UNIVERSIDADE FEDERAL DE GOIÁS

FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS
CONTÁBEIS E ECONOMIA

Marcelo dos Santos Ventura

Monte Carlo simulation studies in log-symmetric regression models

Goiânia

2018

sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Marcelo dos Santos Ventura

Título do trabalho: **Monte Carlo simulation studies in log-symmetric regression models**

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Marcelo dos Santos Ventura
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Heitor Saulo Dias dos Santos
Assinatura do(a) orientador(a)[2]

Data: 26 / 03 / 18

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

[2] A assinatura deve ser escaneada.

Marcelo dos Santos Ventura

Monte Carlo simulation studies in log-symmetric regression models

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Economia da Faculdade de Administração, Ciências Contábeis e Economia da Universidade Federal de Goiás, como requisito parcial para obtenção de título de Mestre em Economia.

Universidade Federal de Goiás – UFG

Faculdade de Administração, Ciências Contábeis e Economia

Orientador: Prof. Dr. Helton Saulo Bezerra dos Santos

Coorientador: Prof. Dr. Victor Leiva

Goiânia

2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Ventura, Marcelo dos Santos
Monte Carlo simulation studies in log-symmetric regressions [manuscrito] / Marcelo dos Santos Ventura. - 2018.
xlii, 42 f.

Orientador: Prof. Dr. Helton Saulo Bezerra dos Santos; co orientador Dr. Victor Eliseo Leiva Sanchez.
Dissertação (Mestrado) - Universidade Federal de Goiás, Faculdade de Administração, Ciências Contábeis e Ciências Econômicas (FACE), Programa de Pós-Graduação em Economia, Goiânia, 2018.
Bibliografia.
Inclui siglas, lista de figuras, lista de tabelas.

1. Distribuições log-simétricas. 2. Critérios de seleção de modelos. 3. Simulações de Monte Carlo. 4. Dados de cinema. 5. Modelos de regressão. I. Saulo Bezerra dos Santos, Helton, orient. II. Título.

CDU 519.22

ATA DE DEFESA PÚBLICA DE DISSERTAÇÃO

No dia 09 de março de 2018, no horário de 10:00 horas às 11:30 horas, foi realizada, em sessão pública na sala 2110 da FACE, a defesa da dissertação *Monte Carlo simulation studies in log-symmetric regression models,* de autoria do Mestrando Marcelo dos Santos Ventura, do Programa de Pós-Graduação em Economia – PPE da Universidade Federal de Goiás. A Comissão Examinadora, constituída pelos professores Victor Eliseo Leiva Sanchez/PPE/UFG (Coorientador/membro interno), Tatiane F.N. Melo da Silva/PPE/UFG (Examinadora/membro interno) e Cynthia Arantes Vieira Tojeiro/UFG/IME (Examinadora/membro externo), emitiu o seguinte parecer/recomendações:

Resultado Final:

[X] Aprovado

[ ] Reprovado

Recomendações:

Aprovado condicionado as correçoes sugeridas pela banca.

Eu, Victor Eliseo Leiva-Sanchez, orientador do discente, lavrei a presente Ata, que segue assinada por mim e pelos demais membros da Comissão Examinadora.

Prof. Dr. Victor Eliseo Leiva-Sanchez (Coorientador/Presidente/PPE/UFG)

Tatiane F do N. Melo da Silva

Profa. Dr. Tatiane F.N. Melo da Silva (Examinadora Interna/PPE/UFG)

Cynthia Arantes Vieira Tojeiro

Profa. Dr. Cynthia Arantes Vieira Tojeiro (Examinadora Externo/UFG/IME)

Este trabalho é dedicado aos meus pais,
e todos que me apoiaram durante o meu mestrado.

Agradecimentos

Durante o mestrado tive o apoio de muitas pessoas, sem as quais não teria chegado ao fim. Em primeiro lugar, agradeço a minha família pelo suporte no período do mestrado. Também agradeço ao Prof. Paulo Scalco, que enquanto coordenador me orientou durante a minha mudança de forma a tornar as coisas mais fáceis na vinda de Brasília. Além disso, obrigado Prof. Sandro por todas as conversas, conselhos, e por ter representado o Rio de Janeiro durante minha jornada. Saudações rubro negras.

Em especial, não tenho palavras para expressar minha gratidão ao Prof. Helton Saulo e Prof. Victor Leiva, que foram sem dúvida minhas maiores inspirações durante o mestrado. Trabalho duro, atenção com o próximo, gentileza, acessibilidade por email, e a amizade de vocês são coisas que levo comigo para minha vida pessoal e profissional. Obrigado por serem excelentes exemplos para mim, por todo o tempo dispensado a realização desse trabalho, e pela confiança.

Aos meus pais e minha irmã, obrigado pela confiança e suporte emocional em todo o processo. Viver dois anos em goiânia foi uma experiência inesquecível para mim, e sem a torcida de voces não teria conseguido.

À minha futura mestra em Educação, Alana, obrigado por ter vivido tudo isso comigo! Parabéns pela coragem em se mudar de estado, por ter acreditado em mim em todos os momentos difíceis, de ausência, adaptação, solidão, cansaço, e pouco lazer. Daqui a pouco será sua defesa e estarei torcendo por você. Te amo muito.

À secretaria do PPE, em especial ao Gilson e Jaqueline. Nunca imaginei encontrar pessoas tão amáveis e solícitas. Sentirei saudade das nossas conversas de manhã cedo Gilson.

Aos amigos do mestrado que represento por Lucas, Paula, e Marcos, obrigado por me ajudarem em momentos difíceis, e tornarem simples comemorações inesquecíveis.

À FAPEG, pelo apoio financeiro e fomento da ciência regional, muito obrigado.

Abstract

This work deals with two Monte Carlo simulation studies in log-symmetric regression models, which are particulary useful for the cases when the response variable is continuous, strictly positive and asymmetric, with the possibility of the existence of atypical observations. In log-symmetric regression models, the distribution of the random errors multiplicative belongs to the log-symmetric class, which encompasses log-normal, log-Student-t, log-power-exponential, log-slash, log-hyperbolic distributions, among others. The first simulation study has as objective to examine the performance for the maximum-likelihood estimators of the model parameters, where various scenarios are considered. The objective of the second simulation study is to investigate the accuracy of popular information criteria as AIC, BIC, HQIC and their respective corrected versions. As illustration, a movie data set obtained and assembled for this dissertation is analyzed to compare log-symmetric models with the normal linear model and to obtain the best model by using the mentioned information criteria.

Keywords: Log-symmetric distributions, Model selection criteria, Monte Carlo simulation, Movie data, Regression models, R software.

Resumo

Este trabalho aborda dois estudos de simulação de Monte Carlo em modelos de regressão log-simétricos, os quais são particularmente úteis para os casos em que a variável resposta é contínua, estritamente positiva e assimétrica, com possibilidade da existência de observações atípicas. Nos modelos de regressão log-simétricos, a distribuição dos erros aleatórios multiplicativos pertence à classe log-simétrica, a qual engloba as distribuições log-normal, log-Student-t, log-exponencial-potência, log-slash, log-hyperbólica, entre outras. O primeiro estudo de simulação tem como objetivo examinar o desempenho dos estimadores de máxima verossimilhança desses modelos, onde vários cenários são considerados. No segundo estudo de simulação o objetivo é investigar a eficácia critérios de informação populares como AIC, BIC, HQIC e suas respectivas versões corrigidas. Como ilustração, um conjunto de dados de filmes obtido e montado para essa dissertação é analisado para comparar os modelos de regressão log-simétricos com o modelo linear normal e para obter o melhor modelo utilizando os critérios de informação mencionados.
Palavras-chave: Critério de seleção de modelos, Dados de cinema, Distribuições log-simétricas, Modelos de regressão, Simulação de Monte Carlo, Software R.

List of Figures

List of Tables

List of abbreviations and acronyms

IMDb	Internet Movie Database
AIC	Akaike information criterion
BIC	Bayesian information criterion
HQIC	Hannan and Quinn information criterion
ML	Maximum likelihood
PDF	Probability density function

Contents

1 Introduction

In this chapter, we provide the bibliographical review, describe the objectives and the composition of chapters, as well as details of the computational support.

1.1 Bibliographical review

The normal linear regression model is widely used in data analysis. Although the normal distribution has been the base for regression models, it is not appropriate for modeling asymmetric and strictly positive data; see Limpert, Stahel, and Abbt (2001) and Puig (2008). In fact, the lack of a structure to accommodate asymmetry, for example, can lead to estimates significantly different than the true values of the model parameters; see Desousa et al. (2018).

The class of log-symmetric distributions is a flexible alternative to deal with asymmetric and strictly positive data when possible atypical observations are present. This class has as members some well-known asymmetric, positive, and continuous distributions; e.g. the log-normal (see CROW; SHIMIZU, 1987), log-Student-*t*, log-power-exponential (VANEGAS; PAULA, 2016b), log-logistic (BENNETT, 1983), log-hyperbolic, log-slash (VANEGAS; PAULA, 2016b), harmonic-law (PUIG, 2008), and log-contaminated-normal (TUKEY, 1960) models. The Vanegas and Paula (2016b) paper is referred to the interested reader for further information about log-symmetric distributions and their properties.

