Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Modelagem baseada na distribuição Birnbaum-Saunders

Renata Villas Boas Dias

Orientador: Professor Helton Saulo Bezerra dos Santos

Brasília

2018

Renata Villas Boas Dias

Modelagem baseada na distribuição Birnbaum-Saunders

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Professor Helton Saulo Bezerra dos Santos

Brasília

2018

Renata Villas Boas Dias
Modelagem baseada na distribuição Birnbaum-Saunders
/ Renata Villas Boas Dias. – Brasília, 2018-
56 p.
Orientador: Professor Helton Saulo Bezerra dos Santos

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

Renata Villas Boas Dias

Modelagem baseada na distribuição Birnbaum-Saunders

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 26 de junho de 2018:

Professor Helton Saulo Bezerra dos Santos
Orientador

Eduardo Yoshio Nakano
Membro da Banca

Roberto Vila Gabriel
Membro da Banca

Brasília
2018

Agradecimentos

Primeiramente, agradeço ao Professor Helton Saulo pela orientação e comentários imprescindíveis ao trabalho, bem como aos conhecimentos de programação e estatística em geral. Não imagino outro professor para me orientar de forma tão construtiva e presente. Agradeço também a todos os meus professores que, de alguma forma, marcaram minha passagem pela UnB e que me ensinaram tanto e tão bem. Também agradeço à Estat por me ajudar a crescer, gostar um pouco mais da profissão de estatístico e conhecer pessoas que para sempre marcarão minha vida. Com certeza minha passagem pela UnB não seria a mesma sem vocês.

Agradeço também aos meus amigos. Alguns de vocês tornaram o dia a dia na universidade mais divertido. Outros tornaram as madrugadas de estudo suportáveis. Aos meus amigos fora da UnB, vocês me salvaram dos assuntos da universidade, das horas de estudos e dos dramas das disciplinas. Além disso, por entenderem (às vezes) quando tive que trocá-los por estudos, empresa júnior, TCC, estágio e etc.

Agradeço aos meus pais, Rubens e Isabela, por investirem no meu futuro e sempre acreditarem em mim (até demais). Por me apoiarem em todos os momentos de estresse, insegurança e desespero, e por terem me ajudado a escolher o curso de Estatística, que me trouxe aonde estou hoje. Além disso, por sempre estarem à disposição para me aconselhar sobre meu futuro e me apoiarem em todas as minhas decisões. Agradeço à minha irmã Marina por também sempre acreditar em mim e no meu potencial. Por, mesmo com a distância, estar sempre ao meu lado quando preciso, até mesmo nas horas de ler e comentar todos os meus trabalhos da universidade. Além disso, por todos os seus conselhos, por sempre pensar no meu futuro e me dar aquele puxão de orelha quando necessário. Você com certeza é um exemplo para mim. Agradeço aos meus avós pela segunda casa, pelo refúgio de estudos e por serem sempre tão prestativos, presentes e acolhedores. E a toda a minha família pelo amor incondicional.

Resumo

Recentemente, modelagens baseadas na distribuição Birnbaum-Saunders têm recebido considerável atenção, com diversos estudos sobre diferentes abordagens de modelos de regressão baseados nessa distribuição. Neste trabalho é feita uma avaliação de três abordagens - modelo BS, BSR e log-BS. Em particular, são realizadas simulações via método de Monte Carlo a fim de avaliar o desempenho dos estimadores de máxima verossimilhança dos modelos e comparar o desempenho deles quando se assume diferentes distribuições para a geração dos dados. Além disso, duas aplicações a dados reais são conduzidas com o objetivo de comparar os ajustes dos três modelos. Dessa forma, este estudo mostra uma comparação das três abordagens e busca identificar aspectos em que cada uma delas apresenta um desempenho melhor em relação às demais. Observa-se que o modelo BSR é o menos robusto, se feita uma comparação entre os três modelos, sendo o mais impactado pelos valores influentes em relação à estimação do parâmetro de forma. Em relação à estimação dos demais parâmetros, não se observa nenhum padrão. Assim, apesar de diversos estudos assumirem apenas a abordagem do modelo de regressão BS ou log-BS, não é evidente qual deles abordar, uma vez que eles mostram desempenhos semelhantes. Além disso, apesar de estudos que abordam o modelo BSR, é mostrado que ele apresenta resultados piores em certos aspectos, se comparado aos modelos BS e log-BS.

Palavras-chave: Distribuição Birnbaum-Saunders, Regressão Birnbaum-Saunders, Simulação de Monte Carlo.

Abstract

Recently, modelling based on the Birnbaum-Saunders distribution has received considerable attention, with many studies using different approaches of regression models based on this distribution. In this work, these approaches are evaluated - BS, BSR and log-BS models. Specifically, Monte Carlo simulations are carried out to assess the performance of the maximum likelihood estimators. Moreover, two applications to real data sets are conducted with the objective of comparing the adjustment of these three models. Therefore, this study shows a comparison of these approaches and tries to identify aspects in which each of them has a better performance if compared to the others. For both applications, regarding the estimation of the shape parameter, the BSR model is the most affected by influential data. Moreover, no pattern is identified regarding the estimation of the other parameters. So, although many studies use only the BS or log-BS approaches, it is not obvious which one should be used, once they show similar performance. Also, despite studies approaching the BSR model, it is shown that it has worst results in certain aspects if compared to the BS and log-BS models.

Keywords: Birnbaum-Saunders distribution, Birnbaum-Saunders Regression, Monte Carlo simulation.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A distribuição normal é o modelo mais importante e mais utilizado em estatística. Porém, tratando-se da modelagem de dados referentes a tempo de vida ela não é adequada. Isso se deve ao fato de que, em geral, distribuições de tempo de vida possuem assimetria à direita e assumem apenas valores maiores que zero. Assim, em análises de tempo de vida, distribuições com suporte positivo são utilizadas, como a distribuição Birnbaum-Saunders (BS).

Motivados por problemas de fadiga em materiais de aeronaves comerciais causada por vibração constante, Birnbaum e Saunders (1969) derivaram a distribuição Birnbaum-Saunders, uma distribuição de probabilidade que descreve tempos de vida associados a materiais expostos à fadiga, a qual é produto de estresse e tensão cíclica. Na derivação dessa distribuição, denotada por BS, o tempo decorrido até um dano cumulativo, produzido pelo desenvolvimento e crescimento de uma rachadura dominante, supera um valor limiar e produz a falha no material.

A distribuição BS tem recebido significativa atenção nos últimos tempos devido a seus argumentos teóricos combinados com a sua relação com a distribuição normal. Ela pode ser obtida por meio de duas abordagens distintas. A primeira tem origem na física dos materiais, o que permite a essa distribuição ser interpretada como uma distribuição de tempo de vida. A segunda abordagem é baseada na definição da própria distribuição. Ela diz que qualquer variável aleatória que segue distribuição BS é uma transformação de uma outra variável aleatória com distribuição normal padrão. Por meio dessa segunda abordagem, algumas generalizações e extensões dessa distribuição podem ser obtidas, como aquelas discutidas por Rieck e Nedelman (1991), Balakrishnan e Zhu (2014) e Leiva et al. (2014). Algumas delas também são atribuídas a Díaz-Garcia e Leiva-Sánchez (2005), Balakrishnan et al. (2011), Gomez, Olivares-Pacheco e Bolfarine (2009), Vilca-Labra e Leiva-Sánchez (2006) e Fierro et al. (2013).

O objetivo deste trabalho consiste em realizar uma avaliação de três abordagens distintas de modelos de regressão baseados na distribuição Birnbaum-Saunders, discutidas por Rieck e Nedelman (1991), Balakrishnan e Zhu (2014) e Leiva et al. (2014). Assim, tem-se como objetivos específicos a realização de simulações de Monte Carlo para avaliar o desempenho dos estimadores de máxima verossimilhança, além de aplicações a dados reais a fim de comparar os ajustes dos três modelos de regressão.

Neste trabalho são avaliadas três abordagens de modelo de regressão baseados na distribuição BS. Assim, primeiramente, o capítulo 2 abrange a metodologia empregada, sendo definidos os três modelos de regressão avaliados: BS, log-BS e BSR, além de suas

respectivas distribuições. No capítulo 3 são apresentados os resultados de duas simulações via método de Monte Carlo. A primeira tem como objetivo avaliar o desempenho dos estimadores de máxima verossimilhança, comparando-se o viés empírico e erro quadrático médio deles. Já a segunda é realizada com o intuito de avaliar esses modelos quando se assume que os dados gerados são provenientes de outras distribuições assimétricas, sendo então utilizadas distribuições log-simétricas na geração dos dados. Já no capítulo 4 são realizadas duas aplicações a dados reais, em que o objetivo consiste em comparar os ajustes dos modelos de regressão aqui apresentados. Por fim, algumas considerações finais são apresentadas no capítulo 5.

2 Metodologia

Neste capítulo é detalhada a metodologia utilizada no trabalho. Assim, é definida a distribuição BS clássica, além de duas de suas reparametrizações, em que se obtém a distribuição log-BS e BSR. Ainda, são definidos os modelos de regressão BS, BSR e log-BS, além dos resíduos de Cox-Snell Generalizado e do Quantil Aleatorizado e os critérios de seleção de modelos AIC, BIC e o Erro Quadrático Médio.

2.1 Distribuições BS

2.1.1 A distribuição BS clássica

A distribuição BS clássica introduzida por Birnbaum e Saunders (1969) é contínua, unimodal e com assimetria à direita, assim como a maioria das distribuições de tempo de vida. Se uma variável aleatória T segue uma distribuição BS com parâmetro de forma $\alpha > 0$ e de escala $\theta > 0$, utiliza-se a notação $T \sim \text{BS}(\alpha, \theta)$. Por outro lado, quando se utiliza a notação $Z \sim \text{N}(0,1)$, é dito que a variável aleatória Z segue distribuição normal padrão. As variáveis aleatórias $T \sim \text{BS}(\alpha, \theta)$ e $Z \sim \text{N}(0,1)$ são relacionadas por uma função monótona. Assim, qualquer variável aleatória T que siga distribuição BS pode ser obtida por uma transformação de outra variável aleatória Z com distribuição normal padrão. Essa relação da distribuição BS com a normal padrão é um fator que chama a atenção para essa distribuição. Obtém-se então que

$$T = \theta \left(\frac{\alpha Z}{2} + \sqrt{\left(\frac{\alpha Z}{2}\right)^2 + 1} \right)^2,$$

em que Z é uma variável aleatória que segue distribuição normal padrão, de modo que

$$Z = \frac{1}{\alpha} \left(\sqrt{\frac{T}{\theta}} - \sqrt{\frac{\theta}{T}} \right).$$

Seja $T \sim \text{BS}(\alpha, \theta)$, então a função de distribuição acumulada (FDA) de T é dada por:

$$F(t, \alpha, \theta) = \Phi \left[\frac{1}{\alpha} \left(\sqrt{\frac{T}{\theta}} - \sqrt{\frac{\theta}{T}} \right) \right], \tag{2.1}$$

em que Φ é a FDA da normal padrão e α e θ os parâmetros de forma e escala, respectivamente. Ainda, a função densidade de probabilidade (FDP) de T é dada por:

$$f_T(t;\alpha;\theta) = \frac{1}{\sqrt{8\pi}}\left(\exp\left(\frac{1}{\alpha^2}\right)\right)\exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\theta}+\frac{\theta}{t}\right)\right)\frac{t^{-3/2}}{\alpha\theta^{1/2}}(t+\theta), \tag{2.2}$$

para $t>0$, $\alpha>0$ e $\theta>0$.

Pela Equação (2.2) pode-se verificar que se $T \sim \text{BS}(\alpha,\theta)$, então para todo $b>0$ a variável aleatória $Y = bT$ segue distribuição BS com parâmetros α e $b\theta$. Além disso, a variável aleatória $Y = 1/T$ segue a mesma distribuição de T, sendo que o parâmetro θ é substituído por $1/\theta$. Assim, se $T \sim \text{BS}(\alpha,\theta)$, então $bT \sim \text{BS}(\alpha, b\theta)$ para b>0 e $1/T \sim \text{BS}(\alpha, 1/\theta)$.

Uma outra representação de uma variável aleatória que segue distribuição BS é relacionada à sua transformação logarítmica. Dessa forma, se $T \sim \text{BS}(\alpha,\theta)$, então $Y = \log(T) \sim \text{BS}(\alpha, \log(\theta))$, em que log-BS$(\alpha, \log(\theta))$ representa uma distribuição conhecida como log-BS.

Dois indicadores muito úteis em análise de tempo de vida são as funções de risco e de sobrevivência. Assim, essas funções de $T \sim \text{BS}(\alpha,\theta)$ são dadas respectivamente por:

$$R_T(t;\alpha;\theta) = 1 - F_T(t;\alpha;\theta) = \Phi\left[-\frac{1}{\alpha}\left(\sqrt{\frac{t}{\theta}}-\sqrt{\frac{\theta}{t}}\right)\right],\ t>0,$$

e

$$h_T(t;\alpha;\theta) = \frac{f_T(t;\alpha;\theta)}{R_T(t;\alpha;\theta)} = \frac{\frac{1}{\sqrt{8\pi}}\left(\exp\left(\frac{1}{\alpha^2}\right)\right)\exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\theta}+\frac{\theta}{t}\right)\right)\frac{t^{-3/2}}{\alpha\theta^{1/2}}(t+\theta)}{\Phi\left[-\frac{1}{\alpha}\left(\sqrt{\frac{t}{\theta}}-\sqrt{\frac{\theta}{t}}\right)\right]},\ t>0,$$

para $0 < R_T(t;\alpha;\theta) < 1$.

