UFAM

MODELO DE REGRESSÃO TOBIT COM A DISTRIBUIÇÃO GENERALIZADA BIRNBAUM-SAUNDERS

Thiago Souza de Melo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: Jeremias da Silva Leão

Manaus

Fevereiro de 2019

MODELOS DE REGRESSÃO TOBIT COM A DISTRIBUIÇÃO GENERALIZADA BIRNBAUM-SAUNDERS

Thiago Souza de Melo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinado por:

Prof. Dr. Jeremias da Silva Leão - UFAM (Orientador)

Prof. Dr. Manoel Ferreira dos Santos Neto - UFCG (Membro Externo)

Prof. Dr. Celso Rômulo Barbosa Cabral - UFAM (Membro)

MANAUS-AM
FEVEREIRO DE 2019

Melo, Thiago Souza de
M528m Modelo de regressão tobit com a distribuição
generalizada Birnbaum-Saunders. / Thiago Souza de Melo. 2019
45 f.: il.; 31 cm.

Orientador: Jeremias da Silva Leão
Dissertação (Mestrado em Matemática - Estatística)
Universidade Federal do Amazonas.

1. Distribuição Generalizada Birnbaum-Saunders. 2. Modelos Tobit.
3. Análise de Sobrevivência. 4. Análise de resíduo e diagnóstico.
I. Leão, Jeremias da Silva II. Universidade Federal do
Amazonas. III. Título

Aos meus pais Cleonice e Desvaldo
e a minha esposa Welem como
forma de gratidão.

Agradecimentos

Agradeço,

Em primeiro lugar Ao Eterno, pela saúde, vida, força, e perseverança para realizar o mestrado.

A minha querida esposa Welem, pelo grande incentivo antes e durante o curso. A minha família, meus pais de maneira especial que me apoiaram e ajudaram durante toda minha vida.

Agradeço aos meus amigos de turma Érico, Roberto e Antônio Azevedo. Também a todos os colegas de classe.

Minha gratidão aos professores do Departamento de Estatística, que sempre estão dispostos a ajudar os alunos. Em especial ao meu orientador Jeremias da Silva Leão, pelo incentivo e paciência durante a orientação, por ter contribuído em meus conhecimentos. Agradeço também aos professores: James Dean, Max Souza , Celso Rômulo, José Raimundo e José Mir, pelas contribuições em meus conhecimentos através das disciplinas.

Agradeço a Capes pelo apoio financeiro durante o mestrado.

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

MODELO DE REGRESSÃO TOBIT COM A DISTRIBUIÇÃO GENERALIZADA BIRNBAUM-SAUNDERS

Thiago Souza de Melo

Fevereiro/2019

Orientador: Jeremias da Silva Leão

Área de Concentração: Estatística

Neste trabalho propomos o modelo de regressão tobit baseado na distribuição generalizada Birnbaum-Saunders (Barros *et al.* (2008)). Implementamos uma abordagem baseada no método da máxima verossimilhança para obter as estimativa dos parâmetros e derivamos medidas para análise de resíduos e diagnóstico. Em seguida fizemos um estudo via simulações de Monte Carlo com o objetivo de avaliar o desempenho dos estimadores de máxima verossimilhança do modelo proposto. Por fim, ilustramos a metodologia proposta usando um conjunto de dados reais.

Palavras-chave: Distribuição Generalizada Birnbaum-Saunders; Modelos Tobit; Análise de Sobrevivência; Análise de resíduo e diagnóstico.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

MODELO DE REGRESSÃO TOBIT COM A DISTRIBUIÇÃO GENERALIZADA BIRNBAUM-SAUNDERS

Thiago Souza de Melo

February/2019

Advisor: Jeremias da Silva Leão

Research lines: Statistics

In this work, we propose the tobit regression model based on the Birnbaum-Saunders generalized distribution (Barros *et al.* (2008)). We implemented an approach based on the maximum likelihood method to obtain the parameter estimates and derive measurements for residue analysis and diagnostics. We then carried out a simulation of Monte Carlo study with the objective of evaluating the performance of the maximum likelihood estimators of the proposed model. Finally, we illustrate the proposed methodology using a set of real data.

Keywords: Generalized Birnbaum-Saunders distribution; Tobit Models; Survival analysis; Residue analysis and diagnosis.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Em estudos com modelagem estatística encontra-se alguns casos em que a variável de interesse possui algumas limitações nas extremidades (inferior ou superiormente). Nesse contexto Tobin (1958) introduziu uma metodologia alternativa ao modelo de regressão, uma vez que esses dados limitados causam violação no pressuposto de linearidade em modelos de regressão. Dessa forma, Tobin (1958) sugeriu interpretar o valor extremo da variável resposta como sendo censura, metodologia conhecida como modelo tobit. O modelo tobit tem sido usado em estudos nas diversas áreas (ver e.g. Amemiya (1984); Mroz (1987); Barros *et al.* (2008)). Considerando-se a abordagem apresentada em Tobin (1958) nota-se que o autor assumiu normalidade para o termo aleatório do modelo, contudo essa suposição pode não ser adequada em algumas aplicações, dada a necessidade de uma modelagem mais robusta. Com isso, alguns autores desenvolveram modelos tobit mais flexíveis, principalmente no contexto de estudos com dados de análise de sobrevivência, na qual a variável resposta censurada normalmente possui assimetria positiva, ou bimodalidade e caudas mais pesadas (ver e.g. Galea *et al.* (2004); Barros *et al.* (2010); Martínez-Flórez *et al.* (2013a); Martínez-Flórez *et al.* (2013b); Leiva *et al.* (2014a)). Barros *et al.* (2018) propuseram uma generalização para o modelo tobit, introduzindo a família de distribuições de contornos elípticos, que oferecem distribuições mais flexíveis para modelar valores extremos, como a distribuição t-Student que em particular é conhecida por possuir caudas mais pesadas, e discutiram uma análise entre o modelo tobit normal e o modelo tobit baseado na distribuição t. Além disso, De Sousa *et al.* (2018) desenvolveram um novo modelo tobit, baseado em modelos assimétricos, a distribuição Biunbaum-Saunders, apresentando uma análise diagnóstica de influência global e local.

A distribuição Birnbaum-Saunders (BS) foi desenvolvida por Birnbaum & Saunders (1969) com o interesse de verificar o tempo até a ocorrência de alguma falha em materiais causada devido a danos. Várias metodologias baseadas na distribuição foram propostas. Por exemplo, Rieck & Nedelman (1991) apresentaram um modelo log-linear Birnbaum-Saunders a qual é um caso particular da classe de distribuições seno-hiperbólico normal, Owen & Padgett (2000) apresentaram modelos BS para tempo de vida acelerado. A BS tem sido aplicada em diversas áreas, na qual destacamos alguns trabalhos no contexto de análise com dados médicos, são estes: Leiva *et al.* (2007); Barros *et al.* (2008); Qu & Xie (2011) e Leiva (2015).

Díaz-García & Leiva-Sánchez (2005) apresentaram uma generalização da distribuição BS, baseado em distribuições de contornos elípticos, também chamada de distribuições simétricas (ver e.g. Anderson (1990); Fang (1990) e Galea *et al.* (2000)). O objetivo dos autores é modelar dados com diferentes graus de assimetria, curtose e com caudas pesadas. Barros *et al.* (2008) apresentaram uma nova classe de modelo de regressão aplicados em dados de sobrevivência, onde é realizado uma análise entre o modelo BS e o modelo BS-t. Barros *et al.* (2009) desenvolveram um pacote na linguagem `R`, chamado `gbs` usado para analisar dados com ou sem censura, usando a distribuição generalizada Birnbaum-Saunders (GBS), veja mais em Leiva *et al.* (2009).

Desta forma, este trabalho tem como objetivo apresentar o modelo de regressão tobit usando a distribuição GBS. De forma específica, apresentamos os principais recursos inferenciais para a distribuição BS-t. Esta dissertação está dividida da seguinte forma. No Capítulo 2 apresentamos uma breve revisão para contextualização da distribuição BS, posteriormente a GBS, e finalizamos o capítulo descrevendo a forma inicial do modelo tobit. No Capítulo 3 apresentamos o modelo tobit-BS-t, sua estimação e alguns resultados assintóticos. Em seguida, descrevemos a análise dos resíduos e alguns métodos de diagnósticos para o modelo. No Capítulo 4 avaliamos um estudo de simulação, e também aplicamos a teoria estudada a um conjunto de dados. No Capítulo 5 descrevemos as principais conclusões e citamos propostas futuras.

Capítulo 2

Preliminares

Neste Capítulo iremos apresentar propriedades da distribuição BS, visto que, como mencionado anteriormente, iremos obter um modelo baseado em uma de suas generalizações. Em seguida apresentamos a distribuição GBS, na qual descrevemos na Tabela 2.1 as expressões das funções densidades que podem ser obtidas para distribuições dessa classe. Dentre estas, focaremos no caso específico BS-t, em que apresentamos algumas de suas propriedades, como por exemplo: geração de números aleatórios, momentos, entre outros. Por fim, introduzimos os principais conceitos sobre modelos tobit, metodologia apresentada em Tobin (1958).

2.1 Distribuição Birnbaum-Saunders

A distribuição BS proposta por Birnbaum & Saunders (1969) é conhecida como modelo de tempo de fadiga. Por conta de suas propriedades matemáticas, como também devido a sua estreita relação com a distribuição normal a BS tem sido objeto de estudo de diversos pesquisadores nos últimos cinquenta anos. Desde então a BS foi aplicada a dados reais de diversas áreas. Aos leitores interessados recomenda-se ver Leiva (2015), visto que o autor faz um apanhado sobre tudo que já foi desenvolvido com a BS até aquele momento.

Seja T uma variável aleatória que representa o tempo até a ocorrência do evento interessado, então podemos assumir que essa variável segue uma distribuição BS se sua

função de distribuição acumulada (CDF) é dada por

$$F_t(t,\alpha,\sigma) = \Phi\left(\frac{1}{\alpha}\left(\sqrt{t/\sigma} - \sqrt{\sigma/t}\right)\right), \tag{2.1}$$

em que $t > 0, \alpha > 0, \sigma > 0$ e Φ é a CDF de uma distribuição normal padrão, iremos denotar por $T \sim BS(\alpha,\sigma)$. Os parâmetros α e σ representam a forma e escala, respectivamente. Além disso, a BS é uma distribuição unimodal e assimétrica a direita. A função densidade de probabilidade (PDF) da variável aleatória T em (2.1) é expressa por

$$f_T(t;\alpha,\sigma) = \frac{1}{2\alpha}\left(\sqrt{t/\sigma t} + \sqrt{\sigma/t^{\frac{2}{3}}}\right)\Phi\left(\frac{1}{\alpha}\left(\sqrt{t/\sigma} - \sqrt{\sigma/t}\right)\right). \tag{2.2}$$

A PDF dada em (2.2) pode ser reescrita como,

$$f_T(t;\alpha,\sigma) = \frac{\exp(\alpha^{-2})}{2\alpha\sqrt{2\pi\sigma}}\exp\left(\frac{1}{2\alpha^2}\left(\frac{t}{\sigma} + \frac{\sigma}{t}\right)\right)t^{-\frac{3}{2}}(t+\sigma). \tag{2.3}$$

A partir de (2.1) e (2.2) obtemos a representação estocástica da variável aleatória T, em termos da distribuição normal padrão, $Z \sim N(0,1)$, dada por

$$T = \sigma\left(\alpha Z/2 + \sqrt{(\alpha Z/2)^2 + 1}\right)^2$$

e

$$Z = \frac{1}{\alpha}\left(\sqrt{T/\sigma} - \sqrt{\sigma/T}\right) \sim N(0,1).$$

Algumas propriedades da distribuição BS são apresentadas a seguir:

i) O r-ésimo momento da distribuição BS é dado por,

$$E(T^r) = \sigma\sum_{j=0}^{r}\binom{2r}{2j}\sum_{i=0}^{j}\binom{j}{i}\frac{(2(r-j+i))!}{2^{r-j+i}(r-j+i)!}\left(\frac{\alpha}{2}\right)^{2(r-j+i)}; \tag{2.4}$$

ii) Para $b > 0$, $bT \sim BS(\alpha, b\sigma)$, mostrando que a BS é fechada sobre um multiplicador escalar (proporcionalidade);

iii) $1/T \sim BS(\alpha, 1/\sigma)$ implicando que a BS é fechada sobre a reciprocidade;

iv) o parâmetro σ é a mediana da BS, que pode ser obtido diretamente quando $q = 0.5$ da sua função quantil dada por

$$t(q;\alpha,\sigma) = F_T^{-1}(q;\alpha,\sigma) = \sigma\left(\frac{\alpha z(q)}{2} + \sqrt{\frac{(\alpha z(q))^2}{2} + 1}\right)^2, \quad 0 < q < 1,$$

em que $z(q)$ é a função quantil da normal padrão.

