Trends in Computational and Applied Mathematics
in Computational and Applied Mathematics

Trends in Computational and Applied Mathematics, **24**, N. 4 (2023), 635-658
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2023.024.04.00635

Análise de Resíduos para o Modelo Logístico Generalizado Dependente do Tempo

L.E.F. OLIVEIRA[1], L.S. SANTOS[2], L.C. FABIO[3], P.H. FERREIRA[4] e J.M.F. CARRASCO[5*]

Recebido em 18 de abril de 2022 / Aceito em 5 de junho de 2023

RESUMO. Pesquisadores de diferentes áreas do conhecimento têm utilizado o modelo de riscos proporcionais de Cox devido à sua simplicidade e fácil interpretação ao estudar situações em que a variável resposta é o tempo até a ocorrência de um evento de interesse. No entanto, o modelo tradicional de Cox não é adequado para descrever conjuntos de dados que violam a suposição de proporcionalidade dos riscos (ou taxas de falha) e os efeitos das covariáveis ao longo do tempo não são detectados. O modelo logístico generalizado dependente do tempo (GTDL, do inglês *generalized time-dependent logistic*) tem sido utilizado como alternativa na modelagem de dados de sobrevivência, levando em consideração a suposição de não proporcionalidade dos riscos. Na literatura é encontrada uma ampla e relevante produção em procedimentos inferenciais, mas nenhuma contribuição em métodos ou técnicas de diagnóstico. Neste artigo, os resíduos de Cox-Snell, modificados de Cox-Snell, *martingale*, *deviance*, quantílicos aleatorizados, NMSP (do inglês *normally-transformed modified survival probabilities*) e NRSP (do inglês *normally-transformed randomized survival probabilities*) são propostos para avaliar a adequabilidade do modelo GTDL aos dados. Um estudo de simulação via Monte Carlo é conduzido com o proposito de investigar a distribuição empírica desses resíduos. Em suma, os resultados de simulação obtidos indicam a adequação, para o modelo GTDL, dos resíduos quantílicos aleatorizados e NRSP, independentemente da proporção de censura nos dados. A biblioteca `GTDL` é construída e disponibilizada na linguagem de programação `R`. Finalmente, a metodologia estudada é aplicada a um conjunto de dados reais, disponível na literatura, envolvendo pacientes diagnosticados com câncer de pulmão em estágio avançado. Os códigos de instalação e uso da biblioteca `GTDL` são exibidos no Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`).

Palavras-chave: análise de resíduos, câncer de pulmão, modelo de riscos proporcionais de Cox, modelo GTDL, simulação de Monte Carlo.

*Autor correspondente: Jalmar Carrasco – E-mail: carrascojalmar@gmail.com

[1]Departamento de Estatística, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, 40170-110, Salvador, BA, Brasil – E-mail: lucaseber@hotmail.com https://orcid.org/0009-0009-3720-7201

[2]Departamento de Estatística, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, 40170-110, Salvador, BA, Brasil – E-mail: lucianno0800@gmail.com https://orcid.org/0000-0000-0002-9455

[3]Departamento de Estatística, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, 40170-110, Salvador, BA, Brasil – E-mail: lizandrafabio@gmail.com https://orcid.org/0000-0003-2910-5634

[4]Departamento de Estatística, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, 40170-110, Salvador, BA, Brasil – E-mail: paulohenri@ufba.br https://orcid.org/0000-0001-6312-6098

[5]Departamento de Estatística, Universidade Federal da Bahia, Avenida Milton Santos s/n, Campus de Ondina, 40170-110, Salvador, BA, Brasil – E-mail: carrascojalmar@gmail.com; https://orcid.org/0000-0002-0983-1316

1 INTRODUÇÃO

Nas últimas quatro décadas, a Análise de Sobrevivência tornou-se uma das áreas da Estatística com maior evolução. A existência de um suporte computacional mais eficiente e de aprimoramentos das técnicas estatísticas favorece a obtenção de resultados cada vez mais precisos e estimula a sua empregabilidade em diversos campos da ciência, como em estudos demográficos, medicina, saúde pública, biologia, engenharia, segurança pública, entre outros. Um fator que descreve o estudo de análise de sobrevivência é a caracterização da variável resposta que, por vezes, é o tempo até a ocorrência do evento de interesse, que pode ser, por exemplo, na medicina, o tempo até a morte de um paciente, ou na engenharia, o tempo até a falha de um produto fabricado. Outro aspecto importante é a presença de censura nos dados, isto é, a presença de observações incompletas ou parciais. Tais observações podem ocorrer por uma variedade de razões, dentre elas, a perda de acompanhamento do paciente no decorrer do estudo e a não ocorrência do evento de interesse até o término do experimento, isto é, a informação obtida sobre estes indivíduos é que o seu tempo até o evento acontecer é superior ao tempo registrado até o último acompanhamento [5].

A modelagem de dados clínicos passou por um avanço com a introdução da metodologia semiparamétrica mediante o modelo de regressão de Cox [6]. Também conhecido como modelo de riscos proporcionais, o modelo semi-paramétrico de Cox, composto pelo produto de dois componentes, um não-paramétrico e outro paramétrico, é amplamente empregado na medicina e na saúde pública. A existência de técnicas para avaliar a sua adequação, a facilidade de interpretação de seus coeficientes e a sua flexibilidade em virtude da presença do componente não-paramétrico, justificam a escolha desse modelo em muitas produções científicas. Entretanto, muitos pesquisadores desconsideram a importante premissa de proporcionalidade dos riscos, o que pode acarretar sérios vícios na estimação dos coeficientes do modelo, segundo [24]. Com a necessidade de analisar conjuntos de dados com propriedades de taxas de falha que não atendem à proporcionalidade pressuposta para o modelo de Cox, muitos trabalhos apresentaram modelos alternativos que acomodam dados com essa violação de forma exitosa, obtendo, assim, resultados mais precisos.

Em meio às diversas técnicas propostas para lidar com o relaxamento do pressuposto de proporcionalidade dos riscos, [19] apresentou um modelo competitivo completamente paramétrico em relação ao modelo de Cox, conhecido como modelo logístico generalizado dependente do tempo (GTDL, sigla do inglês *generalized time-dependent logistic*), que pode facilmente incorporar o efeito da dependência do tempo em seu ambiente. [3] fizeram uso desse modelo em um estudo comparativo com os modelos de fragilidade gama de riscos proporcionais e, posteriormente, [2] avaliaram a performance do modelo GTDL para dados de sobrevivência multivariados com e sem a inclusão do termo de fragilidade. Outros pesquisadores partiram do modelo GTDL para a elaboração de produções relevantes em procedimentos de inferência [17], estimação intervalar de parâmetros do modelo GTDL para amostras de tamanhos médios e pequenos [18] e, mais recentemente, na aplicação da modelagem dos dados de falha de válvulas de segurança de subsuperfície [16]. Entretanto, métodos ou técnicas de diagnóstico para o modelo GTDL não tem sido propostos.

Na modelagem estatística de dados, a verificação de possíveis afastamentos das suposições estabelecidas para o modelo, quer sejam componentes determinísticos, quer sejam estocásticos, é uma etapa essencial, segundo [22]. Ao mesmo tempo, torna-se necessário avaliar a interferência das observações discrepantes ou atípicas nos resultados. A análise de resíduos, portanto, surge como uma técnica valiosa para a identificação de pontos *outliers* e investigação se algumas suposições do modelo estão sendo violadas.

Para dados de sobrevivência, alguns resíduos são frequentemente empregados pela eficiência em se adequar a certos modelos de regressão, como os resíduos de Cox-Snell [7], Cox-Snell modificados [11], *martingale* e *deviance* [13]. Os resíduos de Cox-Snell, *martingale*, *deviance* e de Schoenfeld para uso com modelos de riscos proporcionais para dados de sobrevivência censurados por intervalo, foram empregados por [10]. Recentemente, os resíduos de probabilidades de sobrevivência aleatoriamente transformados, propostos por [14], foram aplicados nos modelos de teste acelerado de falha Weibull, log-normal e log-logístico, demonstrando a sua eficácia na detecção de má especificação de modelos, da família de distribuições e da forma funcional de covariáveis, e na identificação da violação do pressuposto de riscos proporcionais.

Neste trabalho, o objetivo é propor ferramentas que permitam mensurar a adequação do modelo GTDL, mediante técnicas de análise de resíduos. Dentre os tipos de resíduos utilizados, estão: Cox-Snell, Cox-Snell modificados, *deviance*, *martingale*, quantílicos aleatorizados, NMSP (sigla do inglês *normally–transformed modified survival probabilities*) e NRSP (sigla do inglês *normally–transformed randomized survival probabilities*). Vale destacar o fato de que os dois últimos resíduos citados ainda não foram explorados na literatura para o modelo GTDL. Neste trabalho, os procedimentos inferenciais e de diagnóstico para o modelo GTDL são realizados utilizando a biblioteca `GTDL`[1], desenvolvida por parte dos autores e disponibilizada na linguagem de programação `R` [23].

O restante do texto está organizado como segue. Na Seção 2 é definido o modelo GTDL proposto por [19]. Na Seção 3 é discutida a análise de resíduos para o modelo GTDL. Na Seção 4 é conduzido um estudo de simulação de Monte Carlo para investigar a distribuição empírica dos resíduos definidos na Seção 3. Na Seção 5 é analisado um conjunto de dados reais utilizando a biblioteca `GTDL`, sendo que os códigos `R` para a instalação e uso dessa biblioteca são apresentados no Material Suplementar. Finalmente, na Seção 6 são listadas algumas conclusões.