Os coeficientes de curtose e assimetria da distribuição BS são dados respectivamente por:

$$CS(T) = \frac{16\alpha^2(11\alpha^2+6)^2}{(5\alpha^2+4)^3},$$

$$CK(T) = 3 + \frac{6\alpha^2(93\alpha^2+40)}{(5\alpha^2+4)^2}.$$

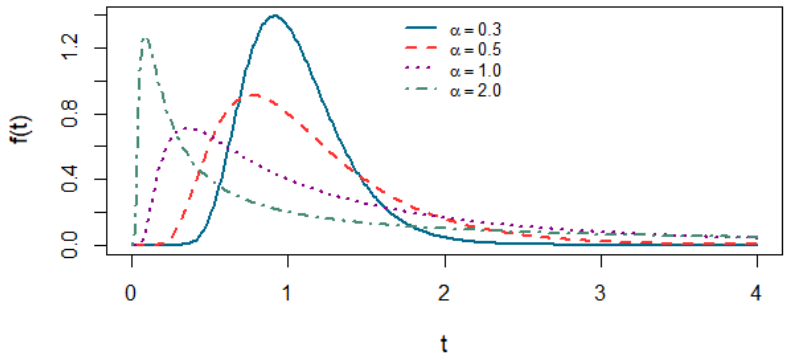
Já a média de $T \sim BS(\alpha,\theta)$ e sua respectiva variância são dadas por:

$$E(T) = \theta\left(1+\frac{\alpha^2}{2}\right);$$

$$Var(T) = \theta^2\alpha^2\left(1+\frac{5}{4}\alpha^2\right).$$

Na Figura 1 se encontra o gráfico da FDP da distribuição BS para diferentes valores do parâmetro de forma α e parâmetro de escala θ igual a 1.

Figura 1 – Gráfico da função densidade de probabilidade da distribuição Birnbaum-Saunders para diferentes valores de α e com $\theta = 1$



Como já mencionado, a distribuição BS é contínua, unimodal e assimétrica à direita. Além disso, observa-se pela figura acima que, à medida que o parâmetro α tende a zero, essa distribuição tende a ser simétrica em torno de θ (mediana). Por outro lado, à medida que o parâmetro de forma α aumenta, a distribuição BS apresenta cauda mais pesada. Dessa forma, o parâmetro α modifica a simetria e curtose da distribuição BS.

2.1.2 A distribuição log-BS

A distribuição log-BS é um caso particular da distribuição seno hiperbólica normal (SHN), a qual foi desenvolvida por Rieck e Nedelman (1991) e pode ser definida por meio de uma transformação de uma distribuição normal padrão. Ou seja, se $Y \sim \text{SHN}(\alpha, \mu, \sigma)$, então

$$Z = \frac{2}{\alpha}\text{senh}\left(\frac{Y-\mu}{\sigma}\right) \sim N(0,1),$$

e

$$Y = \mu + \sigma\text{arcosh}\left(\frac{\alpha Z}{2}\right) \sim SHN(\alpha, \mu, \sigma), \text{ em que } Z \sim N(0,1).$$

Além disso, α é parâmetro de forma, μ de localização e σ de escala. O primeiro e terceiro admitem apenas valores positivos. Já o segundo admite qualquer valor no conjunto dos reais. Então, $Y \sim \text{SHN}(\alpha, \mu, \sigma)$ e sua FDA é dada por:

$$F_y(y;\alpha;\mu;\sigma) = \Phi\left(\frac{2}{\alpha}\text{senh}\left(\frac{y-\mu}{\sigma}\right)\right).$$

A FDP de Y é dada por:

$$f_Y(y;\alpha;\mu;\sigma) = \phi\left(\frac{2}{\alpha}\text{senh}\left(\frac{y-\mu}{\sigma}\right)\right)\frac{2\cosh((y-\mu)/\sigma)}{\alpha\sigma},$$

em que -$\infty < y < \infty$.

Rieck e Nedelman (1991) mostraram que se $T \sim \text{BS}(\alpha, \theta)$, então $\log(T) \sim \text{SHN}(\alpha, \mu, \sigma = 2)$, em que $\mu = \log(\theta)$. Assim, a distribuição SHN também é conhecida como a distribuição log-BS. É interessante ressaltar ainda que a estimação dos parâmetros e geração de números aleatórios da distribuição BS podem ser obtidos com mais eficiência da distribuição log-BS.

2.1.3 A distribuição BS reparametrizada pela média (BSR)

Uma reparametrziação da distribuição BS foi proposta por Santos-Neto et al. (2012) e é indexada pelos parâmetros μ e δ, em que $\mu > 0$, além de ser a média, é o parâmetro de escala e $\delta > 0$ o parâmetro de forma e precisão. Baseado nessa reparametrização, a FDP de Y é dada por:

$$f(y;\mu;\delta) = \frac{exp(\delta/2)\sqrt{\delta+1}}{4y^{3/2}\sqrt{\pi\mu}}\left[y + \frac{\delta\mu}{\delta+1}\right] exp\left(-\frac{\delta}{4}\left[\frac{y[\delta+1]}{\delta\mu} + \frac{\delta\mu}{y[\delta+1]}\right]\right) \ , \ y > 0.$$

Nesse caso utiliza-se a notação $Y \sim \text{BSR}(\mu, \delta)$. A média e variância da variável aleatória Y são dadas por μ e μ^2/ϕ, respectivamente, sendo que $\phi = [\delta+1]^2/[2\delta+5]$.

O parâmetro δ controla a simetria e curtose da distribuição de modo que, à medida que a FDP correspondente é mais concentrada em volta da média, a variabilidade decresce. Além disso, o parâmetro μ modifica a escala da distribuição e, assim, à medida que ele cresce, a variabilidade também cresce.

2.2 Modelagem baseada nas distribuições BS

Nesta seção é feita a descrição de três abordagens de modelos de regressão BS, baseados nas distribuições descritas na seção anterior. Essas três abordagens consistem nos modelos BS, log-BS e BSR.

2.2.1 Modelo de regressão BS

Suponha $T \sim \text{BS}(\alpha, \theta)$, sendo α e θ os parâmetros de forma e escala, respectivamente. Ainda, suponha p covariáveis $\mathbf{x} = (1, x_1, ..., x_p)^T$. Assumindo uma função de ligação log-linear para o parâmetro de escala θ em (2.1), tem-se que:

$$F(t, \alpha, \theta) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\theta}} - \sqrt{\frac{\theta}{t}}\right)\right] \ ,$$

em que $h(\theta) = \beta' x$ corresponde à função de ligação. Quando se tem o link $\log\ \theta = e^{\beta' x}$, sendo o modelo de regressão equivalente a 2.6. Além disso, $\beta' x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.

A FDP correspondente é:

$$f(t,\alpha,\theta)=\frac{1}{2\sqrt{2\pi}\alpha\theta}\left[\left(\frac{\theta}{t}\right)^{\frac{1}{2}}+\left(\frac{\theta}{t}\right)^{\frac{3}{2}}\right]\exp\left[-\frac{1}{2\alpha^2}\left(\frac{\theta}{t}+\frac{t}{\theta}-2\right)\right]\ , \tag{2.3}$$

e a função de risco é dada por:

$$h(t,\alpha,\theta)=\frac{f(t,\alpha,\theta)}{1-F(t,\alpha,\theta)}\ ,t>0,\ \alpha>0,\ \theta>0.$$

Suponha que se observe $n(\geq p+2)$ tempos de falha, denotados por $(t_1,t_2,...,t_n)$ com as covariáveis correspondendo a t_i como $\mathbf{x_i}=(1,x_{1i},x_{2i},...,x_{pi})$. Supondo o parâmetro de escala $\theta_i=\exp[\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+...+\beta_px_{pi}]$, tem-se que a função de verossimilhança é:

$$L=\prod_{i=1}^{n}f_T(t_i,\alpha,\theta_i)\ , \tag{2.4}$$

em que f é a FDP da distribuição BS dada em (2.3). Por (2.4) obtém-se que a função de log-verossimilhança é então dada por:

$$log\ L=-n\log\ \alpha+\sum_{i=1}^{n}log\left[\left(\frac{t_i}{\theta_i}\right)^{\frac{1}{2}}\left(\frac{\theta_i}{t_i}\right)^{\frac{1}{2}}\right]-\frac{1}{2\alpha^2}\sum_{i=1}^{n}\left(\frac{t_i}{\theta_i}+\frac{\theta_i}{t_i}-2\right)\ . \tag{2.5}$$

Dado $\beta_0,\beta_1,...,\beta_p$, o estimador de máxima verossimilhança para α é dado por:

$$\hat{\alpha}=\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i}{\theta_i}+\frac{\theta_i}{t_i}-2\right)\right]^{\frac{1}{2}}\ .$$

Os estimadores de máxima verossimilhança de $\beta_0,\beta_1,...,\beta_p$ podem ser obtidos utilizando um método de otimização não-linear.

2.2.2 Modelo de regressão log-BS

O segundo modelo de regressão BS é baseado na transformação logarítmica das observações, podendo ser bastante útil para dados que seguem distribuição SHN ou suas generalizações, sem que nenhuma transformação na escala seja necessária.

O modelo de regressão BS é dado por:

$$T_i=\theta_i\varphi_i=\exp(\mu_i)\varphi_i=\exp(\mathbf{x_i}^\top\eta)\varphi_i\ ,\quad i=1,2,...,n. \tag{2.6}$$

em que T_i é a variável resposta e $\theta_i=\exp(\mu_i)$ a mediana, ambos para o caso i. Além disso, $\varphi_i\sim\text{BS}(\alpha,1)$ é o erro do modelo de regressão para $i=1,2,...,n$.

Considerando o modelo anterior tal que $T_1, T_2, ..., T_n$ é uma amostra de tamanho n de $T \sim \text{BS}(\alpha, \theta_i)$, então $Y_1 = \log(T_1)$, $Y_2 = \log(T_2)$, ..., $Y_n = \log(T_n)$ pode ser considerada uma amostra de tamanho n de $Y \sim \text{log-BS}(\alpha, \mu_i = \log(\theta_i))$. Assim, ao aplicar o logarítmo em (2.6), obtém-se:

$$Y_i = \mu_i + \epsilon_i = \mathbf{x_i}^\top \eta + \epsilon_i \ , \quad i = 1, 2, ..., n.$$

em que $Y_i = \log(T_i)$ é a resposta logarítmica para o caso i e $\epsilon_i = \log(\varphi_i) \sim log\text{-BS}(\alpha, 0)$ é o erro do modelo de regressão para $i = 1, 2, ...n$. Além disso, η e x_i são os mesmos já especificados acima em (2.6).

Uma das propriedades da distribuição BS citadas anteriormente elenca que $bT \sim \text{BS}(\alpha, b\theta)$ para b> 0. Assim, por meio dessa propriedade para a distribuição BS, $T_i \sim \text{BS}(\alpha, \theta_i)$. Além disso, $\eta = (\eta_0, \eta_1, ..., \eta_p)^\top$ é um vetor $(p+1)\times 1$ de parâmetros desconhecidos a serem estimados e $\mathbf{x_i} = (1, x_{i1}, ..., x_{ip})$ contém os valores de p variáveis explicativas.

A função de log-verossimilhança de $\gamma = (\alpha, \eta^\top)$ é dada por:

$$l(\gamma) = c_3 + \sum_{i=1}^{n} \log(\xi_{i1}) - \frac{1}{2}\sum_{i=1}^{n} \xi_{i2}^2,$$

em que c_3 é uma constante que não depende de θ. Ainda,

$$\xi_{i1} = \frac{2}{\alpha}\cosh\left(\frac{y_i - \mu_i}{2}\right) \quad \xi_{i2} = \frac{2}{\alpha}\text{senh}\left(\frac{y_i - \mu_i}{2}\right),$$

sendo que $\mu_i = \mathbf{x_i}^\top \eta$ para i=1,2...,n. As estimativas de máxima verossimilhança são então obtidas ao derivar as funções de log-verossimilhança em relação a α e η_j e igualá-las a zero. Porém, não há solução analítica para isso e, portanto, métodos iterativos para a solução de otimizações não lineares são necessários.

2.2.3 Modelo de regressão BSR

Pode-se definir um modelo de regressão baseado na FDP de Y e sua reparametrização por meio do componente a seguir:

$$h(\mu_i) = \eta_i = x_i^\top \beta \ , \quad i = 1, ..., n.$$

em que $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$ para $p < n$ é vetor de parâmetros a ser estimado e $\mathbf{x_i} = (1, x_{i1}, ..., x_{ip})$ representa os valores dos p regressores. Além disso, $\mu_i = h^{-1}(\mathbf{x_i}^\top \beta)$, sendo que a função de ligação h é monótona, positiva e diferenciável pelo menos duas vezes.

