Regressão quantílica para dados censurados

Louise Rossi Rasteiro

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística
Orientador: Profª. Drª. Gisela Tunes da Silva

São Paulo, maio de 2017

Regressão quantílica para dados censurados

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Louise Rossi Rasteiro
e aprovada pela Comissão Julgadora.

Comissão Julgadora:

- Profª. Drª. Gisela Tunes da Silva (Presidente) - IME-USP
- Profª. Drª. Silvia Nagib Elian - IME-USP
- Prof. Dr. Antonio Eduardo Gomes - UnB

Agradecimentos

Agradeço à Profª Gisela Tunes da Silva, que desde os tempos de iniciação científica acreditou em mim e me orientou com tanta dedicação. Seus ensinamentos contribuíram para o meu crescimento e foram essenciais para que eu seguisse ao longo dessa jornada.

Agradeço aos meus pais, Albertino e Concettina, que nunca mediram esforços para que eu pudesse realizar os meus sonhos. Sem o apoio de vocês, nada disso seria possível.

Agradeço à minha irmã Lillian por ser o meu exemplo, pelas conversas e bons momentos.

Agradeço à Rayani e todos os amigos do trabalho por me apoiarem e permitirem que eu chegasse até aqui.

Sou profundamente grata a todos os amigos da graduação e pós-graduação que me ajudaram em todos esses anos. Em especial, agradeço à minha amiga Yanfei, pelas conversas e troca de experiências. Ao Victor, meu caro colega, por compartilhar as suas notas de aula, e à Elizabeth, por despender seu tempo para me auxiliar.

Agradeço também à minha amiga Aline, que sempre me apoiou e me motivou a seguir os meus objetivos.

Agradeço aos professores Silvia Nagib Elian e Antonio Eduardo Gomes por terem aceitado o convite para participar da banca e pelas sugestões e correções desta dissertação.

Por fim, agradeço ao meu namorado, Thomaz, pela paciência infinita, pelos conselhos e ajuda valiosa, e por me fazer sempre tão feliz. Sem dúvida, ter você ao meu lado tornou esse processo muito menos árduo.

Resumo

RASTEIRO, L. R. **Regressão quantílica para dados censurados**. 2017. 91 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

A regressão quantílica para dados censurados é uma extensão dos modelos de regressão quantílica que, por levar em consideração a informação das observações censuradas na modelagem, e por apresentar propriedades bastante satisfatórias, pode ser vista como uma abordagem complementar às metodologias tradicionais em Análise de Sobrevivência, com a vantagem de permitir que as conclusões inferenciais sejam tomadas facilmente em relação aos tempos de sobrevivência propriamente ditos, e não em relação à taxa de riscos ou a uma função desse tempo. Além disso, em alguns casos, pode ser vista também como metodologia alternativa aos modelos clássicos quando as suposições destes são violadas ou quando os dados são heterogêneos. Apresentam-se nesta dissertação três técnicas para modelagem com regressão quantílica para dados censurados, que se diferenciam em relação às suas suposições e forma de estimação dos parâmetros. Um estudo de simulação para comparação das três técnicas para dados com distribuição normal, Weibull e log-logística é apresentado, em que são avaliados viés, erro padrão e erro quadrático médio. São discutidas as vantagens e desvantagens de cada uma das técnicas e uma delas é aplicada a um conjunto de dados reais do Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

Palavras-chave: Regressão quantílica; Análise de Sobrevivência; Dados censurados; Estimador de Kaplan-Meier; *Kernel*; Árvore de Sobrevivência.

Abstract

RASTEIRO, L. R. **Censored quantile regression**. 2017. 91 s. Dissertation (Master degree) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Censored quantile regression is an extension of quantile regression, and because it incorporates information from censored data in the modelling, and presents quite satisfactory properties, this class of models can be seen as a complementary approach to the traditional methods in Survival Analysis, with the advantage of allowing inferential conclusions to be made easily in terms of survival times rather than in terms of risk rates or as functions of survival time. Moreover, in some cases, it can also be seen as an alternative methodology to the classical models when their assumptions are violated or when modelling heterogeneity of the data. This dissertation presents three techniques for modelling censored quantile regression, which differ by assumptions and parameter estimation method. A simulation study designed with normal, Weibull and loglogistic distribution is presented to evaluate bias, standard error and mean square error. The advantages and disadvantages of each of the three techniques are then discussed and one of them is applied to a real data set from the Heart Institute of Hospital das Clínicas, University of São Paulo.

Keywords: Quantile regression; Survival Analysis; Censored data; Kaplan-Meier estimator; *Kernel*; Survival Tree.

Sumário

1 Introdução **1**
1.1 Revisão Bibliográfica 2
1.2 Objetivos e Organização do Trabalho 5

2 Uma Revisão sobre Regressão Quantílica **7**
2.1 Definição de Quantil 8
2.2 Regressão Quantílica 11
2.2.1 Propriedades e Inferência 14

3 Regressão Quantílica na Presença de Censura **21**
3.1 Método Recursivo 24
3.1.1 Esquema de Ponderação via Kaplan-Meier 24
3.1.2 Algoritmo de Portnoy, 2003 26
3.2 Abordagens de Pesos Locais 33
3.2.1 Pesos Estimados via função *Kernel* 34
3.2.2 Pesos Estimados via Árvores de Sobrevivência 37

4 Estudo de Simulação para Comparação das Metodologias para Dados Censurados **41**
4.1 Linearização dos Principais Modelos Paramétricos em Análise de Sobrevivência 44

5 Aplicação a Dados Clínicos **51**
5.1 Análises Inferenciais 53

6 Discussão e Considerações Finais **63**

A Estimadores de Densidade *Kernel* **65**

B Árvore de Sobrevivência **71**

C Gráficos do Estudo de Simulação **75**

Referências Bibliográficas **89**

Capítulo 1

Introdução

Um dos grandes interesses em Análise de Sobrevivência é estudar o efeito que uma ou mais covariáveis exerce sobre o tempo até a ocorrência de um evento especificado. Problemas desse tipo não podem ser modelados com técnicas usuais de regressão, uma vez que os dados de sobrevivência, em geral, caracterizam-se pela presença de observações censuradas. Tradicionalmente esse tipo de análise é feito via modelo de riscos proporcionais de Cox, em que as covariáveis são incluídas na taxa de falha. No entanto, em algumas situações, a abordagem de Cox pode não ser tão interessante, já que as conclusões inferenciais também são dadas em termos das taxas de falha, e não diretamente sobre o tempo de sobrevivência. Além disso, a suposição de riscos proporcionais pode não ser verdadeira, e neste caso outras metodologias devem ser estudadas.

A motivação desta dissertação parte de um conjunto de dados reais do Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, em que o objetivo era estudar o tempo de sobrevivência de pacientes com insuficiência cardíaca, sujeito à censura, em função de uma série de covariáveis, em especial, da taxa de linfócitos. O conjunto de dados foi analisado no Centro de Estatística Aplicada do Instituto de Matemática e Estatística da Universidade de São Paulo (CEA IME-USP) e encontra-se descrito em Botter et al. (2012).

Na análise original desse conjunto de dados foi utilizado modelo de Cox para avaliar a relação das covariáveis com a probabilidade de sobrevida dos pacientes. Considere como objetivo secundário do estudo entender o efeito que as covariáveis exercem sobre o tempo de sobrevivência propriamente dito. Na abordagem de Cox as conclusões são feitas em relação à taxa de falha, e nesse sentido propôs-se aplicar a regressão quantílica para dados censurados, cuja técnica de modelagem permite facilmente interpretações em termos dos tempos de sobrevivência.

A rigor, conforme será discutido, para esse conjunto de dados composto por 3.139 observações, e com aproximadamente 56% de censura, o teste global para proporcionalidade dos riscos revelou evidência à violação da suposição do modelo de riscos proporcionais de Cox. Ou seja, tem-se mais uma razão para busca de uma metodologia alternativa ao modelo tradicional para análise desses dados.

Existem na literatura algumas abordagens alternativas ao modelo de Cox. Uma delas são os modelos de vida acelerados, que estão descritos em Colosimo e Giolo (2006), por exemplo. A metodologia, no entanto, esbarra em duas grandes dificuldades. Em primeiro lugar, para ajustar um modelo de vida acelerado é necessário o conhecimento acerca da distribuição dos dados que, na prática, é desconhecida. Em segundo lugar, alguns modelos precisam ser ajustados em função do logaritmo da variável resposta, mas as conclusões na escala original muitas vezes não seguem diretamente, e por isso podem ser comprometidas.

Os modelos de regressão quantílica para dados censurados podem ser vistos como um complemento valioso aos modelos de riscos proporcionais de Cox e modelos de vida acelerados, ou mesmo uma alternativa a essas análises tradicionais. Tratam-se de extensões da metodologia de regressão quantílica para o caso em que a variável resposta pode estar sujeita à censura, cuja ideia geral é ajustar uma regressão para cada quantil condicional de interesse da variável resposta em função das covariáveis. Ao se ajustar um modelo para cada quantil condicional, admite-se que as conclusões possam ser diferentes ao longo da distribuição da variável dependente, conferindo um conhecimento mais abrangente no estudo da variabilidade dos dados.

A regressão quantílica tem algumas boas propriedades que justificam sua aplicação em muitas situações práticas. Por exemplo, os modelos são invariantes sob transformações monótonas. Neste caso, ao contrário dos modelos de vida acelerados, a interpretação dos parâmetros não é comprometida após transformação logarítmica, por exemplo, e é dada diretamente. Além disso, a técnica é extremante útil na análise de dados heterocedásticos, ou ainda no estudo de populações não homogêneas, além de ser robusta a presença de pontos atípicos (*outliers*).

Existem na literatura diversas técnicas para ajustar modelos de regressão quantílica para dados censurados. Apresentam-se nessa dissertação três dessas técnicas. A primeira metodologia para dados censurados que será apresentada é a de Portnoy (2003), estudo pioneiro ao tratar de censura aleatória nos modelos. Trata-se de uma técnica recursiva que assume que a linearidade entre a variável resposta e as covariáveis é verdadeira em todos os quantis condicionais. No entanto, conforme será discutido, essa suposição pode não ser verificada em geral. Então, como alternativa ao modelo recursivo de Portnoy (2003), propôs-se estudar metodologias de pesos locais, que são mais flexíveis no que diz respeito à suposição de linearidade global dos modelos. Nesse sentido, apresentam-se as técnicas de Wang e Wang (2009) e Wey et al. (2014), bastante parecidas entre si, mas com algumas particularidades no processo de estimação dos pesos locais.

1.1 Revisão Bibliográfica

A regressão quantílica foi introduzida originalmente por Koenker e Bassett Jr (1978). Em seu artigo, os autores propõem a classe de modelos no contexto linear, apresentando-a como uma alternativa mais robusta ao método de Estimação de Mínimos Quadrados. Destacam ainda a sua eficiência na estimação dos parâmetros, sobretudo para os casos em que os erros não seguem distribuição da família Gaussiana. A metodologia já foi estudada por diversos autores,

que contribuíram para seu desenvolvimento teórico, progresso nos algoritmos computacionais e aplicações práticas.

Uma importante contribuição foi dada por Powell (1986a). O autor foi o pioneiro ao estender a regressão de mínimos erros absolutos no caso censurado, em Powell (1984), à metodologia de regressão quantílica. Trabalhar com dados censurados permitiu a introdução da técnica em Análise de Sobrevivência, cujos dados, em geral, caracterizam-se por apresentar informação parcial ou incompleta na variável dependente, o tempo de sobrevivência.

Supondo restrições nos quantis condicionais dos erros do modelo, Powell (1986a) trata do caso de censura fixa (também conhecida como *censura de tipo I*), isto é, em que todas as observações não podem ser observadas após um tempo C conhecido. É possível estender a técnica para o caso em que cada observação tem um tempo de censura C_i correspondente, mas também seus valores devem ser conhecidos, inclusive para observações não censuradas. Além disso, a metodologia apresenta algumas complicações computacionais, principalmente para grandes amostras, por envolver a minimização de uma função não convexa nos parâmetros, apesar de apresentar propriedades assintóticas desejáveis sob condições de regularidade.

Vários autores dedicaram-se a encontrar abordagens alternativas na análise de regressão quantílica na presença de censura, computacionalmente mais eficientes ou que também apresentassem boas propriedades para pequenas amostras, por exemplo. Buchinsky e Hahn (1999), sob as mesmas suposições de Powell (1986a), propuseram um estimador cuja função objetivo é globalmente convexa, e que pode ser visto como solução de um problema de programação linear. Chen e Khan (2001) e Khan e Powell (2001) apresentam métodos de estimação em dois estágios, desenhados para superar o viés de estimação que surge em amostras finitas.

Outros autores partiram de cenários em que se têm suposições adicionais para desenvolvimento de suas técnicas. Powell (1986b) e Newey (1991), por exemplo, assumem que os erros seguem distribuição simétrica, ou que os dados têm observações truncadas. Duncan (1986), Honoré e Powell (1994), Horowitz (1986) e Moon (1989) partem de um outro pressuposto, de que os erros do modelo são independentes das variáveis explicativas.

No entanto, todas essas propostas tem uma forte limitação em muitas aplicações práticas, por tratarem apenas do caso de censura de tipo I. Nesse sentido, tem-se na literatura uma série de outros artigos que buscaram generalizar o tipo de censura, incluindo a censura aleatória nas análises.

Ying et al. (1995), por exemplo, propõem um procedimento de estimação semiparamétrico para analisar modelo de regressão para a mediana na presença de censura aleatória à direita. A metodologia, no entanto, assume que os tempos de sobrevivência T_i e os tempos de censura C_i são incondicionalmente independentes. Na prática isso significa que C_i não pode depender das covariáveis, o que não é observado em geral. Mais tarde, McKeague et al. (2001) sugerem um método de estimação para a mediana baseado no *missing information principle* (MIP) que, conforme demonstram, para covariáveis discretas é equivalente ao introduzido por Ying et al. (1995).

Várias outras metodologias foram propostas, com suposições bastante restritivas ou aplicações computacionalmente complexas. Lindgren (1997), por exemplo, apresenta um método

de estimação dos quantis condicionais baseado na técnica de Mínimos Erros Absolutos Ponderados, com pesos estimados não parametricamente via Kaplan Meier local para cada uma das observações. Como discutido em Portnoy (2003), o uso de regressão não paramétrica, no entanto, é computacionalmente inviável a menos que a dimensão dos dados seja pequena, ao passo que as conclusões estatisticamente relevantes requerem um número grande de observações na amostra.

Mais tarde, Yang (1999) desenvolveu uma metodologia baseada na criação de pesos para a função de risco e função de sobrevivência da regressão para a mediana. Porém o método envolve a resolução de equações não lineares complicadas, que nem sempre têm solução única, além de supor que os erros são independentes e identicamente distribuídos. Como outro exemplo, pode-se citar também o trabalho de Honoré et al. (2002), que apresenta uma generalização de Powell (1986a), permitindo o estudo de censura aleatória, mas que assume que T_i é independente de C_i e também do vetor de covariáveis $\mathbf{x}_i$.

Mais recentemente, Portnoy (2003) estabeleceu um método cujas suposições não exigem independência incondicional dos tempos de sobrevivência e censura. O método baseia-se na ideia de redistribuição de massa proposta por Efron (1967), em que a massa de observações censuradas, $P(T_i > C_i|C_i, X_i)$, é redistribuída às não censuradas à direita. Em seu trabalho, Portnoy compara o método com o tradicional modelo de Cox, e discute as vantagens da regressão quantílica na modelagem da heterogeneidade dos dados e também como abordagem natural quando o interesse primário reside nos tempos de sobrevivência.

A natureza recursiva do método de Portnoy (2003), no entanto, é complicada do ponto de vista assintótico e inferencial. Alternativamente, Peng e Huang (2008) desenvolveram um método também baseado na independência condicional dos tempos de censura, mas usando teoria de martingais. A metodologia proposta tem boas propriedades, não apresenta algumas das complicações computacionais de seu precursor.

Porém, ambas as abordagens, de Portnoy e Peng e Huang, assumem linearidade em todos os quantis condicionais da variável resposta, dada as covariáveis, o que pode ser bastante restritivo na prática. Como discutido em Wang e Wang (2009), na análise do tempo de vida de pacientes que tiveram infarto agudo do miocárdio, os quantis condicionais inferiores do tempo de sobrevivência correlacionados com a idade não seguem uma relação linear. Os autores propõem então um método mais flexível, que envolve a recente teoria de estimador M, e funções *kernel*.

A metodologia de Wang e Wang (2009), contudo, tem algumas limitações. Em primeiro lugar, a metodologia foi desenvolvida apenas para variáveis contínuas, e na prática, a presença de variáveis categóricas inviabiliza o uso da técnica. Outro problema decorre do uso do suavizador *kernel* nos modelos de regressão linear, que encontra algumas dificuldades mesmo com um número moderado de covariáveis. Um método alternativo é proposto por Wey et al. (2014) também com boas propriedades, que usa a técnica de particionamento recursivo.

De uma forma geral, a teoria de regressão quantílica é bastante recente, e em Análise de Sobrevivência vem sendo aplicada cada vez mais como alternativa ou complemento à tradicional metodologia de Cox. Na literatura, os artigos buscam generalizar e aprimorar cada

vez mais as técnicas, mas muito ainda pode ser explorado para a sua consolidação.

1.2 Objetivos e Organização do Trabalho

O objetivo principal desta dissertação é motivar o uso de modelos de regressão quantílica para dados censurados, comuns em estudos de Análise de Sobrevivência. Para isso, inicialmente no Capítulo 2 é apresentada uma introdução aos modelos de regressão quantílica na situação em que os dados não estão sujeitos à censura. Nesse capítulo são discutidas algumas propriedades e inferência para essa classe de modelos.

A extensão da regressão quantílica para o contexto em que os dados podem ser censurados é apresentada no Capítulo 3, que está dividido em duas seções. A primeira delas trata da abordagem recursiva para estimação dos parâmetros, segundo a metodologia de Portnoy (2003). A segunda, por sua vez, aborda a metodologia de pesos locais, em que são apresentados os trabalhos de Wang e Wang (2009) e Wey et al. (2014).

Na sequência, no Capítulo 4, é apresentado um estudo de simulação para a comparação das três metodologias em alguns contextos específicos. No Capítulo 5, apresenta-se a aplicação ao conjunto de dados reais do Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, referente ao tempo de sobrevida de pacientes diagnosticados com insuficiência cardíaca. Finalmente, no Capítulo 6 é apresentada uma discussão acerca dos modelos estudados nesta dissertação, em que são enfatizadas as suas vantagens e desvantagens sobre as técnicas usuais para análise de dados sujeitos à censura.

Capítulo 2

Uma Revisão sobre Regressão Quantílica

Entender e inferir sobre os efeitos causais de um conjunto $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$, de p covariáveis observadas em uma sequência $\{Y_i\}_n$ de n variáveis aleatórias de interesse são necessidades comuns a estudos de diversas áreas do conhecimento.

Os modelos de regressão como ferramenta Estatística têm, dessa forma, sua importância devidamente reconhecida, e são amplamente utilizados. Para sua construção, considera-se que cada variável Y_i é, na verdade, resultado da soma de uma parte sistemática, que é função de $\mathbf{x}_i$, e de um erro aleatório ε_i condicionalmente independente dado $\mathbf{x}_i$, $i = 1, \cdots, n$. Em outras palavras, pode-se escrever:

$$Y_i = \mu(\mathbf{x}_i) + \varepsilon_i \quad i = 1, \cdots, n.$$

Além disso, como suposições dos modelos, usualmente assume-se que $E(\varepsilon_i|\mathbf{x}_i) = 0$ e $Var(\varepsilon_i|\mathbf{x}_i) < \infty$.

Nos modelos de regressão mais usuais tem-se interesse em trabalhar com a esperança condicional de Y_i, dado o vetor de covariáveis $\mathbf{x}_i$. Observe que, sob as suposições do modelo, $E(Y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i)$. A escolha de $E(Y_i|\mathbf{x}_i)$ provém, principalmente, dos mesmos motivos que tornam a média a principal medida resumo de uma população ou conjunto de dados. Em particular, $E(Y_i|\mathbf{x}_i)$ apresenta boas propriedades, como linearidade, por exemplo, e é a função de $\mathbf{x}_i$ que minimiza o erro quadrático médio, isto é, $\min_{\mu(\mathbf{x}_i)} E[(Y_i - \mu(\mathbf{x}_i))^2]$, entre todas as funções $\mu(\mathbf{x}_i) : \mathbb{R}^p \to \mathbb{R}$.

Considere, por exemplo, a estrutura $\mu(\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. Ou seja, suponha que $E(Y_i|\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. A regressão linear, como é chamada, além de apresentar simples interpretação para o vetor $\boldsymbol{\beta} = (\beta_1, \cdots \beta_n)^T$, também é atraente do ponto de vista computacional na estimação destes parâmetros, o que contribuiu para o sucesso e sua consolidação como ferramenta de análise. Além disso, quando é possível incluir suposições de normalidade nos erros aleatórios, a regressão linear apresenta propriedades ainda mais satisfatórias, como a eficiência, por exemplo.

Porém, como destaca Koenker (2005), referenciando o trabalho de Mosteller e Tukey (1977), a média como única medida resumo raramente é suficiente na avaliação dos dados. Em geral, *boxplots*, gráfico de quantis, histogramas, medidas de assimetria e de curtose, por exemplo, tam-

bém devem ser avaliadas e apresentadas, ao menos como informação descritiva complementar, para que se tenha uma visão mais completa da variável em estudo.

Nesse sentido, assim como se buscam outras medidas, além da média, para análise descritiva dos dados, também se pode estudar outras abordagens estatísticas do ponto de vista inferencial. Os modelos de regressão quantílica, por exemplo, apresentam os efeitos causais das covariáveis sobre a resposta nos diferentes quantis da distribuição da mesma, e por isso podem ser vistos como uma abordagem alternativa à metodologia de regressão usual. Ou seja, enquanto os modelos clássicos se limitam à análise das médias condicionais, a regressão quantílica permite a análise ao longo de toda a distribuição condicional da variável resposta nas covariáveis.

Os modelos de regressão quantílica surgiram como uma generalização do método de minimização dos resíduos absolutos, desenvolvidos no início do século XIX. Conforme será discutido nas próximas seções, a regressão quantílica esbarrou durante muito tempo na dificuldade de estimação dos parâmetros que, ao contrário dos modelos de regressão lineares usuais, não tem fórmula analítica. Porém, com o advento dos computadores, e também desenvolvimento das técnicas de programação linear, a metodologia vem ganhando cada vez mais espaço nos estudos empíricos e pesquisas acadêmicas.

Na próxima seção apresentam-se as definições de quantil, que são norte para o entendimento da regressão quantílica propriamente dita.

2.1 Definição de Quantil

Seja Y uma variável aleatória com função de distribuição acumulada dada por $F_Y(\cdot)$. O quantil de ordem τ para Y é definido como:

Definição 1 *O quantil de ordem τ para Y, $\tau \in [0,1]$, é o menor valor y tal que $F_Y(y) = \tau$.*

Em outras palavras, o quantil de ordem τ, que será denotado por $Q_Y(\tau)$, pode ser visto como resultado da função inversa $F_Y^{-1}(\tau)$, de modo que:

$$Q_Y(\tau) = F_Y^{-1}(\tau) = inf\{y : F_Y(y) \geq \tau\}, \qquad \tau \in [0,1].$$

De acordo com as propriedades bastante conhecidas da função de distribuição acumulada e de sua função inversa, se $F_Y(\cdot)$ é estritamente crescente, então existe um único número real y tal que $F_Y(y) = \tau$.

Em estudos empíricos, no entanto, a função $F_Y(\cdot)$, não é, em geral, conhecida. Dessa forma, considere uma amostra aleatória $\{y_1, \cdots, y_n\}$ de tamanho n da variável Y. Tem-se então a seguinte definição:

Definição 2 *Uma estimativa para o quantil τ de Y é dada pelo menor valor y tal que:*

$$\hat{F}_Y(y) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(y_i \leq y)\right] \geq \tau, \qquad \mathcal{I}(y_i \leq y) = \begin{cases} 1, & y_i \leq y, \\ 0, & y_i > y, \end{cases}$$

em que $\hat{F}_Y(y)$ é uma estimativa para $F_Y(y)$.

As definições apresentadas anteriormente se baseiam no conceito de *ordenação* dos dados. Apesar de serem as mais usuais, não são as únicas formas de definir, respectivamente, o quantil populacional e amostral de ordem τ de Y, mas também é possível apresentá-las à luz de um problema de *otimização*, conforme descrito a seguir.

Considere a função de perda $\rho_\tau(Y - y)$, em que:

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

sendo que $\rho_\tau(u) \geq 0$, $\forall u$. Considerando Y variável aleatória contínua, observe que:

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

Derivando a expressão em relação à y e igualando-se a zero, obtém-se:

$$(1 - \tau)\int_{-\infty}^{y} dF_Y(t) - \tau\int_{y}^{+\infty} dF(t) = F_Y(y) - \tau = 0.$$

Como a função de distribuição acumulada é não decrescente, todo y em $\{y : F_Y(y) = \tau\}$ minimiza o valor esperado da função $\rho_\tau(Y - y)$. Portanto, de acordo com a Definição 1, $y = E[\rho_\tau(Y - y)]$ é quantil de ordem τ de Y. Assim, uma definição de quantil equivalente pode ser escrita como:

Definição 3 *O quantil de ordem τ para Y é dado por:*

$$\mathcal{Q}_Y(\tau) = arg\min_y E[\rho_\tau(Y - y)].$$

Considere, por exemplo, $\tau = 1/2$, probabilidade associada ao quantil denominado mediana de Y. Observe que:

$$E[\rho_{1/2}(Y - y)] = -\frac{1}{2}\int_{-\infty}^{y}(t - y)dF_Y(t) + \frac{1}{2}\int_{y}^{+\infty}(t - y)dF_Y(t) = \frac{1}{2}E|Y - y|.$$

Ou seja, minimizar $E[\rho_{1/2}(Y-y)]$ é equivalente a minimizar $E|Y-y|$. De um modo geral, o quantil definido de acordo com a Definição 3 pode ser visto como uma generalização do problema de minimizar a esperança dos resíduos absolutos resultantes ao usar y para predizer Y.

Observe que, de acordo com a Lei dos Grandes Números, para n suficientemente grande, tem-se que a média amostral da função $\rho_\tau(y_i - y)$ converge para o seu valor esperado. Dessa forma, pode-se escrever a seguinte definição:

Definição 4 *Uma estimativa consistente para o quantil de ordem τ de Y é dada pelo valor y que minimiza a soma:*

$$S_n(y) = \frac{1}{n}\sum_{i=1}^{n}[\rho_\tau(y_i - y)], n \to +\infty.$$

Para ilustrar o conceito de quantil e suas definições, considere uma amostra aleatória $y_1, \cdots, y_n$ com $n = 1.000$ observações da variável $Y \sim \mathcal{N}(0, 1)$, isto é, que segue uma distribuição normal de média zero e variância igual a um, e suponha que esta distribuição seja desconhecida e que o objetivo é estimar diferentes quantis da distribuição de Y.

Uma representação gráfica bastante utilizada nesse contexto é o *boxplot*, ou gráfico de caixa, que considera o conceito de ordenação dos dados em sua construção, e apresenta os quantis $0,25; 0,50$ e $0,75$. Ainda no contexto de ordenação dos dados, o gráfico de quantis é uma técnica gráfica alternativa e equivalente ao boxplot, mas que permite representar outros quantis de interesse. A Figura 2.1 (a) e 2.1 (b) apresenta, respectivamente, o *boxplot* e gráfico de quantis para o exemplo simples enunciado anteriormente.

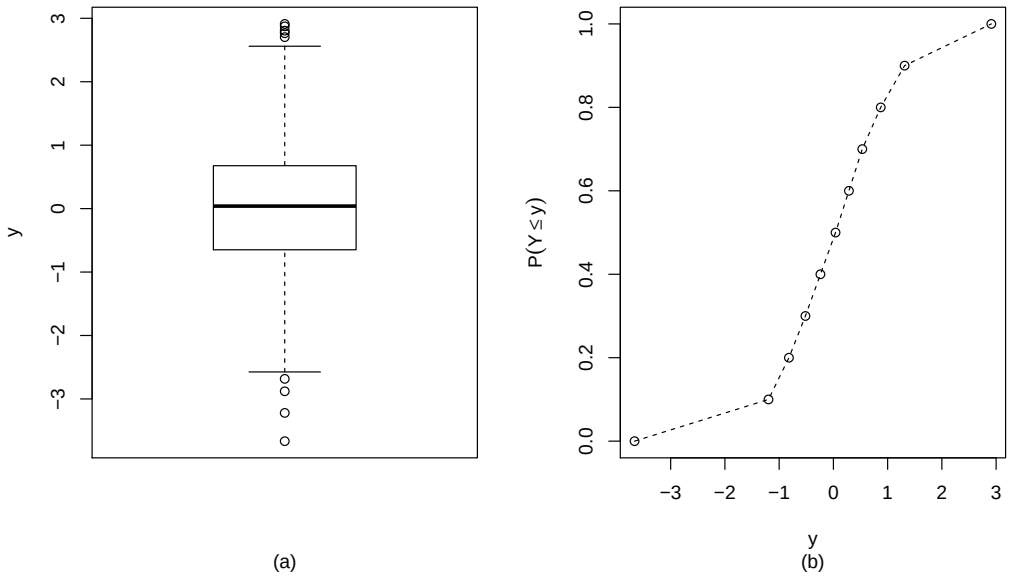

Figura 2.1: (a) Boxplot e (b) gráfico de quantis para uma amostra aleatória de 1.000 observações de $Y \sim N(0, 1)$. Os quantis estimados para $\tau = 0,25, 0,50$ e $0,75$, por exemplo, são, respectivamente, iguais a $-0,683$, $-0,024$ e $0,669$.

No caso da segunda definição de quantil, que considera um problema de otimização, o objetivo é encontrar $y \in \{y_1, \cdots, y_n\}$ que minimiza a função

$$S_n(y) = \sum_{i=1}^{n} [\rho_\tau(y_i - y)].$$

A Figura 2.2 traz o cálculo de $S_n(y)$, para $\tau \in (0,25; 0,50; 0,75)$. Observa-se na Figura 2.2 que o valor que torna $S_n(y)$ mínimo para $\tau = 0,5$, por exemplo, é $-0,024$, que coincide com o valor encontrado no contexto da ordenação dos dados.

A segunda definição de quantil será base para o entendimento de regressão quantílica, que é introduzida a seguir.

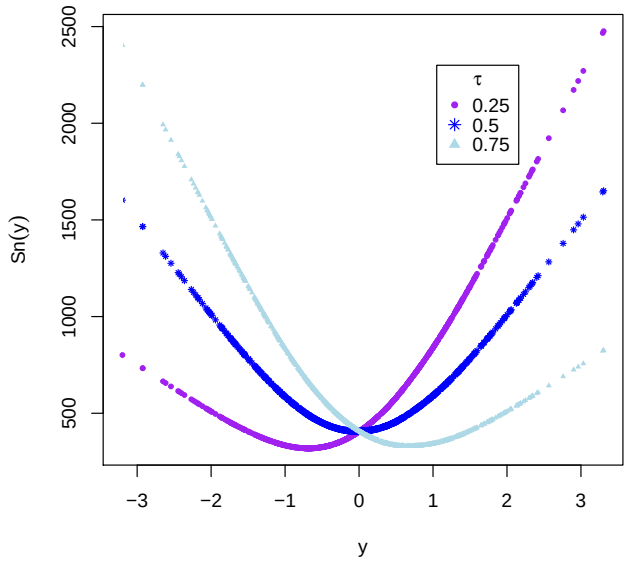

Figura 2.2: Gráfico de $S_n(\tau), \tau \in (0,25, 0,50, 0,75)$ para uma amostra de 1.000 observações de $Y \sim N(0,1)$. Os valores que y que minimizam as curvas S_n são, respectivamente iguais a $-0,683, -0,024$ e $0,669$, que coincidem com estimativas dos quantis encontradas no cálculo via problema de ordenação dos dados.

2.2 Regressão Quantílica

Sejam $Y_i, i = 1, \cdots, n$, variáveis aleatórias que são resposta do estudo. Em Análise de Sobrevivência, por exemplo, Y_i pode ser definida como o tempo até a ocorrência de um evento especificado, ou uma função desse tempo, para a unidade experimental i. Seja $\mathbf{x}_i \in \mathbb{R}^p$ vetor observado de covariáveis. Considere que as variáveis Y_i são condicionalmente independentes dado $\mathbf{x}_i, \forall i = 1, \cdots, n$.

Conforme discutido anteriormente, enquanto a regressão usual limita-se em descrever a relação de Y_i com as covariáveis do estudo em termos de médias condicionais, a regressão quantílica é uma técnica de modelagem estatística que permite analisar essa relação em qualquer quantil de ordem τ de interesse, $\tau \in [0, 1]$.