A demonstração das propriedades acima, pode ser encontrada em Balakrishnan *et al.* (2009). Através da equação (2.4), obtemos a esperança e a variância da distribuição BS com parâmetros α e σ, dada por

$$E(T) = \sigma\left(1 + \frac{\alpha^2}{2}\right) \qquad \text{e} \qquad Var(T) = (\alpha\sigma)^2\left(1 + \frac{5\alpha^2}{4}\right).$$

Podemos definir a função de sobrevivência (SF) para a distribuição BS, usada para calcular a chance de uma observação não falhar até o tempo t,

$$S_T(t;\alpha,\sigma) = 1 - F_T(t;\alpha,\sigma) = 1 - \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\sigma}} - \sqrt{\frac{\sigma}{t}}\right)\right]. \tag{2.5}$$

Outra importante medida usada na caracterização de modelo de sobrevivência é a função taxa de falha (HR). A HR mensura a taxa instantânea de falha da variável, em um determinado tempo t, conforme define Giolo & Colosimo (2006). Para a distribuição BS, podemos obter HR a partir das funções dadas nas equações (2.3) e (2.5), que é dada por

$$h_T(t;\alpha,\sigma) = \frac{f_T(t;\alpha,\sigma)}{S_T(t;\alpha,\sigma)} = \frac{\frac{\exp(\alpha^{-2})}{2\alpha\sqrt{2\pi\sigma}}\exp\left(\frac{1}{2\alpha^2}\left(\frac{t}{\sigma} + \frac{\sigma}{t}\right)\right)t^{-\frac{3}{2}}(t+\sigma)}{1 - \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\sigma}} - \sqrt{\frac{\sigma}{t}}\right)\right]}. \tag{2.6}$$

A Figura 2.1 nos mostra a flexibilidade da distribuição BS, nela encontramos as funções; densidade, sobrevivência e taxa de falha para diferentes valores do parâmetro α. Como mencionado anteriormente, temos que α é o parâmetro de forma e conforme aumentamos o seu valor a assimetria da BS aumenta, ver Figura 2.1. Observamos também que a HR dada na equação (2.6) assume valor zero em $t = 0$, não possui um comporta-

mento monótono, cresce até um valor máximo e em seguida decresce até um certo tempo t.

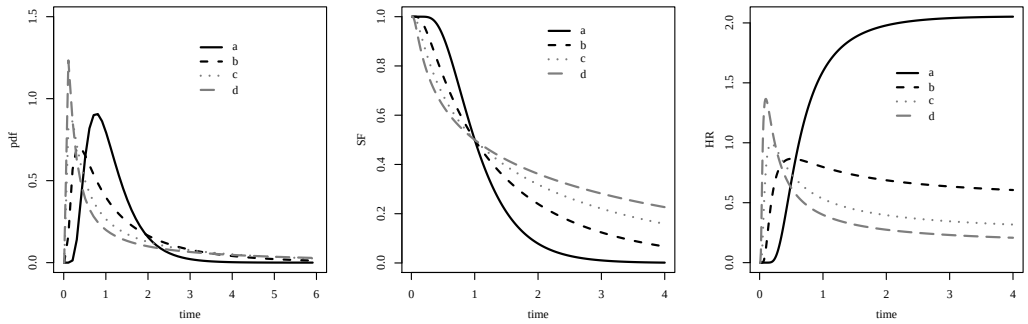


Figura 2.1: Plot da PDF, SF e HR da distribuição Birnbaum-Saunders $(\alpha, \sigma = 1)$.

2.2 Distribuição Generalizada Birnbaum-Saunders

A GBS é definida em termos de distribuições simétricas em $\mathbb{R}$, também conhecidas como distribuições univariadas de contornos elípticos; veja Fang (1990) para mais detalhes. Assim, a variável T segue uma distribuição GBS, denotada por $T \sim GBS(\alpha, \sigma, f)$, em que α é o parâmetro de forma e σ é o parâmetro de escala, e f é uma PDF associada a uma densidade simétrica. Especificamente, temos que

$$T = \frac{\sigma}{4}\left[\alpha z + \sqrt{\alpha^2 z^2 + 4}\right] \sim GBS(\alpha, \sigma, f),$$

com

$$z = \frac{1}{\alpha}\left[\sqrt{\frac{T}{\sigma}} - \sqrt{\frac{\sigma}{T}}\right],$$

em que z segue uma distribuição simétrica em $\mathbb{R}$, que é denotado $z \sim S(f)$. A PDF de T é

$$f_T(t) = f\left(\alpha^{-1}\left[\sqrt{\frac{t}{\sigma}} - \sqrt{\frac{\sigma}{t}}\right]\right)(2\alpha\sigma)^{-1}\left[\sqrt{\frac{\sigma}{t}} - \left(\sqrt{\frac{\sigma}{t}}\right)^3\right], \tag{2.7}$$

em que $f \geq 0, \alpha > 0, \sigma > 0$. Algumas propriedades da distribuição GBS são apresentadas a seguir (veja Sanhueza *et al.* (2008)):

i) Se $T \sim GBS(\alpha, \sigma, f)$ então, $aT \sim GBS(\alpha, a\sigma, f)$, para $a > 0$;

ii) Se $T \sim GBS(\alpha, \sigma, f)$, então a variável aleatória $T^{-1} \sim GBS(\alpha, \sigma^{-1}, f)$, ou seja, a distribuição GBS é fechada sob reciprocidade.

A motivação para o uso da distribuição GBS deve-se ao fato de conter uma família de distribuições mais flexíveis, por possuírem caudas mais ou menos pesadas, quando comparado a distribuição BS. Díaz-García & Leiva-Sánchez (2005) consideraram diferentes distribuições para a densidade f. Por exemplo: Laplace, logística, normal (que cai na distribuição BS), t-Student, entre outras, veja a Tabela 2.1 abaixo.

Tabela 2.1: Se $T \sim GBS(\alpha,\sigma;f)$, com $\alpha > 0$ e $\sigma > 0$. Então, para as distribuições indicadas, a PDF de T é dada por :

Distribuição	$f_T(t)$
BS	$\frac{1}{\sqrt{8\pi}\alpha\sigma}\exp\left(-\frac{1}{2\alpha^2}\left[\frac{t}{\sigma}+\frac{\sigma}{t}-2\right]\right)\left(\left[\frac{t}{\sigma}\right]^{-\frac{1}{2}}+\left[\frac{t}{\sigma}\right]^{-\frac{3}{2}}\right),\ t>0.$
Laplace	$\frac{1}{4\alpha\sigma}\exp\left(-\frac{1}{\alpha}\left\lvert\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right\rvert\right)\left(\left[\frac{t}{\sigma}\right]^{-\frac{1}{2}}+\left[\frac{t}{\sigma}\right]^{-\frac{3}{2}}\right),\ t>0.$
Logístico	$\frac{\exp\left(\frac{1}{\alpha}\left[\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right]\right)}{\left[1+\exp\left(\frac{1}{\alpha}\left[\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right]\right)\right]^2}\left[\sqrt{\frac{\sigma}{t}}-\sqrt{\frac{\sigma^3}{t^3}}\right],\ t>0.$
Exponencial Potência	$\frac{sr^{\frac{1}{2s}}\alpha^{-1}}{2\sigma\Gamma\left(\frac{1}{2s}\right)}\exp\left(-\frac{r}{\alpha}\left\lvert\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right\rvert^{2s}\right)\left(\left[\frac{\sigma}{t}\right]^{-\frac{1}{2}}+\left[\frac{\sigma}{t}\right]^{-\frac{3}{2}}\right),\ t>0,\ r,s>0.$
Cauchy	$\frac{1}{2\pi\alpha\sigma}\left(1+\frac{1}{\alpha^2}\left[\frac{t}{\sigma}+\frac{\sigma}{t}-2\right]\right)^{-1}\left(\left[\frac{t}{\sigma}\right]^{-\frac{1}{2}}+\left[\frac{t}{\sigma}\right]^{-\frac{3}{2}}\right),\ t>0.$

Em particular, se z tem uma distribuição t-Student, onde ν representa os graus de liberdade, denotado por $z \sim t_\nu$. Então

$$T = \frac{\sigma}{4}\left[\alpha z + \sqrt{\alpha^2 z^2 + 4}\right] \sim GBS(\alpha,\sigma,t_\nu),$$

onde a PDF é dada por

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{2\alpha\sqrt{\sigma\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu\alpha^2}\left(\frac{t}{\sigma}+\frac{\sigma}{t}-2\right)\right]^{-(\nu+1)/2}\frac{(t+\sigma)}{\sqrt{t^3}}, \tag{2.8}$$

com $t > 0$.

A CDF da variável aleatória T é dada por

$$F_T(t;\alpha,\sigma,\nu) = \Phi_t\left(\alpha^{-1}\left[\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right]\right),$$

em que Φ_t denota a CDF da distribuição t_ν. As funções SF e HR da variável aleatória T é expressa por

$$S_T(t;\alpha,\sigma,\nu) = \Phi_t\left(-\frac{1}{\alpha}\left[\sqrt{\frac{t}{\sigma}}-\sqrt{\frac{\sigma}{t}}\right]\right)$$

e

$$h_T(t;\alpha,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\left[1+\frac{1}{\nu\alpha^2}\left(\frac{t}{\sigma}+\frac{\sigma}{t}-2\right)\right]^{-(\nu+1)/2}(t+\sigma)}{2\alpha\Gamma(\nu/2)\sqrt{\sigma\pi\nu}\Phi_t\left(\alpha^{-1}\left[\sqrt{t/\sigma}-\sqrt{\sigma/t}\right]\right)\sqrt{t^3}},\ t>0. \qquad (2.9)$$

O valor esperado e a variância da variável T são dadas respectivamente por

$$E(T) = \sigma\left[1+\frac{\alpha^2}{2}\left(\frac{\nu}{\nu-2}\right)\right],$$

com $\nu > 2$ e

$$Var(T) = (\sigma\alpha)^2\left[\left(\frac{\nu}{\nu-2}+\frac{5}{4}\frac{\alpha^2\nu^2\left(\nu-\frac{8}{5}\right)}{(\nu-4)(\nu-2)^2}\right)\right],$$

com $\nu > 4$.

A Figura 2.2 mostra a função densidade $GBS(\alpha = 0.5, \sigma = 1, t_\nu)$, também denotado por BS-$t(\alpha = 0.5, \sigma = 1, \nu)$, com diferentes valores para o grau de liberdade ν. Observa-se que a medida que o parâmetro ν cresce, a distribuição BS-t tende a distribuição BS. A Figura 2.2 mostra o comportamento da HR para a distribuição BS-t dada pela equação (2.9). Observa-se que essa função não é monótona, assume valor zero no tempo $t = 0$, atinge seu valor máximo, depois decresce e estabiliza em um determinado tempo. Além disso, notamos também que quanto maior for o parâmetro ν, maior será a HR.