A variância de Y_i é uma função de μ_i. Assim, apesar de se estar modelando pela média, está também modelando pela variância pelo simples fato de $Var(Y_i) = \frac{\mu^2}{\phi}$.

A função de log-verossimilhança do modelo proposto em (2.6) para $\lambda = (\beta^\top, \delta)^\top$ é dada por:

$$l(\lambda; y) = l(\lambda) = \sum_{i=1}^{n} l_i(\mu_i, \delta, y_i),$$

em que

$$l_i(\mu_i, \delta, y_i) = l_i(\mu_i, \delta) = \frac{\delta}{2} - \frac{\log(16\pi)}{2} - \frac{1}{2}\log\left(\frac{[\delta+1]y_i^3\mu_i}{[\delta y_i + y_i + \delta\mu_i]^2}\right) - \frac{y_i[\delta+1]}{4\mu_i} - \frac{\delta^2\mu_i}{4[\delta+1]y_i}.$$

Para estimar os parâmetros do modelo pelo método de máxima verossimilhança deve-se resolver $l'_\theta = 0$. Porém, não há forma fechada para essas estimativas e, portanto, deve-se utilizar métodos iterativos para otimização não linear, como os algoritmos de Newton ou Quasi-Newton e Scoring de Fisher.

2.3 Critério de informação de Akaike e Bayesiano

Os critérios de informação fornecem uma base para a seleção de um modelo dentre uma lista de vários. A seleção de modelos utilizando os critérios de informação foi desenvolvida com o objetivo de resumir as evidências fornecidas pelos dados a favor de um modelo.

Enquanto diversos critérios para a comparação de modelos de regressão já foram desenvolvidos, neste trabalho são utilizados o critério de informação de Akaike (AIC) e o Bayesiano de Schwarz (BIC), ambos alternativas muito populares quando se tem que escolher entre modelos de regressão. Os dois critérios penalizam a adição de preditores no modelo. O AIC é dado por:

$$AIC = -2\log(L) + 2h \ ,$$

em que L consiste na verossimilhança e h é o número de parâmetros no modelo.

O critério BIC difere do Akaike apenas no fato do segundo termo depender do tamanho da amostra, sendo então dado por:

$$BIC = -2\log(L) + h\log(n),$$

em que n o tamanho amostral.

Ambos os critérios penalizam a adição de parâmetros no modelo, sendo a única diferença entre eles a intensidade dessa penalização. Assim, o critério Bayesiano de Schwartz penaliza mais modelos mais complexos se comparado ao critério AIC. O modelo com menor valor para esses critérios deve ser selecionado.

2.4 Análise de resíduos

Um dos problemas mais complexos em estatística paramétrica é a identificação da distribuição mais adequada aos dados. Assim, alguns métodos para verificar a qualidade do ajuste de um modelo devem ser utilizados.

Os resíduos e, mais especificamente, os gráficos de resíduos, também desempenham um papel fundamental na verificação de um modelo estatístico. Assim, eles são uma ferramenta muito utilizada para verificar a qualidade do ajuste de um modelo de regressão. Uma gama de resíduos já foi proposta devido à dificuldade de escolha de um resíduo ao se modelarem dados de sobrevivência - a facilidade não é a mesma como para modelos lineares generalizados. Dentre eles se encontram o resíduo do Quantil Aleatorizado e de Cox-Snell, os quais são abordados neste trabalho.

Os resíduos de Cox-Snell são um tipo de resíduo normalizado utilizado em análise de sobrevivência. Eles levam em consideração a distribuição e a estimação dos parâmetros do modelo de regressão. Assim, o resíduo Cox-Snell Generalizado (GCS) é dado por:

$$r_i^{GCS} = -\log(\hat{S}(x_i)) \ , \ i = 1, ..., n.$$

em que $\hat{S}$ é a função de sobrevivência ajustada. A função de sobrevivência para o modelo BS reparametrizado pela média é dada por:

$$S(y;\mu;\delta) = \Phi\left(-\sqrt{\frac{\delta}{2}}\left[\sqrt{\frac{(\delta+1)y}{\mu\delta}} - \sqrt{\frac{\mu\delta}{(\delta+1)y}}\right]\right) \ , \ y > 0, \ \mu > 0, \ \delta > 0,$$

enquanto que para o modelo BS reparametrizado pela mediana ela é dada por:

$$S(t;\alpha;\theta) = \Phi\left(-\frac{1}{\alpha}\left[\sqrt{\frac{t}{\theta}} - \sqrt{\frac{\theta}{t}}\right]\right) \ , \ t > 0, \ \alpha > 0, \ \theta > 0.$$

Já a função de sobrevivência para o modelo log-BS é dado por:

$$S(y,\alpha,\mu,\sigma) = \Phi\left(-\frac{2}{\alpha}\text{senh}\left(\frac{y-\mu}{\sigma}\right)\right) \ , \ y \ \epsilon \ \mathbb{R}$$

Um critério de importância para esse tipo de resíduo é que, sendo o modelo corretamente especificado, o resíduo Cox-Snell generalizado será distribuído conforme uma distribuição exponencial uniária (Exp(1)). Utilizando-se desse critério, os resíduos de Cox-Snell podem ser avaliados graficamente, ou formalmente por meio de um teste de adequabilidade do modelo.

Já o segundo tipo de resíduo considerado é o Quantil Aleatorizado, usualmente empregado em modelos generalizados aditivos para locação, escala e forma. Apesar da falta de aplicações relacionadas aos resíduos do Quantil Aleatório, ele também foi desenvolvido

a fim de contornar os problemas dos resíduos usuais e é possível verificar sua qualidade em diversos artigos, como em Saulo et al. (2017). O resíduo Quantil Aleatorizado é definido por:

$$r_i^{RQ} = \Phi^{-1}(\hat{S}(x_i)) \ , \ i = 1, ..., n.$$

em que Φ^{-1} é a função inversa da FDA da normal padrão e $\hat{S}$ é a função de sobrevivência ajustada. O resíduo RQ segue distribuição normal padrão quando o modelo é especificado corretamente.

3 Simulação de Monte Carlo

O objetivo geral deste trabalho consiste em avaliar os modelos de regressão baseados na distribuição BS. Para isso, são estudados e comparados três diferentes modelos de regressão para a distribuição BS. Dessa forma, são realizadas simulações de Monte Carlo para avaliar os diferentes modelos e seus estimadores de máxima verossimilhança, vises e erros quadráticos médios.

O viés de um estimador H consiste na diferença entre seu valor esperado e o valor do parâmetro θ que está sendo estimado. Assim, ele é dado por:

$$viés(H, \theta) = E(H) - \theta$$

Um estimador com viés igual a zero é dito não viesado e satisfaz $E_\theta(H) = \theta$ para todo θ. Um viés pequeno reflete o desejo de, se um experimento for repetido indefinidamente, então a média das estimativas resultantes vai ser próxima ao verdadeiro valor do parâmetro que está sendo estimado.

O erro quadrático médio (EQM) ou *mean squared error* (MSE) consiste em um dos critérios básicos para a avaliação de estimadores. Ele mede a diferença quadrática média entre o estimador H e o parâmetro θ, sendo uma medida de desempenho razoável quando se trata de um estimador pontual. O MSE possui, ao menos, duas vantagens em relação a outras medidas de distância, sendo elas: ser bastante tratável analiticamente e possuir uma interpretação. Dessa forma, ele é dado por:

$$EQM(H, \theta) = E_\theta[(H - \theta)^2] \tag{3.1}$$

Como se pode ver por 3.1, o EQM incorpora um componente que mede a variabilidade do estimador (precisão) e outro que mede seu viés (exatidão). Um estimador com boas propriedades do EQM apresenta, simultaneamente, pequena variância e pequeno viés. A fim de se encontrar um estimador com boas propriedades do erro quadrático médio, é necessário obter estimadores que controlam a variância e o viés.

Estimadores não viesados e que apresentam erro padrão mínimo são sempre almejados. Porém, às vezes, esses objetivos são incompatíveis. Embora estimadores não viesados possam ser também razoáveis, do ponto de vista do erro quadrático médio, controlar vieses não garante que o EQM seja também controlado.

3.1 Simulação 1

Com o objetivo de avaliar o desempenho dos estimadores de máxima verossimilhança dos três modelos de regressão baseados na distribuição Birnbaum-Saunders, são gerados dados de cada um desses modelos, em que se assume que as covariáveis se distribuem conforme uma Uniforme(0,1). Isso é feito por meio de simulações de Monte Carlo, em que se considera 0,2 e 1,5 como valores reais para os parâmetros β_0 e β_1, respectivamente. Além disso, esse estudo por simulação considera três tamanhos de amostra distintos, sendo eles 50, 100 e 200, além de três valores para o parâmetro de forma: 0,5, 1,0 e 2,0 com 500 réplicas para cada tamanho amostral.

É utilizado o link log para os modelos reparametrizados pela média e mediana (BS e BSR). Os resultados para as estimativas por máxima verossimilhança estão dispostos nas Tabelas 1 (modelo BSR), 2 (modelo log-BS) e 3 (modelo BS). Nelas estão incluídos o viés empírico dos estimadores, além de seus respectivos erros quadráticos médios (EQM), que estão entre parênteses.

Tabela 1 – Viés empírico e erro quadrático médio (entre parênteses) provenientes dos dados simulados para os estimadores de máxima verossimilhança indicados para os parâmetros do modelo BSR e tamanho amostral n.

n	$\boldsymbol{\delta}$	$\hat{\boldsymbol{\delta}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$
50	0,5	0,0571 (0,0168)	-0,0621 (0,1479)	0,0006 (0,3855)
	1,0	0,1111 (0,0655)	-0,0469 (0,0983)	0,0065 (0,2626)
	2,0	0,2149 (0,2557)	-0,0324 (0,0577)	0,0097 (0,1579)
100	0,5	0,0240 (0,0060)	-0,0053 (0,0722)	-0,0465 (0,1887)
	1,0	0,0470 (0,0237)	-0,0015 (0,0497)	-0,0387 (0,1324)
	2,0	0,0910 (0,0938)	0,0014 (0,0301)	-0,0302 (0,0820)
200	0,5	0,0117 (0,0028)	-0,0115 (0,0385)	-0,0019 (0,1003)
	1,0	0,0230 (0,0113)	-0,0089 (0,0266)	-0,0015 (0,0713)
	2,0	0,0447 (0,0449)	-0,0063 (0,0161)	-0,0012 (0,0443)

Tabela 2 – Viés empírico e erro quadrático médio (entre parênteses) provenientes dos dados simulados para os estimadores de máxima verossimilhança indicados para os parâmetros do modelo log-BS e tamanho amostral n.

n	$\boldsymbol{\alpha}$	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$
	0,5	-0,0151 (0,0028)	0,0032 (0,0165)	-0,0090 (0,0462)
50	1,0	-0,0322 (0,0114)	0,0041 (0,0550)	-0,0124 (0,1529)
	2,0	-0,0693 (0,0473)	0,0007 (0,1290)	-0,0061 (0,3563)
	0,5	-0,0076 (0,0012)	0,0032 (0,0087)	-0,0043 (0,0250)
100	1,0	-0,0163 (0,0048)	0,0070 (0,0290)	-0,0102 (0,0832)
	2,0	-0,0351 (0,0195)	0,0131 (0,0666)	-0,0208 (0,1913)
	0,5	-0,0038 (0,0006)	0,0040 (0,0044)	-0,0073 (0,0134)
200	1,0	-0,0081 (0,0025)	0,0066 (0,0148)	-0,0129 (0,0444)
	2,0	-0,0175 (0,0100)	0,0075 (0,0347)	-0,0175 (0,1006)

Tabela 3 – Viés empírico e erro quadrático médio (entre parênteses) provenientes dos dados simulados para os estimadores de máxima verossimilhança indicados para os parâmetros do modelo BS e tamanho amostral n.

n	$\boldsymbol{\alpha}$	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$
	0,5	-0,0149 (0,0026)	0,0033 (0,0150)	-0,0081 (0,0439)
50	1,0	-0,0317 (0,0104)	0,0072 (0,0503)	-0,0152 (0,1449)
	2,0	-0,0685 (0,0431)	0,0126 (0,1176)	-0,0225 (0,3310)
	0,5	-0,0082 (0,0012)	0,0025 (0,0085)	0,0041 (0,0237)
100	1,0	-0,0174 (0,0049)	0,0054 (0,0280)	0,0057 (0,0793)
	2,0	-0,0372 (0,0200)	0,0102 (0,0622)	0,0018 (0,1814)
	0,5	-0,0055 (0,0006)	-0,0012 (0,0052)	0,0036 (0,0147)
200	1,0	-0,0115 (0,0025)	-0,0020 (0,0173)	0,0060 (0,0495)
	2,0	-0,0242 (0,0103)	-0,0025 (0,0385)	0,0071 (0,1124)

Por meio dos resultados dispostos nas Tabelas 1, 2 e 3, é possível observar que, em geral, à medida que o tamanho amostral n cresce, o viés e o EQM diminuem, o que é de se esperar. Isso acontece para os três modelos. Além disso, nota-se que, à medida que o valor do parâmetro δ, no caso do modelo BSR, e α, para os modelos BS e log-BS, aumenta, a perfomance do estimador desse parâmetro piora.

