Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Regressão quantílica para dados com censura intervalar

Alessandra Analu Moreira da Silva

Orientador: Prof. Dr. Antonio Eduardo Gomes

Brasília, 07 de março de 2019

Regressão quantílica para dados com censura intervalar

Alessandra Analu Moreira da Silva

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Brasília, 07 de março de 2019

Regressão quantílica para dados com censura intervalar

Dissertação elaborada pela candidata Alessandra Analu Moreira da Silva submetida ao Programa de Pós-graduação em Estatística do Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do grau de Mestre em Estatística.

Texto aprovado por:

Comissão Julgadora:

- Prof. Dr. Antonio Eduardo Gomes - EST/UnB (orientador)
- Prof.ª Dr.ª Gisela Tunes da Silva - IME/USP
- Prof.ª Dr. Eduardo Yoshio Nakano - EST/UnB
- Prof. Dr. Helton Saulo Bezerra dos Santos - EST/UnB (suplente)

À minha família, a pequena Lua e in memoriam de Vladimir Moreira da Silva.

Agradecimentos

Sou profundamente grata ao meu orientador, Prof. Dr. Antonio Eduardo Gomes, por me transmitir seus conhecimentos, por sua dedicação, sua paciência e disponibilidade à minha dissertação.

Agradeço à minha família, a minha mãe Ivone, ao meu pai Egídio e ao meu irmão Fábio, pelo apoio incondicional e pelo incentivo.

Agradeço à família Zingano (Paulo, Janaína e Carolina) pelos conselhos e pelo apoio. Sou muito grata pelo carinho de vocês.

Agradeço aos meus queridos amigos Leonardo de Miranda Pinheiro, Mayara Belló Soares e Sue Helen Wainberg pela amizade e conversas a distância.

Agradeço aos amigos, que a UnB me deu, Adolfo Manoel Dias da Silva, Geiziane Silva de Oliveira, Erique Pereira Neto, Alisson Carlos da Costa Silva, Márcia Araújo Maia e Juliano Cézar Sant'Anna, pelo companheirismo, por compartilharem seus conhecimentos e pelos dias de estudos; especialmente ao Adolfo, a pessoa mais solícita e altruísta que conheci, além de realizar tudo com mestria.

Agradeço a todos que estiveram presente durante estes dois anos de mestrado e participaram de alguma forma nessa minha trajetória. Em particular, agradeço aos demais professores do PPGEST-UnB e, sobretudo, aqueles que contribuíram para o meu crescimento acadêmico e profissional, como o Prof. Dr. Raul Yukihiro Matsushita.

Agradeço ao Instituto Brasileiro de Informação em Ciência e Tecnologia pela contribuição no estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"All victories hide an abdication".

Simone de Beauvoir

"It is by the aid of Statistics that law in the social sphere can be ascertained and codified".

Florence Nightingale

Resumo

DA SILVA, A. A. M. **Regressão quantílica para dados com censura intervalar**. 2019. 101f. Dissertação (Mestrado) - Pós-graduação do Departamento de Estatística, Universidade de Brasília, Brasília. 2019.

Com a extensão da regressão quantílica à análise de dados de sobrevivência, a técnica se mostrou uma possibilidade ou complemento por modelar de forma direta o quantil condicional do tempo de sobrevivência. Também proporcionou uma interpretação mais acessível, visto que as conclusões obtidas são feitas no tocante ao tempo de sobrevivência. A motivação dessa dissertação foi a generalização da regressão quantílica quando a variável resposta apresenta censura intervalar, cujo foco foi a adaptação de um algoritmo proposto para a estimação de máxima verossimilhança não paramétrica da distribuição da variável resposta na presença da censura intervalar. Para análise de sua performance, foi realizado um estudo de simulação, com cenários diferentes e comparando essa metodologia a outra técnica proposta na literatura, aplicado a dados com censura intervalar, avaliando o vício, o erro padrão e o erro quadrático médio. Ademais, realizou-se a aplicação a conjuntos de dados reais com a intenção de também verificar o desempenho dos métodos estudados.

Palavras-chave: Regressão quantílica; Análise de sobrevivência; Censura intervalar; Dados censurados; *Kernel*; Regressão Isotônica.

Abstract

DA SILVA, A. A. M. **Quantile Regression for Interval Censored Data**. 2019. 101s. Dissertation (Master degree) - Department of Statistics, University of Brasília, Brasília, 2019.

With the extension of the quantile regression to the analysis of survival data, the technique proved to be a possibility or complement by directly modeling the conditional quantile of survival time. It also provides a more accessible interpretation, since the conclusions obtained are made regarding survival time. The motivation of this dissertation was the generalization of the quantile regression when the response variable is has interval censorship. The focus was the adaptation of a proposed algorithm for the nonparametric maximum likelihood estimation of the distribution of the response variable in the presence of interval censorship. In order to analyze its performance, a simulation study was carried out, with different scenarios and by comparing this methodology to another technique proposed in the literature, applied to data with interval censorship, evaluating the bias, the standard error and the mean square error. In addition, the application was made to real data sets with the intention of also verifying the performance of the methods studied.

Keywords: Quantile regression; Survival analysis; Interval censored data; Censored data; *Kernel*; Isotonic Regression.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A regressão quantílica é reconhecida como uma extensão mais robusta à tradicional estimação via mínimos quadrados dos modelos de regressão linear, que somente detecta associações com a média condicional. Esse procedimento estatístico oferece uma opção mais ampla para avaliar a influência dos efeitos das covariáveis sobre a variável resposta nos diferentes níveis de quantis, produzindo informações mais detalhadas. Assim, o modelo de regressão quantílica especifica o valor do quantil condicional de uma variável resposta real Y dado o valor $X = x$, um vetor de covariáveis p-dimensional (Koenker e Bassett Jr, 1978):

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}'\beta(\tau),$$

em que $Q_Y(\tau|\mathbf{x})$ é o quantil condicional de $Y|X = x,\ 0 \leq \tau \leq 1$ e $\beta \in \mathbb{R}^p$ (é o vetor de parâmetros) para $p \geq 1$.

Algumas das propriedades da regressão quantílica são bastante atrativas, principalmente diante da violação de pressupostos e de *outliers*. Quando há falha na suposição de homoscedasticidade, a regressão quantílica proporciona uma abordagem menos sensível a essa variabilidade estatística dos dados se comparada à média condicional, havendo uma acomodação da heteroscedasticidade. Dessa maneira, possibilita uma perspectiva mais precisa e abrangente em relação aos diferentes níveis de associações. Esse mesmo comportamento é constatado na presença de *outliers*, em que as estimativas da regressão quantílica também se apresentam robustas. Outro recurso importante do método é com relação às transformações, visto que há invariância dos modelos às transformações monótonas. Portanto, a distribuição original é preservada, permitindo que a interpretação seja feita de forma direta dos parâmetros.

Com a extensão da regressão quantílica à análise de dados de sobrevivência, a técnica

se mostrou uma possibilidade ao tradicional modelo de riscos proporcionais de Cox, por modelar de forma direta o quantil condicional do tempo de sobrevivência. Assim, facilita a interpretação, visto que as conclusões obtidas pela regressão quantílica são feitas no tocante ao tempo de sobrevivência e não em relação a função de risco – como no método de Cox.

Para análise de dados com tempo de censura aleatória (além dos tipos censura à direita e à esquerda), há o tipo chamado de censura intervalar ou sobrevivência intervalar, em que o tempo do evento (Y) somente é conhecido por ter ocorrido num determinado intervalo de tempo, ou seja, $Y \in (t_{1i}, t_{2i}]$, em que t_{1i} e t_{2i} são instantes de observação. A censura intervalar se divide nos casos I e II. Os dados de censura intervalar do caso I, mais conhecido como estado corrente (*current status*), concernem a dados em que todas as observações são censuradas pela direita, $Y \in (t_{1i}, \infty)$, ou pela esquerda, $Y \in (0, t_{1i})$, em que Y e t_{1i} são variáveis independentes. Portanto, no estado corrente, o indivíduo é observado apenas uma vez no experimento num determinado tempo de observação (t_{1i}), verificando a ocorrência ou não do evento de interesse. Já o caso II é o caso geral, quando há exames periódicos, em que se tem duas informações de tempo para um mesmo caso, $t_{1i} < Y \leq t_{2i}$.

Nesta dissertação, a motivação foi a generalização da regressão quantílica quando a variável resposta está em censura intervalar. Um dos enfoques é a adaptação de um algoritmo proposto para a estimação não paramétrica de máxima verossimilhança da distribuição da variável resposta na presença da censura intervalar. Para aplicação do método de Portnoy, descrito em Rasteiro (2017), para dados com censura à direita, realizou-se a imputação dos tempos de falha para os casos em que o intervalo de censura é finito. Para uma observação em que isso ocorre, a função de distribuição condicional do tempo de falha, dado o valor da covariável, foi estimada adaptando-se o algoritmo iterativo da função minorante convexa, introduzindo pesos às observações, de tal forma que esses fossem decrescentes quanto maior fosse a distância entre o valor da covariável para a observação fixada e para o restante das observações, utilizando núcleo estimador (Kernel). Obtida a estimativa, essa foi uma função escada dos valores das observações, utilizando novamente núcleo estimador para que se pudesse obter o valor imputado $\widetilde{y}_i$ para a variável resposta, gerando um valor z_i com distribuição uniforme no intervalo $[\widetilde{F}(t_{1i}), \widetilde{F}(t_{2i})]$ e obtendo $\widetilde{y}_i = \widetilde{F}^{-1}(z_i)$ iterativamente.

O método proposto foi comparado com a metodologia abordada por Zhou et al. (2017),

que trata da estimação do vetor de parâmetros $\beta(\tau)$ para modelos de regressão quantílica com dados censurados por intervalos e a obtenção da propriedade de consistência. Em Zhou et al. (2017), a propriedade da normalidade assintótica é estabelecida com um viés convergindo para zero. Para reduzir o viés, Zhou et al. (2017) propõem o método de correção de polarização baseado na metodologia direta (estimativa inicial). Esses método não exige que os vetores de censura sejam distribuídos de forma idêntica e pode ser aplicado a modelos com várias covariáveis.

A dissertação está organizada da seguinte forma: o Capítulo 2 trata de uma breve revisão sobre os quantis via otimização, apresentando a solução para uma simples minimização de uma soma da função perda, resolvida pelo quantil amostral de ordem τ. Também se refere à definição de modelos de regressão quantílica, propriedades e inferência para esses modelos. No Capítulo 3, há uma abordagem inicial sucinta sobre censura intervalar. Ademais, trata sobre a regressão quantílica aplicada a dados de censura intervalar, casos I e II. Assim, também versa sobre o método proposto que utiliza a suavização da estimativa da distribuição condicional de Y dado o valor das covariáveis contínuas, utilizando a estimação não paramétrica através da regressão isotônica e suavizada com a função kernel, incluindo a aplicação da metodologia usada por Rasteiro (2017). Além disso, há uma explanação sobre a abordagem realizada por Zhou et al. (2017). O Capítulo 4 exibe um estudo de simulação para aplicação do algoritmo proposto para o estimador de máxima verossimilhança não paramétrico da distribuição da variável dependente na presença da censura intervalar, em comparação com a técnica de Zhou et al. (2017). No Capítulo 5, tem-se a aplicação das metodologias a dados empíricos. O Capítulo 6 apresenta as considerações finais sobre o estudo.

Capítulo 2

Regressão Quantílica

O método dos mínimos quadrados (MMQ) vem sendo empregado, particularmente, para realizar a análise de regressão, já que o procedimento versa sobre a relação entre a variável resposta de interesse e as covariáveis, descrevendo a média da variável Y para cada valor fixo de X, através de uma função de média condicional da resposta (Hao e Naiman, 2007). Suponha que Y seja uma variável dependente contínua e X um preditor p-dimensional, considerando o modelo padrão de regressão linear

$$Y_i = \mathbf{x}_i^{'}\beta + \varepsilon_i, \quad i = 1, ..., n,$$

em que ε_i são independentes, identicamente e normalmente distribuídos com média zero e variância desconhecida σ^2. O cerne da análise de regressão clássica está na média, $E(Y|\mathbf{X} = x) = x'\beta$, em que β mensura a variação marginal no valor esperado de Y dada a variação unitária em X, pois assume que $E(\varepsilon_\mathbf{i}|X) = 0$ e a $Var(\varepsilon_\mathbf{i}|X) = \sigma^2$. Ademais, a aplicação de modelos de médias condicionais conduz a estimadores que são acessíveis para computar e interpretar, apresentando, também, propriedades estatísticas atraentes. Os estimadores $(\hat{\beta})$ podem ser obtidos por MMQ, que minimiza o valor da função de erro quadrático médio, isto é, $\min_{\beta \in \mathrm{R}^p} \sum_i^n [Y_i - \mathbf{x}_i'\beta]^2$, em que $\mathbf{x}_i'$ é a i-ésima linha da matriz X.

Entretanto, a estrutura de médias condicionais possui limitações intrínsecas, pois ao resumir a relação entre Y e X, o modelo gerado não pode ser imediatamente estendido a outras medidas, além da média. Outra desvantagem ocorre quando há elevada distorção da distribuição, o que pode dificultar a interpretação da média. No caso de violação da suposição de normalidade, isso pode suscitar imprecisão nos erros padrão. Assim, esse modelo tradicional não considera as propriedades distributivas condicionais completas da variável dependente, apresenta apenas informação sobre a média condicional.

Apesar do modelo de regressão linear e do modelo de regressão quantílica serem análogos em alguns aspectos, já que ambos tratam de uma variável dependente contínua e função linear em parâmetros (desconhecidos), as técnicas diferem na modelagem dos dados e na dependência das suposições pertinentes aos erros, como já mencionado. A regressão quantílica é considerada uma generalização natural da regressão linear clássica, pois é robusta quanto às suposições distributivas, posto que o estimador pesa o comportamento local da distribuição perto do quantil específico, mais do que o comportamento remoto da distribuição (Hao e Naiman, 2007), logo a estimação dos parâmetros é considerada mais eficiente. Na regressão quantílica também há variação no quantil condicional, pois é possível modelar qualquer posição predeterminada da distribuição, para qualquer quantil especificado. Portanto, é uma abordagem mais abrangente para a análise estatística.

Desde o trabalho seminal de Koenker e Bassett Jr (1978), admitiu-se a regressão quantílica como uma extensão mais robusta, especialmente para os erros que não apresentam distribuições Gaussianas. Os primeiros trabalhos de aplicações empíricas dos modelos de regressão quantílica foram dos economistas (Buchinsky, 1994; Chamberlain, 1994), com estudos que pesquisavam a respeito de toda a distribuição condicional dos salários, a fim de precisar se os retornos de grau de escolaridade, a experiência e os efeitos da filiação sindical diferiam entre os quantis salariais.

Outros estudos deram sequência, abordando tópicos adicionais, com a utilização da regressão quantílica para análise salarial, como diferenças de salários entre brancos e minorias (Chay e Honore, 1998), diferença salarial entre homens e mulheres (Fortin e Lemieux, 1998), a transferência intergeracional de rendimentos (Eide e Showalter, 1999), distribuições de salários dentro de indústrias específicas (Budd e McCall, 2001), variação na distribuição salarial (Machado e Mata, 2005; Melly, 2005) e nível educacional e desigualdade salarial (Lemieux, 2006).

O emprego da regressão quantílica também se estendeu para discorrer sobre o impacto demográfico sobre o peso ao nascer do bebê (Abrevaya, 2002) e a qualidade da escolaridade (Bedi e Edwards, 2002; Eide et al., 2002). Além disso, a regressão quantílica se espalhou para outros campos, notadamente a ecologia e as ciências ambientais (Cade et al., 1999; Scharf et al., 1998), a sociologia (Hao, 2005, 2006a,b) e medicina e saúde pública (Austin et al., 2005; Wei et al., 2006). Informações obtidas em Hao e Naiman (2007).

2.1 Quantis via otimização

Nesta seção, os conceitos foram baseados no livro de Koenker e Portnoy (1996).

Seja Y qualquer variável aleatória de valor real com a função de distribuição acumulada

$$F_Y(y) = F(y) = P(Y \leq y).$$

O quantil de ordem τ para Y é definido como

$$Q_Y(\tau) = Q(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}, \tag{2.1}$$

para $\forall \tau \in [0,1]$. Se $F(\cdot)$ é estritamente crescente e contínua, então $F_Y^{-1}(\tau)$ é um único número real y tal que $F(y) = \tau$ (Gilchrist, 2000).

Os quantis ordinários de uma amostra surgem de um problema de otimização elementar. Considere a questão da teoria de decisão em que se necessita uma estimava pontual para uma variável aleatória com função de distribuição $F(\cdot)$. Tome, a função perda definida como

$$\rho_\tau(u) = u\{\tau - I(u < 0)\}, \qquad I(u < 0) = \begin{cases} 1, & u < 0 \\ 0, & u \geq 0 \end{cases} \tag{2.2}$$

para $\forall \tau \in [0,1]$ e $\rho_\tau(u) \geq 0$ para $\forall u$, em que I é a função indicadora e $u = Y - \hat{y}$. Para minimizar a perda esperada, considera-se encontrar $\hat{y}$. Logo, tem-se,

$$E[\rho_\tau(Y - \hat{y})] = \tau \int_{\hat{y}}^{\infty} (y - \hat{y}) dF_Y(y) + (\tau - 1) \int_{-\infty}^{\hat{y}} (y - \hat{y}) dF_Y(y).$$

Para encontrar a condição de primeira ordem com respeito a $\hat{y}$, tem-se

$$0 = \frac{\partial}{\partial \hat{y}} \{E[\rho_\tau(Y - \hat{y})]\} = \tau \int_{\hat{y}}^{\infty} dF_y(y) + (\tau - 1) \int_{-\infty}^{\hat{y}} dF_y(y) = F(\hat{y}) - \tau.$$

Dado que a função de distribuição acumulada é monótona, qualquer elemento de $\{y : F(y) = \tau\}$ minimiza a perda esperada, isto é, $\hat{y}$ é o quantil de ordem τ para Y, $\hat{y} = F_Y^{-1}(\tau)$, que minimiza a observação esperada para a função de perda definida

$$\min_{\hat{y}} E[\rho_\tau(Y - \hat{y})]. \tag{2.3}$$

No caso de a função de distribuição acumulada ser substituída por $F_n(\hat{y})$, uma função de distribuição empírica é

$$F_n(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq \hat{y}).$$

A escolha de $\hat{y}$ para minimizar a perda esperada,

$$\int \rho_\tau(y_i - \hat{y}) dF_n(y) \; = \; \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}) \; ,$$

acarreta no quantil amostral de ordem de τ. O problema de otimização, para encontrar o quantil amostral de ordem τ, dá-se da seguinte forma

$$\hat{q_\tau} = \min_{q_\tau \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - q_\tau). \tag{2.4}$$

Para exemplificar o problema de minimização (2.4), a seguir, construiu-se alguns dados fictícios com auxílio do *software* R.

Exemplo 2.1. Seja $Y_1, ..., Y_n$ uma amostra aleatória com $n = 1.000$ observações de uma variável $Y \sim \mathcal{N}(0,5 \; ; \; 0,5)$. Assim, computou-se a função (2.4) para $q_\tau = y$, em que y representa cada observação da amostra, e para os diferentes valores de quantis, $\tau \in (0,25, \; 0,5, \; 0,75)$.

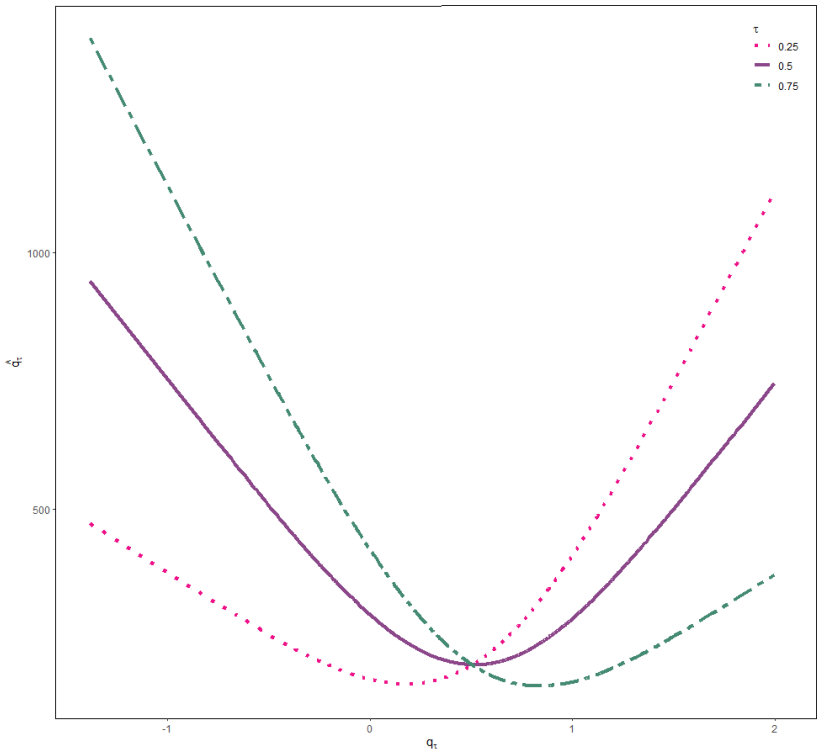

Figura 2.1: Exemplo de $\hat{q_\tau} = \sum \rho_\tau(y_i - q_\tau)$, em que $\tau \in (0,25, \; 0,5, \; 0,75)$.

Os valores de q_τ, que minimizam $\hat{q_\tau}$, são iguais a $0,176$, $0,526$ e $0,828$, conforme os respectivos $\tau = (0,25, \; 0,5, \; 0,75)$, nesse exemplo.

2.2 Estimação para regressão quantílica

A regressão quantílica avalia o impacto de um vetor de preditores $\mathbf{X} = \mathbf{x}$, p-dimensional, em uma variável resposta Y (escalar), definida para $\tau \in [0, 1]$. Assim, a função do quantil de ordem τ condicional é dada por

$$Q_Y(\tau|x) = \inf\{y : P\{Y \leq y \mid X = x\} \geq \tau\}, \tag{2.5}$$

em que $Q_Y(\tau|x) = Q_\tau(Y|X = x)$. Como já foi mencionado, Koenker e Bassett Jr (1978) defenderam a substituição de um modelo linear para a média de resposta pelo modelo

$$Q_Y(\tau|x) = \beta_0(\tau) + x'\beta_1(\tau) + \varepsilon, \tag{2.6}$$

em que $\tau \in [0, 1]$, os coeficientes quantílicos são $\beta_0(\tau) \in \mathrm{R}$ e $\beta_1(\tau) \in \mathrm{R^p}$ e ε é o vetor de erros desconhecidos, supondo que o quantil condicional de ordem τ é zero, isto é, $Q(\tau|x) \equiv \inf\{a : P(\varepsilon \leq a|X = x) \geq \tau\} = 0$ ou $P(\varepsilon \leq 0|X = x) = \tau$ (Rodrigues et al., 2016; Davino et al., 2013). Pode-se interpretar o coeficiente β_τ como a taxa de variação no τ-ésimo quantil da variável resposta Y ao variar em uma unidade o valor do i-ésimo regressor, mantendo os valores fixos das demais variáveis (Rasteiro, 2017). Do mesmo modo, como o quantil da amostra de ordem τ resolve (2.4), logo, leva-se a uma abordagem da equação (2.6), obtendo-se

$$\hat{Q}_Y(\tau|x) = \min_{Q_Y(\tau|x)} \sum_{i=1}^{n} \rho_\tau(y_i - Q_Y(\tau|x)), \tag{2.7}$$

em que $Q_Y(\tau|x) = Q_\tau(Y|X = x)$ denota uma função (geral) do quantil condicional.

Seja $(y_i, x_i), \forall i \geq 1$, com n valores observados de (Y, X). Na inferência do modelo de regressão quantílica linear, a estimação de $\beta(\tau)$ é realizada resolvendo o problema de minimização

$$\hat{\beta}(\tau) = \min_{\beta \in \mathrm{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b), \tag{2.8}$$

ou seja, visando obter uma estimativa de β que minimize essa função.

Agora revisar-se-á alguns métodos, como simplex e ponto interior para resolver a equação (2.8). No caso particular de tendência central, a soma de erros absolutos é minimizada pelo estimador da regressão mediana. As outras funções quantílicas condicionais são estimadas minimizando uma soma de erros absolutos assimetricamente ponderados.

Algoritmo Simplex

O problema de regressão quantílica (2.8), ou seja, o cálculo do estimador $\hat{\beta}(\tau)$, pode ser reescrito como o problema de programação

$$\min_{\beta \in \mathrm{R}^p} \left[\sum_{i \in \{i: y_i \geq x_i \beta\}} \tau |y_i - \mathbf{x_i}' \boldsymbol{\beta}| + \sum_{i \in \{i: y_i < x_i' \beta\}} (1-\tau) |y_i - \mathbf{x_i}' \boldsymbol{\beta}| \right]$$

$$\min_{(\beta, u, v) \in \mathrm{R}^p \times \mathrm{R}^{2n}_+} \{\tau \mathbf{1}_n' \mathbf{u} + (\mathbf{1} - \tau) \mathbf{1}_n' \boldsymbol{\upsilon} \mid \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \boldsymbol{\upsilon}\}, \tag{2.9}$$

$$\text{sujeito a } \beta \in \mathrm{R}^p, \ \mathbf{u} \geq 0 \ \ e \ \ \upsilon \geq 0,$$

em que $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{X} = (x_1, ..., x_n)'$, $\mathbf{1_n}$ é um vetor $(n \times 1)$ de valores iguais a um, $\mathbf{u} = [y - \mathbf{X}\boldsymbol{\beta}]_+$, $\boldsymbol{\upsilon} = [\mathbf{X}\boldsymbol{\beta} - y]_+$ e $[z]_+$ é a parte não-negativa de z. Levando isso alguns passos a frente, tem-se que $\boldsymbol{\theta} = (\boldsymbol{\psi}', \boldsymbol{\phi}', \mathbf{u}', \boldsymbol{\upsilon}')$, $\boldsymbol{\phi} = [\ \beta\ ]_+$, $\boldsymbol{\psi} = [\ -\beta\ ]_+$, $\mathbf{B} = [\mathbf{x} \quad -\mathbf{x} \quad \mathbf{I_n} \quad -\mathbf{I_n}]$, em que I_n é a matriz identidade de ordem n, e $\mathbf{d} = (\mathbf{0}', \mathbf{0}', \tau \mathbf{1}_n', (\mathbf{1} - \tau)\mathbf{1}_n')'$, em que $\mathbf{0}' = (0, 0, ..., 0)_p$. Pode-se escrever o problema como um problema primal de minimização de programação linear na forma padrão

$$(P) \quad \min_{\theta} \mathbf{d}' \boldsymbol{\theta} \tag{2.10}$$

$$\text{s.a. } \mathbf{B}\boldsymbol{\theta} = y,$$

$$\boldsymbol{\theta} \geq 0.$$

Assim, a resolução do problema (2.8) pode se dar através do método simplex, o qual começa em uma solução básica factível (um dos pontos extremos). Se for a solução ótima, parará; mas se não for, continuará deslocando de vértice para vértice no exterior do conjunto de restrições lineares. O algoritmo continua até fornecer o menor valor para a função objetivo (ou maior, para problemas de maximização), comparada com a atual, ou seja, chegar ao ótimo (novo vértice).

O livro de Dantzig (2002) relata que suas ideias sobre o método simplex, surgiram em 1947 como um ensaio para a resolução de uma classe de problema de planejamento militar, empregando métodos semelhantes aos que ele utilizou em estudos anteriores com Wald e Neyman sobre o lema de Neyman-Pearson. No entanto, Barrodale e Roberts (1973) foram quem propuseram o primeiro algoritmo que explora de forma dual as variáveis limitadas do problema de minimização dos erros absolutos do modelo de regressão linear. Já com

relação a regressão quantílica, Koenker e d'Orey (1987) sugeriram um algoritmo modificado do método simplex, mais eficiente, para esse tipo de problema quando o tamanho do conjunto de dados é moderado. Além disso, o algoritmo é o mais estável por sempre encontrar uma solução, apesar de ser lento para um grande número de observações (Chen e Wei, 2005). Essa abordagem de Barrodale e Roberts (1973) está implementada no pacote *quantreg* do *software* R, utilizando o comando *method = "br"*. Para maiores detalhes sobre o método simplex vide Koenker et al. (2017) e Santos (2012).

Algoritmo de Ponto Interior

Já o método de ponto interior é computacionalmente superior se comparado ao tradicional método simplex para grandes problemas de programação linear e para grandes aplicações de regressão quantílica (Portnoy et al., 1997). Esse algoritmo é um conjunto de técnicas iterativas que, em vez de percorrer o exterior do conjunto de restrições, como no simplex, desloca-se para o interior do conjunto de restrições (região factível) em direção a uma solução de vértice.

A abordagem do método ponto interior por Karmarkar (1984) apresenta uma conexão estreita com o trabalho sobre métodos de barreira para otimização restrita, de Fiacco e Mccormick (1968), e também para programas lineares de Frisch (1956). Seguindo a exposição em Portnoy et al. (1997), considere o programa linear canônico

$$\min \{c'x \mid Ax = b, x \geq 0\}. \tag{2.11}$$

Associa-se esse problema à reescrita da barreira logarítmica (método de Frisch)

$$B(x, \mu | Ax = b), \tag{2.12}$$

em que

$$B(x, \mu) = c'x - \mu \sum \log(x_k).$$

A equação (2.12) substitui as restrições de desigualdade em (2.11) com um termo de penalidade da barreira do logaritmo, assim, minimizando (2.12) com uma sequência de parâmetros μ, tal que $\mu \to 0$. Com isso, obtém-se, no limite, uma solução para o problema original (2.11). O problema modificado mostra uma clara vantagem, pois gera passos do método de Newton para qualquer μ fixo. Portnoy et al. (1997) escreveram o problema quadrático (Newton) para uma direção resultante de descida, p, inciando em x

como (Koenker et al., 2017):

$$\min_p = \{c'p - \mu p' X^{-1} 1_n + \frac{1}{2}\mu p' X^{-2} p | Ap = b\},$$

em que

$$X = \boldsymbol{diag}(x) = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_p \end{pmatrix}.$$

Denotando um vetor de multiplicadores de Lagrange para a restrição de igualdade por y, esse problema produz condições de primeira ordem

$$\{c - \mu X^{-1} 1_n + \mu X^{-2} p = A'y, Ap = 0\},$$

que, multiplicando-se por AX^2, pode ser reformulado como

$$AX^2 A'y = AX^2 c - \mu AX^1 1_n.$$

Resolver y e substituir as condições de primeira ordem produz uma direção de Newton, δ. A dificuldade inerente de cada passo desse método de barreira primordial consiste, portanto, em resolver o sistema linear $p \times p$ nessa equação. Koenker et al. (2017) mencionam que alguma melhoria no desempenho pode ser alcançada explorando as formulações primais e duais do problema. O dual do problema canônico pode ser expresso como (Koenker et al., 2017):

$$\max_y \{b'y \mid A'y + z = c, \;\; z \geq 0\}.$$

A otimização na primal implica em $c - \mu X^{-1} 1_n = A'y$, então, pode-se definir $z = \mu X^{-1}$ para satisfazer a restrição dual e obter o sistema

$$Ax = b, x \geq 0,$$
$$A'y + z = c, z \geq 0,$$
$$Xz = \mu 1_n.$$

A trajetória paramétrica $(x(\mu), y(\mu), z(\mu))$ descreve o "caminho central" do centro da restrição definido para uma solução no limite do conjunto de restrições, satisfazendo a clássica condição de folga complementar, $Xz = 0$, quando $\mu = 0$. Essa formulação primal-dual novamente resulta em um sistema linear $p \times p$, que requer o mesmo esforço

computacional em cada iteração (Koenker et al., 2017). Para completar a descrição do método primal-dual, precisaria especificar a distância a percorrer na direção p, como ajustar μ à medida que avança ao longo do caminho central e como parar; esses detalhes podem ser encontrados em Chen e Wei (2005); Portnoy et al. (1997).