The log-symmetric regression model provides a relevant solution and alternative for modeling data where the assumption of normality is not valid; see Vanegas and Paula (2017). In addition, this regression model allows either the median and skewness of the response variable to be explicitly modeled. Both measures are extremely important for asymmetric variables. Therefore, more practical applications of this regression model can be found; see Vanegas and Paula (2016a), for a proposal of extended log-symmetric regression models; Vanegas and Paula (2017), for an extension which allows the presence of non-informative left- or right-censored observations in log-symmetric regression models; Saulo and Leão (2017), for an application to duration models and high frequency financial data; and Balakrishnan et al. (2017), for a development of three simple closed form estimators for a class of log-symmetric distributions.

In regression analysis, model selection criteria play an important role in providing the best adjustment for an estimated model under either of following situations: (i) to select the number of parameters in a regression model (see AKAIKE, 1992) and to choose which distribution better describes the data (see VANEGAS; PAULA, 2016a). Thus, information criteria are a relevant topic of study in statistics. For instance, Akaike (1992) developed the Akaike information criterion (AIC), which is a way to minimize Kullback-Leibler (KL) information (KULLBACK, 1997), the distance between two densities. In addition, some variants of AIC as the Bayesian information criterion (BIC) (SCHWARZ, 1978), and Hannan and Quinn information criterion (HQIC) (HANNAN; QUINN, 1979), provide solutions to estimate the expected predictive log-likelihood by correcting its biased maximization.

1.2 Objectives and description of chapters

Considering the gap of asymmetric modeling in statistics as a motivation for using log-symmetric regression models, this work focuses on three objectives: (i) to examine the performance for the maximum likelihood (ML) estimators of the model parameters; (ii) to investigate the accuracy of AIC, BIC, HQIC and their respective corrected versions; and (iii) to introduce an empirical application with movie data. In the first objective, we use Monte Carlo simulations to examine asymptotic properties of the ML estimators of the parameters of different log-symmetric distributions. In the second objective, we explore information criteria on model order dimensioning and log-symmetric distribution selection. Specifically, this second objective is two folds because, first, allows us to evaluate whether the correct quantity of parameters is detected by the information criterion or not, inducting an over or sub parameterization; and second, it permits us to evaluate if the true distribution, which generates the data, is correctly identified with a smallest value of such criteria in relation to other log-symmetric models or not. In the third objective, we apply a log-symmetric regression model to web-scrapped movie data from IMDb.com. Then, we take the movie box office as the response variable, which is right-skewed. Furthermore, we propose "rating" (IMDb's users ratings), "number of votes", and "budget" as explanatory variables (covariates). Finally, we present the obtained results by a computational routine in `R` code, with the package `ssym`, created by Vanegas and Paula (2016c) for log-symmetric regression modeling.

The remainder of this work proceeds as follows. In Chapter 2, we describe the class of log-symmetric distributions and its corresponding regression models. In Chapter 3, we propose two Monte Carlo simulation studies on ML estimators and information criteria performances for log-symmetric regressions. In Chapter 4, we provide an application to movie data. Finally, in Chapter 5, we discuss the conclusions of this study and suggestions for future works.

1.3 Computational support

According to the Institute of Electrical and Electronics Engineers (IEEE), the `R` language (see R CORE TEAM, 2016, `https://cran.r-project.org`) was in 2017 the sixth most popular programming language in the world. It proves its reliability as a computational language for statistics and its applications around the world, which covers fields like economics, law and medicine, among others. R-Studio (see RSTUDIO TEAM, 2016, `https://www.rstudio.com`) is an IDE for that language and was used in this work. Auxiliary packages as `rvest` (see WICKHAM, 2016) and `ssym` (see VANEGAS; PAULA, 2016c) were also used. The first one is a web-scrapping tool, and the second one is developed by authors to make possible log-symmetric fitting in `R`.

2 Log-symmetric regression models

In this chapter, we review some symmetric and log-symmetric distribution statistical properties. Furthermore, we cover main premises of log-symmetric regression, and highlight median and skewness modeling. In addition, ML estimation method is described. Furthermore, we review here the model selection criteria used in this work.

2.1 Log-symmetric distributions

The symmetric class of distributions is popular in different fields of study such as economics, physics, and statistics; see Almeida et al. (2017), Szu and Hartley (1984), and Fang, Kotz, and Ng (1990). However, its structure does not allow strictly positive and skewed random variables to be represented. Hence, log-versions of each associated distribution have been developed; see Bennett (1983), Crow and Shimizu (1987) and Puig (2008).

Let Y be a continuous random variable following a symmetric distribution with location parameter $-\infty < \mu < \infty$, dispersion parameter $\phi > 0$ and density generator g. We use the notation $Y \sim \mathrm{S}(\mu, \phi^2, g)$. This distribution can be defined in terms of its probability density function (PDF) given by

$$f_Y(y; \mu, \phi, g) = \frac{1}{\phi} g\left(\frac{(y-\mu)^2}{\phi^2}\right), \quad y \in \mathbb{R}, \tag{2.1}$$

with $g(u) > 0$ for $u > 0$ and $\int_0^\infty u^{-1/2} g(u)\mathrm{d}u = 1$; see Fang, Kotz, and Ng (1990). The class of log-symmetric distributions is obtained when we set $T = \exp(Y)$, denoted by $T \sim \mathrm{LS}(\eta, \phi^2, g)$, where $\eta = \exp(\mu) > 0$ is a scale parameter, $\phi > 0$ is a shape (skewness or relative dispersion) parameter, for some function (density generator) g: $\mathbb{R} \to [0, \infty)$, with $\int_0^\infty u^{-1/2} g(u)\mathrm{d}u = 1$. The PDF of T is given by

$$f_T(t; \eta, \phi, g) = \frac{1}{\phi t} g(\widetilde{t}^2), \quad t > 0,$$

where $\widetilde{t} = \log((t/\eta)^{1/\phi})$. Note that the density generator g leads to different distributions and is associated with an extra parameter ζ (or an extra parameter vector $\boldsymbol{\zeta}$).

Table 1 – Density generator functions

Distribution	$g(u)$	ζ	v(t)
Log-normal	$\exp(-\frac{1}{2}u)$	-	1
Log-Student-t	$(1 + u\frac{1}{\zeta}) - (\frac{(\zeta+1)}{2})$	$\zeta > 0$	$\frac{\zeta+1}{\zeta+t^2}$
Log-power-exponential	$\exp(-\frac{1}{2}u^{\frac{1}{(1+\zeta)}})$	$-1 < \zeta \leq 1$	$\frac{\lvert t\rvert^{-\frac{2\zeta}{\zeta+1}}}{1+\zeta}$
Log-hyperbolic	$\exp(-\zeta\sqrt{1+u})$	$\zeta > 0$	$\frac{\zeta}{\sqrt{1+t\hat{}2}}$
Log-slash	$\mathrm{IGF}^1(\zeta + \frac{1}{2})$	$\zeta > 0$	κ
Log-normal-contam.	$\sqrt{\zeta_2}\exp(-\zeta\frac{u}{2}) + (1-\zeta_1)\frac{\exp(-\frac{u}{2})}{\zeta_1}$	$0 < \zeta_1 < 1, 0 < \zeta_2 < 1$	τ

Note that in Table 1 $\mathrm{IGF}(a, x) = \int_0^1 \exp(-tx)t^{a-1}\mathrm{d}t$ for $a > 0$ and $x \geq 0$ is an incomplete gamma function, $\kappa = \mathrm{IGF}(1/2(2\zeta+3, t^2/2)/\mathrm{IGF}(1/2(2\zeta+1, t^2/2)$ and $\tau = \frac{\zeta_2^{3/2}\zeta_1\exp(0.5t^2(1-\zeta_2))+(1-\zeta_1)}{\zeta_2^{1/2}\zeta_1\exp(0.5t^2(1-\zeta_2))+(1-\zeta_1)}$. The $v(t)$ denotes a weight function given $g(\cdot)$. Also, it is important to mention that the extra parameter ζ generally is unknown, and each distribution has a density generator for ζ, which is chosen to minimize an overall measure of adjustment Υ, defined by

$$\Upsilon_\zeta = \frac{\sum_{k=1}^{n}\left|\Phi^{-1}(F_Z(\widehat{\widetilde{z}})) - v^{(k)}\right|}{n}, \tag{2.2}$$

where $\Phi^{-1}(F_Z(\widehat{\widetilde{z}}^{(1)}), \ldots, \Phi^{-1}(F_Z(\widehat{\widetilde{z}}^{(n)}))$ is an ordered random sample and $\widehat{\boldsymbol{\theta}}$ is consistent, with Φ being the cumulative distribution function of a normal distribution, F_Z is the cumulative distribution function of Z, $\widehat{\widetilde{z}}^{(k)}$ is the k-th order statistic of $\widehat{\widetilde{z}}$, and $v^{(k)}$ is its expectation in a random sample with size n. Better goodness-of-fit measures are obtained with minimum possible values of Υ. For a determined ζ and a log-symmetric fitted regression, in the `extra.parameter()` function from the `ssym` package, a graph is generated containing statistics from the log-likelihood function and Υ.