Comparando-se os resultados para as três abordagens distintas de modelos, os resultados obtidos pelos modelos BS e log-BS para o estimador de α são melhores que aqueles obtidos pelo modelo BSR. Já em relação a $\hat{\beta}_0$ e $\hat{\beta}_1$, é importante verificar que nem sempre o modelo com menor viés é também aquele com menor valor para o erro quadrático médio. Assim, por exemplo, para β_0, $\alpha = 0,5$ e tamanho amostral 50, o modelo log-BS apresenta menor viés se comparado aos demais modelos. Porém, nas mesmas circunstâncias, o modelo BS apresenta menor EQM.

3.2 Simulação 2

Uma segunda simulação é realizada com o intuito de avaliar as três abordagens de modelos de regressão quando se assume que os dados gerados são provenientes de outras distribuições assimétricas. Assim, distribuições log-simétricas são utilizadas na geração dos dados.

A classe de distribuições simétricas é popular nas áreas de estudo mais variadas. Porém, sua estrutura não permite a representação de variáveis estritamente positivas que seguem uma distribuição assimétrica. Com isso, versões logarítmicas de distribuições simétricas foram desenvolvidas. Por exemplo, se $T = \exp(Y)$, então T pertence à classe de distribuições log-simétricas, podendo ser representada por $T \sim \text{LS}(\eta, \phi, g)$. Assim, η=exp(μ) e ϕ são os parâmetros de escala e forma, respectivamente, para algum kernel g. A FDP de T é dada por:

$$f_T(t, \eta, \phi) = \frac{1}{\phi_z} g\left(\log\left(\left(\frac{t}{\eta}\right)^{1/\phi}\right)^2\right) ,$$

em que $t > 0$, $\eta > 0$, $\phi > 0$ e g leva à distribuições log-simétricas variadas.

As distribuições log-simétricas aqui utilizadas são: log-normal, log-hiperbólica, log-slash e log-contaminada-normal. Essa é uma maneira de avaliar e comparar o desempenho dos estimadores de máxima verossimilhança quando o processo gerador dos dados não vem dos próprios modelos, diferentemente do que é feito na simulação anterior. Além disso, avalia-se a capacidade preditiva e os critérios AIC, BIC e EQM. São realizadas 500 réplicas de Monte Carlo e são considerados os tamanhos amostrais 50, 100 e 200. Assume-se que as covariáveis são distribuídas conforme uma Uniforme(0,1) e o modelo é dado por:

$$\eta_i = 2,5 + 3x_{1_i} + 0,9x_{2_i} .$$

Assim, os valores dos parâmetros β_0, β_1 e β_2 são fixados como sendo iguais a 1,5, 3,0 e 0,9, respectivamente. Além disso, $\phi = 3$.

Nas Tabelas 4, 5, 6 e 7 estão expostos os valores para os critérios AIC e BIC de cada um dos modelos quando é utilizada uma distribuição log-simétrica na geração dos dados, sendo que os valores expostos para esses critérios correspondem ao valor médio obtido pelas réplicas de Monte Carlo. Além disso, é apresentado o viés obtido da estimação dos parâmetros dos modelos, além de seus respectivos erros quadráticos médios e esse erro para o modelo. Os resultados para o EQM de predição de cada modelo deve ser multiplicado por $10e8$ para se obter os verdadeiros valores.

Tabela 4 – Resultados obtidos para o viés empírico e EQM para os estimadores de máxima verossimilhança para os parâmetros de cada modelo e tamanho amostral n, quando é utilizada a distribuição log-normal na geração dos dados, além do EQM do modelo, AIC e BIC

n	**Distribuição**	β_0	β_1	β_2	**EQM**	**AIC**	**BIC**
	BS	-0,5461 (4,0764)	0,0201 (0,1158)	0,0205 (0,1188)	0,001	458,21	465,86
50	log-BS	-0,0008 (0,1136)	0,0148 (0,1123)	0,0144 (0,1166)	0,001	203,61	211,25
	BSR	1,3722 (2,3550)	-0,0040 (0,1122)	0,0071 (0,1196)	0,0178	453,28	460,93
	BS	-0,1167 (0,9517)	0,0188 (0,0675)	-0,0140 (0,0682)	0,0008	911,70	922,12
100	log-BS	0,0145 (0,0653)	0,0189 (0,0665)	-0,0157 (0,0678)	0,0008	409,08	419,50
	BSR	1,3828 (2,0366)	0,0115 (0,0673)	-0,0184 (0,0685)	0,0008	909,58	920,00
	BS	-0,1134 (0,8817)	0,0230 (0,0436)	-0,0140 (0,0419)	0,0005	1.825,38	1.838,58
200	log-BS	0,0021 (0,0400)	0,0221 (0,0434)	-0,0156 (0,0418)	0,0005	822,52	835,71
	BSR	1,4163 (2,0823)	0,0187 (0,0427)	-0,0153 (0,0417)	0,0005	1.822,44	1.835,63

Quando o processo gerador dos dados é proveniente da distribuição log-normal, pode-se observar que, para todos os modelos, à medida que o tamanho amostral n cresce, o erro quadrático dos estimadores diminui. Já o viés não segue um padrão, algumas vezes diminui e, em outros casos, aumenta. Além disso, em relação ao EQM de cada modelo, os três apresentam resultados iguais para tamanhos amostrais iguais a 100 e 200, diferindo somente quando n = 50, em que os modelos BS e log-BS apresentam o menor valor.

No que diz respeito aos valores obtidos para os critérios AIC e BIC dos três modelos, pode-se observar que seus valores aumentam à medida que o tamanho amostral cresce. Além disso, dado que se deseja valores menores para esses critérios, o modelo log-BS é o melhor.

Em relação ao melhor modelo segundo o viés e EQM dos estimadores, nenhum modelo se destaca apresentando melhor performance para todos os seus estimadores. O modelo log-BS apresenta resultados melhores segundo esses dois aspectos quando se observa β_0. Já em relação a β_1, o modelo BSR apenas não apresenta o melhor resultado quando se observa o EQM desse estimador para $n = 100$. Por fim, em relação a β_2, não há um modelo que apresenta melhores resultados em relação ao erro quadrático médio desse estimador ou viés.

Tabela 5 – Resultados obtidos para o viés empírico e EQM para os estimadores de máxima verossimilhança para os parâmetros de cada modelo e tamanho amostral n, quando é utilizada a distribuição log-hiperbólica na geração dos dados, além do EQM do modelo, AIC e BIC

n	**Distribuição**	β_0	β_1	β_2	**EQM**	**AIC**	**BIC**
	BS	-0,9438 (10,9884)	0,0064 (1,5217)	-0,0608 (1,8720)	663,8024	550,66	558,31
50	log-BS	0,0574 (3,1711)	0,0774 (1,1885)	0,0068 (1,2125)	0,3239	286,32	293,97
	BSR	3,9856 (21,1643)	-0,3163 (1,1055)	-0,1547 (1,2384)	260,2446	541,37	549,02
	BS	-0,6527 (8,6007)	-0,0236 (1,0055)	-0,0270 (1,4019)	266,8299	1.134,11	1,144,53
100	log-BS	0,0244 (3,1718)	0,0164 (0,9442)	-0,0273 (1,1822)	266,8244	615,26	625,68
	BSR	4,3112 (23,2541)	-0,2407 (0,9062)	-0,1285 (1,1140)	2969,962	1.119,01	1.129,44
	BS	-0,2828 (11,6529)	0,0752 (1,1308)	0,0081 (1,1236)	24,2277	2.386,14	2.399,33
200	log-BS	0,0497 (2,8555)	0,0862 (1,0196)	0,0228 (1,1203)	6,4348	1.341,32	1.354,51
	BSR	5,0310 (30,5182)	-0,2597 (0,9167)	-0,0993 (1,0268)	75,1397	2.358,43	2.371,62

Quando se tem dados provenientes da distribuição log-hiperbólica, pode-se observar que o erro quadrático médio dos estimadores e seus vieses nem sempre diminuem à medida que o tamanho amostral cresce, o que não é esperado. Em relação ao EQM de cada modelo, AIC e BIC, pode-se observar que esses valores aumentam à medida que o tamanho amostral também aumenta. Além disso, ao comparar os três modelos, o log-BS apresenta resultado melhor para todos esses critérios.

Em relação ao viés dos estimadores de máxima verossimilhança, o modelo log-BS apresenta melhor resultado na estimação de β_0. Em relação aos demais parâmetros, os modelos BS e log-BS mostram resultados melhores que o modelo BSR. Já em relação ao erro quadrático médio dos estimadores, o modelo log-BS também apresenta melhores resultados na estimação de β_0 e, em relação aos demais parâmetros, o modelo BSR é melhor para todos os tamanhos amostrais considerados, com exceção de $n = 50$ para o parâmetro β_2.

Dessa forma, o modelo log-BS apresenta melhor performance de seu estimador de máxima verossimilhança de β_0. Em relação à performance dos estimadores dos demais parâmetros, não há um modelo que se destaque dentre os três aqui abordados.

Tabela 6 – Resultados obtidos para o viés empírico e EQM para os estimadores de máxima verossimilhança para os parâmetros de cada modelo e tamanho amostral n, quando é utilizada a distribuição log-slash na geração dos dados, além do EQM do modelo, AIC e BIC

n	**Distribuição**	β_0	β_1	β_2	**EQM**	**AIC**	**BIC**
	BS	-0.7937 (6,4317)	0,0457 (0,2165)	-0,0162 (0,2327)	0,0055	477,04	484,69
50	log-BS	-0.0071 (0,3056)	0,0344 (0,2046)	-0,0245 (0,2396)	0,0054	220,92	228,57
	BSR	1,8551 (4,7396)	-0,0343 (0,2035)	-0,0505 (0,2211)	0,2191	471,19	478,84
	BS	-0,3679 (3,0633)	0,0186 (0,1734)	0,0331 (0,1760)	0,0025	964,65	975,07
100	log-BS	0,0092 (0,2782)	0,0126 (0,1754)	0,0435 (0,1858)	0,0024	456,92	467,34
	BSR	1,9486 (4,2864)	-0,0169 (0,1710)	0,0250 (0,1804)	0,0079	958,24	968,66
	BS	-0,1191 (1,1924)	-0,0163 (0,1065)	-0,0071 (0,1173)	0,0044	1.929,53	1.942,73
200	log-BS	0,0026 (0,1854)	-0,0177 (0,1043)	-0,0070 (0,1167)	0,0044	925,62	938,81
	BSR	2,0191 (4,5717)	-0,0319 (0,0985)	-0,0130 (0,1195)	0,0087	1.927,62	1.940,81

Quando a geração dos dados é proveniente da distribuição log-slash, assim como ocorre com a log-hiperbólica, pode-se observar que o erro quadrático médio dos estimadores de máxima verossimilhança e viés nem sempre diminuem à medida que o tamanho amostral cresce. Em relação ao EQM de cada modelo, os modelos BS e log-BS apresentam resultados semelhantes e melhores que aqueles obtidos para o modelo BSR. Já em relação aos critérios AIC e BIC, os modelos BS e BSR apresentam resultados semelhantes e o log-BS apresenta resultados melhores para todos os tamanhos amostrais, se comparado aos demais modelos.

Em relação à estimação dos parâmetros, o modelo log-BS apresenta menor viés e EQM na estimação de β_0 para todo n. Já em relação à estimação de β_1, o modelo BSR apresenta menor erro quadrático médio para todos os tamanhos amostrais. Tratando-se da estimação de β_2, não há nenhum padrão perceptível.

Tabela 7 – Resultados obtidos para o viés empírico e EQM para os estimadores de máxima verossimilhança para os parâmetros de cada modelo e tamanho amostral n, quando é utilizada a distribuição log-contaminada-normal na geração dos dados, além do EQM do modelo, AIC e BIC

n	**Distribuição**	β_0	β_1	β_2	**EQM**	**AIC**	**BIC**
	BS	-0,7056 (5,9455)	0,0085 (0,2456)	0,0656 (0,2989)	0,0052	488,07	495,72
50	log-BS	0,0284 (0,3369)	0,0124 (0,2499)	0,0536 (0,2887)	0,0051	230,09	237,74
	BSR	2,0710 (5,4374)	-0,0722 (0,2293)	0,0264 (0,2882)	0,2196	481,99	489,64
	BS	-0,2968 (2,4119)	0,0143 (0,1966)	-0,0372 (0,2024)	0,0025	976,18	986,60
100	log-BS	-0,0074 (0,2339)	0,0103 (0,1928)	-0,0405 (0,2006)	0,0024	472,18	482,60
	BSR	2,1236 (5,0296)	-0,0152 (0,1778)	-0,0505 (0,1996)	0,0061	971,14	981,56
	BS	-0,2210 (2,0608)	0,0210 (0,1547)	-0,0049 (0,1378)	0,0186	1.974,80	1.987,99
200	log-BS	0,0107 (0,1700)	0,0167 (0,1497)	-0,0063 (0,1374)	0,0186	967,27	980,46
	BSR	2,2889 (5,8350)	0,0024 (0,1429)	-0,0081 (0,1363)	0,042	1.967,64	1.980,83

Por último, é feita a mesma análise quando os dados são gerados da distribuição log-contaminada-normal. Ao se observar o erro quadrático médio e viés dos estimadores de máxima verossimilhança de cada modelo, verifica-se que eles nem sempre diminuem à medida que aumenta o tamanho amostral. Em relação ao EQM dos modelos, observam-se resultados melhores para o modelo log-BS e BS, ambos apresentando resultados semelhantes. Já em relação ao AIC e BIC, o modelo log-BS obtém menores valores, sendo que os modelos BSR e BS apresentam valores semelhantes, mas piores que os do modelo log-BS.

