






ARTICLE

Using the Box-Cox family of distributions to model censored data: a distributional regression approach

 Luiz R. Nakamura,^{*,1}  Thiago G. Ramires,²  Ana J. Righetto,³  Viviane C. Silva,⁴ and  Andréa C. Konrath⁵

¹Department of Statistics, Federal University of Lavras, Lavras, Brazil

²Department of Mathematics, Federal Technological University of Paraná, Apucarana, Brazil

³Alvaz Agritech, Londrina, Brazil

⁴Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Federal University of Lavras, Lavras, Brazil

⁵Department of Informatics and Statistics, Florianópolis, Brazil

*Corresponding author. Email: luiz.nakamura@ufla.br

(Received: June 1,2022; Revised: December 14,2022; Accepted: December 14,2022; Published: December 30,2022)

Abstract

The study of the expected time until an event of interest is a recurring topic in different fields, such as medical, economics and engineering. The Kaplan-Meier method and the Cox proportional hazards model are the most used methodologies to deal with such kind of data. Nevertheless, in recent years, the generalised additive models for location, scale and shape (GAMLSS) models – which can be seen as distributional regression and/or beyond the mean regression models – have been standing out as a result of its highly flexibility and ability to fit complex data. GAMLSS are a class of semi-parametric regression models, in the sense that they assume a distribution for the response variable, and any and all of its parameters can be modelled as linear and/or non-linear functions of a set of explanatory variables. In this paper, we present the Box-Cox family of distributions under the distributional regression framework as a solid alternative to model censored data.

Keywords: GAMLSS; Kidney disease; Renal insufficiency.

1. Introduction

Survival analysis is a set of methods used to understand and model the behaviour of the time until a certain event of interest (e.g., death or component failure). Due to its characteristic, the event of interest may not occur, and then the observations are considered censored, thus making it impossible to measure our main response (Gijbels, 2010).

As stated by Emmert-Streib & Dehmer (2019), there are two main methodologies to analyse such kind of data, namely the Kaplan–Meier method (Kaplan & Meier, 1958) and the Cox proportional hazards model (Cox, 1972). The main difference between both is that the first one is not able to consider multiple features to explain the target variable, whereas the Cox model can.

Another interesting alternative may be the distributional regression models, which, as stated by Heller *et al.* (2022), was first proposed by Rigby & Stasinopoulos (2005) as the generalised additive models for location, scale and shape (GAMLSS), a class of regression models that extends the well-known generalised linear models (Nelder & Wedderburn, 1972) and generalised additive models (Hastie & Tibshirani, 1990). In fact, different works have already considered this framework to model censored data, such as Castro *et al.* (2010), Alizadeh *et al.* (2019), Ramires *et al.*, 2018, Ramires *et al.* (2019), Ramires *et al.* (2021a) and Ramires *et al.* (2021b).

The key factor in this flexible approach is that any and all parameters (not only the location, which is often the mean) of the assumed response variable distribution (that does not necessarily belong to the exponential family) can be modelled as linear and/or non-linear functions of the explanatory variables (Rigby & Stasinopoulos, 2005), i.e., GAMLSS belongs to the beyond mean regression models (Kneib, 2013), since different regression structures (considering different set of covariates) are fitted to explain different (and possibly complex) characteristics of the response (e.g., skewed and platokurtic/leptokurtic data).

Since any distribution may be used in the GAMLSS framework, in this paper, we will present the Box–Cox family of distributions (Rigby *et al.*, 2019) as a strong candidate to model censored data.

2. GAMLSS framework

Generically, let T follow a distribution $\mathcal{D}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector, then a GAMLSS can be defined as

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{jk}(\mathbf{x}_{jk}), \quad (1)$$

where $g_k(\cdot)$, $k = 1, \dots, p$, denotes a link function related to the k th parameter, \mathbf{X}_k is a known model matrix, $\boldsymbol{\beta}_k$ is a parameter vector, $s_{jk}(\cdot)$ are smooth functions of \mathbf{x}_{jk} . Note that if $\sum_{j=1}^{J_k} s_{jk}(\mathbf{x}_{jk}) = 0$,

as in Nakamura *et al.* (2019), model (1) is reduced to its parametric version. It is noteworthy that the assumed distribution for T does not necessarily belong to the exponential family. For a large list of distributions already implemented in the `gamlss` package (Stasinopoulos & Rigby, 2007) in R software (R Core Team, 2022), check Rigby *et al.* (2019).

