Box-Cox symmetric distributions and applications to nutritional data

Silvia L.P. Ferrari*
Department of Statistics, University of São Paulo, São Paulo, Brazil - silviaferrari.usp@gmail.com

Giovana Fumes
Department of Statistics, University of São Paulo, São Paulo, Brazil - gifumesbtu@gmail.com

Abstract

We introduce and study the Box-Cox symmetric class of distributions, which is useful for modeling positively skewed, possibly heavy-tailed, data. The new class of distributions includes the Box-Cox t, Box-Cox Cole-Green, Box-Cox power exponential distributions, and the class of the log-symmetric distributions as special cases. It provides easy parameter interpretation, which makes it convenient for regression modeling purposes. Additionally, it provides enough flexibility to handle outliers. The usefulness of the Box-Cox symmetric models is illustrated in a series of applications to nutritional data.

Keywords: Box-Cox transformation; Box-Cox power exponential distribution; Box-Cox t distribution; nutrients intake.

1. Box-Cox symmetric distributions

Positive continuous data usually present positive skewness and outlier observations. This is the typical situation with survival times, nutrients intake and family income data, among many other examples. Since the Box and Cox (1964) seminal paper, the Box-Cox power transformation has been routinely employed for transforming to normality. Let Y be a positive random variable. The Box-Cox transformation is defined as $V(Y) = (Y^{\nu} - 1)/\nu$, if $\nu \neq 0$, and $V(Y) = \log Y$, if $\nu = 0$. Despite its popularity and ease of implementation, this approach, however, has drawbacks, one of them being the fact that the model parameters cannot be easily interpreted in terms of the original response. A conceptual shortcoming is that the support of the transformed variable is not the whole real line and hence a (non-truncated) normal distribution should not be assumed for the transformed data.

An alternative approach to the Box-Cox transformation that allows the parameters to be interpretable as characteristics of the original data is the Box-Cox Cole-Green distribution (Cole and Green, 1992; Stasinopoulos et al., 2008). It uses the Box-Cox approach, but the parameters are incorporated into the transformation. The Box-Cox Cole-Green distribution has support in $\mathbb{R}^{+}$ and is defined from the transformation

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

where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$, assuming that Z has a standard normal distribution truncated at the interval

$$A(\sigma,\nu) = \begin{cases} \left(-\frac{1}{\sigma\nu}, \infty\right), & \text{if} \quad \nu > 0, \\ \left(-\infty, -\frac{1}{\sigma\nu}\right), & \text{if} \quad \nu < 0, \\ (-\infty, \infty), & \text{if} \quad \nu = 0. \end{cases}$$

The Box-Cox symmetric (BCS) class of distributions replaces the normal distribution by the class of the continuous standard symmetric distributions. A continuous random variable W is said to have a symmetric distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$ and density generating function r if its probability density function (pdf) is given by $v(w,\mu,\sigma;r) = \sigma^{-1}r(\sigma^{-2}(w-\mu)^2)$, $w \in \mathbb{R}$, where $r(\cdot)$ satisfies $r(u) > 0$, for $u > 0$ and $\int_0^{\infty} u^{-1/2} r(u)\mathrm{d}u = 1$. Let Y be a positive continuous random variable. We say that Y has a Box-Cox symmetric distribution with parameters $\mu > 0$, $\sigma > 0$ and $\nu \in \mathbb{R}$ and density generating function r, and write $Y \sim BCS(\mu,\sigma,\nu;r)$, if Z given in (1) has a truncated standard symmetric distribution with support $A(\sigma,\nu)$.

The BCS class of distributions reduces to the log-symmetric class of distributions (Vanegas and Paula, 2014a, 2014b) when ν is fixed at zero. Additionally, it leads to the Box-Cox Cole-Green (Stasinopoulos et al., 2008), Box-Cox t (Rigby and Stasinopoulos, 2006) and Box-Cox power exponential (Rigby and Stasinopoulos, 2004; Voudouris et al., 2012) distributions by taking Z as a truncated standard normal, Student-t and power exponential random variable, respectively. Additionally, it allows the definition of new distributions, such as the Box-Cox slash distribution, which is defined and explored in this paper.

The BCS class of distributions has a number of interesting properties. In particular, it allows easy parameter interpretation from its quantiles. All the quantiles are proportional to μ and, if $\nu = 0$, μ is the median of Y. Also, some distributions in the BCS class produce robust maximum likelihood estimation against outliers. Additionally, the BCS class includes distributions with different right tail heaviness. Tail heaviness is studied from two different approaches, namely the regular variation theory (de Haan, 1970), and the Rigby et al. (2014) criterion.

2. Applications and comparison of different approaches

We present applications of the BCS distributions in the analysis of micro and macronutrients intake. The data refer to observations of nutrients intake based on the first 24-hour dietary recall interview for $n = 368$ individuals. For each nutrient, we assume that the data $Y_1, \ldots, Y_n$ are independent. We fitted the following different models to the data: Box-Cox t (BCT) with fixed degrees of freedom parameter ($\tau = 4$) and with unknown, estimated from the data, degrees of freedom; Box-Cox Cole-Green (BCCG), which corresponds to the BCT model with $\tau \to \infty$; skew-normal (SN) and skew-t (ST) (Azzalini, 2005); and transformed symmetric models with normal (TN) and t (TT) errors (Cordeiro and Andrade, 2011). The TN (TT) model assumes that the Box-Cox transformed data follow a normal (Student-t) distribution. For the SN and the TT models, we only considered the case in which the number of degrees of freedom is fixed, and again we set $\tau = 4$, because estimating τ along with the other parameters often caused numerical problems. In all the cases, the unknown parameters were estimated by the maximum likelihood method. For the BCT, BCCG, SN and ST distributions, we used the `gamlss` package implemented in `R`, while for the TN and TT models we used both the function `optim` in `R` and the `PROC NLP` in `SAS`. Goodness-of-fit was evaluated using the following criteria: Akaike information criterion (AIC) and Anderson-Darling statistics (AD, ADR, and AD2R); see Luceño (2005, Tables 1, 2 and B.1). AD is a global measure of lack-of-fit, while both ADR and AD2R are more sensitive to the lack of fit in the right tail of the distribution; AD2R puts more weight in the right tail than ADR.