Em relação aos estimadores de máxima verossimilhança, o modelo log-BS apresenta resultados melhores que os obtidos pelos demais modelos em relação à estimação de β_0. Porém, isso não ocorre na estimação de todos os parâmetros. Em relação aos estimadores de máxima verossimilhança dos demais, em geral, o modelo BSR apresenta menor erro quadrático médio. Por outro lado, não se observa um padrão em relação ao viés desses estimadores.

Por meio das informações dispostas nas tabelas 4, 5, 6 e 7, é possível observar que o EQM do modelo é sempre menor quando o processo gerador dos dados é proveniente de um modelo log-normal. Em seguida, os modelos log-slash e log-contaminada-normal apresentam resultados bastante semelhantes no que diz respeito ao EQM dos modelos. Quando os dados vêm da distribuição log-hiperbólica os EQM dos estimadores de β_0 são bem altos, além do resultado do EQM do modelo ajustado. Ao comparar os três modelos (BS, log-BS e BSR), nota-se que o EQM obtido para os modelos BS e log-BS é semelhante e, em geral, menor que aquele obtido para o modelo BSR.

Em relação à estimação dos parâmetros, observa-se que, em geral, os três modelos apresentam boa performance dos estimadores de máxima verossimilhança quando o processo gerador dos dados é proveniente das distribuições log-simétricas aqui trabalhadas. Ainda, o modelo log-BS apresenta uma performance melhor do estimador de β_0 se comparado aos demais modelos. Já em relação aos demais estimadores, nenhum padrão pode ser definido.

4 Aplicação a dados reais

A distribuição BS possui diversas aplicações em vários contextos diferentes que podem ser encontradas em Johnson, Kotz e Balakrishnan (1995). Algumas aplicações dessa distribuição publicados fora da área de fadiga de materiais são: análise das características da chuva da cidade de Hiroshima - Mills (1997), Seto Iwase Kosei (1993), Seto Iwase Kosei (1995) e Johnson, Kotz e Balakrishnan (1995); risco de contaminação em rios, lagos e reservatórios pelo homem e atividades agrícolas, devido ao acúmulo de nutrientes vegetais ao longo do tempo - Leiva, Sanhueza e Angulo (2009) e Vilca et al. (2010); qualidade do ar devido à acumulação de poluentes na atmosfera ao longo do tempo - Ferreira, Gomes e Leiva (2012) e Marchant et al. (2013b).

Em geral, as aplicações da distribuição BS não são limitadas apenas a essas áreas mencionadas. Recentemente, seu uso nas áreas de *business*, economia, finanças, indústria, seguros, nutrição, psicologia e controle de qualidade também é considerado, entre outros. Assim, com o propósito de comparar os ajustes dos modelos de regressão apresentados aqui, duas aplicações a dados reais são realizadas: a primeira relacionada a vendas atuais de produtos de consumo e a segunda relacionada à fadiga de materiais.

4.1 Aplicação 1 - Vendas atuais de produtos de consumo

4.1.1 Análise Exploratória

De início, é utilizado um conjunto de dados real obtido do pacote $faraway$ do $software$ estatístico R. Esses dados estão relacionados a vendas projetadas (variável explicativa) e vendas atuais (variável resposta) de 20 produtos de consumo. A fim de se conhecer o banco de dados e antes de análises mais aprofundadas, na Tabela 8 é apresentado um resumo de certas medidas descritivas das vendas atuais.

Tabela 8 – Medidas descritivas para as vendas atuais

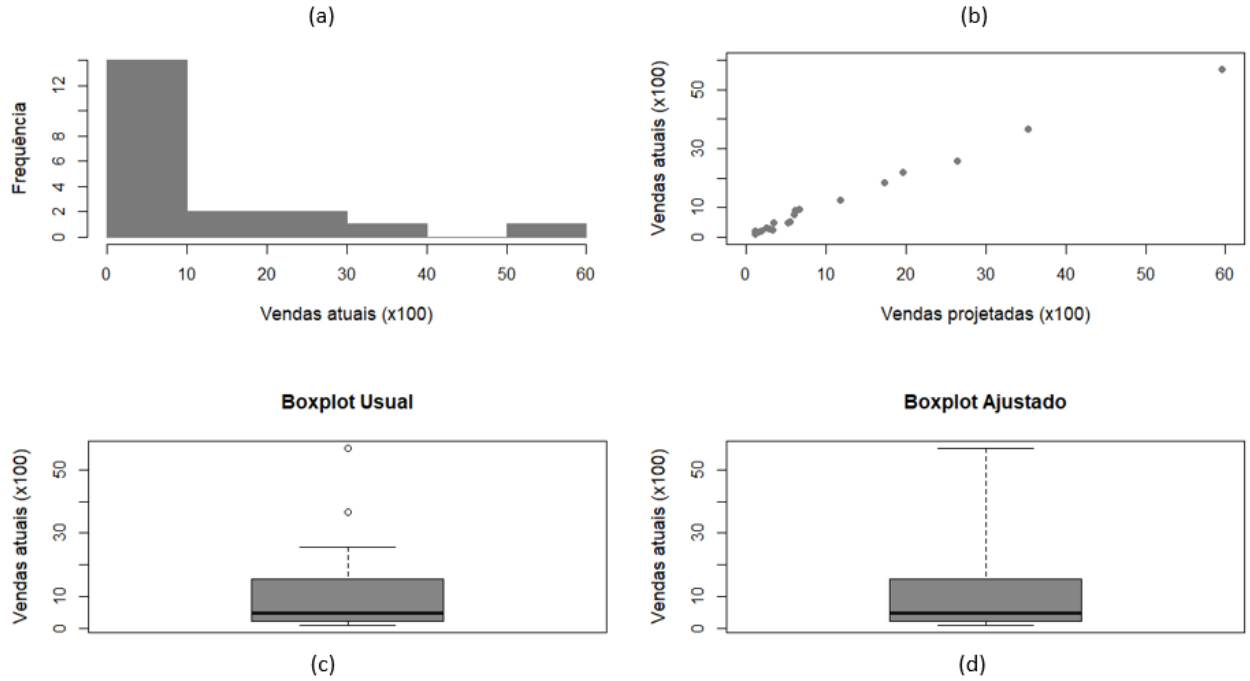
Medida descritiva	Valor
Valor mínimo	99
Média	1.138,55
Mediana	493,50
Desvio Padrão	1.435,91
Coeficiente de Variação	126,12
Coeficiente de Assimetria	1,79
Valor máximo	5.673

Por meio da Tabela 8, observa-se que a mediana está distante da média, estando

bem mais próxima do menor valor observado do que da média ou do maior valor. Esse fato, em conjunto com o coeficiente de assimetria obtido igual a 1,79, indica que os dados referentes às vendas atuais são assimétricos à direita, ou seja, sua cauda direita é mais longa que a esquerda. Essa assimetria consiste em uma das características da distribuição BS e pode ser vista com mais clareza por meio da Figura 2 .

Na Figura 2 encontra-se o gráfico de dispersão da venda projetada versus a atual, o histograma das vendas atuais e o boxplot usualmente utilizado e conhecido, além do ajustado das vendas atuais.

Figura 2 – Histograma (a), gráfico de dispersão (b) e boxplot usual (c) e boxplot ajustado (d) para as vendas atuais



O boxplot é muito popular para a finalidade de visualizar a distribuição de dados contínuos e unimodais. Isso se deve ao fato de essa forma gráfica fornecer uma gama de informações, como assimetria, localização, distribuição e caudas dos dados. Porém, quando se dispõe de dados assimétricos, certos pontos podem ser ditos como discrepantes erroneamente. Dessa forma, deve-se analisar o boxplot ajustado, o que resulta em uma representação mais precisa dos dados e de possíveis *outliers*.

Por meio da Figura 2(b), percebe-se, primeiramente que, à medida que a venda atual cresce, a venda projetada também cresce. Ou seja, a venda atual do j-ésimo produto está linearmente relacionado à j-ésima venda projetada. Ainda, é possível observar que a variabilidade das vendas tende a crescer à medida que o valor da venda também cresce, o que pode ser um indício de variância não constante dos dados.

Por meio do histograma da Figura 2(a), é possível perceber que os valores das

vendas atuais possuem uma distribuição unimodal e assimétrica à direita, com valores concentrados no intervalo de 0 a 1.000. Isso corrobora o que foi visto anteriormente, em que, após o cálculo da média das vendas atuais, foi percebido que esta se encontra consideravelmente maior que a mediana. Já por meio do boxplot usual das vendas atuais (Figura 2(c)), dois valores discrepantes são vistos, enquanto que, ao se construir o boxplot ajustado para distribuições assimétricas, nenhum *outlier* é detectado (Figura 2(d)).

Por meio da Tabela 8 e Figura 2, pode-se perceber que a amplitude elevada dos dados sugere uma alta variabilidade, o que corrobora o coeficiente de variação encontrado de 123%. Dessa forma, segundo o que foi dito, o ideal consiste em um modelo de regressão que consiga descrever, simultaneamente, a média e a variância não constante desse conjunto de dados, além da assimetria detectada.

4.1.2 Estimação dos parâmetros

A análise descritiva mostra uma distribuição unimodal e assimétrica à direita dos dados, além da variância não constante deles. Portanto, um modelo de regressão de BS parece apropriado para descrevê-los. Porém, neste trabalho há três abordagens distintas de modelos de regressão baseados na distribuição BS e, dessa forma, é de interesse comparar os ajustes deles a fim de verificar qual apresenta melhores ajustes para esse conjunto de dados. Em outras palavras, verificar qual modelo é mais robusto a observações atípicas.

A primeira abordagem de modelo de regressão baseado na distribuição BS trata do modelo BSR, proposto por Leiva et al. (2014). Assim, se assume a resposta $Y_i \sim$ BS (μ_i, δ) para os dados de vendas e a componente sistemática do modelo de regressão para a média é expressa da seguinte forma:

$$\mu_i = \beta_0 + \beta_1 x_i, \ \ i = 1, ..., 20,$$

em que β_0 e β_1 são os coeficientes de regressão e x_i é o valor do regressor X.

A segunda abordagem trata do modelo log-BS, baseado na transformação logarítmica das observações. O modelo de regressão é dado por:

$$Y_i = \mu_i + \epsilon_i, \ \ i = 1, ..., 20,$$

em que $Y_i = \log(T_i)$ é a resposta logarítmica i e $\epsilon_i = \log(\varphi_i) \sim \log{-}\text{BS}(\alpha, 0)$ é o erro do modelo de regressão para i=1,...,n. Além disso, vale ressaltar que se $T \sim \text{BS}(\alpha, \theta_i)$, então $Y \sim$ log-BS$(\alpha, \mu_i = \log(\theta_i))$.

Já a terceira e última abordagem de modelo de regressão baseado na distribuição BS trata do modelo de regressão BS, baseado na mediana e proposto por Balakrishnan e Zhu (2014). Assume-se uma função de ligação de forma que $h(\theta) = \eta_i, \ \ i = 1, ..., 20.$

Apenas para o modelo BSR, β_0 (intercepto) não é significativo ao nível de significância de 5%, ou seja, não é possível rejeitar que seu valor é igual zero. Assim, a fim de comparar os ajustes dos três modelos, o intercepto não é levado em consideração em nenhum deles. Dessa forma, as estimativas dos parâmetros obtidas por máxima verossimilhança estão dispostas na Tabela 9 e os valores para o AIC e BIC se encontram na Tabela 10.