As an example, Figure 1 was generated to choose an extra-parameter of the log-Student-*t* distribution. Note that the interval of values for ζ was $2.5 \leq \zeta \leq 4.0$. First graph suggests that the ideal value of ζ for minimizing Υ and information criteria is three. Hence, this value generates better statistics for model selection criteria. However, this iterative process is necessary only if ζ is not given, which was applied in the simulation study presented in the next chapter.

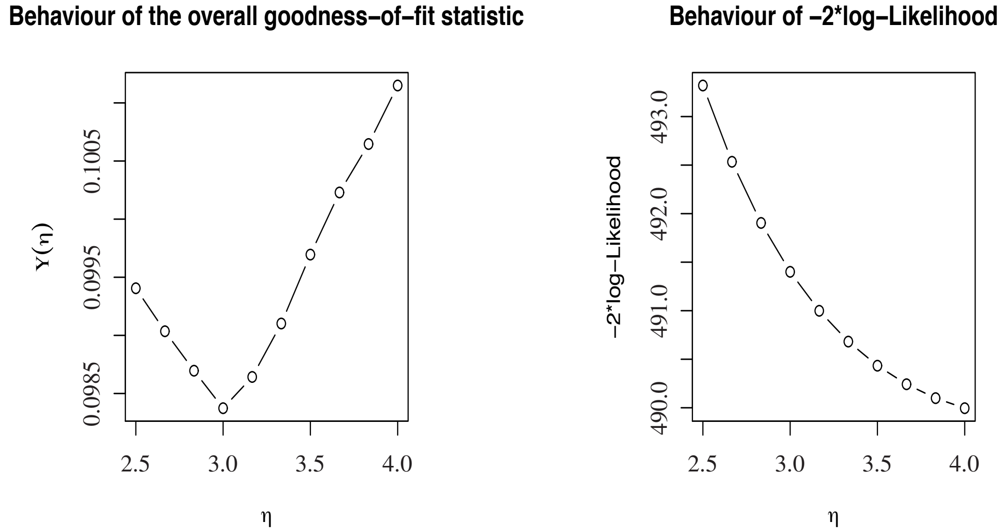

Figure 1 – Graphs generated by function `extra.parameter()`

2.2 Log-symmetric regression models

The log-symmetric regression model provides a unique way to model two important measures for positive skewed distributions: median and skewness (or the relative dispersion). Thus, these models have received attention and applications in recent literature. Boston crime data were analyzed by Vanegas and Paula (2016a). Medeiros and Ferrari (2017) developed a Barlett-type correction application for these models. In addition, Vanegas and Paula (2017) proposed a development to accept the presence of non-informative left or right censored data. Despite the high density and volatility, financial data were fitted by a log-symmetric duration model in Saulo and Leão (2017).

Let $T_1, \ldots, T_n$ be independent random variables such that $T_i \sim \mathrm{LS}(\eta_i, \phi_i^2, g)$. In addition, suppose that $\eta_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, for $i = 1, \ldots, n,$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of unknown parameters and $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the ith observation on a set of p covariates. Then, the log-symmetric linear regression model is given by

$$T_i = \eta_i \, \epsilon_i^{\sqrt{\phi}}, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\epsilon_i \sim \mathrm{LS}(1, 1, g)$. By applying the logarithm on both sides of (2.3), we obtain

$$\underbrace{\log(T_i)}_{Y_i} = \underbrace{\log(\eta_i)}_{\mu_i} + \sqrt{\phi}\,\underbrace{\log(\epsilon_i)}_{\varepsilon_i}, \quad i = 1, \ldots, n, \tag{2.4}$$

where $\log(\eta_i) = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $\phi_i > 0$ is the skewness (or relative dispersion), $\varepsilon_i \overset{\text{IID}}{\sim} \mathrm{S}(0, 1, g)$, that is, independent and identically distributed (IID), ϕ is unknown and $Y_i \sim \mathrm{S}(\mu_i, \phi_i^2, g)$; see Vanegas and Paula (2017).

Vanegas and Paula (2016b) added more flexibility to log-symmetric model regressions by allowing that median or skewness may be described using an undetermined number of parametric functions. Therefore, either possible relations between median, η_k, skewness, ϕ_k, and covariates are represented as

$$\eta_k = \eta(x_k, \beta) \quad \text{or} \quad \log(\eta_k) = x_k^\top \beta \quad \text{and} \quad \log(\phi_k) = \boldsymbol{w}_k^\top \gamma, \tag{2.5}$$

whereas median and skewness covariates are represented by vectors $\boldsymbol{x}_{\boldsymbol{k}}^* = (x_k^t, a_{k,1}, \ldots, a_{k,p'})^\top$ and $\boldsymbol{w}_k^* = (w_k^t, a_{k,1}, \ldots, a_{k,q'})^\top$. Note that the function $\log(\eta(x_k, \cdot))$ is continuous and differentiable twice for $\boldsymbol{\beta} \in \Omega_\beta$ (compact set with interior points). Unknown parameters are given by vectors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$, and estimated via the ML method, given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(g(z_i^2)) - 0.5\log(\phi_i), \tag{2.6}$$

in which $\ell(\boldsymbol{\theta})$ equals the ith contribution to the log-likelihood function, whereas $z_i = (y_i - \mu_i)/\phi$. Note that $\boldsymbol{\theta} = (\eta, \phi)^\top$ has not an analytical closed form solution. Then, it is obtained by a numerical iterative procedure solving the system equations of either score vectors for $(U_\eta(\widehat{\boldsymbol{\theta}}), U_\phi(\widehat{\boldsymbol{\theta}})) = (\partial\ell(\widehat{\boldsymbol{\theta}}/\partial\widehat{\eta}), \partial\ell(\widehat{\boldsymbol{\theta}}/\partial\widehat{\phi})) = (0, 0)$. Note that $\widehat{\eta}$ and $\widehat{\phi}$ are maximum likelihood estimates (MLEs) of η and ϕ, and also are proportional to geometrical and arithmetical mean of the observations, respectively. Regarding (2.5), $U(\boldsymbol{\theta})$ can be represented by

$$U(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{D}_\beta^\top \boldsymbol{\Omega}^{-1} \boldsymbol{D}_{(v)}(\boldsymbol{y} - \boldsymbol{\mu}) \\ \frac{1}{2}\boldsymbol{W}^\top(\boldsymbol{v}_n \boldsymbol{z}_n^2 - \boldsymbol{1}_n) \end{bmatrix}, \tag{2.7}$$

where $\boldsymbol{\mu} = (\mu_1), \ldots, \mu_n)^\top$, $\boldsymbol{\Omega}$ denotes the diagonal matrix of $\boldsymbol{\phi}$ and $\boldsymbol{D}_{(v)}$ represents the diagonal matrix of $\boldsymbol{v}$.

According to Vanegas and Paula (2016b), all log-symmetric distributions mentioned in Table 1 follow the regularity conditions of large sample theory. Thus, the asymptotic distribution of $\widehat{\theta}$ for $\boldsymbol{Z} \sim \mathrm{S}(0, 1, g)$is represented by

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{D} \mathrm{N}\left(0; \begin{bmatrix} \phi\eta^2(\mathrm{E}[v^2(Z)Z^2])^{-1} & 0 \\ 0 & 4\phi^2(\mathrm{E}[v^2(Z)Z^4] - 1)^{-1} \end{bmatrix}\right) \tag{2.8}$$

2.3 Model selection criteria

AIC is a method for model selection. Given $\boldsymbol{\theta}$, a vector of parameters, and $\widehat{\boldsymbol{\theta}}$, its respective estimate, of a distribution with PDF $f(x|\boldsymbol{\theta})$, the estimate is the maximized result of the expected log-likelihood function, that is

$$\mathrm{E}\left[\log(f(X|\widehat{\theta}))\right] = \int f(x|\theta)\log(f(x|\widehat{\theta}))\mathrm{d}x. \tag{2.9}$$

Note that the left hand side of (2.9) is the same of (2.6). Also, X is a random variable independent of $\widehat{\boldsymbol{\theta}}$ which has a PDF $f(x|\boldsymbol{\theta})$. Furthermore, Akaike (1992) showed that (2.9) is equivalent to maximizing an information quantity given by

$$\mathrm{E}\left[\log\left(\frac{f(X|\widehat{\boldsymbol{\theta}})}{f(X|\boldsymbol{\theta})}\right)\right] = \int f(x|\boldsymbol{\theta})\log\left(\frac{f(x|\widehat{\boldsymbol{\theta}})}{f(x|\boldsymbol{\theta})}\right)\mathrm{d}x. \tag{2.10}$$

The right side of (2.10) gives the Kullback-Leibler mean information for discriminating between $f(x|\widehat{\boldsymbol{\theta}})$ and $f(x|\boldsymbol{\theta})$. It is known it is a measure of distance between two PDFs. Also, it a metric for model fitting used to examine how different k fitted models $f(x|\widehat{\boldsymbol{\theta}}_k)$ are in relation to the original PDF $f(x|\boldsymbol{\theta})$. The AIC is given by

$$\mathrm{AIC} = -2\log(f(X|\widehat{\theta}_k)) + 2k, \tag{2.11}$$

where $\log(f(X|\widehat{\theta}_k))$ represents the maximum numerical value of the log-likelihood function, and k denotes the number of estimable parameters. Hurvich and Tsai (1989) highlighted that AIC is biased and tends to select overparameterized models in small samples. Thus, the authors corrected the AIC. The difference between the AIC and its corrected is in the last term of the expression defined below in (2.12). This term treats to correct the small-sample tendencies of the AIC. However, in greater samples, AIC's performance is asymptotically the same as AICc, see Allan McQuarrie and Tsai (1998).