In this paper we will focus on the Box–Cox family of distributions under the GAMLSS framework. Let us consider the following transformed random variable Z , given by

$$Z = \begin{cases} \frac{1}{\sigma^\nu} \left[\left(\frac{T}{\mu} \right)^\nu - 1 \right], & \text{if } \nu \neq 0 \\ \frac{1}{\sigma} \log \left(\frac{T}{\mu} \right), & \text{if } \nu = 0 \end{cases},$$

for $T > 0$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$. If Z follows a truncated standard normal distribution, then T will follow a Box–Cox Cole and Green distribution (Cole & Green, 1992), denoted by BCCG(μ, σ, ν). If Z follows a truncated standard power exponential distribution with power parameter $\tau > 0$, then T follows a Box–Cox power exponential distribution (Rigby & Stasinopoulos, 2004), denoted by BCPE(μ, σ, ν, τ). If Z is assumed to follow a truncated standard t distribution

with $\tau > 0$ degrees of freedom, then T follows a Box-Cox t distribution (Rigby & Stasinopoulos, 2006), denoted by $BCT(\mu, \sigma, \nu, \tau)$.

The key aspect is that the Box-Cox family of distributions is very flexible and reliable (Rigby *et al.*, 2019) and its parameters have straightforward interpretation, which is a critical characteristic in the process of choosing a GAMLSS model (Ramires *et al.*, 2021a). In the three distributions – BCCG, BCPE and BCT – μ is the median, σ is the approximate coefficient of variation and ν is a skewness parameter. Further, in the last two distributions, τ is a kurtosis parameter, in which BCPE is able to model both leptokurtosis and platykurtosis and BCT models leptokurtosis (Rigby *et al.*, 2019).

Regarding the estimation process for a GAMLSS in its parametric version, let us consider that t_1, \dots, t_n is a sample of n independent observations, c_i denotes the censoring time, $\gamma_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$, where $\delta_i = 1$ if t_i is a time-to-event and $\delta_i = 0$ if it is right censored. Hence, we shall maximise the total log-likelihood function under non-informative censoring, which is given by

$$l(\theta) = \sum_{i \in F} \log f(t_i | \theta_i) + \sum_{i \in C} \log S(t_i | \theta_i),$$

where $S(t_i | \theta_i)$ is the survival function related to the distribution \mathcal{D} , F and C denote the sets of individuals for each t_i . Note that we are considering the GAMLSS as its parametric version. The numerical maximisation is performed using the Rigby and Stasinopoulos (RS) algorithm (Rigby & Stasinopoulos, 2005).

In order to properly select different features (explanatory variables) in each of the regression structures in (1), considering the Box-Cox family of distributions, we apply the Strategy A, a stepwise-based procedure described in details by Stasinopoulos *et al.* (2017), Nakamura *et al.* (2017) and Ramires *et al.* (2021b).

Finally, we use the normalised quantile residuals (Dunn & Smyth, 1996) in order to assess the adequacy of the fitted model. The main advantage here is that, regardless of the assumed distribution for the response variable, the residuals will necessarily follow a standard normal distribution if the fitted model is appropriate.

3. The data

The data analysed in this paper were collected from the Maringá Kidney Institute, for 177 patients (80 women and 97 men) classified with chronic renal failure enrolled in a hemodialysis program from 1978 up to 2010, and it was provided by Ramires *et al.* (2018). As specified in the authors' paper, the response variable is observed time (in days), of which the total of failure times (death) is 119 and 58 observations were considered as censored (if the patient did not continue in the program for any reason or if the patient did not die until the end of the study). The explanatory variables that were considered to explain the response are displayed in Table 1.

Ramires *et al.* (*ibid.*) already analysed these data under the GAMLSS framework, based on the exponential, Weibull and log-normal distributions. In their paper, considering both Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978), they found out that the Weibull distribution was the most appropriate. Hence, in this paper, we will compare their results with the Box-Cox family of distributions.

