



Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Regressão quantílica para dados com censura intervalar

Alessandra Analu Moreira da Silva

Orientador: Prof. Dr. Antonio Eduardo Gomes

Brasília, 07 de março de 2019

Regressão quantílica para dados com censura intervalar

Alessandra Analu Moreira da Silva

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

Brasília, 07 de março de 2019

Regressão quantílica para dados com censura intervalar

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

Texto aprovado por:

Comissão Julgadora:

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

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

Agradecimentos

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

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

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

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

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

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

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

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

“All victories hide an abdication”.

Simone de Beauvoir

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

Florence Nightingale

Resumo

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

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

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

Abstract

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

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

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

Lista de Figuras

2.1	Exemplo de $\hat{q}_\tau = \sum \rho_\tau(y_i - q_\tau)$, em que $\tau \in (0, 25, 0, 5, 0, 75)$	8
3.1	Gráficos do método proposto nessa dissertação.	26
3.2	Exemplo de diagrama de soma acumulada e sua função minorante convexa máxima.	32
3.3	Gráfico da função de sobrevivência estimada para os dados do exemplo proposto.	41
5.1	Gráficos da relação entre o Tempo (intervalo para a data de compra do celular atual em dias) e a Idade (em anos).	64
5.2	Gráfico do ajuste das retas de regressão, utilizando o quantil de ordem $\tau = (0, 2; 0, 4; 0, 6; 0, 8)$ e segundo o método proposto nessa dissertação. . .	66
5.3	Gráficos da relação entre o Tempo (em anos) de ocorrência da primeira cárie e a Idade (em anos) do início da escovação.	68
5.4	Gráfico do ajuste das retas de regressão, utilizando o quantil de ordem $\tau = (0, 1; 0, 2)$ e segundo o método proposto nessa dissertação.	69
A.1	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Normal}(0, 0 ; 0, 3)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	84
A.2	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Normal}(0, 0 ; 0, 3)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	85

A.3	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Normal}(0, 0 ; 0, 3)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	86
A.4	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Lognormal}(0, 0 ; 0, 3)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	87
A.5	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Lognormal}(0, 0 ; 0, 3)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	88
A.6	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Lognormal}(0, 0 ; 0, 3)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	89
A.7	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Logística}(0, 0 ; 0, 2)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	90
A.8	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Logística}(0, 0 ; 0, 2)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ .	91

A.9	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Logística}(0, 0 ; 0, 2)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	92
A.10	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Weibull}(3, 0 ; 1, 0)$: Vício do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	93
A.11	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Weibull}(3, 0 ; 1, 0)$: EP do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	94
A.12	Gráficos dos resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Weibull}(3, 0 ; 1, 0)$: EQM do método de Zhou et al. e de Imputação da estimativa de b_0 e b_1 , apresentando os seus respectivos quantis de ordem τ	95
B.1	Gráficos das estimativas não paramétricas da curva de sobrevivência: (a) Curva de sobrevivência global e da idade 20 anos (pontilhada); (b) Curva de sobrevivência global e da idade 30 anos (pontilhada); (c) Curva de sobrevivência global e da idade 40 anos (pontilhada); (d) Curva de sobrevivência global e da idade 50 anos (pontilhada); (e) Curva de sobrevivência global e da idade 60 anos (pontilhada); (f) Curva de sobrevivência global e da idade 72 anos (pontilhada).	98
B.2	Gráficos das estimativas não paramétricas da curva de sobrevivência global e das idades.	99

B.3	Gráficos das estimativas não paramétricas da curva de sobrevivência: (a) Curva de sobrevivência global e do início de escovação com 0,5 anos (pontilhada); (b) Curva de sobrevivência global e do início de escovação com 1,5 anos (pontilhada); (c) Curva de sobrevivência global e do início de escovação com 2,5 anos (pontilhada); (d) Curva de sobrevivência global e do início de escovação com 3,5 anos (pontilhada); (e) Curva de sobrevivência global e do início de escovação com 4,5 anos (pontilhada); (f) Curva de sobrevivência global e do início de escovação com 5,5 anos (pontilhada).	100
B.4	Gráficos das estimativas não paramétricas da curva de sobrevivência global e das idades de início de escovação.	101