As one of numerous proposed alternatives, BIC imposes a higher penalty term than AIC. Its objective is to predict more accurately what dimension is more adequate as the number of model observation increases. Nevertheless, Kass and Wasserman (1995) and Giraud (2014) pointed out that it has limitations. For instance, a better performance is observed when n is much larger than the number of parameters k in the model. Hence, a higher penalty term for small samples was needed and Allan D McQuarrie (1999) proposed a different penalty in BICc.

The HQIC focuses on autoregressive models and can be applied to regression models. Its performance is similar to AIC and it has a corrected version called HQICc; see Allan McQuarrie and Tsai (1998). Therefore, a different penalty term than $2k$ in (2.11) is suggested in all model selection criteria mentioned above as

$$\begin{aligned}
\text{AICc} &= -2\log(f(X|\widehat{\theta}_k)) + 2k\left(\frac{n}{n-k-1}\right), \\
\text{BIC} &= -2\log(f(X|\widehat{\theta}_k)) + k\log(n), \\
\text{BICc} &= -2\log(f(X|\widehat{\theta}_k)) + k\log(n)\frac{k}{n-k-2}, \\
\text{HQIC} &= -2\log(f(X|\widehat{\theta}_k)) + 2k\log(\log(n)), \\
\text{HQICc} &= -2\log(f(X|\widehat{\theta}_k)) + \frac{2k\log(\log(n))}{n-k-2}.
\end{aligned}$$

3 Monte Carlo simulation studies

In this chapter, we carry out two Monte Carlo simulation studies to evaluate the performances of the ML estimators and the information criteria. The first simulation study considers six different log-symmetric distributions and aims to verify whether ML estimators are asymptotically consistent. The second simulation study has two specific objectives related to information criteria: testing model dimensioning and the estimation of a log-symmetric distribution. Thus, simulated results are presented.

3.1 Maximum likelihood estimators

We propose a Monte Carlo simulation study with six log-symmetric distributions: log-contaminated-normal, log-hyperbolic, log-normal, log-power-exponential, log-student-t and log-slash. We intend to examine the asymptotic performance of ML estimators for median and skewness submodels. We define $\zeta \in \{4, 0.9, 1, 4, (0.5, 0.5)\}$, for log-student-$t$, log-power-exponential, log-hyperbolic, log-slash, and log-contaminated-normal distributions. Note that the log-normal distribution does not have a ξ extra-parameter, according to Table 1. In addition, the number of Monte Carlo replications are defined to be 1000 for each sample size, $n \in \{20, 50, 100\}$. We elaborate an algorithm to generate a log-symmetric distribution as follows:

(i) Generate x_1, x_2, x_3 and x_4 from a multivariate normal distribution.

(ii) Define the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for median and skewness (or relative dispersion) submodels. The vectors used in this work are defined as (log-symmetric regression may involve non-parametric approaches as shown in (2.5) and (**??**); however, only parametric functions are covered in this work)

$$\eta_t = 2.5 + 3x_{1_t} + 0.9x_{2_t}, \tag{3.1}$$

$$\phi_t = 0.7 + 0.25x_{3_t} + 1.5x_{4_t}. \tag{3.2}$$

Note that Bayer and Cribari-Neto (2017) presented either easily and weakly identifiable scenarios, in which parameter values β_1 and β_2 are defined to be different and the same. We deal with the second scenario.

(iii) Generate η_t and ϕ_t.

(iv) Get ξ_t from a previously defined log-symmetric distribution as well as the extra-parameter value ζ .

(v) Use η_t and ϕ_t to generate ξ_t from a determined log-symmetric distribution, and extra-parameter value ζ.

(vi) Generate the variable Z after substituting all variables into (2.3).

The results for each log-symmetric distribution are described in Tables 2, 3, 4, 5, 6 and 7, whereas first column represents the following measured statistics: empirical mean, coefficients of skewness(CS) and kurtosis (CK), relative bias (RB), and random mean squared error (RMSE). In addition, values for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from median and skewness submodels are represented and compared to its ML estimates for each n sample. Overall, we can infer from results that ML estimates are asymptotically efficient for all log-symmetric distributions considered.

3.2 Information criteria

We carry out an additional Monte Carlo study to evaluate the performance of information criteria mentioned in Section 2.3. This study addresses two relevant issues. First, we examine model dimension selection, that is, we investigate what criterion is more accurate to suggest the right order of parameters considering median and skewness submodel. The second topic refers to the model suggested by the criteria. We check whether the log-symmetric distribution used to generate the data is correctly suggested by a model selection criteria among all the considered distributions of this same class. More specifically, in the first study, the number of parameters k for median and skewness submodels are six. For over specified models, $(k > 6)$, x_3, x_4, x_5 are fitted. In this way, each submodel may contain $k = 1, \dots, 5$ parameters. Thus, $k = 2$ and $k = 10$ represent the minimum and maximum order of parameters for each log-symmetric model. All

Table 2 – Statistics and estimates for the log-normal distribution.

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.4668	2.4820	2.500	2.9927	2.9805	3.0089	0.9038	0.8875	0.9007
CS	0.0369	-0.1391	-0.0505	-0.1051	0.0569	0.0377	-0.0275	-0.0676	0.0386
CK	0.1564	-0.1027	0.0302	-0.0101	0.1365	-0.1115	0.1933	-0.0095	0.0058
RB	-0.0332	-0.0180	0.001	-0.0073	-0.0195	0.0089	0.0038	-0.0125	0.0007
RMSE	0.4198	0.3273	0.2538	0.4142	0.3856	0.2478	0.4778	0.3655	0.2147
	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.3802	-0.3427	-0.4360	0.0480	0.1835	0.2346	0.9946	1.1309	1.2984
CS	-0.2167	0.0163	0.0469	0.1890	-0.1967	0.1961	-0.0396	0.1585	0.0251
CK	0.0801	0.0973	0.0872	-0.0321	0.0522	0.1775	0.5503	0.3299	0.0508
RB	-1.0802	-1.0427	-1.1360	-0.2020	-0.0665	-0.0154	-0.5054	-0.3691	-0.2016
RMSE	1.1833	1.0988	1.1628	0.4943	0.3251	0.2195	0.6819	0.4949	0.3104

Table 3 – Statistics and estimates for the log-student-t distribution.

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.4827	2.514	2.485	3.0217	3.0296	3.0032	0.9260	0.8829	0.9046
CS	-0.0304	0.1479	-0.2188	0.0562	0.2239	0.0969	-0.2284	-0.3450	0.1045
CK	1.0982	0.9225	0.7403	1.0593	2.8345	0.4233	2.2576	5.9390	0.2086
RB	-0.0173	0.0145	-0.0147	0.0217	0.0296	0.0032	0.0260	-0.0171	0.0046
RMSE	0.6263	0.4829	0.3375	0.6093	0.5498	0.3405	0.6738	0.5306	0.2926
	γ_1			γ_2			γ_3		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.6418	-0.6264	-0.7852	0.0221	0.1782	0.2006	0.9133	1.0186	1.2248
CS	0.1212	0.5195	0.1386	0.0732	-0.3106	0.0881	0.0442	0.1754	-0.1432
CK	0.6408	1.7560	0.2661	0.2470	0.2644	-0.0403	0.5456	0.3222	0.0484
RB	-1.3418	-1.3264	-1.4852	-0.2279	-0.0718	-0.0494	-0.5867	-0.4814	-0.2752
RMSE	1.4598	1.3951	1.5137	0.5290	0.3294	0.2377	0.7533	0.5898	0.3744

Table 4 – Statistics and estimates for the log-power-exponential distribution.