Chen e Wei (2005) mencionam que o algoritmo de ponto interno é simples, facilmente adaptado em outras situações (por exemplo, regressão quantil restrita), e muito rápido para conjuntos de dados que possuem muitas observações e um pequeno número de variáveis explicativas, mas apresenta dificuldade na presença de *outliers* nas covariáveis. O método ponto interior está implementado no pacote *quantreg* do *software* R, utilizando o comando *method = "fn"* ou *method = "pfn"*, este se o caso é o uso do pré-processamento, que melhora substancialmente o desempenho do algoritmo (Santos, 2012; Koenker et al., 2017).

2.3 Propriedades da Regressão quantílica

Teorema 1 (Koenker e Bassett Jr, 1978) - Seja a matriz $A_{p\times p}$ não singular, $\gamma \in \mathrm{R}^p$ e $a > 0$. Então, para qualquer $\tau \in [0,1]$,

(i) Equivariância de escala:

* $\hat{\beta}_\tau(ay, X) = a\hat{\beta}_\tau(y, X)$,
* $\hat{\beta}_\tau(-ay, X) = a\hat{\beta}_{1-\tau}(y, X)$.

(ii) Equivariância de regressão: $\hat{\beta}_\tau(y + X\gamma, X) = \hat{\beta}_\tau(y, X) + \gamma, \ \ \gamma \in \mathrm{R}^p$.

(iii) Equivariância da reparametrização da matriz de *design*: $\hat{\beta}_\tau(y, AX) = A^{-1}\hat{\beta}_\tau(y, X)$, $|A| \neq 0$.

(iv) Equivariância a transformações monótonas: $Q_\tau(h(Y)|X) = h(Q_\tau(Y|X))$.

O item (iv) é outra propriedade que os quantis dispõem, como mencionado anteriormente. Essa propriedade é essencial para uma maior compreensão do potencial da regressão quantílica. Muitas vezes surgem situações em que as transformações são consideradas para obter linearidade ou uma distribuição mais próxima da almejada. As transformações de logaritmo, por exemplo, são muito utilizadas para corrigir a assimetria à direita de uma distribuição. Isso permite uma interpretação das estimativas em termos relativos, no caso

de modelos de regressão linear. Entretanto, em termos absolutos, a média condicional da variável dependente não pode ser obtida a partir da média condicional na escala de log, por exemplo, pois é evidente que não usufrui dessa propriedade,

$$Eh(Y) \neq h(E(Y)).$$

A transformação modifica sobretudo o que está sendo estimado na regressão de mínimos quadrados (Hao e Naiman, 2007).

Já a regressão quantílica obtém o seguinte resultado:

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)), \tag{2.13}$$

em que $h(\cdot)$ é uma função não decrescente em R e $Q_Y(\tau)$ é o quantil de ordem τ de qualquer v.a. Y de interesse apoiada no fato elementar que, para qualquer h monótona,

$$P(Y \leq y) = P(h(Y) \leq h(y)).$$

A robustez também é uma importante característica da regressão quantílica, que está relacionada à violação de pressupostos de modelo relativos à variável resposta e a não sensibilidade a *outliers*. Essa propriedade surge em decorrência da essência da função de distância,

$$\sum_{i=1}^{n} d_\tau(y_i, \hat{y_i}) = \tau \sum_{i:y_i \geq b(\tau)x_i} |y_i - b(\tau)x_i| + (1-\tau) \sum_{i:y_i < b(\tau)x_i} |y_i - b(\tau)x_i|,$$

que é minimizada, pois alterando os valores da variável dependente de interesse, não modificando o sinal do resíduo, a linha ajustada continua a mesma. Do mesmo modo, é muito limitada a influência dos *outliers*, no que se refere aos quantis univariados.

Para o modelo de regressão linear clássico da média, quando se computa a matriz de variância e covariância das estimativas, a suposição de normalidade faz-se necessária. A violação dessa suposição pode proporcionar imprecisão nos erros padrão. Já a regressão quantílica apresenta robustez a suposições distributivas devido ao estimador pesar o comportamento local da distribuição adjacente ao quantil de interesse mais do que o comportamento remoto da distribuição (Hao e Naiman, 2007).

2.4 Inferência para o modelo de regressão quantílica

Nesta seção, abordar-se-á a obtenção de intervalos de confiança para os parâmetros do modelo regressão quantílica, tratando sucintamente os seguintes métodos: o assintótico e o *boostrap*. Também discorrerá sobre o teste de hipótese linear geral para regressão quantílica.

2.4.1 Intervalo de confiança: método assintótico

A teoria assintótica é uma abordagem elementar também oferecida pela visão de otimização dos quantis amostrais ordinários, cujos resultados obtidos em modelo amostral se generalizam como o modelo clássico de regressão linear (Koenker, 2005). Seja o modelo linear:

$$y_i = \beta_0(\tau) + \mathbf{x}_i'\beta_1(\tau) + \varepsilon_i,$$

em que, para algum τ de interesse, supondo que ϵ_i tenha distribuição $F(\cdot)$ e sejam erros independentes e identicamente distribuídos (iid). Então, as funções quantílicas de y_i são (determinado em (2.6)):

$$Q(\tau|x) = \beta_0(\tau) + \mathbf{x}_i'\beta_1(\tau) + F_{\varepsilon_i}^{-1}(\tau),$$

na qual F_{ε_i} denota a função de distribuição comum dos erros, para algum τ de interesse, e que o quantil de ordem τ de ϵ_i seja igual a zero.

Teorema 2 (Koenker e Bassett Jr, 1978) - Seja uma sequência de estimadores $\{\hat{\beta}(\tau_1), \hat{\beta}(\tau_2), ..., \hat{\beta}(\tau_m)\}$, com $0 < \tau_1 < ... < \tau_m < 1$, para os parâmetros do modelo (2.6), em que $F(\cdot)$ é contínua e tem densidade $f(\cdot)$ contínua e positiva em $Q(\tau)$, sendo $Q(\tau) = F^{-1}(\tau_i)$, que é o quantil de ordem τ_i. Verifica-se que, na configuração em que os erros da regressão quantílica são independentes e identicamente distribuídos (iid), o estimador $\hat{\beta}(\tau)$ é não viciado e segue uma distribuição normal assintótica:

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{D} \mathcal{N}(0,\ \mathbf{V}(\tau_1, ..., \tau_m)), \tag{2.14}$$

em que a matriz de covariâncias $\mathbf{V}(\tau)$, pode ser dada como $\Omega(\tau_1, ..., \tau_m,\ \mathbf{F}) \otimes D^{-1}$, sendo $\Omega(\tau_1, ..., \tau_m,\ \mathbf{F})$ a matriz de covariâncias dos m quantis amostrais correspondentes de amostras aleatórias com distribuição F, $\otimes$ é o produto de *Kronecker* e D é uma matriz positiva definida como $\lim_{n \to \infty} n^{-1} \sum \mathbf{x}_i\mathbf{x}_i'$ e $x_{1n} = 1 : n = 1, 2, ...$ Quando há um

único τ, a matriz de variância e covariância na equação (2.14) se reduz a

$$\mathbf{V}(\tau) = \frac{\tau(1-\tau)}{f^2(0)}(\mathbf{X}'\mathbf{X})^{-1}, \tag{2.15}$$

na qual $f_i, \forall i \in [0, n]$ é a função densidade dos erros (Santos, 2012).

No caso de erros independentes, mas não identicamente distribuídos, com as funções de densidade de probabilidade (f_i), a teoria assintótica de $\hat{\beta}(\tau)$ é um pouco mais complexa, segundo Koenker (2005). A distribuição assintótica de $\sqrt{n}(\hat{\beta}(\tau)-\beta(\tau))$ segue uma normal com média zero e matriz de variância e covariância

$$\mathbf{V}(\tau) = \tau(1-\tau)\mathbf{H}_n^{-1}\mathbf{J}_n\mathbf{H}_n^{-1} \tag{2.16}$$

que possui a forma "sanduíche" de Huber et al. (1967), conforme a seguir:

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{D} \mathcal{N}(0,\ \tau(1-\tau)\mathbf{H}_n^{-1}\mathbf{J}_n\mathbf{H}_n^{-1}), \tag{2.17}$$

em que

$$J_n(\tau) = n^{-1}\sum_{i=1}^{n} x_i x_i'$$

e

$$H_n(\tau) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i x_i' f_i(\hat{\beta}(\tau_i)),$$

dado que $f_i(\hat{\beta}(\tau_i))$ é a função de densidade de probabilidade da variável dependente considerada no quantil condicional (de interesse) de ordem τ.

Nos dois casos expostos (2.15) e (2.16), há maneiras diferentes de estimar a matriz de variância e covariância. Para estimar $\mathbf{V}(\tau)$ da equação (2.15), conforme Kocherginsky et al. (2005), pode-se obter uma estimativa de $1/f(0)$ utilizando a diferença entre os quantis empíricos (Santos, 2012):

$$\frac{\hat{F}^{-1}(\tau + h_n) - \hat{F}^{-1}(\tau - h_n))}{2h_n},$$

com $lim_{n\to 0}\ h_n = 0$ e h_n computado segundo Hall e Sheather (1988). Na função *summary.rq* do pacote *quantreg* do *software* R, para usar esse método de inferência, basta tomar o comando *se = "iid"*, que fornecerá os valores das estimativas dos parâmetros do modelo, seus erros padrão e significância de cada estimativa (Santos, 2012).

Já para estimar $\mathbf{V}(\tau)$ na equação (2.16), uma forma é atribuída através de h_n e, na sequência, a obtenção de $\boldsymbol{H}_n$ (Koenker, 2005). Assim, um estimador assintoticamente

não viciado para $f_i(\hat{\beta}(\tau_i))$ computa-se do seguinte modo (Kocherginsky et al., 2005):

$$\frac{2h_n}{(x_i'\hat{\beta}_{\tau+h_n} - x_i'\hat{\beta}_{\tau-h_n})}.$$

Note que se $f_i(\hat{\beta}(\tau_i)) = f(\hat{\beta}(\tau_i)), \forall i,$ i.e., sob a suposição de erros aleatórios independentes identicamente distribuídos, a matriz de variância e covariância corresponde com a exposta na equação (2.15) (Rasteiro, 2017). E tem-se

$$\hat{\boldsymbol{H}}_n(\tau) = n^{-1} \sum_{i=1}^{n} \hat{f}_i \left[\hat{\beta}(\tau_i)\right] x_i x_i',$$

assumindo que $\hat{\boldsymbol{H}}_n \xrightarrow{D} \boldsymbol{H}_n$. Já para esse método de inferência, utiliza-se o comando *se = "nid"* também na função *summary.rq* do pacote *quantreg* do *software* R (Santos, 2012).

Por conseguinte, utilizando o teorema 2, pode-se obter um intervalo de confiança assintótico para os parâmetros β_τ dos modelos de regressão quantílica com a estimativa da matriz de variância-covariância. Para mais detalhes sobre esses resultados assintóticos, vide Koenker e Bassett Jr (1978), Hendricks e Koenker (1992), Koenker e Machado (1999), Koenker e Xiao (2002), Koenker (2005) e Kocherginsky et al. (2005).

2.4.2 Intervalo de confiança: método bootstrap

A reamostragem é uma alternativa evidente à inferência assintótica, por estimar erros-padrão de parâmetros sem requerer suposição alguma no tocante à distribuição de erros (Gould et al., 1993). *Bootstrap* em pares representa uma forma muito eficaz de estimação da matriz de variância-covariância e de obter intervalos de confiança, (Efron e Tibshirani, 1994). Para os modelos de regressão quantílica, as técnicas de *bootstrap* foram consideradas do mesmo modo eficazes (Parzen et al., 1994); entretanto, demandam cálculos repetidos das estimativas do quantil de regressão, o que pode ocasionar em um maior custo computacional, principalmente quando n e p são grandes. Afirma-se que, para obter uma estimativa aceitável da matriz de variância-covariância, são necessárias 50 réplicas de *bootstrap*. (Efron e Tibshirani, 1994).

Ademais, a despeito das estimativas de *bootstrap* serem consideradas imparciais, elas expõem diferentes causas de variabilidade, posto que são baseadas em um número finito de replicações (variabilidade de reamostragem de *bootstrap*) e em uma única amostra de uma determinada população (Davino et al., 2013). Há várias técnicas utilizando reamostragem, como:

- Método pares xy ou *bootstrap* da matriz de design (Kocherginsky, 2003);
- Método baseado em funções de estimativa pivotantes (Parzen et al., 1994);
- *Bootstrap* marginal da cadeia de Markov (He e Hu, 2002; Kocherginsky, 2003; Kocherginsky et al., 2005).

O método pares xy representa a construção de um dado número de amostras (B), usualmente com o mesmo tamanho do conjunto de dados original, em que cada amostra obtida se dá através de um processo de amostragem aleatória com substituição do conjunto original de dados. O processo de reamostragem é aplicado juntamente aos vetores x e y. As regressões quantílicas B são efetuadas nas amostras de *bootstrap* e um vetor das estimativas dos parâmetros é capturado para cada quantil de interesse. Nesse caso, sob a distribuição normal assintótica, um intervalo de confiança para os parâmetros β_τ dos modelos de regressão quantílica é dado por (Davino et al., 2013):

$$\underset{100(1-\alpha)\%}{\texttt{I.C.}}\left[\hat{\beta}_i(\tau)\right] = \left[\hat{\beta}_i(\tau) - \widehat{EP(\hat{\beta}_i(\tau))} \times z_{(1-\alpha/2)};\ \hat{\beta}_i(\tau) + \widehat{EP(\hat{\beta}_i(\tau))} \times z_{(1-\alpha/2)}\right],$$

em que $\widehat{EP(\hat{\beta}_i(\tau))}$ é o estimador do erro padrão do estimador do parâmetro $\beta_i(\tau)$ e $z_{(1-\alpha/2)}$ é o quantil de ordem $(1-\alpha/2)$ da distribuição normal padrão.

No caso de *bootstrap* baseado em funções de estimativa pivotais, também chamado de método pwy, a metodologia consiste na reamostragem da condição de subgradiente, utilizada nos modelos de regressão quantílica, para obter as estimativas dos parâmetros, que é uma grandeza pivotal para o verdadeiro parâmetro do quantil de ordem τ (Davino et al., 2013).

Já *bootstrap* marginal de cadeia de Markov (MCMB) tem como objetivo a redução de exigências computacionais decorrentes do uso dos processos de *bootstrap* mencionados anteriormente. A ideia subjacente é computar equações unidimensionais para cada replicação de *bootstrap* em lugar de um sistema p-dimensional requisitado pelas outras abordagens de *bootstrap*, onde p corresponde ao número de variáveis explicativas (Davino et al., 2013). Para maiores informações sobre os diferentes tipos de métodos *bootstrap*, vide também Santos (2012) e Koenker et al. (2017).

2.4.3 Teste de Hipótese Linear Geral

Os conceitos, nesta subseção, foram baseados no livro de Koenker (2005).

Seja a hipótese linear geral sobre o vetor $\beta = (\beta(\tau_1)', \beta(\tau_2)', ..., \beta(\tau_m)')'$ da forma

$$H_0 : \boldsymbol{R}\beta = \mathbf{r}, \tag{2.18}$$

em que $\boldsymbol{R}_{p\times p}$ é uma matriz de posto completo de constantes conhecidas e $\mathbf{r}_{m\times 1}$ é um vetor de constantes também conhecidas, e a estatística de teste é

$$T_n = n(\boldsymbol{R}\hat{\beta} - \mathbf{r})'[\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\beta} - \mathbf{r}),$$

em que $\mathbf{V}_n$, sendo $n = mp \times mp$, é uma matriz com o bloco i, j dado por

$$\mathbf{V}_n(\tau_i, \tau_j) = [\min(\tau_i, \tau_j) - \tau_i\tau_j]H_n(\tau_i)^{-1}J_n(\tau_i, \tau_j)H_n(\tau_j)^{-1},$$

na qual a estatística T_n, sob H_0, tem distribuição assintótica $\mathcal{X}^2_q$; H_n e J_n foram definidos anteriormente.

Essa formulação acomoda uma vasta diversidade de situações de teste, desde de testes simples com um único coeficiente de regressão quantílica até testes conjuntos compreendendo muitas covariáveis e quantis distintos. Esses testes proporcionam uma alternativa robusta aos testes convencionais baseados em mínimos quadrados, devido às propriedades básicas da regressão quantílica (Koenker e Bassett Jr, 1978). Para ilustrar a Seção 2.4, será descrito o modelo de regressão quantílica e os intervalos de confiança, através do exemplo a seguir.

Exemplo 2.2. Esse estudo de simulação proposto, inspirado no exemplo de Rasteiro (2017), conduzido para abordar o modelo de regressão quantílica, foi construído com uma covariável, considerando uma amostra com $n = 10.000$ observações. Seja a variável resposta, Y, gerada a partir do seguinte modelo padrão de regressão linear:

$$Y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \varepsilon_i, \quad \text{para} \quad i = 1, ..., n,$$

em que ε_i são independentes e identicamente distribuídos com $\varepsilon \sim \mathcal{N}(0, 1)$ e x_i são preditores com $\mathbf{X} \sim \text{U}(3, 5)$. Os parâmetros β_0 e β_1 são definidos como 3 e 4, respectivamente. Para esse exemplo, foram escolhidos os seguintes quantis de ordem

$\tau = (0,10; 0,25; 0,50; 0,75; 0,90)$. Assim, o modelo de regressão quantílica, que modela quantis condicionais como funções de covariáveis, utilizado foi

$$Q_Y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x_i.$$

Para estimação e extração de inferências sobre as funções quantílicas condicionais, nesse estudo, empregou-se o pacote *quantreg* do *software* R, que é uma implementação da regressão quantílica, que auxiliou no cálculo para a obtenção das estimativas b_0 e b_1 dos parâmetros. Os erros padrão foram computados pelo método *bootstrap*, técnica *xy-pair*, fazendo uso da função *summary.rq*, onde se deve acrescentar o comando *se = "boot"* e especificar como *bsmethod = "xy"*, desse modo inferindo sobre o vetor de parâmetros. Os resultados, na tabela a seguir, apresentam as estimativas b_0 e b_1 dos parâmetros β_0 e β_1, segundo os valores de τ especificados e seus erros padrão correspondentes.

Tabela 2.1 - Estimativas dos parâmetros para regressão quantílica.

τ	Parâmetro	Estimativa	Erro Padrão	Wald	p-valor	Intervalo de Confiança	
						L.I.	L.S.
0,10	β_0	1,656	0,057	28,988	$<$ 0,001	1,544	1,768
	β_1	3,545	0,179	19,782	$<$ 0,001	3,194	3,896
0,25	β_0	2,220	0,039	57,497	$<$ 0,001	2,145	2,296
	β_1	4,629	0,071	64,999	$<$ 0,001	4,490	4,769
0,50	β_0	2,925	0,027	107,325	$<$ 0,001	2,872	2,979
	β_1	4,823	0,050	97,421	$<$ 0,001	4,726	4,920
0,75	β_0	3,575	0,035	103,720	$<$ 0,001	3,507	3,642
	β_1	4,956	0,055	90,560	$<$ 0,001	4,849	5,063
0,90	β_0	4,211	0,038	110,076	$<$ 0,001	4,136	4,286
	β_1	4,904	0,072	68,302	$<$ 0,001	4,764	5,045

Os valores das estatísticas de Wald, também, estão exibidos na Tabela 2.1, que verifica a hipótese de interesse se todos os parâmetros $\beta_i(\tau)$ são iguais a zero versus a hipótese alternativa, de que pelo menos um deles difere de zero; segundo resultados, contata-se que, a nível de 5% de significância, os parâmetros diferem significativamente de zero. Em consonância com isso, são apresentados os valores dos intervalos com 95% de confiança (I.C.), conforme suas respectivas estimativas e quantis de ordem τ, que foram calculados a parte, com auxílio dos resultados fornecidos pelos pacotes mencionados. Desse modo,

conforme as estimativas obtidas, pode-se apontar os modelos finais de regressão quantílica, tais como

$$\hat{Q}_Y(0,10|x) = 1,656 + 3,545x_i,$$
$$\hat{Q}_Y(0,25|x) = 2,220 + 4,629x_i,$$
$$\hat{Q}_Y(0,50|x) = 2,925 + 4,823x_i,$$
$$\hat{Q}_Y(0,75|x) = 3,575 + 4,956x_i,$$
$$\hat{Q}_Y(0,90|x) = 4,211 + 4,904x_i.$$

Na regressão quantílica, a interpretação dos parâmetros, desses resultados, dá-se da seguinte forma, por exemplo: aumentando em uma unidade a covariável x_i, estima-se que a mediana (quantil de ordem $\tau = 0,5$) de y_i aumente em 4,823 unidades. Sendo assim, a leitura é feita na forma de taxa de variação no quantil de ordem τ de interesse ao variar o valor da variável explicativa.

No próximo capítulo, será abordado a aplicação da regressão quantílica para dados censurados, denotando como uma alternativa mais abrangente da distribuição de sobrevivência a outras técnicas clássicas, visto que o modelo propicia às covariáveis efeitos variados em cada nível de quantil no tempo de acompanhamento.

Capítulo 3

Regressão quantílica para dados censurados

A atenção ao modelo de regressão quantílica vem aumentando cada vez mais na análise de sobrevivência, devido ao interesse científico nos tempos de evento, e os quantis são ferramentas quantitativas com maior robustez e mais flexíveis para caracterizar os tempos de evento do que os dispositivos baseados em média, como já mencionado. Além disso, o modelo possibilita a análise de características locais, particulares da distribuição condicional, de um tempo de interesse do evento (Koenker et al., 2017).

A utilidade da técnica, para análise de sobrevivência, foi evidenciada por muitas aplicações relatadas na literatura. Powell (1984, 1986) foi o primeiro a introduzir o modelo de regressão quantílica para dados com tempos de censura "fixos", cujo termo fixo, nesse caso, representa que os valores censurados, para a variável dependente, são assumidos como sendo conhecidos por todas as observações. Essa técnica também é conhecida como modelo Tobit. Apesar dessa abordagem ter estabelecido um modo sagaz de corrigir a censura, a função objetivo não era convexa sobre os valores dos parâmetros, dificultando minimização global (Ou et al., 2016).

Para o modelo de regressão quantílica com dados censurados à direita, a estimação dos parâmetros foi estudada por diversos autores, alguns deles são Powell (1984), Pollard (1990), Rao e Zhao (1992), Ying et al. (1995) e Wang e Wang (2009). Em Portnoy (2003), há abordagem de uma técnica recursiva, que estima os parâmetros do modelo de regressão quantílica, aplicado a dados censurados, e assume linearidade dos quantis condicionais da variável resposta dada as variáveis explicativas, o que pode não ser verificada no geral. O Portnoy utiliza uma metodologia de ponderação, na qual, dado a fixação dos valores de τ, são calculados pesos atribuídos aos dados censurados. Já a utilização do modelo com dados duplamente censurados (ou seja, dados sujeitos a censura à esquerda e à direita),

a estimativa de amostra única e o teste de duas amostras foram estudados por Turnbull (1974), Chang et al. (1990), Zhang et al. (1996), Ren et al. (1997), McKeague et al. (2001), Chay e Powell (2001), Cai e Cheng (2004), Lin et al. (2012) e outros.

Em Koenker e Geling (2001), uma análise de regressão quantílica foi detalhada para um conjunto de dados de longevidade média, ilustrando como usar a regressão quantílica para avaliar e interpretar efeitos das covariáveis em diferentes segmentos da distribuição de tempo do evento. Tal capacidade constitui outra grande vantagem da regressão quantílica sobre os modelos de sobrevivência convencionais, como o modelo de riscos proporcionais e o modelo de tempo de falha acelerado, que implicitamente influenciam a mudança de localização pura dos efeitos nos tempos de sobrevivência ou nas suas transformações monótonas (Koenker et al., 2017).

Para dados com censura intervalar, estudos foram considerados por muitos autores, dentre eles: Turnbull (1976), Finkelstein e Wolfe (1985), Gentleman e Geyer (1994), Huang e Wellner (1997) e Li et al. (1998). Kim et al. (2010) versam sobre modelo de regressão mediana com dados censurados em intervalos, tendo proposto um estimador através de uma função que se baseia na adaptação do princípio da falta de informação (MIP) de Efron (1967). No entanto nenhuma propriedade foi obtida em teoria.

Ou et al. (2016) abordam metodologia para dados conhecidos como *current status* e propõem um modelo de regressão quantílica para analisá-los, visto que não exige hipóteses de distribuição e os coeficientes podem ser interpretados como efeitos de regressão direta na distribuição do tempo de falha na escala de tempo original. O modelo apresentado assume que o quantil condicional do tempo de falha é uma função linear das covariáveis. Assumiram a independência condicional entre o tempo de falha e o tempo de observação. Um M-estimador foi desenvolvido para estimação de parâmetros, que é computado utilizando o procedimento côncavo-convexo e seus intervalos de confiança são construídos através de um método de subamostragem. Para o estimador, as propriedades assintóticas são derivadas e comprovadas utilizando a moderna teoria de processos empíricos. O desempenho da amostra pequena do método proposto foi demonstrado através de aplicação via estudos de simulação e na análise de dados do *Mayo Clinic Study of Aging*, que estudou a incidência do comprometimento cognitivo leve (MCI) nos subtipos: amnéstico (aMCI) e não-amnéstico (naMCI) em homens e mulheres separadamente.

Em Zhou et al. (2017), com a generalização da regressão quantílica para dados observa-

dos completos, propôs-se um método de estimação para modelos de regressão quantílica a dados com censura intervalar. O estimador apresentado é definido como o ponto ótimo de solução de um problema de minimização com função objetiva convexa. A propriedade da normalidade assintótica é estabelecida com um viés convergindo para zero. Para reduzir o viés, dois métodos de correção de polarização foram propostos, baseados em *bootstrap* e a estimativa inicial respectivamente, os quais não exigem distribuição de forma idêntica dos vetores de censura e podem ser aplicados a modelos com muitas covariáveis, como covariáveis de desenho aleatório fixo, discreto aleatório ou contínuo. Além de fácil aplicação computacional, esses métodos apresentaram bons resultados, segundo os autores.

O método proposto nesse trabalho consiste em imputar o valor da variável resposta para as observações com intervalo finito, mantendo inalteradas as observações com censura à direita. Utiliza-se, então, a metodologia existente para ajuste do modelo de regressão quantílica para dados com censura à direita descrita em Rasteiro (2017) e, brevemente, tratada na subseção 3.2.4. Para proceder a imputação do valor da variável resposta com intervalo finito, primeiramente obtém-se o estimador não paramétrico de máxima verossimilhança (ENPMV) da função de distribuição (f.d.a.) condicional da variável resposta dado o valor das covariáveis contínuas (subseção 3.2.1.1). Para tal, adapta-se o método descrito em Groeneboom e Wellner (1992) para cálculo do ENPMV, atribuindo-se pesos diferentes às observações, proporcionais à distância dos valores das covariáveis $\{\boldsymbol{x}\}$ para cada observação, em relação aos valores das covariáveis para a observação na qual se está fazendo a imputação. Tais pesos podem ser calculados via núcleo estimador (kernel). Obtido o ENPMV da f.d.a., o valor imputado é obtido via método da bissecção, gerando um valor z_i a partir da distribuição uniforme no intervalo $[\hat{F}(t_{1i} \mid \mathbf{x}_i), \hat{F}(t_{2i} \mid \mathbf{x}_i)]$, e buscando o valor $\widetilde{y}_i$ tal que $\hat{F}(y_i \mid \mathbf{x}_i) = z_i$, em que $\hat{F}\ (\cdot \mid \mathbf{x}_i)$ é uma versão suavizada do ENPMV da f.d.a. condicional da variável resposta dado o valor das covariáveis. A versão suavizada do ENPMV de $F(\cdot \mid \mathbf{x}_i)$ também pode ser obtida via núcleo estimador. A figura 3.1 ilustra essa metodologia.

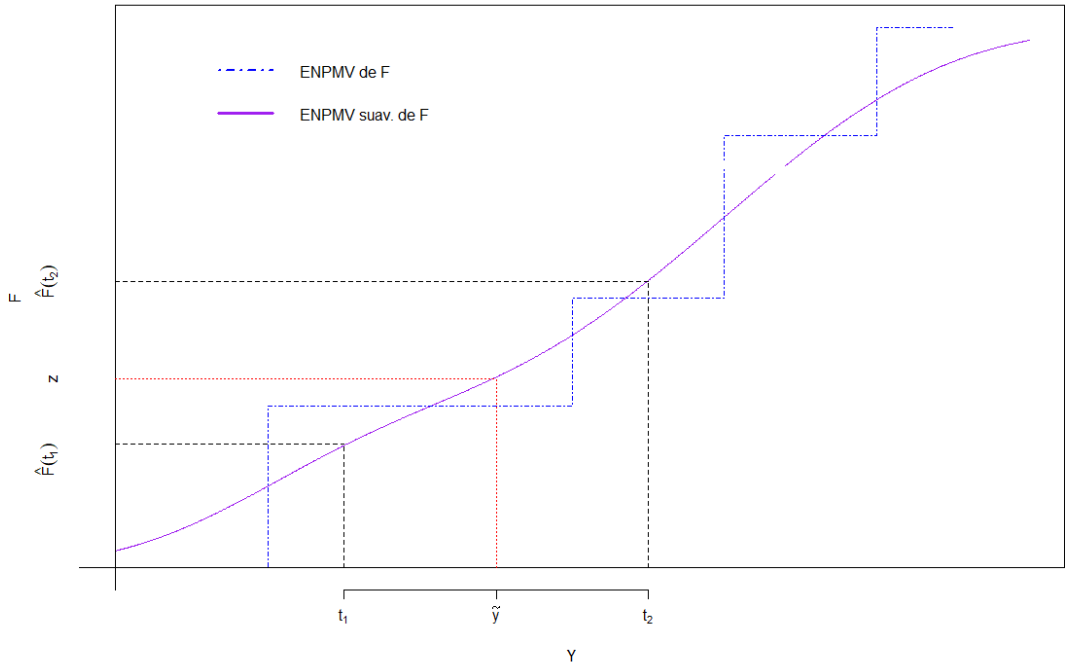

Figura 3.1: Gráficos do método proposto nessa dissertação.

Para covariáveis discretas, a imputação é feita, como a recém descrita, para subconjuntos dos dados determinados pelos valores das covariáveis discretas.

Nas próximas seções, tratar-se-ão os seguintes assuntos: censura intervalar (sucintamente), onde também versará sobre as técnicas que são utilizadas na nova proposta de metodologia, além de descrever, de forma breve, a técnica de Zhou et al. (2017).

3.1 Censura intervalar

Na análise de sobrevivência, os dados são quase inexoravelmente complicados por alguma forma de censura ou não resposta (Koenker, 2005). Os dados de censura intervalar, no tempo de sobrevivência Y, podem ser expressos como n vetores (T_{1i}, T_{2i}) da população (T_1, T_2), ou n intervalos I_i da população I. Dois tipos de dados por censura intervalar (CI) são, na sequência, abordados.