Tabela 9 – Estimativas dos parâmetros por máxima verossimilhança

Modelo	**Parâmetro**	**Estimativa**	**Erro Padrão**	**P-valor**
log-BS	β_1	0,0026	0,0001	$<,001$
	α	11,6078	1,5679	$<,001$
BS	β_1	1,0674	0,0478	$<,001$
	α	0,2011	0,0318	$<,001$
BSR	β_1	1,0890	0,0492	$<,001$
	δ	49,4420	0,3162	$<,001$

Tabela 10 – Valores dos critérios de seleção para cada modelo

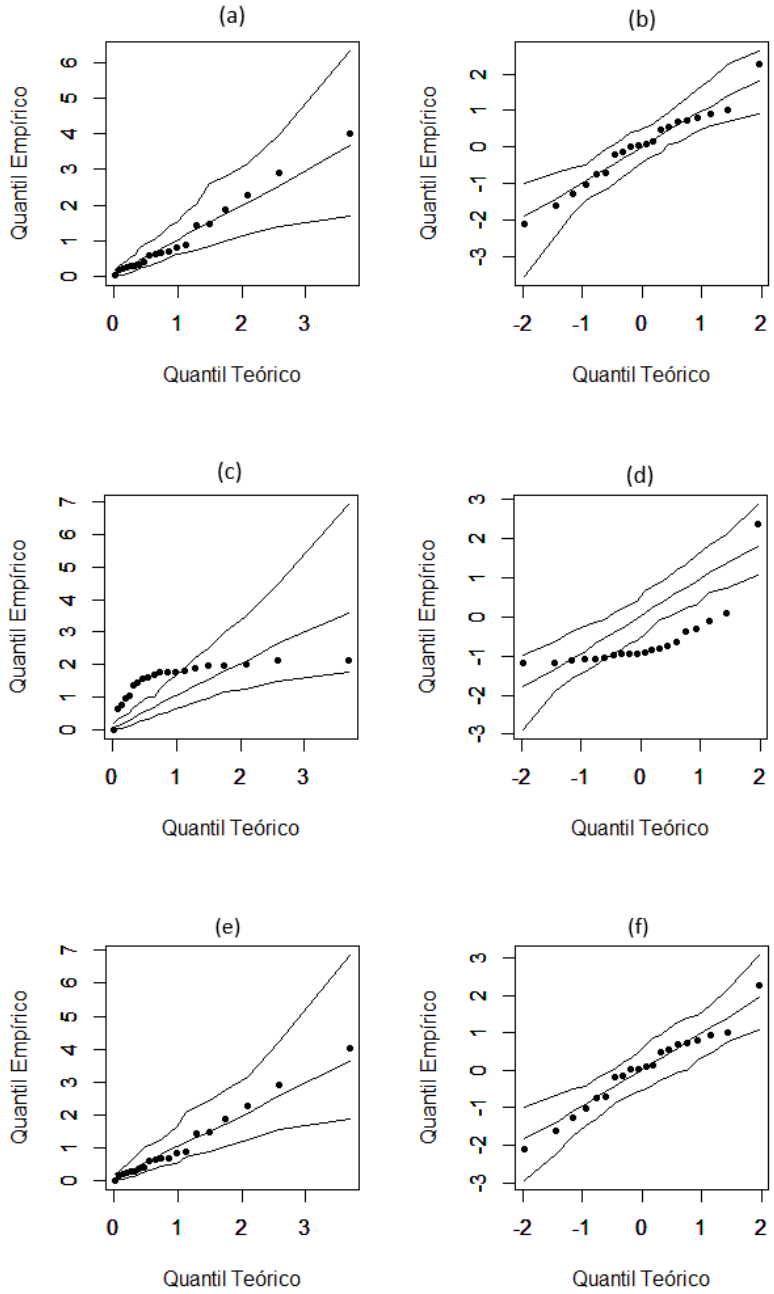
Modelo	**AIC**	**BIC**
log-BS	98,32	100,32
BS	251,78	253,77
BSR	251,78	253,77

Por meio da Tabela 9, observa-se que os modelos BS e BSR apresentam resultados semelhantes no que diz respeito à estimação de β_1 e seu respectivo erro padrão, além de valores iguais para o critérios AIC e BIC. Em relação ao valor desses critérios, o modelo log-BS apresenta valores bem mais baixos. Já as estimativas do parâmetro de forma variam para os três modelos. Todas as estimativas foram significativas ao nível de significância de 5%.

4.1.3 Análise de Resíduos

Em seguida, deve-se verificar a qualidade do ajuste dos modelos aqui apresentados. Dessa forma, na Figura 3 encontram-se os gráficos dos resíduos de Cox-Snell e do Quantil Aleatorizado para as três abordagens de modelos de regressão. As Figuras 3(a) e 3(b) se referem ao modelo BSR, enquanto as Figuras 3(c), 3(d), 3(e) e 3(f) se referem aos modelos log-BS e BS, respectivamente.

Figura 3 – Resíduo de Cox-Snell para o modelo BSR(a), log-BS (c) e BS (e), além do resíduo do Quantil Aleatorizado para o modelo BSR (b), log-BS (d) e BS (f).



Verifica-se que o comportamento de ambos os resíduos para os modelos BSR e BS é bastante semelhante, além de parecer que possuem uma boa qualidade de ajuste, enquanto o modelo log-BS não. Para esse último modelo, os resíduos na verdade ultrapassam os limites de confiança, além de possuírem um comportamento diferente do esperado, assemelhando-se a uma curva. Assim, os gráficos relacionados aos modelos BS e BSR não mostram características inesperadas, indicando que esses modelos se ajustam bem ao conjunto de dados. Por outro lado, o modelo log-BS parece não ser apropriado para

descrevê-lo.

4.1.4 Análise de Diagnóstico

Com o objetivo de verificar o quão sensível são as estimativas de máxima verossimilhança a dados atípicos, uma análise relacionada à influência local é conduzida. Leiva et al. (2014), durante a análise desse conjunto de dados, detecta as observações #13 e #19 como sendo possíveis valores influentes. Assim, com base nesse resultado obtido pelos autores, é analisado o impacto nas inferências dos modelos quando esses valores são retirados do banco de dados. Os parâmetros dos modelos são então novamente estimados, analisando-se a estimação quando apenas a observação #13 ou #19 é retirada e quando ambas são removidas. Na Tabela 11 é possível observar o impacto nas estimativas dos parâmetros quando esses casos são removidos por meio da medida de Mudança Relativa (RC). Essa medida é calculada da seguinte forma:

$$RC_\theta = \left| \frac{\hat{\theta}_j - \hat{\theta_{j(i)}}}{\hat{\theta}_j} \right| , \tag{4.1}$$

sendo $\hat{\theta_{j(i)}}$ a estimativa de máxima verossimilhança de θ_j após a retirada da i-ésima observação, para $j = 1, ..., 20$ e $i = 1, 2$. Além disso, $\theta_1 = \beta_1$ e $\theta_2 = \delta$ (ou α).

Tabela 11 – Mudança relativa (em porcentagem) das estimativas de máxima verossimilhança para os casos removidos e p-valores para o modelo relacionado às vendas

Modelo	**Casos Removidos**		β_1 **(vendas projetadas)**	δ
BSR	Nenhum	RC($\hat{\theta}$)	-	-
		RC($\hat{SE}$)	-	-
		p-valor	<,001	<,001
	{13}	RC($\hat{\theta}$)	1,98	30,13
		RC($\hat{SE}$)	8,35	2,62
		p-valor	<,001	<,001
	{19}	RC($\hat{\theta}$)	2,58	23,71
		RC($\hat{SE}$)	10,16	2,59
		p-valor	<,001	<,001
	{13, 19}	RC($\hat{\theta}$)	0,63	71,84
		RC($\hat{SE}$)	20,22	5,44
		p-valor	<,001	<,001

	Casos Removidos		β_1 **(vendas projetadas)**	α
log-BS	Nenhum	RC($\hat{\theta}$)	-	-
		RC($\hat{SE}$)	-	-
		p-valor	<,001	<,001
	{13}	RC($\hat{\theta}$)	0,78	0,75
		RC($\hat{SE}$)	13,33	8,81
		p-valor	<,001	<,001
	{19}	RC($\hat{\theta}$)	0,36	0,81
		RC($\hat{SE}$)	0,99	0,76
		p-valor	<,001	<,001
	{13, 19}	RC($\hat{\theta}$)	1,25	1,55
		RC($\hat{SE}$)	14,41	9,50
		p-valor	<,001	<,001
	Casos Removidos		β_1 **(vendas projetadas)**	α
BS	Nenhum	RC($\hat{\theta}$)	-	-
		RC($\hat{SE}$)	-	-
		p-valor	<,001	<,001
	{13}	RC($\hat{\theta}$)	2,45	12,35
		RC($\hat{SE}$)	7,76	10,10
		p-valor	<,001	<,001
	{19}	RC($\hat{\theta}$)	2,21	10,09
		RC($\hat{SE}$)	9,71	7,76
		p-valor	<,001	<,001
	{13,19}	RC($\hat{\theta}$)	0,19	23,72
		RC($\hat{SE}$)	19,27	19,60
		p-valor	<,001	<,001

Na Tabela 11 é possível observar o impacto da retirada de cada observação (ou de ambas) nas estimativas dos parâmetros para cada uma das três abordagens de modelo de regressão apresentados neste trabalho. Por meio dela, se percebe que as mudanças relativas mais significativas são referentes às estimativas do parâmetro δ (ou α, tratando-se dos modelos BS e log-BS). Além disso, essas maiores mudanças estão relacionadas à retirada, de forma simultânea, das observações #13 e #19. Esse fenômeno ocorre para as três abordagens de modelo de regressão (log-BS, BS e BSR) em relação à estimação desse parâmetro.

Como dito, para os três modelos, a maior mudança relativa na estimação de δ (α no caso dos modelos log-BS e BS) é obtida ao se retirar as observações #13 e #19 simultaneamente. Porém, em relação à estimação de β_1 esse fato é visto apenas para o modelo log-BS. Assim, para o modelo BSR a maior *relative change* (RC) na estimação de β_1 é vista ao se retirar a observação #19, enquanto que para o modelo BS a maior mudança relativa se deve à retirada da observação #13.

O modelo log-BS apresenta menores mudanças relativas na estimação de seus parâmetros se comparado aos demais modelos. A única excessão ocorre na estimação de β_1 quando são retirados de forma simultânea os valores #13 e #19, em que o modelo que

apresente menor RC é o BS. Dessa forma, em geral, pode-se dizer que o modelo log-BS é o menos impactado pela retirada das observações influentes do banco de dados.

Apesar de o modelo log-BS ser o menos impactado pela retirada dos valores influentes, seus resíduos mostram que ele não é adequado para modelar o conjunto de dados relacionado às vendas atuais. Dessa forma, deixando de lado o modelo log-BS e comparando apenas os modelos BS e BSR, apesar de possíveis valores influentes serem detectados, estes não afetam as inferências de nenhum dos dois modelos. Assim, ambos os modelos são robustos a dados atípicos, sendo que, para a estimação de δ (ou α) o modelo BSR sofre o maior impacto desses valores.

4.1.5 Aplicação 2 - Dados relacionados à fadiga de materiais

Para uma segunda aplicação, dispõe-se de um banco de dados relacionado a tempos de falha de dez peças de aço endurecido testadas em quatro níveis de estresse distintos. Os dados foram obtidos no *Princeton Laboratories of Mobil Research and Development Co* da Universidade de Princeton, nos Estados Unidos e podem ser encontrados em McCool (1980).

Primeiramente, uma análise exploratória dos dados é aplicada. Assim, na Tabela 12 se encontram algumas medidas descritivas relacionadas ao tempo de falha dessas peças.

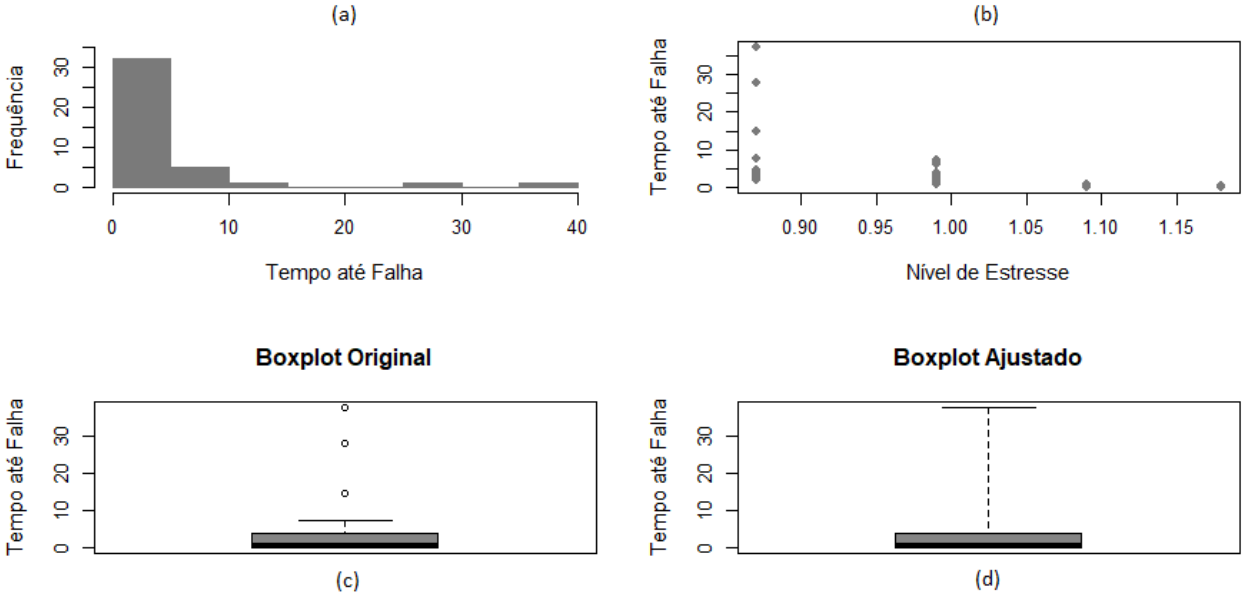
Tabela 12 – Medidas descritivas para os tempos de falha

Medida descritiva	**Valor**
Valor mínimo	0,012
Média	3,75
Mediana	0,84
Desvio Padrão	7,44
Coeficiente de Variação	198,12
Coeficiente de Assimetria	3,19
Valor máximo	37,4

É obtido o valor de 3,19 para o coeficiente de assimetria, indicando que esse conjunto de dados é assimétrico à direita e com uma assimetria mais acentuada que aquela vista nos dados apresentados na Aplicação 1, pelo valor desse coeficiente ser maior. Essa assimetria é corroborada pela mediana mais próxima do menor valor observado do que da média. Como dito, essa assimetria consiste em uma das características da distribuição Birnbaum-Saunders e pode ser vista também pela Figura 4(a).