[1] `https://CRAN.R-project.org/package=GTDL`

2 O MODELO GTDL

Seja T uma variável aleatória não-negativa que representa o tempo até a ocorrência do evento de interesse. O modelo GTDL é caracterizado por apresentar as seguintes funções de densidade, de sobrevivência e de risco, respectivamente:

$$\begin{aligned} f(t \mid \boldsymbol{\theta}) &= \lambda \left\{ \frac{\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})} \right\} \times \left\{ \frac{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta})} \right\}^{-\lambda/\alpha}, \\ S(t \mid \boldsymbol{\theta}) &= \left\{ \frac{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta})} \right\}^{-\lambda/\alpha} \quad \text{e} \\ h(t \mid \boldsymbol{\theta}) &= \lambda \left\{ \frac{\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})} \right\}, \end{aligned}$$

em que $\boldsymbol{\theta} = (\lambda, \alpha, \boldsymbol{\beta})^\top$, sendo $\lambda > 0$ um escalar, $\alpha \in \mathbb{R}$ é o parâmetro que mede o efeito do tempo e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ é o vetor de k parâmetros que mensura a influência das k covariáveis $\boldsymbol{X}^\top = (X_1, \ldots, X_k)$. Se $\lambda = 1$, o modelo em questão se reduz à família logística tempo-dependente [19]. De acordo com [19], o modelo GTDL não pode ser considerado como um modelo de riscos proporcionais ou como um modelo de vida acelerada. Essa diferença pode ser percebida considerando-se, por exemplo, duas observações/indivíduos quaisquer, i e j, que possuem valores distintos de X_1. Desta forma, a razão de riscos obtida entre essas duas observações passa a ser:

$$\frac{h(t \mid X_{1,i})}{h(t \mid X_{1,j})} = \exp((X_{1,i} - X_{1,j})\beta_1) \times \frac{1+\exp(\alpha t + X_{1,j}\beta_1)}{1+\exp(\alpha t + X_{1,i}\beta_1)}.$$

Analogamente, a razão de riscos ou risco relativo, para o modelo GTDL, entre dois níveis (zero e um, considerando o zero como referência) de uma l-ésima covariável, X_l, sendo $l = 1, \ldots, k$, mantendo fixas as outras $(k-1)$ covariáveis, é definida como:

$$HR_{\text{GTDL}}(t) = \exp(\beta_l) \times \frac{1+\exp(\alpha t)}{1+\exp(\alpha t + \beta_l)}, \quad \forall t \geq 0. \tag{2.1}$$

Facilmente, é possível observar em (2.1) que $HR_{\text{GTDL}}(t) = 2 \times \exp(\beta_l)/\{1+\exp(\beta_l)\}$ quando $t = 0$. Para $\alpha > 0$ (caso da aplicação real deste trabalho; ver Seção 5), $\lim_{t\to\infty} HR_{\text{GTDL}}(t) = 1$, e para $\alpha < 0$, $\lim_{t\to\infty} HR_{\text{GTDL}}(t) = \exp(\beta_l) = HR_{\text{Cox}}$, em que HR_{Cox} denota a razão de riscos para o modelo de Cox [5, p.163]. Além disso, se $\beta_l < 0$, $HR_{\text{GTDL}}(t) \geq HR_{\text{Cox}}$, e se $\beta_l > 0$, $HR_{\text{GTDL}}(t) \leq HR_{\text{Cox}}$, $\forall t \geq 0$. Quando a l-ésima covariável é contínua (idade, por exemplo), ao incrementá-la em uma unidade, a razão de riscos fica dada por:

$$HR_{\text{GTDL}}(t) = \exp(\beta_l) \times \frac{1+\exp(\alpha t + X_l\beta_l)}{1+\exp(\alpha t + (X_l+1)\beta_l)}, \quad \forall t \geq 0.$$

No modelo GTDL, a vantagem, segundo [16], é que não é considerada a suposição prévia de existência de fração de cura, deixando os dados indicarem a presença (para $\alpha < 0$) ou não (para $\alpha > 0$) de uma proporção de curados, sem requerer parâmetros extras como ocorre em modelos

de fração de cura tradicionais. A probabilidade de sobreviventes de longa duração, mediante o cálculo do limite da função de sobrevivência quando $\alpha < 0$, é obtida pela expressão [21]:

$$\begin{aligned} p(\boldsymbol{X}) &= \lim_{t\to\infty} S(t \mid \boldsymbol{\theta}) = \lim_{t\to\infty} \left\{ \frac{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta})} \right\}^{-\lambda/\alpha} \\ &= \{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta})\}^{\lambda/\alpha}. \end{aligned}$$

A função de risco acumulado, conforme [20], é dada por:

$$H(t \mid \boldsymbol{X}) = \int_0^t h(s \mid \boldsymbol{X})\, ds = \frac{\lambda}{\alpha} \log\left(\frac{1+\exp(\alpha t + \boldsymbol{X}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta})} \right). \tag{2.2}$$

Dada uma amostra aleatória de $(k+2)$-uplas, $(t_1, \delta_1, \boldsymbol{X}_1), \ldots, (t_n, \delta_n, \boldsymbol{X}_n)$, de tamanho n de uma variável aleatória T caracterizada pelo modelo GTDL, o indicador de censura $\delta_i = 1$ se ocorre o evento (falha) e $\delta_i = 0$ se não ocorre o evento (censura) e variáveis independentes $\boldsymbol{X}$ (fixas), a função de verossimilhança é definida como:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \lambda \frac{\exp(\alpha t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta})}{1+\exp(\alpha t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta})} \right\}^{\delta_i} \left\{ \frac{1+\exp(\alpha t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} \right\}^{-\lambda/\alpha}.$$

Estimadores de máxima verossimilhança de $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, são obtidos derivando parcialmente o logaritmo da função de verossimilhança, $\ell(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$, com respeito a $\boldsymbol{\theta}$ e igualando a zero. Na prática, métodos numéricos, como Newton-Raphson, escore de Fisher, Quasi-Newton, entre outros, são necessários para encontrar as estimativas de máxima verossimilhança. Particularmente, neste trabalho será utilizado o método Quasi-Newton desenvolvido por Broyden-Fletcher-Goldfarb-Shanno (método BFGS), que está implementado na função $optim(\cdot, method =$ "$BFGS$") do *software* R.

3 ANÁLISE DE RESÍDUOS

Na análise dos dados de sobrevivência, a avaliação da adequação dos modelos ajustados é uma etapa fundamental na análise de dados. A análise de resíduos é um procedimento que valida a escolha do modelo utilizado. Assim como ocorre com os modelos de regressão, esta técnica permite mensurar a qualidade da aproximação do modelo postulado aos dados observados, avaliar se as suposições do modelo são aplicáveis aos dados disponíveis e indicar a presença de *outliers* e de pontos de alavancagem. Segundo [5], as técnicas gráficas que utilizam resíduos podem ser empregadas como uma ferramenta para rejeitar determinados modelos que são inapropriados, não provando, entretanto, que certo modelo é o ideal. Para considerar um modelo como adequado, na prática comumente espera-se visualizar um comportamento aleatório em torno do zero no gráfico de resíduos *versus* a ordem das observações ou os valores preditos. Algumas exceções são encontradas na literatura quando há a presença de dados censurados, tais como os resíduos de Cox-Snell e extensões, cuja distribuição é limitada no espaço dos números reais positivos. Nesta seção, o objetivo é desenvolver uma análise de resíduos para o modelo GTDL.

3.1 RESÍDUOS DE COX-SNELL E EXTENSÕES

Os resíduos de Cox-Snell [7] são amplamente utilizados em Análise de Sobrevivência, sendo muito úteis para auxiliar no exame do ajuste global do modelo final. Tais resíduos são definidos como: $\widehat{r}_{\mathrm{CS}_i} = \widehat{H}(t_i \mid \boldsymbol{X}_i) = -\log\left(\widehat{S}(t_i \mid \boldsymbol{X}_i)\right)$, para $i = 1,2,\ldots,n$, em que $\widehat{H}(t_i \mid \boldsymbol{X}_i)$ e $\widehat{S}(t_i \mid \boldsymbol{X}_i)$ representam, respectivamente, a função de risco acumulado e a função de sobrevivência estimadas do modelo ajustado. Considerando as funções de risco acumulado e de sobrevivência do modelo GTDL, os resíduos de Cox-Snell para o modelo GTDL são dados por:

$$\widehat{r}_{\mathrm{CS}_i} = \frac{\widehat{\lambda}}{\widehat{\alpha}}\Big\{\log(1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})) - \log(1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}))\Big\},$$

para $i = 1,2,\ldots,n$. Segundo [4], os resíduos de Cox-Snell não podem ser negativos, nem simetricamente distribuídos em torno do zero. É assumido que esses resíduos tenham uma distribuição exponencial com média igual a um, isto é, $\widehat{r}_{\mathrm{CS}_i} \sim \mathrm{Exp}(1)$. [12, p.313] recomendaram que os resíduos de Cox-Snell sejam utilizados com muito cuidado, devido ao fato de que a distribuição exponencial para esses resíduos é válida somente quando os valores reais dos parâmetros são usados (ver também [5]). [8] descobriram que a adição de uma unidade aos resíduos de Cox-Snell infla os seus valores. Visto que os resíduos oriundos de observações não-censuradas não podem ser considerados da mesma forma que os resíduos provenientes de observações censuradas, torna-se necessário adotar uma modificação destes para que a censura seja considerada na interpretação dos resultados, sendo propostas pelos pesquisadores diversas modificações para o resíduo de Cox-Snell.

Conforme destacado em [4], as modificações podem envolver a inclusão da mediana, de uma unidade ou até da média harmônica. Neste artigo, é incorporado aditivamente um excesso residual nas observações censuradas, ou seja, uma constante positiva $\Delta = \log(2) \approx 0{,}693$, que trata-se da mediana da distribuição exponencial com média 1. Assim, os resíduos modificados de Cox-Snell para o modelo GTDL são dados por:

$$\widehat{r}_{\mathrm{CSm}_i} = \begin{cases} \frac{\widehat{\lambda}}{\widehat{\alpha}}\Big\{\log(1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})) - \log(1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}))\Big\}, & \textit{se } i \in F, \\ 0{,}693 + \frac{\widehat{\lambda}}{\widehat{\alpha}}\Big\{\log(1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})) - \log(1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}))\Big\}, & \textit{se } i \in C, \end{cases} \tag{3.1}$$

em que F e C representam os indicadores de falha e censura, respectivamente. Analogamente, os resíduos *martingale* ajustam os resíduos de Cox-Snell de forma que estes tenham média em torno de zero quando as observações forem censuradas, isto é, $\widehat{r}_{\mathrm{M}_i} = \delta_i + \log\left(\widehat{S}(t_i \mid \boldsymbol{X}_i)\right)$, para $i = 1,2,\ldots,n$, em que δ_i denota a variável indicadora de falha (1 se t_i é tempo de falha e 0 se t_i é tempo de censura). Os resíduos $\widehat{r}_{\mathrm{M}_i}$ possuem uma distribuição assimétrica, variando de $-\infty$ a 1, e podem auxiliar na identificação de pontos atípicos, na adequação do modelo ao pressuposto de riscos proporcionais, na precisão preditiva ou na determinação da melhor forma funcional para uma covariável de um modelo de regressão (ver, por exemplo, [26]). Os resíduos *martingale* para o modelo GTDL são definidos como:

$$\widehat{r}_{\mathrm{M}_i} = \begin{cases} -\frac{\widehat{\lambda}}{\widehat{\alpha}}\Big\{\log(1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})) - \log(1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}))\Big\}, & \textit{se } i \in C, \\ 1 - \frac{\widehat{\lambda}}{\widehat{\alpha}}\Big\{\log(1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})) - \log(1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}))\Big\}, & \textit{se } i \in F. \end{cases}$$