Caso I. Também conhecido como estado corrente (*current status*). Seja C o tempo de

censura, defina δ como o indicador de censura à esquerda, tal que

$$\delta_i = \begin{cases} 1, & \text{se } Y_i \leq C_i, \\ 0, & \text{se } Y_i > C_i. \end{cases}$$

Se $\delta = 1$, tem-se censura à esquerda e, se $\delta = 0$, tem-se censura à direita. Denota-se que a censura à esquerda equivale a $Y \in (0, t_{1i}]$ e a censura à direita a $Y \in [t_{1i}, \infty)$, sendo Y e t_{1i} variáveis independentes, em que Y não é observado e só um dos tempos de observação é informado (t_{1i}) (Koenker, 2005). Considere um estudo em que um animal de laboratório deve ser dissecado para verificar se um tumor se desenvolveu e para tanto deverá ser sacrificado. Nesse caso, Y é o início do tumor e C é o tempo da dissecção, portanto, somente se pode inferir, no momento da dissecção, se há presença de tumor ou se ainda não se desenvolveu, exemplo mencionado por Ayer et al. (1955).

Caso II. É considerado o tipo geral de censura intervalar, no qual

$$\begin{cases} \delta_i = 1, \ \gamma_i = 0, & \text{se } Y_i \leq t_{1i} \\ \delta_i = 0, \ \gamma_i = 1, & \text{se } t_{1i} < Y_i \leq t_{2i} \\ \delta_i = 0, \ \gamma_i = 0, & \text{se } Y_i > t_{2i} \end{cases},$$

em que δ_i e γ_i são indicadores de censura para o i-ésimo indivíduo, $i = 1, ..., n$. No caso geral, há registro de dois tempos de observação para o i-ésimo indivíduo, sendo $C = (t_{1i}, t_{2i})$ e $t_{1i} < t_{2i}$, (Koenker et al., 2017). Em estudos de acompanhamento médico, cada paciente realiza diversos exames periódicos e o evento de interesse só é conhecido antes do primeiro retorno, $Y \in (0, t_{1i}]$, ou entre dois instantes de observação consecutivos, $Y \in (t_{1i}, t_{2i}]$, ou após o último, $Y \in [t_{2i}, \infty)$, exemplo citado por Becker e Melbye (1991).

3.1.1 Função de verossimilhança

A função de verossimilhança, incorporando a censura intervalar, é descrita segundo os casos explanados a seguir.

Caso I. Seja $(c_i, \delta_i)_{i=1,...,n}$ e tomando Y e C variáveis independentes (não negativas), obtém-se a função de verossimilhança de F_Y a partir da distribuição conjunta de C e de δ, expressa a seguir (Silva, 2011). Aqui g representa a função densidade de probabilidade (f.d.p) de C.

Para $\delta = 1$, obtém-se

$$\begin{aligned} P(C \leq c,\ \delta = 1) &= P(C \leq c,\ Y \leq C) \\ &= \int_0^c \int_0^x f(y)g(x)dydx \\ &= \int_0^c g(t) \int_0^x f(y)dydx \\ &= \int_0^c g(x)F(x)dx \end{aligned}$$

e diferenciando essa equação com relação a c, tem-se que

$$\frac{\partial}{\partial c} \int_0^c g(x)F(x)dx = g(c)F(c).$$

Já em $\delta = 0$, obtém-se

$$\begin{aligned} P(C \leq c,\ \delta = 0) &= P(C \leq c,\ Y > C) \\ &= \int_0^c \int_x^\infty f(y)g(x)dydx \\ &= \int_0^c g(x) \int_x^\infty f(y)dydx \\ &= \int_0^c g(x)[1 - F(x)]dx, \end{aligned}$$

e diferenciando essa equação com relação a c, tem-se que

$$\frac{\partial}{\partial c} \int_0^c g(x)[1 - F(x)]dx = g(c)[1 - F(c)].$$

À vista disso, para o caso I, a função verossimilhança de F é

$$\mathrm{L}(F) \propto \prod_{i=1}^{n} [F(c_i)]^{\delta_i} [1 - F(c_i)]^{1-\delta_i}$$

e a função de log-verossimilhança é dada por

$$\mathcal{L}(F) = \sum_{i=1}^{n} \delta_i \log F(c_i) + (1 - \delta_i) \log[1 - F(c_i)], \tag{3.1}$$

na qual $F(c_i) = P(Y_i \leq c_i)$ é estritamente contínua.

Caso II. Seja $(t_{1i}, t_{2i}, \delta_i, \gamma_i)_{i=1,\ldots,n}$, assumindo que Y é independente de $(T_1,\ T_2)$ (que são variáveis não negativas) e $P(T_1 \leq T_2) = 1$, obtém-se a função de verossimilhança de F_Y usando a distribuição conjunta de T_1, T_2, δ e γ, expressa a seguir (Silva, 2011). Aqui h representa a função densidade de probabilidade (f.d.p) conjunta de T_1 e T_2.

Para $\delta = 1$ e $\gamma = 0$ (censura à esquerda), obtém-se

$$\begin{aligned} P(T_1 \leq t_1,\ T_2 \leq t_2,\ \delta_i = 1,\ \gamma_i = 0) &= P(T_1 \leq t_1,\ T_2 \leq t_2,\ Y \leq t_1) \\ &= \int_0^{t_1} \int_x^{t_2} \int_0^x f(y)h(x,z)dydzdx \\ &= \int_0^{t_1} \int_x^{t_2} h(x,z) \int_0^x f(y)dydzdx \\ &= \int_0^{t_1} \int_x^{t_2} h(x,z)F(x)dzdx \\ &= \int_0^{t_1} F(x) \int_x^{t_2} h(x,z)dzdx, \end{aligned}$$

e diferenciando essa equação com relação a t_1 e a t_2, tem-se que

$$\frac{\partial}{\partial t_1} \int_0^{t_1} F(x) \int_x^{t_2} h(x,z)dzdx = \int_{t_1}^{t_2} h(t_1,z)dz$$

e

$$\begin{aligned} \frac{\partial^2}{\partial t_1 \partial t_2} \int_0^{t_1} F(x) \int_x^{t_2} h(x,z)dzdx &= \frac{d}{dt_2} F(t_1) \int_{t_1}^{t_2} h(t_1,z)dz \\ &= F(t_1)h(t_1,t_2). \end{aligned}$$

Em $\delta = 0$ e $\gamma = 1$ (censura no intervalo $(t_1, t_2]$), obtém-se

$$\begin{aligned} P(T_1 \leq t_1,\ T_2 \leq t_2,\ \delta_i = 0,\ \gamma_i = 1) &= P(T_1 \leq t_1,\ T_2 \leq t_2,\ t_1 < Y \leq t_2) \\ &= \int_0^{t_1} \int_x^{t_2} \int_x^z f(y)h(x,z)dydzdx \\ &= \int_0^{t_1} \int_x^{t_2} h(x,z) \int_x^z f(y)dydtdx \\ &= \int_0^{t_1} \int_x^{t_2} h(x,z)[F(z) - F(x)]dzdx, \end{aligned}$$

e diferenciando essa equação com relação a t_1 e a t_2, tem-se que

$$\frac{\partial}{\partial t_1} \int_0^{t_1} \int_x^{t_2} h(x,z)[F(z) - F(x)]dzdx = \int_{t_1}^{t_2} h(t_1,z)[F(z) - F(x)]dz$$

e

$$\begin{aligned} \frac{\partial^2}{\partial t_1 \partial t_2} \int_0^{t_1} \int_x^{t_2} h(x,z)[F(z) - F(x)]dzdx &= \frac{d}{dt_2} \int_{t_1}^{t_2} h(t_1,z)[F(z) - F(t_1)]dz \\ &= h(t_1,t_2)[F(t_2) - F(t_1)]. \end{aligned}$$

Já para $\delta_i = 0$ e $\gamma_i = 0$ (censura à direita), obtém-se

$$\begin{aligned} P(T_1 \leq t_1,\ T_2 \leq t_2,\ \delta_i = 0,\ \gamma_i = 0) &= P(T_1 \leq t_1,\ T_2 \leq t_2,\ Y > t_2) \\ &= \int_0^{t_1}\int_x^{t_2}\int_z^{\infty} f(y)h(x,z)dydzdx \\ &= \int_0^{t_1}\int_x^{t_2} h(x,z)\int_z^{\infty} f(y)dydtdx \\ &= \int_0^{t_1}\int_x^{t_2} h(x,z)[1-F(z)]dzdx, \end{aligned}$$

e diferenciando essa equação com relação a t_1 e a t_2, tem-se que

$$\frac{\partial}{\partial t_1}\int_0^{t_1}\int_x^{t_2} h(x,z)[1-F(z)]dzdx = \int_{t_1}^{t_2} h(t_1,z)[1-F(z)]dz$$

e

$$\begin{aligned} \frac{\partial^2}{\partial t_1 \partial t_2}\int_0^{t_1}\int_x^{t_2} h(x,z)[1-F(z)]dzdx &= \frac{d}{dt_2}\int_{t_1}^{t_2} h(t_1,z)[1-F(z)]dz \\ &= h(t_1,t_2)[1-F(t_2)]. \end{aligned}$$

Em vista disso, a função de verossimilhança de F, desse caso, é dada por

$$\mathrm{L}(F) \propto \prod_{i=1}^{n}[F(t_{1i})]^{\delta_i}[F(t_{2i}) - F(t_{1i})]^{\gamma_i}[1-F(t_{2i})]^{1-\delta_i-\gamma_i}$$

e a função log-verossimilhança de F é

$$\mathcal{L}(F) = \sum_{i=1}^{n}\delta_i \log F(t_{1i}) + \gamma_i \log[F(t_{2i}) - F(t_{1i})] + (1-\delta_i-\gamma_i)\log(1-F(t_{2i})). \qquad (3.2)$$

A estimação dos parâmetros da distribuição dos tempos de falha dá-se através da maximização da função de verossimilhança, com a substituição da expressão de F na função log-verossimilhança, no respectivo caso estudado, segundo equação (3.1) ou equação (3.2).

Para estimar os parâmetros, utilizar-se-á um método iterativo baseado em regressões isotônicas proposto por Groeneboom e Wellner (1992) e Barlow (1972).

3.2 Regressão quantílica para censura intervalar

Nos casos mais simples de dados de sobrevivência, observa-se um tempo de censura de interesse a cada indivíduo e a resposta observada é o mínimo do tempo real do evento e do

tempo de censura (Koenker, 2005). Sendo Y o tempo de falha de interesse e X o vetor de covariável com dimensão $k \times 1$, considere um modelo de regressão quantílica, que se ajusta ao quantil condicional em função da covariável

$$Q_Y(\tau|\mathbf{X}) = \mathbf{X}'\beta(\tau),$$

no qual $0 < \tau < 1$ e $\beta(\tau)$ é o vetor de coeficientes de regressão desconhecidos, que representa os efeitos da covariável X no quantil de ordem τ de Y, que depende de τ. Cada elemento de $\beta(\tau)$ pode ser interpretado como uma mudança estimada no quantil de ordem τ de Y, dada uma mudança de uma unidade na covariável correspondente, enquanto outras variáveis no modelo são mantidas constantes. O foco reside na estimativa e na inferência de $\beta(\tau)$ (Koenker, 2005; Ou et al., 2016).

Em regressão quantílica, para os casos em que y_i são observados, o estimador é definido conforme a equação (2.8), $\hat{\beta}(\tau) = \min\limits_{\beta \in \mathbb{R}^p} \sum\limits_{i=1}^{n} \rho_\tau(y_i - x_i'b)$, e a função perda é $\widetilde{\beta}(\tau) = \rho_\tau(y_i - x_i'b)$, isto é,

$$\widetilde{\beta}(\tau) = \begin{cases} \tau|y_i - x_i'b|, & \text{se } y_i \geq x_i'b \\ (1-\tau)|y_i - x_i'b|, & \text{se } y_i < x_i'b. \end{cases}$$

Nos casos de censura intervalar (caso 2), note que $y_i - x_i'b > 0$ é válido para $x_i'b \leq t_{1i}$ e $y_i - x_i'b < 0$ é válido para $x_i'b > t_{2i}$. Seja a função perda quantílica definida como:

$$\begin{cases} \tau|y_i - x_i'b(\tau)|, & \text{se } t_{1i} \geq x_i'b(\tau) \\ \psi_i(\tau, \beta), & \text{se } t_{1i} < x_i'b(\tau) \leq t_{2i} \\ (1-\tau)|y_i - x_i'b(\tau)|, & \text{se } t_{2i} < x_i'b(\tau) \end{cases},$$

para algum $\psi_i(\tau, \beta)$.

Na subseções 3.2.1, 3.2.2, 3.2.3 e 3.2.4, a seguir, abordar-se-ão as técnicas que foram utilizadas no método proposto nessa dissertação, onde, respectivamente, as três primeiras Subsubseções mencionadas referem-se a dados com censura intervalar (caso II) e a última a dados com censura à direita.

3.2.1 Regressão isotônica

Para estimação não paramétrica de máxima verossimilhança, utilizou-se a regressão isotônica, explanada a seguir, como em Barlow et al. (1972).

Seja Y um conjunto finito $\{y_1, ..., y_k\}$ com a ordenação simples $y_1 \prec y_2 \prec ... \prec y_k$. Uma função $f : Y \to \mathrm{R}$ é isotônica se $y,\ x \in Y\ e\ y \prec x$ implicam $f(y) \leq f(x)$. Seja g uma dada função em Y e w uma dada função positiva em Y. Uma função isotônica g^* em Y é uma regressão isotônica de g com pesos w em relação à ordenação simples $y_1 \prec y_2 \prec ... \prec y_k$, se minimiza, na classe de funções isotônicas f em Y, a soma

$$\sum_{y\,\in\,Y} [g(y_i) - f(y_i)]^2 w(y_i).$$

Assumindo ainda a ordenação simples $y_1 \prec y_2 \prec ... \prec y_k$, considere as somas acumuladas

$$G_j = \sum_{i=1}^{j} w(y_i)g(y_i) \qquad e \qquad W_j = \sum_{i=1}^{j} w(y_i), \qquad j = 1, ..., k.$$

O gráfico dos pontos $P_j = (W_j, G_j),\ j = 1, ..., k,\ \ e\ \ P_0 = (0, 0)$ no plano cartesiano, constitui o diagrama de soma acumulada (DSA) da função dada g com pesos w. A regressão isotônica de g é dada pela inclinação da função minorante convexa máxima do diagrama de soma acumulada. A função minorante convexa máxima (MCM) é o gráfico do supremo de todas as funções convexas, cujos gráficos se encontram abaixo do diagrama de soma acumulada, conforme exemplificado na Figura 3.2.

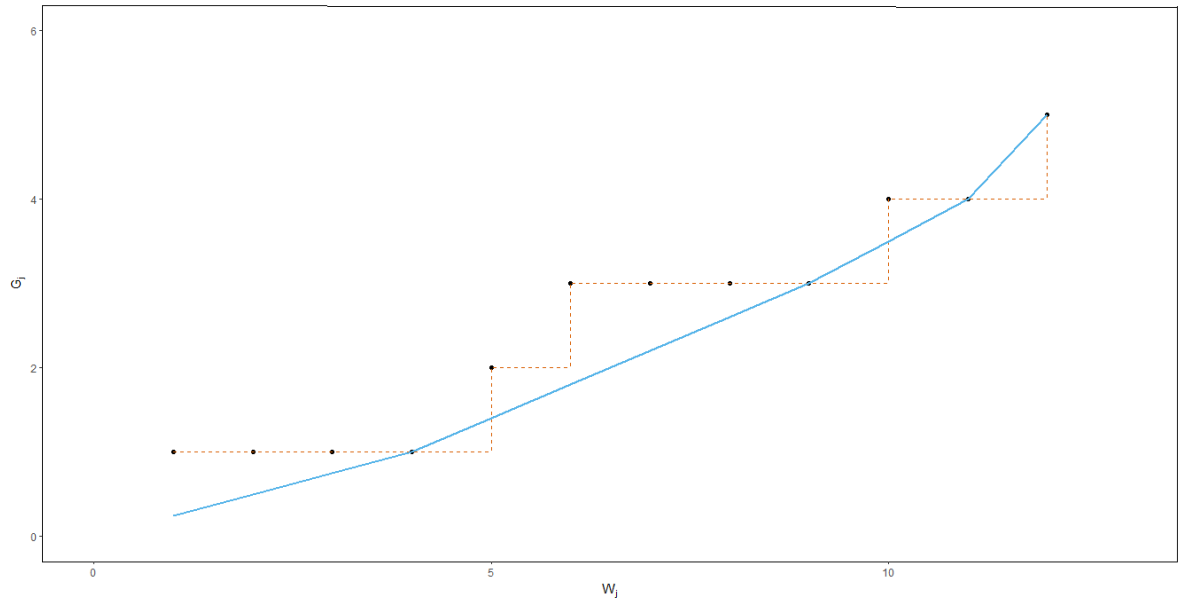

Figura 3.2: Exemplo de diagrama de soma acumulada e sua função minorante convexa máxima.

Em Groeneboom e Wellner (1992), a regressão isotônica g^* pode também ser obtida

por:

$$g^*(y_i) = \max_{r \leq i} \min_{k \geq i} \frac{\sum\limits_{r \leq j \leq k} g(y_i) w(y_i)}{\sum\limits_{r \leq j \leq k} w(y_i)}.$$

O Teorema 1.10, em Barlow et al. (1972) afirma que a regressão isotônica em g^* maximiza

$$\sum_y \{\Phi[f(y)] + [g(y) - f(y)]\varphi[f(y)]\}\, w(y) \tag{3.3}$$

na classe de funções isotônicas f, se Φ é estritamente convexa e $\varphi(x) = \Phi'(x) = \frac{d\Phi(x)}{dx}$.

3.2.1.1 Estimador não paramétrico de máxima verossimilhança (ENPMV) - censura intervalar

Caso I. Para estimação não paramétrica de máxima verossimilhança, utiliza-se a regressão isotônica, discorrida na subseção anterior.

No caso de estado corrente, a função minorante máxima convexa é definida como a função de $H^* : [0, n] \to \mathrm{R}$ tal que

$$H^* : [0, \sum_{i=1}^{j} w(y_i)] \to \mathrm{R},$$

dado que

$$H^*(t) = \sup \left\{ H(t) : H(i) \leq \sum_{j \leq i} \delta_j, \text{ para cada } i,\ 0 \leq i \leq n,\ H(0) = 0,\ e \text{ H convexa} \right\}.$$

Considere que se $\delta_i = 0,\ 1 \leq i \leq k_1$ e $\delta_i = 1,\ k_2 \leq i \leq n$, para quaisquer $0 < k_1 < k_2 < n$, então $g^*(y_i) = 0,\ 1 \leq i \leq k_1$ e $g^*(y_i) = 1,\ k_2 \leq i \leq n$. O valor de g^* é dada pela derivada à esquerda de H^* do DSA, em i para $i = 1, ..., n$, no ponto $\sum_{j \leq i} w(y_i)$ (Groeneboom e Wellner, 1992).

A determinação do estimador não paramétrico de máxima verossimilhança de F, no caso I, dá-se através da obtenção de uma $\hat{F}$ que maximize a equação (3.1), tal que $0 \leq \hat{F}(c_1) \leq ... \leq \hat{F}(c_n) \leq 1$. Fazendo $f = F,\ y_i = c_i\ (i = 1, ..., n),\ \Phi[F(c_i)] = F(c_i)\log F(c_i) + [1 - F(c_i)]\log[1 - F(c_i)],\ \varphi[F(c_i)] = \Phi'(t) = \log F(c_i) - \log[1 - F(c_i)]$ e

substituindo na equação (3.3), tem-se:

$$\begin{aligned}
&\sum_{i=1}^{n}\{\Phi[F(c_i)] + [g(c_i) - F(c_i)]\varphi[F(c_i)]\}w(c_i) \\
&= \sum_{i=1}^{n}\{F(c_i)\log F(c_i) + [1 - F(c_i)]\log[1 - F(c_i)] + [\delta_i - F(c_i)][\log F(c_i) - \log(1 - F(c_i))]\} \\
&= \sum_{i=1}^{n}\{\delta_i \log F(c_i) + (1 - \delta_i)\log[1 - F(c_i)]\} \qquad (3.4) \\
&= \mathcal{L}(F).
\end{aligned}$$

Portanto, o ENPMV de F é dado pela regressão isotônica g^* de $g(c_i) = \delta_i,\ w_i(c_i) = 1,$ e pode ser encontrada graficamente plotando os pontos:

$$\left(\sum_{j=1}^{i} w(c_i), \sum_{j=1}^{i} w(c_i)g(c_i)\right) = \left(i, \sum_{j \le i} \delta_i\right), \quad i = 1, ..., n,$$

no plano cartesiano e computando a função minorante convexa em $[0, n]$, pois $\hat{F}(c_i)$ é dada pela derivada à esquerda no ponto i do diagrama de soma acumulada, que é formado por esses pontos (Groeneboom e Wellner, 1992).

Caso II. Para obter o estimador não paramétrico de máxima verossimilhança no caso geral de censura intervalar, utiliza-se os resultados de regressão isotônica, porém de forma iterativa.

Sejam $Q_j,\ j = 1, ..., m,\ m = n + \sum_{i=1}^{n} \gamma_i,$ os valores ordenados do conjunto

$$\begin{aligned}
J &= J^{(1)} \bigcup J^{(2)}, \text{ em que} \qquad (3.5) \\
J^{(1)} &= \{t_{1i} : \delta_i = 1 \text{ ou } \gamma_i = 1, i = 1, ..., n\} \text{ e} \\
J^{(2)} &= \{t_{2i} : \gamma_i = 1 \text{ ou } \delta_i = \gamma_i = 0, i = 1, ..., n\}.
\end{aligned}$$

Em Groeneboom e Wellner (1992), a Proposição 1.4 menciona que: considere Q_1 correspondente a uma observação $t_{1i},$ tal que $I\{Y_i \le t_{1i}\} = 1,$ e seja a maior estatística de ordem Q_m correspondente a uma observação $t_{2i},$ tal que $I\{Y_i \ge t_{2i}\} = 1.$ Então, $\hat{F}$ é o estimador não paramétrico de máxima verossimilhança de $F,$ se, e somente se, $\hat{F}$ é a derivada à esquerda da função minorante convexa máxima do diagrama de soma acumulada, que consiste nos pontos $P_j = (G_{\hat{F}}(Q_j), H_{\hat{F}}(Q_j)),\ j = 1, ..., m,\ e\ P_0 = (0, 0).$

Se, por exemplo, Q_1 corresponde a um intervalo $[t_{1i}, t_{2i}]$ com $\delta_i = 0$ com $\gamma_i = 1,$ então, o termo $n^{-1}\log\{F(t_{2i}) - F(t_{1i})\}$ na Equação (3.2) é maximizado fazendo $\hat{F}(t_{1i}) = 0.$

Similarmente, se, por exemplo, a maior estatística de ordem $t_{(m)} \in J$ corresponderia a um tempo de observação t_{2i} tal que $I\{t_{1i} < Y_i \leq t_{2i}\} = 1$, assim a $\hat{F}$ que maximiza Equação (3.2) deveria satisfazer $\hat{F}(t_{2i}) = 1$.

Se existe $K_1 \geq 1$ tal que $Q_1, \ldots, Q_{K_1}$ correspondem a valores de t_{1i} com $\gamma_i = 1$, então $\hat{F}(Q_j) = 0,\ j = 1, \ldots, K_1$. Se existe $K_2 \leq m$ tal que $Q_{K_2}, \ldots, Q_m$ correspondem a valores de t_{2i} com $\gamma_i = 1$ ou de t_{1i} com $\gamma_i = 1$, logo $\hat{F}(Q_j) = 1,\ j = K_2, \ldots, m$. Os valores de $\hat{F}(Q_j),\ K_1 < j < K_2$, são obtidos utilizando o algoritmo iterativo da função minorante convexa descrita a seguir (Groeneboom e Wellner, 1992).

Considere as funções $G_F(y)$ e $H_F(y)$ dada por:

$$G_F(y) = \frac{1}{n}\left\{\sum_{i:t_{1i}\leq y} \frac{\delta_i}{[F(t_{1i})]^2} + \sum_{i:t_{1i}\leq y} \frac{\gamma_i}{[F(t_{2i}) - F(t_{1i})]^2} + \sum_{i:t_{2i}\leq y} \frac{\gamma_i}{[F(t_{2i}) - F(t_{1i})]^2} + \sum_{i:t_{2i}\leq y} \frac{1-\delta_i-\gamma_i}{[1-F(t_{2i})]^2}\right\}, \tag{3.6}$$

e

$$H_F(y) = W_F(y) + \sum_{j:Q_j\leq y} F(Q_j)[G_F(Q_j) - G_F(Q_{j-1})], \tag{3.7}$$

sendo

$$W_F(y) = \frac{1}{n}\left\{\sum_{i:t_{1i}\leq y} \frac{\delta_i}{F(t_{1i})} + \sum_{i:t_{1i}\leq y} \frac{\gamma_i}{F(t_{2i}) - F(t_{1i})} + \sum_{i:t_{2i}\leq y} \frac{\gamma_i}{F(t_{2i}) - F(t_{1i})} + \sum_{i:t_{2i}\leq y} \frac{1-\delta_i-\gamma_i}{1-F(t_{2i})}\right\}. \tag{3.8}$$

Sendo assim, os passos do algoritmo iterativo são:

(i) Tome $F^{(0)}(Q_j) = \frac{j}{m},\ K_1 < j < K_2$.

(ii) Construa o DSA com os pontos $P_0 = (0,0)$ e $P_j = (G_{\hat{F}^{(k)}}(Q_j),\ H_{\hat{F}^{(k)}}(Q_j)),\ j = 1, ..., m$ e obtenha $\hat{F}^{(k+1)}(Q_j)$ como sendo a derivada à esquerda no ponto $G_{\hat{F}^{(k)}}(Q_j)$ da função minorante convexa máxima do diagrama de soma acumulada.

(iii) Critério de parada: $||\hat{F}^{(k+1)} - \hat{F}^{(k)}|| < \epsilon$, para alguma norma $||\cdot||$.

3.2.2 Núcleo estimador

O núcleo estimador (*kernel estimator*) é um método estatístico não paramétrico baseado na estimação da função de kernel, para interpolação de dados. As estimativas de densi-

dade são aproximações locais da função alvo, que utilizam somente informações próximas ao ponto a ser interpolado, sendo assim, cada observação é ponderada localmente. Isto é, os estimadores de kernel são médias móveis ponderadas de uma função alvo, em que o peso é determinado por meio de uma função kernel com uma suavização de kernel (*bandwidth*). A suavização de kernel é local e se estende apenas para a observação do vizinho mais próximo, fornecendo a cada valor uma estimativa (Ali, 1998).

3.2.2.1 Dados não censurados

A função kernel k é descrita em Hollander et al. (2013) como uma função tal que

$$\begin{aligned} \mathrm{k}(y) \geq 0, \quad -\infty < y < \infty, \quad \mathrm{k}(-y) = \mathrm{k}(y), \\ \int_{-\infty}^{\infty} \mathrm{k}(y)dy = 1 \\ \text{e} \\ \int_{-\infty}^{\infty} y\mathrm{k}dy = 0, \ \text{ para } \ \forall y \in \mathrm{R}. \end{aligned} \tag{3.9}$$

Então, k é uma função não negativa simétrica em torno de zero, que integra um, que apresenta propriedades de uma função densidade de uma variável aleatória. Várias funções kernel foram propostas para uso com estimativa de densidade, uma simples é o kernel uniforme ou *box* kernel. O kernel uniforme é definido como

$$\mathrm{k}(y) = \begin{cases} 1, & \text{se } -1/2 \leq y < 1/2 \\ 0, & \text{se caso contrário} \end{cases}$$

e cumpre as três restrições explicitadas acima. Outra função do kernel comum é o kernel normal, $f(y) = (2\pi\sigma^2)^{-1/2}e^{-(y-\mu)^2/(2\sigma^2)}$, com μ definido como zero para garantir simetria e σ uma constante. Se k é uma função do kernel, então também é a versão em escala

$$\frac{1}{h}\mathrm{k}\left(\frac{y}{h}\right).$$

para $h > 0$. Esse kernel em escala pode ser centrado em qualquer ponto Y_i dos dados e a simetria não é em torno de zero, como acima, mas em torno de Y_i

$$\frac{1}{h}\mathrm{k}\left(\frac{y - Y_i}{h}\right).$$

Considere os dados, uma amostra $Y_1, ...Y_n$ de observações i.i.d de uma distribuição univariada contínua com função densidade de probabilidade f. A estimativa da densidade f do núcleo no ponto y é

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{k}\left(\frac{y - Y_i}{h}\right), \tag{3.10}$$

na qual h é a *bandwidth*, também chamado de janela ou parâmetro de suavização e k(·) é uma função de densidade de probabilidade qualquer, geralmente simétrica, denominada núcleo. A janela de h controla o grau de suavização. Logo, $\hat{f}$ é uma função densidade de probabilidade, se k é não negativa e se satisfaz a condição (3.9). Além disso, $\hat{f}$ herda todas as propriedades de diferenciabilidade e continuidade do Kernel k (Silverman, 1986).

3.2.2.2 Dados censurados

Suponha que $\hat{S}$ seja o ENPMV da função de sobrevivência, então, em Pan (2000) para caso de dados censurados, a estimativa do kernel de sua função de densidade é

$$\hat{f}(y) = -\frac{1}{h}\int_{-\infty}^{\infty} \mathrm{k}\left(\frac{y - Y}{h}\right) d\hat{S}(y), \tag{3.11}$$

em que k(·) é a função núcleo e h é o parâmetro de suavização. (Silva, 2011).

O estimador da densidade em (3.11), para dados com censura intervalar, pode ser reescrito como

$$\widetilde{f}_y(y) = \frac{1}{h}\sum_{j=1}^{n} s_j \mathrm{k}\left(\frac{y - Q_j}{h}\right), \tag{3.12}$$

com $s_j = \hat{S}(Q_j) - \hat{S}(Q_{j+1}) = \hat{F}(Q_{j+1}) - \hat{F}(Q_j)$, os Q_j são os valores ordenados do conjunto J na equação (3.5) e, na Subseção 3.2.1.1, mostra-se como $\hat{F}$ é computado. Maiores informações em Pan (2000) e Silva (2011).

