Modelos de Regressão Quantílica

Bruno Ramos dos Santos

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

Programa: Estatística

Orientadora: Profa. Dra. Silvia Nagib Elian

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, março de 2012

Modelos de Regressão Quantílica

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Bruno Ramos dos Santos
e aprovada pela Comissão Julgadora.

Comissão Julgadora:

- Profa. Dra. Silvia Nagib Elian (orientadora) - IME-USP
- Prof. Dr. Gilberto Alvarenga Paula - IME-USP
- Prof. Dr. Ronaldo Dias - IMECC-UNICAMP

Agradecimentos

Eu agradeço, em primeiro lugar, à minha orientadora, Profa. Silvia Nagib Elian, sem a qual este trabalho não teria sido feito. Sua disponibilidade e dedicação em me orientar foram primordiais na execução dessa dissertação.

Eu agradeço também aos meus pais, Heleno e Ana, e aos meus irmãos, Heleana, Mariane e Luís Henrique, pelo apoio incondicional e incentivo aos estudos, desde a minha infância até os dias atuais.

Não posso deixar de lembrar de meus colegas do programa de Mestrado, companheiros de estudos nos finais de semana, Akira Uematsu, Gleyce Noda, Karina Nakamura, Lina Thomas, Sérgio Coichev, Tiago Almeida, aos quais eu sou muito agradecido pela companhia durante essa etapa da minha formação.

E, por último, gostaria de agradecer ao apoio da minha namorada, Anouch Kurkdjian, que esteve presente em todos momentos desse mestrado, desde a decisão em deixar o emprego para me dedicar integralmente aos estudos até a entrega da dissertação, sempre me ajudando com conselhos e me incentivando nos momentos de incertezas. Suas palavras foram muito importantes para a concretização desse trabalho.

Resumo

Este trabalho trata de modelos de regressão quantílica. Foi feita uma introdução a essa classe de modelos para motivar a discussão. Em seguida, conceitos inferenciais, como estimação, intervalos de confiança, testes de hipóteses para os parâmetros são discutidos, acompanhados de alguns estudos de simulação. Para analisar a qualidade do ajuste, são apresentados o coeficiente de determinação e um teste de falta de ajuste para modelos de regressão quantílica. Também é proposta a utilização de gráficos para análise da qualidade do ajuste considerando a distribuição Laplace Assimétrica. Uma aplicação utilizando um banco de dados com informação sobre renda no Brasil foi utilizado para exemplificar os tópicos discutidos durante o texto.
Palavras-chave: Regressão Quantílica; Conceitos Inferenciais; Qualidade do Ajuste; Modelos de Renda.

Abstract

This work is about quantile regression models. An introduction was made to this class of models to motivate the discussion. Then, inferential concepts, like estimation, confidence intervals, tests of hypothesis for the parameters are discussed, followed by some simulation studies. To analyse goodness of fit, a coefficient of determination and a lack-of-fit test for quantile regression models are presented. It's also proposed the use of graphs for the goodness of fit analysis considering the Asymmetric Laplace Distribution. An application using a data base with information about income in Brazil was used to exemplify the topics discussed during the text.

Keywords: Quantile Regression; Inferential Concepts; Goodness of fit; Income Models.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Segundo Montgomery et al. (2001), a análise de regressão pode ser descrita como uma técnica estatística utilizada para investigar e modelar o relacionamento entre variáveis. Como exemplo, um pesquisador poderia estar interessado em saber se a variável idade influencia e de que forma influencia a variável renda. Nesse caso, uma amostra de pessoas de diferentes idades com suas respectivas rendas poderia gerar um ajuste de um modelo de regressão, que auxiliaria a explicar a relação entre essas duas variáveis.

Se considerado o método de minimização de mínimos quadrados para estimação dos parâmetros do modelo, alguns trabalhos podem ser utilizados para consulta, por exemplo Searle (1971) e Rao (1973).

No entanto, este texto se foca em outra técnica, que pode ser chamada de minimização de erros absolutos ponderados, a qual, veremos, resulta nos modelos de regressão quantílica, que é o tema da presente dissertação. Sobre esse tema, apenas para motivar a discussão, poderíamos usar um trecho do texto de Mosteller e Tukey (1977), citado em Koenker (2005):

> O que a curva de regressão faz é dar um grande resumo das médias das distribuições correspondentes ao conjunto dos x's observados. Nós poderíamos ir além e computar diversas curvas de regressão correspondendo aos vários pontos percentuais da distribuição e dessa forma ter uma visão mais completa desse conjunto. Usualmente isso não é feito, e logo a regressão frequentemente dá uma visão mais incompleta. Assim como a média dá uma visão incompleta de uma única distribuição, também a curva de regressão dá uma visão incompleta correspondente para um conjunto de distribuições. [1]

Nas próximas seções, segue uma introdução ao uso da técnica de regressão quantílica linear, assim como a sua definição e alguns exemplos para melhor elucidar o tema.