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.5663	2.5018	2.4861	3.0715	3.0124	2.9878	0.8815	0.9040	0.9026
CS	0.0469	0.0358	0.0440	0.0688	-0.0343	-0.0095	0.0417	-0.0657	0.0659
CK	0.0594	0.0589	0.0689	0.2940	1.0404	0.5932	0.5199	0.7307	0.9375
RB	0.0663	0.0018	-0.0139	0.0715	0.0124	-0.0122	-0.0185	0.0040	0.0026
RMSE	1.1230	0.8163	0.6464	1.0692	0.9541	0.6375	1.2383	0.9192	0.5308
	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.6401	-0.6785	-0.7790	0.0334	0.1698	0.2145	0.8512	1.0013	1.1871
CS	-0.1690	0.0742	0.1263	0.1129	-0.0684	-0.0250	0.0176	-0.0022	-0.0448
CK	0.1988	0.2835	-0.0136	0.3044	0.1158	0.0642	0.5036	0.1423	0.4078
RB	-1.3401	-1.3785	-1.4790	-0.2166	-0.0802	-0.0355	-0.6488	-0.4987	-0.3129
RMSE	1.4713	1.4519	1.5092	0.5104	0.3494	0.2372	0.8130	0.6037	0.4042

Table 5 – Statistics and estimates for the log-hyperbolic distribution.

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.4987	2.5281	2.4870	2.9936	2.9853	3.0167	0.8772	0.8864	0.8868
CS	-0.0229	-0.0535	-0.1243	0.1198	-0.0059	0.0913	0.0790	-0.0264	0.0732
CK	-0.0175	0.0372	0.0296	0.3010	0.2265	0.1898	1.4154	0.7078	0.2260
RB	-0.0013	0.0281	-0.0130	-0.0064	-0.0147	0.0167	-0.0228	-0.0136	-0.0132
RMSE	0.7418	0.5375	0.3949	0.6976	0.6117	0.4057	0.8596	0.6174	0.3362
	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.5482	-0.5603	-0.6688	0.0323	0.1839	0.2126	0.9094	1.0366	1.2516
CS	-0.0666	0.0522	0.1153	0.1646	-0.2219	-0.0914	0.1108	0.0984	0.0098
CK	0.1399	0.1066	0.1228	0.0479	0.0897	0.0524	0.3357	0.1500	-0.0435
RB	-1.2482	-1.2603	-1.3688	-0.2177	-0.0661	-0.0374	-0.5906	-0.4634	-0.2484
RMSE	1.3690	1.3247	1.3966	0.5057	0.3213	0.2408	0.7553	0.5735	0.3557

Table 6 – Statistics and estimates for the log-slash distribution.

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.4894	2.5103	2.4975	2.9946	2.9995	3.0070	0.9046	0.9021	0.8978
CS	0.1347	-0.0441	0.1858	-0.0253	0.0096	-0.0694	-0.0100	0.0097	-0.0237
CK	0.0607	-0.0356	0.1422	0.0405	-0.0970	-0.0213	0.0864	-0.0181	0.2163
RB	-0.0106	0.0103	-0.0025	-0.0054	-0.0005	0.0070	0.0046	0.0021	-0.0022
RMSE	0.4953	0.3831	0.2866	0.4857	0.4354	0.2883	0.5731	0.4296	0.2326
	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.3911	-0.3923	-0.4785	0.0173	0.1816	0.2161	0.9648	1.1000	1.2782
CS	-0.1234	-0.0832	0.0724	0.0411	-0.2699	-0.0869	0.0997	0.12338	-0.0235
CK	-0.0720	-0.0037	-0.1032	0.4084	0.0007	0.2750	0.1932	-0.1192	0.1260
RB	-1.0911	-1.0923	-1.1785	-0.2327	-0.0684	-0.0339	-0.5352	-0.3999	-0.2218
RMSE	1.1984	1.1516	1.2027	0.5030	0.3466	0.2224	0.6946	0.5240	0.3231

Table 7 – Statistics and estimates for the log-contaminated-normal distribution

Statistic	n								
	20	50	100	20	50	100	20	50	100
	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
True value	2.5	2.5	2.5	3	3	3	0.9	0.9	0.9
Mean	2.5231	2.5175	2.5149	3.0043	2.9975	2.9998	0.9018	0.9095	0.8927
CS	0.0488	-0.0357	-0.2188	-0.0628	0.0209	-0.0514	0.1331	0.0423	0.0606
CK	-0.1386	-0.2959	0.2829	0.0858	0.4446	-0.1295	0.1465	0.0901	0.2561
RB	0.0231	0.0175	0.0149	0.0043	-0.0025	-0.0002	0.0018	0.0095	-0.0073
RMSE	0.5520	0.4229	0.2935	0.5050	0.4690	0.2920	0.6014	0.4483	0.2531
	$\hat{\gamma}_1$			$\hat{\gamma}_2$			$\hat{\gamma}_3$		
True value	0.7	0.7	0.7	0.25	0.25	0.25	1.5	1.5	1.5
Mean	-0.4230	-0.4114	-0.4978	0.0253	0.2030	0.2197	0.9521	1.0908	1.2872
CS	-0.1953	-0.0368	-0.1058	0.3547	0.0376	-0.0991	-0.0218	0.1980	0.0602
CK	0.2318	0.0253	0.2201	0.8495	0.3569	-0.0403	0.5091	0.1684	0.0013
RB	-1.1230	-1.1114	-1.1978	-0.2247	-0.0470	-0.0303	-0.5479	-0.4092	-0.2128
RMSE	1.2250	1.1740	1.2252	0.5119	0.3230	0.2246	0.7083	0.5297	0.3179

possible combinations are generated and 35 alternative pairs of submodels are analyzed. In total, 1000 Monte Carlo replications, 36 regression model and four different samples generate 144000 results for each model selection criteria. We consider the following step-by-step procedure:

(i) Generate x_1, x_2, x_3, x_4, x_5 from a multivariate normal distribution.

(ii) Define median and skewness (or relative dispersion) submodels as

$$\eta_t = 3 + 0.5x_{1_t} + 1.2x_{2_t}, \tag{3.3}$$

$$\phi_t = 0.3 + 0.1x_{1_t} + 0.7x_{2_t}. \tag{3.4}$$

(iii) Obtain η_t and ϕ_t.

(iv) Get ξ_t from a log-symmetric distribution, and the extra-parameter value ζ (we define the extra-parameter as $\zeta = 4$ to generate a log-slash distribution).

(v) Generate the variable Z after substituting all variables into (2.3).

(vi) Fit Z with each possible combinations of X_k parameters for either median and skewness submodels.

(vii) Compute AIC, AICc, BIC, BICc HQIC, and HQICc.

The second approach aims to inspect the accuracy of information criteria on choosing log-symmetric distributions. Six log-symmetric distributions are evaluated; log-normal, log-student-t, log-slash, log-hyperbolic, log-power-exponential and log-contaminated-normal. The objective is to count how many times the model selection criteria suggests the same distribution used to generate the data, instead of other five candidates from the log-symmetric class. As mentioned, 1000 Monte Carlo replications, six regression models and four different samples generate 24000 results for each model selection criteria. The following algorithm describes the process used in this work (note that additional steps are required to provide the best adjustment and parameter estimation, regarding (2.6); the results for both information criteria applications are provided in Tables 8 and 9):

(i) Generate x_1, x_2, x_3, x_4 from a multivariate normal distribution.

(ii) Define median and skewness (or relative dispersion) submodels as

$$\eta_t = 2.5 + 3x_{1_t} + 0.9x_{2_t}, \quad (3.5)$$
$$\phi_t = 0.7 + 0.25x_{3_t} + 1.5x_{4_t}. \quad (3.6)$$

iii) Obtain η_t and ϕ_t.

iv) Get ξ_t from six log-symmetric distributions and the defined extra-parameter value ζ (note that we define the extra-parameter as $\zeta \in \{4, 0.9, 1, 4, (0.5, 0.5)\}$ to generate respectively the log-student-t, log-power-exponential, log-hyperbolic, log-slash, and log-contaminated-normal distributions).

(v) Generate the variable Z after substituting all variables into (2.3).

(vi) Fit Z with each possible combination of different distributions, using the extra-parameter values in (iv).

vii) Re-fit Z with a parameter ξ given by a range of values defined for each distribution, in order to minimize the log-likelihood statistics.

viii) Compute the correct and incorrect distribution specifications.