Lista de Tabelas

2.1	Estimativas dos parâmetros para regressão quantílica.	20
3.1	Tempos de sobrevivência de 10 observações, com censura dos tempos ordenados y_2 , y_7 e y_8 e estimativas da função de sobrevivência e função distribuição acumulada, $S(y_i)$ e $F(y_i)$, respectivamente.	41
3.2	Resultado do ajuste para o modelo de regressão quantílica.	45
4.1	Resultados da simulação para o exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Normal}(0, 0 ; 0, 3)$	57
4.2	Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Lognormal}(0, 0 ; 0, 3)$	58
4.3	Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Logística}(0, 0 ; 0, 2)$	59
4.4	Resultados da simulação para exemplo 4.1, baseados no tamanho da amostra ($n = 200$), no número de amostras ($n^* = 20$) e considerando $\tilde{e}_i \sim \text{Weibull}(3, 0 ; 1, 0)$	60
5.1	Informações sobre o celular atual fornecidas pelos entrevistados.	62
5.2	Medidas resumo dos limites superior e inferior das datas desde a aquisição do celular (em dias).	63
5.3	Frequência das faixas etárias dos entrevistados.	63

5.4	Estimativas para os parâmetros dos modelos de regressão quantílica, os erros padrão e intervalos de confiança.	65
5.5	Frequência das faixas de idade das crianças de quando iniciaram a escovação dos dentes.	67
5.6	Medidas resumo dos limites superior e inferior das idades da primeira ocorrência de cárie (em anos).	67
5.7	Estimativas para os parâmetros dos modelos de regressão quantílica, os erros padrão e e intervalos de confiança.	68

Sumário

1. Introdução	1
2. Regressão Quantílica	5
2.1 Quantis via otimização	7
2.2 Estimação para regressão quantílica	9
2.3 Propriedades da Regressão quantílica	13
2.4 Inferência para o modelo de regressão quantílica	15
2.4.1 Intervalo de confiança: método assintótico	15
2.4.2 Intervalo de confiança: método <i>bootstrap</i>	17
2.4.3 Teste de Hipótese Linear Geral	19
3. Regressão quantílica para dados censurados	23
3.1 Censura intervalar	26
3.1.1 Função de verossimilhança	27
3.2 Regressão quantílica para censura intervalar	30
3.2.1 Regressão isotônica	31
3.2.1.1 Estimador não paramétrico de máxima verossimilhança (ENPMV)	
- censura intervalar	33
3.2.2 Núcleo estimador	35
3.2.2.1 Dados não censurados	36
3.2.2.2 Dados censurados	37
3.2.3 Suavização da estimativa da distribuição condicional de Y dado valor	
das covariáveis contínuas	38
3.2.4 Método de Portnoy	38

3.2.4.1	Ponderação via Kaplan-Meier	40
3.2.4.2	Algoritmo recursivo de Portnoy	43
3.2.5	Método proposto	47
3.2.6	Método de Zhou	48
3.2.6.1	Correção do vício	52
4.	<i>Estudo de simulação para comparação dos métodos</i>	55
5.	<i>Aplicação das metodologias a dados empíricos</i>	61
5.1	Bancos de dados	61
6.	<i>Considerações finais</i>	71
	<i>Referências</i>	73
	<i>Apêndice</i>	81
A.	<i>Gráficos do estudo de simulação para comparação dos métodos</i>	83
B.	<i>Gráficos da aplicação das metodologias a dados empíricos</i>	97