[1] Traduzido do seguinte trecho em inglês: *What the regression curve does is give a grand summary for the averages of the distribution corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distribution and thus get a more complete picture of the set. Ordinarily this is not done, and so the regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.*

1.1 Erros Quadráticos ou Erros Absolutos

Modelos de regressão são de extrema utilidade em estudos estatísticos, devido tanto à sua facilidade de interpretação quanto à grande diversidade de programas estatísticos que hoje são capazes de fazer esse tipo de análise. E dentre os métodos de estimação dos parâmetros do modelo, podemos citar o método de minimização dos quadrados dos erros como o mais utilizado. Isso talvez se deva ao fato da facilidade computacional para implementar tal cálculo e, além disso, em caso de distribuição normal dos erros do modelo, o estimador obtido por este método possui boas propriedades.

De acordo com Stigler (1986), o método dos mínimos quadrados tem origem no início do século XIX, com o trabalho de Legendre. Antes disso porém, no século XVIII, Boscovich teria sugerido uma maneira de estimar os parâmetros do modelo por um método que pode ser considerado como o precursor do método de minimização dos erros absolutos. Boscovich tinha interesse em estimar a elipcidade da Terra e para alcançar esse resultado propôs um modelo de regressão que era aceito na época para descrever a relação das medidas que utilizava em seu cálculo. Inicialmente, tendo obtido as medidas associadas ao modelo em cinco locais diferentes, Boscovich traçou todas as retas possíveis que passavam por pelo menos dois pontos coletados e sugeriu como estimativa do parâmetro da inclinação da sua reta de regressão, o valor médio de todas as inclinações possíveis de retas e uma outra estimativa que retirava duas dessas possibilidades para calcular a média. Em uma segunda tentativa de propor um modelo para calcular a medida de interesse, o cientista propôs estimar os parâmetros do modelo a partir da minimização dos erros absolutos, porém com a restrição de que a soma de tais erros fosse igual a zero.

É importante então ressaltar porque a minimização dos quadrados dos erros alcançou maior relevância na estimação dos parâmetros nos modelos de regressão em detrimento da minimização dos erros absolutos, ainda que este último tenha surgido antes do primeiro. Ocorre que mesmo quando do seu surgimento, Boscovich encontrou dificuldades para computar os valores dos estimadores propostos a partir desse método. Somente com o avanço dos computadores e a utilização de programação linear, principalmente, é que a utilização dessa técnica começou a crescer.

Entretanto, embora o método dos mínimos quadrados seja o mais utilizado, este tem algumas limitações que levaram à busca por outros métodos. Em primeiro lugar, esta metodologia está fortemente associada à distribuição normal dos erros. Quando essa não é alcançada, ou seja, quando os erros estão distribuídos de uma forma assimétrica ou possuem uma cauda mais pesada que a da distribuição normal, então a performance deste método na estimação dos parâmetros é ruim. Na verdade, o que ocorre é que as suposições básicas do modelo não são verificadas. Nesse caso, Box e Cox (1964) sugerem transformar a variável resposta na tentativa de satisfazer às suposições do modelo, porém esta alternativa pode dificultar a interpretação dos parâmetros do modelo ajustado. Outro trabalho importante nesse sentido é Nelder e Wedderburn (1972), que definem uma nova classe de modelos: os Modelos

Lineares Generalizados, que caracterizam a relação entre a variável resposta pertencente à família exponencial e suas variáveis preditoras.

Ainda com relação aos problemas na utilização do método de mínimos quadrados, há a questão da influência que *outliers* exercem nas estimativas dos parâmetros do modelo. Isso faz com que seja necessário sempre que se utiliza essa técnica, uma criteriosa avaliação de quanto cada ponto influencia no ajuste do modelo, o que pode se tornar bastante trabalhoso, uma vez que tanto *outliers* na variável resposta quanto nas variáveis preditoras podem atrapalhar na identificação da verdadeira relação entre as variáveis de interesse.