4. Results and Discussion

Table 2 displays values of the global deviance – minus twice the fitted log-likelihood –, AIC and BIC for each fitted model selected through the Strategy A. As can be seen, the best fitted model

Table 1. Covariates of the chronic renal failure data

Variable	Label	Type	Range
Age at the beginning of treatment	age	Quantitative	17 – 88 years
Sex	sex	Factor	male, female
Marital status	mar	Factor	living together, married, separated, divorced single, widowed, others
Skin colour indicator	col	Factor	yellow, white, black, pardo
Antibodies to hepatitis B	hepB	Factor	true, false
Antibodies to hepatitis C	hepC	Factor	true, false
Diabetic indicator	diab	Factor	true, false
Kidney transplant indicator	transp	Factor	true, false

is the GAMLSS based on the BCPE distribution (presenting AIC and BIC equal to 1,913.69 and 1,961.33, respectively). Furthermore, Table 3 shows which covariates are present (or not) in each final fitted model for each distribution considered in this paper.

Table 2. Statistics from the best fitted models for each used distribution

Distribution	Deviance	AIC	BIC
BCPE	1,883.69	1,913.69	1,961.33
Weibull	1,909.99	1,931.99	1,966.93
BCCG	1,904.60	1,932.60	1,977.07
BCT	1,911.96	1,939.96	1,984.42

Table 3. Covariate presence (or absence) in each of the distributions' parameters

Model	Parameter	age	sex	mar	col	hepB	hepC	diab	transp
BCPE	μ	×				×			×
	σ		×			×	×		×
	ν						×		
	τ		×					×	×
Weibull	μ	×	×			×	×		×
	σ	×				×	×		×
BCCG	μ	×	×			×			×
	σ	×	×			×			×
	ν					×	×	×	
BCT	μ	×	×			×	×		×
	σ	×					×		×
	ν	×							
	τ	×							×

The final fitted (parametric) GAMLSS model based on the BCPE distribution is given by

$$\begin{aligned}
 \hat{\mu} &= \exp \{7.060 - 0.041 \text{ age} + 0.797 [\text{if hepB=true}] + 2.138 [\text{if transp=true}]\} \\
 \hat{\sigma} &= \exp \{1.828 + 0.367 [\text{if sex=ma1e}] - 0.088 [\text{if hepB=true}] + 2.400 [\text{if hepC=true}] \\
 &\quad + 21.222 [\text{if transp=true}]\} \\
 \hat{\nu} &= 0.724 + 1.064 [\text{if hepC=true}] \\
 \hat{\tau} &= \exp \{0.232 + 0.108 [\text{if sex=ma1e}] + 0.136 [\text{if diab=true}] - 3.980 [\text{if transp=true}]\}.
 \end{aligned} \tag{2}$$

All parameters are significant at the 5% significance level, apart from the ones related to the variables sex and diabetic indicator in the kurtosis parameter τ . However, Lee *et al.* (2016) highlight the danger of using naive p-values after the model selection stage and thus we will keep both variables in the final model.

Based on equation (2), we can note that only three variables (from the original eight presented in Table 1) were considered to model the median μ . For each additional year at the beginning of the treatment, the median of the patient's lifetime expectancy decreases 0.041 days. Under the same circumstances, patients who have antibodies to hepatitis B have more than double the median of lifetime expectancy (121,89%) than the ones who have not. Further, if a patient has undergone kidney transplantation, its lifetime expectancy increases more than seven times. In the one hand, the presence of such variables in the regression model for μ is expected as can be seen in a wide literature, e.g., Tangri *et al.* (2016) and Fabrizi *et al.* (2017). On the other hand, the absence of the other features may not be expected, which may be being caused by the different fitted regression structures in the GAMLSS framework.

Patients who have antibodies to hepatitis B present a smaller variability, measured here as the coefficient of variation σ . Moreover, the lifetime expectancy variability is 1.44 times greater in male patients and 11.02 times greater in patients who have antibodies to hepatitis C. This variability is way large in patients who have undergone kidney transplant, as can be seen by the coefficient associated to this feature (21.222). Such behaviour might be explained by the great number of factors associated with post-kidney transplant success rate (Sprangers *et al.*, 2018).