A versão suavizada do estimador não paramétrico de $F(y)$ é dada por:

$$\widetilde{F}_Y(y) = \sum_{j=1}^{n} \mathrm{K}\left(\frac{y - Q_j}{h_y}\right)\left[\hat{F}(Q_j) - \hat{F}(Q_{j-1})\right],$$

sendo $\hat{F}(Q_j)$ $j = 1, \ldots, n$, obtido via algoritmo iterativo da função minorante convexa (Groeneboom e Wellner, 1992), e

$$\mathrm{K}\left(\frac{y - Q_j}{h_y}\right)\left[\hat{F}(Q_j) - \hat{F}(Q_{j-1})\right] = \int_{-\infty}^{y} \frac{1}{h_y}\mathrm{k}\left(\frac{t - Q_j}{h_y}\right)\left[\hat{F}(Q_j) - \hat{F}(Q_{j-1})\right] dt.$$

3.2.3 Suavização da estimativa da distribuição condicional de Y dado valor das covariáveis contínuas

Uma versão suavizada da distribuição condicional de Y dado o valor de x de uma covariável X é dado por:

$$\widetilde{F}_{Y|X}(y|x) = \sum_{j=1}^{n} \mathrm{K}\left(\frac{y - Q_j}{h_y}\right)\left[\hat{F}(Q_j|x) - \hat{F}(Q_{j-1}|x)\right],$$

sendo $\hat{F}(Q_j|x)$ obtida via algoritmo iterativo da função minorante convexa substituindo n^{-1} por $\mathrm{k}\left(\frac{x-x_i}{h_x}\right)$, i.e., fazendo

$$G_{\hat{F}}(y|x) = \sum_{i=1}^{n} \mathrm{k}\left(\frac{x - x_i}{h_x}\right)\left\{\sum_{i:T_{1i}\leq y} \frac{\delta_i}{[\hat{F}(T_{1i}|x)]^2} + \sum_{i:T_{1i}\leq y} \frac{\gamma_i}{[\hat{F}(T_{2i}|x) - \hat{F}(T_{1i}|x)]^2}\right.$$
$$\left. + \sum_{i:T_{2i}\leq y} \frac{\gamma_i}{[\hat{F}(T_{2i}|x) - \hat{F}(T_{1i}|x)]^2} + \sum_{i:T_{2i}\leq y} \frac{1 - \delta_i - \gamma_i}{[1 - \hat{F}(T_{2i}|x)]^2}\right\} \tag{3.13}$$

e

$$H_{\hat{F}}(y|x) = W_{\hat{F}}(y|x) + \sum_{j:Q_j\leq y} \hat{F}(Q_j|x)[G_{\hat{F}}(Q_j|x) - G_{\hat{F}}(Q_{j-1}|x)],$$

sendo

$$W_{\hat{F}}(y|x) = \sum_{i=1}^{n} \mathrm{k}\left(\frac{x - x_i}{h_x}\right)\left\{\sum_{i:T_{1i}\leq y} \frac{\delta_i}{\hat{F}(T_{1i}|x)} + \sum_{i:T_{1i}\leq y} \frac{\gamma_i}{\hat{F}(T_{2i}|x) - \hat{F}(T_{1i}|x)}\right.$$
$$\left. + \sum_{i:T_{2i}\leq y} \frac{\gamma_i}{\hat{F}(T_{2i}|x) - \hat{F}(T_{1i}|x)} + \sum_{i:T_{2i}\leq y} \frac{1 - \delta_i - \gamma_i}{1 - \hat{F}(T_{2i}|x)}\right\}. \tag{3.14}$$

Há a possibilidade de generalização, pois se houver um vetor multidimensional, mais de uma covariável contínua, pode-se substituir por um Kernel multidimensional, $\mathrm{k}\|\frac{x-x_j}{h_x}\|$.

Na sequência, será abordado o método recursivo de Portnoy (2003), aplicado a dados com censura à direita., que trata da teoria de redistribuição da massa de probabilidade para estimação da função de distribuição acumulada de Y_i, sugerida por Efron (1967).

3.2.4 Método de Portnoy

Para os dados na presença de censura, a premissa da generalização da RQ está no fato dos subgradientes de $\hat{\beta}(\tau)$ (equação 2.8), ou seja, das derivadas parciais direcionais com relação $b(\tau)$, somente depende do valor observado de Y_i através da função indicadora $I(y_i - x_i'b(\tau) \leq 0)$ (Koenker, 2005; Rasteiro, 2017).

No caso de observação de tempo de falha ou censura à direita, Rasteiro (2017) mostra que, para estimar o vetor $\beta(\tau)$, não há necessidade de se saber o valor exato que Y_i assume, e sim somente quando seu valor é menor ou maior que $\boldsymbol{x}_i'\mathbf{b}(\tau)$. Relembre que a variável aleatória observada, por estar sujeita a censura à direita, é definida como $\widetilde{Y} = \min(Y_i, C_i)$. Na ocorrência de censura com $C_i \geq \boldsymbol{x}_i'\mathbf{b}(\tau)$, considere que $F_{Y_i|\mathbf{x}_i}(y_i)$, seja conhecida. Veja que, quando há censura de Y_i (i.e. $Y_i > C_i$), alguns cenários podem ocorrer no estudo da função $I(y_i - \boldsymbol{x}_i'\mathbf{b}(\tau))$, conforme a seguir:

1. $C_i > \boldsymbol{x}_i'\mathbf{b}(\tau)$: nesse caso, $I(y_i - \boldsymbol{x}_i'\mathbf{b}(\tau) \leq 0) = 0$, mesmo não observando a variável Y_i.

2. $C_i < \boldsymbol{x}_i'\mathbf{b}(\tau)$: ao contrário do caso anterior, não é possível saber qual o valor a função indicadora assume, somente com a observação da variável C_i , pois há outras duas situações em que o valor da função está associado, com as seguintes probabilidades de ocorrência:

    a. $Y_i \in (C_i, \mathbf{x}_i'\mathbf{b}(\tau)]$: nesse caso, $I(y_i - \boldsymbol{x}_i'\mathbf{b}(\tau) \leq 0) = 1$ e sua probabilidade de ocorrência é dada por:

    $$\widetilde{w}_i(\tau) = P\{Y_i \in (C_i, \boldsymbol{x}_i'\boldsymbol{b}(\tau)] | Y_i > C_i\} = \frac{P[C_i < Y_i \leq \boldsymbol{x}_i'\boldsymbol{b}(\tau)]}{P[Y_i > C_i]} = \frac{\tau - F_{Y_i|\mathbf{x}_i}(C_i)}{1 - F_{Y_i|\mathbf{x}_i}(C_i)}.$$

    b. $Y_i \in (\mathbf{x}_i'\mathbf{b}(\tau), \infty)$: já aqui é complementar ao caso anterior, dado que $Y_i > C_i$. Logo, a probabilidade observada nessa situação é de $1 - \widetilde{w}_i(\tau)$, com $I(y_i - \boldsymbol{x}_i'\mathbf{b}(\tau) \leq 0) = 0$.

Na prática, não se conhece $F_{Y_i|\mathbf{x}_i}(y_i)$, sendo assim, deve-se estimá-la. Por conseguinte, suponha uma amostra de n observações da variável aleatória "tempo até a ocorrência da falha" (especificada), sujeita à censura aleatória à direita, ou seja, suponha as variáveis $(\widetilde{y}_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, com $\widetilde{y}_i = \min(y_i, c_i)$ e $\delta_i = I(y_i < c_i)$. Para as observações censuradas, $\delta_i = 0$, a estimativa de $\widetilde{w}_i(\tau)$ é definida por:

$$w_i = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i},$$

quando $\hat{\tau}_i$ é uma estimativa para $F_{Y_i|\mathbf{x}_i}(C_i)$.

Portnoy (2003), instigando a proposta de Efron (1967) sobre a teoria de redistribuição da massa de probabilidade, sugeriu uma nova abordagem para estimação de $F_{Y_i|\mathbf{x}_i}(C_i)$

com a introdução de pesos à função de perda (2.4), de maneira a incorporar, na estimação dos parâmetros $\beta(\tau)$, a informação de censura aleatória (Rasteiro, 2017). O esquema de ponderação via Kaplan-Meier discutido, a seguir, é importante para a compreensão do método e também para estimar os parâmetros da função quantílica para dados com censura à direita, discorrido em Rasteiro (2017).

3.2.4.1 Ponderação via Kaplan-Meier

Considere o estimador não-paramétrico de Kaplan-Meier (KM) para a estimar a função de sobrevivência de uma variável aleatória, que pode estar sujeita a censura à direita. Esse estimador é uma adaptação da função de sobrevivência empírica, que é escrita como (Giolo e Colosimo, 2006; Rasteiro, 2017):

$$\hat{S}(y) = 1 - \hat{F}(y) = \frac{\text{número de falha até o tempo y}}{\text{número total de observações na amostra}},$$

$\hat{S}(y)$ é uma função escada com saltos nos instantes de tempo k, nos quais, de fato, são observados o evento de interesse.

A fórmula clássica do estimador pode ser definida após estas premissas:

- Sejam $y_1 < \ldots < y_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em y_j, $j = 1, \ldots, k$;
- n_j o número de indivíduos sob risco em y_j.

Então, o estimador de Kaplan-Meier é expresso como:

$$\hat{S}(y) = \prod_{j:y_j<y} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:y_j<y} \left(1 - \frac{d_j}{n_j} \right)$$

Maiores informações podem ser obtidas em Giolo e Colosimo (2006).

Exemplo 3.1. Para ilustrar o estimador supracitado, considere uma amostra de 10 observações das variáveis aleatórias independentes, quando os tempos y_2, y_7 e y_8 são censurados à direita, exemplo proposto inspirado em Rasteiro (2017).

Tabela 3.1 - Tempos de sobrevivência de 10 observações, com censura dos tempos ordenados y_2, y_7 e y_8 e estimativas da função de sobrevivência e função distribuição acumulada, $S(y_i)$ e $F(y_i)$, respectivamente.

i	y_i	**Risco**	**Evento**	$\hat{S}(y_i)$	$\hat{F}(y_i)$
1	23	10	1	0,900	0,100
3	68	8	1	0,787	0,213
4	70	7	1	0,675	0,325
5	71	6	1	0,562	0,438
6	100	3	1	0,450	0,550
9	181	2	1	0,225	0,775
10	199	1	1	0,000	1,000

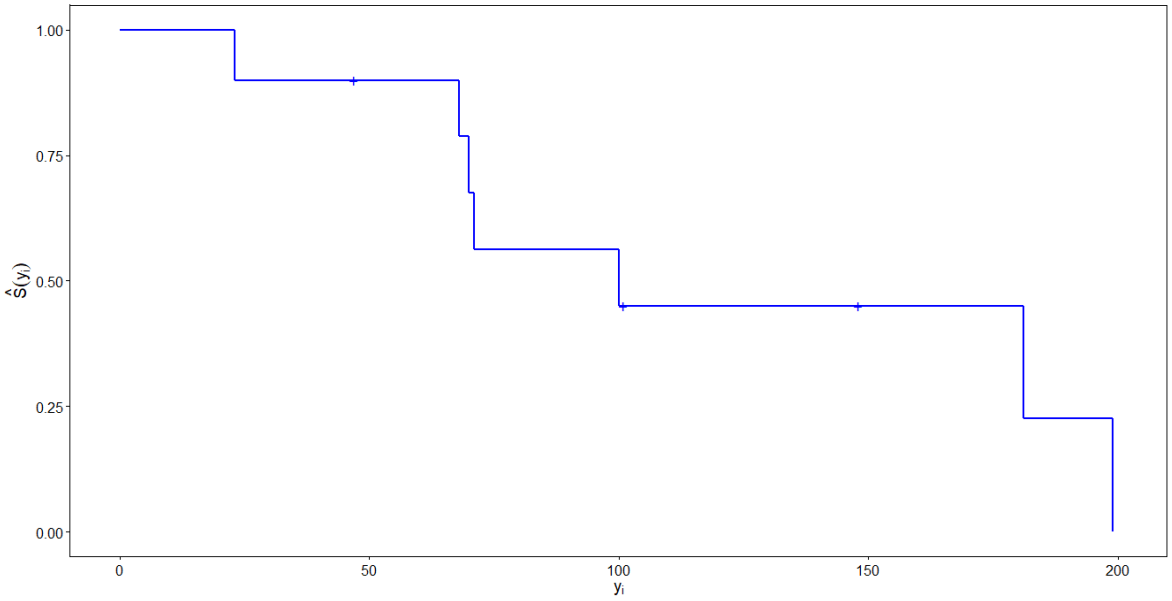

Figura 3.3: Gráfico da função de sobrevivência estimada para os dados do exemplo proposto.

Conforme em Rasteiro (2017), o estimador de KM pode também ser computado de outra maneira, menos comum, mas que o considera como resultado de um esquema de ponderação, que atribui pesos a cada uma das observações. No exemplo, perceba que como a primeira observação não é censurada, logo a função de sobrevivência empírica tem salto de dimensão $1/10$. Ou melhor, isso corresponde a dizer que cada observação tem peso igual a 1. Por isso, a distribuição acumulada no ponto i dá-se pela razão entre os pesos anteriores a observação y_i e o tamanho n da amostra.

Na Figura 3.3, no entanto, a massa de probabilidade não é atribuída, pelo estimador de

KM, a observações censuradas. Por conseguinte, a distribuição acumulada na observação y_2 é igual a $0,100$, o que corresponde à função de sobrevivência estimada em y_2 igual a $0,900$. O peso de y_2 é essencial para computar a função de distribuição acumulada nos instantes de falha posteriores, o qual não pode ser igual a 1, porque isso seria o mesmo que dizer que a observação é não censurada. Portanto, o intuito é repartir o peso da observação y_2 considerando as duas possíveis condições: o verdadeiro valor do tempo de falha ocorre entre $i = 2$ e $i = 3$, ou o verdadeiro tempo é posterior a 3. Essas condições têm probabilidade $(\tau_i^* - 0,100)/(1 - 0,100)$ e $(1 - \tau_i^*)/(1 - 0,100)$, respectivamente, sendo τ_i^*, estimativa de τ_i, a probabilidade acumulada no tempo de falha y_3 e y_4, segundo apresentando nessa seção. Então, em conformidade com o que foi exposto, τ_i^* deve satisfazer:

$$\tau_3^* = \frac{1 + (\tau_3^* - 0,100)/(1 - 0,100) + 1}{10},$$

em que 1 é o peso da observação y_1, que é não censurada, $(\tau_3^* - 0,100)/(1 - 0,100)$ é o peso de y_2 e 1, de y_3 em τ_3^*. Calculando a equação, encontra-se $\tau_3^* = 0,213$, que coincide com o valor de $\hat{F}(y_3)$ da Tabela 3.1.

Então, para as observações y_4, y_5 e y_6, que também não são censuradas:

$$\tau_4^* = \frac{1 + (\tau_4^* - 0,100)/(1 - 0,100) + 1 + 1}{10},$$
$$\tau_5^* = \frac{1 + (\tau_5^* - 0,100)/(1 - 0,100) + 1 + 1 + 1}{10}$$

e

$$\tau_6^* = \frac{1 + (\tau_6^* - 0,100)/(1 - 0,100) + 1 + 1 + 1 + 1}{10},$$

computando as equações, tem-se que $\tau_4^* = 0,325$ para y_4, $\tau_5^* = 0,438$ para y_5 e $\tau_6^* = 0,550$ para y_6. Em contrapartida, y_7 e y_8 são censurados, portanto, a função distribuição acumulada nas observações y_7 e y_8 é igual a $0,550$, pois não há massa de probabilidade associada. Para a observação em y_9,

$$\tau_9^* = \frac{1 + (\tau_9^* - 0,100)/(1 - 0,100) + 4 + 2(\tau_9^* - 0,550)/(1 - 0,550) + 1}{10},$$

e calculando tem-se que $\tau_9^* = 0,775$. Do mesmo modo, a função distribuição acumulada no tempo de falha y_{10} é igual a 1. Quando se introduz uma nova observação ao conjunto de dados, por exemplo entre $i = 1$ e $i = 3$, independentemente da escolha de y_{11}, a

estimativa de τ_i^* não é alterada, por exemplo, $y_{11} = +\infty$, com peso $(1 - 0,213)/(1 - 0,100) = 0,874$. Assim, introduz-se uma nova observação fictícia no conjunto de dados para cada observação censurada, que foi ponderada, optando-se por qualquer valor além do escopo dos dados (Rasteiro, 2017).

Assim, para o cálculo da função de distribuição acumulada, esse método de atribuição de pesos às observações remete ao conceito de redistribuição da massa de probabilidades proposto por Efron (1967), que resulta em 1 menos a estimativa de KM, e é uma maneira alternativa de computar suas estimativas (Rasteiro, 2017).

3.2.4.2 Algoritmo recursivo de Portnoy

Portnoy (2003) propõe a utilização do esquema de ponderação de KM, com algumas alterações, que visa computar os pesos, $w_i(\tau)$, das observações censuradas, mas com fixação prévia dos valores de τ_j^*. Também deve calcular a estimativa de $F_{Y_i|\mathbf{x}_i}(C_i)$, $\hat{\tau}_i$. Então, os pesos w_i podem ser definidos como

$$w_i \equiv w_i(\tau) = \begin{cases} 1, & \text{se } \delta_i = 1 \text{ ou } \tau_j^* \leq \hat{\tau}_i \\ \frac{\tau_j^* - \hat{\tau}_i}{(1-\hat{\tau}_i)}, & \text{se } \delta_i = 0 \text{ e } \tau_j^* > \hat{\tau}_i \end{cases},$$

com $\tau_j^* : j = 1, \ldots, M$.

Os passos do algoritmo de Portnoy (2003) compreendem (Rasteiro, 2017):

1. Estima-se $b(\tau_1^*)$, dado o valor τ_1^*, aplicando a regressão quantílica linear usual a todos os dados, incluindo as observações censuradas (x_i, C_i).

2. Para estimar $b(\tau_2^*)$, opera-se da mesma forma que no caso não censurado. Assim, verifica-se se há valores c_i, $i = 1, \ldots, n$, tais que $c_i \in \left[\mathbf{x}_i^T \mathbf{b}(\tau_1^*);\ \mathbf{x}_i^T \mathbf{b}(\tau_2^*)\right]$. Se houver esses valores, $\mathbf{b}(\tau_2^*)$ deverá ser reestimado, abrangendo as observações censuradas. Seja K o conjunto de índices de tais observações censuradas. Atribui-se $\hat{\tau}_i = \tau_1^*$ para as observações em K. Igualmente, $\mathbf{b}(\tau_2^*)$ é o vetor de parâmetros que minimiza a função:

$$\sum_{i \notin K} \rho_{\tau_2^*} \left[\widetilde{y}_i - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_2^*)\right] + \sum_{i \in K} \left\{ w_i(\tau_2^*) \rho_{\tau_2^*} \left[c_i - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_2^*)\right] + [1 - w_i(\tau_2^*)] \rho_{\tau_2^*} \left[y^{+\infty} - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_2^*)\right] \right\},$$

em que $y^{+\infty}$ é um valor suficientemente grande incluído ao conjunto de dados, além do escopo dos valores observados de $\widetilde{y}_i$.

3. Dado que já se tenha computado $\mathbf{b}(\tau_j^*)$ e defina K o índices das observações censuradas que colaboram para o seu cálculo. O passo seguinte é a estimação $\mathbf{b}(\tau_{j+1}^*)$, admitindo que não há outras censuras além das consideradas em K. Entretanto, tome que após a estimação do parâmetro $\mathbf{b}(\tau_{j+1}^*)$, há uma observação censurada c_k tal que

$$c_k \in \left[\boldsymbol{x}_k^T\boldsymbol{b}(\tau_j^*); \boldsymbol{x}_k^T\boldsymbol{b}(\tau_{j+1}^*)\right],$$

ou seja, o vetor $\mathbf{b}(\tau_{j+1}^*)$ necessita ser recalculado, tomando-se também a informação dessa censura. De outra maneira, o índice k deve ser incorporado ao conjunto K. Atribui-se, logo, que $\hat{\tau}_k = \tau_j^*$ para essa observação que precisa ser ponderada com peso $w_k(\tau_j^*)$. Além disso, é necessário introduzir uma observação em $+\infty$ com peso $1 - w_k(\tau_j^*)$, de modo a cobrir todas as possibilidades para o verdadeiro tempo de sobrevivência da k-ésima observação. Então, de um modo geral, $\mathbf{b}(\tau_{j+1}^*)$ é o vetor de parâmetros que minimiza a função:

$$\sum_{i\notin K} \rho_{\tau_{j+1}^*}[\widetilde{y}_i - \boldsymbol{x}_i^T\boldsymbol{b}(\tau_{j+1}^*)] + \sum_{i\in K}\Big\{w_i(\tau_{j+1}^*)\rho_{\tau_{j+1}^*}\left[c_i - \boldsymbol{x}_i^T\boldsymbol{b}(\tau_{j+1}^*)\right] + [1 - w_i(\tau_{j+1}^*)]\rho_{\tau_{j+1}^*}\left[y^{+\infty} - \boldsymbol{x}_i^T\boldsymbol{b}(\tau_{j+1}^*)\right]\Big\},$$

no qual $y^{+\infty}$ é um tempo de falha suficientemente grande introduzido ao conjunto de dados.

4. O passo 3 deve ser repetido até que τ_{j+1}^* seja igual a 1 ou se restarem apenas observações censuradas à direita de $\mathbf{x}_i^T\mathbf{b}(\tau_{j+1}^*)$.

Segundo Rasteiro (2017), o algoritmo de Portnoy (2003) implementado nos principais *softwares*, como no R, é eficiente e exibe rapidamente os resultados. Entretanto, a principal desvantagem do método é o pressuposto de que todos os quantis se relacionam linearmente com as covariáveis, pois, na prática, isso nem sempre se verifica. Por exemplo, em geral, os primeiros quantis não seguem uma relação linear, por apresentarem poucas observações.

Para exemplificar a aplicação do método recursivo de Portnoy (2003), usou-se um conjunto de dados cedido por um instituto em Brasília, apenas para demonstrar o uso da

função "*crq*" do pacote *quantreg* do *software* R. Trata-se de um estudo sobre o tempo de execução de um processo (em minutos) com relação a cinco localidades, onde foram avaliados 1.522 processos; 35% desses dados são censurados à direita. Utilizou-se, para ajuste do modelo de regressão quantílica, o quantil de ordem $\tau = 0,5$. O interesse foi avaliar o tempo de execução mediano de um processo.

Tabela 3.2 - Resultado do ajuste para o modelo de regressão quantílica.

Variável	Estimativa	Erro Padrão	Valor p	IC(95%)
Intercepto	57,74	0,68	0,00	[56,27 ; 58,93]
bd$locV1	-	-	-	-
bd$locV2	-29,99	4,12	0,00	[-40,49 ; -24,35]
bd$locV3	-18,54	5,37	0,00	[-30,91 ; -9,87]
bd$locV4	-7,81	1,75	0,00	[-10,42 ; -3,57]
bd$locV5	-6,69	3,97	0,09	[-14,68 ; 0,88]

Na tabela 3.2, verifica-se que somente a variável localidade ('loc") no grupo V5 ("locV5") não é significativa a nível de 5%. Um exemplo de interpretação seria que, ao mudar de localidade, o tempo mediano de execução do processo é 18,5 minutos menor para um processo que tem a localidade alterada de V1 para V3.

A seguir serão abordados alguns pressupostos e propriedades dos estimadores obtidos pelo método de Portnoy (2003), que são discutidas e retratadas em seu artigo.

Propriedades Assintóticas

Portnoy (2003) propõe as seguintes condições:

C1. Seja $\epsilon > 0$ tal que $\widetilde{\tau}_1 \geq \epsilon$, em que $\widetilde{\tau}_1$ é o único e menor valor de $\hat{\tau}_1$ definido em $x_i'\beta(\widetilde{\tau}_i) = C_i$ e $\delta_i = 0$.

C2. A função densidade de probabilidade Y (dado x_i) e suas derivadas satisfazem: $a \leq f_i(u) \leq b$, $|f_i'(u)| \leq c$, uniformemente para $\epsilon \leq F_i(u) \leq 1-\epsilon$ e uniformemente em $i = 1, \ldots, n$, quando $a > 0$, $b < +\infty$ e c são constantes que podem depender de ϵ.

C3. Seja a constante B tal que $\| x_i \| \leq B$ uniformemente em $i = 1, \ldots, n$. Portnoy (2003) menciona que deve ser possível permitir que o limite em x_i dependa de n

(contanto que n cresça lentamente). Entretanto, as propriedades assintóticas parecem complicadas, de modo que a simplificação oferecida pela presunção da limitação parece ser razoável, relaxando o pressuposto. Além disso, geralmente há poucas observações com x_i grande; e assim as estimativas dos quantis condicionais seriam ruins.

C4. Defina

$$Z_n(\tau) = \frac{1}{n} X' \left(\text{diag} \left\{ \widetilde{w}_i(\tau) f_i \left(x_i' \beta(\tau) \right) \right\} \right) X.$$

Então, seja uma matriz definida (não aleatória), $Z(\tau)$ e a constante c tal que, para n suficientemente grande,

$$\| Z_n(\tau) - Z(\tau) \| \leq c n^{-1/4}.$$

Segundo o Condição 1, defina $D_n \equiv \text{diag}(d_i)$, com

$$d_i \equiv \tau(1-\tau) - (1-\tau) \left\{ I(C_i \leq x_i' \beta(\tau) \left[\frac{\tau - P_{x_i} Y_i \leq C_i}{1 - P_{x_i} Y_i \leq C_i} \right] \right\}.$$

Assuma que

$$X' D_n X \to V(\tau), \text{ quando } n \to \infty,$$

em que $V(\tau)$ é a matriz de variância e covariância (2.17); então, com $Z(\tau)$ sob as Condições apresentadas, de $C1$ a $C4$, o estimador é não viesado e segue assintoticamente a distribuição normal multivariada:

$$\sqrt{n} \left(\hat{\boldsymbol{b}}(\tau, \hat{w}) - \boldsymbol{\beta}(\tau) \right) \xrightarrow{D} \mathcal{N} \left(\boldsymbol{Z}^{-1}(\tau) \boldsymbol{V}(\tau) \boldsymbol{Z}^{-1}(\tau) \right).$$

Na prática, bem como no contexto em dados sem censura, abordado em Rasteiro (2017), é inviável a estimação da função variância de $\mathbf{b}(\tau, w)$ de modo direto, já que requer a função de distribuição de Y_i, que é desconhecida. Salienta que uma solução viável para estimação da função variância seria a utilização de uma implementação via *bootstrap*, que há no *software* R. Assim, uma estatística de teste para a hipótese, como a distribuição assintótica do estimador de $\boldsymbol{\beta}(\tau)$ é normal: $H_0 : \beta_s(\tau) : 0, 1 \;\; s = 1, \ldots, p$, por exemplo, dá-se pela estatística t:

$$T = \frac{b_s(\tau) - 0}{\sqrt{\hat{Z}_e / n}} \sim t_{n-1},$$

em que, para a variância de $b_s(\tau)$, $\hat{Z}_e$ é a estimativa de *bootstrap*. Da mesma maneira, pode-se escrever o intervalo de confiança *bootstrap*-t como, vide Efron e Tibshirani (1994):

$$\underset{100(1-\alpha)\%}{\texttt{I.C.}\ [b_s(\tau)]} = \left[b_s(\tau) \mp t_{\alpha/2;n-1} \sqrt{\frac{\hat{Z}_e}{n}} \right], \text{ com nível de confiança } 1-\alpha.$$

Contudo, Efron e Tibshirani (1994) sugerem a utilização de um método híbrido de *bootstrap* para estabelecer os intervalos de confiança para o parâmetro $\beta(\tau)$. A proposta, nesse caso, baseia-se em definir via *bootstrap* as distâncias interquartílicas $b_s^*(\tau)_{0,75} - b_s^*(\tau)_{0,50}$ e $b_s^*(\tau)_{0,50} - b_s^*(\tau)_{0,25}$, em que o quantil de ordem k das estimativas de $b_s(\tau)$ determinadas via *bootstrap* é $b_s^*(\tau)_k$. Na sequência, multiplica-se essas medidas estimadas por $2,906$, cujo valor é usado para assegurar a consistência do estimador em Heritier et al. (2009), e se adiciona o valor $\beta^*(\tau)_{0,50}$. Portanto, os intervalos de confiança são obtidos com coeficiente de confiança de 95% para $b_s(\tau)$; desse modo, verifica-se que eles não são absolutamente simétricos (Rasteiro, 2017).

A vantagem, de uma forma geral, em realizar inferência de regressão quantílica com dados censurados via *bootstrap* é que metodologias baseadas em reamostragem das variáveis (Y_i, C_i, δ_i) podem ser explicadas pela teoria clássica de *bootstrap*, segundo autores. Sob outra perspectiva, métodos que incluem os resíduos do modelo, ou que tratam do contexto em que as variáveis (Y_i, C_i, δ_i) não são independentes e identicamente distribuídas para $\forall i = 1, \ldots, n$, ainda demandam estudos e novas teorias (Rasteiro, 2017). Maiores informações vide Portnoy (2003) e Rasteiro (2017).

A seguir será apresentado o algoritmo da metodologia proposta nessa dissertação, a qual é a aplicação do método de Portnoy, utilizado em Rasteiro (2017), adaptado ao algoritmo iterativo da função minorante convexa, este discorrido nessa Subsubseção 3.2.3.

3.2.5 Método proposto

Aqui são os passos do algoritmo proposto nessa dissertação. $\tilde{F}(\cdot \mid x_i)$ representa o estimador não paramétrico de máxima verossimilhança (ENPMV) condicional suavizado de $F(\cdot \mid x_i)$. Para o caso em que há censura intervalar:

1. Faça $i = 1$;
2. Enquanto $i \leq n$;

3. Se $\delta_i = \gamma_i = 0$, faça $i \leftarrow i + 1$; vá para (2);

4. Se $\delta_i = 1$, faça $\tilde{F}(L_i \mid x_i) = 0$ e calcule $\tilde{F}(R_i \mid x_i)$;

5. Se $\gamma_i = 1$, calcule $\tilde{F}(L_i \mid x_i)$ e $\tilde{F}(R_i \mid x_i)$;

6. Gere $z_i \sim U(\tilde{F}(L_i \mid x_i), \tilde{F}(R_i \mid x_i))$;

7. Faça $m_i = (L_i + R_i)/2$;

8. Calcule $\tilde{F}(m_i \mid x_i)$;

9. Se $\tilde{F}(m_i \mid x_i) > z_i$, faça $R_i = m_i$;

10. Se $\tilde{F}(m_i \mid x_i) \leq z_i$, faça $L_i = m_i$;

11. Se $R_i - L_i < \varepsilon$, faça $\tilde{y}_i = m_i$; caso contrário, vá para (7).

Já no caso em que, também, há censura à direita, aplicou-se o método de Portnoy (2003).

Na sequência, será apresentado o método de Zhou et al. (2017), o qual foi comparado ao método proposto.

3.2.6 Método de Zhou

No caso II de censura intervalar, será abordado a metodologia de Zhou et al. (2017), que propõem métodos de estimação do vetor de parâmetros $\beta(\tau)$ para o modelo de regressão quantílica censurado por intervalos com $\tau \in [0, 1]$. Num primeiro momento, os autores propõem um método de estimação, que é uma generalização direta da regressão quantílica para dados observados completos. O estimador proposto é definido como o ponto ótimo de solução de um problema de minimização com função objetivo convexa. A propriedade da normalidade assintótica é estabelecida com um viés convergindo para zero. Para reduzir o viés, dois métodos de correção de polarização são propostos, baseados em *bootstrap* e estimativa inicial respectivamente. Ademais, os métodos propostos não exigem que os vetores de censura sejam distribuídos de forma idêntica e podem ser aplicados facilmente a modelos com várias covariáveis, como covariáveis de desenho aleatório fixo, discreto aleatório ou contínuo.