Por outro lado, conforme veremos na continuação deste texto, o método de minimização dos erros absolutos é robusto na presença de *outliers* na variável resposta. Além disso, quando a distribuição dos erros não é normal, esse método se mostra melhor para descrever uma posição central da distribuição condicional da variável resposta, ao estimar o valor mediano da distribuição. Já a regressão quantílica, conforme veremos na próxima seção, se baseia no método dos erros absolutos, porém para estimar os diversos quantis de interesse é feita uma ponderação na minimização desses erros.

1.2 Definição de quantis

Os quantis de uma população ou de uma amostra podem ser definidos da seguinte forma:

> O quantil de ordem τ de uma população ou de uma amostra é o valor m tal que $100\tau\%$ dos valores populacionais ou amostrais são inferiores a ele, com $0 < \tau < 1$.

A definição para o caso populacional também pode ser enunciada utilizando a função de distribuição acumulada da variável aleatória X, em que

$$F(x) = P(X \leqslant x).$$

Então utilizando a função inversa da distribuição acumulada no ponto τ, define-se que

$$F^{-1}(\tau) = \inf\{x : F(x) \geqslant \tau\} \tag{1.1}$$

é o quantil de ordem τ da variável aleatória X. A mediana, nesse caso, seria definida como $F^{-1}(1/2)$. O primeiro quartil e o terceiro quartil seriam $F^{-1}(1/4)$ e $F^{-1}(3/4)$, respectivamente.

No entanto, podemos definir o quantil de ordem τ ainda de uma terceira forma, que é essencial no entendimento dos modelos de regressão quantílica.

Inicialmente para a mediana, podemos pensar da seguinte forma. Seja Y com função de distribuição acumulada F. Estamos interessados no valor m que minimiza $E|Y - m|$. Esse valor é a mediana de Y. A prova desse resultado é simples e pode ser encontrada em Hao e Naiman (2007).

O resultado anterior pode ser generalizado para todos os quantis da seguinte maneira. Consideremos o problema da teoria da decisão de prever um valor da variável aleatória X com função distribuição de probabilidades F. Adotada a função de perda

$$\rho_\tau(u) = u(\tau - I(u < 0)), \quad 0 < \tau < 1, \tag{1.2}$$

em que I é a função indicadora, considere o problema de encontrar $\hat{x}$, um previsor de X, que minimize a perda esperada.

Então, temos

$$E\big[\rho_\tau(X - \hat{x})\big] = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x})dF(x).$$

Diferenciando esta expressão com relação a $\hat{x}$ e igualando a zero, obtemos

$$(1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau = 0.$$

Como F é monótona, qualquer elemento do conjunto $\{x : F(x) = \tau\}$ minimiza a perda esperada, ou seja, $\hat{x} = F^{-1}(\tau)$ minimiza a perda esperada para função de perda definida em (1.2), e $\hat{x}$ é o quantil de ordem τ segundo a definição (1.1).

Em particular, para $\tau = 1/2$,

$$\begin{aligned} E\big[\rho_{1/2}(X - \hat{x})\big] &= -\frac{1}{2}\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \frac{1}{2}\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x) \\ &= \frac{1}{2}E\big[|X - \hat{x}|\big] \end{aligned}$$

e, portanto, a minimização de $E\big[\rho_{1/2}(X - \hat{x})\big]$ é equivalente à minimização de $E\big[|X - \hat{x}|\big]$, que resulta em $\hat{x}$ igual à mediana.

Com essa definição para os quantis, podemos enunciar a idéia da regressão quantílica. Para isso, vamos inicialmente fazer um paralelo com o modelo linear com variável resposta com distribuição normal, pois essa técnica oferece um caminho para o desenvolvimento dos modelos de regressão quantílica.

Um resultado bastante conhecido na Estatística é que, dada uma amostra de n observações de uma variável aleatória Y, a média amostral é a solução do seguinte problema de minimização

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Logo, se a intenção é expressar a média condicional de Y dado $\boldsymbol{x}$ como uma função linear nos parâmetros $\boldsymbol{\beta}$, isto é, $\mu(x) = \boldsymbol{x}'\boldsymbol{\beta}$, então o estimador $\hat{\boldsymbol{\beta}}$ pode ser obtido pelo método de mínimos quadrados, ou seja, calculando

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i}'\boldsymbol{\beta})^2,$$

em que $\boldsymbol{x_i}'$ é a i-ésima linha da matriz X de valores não-aleatórios conhecidos.

