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

New flexible parametric and semiparametric models for survival analysis

Thiago Gentil Ramires

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

Piracicaba
2017

Thiago Gentil Ramires
Degree in Statistics

New flexible parametric and semiparametric models for survival analysis

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

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

Piracicaba
2017

RESUMO

Novos modelos flexíveis paramétricos e semi-parametricos para análise de sobrevivência

Nesse trabalho foi proposto uma nova distribuição, denominada de exponentiated log-sinh Cauchy, a qual possui densidades bimodais e pode ser utilizada como alternativa aos modelos de mistura. Com base na nova distribuição, foram propostos: modelos de regressão baseados nos modelos GAMLSS; modelos com fraąão de cura baseados em modelos de mistura e tempo de promoção; modelo semi-paramétrico modelando os parâmetros com splines penalizados; modelo semi-paramétrico com fração de cura utilizando splines para modelar efeitos não lineares na proporção de curados. Para todos os modelos propóstos, toda parte computacional foi implementada no *software* R, sendo disponibilizada ao longo do documento assim como breve descrições de uso.

Palavras-chave: Bimodalidade, GAMLSS, P-splines, Fração de cura

ABSTRACT

New flexible parametric and semiparametric models for survival analysis

In this work was proposed a new distributions, called log-sinh Cauchy, with has bimodal shapes and can be used as alternative to the mixture models. Based in the proposed distribution, the following models were proposed: Regression model based in the GAMLSS framework; models with cure rate based in the mixture and promotion time models; semiparametric models, modeling the parameters using penalized splies; semiparametric models, using the penalized splines to model the non-linear effects present in the cure rate. For all proposed models, the computational codes were implemented in the R software, with is available along of the document as well as some brief introduction on how to use them.

Keywords: Bimodality, GAMLSS, P-splines, Cure rate

1 INTRODUCTION

Present in virtually all areas, statistics is an outstanding tool for data analysis. Among these areas is the survival analysis, which has applications in several areas of research, like medicine, agronomy, engineering, biology, economics and other areas related to health and finance. The increasing use of statistics is due in part to the development of more efficient techniques and methods along with technological and computational advances that allow the creation of more sophisticated models for analyzing data with different behavior than usually found in case studies in the literature.

With the ease of database construction, new density behaviors related to variable responses are emerging, which in some situations require extremely complicated shapes and more complex models. Recently, several models have been proposed in the survival analysis literature, which have greater flexibility, resulting in more accurate estimates and analyses. Mixtures and transformations between distributions generate interesting results when applying probability density failure or risk rate functions. Over the past 10 years, hundreds of new models have been proposed in the survival analysis literature, for which a brief discussion can be found in Tahir and Nadarajah (2015).

Among the different proposed models, it is notable that only a small number take bimodal forms. Data that exhibit bimodal behavior arise in many different disciplines. In medicine, urine mercury excretion has two peaks, see for example, Ely *et al.* (1999). In material characterization, in a study conducted by Dierickx *et al.* (2000), grain size distribution data revealed a bimodal structure. In meteorology, Zhang *et al.* (2003) found that water vapor levels in tropical regions commonly have bimodal distributions. Furthermore, most models that are able to assume bimodal forms have positive skewness, and are inefficient to fit symmetric or negative symmetry. Alternatively, many authors have used mixture distributions to model data with bimodal behavior, associating a specific distribution for each modal region.

Due to the need for new models capable of capturing bimodal forms, present a new model for survival analysis called "exponentiated log-sinh Cauchy", which has four parameters and is able to take symmetrical bimodal or positive or negative asymmetrical shape. Properties, applications, simulations and computational implementation of the new model are also presented.

In many cases, the response variable's behavior is influenced by other variables, called explanatory variables. In such cases, it is necessary to add these variables in statistical models to achieve better interpretation. One of the most common methods of relating the explanatory variables with the response variable is to use the class of location-scale models. But location-scale models relate only the location parameter with the explanatory variables, so in many cases it is necessary to use more complex models to get a good fit, which would not be needed if the scaling, kurtosis or others parameters were also modeled by explanatory variables. In this sense, we present a regression model, based in the exponentiated log-sinh Cauchy model, which belongs to the "generalized additive models for location, scale and shape" (GAMLSS) class of models (Rigby and Stasinopouls , 2005).

The advantage of the class when compared to the location-scale class of models is that all parameters can be explained by explanatory variables, which in the case of exponentiated log-sinh Cauchy model are the location, scale, bimodality and skewness parameters. All computational scripts of the new regression model were implemented in the R software (Team, 2013) using the GAMLSS package (Stasinopoulos and Rigby, 2007) and are available, for easy use by anyone familiar with the R software.