Em outras palavras, trata-se de uma metodologia capaz de descrever a função $f(\cdot, \tau)$ tal que

$$Q_{Y_i|\mathbf{x}_i}(\tau) = f(\mathbf{x}_i, \tau), \tag{2.1}$$

para todo $\tau \in [0, 1]$. A função $f(\cdot, \tau)$ é dita parte sistemática do modelo de regressão. Observe que a função $f(\cdot, \tau)$ pode ser diferente para cada τ.

Uma forma intuitiva de entender a regressão quantílica, e que é apresentada usualmente na literatura da área, é uma analogia aos modelos de regressão clássica (ver, por exemplo, Koenker (2005) e Santos (2012)). Neste caso, cada valor observado da variável resposta do estudo é dado pela soma de uma parte sistemática, que é quantil de ordem τ de Y_i, $f(\mathbf{x}_i, \tau)$, e

de um erro aleatório u_i. Isto é:

$$y_i = f(\mathbf{x}_i, \tau) + u_i, \tag{2.2}$$

com u_i independentes e identicamente distribuídas, $i = 1, \cdots, n$. Supondo-se que o quantil de ordem τ de u_i, condicional a $\mathbf{x}_i$, é igual a zero, observe que a função a ser modelada pode ser expressa como apresentado em (2.1). Essa forma de entender o modelo de regressão quantílica é importante para o desenvolvimento da teoria inferencial, que será discutida mais adiante.

No entanto, conforme discutido em Koenker (2005), por exemplo, a suposição de erros identicamente distribuídos não é uma condição necessária para ajuste da regressão quantílica. Ao contrário da metodologia clássica de regressão, os modelos de regressão quantílica são capazes de incorporar a informação de heterocedasticidade dos erros aleatórios independentes.

Definido o conceito de regressão quantílica, é necessário entender como é feita a interpretação dos parâmetros de seus modelos. Considere, por exemplo, que $f(\mathbf{x}_i, \tau) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau)$ para um τ fixado. Neste caso, a interpretação dos parâmetros $\boldsymbol{\beta}(\tau)$ é essencialmente a mesma de qualquer outro modelo linear, no sentido de se dar em função da taxa de variação. Ou seja, o coeficiente $\beta_j(\tau)$, $j = 1, \cdots, p$, pode ser interpretado como a taxa de variação no $\tau-$ésimo quantil da variável resposta Y ao variar-se em uma unidade o valor da $j-$ésima covariável mantendo-se os valores das demais variáveis fixos. Isto é,

$$\beta_j(\tau) = \frac{\partial Q_Y|\mathbf{x}(\tau)}{\partial \mathbf{x}_j}.$$

Para estimação dos parâmetros $\boldsymbol{\beta}(\tau)$, relembre inicialmente que, no caso univariado, de acordo com Definição 4, o quantil de ordem τ pode ser consistentemente estimado encontrando-se o valor y da amostra que minimiza a função $\sum_{i=1}^{n} \rho_\tau(y_i - y)$. Na presença das covariáveis, o valor y é modelado por $\mathcal{Q}_{Y_i}(\tau|\mathbf{x}_i)$, que no caso linear é dado por $\mathbf{x}_i^T \boldsymbol{\beta}(\tau)$. Então, o interesse é encontrar $\mathbf{b}(\tau)$, estimativa de $\boldsymbol{\beta}(\tau)$, que minimiza a função:

$$S_n[\boldsymbol{b}(\tau)] = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \mathbf{b}(\tau)). \tag{2.3}$$

Pelas propriedades bastante conhecidas de cálculo, tem-se que o valor $\mathbf{b}(\tau)$ que minimiza a função (2.3) é também raiz da seguinte função de estimação:

$$D_n[\mathbf{b}(\tau)] = \frac{dS_n[\mathbf{b}(\tau)]}{d\mathbf{b}(\tau)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \{\tau - \mathcal{I}(y_i - \mathbf{x}_i^T \mathbf{b}(\tau) \leq 0)\}. \tag{2.4}$$

No entanto, não é trivial encontrar a raiz da equação (2.4) que, por envolver uma função indicadora, não assume fórmula analítica. Como alternativa, a literatura sugere a reformulação da função (2.3) para uma equação equivalente, entendendo regressão quantílica como resultado de um problema de programação linear, em que é possível encontrar a solução usando métodos já bastante conhecidos e consolidados.

Como motivação à interpretação de regressão quantílica como um problema de programa-

ção linear, defina inicialmente:

$$e_i = y_i - \mathbf{x}_i^T \mathbf{b}(\tau), \mu_i = \max\{e_i, 0\} \text{ e } \nu_i = \max\{-e_i, 0\}, i = 1, \cdots, n,$$

e observe que $S_n[\boldsymbol{b}(\tau)]$ pode ser reescrita em função de $\mu_1, \cdots, \mu_n$ e $\nu_1, \cdots, \nu_n$ assim definidos, como:

$$S_n[\boldsymbol{b}(\tau)] = \sum_{i=1}^{n} e_i\{\tau - 0\}\mathcal{I}(e_i > 0) + e_i\{\tau - 1\}\mathcal{I}(e_i \leq 0) = \sum_{i=1}^{n}\{\tau\mu_i + (1-\tau)\nu_i\}.$$

Ou seja, tem-se interesse em minimizar $S_n[\boldsymbol{b}(\tau)]$ assim definida em função das restrições sobre μ_i e ν_i, as partes positivas e negativas dos resíduos, respectivamente. Minimizar uma função linear com restrições de equações ou inequações lineares é essencialmente um problema de programação linear. Mais especificamente, considere o problema:

$$\min_{(b,\mu,\nu)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \{\tau\mathbf{1}_n^T\boldsymbol{\mu} + (1-\tau)\mathbf{1}_n^T\boldsymbol{\nu} | \mathbf{X}\mathbf{b}(\tau) + \boldsymbol{\mu} - \boldsymbol{\nu} = \mathbf{y}\},$$

em que $\mathbf{1}_n$ denota um vetor $n \times 1$ de valores iguais a 1, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$, $\mathbf{y} = (y_1, \cdots, y_n)^T$, e $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$ e $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_n)^T$ definidos, respectivamente, como:

$$\mu_i = \begin{cases} y_i - \mathbf{x}_i^T\mathbf{b}(\tau), & \text{se } y_i - \mathbf{x}_i^T\mathbf{b}(\tau) > 0, \\ 0, & \text{caso contrário;} \end{cases} \qquad \nu_i = \begin{cases} -y_i + \mathbf{x}_i^T\mathbf{b}(\tau), & \text{se } y_i - \mathbf{x}_i^T\mathbf{b}(\tau) < 0, \\ 0, & \text{caso contrário,} \end{cases}$$

para $i = 1, \cdots, n$.

Então, conforme discutido em Chen e Wei (2005), pode-se escrever o seguinte problema padrão de programação linear (P):

$$\begin{aligned} (P) \qquad & \min_{\boldsymbol{\theta}} \mathbf{d}^T\boldsymbol{\theta} \\ \text{sujeito a} \quad & B\boldsymbol{\theta} = \mathbf{y}, \\ & \boldsymbol{\theta} \geq 0, \end{aligned}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\phi}^T, \boldsymbol{\varphi}^T, \boldsymbol{\mu}^T, \boldsymbol{\nu}^T)^T$, $\boldsymbol{\phi} = [\boldsymbol{b}(\tau)]_+$, $\boldsymbol{\varphi} = [-\boldsymbol{b}(\tau)]_+$, $\boldsymbol{\mu}$ e $\boldsymbol{\nu}$ são conforme definidos anteriormente, $\mathbf{d} = (\mathbf{0}_p^T, \mathbf{0}_p^T, \tau\mathbf{1}_n^T, (1-\tau)\mathbf{1}_n^T)$, em que $\mathbf{0}_p$ é o vetor $p \times 1$ de valores iguais a zero. A matriz B, por sua vez, pode ser definida como $B = [\mathbf{X} \ -\mathbf{X} \ \ \mathbf{I}_n \ -\mathbf{I}_n]$, em que $\mathbf{I}_n$ é a matriz identidade de ordem n.

O uso das ferramentas de programação linear permitiu, portanto, o desenvolvimento da regressão quantílica. Entre as técnicas utilizadas para a resolução destes problemas, pode-se citar o método *simplex*, processo iterativo que se inicia com uma solução que satisfaz as restrições lineares, e faz a busca pela solução que resulta no menor valor da função objetivo (ou maior, em problemas de maximização). Uma interpretação geométrica do método *simplex* e maiores detalhes sobre a técnica podem ser encontradas, por exemplo, em Koenker (2005) e Davino et al. (2013).

Em problemas de minimização dos erros absolutos do modelo, o primeiro algoritmo eficiente que fez uso de programação linear foi o proposto por Barrodale e Roberts (1973). Mais

tarde, já no contexto de regressão quantílica, Koenker e d'Orey (1987) propuseram uma adaptação do método *simplex*, que é bastante conveniente para um número moderado de observações. Para grandes amostras, a literatura sugere o uso de outra técnica computacionalmente mais eficiente, a do ponto interior, proposta por Portnoy e Koenker (1997). Uma introdução a essa técnica e uma comparação com o método *simplex* podem ser encontradas em Chen e Wei (2005).

Conforme mencionado anteriormente, a regressão quantílica é uma técnica que permite a modelagem de qualquer quantil de ordem τ de interesse, $\tau \in [0, 1]$. Em alguns casos, inclusive, tem-se interesse em estudar todos os quantis, de modo a compreender toda a distribuição da variável resposta em função das covariáveis, o que é denominado *processo de regressão quantílica*.

Quando o interesse reside no processo como um todo, usualmente determina-se uma sequência de pontos $\tau_1^*, \tau_2^*, \cdots$, com $\tau_i^* \in [0, 1]$ igualmente espaçados, de modo que a distância entre cada ponto seja tão pequena quanto se queira. Em outras palavras, divide-se o intervalo $[0, 1]$ em uma grade (*grid*, em inglês) densa, com pontos equidistantes. Então, calcula-se a estimativa dos parâmetros do modelo para cada um dos τ_i^*, no caso, assumindo que a linearidade é presente em todos os quantis. O número de quantis distintos que é possível obter em um processo de regressão quantílica, como destaca Davino et al. (2013), está relacionado com o número de observações e o tamanho da amostra. Mais especificamente, o número de quantis distintos está relacionado positivamente com o tamanho da amostra e negativamente com o número de covariáveis.

Uma importante observação que deve ser feita no estudo de regressão quantílica é que os quantis não necessariamente têm a mesma ordenação de y_i, já que o quantil é condicional às covariáveis. A Tabela 2.1 apresenta um exemplo que ilustra tal fato. Nela são apresentados os quantis associados às probabilidades $\tau \in (0,25; 0,50; 0,75)$, calculados com auxílio do *software* **R**, para uma amostra extremamente simples e fictícia. Observe então, por exemplo, que $y_5 > y_4$, mas $\hat{Q}_{y_5|x_5}(\cdot) < \hat{Q}_{y_4|x_4}(\cdot)$, para $\tau \in (0,25; 0,50)$.

Tabela 2.1: Quantis associados às probabilidades $\tau \in (0,25; 0,50; 0,75)$, para uma amostra fictícia.

i	1	2	3	4	5	6	7	8	9	10
y_i	1	2	3	4	5	6	7	8	9	10
x_i (binária)	1	1	0	0	1	0	0	1	0	1
$\hat{Q}_{y_i\|x_i}(0,25)$	2	2	4	4	2	4	4	2	4	2
$\hat{Q}_{y_i\|x_i}(0,50)$	5	5	6	6	5	6	6	5	6	5
$\hat{Q}_{y_i\|x_i}(0,75)$	8	8	7	7	8	7	7	8	7	8

2.2.1 Propriedades e Inferência

Conforme discutido, o problema de estimação dos parâmetros $\boldsymbol{\beta}(\tau)$ foi resolvido aplicando-se ferramenta de programação linear. Uma vez estimados os parâmetros, é importante conhecer

suas propriedades e métodos inferenciais disponíveis.

Observe que o vetor de parâmetros estimados $\mathbf{b}(\tau)$ depende de τ, claramente de $\mathbf{y}$ e da matriz de covariáveis $\mathbf{X}$ observados na amostra, sendo $\mathbf{b}(\tau)$, $\mathbf{y}$ e $\mathbf{X}$ conforme definidos anteriormente. Dessa forma, para enunciar as propriedades que seguem, denote $\mathbf{b}(\tau) = \mathbf{b}(\tau, \mathbf{y}, \mathbf{X})$.

Teorema (Koenker e Bassett, 1978). *Seja $\boldsymbol{A}$ matriz não singular de dimensão $p \times p$, $\boldsymbol{\gamma} \in \mathbb{R}^p$, e $a > 0$. Então, para qualquer $\tau \in [0, 1]$, pode-se mostrar que:*

(i) $\mathbf{b}(\tau, a\mathbf{y}, \mathbf{X}) = a\mathbf{b}(\tau, \mathbf{y}, \mathbf{X})$

(ii) $\mathbf{b}(\tau, -a\mathbf{y}, \mathbf{X}) = a\mathbf{b}(1 - \tau, \mathbf{y}, \mathbf{X})$

(iii) $\mathbf{b}(\tau, \mathbf{y} + \mathbf{X}\boldsymbol{\gamma}, \mathbf{X}) = \mathbf{b}(\tau, \mathbf{y}, \mathbf{X}) + \boldsymbol{\gamma}$

(iv) $\mathbf{b}(\tau, \mathbf{y}, \mathbf{X}A) = A^{-1}\mathbf{b}(\tau, \mathbf{y}, \mathbf{X})$

As propriedades (*i*) e (*ii*) tratam de equivariância de escala, enquanto que a propriedade (*iii*) aborda o contexto conhecido como equivariância de regressão e a (*iv*) é chamada de equivariância da reparametrização da matriz de planejamento (Koenker, 2005).

Outra importante propriedade da regressão quantílica é a equivariâcia sob transformações monótonas. Relembre que, nos problemas de regressão usuais, quando a transformação da variável resposta se faz necessária para obtenção de propriedades desejáveis dos estimadores, como linearidade, por exemplo, a interpretação dos parâmetros é comprometida, uma vez que deve ser feita em função da variável transformada. Isso pode ser demonstrado pela desigualdade de Jensen, em que:

$$E(g(Y)) \neq g(E(Y)).$$

Por outro lado, em regressão quantílica tem-se que:

$$Q_{g(Y)}(\tau) = g(Q_Y(\tau)),$$

que deriva diretamente do fato de que $P(Y \leq y) = P(g(Y) \leq g(y))$. Conforme será discutido no Capítulo 3, essa propriedade é bastante importante para aplicação da metodologia de regressão quantílica no contexto de Análise de Sobrevivência, uma vez que os tempos de sobrevivência, em geral, não seguem uma relação linear com as covariáveis.

Intervalos de confiança e teste de hipóteses

Considere agora o modelo linear definido conforme (2.2):

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + u_i.$$

Suponha que os erros u_i são independentes e identicamente distribuídos com função de distribuição $F(\cdot)$, e que o quantil de ordem τ de u_i seja igual a zero. Considere ainda uma sequência $\tau_1, \cdots, \tau_m$ de probabilidades de interesse (Koenker, 2005):

1. A função densidade $f(\cdot)$, associada à função de distribuição acumulada $F(\cdot)$, é tal que $f(F^{-1}(\tau_j)) > 0, j = 1, \cdots, m$.

2. O modelo é ajustado com intercepto.

3. $\lim_{n\to+\infty} \sum \mathbf{x}_i \mathbf{x}_i^T = Q$, em que Q é matriz positiva definida.

Nessas condições, pode-se mostrar que:

$$\sqrt{n}(\mathbf{b}(\tau_1) - \boldsymbol{\beta}(\tau_1), \cdots, \mathbf{b}(\tau_m) - \boldsymbol{\beta}(\tau_m)) \xrightarrow{D} \mathcal{N}(0, V(\tau_1, \cdots, \tau_m)),$$

em que $V(\tau_1, \cdots, \tau_m) = \Omega(\tau_1, \cdots, \tau_m, \mathbf{F}) \otimes Q^{-1}$, e $\Omega(\tau_1, \cdots, \tau_m, \mathbf{F})$ é a matriz de covariâncias entre os m quantis amostrais, e $\otimes$ representa o produto de *Kronecker*. Ou seja, sob as suposições do modelo, os estimadores dos parâmetros da regressão quantílica são não viesados e seguem distribuição normal assintótica. Sob condições adicionais, é possível mostrar a consistência do estimador. Ver, por exemplo, Koenker (2005).

Conforme discutido em Santos (2012), alguns autores propõem formas de estimar a função de covariâncias, que para o caso de um único τ é dada por:

$$V(\tau) = \frac{\tau(1-\tau)}{f^2(0)}(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.5}$$

Por exemplo, uma estimativa para $1/f(0)$ pode ser calculada a partir de diferenças entre os quantis empíricos:

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

com $\lim_{h_n\to+\infty} h_n = 0$, e h_n calculado conforme Hall e Sheather (1988).

Se, por outro lado, os erros do modelo não são identicamente distribuídos, também é possível mostrar a normalidade e boas propriedades do estimador. Conforme discutido em Koenker (2005), neste caso, é possível mostrar que para determinado τ de interesse,

$$\sqrt{n}(\mathbf{b}(\tau) - \boldsymbol{\beta}(\tau)) \xrightarrow{D} \mathcal{N}(0, V(\tau)),$$

em que

$$V(\tau) = \tau(1-\tau)H_n^{-1}J_nH_n^{-1},$$

com $J_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, $H_n(\tau) = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T f_i(\xi_i(\tau))$, e $f_i(\xi_i(\tau))$ denota a função densidade de probabilidade da variável resposta avaliada no quantil condicional de ordem τ. O autor discute ainda que, neste caso, uma estimativa não viesada para $f_i(\xi_i(\tau))$ é dada por:

$$\frac{2h_n}{\mathbf{x}_i^T\boldsymbol{b}(\tau + h_n) - \mathbf{x}_i^T\boldsymbol{b}(\tau - h_n)}.$$

Observe que se $f_i(\xi_i(\tau)) = f(\xi_i(\tau)), \forall i$, isto é, sob a suposição de erros aleatórios independentes e identicamente distribuídos, a matriz de covariâncias coincide com a apresentada anteriormente em (2.5).

Estimada a matriz de covariância, e sob a distribuição normal assintótica, é possível construir intervalos de confiança para os parâmetros a fim de avaliar se podem ser considerados diferente de zero. Neste caso, tem-se que:

$$IC(b_i(\tau), 1-\alpha) = b_i(\tau) \mp t_{\alpha/2;n-1} \sqrt{\frac{\hat{V}(\tau)}{n}},$$

em que $\hat{V}(\tau)$ é uma estimativa de $V(\tau)$, e $t_{\alpha/2;n-1}$ é o quantil de ordem $\alpha/2$ da distribuição *t-Student* com $n-1$ graus de liberdade.

Outra metodologia de construção de intervalos de confiança se baseia na técnica de *bootstrap*. Trata-se de um esquema de reamostragem que consiste em selecionar n pares $(y_i, \mathbf{x}_i)$ com reposição da amostra original de tamanho n, de modo que cada par tenha probabilidade de $1/n$ de ser sorteado. Esse procedimento é repetido B vezes e, para cada uma delas, o vetor de parâmetros $\mathbf{b}(\tau)$ é calculado. Cada uma dessas B estimativas contribui para a estimação do erro padrão dos parâmetros.

Então, um intervalo de confiança para o parâmetro $b_i(\tau)$, $i = 1, \cdots, p$, com coeficiente de confiança $1-\alpha$ é dado por:

$$IC(b_i(\tau), 1-\alpha) = b_i(\tau) \mp t_{\alpha/2;n-1} \widehat{EP[b_i(\tau)]}.$$

A desvantagem em adotar a metodologia de *bootstrap* é que, para grandes amostras, o custo operacional é bastante intenso. Para outras formas de estimação dos intervalos de confiança referencia-se a dissertação de Santos (2012).

Em relação a testes de hipóteses, dois tipos de testes são de interesse. Em primeiro lugar, quer-se testar se parâmetros dentro de um mesmo quantil são iguais a constantes conhecidas, como zero por exemplo. Neste caso, a literatura sugere a aplicação do teste de Wald, que não apresenta grandes complicações uma vez estimada a matriz de covariâncias. Outro teste de interesse é avaliar e comparar parâmetros de diferentes quantis. Conforme apresentado em Koenker (2005), é possível escrever um único teste de hipóteses para avaliar essas duas situações, que será discutido a seguir.

Considere $\zeta = (\boldsymbol{\beta}(\tau_1)^T, \cdots, \boldsymbol{\beta}(\tau_m)^T)^T$. Seja $\boldsymbol{R}$ uma matriz de posto completo de ordem q de constantes conhecidas e r um vetor de constantes também conhecidas, de dimensão $m \times 1$. Pode-se escrever o seguinte teste de hipóteses geral:

$$H_0 : \boldsymbol{R\zeta} = \boldsymbol{r}$$

e a estatística de teste é dada por:

$$T_n = n(\boldsymbol{R\hat{\zeta}} - \boldsymbol{r})^T [\boldsymbol{RV}^{-1}\boldsymbol{R}^T]^{-1} (\boldsymbol{R\hat{\zeta}} - \boldsymbol{r}), \text{ com } \hat{\zeta} = (\boldsymbol{b}(\tau_1)^T, \cdots, \boldsymbol{b}(\tau_m)^T)^T,$$

em que V é uma matriz de dimensão $mp \times mp$ em que o bloco i, j é dado por:

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

com J_n e H_n definidos conforme anteriormente. Sob a hipótese nula, T_n tem distribuição qui-quadrado assintótica com q graus de liberadade.

Koenker (2005) aponta esse teste de hipóteses geral para regressão quantílica como uma alternativa robusta aos convencionais testes para detectar heterocedasticidade dos parâmetros, uma vez que a metodologia de regressão quantílica é robusta à presença de valores discrepantes na variável resposta.

Para ilustrar o modelo de regressão quantílica, considere o exemplo a seguir, construído com dados fictícios gerados com auxílio do *software* R.

Exemplo

Considere uma amostra de $n = 1.000$ observações da variável aleatória Y tal que:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

com ε_i independentes e identicamente distribuídos tais que $\varepsilon \sim \mathcal{N}(0, 1)$. Considere $\beta_0 = 4, \beta_1 = 2$ e x_i gerada a partir de uma amostra de uma variável aleatória X_i com distribuição uniforme no intervalo $(3, 4)$. Suponha que x_i é dada, que Y_i, condicional a x_i, sejam independentes, $i = 1, \cdots, n$, e que o interesse seja estimar os quantis de ordem $\tau = (0,25; 0,50; 0,75)$. Isto é, tem-se interesse em estimar $\beta_0(\tau)$ e $\beta_1(\tau)$ tais que

$$Q_{Y_i|x_i}(\tau) = \beta_0(\tau) + \beta_1(\tau).$$

Com auxílio do pacote *quantreg* do *software* R, foram calculadas as estimativas $b_0(\tau)$ e $b_1(\tau)$ dos parâmetros $\beta_0(\tau)$ e $\beta_1(\tau)$, respectivamente, para cada um dos quantis de interesse. Os erros padrão das estimativas foram estimados via *bootstrap* e os resultados encontram-se na Tabela 2.2.

Tabela 2.2: Estimativas dos parâmetros no modelo de regressão quantílica para $\tau \in (0,25; 0,50; 0,75)$.

Parâmetro	Estimativa	Erro Padrão	Wald	Valor p
$\beta_0(0{,}25)$	3,29	0,52	6,35	<0,001
$\beta_1(0{,}25)$	1,99	0,15	13,49	<0,001
$\beta_0(0{,}50)$	4,30	0,48	8,87	<0,001
$\beta_1(0{,}50)$	1,90	0,14	13,57	<0,001
$\beta_0(0{,}75)$	4,53	0,46	9,76	<0,001
$\beta_1(0{,}75)$	2,02	0,13	15,42	<0,001

Observe na Tabela 2.2 que também já são apresentados os valores da estatística de Wald para testar se os parâmetros são significativamente diferentes de zero ou não. Observe que, ao nível de significância de 5%, os parâmetros são significativos, e os intervalos de confiança

correspondentes, com coeficiente de confiança igual a 95%, são dados por:

$$IC(\beta_0(0,25);0,95) = 3,29 \mp 1,96 \times 0,52 = [2,27;4,31]\,,$$
$$IC(\beta_1(0,25);0,95) = 1,99 \mp 1,96 \times 0,15 = [1,70;2,28]\,,$$

$$IC(\beta_0(0,50);0,95) = 4,30 \mp 1,96 \times 0,48 = [3,36;5,24]\,,$$
$$IC(\beta_1(0,50);0,95) = 1,90 \mp 1,96 \times 0,14 = [1,63;2,17]\,,$$

$$IC(\beta_0(0,75);0,95) = 4,53 \mp 1,96 \times 0,46 = [3,63;5,43]\,,$$
$$IC(\beta_1(0,75);0,95) = 2,02 \mp 1,96 \times 0,13 = [1,77;2,27]\,.$$

Os modelos finais, para cada um dos quantis de ordem τ, podem ser escritos como:

$$\hat{Q}_{Y_i|x_i}(0,25) = 3,29 + 1,99x_i.$$
$$\hat{Q}_{Y_i|x_i}(0,50) = 4,30 + 1,90x_i.$$
$$\hat{Q}_{Y_i|x_i}(0,75) = 4,53 + 2,02x_i.$$

A interpretação dos parâmetros, conforme discutido ao longo deste capítulo, é dada em termos da taxa de variação no quantil em análise ao se variar o valor da covariável x_i. Por exemplo, aumentando-se uma unidade o valor da covariável x_i, estima-se que a mediana de y_i (quantil de ordem $\tau = 0,50$), por exemplo, aumenta em 1,90 unidades. Por outro lado, nesta situação, estima-se que o quantil de ordem $\tau = 0,75$ aumenta 2,02 unidades.

Para esse caso simples, é possível entender a relação entre os parâmetros da regressão quantílica, $\beta_0(\tau)$ e $\beta_1(\tau)$, com β_0 e β_1, parâmetros da regressão linear usual. Observe que

$$F_{Y_i|\mathbf{x}_i}(y_i) = \tau \Rightarrow F^{-1}_{Y_i|\mathbf{x}_i}(y_i) = z_\tau + \beta_0 + \beta_1 x_i,$$

em que z_τ é o quantil de ordem τ da distribuição normal padrão. Então, $\beta_1(\tau) = \beta_1$ e $\beta_0(\tau) = \beta_0 + z_\tau$. No exemplo, $\beta_0(0,25) = 4 - 0,67 = 3,33$; $\beta_0(0,50) = 4 + 0 = 4$ e $\beta_0(0,75) = 4 + 0,67 = 4,67$.

A ideia geral apresentada para a modelagem com regressão quantílica requer que as variáveis Y_i sejam completamente observáveis,$\forall i = 1, \cdots, n$. Na prática, no entanto, a variável dependente pode estar sujeita à censura, e por isso alguns autores dedicaram-se à extensão dos modelos de regressão quantílica para o caso mais geral, em que se têm censuras, isto é, informações incompletas ou parciais acerca da resposta do estudo. Uma introdução aos modelos de regressão quantílica lineares para dados censurados, aplicados à Análise de Sobrevivência, é apresentada no próximo capítulo.

Capítulo 3

Regressão Quantílica na Presença de Censura

Em Análise de Sobrevivência, os dados, em geral, caracterizam-se pela presença de censura, definida como a informação incompleta ou parcial da variável resposta nas unidades amostrais. Existem vários tipos de censura, e sua classificação depende da informação que se tem acerca do momento de ocorrência do evento de interesse (denominado falha, em geral). Considera-se nesta dissertação o caso de censura aleatória à direita, que conforme definido em Klein e Moeschberger (2005), por exemplo, ocorre quando as unidades experimentais deixam de ser observadas após um tempo C_i (variável aleatória), porém antes de apresentarem o evento, isto é, $Y_i > C_i$.[1]

Note que, por estar sujeita à censura à direita, a variável aleatória observada é, na verdade, $\tilde{Y}_i$, definida como mínimo entre Y_i e C_i, $\tilde{Y}_i = \min(Y_i, C_i)$. Defina a variável $\delta_i = \mathbb{I}(Y_i < C_i)$ como indicadora do evento, isto é:

$$\delta_i = \begin{cases} 1, & Y_i < C_i, \\ 0, & Y_i \geq C_i. \end{cases}$$

Da mesma forma que Y_i, a variável C_i também pode estar correlacionada com as covariáveis. Por exemplo, no estudo do tempo de vida de pacientes com câncer, a não observação do evento *morte* em um grupo de pacientes pode estar associado ao tipo de tratamento a que foram submetidos. Em outro exemplo, no mercado segurador, não se observar o evento *sinistro* para um grupo de segurados da carteira pode estar associado às suas variáveis de perfil. A suposição que normalmente é feita em relação a C_i, e que será assumida doravante, é que Y_i e C_i são condicionalmente independentes dado o vetor de covariáveis $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T, \forall i = 1, \cdots, n$.

Relembre que, no modelo de regressão quantílica linear, o objetivo é estimar o quantil de ordem τ da distribuição de Y_i, dada a suposição de que as covariáveis se relacionam com a

[1] As metodologias aqui apresentadas, no entanto, podem ser aplicadas também ao contexto de *censura de Tipo I*, que ocorre quando se pré-estabelece um tempo máximo de observação para cada um dos itens em estudo (ou seja, para C_i conhecido e fixado para todo i).

variável resposta linearmente em τ, isto é,

$$Q_{Y_i|\mathbf{x}_i}(\tau) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau),$$

em que $\boldsymbol{\beta}(\tau)$ é o vetor de parâmetros do modelo, que depende de τ. Além disso, conforme apresentado no capítulo anterior, uma estimativa de $\boldsymbol{\beta}(\tau)$ é dada pelo vetor $\mathbf{b}(\tau)$ que minimiza a expressão (2.3). Conforme discutido em Koenker (2005), a premissa para a generalização da regressão quantílica para dados na presença de censura está no fato dos subgradientes de $S_n(\mathbf{b}(\tau))$, isto é, de suas derivadas parciais direcionais com relação a $\mathbf{b}(\tau)$, só dependerem do valor observado de Y_i através da função indicadora $\mathcal{I}(y_i - \mathbf{x}_i^T\mathbf{b}(\tau) \leq 0)$. [2] Dessa forma, para a estimação do vetor $\boldsymbol{\beta}(\tau)$ não é necessário saber o valor exato que Y_i assume, mas apenas se seu valor é menor ou maior do que $\mathbf{x}_i^T\mathbf{b}(\tau)$. A Figura 3.1 esquematiza as possíveis localizações do verdadeiro valor de Y_i em relação à $\mathbf{x}_i^T\mathbf{b}(\tau)$.

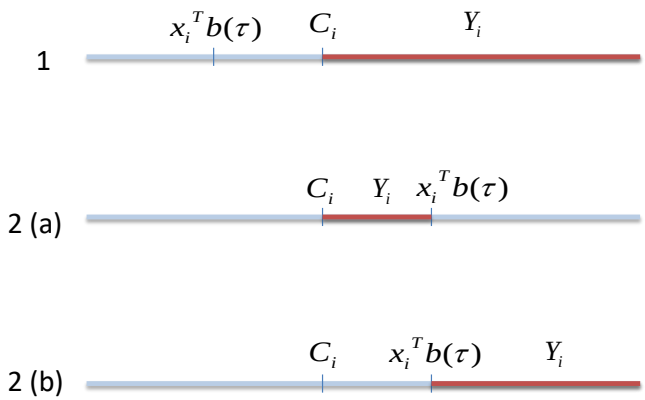

Figura 3.1: Esquema com as possíveis localizações do verdadeiro valor de Y_i em relação à $\mathbf{x}_i^T\mathbf{b}(\tau)$ em um cenário de censura aleatória à direita.

Mais especificamente, suponha que $F_{Y_i|\mathbf{x}_i}(y_i)$, função de distribuição acumulada de Y_i, dado $\mathbf{x}_i$, seja conhecida. Observe que, quando Y_i é censurada, isto é, $Y_i > C_i$, podem existir as seguintes situações no estudo da função indicadora $\mathcal{I}(y_i - \mathbf{x}_i^T\mathbf{b}(\tau) \leq 0)$:

1. $C_i > \mathbf{x}_i^T\mathbf{b}(\tau)$. Neste caso, mesmo sem observar a variável Y_i, é fácil ver que $\mathcal{I}(Y_i - \mathbf{x}^T\mathbf{b}(\tau) \leq 0) = 0$.

2. $C_i < \mathbf{x}_i^T\mathbf{b}(\tau)$. Diferente do caso anterior, apenas com a observação da variável C_i não é possível saber qual valor a função indicadora assume. O valor da função indicadora está associado às seguintes duas situações, cujas probabilidades de ocorrência devem ser avaliadas:

[2] Em termos de programação linear, nota-se que a contribuição de cada observação para a estimação de $\boldsymbol{\beta}(\tau)$ depende apenas do sinal dos resíduos, definidos como $y_i - \mathbf{x}_i^T\mathbf{b}(\tau)$. Uma explicação mais completa de programação linear em regressão quantílica pode ser encontrada, por exemplo, em Davino et al. (2013).