The only variable considered in the Strategy A selection method that affects the skewness parameter ν is the antibodies to hepatitis C. When a patient does not present such indicator, the distribution of the failure times is slightly positively skewed – since $\hat{\nu} = 0.724 < 1$ –, while when this indicator is observed, such distribution is negatively skewed – since $\hat{\nu} = 1.788 > 1$ (Rigby & Stasinopoulos, 2004; Rigby *et al.*, 2019).

The fitted model for τ indicates that the distribution of the failure times will always be leptokurtic – since $\hat{\tau} < 2$ (Rigby & Stasinopoulos, 2004; Rigby *et al.*, 2019) – regardless the level of any of the explanatory variables. Nevertheless, such distribution present heavier tails when the patient has undergone kidney transplantation.

Finally, we check the model adequacy using the normalised quantile residuals (Dunn & Smyth, 1996) through the worm plot (Buuren & Fredriks, 2001). Basically, if a linear, quadratic or cubic shape is observed, then the fitted model is not able to proper fit the variance, skewness or kurtosis of the data, respectively (Stasinopoulos *et al.*, 2017). Since there is no particular shape in Figure 1, we can say that the final fitted GAMLSS model based on the BCPE distribution provided a reasonable fit to the data.

5. Conclusions

In this paper, we have presented the Box-Cox family of distributions as an interesting alternative to fit censored data under the generalised additive models for location, scale and shape (GAMLSS) framework. Among these distributions, the Box-Cox power exponential (BCPE) returned the best

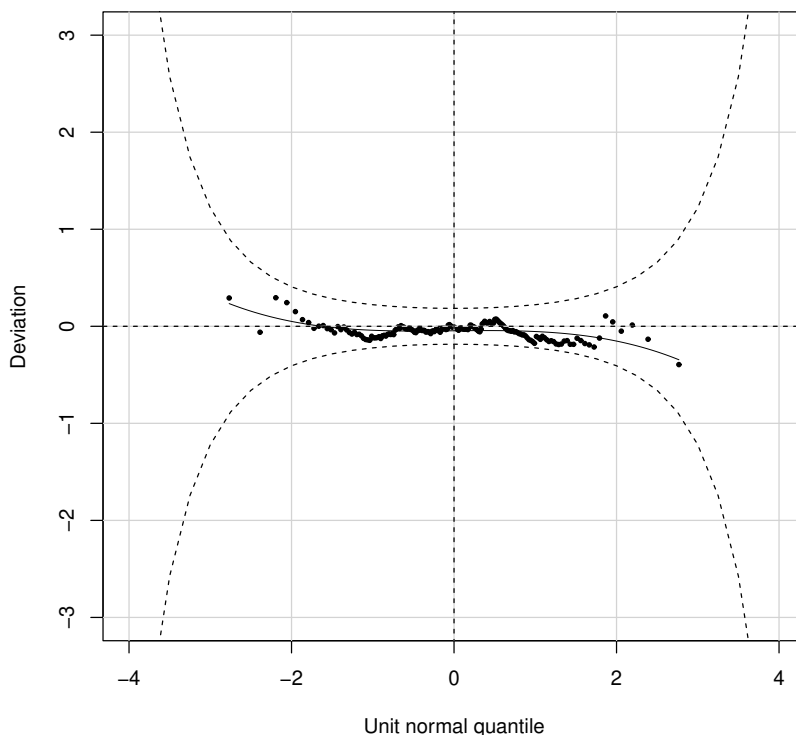


Figure 1. Worm plot of the final fitted model based on the BCPE distribution.

(smallest) AIC and BIC values when compared to the Box-Cox t (BCT) and Box-Cox Cole and Green (BCGG) distributions. Moreover, the BCPE distribution also outperformed the Weibull distribution, previously fitted to these data. All models were selected through the so-called Strategy A, a stepwise-based method which is used to select the best set of covariates in each of the parameters of the response variable distributions. Although eight explanatory variables were initially available to explain the response (failure times – death), after Strategy A was performed, only six variables were selected (not necessarily in all parameters) in the final fitted GAMLSS model based on the BCPE distribution, namely: age at the beginning of treatment, antibodies to hepatitis B, antibodies to hepatitis C, kidney transplant indicator, sex and diabetic indicator. Overall, such procedure considering this family of distributions can be applied in similar data, i.e., when we are interested in modelling time until a certain event of interest.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).