Em casos de censura em intervalos, observa-se que $y_i - x_i'\beta(\tau) > 0$ é válido para $x_i'\beta(\tau) \leq t_{1i}$ e $y_i - x_i'\beta(\tau) < 0$ para $x_i'\beta(\tau) > t_{2i}$, assim, nesse caso, pode-se definir a função de perda de quantis $\zeta_i(\tau, \beta)$ como

$$\zeta_i(\tau,\beta) = \begin{cases} \tau|y_i - x_i'\beta(\tau)|, & \text{se } t_{1i} \geq x_i'\beta(\tau) \\ \psi_i(\tau,\beta), & \text{se } t_{1i} < x_i'\beta(\tau) \leq t_{2i} \\ (1-\tau)|y_i - x_i'\beta(\tau)|, & \text{se } t_{2i} < x_i'\beta(\tau) \end{cases}, \tag{3.15}$$

para alguns ψ_i. $\zeta_i(\tau, \beta)$ não pode ser usada diretamente para obter o estimador de $\beta_i(\tau)$ pois ψ_i não é definida e y_i é não observável. Como a contribuição de cada ponto para o estimador, depende apenas do sinal do resíduo, na equação (3.15), y_i pode ser substituído por t_{1i}, se $t_{1i} \geq x_i'\beta(\tau)$, e por t_{2i}, se $t_{2i} < x_i'\beta(\tau)$. Os autores salientam que não é fácil definir ψ_i, pois se $x_i'\beta(\tau) \in (t_{1i}, t_{2i})$ é válido, não temos ideia do sinal de $y_i - x_i'\beta(\tau)$, contudo observe que se o intervalo de censura $t_{2i} - t_{1i}$ for pequeno, y_i estará próximo de t_{1i} e t_{2i} e, concomitantemente, $P(x_i'\beta(\tau) \in (t_{1i}, t_{2i}])$ estará próximo de zero sob algumas condições de regularidade. Dessa forma, pode-se desconsiderar a contribuição desses pontos nesses casos. Por conseguinte, pode-se modificar $\zeta_i(\tau, \beta)$ para

$$\begin{aligned} \widetilde{\zeta}_i(\tau,\beta) &= \begin{cases} \tau|y_i - x_i'\beta(\tau)|, & \text{se } t_{1i} \geq x_i'\beta(\tau) \\ 0, & \text{se } t_{1i} < x_i'\beta(\tau) \leq t_{2i} \\ (1-\tau)|y_i - x_i'\beta(\tau)|, & \text{se } t_{2i} < x_i'\beta(\tau) \end{cases} \\ &= \begin{cases} \tau|t_{1i} - x_i'\beta(\tau)|, & \text{se } t_{1i} \geq x_i'\beta(\tau) \\ 0, & \text{se } t_{1i} < x_i'\beta(\tau) \leq t_{2i} \\ (1-\tau)|t_{2i} - x_i'\beta(\tau)|, & \text{se } t_{2i} < x_i'\beta(\tau) \end{cases}. \end{aligned} \tag{3.16}$$

$\widetilde{\zeta}_i(\tau, \beta)$ é reescrito como $\widetilde{\zeta}_i(\tau,\beta) = (1-\tau)|t_{2i} - \max(t_{2i}, x_i'\beta(\tau))| + \tau|t_{1i} - \min(t_{1i}, x_i'\beta(\tau))|$.

Defina o estimador do quantil de censura intervalar, $\hat{\beta}_n(\tau)$ do vetor de parâmetro de regressão como o ponto de solução ideal do problema de minimização:

$$\min_{\beta\in\Theta} \left\{ n^{-1} \sum_{i=1}^{n} \widetilde{\zeta}_i(\tau,\beta) \right\}, \tag{3.17}$$

isto é,

$$\hat{\beta}_n(\tau) = \arg\min_{\beta\in\Theta} \left\{ n^{-1} \sum_{i=1}^{n} \widetilde{\zeta}_i(\tau,\beta) \right\}. \tag{3.18}$$

Se $\{y_i\}$ for observado, ou seja, $t_{1i} = t_{2i}$ para cada i, o estimador $\hat{\beta}_n(\tau)$, definido na equação (3.18), será transformado em estimador quantílico $\widetilde{\beta}_n(\tau)$ para dados observados completos. Em vista disso, o método acima é uma generalização direta da estimação do quantil para modelos de regressão com dados observados completos. Note que $\widetilde{\zeta}_i(\tau, \beta) = (1-\tau)|\max(t_{2i}, x_i'\beta(\tau)) - t_{2i}| + \tau|t_{1i} - \min(t_{1i}, x_i'\beta(\tau))|$, visto que $\max(c_1 + c_2,\ c_3 + c_4) \leq \max(c_1, c_3) + \max(c_2, c_4)$ e $\min(c_1 + c_2,\ c_3 + c_4) \geq \min(c_1, c_3) + \min(c_2, c_4)$ para quaisquer números reais c_1, c_2, c_3 e c_4. Os autores mencionam que é fácil provar que a função objetivo da equação (3.17) é convexa; isso faz com que o estimador quantílico seja fácil de computar.

Seja $\|\cdot\|$ a norma L_2 do vetor correspondente. Para atribuir as propriedades assintóticas do estimador quantílico $\hat{\beta}_n(\tau)$ (Zhou et al., 2017), primeiro assume-se que as seguintes suposições estão satisfeitas:

A1. Para qualquer $\tau \in (0, 1)$, o verdadeiro vetor do parâmetro $\beta_0(\tau)$ é um ponto interior do espaço paramétrico Θ.

A2. Para qualquer $\tau \in (0, 1)$, a função de distribuição condicional de y_i dado x_i é contínua em $x_i'\beta_0(\tau)$.

A3. (t_{1i}, t_{2i}), $i = 1, \ldots, n$, são vetores aleatórios independentes (não é necessário que sejam identicamente distribuídos), que satisfaçam $\sup_i |t_{2i} - t_{1i}| \leq \varrho_n$ para alguma sequência de $\varrho_n \to 0$, quando $n \to \infty$. Entretanto, tanto $G_i^1(\cdot)$ como $G_i^2(\cdot)$, que são funções de distribuição marginal de t_{1i} e t_{2i} respectivamente, têm derivadas contínuas e limitadas próximas ao ponto $x_i'\beta_0(\tau)$.

A4. Para cada $\epsilon > 0$, existe um M finito tal que

$$E\left[\frac{1}{n}\sum_{i=1} n \parallel x_i \parallel^2 \mathcal{I}(\parallel x_i \parallel > M)\right] < \epsilon,$$

para todos n suficientemente grande.

A5. A sequência dos menores autovalores das matrizes:

$$K_n = E\left\{\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \left[(1-\tau)\left.\frac{\partial G_i^2(r)}{\partial r}\right|_{r=x_i'\beta_0(\tau)} + \tau \left.\frac{\partial G_i^1(l)}{\partial l}\right|_{l=x_i'\beta_0(\tau)}\right]\right\}$$

está limitado de zero para n suficientemente grande.

Sob essas premissas, os principais resultados de Zhou et al. (2017) são os seguintes teoremas:

T1 Para qualquer $\tau \in (0,1)$, sob as premissas $A1 - A5$, $\hat{\beta}_n(\tau) \xrightarrow{P} \beta_0(\tau)$ quando $n \to \infty$, em que "$\xrightarrow{P}$" significa convergência em probabilidade.

T2 Para qualquer $\tau \in (0,1)$, sob as premissas $A1 - A5$,

$$2\sqrt{n}\widetilde{K}_n^{-1/2} K_n \left(\hat{\beta}_n(\tau) - \beta_0(\tau) + K_n^{-1} L_n\right) \xrightarrow{D} \mathcal{N}\left(0, \boldsymbol{I}_m\right),$$

quando $n \to \infty$, em que $\boldsymbol{I}_m$ denota a matriz identidade de ordem m, "$\xrightarrow{D}$" significa convergência em distribuição e

$$\begin{aligned}
\widetilde{K}_n &= \frac{1}{n}\sum_{i=1}^{n} E\left\{x_i x_i' \left[\tau^2 P_{1i} + (1-\tau)^2 P_{2i} + 2\tau(1-\tau)P_i\right]\right\}, \\
P_{1i} &= P\left(x_i'\beta_0(\tau) \leq t_{1i} | x_i\right) P\left(x_i'\beta_0(\tau) > t_{1i} | x_i\right), \\
P_{2i} &= P\left(x_i'\beta_0(\tau) \leq t_{2i} | x_i\right) P\left(x_i'\beta_0(\tau) > t_{2i} | x_i\right), \\
P_i &= P\left(x_i'\beta_0(\tau) \leq t_{1i} | x_i\right) P\left(x_i'\beta_0(\tau) > t_{2i} | x_i\right), \\
L_n &= \frac{1}{n} E\left\{x_i \left[(1-\tau)\mathcal{I}\left(t_{2i} < x_i'\beta_0(\tau)\right) - \tau\mathcal{I}\left(x_i'\beta_0(\tau) \leq t_{1i}\right)\right]\right\}.
\end{aligned}$$

T3 Suponha $\varrho_n = O(n^{-1/2})$, logo sob as premissas $A1 - A5$,

$$2\sqrt{n}\widetilde{K}_n^{-1/2} K_n \left(\hat{\beta}_n(\tau) - \beta_0(\tau) + K_n^{-1} L_n\right) \xrightarrow{D} \mathcal{N}\left(0, \boldsymbol{I}_m\right),$$

para qualquer $\tau \in (0,1)$ quando $n \to \infty$.

Zhou et al. (2017) salientam que, para os vetores de censura (t_{1i}, t_{2i}), $i = 1, \ldots, n$, não se presume que sejam identicamente distribuídos e que os vetores de covariáveis x_i, $i = 1, \ldots, n$, podem ser variáveis aleatórias discretas fixas ou variáveis aleatórias contínuas. A suposição A4 impede que um único ponto de amostra com $\| x_i \|$ muito grande tenha um efeito dominante sobre o estimador $\hat{\beta}_n(\tau)$, essa suposição assegurará que n seja suficientemente grande, e $\widetilde{K}_n$ são matrizes definidas não negativas com autovalores limitados acima, positivamente.

Para modelos de regressão quantílica com dados observados, dados censurados à direita ou dados duplamente censurados, em Zhou et al. (2017), também se observa que as suposições sobre a função de distribuição condicional de y_i dado x_i (ou função de distribuição de erros de regressão) são, constantemente, necessárias para assegurar a unicidade

do estimador de parâmetros. No entanto, para dados com censura intervalar, uma vez que não há y_i sendo observado, as suposições sobre a função de distribuição condicional de y_i não são suficientes. As suposições A3 e A5, que são sobre as propriedades das funções de distribuição condicional das variáveis de censura observadas, são requisitos adicionais para garantir a unicidade do estimador de quantil para amostras grandes. Na prática, suposições similares a A5 são, constantemente, necessárias na regressão quantílica para censura à direita ou dados duplamente censurados, maiores detalhes vide Pollard (1990), Rao e Zhao (1993), Xiuqing e Jinde (2005) e Wang e Wang (2009). A suposição A5 é um requisito em que t_{1i} ou t_{2i} é próximo de $x_i'\beta(\tau)$ para pontos de amostra "suficientes" e que os regressores $x_i'\beta(\tau)$ são não colineares para essas observações.

Contudo, Zhou et al. (2017) mencionam uma limitação importante à estimação de quantil proposta, que não pode ser aplicada a dados com duração infinita de intervalo de censura, como dados censurados à direita ou censurados duplamente, o que se verifica constantemente em estudos empíricos, dados experimentais.

3.2.6.1 Correção do vício

Em Zhou et al. (2017), no Teorema 3, mostra-se que sob a condição $\varrho_n = O(n^{-1/2})$ e as suposições A1-A5, a propriedade da normalidade assintótica padrão é verdadeira, quando $n \to \infty$. Mas, de fato, ϱ_n provavelmente convergiria muito mais lentamente que $n^{-1/2}$. Pela suposição A5 e equação $\| L_n \|^2 = O(\varrho_n^2)$, demonstrada em Zhou et al. (2017), tem-se que o estimador quantílico convergirá para distribuição normal com um viés zero.

Posto isso, para obtenção da estimação do quantil com as correções, utiliza-se os dados observados $\{(t_{1i}, t_{2i}, x_i)\}$, resolvendo o problema de minimização (3.17) primeiro e depois fazendo as correções de polarização apresentadas, a correção de *bootstrap* e a correção pelo método direto. Zhou et al. (2017) concluem que a metodologia *bootstrap*, devido ao cálculo, é mais complexa que o método direto, sendo esse o mais recomendado, principalmente pela rapidez. Em virtude disso, foi tratado somente o método direto,que está sumariamente descrito em seguida. Sugerem também que o erro padrão deve ser obtido via *bootstrap*.

A. Método direto

A partir da prova dos teoremas em Zhou et al. (2017), sabe-se que o vício existe,

principalmente, porque $E(I_5) \neq 0$, em que

$$I_5 = n^{-1} \sum_{i=1}^{n} z' x_{ni} \left[(1-\tau) \mathcal{I}(x_i' \beta_0 > t_{2i}) - \tau \mathcal{I}(x_i' \beta_0 \leq t_{1i}) \right].$$

Portanto, propôs-se o método direto de correção de viés.

Seja

$$\Psi_i^{\beta_0(\tau)}(\beta) = x_i' \beta \left[(1-\tau) \mathcal{I}(x_i' \beta_0(\tau) > t_{2i}) - \tau \mathcal{I}(x_i' \beta_0(\tau) \leq t_{1i}) \right],$$
$$\Psi_i^{\beta_n(\tau)}(\beta) = x_i' \beta \left[(1-\tau) \mathcal{I}(x_i' \beta_n(\tau) > t_{2i}) - \tau \mathcal{I}(x_i' \beta_n(\tau) \leq t_{1i}) \right]$$

e defina que

$$\beta_n^*(\tau) = arg \min_{\beta \in \Theta} \left\{ n^{-1} \sum_{i=1}^{n} \left[\widetilde{\zeta}_i(\tau, \beta) - \Psi_i^{\beta_0(\tau)}(\beta) \right] \right\}.$$

Pela prova do Lema 1 (em Zhou et al. (2017)), é fácil ver que

$$E \left\{ n^{-1} \sum_{i=1}^{n} \left[\widetilde{\zeta}_i(\tau, \beta) - \Psi_i^{\beta_0(\tau)}(\beta) \right] - n^{-1} \sum_{i=1}^{n} \left[\widetilde{\zeta}_i(\tau, \beta_0(\tau)) - \Psi_i^{\beta_0(\tau)}(\beta) \right] \right\}$$
$$= \frac{1}{2} z' z + O(\| z \|^2)$$

uniformemente em n e uniformemente sobre $\| z \| \leq v$, com $v \to 0$, no qual $z = K_n^{1/2}(\beta - \beta_0(\tau))$. Então, pelo pressuposto A1-A5 e a prova do Lema 1 apresentada em Zhou et al. (2017), pode-se escrever:

$$2\sqrt{n} \widetilde{K}_n^{-1/2} K_n (\beta^*(\tau) - \beta_0(\tau)) \xrightarrow{D} N(0, I_m),$$

quando $n \to +\infty$ para qualquer $\tau \in (0,1)$.

Portanto, $\beta_n^*(\tau)$ não é um estimador do vetor de parâmetros, porque $\beta_0^*(\tau)$ é incluído em $\Psi_i^{\beta_0(\tau)}(\beta)$. Como $\hat{\beta}_n(\tau)$, obtido pela equação (3.18) converge para $\beta_0(\tau)$ em probabilidade, é natural definir o estimador corrigido por viés como

$$\hat{\beta}_n^D(\tau) = \arg \min_{\beta \in \Theta} \left\{ n^{-1} \sum_{i=1}^{n} \left[\widetilde{\zeta}_i(\tau, \beta) - \Psi_i^{\hat{\beta}_n(\tau)}(\beta) \right] \right\}. \tag{3.19}$$

Como $\widetilde{\zeta}_i(\tau, \beta)$ e $\Psi_i^{\hat{\beta}_n(\tau)}(\beta)$ são convexos e lineares em β, respectivamente, a função objetiva de minimização no problema (3.19) também é convexa.

Maiores informações, especialmente sobre as provas, os teoremas e os lemas, vide Zhou et al. (2017).

Capítulo 4

Estudo de simulação para comparação dos métodos

Neste capítulo, apresentar-se-á um estudo de simulação, que foi realizado para avaliar a performance dos métodos descritos no capítulo 3, para ajuste do modelo de regressão quantílica para dados com censura intervalar. Esses métodos são: as implementações feitas com a metodologia desenvolvida por Zhou et al. (2017), descrita na subseção 3.2.6, e a proposta de algoritmo, que foi tratado na seção 3.1.

O exemplo apresentado, a seguir, foi extraído de Zhou et al. (2017). Nele é considerado o modelo de regressão quantílica com censura intervalar e com covariável aleatória contínua.

Exemplo 4.1. Neste exemplo, apresentado por Zhou et al. (2017), o modelo de regressão é definido como:

$$y_i = b_0 + b_1 x_1 + e_i,$$

para $b_0 = 1,0 + F^{-1}(\tau)$ e $b_1 = 1,0$, considerando nove valores diferentes de τ (0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 e 0,9), onde F é a função distribuição de e_i. As covariáveis do banco de dados $\{(t_{1i}, t_{2i}, x_i)\}$ foram geradas da seguinte forma:

1. $\{x_i\}$ segue uma distribuição Normal com média 0,5 e desvio padrão 0,5.

2. Os erros da regressão $\{x_i\}$ são derivados de $e_i = \tilde{e}_i - \text{mediana}(\tilde{e})$ para assegurar sua mediana no ponto zero, sendo $\{\tilde{e}_i\}$ gerados independentemente, segundo as quatro distribuições a seguir:

   (a) Distribuição Normal $(0,0\ ;\ 0,3)$;

   (b) Distribuição Lognormal $(0,0\ ;\ 0,3)$;

(c) Distribuição Logística $(0,0\ ;\ 0,2)$;

(d) Distribuição Weibull $(3,0\ ;\ 1,0)$.

3. Para cada i, gerar intervalo de censura $\{(t_{1i}, t_{2i}]\}$, primeiro determina-se $U_i = \min\{y_i\} - 0,3 + r_i$, para $r_i \sim U(0\ ,\ 0,3)$. Em seguida, define-se

$$t_{1i} = U_i + \sum_{j=0}^{k-1} l_j$$

e

$$t_{2i} = U_i + \sum_{j=0}^{k} l_j,$$

para $l_0 = 0$, $l_j \sim U(0\ ,\ 0,3)$ independentemente para $j = 1, \ldots, k$, e k é inteiro não negativo que satisfaz $U_i + \sum_{j=0}^{k-1} l_j < y_i \leq U_i + \sum_{j=0}^{k} l_j$.

Em relação a cada cenário, relatou-se o viés, erro padrão (EP) e o erro quadrático médio (EQM) dos estimadores dos parâmetros com base no tamanho da amostra, n = 200, e no número de amostras, $n^* = 20$, para as execuções da simulação.

Para maior precisão, a imputação não deve ser única; posto isso, para cada observação, foi realizada imputação, ou seja, deve-se completar os dados uma vez e, assim, calcular os seus estimadores com base naqueles dados (Van Buuren, 2012). Portanto, no algoritmo proposto, foram consideradas cinco imputações, tomando a média delas e após calculando as estimativas dos parâmetros. Os resultados das simulações implementadas são apresentados na tabela 4.1, na qual é possível avaliar o desempenho de cada método avaliado, considerando as diferentes suposições.

Tabela 4.1 - Resultados da simulação para o exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Normal}(0,0\ ;\ 0,3)$.

τ	Método	Estimativa de b_0			Estimativa de b_1		
		Vício	**Erro Padrão**	**EQM**	**Vício**	**Erro Padrão**	**EQM**
0,1	Imputação	-0,0088	0,0663	0,0045	0,0061	0,0846	0,0072
	Zhou et al.	0,0690	0,0671	0,0093	0,0164	0,0838	0,0073
0,2	Imputação	0,0041	0,0546	0,0030	-0,0088	0,0815	0,0067
	Zhou et al.	0,0557	0,0485	0,0055	-0,0019	0,0850	0,0072
0,3	Imputação	0,0080	0,0363	0,0014	-0,0103	0,0683	0,0048
	Zhou et al.	0,0403	0,0450	0,0037	-0,0108	0,0753	0,0058
0,4	Imputação	0,0116	0,0364	0,0015	-0,0167	0,0692	0,0051
	Zhou et al.	0,0267	0,0379	0,0021	-0,0118	0,0629	0,0041
0,5	Imputação	0,0112	0,0339	0,0013	-0,0158	0,0642	0,0044
	Zhou et al.	0,0075	0,0329	0,0011	-0,0137	0,0622	0,0041
0,6	Imputação	0,0110	0,0358	0,0014	-0,0128	0,0651	0,0044
	Zhou et al.	-0,0112	0,0302	0,0010	-0,0090	0,0571	0,0033
0,7	Imputação	0,0174	0,0396	0,0019	-0,0157	0,0693	0,0025
	Zhou et al.	-0,0253	0,0434	0,0051	-0,0059	0,0700	0,0049
0,8	Imputação	0,0158	0,0491	0,0027	-0,0135	0,0731	0,0042
	Zhou et al.	-0,0406	0,0509	0,0055	-0,0037	0,0713	0,0051
0,9	Imputação	0,0155	0,0541	0,0032	-0,0065	0,0771	0,0060
	Zhou et al.	-0,0702	0,0463	0,0071	0,0026	0,0607	0,0037

Com relação às avaliações denotadas na tabela 4.1, conclui-se, de modo geral, que há semelhança entre as duas técnicas, pois os valores absolutos resultantes dos vícios são próximos, bem como os valores dos erros padrão. O vício é equivalente ou melhor para alguns $\boldsymbol{\tau}$ iniciais, no caso do algoritmo proposto (com imputações). Entretanto, para $\tau = (0,6,\ 0,7,\ 0,8,\ 0,9)$, esse método apresenta estimativas de b_1 com valores de vício maiores que os da metodologia de Zhou et al. (2017). Quanto aos erros quadráticos médios, para os métodos analisados nos diferentes $\boldsymbol{\tau}$, eles foram valores baixos e também bem próximos entre si, com mudança somente na terceira ou quarta casa decimal, evidenciando a eficiência de ambas as técnicas.

Tabela 4.2 - Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Lognormal$(0,0\ ;\ 0,3)$.

τ	Método	Estimativa de b_0			Estimativa de b_1		
		Vício	**Erro Padrão**	**EQM**	**Vício**	**Erro Padrão**	**EQM**
0,1	Imputação	-0,0178	0,0459	0,0024	0,0133	0,0570	0,0034
	Zhou et al.	0,0661	0,0394	0,0059	0,0045	0,0587	0,0035
0,2	Imputação	-0,0046	0,0404	0,0017	0,0063	0,0494	0,0025
	Zhou et al.	0,0504	0,0335	0,0037	0,0066	0,0446	0,0020
0,3	Imputação	0,0052	0,0390	0,0015	-0,0054	0,0521	0,0027
	Zhou et al.	0,0389	0,0339	0,0027	0,0059	0,0485	0,0024
0,4	Imputação	0,0098	0,0294	0,0010	-0,0090	0,0483	0,0024
	Zhou et al.	0,0275	0,0287	0,0016	-0,0031	0,0465	0,0022
0,5	Imputação	0,0089	0,0254	0,0007	-0,0045	0,0490	0,0024
	Zhou et al.	0,0098	0,0233	0,0006	0,0023	0,0484	0,0023
0,6	Imputação	0,0051	0,0258	0,0007	0,0046	0,0487	0,0024
	Zhou et al.	-0,0034	0,0254	0,0007	0,0003	0,0496	0,0025
0,7	Imputação	0,0084	0,0380	0,0015	0,0098	0,0648	0,0043
	Zhou et al.	-0,0210	0,0381	0,0019	0,0120	0,0645	0,0043
0,8	Imputação	0,0070	0,0522	0,0028	0,0215	0,0763	0,0063
	Zhou et al.	-0,0478	0,0529	0,0051	0,0210	0,0718	0,0056
0,9	Imputação	0,0125	0,0730	0,0055	0,0283	0,1046	0,0117
	Zhou et al.	-0,0728	0,0773	0,0113	0,0316	0,1043	0,0119

Na tabela 4.2, em que $\tilde{e}_i$ segue uma distribuição Lognormal, observa-se que as metodologias, com relação à estimativa de b_1, demonstram comportamentos, na maior parte, parecidos considerando os vícios, mas há valores de $\boldsymbol{\tau}$, como $0,1$ e $0,6$, que exibem valores para o vício ainda um pouco menores para o método de Zhou et al. (2017). No entanto, para o erro quadrático médio, os valores das técnicas são equiparáveis, pequenos e exibindo, na grande maioria, diferença somente na quarta casa decimal, sendo assim, com desempenho similares.

Tabela 4.3 - Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Logística$(0,0\ ;\ 0,2)$.

τ	Método	Estimativa de b_0			Estimativa de b_1		
		Vício	**Erro Padrão**	**EQM**	**Vício**	**Erro Padrão**	**EQM**
0,1	Imputação	-0,0256	0,0459	0,0028	-0,0110	0,0714	0,0052
	Zhou et al.	0,0572	0,0523	0,0060	-0,0059	0,0679	0,0046
0,2	Imputação	-0,0170	0,0372	0,0017	-0,0015	0,0534	0,0029
	Zhou et al.	0,0393	0,0331	0,0026	0,0029	0,0426	0,0018
0,3	Imputação	-0,0063	0,0320	0,0011	-0,0053	0,0517	0,0019
	Zhou et al.	0,0263	0,0348	0,0027	-0,0023	0,0511	0,0026
0,4	Imputação	0,0004	0,0298	0,0009	-0,0064	0,0542	0,0030
	Zhou et al.	0,0166	0,0352	0,0015	-0,0054	0,0502	0,0025
0,5	Imputação	0,0022	0,0272	0,0007	-0,0084	0,0450	0,0021
	Zhou et al.	0,0001	0,0359	0,0013	-0,0070	0,0520	0,0028
0,6	Imputação	0,0052	0,0319	0,0010	-0,0119	0,0476	0,0024
	Zhou et al.	-0,0120	0,0337	0,0013	-0,0055	0,0481	0,0023
0,7	Imputação	0,0083	0,0450	0,0021	-0,0087	0,0584	0,0035
	Zhou et al.	-0,0288	0,0426	0,0026	-0,0060	0,0540	0,0030
0,8	Imputação	0,0165	0,0542	0,0032	-0,0153	0,0668	0,0047
	Zhou et al.	-0,0419	0,0486	0,0041	-0,0070	0,0729	0,0054
0,9	Imputação	0,0220	0,0533	0,0033	-0,0007	0,0859	0,0074
	Zhou et al.	-0,0595	0,0527	0,0063	0,0067	0,0747	0,0056

Para estimativa do coeficiente b_1 na tabela 4.3, verifica-se que os erros quadráticos médios, que levam em consideração tanto o vício quanto o erro padrão, apresentam valores bem próximos e baixos, demonstrando que os métodos abordados são equiparáveis em performance.

Tabela 4.4 - Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Weibull}(3,0\ ;\ 1,0)$.

τ	Método	Estimativa de b_0			Estimativa de b_1		
		Vício	**Erro Padrão**	**EQM**	**Vício**	**Erro Padrão**	**EQM**
0,1	Imputação	-0,0037	0,0358	0,0013	-0,0233	0,0609	0,0042
	Zhou et al.	0,0743	0,0332	0,0066	-0,0205	0,0481	0,0027
0,2	Imputação	-0,0011	0,0401	0,0016	-0,0217	0,0744	0,0060
	Zhou et al.	0,0620	0,0400	0,0054	-0,0242	0,0680	0,0052
0,3	Imputação	0,0032	0,0420	0,0018	-0,0146	0,0769	0,0061
	Zhou et al.	0,0426	0,0405	0,0035	-0,0169	0,0692	0,0051
0,4	Imputação	0,0118	0,0363	0,0015	-0,0206	0,0715	0,0055
	Zhou et al.	0,0211	0,0371	0,0018	-0,0180	0,0655	0,0046
0,5	Imputação	0,0180	0,0382	0,0018	-0,0253	0,0661	0,0050
	Zhou et al.	0,0102	0,0317	0,0011	-0,0200	0,0592	0,0039
0,6	Imputação	0,0155	0,0315	0,0012	-0,0209	0,0560	0,0036
	Zhou et al.	-0,0048	0,0300	0,0009	-0,0216	0,0572	0,0037
0,7	Imputação	0,0183	0,0305	0,0013	-0,0254	0,0478	0,0029
	Zhou et al.	-0,0195	0,0303	0,0013	-0,0173	0,0514	0,0029
0,8	Imputação	0,0212	0,0433	0,0023	-0,0228	0,0547	0,0035
	Zhou et al.	-0,0379	0,0470	0,0037	-0,0161	0,0549	0,0033
0,9	Imputação	0,0305	0,0588	0,0044	-0,0109	0,0754	0,0058
	Zhou et al.	-0,0517	0,0579	0,0060	-0,0065	0,0649	0,0043

No caso em que $\tilde{e}_i$ segue uma distribuição Weibull, tabela 4.4, é possível notar que as metodologias apresentaram, de uma forma geral, valores equivalentes quanto às avaliações expostas. Os resultados alcançados para os erros quadráticos médios também foram muito próximos, como constatado nas tabelas anteriores para outros cenários.

Os gráficos gerados, para auxiliar na visualização dos resultados da simulação, abordados previamente, estão expostos no Apêndice A.

Capítulo 5

Aplicação das metodologias a dados empíricos

Neste capítulo, como já mencionado na introdução, abordar-se-ão os comportamentos das metodologias estudadas (de Zhou et al. e do algoritmo proposto) aplicados a dados reais. Para isso, foram utilizados os seguintes conjuntos de dados, que apresentam censura intervalar: compras de celulares e ocorrência de cáries.

5.1 Bancos de dados

Compras de Celulares

Um dos bancos de dados, a ser utilizado, é sobre compras de celulares e está disponível no pacote *icensBKL* do *software* R, nomeado como *mobile*. Conforme Bogaerts et al. (2017), em fevereiro de 2013, uma pesquisa sobre compras de celulares foi realizada na Finlândia, entre proprietários de 15 a 79 anos de idade de um telefone celular. Os participantes foram amostrados aleatoriamente, a partir de um diretório de números de telefone disponíveis publicamente, estabelecendo quotas no sexo, idade e região dos entrevistados. Um total de 536 entrevistas concluídas foram gravadas usando um sistema de entrevista telefônica assistida por computador (CATI). A quantidade de proprietários do sexo feminino, mas também de proprietários de 15 a 24 anos, estava sub-representada nos dados, enquanto os homens e os proprietários de 65 a 79 anos estavam super-representados no estudo, em comparação com as estatísticas oficiais finlandesas de 2012. Os entrevistados foram solicitados a responder perguntas sobre a compra de seus celulares (atuais e anteriores) e a relatar algumas características familiares. As entrevistas se concentraram nas seguintes questões:

Q1. Quando você comprou o seu celular? (mês e ano; se o mês não foi lembrado, a estação

do ano foi solicitada);

Q2. Quando você comprou o seu celular anterior? (ano e mês; se mês não foi lembrado, a estação do ano foi solicitada);

Q3. Qual é o seu gênero? (masculino; feminino);

Q4. Qual é o seu grupo etário? (15–24; 25–34; 35–44; 45–54; 55–64; 65–79 anos de idade);

Q5. Qual é o tamanho da sua casa? (quantidade de pessoas por domicílio: 1, 2, 3, 4, 5 ou mais pessoas; sem resposta);

Q6. Qual é o seu rendimento familiar antes dos impostos? (30.000 ou menos; 30.001–50.000; 50.001–70.000; mais de 70.000 euros; sem resposta).

Os tempos de compra (mês e ano) são censurados por intervalo, porque apenas o mês, e não o dia da compra, foi solicitado. Além disso, muitos entrevistados não conseguiram lembrar o momento da compra, conforme exibido na tabela 5.1.

Tabela 5.1 - Informações sobre o celular atual fornecidas pelos entrevistados.