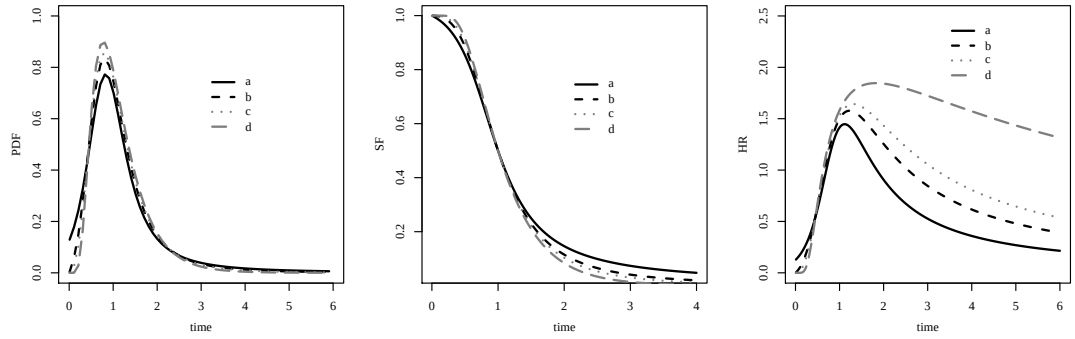


Figura 2.2: Plot da PDF, SF e HR da distribuição $GBS(\alpha = 0.5, \sigma = 1, t_\nu)$.

2.3 O modelo tobit

Considere $\{Y_1, \ldots, Y_m, Y_{m+1}, \ldots, Y_n\}$ uma amostra aleatória de tamanho n da variável Y com censura à esquerda (de zero, ou de algum limite de detecção mínimo), com observações independentes (ID), não necessariamente independentes e identicamente distribuídas

(IID). Assumimos ainda que esta amostra possui m observações censuradas à esquerda, e $(n-m)$ observações (completas ou não censuradas). Desta forma, este esquema de visualização da censura pode ser utilizado para construção do modelo de regressão com uma variável resposta censurada Y^*, que é chamada de variável latente (não observada). Assim, os dados censurados m (não observados) correspondentes aos valores de Y^* menores ou iguais a um ponto limitante y_0 (censura à esquerda), de maneira que todas essas observações tomam o valor y_0. Os outros dados $(n-m)$(observados) estão ligados com os valores de Y^* maior que y_0, que pode ser escrita por uma estrutura de regressão linear do tipo $\mathbf{x}_i^\top \boldsymbol{\beta}$. Essa estrutura de modelagem, apresentada pelo modelo tobit normal com resposta censurada à esquerda é dada da seguinte forma

$$Y_i = \begin{cases} y_0, & Y_i^* \leq y_0, \ \ i = 1, \ldots, m, \\ \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, & Y_i^* > y_0, \ \ i = m+1, \ldots, n, \end{cases} \tag{2.10}$$

em que $\varepsilon_i \sim N(0, \sigma^2)$ representa o erro aleatório do modelo, $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão, que representa os parâmetros desconhecidos a serem estimados, e $\mathbf{x}_i$ é um vetor contendo os valores das covariáveis. Observe que y_0 dado em (2.10) é um valor limitante pré-fixado, responsável por tornar a resposta do modelo de regressão definido em (2.10) ser limitado (ou censurado), como estabelecido originalmente por Tobin (1958).

Observe a similaridade existente entre o modelo probit normal (ver e.g, Aldrich & Nelson (1984)) e o modelo tobit normal descrito em (2.10). No modelo probit normal a resposta é uma variável latente (não observada) descrita por

$$Y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2.11}$$

em que $\mathbf{x}_i^\top$, $\boldsymbol{\beta}$ e ε_i são definido analogamente em (2.10). Como não é possível observar a variável latente Y_i^*, a variável indicadora é

$$Y_i = \begin{cases} 0, & Y_i^* \leq y_0, \ \ i = 1, \ldots, m, \\ 1, & Y_i^* > y_0, \ \ i = m+1, \ldots, n. \end{cases} \tag{2.12}$$

Os modelos probit e tobit são os mesmos para a variável latente Y^*, mas os modelos para a variável resposta observada Y são diferentes. No modelo tobit, sabemos o valor de Y^* quando $Y^* > y_0$, enquanto que no modelo probit apenas sabemos que $Y^* > y_0$, mas

não conhecemos o seu valor. Desta forma, há mais informação no modelo tobit, do que o modelo probit. Além disso, as estimativas de β pelo modelo tobit são mais eficientes que pelo modelo probit. Além das vantagens dos modelos censurados permitirem estimar a variação de Y^* que não é possível para observações censuradas pelo modelo probit (De Sousa *et al.* (2018)). Veja Scott Long (1997) para mais detalhes sobre logit, probit e tobit.

Capítulo 3

Modelo de regressão tobit-BS-*t*

Neste Capítulo introduzimos o modelo de regressão tobit em que a variável resposta possui distribuição GBS, ou seja, vamos considerar o caso em que a variável com distribuição tobit-GBS possa ser modelada por covariáveis explicativas e parâmetros desconhecidos. Em seguida, apresentamos a parte inferencial para um caso particular, o modelo tobit-BS-*t*. Descrevemos a estimação dos parâmetros do modelo proposto, além disso, alguns resultados assintóticos bem como a função escore e a matriz hessiana são mostrados. Por fim, apresentamos algumas medidas de avaliação residual e alguns métodos de análise diagnóstica baseados na influência local proposta por Cook (1986), porém agora baseada na distribuição GBS, conforme descrita em Leiva *et al.* (2009).

3.1 Modelo tobit-BS-*t*

Considere a representação para o modelo BS-*t*

$$T_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\delta_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

em que T_i é a variável resposta e $\delta_i \sim GBS(\alpha, 1, t_\nu)$ é o erro do modelo. Então, com base na propriedade (i) da GBS, temos que $T_i = \exp(\mathbf{x}^\top \boldsymbol{\beta}) \sim GBS(\alpha, \exp(\mathbf{x}^\top \boldsymbol{\beta}), t_\nu)$. Aplicando o logaritmo em (3.1), nós obtemos

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3.2}$$

em que $Y_i = \log(T_i)$ é a log-resposta, $\mathbf{x}_i^\top$ é o vetor de covariáveis, $\boldsymbol{\beta}$ é um vetor de coeficientes associados às covariáveis e $\varepsilon_i = \log(\delta_i) \sim \log-\text{GBS}(\alpha, 0, t_\nu)$.

Com base nessa descrição da distribuição BS-t, que é dada pela equação (2.8), temos o seguinte resultado: se $Y = \log(T)$, em que $T \sim$ BS-$t(\alpha, \beta, \nu)$, então $Y = \log(T) \sim SH(\alpha, \gamma = \log(\beta), 2, t_\nu) \equiv \log-\text{BS-}t(\alpha, \beta, \nu)$, em que SH é a notação para distribuições seno-hiperbólicos (veja e.g. Dıaz-Garcıa & Domınguez-Molina (2006) e Cancho *et al.* (2010)). Portanto a PDF de Y é dada por

$$f_Y(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\alpha(\nu\pi)^{1/2}\Gamma(\nu/2)} \cosh\left(\frac{y-\mu}{2}\right) \left\{1 + \frac{4}{\nu\alpha^2}\left[\sinh\left(\frac{y-\mu}{2}\right)\right]^2\right\}^{-(\nu+1)/2}, \quad (3.3)$$

em que $y \in \mathbb{R}$ e $\mu \in \mathbb{R}$. A CDF e a SF de Y são dadas respectivamente por,

$$F_Y(y) = \Phi_t\left[\frac{2}{\alpha}\sinh\left(\frac{y-\mu}{2}\right)\right]$$

e

$$S_Y(y) = \Phi_t\left[-\frac{2}{\alpha}\sinh\left(\frac{y-\mu}{2}\right)\right],$$

em que Φ_t é a CDF da distribuição t-Student. Os parâmetros α e μ são responsáveis pela forma e locação (média), respectivamente. Enquanto ν é o grau de liberdade da distribuição t, responsável pela forma da curtose da distribuição. Além disso, α está relacionada com a modalidade da distribuição, conforme descreve Barros *et al.* (2008) .

Desta forma, baseado na definição do modelo tobit, nós obtemos o modelo tobit-BS-t

$$Y_i = \begin{cases} y_0, & y_i^* \leq y_0, \ \ i = 1, \ldots, m, \\ \mathbf{x}_i^\top\boldsymbol{\beta} + \varepsilon_i, & y_i^* > y_0, \ \ i = m+1, \ldots, n, \end{cases} \quad (3.4)$$

em que $y_i^* = \log T_i^*$, T_i^* é a variável latente observada para valores maiores que y_0 e censurada caso contrário, $\boldsymbol{\beta}$, $\mathbf{x}$ e ε_i são definidos anteriormente.

3.1.1 Estimação

A estimação dos parâmetros do modelo tobit-BS-$t(\alpha, \boldsymbol{\beta}, \nu)$, definido em (3.4), é realizada pelo método de máxima verossimilhança (ML). A função de log-verossimilhança

do modelo proposto com vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})^\top$ obtido de (3.3), é dado por

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= (n-m)\log\Gamma\left(\frac{\nu+1}{2}\right) - (n-m)\log(2) - \frac{(n-m)}{2}\log(\pi\nu) \\ &\quad -(n-m)\log\Gamma\left(\frac{\nu}{2}\right) + \sum_{i=m+1}^{n}\log(\xi_{i1}) - \frac{(\nu+1)}{2}\sum_{i=m+1}^{n}\log\left(\nu+\xi_{i2}^2\right) \\ &\quad + \frac{(\nu+1)}{2}(n-m)\log(\nu) + \sum_{i=1}^{m}\log\left[\Phi_t(\xi_{i2}^c)\right], \end{aligned} \tag{3.5}$$

em que

$$\xi_{i1} = \cosh\left(\frac{y_i - \mathbf{x}_i^\top\boldsymbol{\beta}}{2}\right), \;\; \xi_{i2} = \frac{2}{\alpha}\sinh\left(\frac{y_i - \mathbf{x}_i^\top\boldsymbol{\beta}}{2}\right) \text{ e } \;\; \xi_{i2}^c = \frac{2}{\alpha}\sinh\left(\frac{y_0 - \mathbf{x}_i^\top\boldsymbol{\beta}}{2}\right).$$

Para obter o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ é necessário maximizar a função de log-verossimilhança dada em (3.5). O vetor escore é $\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = (\dot{\ell}_\alpha, \dot{\ell}_{\boldsymbol{\beta}}^\top)^\top$ que contém as primeiras derivadas da função de log-verossimilhança (3.5), onde

$$\dot{\ell}_\alpha = \begin{cases} -\frac{\lambda(\xi_{i2}^c)\xi_{i2}^c}{\alpha}, & i = 1,\ldots,m, \\ \frac{w(\xi_{i2}^2)\xi_{i2}^2 - 1}{\alpha}, & i = m+1,\ldots,n, \end{cases} \tag{3.6}$$

e

$$\dot{\ell}_{\boldsymbol{\beta}} = \begin{cases} -\frac{\mathbf{x}_i^\top}{\alpha}\lambda(\xi_{i2}^c)\cosh(\delta_i), & i = 1,\ldots,m, \\ -\frac{\mathbf{x}_i^\top}{2}\tanh(\delta_i) + \frac{\mathbf{x}_i^\top}{\alpha^2}w(\xi_{i2}^2)\sinh(2\delta_i), & i = m+1,\ldots,n, \end{cases} \tag{3.7}$$

em que $\delta_i = (y_0 - \mathbf{x}_i^\top\boldsymbol{\beta})/2$, com $i = 1,\ldots,m$, $\delta_i = (y_i - \mathbf{x}_i^\top\boldsymbol{\beta})/2$, com $i = m+1,\ldots,n$ e $\lambda(\xi_{i2}^c) = \phi_t(\xi_{i2}^c)/\Phi_t(\xi_{i2}^c)$ e $w(\xi_{i2}^2) = (\nu+1)/(\nu+\xi_{i2}^2)$. O estimador ML de $\boldsymbol{\theta}$ é obtido igualando as equações (3.6) e (3.7) a zero. Porém, os dois sistemas de equações não apresentam uma solução analítica fechada, sendo necessário a utilização de métodos numéricos para maximizarmos o logaritmo da função de verossimilhança. Leiva *et al.* (2007) sugerem o uso do algoritmo quasi-Newton, Broyden-Fletcher-Goldfarb-Shanno (BFGS), usando como valores iniciais para o procedimento numérico, $\alpha^2 = 4(\sinh((y_i - \mathbf{x}^\top\boldsymbol{\beta})/2)^2)/(n-m)$ e $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, onde $\boldsymbol{X}$ é composta pelas linhas com $\mathbf{x}_i$.