(a) $Y_i \in (C_i, \mathbf{x}_i^T \boldsymbol{b}(\tau)]$. Neste caso, $\mathcal{I}(Y_i - \mathbf{x}_i^T \mathbf{b}(\tau) \leq 0) = 1$, e a probabilidade de se observar essa situação é dada por

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

(b) $Y_i \in (\mathbf{x}_i^T \boldsymbol{b}(\tau), +\infty)$. Este caso é complementar ao anterior, dado que $Y_i > C_i$. Então a sua probabilidade de ocorrência é $1 - \tilde{w}_i(\tau)$, com $\mathcal{I}(Y_i - \mathbf{x}_i^T \mathbf{b}(\tau) \leq 0) = 0$.

Na prática, a função $F_{Y_i|\mathbf{x}_i}(y_i)$ não é conhecida, e portanto, deve ser estimada. Dessa forma, considere uma amostra de n observações da variável aleatória *tempo até a ocorrência da falha* (especificada), sujeita à censura aleatória à direita. Em outras palavras, considere a tripla de variáveis $\{\tilde{y}_i, \delta_i, \mathbf{x}_i\}$, $i = 1 \cdots, n$, em que $\tilde{y}_i = \min(y_i, c_i)$ e $\delta_i = \mathcal{I}(y_i < c_i)$. Para as observações censuradas, $\delta_i = 0$, denote por $w_i(\tau)$ a estimativa de $\tilde{w}_i(\tau)$, dada por:

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i}, \tag{3.1}$$

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

Motivado pela teoria de redistribuição da massa de probabilidade proposta por Efron (1967), Portnoy (2003) propôs uma forma de estimar $F_{Y_i|\mathbf{x}_i}(C_i)$ e introduzir os pesos w_i à função (2.3), de modo a incorporar a informação da censura aleatória na estimação dos parâmetros $\boldsymbol{\beta}(\tau)$. Sua contribuição permitiu avanços no estudo da regressão quantílica na presença de censura, e impulsionou seu uso em aplicações práticas em Análise de Sobrevivência.

A ideia geral da metodologia de Portnoy (2003), e que será discutida com mais detalhes na próxima seção, é que, se a $i-$ésima observação é censurada, $\delta_i = 0$, mas se $c_i > \mathbf{x}_i^T \mathbf{b}(\tau)$, então a sua contribuição é a mesma de uma observação não censurada: em ambos os casos o resultado da função indicadora é conhecido.

Por outro lado, se $\delta_i = 0$ e $c_i < \mathbf{x}_i^T \mathbf{b}(\tau)$, a contribuição da observação censurada não pode ser a mesma de uma não censurada, sob o risco de ter um modelo viesado, já que o resultado da função indicadora não pode ser determinado. Neste caso, deve-se levar em consideração as duas possibilidades enunciadas no **item 2** anterior. Mais especificamente, deve-se atribuir peso $w_i(\tau)$ à $i-$ésima observação, ou seja, ao instante c_i, e $1 - w_i(\tau)$ a um tempo de falha qualquer, desde que maior do que $\mathbf{x}_i^T \mathbf{b}(\tau)$. Na prática, como será discutido, escolhe-se um valor grande o suficiente para estar além do escopo dos tempos de falha. Em outras palavras, uma nova observação "fictícia" é introduzida, com o mesmo vetor de covariáveis.

Portnoy (2003) apresenta, então, em seu trabalho um método recursivo para a estimação dos parâmetros do modelo, partindo da ideia da estimação do processo de regressão quantílica e assumindo que em todos os quantis da variável resposta a estrutura com as covariáveis é linear. Mais tarde, Wang e Wang (2009) propuseram uma metodologia de pesos locais, em que se o interesse é estudar a mediana, por exemplo, não é necessário calcular regressões para todos os quantis anteriores. Isso relaxa a suposição de linearidade global: a linearidade é necessária e suficiente apenas no quantil de interesse. Wey et al. (2014) também apresentam

uma metodologia de pesos locais, mas usando a teoria de árvore de sobrevivência, conferindo flexibilização na modelagem no que diz respeito a estudos com muitas covariáveis. Essas três abordagens serão discutidas com mais detalhes nas próximas seções.

3.1 Método Recursivo

Como motivação à metodologia de Portnoy (2003), considere o esquema de ponderação descrito a seguir, baseado na teoria de redistribuição da massa de probabilidade proposta por Efron (1967), para estimar a função de distribuição acumulada de Y_i.

3.1.1 Esquema de Ponderação via Kaplan-Meier

Considere inicialmente o estimador de Kaplan-Meier para a estimação da função de sobrevivência de uma variável aleatória que pode estar sujeita a censura. Como resultado, tem-se uma função escada com saltos nos instantes de tempo em que, de fato, são observados o evento de interesse. Relembre que o estimador de Kaplan-Meier é uma generalização da função de sobrevivência empírica, definida por:

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

Como é bem conhecida, a fórmula clássica para estimar a função de sobrevivência via Kaplan-Meier envolve analisar o número de falhas até o instante de tempo de interesse, e o número de indivíduos sob risco neste tempo, ou seja, número de indivíduos que não falharam e não foram censurados até o tempo imediatamente anterior. Mais especificamente, considere:

- $y_1 < \cdots < y_m$, os m tempos de falha distintos e ordenados,
- d_j o número de falhas em y_j, $j = 1, \cdots, m$, e
- n_j o número de indivíduos sob risco em y_j.

O estimador de Kaplan-Meier pode ser então definido como:

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

Maiores detalhes e uma justificativa para a expressão podem ser encontrados em Colosimo e Giolo (2006).

Para ilustrar o estimador de Kaplan-Meier, suponha uma amostra de 10 observações das variáveis aleatórias independentes e identicamente distribuídas $Y_1, \cdots, Y_{10}$, em que $y_1 = 1, \cdots, y_{10} = 10$. Considere ainda que as observações y_3, y_6 e y_7 são censuradas à direita, isto é, o verdadeiro tempo de falha é posterior ao tempo observado. A Figura 3.2 apresenta a função de sobrevivência estimada e a Tabela 3.1 apresenta os valores estimados da função de sobrevivência e da função de distribuição acumulada para este exemplo simples. Na tabela,

as colunas Risco e Evento representam, respectivamente, o número de indivíduos sob risco e o número de eventos (falhas) observados em y_i.

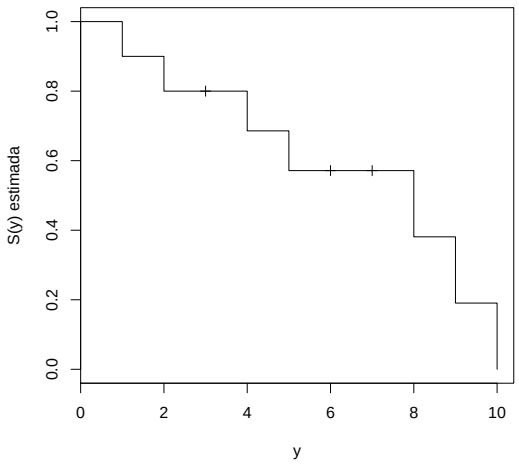

Figura 3.2: Curva de sobrevivência estimada para um exemplo simples e fictício, em que $y_1 = 1, \cdots, y_{10} = 10$, com y_3, y_6 e y_7 censuradas.

Tabela 3.1: Estimativas da função de sobrevivência e função de distribuição acumulada, $S(y_i)$ e $F(y_i)$, respectivamente, para o exemplo em que $y_1 = 1, \cdots, y_{10} = 10$, com y_3, y_6 e y_7 censuradas.

y_i	Risco	Evento	$\hat{S}(y)_i$	$\hat{F}(y)_i$
1	10	1	0,900	0,100
2	9	1	0,800	0,200
4	7	1	0,686	0,314
5	6	1	0,571	0,429
8	3	1	0,381	0,619
9	2	1	0,190	0,810
10	1	1	0,000	1,000

Outra forma de calcular o estimador de Kaplan-Meier, não tão usual, é considerá-lo como resultante de um esquema de ponderação que atribui pesos a cada uma das observações. No exemplo, note que como as duas primeiras observações são não censuradas, então a função de sobrevivência empírica tem saltos de dimensão 1/10. Visto de outra forma, isso é equivalente a dizer que cada observação tem peso igual a 1, e que, portanto, a distribuição acumulada no ponto i é dada pela razão entre os pesos anteriores a observação y_i (inclusive) e o tamanho n da amostra, no caso, $i = 1, 2$ e $n = 10$.

Por outro lado, conforme evidenciado na Figura 3.2, o estimador de Kaplan-Meier não atribui massa de probabilidade a observações censuradas. A distribuição acumulada no instante y_3 é, portanto, igual a $0,2$ (ou, equivalentemente, a função de sobrevivência estimada em y_3 é igual a $0,8$). Neste caso, o peso de y_3, necessário para computar a função de distribuição

acumulada nos instantes de falha posteriores, não pode ser igual a 1, pois isto seria equivalente a dizer que a observação é não censurada. A ideia então é repartir o peso da observação y_3 levando-se em conta as duas possíveis situações: o verdadeiro valor do tempo de falha ocorre entre os instantes 3 e 4, ou o verdadeiro tempo é posterior a 4. Conforme apresentado no início deste capítulo, essas situações tem probabilidades $(\tau^* - 0,2)/(1 - 0,2)$ e $(1 - \tau^*)/(1 - 0,2)$, respectivamente, sendo τ^*, estimativa de τ, a probabilidade acumulada no tempo de falha y_4. Então, de acordo com o que foi apresentado, τ^* deve satisfazer:

$$\tau^* = \frac{2 + (\tau^* - 0,2)/(1 - 0,2) + 1}{10},$$

em que 2 é o peso das observações y_1 e y_2 (não censuradas), $(\tau^* - 0,2)/(1 - 0,2)$ é o peso de y_3, e 1, de y_4. Resolvendo a equação, encontra-se $\tau^* = 0,314$, que coincide com o valor de $\hat{F}(y_4)$ apresentado na Tabela 3.1. Então, uma nova observação é introduzida ao conjunto de dados, por exemplo, $y_{11} = +\infty$, com peso $(1 - 0,314)/(1 - 0,2) = 0,857$. Observe que, independentemente da escolha de y_{11}, a estimativa de τ^* não é alterada.

Seguindo no exemplo, a observação y_5 também não é censurada, então:

$$\tau^* = \frac{2 + (\tau^* - 0,2)/(1 - 0,2) + 1 + 1}{10}.$$

Resolvendo a equação, tem-se que $\tau^* = 0,429$. Por outro lado, y_6 e y_7 são censuradas, então a função de distribuição acumulada nos instantes y_6 e y_7 é igual $0,429$, já que não tem massa de probabilidade associada. Para a observação em y_8 tem que:

$$\tau^* = \frac{2 + (\tau^* - 0,2)/(1 - 0,2) + 1 + 1 + 2 \times (\tau^* - 0,429)/(1 - 0,429) + 1}{10},$$

em que $\tau^* = 0,619$. Prosseguindo da mesma forma, é fácil ver que para o tempo de falha y_9 tem-se que $\tau^* = 0,810$ e que no tempo y_{10} a função de distribuição acumulada é igual a 1.

Conforme discutido anteriormente, para cada observação censurada que foi ponderada, uma nova observação fictícia é introduzida no conjunto de dados. Escolhe-se qualquer valor além do escopo dos dados, pois conforme apresentado, a introdução dessa observação não afeta a estimativa de τ.

Observe que essa forma de atribuir pesos às observações para o cálculo da função de distribuição acumulada, que remete à ideia de redistribuição da massa de probabilidades proposta por Efron (1967), resulta em 1 menos a estimativa de Kaplan-Meier, e portanto, é uma forma alternativa de computar suas estimativas.

3.1.2 Algoritmo de Portnoy, 2003

O esquema de ponderação via Kaplan-Meier apresentado anteriormente é essencial para entendimento do algoritmo proposto por Portnoy (2003) para estimação dos parâmetros do modelo em regressão quantílica, no cenário em que a variável resposta está sujeita a censura.

Na metodologia proposta pelo autor supracitado, o esquema de ponderação é utilizado, no entanto, com algumas modificações. O objetivo é calcular os pesos $w_i(\tau)$ que serão atribuídos

às observações censuradas, mas os valores de τ, ou seja, os quantis de interesse da distribuição, são fixados previamente. Apenas $\hat{\tau}_i$, estimativa de $F_{Y_i|\mathbf{x}_i}(C_i)$, deve ser calculada. Dessa forma, defina $\tau_j^* : j = 1, 2, \cdots$, uma sequência de probabilidades associadas a uma partição fixada da variável resposta. A sequência τ_j^* pode ser definida, por exemplo, como uma grade de probabilidades igualmente espaçadas $0 < \tau_1^* < \cdots < \tau_M^* < 1$, em que M é fixado. Os pesos $w_i(\tau_m^*), m = 1, \cdots, M$, são dados por:

$$w_i(\tau_m^*) = \begin{cases} 1, & \delta_i = 1 \text{ ou } \hat{\tau}_i > \tau_m^*, \\ \frac{\tau_m^* - \hat{\tau}_i}{1 - \hat{\tau}_i}, & \delta_i = 0 \text{ e } \hat{\tau}_i < \tau_m^*. \end{cases}$$

A metodologia de Portnoy (2003) envolve então os seguintes passos:

1. *Dado o valor τ_1^*, estima-se $\boldsymbol{b}(\tau_1^*)$ aplicando regressão quantílica linear usual, isto é, como no caso sem censura.*

   Relembre que, na estimação da função de sobrevivência via Kaplan-Meier, são retirados do estudo os tempos censurados c_k que são estritamente menores do que o primeiro tempo de falha. Isso porque o método de Kaplan-Meier não atribui saltos à função de sobrevivência estimada nos instantes de censura.

   Analogamente, a suposição que é feita em regressão quantílica é que não existe nenhuma observação censurada c_k tal que $c_k \leq \mathbf{x}_i^T \mathbf{b}(\tau_1^*), \forall\ i = 1, \cdots, n$. Se existir algum c_k nessa situação, então a observação deve ser retirada da amostra, e o vetor de parâmetros $\mathbf{b}(\tau_1^*)$ deve ser recalculado.

   Observe que, como resultado do primeiro passo do algoritmo, tem-se que $\mathcal{I}(\tilde{y}_i - \mathbf{x}_i^T \mathbf{b}(\tau_1^*) \leq 0) = 0$, para toda observação (restante) da amostra.

2. *Para estimar $\boldsymbol{b}(\tau_2^*)$, procede-se exatamente igual ao caso sem censura. Então verifica-se se existem valores c_i, $i = 1, \cdots, n$, tais que:*

$$c_i \in \left[\boldsymbol{x}_i^T \boldsymbol{b}(\tau_1^*); \boldsymbol{x}_i^T \boldsymbol{b}(\tau_2^*) \right].$$

   Se existirem, $\boldsymbol{b}(\tau_2^*)$ deve ser reestimado, levando-se em consideração a existência de tais observações censuradas. Denote por K o conjunto de índices de tais observações censuradas. Atribui-se $\hat{\tau}_i = \tau_1^*$ para as observações em K. Dessa forma, $\boldsymbol{b}(\tau_2^*)$ é o vetor de parâmetros que minimiza a função:

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

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

   Proceder exatamente igual ao caso sem censura significa supor que todos os valores $\tilde{y}_i$, $i = 1, \cdots, n$, são, na verdade, tempos de falha. O algoritmo é baseado na avaliação

das censuras e reestimação dos parâmetros, objetivando uma estimativa não viesada, que seria obtida se não fosse atribuído o esquema de ponderação. De fato, a escolha do valor de $y^{+\infty}$ é arbitrária, mas normalmente se escolhe um valor além do escopo de $\tilde{y}_i$, para todo i, pois dessa forma evita-se que sejam necessárias novas definições de $y^{+\infty}$ para o cálculo das estimativas dos parâmetros em quantis maiores.

3. *Suponha que já se tenha calculado $\boldsymbol{b}(\tau_j^*)$, e denote por K o índice das observações censuradas que contribuem para o seu cálculo. O próximo passo é estimar $\boldsymbol{b}(\tau_{j+1}^*)$, admitindo que não existam outras censuras além das já consideradas em K. Suponha, no entanto, que a após estimar o parâmetro $\boldsymbol{b}(\tau_{j+1}^*)$, exista uma observação censurada c_k tal que:*

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

Isso significa que o vetor $\boldsymbol{b}(\tau_{j+1}^*)$ precisa ser recalculado, considerando-se também a informação dessa censura. Em outras palavras, o índice k deve ser incorporado ao conjunto K. Assume-se então que $\hat{\tau}_k = \tau_j^*$ para essa observação que precisa ser ponderada com peso $w_k(\tau_j^*)$. Além disso, deve-se 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. Assim, de uma forma geral, $\boldsymbol{b}(\tau_{j+1}^*)$ é o vetor de parâmetros que minimiza a função:

$$\sum_{i \notin K} \rho_{\tau_{j+1}^*}(\tilde{y}_i - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_{j+1}^*)) +$$

$$+ \sum_{i \in K} \{w_i(\tau_{j+1}^*) \times \rho_{\tau_{j+1}^*}(c_i - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_{j+1}^*)) + (1 - w_i(\tau_{j+1}^*)) \times \rho_{\tau_{j+1}^*}(y^{+\infty} - \boldsymbol{x}_i^T \boldsymbol{b}(\tau_{j+1}^*))\}$$

em que $y^{+\infty}$ é um tempo de falha suficientemente grande introduzido ao conjunto de dados.

Observe que a metodologia está baseada no processo de regressão quantílica, que cobre toda a distribuição do tempo de sobrevivência. A cada passo, os pesos do modelo de todas as observações censuradas e que já foram analisadas anteriormente são reestimados, considerando-se o τ de interesse.

4. *Repete-se o passo 3 até que τ_{j+1}^* seja igual a 1, ou quando restarem apenas observações censuradas à direita de $\boldsymbol{x}^T \boldsymbol{b}(\tau_{j+1}^*)$.*

Relembre que, no estimador de Kaplan-Meier, quando as maiores observações são censuradas, tem-se exatamente o mesmo caso, em que a função de sobrevivência estimada é incompleta.

Em termos computacionais, o algoritmo implementado nos principais *softwares*, como R e SAS, por exemplo, é extremante eficiente e apresenta os resultados rapidamente. No entanto, a principal desvantagem da metodologia é a suposição de que todos os quantis se relacionam linearmente com as covariáveis. Na prática, isso nem sempre é verificado: em geral, por

apresentarem poucas observações, os primeiros quantis não seguem uma relação linear, por exemplo.

A seguir são apresentadas algumas suposições e propriedades da metodologia de Portnoy (2003), e que são discutidas e provadas no artigo do autor.

Propriedades Assintóticas e Inferência

A metodologia de Portnoy (2003) considera as seguintes suposições:

S1. *Seja $\tilde{\tau}_1$ o "verdadeiro" (único) valor de $\hat{\tau}_1$ tal que $\boldsymbol{x}_i^T\boldsymbol{\beta}(\tilde{\tau}_1) = C_i$. Existe $\varepsilon > 0$, tal que $\tilde{\tau}_1 \geq \varepsilon$.*

Trata-se de uma condição necessária para garantir que o primeiro quantil não tem informações censuradas. Também é a premissa da estimação da função de sobrevivência via Kaplan-Meier.

S2. *A função densidade de probabilidade de Y_i, dado $\boldsymbol{x}_i$, e sua derivada satisfazem:*

$$a \leq f_i(u) \leq b, \qquad |f_i'(u)| \leq c,$$

uniformemente para $\varepsilon \leq F_i(u) \leq 1 - \varepsilon$ e uniformemente em $i = 1, \cdots, n$, em que $a > 0, b < +\infty$ e c é uma constante que pode depender de ε.

S3. *Existe uma constante B tal que $\|\boldsymbol{x}_i\| \leq B$ uniformemente em $i = 1, \cdots, n$,* em que $\|v\|$ representa a norma do vetor v.

Em seu artigo, Portnoy (2003) ressalta que é possível relaxar esta suposição, permitindo que o limite de $\mathbf{x}_i$ dependa do tamanho da amostra n, desde que n cresça lentamente. No entanto, o autor discute que as propriedades assintóticas tornam-se complicadas a ponto da suposição de limite fixo ser razoável. Outra justificativa é que, em geral, se tem poucas observações com grandes valores em $\mathbf{x}_i$, e para estes casos, as estimativas dos quantis seriam ruins.

S4. *Defina*

$$\boldsymbol{S}_n(\tau) = \frac{1}{n}\boldsymbol{X}^T(diag\{\tilde{w}_i(\tau)f_i(\boldsymbol{x}_i\boldsymbol{\beta}(\tau))\})\boldsymbol{X}.$$

Existe uma matriz (não aleatória) positiva definida, $S(\tau)$, e uma constante c tal que, para n suficientemente grande,

$$\|\boldsymbol{S}_n(\tau) - \boldsymbol{S}(\tau)\| \leq cn^{-1/4}.$$

Defina agora a matriz $\mathbf{D}_n \equiv \text{diag}(\{d_i\})$, em que

$$d_i \equiv \tau(1-\tau) - (1-\tau)\left\{\mathbb{I}(C_i \leq \mathbf{x}_i^T\boldsymbol{\beta}(\tau))\left[\frac{\tau - \mathbb{P}_{\mathbf{x}_i}(Y_i \leq C_i)}{1 - \mathbb{P}_{\mathbf{x}_i}(Y_i \leq C_i)}\right]\right\}.$$

Assuma que

$$\mathbf{X}^T\mathbf{D}_n\mathbf{X} \to \mathbf{V}(\tau),$$

quando $n \to \infty$. Quando as suposições S1-S4 são válidas, Portnoy (2003) mostra que:

$$\sqrt{n}(\mathbf{b}(\tau, w) - \boldsymbol{\beta}(\tau))) \xrightarrow{D} \mathcal{N}(0, \mathbf{S}^{-1}(\tau)\mathbf{V}(\tau)\mathbf{S}^{-1}(\tau)).$$

Ou seja, sob as suposições acima, o estimador é não viesado e segue assintoticamente distribuição normal multivariada. Na prática, assim como no contexto sem censura, é impossível estimar a função de variância de $\mathbf{b}(\tau, w)$ diretamente, já que envolve a função de distribuição de Y_i, que é desconhecida.

Uma possível solução para inferência, implementada no *software* R, é estimar via *bootstrap* a função de variância. Então, como a distribuição assintótica do estimador de $\boldsymbol{\beta}(\tau)$ é normal, uma estatística de teste para a hipótese:

$$H_0 : \beta_l(\tau) = 0, l = 1, \cdots, p,$$

por exemplo, é dada pela estatística t :

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

em que $\hat{s}_e$ é a estimativa de *bootstrap* para a variância de $b_l(\tau)$. Equivalentemente, o intervalo de confiança *bootstrap-t*, com nível de confiança $1 - \alpha$ (ver, por exemplo, Efron e Tibshirani (1994)), pode ser escrito como:

$$IC(b_l(\tau), 1 - \alpha) = b_l(\tau) \mp t_{\alpha/2;n-1}\sqrt{\frac{\hat{s}_e}{n}}.$$

No entanto, o autor propõe o uso de uma metodologia híbrida de *bootstrap* para determinação de intervalos de confiança para o parâmetro $\beta(\tau)$. Neste caso, a proposta consiste em determinar via *bootstrap* as distâncias interquartílicas $b_l(\tau)^*_{,75} - b_l(\tau)^*_{,50}$ e $b_l(\tau)^*_{,50} - b_l(\tau)^*_{,25}$, em que $b_l(\tau)^*_k$ é o quantil de ordem k das estimativas de $b_l(\tau)$ determinadas via *bootstrap*. Em seguida, multiplicar essas medidas estimadas por 2,906 e adicionar o valor $\beta(\tau)^*_{,50}$. Conforme discutido em Heritier et al. (2009), o valor 2,906 é utilizado para garantir a consistência do estimador. Dessa forma, obtêm-se os intervalos de confiança com coeficiente de confiança de 95% para $b_l(\tau)$. Observe que, dessa forma, os intervalos não são necessariamente simétricos.

Outra possível abordagem para intervalos baseados em *bootstrap* para os parâmetros da regressão quantílica é sugerida nos trabalhos de Wang e Wang (2009) e Wey et al. (2014). A técnica de construção consiste em gerar B amostras *bootstrap* e estimar em cada uma delas os parâmetros correspondentes. Posteriormente, tais estimativas são ordenadas e tomam-se os quantis $0,025$ e $0,975$ para a construção de um intervalo com confiança de 95% para cada um dos parâmetros, por exemplo.

De um modo geral, a vantagem de se adotar *bootstrap* para inferência em regressão quantílica com censura, conforme destacam os autores, é que métodos baseados em reamostragem

da tripla (Y_i, C_i, δ_i) podem ser justificados pela teoria clássica de *bootstrap*. Por outro lado, metodologias que envolvem os resíduos do modelo, ou que abordam o contexto em que as triplas (Y_i, C_i, δ_i) não são independentes e identicamente distribuídas para todo $i = 1, \cdots, n$, ainda carecem de estudos e novas teorias.

Exemplo

Para ilustrar o algoritmo recursivo de Portnoy (2003), considere como exemplo o conjunto de dados *rdata* disponível no pacote *relsurv* do *software* R. Trata-se de um estudo conduzido em Ljubljana, na Eslovênia, cujo objetivo era avaliar o tempo de sobrevivência (em dias) de pacientes após infarto agudo do miocárdio. Uma análise para este conjunto de dados também é apresentada em Wang e Wang (2009), em que foram avaliados 972 pacientes com idades entre 40 e 80 anos, de ambos os sexos, sendo que 507 apresentaram o evento "morte".

As curvas de Kaplan-Meier para as variáveis em estudo são apresentadas na Figura 3.3. Observe que as curvas de Kaplan-Meier sugerem haver diferenças no tempo de sobrevivência entre homens e mulheres, mas não tão acentuadas. Por outro lado, parece haver diferença significativa nos tempos de sobrevivência entre os grupos de idade analisados, tal que, quanto maior a idade, menor o tempo de sobrevivência.

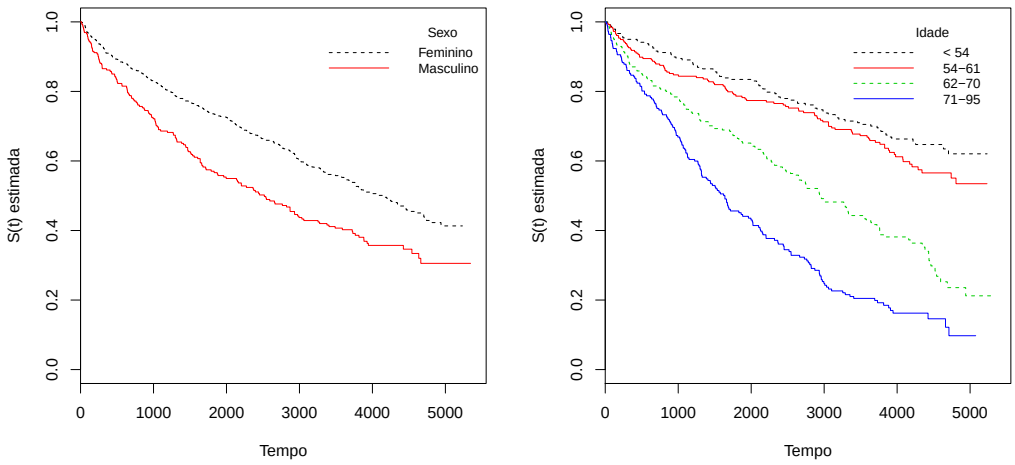

Figura 3.3: Curvas de Kaplan-Meier para as variáveis sexo e idade de um estudo de tempo de sobrevivência após infarto agudo do miocárdio.

Mais especificamente, suponha que seja de interesse avaliar o tempo de vida mediano dos pacientes após infarto agudo do miocárdio.

A literatura sugere que os tempos de sobrevivência não podem ser modelados por uma função linear das covariáveis. Uma possível solução é aplicar transformação logarítmica na variável dependente que, em geral, lineariza essa relação. Relembre que a regressão quantílica tem a propriedade de equivariância sob transformações monótonas e, portanto, a interpretação na escala original dos dados não sofre nenhuma perda.

Wang e Wang (2009) apresentam um gráfico descritivo para avaliar se a suposição de linearidade no quantil $\tau = 0,5$ é válida, e tal gráfico é reproduzido na Figura 3.4. A construção do gráfico é baseada em suavizadores do tipo *splines*. De acordo com os autores, é razoável assumir a linearidade para o quantil de ordem $\tau = 0,50$. Supôs-se que a linearidade é válida para todos os quantis e então se ajustou a metodologia recursiva de Portnoy. Os resultados do ajuste do modelo são apresentados na Tabela 3.2.

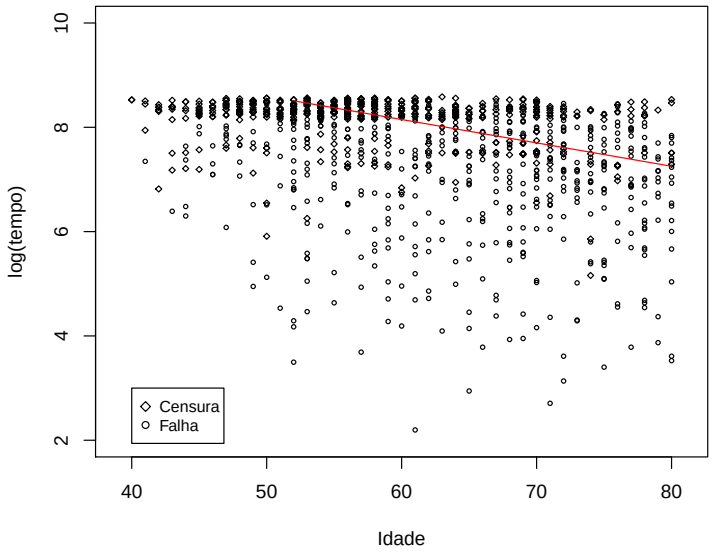

Figura 3.4: Idade *versus* logaritmo dos tempos de sobrevivência de pacientes com infarto agudo do miocárdio. A reta estimada é o suavizador do tipo *spline* para avaliar a suposição de linearidade do modelo.

Tabela 3.2: Estimativas dos parâmetros calculadas via método recursivo para quantil de ordem $\tau = 0,50$. Os erros padrão foram estimados via *bootstrap*.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	Wald	Valor p
Intercepto	11,95	155.520,06	0,68	17,58	<0,001
Sexo Feminino	-	-	-	-	-
Sexo Masculino	0,13	1,14	0,16	0,85	0,397
Idade	-0,06	0,94	0,01	-5,86	<0,001

Observe que apenas a variável Idade e o intercepto do modelo são significativos ao nível de significância de 5%, determinado via teste de Wald, considerando a estimativa de *bootstrap* para os erros padrão do modelo.

Para a construção dos intervalos de confiança, foram geradas 500 subamostras do conjunto de dados, e para cada uma delas, foi ajustada a metodologia recursiva de Portnoy para estimação dos parâmetros do modelo. As estimativas de cada um dos parâmetros foram ordenadas e

então se considerou o quantil de ordem $\tau = 0,025$ e $\tau = 0,975$ da distribuição das estimativas. Os intervalos obtidos são apresentados na Tabela 3.3.

Tabela 3.3: Intervalos de confiança *bootstrap* para os parâmetros estimados via metodologia recursiva de Portnoy.

Variável	Estimativa	IC(95%)
Intercepto	11,95	[10.77; 12.86]
Sexo Feminino	-	-
Sexo Masculino	0,13	$[-0,11; 0,38]$
idade	-0,06	$[-0,07; -0,04]$

A interpretação dos parâmetros é feita da mesma forma que em regressão quantílica para dados não censurados: em termos da taxa de variação das estimativas. Assim, levando-se em consideração a significância dos parâmetros e a escala original do tempo de sobrevivência, pode-se concluir que o tempo de sobrevida mediano de um paciente após infarto agudo do miocárdio é 0,06 vezes menor ao se aumentar em um ano a sua idade $(1 - 0,94)$.

Conforme discutido anteriormente, modelos de regressão quantílica podem ser vistos como complementos valiosos às análises tradicionais. Para esse exemplo, ajustou-se também o modelo de Cox, cujas estimativas dos parâmetros são apresentadas na Tabela 3.4.

Tabela 3.4: Resultados do ajuste do modelo de riscos proporcionais de Cox para os dados de pacientes após infarto agudo do miocárdio.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	Valor p
Sexo Feminino	-	-	-	-
Sexo Masculino	-0,04	0,96	0,10	0,680
Idade	0,06	1,06	0,00	<0,001

Observe que as conclusões inferenciais são as mesmas do modelo de regressão quantílica: apenas a variável idade é significativa ao nível de significância de 5%, e quanto maior a idade, maior é a taxa de risco. Então, neste caso, é natural pensar que menor é o tempo de sobrevivência, o que está de acordo com o apresentado pelo modelo de regressão quantílica para dados censurados.