Na Figura 4, é possível observar o gráfico de dispersão dos tempos até a falha das dez peças de aço endurecido (b), além de seu histograma (a), boxplot usualmente utilizado (c) e o ajustado (d).

Figura 4 – Histograma (a), gráfico de dispersão (b) e boxplot usual (c) e boxplot ajustado (d) dos tempos até a falha



Por meio da Figura 4(b), observa-se que para níveis de estresse mais baixos, a variabilidade dos tempos de falha do material é maior, variando desde 1,67 até 37,4 quando o nível de estresse é de 0,87. Assim, é possível observar que a variabilidade dos tempos de falha tende a diminuir à medida que o nível de estresse aumenta, podendo esse ser um indício de variância não constante desse conjunto de dados.

Por meio da Figura 4(a) se verifica o que foi visto na Tabela 12: a distribuição dos tempos até a falha são assimétricos à direita, com valores concentrados no intervalo de 0 a 5. Além disso, é importante observar que, por se ter dados assimétricos, a Figura 4(c) aponta a existência de três *outliers*, enquanto a Figura 4(d) não mostra nenhum valor como discrepante - o boxplot ajustado é mais adequado para a visualização de distribuições assimétricas.

4.1.6 Estimação dos parâmetros

Dados relacionados ao processo de fadiga são idealmente modelados pela distribuição BS, sendo esse processo o que deu lhe origem. Neste trabalho, são abordados três modelos de regressão distintos baseados nessa distribuição e, dessa forma, há o interesse em se estimar os parâmetros para esse conjunto de dados relacionados a cada um desses três modelos.

A primeira abordagem trata do modelo log-BS, baseado na transformação logarítmica das observações, ou seja, $Y_i = log(T_i)$. Já a segunda abordagem trata do modelo

BSR, em que se assume a resposta $Y_i \sim \text{BS}\ (\mu_i, \delta)$ e uma componente sistemática para a média. Nesta aplicação, é considerado o modelo de regressão BSR definido por:

$$Y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i,\ \ i = 1, ..., 40.$$

Por fim, a terceira e última abordagem trata do modelo BS e é baseada na mediana com uma função de ligação log-linear para o parâmetro de escala θ. Assim, $\theta_i = \exp[\beta_0 + \beta_1 \log(x_1) + ... + \beta_p \log(x_{40})]$.

Na Tabela 13 encontram-se as estimativas dos parâmetros para os três modelos, obtidos por meio da máxima verossimilhança, além dos p-valores e erro padrão dos respectivos parâmetros. Além disso, os valores dos critérios AIC (de Akaike) e BIC (Bayesiano) - Tabela 14.

Tabela 13 – Estimativas dos parâmetros por máxima verossimilhança

Modelo	**Parâmetro**	**Estimativa**	**Erro Padrão**	**P-valor**
log-BS	β_0	0,0978	0,1707	0,566
	β_1	-14,1164	1,5714	$<,001$
	α	1,2791	0,1438	$<,001$
BS	β_0	0,0978	0,1707	0,566
	β_1	-14,1163	1,5714	$<,001$
	α	1,2791	0,1438	$<,001$
BSR	β_0	0,6962	0,1935	$<,001$
	β_1	-14,1170	1,5718	$<,001$
	δ	1,222	0,2748	$<,001$

Tabela 14 – Valores dos critérios de seleção para cada modelo

Modelo	**AIC**	**BIC**
log-BS	129,24	134,30
BSR	125,12	130,18
BS	125,12	130,18

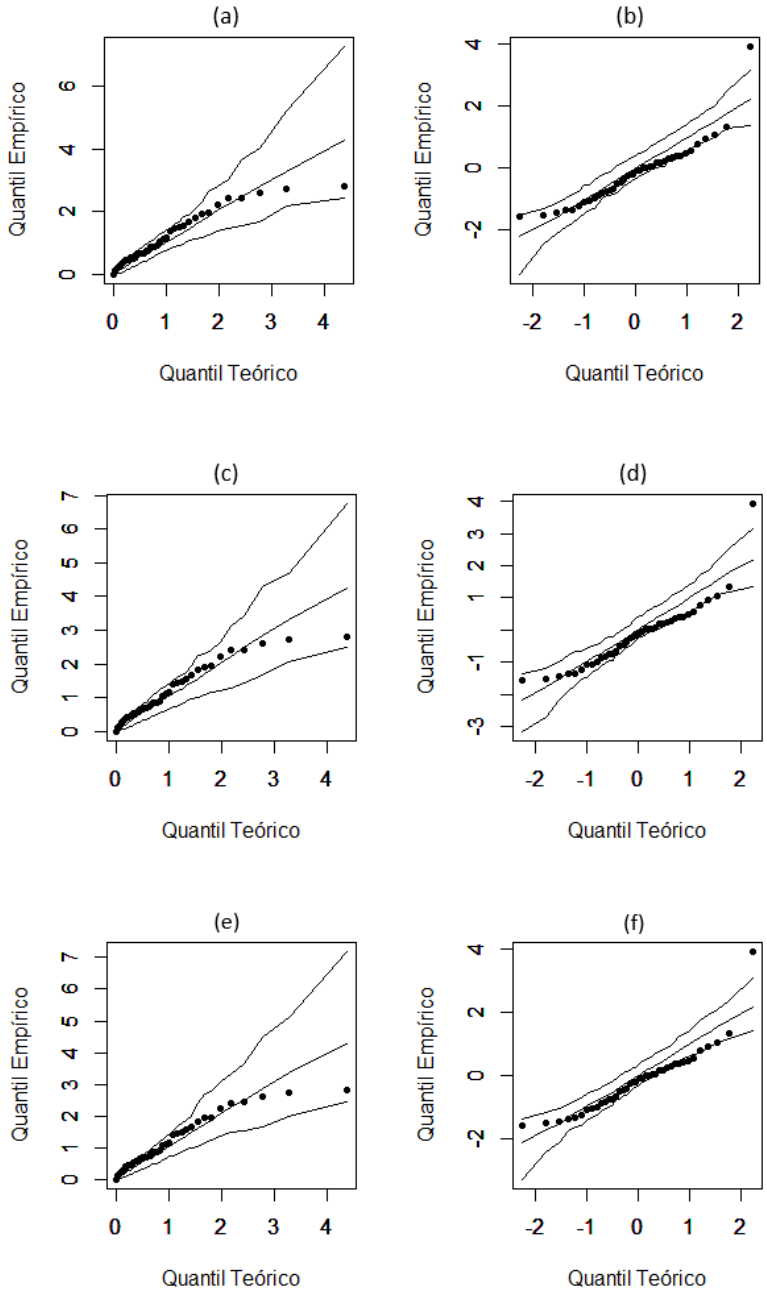
Os três modelos apresentam estimativas semelhantes de β_1 e do parâmetro de precisão. Porém, as estimativas do intercepto são distintas: os modelos log-BS e BS apresentam valores iguais (0,0978), enquanto o modelo BSR apresenta uma estimativa de 0,6962. Importante perceber que apenas para esse modelo o intercepto é significativo ao nível de significância de 5%.

Sabe-se que quanto menor o valor obtido para os critérios AIC e BIC, melhor é o ajuste do modelo. Dito isso, observam-se valores iguais para ambos os critérios para os modelos BSR e BS, sendo eles menores que os observados para o modelo log-BS.

4.1.7 Análise de Resíduos

A qualidade do ajuste dos modelos apresentados neste trabalho deve ser verificada, utilizando-se para isso os resíduos de Cox-Snell e Quantil Aleatorizado. Assim, na Figura 5 se encontram os gráficos dos resíduos para cada um dos três modelos.

Figura 5 – Resíduo de Cox-Snell para os modelos BSR (a), log-BS, (c) e BS (e), além do Quantil Aleatorizado para os modelos BSR (b), log-BS (d) e BS (f).



Por meio dos gráficos, verifica-se que o comportamento do resíduo de Cox-Snell é semelhante para os três modelos. O mesmo pode ser visto para o resíduo Quantil

Aleatorizado. Tendo como base primeiramente o resíduo Cox-Snell, os dados parecem estar mais bem ajustados se comparado ao resíduo Quantil Aleatorizado. Todos os valores estão dentro do intervalo de confiança, mas, em contrapartida, os resíduos Quantil Aleatorizado parecem ter um comportamento mais semelhante à linha central dos resíduos, possuindo, entretanto, valores que ultrapassam (ligeiramente) os limites de confiança.

Dessa forma, os gráficos fornecem indícios de que os três modelos estão bem ajustados aos dados e de forma similar, não sendo possível, pelos gráficos de resíduos, afirmar qual modelo possui um melhor ajuste.

4.1.8 Análise de Diagnóstico

Assim como feito para o conjunto de dados anteriormente utilizado, é de interesse realizar uma análise relacionada às observações influentes. Assim, tendo como objetivo verificar a sensibilidade das estimativas de máxima verossimilhança a dados atípicos, estimar novamente os parâmetros do modelo após a retirada dos valores influentes do banco de dados. Dessa forma, está sendo analisado o impacto nas inferências do modelo quando esses valores são removidos, observando qual modelo é mais robusto. Isso é feito por meio da medida Mudança Relativa (4.1).

Como mencionado anteriormente, o conjunto de dados pode ser encontrado em McCool (1980). Uma aplicação desses dados pode ser vista em Lemonte (2012), em que é introduzido um modelo de regressão log-BS assimétrico. A fim de identificar observações influentes nesse modelo, o autor aplica a medida *generalized leverage* e métodos de influência local. Assim, ele verifica que as observações #2, #3, #10, #18 e #40 possuem maior influência se comparadas às demais observações. Os casos #9, #10 e #21 possuem influência sobre seus próprios valores ajustados e, além disso, a observação #1 também é detectada influente.

Assim, tendo como referência Lemonte (2012) na identificação dos valores influentes, são retirados do banco de dados cada um desses valores (#1, #2, #3, #9, #10, #18, #21 e #40), um de cada vez. Por meio da Tabela 15 se percebe o impacto nas estimativas dos parâmetros de cada um dos modelos quando essas observações são removidas do banco de dados.

Tabela 15 – Mudança relativa (em porcentagem) das estimativas de máxima verossimilhança para os casos removidos e p-valores para o modelo relacionado à fadiga de materiais expostos a certos níveis de estresse.