3.1.2 Inferência

A inferência assintótica para o vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \beta)^\top$, é baseado na distribuição normal p-variada. Sob as condições de regularidades definidas em Hinkley & Cox (1979), os estimadores de máxima verossimilhança $\hat{\alpha}$ e $\widehat{\beta}$ são consistentes e possuem distribuição assintótica normal multivariada. Esta distribuição tem um vetor de média assintótica, com elementos α e β e tem uma matriz de covariância assintótica igual a $J(\boldsymbol{\theta})^{-1}$, em que pode ser aproximado pela esperança da matriz de Informação de Fisher (IF). Então, para $n \to \infty$, nós temos que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_{(p+1)}\left[\mathbf{0}_{(p+1)}, J(\boldsymbol{\theta})^{-1}\right],$$

em que a média converge em distribuição para $\mathbf{0}_{(p+1)}$, um vetor de zeros $(p+1) \times 1$, e

$$J(\boldsymbol{\theta}) = \lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{\theta}),$$

com $I(\boldsymbol{\theta})$ sendo o valor esperado da matriz IF. Observe que $I(\boldsymbol{\theta})^{-1}$ é um estimador consistente para a matriz de variância-covariância assintótica de $\hat{\boldsymbol{\theta}}$, dito $J(\boldsymbol{\theta})^{-1}$. Na prática, pode-se aproximar o valor esperado da matriz IF, pela sua versão aproximada (ver e.g, Efron & Hinkley (1978)), considerando os elementos da matriz de informação inversa observada, usada para aproximar os erros padrões (SEs) correspondentes. A matriz IF observada é obtida a partir da matriz hessiana, que contém as derivadas de segunda ordem da equação (3.5), dada por

$$\boldsymbol{\ell} = \begin{pmatrix} \text{tr}(\boldsymbol{G}) & \mathbf{k}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{k} & \mathbf{X}^\top V \mathbf{X} \end{pmatrix}, \tag{3.8}$$

em que $\boldsymbol{V} = \text{diag}(\boldsymbol{v}_1(\boldsymbol{\theta}), \boldsymbol{v}_2(\boldsymbol{\theta}), \boldsymbol{v}_3(\boldsymbol{\theta}), \ldots, \boldsymbol{v}_n(\boldsymbol{\theta}))$, $\boldsymbol{k} = (k_1(\boldsymbol{\theta}), k_2(\boldsymbol{\theta}), k_3(\boldsymbol{\theta}), \ldots, k_n(\boldsymbol{\theta}))^\top$, e $\boldsymbol{G} = \text{diag}(\boldsymbol{g}_1(\theta), \boldsymbol{g}_2(\theta), \boldsymbol{g}_3(\theta), \ldots, \boldsymbol{g}_n(\theta))$, com

$$
\begin{aligned}
v_i(\boldsymbol{\theta}) &= \begin{cases} \frac{\lambda'(\xi_{i2}^c)[\cosh(\delta_i)]^2}{\alpha^2} + \frac{\lambda(\xi_{i2}^c)\xi_{i2}^c}{4}, & i = 1,\ldots,m, \\ \frac{(\operatorname{sech}(\delta_i))^2}{4} - \frac{2w'(\xi_{i2}^2)}{\alpha^4}(\sinh(2\delta_i))^2 - \frac{w(\xi_{i2}^2)\cosh(2\delta_i)}{\alpha^2}, & i = m+1,\ldots,n, \end{cases} \\
k_i(\boldsymbol{\theta}) &= \begin{cases} \frac{\lambda'(\xi_{i2}^c)\sinh(2\delta_i)}{\alpha^3} + \frac{\lambda(\xi_{i2}^c)\cosh(\delta_i)}{\alpha^2}, & i = 1,\ldots,m, \\ -\frac{2\sinh(2\delta_i)}{\alpha^3}[w'(\xi_{i2}^2)\xi_{i2}^2 + w(\xi_{i2}^2)], & i = m+1,\ldots,n, \end{cases} \\
g_i(\boldsymbol{\theta}) &= \begin{cases} \frac{1}{\alpha}[\lambda'(\xi_{i2}^c)(\xi_{i2}^c)^2 + 2\xi_{i2}^c\lambda(\xi_{i2}^c)], & i = 1,\ldots,m, \\ \frac{1}{\alpha^2} - \frac{2}{\alpha^2}w'(\xi_{i2}^2)\xi_{i2}^4 - \frac{3}{\alpha^2}w(\xi_{i2}^2)\xi_{i2}^2, & i = m+1,\ldots,n, \end{cases}
\end{aligned}
$$

em que $\lambda'(\cdot)$ e $w'(\cdot)$ são as derivadas de $\lambda(\cdot)$ e $w(\cdot)$ dadas na equação (3.7), respectivamente.

3.2 Análise de resíduo e diagnóstico

A análise residual tem como objetivo avaliar as suposições feitas sobre o modelo, e detectar observações atípicas. De acordo com Paula (2015) a análise residual é uma etapa importante para verificar o ajuste do modelo, e os possíveis afastamentos dos pressupostos. Em modelos de regressão normalmente são utilizados os resíduos padronizados e os de Pearson, contudo de acordo com Barros *et al.* (2010) em aplicações de modelos tobit, mesmo sobre normalidade, esses tipos de resíduos parecem não ser adequados. Para o modelo tobit-BS-*t* utilizamos o resíduo de Cox-Snell generalizado (CSG) que é dado por

$$
r_i^{CSG} = -\log\left(\hat{S}_Y(y_i;\alpha,\boldsymbol{\beta},\mathbf{x})\right), \quad i = 1,\ldots,n,
$$

em que $\hat{S}_Y$ é o estimador da função de sobrevivência, no modelo definido em (3.4). Observe que o estimador da SF do modelo log-GBS, avaliado no caso *i*, é dado por

$$
S_Y(y_i;\alpha,\boldsymbol{\beta},\mathbf{x}) = \Phi_t\left(-\frac{2}{\widehat{\alpha}}\sinh\left(\frac{y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}}{2}\right)\right), \quad i = 1,\ldots,n. \tag{3.9}
$$

Uma observação importante sobre o resíduo CSG, é que independente da especificação do modelo, ele segue uma distribuição Exp (1) (ver e.g. Bhatti (2010); Leiva *et al.* (2014b)).

3.2.1 Influência Global

Uma medida muito usada para avaliar a exclusão da i-ésima observação é o desvio entre log-verossimilhanças (LD), definido por

$$LD_i(\boldsymbol{\theta}) = 2\left[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{(i)})\right], \tag{3.10}$$

em que $\widehat{\boldsymbol{\theta}}_{(i)}$ denota o estimador ML obtido após a eliminação da i-ésima observação, $i = 1,\ldots,n$. A ideia básica desse método é comparar a diferença entre $\widehat{\boldsymbol{\theta}}_{(i)}$ e $\widehat{\boldsymbol{\theta}}$, e avaliar se $\widehat{\boldsymbol{\theta}}_{(i)}$ está de alguma forma afastado de $\widehat{\boldsymbol{\theta}}$, então a i-ésima observação é considerada como influente. Com base nessa medida de comparação entre $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_{(i)}$, Cook (1977) propôs outra métrica chamada distância de Cook, usada na verificação do efeito que cada caso produz na estimação dos parâmetros. Cook (1977) sugeriu a eliminação de cada caso e a avaliação da função de log-verossimilhança para o caso i removido. A distância de Cook generalizada (GCD) para o modelo tobit é mostrada por Barros *et al.* (2018), e possui a seguinte forma

$$GCD_i(\boldsymbol{\theta}) = \frac{1}{p+1}\left[(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})\ddot{\ell}^{-1}(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})\right], \ \ i = 1,\ldots,n, \tag{3.11}$$

em que p é o número de coeficientes da regressão do modelo. Para facilitar os cálculos, usa-se as aproximações de primeira ordem em $\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{(i)} \approx \ddot{\ell}_{(i)}^{-1}\dot{\ell}_{(i)}$ na equação (3.11) e reescrevemos

$$GCD_i(\boldsymbol{\theta}) = \frac{1}{p+1}\left[\dot{\ell}_{(i)}^{\top}\ddot{\ell}_{(i)}^{-1}(-\ddot{\ell})\ddot{\ell}_{(i)}^{-1}\dot{\ell}_{(i)}\right], \ \ i = 1,\ldots,n, \tag{3.12}$$

em que $\dot{\ell}_{(i)}$ e $\ddot{\ell}_{(i)}$ são o vetor escore e a matriz hessiana do modelo definida nas equações (3.7) e (3.8), respectivamente, avaliados $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ e considerando a eliminação do i-ésimo caso. Normalmente, a análise diagnóstica é realizada no vetor de coeficientes $\boldsymbol{\beta}$. Nesse caso, temos

$$GCD_i(\boldsymbol{\beta}) = \frac{1}{p}\left[(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})\ddot{\ell}^{-1}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})\right], \ \ i = 1,\ldots,n. \tag{3.13}$$

Alguns autores sugeriram considerar um caso i como potencialmente influente em $\boldsymbol{\beta}$, se o valor da GCD para o caso i for maior que $2/n$, nesse caso a observação é considerada

potencialmente influente na estimação do vetor de parâmetros, veja mais detalhes em Zhu & Zhang (2004); Barros *et al.* (2010); Barros *et al.* (2018) e De Sousa *et al.* (2018).

3.2.2 Influência Local

O método de influência local tem o objetivo de verificar a existência de observações (pontos) que, sob pequenas modificações, causam alguma interferência nos resultados do ajuste proposto, ou seja, pontos que estejam causando afastamento das suposições feitas no modelo (Paula (2015)). Este método utiliza a análise do gráfico de influência baseado no conceito de curvatura normal, bastante conhecido em literaturas de geometria diferencial. Essa técnica usa o afastamento pela função de verossimilhança em torno de um ponto particular (ver e.g. Cook (1986)). Esse método não necessita de nenhuma eliminação. Existem muitas maneiras de realizar a análise de influência local, vamos usar ponderações de casos, pertubação de casos na variável resposta e na covariável. Seja $\ell(\boldsymbol{\theta})$ a função de log-verossimilhança, em que $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^\top$ é o vetor de parâmetros em interesse. Denotamos $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ a função de log-verossimilhança, definida em (3.5) para o modelo perturbado, em que $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)^\top$ é um vetor de pertubação em um subconjunto $\Omega \in \mathbb{R}^n$. Com o objetivo de avaliar a influência da pertubação sobre as estimativas de $\boldsymbol{\theta}$, Cook (1977) propôs uma generalização da LD definida em (3.10), o afastamento pela verossimilhança $LD(\boldsymbol{\omega})$, definido dado por

$$LD(\boldsymbol{\omega}) = 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})), \tag{3.14}$$

em que $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denota o estimador ML sobre o modelo $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$.