Por sua vez, os resíduos *deviance*, introduzidos por [26], tornam a distribuição dos resíduos *martingale* mais simétrica em torno de zero, possibilitando a identificação de pontos atípicos e melhorando a interpretabilidade dos resultados. Para o modelo GTDL, os resíduos *deviance* são dados por:

$$\widehat{r}_{\mathrm{D}_i} = \begin{cases} \operatorname{sgn}(\widehat{r}_{\mathrm{M}_i})\left[-2\widehat{r}_{\mathrm{M}_i}\right]^{1/2}, & \textit{se } i \in C, \\ \operatorname{sgn}(\widehat{r}_{\mathrm{M}_i})\left[-2\{\widehat{r}_{\mathrm{M}_i} + \log(1-\widehat{r}_{\mathrm{M}_i})\}\right]^{1/2}, & \textit{se } i \in F, \end{cases}$$

em que $\operatorname{sgn}(\cdot)$ denota a função sinal.

3.2 RESÍDUOS QUANTÍLICOS ALEATORIZADOS

Propostos por [9] para avaliar a qualidade do ajuste de modelos, os resíduos quantílicos aleatorizados podem ser utilizados para dados de sobrevivência, conforme indicado por [16]. Seja uma variável aleatória T com função de distribuição acumulada (FDA) $\bar{S}$, então os resíduos quantílicos aleatorizados podem ser definidos como: $\widehat{r}_{\mathrm{Q}_i} = \Phi^{-1}\left(\bar{S}(t_i \mid \widehat{\boldsymbol{\theta}})\right) = \Phi^{-1}\left(1 - S(t_i \mid \widehat{\boldsymbol{\theta}})\right)$, para $i = 1, 2, \ldots, n$, em que $\Phi(\cdot)$ denota a FDA da normal padrão e $\widehat{\boldsymbol{\theta}}$, as estimativas de máxima verossimilhança para $\boldsymbol{\theta}$. Para o modelo GTDL, os resíduos quantílicos aleatorizados são dados por:

$$\widehat{r}_{\mathrm{Q}_i} = \begin{cases} \Phi^{-1}\left(1 - \left\{\frac{1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right), & \textit{se } i \in F, \\ \Phi^{-1}\left(U_i \times \left[1 - \left\{\frac{1+\exp(\widehat{\alpha}t_i + \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right]\right), & \textit{se } i \in C, \end{cases}$$

em que U_i representa um número aleatório no intervalo $(0, a)$, com $a = 1 - S(t_i \mid \widehat{\boldsymbol{\theta}})$.

3.3 RESÍDUOS NMSP

Considere um i-ésimo indivíduo qualquer, T_i^* uma variável aleatória positiva e contínua que denota o verdadeiro tempo de falha e C_i uma variável aleatória positiva e contínua representando o tempo de censura, para $i = 1, 2, \ldots, n$. A função de sobrevivência de T_i^*, baseada num modelo postulado, é definida como $S_i(T_i^*) = P(T_i^* > t_i^*)$. Visto que não é possível observar o verdadeiro tempo de falha quando $T_i^* > C_i$, tem-se que os tempos de falha observados são realizações da variável aleatória definida como $T_i = \min(T_i^*, C_i)$. Além disso, o *status* de falha é definido como $\delta_i = I(T_i^* < C_i)$, em que δ_i é igual a 1 se T_i não for censurado e 0, caso contrário. Ocorre que, ao haver tempos censurados, as distribuições de probabilidade não-modificadas, calculadas com $S_i(T_i^*)$, apresentam os valores de probabilidade mais próximos de 1 do que de 0, pois estas não são uniformemente distribuídas sob o verdadeiro modelo. Isto acontece porque não se leva em consideração o efeito da censura à direita, tornando $T_i = C_i$ um valor menor do que o verdadeiro tempo de falha T_i^*, fazendo, consequentemente, com que a probabilidade de sobrevivência $S_i(T_i)$ seja maior do que $S_i(T_i^*)$.

Com a finalidade de lidar com esta limitação, métodos foram propostos para reduzir as probabilidades de sobrevivência não-modificadas (PSNM), passando, assim, a utilizar probabilidades de sobrevivência modificadas (PSM), S_i^{M}, diferindo-se com a inserção de um fator, η, que assume os valores 0 ou 1, conforme expresso a seguir:

$$S_i^{\mathrm{M}}(T_i,\delta_i,\eta)=\begin{cases} S_i(T_i), & \text{se } \delta_i=1,\\ \eta S_i(T_i), & \text{se } \delta_i=0.\end{cases}$$

Alguns autores, como [4] e [25], abordam diferentes formas para a utilização de η. Para os resíduos modificados de Cox-Snell, definidos em (3.1), com o quantil da distribuição exponencial de parâmetro 1, e utilizando a constante $\Delta=-\log(\eta)$, tem-se:

$$r_i^{\mathrm{cM}}(T_i,\delta_i,\Delta)=-\log\left(S_i^{\mathrm{M}}(T_i,\delta_i,\eta)\right)=\begin{cases} -\log\left(S_i(T_i)\right), & \text{se } \delta_i=1,\\ -\log\left(S_i(T_i)\right)+\Delta, & \text{se } \delta_i=0.\end{cases}$$

De forma particular, considerar $\Delta=1$ equivale a encontrar $\eta=1/e\approx 0{,}368$. Transformando as PSM pelo quantil da distribuição normal, chega-se aos resíduos NMSP, propostos por [14] e definidos pela seguinte expressão:

$$\widehat{r}_i^{\mathrm{NMSP}}(T_i,\delta_i,\eta)=\Phi^{-1}\left(\widehat{S}_i^{\mathrm{M}}(T_i,\delta_i,\eta)\right).$$

Para o modelo GTDL, os resíduos NMSP são dados por:

$$\widehat{r}_i^{\mathrm{NMSP}}(T_i,\delta_i,\eta)=\begin{cases} \Phi^{-1}\left(\left\{\frac{1+\exp(\widehat{\alpha}t_i+\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right), & \text{se } \delta_i=1,\\ \Phi^{-1}\left(\eta\times\left\{\frac{1+\exp(\widehat{\alpha}t_i+\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right), & \text{se } \delta_i=0.\end{cases}$$

Neste trabalho, assim como em [14], é utilizado $\eta=1/e$ (ou, equivalentemente, $\Delta=1$).

3.4 RESÍDUOS NRSP

O conceito principal dos resíduos NRSP, segundo [14], reside na aleatorização das probabilidades de sobrevivência dos tempos de censura para a obtenção de um número uniforme entre 0 e $S_i(T_i)$, para T_i censurado, ao invés da escolha de um fator de redução fixo, por exemplo, η ou Δ, como nos resíduos NMSP. A probabilidade de sobrevivência aleatorizada é definida como:

$$S_i^{\mathrm{R}}(T_i,\delta_i,U_i)=\begin{cases} S_i(T_i), & \text{se } \delta_i=1,\\ U_i S_i(T_i), & \text{se } \delta_i=0,\end{cases}$$

em que U_i é um número aleatório constante no intervalo $(0,1]$ e $S_i(T_i)$, a PSNM. Adotando-se a transformação para a FDA da normal padrão, obtém-se:

$$\widehat{r}_i^{\mathrm{NRSP}}(T_i,\delta_i,U_i)=\Phi^{-1}\left(\widehat{S}_i^{\mathrm{R}}(T_i,\delta_i,U_i)\right).$$

Para o modelo GTDL, os resíduos NRSP são dados por:

$$\widehat{r}_i^{\mathrm{NRSP}}(T_i,\delta_i,U_i)=\begin{cases}\Phi^{-1}\left(\left\{\frac{1+\exp(\widehat{\alpha}t_i+\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right), & \text{se } \delta_i=1,\\ \Phi^{-1}\left(U_i\times\left\{\frac{1+\exp(\widehat{\alpha}t_i+\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}})}\right\}^{-\widehat{\lambda}/\widehat{\alpha}}\right), & \text{se } \delta_i=0.\end{cases}$$

4 ESTUDO DE SIMULAÇÃO

Foi realizado um estudo de simulação via Monte Carlo, utilizando o *software* R, com o proposito de investigar a distribuição empírica dos resíduos definidos para o modelo GTDL. Foram consideradas $R=10.000$ réplicas de Monte Carlo, tamanhos de amostra de $n=50,100,150$ e 200, percentuais de censura de 10%, 30% e 50%, e uma única variável explicativa $x\sim$ Uniforme$(0,1)$, para três cenários distintos (descritos nas subseções seguintes), além de um outro cenário considerando duas variáveis explicativas, $x_1\sim$ Uniforme$(0,1)$ e $x_2\sim$ Bernoulli$(0{,}5)$. Amostras aleatórias foram obtidas assumindo $t=\{\log\left(\{1+\exp(\gamma)\}(1-u)^{-\alpha/\lambda}-1\right)-\gamma\}/\alpha$, sendo $u\sim$ Uniforme$(0,1)$ e $\gamma=\beta x$. Os respectivos tempos de falha foram gerados segundo a distribuição (modelo) GTDL, e os tempos de censura segundo a distribuição exponencial com média m. No Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`) são disponibilizados os códigos, na linguagem de programação `R` [23], para a construção dos gráficos apresentados nas Figuras 1 e 2.