3.2 Abordagens de Pesos Locais

Conforme apresentado na seção anterior, uma das suposições para a modelagem via algoritmo de Portnoy (2003) é a relação de linearidade dos parâmetros com o quantil de ordem τ da variável resposta, para todo $\tau \in [0,1]$. No entanto, quando essa suposição é violada, as

estimativas dos parâmetros do modelo para um determinado quantil de interesse podem ser comprometidas.

Exemplo

Ainda como exemplo, considere o conjunto de dados referente ao tempo de sobrevida de pacientes após infarto agudo do miocárdio, apresentado na seção anterior. A Figura 3.4 foi apresentada para justificar a aplicação da regressão quantílica linear para o logaritmo do tempo de sobrevivência em função da idade para um quantil específico, $\tau = 0,50$. Naquela ocasião, supôs-se que os demais quantis também seguiam a relação linear. No entanto, Wang e Wang (2009) apresentam em seu artigo a avaliação dos demais quantis e argumentam que a suposição de linearidade é rejeitada para os quantis menores. A Figura 3.5 reproduz o gráfico original dos autores, também baseado em suavizadores do tipo *splines*, que ilustra tal fato.

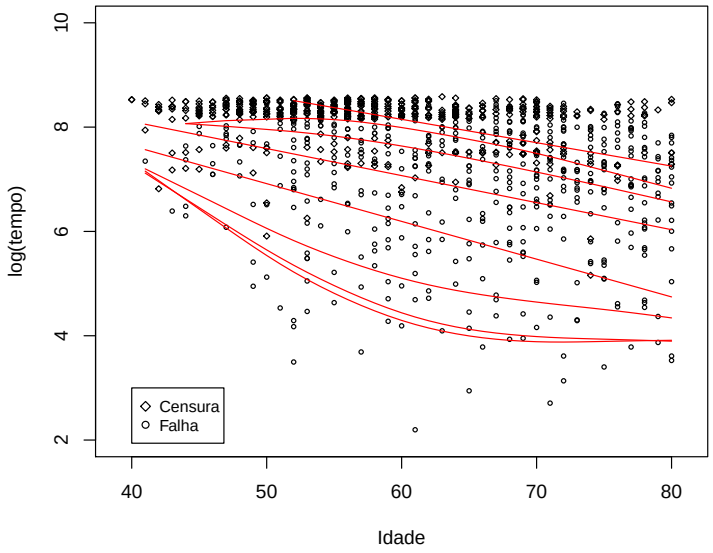

Figura 3.5: Idade *versus* logaritmo dos tempos de sobrevivência de pacientes com infarto agudo do miocárdio. As curvas estimadas são suavizadores do tipo *spline* para avaliar a suposição de linearidade do modelo nos quantis 0,01; 0,03; 0,05; 0,1; 0,2; 0,3; 0,4 e 0,5.

Nesta seção serão discutidas duas abordagens de pesos locais, propostas por Wang e Wang (2009) e Wey et al. (2014), e que permitem ajustar modelos de regressão quantílica para dados censurados para um determinado quantil de interesse sem impor a suposição de linearidade global, bastante restritiva na prática.

3.2.1 Pesos Estimados via função *Kernel*

Similarmente a Portnoy (2003), Wang e Wang (2009) propõem uma metodologia para estimar os parâmetros da regressão quantílica também partindo da ideia da redistribuição de massa

de probabilidade proposta por Efron (1967). Essencialmente, os pesos de cada observação são definidos de forma análoga, isto é,

$$w_i(\tau) = \begin{cases} 1, & \delta_i = 1 \text{ ou } \hat{\tau}_i > \tau, \\ \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i}, & \delta_i = 0 \text{ e } \hat{\tau}_i < \tau. \end{cases} \tag{3.2}$$

Porém, a principal diferença entre as metodologias é que, para a de Wang e Wang (2009) não é necessário calcular todo o processo de regressão quantílica, mas apenas o quantil (ou quantis, se for o caso) de interesse. Relembre que a metodologia de Portnoy (2003) envolve o processo de regressão quantílica de modo recursivo, em que para estimar cada um dos pesos deve-se levar em consideração os quantis anteriores, que determinam o valor de $\hat{\tau}_i$.

A ideia da metodologia de pesos locais é fixar um quantil τ de interesse, a mediana, por exemplo, e minimizar a função objetivo ponderada:

$$\sum_{i=1}^{n} \{w_i(\tau) \times \rho_\tau(\tilde{y}_i - \mathbf{x}_i^T \mathbf{b}(\tau)) + (1 - w_i(\tau)) \times \rho_\tau(y^{+\infty} - \mathbf{x}_i^T \mathbf{b}(\tau))\},$$

em que, da mesma forma, $y^{+\infty}$ é um tempo de falha suficientemente grande para estar além do escopo de $\tilde{y}$, e que é inserido ao conjunto de dados.

Como o processo de estimação não é recursivo, é necessário outra forma de determinar as probabilidades $\hat{\tau}_i$, que, conforme já discutido anteriormente, é uma estimativa para a função de distribuição acumulada da variável dependente no ponto c_i, dadas as covariáveis. Os autores propõem então o uso do estimador Kaplan-Meier local, que pode ser escrito como:

$$\hat{F}_{Y_i|\mathbf{x}_i}(y) = 1 - \prod_{j=1}^{n} \left[1 - \frac{B_{nj}(\mathbf{x})}{\sum_{k=1}^{n} \mathcal{I}(\tilde{y}_k \geq \tilde{y}_j) B_{nk}(\mathbf{x})} \right]^{\nu_j(y)}, \tag{3.3}$$

em que $\nu_j(y) = \mathcal{I}(\tilde{y}_j \leq y, \delta_j = 1)$ e $B_{nj}(\mathbf{x})$ é uma função de pesos para as observações.

Observe em (3.3) que, se $B_{nk}(\mathbf{x}) = 1/n$, então $1 - \hat{F}_{Y_i|\mathbf{x}_i}(y)$ coincide com o estimador clássico de Kaplan-Meier. Os autores propõem ainda uma modificação no estimador de pesos locais, incluindo pesos de Nadaraya-Watson na estimação da função de distribuição acumulada. Dessa forma,

$$B_{nk} = \frac{K(\frac{\mathbf{x} - \mathbf{x}_k}{h_n})}{\sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{x}_i}{h_n})},$$

em que $K(\cdot)$ é a função de densidade *kernel*, e $h_n \in \mathbb{R}^+$ é o *bandwidth*, com $h_n \to 0$, quando n converge para infinito. Maiores detalhes sobre a densidade *kernel* podem ser encontrados no Apêndice A.

Diferentes funções *kernel* podem ser usadas na estimação de $\hat{F}_{Y_i|\mathbf{x}_i}(y)$. Entretanto, é possível mostrar que a escolha da função *kernel* não é tão influente no resultado. Assim como os autores, usa-se nesta dissertação a função *kernel* biquadrática, isto é,

$$K(x) = \frac{15}{16}(1 - x^2)^2 \mathcal{I}(|x| \leq 1).$$

Assume-se que as covariáveis são aleatórias, e independentes, de modo que dadas duas covariáveis x_1 e x_2, $K(x_1, x_2) = K(x_1) \times K(x_2)$.

Conforme discutido em Hastie e Tibshirani (1990), a escolha do *bandwidth*, h_n, no contexto de estimação via *kernel* pode produzir mudanças consideráveis na estimação da função de distribuição acumulada. Um dos critérios para seleção de h_n é a *validação cruzada*. Existem várias abordagens para aplicação de validação cruzada, e uma delas é brevemente discutida no Apêndice A desta dissertação. Wang e Wang (2009) apresentam um estudo de simulação em seu artigo em que mostram que o estimador proposto para estimar os parâmetros da regressão quantílica não é sensível à escolha de h_n, quando se tem apenas uma covariável. Nesta dissertação considerou-se um h_n para cada covariável, utilizando-se o pacote *kedd* do *software* R, que faz um método exaustivo de validação cruzada. Conforme será discutido mais adiante, nos estudos de simulação, de fato, a escolha de h_n não se mostra tão influente na estimação dos parâmetros.

Propriedades Assintóticas e Inferência

A metodologia considera as seguintes suposições, que são necessárias para demonstração das propriedades de seu estimador:

A1 *Existe uma constante k_x tal que $E\|\boldsymbol{x}\|^3 <= k_x$. Além disso, $\max_{1\leq i\leq n} \|\boldsymbol{x}_i\| = \mathcal{O}(n^{1/2}(\log n)^{-1})$, e $E(\boldsymbol{x}\boldsymbol{x}^T)$ é uma matriz de dimensão $p \times p$ positiva definida.*

A2 *Seja $F_0(y|\boldsymbol{x}) = P(Y_i \leq y|\boldsymbol{x})$ e $G(y|\boldsymbol{x}) = P(C_i \leq y|\boldsymbol{x})$. As funções $F_0(y|\boldsymbol{x})$ e $G(y|\boldsymbol{x})$ tem primeira derivada em y, denotadas por $f_0(y|\boldsymbol{x})$ e $g(y|\boldsymbol{x})$, respectivamente, que são limitadas uniformemente por um $b < +\infty$. Além disso, $F_0(y|\boldsymbol{x})$ e $G(y|\boldsymbol{x})$ tem derivadas parciais de segunda ordem com respeito a $\boldsymbol{x}$ limitadas uniformemente em y.*

A3 *Para $\boldsymbol{b}(\tau)$ na vizinhança de $\boldsymbol{\beta}(\tau)$, $E[\boldsymbol{x}\boldsymbol{x}^T \times f_0(\boldsymbol{x}^T\boldsymbol{b}|\boldsymbol{x}) \times \{1 - G(\boldsymbol{x}^T\boldsymbol{b}|\boldsymbol{x})\}]$ é positiva definida.*

A4 *O bandwidth é tal que $h_n = \mathcal{O}(n^{-1/2+\gamma_0})$, em que $0 < \gamma_0 < 1/4$.*

A5 *A função kernel $K(\cdot) \geq 0$ tem suporte em um compacto. Tem continuidade Lipschitz de ordem 1 e satisfaz $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int K^2(u)du < \infty$, e $\int |u|^2K(u)du < \infty$.*

De posse das propriedades de A1-A5 listadas, os autores demonstram a convergência em probabilidade dos estimadores dos parâmetros do modelo. Em outras palavras, demonstram que

$$\mathbf{b}(\tau) \to \boldsymbol{\beta}(\tau),$$

em probabilidade, quando $n \to \infty$. Além disso, para $1/6 < \gamma_0 < 1/4$, mostram que:

$$n^{1/2}(\mathbf{b}(\tau) - \boldsymbol{\beta}(\tau)) \to \mathcal{N}(0, \boldsymbol{\Gamma_1}^{-1}\mathbf{V}\boldsymbol{\Gamma}_1^{-1}),$$

em distribuição, em que:

$$\boldsymbol{\Gamma}_1 = E[\mathbf{x}\mathbf{x}^T\{1 - G(\mathbf{x}^T\mathbf{b}(\tau)|\mathbf{x})\} \times f_0(\mathbf{x}^T\mathbf{b}(\tau)|\mathbf{x})]$$

e

$$\mathbf{V} = Cov(\mathbf{m}_i(\mathbf{b}, F_0) + (1-\tau)\boldsymbol{\phi}(\tilde{y}_i, \delta_i, \mathbf{x}_i^T\mathbf{b}(\tau), \mathbf{x}_i)),$$

com $\mathbf{m}_i(\mathbf{b}, F_0) = \mathbf{x}_i\{\tau - w_i(F_0)\mathcal{I}(\tilde{y}_i < \mathbf{x}_i^T\mathbf{b}(\tau))\}$ e $\boldsymbol{\phi}(\tilde{y}_i, \delta_i, \mathbf{x}_i^T\mathbf{b}(\tau), \mathbf{x}_i))$ é uma função que depende de $F_0(y|\mathbf{x}_i)$ e $G(y|\mathbf{x}_i)$.

Observe que, assim como a metodologia de Portnoy (2003), a matriz de covariâncias do estimador depende das funções de distribuição acumulada $F_0(\cdot)$ e $G(\cdot)$, que são desconhecidas.

Os autores propõem que a inferência da significância dos parâmetros seja a partir de intervalos de confiança, construídos à luz da teoria de *bootstrap*. Neste caso, os autores sugerem que os limites do intervalo com nível de significância de 5%, por exemplo, sejam calculados a partir do 2,5 e 97,5 percentis dos coeficientes estimados via *bootstrap*. Em seu artigo, Wang e Wang (2009) mostram que a performance dessa abordagem é satisfatória em estudo de Monte Carlo.

No *software* R, a metodologia de Wang e Wang (2009) pode ser ajustada com o uso da mesma função que a utilizada para ajuste sem censura, apenas incluindo os pesos como parâmetro (função *rq* do pacote *quantreg*). Então a saída do modelo apresenta o teste *t* para significância dos parâmetros. Observe que o estimador de $\boldsymbol{\beta}(\tau)$ tem distribuição assintótica normal, então o teste *t* também pode ser utilizado como alternativa para grandes amostras.

3.2.2 Pesos Estimados via Árvores de Sobrevivência

Apesar de conferir algumas vantagens, como flexibilização da suposição de linearidade global e não requerer a estimação recursiva dos parâmetros, a metodologia de Wang e Wang (2009) tem algumas desvantagens. A aplicação dos estimadores *kernel* torna-se impraticável quando muitas variáveis aleatórias fazem parte do estudo, conforme destaca Wey et al. (2014), ressaltando que os estimadores *kernel* não são recomendáveis para dimensões maiores do que duas variáveis. A razão para tal fato é que quanto maior a dimensão do vetor de variáveis $\mathbf{x}$, mais espaçados ficarão os pontos amostrais. Em outras palavras, em um espaço de dimensão mais elevado haverá poucas observações em torno de um vetor $\mathbf{x}$ qualquer, a menos que o tamanho da amostra seja extremamente grande. Caso a amostra não seja suficientemente grande, os valores de *bandwidth* devem ser cada vez maiores, o que resulta no aumento do viés do estimador. Uma discussão acerca do assunto pode ser encontrada em Scott e Sain (2005). Outra desvantagem em relação à abordagem de Wang e Wang (2009), ainda de acordo com Wey et al. (2014), é que a metodologia foi desenvolvida apenas para variáveis contínuas, o que é bastante restritivo para aplicações práticas.

Nesse sentido, Wey et al. (2014) apresentam um estimador alternativo para o cálculo dos pesos $w_i(\tau)$, definidos em (3.2), baseado na metodologia de árvores de sobrevivência. As árvores são uma ferramenta flexível, sendo bastante poderosas do ponto de vista preditivo, bem como um valioso método descritivo.

Uma breve introdução às árvores de sobrevivência é apresentada no Apêndice B. De um modo geral, trata-se de um esquema de partição disjunta do conjunto de dados em grupos homogêneos entre si de acordo com um critério de avaliação das curvas de sobrevivência.

O número de grupos é determinado por algumas condições que avaliam o número mínimo de indivíduos em risco e número mínimo de eventos dentro de cada grupo, necessários para aplicação do estimador de Kaplan-Meier nos grupos resultantes.

Em outras palavras, particionada a amostra em grupos disjuntos entre si, para um dado $\mathbf{x}_i$ fixado é possível calcular o correspondente valor de $F_{Y_i|\mathbf{x}_i}(C_i)$ utilizando o estimador de Kaplan-Meier.

Conforme discutido no Apêndice B, as árvores, no entanto, são bastante sensíveis em relação à amostra, no sentido de que dados ligeiramente diferentes podem resultar em árvores completamente distintas. Nesse sentido é aplicada a ideia de *bagging*, proposta por Breiman (1996) , que consiste em retirar L amostras com reposição da amostra original, construir a árvore para cada uma delas, e calcular a média das estimativas, de modo a obter uma estimativa mais estável. Isto é,

$$\hat{\tau} = \hat{F}_{Y_i|\mathbf{x}_i}(y) = \frac{1}{L}\sum_{b=1}^{L} \tilde{F}^b_{Y_i|\mathbf{x}_i}(y),$$

em que $\tilde{F}^b_{Y_i|\mathbf{x}_i}(y)$ é a $b-$ésima subamostra *bootstrap*.

Propriedades Assintóticas e Inferência

É importante destacar que as metodologias de Wey et al. (2014) e Wang e Wang (2009) diferenciam-se apenas na estimação de $\hat{\tau}_i$ (e consequentemente, nos pesos $w_i(\tau)$). Assim, algumas das suposições dos modelos coincidem. Mais especificamente, as condições A1-A3 são as mesmas, enquanto que A4 e A5, que dizem respeito aos estimadores *kernel*, são substituídas por algumas suposições que dizem respeito a árvores, e que podem ser encontradas no artigo dos autores.

Assumindo que $(Y_i, \delta_i, \mathbf{x}_i), i = 1, \cdots, n$, são independentes e identicamente distribuídas com $\min(Y_i, C_i)$ independente de C_i condicional a $\mathbf{x}_i$, e sob algumas suposições adicionais, é possível mostrar que, se a estimativa $\boldsymbol{b}(\tau)$ de $\boldsymbol{\beta}(\tau)$ que minimiza a expressão

$$\sum_{i=1}^{n}\{w_i(\tau) \times \rho_\tau(\tilde{y}_i - \mathbf{x}_i^T\mathbf{b}(\tau)) + (1 - w_i(\tau)) \times \rho_\tau(y^{+\infty} - \mathbf{x}_i^T\mathbf{b}(\tau))\},$$

é tal que $\hat{\tau}_i$ de $w_i(\tau)$ é calculado usando a metodologia de árvore de sobrevivência, então

$$\mathbf{b}(\tau) \xrightarrow{P} \boldsymbol{\beta}(\tau),$$

em probabilidade, quando $n \to \infty$.

No entanto, conforme os autores discutem no artigo, a normalidade assintótica não pode ser provada diretamente, por envolver partição recursiva de uma amostra censurada. A maioria dos textos da área de árvores, ainda de acordo com os autores, foca em mostrar a consistência do estimador, mas a literatura ainda carece de estudos que tratem da convergência assintótica. Como alternativa para a inferência, no que diz respeito à significância dos parâmetros, a sugestão de Wey et al. (2014) também é utilizar a metodologia de *bootstrap* para construção

de intervalos de confiança, da mesma forma que é feito em Wang e Wang (2009). Em outras palavras, utilizando os $\alpha/2$ e $1-\alpha/2$ percentis dos coeficientes estimados via *bootstrap* para calcular os limites do intervalo.

Exemplo

Voltando ao exemplo de tempo de sobrevida de pacientes após infarto agudo do miocárdio, ajustando-se o modelo de regressão quantílica para o quantil de ordem $\tau = 0,50$ segundo a metodologia de Wang e Wang (2009), encontram-se as estimativas dos parâmetros do modelo apresentadas na Tabela 3.5. O cálculo foi feito com auxílio do *software* R, em que os pesos, estimados de acordo com o apresentado ao longo deste capítulo, foram inseridos na função *rq* do *software* R.

Tabela 3.5: Estimativas dos parâmetros calculadas com método de pesos locais via estimador *kernel* para quantil de ordem $\tau = 0,50$. Os erros padrão foram estimados via *bootstrap*.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	Wald	Valor p
Intercepto	11,30	81.116,69	0,85	13,35	<0,001
Sexo Feminino	-	-	-	-	-
Sexo Masculino	0,21	1,23	0,12	1,75	0,080
idade	-0,05	0,95	0,01	-4,29	<0,001

Observe que a conclusão inferencial é a mesma do modelo de Portnoy, com a observação de que, neste caso, a variável Sexo é marginalmente significativa de acordo com o teste de Wald. Os intervalos de confiança, baseados em reamostragem do conjunto de dados e nos quantis dessas estimativas, foram construídos e são apresentados a seguir, na Tabela 3.6. Para a sua construção, foram geradas 500 subamostras do conjunto de dados e para cada uma delas foi ajustada a metodologia em estudo. As estimativas de cada um dos parâmetros foram ordenadas e então se considerou os quantis de ordem $0,025$ e $0,975$ da distribuição das estimativas.

Tabela 3.6: Intervalos de confiança *bootstrap* para os parâmetros estimados via metodologia de pesos locais com estimadores *kernel*

Variável	Estimativa	IC(95%)
Intercepto	11,30	$[10,03; 11,65]$
Sexo Feminino	-	-
Sexo Masculino	0,21	$[-0,02; 0,36]$
idade	-0,05	$[-0,06; -0,03]$

Por outro lado, ajustando-se a metodologia de Wey et al. (2014), cujos pesos das observações censuradas são calculados via árvore de sobrevivência, encontram-se as estimativas para os parâmetros e respectivos intervalos de confiança apresentados na Tabela 3.7. Apesar de ser uma opção da função de regressão quantílica no *software* R, omitiu-se nesta dissertação o teste de Wald, visto que a literatura não demonstra a normalidade assintótica dos estimadores. O critério para a construção dos intervalos de confiança foi o mesmo que o adotado para a metodologia de Wang e Wang (2009), apresentado anteriormente. As estimativas dos parâmetros foram obtidas com o pacote *RPcrq*, disponível na página do autor e referenciada em seu artigo. Como critério de parada, considerou-se um número mínimo de 15 indivíduos em risco em cada ramo final da árvore de sobrevivência.

Tabela 3.7: Estimativas dos parâmetros calculadas com método de pesos locais via árvore de sobrevivência para quantil de ordem $\tau = 0,50$. Os erros padrão foram estimados via *bootstrap*.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	IC(95%)
Intercepto	9,65	15.569,36	0,18	$[9,35; 10,2]$
Sexo Feminino	-	-	-	-
Sexo Masculino	0,19	1,21	0,09	$[-0,04; 0,30]$
Idade	-0,03	0,97	0,00	$[-0,04; -0,02]$

Observe que os intervalos com coeficiente de confiança de 95% gerados com a metodologia de pesos locais com pesos estimados via árvore de sobrevivência apresentaram a menor amplitude, comparado com o modelo de pesos *kernel* e metodologia recursiva.

De um modo geral, na comparação das metodologias, observam-se as mesmas conclusões inferenciais, mas a de Portnoy foi a que apresentou o maior intervalo de confiança para os parâmetros entre as três abordagens. Como possível razão, atribuiu-se o fato de que a suposição de linearidade global é violada. Em relação às metodologias de pesos locais, observa-se que a amplitude dos intervalos é parecida, sendo a da metodologia de árvores de sobrevivência ligeiramente menor. Por outro lado, na análise dos intervalos de confiança parece haver mais evidências para considerar a variável Sexo como marginalmente significativa na abordagem com pesos locais de *kernel*.

Capítulo 4

Estudo de Simulação para Comparação das Metodologias para Dados Censurados

No Capítulo 3, foram apresentadas três metodologias para ajuste do modelo de regressão quantílica linear para dados censurados: o método recursivo de Portnoy (2003) e as metodologias de pesos locais de Wang e Wang (2009) e de Wey et al. (2014). Conforme discutido, o método recursivo de Portnoy diferencia-se dos outros dois por requerer a suposição de linearidade global, enquanto que as duas abordagens de pesos locais, que requerem a suposição de linearidade apenas no quantil de interesse, se diferenciam pela forma de estimação dos pesos das observações censuradas.

É de interesse comparar as três abordagens quando os dados seguem relação linear em todos os quantis condicionais, pois conforme será discutido na Seção 4.1, os principais modelos paramétricos utilizados para modelagem em Análise de Sobrevivência podem ser linearizados com a aplicação da função logarítmica na variável resposta. Além disso, é de interesse avaliar se, de fato, os três modelos são igualmente satisfatórios nesse contexto mais trivial. Um estudo similar é apresentado em Wey et al. (2014), mas restrito aos quantis $\tau = 0,25$ e $\tau = 0,50$. Apresenta-se nesta dissertação também a comparação para o quantil $\tau = 0,75$, completando-se a avaliação em todos os quartis condicionais da variável resposta.

Na literatura de regressão quantílica em geral, e em particular em Wey et al. (2014), os estudos de simulação para os parâmetros são feitos a partir de amostras aleatórias do próprio quantil de interesse: os parâmetros dependem de τ. Isto é, geram-se amostras aleatórias que seguem estrutura da forma:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + u_i,$$

com u_i tal que seu quantil de ordem τ seja igual a zero. Conforme discutido ao longo desta

dissertação, neste caso, o quantil de ordem τ para a variável aleatória Y_i, dado $\mathbf{x}_i$, é dado por:

$$Q_{Y_i|\mathbf{x}_i}(\tau) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau).$$

A simulação é feita dessa forma apenas por conveniência de avaliação do viés dos estimadores, ilustração da própria definição de regressão quantílica, e possibilidade de fixar diferentes percentuais de censura em cada um dos quantis condicionais para estudo. No entanto, também é possível gerar dados independentemente de τ, de modo mais intuitivo para entender como funciona regressão quantílica em aplicações práticas, mas com algumas suposições adicionais. Por exemplo, em regressão clássica é usual gerar amostras satisfazendo a estrutura:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

com ε_i erros aleatórios independentes e identicamente distribuídos com distribuição normal padrão, por exemplo, isto é, $\varepsilon_i \sim \mathcal{N}(0,1)$. Observe, neste caso, que:

$$F_{Y_i|\mathbf{x}_i}(y_i) = \tau \Rightarrow F^{-1}_{Y_i|\mathbf{x}_i}(\tau) = z_\tau + \mathbf{x}_i^T \boldsymbol{\beta},$$

em que z_τ é o quantil de ordem τ da distribuição normal padrão. Suponha, por exemplo, que se tenha apenas uma covariável no estudo. Neste caso, pode-se escrever:

$$Q_{Y_i|\mathbf{x}_i}(\tau) = \beta_0^* + \beta_1 x_i,$$

com $\beta_0^* = \beta_0 + z_\tau$. Observe que dados simulados dessa forma impõem que todos os quantis da variável resposta na presença das covariáveis são lineares. Essa abordagem para a geração dos dados é utilizada nesta dissertação.

Mais especificamente, considerou-se o seguinte modelo:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

em que $x_{1i} \sim$ Uniforme$(3,4)$, $x_{2i} \sim$ Binomial$(n; \frac{1}{2})$, $i = 1, \cdots, n$, e $\varepsilon_i \sim \mathcal{N}(0,1)$, com $\beta_0 = 4, \beta_1 = 2$ e $\beta_2 = 1$. Observe que, com a escolha desses parâmetros, a probabilidade de $\tilde{Y}_i$ ser negativa é muito pequena (considerando $x_{1i} = 3$ e $x_{2i} = 0$, por exemplo, a probabilidade de $\tilde{Y}_i$ ser negativa é $7,62\text{e}^{-24}$).

Para introduzir censura no modelo, foi gerada uma variável aleatória C_i com distribuição uniforme no intervalo $(0;50)$ e considerou-se como resposta a variável $\tilde{Y}_i$, definida como o mínimo entre Y_i e C_i, isto é, $\tilde{Y}_i = \min(Y_i, C_i)$. Dessa forma, a proporção de censura resultou em cerca de 25% para cada amostra. No total foram geradas 1.000 amostras com tamanhos $N = 400$ e $N = 800$ observações. Posteriormente, gerou-se novamente 1.000 amostras, também de tamanhos $N = 400$ e $N = 800$, mas com C_i com distribuição uniforme no intervalo $(0,25)$. Neste caso, o percentual de censura de $\tilde{Y}_i = \min(Y_i, C_i)$ resultou em cerca de 50% para cada amostra. Em todas as amostras observaram-se valores positivos para a variável $\tilde{Y}_i$, coerentemente com o contexto de Análise de Sobrevivência, em que os tempos são sempre positivos.

Em seguida, para cada uma das amostras aplicou-se as três metodologias avaliadas nos quartis da distribuição de Y_i condicional às covariáveis. Isto é, condicional às covariáveis, foram avaliados os quantis $\tau = (0,25; 0,50; 0,75)$. Para o ajuste da metodologia de Wang e Wang (2009) em cada uma das amostras calculou-se o *bandwidth* via validação cruzada para cada uma das variáveis, e então calculou-se a função *kernel* como o produto de funções *kernel* das covariáveis, conforme discutido no Apêndice A. Os pesos foram calculados conforme apresentado no capítulo anterior, e introduzidos na função *rq* do pacote *quantreg* do *software* R. Para a metodologia de Wey et al. (2014), considerou-se como regra de parada para construção das árvores de sobrevivência, além dos testes para a comparação das curvas de sobrevivência, o número mínimo de indivíduos em risco em cada ramo final da árvore igual a 15, que é a opção padrão implementada no pacote *RPcrq* disponibilizado pelos autores para o *software* R (que se encontra disponível online, na página referenciada em seu artigo). Os resultados das simulações para as amostras com proporção de censura 25% e 50% estão registrados, respectivamente, nas Tabelas 4.1 e 4.2. Nelas, são apresentados o viés, erro padrão e erro quadrático médio (EQM) estimados a partir da avaliação das 1.000 amostras (#Estimativas) de tamanhos N=400 e N=800, sendo que o erro quadrático médio foi calculado da seguinte forma:

$$EQM(\hat{\beta}(\tau)) = \sum_{i=1}^{\#\text{Estimativas}} \frac{(\hat{\beta}(\tau)^{(i)} - \beta(\tau))^2}{\#\text{Estimativas}}.$$

Tabela 4.1: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 25%, para a comparação das três metodologias de regressão quantílica para dados censurados, avaliadas nos quantis $\tau = 0,25$; $0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Erro Padrão			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	1.000	-0,065	0,018	-0,011	0,918	0,262	0,149	0,847	0,069	0,022
		Wang & Wang	1.000	-0,068	0,020	-0,010	0,932	0,265	0,150	0,873	0,071	0,023
		Wey et al.	1.000	-0,058	0,018	-0,010	0,929	0,265	0,152	0,866	0,070	0,023
	0,50	Portnoy	1.000	-0,082	0,022	-0,001	0,839	0,239	0,140	0,710	0,058	0,020
		Wang & Wang	1.000	-0,074	0,020	<0,001	0,839	0,239	0,141	0,709	0,058	0,020
		Wey et al.	1.000	-0,072	0,020	<0,001	0,841	0,240	0,140	0,711	0,058	0,020
	0,75	Portnoy	1.000	-0,075	0,021	-0,004	0,922	0,264	0,156	0,855	0,070	0,024
		Wang & Wang	1.000	-0,071	0,021	-0,005	0,927	0,266	0,159	0,863	0,071	0,025
		Wey et al.	1.000	-0,036	0,011	-0,013	0,929	0,266	0,157	0,863	0,071	0,025
800	0,25	Portnoy	1.000	0,023	-0,008	0,003	0,662	0,187	0,104	0,439	0,035	0,011
		Wang & Wang	1.000	0,025	-0,007	0,003	0,661	0,187	0,106	0,437	0,035	0,011
		Wey et al.	1.000	0,029	-0,008	0,003	0,666	0,189	0,106	0,444	0,036	0,011
	0,50	Portnoy	1.000	0,026	-0,008	-0,001	0,599	0,170	0,100	0,359	0,029	0,010
		Wang & Wang	1.000	0,030	-0,008	-0,001	0,603	0,171	0,101	0,364	0,029	0,010
		Wey et al.	1.000	0,031	-0,008	-0,002	0,604	0,171	0,101	0,366	0,029	0,010
	0,75	Portnoy	1.000	0,021	-0,007	0,002	0,657	0,185	0,109	0,431	0,034	0,012
		Wang & Wang	1.000	0,018	-0,005	0,003	0,666	0,188	0,110	0,444	0,035	0,012
		Wey et al.	1.000	0,049	-0,013	-0,005	0,667	0,189	0,110	0,447	0,036	0,012

Para ajudar na visualização dos resultados da simulação, foram construídos gráficos com viés, erro padrão e erro quadrático médio de cada um dos parâmetros em cada um dos quartis. Esses gráficos são apresentados no Apêndice C, nas Figuras C.1, C.2, C.3 e C.4.

De um modo geral, observa-se que os três modelos, conforme esperado, ajustaram bem os dados e são bastante semelhantes quanto à avaliação do viés, erro padrão e erro quadrático médio. Apenas o de Wey et al. (2014) avaliado no quantil $\tau = 0,75$ para a proporção de censura igual a 50% apresentou estimativas dos parâmetros com viés maior do que as outras duas metodologias. Neste cenário, ajustou-se nova simulação para a metodologia de Wey et al. (2014), mas considerando número maior de indivíduos em risco nos ramos finais da árvore de sobrevivência (foram considerados 30 e 60 indivíduos em risco). No entanto, os resultados ficaram bastante parecidos com o apresentado na Tabela 4.2, e por isso foram omitidos nessa dissertação.