Introdução

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

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

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

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

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

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

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

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

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

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

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

Regressão Quantílica

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

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

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

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

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

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

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

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

2.1 Quantis via otimização

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

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

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

O quantil de ordem τ para Y é definido como

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

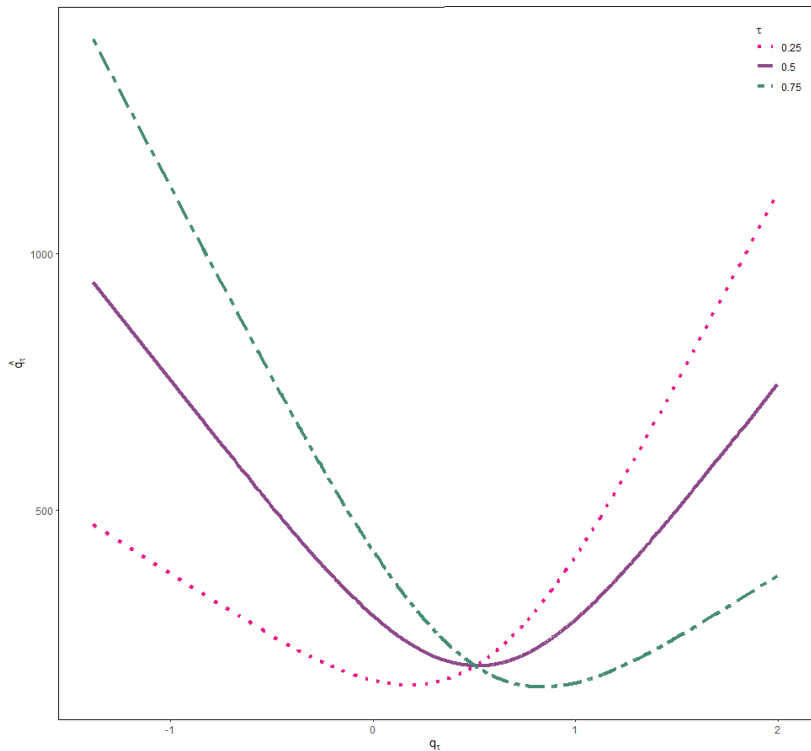


Figura 2.1: Exemplo de $\hat{q}_\tau = \sum \rho_\tau(y_i - q_\tau)$, em que $\tau \in (0, 25, 0, 5, 0, 75)$.

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

2.2 Estimação para regressão quantílica

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

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

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

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

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

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

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

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

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

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

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

Algoritmo Simplex

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

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} & \left[\sum_{i \in \{i: y_i \geq x_i \beta\}} \tau |y_i - \mathbf{x}_i' \beta| + \sum_{i \in \{i: y_i < x_i' \beta\}} (1 - \tau) |y_i - \mathbf{x}_i' \beta| \right] \\ \min_{(\beta, \mathbf{u}, \mathbf{v}) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} & \{ \tau \mathbf{1}'_n \mathbf{u} + (1 - \tau) \mathbf{1}'_n \mathbf{v} \mid \mathbf{y} = \mathbf{X} \beta + \mathbf{u} - \mathbf{v} \}, \\ & \text{sujeito a } \beta \in \mathbb{R}^p, \mathbf{u} \geq 0 \text{ e } \mathbf{v} \geq 0, \end{aligned} \quad (2.9)$$

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

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

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

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

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

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

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

Algoritmo de Ponto Interior

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

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

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

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

$$B(x, \mu \mid Ax = b), \quad (2.12)$$

em que

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

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

como (Koenker et al., 2017):

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

em que

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

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

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

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

$$AX^2A'y = AX^2c - \mu AX^11_n.$$

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

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

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

$$Ax = b, x \geq 0,$$

$$A'y + z = c, z \geq 0,$$

$$Xz = \mu 1_n.$$

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

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

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

2.3 Propriedades da Regressão quantílica

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

(i) Equivariância de escala:

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

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

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

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

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

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

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

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

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

Já a regressão quantílica obtém o seguinte resultado:

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

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

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

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

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

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

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

2.4 Inferência para o modelo de regressão quantílica

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

2.4.1 Intervalo de confiança: método assintótico

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

em que

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

e

$$\mathbf{H}_n(\tau) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n x_i x_i' f_i(\hat{\beta}(\tau_i)),$$

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

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

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

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

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

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

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

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

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

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

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

2.4.2 Intervalo de confiança: método *bootstrap*

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

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

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

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

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

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

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

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

2.4.3 Teste de Hipótese Linear Geral

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

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

$$H_0 : \mathbf{R}\beta = \mathbf{r}, \quad (2.18)$$

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

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

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

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

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

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

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

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

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

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

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

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

Tabela 2.1 - Estimativas dos parâmetros para regressão quantílica.