Com o interesse de estudar o comportamento da função $LD(\boldsymbol{\omega})$ em torno de uma vizinhança, Cook (1986) sugeriu o estudo da curvatura normal em uma vizinhança no ponto $\boldsymbol{\omega}_0 = (1, 1, \ldots, 1)^\top$, chamado de vetor de não perturbação, na direção arbitrária de um vetor unitário $\imath$, com $\| \imath \| = 1$. Em geral, a curvatura normal possui a seguinte forma,

$$C_\ell(\boldsymbol{\theta}) = 2|\imath^\top \boldsymbol{\Delta}^\top \ddot{\ell}^{-1} \boldsymbol{\Delta} \imath|,$$

em que $\boldsymbol{\Delta}$ é uma matriz de perturbação $(p \times n)$ e $\ddot{\ell}$ é a matriz hessiana, apresentada na equação (3.8). A matriz Δ, depende do esquema de perturbação usado e seus elementos são dados por

$$\Delta_{ij} = \frac{\partial \ell(\theta|\omega)}{\partial \theta_i \partial \omega_j},$$

para $i = 1,\ldots,n$ e $j = 1,\ldots,p+1$ avaliados em $\theta = \widehat{\theta}$ e $\omega = \omega_0$. Com o objetivo de buscar observações influentes sob pequenas perturbações, Barros *et al.* (2010) propuseram o gráfico baseado no autovalor ι_{max} associado à curvatura normal, que pode ser obtido usando o autovalor máximo de

$$B(\theta) = \mid \Delta^\top \ddot{\ell}^{-1} \Delta \mid . \tag{3.15}$$

Quando o interesse é avaliar a influência parcial do vetor de parâmetros β, então (3.15) fica dado por

$$C_\iota(\theta) = 2 \mid \iota^\top \Delta^\top (\ddot{\ell}^{-1} - B_1) \Delta \iota \mid . \tag{3.16}$$

Observa-se que, no caso de (3.16), o parâmetro α é removido da análise. Portanto, a verificação de pontos influentes é realizada somente sob β e B_1 assume a forma

$$B_1 = \begin{pmatrix} \mathrm{tr}(\boldsymbol{G})^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.17}$$

Da mesma forma, se o interesse é apenas em α, então (3.16) é expressa por

$$C_\iota(\theta) = 2 \mid \iota^\top \Delta^\top (\ddot{\ell}^{-1} - B_2) \Delta \iota \mid, \tag{3.18}$$

em que B_2 para o modelo é dado por

$$B_2 = \begin{pmatrix} 0 & 0 \\ 0 & (\mathrm{X}^\top \mathrm{V} \mathrm{X})^{-1} \end{pmatrix}, \tag{3.19}$$

em que tanto tr(G) e $(\mathrm{X}^\top \mathrm{V} \mathrm{X})^{-1}$ são obtidos a partir da matriz hessiana para o modelo, dada por (3.8).

Além de considerar a direção do vetor máximo da curvatura normal $\boldsymbol{\iota}_{\max}$, uma im-

portante direção pode ser utilizada para avaliar a influência local em $\widehat{\theta}$. O vetor $\imath = e_{in}$ corresponde à direção da i-ésima observação, em que e_{in} é um vetor $n \times 1$, com base canônica de $\mathbb{R}^n$, que assume o valor um. Denotamos a curvatura normal, com o objetivo de verificar a influência local total do i-ésimo caso dada por

$$C_i(\theta) = 2 \mid b_{ii} \mid, \;\; i = 1, \ldots, n,$$

em que b_{ii} representa o i-ésimo elemento da matriz definida em (3.15) para cada caso especificado. Verbeke & Molenberghs (2000) propõem considerar que uma observação é potencialmente influente em θ, se $C_i(\theta) > 2C(\theta)$, em que

$$C(\widehat{\theta}) = \frac{1}{n}\sum_{=1}^{n} C_i(\theta),$$

como potenciais pontos influentes, veja detalhes em Zhu & Zhang (2004).

Pertubação de casos: Este esquema tem como objetivo avaliar se casos com diferentes pesos causam algum impacto nas estimativas ML do parâmetro θ. Este esquema é um dos mais utilizados na avaliação de influencia sobre um modelo. Considere a função de log-verossimilhança $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \ell_i$, com ℓ_i dado em (3.5) e $\boldsymbol{\omega} \in [0,1]$. Tomando a derivada parcial com relação a $\boldsymbol{\omega}^\top$, obtemos,

$$\frac{\partial \ell(\theta|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}^\top} = \sum_{i=1}^{n} \ell_i(\theta) e_{in}^\top,$$

em que $e_{in}^\top$ é um vetor $(n \times 1)$. Após avaliar θ a $\widehat{\theta}$ e $\boldsymbol{\omega}$ a $\widehat{\omega}_0$, obtemos a seguinte matriz de pertubação para este esquema

$$\Delta = \sum_{i=1}^{n} h_i e_{in}^\top, \tag{3.20}$$

em que h_i é dado por

$$h_i = \begin{pmatrix} \frac{\partial \ell_i(\theta)}{\partial \alpha} \\ \frac{\partial \ell_i(\theta)}{\partial \beta} \end{pmatrix}.$$

Através da função de log-verossimilhança definida em (3.5), obtemos uma expressão explicita para h_i, dada por

$$\dot{\ell}_{\alpha_i}(\theta|\boldsymbol{\omega}) = a_i = \begin{cases} -\frac{\lambda(\xi_{i2}^c)\xi_{i2}^c}{\alpha}, & i = 1,\ldots,m, \\ \frac{1}{\alpha}[w(\xi_{i2}^2)\xi_{i2}^2 - 1], & i = m+1,\ldots,n, \end{cases} \tag{3.21}$$

e

$$\dot{\ell}_{\beta_i}(\theta|\boldsymbol{\omega}) = b_i = \begin{cases} -\frac{1}{\alpha}\lambda(\xi_{i2}^c)\cosh(\delta_i), & i = 1,\ldots,m, \\ -\frac{1}{2}\tanh(\delta_i) + \frac{1}{\alpha^2}w(\xi_{i2}^2)\sinh(2\delta_i), & i = m+1,\ldots,n. \end{cases} \tag{3.22}$$

A matriz de perturbação de casos ponderados dada por Δ e definida em (3.20), é decomposta em $\Delta_\alpha = (a_1,\ldots,a_n)$ e $\Delta_\beta = \mathrm{X}^\top diag\{b_1,\ldots,b_n\}$.

Perturbação na resposta: Muitas técnicas são usadas para considerar uma perturbação na variável resposta. Consideramos um esquema de perturbação na resposta, conhecida como perturbação aditiva, que é definida por $Y_{i\boldsymbol{\omega}} = Y_i + \boldsymbol{\omega}_i S_Y$, para $i = m+1,\ldots,n$, em que S_Y é um componente escalar que pode ser considerado, o desvio padrão (SD) da variável resposta. A função de log-verossimilhança do modelo sob perturbação na variável resposta é dada por

$$\ell(\theta|\boldsymbol{\omega}) \propto \begin{cases} \sum_{i=1}^m \boldsymbol{\omega}\log\left[\Phi_t(\xi_{i2}^c)\right], & i = 1,\ldots,m; \\ \sum_{i=m+1}^n \boldsymbol{\omega}\left[\log(\xi_{i1}) - \left\{\frac{(\nu+1)}{2}\right\}\log\left(\nu + \xi_{i2}^2\right)\right], & i = m+1,\ldots,n, \end{cases} \tag{3.23}$$

em que ξ_{i1}, ξ_{i2} e ξ_{i2}^c estão definidos em (3.7) depois de mudar Y por $Y_{i\boldsymbol{\omega}}$. Neste caso nós temos os seguintes elementos para a matriz Δ de perturbação na resposta,

$$\Delta_\alpha = (c_{m+1},\ldots,c_n),\ \ \Delta_\beta = \mathrm{X}^\top diag\{d_{m+1},\ldots,d_n\},$$

em que

$$c_i = \begin{cases} \frac{S_Y}{\alpha^2}\left\{\cosh(\delta_i)\lambda(\xi_{i2}^c) + \frac{1}{\alpha}\sinh(2\delta_i)\lambda'(\xi_{i2}^c)\right\}, & i = 1,\ldots,m, \\ \frac{S_Y}{\alpha}\left[\xi_{i1}\xi_{i2}w(\xi_{i2}^2) + \xi_{i1}\xi_{i2}^3 w'(\xi_{i2}^2)\right], & i = m+1,\ldots,n, \end{cases}$$

e

$$d_i = \begin{cases} \frac{S_Y}{4}\left\{\xi_{i2}^c\lambda(\xi_{i2}^c) + \frac{4}{\alpha^2}(\cosh(\delta_i))^2\lambda'(\xi_{i2}^c)\right\} & , \ i = 1,\ldots,m, \\ S_Y\left[\frac{1}{\alpha^2}\cosh(2\delta_i)w(\xi_{i2}^c) - \frac{1}{4}(\text{sech}(\delta_i))^2 - \frac{2}{\alpha^4}\sinh(2\delta_i)w'(\xi_{i2}^2)\right] & , \ i = m+1,\ldots,n, \end{cases}$$

avaliados em $\theta = \widehat{\theta}$ e $\omega = \omega_0$. Vale observar que sob o esquema de perturbação na resposta em modelos tobit, só faz sentido a parte não censurada dos dados. Isso ocorre pelo fato de que a parte censurada está abaixo do limite y_0, nesse caso cada observação censurada recebe o mesmo valor y_0, conforme De Sousa *et al.* (2018).

Perturbação na covariável: Da mesma forma do esquema da variável resposta, consideramos o caso de perturbação aditiva na covariável tomando a seguinte forma

$$x_{it\omega} = x_{it} + \omega_i S_X, \ \ i = 1,\ldots,n,$$

em que S_X pode ser o SD da covariável correspondente x_t. Considere a mesma função de log-verossimilhança $\ell(\theta|\omega)$ mostrada no esquema anterior dada por (3.23). Obtemos a matriz Δ de perturbação na covariável, avaliando $\theta = \widehat{\theta}$ e $\omega = \omega_0$, com

$$\Delta_{\beta_{ij}} = \begin{cases} -\frac{S_X\beta_t x_{ij}}{4}\left\{\xi_{i2}^c\lambda(\xi_{i2}^c) + \frac{4}{\alpha^2}(\cosh(\delta_i))^2\lambda'(\xi_{i2}^c)\right\}, & i = 1,\ldots,m, \\ S_X\beta_t x_{ij}\left[\frac{1}{4}\text{sech}^2(\delta_i) - \frac{1}{\alpha^2}\cosh(2\delta_i)w(\xi_{i2}^2) - \frac{2}{\alpha^2}\sinh^2(2\delta_i)w'(\xi_{i2}^2)\right], & i = m+1,\ldots,n, \end{cases}$$

e para $j = t$ temos a forma

$$\Delta_{\beta_{ij}} = \begin{cases} -\frac{S_X\beta_t x_{ij}}{4}\left\{\xi_{i2}^c\lambda(\xi_{i2}^c) + \frac{4}{\alpha^2}(\cosh(\delta_i))^2\lambda'(\xi_{i2}^c)\right\} + \frac{S_X}{\alpha}\cosh(\delta_i)\lambda(\xi_{i2}^c), & i = 1,\ldots,m, \\ S_X\beta_t x_{ij}\left[\frac{1}{4}\text{sech}^2(\delta_i) - \frac{1}{\alpha^2}\cosh(2\delta_i)w(\xi_{i2}^2) - \frac{2}{\alpha^2}\sinh^2(2\delta_i)w'(\xi_{i2}^2)\right] & \\ -S_X\left[\frac{1}{\alpha^2}\sinh(2\delta_i)w(\xi_{i2}^2) - \frac{1}{2}\tanh(\delta_i)\right], & i = m+1,\ldots,n, \end{cases}$$

e $\Delta_\alpha = (\phi_1,\ldots,\phi_n)$, onde

$$\phi_i = \begin{cases} -\frac{S_X\beta_t}{\alpha^2}\left[\cosh(\delta)\lambda(\xi_{i2}^c) + \frac{1}{\alpha}\sinh(2\delta_i)\lambda(\xi_{i2}^c)\right], & i = 1,\ldots,m, \\ -\frac{2}{\alpha^3}S_X\beta_t\sinh(2\delta_i)\left[w(\xi_{i2}^2) + \xi_{i2}^2 w'(\xi_{i2}^2)\right], & i = m+1,\ldots,n. \end{cases}$$

Capítulo 4

Aplicações numéricas

A fim de avaliar o desempenho das estimativas ML do modelo de regressão tobit-BS-*t*, apresentamos uma breve avaliação numérica, através do estudo de simulação. Todas as avaliações numéricas e o processo de estimação dos parâmetros no modelo proposto nesta dissertação, foram realizadas pela linguagem `R` disponível de forma gratuita em `www.r-project.org/`.