Tabela 4.2: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 50%, para a comparação das três metodologias de regressão quantílica para dados censurados, avaliadas nos quantis $\tau = 0,25; 0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Erro Padrão			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	1.000	0,055	-0,016	-0,009	1,112	0,315	0,178	1,238	0,099	0,032
		Wang & Wang	1.000	0,057	-0,012	-0,008	1,127	0,320	0,179	1,273	0,102	0,032
		Wey et al.	1.000	0,065	-0,013	-0,008	1,116	0,316	0,180	1,248	0,100	0,032
	0,50	Portnoy	1.000	0,068	-0,020	-0,008	1,010	0,289	0,164	1,024	0,084	0,027
		Wang & Wang	1.000	0,073	-0,019	-0,006	1,018	0,291	0,164	1,040	0,085	0,027
		Wey et al.	1.000	0,086	-0,022	-0,011	1,003	0,287	0,165	1,013	0,083	0,027
	0,75	Portnoy	1.000	0,052	-0,014	-0,006	1,168	0,331	0,191	1,365	0,110	0,036
		Wang & Wang	1.000	0,066	-0,016	-0,005	1,185	0,336	0,192	1,406	0,113	0,037
		Wey et al.	1.000	0,317	-0,086	-0,042	1,183	0,337	0,196	1,499	0,121	0,040
800	0,25	Portnoy	1.000	-0,015	0,004	0,000	0,771	0,220	0,121	0,594	0,048	0,015
		Wang & Wang	1.000	-0,017	0,008	0,003	0,764	0,218	0,123	0,583	0,048	0,015
		Wey et al.	1.000	-0,010	0,008	0,003	0,771	0,220	0,123	0,594	0,048	0,015
	0,50	Portnoy	1.000	-0,019	0,006	0,002	0,719	0,205	0,118	0,517	0,042	0,014
		Wang & Wang	1.000	-0,017	0,007	0,002	0,717	0,204	0,118	0,514	0,042	0,014
		Wey et al.	1.000	0,001	0,003	-0,001	0,720	0,205	0,118	0,518	0,042	0,014
	0,75	Portnoy	1.000	0,041	-0,010	-0,007	0,776	0,221	0,128	0,603	0,049	0,017
		Wang & Wang	1.000	0,042	-0,008	-0,006	0,786	0,224	0,130	0,619	0,050	0,017
		Wey et al.	1.000	0,283	-0,075	-0,040	0,805	0,230	0,134	0,727	0,058	0,019

De um modo geral, observa-se nas Tabelas 4.1 e 4.2 anteriores que, conforme esperado, o aumento da proporção de censura implica no aumento do erro padrão dos estimadores, e este diminui ao se aumentar o tamanho da amostra.

4.1 Linearização dos Principais Modelos Paramétricos em Análise de Sobrevivência

O tempo de sobrevivência, objeto de estudo de Análise de Sobrevivência, em geral não segue uma relação linear com as covariáveis. Neste caso, é inviável ajustar as técnicas apresentadas nesta dissertação, que assumem linearidade da variável resposta em seus quantis condicionais. Mesmo assim, ao menos para os principais modelos paramétricos utilizados, é possível transformar a variável resposta de modo que os quantis passem a ser lineares. Relembre que a regressão quantílica tem a propriedade de equivariância sob transformações

monótonas que, diferentemente da regressão linear clássica, permite interpretar os parâmetros após transformações de escala, por exemplo.

A seguir são listados os principais modelos paramétricos utilizados em Análise de Sobrevivência e a fórmula de seus quantis. Recomenda-se a leitura de Colosimo e Giolo (2006) para maiores detalhes.

- **Modelo de Regressão Exponencial**

  Se Y_i segue um modelo exponencial dadas as covariáveis, então, a função de distribuição acumulada de Y_i, dado $\mathbf{x}_i$ pode ser expressa como:

$$F_{Y_i|\mathbf{x}_i}(y_i) = 1 - \exp\left(-\frac{y_i}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}\right).$$

  Claramente, os quantis de Y_i não são lineares, pois:

$$F_{Y_i|\mathbf{x}_i}(y_i) = \tau \Rightarrow y_i = -\exp(\mathbf{x}_i^T\boldsymbol{\beta})\log(1-\tau).$$

  No entanto, observe que $Z_i = \log(Y_i)$ segue distribuição valor extremo, e neste caso:

$$F_{Z_i|\mathbf{x}_i}(z_i) = 1 - \exp(-\exp(z_i - \mathbf{x}_i^T\boldsymbol{\beta})),$$

  com

$$F_{Z_i|\mathbf{x}_i}(z_i) = \tau \Rightarrow z_i = \mathbf{x}_i^T\boldsymbol{\beta} + \log(-\log(1-\tau)),$$

  que é linear nos parâmetros. Ou seja, sob a transformação logarítmica é possível ajustar um modelo de regressão quantílica linear para dados que seguem um modelo exponencial.

- **Modelo de Regressão Weibull**

  Da mesma forma que no caso exponencial, para o modelo de regressão Weibull tem-se que:

$$F_{Y_i|\mathbf{x}_i}(y_i) = 1 - \exp\left(-\left(\frac{y}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}\right)^{1/\sigma}\right),$$

  e

$$F_{Y_i|\mathbf{x}_i}(y_i) = \tau \Rightarrow y_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})(-\log(1-\tau))^{\sigma}.$$

  Agora, para $Z_i = \log(Y_i)$, observe que:

$$F_{Z_i|\mathbf{x}_i}(z_i) = 1 - \exp\left(-\exp\left(\frac{z_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right),$$

e

$$F_{Z_i|\mathbf{x}_i}(z_i) = \tau \Rightarrow z_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma \log(-\log(1-\tau)).$$

Ou seja, sob a transformação logarítmica é possível ajustar um modelo de regressão quantílica linear para dados que seguem um modelo Weibull.

- **Modelo de Regressão Log-Logístico**

  Observe que se Y_i dado $\mathbf{x}_i$ tem distribuição log-logística, então:

$$F_{Y_i|\mathbf{x}_i}(y_i) = 1 - \frac{1}{1 + (y_i/\mathbf{x}_i^T\boldsymbol{\beta})^{\gamma}}.$$

  Seus quantis de ordem τ são dados por

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}\left(\frac{\tau}{1-\tau}\right)^{1/\gamma}.$$

  Ou seja, para dados que seguem distribuição log-logística, dadas as covariáveis, é possível usar a metodologia de regressão quantílica linear, pois seus quantis são lineares.

Um estudo de simulação para dados com distribuição Weibull e log-logística é apresentado a seguir, considerando-se ainda os dois cenários de censura abordados anteriormente, isto é, proporção igual a 25% e 50%, também para tamanhos de amostra iguais a $N = 400$ e $N = 800$. Os resultados foram obtidos a partir da estimação dos parâmetros da regressão quantílica para $n = 1.000$ amostras aleatórias em cada cenário, obtidas com auxílio do *software* R. Mais especificamente, foram considerados os seguintes modelos:

1. *Y_i com distribuição Weibull com parâmetros $\sigma = 2$; $\beta_0 = 1,5; \beta_1 = 0,50$ e $\beta_2 = 1,3, i = 1, \cdots, 1.000$, e com x_{1i} uniformemente distribuída no intervalo $(0,1)$ e x_{2i} com distribuição Bernoulli com probabilidade de sucesso igual a 0,50.*

   Para introduzir censura no modelo, foi gerada uma variável aleatória C_i com distribuição uniforme no intervalo $(0; \exp(3,9))$ para proporção de censura na amostra de 25% e distribuição uniforme no intervalo $(0; \exp(3,14))$ para obter cerca de 50% de observações censuradas. Aplicou-se transformação logarítmica na variável resposta $\tilde{Y}_i = \min(Y_i, C_i)$ para ajuste dos parâmetros, para linearização do modelo. Os resultados da avaliação do viés, desvio padrão e erro quadrático médio são apresentados, respectivamente, nas Tabelas 4.3 e 4.4 e nas Figuras C.5, C.6, C.7 e C.8 do Apêndice C.

   Na estimação dos parâmetros, o algoritmo de Portnoy disponível no *software* R apresentou erro para algumas amostras e, por isso, o número de estimativas não é sempre igual a 1.000. Por outro lado, a metodologia de árvores de sobrevivência para o quantil de ordem $\tau = 0,75$ quando a proporção de censura é 50% não apresentou resultados satisfatórios para nenhuma das amostras, mas estimativas de viés e erro padrão bastante inflacionadas. Observe na Tabela 4.4 que estes casos são apresentados com ∞.

De um modo geral, para o cenário com proporção de censura de 25% os resultados das três abordagens são bastante semelhantes, sendo que a de Wey et al. (2014) mostrou-se um pouco mais viesada que as demais na avaliação do quantil de ordem $\tau = 0,75$. Para a proporção de censura de 50%, no entanto, a metodologia recursiva de Portnoy foi a que apresentou menores estimativas de viés, desvio padrão e erro quadrático médio.

Tabela 4.3: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 25%, para a comparação das três metodologias de regressão quantílica para dados censurados da distribuição Weibull, avaliadas nos quantis $\tau = 0,25; 0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Desvio			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	1.000	-0,001	0,001	-0,001	0,057	0,091	0,054	0,003	0,008	0,003
		Wang & Wang	1.000	-0,001	0,005	0,010	0,060	0,094	0,056	0,004	0,009	0,003
		Wey et al.	1.000	<0,001	0,005	0,011	0,060	0,095	0,056	0,004	0,009	0,003
	0,50	Portnoy	1.000	-0,001	<0,001	0,001	0,043	0,071	0,040	0,002	0,005	0,002
		Wang & Wang	1.000	-0,001	0,002	0,006	0,044	0,072	0,041	0,002	0,005	0,002
		Wey et al.	1.000	<0,001	0,002	0,006	0,044	0,072	0,041	0,002	0,005	0,002
	0,75	Portnoy	1.000	<0,001	-0,001	0,001	0,037	0,062	0,036	0,001	0,004	0,001
		Wang & Wang	1.000	<0,001	0,001	0,005	0,038	0,062	0,037	0,001	0,004	0,001
		Wey et al.	1.000	0,011	-0,017	-0,064	0,040	0,069	0,046	0,002	0,005	0,006
800	0,25	Portnoy	1.000	0,003	-0,005	<0,001	0,043	0,067	0,041	0,002	0,004	0,002
		Wang & Wang	1.000	0,003	-0,001	0,007	0,043	0,067	0,041	0,002	0,004	0,002
		Wey et al.	1.000	0,003	0,001	0,011	0,043	0,067	0,041	0,002	0,005	0,002
	0,50	Portnoy	1.000	0,002	-0,003	-0,001	0,031	0,048	0,029	0,001	0,002	0,001
		Wang & Wang	1.000	0,002	-0,001	0,002	0,031	0,048	0,029	0,001	0,002	0,001
		Wey et al.	1.000	0,003	-0,002	0,003	0,031	0,048	0,029	0,001	0,002	0,001
	0,75	Portnoy	1.000	<0,001	0,001	<0,001	0,026	0,043	0,026	0,001	0,002	0,001
		Wang & Wang	1.000	0,000	0,003	0,003	0,027	0,043	0,026	0,001	0,002	0,001
		Wey et al.	1.000	0,011	-0,014	-0,062	0,029	0,048	0,033	0,001	0,002	0,005

Tabela 4.4: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 50%, para a comparação das três metodologias de regressão quantílica para dados censurados da distribuição Weibull, avaliadas nos quantis $\tau = 0,25; 0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Desvio			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	997	0,002	0,001	-0,001	0,067	0,110	0,065	0,004	0,012	0,004
		Wang & Wang	1.000	-0,003	0,023	0,039	0,067	0,110	0,065	0,004	0,013	0,006
		Wey et al.	1.000	-0,096	0,226	0,246	0,072	0,138	0,107	0,014	0,070	0,072
	0,50	Portnoy	997	0,002	-0,002	-0,003	0,052	0,088	0,060	0,003	0,008	0,004
		Wang & Wang	1.000	-0,006	0,020	0,035	0,052	0,088	0,054	0,003	0,008	0,004
		Wey et al.	1.000	-0,060	0,139	0,199	0,051	0,088	0,077	0,006	0,027	0,045
	0,75	Portnoy	852	0,003	-0,001	0,005	0,047	0,081	0,066	0,002	0,007	0,004
		Wang & Wang	1.000	-0,004	0,015	0,045	0,049	0,085	0,060	0,002	0,007	0,006
		Wey et al.	1.000	-0,152	0,313	∞	0,076	0,137	∞	0,029	0,116	∞
800	0,25	Portnoy	1.000	<0,001	-0,002	-0,002	0,046	0,078	0,049	0,002	0,006	0,002
		Wang & Wang	1.000	-0,004	0,017	0,029	0,046	0,078	0,048	0,002	0,006	0,003
		Wey et al.	1.000	-0,115	0,263	0,246	0,050	0,097	0,075	0,016	0,078	0,066
	0,50	Portnoy	1.000	0,001	<0,001	-0,003	0,036	0,061	0,041	0,001	0,004	0,002
		Wang & Wang	1.000	-0,005	0,016	0,027	0,035	0,061	0,037	0,001	0,004	0,002
		Wey et al.	1.000	-0,060	0,138	0,212	0,034	0,062	0,055	0,005	0,023	0,048
	0,75	Portnoy	960	0,001	-0,002	-0,003	0,034	0,059	0,052	0,001	0,003	0,003
		Wang & Wang	1.000	-0,003	0,009	0,032	0,035	0,060	0,045	0,001	0,004	0,003
		Wey et al.	1.000	-0,158	0,320	∞	0,057	0,101	∞	0,028	0,113	∞

2. *Y_i tem distribuição log-logística com parâmetros $\sigma = 2$; $\beta_0 = 0,3$; $\beta_1 = 1$ e $\beta_2 = 0,1$, $i = 1, \cdots, 1.000$, e com x_{1i} uniformemente distribuída no intervalo $(2, 3)$ e x_{2i} com distribuição Bernoulli com probabilidade de sucesso igual a 0,50.*

   Para introduzir censura no modelo, foi gerada uma variável aleatória C_i com distribuição uniforme no intervalo $(0; 16)$ para proporção de censura de 25% na amostra e distribuição uniforme no intervalo $(0; 6, 5)$ para 50%. Relembre que, quando os dados seguem distribuição log-logística com a parametrização apresentada nesta seção, então seus quantis seguem relação linear, não sendo necessária a aplicação da transformação logarítmica. Os resultados do estudo de simulação são apresentados, respectivamente, nas Tabelas 4.5 e 4.6 e nas Figuras C.9, C.10, C.11 e C.12 do Apêndice C.

   Similar à modelagem para distribuição Weibull, na estimação dos parâmetros do modelo log-logístico, o algoritmo de Portnoy disponível no *software* R apresentou erro para algumas amostras e, por isso, o número de estimativas não é sempre igual a 1.000.

   Para algumas amostras modeladas com a metodologia de árvore de sobrevivência, a estimativa dos parâmetros mostrou-se bastante inflacionada comparada com as demais. Estabeleceu-se um valor máximo para cada uma das estimativas igual a 10, um valor atribuído arbitrariamente apenas para desconsiderar estimativas extremante discrepantes, e avaliou-se o resultado das três abordagens para as estimativas inferiores ao ponto de corte. Observe, por exemplo, que todas as estimativas da abordagem de pesos locais de *kernel* foram inferiores ao ponto de corte, ao passo que na metodologia de árvores de sobrevivência isso nem sempre acontece.

   De um modo geral, as metodologias são bastante semelhantes, sendo que a de pesos locais estimados via árvore de sobrevivência apresenta estimativas mais viesadas, comparada às outras, para o quantil 0,75.

Apesar da metodologia de Wey et al. (2014) apresentar o pior desempenho em quantis altos, os resultados para os quantis 0,25 e 0,50 se mostram bastante razoáveis para aplicações práticas. Em seu artigo, os autores apresentam um estudo de simulação para justificar o uso da metodologia quando os quantis inferiores não seguem distribuição linear. Em relação à metodologia de pesos locais estimados via *kernel*, o método de árvores de sobrevivência é mais indicado quando na presença de muitas covariáveis, situação comum nas aplicações práticas.

Tabela 4.5: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 25%, para a comparação das três metodologias de regressão quantílica para dados censurados da distribuição log-logística, avaliadas nos quantis $\tau = 0,25; 0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Desvio			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	1.000	-0,015	-0,001	0,007	0,821	0,334	0,191	0,674	0,112	0,037
		Wang & Wang	1.000	-0,023	0,012	0,009	0,826	0,335	0,190	0,682	0,112	0,036
		Wey et al.	1.000	-0,035	0,018	0,009	0,818	0,332	0,188	0,670	0,110	0,035
	0,50	Portnoy	1.000	0,050	-0,029	0,003	1,324	0,537	0,296	1,753	0,289	0,088
		Wang & Wang	1.000	0,006	0,001	0,010	1,295	0,524	0,292	1,676	0,274	0,085
		Wey et al.	1.000	0,009	0,003	0,011	1,298	0,526	0,288	1,684	0,277	0,083
	0,75	Portnoy	1.000	0,004	-0,022	-0,009	2,639	1,072	0,621	6,960	1,148	0,386
		Wang & Wang	1.000	-0,136	0,078	-0,004	2,501	1,013	0,582	6,269	1,031	0,339
		Wey et al.	1.000	0,148	-0,053	-0,013	2,597	1,056	0,600	6,761	1,117	0,359
800	0,25	Portnoy	1.000	-0,023	0,008	-0,001	0,608	0,245	0,144	0,370	0,060	0,021
		Wang & Wang	1.000	-0,039	0,020	0,001	0,605	0,243	0,143	0,368	0,060	0,020
		Wey et al.	1.000	-0,038	0,023	0,002	0,598	0,241	0,138	0,359	0,058	0,019
	0,50	Portnoy	1.000	<0.001	-0,002	-0,002	0,937	0,378	0,209	0,877	0,142	0,044
		Wang & Wang	1.000	-0,022	0,018	-0,001	0,913	0,368	0,206	0,834	0,136	0,042
		Wey et al.	1.000	-0,032	0,028	-0,001	0,913	0,367	0,205	0,835	0,136	0,042
	0,75	Portnoy	1.000	0,004	-0,013	-0,005	1,996	0,809	0,444	3,982	0,653	0,197
		Wang & Wang	1.000	-0,057	0,042	0,001	1,915	0,775	0,424	3,665	0,601	0,180
		Wey et al.	1.000	0,161	-0,052	-0,014	1,964	0,795	0,437	3,881	0,635	0,191

Tabela 4.6: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos $N = 400$ e $N = 800$, com proporção de censura igual a 50%, para a comparação das três metodologias de regressão quantílica para dados censurados da distribuição log-logística, avaliadas nos quantis $\tau = 0,25; 0,50$ e $0,75$.

N	τ	Método	#Estimativas	Viés			Desvio			EQM		
				$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$	$b_0(\tau)$	$b_1(\tau)$	$b_2(\tau)$
400	0,25	Portnoy	1.000	-0,039	0,010	-0,002	0,888	0,357	0,204	0,790	0,127	0,041
		Wang & Wang	1.000	-0,087	0,054	0,003	0,861	0,345	0,200	0,748	0,122	0,040
		Wey et al.	1.000	-0,067	0,047	0,006	0,866	0,347	0,195	0,754	0,123	0,038
	0,50	Portnoy	1.000	-0,108	0,032	0,001	1,469	0,592	0,343	2,169	0,351	0,117
		Wang & Wang	1.000	-0,223	0,124	0,004	1,370	0,548	0,310	1,925	0,316	0,096
		Wey et al.	1.000	-0,188	0,093	0,005	1,339	0,534	0,309	1,826	0,293	0,095
	0,75	Portnoy	769	0,065	-0,065	0,040	3,474	1,717	0,790	12,055	2,948	0,624
		Wang & Wang	998	-0,784	0,464	0,041	2,496	1,009	0,594	6,840	1,232	0,355
		Wey et al.	941	0,684	-0,281	-0,032	3,155	1,510	0,781	10,408	2,355	0,611
800	0,25	Portnoy	1.000	0,029	-0,016	0,002	0,637	0,262	0,146	0,406	0,069	0,021
		Wang & Wang	1.000	-0,019	0,021	0,005	0,617	0,253	0,140	0,380	0,064	0,020
		Wey et al.	1.000	-0,035	0,036	0,008	0,603	0,247	0,134	0,365	0,062	0,018
	0,50	Portnoy	1.000	0,016	-0,012	0,001	1,052	0,433	0,238	1,106	0,187	0,056
		Wang & Wang	1.000	-0,095	0,065	0,009	0,957	0,393	0,221	0,924	0,158	0,049
		Wey et al.	1.000	-0,067	0,047	0,006	0,924	0,374	0,211	0,857	0,142	0,044
	0,75	Portnoy	945	0,225	-0,135	0,028	2,898	1,241	0,587	8,440	1,556	0,346
		Wang & Wang	1.000	-0,634	0,365	0,054	1,907	0,764	0,446	4,035	0,717	0,202
		Wey et al.	986	1,026	-0,356	-0,016	2,237	0,949	0,545	6,051	1,027	0,297

Capítulo 5

Aplicação a Dados Clínicos

O conjunto de dados que motivou esta dissertação é um estudo do Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo. O objetivo era avaliar a relação entre o tempo de sobrevivência de pacientes diagnosticados com insuficiência cardíaca e uma série de variáveis demográficas e clínicas disponíveis, em especial da taxa de linfócitos. Isso porque, conforme citado em Botter et al. (2012), diferentes estudos na literatura indicam que pacientes com insuficiência cardíaca podem sofrer alterações em diferentes funções orgânicas. Em particular, tem-se interesse em interpretar, em termos da mediana da distribuição dos tempos, a variação no número de dias de sobrevida dos pacientes em função da variação dos valores das covariáveis. A seguir são listadas as variáveis que foram analisadas nesta dissertação com suas descrições e codificações entre parênteses.

Variáveis demográficas:

- **Idade** (em anos): Idade do paciente, calculada com referência à data da primeira consulta.

- **Sexo**: Masculino (M) ou Feminino (F).

Variáveis Clínicas:

- **IMC** (em Kg/m^2): Índice de massa corporal, assumindo os valores: (1) $18{,}5 \leq \text{IMC} < 25$; (2) $\text{IMC} < 18{,}5$; (3) $\text{IMC} \geq 25$ e (4) IMC não avaliado.

- **PAD** (em mmHg): Pressão arterial diastólica, classificada nos seguintes grupos: (1) $40 \leq \text{PAD} < 85$; (2) $85 \leq \text{PAD} < 89$; (3) $89 \leq \text{PAD} < 98$; (4) $98 \leq \text{PAD} < 108$; (5) $\text{PAD} \geq 108$ e (NA) PAD não avaliada.

- **Hipertensão**: Hipertensão arterial, categorizada em três grupos: (1) apresenta hipertensão arterial; (2) não apresenta hipertensão arterial e (NA) hipertensão arterial não avaliada.

- **Classe Funcional**: Classe Funcional do paciente, dividida nos grupos (1) Classe Funcional I; (2) Classe Funcional II; (3) Classe Funcional III; (4) Classe Funcional IV e (NA) Classe Funcional não avaliada.
- **DAE** (em mm): Diâmetro do átrio esquerdo. Classificada nos seguintes grupos: (1) $0 \leq$ DAE ≤ 40; (2) DAE > 40; (3) DAE não avaliada.
- **DDVE** (em mm): Diâmetro diastólico do ventrículo esquerdo. Dividido nos grupos: (1) DDVE ≤ 42; (2) DDVE > 42 e (NA) DDVE não avaliado.
- **DSVE** (em mm): Diâmetro sistólico do ventrículo esquerdo. Dividido nos grupos: (1) DSVE ≤ 39; (2) DSVE > 39 e (NA) DSVE não avaliado.
- **Linfócitos** (em céls/mm^3): taxa de linfócitos, classificada nos seguintes grupos: (1) Linfócitos < 900; (2) Linfócitos ≥ 900 e (NA) taxa de linfócitos não avaliada.
- **HDL** (em mg/dL): Nível de HDL, dividido nas categorias: (1) HDL < 40; (2) HDL ≥ 40 e (NA) HDL não avaliada.
- **Creatinina** (em mEq/L): Taxa de creatinina sérica, assumindo as categorias: (1) Creatinina $< 0{,}8$; (2) $0{,}8 \leq$ Creatinina $< 1{,}3$; (3) Creatinina $\geq 1{,}3$ e (NA) Creatinina não avaliada.

Variáveis de prognóstico:

- **Evento**: (1) Óbito é observado e (0) óbito não é observado (censura).
- **Tempo de vida** (em dias): diferença entre a data da primeira consulta e a data de óbito. Para pacientes que não tiveram óbito, foi considerado o tempo decorrido entre a data da primeira consulta e a data da última consulta do paciente.

O conjunto de dados é composto por 3.139 pacientes, dos quais 1.386 apresentaram o evento morte. Considera-se que os demais pacientes estão sob censura aleatória à direita. Note que a não observação do evento morte não está associada ao término do estudo, mas a última consulta do paciente, e por isso a censura está sendo tratada como aleatória.

Todas as covariáveis, com exceção da Idade, encontram-se categorizadas em função da grande quantidade de observações que não tiveram avaliação, que foram então alocadas no grupo "NA", referente a valores faltantes das covariáveis. A variável Linfócito, por exemplo, apresenta cerca de 50% das observações sem avaliação.

Inicialmente foram construídos os gráficos de Kaplan-Meier para cada uma das covariáveis, mas nem todos são apresentados nesta dissertação. A Figura 5.1 apresenta as curvas de sobrevivência para as variáveis Linfócitos, PAD e HDL, selecionadas apenas como exemplo. Observe que para a variável PAD as curvas de sobrevivência se cruzam, o que é um indicativo de que os riscos para essa variável não são proporcionais. Nos três gráficos parece haver uma relação entre os registros não avaliados e o tempo de sobrevivência; para o grupo em que não foram avaliadas as variáveis, o tempo de sobrevivência é menor do que para o grupo em que foram observadas essas medidas. De um modo geral, esse comportamento é observado para todas as covariáveis.

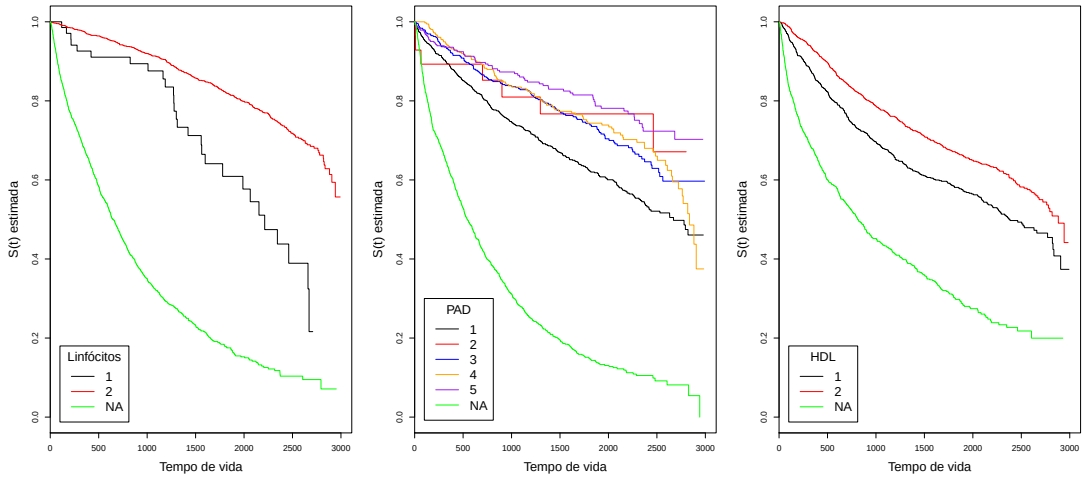

Figura 5.1: Curvas de Kaplan-Meier para as variáveis Linfócitos, PAD e HDL, respectivamente, do estudo do tempo de sobrevivência de pacientes diagnosticados com insuficiência cardíaca.

5.1 Análises Inferenciais

Para a inferência sobre a relação do tempo de sobrevivência de pacientes diagnosticados com insuficiência cardíaca e as covariáveis disponíveis, propôs-se inicialmente a análise dos dados via modelo de riscos proporcionais de Cox que, conforme já discutido, é o mais usual em Análise de Sobrevivência, e também foi a técnica utilizada em Botter et al. (2012). As covariáveis categóricas foram transformadas em variáveis dummies, para a construção de modelos do tipo casela de referência. As estimativas do modelo final de Cox são apresentadas na Tabela 5.1. No modelo, foram testadas outras interações, a saber: DDVE, DSVE e DAE com Sexo e com Idade, mas tais parâmetros não se mostraram significativos. Os passos intermediários para seleção do modelo são omitidos nesta dissertação.

Para avaliar a qualidade do ajuste do modelo final de Cox no que diz respeito à suposição de riscos proporcionais, foram gerados os resíduos de Schoenfeld (Colosimo e Giolo, 2006), utilizando a transformação identidade. As transformações *rank* e de Kaplan-Meier também foram avaliadas à título de validação dos resultados. No entanto, em todos os casos, a análise dos resíduos revelou que a suposição global de riscos proporcionais do modelo é violada. Conforme discutido em Colosimo e Giolo (2006), a violação desta suposição pode implicar em vícios nas estimações dos coeficientes, e que por isso podem não ser confiáveis. O coeficiente de correlação de Pearson ρ entre os resíduos padronizados de Schoenfeld e a função $g(y) = y$, o teste estatístico para hipótese de proporcionalidade dos riscos de cada covariável individualmente, e o teste conjunto considerando todas as covariáveis são apresentados na Tabela 5.2.

Alternativamente ao modelo de Cox, propôs-se estudar os modelos de vida acelerados. Assim, ajustaram-se modelos considerando as distribuições: exponencial e Weibull. No entanto, conforme esperado, os ajustes com essas distribuições não ficaram adequados, justamente por se observar a não proporcionalidade dos riscos no conjunto de dados. Em geral, essas distribuições são adequadas quando a proporcionalidade dos riscos não é violada. Nesse contexto,

distribuições recomendadas são a log-normal e a log-logística. No entanto, a análise de resíduos para essas duas distribuições também revelou falta de ajuste dos modelos, que foram então omitidos desta dissertação.

Modelos de vida acelerado têm a desvantagem de requererem o conhecimento sobre a distribuição dos dados. Na prática, e em particular neste caso, em que as distribuições mais conhecidas não se mostraram adequadas, nem sempre é fácil encontrar uma distribuição que produza um ajuste satisfatório. Como o objetivo do estudo em análise também é entender qual a variação no tempo de sobrevivência mediano ao se alterar o valor das covariáveis, optou-se por avaliar a regressão quantílica para dados censurados.

Nesta dissertação foram apresentados três modelos. Relembre nos exemplos apresentados no Capítulo 3 que a suposição de linearidade global, necessária para ajuste da metodologia recursiva de Portnoy (2003), é bastante restritiva, principalmente em quantis menores, em que se observam poucos dados. Nesse sentido, as metodologias de pesos locais são muito mais flexíveis, e atendem ao objetivo da análise, que é avaliar a relação entre o tempo de vida e as covariáveis em apenas um quantil, $\tau = 0,50$.

Por outro lado, a metodologia de pesos locais de Wang e Wang (2009) apresenta algumas dificuldades associadas ao uso de suavizador *kernel* na estimação dos pesos das observações censuradas. Conforme já discutido, a função *kernel* não é recomendada para problemas com muitas covariáveis, como é o caso deste conjunto de dados.

Nesse sentido, para o ajuste do modelo foi proposta a utilização da metodologia de pesos locais de Wey et al. (2014), com pesos estimados via árvore de sobrevivência. No entanto, o ajuste para a mediana resultou em erros padrão extremamente altos. Optou-se então pela escolha do quantil 0,45, que produz estimativas não inflacionadas do erro padrão. Aqui, a metodologia de Portnoy (2003) é apresentada apenas como referência. Os resultados do ajuste do modelo completo de pesos locais via árvore de sobrevivência e método recursivo de Portnoy são apresentados, respectivamente, nas Tabelas 5.3 e 5.4.

Uma importante observação deve ser feita sobre a modelagem da regressão quantílica para esse conjunto de dados. Assim como seria natural em grande parte dos modelos de vida acelerado, para a modelagem de regressão quantílica também foi utilizada a transformação logarítmica da variável resposta, com a vantagem de ser uma metodologia invariante sob transformações monótonas. Optou-se por usar a escala logarítmica já que estudos da área mostram que os tempos de sobrevivência, em geral, não seguem relação linear com as covariáveis, conforme discutido anteriormente.

Comparando os ajustes da modelagem de pesos locais via árvore de sobrevivência com a metodologia recursiva, observa-se que esta última falha em detectar a significância de alguns parâmetros, como PAD e HDL, por exemplo. Observa-se, também, um peso muito maior para o intercepto. No entanto, de um modo geral, o sinal das estimativas dos parâmetros é respeitado - não é observada nenhuma inversão entre os parâmetros significativos.

Uma das grandes desvantagens da regressão quantílica para dados censurados é que pouco foi desenvolvido para a inferência dos modelos finais. Não é possível fazer, por exemplo, testes conjuntos para os parâmetros dos modelos, que é um procedimento necessário para seleção de

modelos. Mesmo assim, para simplificar o modelo final, foi proposta a retirada de variáveis cujos intervalos de confiança para os parâmetros contivessem o valor zero, considerando o coeficiente de confiança igual a 95%.