Conseguiram informar o mês e o ano da compra	310
Conseguiram fornecer a estação do ano e o ano	115
Não conseguiram lembrar nem o ano	37

Dos 517 entrevistados que responderam às perguntas sobre o telefone anterior, 117 conseguiram informar o mês e o ano da compra, 91 foram capazes de relatar a estação do ano e o ano, 146 forneceram apenas o ano e 163 não conseguiram lembrar nem mesmo o ano. Os seguintes entrevistados foram excluídos da análise: três entrevistados que informaram que seu telefone anterior foi comprado após o telefone atual, 30 entrevistados para os quais os intervalos de compra do telefone anterior e atual são completamente sobrepostos e 6 entrevistados que não informaram o tamanho do domicílio. Como resultado, o conjunto de dados usado em nossas análises contém 478 entrevistados, os quais informaram corretamente que tinham um telefone anterior e têm intervalos não sobrepostos para os tempos de compra do telefone anterior e atual, com 258 homens e 220 mulheres. Mais detalhes sobre a pesquisa podem ser encontrados em Karvanen et al. (2014) e Bogaerts et al. (2017).

O conjunto de dados apresenta 16 variáveis. Para a aplicação, no entanto, somente essas foram utilizadas:

1. CLL: limite inferior numérico da data de compra do telefone atual;

2. CUL: limite superior numérico da data de compra do telefone atual;

3. agegrp (faixa etária): 1 = "15–24 anos", 2 = "25–34 anos", 3 = "35–44 anos", 4 = "45–54 anos", 5 = "55–64 anos" e 6 = "65–79 anos". Através da variável "agegrp", obtém-se, por construção, uma outra variável denominada "age", tomando o ponto médio dos intervalos de "agegrp". Portanto, a variável "age", que será utilizada no estudo, assume os seguintes valores: 20 anos, 30 anos, 40 anos, 50 anos, 60 anos e 72 anos.

Tanto "CLL" como "CUL", tiveram as datas modificadas para mostrar exatamente a quantidade de tempo transcorrido após a compra do celular atual, passando a serem denominadas como "L" e "R", respectivamente, sendo assim, foi colocando a idade do celular como idade igual a zero em 31/01/2013. Verificou-se que o intervalo máximo da data de aquisição do telefone foi de 196 meses.

Tabela 5.2 - Medidas resumo dos limites superior e inferior das datas desde a aquisição do celular (em dias).

Medidas	Limite Inferior (L)	Limite Superior (R)
Mín.	0,0	30,0
1º Q.	215,0	224,0
Mediana	489,0	580,0
Média	726,7	817,6
3º Q.	938,3	1047,0
Máx.	5510,0	5874,0

Tabela 5.3 - Frequência das faixas etárias dos entrevistados.

Faixa etária	15–24	25–34	35–44	45–54	55–64	65–79	Total
N	57	83	66	81	95	96	478

Na Figura 5.1, pode-se visualizar a relação entre o tempo transcorrido da compra do celular (considerando as variáveis "CLL" e "CUL", que foram modificadas para contabilizar exatamente o tempo transcorrido desde a aquisição do celular, segundo já mencionado) e

a idade. Para facilitar a visualização da relação do tempo transcorrido desde a compra do celular com a "Idade", foi realizada uma distribuição das idades ao longo do intervalo de faixa etária aleatoriamente na Figura 5.1, exibindo linhas individuais para cada observação. Percebe-se que o tempo em que a pessoa está com o celular depende da idade, quanto mais nova a pessoa, mais novo é o celular. Os gráficos das curvas de sobrevivência global e por idade estão no apêndice B.

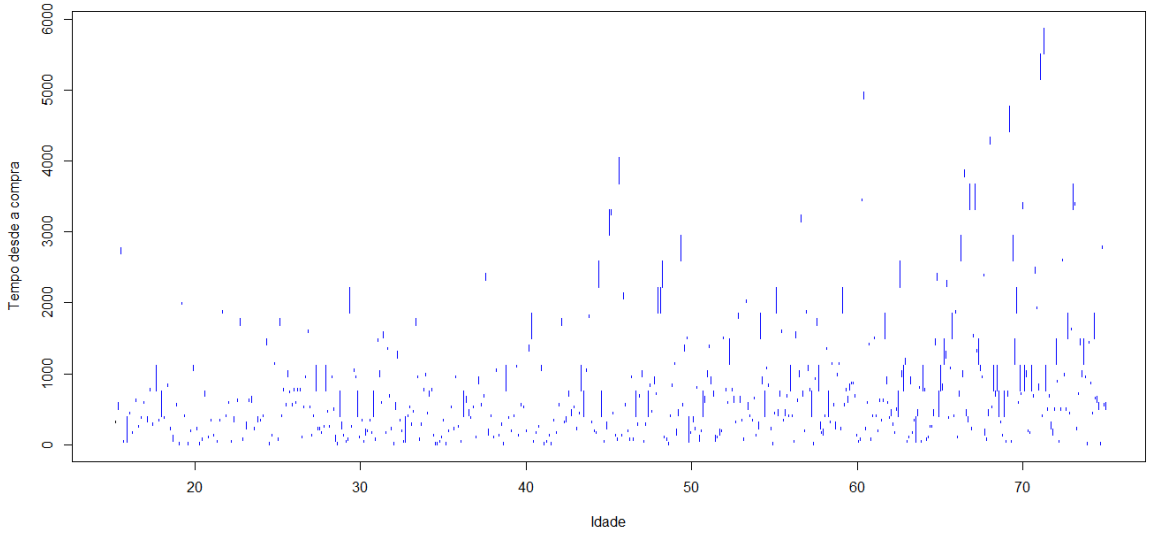

Figura 5.1: Gráficos da relação entre o Tempo (intervalo para a data de compra do celular atual em dias) e a Idade (em anos).

Para esse conjunto de dados, visa-se modelar o tempo transcorrido desde a aquisição do celular atual analisando os efeitos da covariável idade nas respostas quantílicas, utilizando o seguinte modelo de regressão quantílica:

$$Q_{Tempo}(\tau|Age) = \hat{\beta}_0(\tau) + \hat{\beta}_1(\tau)Age.$$

As estimativas dos coeficientes, $\hat{\beta}_0$ e $\hat{\beta}_1$, foram obtidas através do método de Zhou et al. (2017), corrigido pela metodologia direta (subseção 3.2.6) e da proposta de algoritmo, com imputação (seção 3.1).

Tabela 5.4 - Estimativas para os parâmetros dos modelos de regressão quantílica, os erros padrão e intervalos de confiança.

τ	Método	$\hat{\beta}_0$	Erro Padrão	IC Inf	IC Sup	$\hat{\beta}_1$	Erro Padrão	IC Inf	IC Sup
0,1	Imputação	0,0189	39,7475	-79,4761	79,5139	1,9133	1,2612	-0,6091	4,4357
	Zhou et al.	-4,6652	32,7589	-70,1830	60,8526	2,2417	0,8769	0,4879	3,9955
0,2	Imputação	2,6710	44,5615	-86,452	91,794	4,0964	1,1158	1,8648	6,328
	Zhou et al.	1,4694	66,2244	-130,9794	133,9182	4,4831	1,2912	1,9007	7,0655
0,3	Imputação	13,5882	65,0442	-116,5002	143,6766	6,3918	1,2981	3,7956	8,9880
	Zhou et al.	31,4064	90,9770	-150,5476	213,3604	6,0799	1,8306	2,4187	9,7411
0,4	Imputação	101,1631	67,7850	-34,4069	236,7331	6,7703	1,5550	3,6603	9,8803
	Zhou et al.	100,5565	78,4625	-56,3685	257,4815	6,8514	1,5044	3,8426	9,8602
0,5	Imputação	181,6941	39,2807	103,1327	260,2555	8,5864	1,1671	6,2522	10,9206
	Zhou et al.	142,3049	124,3365	-106,3681	390,9779	8,8419	2,4351	3,9717	13,7121
0,6	Imputação	177,2252	54,2538	68,7176	285,7328	10,6087	1,3727	7,8633	13,3541
	Zhou et al.	194,8319	103,9302	-13,0285	402,6923	9,9221	2,1311	5,6599	14,1843
0,7	Imputação	316,1764	163,3257	-10,475	642,8278	11,2510	3,6269	3,9972	18,5048
	Zhou et al.	240,5359	183,0574	-125,5789	606,6507	12,5493	3,6499	5,2495	19,8491
0,8	Imputação	258,0748	119,4991	19,0766	497,073	18,7316	4,4742	9,7832	27,6800
	Zhou et al.	250,5780	204,8714	-159,1648	660,3208	18,0117	4,1805	9,6507	26,3727
0,9	Imputação	243,5727	465,7878	-688,0029	1175,1483	32,9519	10,1815	12,5889	53,3149
	Zhou et al.	146,1069	374,5354	-602,9639	895,1777	34,2196	7,3205	19,5786	48,8606

Na tabela 5.4, percebe-se que há relativa proximidade dos valores de $\hat{\beta}_1$ quando comparadas as metodologias de imputação e de Zhou et al.. O mesmo se verifica para seus erros padrão, que apresentaram, no geral, comportamentos similares.

Para os dois métodos estudados, pode-se observar que há variação dos interceptos, porém há erros padrão muito elevados, o que os tornam não significativos. Nas inclinações das retas ajustadas, também se verifica uma variação, sugerindo que os dados sejam heterocedásticos. Portanto, percebe-se que os efeitos, ocasionados pela idade, aumentam dos quantis de ordem baixa para os de ordem alta, confirmando que se refere a um modelo escala-locação, pois aponta alteração na tendência central e variabilidade no tempo desde a compra do celular. A maior variação no tempo de aquisição do celular, gerado pela mudança de idade, está no quantil de ordem 0,9, sugerindo que, entre indivíduos que têm mais tempo desde a aquisição de celular, a idade é uma condição relevante. Na Figura 5.2, foram plotadas apenas algumas das retas ajustadas para melhor visualização.

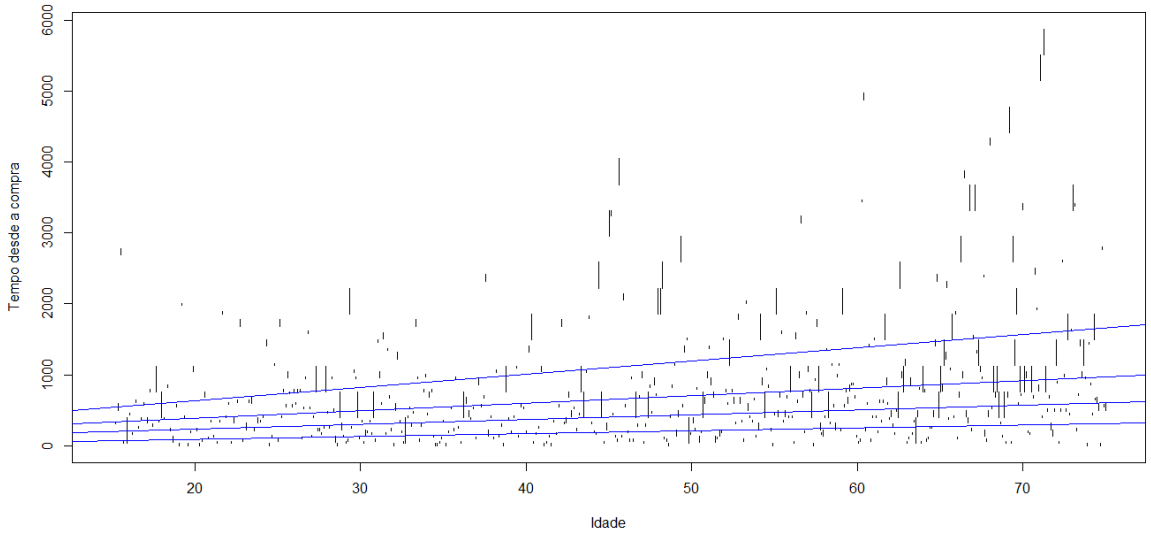

Figura 5.2: Gráfico do ajuste das retas de regressão, utilizando o quantil de ordem $\tau = (0,2; 0,4; 0,6; 0,8)$ e segundo o método proposto nessa dissertação.

Ocorrência de Cáries

Outro conjunto de dados também usado é resultante de um estudo odontológico prospectivo longitudinal, que combina o registro de dados de saúde bucal e a promoção da saúde bucal, realizado em Flandres (norte da Bélgica) de 1996 a 2001. Nesse estudo coorte, 4.468 crianças de idade escolar primária (primeiro ano da escola básica no início do estudo) foram aleatoriamente amostradas e com exames dentais realizados anualmente por um dos 16 dentistas treinados. O banco de dados apresentado contém, principalmente, as informações sobre os tempos de surgimento dentário e de cárie resumido nas observações censuradas por intervalo. Algumas covariáveis da linha de base também estão incluídas.

Informações sobre hábitos de saúde bucal, atendimento odontológico, histórico de traumatismo dentário e dor de dente relacionada às crianças foram obtidas dos pais das crianças por meio de questionário estruturado. No grupo de intervenção, esse questionário foi repetido anualmente com pequenas adaptações a cada ano. Os pais das crianças do grupo controle receberam um questionário, idêntico ao usado no grupo de intervenção, mas apenas no início e no final do período de 6 anos. O questionário foi validado durante a fase de pré-teste. Somente dentro do grupo de intervenção o programa de educação em saúde bucal foi entregue. Isso incluiu uma sessão de educação em saúde bucal para crianças e professores uma vez por ano, precedendo o exame individual de saúde bucal. O conjunto de dados está disponível no pacote *icensBKL* do *software* R, nomeado como *tandmobAll*.

Para mais detalhes sobre o desenho do estudo, ver Vanobbergen et al. (2000).

O conjunto de dados apresenta 143 variáveis. Entretanto, para a aplicação, extraiu-se uma amostra de tamanho $n = 500$ e somente essas variáveis foram utilizadas:

1. STARTBR: fator, indica a idade que iniciou a escovação dos dentes (relatado pelos pais) com: 1 =<= 1 = $[0, 1]$ anos, 2 = $(1, 2]$ = $(1, 2]$ anos, 3 = $(2, 3]$ = $(2, 3]$ anos, 4 = $(3, 4]$ = $(3, 4]$ anos, 5 = $(4, 5]$ = $(4, 5]$ anos e 6 => 5 = depois que com 5 anos de idade. Através da variável "STARTBR", obtém-se, por construção, uma outra variável denominada "age", tomando o ponto médio dos intervalos de "STARTBR". Portanto, a variável "age", que será utilizada no estudo, assume os seguintes valores: 0,5 anos, 1,5 anos, 2,5 anos, 3,5 anos, 4,5 anos e 5,5 anos;

2. FBEG.16: limite inferior para o tempo de cárie (em anos de idade, 'F' significa 'falha') do dente permanente 16 (primeiro molar permanente). "NA" se o tempo de cárie fosse censurado à esquerda;

3. FEND.16: limite superior para o tempo de cárie (em anos de idade, 'F' significa 'falha') do dente permanente 16. "NA" se o tempo de cárie fosse censurado à direita.

Tabela 5.5 - Frequência das faixas de idade das crianças de quando iniciaram a escovação dos dentes.

Idade	<=1	(1,2]	(2,3]	(3,4]	(4,5]	>5	Total
N	59	212	145	54	25	5	500

Tabela 5.6 - Medidas resumo dos limites superior e inferior das idades da primeira ocorrência de cárie (em anos).

Medidas	Limite Inferior (L)	Limite Superior (R)
Mín.	0,0	6,6
1º Q.	8,4	11,5
Mediana	11,0	20,0
Média	9,9	17,15
3º Q.	11,6	20,0
Máx.	12,3	20,0

A amostra apresenta, assim como o conjunto de dados original, um elevado número de dados censurados, cerca de 73%, sendo a maior parte de censura à direita. Os valores 0 e 20

são valores artificiais colocados, referentes as variáveis “L” e “R”, para os casos de censura à esquerda e à direita, respectivamente; para esse último, é um valor do limite superior arbitrário e maior que todos os limites dos intervalos. A idade máxima de surgimento de cárie encontrada no estudo foi em um adolescente de 12 anos e quatro meses e a menor foi em uma criança de seis anos e dois meses. A figura 5.3 apresenta os casos de censura à esquerda, intervalar e à direita (esse com os pontos em cor roxa). Os gráficos das curvas de sobrevivência global e por idade do início da escovação estão no apêndice B.

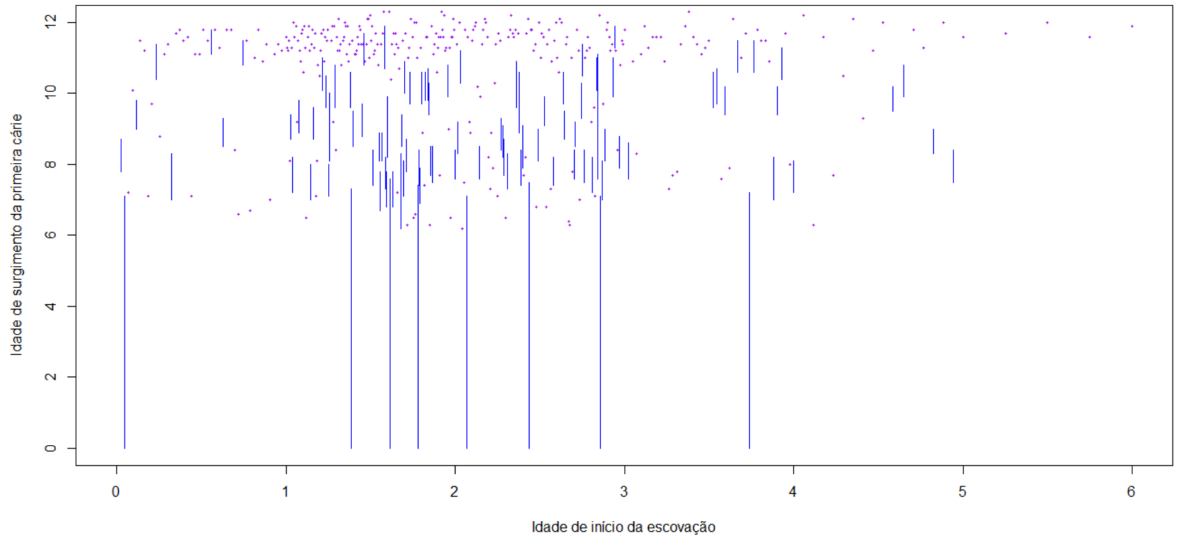

Figura 5.3: Gráficos da relação entre o Tempo (em anos) de ocorrência da primeira cárie e a Idade (em anos) do início da escovação.

Tabela 5.7 - Estimativas para os parâmetros dos modelos de regressão quantílica, os erros padrão e e intervalos de confiança.

τ	Método	$\hat{\beta_0}$	Erro Padrão	IC Inf	IC Sup	$\hat{\beta_1}$	Erro Padrão	IC Inf	IC Sup
0,1	Imputação	7,6730	0,5944	6,4842	8,8618	0,2162	0,3080	-0,3998	0,8322
	Zhou et al.	8,2606	0,3760	7,5086	9,0126	0,0871	0,1753	-0,2635	0,4377
0,2	Imputação	10,1669	0,3227	9,5215	10,8123	-0,0576	0,0861	-0,2298	0,1146
	Zhou et al.	9,7466	0,6260	8,4946	10,9986	-0,0453	0,2587	-0,5627	0,4721

O algoritmo proposto, aplicado à amostra dos dados sobre ocorrência de cáries, retorna estimativas apenas para $\tau = (0,1\ ;\ 0,2)$. No entanto, o método de Zhou et al. (2017)

fornece estimativas para todos os decis, pois se colocou um valor arbitrário maior que todos os limites dos intervalos para as observações com censura à direita, como já dito. Testou-se dois valores para este limite superior artificial, 20 e 99. As estimativas de β_0 e β_1 não se alteram para $\tau = 0,1$ e $\tau = 0,2$, mas foram influenciadas pelo limite superior artificial para τ de ordem maior que $0,2$, o que é esperado por termos cerca de 73% de censura à direita. Isso corrobora com o que foi mencionado por Zhou et al. (2017), que seu método só vale para intervalos finitos e não funciona para dados com censura à direita. O erro padrão foi computado via *bootstrap*, segundo sugerido em Zhou et al. (2017).

No entanto, na tabela tabela 5.7, verifica-se, para as técnicas estudadas, que o efeito da idade de início de escovação não influencia na idade de início da primeira cárie (não é significativo), considerando os dois decis analisados em virtude da magnitude do erro padrão. A Figura 5.4 corrobora com essa afirmação.

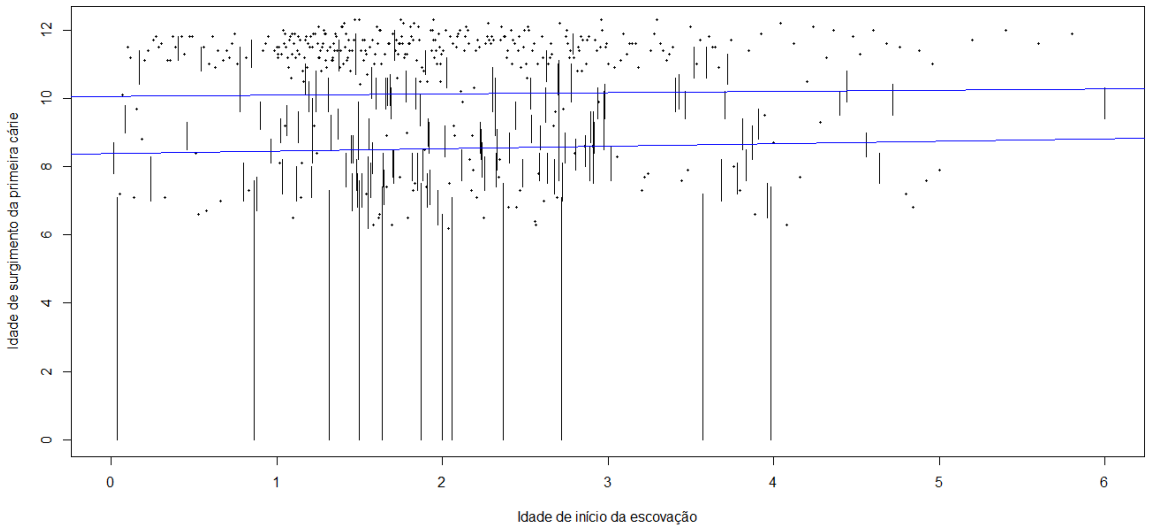

Figura 5.4: Gráfico do ajuste das retas de regressão, utilizando o quantil de ordem $\tau = (0,1; 0,2)$ e segundo o método proposto nessa dissertação.

Capítulo 6

Considerações finais

Nessa dissertação, foram estudados os modelos de regressão quantílica aplicados a dados com censura intervalar. Na introdução, apresentou-se uma contextualização sobre a robustez dos modelos de regressão quantílica, que proporcionam uma perspectiva mais ampla das distribuições condicionais da variável resposta, percebendo o efeito das covariáveis nos diferentes níveis de quantil (Koenker e Bassett Jr, 1978). Além disso, realizou-se uma sucinta comparação à tradicional estimação dos mínimos quadrados dos modelos de regressão linear.

As propriedades da regressão quantílica foram brevemente abordadas, referentes à robustez aos pressupostos e à equivariância. Ademais, para a análise de sobrevivência, a facilidade de interpretação direta das estimativas, como o efeito no tempo de sobrevivência (Koenker e Bassett Jr, 1978; Koenker et al., 2017).

No terceiro capítulo, tratou-se dos modelos de regressão quantílica a dados censurados, mostrando que diversos autores têm estudado a técnica estendida à análise de sobrevivência para o seu desenvolvimento. Entre as metodologias tratadas, está a adaptação do algoritmo proposto para a estimação de máxima verossimilhança não paramétrica da distribuição da variável resposta na presença da censura intervalar, considerando a aplicação do método de Portnoy para dados com censura à direita, descrito em Rasteiro (2017), onde se fez a imputação dos tempos de falha, para os casos em que o intervalo de censura é finito. A outra metodologia descrita, com a qual foi comparada o método proposto na dissertação, foi a de Zhou et al. (2017), que visa a generalização da regressão quantílica para dados observados completos, apresentando um método de estimação para modelos de regressão quantílica a dados com censura intervalar, com aplicação de correção de viés, no qual foi realizado o método direto.

No estudo de simulação, realizado no capítulo quatro, utilizou-se um exemplo proposto em Zhou et al. (2017), que evidenciou equivalência das duas técnicas, para as avaliações expostas, o viés, o erro padrão e, especialmente, o erro quadrático médio, que apresentou valores pequenos e muito próximos entre os dois métodos. Entretanto, quando comparada a velocidade da performance das metodologias estudadas, o método proposto na dissertação é mais intensivo computacionalmente, o que o torna bem mais lento que o do Zhou et al. (2017).

A aplicação a conjuntos de dados reais, no capítulo 5, acabou salientando uma limitação de Zhou et al. (2017) quando há dados com censura à direita, tendo as estimativas geradas, particularmente nos últimos quantis de ordem τ, influenciadas pelo limite superior artificial, o que não ocorre com a técnica proposta nesse estudo. Então, entende-se que a metodologia exposta nesse trabalho, apesar de ser mais complexa computacionalmente, pode ser utilizada em estudos onde não há somente intervalos finitos.

Como sugestão de melhoria, pode-se trabalhar mais com relação a velocidade do algoritmo, talvez analisar a questão da bissecção, que o faz ser mais intensivo e, possivelmente, ver uma outra forma de torná-lo mais rápido nessa parte da técnica.

Referências Bibliográficas

Abrevaya J., Economic applications of quantile regression. Springer, 2002.

Ali A., Nonparametric spatial rainfall characterization using adaptive kernel estimator, Journal of Geographic Information and Decision Analysis, 1998, vol. 2, p. 34

Austin P. C., Tu J. V., Daly P. A., Alter D. A., The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy, Statistics in medicine, 2005, vol. 24, p. 791.

Ayer M., Brunk H. D., Ewing G. M., Reid W. T., Silverman E., An empirical distribution function for sampling with incomplete information, The annals of mathematical statistics, 1955, pp 641–647.

Barlow R., Bartholomew D., Bremner J., Brunk H., Statistical inference under order restrictions (the theory and application of isotonic regression). John Wiley & Sons, 1972.

Barrodale I., Roberts F. D., An improved algorithm for discrete l_1 linear approximation, SIAM Journal on Numerical Analysis, 1973, vol. 10, p. 839.

Becker N. G., Melbye M., Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for hiv positivity, Australian & New Zealand Journal of Statistics, 1991, vol. 33, p. 125.

Bedi A. S., Edwards J. H., The impact of school quality on earnings and educational returns-evidence from a low-income country, Journal of Development Economics, 2002, vol. 68, p. 157.

Bogaerts K., Komarek A., Lesaffre E., Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS. Chapman and Hall/CRC, 2017

Buchinsky M., Changes in the US wage structure 1963-1987: Application of quantile regression, Econometrica: Journal of the Econometric Society, 1994, pp 405–458.

Budd J. W., McCall B. P., The Grocery stores wage distribution: A semi-parametric analysis of the role of retailing and labor market institutions, ILR Review, 2001, vol. 54, p. 484.

Cade B. S., Terrell J. W., Schroeder R. L., Estimating effects of limiting factors with regression quantiles, Ecology, 1999, vol. 80, p. 311.

Cai T., Cheng S., Semiparametric regression analysis for doubly censored data, Biometrika, 2004, vol. 91, p. 277.

Chamberlain G., Quantile regression, censoring, and the structure of wages, Advances in econometrics, 1994, vol. 1, p. 171.

Chang M. N., et al., Weak convergence of a self-consistent estimator of the survival function with doubly censored data, The Annals of Statistics, 1990, vol. 18, p. 391.

Chay K. Y., Honore B. E., Estimation of semiparametric censored regression models: an application to changes in black-white earnings inequality during the 1960s, Journal of Human Resources, 1998, pp 4–38.

Chay K. Y., Powell J. L., Semiparametric censored regression models, Journal of Economic Perspectives, 2001, vol. 15, p. 29.

Chen C., Wei Y., Computational issues for quantile regression, Sankhyā: The Indian Journal of Statistics, 2005, pp 399–417.

Dantzig G. B., Linear programming, Operations research, 2002, vol. 50, p. 42.

Davino C., Furno M., Vistocco D., Quantile regression: theory and applications. John Wiley & Sons, 2013.

Efron B., The two sample problem with censored data. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , vol. 4, 1967, p. 831.

Efron B., Tibshirani R. J., An introduction to the bootstrap. CRC press, 1994.

Eide E. R., Showalter M. H., Factors affecting the transmission of earnings across generations: A quantile regression approach, Journal of Human Resources, 1999, pp 253–267.

Eide E. R., Showalter M. H., Sims D. P., The effects of secondary school quality on the distribution of earnings, Contemporary Economic Policy, 2002, vol. 20, p. 160.

Fiacco A., Mccormick G., Nonlinear Programming: Sequential Unconstrained Minimization Techniques Wiley New York NY Google Scholar, 1968.

Finkelstein D. M., Wolfe R. A., A semiparametric model for regression analysis of interval-censored failure time data, Biometrics, 1985, pp 933–945.

Fortin N. M., Lemieux T., Rank regressions, wage distributions, and the gender gap, Journal of Human Resources, 1998, pp 610–643.

Frisch R., La résolution des problèmes de programme linéaire par la méthode du potentiel logarithmique, Cahiers du Seminaire D'Econometrie, 1956, pp 7–23.

Gentleman R., Geyer C. J., Maximum likelihood for interval censored data: Consistency and computation, Biometrika, 1994, vol. 81, p. 618.

Gilchrist W., Statistical modelling with quantile functions. CRC Press, 2000.

Giolo S. R., Colosimo E. A., Análise de sobrevivência aplicada, Edgard Blucher, 2006.

Gould W., et al., Quantile regression with bootstrapped standard errors, Stata Technical Bulletin, 1993., vol. 2

Groeneboom P., Wellner J. A., Information bounds and nonparametric maximum likelihood estimation. Birkhäuser, 1992.

Hall P., Sheather S. J., On the distribution of a studentized quantile, Journal of the Royal Statistical Society. Series B (Methodological), 1988, pp 381–391

Hao L., Immigration and wealth inequality: A distributional approach. In Invited seminar at The Center for the Study of Wealth and Inequality, Columbia University , 2005.

Hao L., Sources of wealth inequality: Analyzing conditional distribution. In Invited seminar at The Center for Advanced Social Science Research , 2006.a

Hao L., Sources of wealth inequality: Analyzing conditional location and shape shifts. In Research Committee on Social Stratification and Mobility (RC28) of the International Sociological Association (ISA) Spring meeting 2006 in Nijmegen , 2006.b

Hao L., Naiman D. Q., Quantile regression. Quantitative applications in the social sciences, 2007.

He X., Hu F., Markov chain marginal bootstrap, Journal of the American Statistical Association, 2002, vol. 97, p. 783.

Hendricks W., Koenker R., Hierarchical spline models for conditional quantiles and the demand for electricity, Journal of the American statistical Association, 1992, vol. 87, p. 58

Heritier S., Cantoni E., Copt S., Victoria-Feser M.-P., Robust methods in Biostatistics. vol. 825, John Wiley & Sons, 2009.

Hollander M., Wolfe D. A., Chicken E., Nonparametric statistical methods. vol. 751, John Wiley & Sons, 2013.

Huang J., Wellner J. A., Interval censored survival data: a review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics , 1997, p. 123.

Huber P. J., et al., The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , vol. 1, 1967, p. 221.

Karmarkar N., A new polynomial-time algorithm for linear programming. In Proceedings of the sixteenth annual ACM symposium on Theory of computing , 1984, p. 302.