The complete version of this paper contains tables with the goodness-of-fit statistics for all the fitted models to 22 and 11 micro and macronutrients intakes data. The tables convey important information. First, the datasets cover a wide range of light-tailed to heavy-tailed data. This can be seen by the estimated values of the degrees of freedom parameter (τ) under the Box-Cox t model ($\widehat{\tau}$ ranges from 2.2 to 187.4). Second, no convergence problem was observed while fitting the Box-Cox t model (with fixed or estimated degrees of freedom), the transformed t model with fixed τ and the transformed normal model. The maximum likelihood estimation under the skew-normal model did not achieve convergence in 14 cases, followed by the Box-Cox Cole-Green model (10 cases) and the skew-t model (9 cases). Third, the Box-Cox t model with estimated τ performed better than with fixed τ in almost all the cases. Fourth, for the micronutrients data and according to the AIC criterion, the Box-Cox t model with estimated τ achieved the best fit in 12 cases followed by the Box-Cox t model with fixed τ (5 cases). Similar pattern is observed for the macronutrients data. Fifth, according to all the Anderson-Darling criteria, the Box-Cox Cole-Green, skew-normal and transformed normal models were not the best fitting model in any of the cases. In contrast, the Box-Cox t model with estimated τ was the best fitting model in most of the cases. Overall, we conclude that the Box-Cox t model with estimated τ performed better than the other models.

A detailed analysis of the data on the intake of animal protein and energy using the Box-Cox t, Box-Cox Cole-Green and Box-Cox power exponential distributions is presented in the complete version of this paper.

3. Conclusions

This paper proposed a new class of distributions, the Box-Cox symmetric distributions. It contains some well known distributions as special cases and allows the definition of new distributions, such as the Box-Cox slash distribution. It is particularly suitable for inference on positively skewed, possibly heavy-tailed, data.

It permits easy parameter interpretation, a desirable feature for modeling.
There is clear possibility for extension to regression models. Some or all the four parameters of the BCS distributions may be modeled by a link function and a linear or nonlinear regression model structure. The GAMLSS framework (Rigby & Stasinopoulos, 2005) is a natural tool for implementing BCS regression models. It allows the regression structure to include parametric and nonparametric terms and random effects. Box-Cox t, Box-Cox Cole-Green, and Box-Cox power exponential models are already implemented in `gamlss` package in `R`.
Some BCS distributions include an extra parameter; e.g., the degrees of freedom parameter of the BCT distribution. We have not faced convergence problems or unrealistic estimation when the additional parameter were estimated simultaneously with the others. It should be noticed that the sample sizes in our applications were large ($n = 368$). In small samples, it may be advisable to set a grid of values for the extra parameter and choose the value that provides the best fit according to the chosen criteria.
Applications to data on intake of several nutrients illustrated that the BCS distributions are useful in practice. The data correspond to the first 24-hour dietary recall interview for the individuals in the sample. It is part of our current research to develop Box-Cox symmetric models with random and mixed effects to model nutrients intake data taken from repeated 24-hour recalls.

Acknowledgments. We thank José Eduardo Corrente for providing the data used in this study, and Eliane C. Pinheiro for helpful discussions. We acknowledge the financial support of FAPESP and CNPq (Brazil).

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159 – 188.

Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*,**26** 211 – 252.

Cole, T.J. & Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305 – 1319.

Cordeiro, G.M. & Andrade, M.G. (2011). Transformed symmetric models. *Statistical Modelling*, **11**, 371 – 388.

de Haan, L. (1970). On Regular Variation and Its Application to the Weak Convergence of Sample Extremes. Amsterdam: Mathematics Centre.

Luceño, A. (2005). Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis*, **51**, 904 – 917.

Rigby, R.A. & Stasinopoulos, D.M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine*,**23**, 3053 - 3076.

Rigby, R.A. & 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.

Rigby, R.A. and Stasinopoulos, D.M. (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**, 209 – 229.

Rigby, R. A., Stasinopoulos, D.M., Heller, G., & Voudouris, V. (2014). *The Distribution Toolbox of GAMLSS*. `http://www.gamlss.org`.

Stasinopoulos, D.M., Rigby, R.A., & Akantziliotou, C. (2008). *Instructions on how to use the GAMLSS package in R.* `http://www.gamlss.org`

Vanegas, L.H. & Paula, G.A. (2014a). A semiparametric approach for joint modeling of median and skewness. *Test.* DOI:10.1007/s11749-014-0401-7.

Vanegas, L.H. & Paula, G.A. (2014b). Log-symmetric distributions: statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics.*
`http://www.imstat.org/bjps/papers/BJPS272.pdf`.

Voudouris, V., Gilchrist, R., Rigby, R.A., Sedgwick, J., & Stasinopoulos, D.M. (2012). *Modelling skewness and kurtosis with BCPE density in GAMLSS. Journal of Applied Statistics*, **39**, 1279 − 1293.