A primeira variável retirada da análise foi a interação entre DDVE e Sexo, pois considerando coeficiente de confiança de 95%, a estimativa do seu parâmetro ficou em um intervalo que contém o valor zero - que neste caso é considerada como não significativa. Observe que esses parâmetros são apenas marginalmente significativos na análise tradicional de Cox. No modelo sem a interação, retirou-se uma a uma as variáveis Idade e Sexo. Posteriormente, avaliaram-se quais eram as variáveis que apresentavam todas as categorias não significativas (de acordo com o critério do intervalo de confiança discutido anteriormente). As variáveis DSVE e DDVE apresentavam essas características, e também foram, uma a uma, retiradas da análise. Relembre que o objetivo é entender a relação das covariáveis com o tempo de sobrevivência, e não fazer qualquer tipo de previsão. Foram propostas também outras ordenações para a retirada das variáveis, por exemplo, retirando primeiro a variável Idade e mantendo-se a interação. No entanto, os ajustes produzidos nessa primeira etapa mostraram-se, em sua maioria, não significativos, e por isso optou-se por essa ordenação mencionada para redução do modelo.

O modelo final é apresentado na Tabela 5.5, em que pelo menos uma das categorias das covariáveis se mostrou significativa. A interpretação dos parâmetros significativos desse modelo é feita a seguir, considerando sempre que as demais variáveis estão fixadas:

- O aumento da taxa de linfócitos, de menor do que 900céls/mm^3 para maior ou igual a 900céls/mm^3, não se mostrou significativa.
- A variação do IMC entre 18,5Kg/m^2 e 25Kg/m^2 para maior ou igual a 25Kg/m^2 aumenta em 0,05 vezes o tempo de sobrevivência entre pacientes diagnosticados com insuficiência cardíaca (exp (Estimativa) = $1,05$).
- Considerando PAD entre 40mmHg e 85mmHg como referência, o aumento para as demais faixas implica no aumento do tempo de sobrevivência. Por exemplo, o tempo de sobrevivência com PAD entre 89mmHg e 98mmHg ou entre 98mmHg e 108mmHg é 1,09 vezes o tempo se tivesse os valores de PAD da referência. Para o grupo com PAD não avaliado, no entanto, o tempo de sobrevivência é 0,85 vezes o tempo de sobrevivência do grupo com PAD entre 40mmHg e 85mmHg.
- Ao variar-se a Classe Funcional de I para a III o Tempo de vida é multiplicado por 0,89. Em outras palavras, o tempo de sobrevivência é 0,11 vezes menor para um paciente que tem a Classe Funcional alterada de I para a III.
- O aumento da Creatinina, de menor do que 0,8mEq/L para menor do que 1,3mEq/L implica na redução do tempo de sobrevivência. Mais especificamente, uma redução dessa proporção implica em um tempo de sobrevivência 0,14 vezes menor (1-0,86).

Observe que, desconsiderando a análise de significância dos parâmetros do modelo de Cox, e se atentando apenas ao sinal das estimativas, observe que os resultados da regressão

quantílica parecem de acordo com o sentido estimado via modelo de riscos proporcionais de Cox, no sentido de que a redução do tempo de sobrevivência significa aumento do risco no modelo de Cox, ao mesmo tempo em que a aumento do tempo de sobrevivência significa redução do risco.

Observe que, no modelo de riscos proporcionais de Cox, não foi detectada a significância das categorias NA das variáveis PAD e HDL. Observe também, na Tabela 5.2, que para essas categorias a suposição de riscos proporcionais é violada. Por outro lado, no modelo de regressão quantílica com pesos locais, estimados via árvore de sobrevivência, a significância dos parâmetros é detectada e as conclusões são coerentes com os gráficos de Kaplan-Meier apresentados na Figura 5.1: o tempo de sobrevivência é menor para esses indivíduos. Do ponto de vista da aplicação, esse resultado indica que o processo de geração de dados faltantes não é completamente ao acaso, o que daria respaldo para o pesquisador investigar tais indivíduos, no sentido de tentar identificar qual característica do grupo NA implica em tempo de sobrevivência menor. No contexto deste trabalho, a detecção da significância desses parâmetros a partir do modelo de regressão quantílica evidencia que essa abordagem é uma alternativa interessante ao modelo de Cox, especialmente quando a suposição de proporcionalidade dos riscos é violada.

Tabela 5.1: Estimativas obtidas a partir do modelo de riscos proporcionais de Cox.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	Valor p
Idade	0,01	1,01	0,002	<0,001
Sexo (F)	-	-	-	-
Sexo (M)	0,97	2,64	0,468	0,038
IMC (1)	-	-	-	-
IMC (2)	0,57	1,76	0,229	0,013
IMC (3)	-0,27	0,77	0,087	0,002
IMC (NA)	0,10	1,10	0,102	0,333
PAD (1)	-	-	-	-
PAD (2)	-0,16	0,85	0,384	0,676
PAD (3)	-0,30	0,74	0,105	0,005
PAD (4)	-0,26	0,77	0,116	0,026
PAD (5)	-0,37	0,69	0,144	0,009
PAD (NA)	0,16	1,17	0,141	0,258
Hipertensão (1)	-	-	-	-
Hipertensão (2)	0,19	1,21	0,092	0,037
Hipertensão (NA)	0,26	1,29	0,142	0,070
Classe Funcional (1)	-	-	-	-
Classe Funcional (2)	0,18	1,20	0,133	0,172
Classe Funcional (3)	0,62	1,85	0,132	<0,001
Classe Funcional (4)	0,51	1,66	0,147	0,001
Classe Funcional (NA)	0,32	1,37	0,180	0,079
DAE (1)	-	-	-	-
DAE (2)	0,33	1,40	0,100	0,001
DAE (NA)	0,29	1,34	0,136	0,032
DDVE (1)	-	-	-	-
DDVE (2)	-0,38	0,68	0,338	0,254
DDVE (NA)	-0,11	0,90	0,345	0,760
DSVE (1)	-	-	-	-
DSVE (2)	0,42	1,52	0,124	0,001
DSVE (NA)	0,36	1,43	0,131	0,006
Linfócitos (1)	-	-	-	-
Linfócitos (2)	-0,67	0,51	0,199	0,001
Linfócitos (NA)	0,99	2,68	0,198	<0,001
HDL (1)	-	-	-	-
HDL (2)	-0,30	0,74	0,074	<0,001
HDL (NA)	-0,14	0,87	0,075	0,063
Creatinina (1)	-	-	-	-
Creatinina (2)	0,21	1,23	0,127	0,103
Creatinina (3)	0,55	1,73	0,134	<0,001
Creatinina (NA)	0,55	1,73	0,163	0,001
Sexo (M) x DDVE (1)	-	-	-	-
Sexo (M) x DDVE (2)	-0,92	0,40	0,475	0,053
Sexo (M) x DDVE (3)	-0,92	0,40	0,476	0,053

Tabela 5.2: Teste de proporcionalidade dos riscos no modelo de Cox.

Variável	ρ	χ^2	Valor p
Idade	0,10	16,64	<0,001
Sexo (F)	-	-	-
Sexo (M)	0,03	1,24	0,266
IMC (1)	-	-	-
IMC (2)	-0,02	0,79	0,375
IMC (3)	0,00	0,02	0,899
IMC (NA)	-0,03	0,87	0,350
PAD (1)	-	-	-
PAD (2)	-0,04	1,98	0,160
PAD (3)	0,02	0,51	0,477
PAD (4)	0,06	5,70	0,017
PAD (5)	-0,03	1,62	0,203
PAD (NA)	0,06	5,32	0,021
Hipertensão (1)	-	-	-
Hipertensão (2)	-0,01	0,11	0,744
Hipertensão (NA)	0,04	1,68	0,195
Classe Funcional (1)	-	-	-
Classe Funcional (2)	-0,01	0,08	0,773
Classe Funcional (3)	-0,03	1,30	0,254
Classe Funcional (4)	-0,05	3,82	0,051
Classe Funcional (NA)	-0,06	4,42	0,036
DAE (1)	-	-	-
DAE (2)	-0,02	0,76	0,382
DAE (NA)	-0,02	0,36	0,551
DDVE (1)	-	-	-
DDVE (2)	0,02	0,85	0,356
DDVE (NA)	0,02	0,31	0,575
DSVE (1)	-	-	-
DSVE (2)	0,01	0,19	0,665
DSVE (NA)	0,01	0,08	0,775
Linfócitos (1)	-	-	-
Linfócitos (2)	-0,05	2,97	0,085
Linfócitos (NA)	-0,09	10,17	0,001
HDL (1)	-	-	-
HDL (2)	0,03	1,03	0,310
HDL (NA)	-0,07	7,10	0,008
Creatinina (1)	-	-	-
Creatinina (2)	-0,03	1,61	0,204
Creatinina (3)	-0,03	0,95	0,329
Creatinina (NA)	-0,02	0,73	0,393
Sexo (M) x DDVE (1)	-	-	-
Sexo (M) x DDVE (2)	-0,03	1,14	0,287
Sexo (M) x DDVE (3)	-0,03	1,48	0,224
GLOBAL	-	104,92	<0,001

Tabela 5.3: Estimativas obtidas para o modelo completo de regressão quantílica com pesos locais, $\tau = 0,45$, ajustado via árvore de sobrevivência

Variável	Estimativa	exp(Estimativa)	Erro Padrão	IC(95%)
Intercepto	7,64	2.073,51	0,349	[7,152; 8,397]
Idade	0,00	1,00	0,001	[−0,004; 0,001]
Sexo (F)	-	-	-	-
Sexo (M)	-0,72	0,49	0,502	[−1,555; 0,576]
IMC (1)	-	-	-	-
IMC (2)	-0,33	0,72	0,257	[−0,771; 0,070]
IMC (3)	0,06	1,06	0,022	[0,013; 0,109]
IMC (NA)	-0,03	0,97	0,039	[−0,108; 0,046]
PAD (1)	-	-	-	-
PAD (2)	-0,02	0,98	0,135	[−0,204; 0,285]
PAD (3)	0,04	1,05	0,031	[−0,008; 0,099]
PAD (4)	0,06	1,06	0,031	[0,021; 0,140]
PAD (5)	0,06	1,06	0,037	[0,003; 0,138]
PAD (NA)	-0,19	0,83	0,116	[−0,413; −0,016]
Hipertensão (1)	-	-	-	-
Hipertensão (2)	-0,03	0,97	0,031	[−0,077; 0,037]
Hipertensão (NA)	-0,13	0,87	0,090	[−0,327; −0,023]
Classe Funcional (1)	-	-	-	-
Classe Funcional (2)	-0,03	0,97	0,031	[−0,090; 0,027]
Classe Funcional (3)	-0,12	0,89	0,035	[−0,187; −0,045]
Classe Funcional (4)	-0,04	0,96	0,040	[−0,120; 0,039]
Classe Funcional (NA)	-0,14	0,87	0,073	[−0,261; 0,034]
DAE (1)	-	-	-	-
DAE (2)	-0,02	0,98	0,027	[−0,086; 0,020]
DAE (NA)	-0,08	0,93	0,054	[−0,185; 0,000]
DDVE (1)	-	-	-	-
DDVE (2)	0,25	1,28	0,321	[−0,511; 0,665]
DDVE (NA)	0,18	1,19	0,328	[−0,588; 0,661]
DSVE (1)	-	-	-	-
DSVE (2)	-0,09	0,92	0,051	[−0,203; 0,007]
DSVE (NA)	-0,06	0,94	0,051	[−0,179; 0,043]
Linfócitos (1)	-	-	-	-
Linfócitos (2)	0,01	1,01	0,099	[−0,171; 0,216]
Linfócitos (NA)	-0,98	0,37	0,119	[−1,181; −0,756]
HDL (1)	-	-	-	-
HDL (2)	0,07	1,08	0,025	[0,022; 0,114]
HDL (NA)	-0,13	0,88	0,047	[−0,233; −0,045]
Creatinina (1)	-	-	-	-
Creatinina (2)	0,04	1,04	0,033	[−0,044; 0,079]
Creatinina (3)	-0,11	0,89	0,055	[−0,262; −0,034]
Creatinina (NA)	-0,39	0,68	0,182	[−0,729; −0,028]
Sexo (M) x DDVE (1)	-	-	-	-
Sexo (M) x DDVE (2)	0,70	2,01	0,501	[−0,575; 1,549]
Sexo (M) x DDVE (3)	0,62	1,86	0,500	[−0,613; 1,503]

Tabela 5.4: Estimativas obtidas para o modelo completo de regressão quantílica, $\tau = 0,45$, ajustado com o método recursivo de Portnoy (2003)

Variável	Estimativa	exp(Estimativa)	Erro Padrão	Valor p
Intecepto	8,42	4.543,44	0,579	<0,001
Idade	0,00	1,00	0,004	0,254
Sexo (F)	-	-	-	-
Sexo (M)	-0,86	0,42	0,222	<0,001
IMC (1)	-	-	-	-
IMC (2)	-0,64	0,53	0,251	0,011
IMC (3)	0,28	1,32	0,156	0,075
IMC (NA)	0,02	1,02	0,281	0,950
PAD (1)	-	-	-	-
PAD (2)	0,17	1,19	0,289	0,552
PAD (3)	0,28	1,33	0,294	0,338
PAD (4)	0,04	1,04	0,142	0,795
PAD (5)	0,23	1,26	0,241	0,333
PAD (NA)	-0,39	0,68	0,302	0,194
Hipertensão (1)	-	-	-	-
Hipertensão (2)	-0,14	0,87	0,123	0,261
Hipertensão (NA)	-0,38	0,69	0,368	0,305
Classe Funcional (1)	-	-	-	-
Classe Funcional (2)	-0,22	0,80	0,375	0,562
Classe Funcional (3)	-0,54	0,58	0,338	0,113
Classe Funcional (4)	-0,38	0,68	0,419	0,365
Classe Funcional (NA)	0,02	1,02	0,282	0,944
DAE (1)	-	-	-	-
DAE (2)	-0,26	0,77	0,215	0,231
DAE (NA)	-0,21	0,81	0,254	0,409
DDVE (1)	-	-	-	-
DDVE (2)	0,45	1,57	0,184	0,014
DDVE (NA)	0,23	1,26	0,365	0,522
DSVE (1)	-	-	-	-
DSVE (2)	-0,35	0,70	0,145	0,016
DSVE (NA)	-0,33	0,72	0,193	0,087
Linfócitos (1)	-	-	-	-
Linfócitos (2)	0,30	1,34	0,250	0,237
Linfócitos (NA)	-1,02	0,36	0,177	<0,001
HDL (1)	-	-	-	-
HDL (2)	0,19	1,21	0,109	0,077
HDL (NA)	-0,07	0,93	0,159	0,660
Creatinina (1)	-	-	-	-
Creatinina (2)	-0,02	0,98	0,261	0,928
Creatinina (3)	-0,39	0,68	0,294	0,184
Creatinina (NA)	-0,44	0,64	0,273	0,103
Sexo (M) x DDVE (1)	-	-	-	-
Sexo (M) x DDVE (2)	0,82	2,28	0,296	0,005
Sexo (M) x DDVE (3)	0,75	2,12	0,258	0,003

Tabela 5.5: Estimativas obtidas para o modelo final de regressão quantílica com pesos locais, $\tau = 0,45$, ajustado via árvore de sobrevivência.

Variável	Estimativa	exp(Estimativa)	Erro Padrão	IC(95%)
Intercepto	7,73	2.285,50	0,114	$[7,462; 7,935]$
IMC (1)	-	-	-	-
IMC (2)	-0,18	0,83	0,254	$[-0,753; 0,024]$
IMC (3)	0,05	1,05	0,022	$[0,013; 0,108]$
IMC (NA)	-0,03	0,97	0,037	$[-0,101; 0,045]$
PAD (1)	-	-	-	-
PAD (2)	0,01	1,01	0,135	$[-0,159; 0,288]$
PAD (3)	0,05	1,05	0,027	$[-0,001; 0,104]$
PAD (4)	0,09	1,09	0,031	$[0,018; 0,145]$
PAD (5)	0,08	1,09	0,035	$[0,016; 0,151]$
PAD (NA)	-0,20	0,82	0,102	$[-0,414; -0,015]$
Hipertensão (1)	-	-	-	-
Hipertensão (2)	-0,01	0,99	0,030	$[-0,063; 0,045]$
Hipertensão (NA)	-0,16	0,85	0,086	$[-0,327; -0,019]$
Classe Funcional (1)	-	-	-	-
Classe Funcional (2)	-0,04	0,96	0,027	$[-0,083; 0,024]$
Classe Funcional (3)	-0,10	0,90	0,035	$[-0,169; -0,040]$
Classe Funcional (4)	-0,04	0,96	0,045	$[-0,118; 0,029]$
Classe Funcional (NA)	-0,10	0,91	0,081	$[-0,239; 0,059]$
DAE (1)	-	-	-	-
DAE (2)	-0,04	0,96	0,025	$[-0,095; 0,006]$
DAE (NA)	-0,14	0,87	0,052	$[-0,273; -0,073]$
Linfócitos (1)	-	-	-	-
Linfócitos (2)	0,02	1,02	0,107	$[-0,187; 0,271]$
Linfócitos (NA)	-0,99	0,37	0,125	$[-1,215; -0,686]$
HDL (1)	-	-	-	-
HDL (2)	0,05	1,05	0,025	$[0,016; 0,107]$
HDL (NA)	-0,16	0,85	0,048	$[-0,236; -0,058]$
Creatinina (1)	-	-	-	-
Creatinina (2)	0,02	1,02	0,032	$[-0,048; 0,071]$
Creatinina (3)	-0,15	0,86	0,051	$[-0,246; -0,051]$
Creatinina (NA)	-0,34	0,71	0,155	$[-0,720; -0,048]$

Capítulo 6

Discussão e Considerações Finais

Nesta dissertação, foram apresentados os modelos de regressão quantílica para dados censurados como uma abordagem inferencial complementar ou mesmo alternativa aos tradicionais modelos de Análise de Sobrevivência. Sem dúvida, a principal vantagem da regressão quantílica sobre as técnicas usuais é a possibilidade de interpretar direta e facilmente os tempos de sobrevivência (e não em termos da taxa de falha ou de uma função do tempo), sem requerer o conhecimento acerca da distribuição dos dados.

Inicialmente, apresentou-se uma breve introdução aos modelos de regressão quantílica no contexto em que a variável resposta é completamente observável. Nesse ramo da regressão quantílica, a parte inferencial se mostra muito mais desenvolvida, com a existência de testes de hipóteses conjunto para os parâmetros, por exemplo. Existe também teste para detectar a falta de ajuste dos modelos, bastante importante, já que a suposição de linearidade pode ser bastante restritiva. Uma apresentação bastante didática desse teste é apresentado em Santos (2012).

Muitos autores se dedicaram à extensão dos modelos de regressão quantílica para conjuntos de dados com censura. Várias abordagens foram propostas na literatura, que inicialmente tratavam apenas de censura fixa e conhecida para todas as observações. Portnoy (2003) foi pioneiro ao introduzir o conceito de censura aleatória para essa classe de modelos, e a sua metodologia foi apresentada nessa dissertação. Apesar de revolucionar os estudos de regressão quantílica para dados censurados, sua metodologia apresenta uma suposição bastante restritiva na prática: a linearidade global do quantil nas covariáveis em análise. Ao longo desta dissertação, apresentou-se um exemplo em que essa suposição é violada, e discutiu-se que tal fato interfere na amplitude dos intervalos de confiança dos parâmetros, o que, em alguns casos, poderia comprometer, inclusive, as conclusões inferenciais.

Entre as metodologias alternativas ao método recursivo de Portnoy, estudou-se nesta dissertação a abordagem de pesos locais. Modelos de regressão quantílica para dados censurados com pesos locais são mais flexíveis no sentido de não requererem a linearidade global entre covariáveis e o tempo de sobrevivência: a linearidade é necessária e suficiente apenas no quantil de interesse. Em particular, foram estudadas duas técnicas para estimação dos pesos: o uso de

suavizadores *kernel* e a metodologia de árvores de sobrevivência.

De um modo geral, para os casos mais simples com duas covariáveis, o estudo de simulação revelou que as três metodologias são bastante semelhantes, no que diz respeito ao viés, erro padrão e erro quadrático médio. Para o estudo simulado com distribuição comum às análises paramétricas de Análise de Sobrevivência, Weibull e log-logística, as metodologias mostraram-se semelhantes para quantis menores (0,25 e 0,50), sendo que a de Portnoy apresentou o melhor desempenho, como era esperado, por se tratarem de modelos lineares (ou linearizados após transformação, como é o caso da distribuição Weibull).

Apresentou-se, também, uma aplicação a um conjunto de dados reais, em que o uso da técnica de regressão quantílica se mostra adequado para solucionar um dos objetivos do estudo, que era determinar a relação entre as covariáveis diretamente no tempo de sobrevivência. Como comparação, foi apresentado o modelo de Cox, e observou-se que a regressão quantílica foi capaz de detectar a significância de categorias de variáveis que tinham suposição de riscos proporcionais violada, significância que não foi observada no modelo tradicional. Nesse sentido, sugere-se realizar estudos de simulação mais aprofundados que investiguem o comportamento da regressão quantílica quando não existem evidências que sustentem a suposição de proporcionalidade dos riscos para aplicação do modelo de Cox.

A principal desvantagem da regressão quantílica, no entanto, sem dúvida recai sobre a falta de estudos específicos que considerem técnicas de diagnósticos para avaliar a qualidade do ajuste dos parâmetros do modelo. Existe na literatura uma abordagem para avaliar a falta de ajuste proposta por Wang (2008), uma técnica não paramétrica que envolve suavizores. Conforme discutido para os pesos estimados via *kernel* nesta dissertação, os métodos baseados em suavizadores não são recomendados a menos que se esteja trabalhando com um número pequeno de covariáveis (em geral, menor do que três). Na prática, os conjuntos de dados, envolvem dezenas, ou mesmo centenas de covariáveis, tornando a aplicação do método inviável.

Conforme discutido, muito ainda precisa ser estudado no que diz respeito a inferência em modelos de regressão quantílica para dados censurados. Sugere-se investigar a possibilidade de generalizações dos testes de qualidade do ajuste e teste de hipóteses conjunto para os parâmetros para dados censurados, que já estão desenvolvidos para o contexto sem censura.

Além disso, sugere-se o estudo de regressão quantílica para dados censurados com estrutura não linear nos parâmetros, que parece ser muito mais flexível do ponto de vista prático.

Apêndice A

Estimadores de Densidade *Kernel*

Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias independentes e identicamente distribuídas com função densidade $f(\cdot)$ absolutamente contínua, porém desconhecida. Suponha que, para um dado $y \in \mathbb{R}$, seja de interesse estimar a função $f(y)$ a partir de uma amostra aleatória $y_1, \cdots, y_n$.

Para tanto, relembre inicialmente, das definições clássicas de Probabilidade, que a função densidade deriva da função de distribuição acumulada, definida como $F(y) = \int_{-\infty}^{y} f(u)du$. Pensar no relacionamento das definições é bastante conveniente, uma vez que se conhece um estimador intuitivo para $F(y)$, conhecido como função de distribuição acumulada empírica, que pode ser escrito como:

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

Além de intuitivo, o estimador $F_n(y)$ apresenta boas propriedades: de acordo com a Lei Forte dos Grandes Números, $F_n(y)$ converge quase certamente para $F(y)$, quando $n \to +\infty$.

Então, conforme discutido em Tsybakov (2009), uma forma de estimar $f(y)$ surge do argumento de que, para um dado $h \geq 0$ suficientemente pequeno, tem-se que:

$$f(y) \approx \frac{F(y+h) - F(y-h)}{2h}.$$

Finalmente, substituindo $F(\cdot)$ por sua estimativa $F_n(\cdot)$, um estimador para a função densidade é dado por:

$$\hat{f_n}(y) = \frac{F_n(y+h) - F_n(y-h)}{2h} = \frac{1}{2nh}\sum_{i=1}^{n} \mathcal{I}(y-h < y_i \leq y+h) = \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{y_i - y}{h}\right),$$

em que $K_0(u) = \frac{1}{2}\mathcal{I}(-1 < u \leq 1)$.

O estimador $\hat{f_n}(y)$ é chamado *estimador Rosenblatt* e pode ser estendido para o caso mais geral, em que K_0 pode ser substituído por $K : \mathbb{R} \to \mathbb{R}$, uma função integrável qualquer (usualmente simétrica), que satisfaz as condições $\int K(u)du = 1$ e $\int uK(u)du = 0$, de modo que:

$$\hat{f_n}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{y_i - y}{h}\Big).$$

A função K é chamada *função de densidade kernel*. Alguns exemplos de funções de densidade *kernel* habitualmente usadas são:

1. *Kernel* retangular: $K(u) = \frac{1}{2}\mathcal{I}(|u| \leq 1)$,
2. *Kernel* triangular: $K(u) = (1 - |u|)\mathcal{I}(|u| \leq 1)$,
3. *Kernel* parabólico, ou *Kernel* Epanechikov: $K(u) = \frac{3}{4}(1 - u^2)\mathcal{I}(|u| \leq 1)$,
4. *Kernel* biquadrático: $K(u) = \frac{15}{16}(1 - u^2)^2\mathcal{I}(|u| \leq 1)$,
5. *Kernel* Gaussiano: $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2), u \in \mathbb{R}$,
6. *Kernel* Silverman: $K(u) = \frac{1}{2}\exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2} + \pi/4), u \in \mathbb{R}$.

Como uma medida de acurácia do estimador $\hat{f_n}(y)$, a literatura sugere a avaliação do Erro Quadrático Médio (EQM) que, para um $y_0 \in \mathbb{R}$ fixado, é dado por:

$$EQM(y_0) = E\Big[(\hat{f_n}(y_0) - f(y_0))^2\Big].$$

Relembre que o EQM pode ser entendido como resultado da soma de duas componentes:

$$EQM(y_0) = b^2(y_0) + \sigma^2(y_0),$$

em que $b(y_0)$ é a função viés do estimador e $\sigma^2(y_0)$ é a função de variância. Para maiores detalhes, ver Neter et al. (1996), por exemplo. As definições e proposições a seguir trazem alguns resultados das funções viés e de variância do estimador $\hat{f_n}(y)$.

Proposição 1 *Suponha que $f(y) \leq f_{max} < \infty$ para todo $y \in \mathbb{R}$. Seja $K : \mathbb{R} \to \mathbb{R}$ uma função tal que*

$$\int K^2(u)du < \infty.$$

Então, para qualquer $y_0 \in \mathbb{R}, h > 0$ e $n \geq 1$, tem-se que

$$\sigma^2(y_0) \leq \frac{C_1}{nh},$$

em que $C_1 = f_{max}\int K^2(u)du$.

Defina $\lfloor\gamma\rfloor$ o maior inteiro estritamente menor do que o número $\gamma \in \mathbb{R}$.

Definição 1 *Seja T um intervalo definido em $\mathbb{R}$, e sejam γ e L dois números inteiros positivos. A classe Hölder $\Sigma(\gamma, L)$ em T é definida como o conjunto de funções diferenciáveis $l = \lfloor\gamma\rfloor$ vezes, com $f : T \to \mathbb{R}$,*

cuja derivada $f^{(l)}$ satisfaz:

$$|f^{(l)}(y) - f^{(l)}(y')| \leq L|y - y'|^{\gamma - l}, \forall y, y' \in T.$$

Definição 2 *Seja $l \geq 1$ um número inteiro. Diz-se que $K : \mathbb{R} \to \mathbb{R}$ é Kernel de ordem l se as funções $u \longmapsto u^j K(u)$, $j = 1, \cdots, l$, são integráveis e satisfazem as seguintes condições:*

$$\int K(u)du = 1, \ e \ \int u^j K(u)du = 0, j = 1, \cdots, l.$$

Suponha agora que a função densidade $f(\cdot)$ esteja definida em uma classe $\mathcal{P} = \mathcal{P}(\gamma, L)$ de funções, com:

$$\mathcal{P}(\gamma, L) = \{f \mid f \geq 0, \int f(u)du = 1 \text{ e } f \in \Sigma(\gamma, L) \in \mathbb{R}\}.$$

Então, pode-se enunciar a seguinte proposição:

Proposição 2 *Assuma que $f \in \mathcal{P}(\gamma, L)$ e seja K função kernel de ordem $l = \lfloor \gamma \rfloor$ satisfazendo*

$$\int |u|^\gamma |K(u)|du < \infty.$$

Para todo $y_0 \in \mathbb{R}$ e $n \geq 1$ tem-se que

$$|b(x_0)| \leq C_2 h^\gamma,$$

com $C_2 = \frac{L}{l!} \int |u|^\gamma |K(u)|du$.

Observe que, de acordo com as Proposições 1 e 2, o estimador $\hat{f_n}(y_0)$ não é consistente para um h fixado. Por um lado, quanto menor o valor de h, menor é o viés do estimador. Por outro lado, maior é a sua variância $\sigma^2(y_0)$. As demostrações das proposições acima podem ser encontradas em Tsybakov (2009).

Portanto, a escoha de h, conhecido como *bandwidth*, é muito importante para a estimação. Conforme discutido em Hastie e Tibshirani (1990), a escolha da função *kernel* em si (*kernel* retangular, quadrático, etc), no entanto, não é tão influente.

Para a escolha do valor h, observe inicalmente que o EQM é uma medida de acurácia de $\hat{f_n}(y_0)$ apenas para um número $y_0 \in \mathbb{R}$ fixado. Porém, em geral, tem-se o objetivo de analisar a acurácia do estimador para uma sequência de valores. Considere então o Erro Quadrático Médio Integrado, EQMI, uma medida global de avaliação de $\hat{f_n}(\cdot)$ dada por:

$$EQMI(h) = E \int \left[\hat{f_n}(y) - f(y)\right]^2 dy$$

De acordo com o teorema de Tonelli-Fubini,

$$EQMI(h) = \int EQM(y)dy = \int b^2(y)dy + \int \sigma^2(y)dy.$$

Então, o h ideal é o valor h_{id} tal que

$$h_{id} = arg\min_{h\geq 0} EQMI(h).$$

No entanto, conforme discutido em Tsybakov (2009), como a função EQMI depende da função $f(\cdot)$, que é desconhecida, ela não pode ser utilizada como ferramenta de avaliação do desempenho de $\hat{f_n}(\cdot)$. Uma forma alternativa é utilizar a popular metodologia *validação cruzada* que será discutida a seguir.

Observe que $EQMI(h)$ pode ser escrito como:

$$EQMI(h) = E\int \left[\hat{f_n}(y) - f(y)\right]^2 dy = E\left\{\int [\hat{f_n}(y)]^2 dy - 2\int \hat{f_n}(y)f(y)dy\right\} + \int [f(y)]^2 dy.$$

Como $\int [f(y)]^2 dy$ não depende de h, o h_{id} resultante é o mesmo ao se minimizar seguinte função:

$$J(h) = E\left[\int [\hat{f_n}(y)]^2 dy - 2\int \hat{f_n}(y)f(y)dy\right].$$

Tsybakov (2009) mostra que $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(y)$ é estimador não viesado para $E\left[\int \hat{f_n}(y)f(y)dy\right]$, em que

$$\hat{f}_{n,-i} = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{y_j - y}{h}\right).$$

Assim, um estimador não viesado para $J(h)$ é dado por:

$$CV(h) = \int [\hat{f_n}(y)]^2 dy - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(y),$$

em que $CV(h)$ é o estimador de validação cruzada no ponto h (*Cross Validation*, em inglês).

Dessa forma, o h ideal é um valor h_{CV} que minimiza a estimativa do EQMI entre todos os $h \geq 0$. Em outras palavras:

$$h_{CV} = arg\min_{h>0} CV(h).$$

Na prática, fixa-se um conjunto de h's para avaliação, e é escohido aquele que resultar no menor valor de $CV(\cdot)$.

Uma segunda abordagem, implementada no *software* R e utilizada nesta dissertação, consiste em escolher o h que maximiza a seguinte expressão:

$$MLCV(h) = \left(n^{-1}\sum_{i=1}^{n} \log\left[\sum_{j\neq i} K\left(\frac{y_j - y_i}{h}\right) - log[(n-1)/h]\right]\right),$$

que é, na verdade, uma proposta não trivial de maximizar a pseudo- verossimilhança $\prod_{i=1}^{n} f_h(y_i)$. A função utilizada está disponível no pacote *kedd*.

A validação cruzada, no entanto, não é a única forma de avaliação do estimador $\hat{f_n}(y)$. Uma outra abordagem, por exemplo, baseada na análise de Fourier, é apresentada em Tsybakov (2009).