Para a estimação dos parâmetros do modelo tobit normal, utilizamos a função `tobit()` do pacote `AER`, usada para ajustar modelos de regressão tobit (ver e.g. Kleiber & Zeileis (2008) e Kleiber & Zeileis (2015)). Para a avaliação do diagnóstico de influência usamos o pacote `tobitdiag` (Santos-Neto (2016) e Santos-Neto *et al.* (2016)).

Para a estimação e os principais resultados da avaliação do diagnóstico do modelo tobit-BS, usamos as funções implementadas em `R` (De Sousa *et al.* (2018)). A estimação dos parâmetros do modelo tobit-BS-*t* bem como a avaliação do diagnóstico de influência foram implementadas por funções em `R`. Além disso, fizemos uso do pacote `lattice` em conjunto com o pacote `robustbase` para construção de box-plots para dados assimétricos (Rousseeuw *et al.* (2016)).

Por fim, apresentamos uma aplicação em um conjunto de dados de vacinas no Haiti, usando o modelo tobit-BS-*t*. Em particular, realizamos uma comparação entre alguns modelos na família tobit-GBS. De Sousa *et al.* (2018) propuseram o modelo tobit-BS para esse conjunto de dados e realizaram uma análise residual e diagnóstica para esse modelo. Desta forma, introduzimos uma generalização do que foi proposto em De Sousa *et al.* (2018).

4.1 Estudo de simulação

Apresentamos nesta seção um estudo de simulação de Monte Carlo com 5000 réplicas, com o objetivo de avaliar o desempenho das estimativas de ML dos parâmetros do modelo tobit-BS-t, e compará-las com os modelos na classe GBS, descritos anteriormente. Consideramos os seguintes tamanhos amostrais $n = 50, 100, 300, 500$, com as seguintes variações para o parâmetro $\alpha = 0.25, 0.5, 1.0, 1.5, 3$, e $\beta = (0.2, 0.5)^\top$, considerando também as proporções de censuras iguais a $\rho = 0.2, 0.4, 0.6, 0.8$. Considere uma covariável $X \sim U(0,1)$. A medida calculada para avaliação do desempenho é o viés empírico, e posteriormente obtemos o erro quadrático médio (EQM). A Tabela 4.1 mostra os resultados avaliados para os tamanhos amostrais indicados, valores indicados dos parâmetros e as proporções de censuras citadas. Podemos observar que para $\alpha = 0.25, 0.5, 1, 1.5$ e $\rho = 0.2, 0.4, 0.6$, o viés empírico e o EQM diminuem conforme n aumenta, resultado esperado. Observe também que o modelo não é adequado para modelar dados com proporção de censura $\rho = 0.8$, com $\alpha = 3$, nestes casos os estimadores não são considerados consistentes. Em geral, os resultados realizados pelo estudo de simulação fornecem um bom desempenho nas estimativas ML do modelo tobit-BS-t.

Tabela 4.1: Viés empírico e EQM (em parênteses) dos dados simulados para as estimativas de ML dos parâmetros do modelo tobit-BS-t, n e ρ.

n		$\rho = 0.20$			$\rho = 0.40$		
	α	$\widehat{\alpha}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\alpha}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
50	0.25	−0.0014 (0.0013)	−0.0066 (0.0080)	−0.0100 (0.0240)	−0.0015 (0.0018)	−0.0122 (0.0137)	0.0153 (0.0354)
	0.50	−0.0017 (0.0051)	−0.0103 (0.0305)	0.0149 (0.0905)	−0.0005 (0.0075)	−0.0166 (0.0390)	0.0172 (0.1081)
	1.00	−0.0048 (0.0211)	−0.0147 (0.1057)	0.0056 (0.3192)	0.0002 (0.0323)	−0.0241 (0.1280)	0.0194 (0.3658)
	1.50	−0.0103 (0.0483)	−0.0148 (0.2165)	0.0002 (0.6603)	0.0083 (0.0735)	−0.0412 (0.2509)	0.0216 (0.7257)
	3.00	0.0086 (0.2064)	−0.0433 (0.4621)	−0.0006 (1.3631)	0.0792 (0.4845)	−0.0815 (0.6832)	0.0110 (1.7210)
100	0.25	−0.0010 (0.0007)	−0.0013 (0.0040)	0.0014 (0.0116)	0.0002 (0.0009)	−0.0065 (0.0063)	0.0065 (0.0161)
	0.50	−0.0004 (0.0026)	0.0001 (0.0146)	0.0001 (0.0434)	−0.0003 (0.0039)	−0.0076 (0.0186)	0.0062 (0.0506)
	1.00	− 0.0035 (0.0101)	−0.0082 (0.0529)	−0.0033 (0.1604)	0.0001 (0.0154)	−0.0086 (0.0619)	0.0036 (0.1775)
	1.50	−0.0007 (0.0237)	−0.0125 (0.1058)	0.0032 (0.3205)	0.0051 (0.0383)	−0.0262 (0.1278)	0.0211 (0.3584)
	3.00	−0.0014 (0.1043)	−0.0277 (0.2232)	0.0117 (0.6288)	0.0291 (0.2134)	−0.0480 (0.3169)	0.0323 (0.7950)
300	0.25	−0.0001 (0.0002)	−0.0012 (0.0013)	0.0022 (0.0037)	0.0006 (0.0003)	−0.0021 (0.0021)	0.0031 (0.0052)
	0.50	−0.0001 (0.0009)	−0.0021 (0.0048)	0.0025 (0.0147)	0.0003 (0.0013)	−0.0030 (0.0059)	0.0042 (0.0164)
	1.00	0.0002 (0.0035)	−0.0060 (0.0174)	0.0056 (0.0510)	−0.0002 (0.0052)	−0.0020 (0.0199)	−0.0013 (0.0548)
	1.50	−0.0009 (0.0075)	−0.0049 (0.0333)	0.0043 (0.0972)	−0.0002 (0.0123)	−0.0103 (0.0404)	0.0077 (0.1138)
	3.00	0.0062 (0.0324)	−0.0201 (0.0697)	0.0146 (0.1994)	0.0058 (0.0645)	−0.0051 (0.0990)	−0.0030 (0.2425)
500	0.25	0.0001 (0.0001)	−0.0004 (0.0008)	0.0004 (0.0023)	−0.0004 (0.0002)	−0.0004 (0.0012)	−0.0004 (0.0030)
	0.50	0.0004 (0.0005)	0.0004 (0.0028)	−0.0004 (0.0086)	0.0000 (0.0007)	−0.0011 (0.0035)	0.0014 (0.0097)
	1.00	−0.0007 (0.0021)	−0.0021 (0.0103)	0.0015 (0.0316)	−0.0013 (0.0032)	−0.0008 (0.0122)	0.0019 (0.0341)
	1.50	−0.0024 (0.0045)	−0.0020 (0.0201)	0.0012 (0.0608)	0.0005 (0.0077)	−0.0042 (0.0241)	0.0035 (0.0657)
	3.00	0.0002 (0.0193)	−0.0079 (0.0418)	0.0048 (0.1199)	0.0058 (0.0385)	−0.0110 (0.0609)	0.0085 (0.1535)

n		$\rho = 0.60$			$\rho = 0.80$		
	α	$\widehat{\alpha}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\alpha}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
50	0.25	0.0014 (0.0030)	−0.0251 (0.0304)	0.0285 (0.0617)	0.0056 (0.0075)	−0.0858 (0.1351)	0.0885 (0.1950)
	0.50	0.0019 (0.0122)	−0.0272 (0.0703)	0.0242 (0.1556)	0.0319 (0.0479)	−0.1159 (0.2646)	0.0863 (0.3818)
	1.00	0.0172 (0.0658)	−0.0583 (0.2051)	0.0474 (0.4869)	0.1876 (0.8698)	−0.1925 (0.8101)	0.0584 (0.9056)
	1.50	0.0560 (0.2427)	−0.0635 (0.4041)	0.0017 (0.9499)	0.2312 (1.8143)	−0.1597 (1.2442)	0.0341 (1.3631)
	3.00	0.5258 (4.9591)	−0.1835 (1.2763)	−0.0039 (1.9216)	−0.5727 (3.7530)	0.4650 (2.0757)	0.0421 (2.0288)
100	0.25	0.0008 (0.0014)	−0.0152 (0.0145)	0.0174 (0.0290)	0.0029 (0.0034)	−0.0405 (0.0569)	0.0411 (0.0827)
	0.50	0.0033 (0.0066)	−0.0186 (0.0335)	0.0188 (0.0742)	0.0105 (0.0185)	−0.0446 (0.1089)	0.0339 (0.1558)
	1.00	0.0064 (0.0333)	−0.0247 (0.0977)	0.0142 (0.2235)	0.0791 (0.2219)	−0.0973 (0.3405)	0.0309 (0.3931)
	1.50	0.0217 (0.0930)	−0.0320 (0.1855)	0.0018 (0.4211)	0.2731 (1.9699)	−0.1571 (0.7807)	0.0234 (0.6219)
	3.00	0.4029 (4.1910)	−0.1477 (0.7193)	0.0180 (0.8975)	−0.1630 (4.6777)	0.2860 (1.4154)	−0.0001 (0.9176)
300	0.25	0.0007 (0.0005)	−0.0043 (0.0045)	0.1569 (0.0941)	0.0013 (0.0011)	−0.0114 (0.0167)	0.0112 (0.0238)
	0.50	0.0003 (0.0021)	−0.0049 (0.0107)	0.0054 (0.0229)	0.0026 (0.0056)	−0.0104 (0.0340)	0.0029 (0.0491)
	1.00	0.0035 (0.0106)	−0.0113 (0.0318)	0.0058 (0.0724)	0.0220 (0.0401)	−0.0304 (0.0983)	0.0077 (0.1211)
	1.50	0.0090 (0.0292)	−0.0139 (0.0595)	0.0051 (0.1289)	0.0915 (0.2822)	−0.0635 (0.2470)	0.0003 (0.1896)
	3.00	0.0881 (0.3611)	−0.0396 (0.1815)	0.0074 (0.2721)	0.3860 (6.4322)	−0.0092 (0.9416)	0.0064 (0.2926)
500	0.25	−0.0002 (0.0003)	−0.0025 (0.0028)	0.0026 (0.0055)	0.0005 (0.0007)	−0.0069 (0.0102)	0.0070 (0.0144)
	0.50	0.0008 (0.0013)	−0.0033 (0.0062)	0.0027 (0.0132)	0.0007 (0.0033)	−0.0047 (0.0194)	0.0024 (0.0277)
	1.00	0.0018 (0.0061)	−0.0063 (0.0189)	0.0024 (0.0433)	0.0138 (0.0226)	−0.0234 (0.0567)	0.0137 (0.0700)
	1.50	0.0032 (0.0169)	−0.0062 (0.0367)	−0.0007 (0.0816)	−0.0466 (0.1221)	−0.0340 (0.1308)	−0.0016 (0.1109)
	3.00	0.0463 (0.1970)	−0.0211 (0.1158)	0.0068 (0.1687)	0.5209 (7.1079)	−0.0952 (0.7762)	0.0036 (0.1762)

4.2 Vacinas no Haiti

Este conjunto de dados foi usado em um estudo de casos fornecido por Moulton & Halsey (1995), trata-se de uma avaliação imunológica de vacinas contra o sarampo, realizadas no Haiti durante 1987-1990. Anticorpos de neutralização foram coletadas em um grupo de 330 crianças com até um ano de idade, logo depois de serem vacinadas contra o sarampo. O estudo tem como objetivo verificar se vacinas com níveis mais altos podem efetivamente imunizar crianças. As medições de concentração são feitas por ensaios laboratoriais, que estabelecem um limite de detecção mínimo (LDM), especificado por 0.1 mm/l em unidade internacional (UI) ou igual a -2.16 na escala logarítmica. Neste conjunto de dados, cerca de 86 (26.1%) observações estão abaixo do LDM, portanto estes valores são gravados como sendo 0.1. As seguintes variáveis são descritas neste conjunto de dados: níveis de concentração de anticorpos ($Y-$ variável resposta), X_1 é o tipo de vacina ($0-$Schwarz e $1-$Edmonton - Zagreb), X_2 indica os níveis de dosagens ($0-$médio e $1-$alto) e X_3 descreve o sexo da criança ($0-$masculino e $1-$feminino).