Models for survival analysis typically consider that every subject in the study population is susceptible to the event under study and will eventually experience such event if follow-up is sufficiently long. However, there are situations when a fraction of individuals are not expected to experience the event of interest, that is, those individuals are cured or not susceptible. Based in the mixture models (MMs) pioneered by Boag (1949), Berkson and Gage (1952), we propose a new cure rate model based

on the "exponentiated log-sinh Cauchy" distribution. Using the GAMLSS framework, we can model the location, scale, bimodality, skewness and cure rate parameters. Base on the promotion time cure models (Yakovlev and Tsodikov, 1996), we also proposed a new model to estimate breast carcinoma mortality, assuming that the number of competing causes that can influence the survival time follows a Poisson distribution.

When using the parametric regression models belonging to the class of location-scale or GAMLSS models, in many situations the explanatory variables do not have a linear relation with the dependent variable, requiring the use of nonlinear functions to explain its behavior. Among various nonlinear functions, the splines (the focus of this paper) stand out for being extremely flexible in capturing various types of behavior. Currently splines are used especially considering the Cox models (Cox, 1972). Although becoming more popular in the literature, there are few references on the use of splines in the class of location-scale and GMLSS models.

In this context, we propose a new semiparametric heteroscedastic regression model allowing for positive and negative skewness and bimodal shapes using the B-spline basis for nonlinear effects. The proposed distribution is based on the generalized additive models for location, scale and shape framework in order to model any or all the parameters of the distribution using parametric linear and/or nonparametric smooth functions of explanatory variables. Finally the idea of the semiparametric models are extended for the new cure rate models, being possible to estimate nonlinear effects of explanatory variable in the cure rate parameter.

References

Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*,**47**, 501–515.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B,* **11**, 15–53.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.

Dierickx, D., Basu, B., Vleugels, J. and Van der Biest, O. (2000). Statistical extreme value modeling of particle size distributions: experimental grain size distribution type estimation and parameterization of sintered zirconia. *Materials characterization*, **45**, 61–70.

Ely, J.T.A., Fudenberg, H.H., Muirhead, R.J., LaMarche, M.G., Krone, C.A., Buscher, D. and Stern, E.A. (1999). Urine mercury in micromercurialism: bimodal distribution and diagnostic implications. *Bulletin of environmental contamination and toxicology*, **63**, 553–559.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–554.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1–46.

Tahir, M.H. and Nadarajah, S. (2015). Parameter induction in continuous univariate distributions: Well-established G families. *Anais da Academia Brasileira de Ciências*, **87**, 539–568.

Team, R.C. (2000). R Language Definition.

Zhang, C., Mapes, B.E. and Soden, B. J. (2003). Bimodality in tropical water vapour. *Quarterly Journal of the Royal Meteorological Society*, **129**, 2847–2866.

Yakovlev A and Tsodikov AD. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications. Mathematical Biology and Medicine, Vol. 1.* World Scientific, New Jersey.

2 CONCLUSION

The paper proposes the exponentiated log-sinh Cauchy (ELSC) distribution that can be used as an alternative to mixture distributions in modeling bimodal data. Various mathematical properties of the ELSC distribution are investigated. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality.

Based on the ELSC distribution, we propose a general class of exponentiated sinh Cauchy (ESC) regression models, where the mean, dispersion, skewness and bimodal parameters vary across observations through regression structures. The former class of regression models is very suitable for modeling censored and uncensored lifetime data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. 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. We offer some interesting insights, especially regarding model checking, and provide applications of influence diagnostics (global, local and total influence) in the proposed class of regression models with censored data.

In the context of cure rate models, we introduce the exponentiated log-sinh Cauchy cure rate (ELSCcr) model that can be used as an alternative to mixture distributions in modeling bimodal data with or without the presence of immune proportion of individuals. Three real data examples prove empirically that the ELSCcr distribution is very flexible, parsimonious, and a competitive model that deserves to be added to existing distributions in modeling bimodal data. We also presents the parametric log-sinh Cauchy promotion time generalized additive model for location, scale and shape (LSCp GAMLSS) to estimate breast carcinoma mortality, assuming that the number of competing causes that can influence the survival time follows a Poisson distribution.

Considering the presence of non-linear effects occurred by explanatory variables, we present the semiparametric ESC regression model, where the parameters of the model can be modelled as parametric or smooth nonparametric functions of explanatory variables. Two real data sets are used to illustrate the importance of the semiparametric ESC regression model, showing that it provides better performance than the usual methods in the presence of bimodal and asymmetric random errors.

Finally, the semi-parametric *log-sinh Cauchy cure rate* (LSCcr) regression model was proposed, where the cure rate parameter can also be modeled using parametric or smooth nonparametric functions of explanatory variables. A real data set is used to illustrate the usefulness of the semi-parametric LSCcr regression model, showing that it provides better performance than the usual methods in the presence of nonlinear effects in the cure rate proportion.