Table 8 – Model selection criteria results on order parameterization

	$n = 25$			$n = 30$			$n = 40$			$n = 50$		
	$< k$	$= k$	$> k$	$< k$	$= k$	$> k$	$< k$	$= k$	$> k$	$< k$	$= k$	$> k$
AIC	336	132	532	274	136	590	359	146	495	315	153	532
AICc	375	143	482	306	150	544	383	150	467	321	157	522
BIC	363	138	499	307	141	552	397	150	453	339	156	505
BICc	289	128	583	238	124	638	303	140	557	290	150	560
HQIC	343	134	523	286	137	577	376	146	478	321	155	524
HQICc	270	126	604	225	120	655	291	134	575	285	148	567

Table 9 – Model selection criteria results on distribution specification

		$n = 25$		$n = 30$		$n = 40$		$n = 50$	
Distribution	Criteria	inc.	c.	inc.	c.	inc.	c.	inc.	c.
Log-normal	AIC	855	145	893	107	892	108	913	87
	AICc	855	145	893	107	892	108	913	87
	BIC	855	145	893	107	892	108	913	87
	BICc	855	145	893	107	892	108	913	87
	HQIC	855	145	893	107	892	108	913	87
	HQICc	855	145	893	107	892	108	913	87
Log-student-t	AIC	413	587	405	595	342	658	277	723
	AICc	413	587	405	595	342	658	277	723
	BIC	413	587	405	595	342	658	277	723
	BICc	413	587	405	595	342	658	277	723
	HQIC	413	587	405	595	342	658	277	723
	HQICc	413	587	405	595	342	658	277	723
Log-power-exponential	AIC	904	96	855	145	894	106	905	95
	AICc	904	96	855	145	894	106	905	95
	BIC	904	96	855	145	894	106	905	95
	BICc	904	96	855	145	894	106	905	95
	HQIC	904	96	855	145	894	106	905	95
	HQICc	904	96	855	145	894	106	905	95
Log-hyperbolic	AIC	885	115	904	96	887	113	908	92
	AICc	885	115	904	96	887	113	908	92
	BIC	885	115	904	96	887	113	908	92
	BICc	885	115	904	96	887	113	908	92
	HQIC	885	115	904	96	887	113	908	92
	HQICc	885	115	904	96	887	113	908	92
Log-slash	AIC	980	20	995	5	993	7	997	3
	AICc	980	20	995	5	993	7	997	3
	BIC	980	20	995	5	993	7	997	3
	BICc	980	20	995	5	993	7	997	3
	HQIC	979	21	995	5	993	7	997	3
	HQICc	980	20	995	5	993	7	997	3
Log-contaminated-normal	AIC	889	111	902	98	911	89	933	67
	AICc	889	111	902	98	911	89	933	67
	BIC	889	111	902	98	911	89	933	67
	BICc	889	111	902	98	911	89	933	67
	HQIC	889	111	902	98	911	89	933	67
	HQICc	889	111	902	98	911	89	933	67

Considering results from Table 8, note that AIC tends to select overparameterized models, $n = 25(532), n = 30(590), n = 40(495), n = 50(532)$, in comparison to AICc, $n = 25(482), n = 30(544), n = 40(467), n = 50(522)$. Hence, AICc outperforms AIC in small samples. In fact, AICc is the best model selection criterion, with the highest correct specifications for all samples $n = 25(143), n = 30(150), n = 40(150), n = 50(157)$. In addition, BIC is the second best information criterion, with similar results, $n = 25(138), n = 30(141), n = 40(150), n = 50(156)$, if compared to AICc. However, HQICc is the worst performing model selection criterion, with the highest overparameterization, $n = 25(604), n = 30(655), n = 40(575), n = 50(567)$, and lowest correct order dimensioning specifications, $n = 25(126), n = 30(120), n = 40(134), n = 50(148)$. Furthermore, from Table 9 and its columns "inc." and "c." denoting respectively the incorrect and correct distribution specifications, we detect that the model selection criteria accuracy is better for log-student-t and log-power-exponential distributions. Nevertheless, their performances are equal for each sample and distribution. Also, note that the log-slash and log-contaminated-normal distributions are harder to identify than the others. Thus, a generated log-symmetric distribution will not necessarily be suggested despite of other log-symmetric candidates.

4 An application to movie data

In this chapter, we analyze movie data. Data from IMDb were scrapped with a computational routine created in `R`. The objective is to propose a model for box-offices, whereas the covariates to be considered are: budgets, IMDb's ratings mean and total number of notes given by each registered IMDb's user. We compute the ML estimates of the each log-symmetric submodel parameter. Then, two residual envelopes are constructed to compare log-normal and chosen distributions. Model selection criteria are also used to choose best possible log-symmetric model.

4.1 Data description

Movies and econometrics are related since Litman (1983), which proposed a theatrical renting model, according to Greene (2011). Also, Ravid (1999) developed a study of movie stars salaries, revenues, and return on investment. Vany and Walls (1999) investigated the risk modelling of movie business, and later the relation between box-offices and movie main actors in Walls and De Vany (2002). In addition, the probability distributions of budgets and revenues for different rating movies' classes were estimated in Walls and De Vany (2002).

In a book which covers financial movies structure, Epstein (2012) emphasized that data available on internet are superficial, considering that studios cash flow raises a lot of incomes, besides movies tickets, such as renting and selling DVD's, blu-ray's, advertisement, and on-demand services. Furthermore, Epstein (2012) revealed that officially reported movie budgets may not be real for studios. The "latter budget" is rarely seen by anyone outside of a studio. The book provides the example of Lara Croft's Tomb Raider (2001). Its reported budget was US$94 million, but studio's spent was only US$8.7 million. This was possible with several distribution rights contracts and tax reliefs. Due to those adversities, the present work focuses on examining only costs of movies production, and box offices in United States presented on IMDb website.

The movie data were scrapped from `imdb.com`IMDb.com, which contains data about movies, games, audiovisual series, and TV shows. More specifically, this database begins in 1989 with Col Needham and his colleagues at Usenet. A year later they accumulated data about cinema professionals, in more than 10000 TV series and movies. Due to the nonexistence of web browsers, data were divulged by shell scripts for UNIX systems and its name was “rec.arts.movies movie database”. In 1993, database was available with World Wide Web development and three years later Internet Movie Database has become a commercial business. Nowadays, the website belongs to Amazon enterprise.

The main purpose here is to model movie’s tickets incomes based on the covariates: budgets, IMDb rating (in a scale of 0 to 10), and number of notes. Thus, 250 top rated movies were extracted using `rvest` package in RStudio Team (2016). Although the date of collecting was January 31, 2017, there were a lot of non US movies, as missing data about movies budgets and its gross. In addition, the database has movies from 1929 to 2016. Old movies can still be considered popular ones, but their financial data may be more difficult to find if they were not produced in the US. Therefore, database was reduced from 250 movies to 155.

Inflation is important on mensuration and comparison of incomes and budgets for different movies and years. For instance, success movies are re-released, and its incomes are biased. Based on these adjustments gross and incomes were modified considering time registered of gross to 2017’s prices. For example, Matrix was released on March 31, 1999, and its income in September 17, 1999, was US$171.383.253,17 and its budget estimated in US$63.000.000,00. To deflation both, the deflater chosen was average price ticket’s annual variation, considering incomes and productions respective date. This methodology was found in `boxofficemojo.com`, which is a partner of IMDb.

4.2 Exploratory data analysis

First, Table 10 provides a descriptive summary of the box offices (in million US$) including mean ($\bar{y}$), minimum ($y_1$), maximum ($y_n$), median (MD), standard deviation (SD), coefficient of variation (CV), CS and CK. Also, Figure 2 presents the histogram of the data in original and logarithmic scale as well as the boxplot. Figure 3 shows scatterplots and coefficients of correlation between all variables proposed in this work, that is, Y_1 (log(box offices)), X_1 (ratings), X_2 (budget, in thousand US$), X_3 (number of votes).

$\bar{y}$	n	y_1	y_n	MD	SD	CV	CS	CK
103.10	155.00	0.09	936.60	46.84	136.74	132.60	2.41	11.55

Table 10 – Descriptive statistics for box-offices (in million US$)

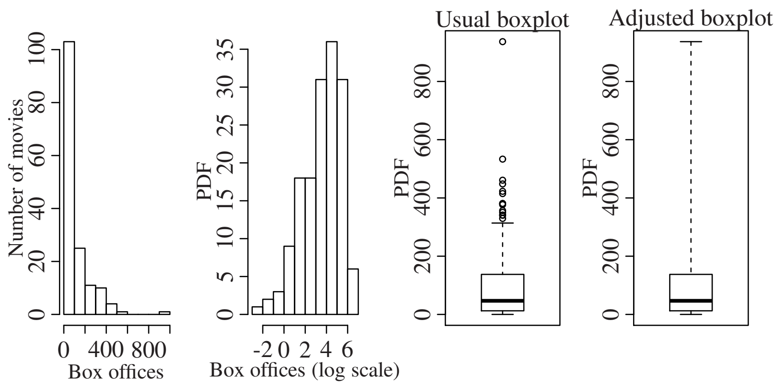

Figure 2 – Histograms (left) and boxplots (right) for the box-offices.

From Figure 2, note the presence of skewness in the response variable. More specifically, there is a concentration of 128 box-offices under US$200.000.000. In addition, the movie with lowest box-office was Infernal Affairs (2002), produced in Hong Kong. Its average rating was 8, and its budget US$8.242.555. Furthermore, highest income is of Star Wars Force Awakens (2015) with rating 8.1, budget of US$251.231.214, and box-office of US$936.627.416. From Table 10, note that the mean is approximately 50% greater than the median, which suggests an asymmetric distribution with positive skewness. Also, the high value of its CV (13260%) indicates the high variability presented. From Figure 3, note that the Pearson correlation is low between (X_1, X_2) and (X_1, Y_1). However, we detect that between (X_1, X_3), (X_2, X_3), (X_2, Y_1), and (X_3, Y_1) there is a medium correlation. Thus, there is no signal of multicolinearity, since covariates are not linearly correlated.