4.1 Cenário 1

Neste cenário é considerada a situação em que o parâmetro α é maior do que zero. Os valores verdadeiros dos parâmetros do modelo GTDL são: $\lambda=0{,}5$, $\alpha=0{,}1$ e $\beta=-3$. Os valores de m que controlam as porcentagens de censura de 10%, 30% e 50% são 0,0075, 0,053 e 0,12, respectivamente. Gráficos que comparam os resíduos com os quantis das distribuições exponencial de média 1 (casos dos resíduos de Cox-Snell e modificados de Cox-Snell) e normal padrão (casos dos resíduos *martingale*, *deviance*, quantílicos aleatorizados, NMSP e NRSP) são apresentados nas Figuras 1 e 2, considerando $n=100$ (os resultados para $n=50,150$ e 200 encontram-se no Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`)). É possível notar que, independentemente da proporção de censura, os resíduos de Cox-Snell e modificados de Cox-Snell mostram uma concordância com a distribuição exponencial padrão (isto é, com média igual a 1). Também é observado que a concordância com a distribuição normal padrão, para os resíduos *martingale*, *deviance* e NMSP, é afetada à medida que a proporção de censura aumenta. Observa-se, ainda, que os resíduos *deviance*, para 10% de censura, encontram-se acima da reta de igualdade entre tais resíduos e os quantis da distribuição normal padrão, isto devido a que a média da distribuição normal padrão pode estar sendo superestimada (isto é, maior do que zero). Os resíduos NRSP e quantílicos aleatorizados, na presença de observações censuradas, apresentam um comportamento adequado em relação à concordância

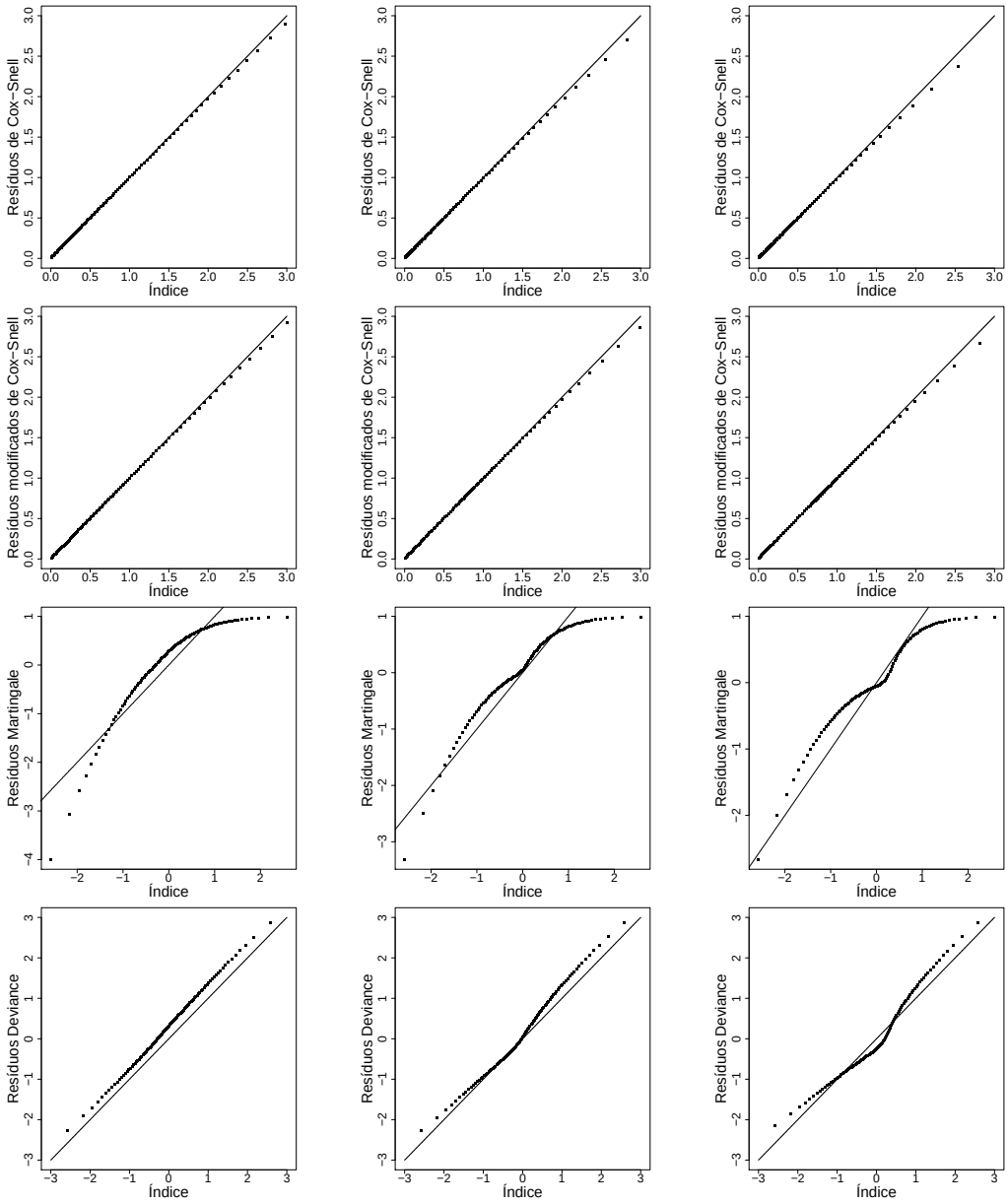

Figura 1: Distribuição empírica dos resíduos de Cox-Snell, modificados de Cox-Snell, *martingale* e *deviance*, considerando $n = 100$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 1).

com a distribuição normal padrão. Finalmente, é possível concluir que, para baixas proporções de censura, os resíduos *deviance* e NMSP exibem um comportamento similar à distribuição normal

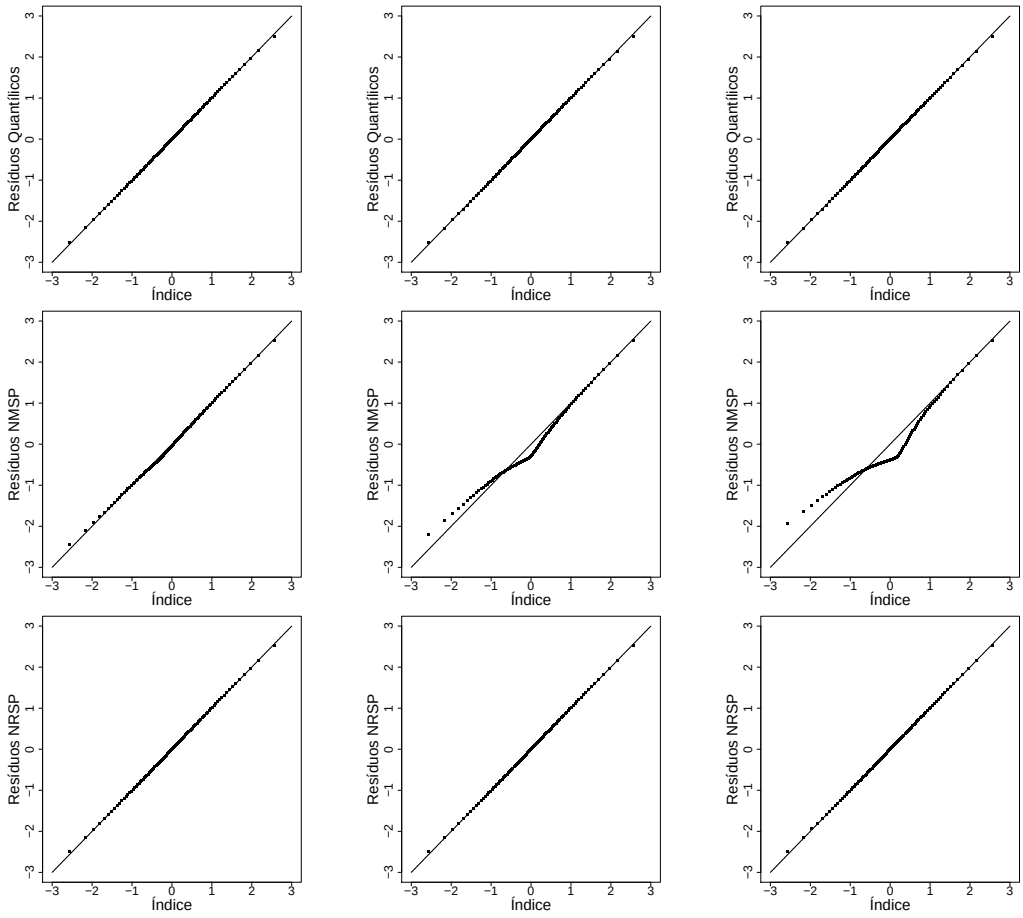

Figura 2: Distribuição empírica dos resíduos quantílicos aleatorizados, NMSP e NRSP, considerando $n = 100$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 1).

padrão. Resultados similares são observados nas figuras apresentadas no Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`).

4.2 Cenário 2

Neste cenário é considerado o caso em que o parâmetro α é menor do que zero. Os valores verdadeiros dos parâmetros do modelo GTDL são: $\lambda = 0{,}45$, $\alpha = -0{,}3$ e $\beta = 3$. Os valores de m que controlam as porcentagens de censura de 10%, 30% e 50% são 0,0025, 0,01 e 0,03, respectivamente.