2. Alizadeh, M, Ramires, T. G., MirMostafaei, S.-K., Samizadeh, M & Ortega, E. M. M. A new useful four-parameter extension of the Gumbel distribution: Properties, regression model and applications using the GAMLSS framework. *Communications in Statistics – Simulation and Computation* **48**, 1746–1767 (2019).
3. Buuren, S & Fredriks, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine* **20**, 1259–1277 (2001).
4. Castro, M, Cancho, V. G. & Rodrigues, J. A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer Methods and Programs in Biomedicine* **97**, 168–177 (2010).
5. Cole, T. J. & Green, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine* **11**, 1305–1319 (1992).
6. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).
7. Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236–244 (1996).
8. Emmert-Streib, F. & Dehmer, M. Introduction to Survival Analysis in Practice. *Machine Learning and Knowledge Extraction* **1**, 1013–1038 (2019).
9. Fabrizi, F., Donato, F. M. & Messa, P. Association Between Hepatitis B Virus and Chronic Kidney Disease: a Systematic Review and Meta-analysis. *Annals of Hepatology* **16**, 21–47 (2017).
10. Gijbels, I. Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 178–188 (2010).
11. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Chapman and Hall/CRC, 1990).
12. Heller, G. Z., Robledo, K. P. & Marschner, I. C. Distributional regression in clinical trials: treatment effects on parameters other than the mean. *BMC Medical Research Methodology* **22**, 56 (2022).
13. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457–481 (1958).
14. Kneib, T. Beyond mean regression. *Statistical Modelling* **13**, 275–303 (2013).
15. Lee, J. D., Sun, D. L., Sun, Y & Taylor, J. E. Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 907–927 (2016).
16. Nakamura, L. R., Cerqueira, P. H. R., Ramires, T. G., Pescim, R. R., Rigby, R. A. & Stasinopoulos, D. M. A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. *Journal of Applied Statistics* **46**, 416–431 (2019).
17. Nakamura, L. R., Rigby, R. A., Stasinopoulos, D. M., Leandro, R. A., Villegas, C & Pescim, R. R. Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized t distribution. *Journal of Data Science* **15**, 221–238 (2017).
18. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)* **135**, 370–384 (1972).
19. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022). <https://www.R-project.org/>.
20. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Carvalho, R. J., Vieira, L. A. & Pereira, C. A. B. Comparison between highly complex location models and GAMLSS. *Entropy* **23**, 469 (2021a).

21. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Ortega, E. M. M. & Cordeiro, G. M. Predicting survival function and identifying associated factors in patients with renal insufficiency in the metropolitan area of Maringá, Paraná State, Brazil. *Cadernos de Saúde Pública* **34**, e00075517 (2018).
22. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J & Cordeiro, G. M. A new semiparametric Weibull cure rate model: fitting different behaviors within GAMLSS. *Journal of Applied Statistics* **46**, 2744–2760 (2019).
23. Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J, Stasinopoulos, D. M. & Rigby, R. A. Validation of stepwise-based procedure in GAMLSS. *Journal of Data Science* **19**, 96–110 (2021b).
24. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 507–554 (2005).
25. Rigby, R. A. & Stasinopoulos, D. M. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine* **23**, 3053–3076 (2004).
26. Rigby, R. A. & Stasinopoulos, D. M. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling* **6**, 209–229 (2006).
27. Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z. & De Bastiani, F. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R* (CRC Press, 2019).
28. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461–464 (1978).
29. Sprangers, B, Nair, V, Launay-Vacher, V, Riella, L. V. & Jhaveri, K. D. Risk factors associated with post-kidney transplant malignancies: an article from the Cancer-Kidney International Network. *Clinical Kidney Journal* **11**, 315–329 (2018).
30. Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location, scale and shape (GAMLSS) in R. *Journal of Statistical Software* **23**, 1–46 (2007).
31. Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V & De Bastiani, F. *Flexible Regression and Smoothing: Using GAMLSS in R* (CRC Press, 2017).
32. Tangri, N. *et al.* Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis. *The Journal of the American Medical Association* **315**, 164–174 (2016).