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

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

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

$$\hat{Q}_Y(0, 10|x) = 1,656 + 3,545x_i,$$

$$\hat{Q}_Y(0, 25|x) = 2,220 + 4,629x_i,$$

$$\hat{Q}_Y(0, 50|x) = 2,925 + 4,823x_i,$$

$$\hat{Q}_Y(0, 75|x) = 3,575 + 4,956x_i,$$

$$\hat{Q}_Y(0, 90|x) = 4,211 + 4,904x_i.$$

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

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

Referências Bibliográficas

- Abrevaya J., Economic applications of quantile regression. Springer, 2002.
- Ali A., Nonparametric spatial rainfall characterization using adaptive kernel estimator, Journal of Geographic Information and Decision Analysis, 1998, vol. 2, p. 34
- Austin P. C., Tu J. V., Daly P. A., Alter D. A., The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy, Statistics in medicine, 2005, vol. 24, p. 791.
- Ayer M., Brunk H. D., Ewing G. M., Reid W. T., Silverman E., An empirical distribution function for sampling with incomplete information, The annals of mathematical statistics, 1955, pp 641–647.
- Barlow R., Bartholomew D., Bremner J., Brunk H., Statistical inference under order restrictions (the theory and application of isotonic regression). John Wiley & Sons, 1972.
- Barrodale I., Roberts F. D., An improved algorithm for discrete l_1 linear approximation, SIAM Journal on Numerical Analysis, 1973, vol. 10, p. 839.
- Becker N. G., Melbye M., Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for hiv positivity, Australian & New Zealand Journal of Statistics, 1991, vol. 33, p. 125.
- Bedi A. S., Edwards J. H., The impact of school quality on earnings and educational returns-evidence from a low-income country, Journal of Development Economics, 2002, vol. 68, p. 157.

- Bogaerts K., Komarek A., Lesaffre E., *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. Chapman and Hall/CRC, 2017
- Buchinsky M., Changes in the US wage structure 1963-1987: Application of quantile regression, *Econometrica: Journal of the Econometric Society*, 1994, pp 405–458.
- Budd J. W., McCall B. P., The Grocery stores wage distribution: A semi-parametric analysis of the role of retailing and labor market institutions, *ILR Review*, 2001, vol. 54, p. 484.
- Cade B. S., Terrell J. W., Schroeder R. L., Estimating effects of limiting factors with regression quantiles, *Ecology*, 1999, vol. 80, p. 311.
- Cai T., Cheng S., Semiparametric regression analysis for doubly censored data, *Biometrika*, 2004, vol. 91, p. 277.
- Chamberlain G., Quantile regression, censoring, and the structure of wages, *Advances in econometrics*, 1994, vol. 1, p. 171.
- Chang M. N., et al., Weak convergence of a self-consistent estimator of the survival function with doubly censored data, *The Annals of Statistics*, 1990, vol. 18, p. 391.
- Chay K. Y., Honore B. E., Estimation of semiparametric censored regression models: an application to changes in black-white earnings inequality during the 1960s, *Journal of Human Resources*, 1998, pp 4–38.
- Chay K. Y., Powell J. L., Semiparametric censored regression models, *Journal of Economic Perspectives*, 2001, vol. 15, p. 29.
- Chen C., Wei Y., Computational issues for quantile regression, *Sankhyā: The Indian Journal of Statistics*, 2005, pp 399–417.
- Dantzig G. B., *Linear programming, Operations research*, 2002, vol. 50, p. 42.
- Davino C., Furno M., Vistocco D., *Quantile regression: theory and applications*. John Wiley & Sons, 2013.