As Figuras 3 e 4 mostram os resultados para este cenário, quando $n = 100$ (os resultados para os demais tamanhos de amostra estão apresentados no Material Suplementar). É possível observar

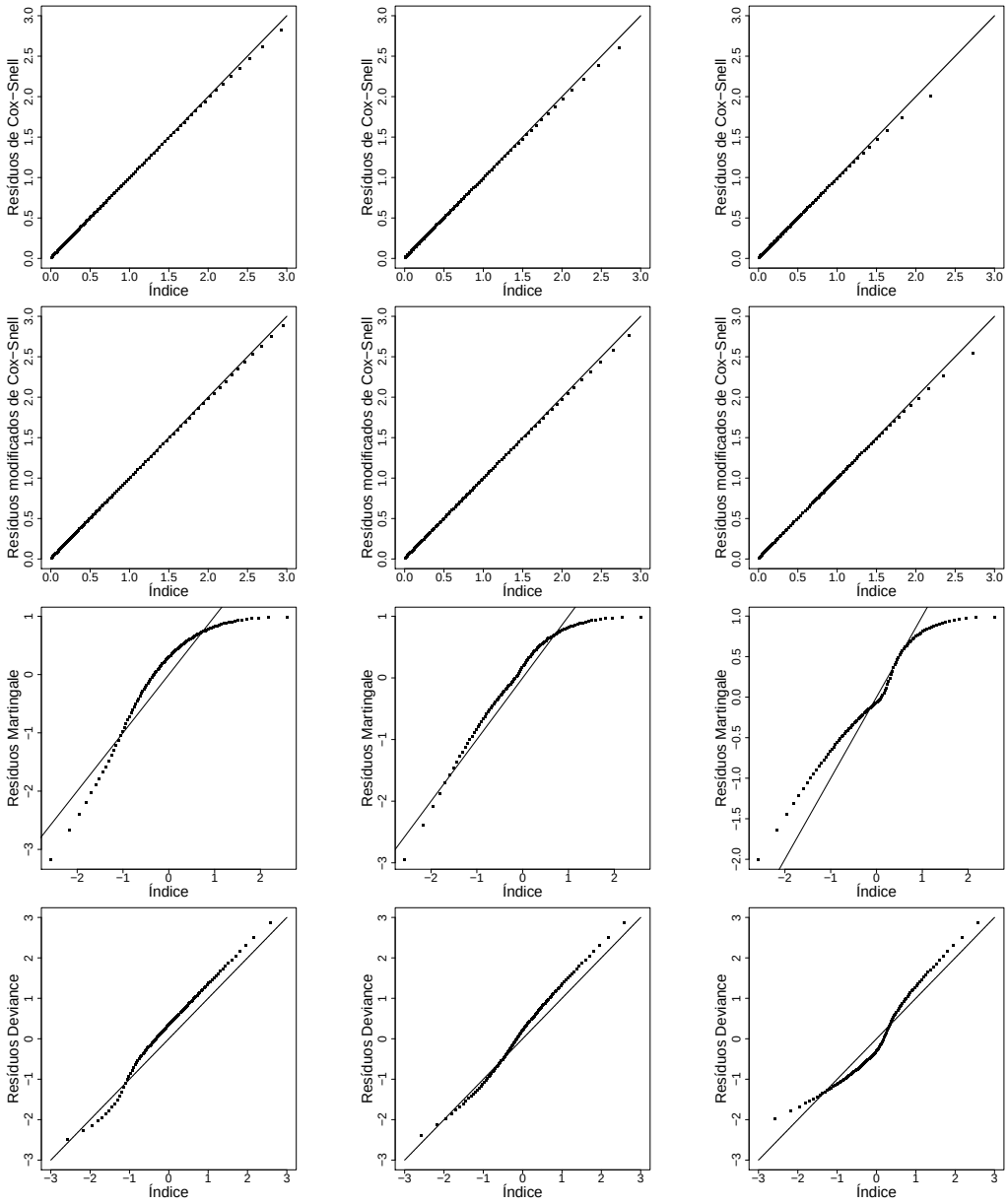

Figura 3: Distribuição empírica dos resíduos de Cox-Snell, modificados de Cox-Snell, *martingale* e *deviance*, considerando $n = 100$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 2).

um comportamento similar ao anterior (cenário 1). Entretanto, os resíduos *deviance*, *martingale* e NMSP, mesmo para pequenas proporções de censura, não apresentam concordância com a distribuição normal padrão. Comportamento semelhante é observado para tamanhos de

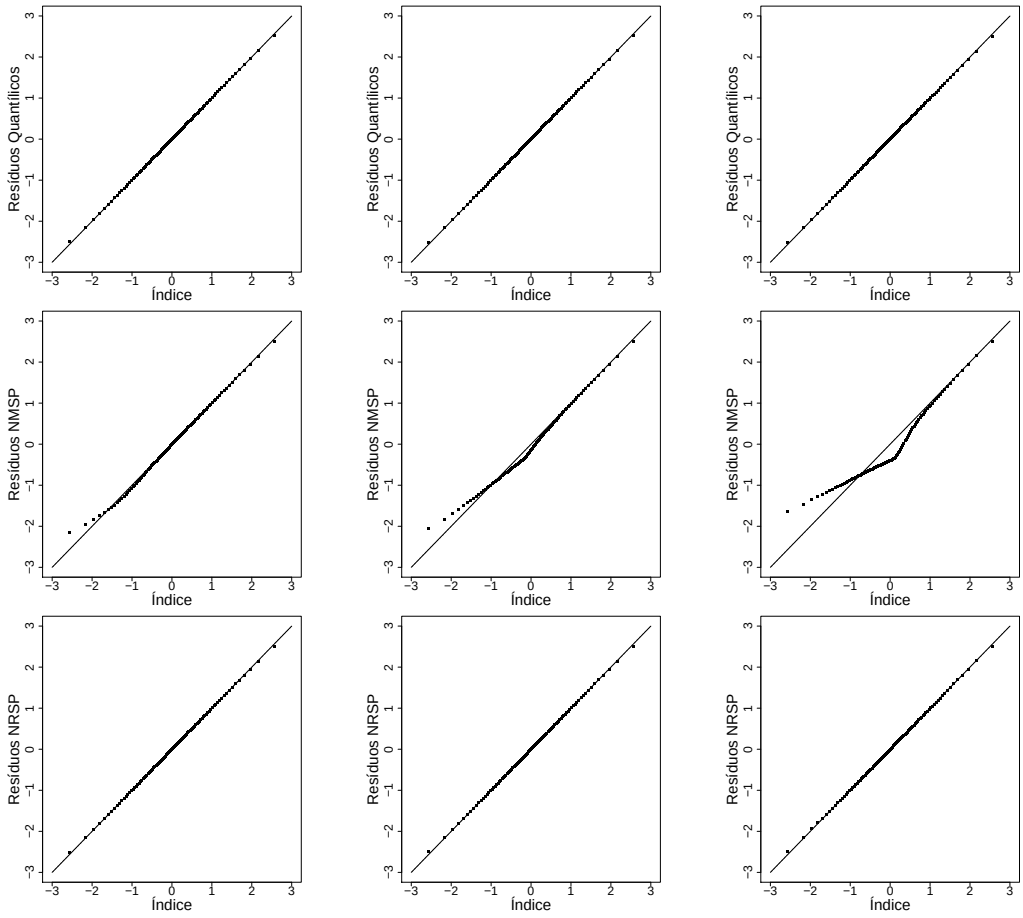

Figura 4: Distribuição empírica dos resíduos quantílicos aleatorizados, NMSP e NRSP, considerando $n = 100$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 2).

amostra de $n = 50, 150$ e 200, cujos resultados estão apresentados no Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`).

4.3 Cenário 3

Os resultados (gráficos) apresentados nas duas subseções anteriores podem mascarar alguma característica importante relacionada à distribuição empírica dos resíduos, como por exemplo, seus momentos amostrais. Desta forma, o terceiro cenário tem como objetivo estudar o comportamento médio dos resíduos em relação aos momentos esperados das distribuições exponencial padrão (casos dos resíduos de Cox-Snell e modificados de Cox-Snell) e normal padrão (casos dos resíduos *martingale*, *deviance*, quantílicos aleatorizados, NMSP e NRSP).

Neste cenário, algumas medidas descritivas, como média, desvio-padrão, assimetria e curtose, foram calculadas para os resíduos em estudo. A Tabela 1 mostra os resultados obtidos para este cenário quando a proporção de censura é 30%. Para esta situação, em particular, é possível observar, através das medidas descritivas calculadas, que os resíduos quantílicos aleatorizados, NMSP e NRSP são os que apresentam maior concordância com a distribuição padrão correspondente (normal).

Tabela 1: Média, desvio-padrão, assimetria e curtose dos resíduos de Cox-Snell (C-S), modificados de Cox-Snell (C-S Mod.), *martingale* (Mart.), *deviance* (Dev.), quantílicos aleatorizados (Quant.), NMSP e NRSP, para $n = 50$ e 200, e porcentagem de censura de 30% (cenário 3).

α	n	Estatística	C-S	C-S Mod.	Mart.	Dev.	Quant.	NMSP	NRSP
0,1	50	Média	0,70	0,91	0,00	0,15	-0,05	0,00	0,00
		Desvio-Padrão	0,75	0,76	0,81	1,09	0,94	1,00	1,00
		Assimetria	1,68	1,36	-1,09	0,23	0,41	0,03	-0,04
		Curtose	2,80	2,01	1,20	-0,71	-0,40	-0,49	-0,49
	200	Média	0,70	0,91	0,00	0,15	-0,05	0,00	0,00
		Desvio-Padrão	0,76	0,78	0,83	1,09	0,94	1,00	1,00
		Assimetria	2,03	1,68	-1,36	0,22	0,41	0,01	-0,01
		Curtose	5,20	4,03	2,79	-0,52	-0,13	-0,19	-0,19
-0,3	50	Média	0,69	0,91	0,00	0,13	-0,05	-0,01	0,01
		Desvio-Padrão	0,61	0,73	0,81	1,15	0,95	0,98	0,98
		Assimetria	1,19	0,98	-0,88	0,03	0,35	0,01	-0,01
		Curtose	0,92	0,48	0,13	-0,86	-0,61	-0,44	-0,45
	200	Média	0,70	0,91	0,00	0,14	-0,04	0,00	0,00
		Desvio-Padrão	0,62	0,74	0,83	1,16	0,96	1,00	1,00
		Assimetria	1,36	1,11	-0,99	0,02	0,36	0,00	0,00
		Curtose	1,93	1,12	0,65	-0,71	-0,41	-0,17	-0,17

As tabelas do Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`) exibem todos os resultados obtidos para este cenário. É esperado que os resíduos de Cox-Snell e modificados de Cox-Snell apresentem uma concordância com a distribuição exponencial padrão. Porém, é observado que, com o aumento da proporção de censura, a média amostral desses resíduos se afasta de 1. Considerando que o valor zero é o esperado para as medidas de assimetria e curtose dos resíduos com distribuição normal padrão, é observada uma maior aproximação de zero para os resíduos quantílicos aleatorizados, NMSP e NRSP, quando a proporção de censura é 10%. Ainda é constatado que os resíduos de Cox-Snell apresentam assimetria e curtose cada vez maiores à medida que a proporção de censura aumenta. As médias e desvios-padrão amostrais dos resíduos *martingale*, *deviance* e NMSP aproximam-se dos momentos populacionais da distribuição normal padrão, independentemente da proporção de censura na amostra. Finalmente, os resíduos quantílicos aleatorizados e NRSP apresentaram

resultados com boa (mas não exata) concordância à distribuição normal padrão, para todas as proporções de censura consideradas.