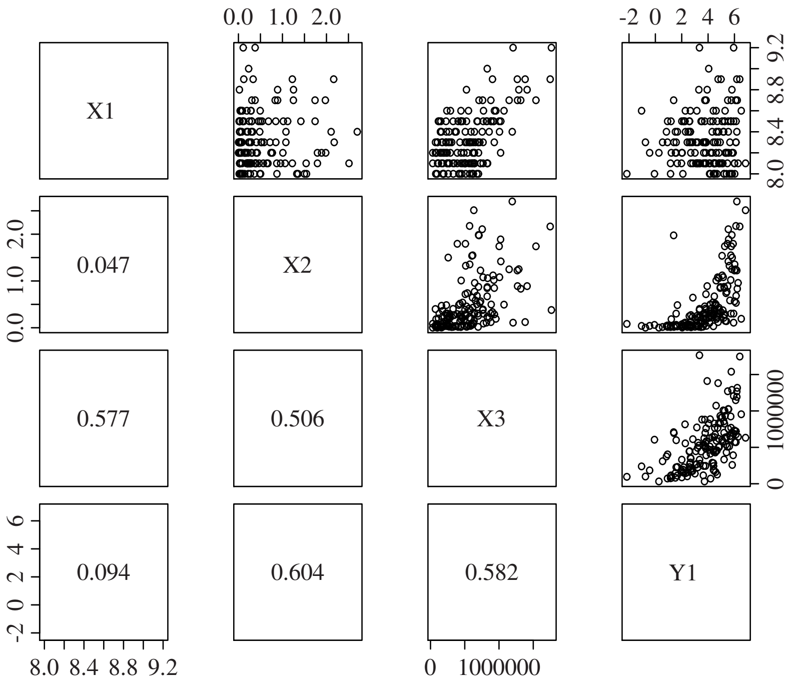

Figure 3 – Scatterplots and correlations of the model variables

4.3 Regression analysis

We consider movies' box-offices (in million US$) as the response variable, which is continuous and follows an asymmetric distribution with positive skewness. The covariates that could explain this response are: rating, budget (in thousand US$) and notes in the median submodel; and the last two can be used for modeling the skewness. Given this structure, median and skewness submodels are proposed as

$$\log(\eta_k) = \beta_1 + \beta_2 X_{1k} + \beta_3 X_{2k} + \beta_4 X_{3k}, \tag{4.1}$$

$$\log(\phi_k) = \gamma_1 + \gamma_2 X_{2k} + \gamma_3 X_{3k}{}^{1}. \tag{4.2}$$

The response variable Z is modeled with median, skewness, and errors as shown in (2.4). Hence, in its logarithm scale, it is represented by

$$\log(Z_k) = \sqrt{\exp(\phi)}\log(\exp(\eta)\xi_k), k = 1, \ldots, 155. \tag{4.3}$$

We estimate the parameters of movie data regression model via the ML method for different log-symmetric distributions. We use an `R` code with the function `ssym.l()` from the `ssym` package. Six distributions are considered: log-normal, log-Student-*t*, log-power-exponential, log-hyperbolic, log-slash and log-contaminated-normal. The results are presented in Table 11.

Table 11 – Extra-parameter and goodness-of-fit statistics for each distribution fitted to movie data.

Distribution	ζ	$-2\log(\hat{\boldsymbol{\theta}})$	AIC	AICc	BIC	BICc	HQIC	HQICc
Log-normal	-	501.724	515.72	516.49	537.03	497.31	524.38	492.87
Log-Student-*t*	3	491.4	505.40	506.16	526.70	486.99	514.05	482.55
Log-power-exponential	0.68	489.844	503.84	504.61	525.15	485.43	512.50	480.99
Log-hyperbolic	0.8	490.299	506.68	507.44	527.99	488.27	515.34	483.83
Log-slash	1	492.681	504.30	505.06	525.60	485.89	512.95	481.45
Log-contam.-normal	0.12 , 0.17	488.721	502.72	503.48	524.02	484.31	511.37	479.87

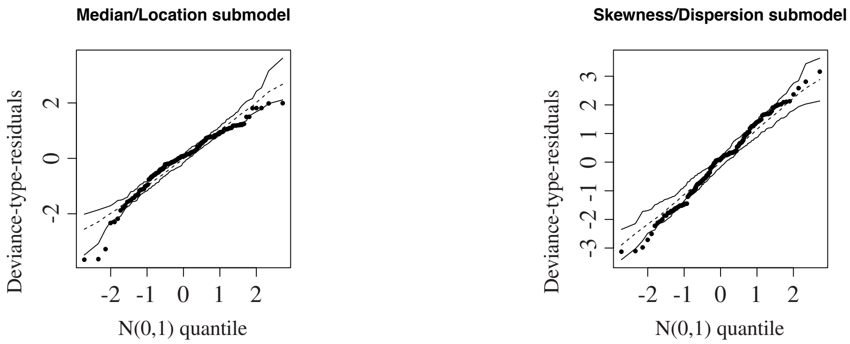

Figure 4 – Median (left) and skewness (right) residual envelopes for box-offices (log-normal)

Table 11 exhibits the value of ζ which minimizes information criteria, selecting the best possible log-symmetric distribution to represent the box-offices. Note that the log-normal distribution was the worst. Hence, it has the highest values of AIC (515.72), AICc (516.49), BIC (537.03), BICc (497.31), HQIC (524.38) and HQICc (492.87). Comparing it to the other six rows, we note that the log-normal distribution is outperformed by distributions with heavier tails on its goodness-of-fit measures. However, the log-contaminated-normal distribution, a mixture of normals, was chosen with the lowest values of AIC (502.72), AICc (503.48), BIC (524.02), BICc (484.31), HQIC (511.37)

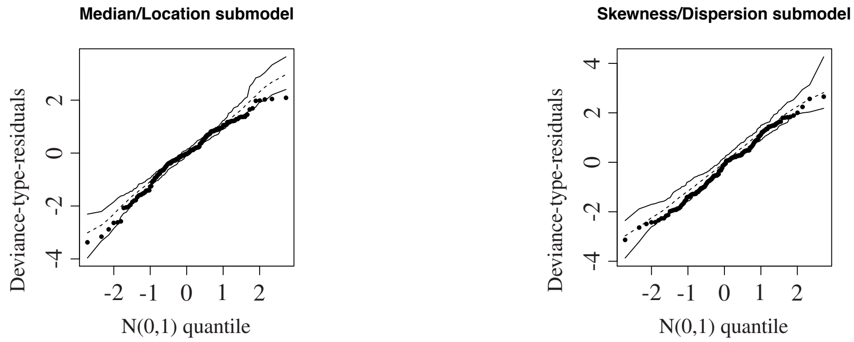

Figure 5 – Median (left) and skewness (right) residual envelopes for box-offices (log-contaminated-normal)

and HQICc (479.87). In fact, Figures 4 and 5 illustrate this difference with residual envelopes for either distributions.

Table 12 displays the parameter estimates for either median and skewness submodel. First, note that median submodel variables are individually statistically significant at 1%, with estimates $\widehat{\beta}_1 = 16.28$, $\widehat{\beta}_2 = 2.2337 \cdot 10^{-6}$, $\widehat{\beta}_3 = 0.9679$, $\widehat{\beta}_4 = -1.6882$. Also, skewness parameter are more statistically significant for the variable budget than for the intercept and number of votes. Second, we detect that estimates of parameters from the median submodel, $2.2337 \cdot 10^{-6}$ and 0.9679, suggest a positive relation between votes, budget and median of box offices. However, rating has a negative relation represented by -1.6882. Third, increasing budgets may decrease asymmetry of box offices.

In relation to marginal analysis, a unit increase of voting changes box offices median US\$26,322.28 (in thousand of dollars). This is possibly corroborated with Oscars nominees popularity, a good marketing or even a highly expected movie sequence, such as Star Wars: The Force Awakens. Movie costs also have a positive relation with incomes, since one more dollar budget means an increase of US\$11,416,433.40 in box offices median. Technology and movie special effects may turn a film very expensive, however more watchable. IMDb's rating increases in one point means a decrease in US\$3,677,423,424.96 for an income. Nevertheless, since lowest and highest ratings are 8 and 9.2, it is more representative to examine an effect of 0.1 point increase reaching a value of: US\$367,742,342.49 less.