- Efron B., The two sample problem with censored data. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , vol. 4, 1967, p. 831.
- Efron B., Tibshirani R. J., An introduction to the bootstrap. CRC press, 1994.
- Eide E. R., Showalter M. H., Factors affecting the transmission of earnings across generations: A quantile regression approach, *Journal of Human Resources*, 1999, pp 253–267.
- Eide E. R., Showalter M. H., Sims D. P., The effects of secondary school quality on the distribution of earnings, *Contemporary Economic Policy*, 2002, vol. 20, p. 160.
- Fiacco A., McCormick G., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* Wiley New York NY Google Scholar, 1968.
- Finkelstein D. M., Wolfe R. A., A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics*, 1985, pp 933–945.
- Fortin N. M., Lemieux T., Rank regressions, wage distributions, and the gender gap, *Journal of Human Resources*, 1998, pp 610–643.
- Frisch R., La résolution des problèmes de programme linéaire par la méthode du potentiel logarithmique, *Cahiers du Seminaire D'Econometrie*, 1956, pp 7–23.
- Gentleman R., Geyer C. J., Maximum likelihood for interval censored data: Consistency and computation, *Biometrika*, 1994, vol. 81, p. 618.
- Gilchrist W., *Statistical modelling with quantile functions*. CRC Press, 2000.
- Giolo S. R., Colosimo E. A., *Análise de sobrevivência aplicada*, Edgard Blucher, 2006.
- Gould W., et al., Quantile regression with bootstrapped standard errors, *Stata Technical Bulletin*, 1993., vol. 2
- Groeneboom P., Wellner J. A., *Information bounds and nonparametric maximum likelihood estimation*. Birkhäuser, 1992.
- Hall P., Sheather S. J., On the distribution of a studentized quantile, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1988, pp 381–391

- Hao L., Immigration and wealth inequality: A distributional approach. In Invited seminar at The Center for the Study of Wealth and Inequality, Columbia University , 2005.
- Hao L., Sources of wealth inequality: Analyzing conditional distribution. In Invited seminar at The Center for Advanced Social Science Research , 2006.a
- Hao L., Sources of wealth inequality: Analyzing conditional location and shape shifts. In Research Committee on Social Stratification and Mobility (RC28) of the International Sociological Association (ISA) Spring meeting 2006 in Nijmegen , 2006.b
- Hao L., Naiman D. Q., Quantile regression. Quantitative applications in the social sciences, 2007.
- He X., Hu F., Markov chain marginal bootstrap, Journal of the American Statistical Association, 2002, vol. 97, p. 783.
- Hendricks W., Koenker R., Hierarchical spline models for conditional quantiles and the demand for electricity, Journal of the American statistical Association, 1992, vol. 87, p. 58
- Heritier S., Cantoni E., Copt S., Victoria-Feser M.-P., Robust methods in Biostatistics. vol. 825, John Wiley & Sons, 2009.
- Hollander M., Wolfe D. A., Chicken E., Nonparametric statistical methods. vol. 751, John Wiley & Sons, 2013.
- Huang J., Wellner J. A., Interval censored survival data: a review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics , 1997, p. 123.
- Huber P. J., et al., The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , vol. 1, 1967, p. 221.
- Karmarkar N., A new polynomial-time algorithm for linear programming. In Proceedings of the sixteenth annual ACM symposium on Theory of computing , 1984, p. 302.
- Karvanen J., Rantanen A., Luoma L., Survey data and Bayesian analysis: a cost-efficient way to estimate customer equity, Quantitative Marketing and Economics, 2014, vol. 12, p. 305

- Kim Y. J., Cho H., Kim J., Jhun M., Median regression model with interval censored data, *Biometrical Journal*, 2010, vol. 52, p. 201.
- Kocherginsky M., He X., Mu Y., Practical confidence intervals for regression quantiles, *Journal of Computational and Graphical Statistics*, 2005, vol. 14, p. 41.
- Kocherginsky M. N., Extensions of Markov chain marginal bootstrap, University of Illinois at Urbana-Champaign, 2003, Tese de Doutorado
- Koenker R., Quantile regression, *Econometric Society Monographs*, 2005, vol. 38.
- Koenker R., Bassett Jr G., Regression quantiles, *Econometrica: journal of the Econometric Society*, 1978, pp 33.–50.
- Koenker R., Chernozhukov V., He X., Peng L., *Handbook of Quantile Regression*. CRC Press, 2017.
- Koenker R., Geling O., Reappraising medfly longevity: a quantile regression survival analysis, *Journal of the American Statistical Association*, 2001, vol. 96, p. 458.
- Koenker R., Machado J. A., Goodness of fit and related inference processes for quantile regression, *Journal of the american statistical association*, 1999, vol. 94, p. 1296
- Koenker R., Portnoy S., *Quantile regression*. ABE, 1996.
- Koenker R., Xiao Z., Inference on the quantile regression process, *Econometrica*, 2002, vol. 70, p. 1583.
- Koenker R. W., d'Orey V., Algorithm AS 229: Computing regression quantiles, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1987, vol. 36, p. 383.
- Lemieux T., Postsecondary education and increasing wage inequality, *American Economic Review*, 2006, vol. 96, p. 195.
- Li G., Zhang C. H., et al., Linear regression with interval censored data, *The Annals of Statistics*, 1998, vol. 26, p. 1306.
- Lin G., He X., Portnoy S., Quantile regression with doubly censored data, *Computational Statistics & Data Analysis*, 2012, vol. 56, p. 797.