4.4 Cenário 4

Neste cenário são consideradas duas variáveis explicativas, $x_1 \sim \text{Uniforme}(0,1)$ e $x_2 \sim \text{Bernoulli}(0{,}5)$, para situações em que o parâmetro α é maior e menor do que zero. Os valores reais dos parâmetros λ e β (neste cenário, β_1) são os mesmos dos cenários 1 e 2. Além disso, é considerado $\beta_2 = 2$, associado a x_2, com valores de m iguais a 1,7, 0,38 e 0,18, para $\alpha > 0$, e 2,2, 0,8 e 0,35, para $\alpha < 0$. Os resultados obtidos para o cenário 4 são apresentados em sua totalidade no Material Suplementar (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`). As Figuras 5 e 6 mostram o comportamento da distribuição empírica para os resíduos quando $n = 100$ e $\alpha > 0$. Como esperado, os resíduos quantílicos aleatorizados e NRSP apresentaram uma melhor concordância com a distribuição normal padrão. Comportamento similar é observado para situações em que o tamanho da amostra $n = 50$, 150 e 200, assim como quando o parâmetro $\alpha < 0$.

Em resumo, os resíduos quantílicos aleatorizados e NRSP parecem ser adequados para o modelo GTDL, independente da proporção de censura presente nos dados. Ainda é possível concluir que a existência de uma leve assimetria nestes resíduos pode auxiliar na identificação de observação atípicas.

5 APLICAÇÃO

Nesta seção é analisado um conjunto de dados, disponível na biblioteca `survival` do `R` sob o nome *lung*, a fim de ilustrar a metodologia proposta. Os dados são de um estudo envolvendo 228 pacientes diagnosticados com câncer de pulmão em estágio avançado [15], oriundos do *North Central Cancer Treatment Group* (NCCTG), dos Estados Unidos. A censura é observada em 27,6% dos casos. Para esta aplicação, foi feita a recategorização da variável *ph.ecog*, devido à existência de apenas um caso referente à categoria 4 (paciente confinado à cama). Tal variável foi então redefinida como *ph.ecog.cat*, considerando as categorias 3 e 4 como uma única (nova) categoria, a saber: paciente acamado por mais de 50% do dia mas não confinado à cama *e* paciente confinado à cama. Na Figura 25 do Material Suplementar são apresentados os gráficos do logaritmo da função de risco acumulado obtida pelo estimador de Nelson-Aalen, que dão indicativos de não-proporcionalidade de riscos para as seguintes covariáveis: escore de desempenho Karnofsky pelo médico, escore de desempenho Karnofsky pelo paciente, calorias consumidas nas refeições, perda de peso nos últimos seis meses e escore de desempenho ECOG (sigla do inglês *Eastern Cooperative Oncology Group*) pelo médico.

Adicionalmente, foi investigada a propriedade de não-proporcionalidade dos riscos através dos resíduos de Schoenfeld [4]. Considerando uma significância de 5%, a covariável calorias consumidas nas refeições apresentou significância estatística (valor-p = 0,04) e, embora não-significativas, as covariáveis escore de desempenho ECOG pelo médico e escore de desempe-

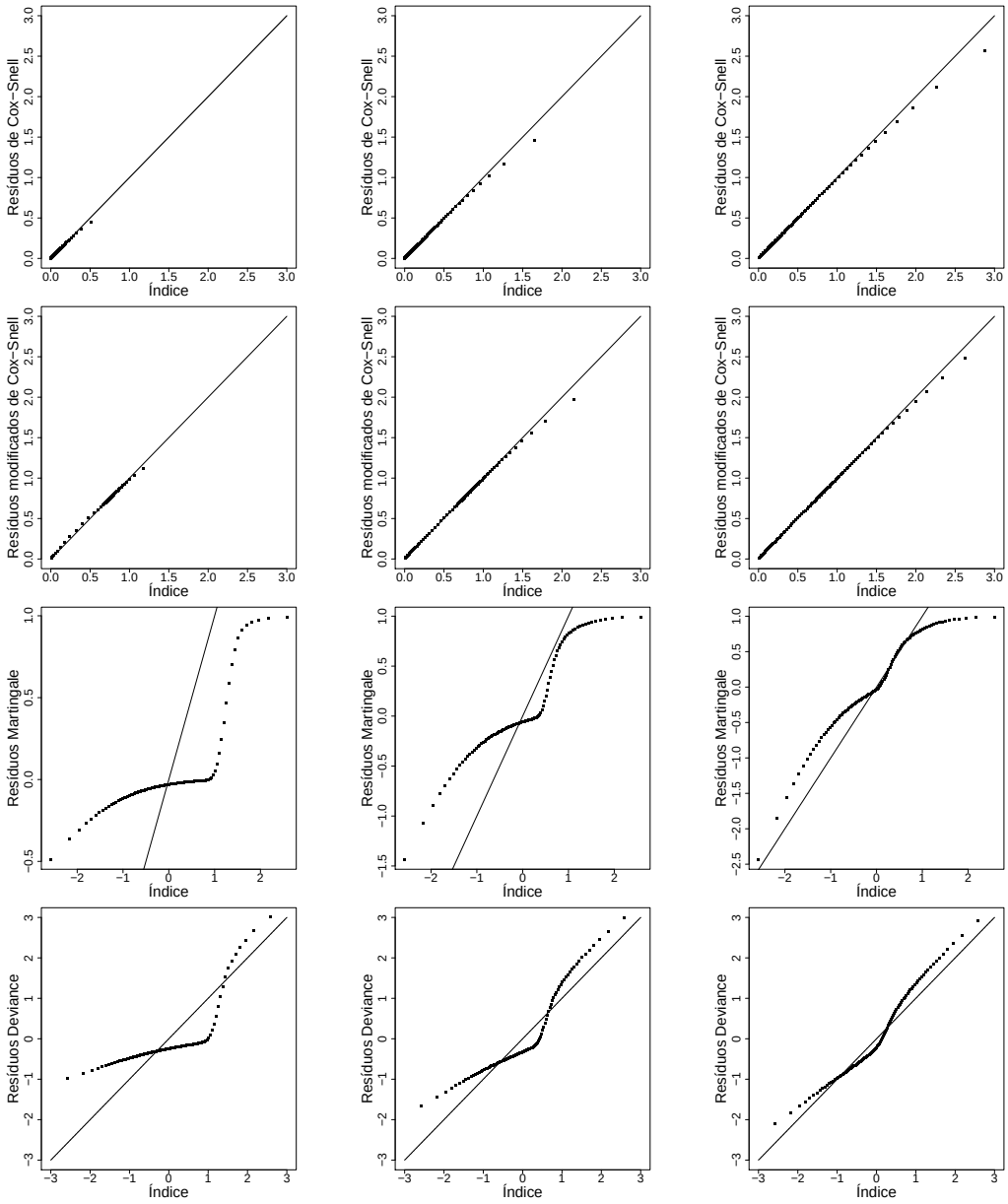

Figura 5: Distribuição empírica dos resíduos de Cox-Snell, modificados de Cox-Snell, *martingale* e *deviance*, considerando $n = 100$, $\alpha > 0$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 4).

nho Karnofsky pelo médico são marginalmente significativas (valor-p = 0,11 e 0,17, respectivamente), sugerindo uma possível violação da suposição de riscos (ou taxas de falha) proporcionais para este nível das covariáveis.

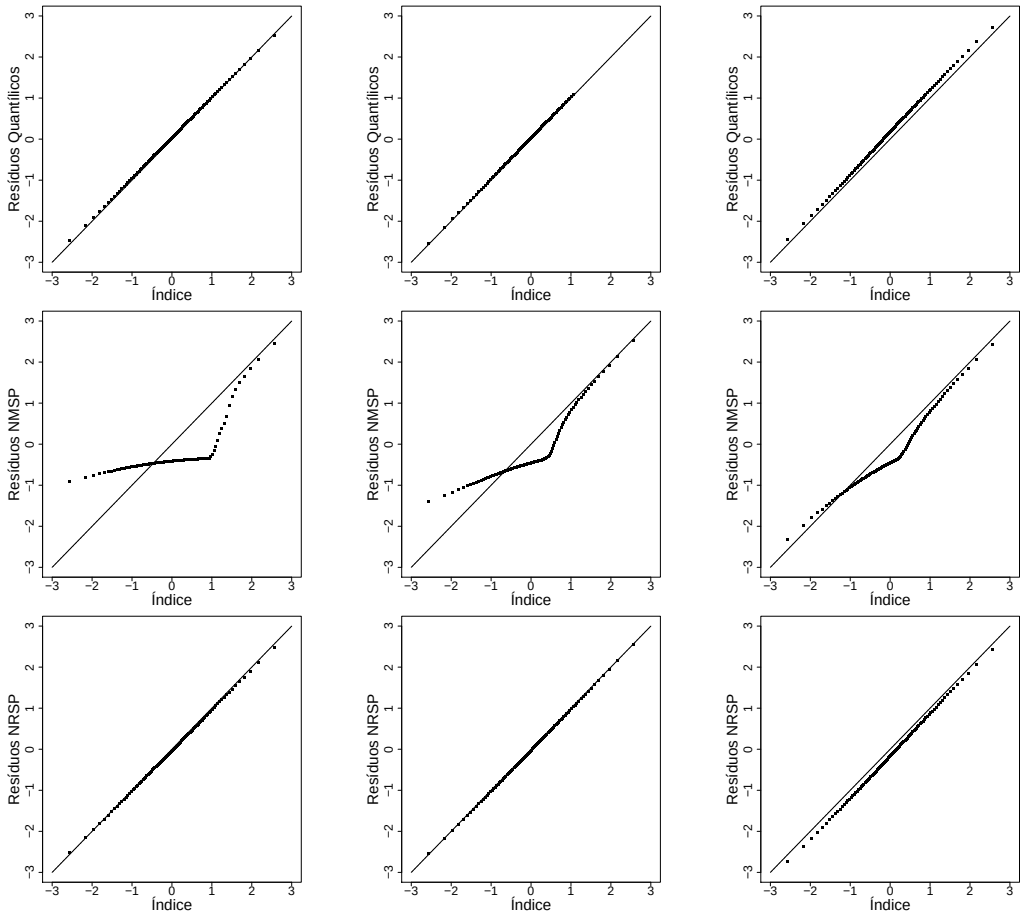

Figura 6: Distribuição empírica dos resíduos quantílicos aleatorizados, NMSP e NRSP, considerando $n = 100$, $\alpha > 0$ e porcentagens de censura (da esquerda para a direita) de 10%, 30% e 50% (cenário 4).