Karvanen J., Rantanen A., Luoma L., Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity, Quantitative Marketing and Economics, 2014, vol. 12, p. 305

Kim Y. J., Cho H., Kim J., Jhun M., Median regression model with interval censored data, Biometrical Journal, 2010, vol. 52, p. 201.

Kocherginsky M., He X., Mu Y., Practical confidence intervals for regression quantiles, Journal of Computational and Graphical Statistics, 2005, vol. 14, p. 41.

Kocherginsky M. N., Extensions of Markov chain marginal bootstrap, University of Illinois at Urbana-Champaign, 2003, Tese de Doutorado

Koenker R., Quantile regression, Econometric Society Monographs, 2005, vol. 38.

Koenker R., Bassett Jr G., Regression quantiles, Econometrica: journal of the Econometric Society, 1978, pp 33.–50.

Koenker R., Chernozhukov V., He X., Peng L., Handbook of Quantile Regression. CRC Press, 2017.

Koenker R., Geling O., Reappraising medfly longevity: a quantile regression survival analysis, Journal of the American Statistical Association, 2001, vol. 96, p. 458.

Koenker R., Machado J. A., Goodness of fit and related inference processes for quantile regression, Journal of the american statistical association, 1999, vol. 94, p. 1296

Koenker R., Portnoy S., Quantile regression. ABE, 1996.

Koenker R., Xiao Z., Inference on the quantile regression process, Econometrica, 2002, vol. 70, p. 1583.

Koenker R. W., d'Orey V., Algorithm AS 229: Computing regression quantiles, Journal of the Royal Statistical Society. Series C (Applied Statistics), 1987, vol. 36, p. 383.

Lemieux T., Postsecondary education and increasing wage inequality, American Economic Review, 2006, vol. 96, p. 195.

Li G., Zhang C. H., et al., Linear regression with interval censored data, The Annals of Statistics, 1998, vol. 26, p. 1306.

Lin G., He X., Portnoy S., Quantile regression with doubly censored data, Computational Statistics & Data Analysis, 2012, vol. 56, p. 797.

Machado J. A., Mata J., Counterfactual decomposition of changes in wage distributions using quantile regression, Journal of applied Econometrics, 2005, vol. 20, p. 445.

McKeague I. W., Subramanian S., Sun Y., Median regression and the missing information principle, Journal of nonparametric statistics, 2001, vol. 13, p. 709.

Melly B., Decomposition of differences in distribution using quantile regression, Labour economics, 2005, vol. 12, p. 577.

Ou F. S., Zeng D., Cai J., Quantile regression models for current status data, Journal of statistical planning and inference, 2016, vol. 178, p. 112.

Pan W., Smooth estimation of the survival function for interval censored data, Statistics in Medicine, 2000, vol. 19, p. 2611.

Parzen M., Wei L., Ying Z., A resampling method based on pivotal estimating functions, Biometrika, 1994, vol. 81, p. 341.

Pollard D., Empirical processes: theory and applications. In NSF-CBMS regional conference series in probability and statistics , 1990, p. i.

Portnoy S., Censored regression quantiles, Journal of the American Statistical Association, 2003, vol. 98, p. 1001.

Portnoy S., Koenker R., et al., The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, Statistical Science, 1997, vol. 12, p. 279.

Powell J. L., Least absolute deviations estimation for the censored regression model, Journal of Econometrics, 1984, vol. 25, p. 303.

Powell J. L., Censored regression quantiles, Journal of econometrics, 1986, vol. 32, p. 143.

Rao C., Zhao L., Asymptotic normality of LAD estimator in censored regression models, Mathematical methods of statistics, 1993, vol. 2, p. 228.

Rao C. R., Zhao L., Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap, Sankhyā: The Indian Journal of Statistics, Series A, 1992, pp 323–331.

Rasteiro L. R., Regressão quantílica para dados censurados, Universidade de São Paulo, 2017, Dissertação de Mestrado

Ren J. J., Gu M., et al., Regression M-estimators with doubly censored data, The Annals of Statistics, 1997, vol. 25, p. 2638.

Rodrigues T., Dortet-Bernadet J.-L., Fan Y., Pyramid quantile regression, arXiv preprint arXiv:1606.05407, 2016

Santos B. R. d., Modelos de regressão quantílica, Universidade de São Paulo, 2012, Dissertação de Mestrado

Scharf F. S., Juanes F., Sutherland M., Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques, Ecology, 1998, vol. 79, p. 448.

Silva A. L., Estimação da função quantílica para dados com censura intervalar, Universidade de Brasília, 2011, Dissertação de Mestrado

Silverman B. W., Density estimation for statistics and data analysis. Routledge, 1986.

Turnbull B. W., Nonparametric estimation of a survivorship function with doubly censored data, Journal of the American statistical association, 1974, vol. 69, p. 169.

Turnbull B. W., The empirical distribution function with arbitrarily grouped, censored and truncated data, Journal of the Royal Statistical Society. Series B (Methodological), 1976, pp 290–295.

Van Buuren S., Flexible imputation of missing data. Chapman and Hall/CRC, 2012

Vanobbergen J., Martens L., Lesaffre E., Declerck D., The Signal-Tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results, European Journal of Paediatric Dentistry, 2000, vol. 2, p. 87

Wang H. J., Wang L., Locally weighted censored quantile regression, Journal of the American Statistical Association, 2009, vol. 104, p. 1117.

Wei Y., Pere A., Koenker R., He X., Quantile regression methods for reference growth charts, Statistics in medicine, 2006, vol. 25, p. 1369.

Xiuqing Z., Jinde W., LAD estimation for nonlinear regression models with randomly censored data, Science in China Series A: Mathematics, 2005, vol. 48, p. 880.

Ying Z., Jung S. H., Wei L. J., Survival analysis with median regression models, Journal of the American Statistical Association, 1995, vol. 90, p. 178.

Zhang C. H., Li X., et al., Linear regression with doubly censored data, The Annals of Statistics, 1996, vol. 24, p. 2720.

Zhou X., Feng Y., Du X., Quantile regression for interval censored data, Communications in Statistics-Theory and Methods, 2017, vol. 46, p. 3848.

Apêndice

Apêndice A

Gráficos do estudo de simulação para comparação dos métodos

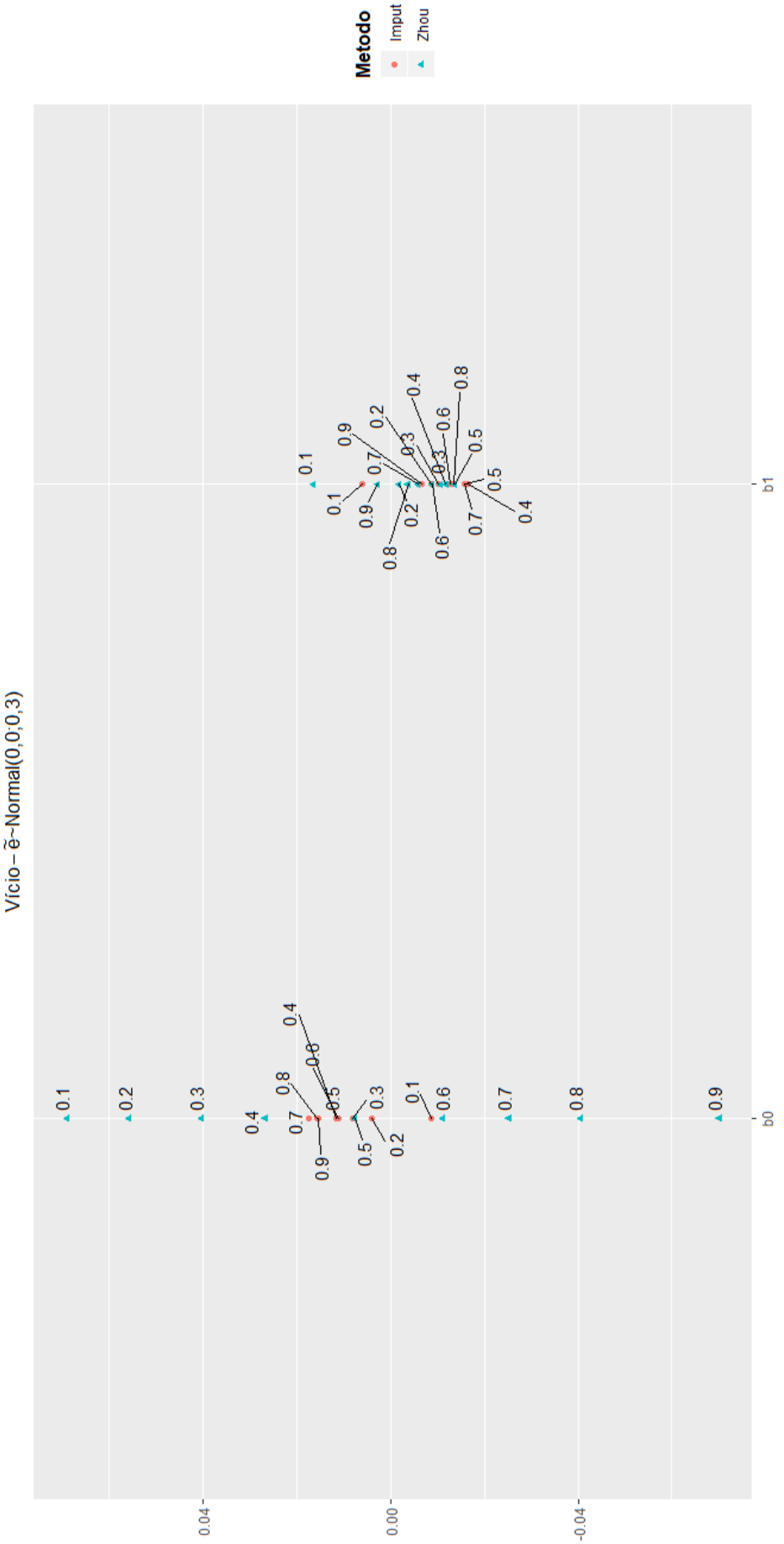

Figura A.1: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Normal$(0,0\ ;\ 0,3)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

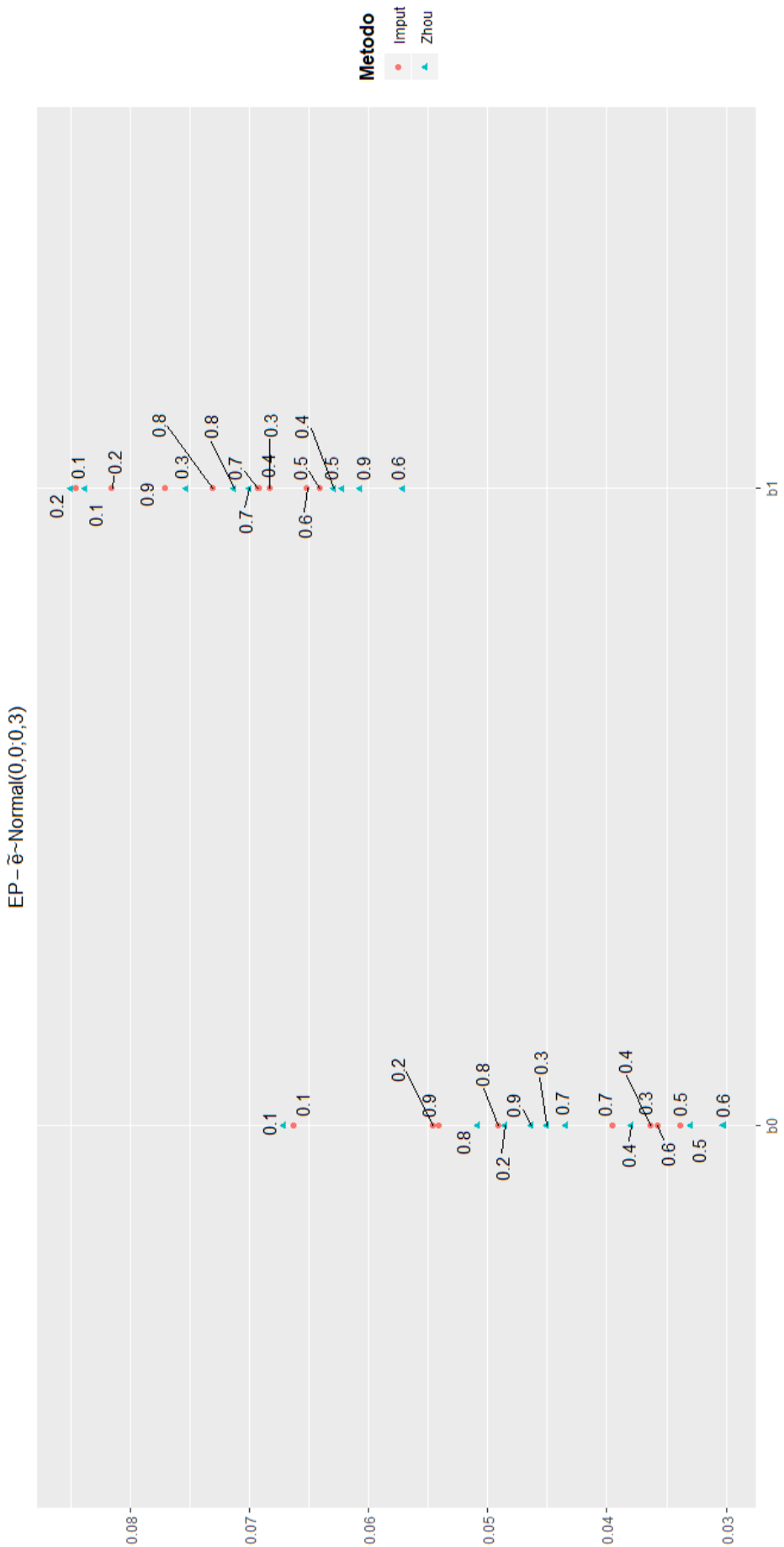

Figura A.2: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Normal$(0,0\ ;\ 0,3)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

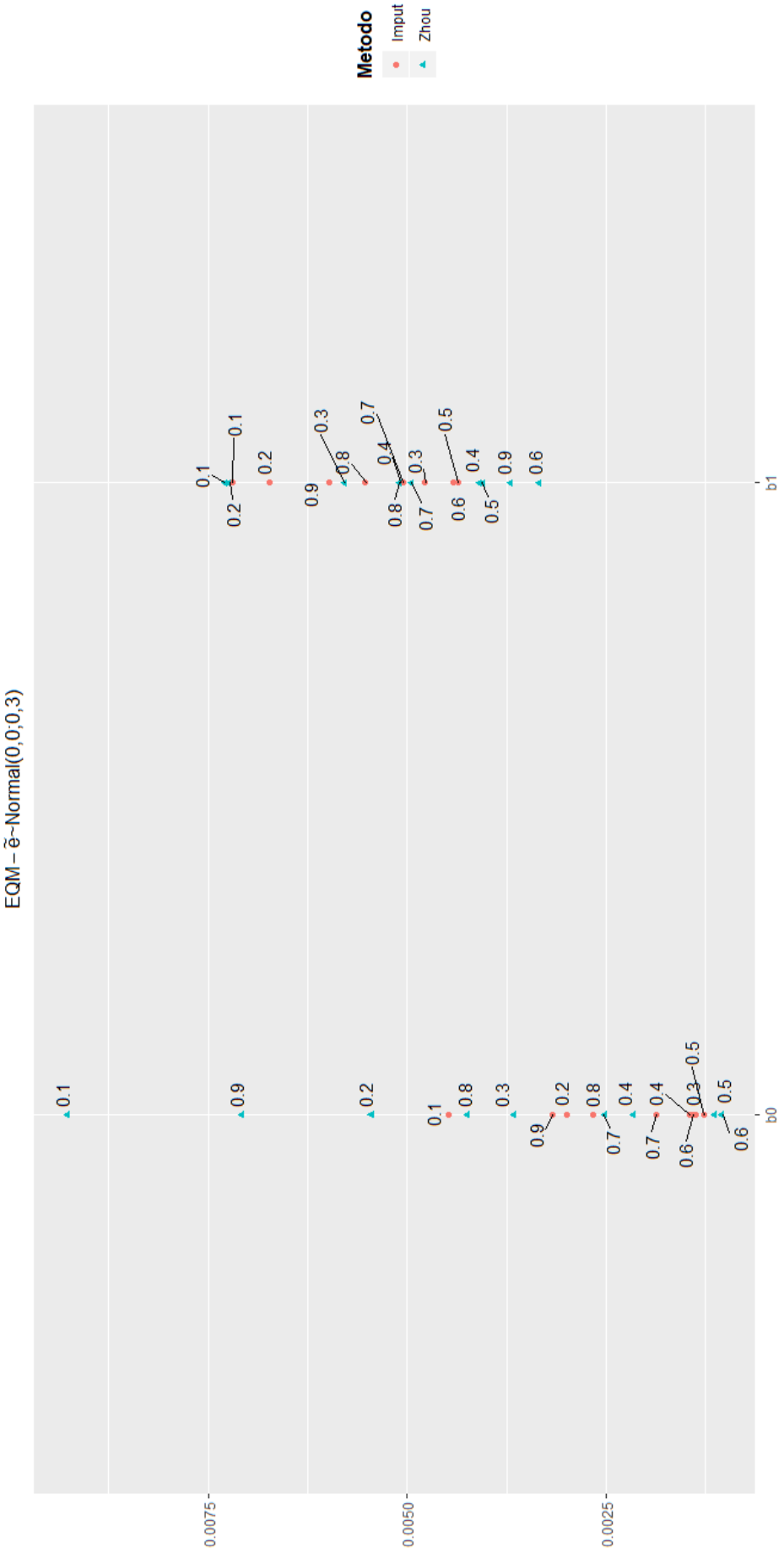

Figura A.3: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Normal$(0,0\ ;\ 0,3)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

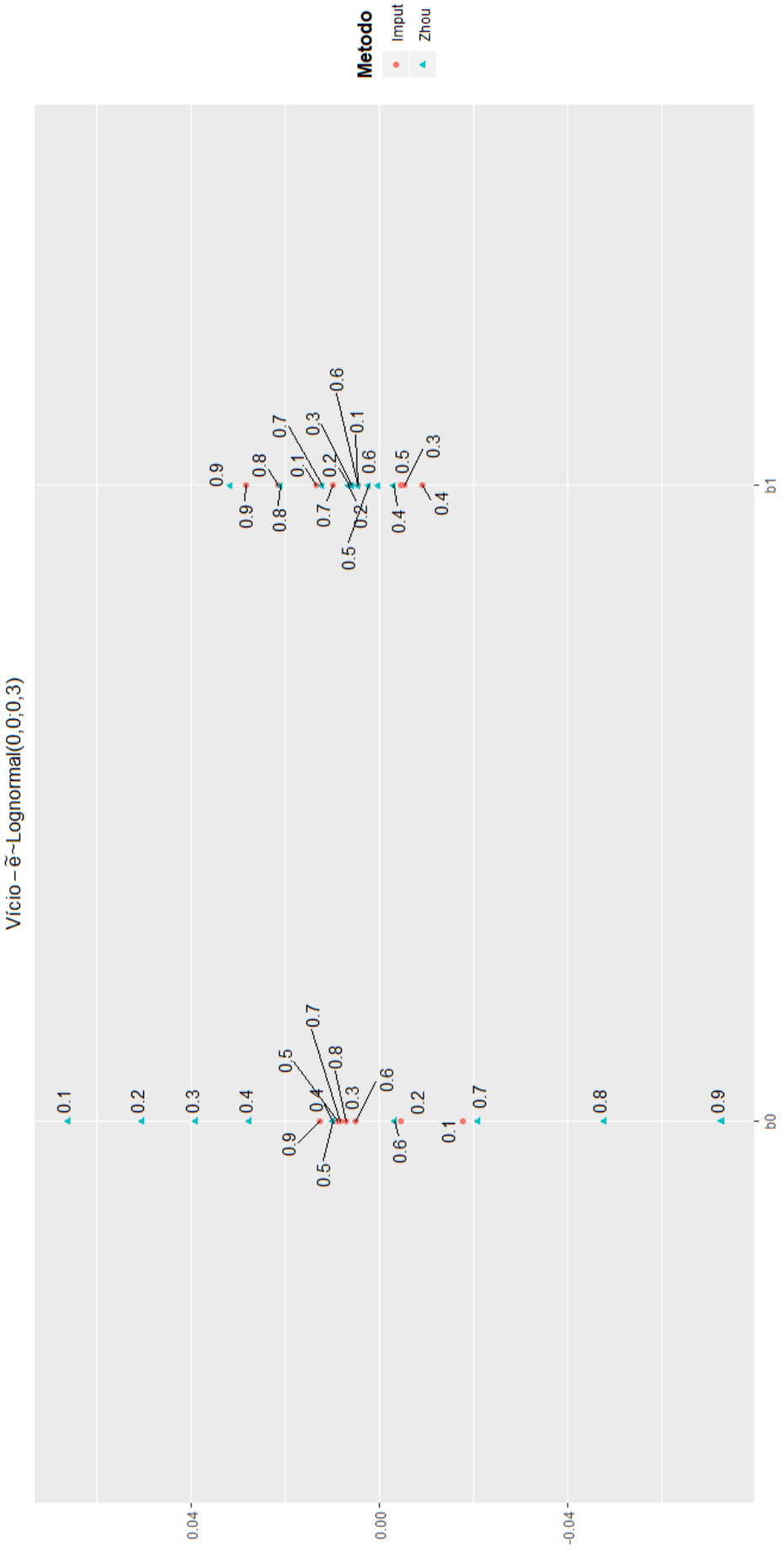

Figura A.4: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Lognormal$(0,0\ ;\ 0,3)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

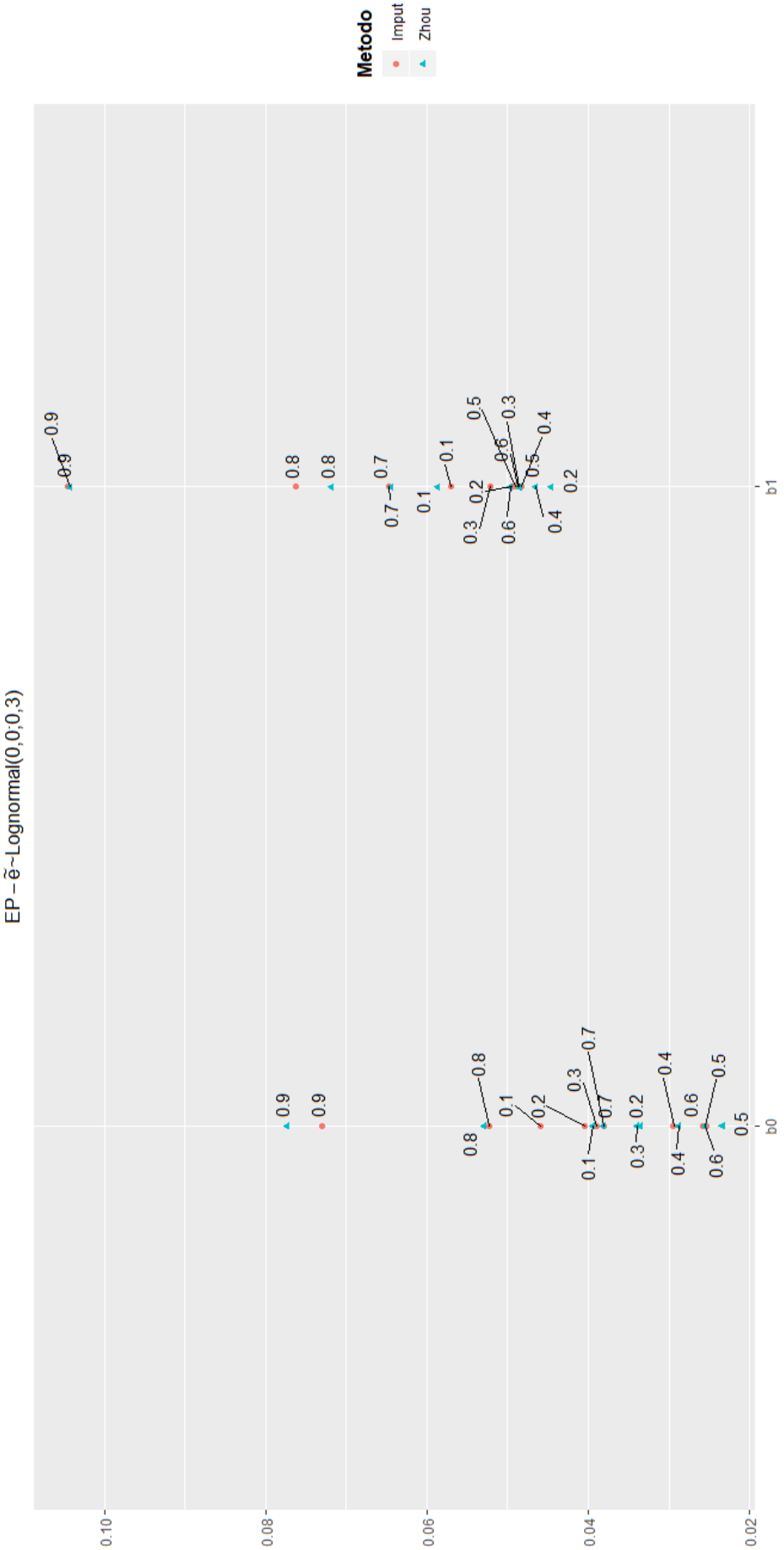

Figura A.5: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Lognormal$(0,0\ ;\ 0,3)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

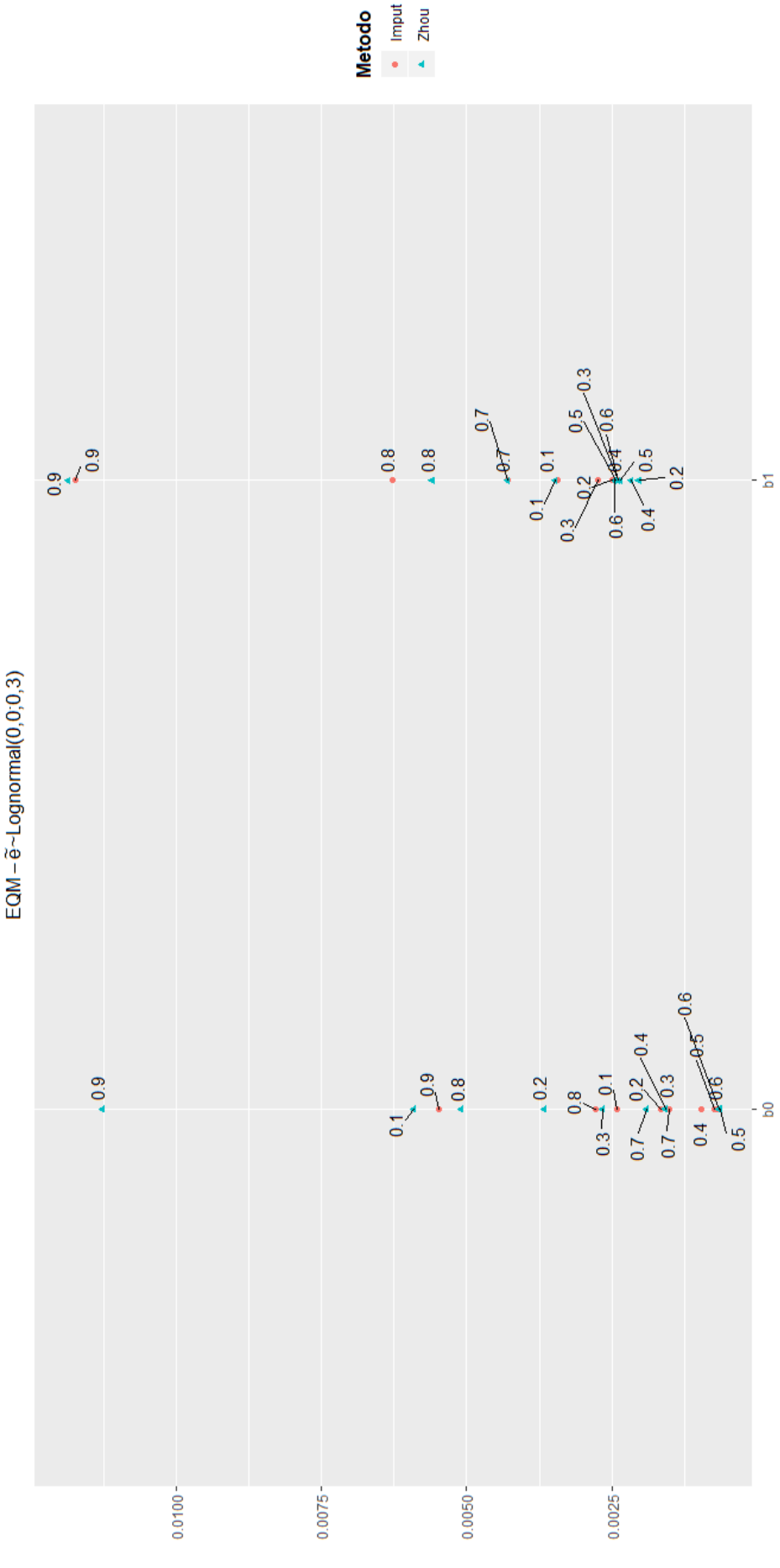

Figura A.6: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Lognormal$(0,0\ ;\ 0,3)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

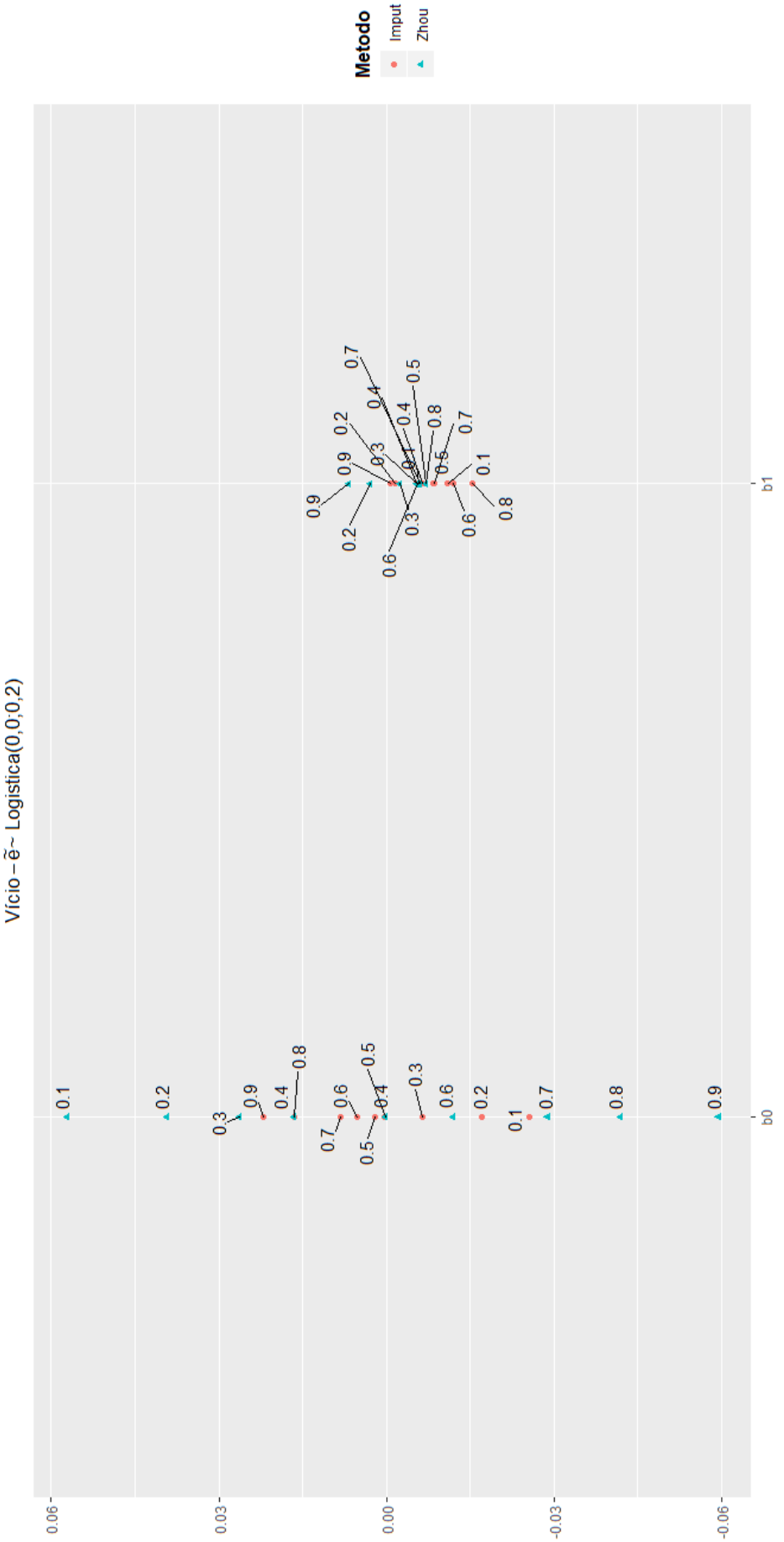

Figura A.7: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Logística$(0,0\ ;\ 0,2)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