- Machado J. A., Mata J., Counterfactual decomposition of changes in wage distributions using quantile regression, *Journal of applied Econometrics*, 2005, vol. 20, p. 445.
- McKeague I. W., Subramanian S., Sun Y., Median regression and the missing information principle, *Journal of nonparametric statistics*, 2001, vol. 13, p. 709.
- Melly B., Decomposition of differences in distribution using quantile regression, *Labour economics*, 2005, vol. 12, p. 577.
- Ou F. S., Zeng D., Cai J., Quantile regression models for current status data, *Journal of statistical planning and inference*, 2016, vol. 178, p. 112.
- Pan W., Smooth estimation of the survival function for interval censored data, *Statistics in Medicine*, 2000, vol. 19, p. 2611.
- Parzen M., Wei L., Ying Z., A resampling method based on pivotal estimating functions, *Biometrika*, 1994, vol. 81, p. 341.
- Pollard D., Empirical processes: theory and applications. In NSF-CBMS regional conference series in probability and statistics , 1990, p. i.
- Portnoy S., Censored regression quantiles, *Journal of the American Statistical Association*, 2003, vol. 98, p. 1001.
- Portnoy S., Koenker R., et al., The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, *Statistical Science*, 1997, vol. 12, p. 279.
- Powell J. L., Least absolute deviations estimation for the censored regression model, *Journal of Econometrics*, 1984, vol. 25, p. 303.
- Powell J. L., Censored regression quantiles, *Journal of econometrics*, 1986, vol. 32, p. 143.
- Rao C., Zhao L., Asymptotic normality of LAD estimator in censored regression models, *Mathematical methods of statistics*, 1993, vol. 2, p. 228.
- Rao C. R., Zhao L., Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap, *Sankhyā: The Indian Journal of Statistics, Series A*, 1992, pp 323–331.

- Rasteiro L. R., Regressão quantílica para dados censurados, Universidade de São Paulo, 2017, Dissertação de Mestrado
- Ren J. J., Gu M., et al., Regression M-estimators with doubly censored data, *The Annals of Statistics*, 1997, vol. 25, p. 2638.
- Rodrigues T., Dortet-Bernadet J.-L., Fan Y., Pyramid quantile regression, arXiv preprint arXiv:1606.05407, 2016
- Santos B. R. d., Modelos de regressão quantílica, Universidade de São Paulo, 2012, Dissertação de Mestrado
- Scharf F. S., Juanes F., Sutherland M., Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques, *Ecology*, 1998, vol. 79, p. 448.
- Silva A. L., Estimação da função quantílica para dados com censura intervalar, Universidade de Brasília, 2011, Dissertação de Mestrado
- Silverman B. W., Density estimation for statistics and data analysis. Routledge, 1986.
- Turnbull B. W., Nonparametric estimation of a survivorship function with doubly censored data, *Journal of the American statistical association*, 1974, vol. 69, p. 169.
- Turnbull B. W., The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1976, pp 290–295.
- Van Buuren S., Flexible imputation of missing data. Chapman and Hall/CRC, 2012
- Vanobbergen J., Martens L., Lesaffre E., Declerck D., The Signal-Tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results, *European Journal of Paediatric Dentistry*, 2000, vol. 2, p. 87
- Wang H. J., Wang L., Locally weighted censored quantile regression, *Journal of the American Statistical Association*, 2009, vol. 104, p. 1117.
- Wei Y., Pere A., Koenker R., He X., Quantile regression methods for reference growth charts, *Statistics in medicine*, 2006, vol. 25, p. 1369.

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

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

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

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