Contudo, há indícios de não-proporcionalidade dos riscos para algumas das covariáveis consideradas, o que, por sua vez, justifica a utilização do modelo GTDL como alternativa eficiente para o modelo de Cox. Na Tabela 2 são apresentadas as estimativas dos parâmetros, obtidas via máxima verossimilhança e máxima verossimilhança parcial, para os modelos GTDL e de riscos proporcionais de Cox, respectivamente. Para o ajuste do modelo GTDL no `R`, foi utilizada a função `mle2.GTDL`da biblioteca `GTDL`.

O processo de obtenção do modelo final consistiu inicialmente na exclusão das observações que continham valores faltantes (*missing data*), resultando em uma amostra de 167 pacientes, seguidamente passamos a retirar sequencialmente as covariáveis cujos coeficientes foram não-significativos, ou seja, com valor-$p > 0{,}05$. As variáveis selecionadas para o modelo final foram sexo, escore ECOG pelo médico e idade. No conjunto de dados completos para as variáveis sexo,

Tabela 2: Estimativas, erros-padrão e valores-p para os parâmetros dos modelos GTDL e de Cox.

Modelo	Parâmetro*	Estimativa	Erro-Padrão	Valor-p
GTDL	λ	0,005	0,001	–
	α	0,006	0,002	–
	β_1	-1,603	0,406	0,001
	$\beta_{2(1)}$	0,611	0,414	0,141
	$\beta_{2(2)}$	2,541	0,731	0,001
	β_3	-0,021	0,007	0,003
Cox	β_1	-0,551	0,168	0,001
	$\beta_{2(1)}$	0,409	0,200	0,040
	$\beta_{2(2)}$	0,916	0,227	0,001
	β_3	0,011	0,009	0,235

* β_1 : sexo feminino;
$\beta_{2(1)}$: escore ECOG pelo médico - acamado $<$ 50% do dia;
$\beta_{2(2)}$: escore ECOG pelo médico - acamado $>$ 50% do dia;
β_3 : idade em anos completos.

escore ECOG pelo médico e idade, é observada a presença de uma única observação contendo valores faltantes; deletamos esta observação e utilizamos uma amostra de 227 pacientes para a análise dos dados.

Na Tabela 2 é possível observar que, para o modelo GTDL, existe diferença significativa: no risco de morte entre as mulheres e os homens (valor-p = 0,001), e como definido em (2.1), com uma menor chance de óbito das mulheres relativamente aos homens, em uma proporção inferior a $100 \times [1-2 \times \exp(-1{,}603)/\{1+\exp(-1{,}603)\}] = 66{,}5\%, \forall t \geq 0$; entre os pacientes confinados à cama mais da metade do dia, com respeito aos pacientes sem restrições de desempenho[2] (valor-p = 0,001), com até $2 \times \exp(2{,}541)/\{1+\exp(2{,}541)\} = 1{,}85$ vezes mais chances de óbito; idade em anos completos (valor-p = 0,003), sendo que com o acréscimo de um ano de vida, para pacientes com idade igual a 62,4 anos (média amostral), por exemplo, o risco de óbito fica aumentado em até $100 \times \{1-\exp(-0{,}021) \times [\{1+\exp(-0{,}021 \times 62{,}4)\}/\{1+\exp(-0{,}021 \times (62{,}4+1))\}]\} = 1{,}6\%$. Na Figura 26 do Material Suplementar são apresentados os gráficos do comportamento da razão de riscos, para os modelos GTDL e Cox, das variáveis em estudo. O modelo GTDL revela também que os dados apontam para a existência de uma tempo-dependência positiva ($\widehat{\alpha} = 0{,}006 > 0$), confirmando que o tempo apresenta um efeito acelerador da morte dos pacientes com câncer de pulmão. Assumindo (erroneamente) riscos proporcionais, isto é, considerando o modelo de Cox como uma potencial alternativa na análise dos dados, são observadas, no entanto, diferenças entre os resultados (estimativas dos parâmetros associados às variáveis sexo, escore ECOG pelo médico e idade) obtidos por este modelo e pelo modelo GTDL (ver Tabela 2). No modelo de Cox, a variável idade não é significativa a um nível de significância de 5%. Ainda é possível afirmar erroneamente que a chance de óbito entre os pacientes do sexo

[2]Nível (ou categoria) de referência.

feminino é $100 \times \{1 - \exp(-0{,}551)\} = 42{,}4\%, \forall t \geq 0$, menor do que os pacientes do sexo masculino (valor-p = 0,001), e que os pacientes confinados à cama menos ou mais da metade do dia possuem, respectivamente, 1,51 e 2,50 vezes mais chances de óbito do que os pacientes sem restrições de desempenho (valor-p = 0,040 e 0,001, respectivamente).

A análise de diagnóstico é uma etapa muito importante na modelagem estatística. Na Figura 7 são apresentados os gráficos dos resíduos NRSP e quantílicos aleatorizados, para o modelo GTDL. Tais gráficos podem ser obtidos no `R` utilizando as funções `nrsp.GTDL` e `random.quantile.GTDL` da biblioteca `GTDL`. A escolha desses resíduos se deve aos resultados de simulação obtidos na Seção 4. Nesta figura é possível observar que ambos os resíduos distribuem-se de forma aleatória em torno de zero e entre as bandas de $[-3, 3]$.

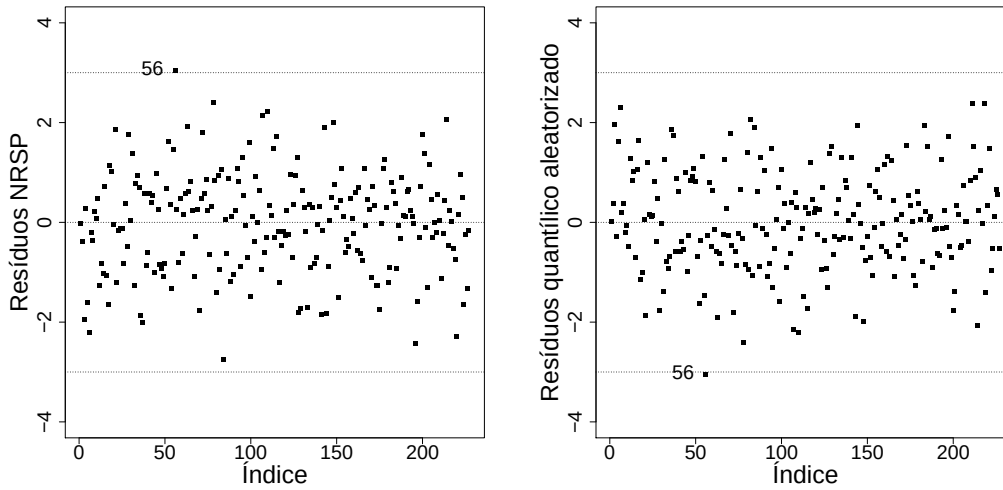

Figura 7: Resíduos NRSP e quantílicos aleatorizados para o modelo GTDL, utilizando o conjunto de dados *lung*.

Contudo, a observação #56 é identificada como observação atípica. Esta se refere a um paciente do sexo masculino, com 65 anos de idade, que teve cinco dias de sobrevida, embora assintomático, com boas classificações no escore de desempenho de Karnofsky, do médico (100) e do paciente (80), que consumiu 338 calorias e perdeu cinco quilos no período. Portanto, essa observação apresenta fortes indícios de ser um ponto atípico do conjunto de dados, além do fato de se destacar nos resíduos NRSP e quantílicos aleatorizados.

Para investigar o impacto da observação atípica no processo de estimação, foi removida a observação #56 do conjunto de dados e calculado o desvio relativo percentual (DRP) para avaliar a magnitude do impacto exercido pela observação retirada. Para $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_6)^\top = (\lambda, \alpha, \beta_1, \beta_{2(1)}, \beta_{2(2)}, \beta_3)^\top$, tem-se que $\text{DRP}_j = \{(\widehat{\psi}_j - \widehat{\psi}_j^*)/\widehat{\psi}_j\} \times 100\%$, para $j = 1, \ldots, 6$ e $\widehat{\psi}_j^*$ a estimativa do parâmetro ψ_j obtida ao retirar a(s) observação(ões) atípica(s).

Tabela 3: Estimativas, erros-padrão, desvios relativos percentuais (DRP) e valores-p para os modelos GTDL e de Cox, após a exclusão da observação #56.

Caso Deletado	Parâmetro	Estimativa	Erro-Padrão	Valor-p	DRP
GTDL - {#56}	λ	0,004	0,001	–	20,00
	α	0,007	0,002	–	−16,67
	β_1	−1,731	0,406	0,000	−7,99
	$\beta_{2(1)}$	0,713	0,425	0,094	−16,69
	$\beta_{2(2)}$	2,299	0,741	0,000	9,52
	β_3	−0,024	0,007	0,001	−14,29
Cox - {#56}	β_1	−0,574	0,169	0,001	−4,17
	$\beta_{2(1)}$	0,441	0,201	0,029	−7,82
	$\beta_{2(2)}$	0,950	0,228	0,000	−3,71
	β_3	0,011	0,009	0,246	0,00

Ao nível de 5% de significância, é observado que não houve alteração considerável nos valores-p referentes aos parâmetros β_1, $\beta_{2(1)}$, $\beta_{2(2)}$ e β_3 do modelo de Cox. No entanto, a remoção da observação #56 tornou mais significante o parâmetro $\beta_{2(1)}$ do modelo GTDL, que passou a apresentar significância estatística a 10%. Não houve grande mudança nos valores de DRP obtidos após a exclusão da observação #56. Finalmente, gráficos de probabilidade normal com envelopes simulados [1], obtidos no `R` utilizando as funções `nrsp.GTDL` e `random.quantile.GTDL` da biblioteca `GTDL`, são apresentados na Figura 8. É observada uma boa aderência do modelo GTDL ao conjunto de dados *lung*, visto que os resíduos quantílicos aleatorizados e NRSP encontram-se dentro das bandas de 95% de confiança.