Esses dados foram analisados por Moulton & Halsey (1995), que usaram um modelo de mistura considerando uma regressão log-normal para observações acima do LDM e um modelo logito na modelagem do excesso de zeros, fazendo uma extensão do modelo proposto por Cragg (1971).

4.2.1 Análise dos dados

Iniciamos a aplicação dos dados com as principais medidas descritivas, com o intuito de verificar o comportamento dos níveis de anticorpos observados no estudo de caso das vacinas contra o sarampo. A Tabela 4.2 ilustra as medidas descritivas, que incluem a média, mediana, SD, e também os coeficiente de variação (CV), assimetria (CS) e curtose (CK). Observa-se assimetria positiva e um alto nível de curtose na distribuição dos dados. A Figura 4.1 mostra o histograma, o gráfico do tempo total em teste (TTT) e os box-plots para os dados de vacina contra o sarampo.

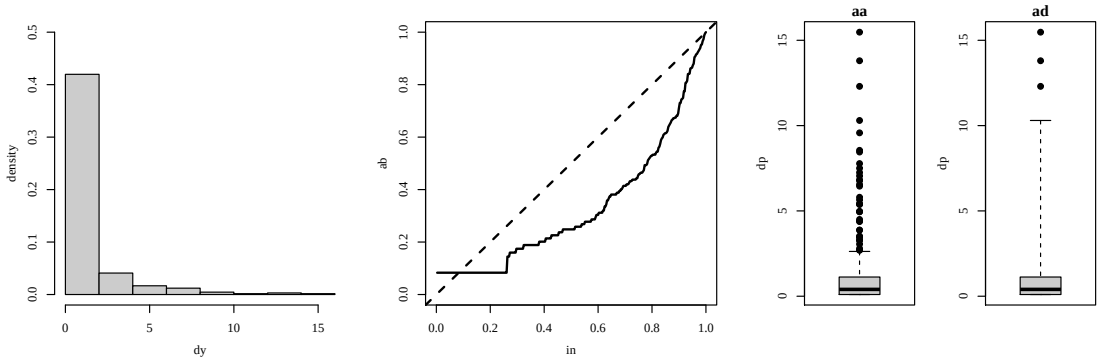
Nos estudos recentes, o gráfico TTT tem sido uma ferramenta importante na verificação do comportamento de uma determinada distribuição. A adequação do modelo nos dados é realizada pela forma e principais característica da HR. Nesta verificação, o ideal é detectar o tipo de HR que os dados possuem, e depois escolher a

melhor distribuição. Relembrando que $h_T(t) = f_T(t)/(1-F_T(t))$ é a função de risco da variável aleatória T, em que $f_T(t)$ e $F_T(t)$ são as funções PDF e CDF, respectivamente. Dessa forma, o gráfico TTT é a função $W(u) = H^u/H^{-1}(1)$, para $0 \leq u \leq 1$, em que $H^{-1}(u) = \int_0^{F_T^{-1}(u)}(1-F_Z(z))dz$ com F_T^{-1} denotando a função inversa da fda de T. Os gráficos dos pontos $(k/n, W_n(k/n))$ podem aproximar a função W, com $W_n(k/n) = (\sum_{i=1}^{k} t(i) + (n-k)(t_k)/\sum_{i=1}^{n} t(i)$, para $k = 1,\ldots,n$, e $t(i)$ denotando a i-ésima estatística de ordem observada. Para mais detalhes sobre algumas formas teóricas das curvas do TTT, ver a Figura 1 em Azevedo *et al.* (2012).

O histograma da Figura 4.1(b) mostra uma concentração muito grande das observações em torno do LDM, indicando o comportamento de assimetria à direita (ou positiva) evidenciado pela Tabela 4.2. O gráfico TTT presente na Figura 4.1 (centro) indicam que algumas observações consideradas atípicas pelo boxplot usual, podem não ser avaliadas como discrepantes, quando consideramos o boxplot ajustado (4.1(c)).

Tabela 4.2: Medidas descritivas para os dados de vacina no Haiti.

n	Min	Max	Media	Mediana	DP	CV	CS	CK
330	0.10	15.47	1.20	0.40	2.10	174.74	3.46	14.37



(a) Histograma (b) TTT plot (c) Boxplots

Figura 4.1: Histograma, TTT plot e boxplots para os dados de vacinas no Haiti.

Definimos o modelo tobit-BS-t para os dados de vacinas no Haiti da seguinte forma,

$$Y_i = \begin{cases} 0.1, & Y_i^* \leq 0.1, i = 1,\ldots,85, \\ Y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, & Y_i^* > 0.1, i = 86,\ldots,330, \end{cases} \tag{4.1}$$

em que $\varepsilon_i \overset{\text{iid}}{\sim} \log - GBS(\alpha, 0, t_\nu)$. Observe que a resposta Y_i^* é uma variável latente ob-

servada para valores maiores que 0.1 e censurada caso contrário. Para a estimação dos parâmetros do modelo utilizou-se o método de máxima verossimilhança. Desta forma, o logaritmo da função de verossimilhança é maximizado pelo método BFGS da função `optim` do `R`.

Uma importante observação, é que muitos autores têm discutido a respeito da estimação do parâmetro ν, o grau de liberdade da distribuição t (ver e.g. Fernandez & Steel (1999) e Lange *et al.* (1989)). Pela dificuldade de estimação e com problemas de maximização da função de verossimilhança, alguns autores recomendam fixar um valor para o parâmetro em questão, alguns autores também sugerem considerar $\nu = 4$ ou até mesmo utilizar os dados para obter informações de ν (ver e.g Berkane *et al.* (1994), Barros *et al.* (2008) e Paula *et al.* (2012)).

A Tabela 4.3 mostra as estimativas de ML, juntamente com seus SEs, e os p-valores associados aos testes para verificar a significância dos parâmetros. Também podemos encontrar os valores para o critério de informação de Akaike (AIC) e o critério de informação bayesiano (BIC). Nota-se através dos resultados da tabela que o modelo tobit-BS-t apresentou os menores valores de AIC e BIC, comparado aos demais modelos ajustados.

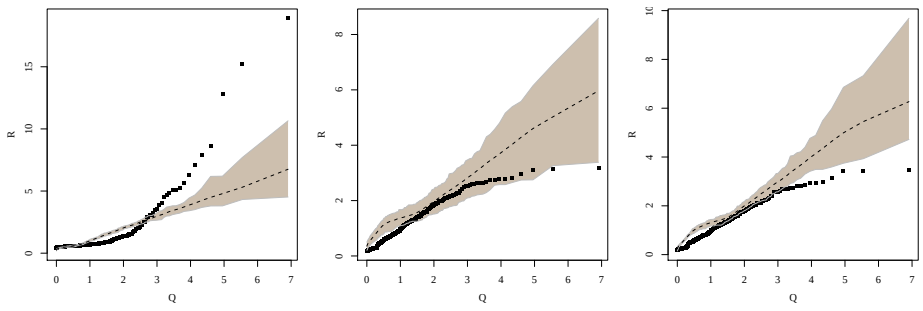
Tabela 4.3: Estimativas ML (SE em parenteses) e os valores AIC e BIC para cada modelo indicado com os dados de vacinas no Haiti.

Modelo	AIC	BIC	σ	α	ν	β_0	β_1	β_2	β_3
tobit-NO	1299.27	1318.27	0.945			0.597	0.225	-0.228	0.271
			(0.047)			(0.288)	(0.297)	(0.295)	(0.296)
p-value						[0.038]	[0.449]	[0.440]	[0.360]
tobit-Lt	1130.68	1153.47	1.474		5	-1.207	0.319	0.208	0.077
			(0.081)			(0.183)	(0.189)	(0.188)	(0.189)
p-value						[< 0.001]	[0.092]	[0.270]	[0.682]
tobit-LPE	1133.79	1159.43	1.311		0.30	-1.182	0.260	0.178	0.070
			(0.070)			(0.173)	(0.180)	(0.175)	(0.181)
p-value						[< 0.001]	[0.149]	[0.316]	[0.697]
tobit-BS	1168.38	1187.37		1.545		-0.910	0.178	0.073	0.121
				(0.081)		(0.105)	(0.127)	(0.126)	(0.126)
p-value						[< 0.001]	[0.160]	[0.560]	[0.335]
tobit-BS-t	1126.16	1148.96		1.662	4	-1.241	0.305	0.086	0.113
				(0.102)		(0.186)	(0.191)	(0.190)	(0.190)
p-value						[< 0.001]	[0.110]	[0.651]	[0.552]

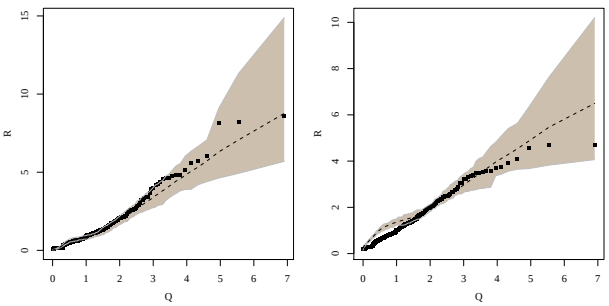
A Figura 4.2 mostra os gráficos normais de probabilidade (conhecidos como QQ plot) para o resíduo GCS com envelopes simulados, para os modelos ajustados. Podemos notar que de acordo com estes, parece haver uma indicação de que os modelos tobit-BS, tobit-BS-*t* e tobit-L*t* conseguem "atingir"a distribuição alvo que é a distribuição Exp(1). Diante desta indicação, como também através dos resultados da Tabela 4.3 há uma indicação de que o modelo tobit-BS-*t* parece ser o mais adequado.

Como o modelo tobit-BS-*t* apresentou um melhor ajuste considerando os critérios de informação e também obteve resultado satisfatório no gráfico das probabilidades normais, realizamos uma avaliação diagnóstica desse modelo.

A Figura 4.3 mostra o gráfico de influência global, com o interesse em avaliar as estimativas de ML sob remoção de um caso. Nota-se que os índices GCD no modelo tobit-BS-*t* não apresentam evidências de pontos influentes.



(a) tobit-NO (b) tobit-L*t* (c) tobit-LPE



(d) tobit-BS (e) tobit-BS-*t*

Figura 4.2: QQ plot e seus envelopes para o resíduo GCS do modelo tobit-GBS aplicado nos dados de vacinas no Haiti.

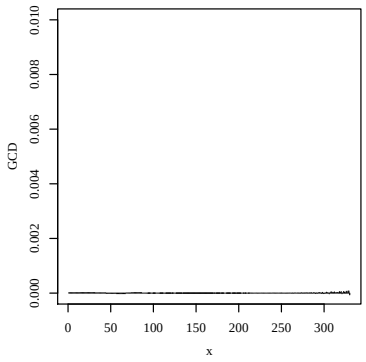


Figura 4.3: Índices das GCD para o modelo tobit-BS-t aplicado nos dados de vacina no Haiti.

Na Figura 4.4, são mostrados os índices C_i sob o esquema de ponderação de casos. Pode-se notar que as observações #326 e #329 destacam-se potencialmente influentes para θ e β, e as observações #326, #328 e #329 influentes para α, todos baseados no modelo tobit-BS-t.