Modelo	**Caso Removido**		β_0 **(intercepto)**	β_1 **(estresse)**	δ
BSR	Nenhum	RC($\hat{\theta}$)	-	-	-
		RC($\hat{SE}$)	-	-	-
		p-valor	$<,001$	$<,001$	$<,001$
	{1}	RC($\hat{\theta}$)	5,14	3,33	2,35
		RC($\hat{SE}$)	1,81	2,56	3,64
		p-valor	$<,001$	$<,001$	$<,001$
	{2}	RC($\hat{\theta}$)	4,09	2,26	0,44
		RC($\hat{SE}$)	1,91	2,60	1,75
		p-valor	$<,001$	$<,001$	$<,001$
	{3}	RC($\hat{\theta}$)	3,69	1,87	0,24
		RC($\hat{SE}$)	2,02	2,61	1,06
		p-valor	$<,001$	$<,001$	$<,001$
	{9}	RC($\hat{\theta}$)	4,68	2,30	0,64
		RC($\hat{SE}$)	2,48	4,31	1,93
		p-valor	0,002	$<,001$	$<,001$
	{10}	RC($\hat{\theta}$)	8,33	3,68	2,78
		RC($\hat{SE}$)	2,22	5,20	4,08
		p-valor	0,003	$<,001$	$<,001$
	{18}	RC($\hat{\theta}$)	6,97	0,86	2,82
		RC($\hat{SE}$)	0,57	1,30	4,11
		p-valor	0,002	$<,001$	$<,001$
	{21}	RC($\hat{\theta}$)	5,64	13,76	96,43
		RC($\hat{SE}$)	23,41	27,11	98,00
		p-valor	$<,001$	$<,001$	$<,001$
	{40}	RC($\hat{\theta}$)	3,33	2,44	0,97
		RC($\hat{SE}$)	0,21	1,72	2,29
		p-valor	0,001	$<,001$	$<,001$
	Caso Removido		β_0 **(intercepto)**	β_1 **(estresse)**	α
log-BS	Nenhum	RC($\hat{\theta}$)	-	-	-
		RC($\hat{SE}$)	-	-	-
		p-valor	0,566	$<,001$	$<,001$
	{1}	RC($\hat{\theta}$)	47,34	3,34	1,16
		RC($\hat{SE}$)	2,34	2,53	0,07
		p-valor	0,4095	$<,001$	$<,001$
	{2}	RC($\hat{\theta}$)	31,08	2,26	0,23
		RC($\hat{SE}$)	2,46	2,58	1,04
		p-valor	0,4636	$<,001$	$<,001$
	{3}	RC($\hat{\theta}$)	25,15	1,87	0,11
		RC($\hat{SE}$)	2,46	2,60	1,39
		p-valor	0,4842	$<,001$	$<,001$
	{9}	RC($\hat{\theta}$)	30,06	2,29	0,30
		RC($\hat{SE}$)	2,60	4,33	0,97
		p-valor	0,6958	$<,001$	$<,001$

	{10}	RC($\hat{\theta}$)	46,11	3,65	1,37
		RC($\hat{SE}$)	2,60	5,22	0,14
		p-valor	0,7634	$<,001$	$<,001$
	{18}	RC($\hat{\theta}$)	37,01	0,86	1,38
		RC($\hat{SE}$)	0,70	1,31	0,14
		p-valor	0,7199	$<,001$	$<,001$
	{21}	RC($\hat{\theta}$)	294,79	13,75	28,68
		RC($\hat{SE}$)	23,08	27,14	28,16
		p-valor	0,003	$<,001$	$<,001$
	{40}	RC($\hat{\theta}$)	19,33	2,44	0,49
		RC($\hat{SE}$)	0,06	1,71	0,76
		p-valor	0,6443	$<,001$	$<,001$
	Caso Removido		β_0 **(intercepto)**	β_1 **(estresse)**	α
BS	Nenhum	RC($\hat{\theta}$)	-	-	-
		RC($\hat{SE}$)	-	-	-
		p-valor	0,566	$<,001$	$<,001$
	{1}	RC($\hat{\theta}$)	47,34	3,34	1,16
		RC($\hat{SE}$)	2,40	2,54	0,07
		p-valor	0,4637	$<,001$	$<,001$
	{2}	RC($\hat{\theta}$)	30,98	2,26	0,23
		RC($\hat{SE}$)	2,46	2,59	1,04
		p-valor	0,1281	$<,001$	$<,001$
	{3}	RC($\hat{\theta}$)	25,15	1,87	0,11
		RC($\hat{SE}$)	2,46	2,61	1,39
		p-valor	0,1224	$<,001$	$<,001$
	{9}	RC($\hat{\theta}$)	30,06	2,29	0,30
		RC($\hat{SE}$)	2,46	4,33	0,97
		p-valor	0,0684	$<,001$	$<,001$
	{10}	RC($\hat{\theta}$)	46,11	3,65	0,36
		RC($\hat{SE}$)	2,50	5,22	0,14
		p-valor	0,0527	$<,001$	$<,001$
	{18}	RC($\hat{\theta}$)	37,01	0,86	1,38
		RC($\hat{SE}$)	0,70	1,31	0,14
		p-valor	0,0616	$<,001$	$<,001$
	{21}	RC($\hat{\theta}$)	294,68	13,76	28,68
		RC($\hat{SE}$)	23,08	27,14	28,16
		p-valor	0,3860	$<,001$	$<,001$
	{40}	RC($\hat{\theta}$)	19,33	2,44	0,49
		RC($\hat{SE}$)	0,06	1,71	0,76
		p-valor	0,0789	$<,001$	$<,001$

Como é visto na Seção 4.1.6, as estimativas por máxima verossimilhança de β_1 para os três modelos de regressão são quase idênticas. Isso ocorre na estimativa desse parâmetro mesmo quando as observações influentes são retiradas do conjunto de dados. Dessa forma, os três modelos apresentam mudança relativa para β_1 igual, para todos os casos removidos.

Os modelos log-BS e BS, além de apresentarem estimativas iguais para β_1, também

apresentam para β_0 e α, sendo suas mudanças relativas iguais em relação à estimação de todos os parâmetros. Dessa forma, a retirada das observações influentes impacta os dois modelos na mesma intensidade.

Conforme o que foi exposto, os três modelos apresentam estimativas quase iguais para β_1. As estimativas dos outros dois parâmetros dos modelos log-BS e BS tambem são iguais, sendo distintas daquelas obtidas para o modelo BSR. Dessa forma, comparando as mudanças relativas desse modelo com as dos demais, observa-se que ele apresenta valores mais baixos para a *relative change* ao se analisar as estimativas de β_0 e mudanças relativas maiores ao se analisar δ.

Pode-se, então, chegar às seguintes conclusões: os três modelos são igualmente impactados pelas observações influentes na estimação de β_1; o modelo BSR é o menos impactado na estimação de β_0; e, por fim, os modelos BS e log-BS sofrem menos impacto das observações influentes na estimação do parâmetro δ (α no caso dos modelos log-BS e BS). Com isso, nenhum dos três modelos de regressão mostra ser mais robusto que os demais ao se analisar a estimação de todos os parâmetros simultaneamente.

5 Considerações Finais

Diante das generalizações e extensões da distribuição Birnbaum-Saunders, neste trabalho são avaliadas três abordagens distintas de modelos de regressão baseados nessa distribuição (BSR, log-BS e BS). Em específico, é avaliado o desempenho dos três modelos em termos de predição e ajuste por meio de duas simulações de Monte Carlo distintas. Além disso, são conduzidas duas aplicações a dados reais com o objetivo de comparar os ajustes dos três modelos.

Após a primeira simulação realizada para diferentes tamanhos amostrais e distintos valores do parâmetro de forma, são observados resultados piores para o modelo BSR, enquanto os resultados para os modelos BS e log-BS são semelhantes. Em relação à segunda simulação realizada, observa-se, em geral, melhores resultados para os critérios AIC e BIC para o modelo log-BS e, para o EQM do modelo, resultados melhores para o log-BS e BS. Além disso, o modelo log-BS possui melhor performance na estimação do intercepto. Em geral, não há um modelo que apresenta melhor performance em todos os aspectos analisados. Porém, há indícios de que o modelo log-BS é relativamente melhor que os demais. Por fim, em relação à essa simulação, é vista uma performance pior quando os dados são gerados da distribuição log hiperbólica e melhor quando são gerados da log normal.

Duas aplicações a dados reais são conduzidas. A primeira considera um conjunto de dados relacionado à venda de produtos. Observa-se para esses dados que, apesar do modelo log-BS ser o menos impactado pela retirada das observações influentes, seus resíduos mostram que esse modelo não está bem ajustado. Assim, comparando-se apenas os modelos BS e BSR, se observa que, em relação à estimação de β_1 nenhum padrão pode ser identificado. Por outro lado, ao se analisar a estimação de δ (ou α, no caso dos modelos BS e log-BS), o modelo BSR mostra ser o mais impactado com a retirada das observações influentes. Nenhuma mudança inferencial é detectada na retirada desses valores.

A segunda aplicação está relacionada a tempos de fadiga de materiais expostos a certos níveis de estresse. Para esse conjunto de dados, observa-se que, em relação à estimaçao de β_1, todos os modelos são igualmente impactados pelas observações influentes. Por outro lado, o modelo BSR sofre maior impacto na estimação de δ, enquanto os modelos log-BS e BS sofrem maior impacto na estimação de β_0.

Ao se comparar as três abordagens de modelos de regressão apresentadas neste trabalho, observa-se que o modelo BSR parece sofrer maior impacto na retirada de valores influentes quando se trata da estimação de δ. Além disso, em relação aos demais modelos, apresenta valores mais altos para os critérios AIC, BIC e EQM do modelo. Assim, esse

modelo mostra ser o menos robusto se comparado aos demais. Tratando-se da estimação dos demais parâmetros, não se percebe nenhum padrão em relação ao modelo mais robusto a valores influentes ou qual modelo apresenta melhor performance desses estimadores.

Referências

BALAKRISHNAN, N. et al. On some mixture models based on the birnbaum–saunders distribution and associated inference. *Journal of Statistical Planning and Inference*, Elsevier, v. 141, n. 7, p. 2175–2190, 2011. Citado na página 17.

BALAKRISHNAN, N.; ZHU, X. Inference for the birnbaum-saunders lifetime regression model with applications. *Communications in Statistics - Simulation and Computation*, v. 128, 2014. Citado 2 vezes nas páginas 17 e 39.

BIRNBAUM, Z.; SAUNDERS, S. A new family of life distributions. *Journal of Applied Probability*, Cambridge University Press, v. 6, p. 319–327, 1969. Citado 2 vezes nas páginas 17 e 19.

DíAZ-GARCIA, J. A.; LEIVA-SáNCHEZ, V. A new family of life distributions based on the elliptically contoured distributions. *Journal of Statistical Planning and Inference*, Elsevier, v. 128, n. 2, 2005. Citado na página 17.

FERREIRA, M.; GOMES; LEIVA, V. On an extreme value version of the birnbaum-saunders distribution. *Revstat*, INE, v. 10, n. 2, p. 181–210, 2012. Citado na página 37.

FIERRO, R. et al. On a birnbaum-saunders distribution arising from a non-homogeneous poisson process. *Statistics & Probability Letters*, Elsevier, v. 83, n. 4, p. 1233–1239, 2013. Citado na página 17.

GOMEZ, H. W.; OLIVARES-PACHECO, J. F.; BOLFARINE, H. An extension of the generalized birnbaum–saunders distribution. *Statistics & Probability Letters*, Elsevier, v. 79, n. 3, p. 331–338, 2009. Citado na página 17.

JOHNSON, N.; KOTZ, S.; BALAKRISHNAN, N. *Continuous univariate distributions*. [S.l.]: New York, US: Wiley, 1995. Citado na página 37.

LEIVA, V. *The Birnbaum-Saunders Distribution*. [S.l.]: Elsevier, 2016. Nenhuma citação no texto.

LEIVA, V.; SANHUEZA, A.; ANGULO, J. M. A length-biased version of the birnbaum-saunders distribution with application in water quality. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 23, n. 3, p. 299–307, 2009. Citado na página 37.

LEIVA, V. et al. Birnbaum–saunders statistical modelling: a new approach. *Statistical Modelling*, Sage Publications Sage India: New Delhi, India, v. 14, n. 1, p. 21–48, 2014. Citado 3 vezes nas páginas 17, 39 e 42.

LEMONTE, A. J. A log-birnbaum-saunders regression model with assymetric errors. *Journal of Statistical Computation and Simulation*, Taylor & Francis Group, v. 82, n. 12, p. 1775–1787, 2012. Citado na página 48.

MARCHANT, C. et al. Air contaminant statistical distributions with application to pm10 in santiago, chile. In: *Reviews of Environmental Contamination and Toxicology Volume 223*. [S.l.]: Springer, 2013b. p. 1–31. Citado na página 37.

MCCOOL, J. I. Confidence limits for weibull regression with censored data. *IEEE Transactions of Reliability*, IEE, v. 29, n. 2, p. 145–150, 1980. Citado 2 vezes nas páginas 44 e 48.

MILLS, J. *Robust Estimation of the Birnbaum-Saunders Distribution*. Tese (Doutorado) — Technical Univesity of Nova Scotia, 1997. Citado na página 37.

RIECK, J. R.; NEDELMAN, J. R. A log-linear model for the birnbaum—saunders distribution. *Technometrics*, Taylor & Francis Group, v. 33, n. 1, p. 51–60, 1991. Citado 3 vezes nas páginas 17, 21 e 22.

SANTOS-NETO, M. et al. On new parameterizations of the birnbaum-saunders distribution. *Pakistan Journal of Statistics*, v. 28, n. 1, p. 1–26, 2012. Citado na página 22.

SAULO, H. et al. Birnbaum-saunders autoregressive conditional duration models applied to high-frequency financial data. *Statistical Papers*, Springer, p. 1–25, 2017. Citado na página 27.

SETO IWASE KOSEI, M. O. S. *Characteristics of Rainfall for a Single Event*. Tese (Doutorado) — Department of Applied Mathematics, Hiroshima University , Hiroshima, Japan, 1993. Citado na página 37.

SETO IWASE KOSEI, M. O. S. Characteristics of rainfall for a single event in hiroshima city. *Tenki (Meteorol. Soc. Jpn.)*, v. 42, p. 147–158, 1995. Citado na página 37.

VILCA, F. et al. An extended birnbaum–saunders model and its application in the study of environmental quality in santiago, chile. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 24, n. 5, p. 771–782, 2010. Citado na página 37.

VILCA-LABRA, F.; LEIVA-SáNCHEZ, V. A new fatigue life model based on the family of skew-elliptical distributions. *Communications in Statistics—Theory and Methods*, Taylor & Francis, v. 35, n. 2, p. 229–244, 2006. Citado na página 17.