Table 12 – Median and skewness estimates for movie data

	Median			Skewness		
	Estimate	SD	Pr($>\|z\|$)	Estimate	SD	Pr($>\|z\|$)
Intercept	16.2835	3.6386	< 0.01	0.1565	0.2501	< 0.53
Votes	$2.2337 \cdot 10^{-6}$	0.0001	< 0.01	$1.2937 \cdot 10^{-7}$	0.0001	< 0.77
Budget	0.9679	0.1444	< 0.01	-0.5616	0.2670	< 0.04
Rating	-1.6882	0.4493	< 0.01	NA	NA	NA

5 Conclusion and future works

Non-normality, and particularly asymmetry, is a relevant topic in statistics and may be present in data from many fields of study. In addition, the information criteria play a definitive role in model selection. The log-symmetric regression model has a flexible statistical structure, providing two important measures for asymmetric data: median and skewness. In this work, we have proposed Monte Carlo simulation studies and a real-data application for log-symmetric regression models. We have provided a robust solution for model selection criteria, diagnostics and modeling of movie box-offices. Furthermore, we have illustrated model flexibility on error assumptions by choosing six distributions from an full skewed family. First, we have carried out a study about the consistency of maximum-likelihood estimators from six log-symmetric distributions. Then, we have addressed the issue of parameter dimensioning in a regression model. Also, we have investigated if a log-symmetric distribution used to generate a variable is selected adequately by some commonly used model selection criteria, instead of other log-symmetric candidates. We have noted from results that the Akaike information criterion tends to select overparameterized models in small samples in comparison to corrected version, which outperformed all model selection criteria. We have detected that a distribution used to generate a log-symmetric variable not necessarily is suggested by all selection criteria. Second, we have proposed an application of log-symmetric regression models to movie data. Epstein (2012) studied budgets and box offices large structures and complexity. However, this work focused on American box offices and budgets published by Internet Movie Database. Statistics presented evidence that even with data boundaries results are satisfactory for analytics. More specifically, we have detected that votes and budgets were directly related to box-offices, having a negative relation between rating and box-offices. We have identified that votes and budgets were directly related to box-offices, having a negative relation between rating and box-offices. We have also noted that intercept and votes were not statistically significant for the skewness submodel. Meanwhile, budgets have had a negative relation with skewness of box-offices. In addition, we have observed that the log-normal distribution was outperformed by all others, being the worst candidate. Thus, distributions with heavier tails than the log-normal one should definitively be considered for data following asymmetric distributions, which is a strong reason to use log-symmetric regression models. Moreover, we have shown that Internet Movie Database contains variables of interest, which may be relevant for the movie industry in countries like Brazil.

Given the previous conclusions, we suggest possible future research related to:

i) Bootstrapped model selection criteria for log-symmetric regression models.

ii) Bootstrapped prediction interval for movies box offices.

iii) Reset test for log-symmetric regression models.

Essays on these problems are being developed and we hope to report their findings in future publications.

Bibliography

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: BREAKTHROUGHS in Statistics. [S.l.]: Springer, 1992. p. 610–624. Cit. on pp. 11, 17.

ALMEIDA, V.; DIVINO, J. A.; GADELHA, S. R. d. B.; MARANHÃO, A. Alíquotas Tributárias Efetivas Médias para a Economia Brasileira: Uma Abordagem Macroeconômica. **Revista Brasileira de Economia**, SciELO Brasil, v. 71, n. 2, p. 153–175, 2017. Cit. on p. 13.

BALAKRISHNAN, N.; SAULO, H.; BOURGUIGNON, M.; ZHU, X. On moment-type estimators for a class of log-symmetric distributions. **Computational Statistics**, p. 1–17, 2017. Cit. on p. 11.

BAYER, F. M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. **Communications in Statistics-Simulation and Computation**, v. 46, p. 729–746, 2017. Cit. on p. 20.

BENNETT, S. Log-logistic regression models for survival data. **Applied Statistics**, JSTOR, p. 165–171, 1983. Cit. on pp. 10, 13.

CROW, E. L.; SHIMIZU, K. **Lognormal distributions**. [S.l.]: Marcel Dekker New York, 1987. Cit. on pp. 10, 13.

DESOUSA, M.; SAULO, H.; LEIVA, V.; SCALCO, P. On a tobit-Birnbaum-Saunders model with an application to antibody response to vaccine. **Journal of Applied Statistics**, v. 45, p. 932–955, 2018. Cit. on p. 10.

EPSTEIN, E. J. **The Hollywood Economist: The Hidden Financial Reality Behind the Movies**. [S.l.]: Melville House, 2012. Cit. on pp. 28, 35.

FANG, K. T.; KOTZ, S.; NG, K. W. **Symmetric Multivariate and Related Distributions**. London, UK: Chapman and Hall, 1990. Cit. on p. 13.

GIRAUD, C. **Introduction to high-dimensional statistics**. [S.l.]: CRC Press, 2014. v. 138. Cit. on p. 18.

GREENE, W. H. **Econometric Analysis, 7th Edition**. 7th Edition. [S.l.]: Prentice Hall, 2011. Cit. on p. 28.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society B**, v. 41, p. 190–195, 1979. Cit. on p. 11.

HURVICH, C. M.; TSAI, C.-L. Regression and time series model selection in small samples. **Biometrika**, v. 76, p. 297–307, 1989. Cit. on p. 17.

KASS, R. E.; WASSERMAN, L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. **Journal of the american statistical association**, Taylor & Francis, v. 90, n. 431, p. 928–934, 1995. Cit. on p. 18.

KULLBACK, S. **Information theory and statistics**. [S.l.]: Courier Corporation, 1997. Cit. on p. 11.

LIMPERT, E.; STAHEL, W. A.; ABBT, M. Log-normal Distributions across the Sciences: Keys and Clues. **AIBS Bulletin**, American Institute of Biological Sciences, v. 51, n. 5, p. 341–352, 2001. Cit. on p. 10.

LITMAN, B. R. Predicting Success of Theatrical Movies: An Empirical Study. **The Journal of Popular Culture**, v. 16, 4 1983. Cit. on p. 28.

MCQUARRIE, A. D. A small-sample correction for the Schwarz SIC model selection criterion. **Statistics & probability letters**, v. 44, p. 79–86, 1999. Cit. on p. 18.

MCQUARRIE, A.; TSAI, C.-L. **Regression and time series model selection**. [S.l.]: World Scientific, 1998. Cit. on pp. 17, 18.

MEDEIROS, F. M. C.; FERRARI, S. L. P. Small-sample testing inference in symmetric and log-symmetric linear regression models. **Statistica Neerlandica**, v. 71, n. 3, p. 200–224, 2017. 10.1111/stan.12107. ISSN 1467-9574. Cit. on p. 15.

PUIG, P. A note on the harmonic law: a two-parameter family of distributions for ratios. **Statistics & Probability Letters**, v. 78, p. 320–326, 2008. Cit. on pp. 10, 13.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2016. Cit. on p. 12.

RAVID, S. A. Information, Blockbusters, and Stars: A Study of the Film Industry. **The Journal of Business**, University of Chicago Press, v. 72, 4 1999. Cit. on p. 28.

RSTUDIO TEAM. **RStudio: Integrated Development Environment for R**. Boston, MA, 2016. Cit. on pp. 12, 29.

SAULO, H.; LEÃO, J. On log-symmetric duration models applied to high frequency financial data. **Economics Bulletin**, v. 37, p. 1089–1097, 2017. Cit. on pp. 11, 15.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, v. 6, p. 461–464, 1978. Cit. on p. 11.

SZU, H.; HARTLEY, R. Fast simulated annealing. v. 122, p. 721–741, Jan. 1984. Cit. on p. 13.

TUKEY, J. W. A survey of sampling from contaminated distributions. **Contributions to probability and statistics**, Stanford University Press, p. 448–485, 1960. Cit. on p. 10.

VANEGAS, L.; PAULA, G. A. An extension of log-symmetric regression models: R codes and applications. **Journal of Statistical Computation and Simulation**, v. 86, p. 1709–1735, 2016. Cit. on pp. 10, 11, 15.

VANEGAS, L.; PAULA, G. A. Log-symmetric distributions: statistical properties and parameter estimation. **Brazilian Journal of Probability and Statistics**, v. 30, p. 196–220, 2016. Cit. on pp. 10, 16, 17.

VANEGAS, L.; PAULA, G. A. Log-symmetric regression models under the presence of non-informative left-or right-censored observations. **TEST**, v. 26, n. 2, p. 405–428, 2017. Cit. on pp. 10, 15, 16.

VANEGAS, L.; PAULA, G. A. **ssym: Fitting Semi-Parametric log-Symmetric Regression Models**. [S.l.], 2016. R package version 1.5.7. Cit. on p. 12.

VANY, A. D.; WALLS, W. D. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office. **Journal of Cultural Economics**, Springer US, v. 23, 4 Nov. 1999. Cit. on p. 28.

WALLS, W. D.; DE VANY, A. Big budgets, movie stars, and wide releases: Empirical analysis of the blockbuster strategy. In: PROCEEDINGS of the XIX Latin American meeting of the Econometric Society, San Paulo. [S.l.: s.n.], 2002. Cit. on p. 28.

WICKHAM, H. **rvest: Easily Harvest (Scrape) Web Pages**. [S.l.], 2016. R package version 0.3.2. Available from:
<`https://CRAN.R-project.org/package=rvest`>. Cit. on p. 12.