Os casos apresentados anteriromente referem-se a um conjunto de dados com apenas uma variável aleatória. Quando, no entanto, tem-se um conjunto de variáveis aleatórias independentes, o estimador *kernel* resultante é dado pelo produto dos *kernel*, isto é,

$$K(x_1, x_2) = K_1(x_1) \times K_2(x_2),$$

em que x_1 e x_2 são duas observações de variáveis aleatórias X_1 e X_2 independentes quaisquer. Maiores detalhes podem ser encontrados em Li e Racine (2003).

A aplicação dos estimadores *kernel* torna-se impraticável, no entanto, quando muitas variáveis aleatórias fazem parte do estudo, conforme destaca Wey et al. (2014), ressaltando que os estimadores *kernel* não são recomendáveis para dimensões maiores do que duas variáveis. A razão para tal fato é que, quanto maior a dimensão do vetor de variáveis $\mathbf{x}$, mais espaçados ficarão os pontos amostrais. Em outras palavras, num espaço de dimensão mais elevado haverá poucas observações em torno de um vetor $\mathbf{x}$ qualquer, a menos que o tamanho da amostra seja extremamente grande. Caso a amostra não seja suficientemente grande, os valores de *bandwidth* devem ser cada vez maiores, o que resulta no aumento do viés do estimador. Uma discussão acerca do assunto pode ser encontrada em Scott e Sain (2005).

Apêndice B

Árvore de Sobrevivência

Os modelos tradicionais para análise do tempo de sobrevivência, que pode estar sujeito a censura, apresentam algumas suposições que podem não ser verificadas, ou ainda apresentam limitações metodológicas que justificam o estudo de técnicas mais flexíveis, como árvores de sobrevivência, por exemplo.

Conforme discutido por Bou-Hamad et al. (2011), ao contrário do modelo de Cox, as árvores de sobrevivência não requerem riscos proporcionais e não partem de uma determinada função de ligação para os parâmetros do modelo. Além disso, alguns tipos de interação entre as covariáveis podem ser identificados automaticamente, sem a necessidade de serem especificadas previamente pelo pesquisador. Os autores ressaltam, ainda, que nos modelos tradicionais, a inferência é feita após vários modelos serem testados, e que as propriedades estatísticas do modelo, após essa seleção, são desconhecidas. Nesse sentido, as árvores de sobrevivência são atrativas também do ponto de vista inferencial.

Trata-se de uma técnica não paramétrica que surgiu em meados dos anos 1980 como uma extensão das árvores de regressão para o contexto em que a variável resposta pode ser censurada. O estudo de árvores de regressão, por sua vez, surgiu no início dos anos de 1960, com o trabalho de Morgan e Sonquist (1963). Mais tarde, a metodologia ganhou popularidade com o desenvolvimento de algoritmos mais eficientes, como por exemplo, o trabalho de Breiman et al. (1984) e o algoritmo *CART*, hoje implementado em vários *softwares* estatísticos, como o R. Sugere-se a leitura de Bou-Hamad et al. (2011) para uma revisão bibliográfica mais detalhada.

Neste Apêndice, é apresentada uma breve descrição de árvores de sobrevivência binárias, as mais populares entre as árvores de sobrevivência. Como critério para a sua construção, são apresentadas as mesmas técnicas descritas no artigo de Wey et al. (2014). O objetivo deste Apêndice não é, portanto, discutir todas as propriedades e possíveis algoritmos, mas descrever as ideias mais gerais das árvores de sobrevivência necessárias para entendimento do cálculo dos pesos das observações censuradas na metodologia de regressão quantílica linear apresentada pelos autores supracitados.

Dessa forma, considere inicialmente a seguinte definição de árvore binária, apresentada em Hothorn et al. (2004), referenciando Breiman et al. (1984):

Definição 1 *Uma árvore binária é um conjunto de q nós e suas arestas. Os nós, que serão denotados por* t_j, $j = 1, \cdots, q$, *são subconjuntos do espaço amostral* χ. *Baseado numa amostra* $\mathcal{L}$ *e em uma regra de partição e parada, a árvore* $T(\mathcal{L}) = \{1, \cdots, q\}$ *é construída. Os elementos de* $T(\mathcal{L})$ *representam os nós pelos seus índices. O nó terminal, isto é, o nó que não pode mais ser repartido, será denotado por um subconjunto da árvore,* $\tilde{T}(\mathcal{L}) \subset T(\mathcal{L})$. *Os nós terminais são partições disjuntas do espaço amostral* χ *de covariáveis, isto é:*

$$\chi = \bigcup_{j \in \tilde{T}(\mathcal{L})} t_j, \text{ e } t_j \cap t_k = \emptyset, \forall j \neq k \in \tilde{T}(\mathcal{L}).$$

Exemplo

Para ilustrar a definição de árvore binária, suponha um exemplo simples e fictício, de uma amostra de pessoas com idade entre 18 e 85 anos, de ambos os sexos, que apresentam determinada enfermidade. Suponha que o interesse seja avaliar o tempo de sobrevivência da população que tem a enfermidade. A Figura B.1 apresenta um desenho esquemático de uma árvore construída para esse exemplo.

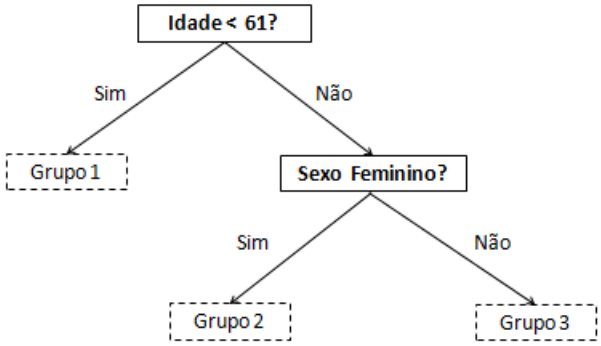

Figura B.1: Exemplo de Árvore binária.

Observe que neste exemplo se têm três nós terminais, que neste caso estão representados pelos Grupos 1, 2 e 3. Observe que o nó terminal 1 corresponde ao grupo de pessoas na amostra com idade inferior a 61 anos. O nó terminal 2, por sua vez, corresponde ao grupo de pessoas com mais de 61 anos (inclusive) e do sexo feminino, enquanto que o terceiro nó terminal são as pessoas com mais do que 61 anos (inclusive) do sexo masculino. Observe ainda que, conforme definição, os nós são disjuntos e o conjunto de nós terminais corresponde à amostra completa.

A metodologia de árvores de sobrevivência binária divide o espaço de covariáveis nos nós e vértices de uma árvore binária. O essencial na metodologia é entender como e quando repartir um nó, sempre levando em consideração que a resposta do estudo, no caso o tempo de sobrevivência, pode estar sujeito a censura.

Em primeiro lugar, em árvores de sobrevivência binárias, cada uma das covariáveis é analisada individualmente. Para tanto, são utilizados algoritmos de partição recursiva que dividem o espaço das covariáveis em regiões que se diferenciam de acordo com algum critério estatístico. Todas as divisões possíveis para cada uma das covariáveis são avaliadas, de modo a aumentar a homogeneidade das respostas dentro de um mesmo nó, e a heterogeneidade para nós diferentes.

Um critério para divisão dos nós, apresentado em Rudser et al. (2012), é dado pelo máximo das quatro estatísticas $G^{\rho,\gamma}$, isto é, para $(\rho, \gamma) = \{(0,0), (1,0), (0,1), (1,1)\}$:

$$G^{\rho,\gamma} = \frac{M_1 + M_0}{M_1 M_0} \sum_{y \in \mathcal{F}} \frac{n_{1y} n_{0y}}{n_{1y} + n_{0y}} \hat{S}(y-)^{\rho} [1 - \hat{S}(y-)]^{\gamma} [\hat{\lambda}_1(y) - \hat{\lambda}_0(y)],$$

em que M_j é o número de indivíduos inicialmente em risco no grupo j, $j = 0, 1$, $\mathcal{F}$ é o conjunto de tempos de falhas únicos, n_j é o número de indivíduos em risco no grupo j no tempo t, e $\hat{\lambda}_j(t)$ é o risco estimado do grupo j no tempo t e $\hat{S}(t-)$ denota a estimativa Kaplan-Meier da curva de sobrevivência considerando ambos os grupos juntos. Observe que $G^{0,0}$ coincide com o teste de *logrank*, enquanto que $G^{1,0}$ é o teste de *Wilcoxon* ponderado, ambos testes conhecidos em Análise de Sobrevivência para detectar diferenças entre funções de sobrevivência. As quantidades $G^{0,1}$ e $G^{1,1}$ não tem nomes conhecidos, mas também testam diferenças entre as funções. A combinação das estatísticas $G^{\rho,\gamma}$, de acordo com Lee (1996) e discutidas por Wey et al. (2014), são usadas para aumentar o poder de detecção das diferenças entre as funções de sobrevivência.

Conforme destaca Rudser et al. (2012), dividindo-se a estatística $G^{\rho,\gamma}$ por uma estimativa consistente da variância, e sob a hipótese nula $H_0 : S_1(t) = S_0(t)$, obtém-se uma estatística consistente e assintoticamente normal, e que pode ser avaliada sob a distribuição normal padrão.

Dessa forma, o primeiro passo em árvores de sobrevivência é encontrar a melhor divisão para cada uma das covariáveis. A melhor divisão para cada covariável é aquela que apresenta o valor máximo entre as estatísticas $G^{\rho,\gamma}$. Para variáveis binárias, por exemplo, só existe uma possível divisão da variável. Por outro lado, se a variável é discreta não ordinal, então o objetivo é dividir os dados em dois conjuntos disjuntos de observações. Para este caso, também é possível trabalhar com variáveis *dummies*, de modo que, se a variável tem k categorias, então criam-se $k-1$ variáveis binárias. Por fim, para variáveis ordinais, o objetivo é encontrar o valor c que mais discrimina os dados de acordo com o valor máximo entre as estatísticas $G^{\rho,\gamma}$.

Dada a melhor divisão de cada uma das variáveis, é escolhida aquela que apresenta a estatística $G^{\rho,\gamma}$ máxima. Se os grupos formados a partir da melhor quebra dessa variável são significativos a um nível de significância α, avaliado sob a distribuição normal padrão, então se repete a busca da melhor variável para cada uma das subamostras formadas a partir dessa divisão. O procedimento é repetido até que não se encontre grupos significativamente diferentes. Além desse critério estatístico para a divisão dos grupos, mostram-se necessários dois outros critérios de parada (Rudser et al., 2012):

1. Número mínimo de indivíduos em risco. Neste caso, cada nó deve ter um número mínimo de indivíduos em risco no tempo especificado.

2. Número mínimo de eventos. Cada nó precisa ter um número mínimo de eventos.

Essas condições são necessárias para o procedimento que segue. Para cada nó terminal, isto é, para cada $\tilde{T}(\mathcal{L})$ são aplicados o estimador de Kaplan-Meier para estimação da função de sobrevivência. Sem essas condições, as estimativas de Kaplan-Meier ficariam comprometidas. Os autores ressaltam que o uso do estimador de Kaplan-Meier pode ser substituído por qualquer outro método não paramétrico para a estimação de funções de sobrevivência na presença de censura. Ou seja, neste caso as árvores são apenas um critério para subdivisão do espaço amostral. Intervalos de confiança e inferência para o tempo de sobrevivência são feitos a partir dos conceitos teóricos do estimador de Kaplan-Meier.

Finalmente, conforme discutido em Wey et al. (2014) e Hothorn et al. (2004), uma das críticas da metodologia de árvores em geral é que pequenas mudanças na amostra poderiam gerar árvores completamente diferentes. Breiman (1996) propôs então um método de *baggin*. Na verdade, o método consiste em gerar um número L de subamostras de mesmo tamanho da amostra original e calcular a função de sobrevivência estimada para cada uma delas. Então, calcular a média dessas estimativas, isto é,

$$\hat{F}_{Y|\mathbf{x}}(y) = \frac{1}{L}\sum_{b=1}^{L} \tilde{F}^{b}_{Y|\mathbf{x}}(y),$$

de modo a obter uma estimativa mais estável e confiável da função de sobrevivência.

Apesar de ser uma estratégia necessária para a aplicação de árvores, o método de *baggin* tem a desvantagem de não ser replicável. Ou seja, dois pesquisadores diferentes que decidam usar o mesmo conjunto de dados podem chegar a uma estimativa diferente para a função de sobrevivência, dado o vetor de covariáveis, pois a técnica envolve um processo de reamostragem aleatória.

Por outro lado, como pontos positivos da técnica, pode-se citar que, além de apresentarem um alto poder preditivo, as árvores apresentam a vantagem de ser uma ferramenta de análise bastante visual, que permite o pesquisador compreender a relação entre as covariáveis e a variável resposta. Além disso, as árvores são invariantes sob transformações monótonas nos preditores e, conforme discutido anteriormente, apresentam flexibilidade para ajustar efeitos das covariáveis não lineares ou não aditivos para estimar a resposta.

Apêndice C

Gráficos do Estudo de Simulação

Neste apêndice são apresentados os gráficos com os resultados do estudo de simulação baseados em amostras de tamanhos $N = 400$, $N = 800$ e proporção de censuras iguais a 25% e 50%, discutidos no Capítulo 4, para comparação das três metodologias de regressão quantílica para dados censurados.

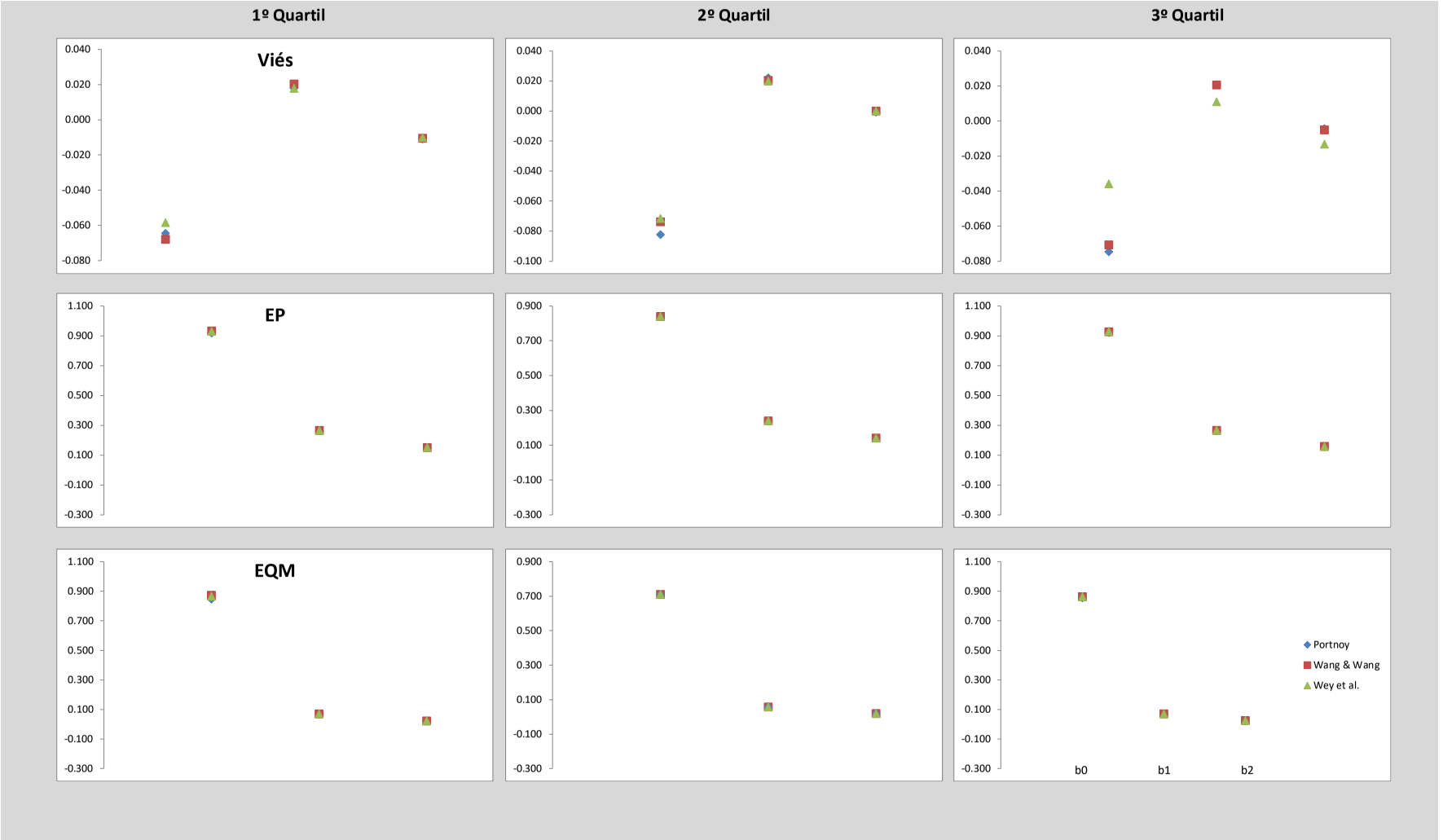

Figura C.1: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição normal.

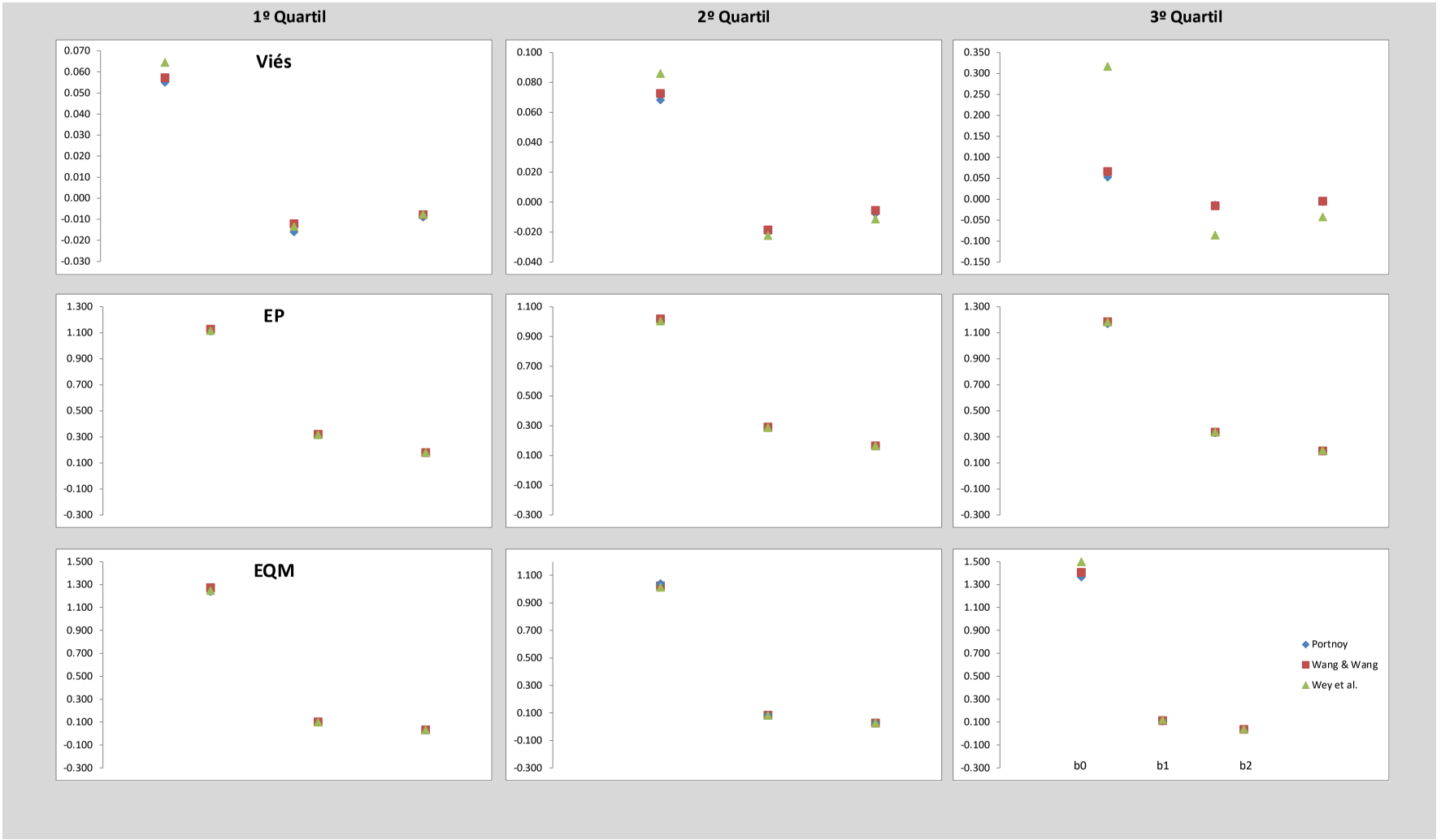

Figura C.2: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição normal.

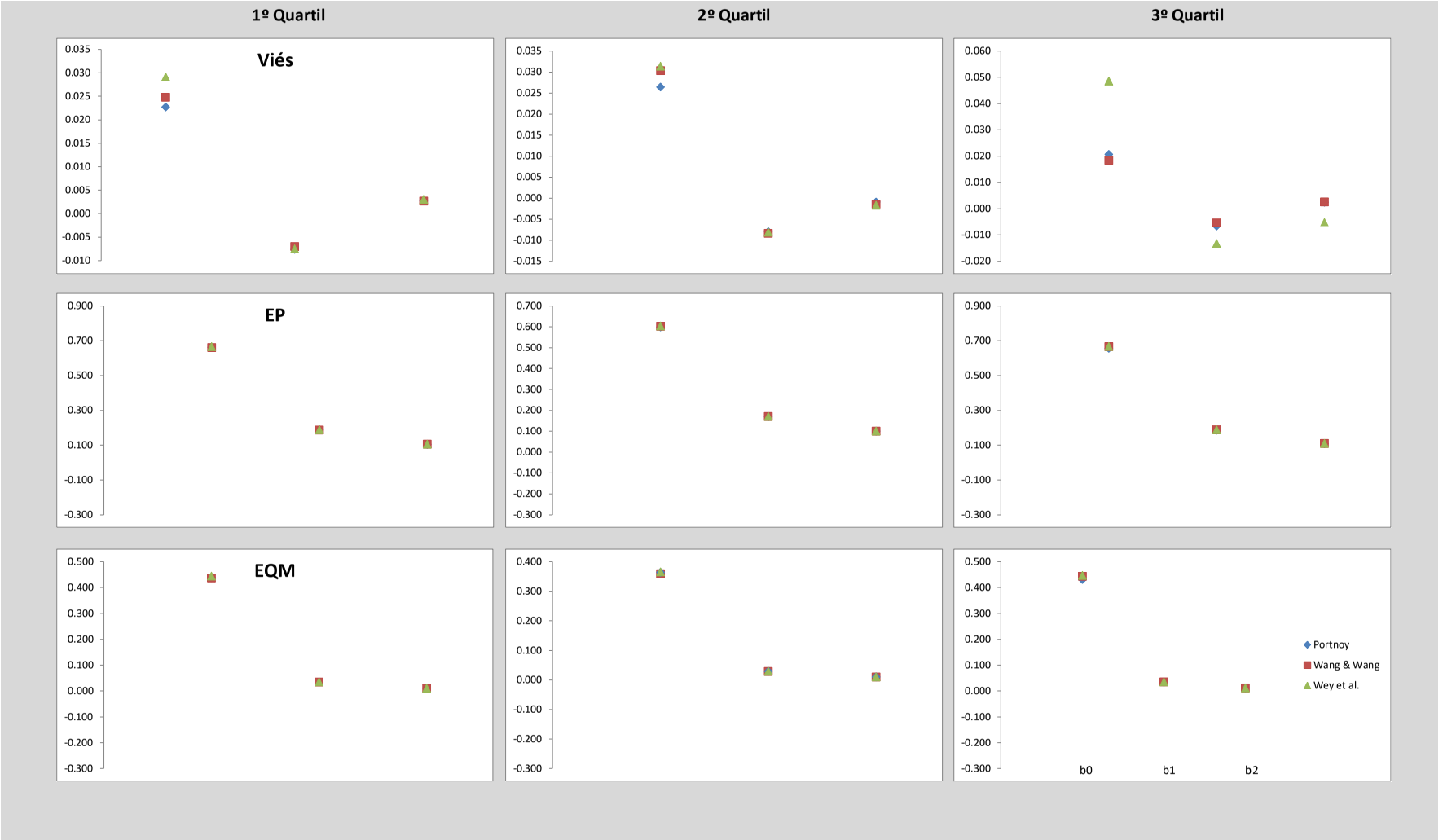

Figura C.3: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição normal.

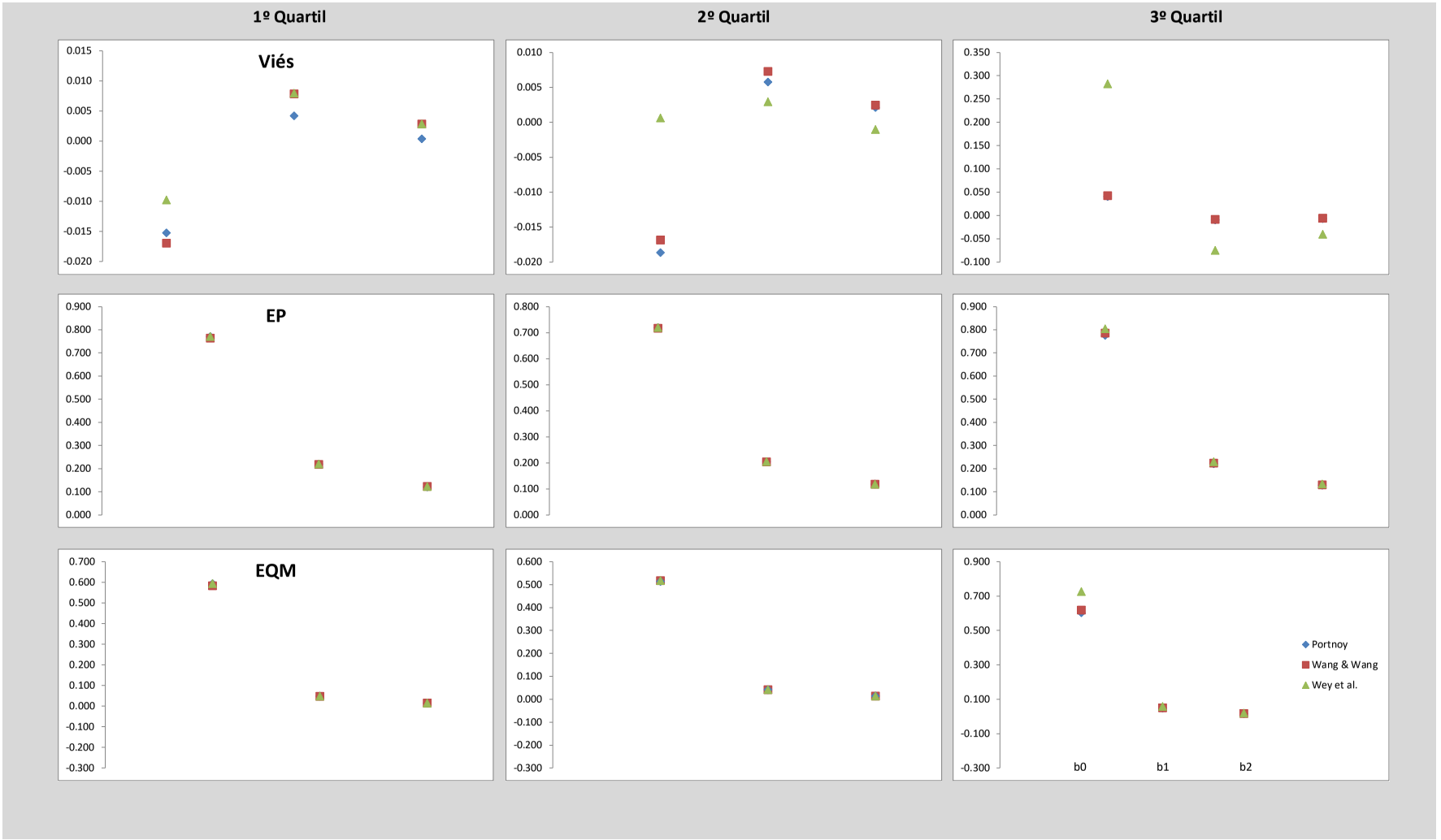

Figura C.4: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição normal.

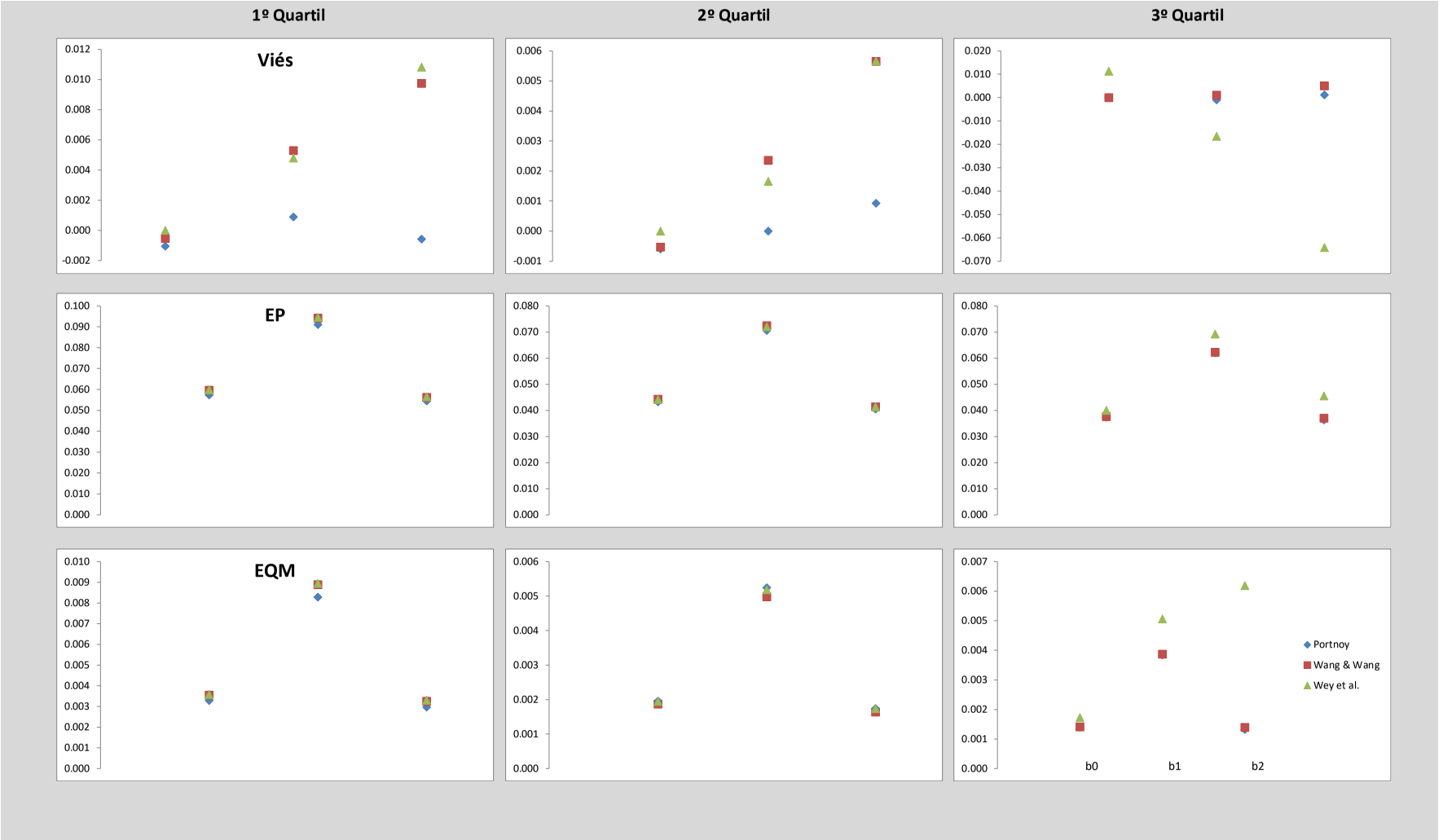

Figura C.5: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição Weibull.