Figura A.8: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Logística$(0,0\ ;\ 0,2)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

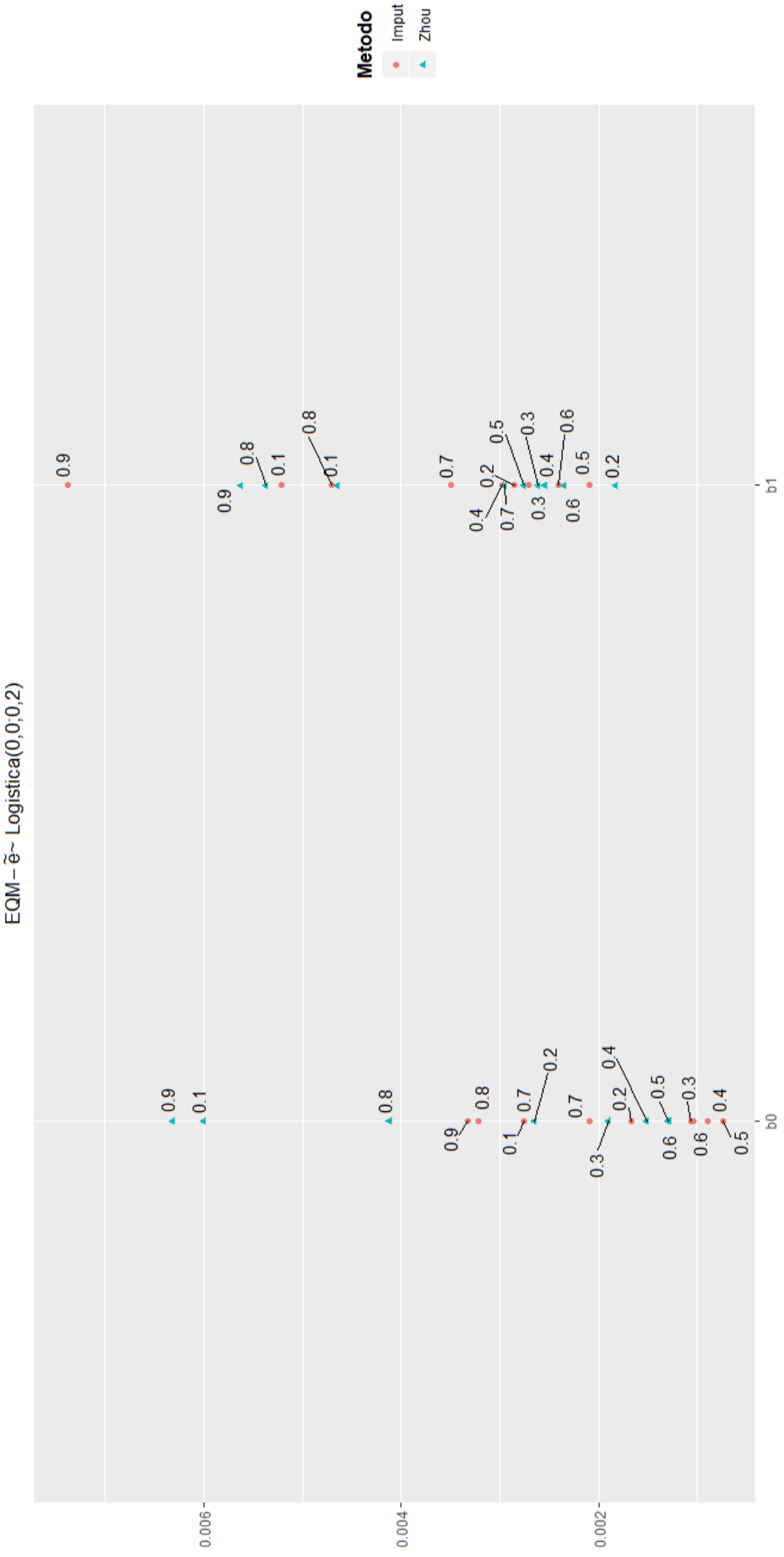

Figura A.9: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Logística$(0,0\ ;\ 0,2)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

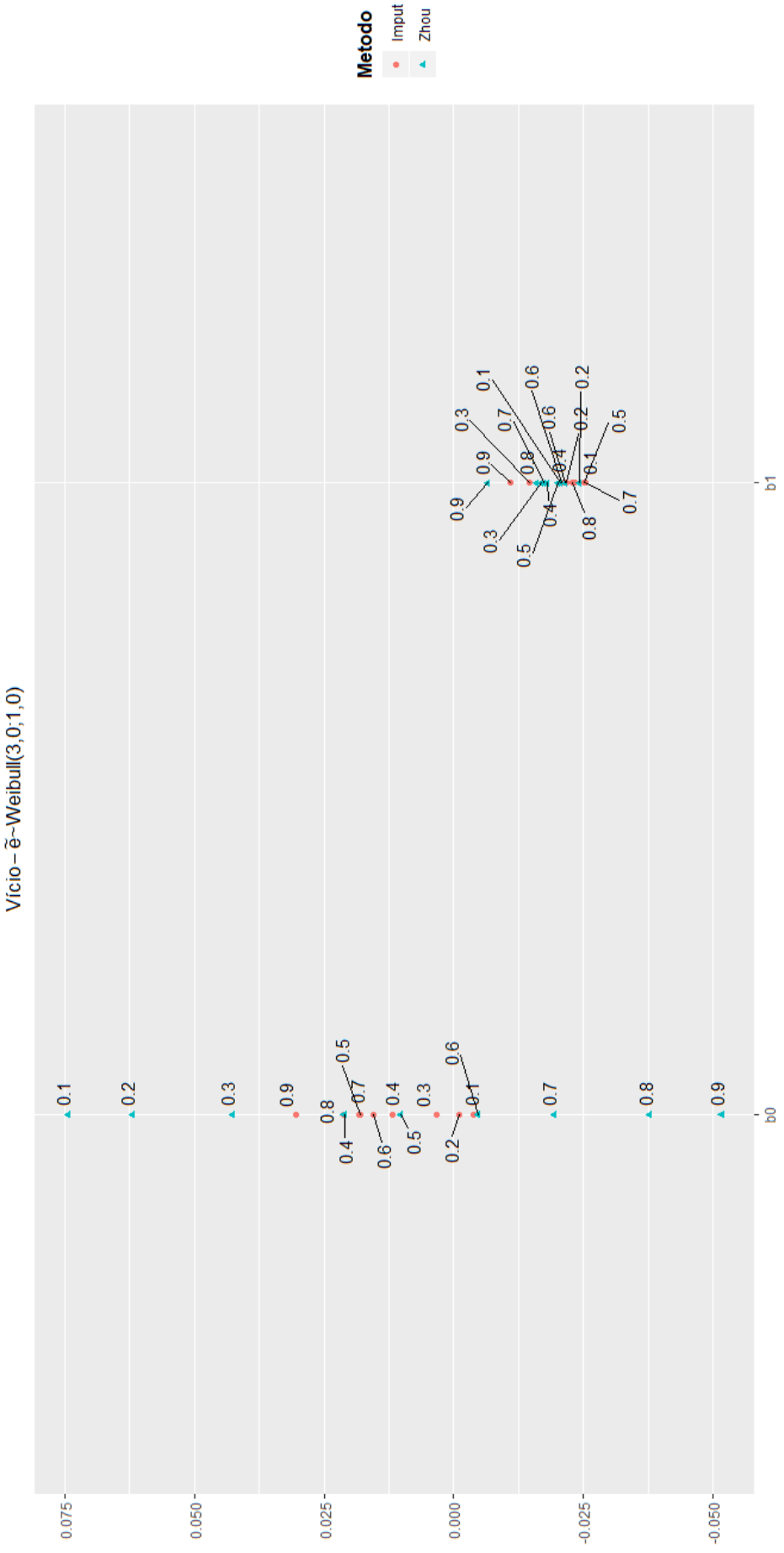

Figura A.10: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Weibull$(3,0\ ;\ 1,0)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

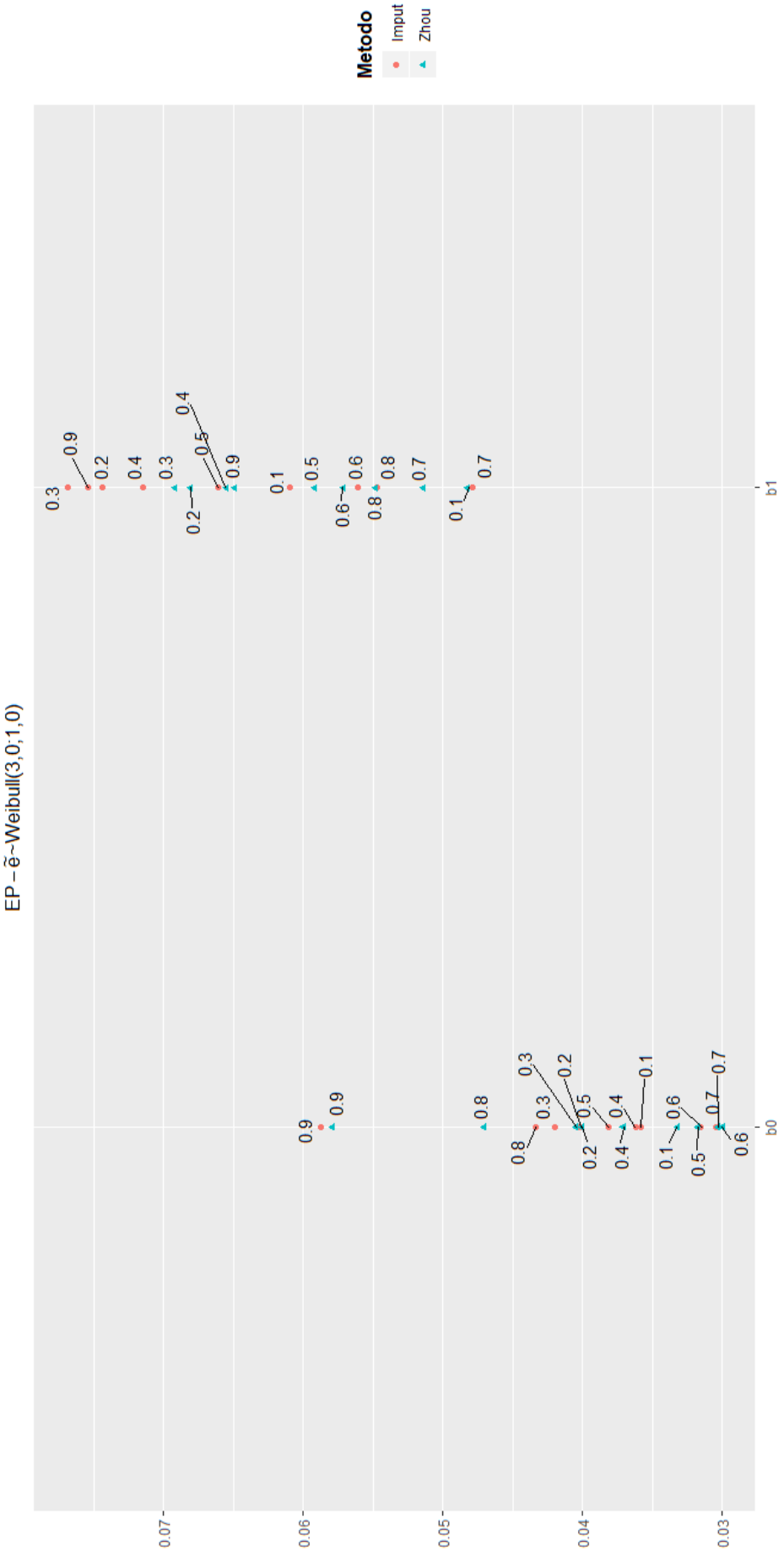

Figura A.11: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Weibull$(3, 0\ ;\ 1, 0)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

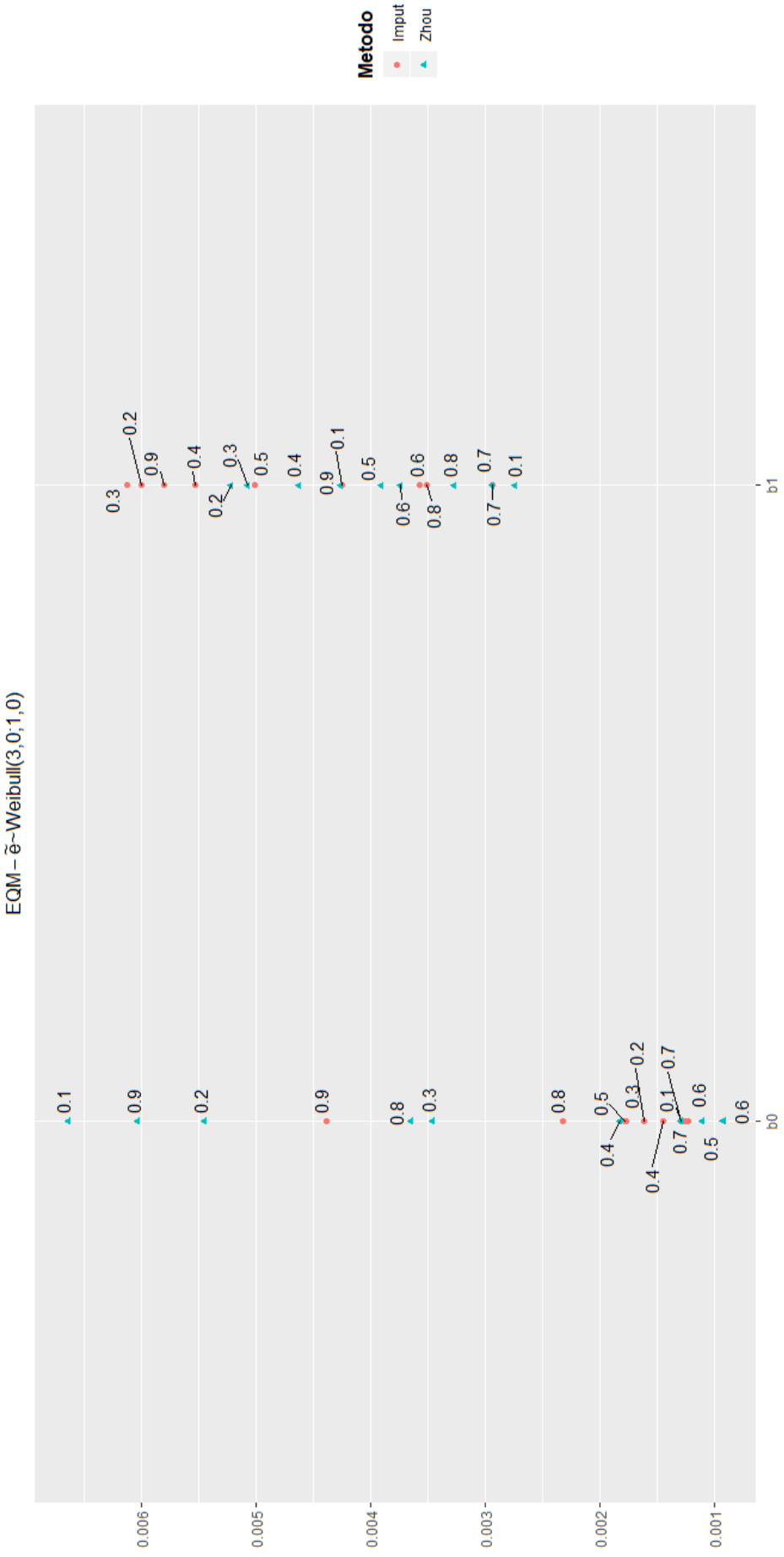

Figura A.12: Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra (n = 200), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim$ Weibull(3,0 ; 1,0): EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1, apresentando os seus respectivos quantis de ordem τ.

Apêndice B

Gráficos da aplicação das metodologias a dados empíricos

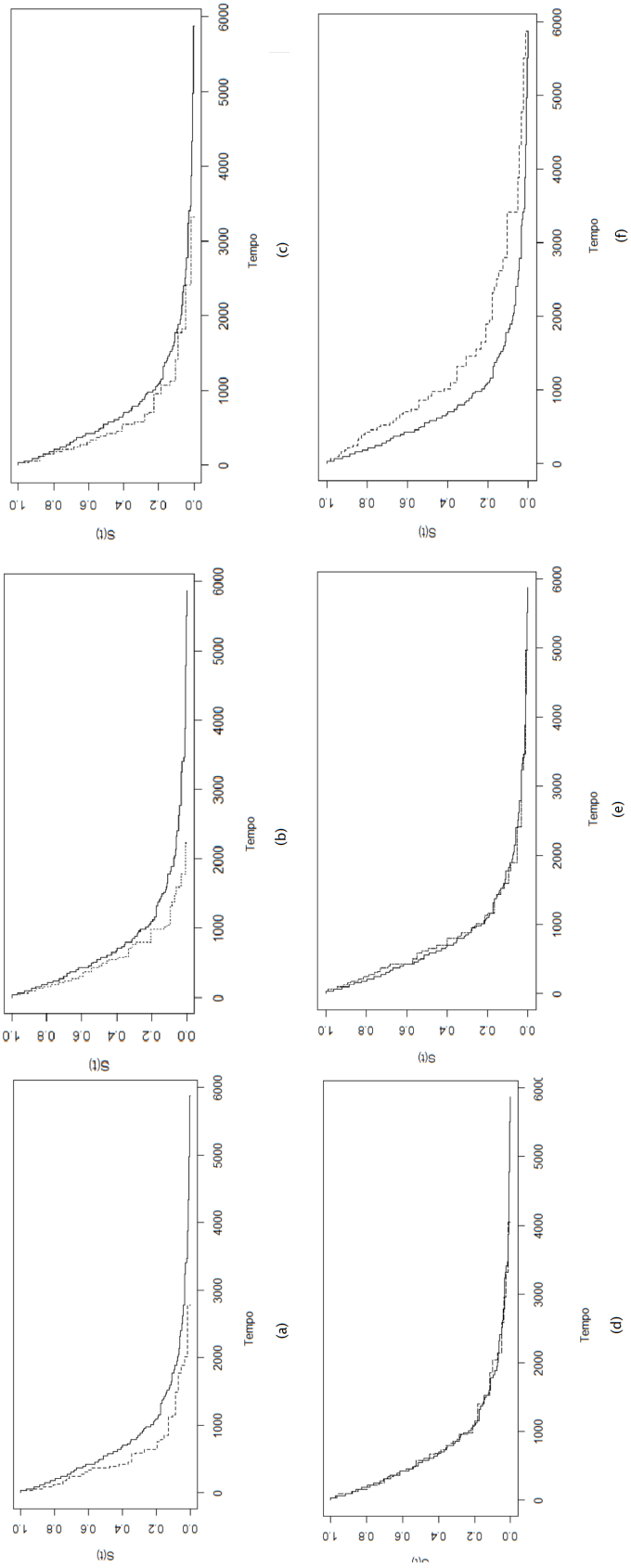

Figura B.1: Gráficos das estimativas não paramétricas da curva de sobrevivência: (a) Curva de sobrevivência global e da idade 20 anos (pontilhada); (b) Curva de sobrevivência global e da idade 30 anos (pontilhada); (c) Curva de sobrevivência global e da idade 40 anos (pontilhada); (d) Curva de sobrevivência global e da idade 50 anos (pontilhada); (e) Curva de sobrevivência global e da idade 60 anos (pontilhada); (f) Curva de sobrevivência global e da idade 72 anos (pontilhada).

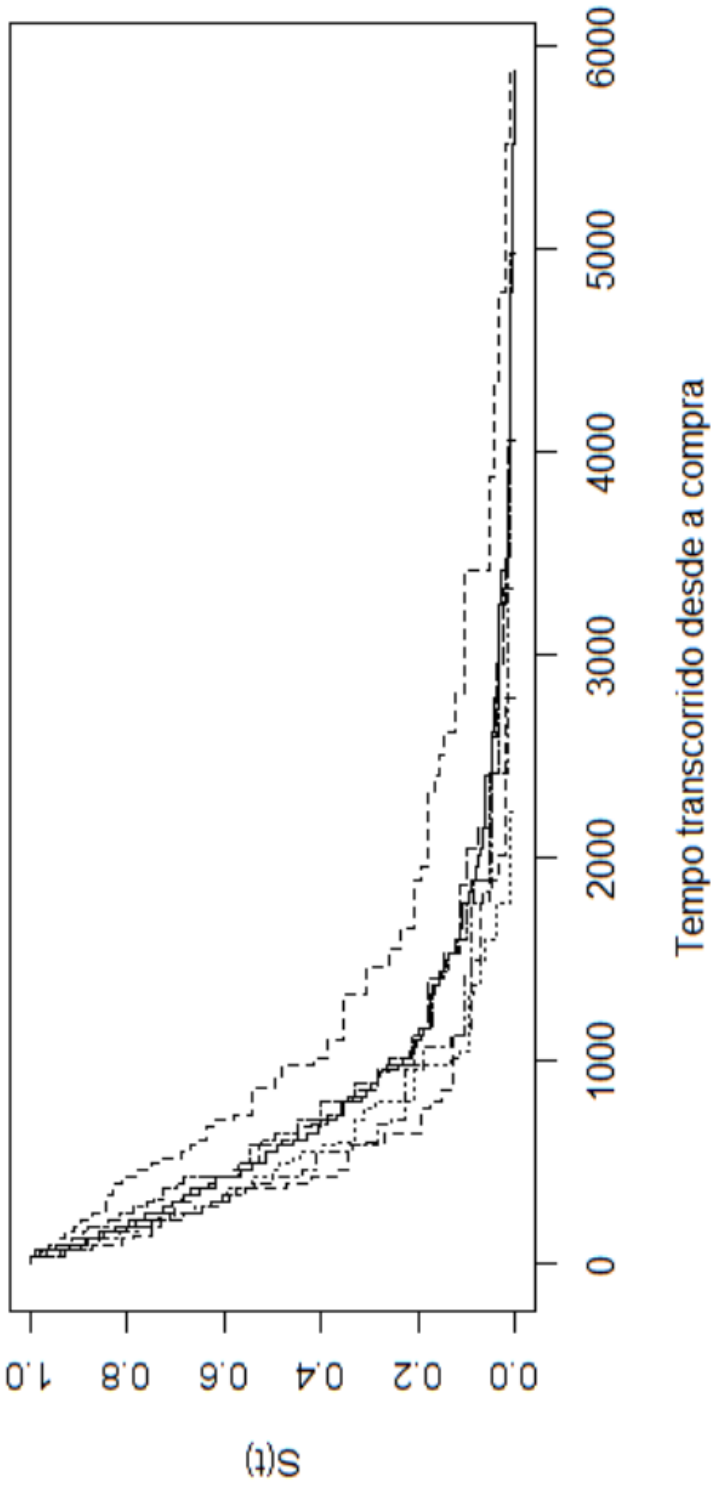

Figura B.2: Gráficos das estimativas não paramétricas da curva de sobrevivência global e das idades.

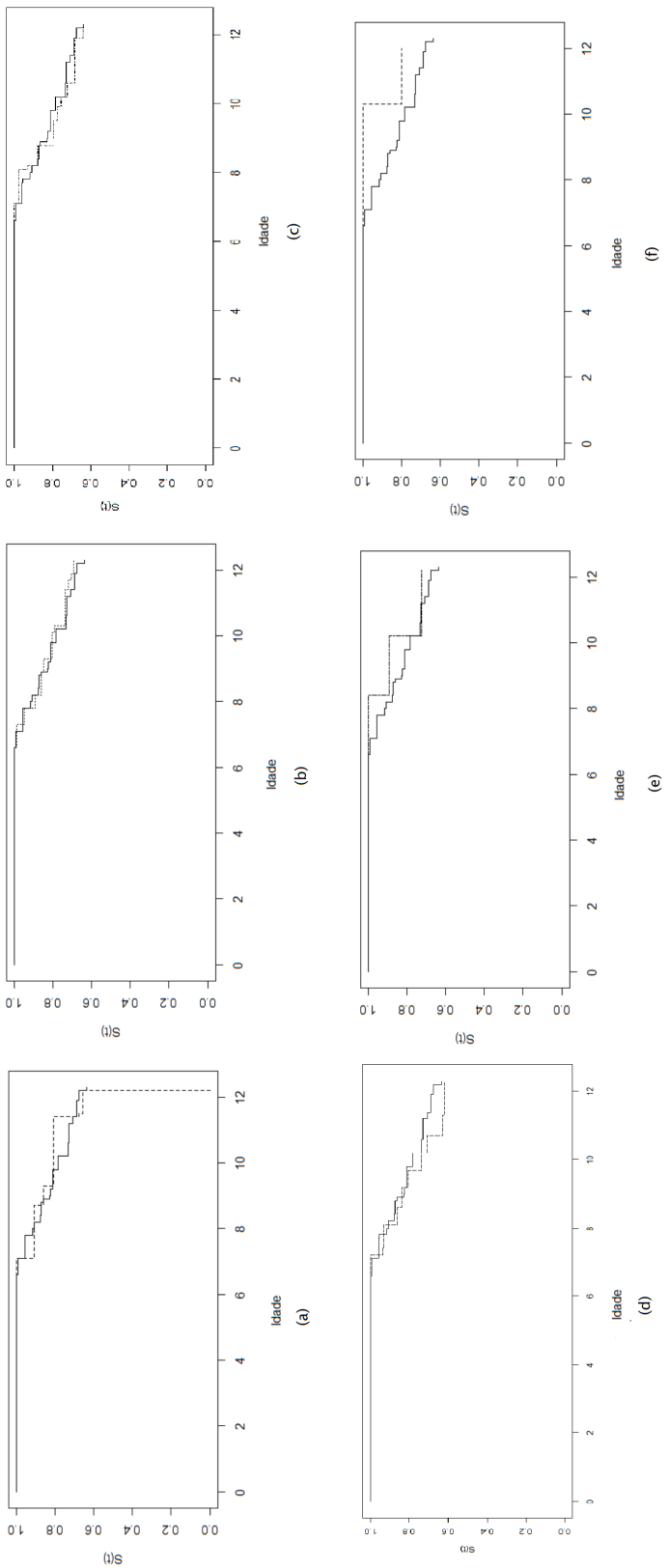

Figura B.3: Gráficos das estimativas não paramétricas da curva de sobrevivência: (a) Curva de sobrevivência global e do início de escovação com 0,5 anos (pontilhada); (b) Curva de sobrevivência global e do início de escovação com 1,5 anos (pontilhada); (c) Curva de sobrevivência global e do início de escovação com 2,5 anos (pontilhada); (d) Curva de sobrevivência global e do início de escovação com 3,5 anos (pontilhada); (e) Curva de sobrevivência global e do início de escovação com 4,5 anos (pontilhada); (f) Curva de sobrevivência global e do início de escovação com 5,5 anos (pontilhada).

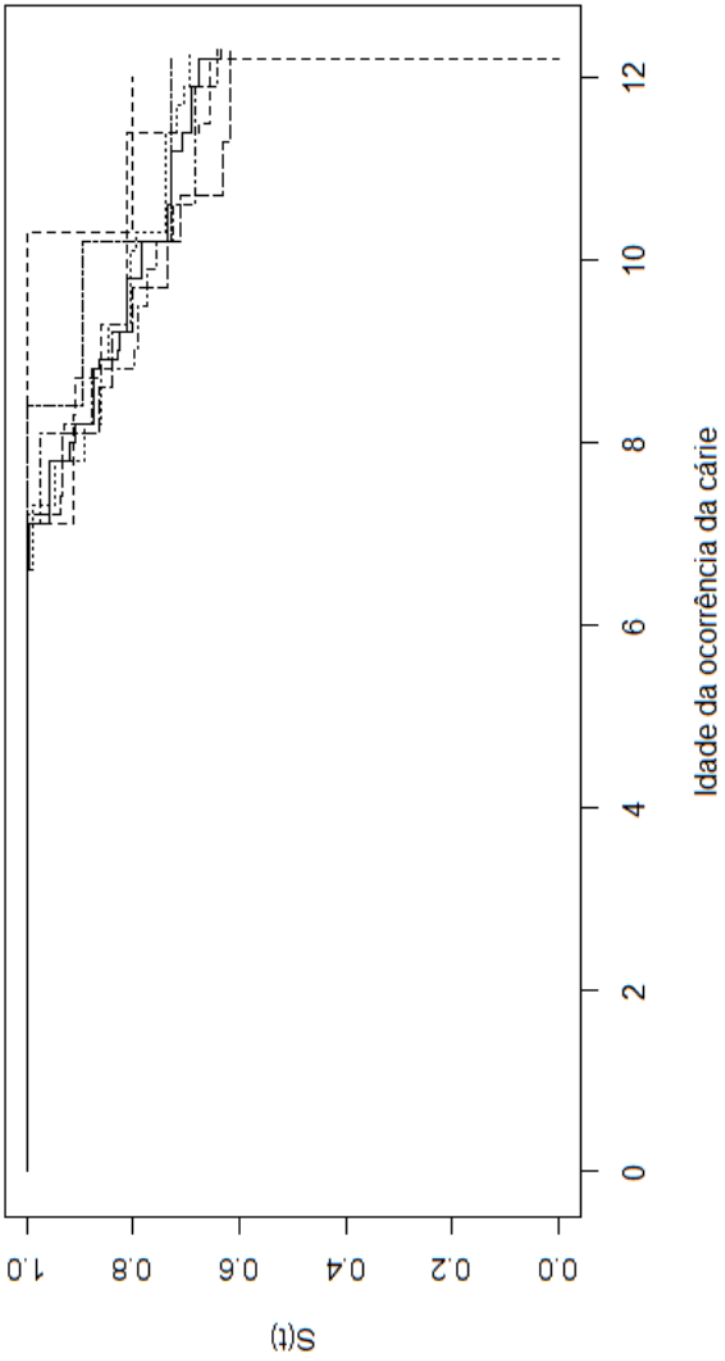

Figura B.4: Gráficos das estimativas não paramétricas da curva de sobrevivência global e das idades de início de escovação.