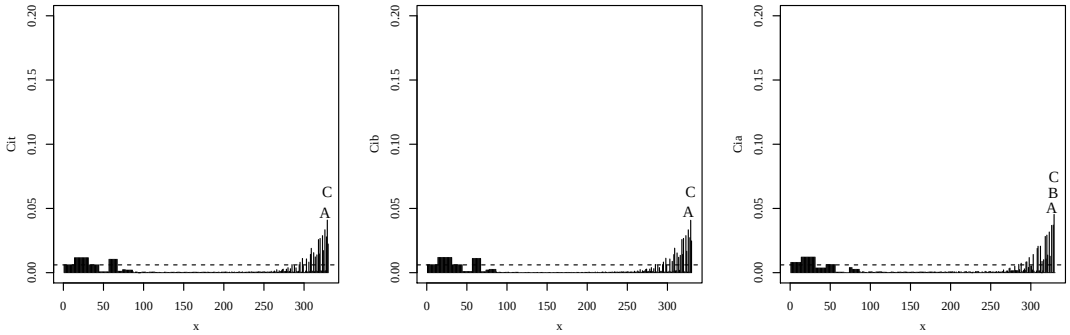


Figura 4.4: Gráfico com índices $C_i(\theta)$(esquerda), $C_i(\beta)$(centro) e $C_i(\alpha)$(direita) sob esquema de ponderação dos casos no modelo tobit-BS-t para os dados de vacina no Haiti.

Após a identificação de possíveis observações potencialmente influentes, utilizamos uma medida muito importante para verificar se cada caso individualmente (ou em grupos) influencia nos resultados inferenciais do modelo. Na Tabela **??** constam as mudanças relativas (RC) em porcentagem RC_{θ_j} de cada estimativa, que é definida por

$$RC_{\theta_j} = \left| \frac{\widehat{\theta}_j - \widehat{\theta}_{j(i)}}{\widehat{\theta}_j} \right| \times 100,$$

em que $\widehat{\theta}_{j(i)}$ é a estimativa de máxima verossimilhança de θ_j, obtida após a eliminação da i-ésima observação, para $j = 1, \ldots, 4$, com $\theta_1 = \beta_0$, $\theta_2 = \beta_1$, $\theta_3 = \beta_2$ e $\theta_4 = \beta_3$.

Pela Tabela **??** nota-se que em geral, o coeficiente β_2 é o que apresenta maiores mu-

danças relativas após a eliminações dos casos indicados. Quando comparamos individualmente as observações, observarmos que #329 apresenta maiores RCs que os casos #326 e #328. Já quando avaliamos as duplas observações, percebemos os casos $\{\#326, \#329\}$ apresentando maiores RCs. Quando avaliamos o grupo contendo as três observações, nota-se um elevado nível nas RCs. Como não verificamos alterações nas significâncias dos coeficientes, não consideramos esses casos como sendo influentes.

Tabela 4.4: *RCs* de cada estimativa indicada dos parâmetros do modelo tobit-BS-t, para cada caso removido com os dados de vacina no Haiti.

Caso(os)	Coeficientes			
Eliminado(s)	β_0	β_1	β_2	β_3
{326}	1.7083	8.4515	32.3662	24.8277
p-value	[< 0.0001]	[0.0821]	[0.5477]	[0.6541]
{328}	4.0852	7.8649	35.8595	23.7974
p-value	[< 0.0001]	[0.0833]	[0.5366]	[0.4602]
{329}	1.8523	9.2155	35.4122	27.1161
p-value	[< 0.0001]	[0.0796]	[0.5380]	[0.6636]
{326,328}	5.7820	16.2958	68.0700	1.2189
p-value	[<0.0001]	[0.0607]	[0.4423]	[0.5538]
{326,329}	3.5976	17.7839	68.1228	52.2573
p-value	[<0.0001]	[0.0576]	[0.4422]	[0.7747]
{328,329}	5.9171	17.0231	70.9626	3.4437
p-value	[<0.0001]	[0.0588]	[0.4339]	[0.5622]
{326,328,329}	7.6483	25.5587	103.4713	28.7395
p-value	[<0.0001]	[0.0846]	[0.3438]	[0.4572]

Capítulo 5

Conclusão e Trabalhos Futuros

Neste trabalho, desenvolvemos o modelo de regressão tobit baseado na distribuição GBS. Em especial, apresentamos os principais resultados inferenciais para o modelo tobit-BS-t, bem como os estimadores de ML, análise diagnóstica e análise de influência global e local através de alguns esquemas de pertubações. Efetuamos um estudo de simulação Monte Carlo, que apresentou resultados satisfatórios, tendo em vista os estimadores de ML dos parâmetros do modelo aqui proposto. Por fim, desenvolvemos uma extensão do estudo realizado por De Sousa *et al.* (2018), na qual comparamos os modelos tobit-NO, tobit-BS com o modelo tobit-BS-t aplicados nos dados de vacinas no Haiti. Observamos que para esse conjunto de dados o modelo tobit-BS-t apresentou um melhor desempenho, considerando suas principais características, tais como assimetria positiva e caudas mais pesadas, acomodando valores atípicos.

Como proposta de trabalho futuro iremos desenvolver o modelo tobit-BS-t considerado sob o enfoque Bayesiano. Utilizar recursos inferências baseado em métodos de Monte Carlo via Cadeia de Markov (MCMC). Para detectar possíveis observações influentes no modelo, usar o método bayesiano de influência caso a caso, baseado na divergência de Kullback-Leibler.

Referências Bibliográficas

Aldrich, J. H. & Nelson, F. D. (1984). *Linear probability, logit, and probit models*, volume 45. Sage.

Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics*, **24**(1-2), 3–61.

Anderson, T. W. (1990). *Statistical inference in elliptically contoured and related distributions*. Allerton Pr.

Azevedo, C., Leiva, V., Athayde, E. & Balakrishnan, N. (2012). Shape and change point analyses of the birnbaum–saunders-t hazard rate and associated estimation. *Computational Statistics & Data Analysis*, **56**(12), 3887–3897.

Balakrishnan, N., Leiva, V., Sanhueza, A., Vilca, F. et al. (2009). Estimation in the birnbaum-saunders distribution based on scale-mixture of normals and the em-algorithm. *SORT-Statistics and Operations Research Transactions*.

Barros, M., Paula, G. A. & Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, **14**(3), 316–332.

Barros, M., Paula, G. A. & Leiva, V. (2009). An r implementation for generalized birnbaum–saunders distributions. *Computational Statistics & Data Analysis*, **53**(4), 1511–1528.

Barros, M., Galea, M., González, M. & Leiva, V. (2010). Influence diagnostics in the tobit censored response model. *Statistical Methods & Applications*, **19**(3), 379–397.

Barros, M., Galea, M., Leiva, V. & Santos-Neto, M. (2018). Generalized tobit models: diagnostics and application in econometrics. *Journal of Applied Statistics*, **45**(1), 145–167.

Berkane, M., Kano, Y. & Bentler, P. M. (1994). Pseudo maximum likelihood estimation in elliptical theory: effects of misspecification. *Computational Statistics & Data Analysis*, **18**(2), 255–267.

Bhatti, C. R. (2010). The birnbaum–saunders autoregressive conditional duration model. *Mathematics and Computers in Simulation*, **80**(10), 2062–2078.

Birnbaum, Z. W. & Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied probability*, **6**(2), 319–327.

Cancho, V. G., Ortega, E. M. & Paula, G. A. (2010). On estimation and influence diagnostics for log-birnbaum–saunders student-t regression models: Full bayesian analysis. *Journal of Statistical Planning and Inference*, **140**(9), 2486–2496.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–169.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

De Sousa, M. F., Saulo, H., Leiva, V. & Scalco, P. (2018). On a tobit–birnbaum–saunders model with an application to medical data. *Journal of Applied Statistics*, **45**(5), 932–955.

Dıaz-Garcıa, J. A. & Domınguez-Molina, J. R. (2006). Some generalisations of birnbaum-saunders and sinh-normal distributions. In *International Mathematical Forum*, volume 1, pages 1709–1727. Citeseer.

Díaz-García, J. A. & Leiva-Sánchez, V. (2005). A new family of life distributions based on the elliptically contoured distributions. *Journal of Statistical Planning and Inference*, **128**(2), 445–457.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, **65**(3), 457–483.

Fang, K.T., K. S. N. K. (1990). Symmetric multivariate and related distributions. *Chapman and Hall, London*.

Fernandez, C. & Steel, M. F. (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika*, **86**(1), 153–167.

Galea, M., Riquelme, M. & Paula, G. A. (2000). Diagnostic methods in elliptical linear regression models. *Brazilian Journal of Probability and Statistics*, pages 167–184.

Galea, M., Leiva-Sánchez, V. & Paula, G. (2004). Influence diagnostics in log-birnbaum-saunders regression models. *Journal of Applied Statistics*, **31**(9), 1049–1064.

Giolo, S. R. & Colosimo, E. A. (2006). Análise de sobrevivência aplicada. *Edgard Blucher*.

Hinkley, D. V. & Cox, D. (1979). *Theoretical statistics*. Chapman and Hall/CRC.

Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2.

Kleiber, C. & Zeileis, A. (2015). R package are: Applied econometrics with r. <https://CRAN.R-project.org/package=AER>.

Lange, K. L., Little, R. J. & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**(408), 881–896.

Leiva, V. (2015). *The Birnbaum-Saunders Distribution*. Academic Press.

Leiva, V., Barros, M., Paula, G. A. & Galea, M. (2007). Influence diagnostics in log-birnbaum–saunders regression models with censored data. *Computational Statistics & Data Analysis*, **51**(12), 5694–5707.

Leiva, V., Barros, M. & Paula, G. (2009). Generalized birnbaum-saunders models using r. *São Paulo: ABE-Associação Brasileira de Estatística*.

Leiva, V., Rojas, E., Galea, M. & Sanhueza, A. (2014a). Diagnostics in birnbaum–saunders accelerated life models with an application to fatigue data. *Applied Stochastic Models in Business and Industry*, **30**(2), 115–131.

Leiva, V., Saulo, H., Leão, J. & Marchant, C. (2014b). A family of autoregressive conditional duration models applied to financial data. *Computational Statistics & Data Analysis*, **79**, 175–191.

Martínez-Flórez, G., Bolfarine, H. & Gómez, H. W. (2013a). The alpha-power tobit model. *Communications in Statistics-Theory and Methods*, **42**(4), 633–643.

Martínez-Flórez, G., Bolfarine, H. & Gómez, H. W. (2013b). Asymmetric regression models with limited responses with an application to antibody response to vaccine. *Biometrical Journal*, **55**(2), 156–172.

Moulton, L. H. & Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, pages 1570–1578.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, pages 765–799.

Owen, W. J. & Padgett, W. J. (2000). A birnbaum-saunders accelerated life model. *IEEE Transactions on Reliability*, **49**(2), 224–229.

Paula, G. A. (2015). Modelos de regressão com apoio computacional. 2013. *Citado na pág*, **1**(9), 10.

Paula, G. A., Leiva, V., Barros, M. & Liu, S. (2012). Robust statistical modeling using the birnbaum-saunders-t distribution applied to insurance. *Applied Stochastic Models in Business and Industry*, **28**(1), 16–34.

Qu, H. & Xie, F.-C. (2011). Diagnostics analysis for log-birnbaum–saunders regression models with censored data. *Statistica Neerlandica*, **65**(1), 1–21.

Rieck, J. R. & Nedelman, J. R. (1991). A log-linear model for the birnbaum-saunders distribution. *Technometrics*, **33**(1), 51–60.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T. & Maechler, M. (2016). robustbase: Basic robust statistics [software]. Disponível em : <http://CRAN.R-project.org/package=robustbase>.

Sanhueza, A., Leiva, V. & Balakrishnan, N. (2008). The generalized birnbaum-saunders distribution and its theory, methodology, and application. *Communications in Statistics-Theory and Methods*, **37**(5), 645–670.

Santos-Neto, M. (2016). tobitdiag: local influence for tobit models. [s.l.]. R package version 0.0.1.

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V., Barros, M. et al. (2016). Reparameterized birnbaum-saunders regression models with varying precision. *Electronic Journal of Statistics*, **10**(2), 2825–2855.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, **7**.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36.

Verbeke, G. & Molenberghs, G. (2000). Linear mixed models for longitudinal data.. new york. *NY Springer*.

Zhu, H. & Zhang, H. (2004). A diagnostic procedure based on local influence. *Biometrika*, **91**(3), 579–589.