, volume 337, page 345. University of Southern Denmark.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

2 FINAL CONSIDERATIONS

In this thesis, I develop flexible parametric and semiparametric regression models based on an extension of the generalized inverse Gaussian distribution and on the odd log-logistic generator of distributions. The new proposed distribution is called the odd log-logistic generalized inverse Gaussian (OLLGIG). The main characteristic of this distribution is that it permits modeling bimodal data without the need to use a mixture of distributions. I use the gamlss package available in the R software to obtain the maximum likelihood and penalized maximum likelihood estimates (MLE and PMLE), as well as to evaluate the sensitivity (global influence and analysis of residuals) of the proposed regression models. In Chapter 2, I define the OLLGIG distribution and describe various structural properties. I then present various simulation studies to evaluate the performance of the MLEs and to study the distribution and residuals utilized empirically to validate the assumptions of the proposed regression models. Finally, I present two applications to real data, the first without considering covariables and the second considering covariables in two systematic components. In Chapter 3, I propose additive, partial and semiparametric regression models based on the OLLGIG distribution and consider three types of penalized smoothers, as well as discussing selection criteria, sensitivity analysis and residuals. Finally, data on climatology, ethanol use and air quality are used to verify the versatility of the proposed models. In Chapter 4 I propose the regression model with fixed effect in the intercept based on the OLLGIG distribution, applying it to a dataset on land values per hectare in some cities in the state of São Paulo, Brazil.