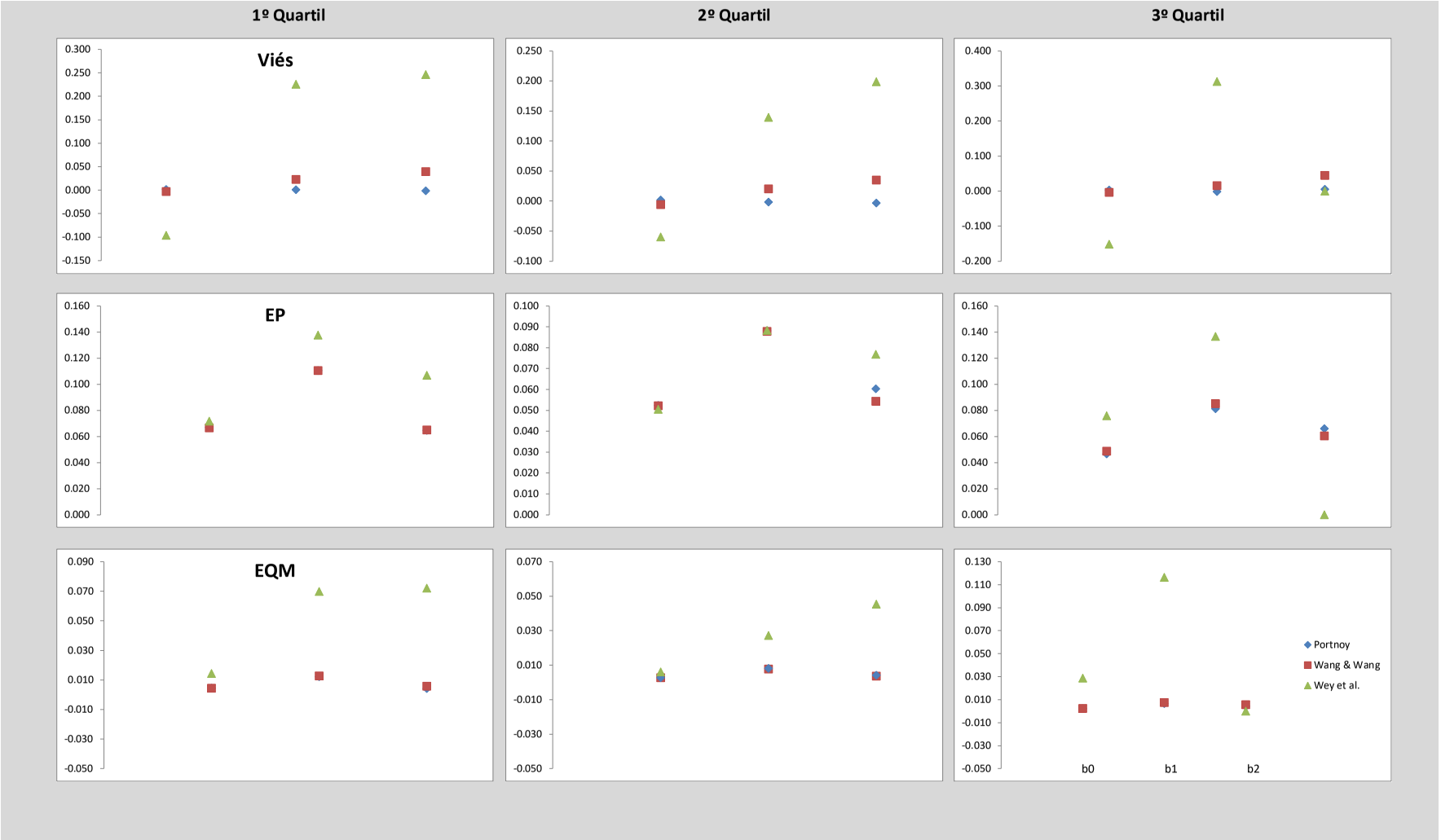

Figura C.6: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição Weibull.

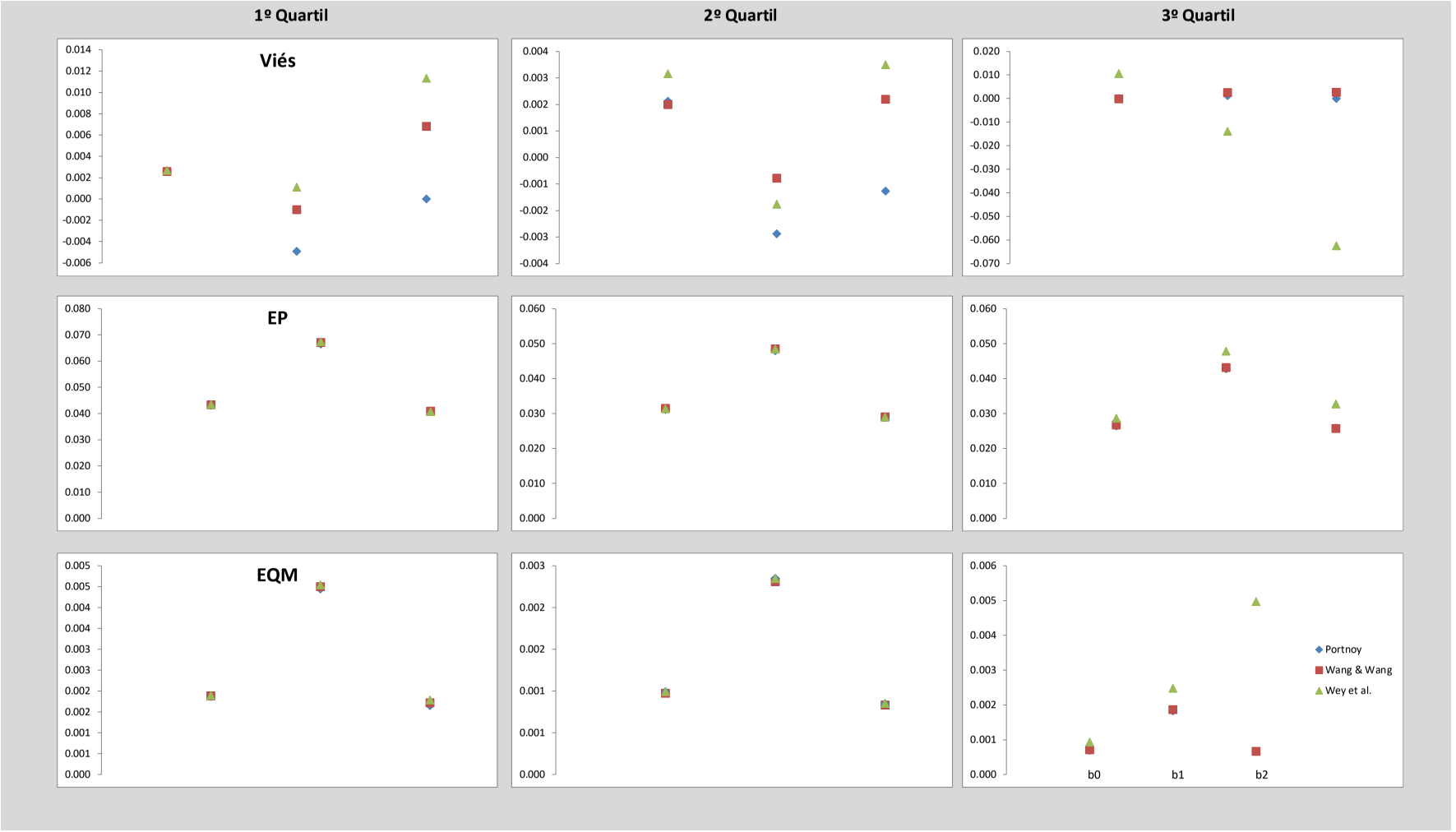

Figura C.7: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição Weibull.

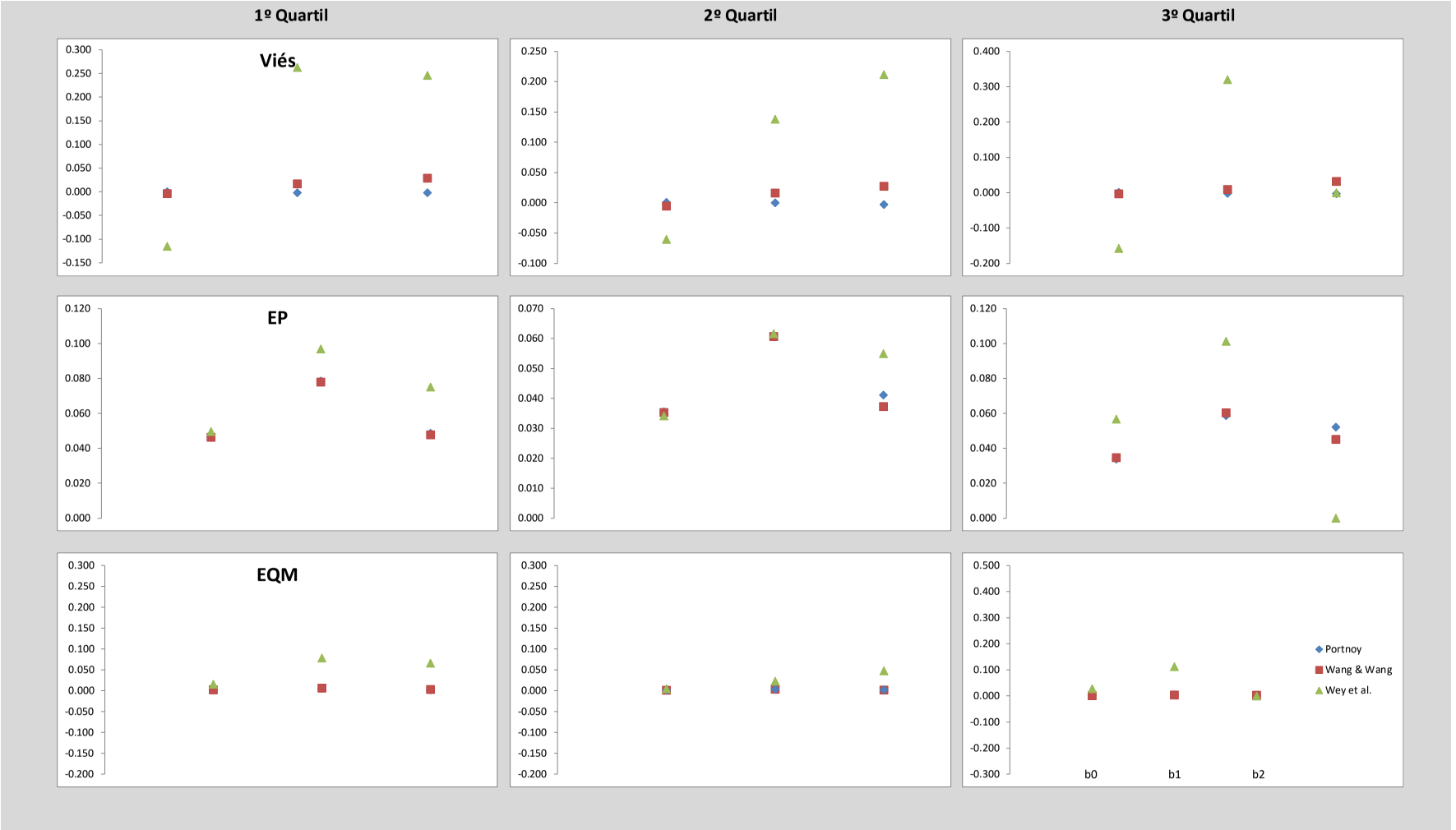

Figura C.8: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição Weibull.

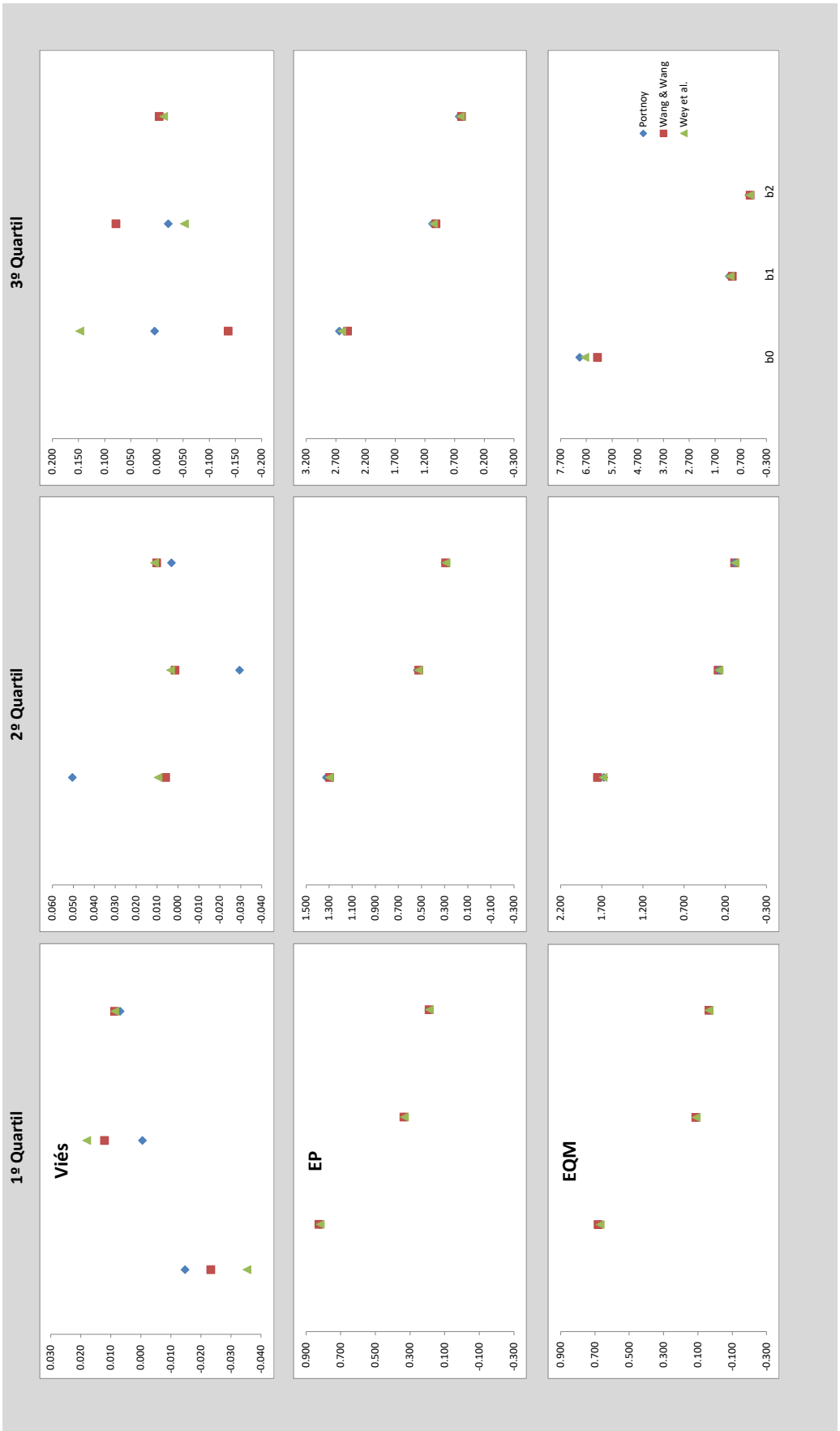

Figura C.9: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição log-logística.

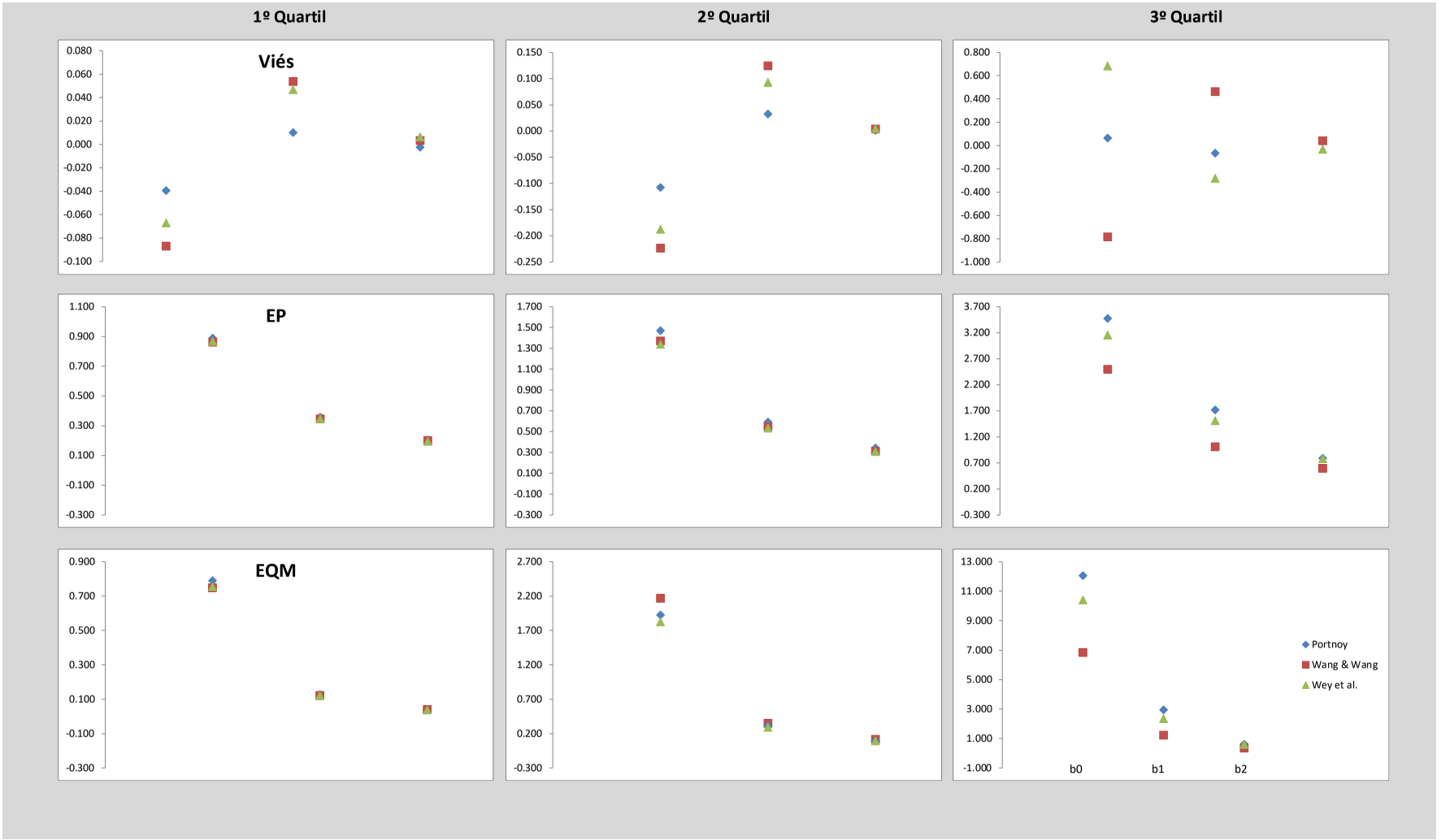

Figura C.10: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=400, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição log-logística.

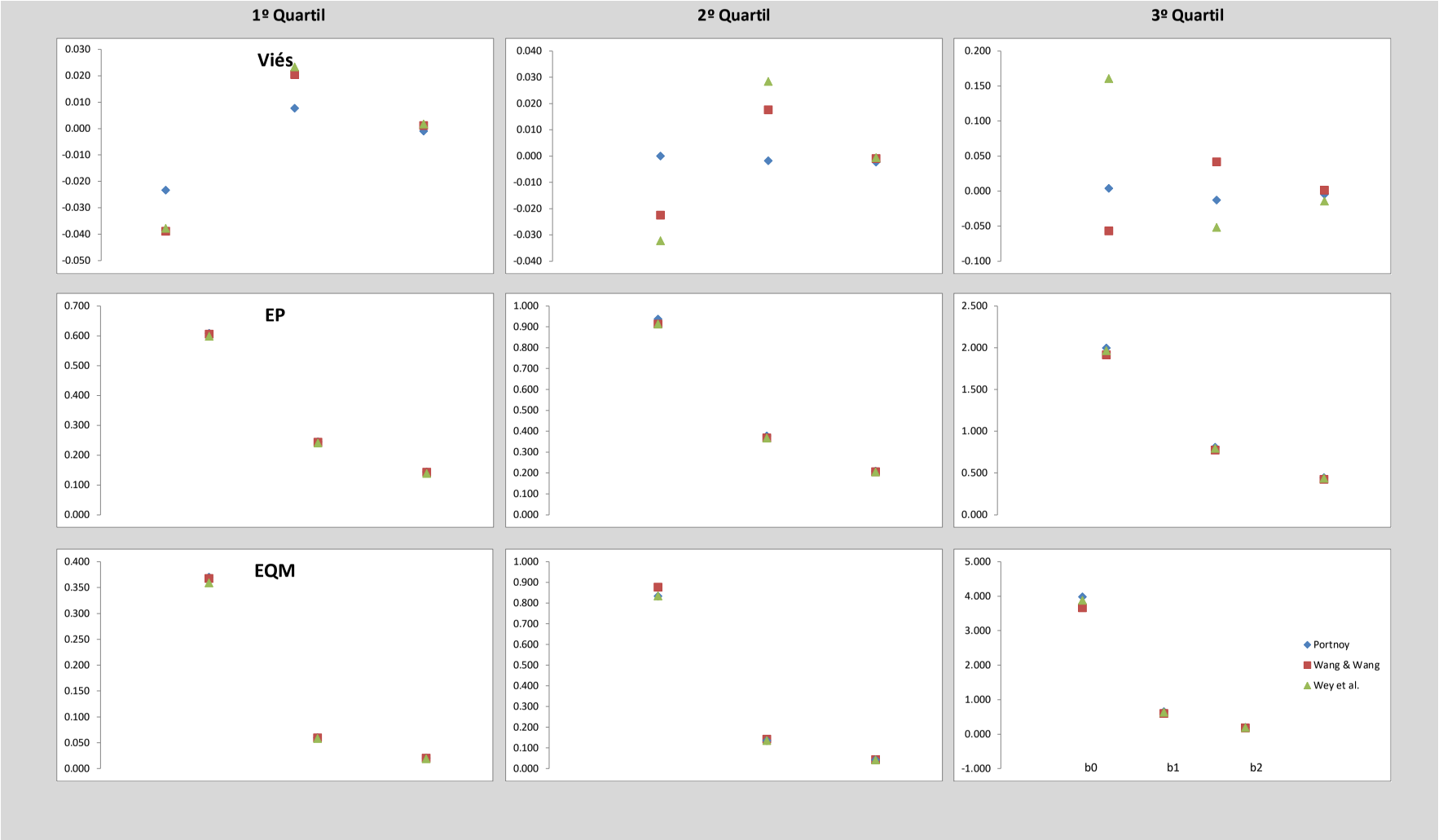

Figura C.11: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 25%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição log-logística.

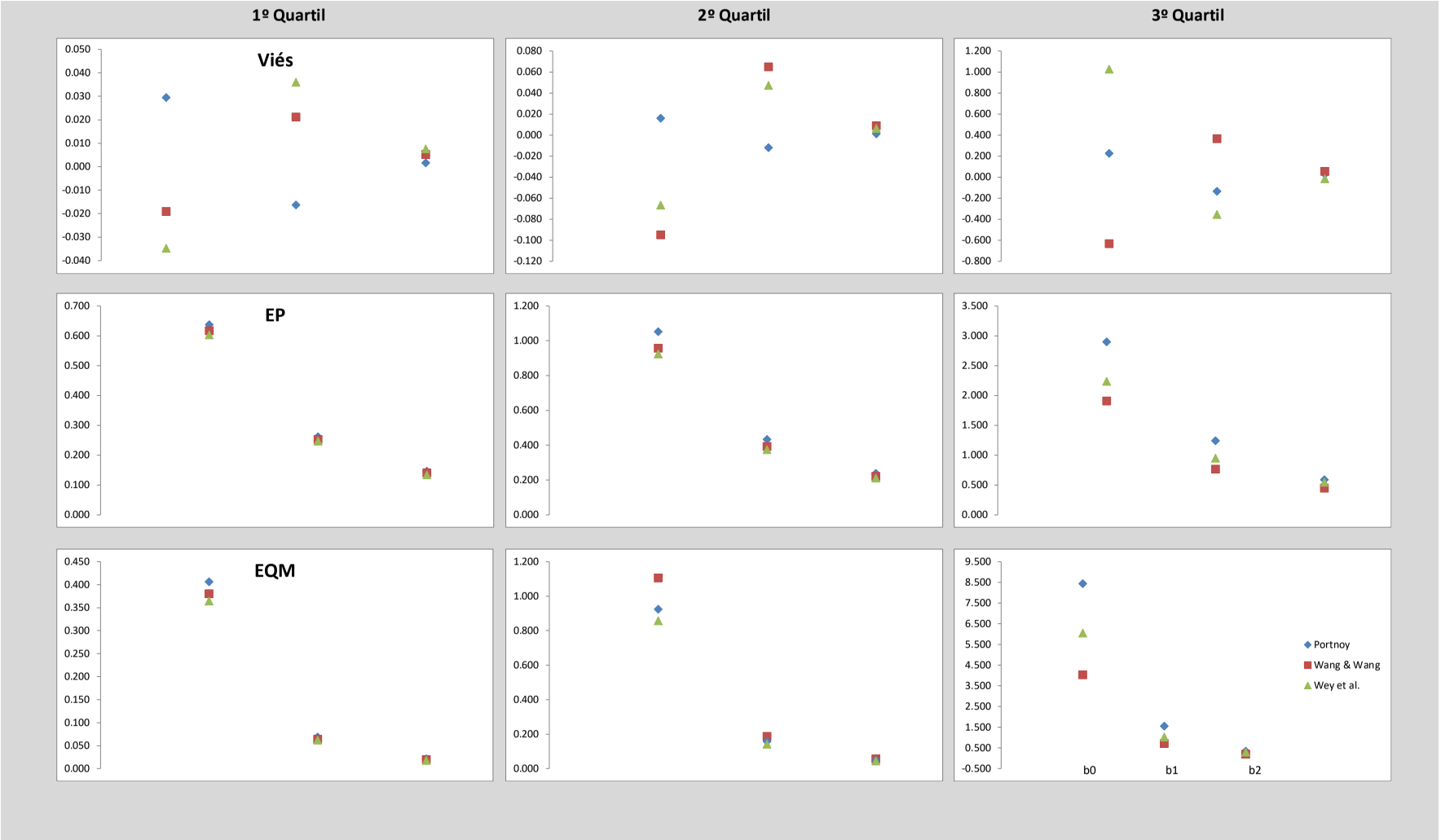

Figura C.12: Resultado do estudo de simulação baseado em 1.000 amostras de tamanhos N=800, com proporção de censura igual a 50%, para comparação das três metodologias de regressão quantílica para dados censurados, avaliados nos quantis $\tau \in (0,25; 0,50; 0,75)$ no modelo linear com distribuição log-logística.

Referências Bibliográficas

Barrodale, I. e Roberts, F. D. K. (1973). "An improved algorithm for discrete L_1 linear approximation". *SIAM Journal on Numerical Analysis*, 10(5), pp. 839–848.

Botter, D. A., Sandoval, M. C., Fujiwara, L. M. e Melo, M. G. N. (2012). *Relação entre as taxas de leucócitos, linfócitos e monócitos com o prognóstico de paciêntes com insuficiência cardíaca de diferentes etiologias*. Rel. téc. Universidade de São Paulo.

Bou-Hamad, I., Larocque, D. e Ben-Ameur, H. (2011). "A review of survival trees". *Statistics Surveys*, 5, pp. 44–71.

Breiman, L. (1996). "Bagging predictors". *Machine learning*, 24(2), pp. 123–140.

Breiman, L., Friedman, J. H., Stone, C. J. e Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Buchinsky, M. e Hahn, J. (1999). "An alternative estimator for the censored quantile regression model". *Econometrica*, 66(3), pp. 653–671.

Chen, C. e Wei, Y. (2005). "Computational issues for quantile regression". *Sankhyā: The Indian Journal of Statistics*, 67(Part 2), pp. 399–417.

Chen, S. e Khan, S. (2001). "Semiparametric estimation of a partially linear censored regression model". *Econometric Theory*, 17(3), pp. 567–590.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. ABE – Projeto Fisher. Edgard Blücher.

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

Duncan, G. M. (1986). "A semi-parametric censored regression estimator". *Journal of Econometrics*, 32(1), pp. 5–34.

Efron, B. (1967). "The two sample problem with censored data". Em: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 4. University of California Press, Berkeley, CA, pp. 831–853.

Efron, B. e Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Hall, P. e Sheather, S. J. (1988). "On the distribution of a studentized quantile". *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3), pp. 381–391.

Hastie, T. J. e Tibshirani, R. J. (1990). *Generalized additive models*. Vol. 43. CRC Press.

Heritier, S., Cantoni, E., Copt, S. e Victoria-Feser, M.-P. (2009). *Robust methods in Biostatistics*. Vol. 825. John Wiley & Sons.

Honoré, B. e Powell, J. L. (1994). "Pairwise difference estimators of censored and truncated regression models". *Journal of Econometrics*, 64(1), pp. 241–278.

Honoré, B., Khan, S. e Powell, J. L. (2002). "Quantile regression under random censoring". *Journal of Econometrics*, 109(1), pp. 67–105.

Horowitz, J. L. (1986). "A distribution-free least squares estimator for censored linear regression models". *Journal of Econometrics*, 32(1), pp. 59–84.

Hothorn, T., Lausen, B., Benner, A. e Radespiel-Tröger, M. (2004). "Bagging survival trees". *Statistics in Medicine*, 23(1), pp. 77–91.

Khan, S. e Powell, J. L. (2001). "Two-step estimation of semiparametric censored regression models". *Journal of Econometrics*, 103(1), pp. 73–110.

Klein, J. P. e Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Koenker, R. (2005). *Quantile regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. e Bassett Jr, G. (1978). "Regression quantiles". *Econometrica*, 46(1), pp. 33–50.

Koenker, R. W. e d'Orey, V. (1987). "Algorithm AS 229: Computing regression quantiles". *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3), pp. 383–393.

Lee, J. W. (1996). "Some versatile tests based on the simultaneous use of weighted log-rank statistics". *Biometrics*, 52, pp. 721–725.

Li, Q. e Racine, J. (2003). "Nonparametric estimation of distributions with categorical and continuous data". *Journal of Multivariate Analysis*, 86(2), pp. 266–292.

Lindgren, A. (1997). "Quantile regression with censored data using generalized L_1 minimization". *Computational Statistics & Data Analysis*, 23(4), pp. 509–524.

McKeague, I. W., Subramanian, S. e Sun, Y. (2001). "Median regression and the missing information principle". *Journal of Nonparametric Statistics*, 13(5), pp. 709–727.

Moon, C.-G. (1989). "A Monte Carlo comparison of semiparametric Tobit estimators". *Journal of Applied Econometrics*, 4(4), pp. 361–382.

Morgan, J. N. e Sonquist, J. A. (1963). "Problems in the analysis of survey data, and a proposal". *Journal of the American Statistical Association*, 58(302), pp. 415–434.

Mosteller, F. e Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Addison-Wesley Series in Behavioral Science. Reading, MA: Addison-Wesley.

Neter, J., Kutner, M. H., Nachtsheim, C. J. e Wasserman, W. (1996). *Applied linear statistical models*. Vol. 4. Irwin Chicago.

Newey, W. K. (1991). "Efficient estimation of Tobit models under conditional symmetry". Em: *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*. Ed. por W. A. Barnett, J. Powell e G. E. Tauchen. International Symposia in Economic Theory and Econometrics. Cambridge University Press, pp. 291–336.

Peng, L. e Huang, Y. (2008). "Survival analysis with quantile regression models". *Journal of the American Statistical Association*, 103(482), pp. 637–649.

Portnoy, S. (2003). "Censored regression quantiles". *Journal of the American Statistical Association*, 98(464), pp. 1001–1012.

Portnoy, S. e Koenker, R. (1997). "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators". *Statistical Science*, 12(4), pp. 279–300.

Powell, J. L. (1984). "Least absolute deviations estimation for the censored regression model". *Journal of Econometrics*, 25(3), pp. 303–325.

— (1986a). "Censored regression quantiles". *Journal of Econometrics*, 32(1), pp. 143–155.

— (1986b). "Symmetrically trimmed least squares estimation for Tobit models". *Econometrica*, 54(6), pp. 1435–1460.

Rudser, K. D., LeBlanc, M. L. e Emerson, S. S. (2012). "Distribution-free inference on contrasts of arbitrary summary measures of survival". *Statistics in Medicine*, 31(16), pp. 1722–1737.

Santos, B. R. (2012). "Modelos de Regressão Quantílica". Dissertação de Mestrado. Universidade de São Paulo.

Scott, D. W. e Sain, S. R. (2005). "9-Multidimensional Density Estimation". Em: *Data Mining and Data Visualization*. Ed. por C. R. Rao, E. J. Wegman e J. L. Solka. Handbook of Statistics 24. Elsevier, pp. 229 –261.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York.

Wang, H. J. e Wang, L. (2009). "Locally weighted censored quantile regression". *Journal of the American Statistical Association*, 104(487), pp. 1117–1128.

Wang, L. (2008). "Nonparametric test for checking lack of fit of the quantité regression model under random censoring". *Canadian Journal of Statistics*, 36(2), pp. 321–336.

Wey, A., Wang, L. e Rudser, K. (2014). "Censored quantile regression with recursive partitioning-based weights". *Biostatistics*, 15(1), pp. 170–181.

Yang, S. (1999). "Censored median regression using weighted empirical survival and hazard functions". *Journal of the American Statistical Association*, 94(445), pp. 137–145.

Ying, Z., Jung, S.-H. e Wei, L.-J. (1995). "Survival analysis with median regression models". *Journal of the American Statistical Association*, 90(429), pp. 178–184.