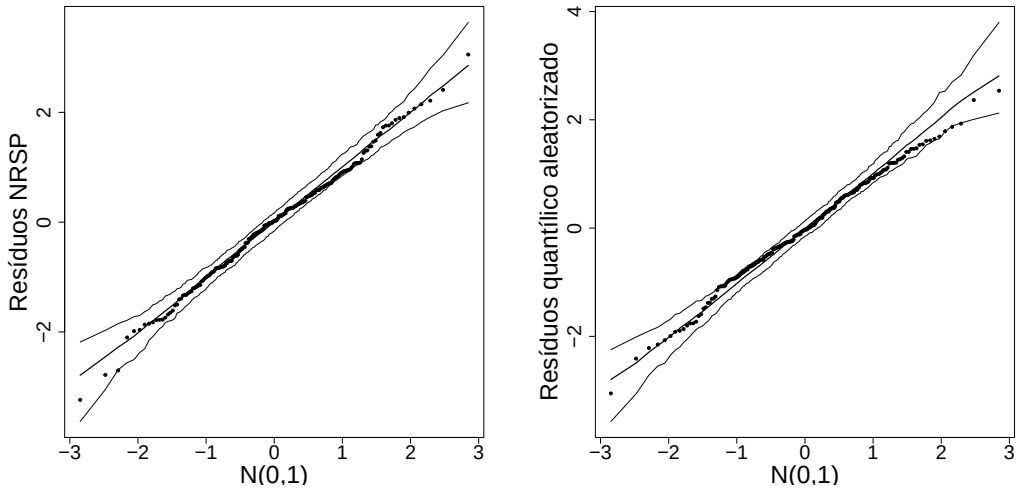

Figura 8: Gráfico de probabilidade normal e envelope simulado dos resíduos NRSP e quantílicos aleatorizados do modelo GTDL, para o conjunto de dados *lung*.

6 CONCLUSÕES

Neste estudo foram desenvolvidos aprofundamentos em Análise de Sobrevivência, relativos à análise de diagnóstico (análise de resíduos) para o modelo GTDL, o qual tem como característica principal a modelagem de conjuntos de dados com covariáveis que não atendem ao pressuposto de riscos proporcionais do modelo de Cox tradicional. Além disso, o modelo GTDL pode indicar a presença ou não de uma proporção de curados na população, sem requerer parâmetros extras, como ocorre em modelos de fração de cura tradicionais.

Para a análise de diagnóstico do modelo GTDL, foram considerados os resíduos de Cox-Snell, modificados de Cox-Snell, *martingale*, *deviance* e quantílicos aleatorizados, além da utilização (de forma inédita no contexto do modelo GTDL) dos resíduos NMSP e NRSP. Foi conduzido um estudo de simulação via Monte Carlo, com o objetivo de avaliar a distribuição empírica dos resíduos outrora propostos. Os resultados obtidos no estudo de simulação indicaram a adequação, para o modelo GTDL, dos resíduos quantílicos aleatorizados e NRSP, independentemente da proporção de censura nos dados. Ademais, a presença de leve assimetria nesses resíduos poderia auxiliar na detecção de observações atípicas (*outliers*).

A biblioteca `GTDL`, implementada na linguagem de programação `R`, é disponibilizada para fins de ilustração da metodologia proposta. Uma aplicação a um conjunto de dados reais de sobrevivência, comparando o desempenho do modelo GTDL com o do modelo de Cox foi realizada. Os resultados obtidos foram satisfatórios e melhores na utilização do modelo GTDL, fornecendo uma melhor interpretabilidade dos riscos e da tempo-dependência existente. A análise de resíduos pôde comprovar a adequabilidade do modelo GTDL e foi essencial na identificação de uma observação atípica.

Assim, o uso da análise de diagnóstico para o modelo GTDL pode aumentar a eficiência da modelagem para os casos que envolvam a existência de riscos não-proporcionais entre as variáveis nos conjuntos de dados. Também é esperado que este estudo contribua na promoção da extensão desta metodologia para outras áreas da ciência, aumentando a empregabilidade da técnica em trabalhos futuros.

Agradecimentos

À Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelas bolsas, respectivamente, de mestrado e de iniciação científica outorgadas aos dois primeiros autores. Agradecemos ao editor e revisores pelas sugestões e comentários.

ABSTRACT. Researchers from different areas of knowledge have used the Cox proportional-hazards model, due to its simplicity and easy interpretation when studying situations in which the response variable is the time until the occurrence of an event of interest. However, the traditional Cox proportional-hazards model is not suitable for modeling data sets that violate the assumption of proportionality of the risks (or failure rates)

and the effects of covariates over time are not detected. The generalized time-dependent logistic (GTDL) model has been used as an alternative in the modeling of survival data, taking into account the assumption of non-proportionality of the risks. In the literature, we found a wide and relevant production in inferential procedures, but no contribution in diagnostic methods or techniques. In this paper, Cox-Snell, modified Cox-Snell, martingale, deviance, randomized quantiles, NMSP (normally-transformed modified survival probabilities) and NRSP (normally-transformed randomized survival probabilities) residuals are proposed to assess the suitability of the GTDL model to the data. A Monte Carlo simulation study is conducted in order to investigate the empirical distribution of these residuals. In summary, the obtained simulation results indicate the adequacy, for the GTDL model, of the randomized quantile and NRSP residuals, regardless of the proportion of censorship in the data. The `GTDL` library is built and made available in the `R` programming language. Finally, the methodology studied is applied to a set of real data, available in the literature, involving patients diagnosed with advanced-stage lung cancer. Codes for installing and using the `GTDL` library are shown in the Supplementary Material (`https://github.com/carrascojalmar/GTDL-Material-Suplementar`).

Keywords: residual analysis, lung cancer, Cox proportional-hazards model, GTDL model, Monte Carlo simulation.

REFERÊNCIAS

[1] A.C. Atkinson. "Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis". Oxford University Press (1985).

[2] M. Blagojevic & G. MacKenzie. PH and non-PH frailty models for multivariate survival data. *In: Proceedings of the 19th International Workshop on Statistical Modelling*, (2004), 330–334.

[3] M. Blagojevic, G. MacKenzie & I.D. Ha. A Comparison of Non-PH and PH Gamma Frailty Models. *In: Proceedings of the 18th International Workshop on Statistical Modelling*, **18** (2003), 39–44.

[4] D. Collett. "Modelling Survival Data in Medical Research". Chapman and Hall/CRC, 3rd ed. (2015).

[5] E.A. Colosimo & S.R. Giolo. "Análise de Sobrevivência Aplicada". Edgard Blücher (2006).

[6] D.R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2) (1972), 187–220.

[7] D.R. Cox & E.J. Snell. A General Definition of Residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **30**(2) (1968), 248–275.

[8] J. Crowley & M. Hu. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72**(357) (1977), 27–36.

[9] P.K. Dunn & G.K. Smyth. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**(3) (1996), 236–244.

[10] C.P. Farrington. Residuals for Proportional Hazards Models with Interval-censored Survival Data. *Biometrics*, **56**(2) (2000), 473–482.

[11] T.R. Fleming & D.P. Harrington. "Counting Processes and Survival Analysis". John Wiley & Sons (2005).

[12] J.P. Klein & M.L. Moeschberger. "Survival Analysis: Techniques for Censored and Truncated Data". Springer, 2nd ed. (2003).

[13] J.F. Lawless. "Statistical Models and Methods for Lifetime Data". John Wiley & Sons, 2nd ed. (2003).

[14] L. Li, T. Wu & C. Feng. Model diagnostics for censored regression via randomized survival probabilities. *Statistics in Medicine*, **40**(6) (2021), 1482–1497.

[15] C.L. Loprinzi, J.A. Laurie, H.S. Wieand, J.E. Krook, P.J. Novotny, J.W. Kugler, B. J., M. Law, M. Bateman, N.E. Klatt, A.M. Dose, P.S. Etzell, R.A. Nelimark, J.A. Mailliard & C.G. Moertel. Prospective evaluation of prognostic variables from patient-completed questionnaires. *Journal of Clinical Oncology*, **12**(3) (1994), 601–607.

[16] F. Louzada, J.A. Cuminato, O.M.H. Rodriguez, V.L.D. Tomazella, E.A. Milani, P.H. Ferreira, P.L. Ramos, G. Bochio, I.C. Perissini, O.A.G. Junior, A.L. Mota, L.F.A. Alegría, D. Colombo, P.G.O. Oliveira, H.F.L. Santos & M.V.C. Magalhães. Incorporation of frailties into a non-proportional hazard regression model and its diagnostics for reliability modeling of downhole safety valves. *IEEE Access*, **8** (2020), 219757–219774.

[17] F. Louzada-Neto, C.P. Cremasco & G. MacKenzie. Sampling-based inference for the generalized time-dependent logistic hazard model. *Journal of Statistical Theory and Applications*, **9**(2) (2010), 169–184.

[18] F. Louzada Neto, G. MacKenzie, C.P. Cremasco & P.H. Ferreira-Silva. On the interval estimation of the parameters of a generalized time-dependent logistic model. *Revista Brasileira de Biometria*, **29**(3) (2011), 512–519.

[19] G. Mackenzie. Regression models for survival data: The generalized time-dependent logistic family. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **45**(1) (1996), 21–34.

[20] G. MacKenzie & D. Peng (eds.). "Statistical Modelling in Biostatistics and Informatics: Selected Papers". Springer (2014).

[21] E.A. Milani, V.L.D. Tomazella, T.C.M. Dias & F. Louzada. The generalized time-dependent logistic frailty model: An application to a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland. *Brazilian Journal of Probability and Statistics*, **29**(1) (2015), 132–144.

[22] G.A. Paula. "Modelos de regressão com apoio computacional". IME/USP, São Paulo (2013).

[23] R Core Team. "R: A Language and Environment for Statistical Computing". R Foundation for Statistical Computing, Vienna, Austria (2023). URL `https://www.R-project.org/`.

[24] C.A. Struthers & J.D. Kalbfleisch. Misspecified Proportional Hazard Models. *Biometrika*, **73**(2) (1986), 363–369.

[25] T.M. Therneau & P.M. Grambsch. "Modeling Survival Data: Extending the Cox Model". Springer-Verlag. Statistics for Biology and Health (2000).

[26] T.M. Therneau, P.M. Grambsch & T.R. Fleming. Martingale-based residuals for survival models. *Biometrika*, **77**(1) (1990), 147–160.