Por outro lado, no problema da teoria da decisão de prever X por $\hat{x}$ com função de perda $\rho_\tau(\mu)$, se F é substituída por sua função de distribuição empírica

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leqslant x),$$

a minimização da perda esperada

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(x_i - \hat{x})$$

produz, nesse caso, o quantil amostral de ordem τ. Dessa forma, dada uma amostra de n observações da variável Y, o quantil amostral de ordem τ resolve o problema de minimização a seguir

$$\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - q). \tag{1.3}$$

Inclusive, podemos verificar esse resultado de uma forma bastante prática com um exemplo simples. Para ilustrar esse resultado, vamos considerar uma amostra de 1000 observações com distribuição Uniforme[0,1]. Calculemos então o valor de (1.3) para q igual a cada valor de x presente na amostra para três diferentes valores de τ. Utilizamos no exemplo τ igual a 0,25, 0,50 e 0,75. O resultado se encontra na Figura 1.1.

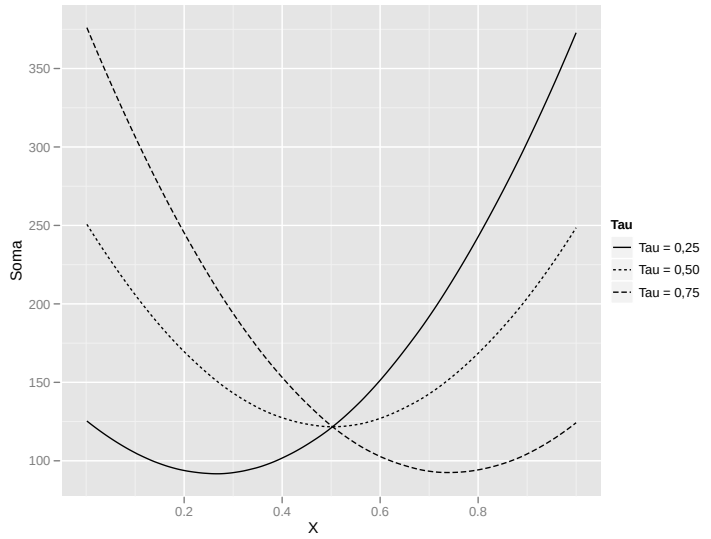

Figura 1.1: *Valores de* $\sum \rho_\tau(x_i - q)$ *para* $q = x$*, x pertencente a uma amostra de 1000 observações com distribuição Uniforme[0,1], com* $\tau = 0,25, 0,50, 0,75$.

Podemos notar que o valor amostral que faz com que a soma em (1.3) obtenha o menor valor possível para $\tau = 0,5$ se encontra perto do valor 0,5, ou seja, praticamente coincidindo com a mediana de uma amostra com distribuição Uniforme[0,1], o que era esperado. O mesmo acontece para os outros valores de τ utilizados no exemplo.

Dessa forma, Koenker e Bassett (1978) sugeriram em seu artigo seminal o seguinte procedimento. Se a intenção é especificar o quantil condicional de Y dado $\boldsymbol{x}$ como uma função linear nos parâmetros da forma $Q_\tau(Y|x) = \boldsymbol{x}'\boldsymbol{\beta}(\boldsymbol{\tau})$, em que $\boldsymbol{\beta}(\boldsymbol{\tau})$ é um vetor de parâmetros, para estimar $\boldsymbol{\beta}(\boldsymbol{\tau})$ basta encontrar então $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ que seja a solução do problema de minimização

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i}'\boldsymbol{\beta}). \tag{1.4}$$

Com isso, no caso do método dos mínimos quadrados, se escrevemos a relação linear entre as duas variáveis da seguinte forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{1.5}$$

em que ε_i tem média 0, logo podemos dizer que a média condicional da variável $Y|X$ pode ser escrita como

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Então, se o interesse é estudar diversos quantis da distribuição condicional da variável resposta Y, supondo que valem relações lineares do tipo

$$y_i = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \cdots + \beta_p(\tau) x_{ip} + u_i, \tag{1.6}$$

em que u_i são variáveis aleatórias independentes e identicamente distribuídas com quantil de ordem τ igual a zero, podemos dizer então que o quantil condicional de ordem τ de $Y|X$ é

$$Q_\tau(Y|x) = \beta_0(\tau) + \beta_1(\tau) x_1 + \cdots + \beta_p(\tau) x_p. \tag{1.7}$$

Se considerarmos a distribuição Laplace Assimétrica para os erros em (1.6), verificamos então que o estimador de máxima verossimilhança para o vetor de parâmetros $\boldsymbol{\beta}(\boldsymbol{\tau})$ coincide com o estimador apresentado em (1.4), conforme discutido no Apêndice C.

Devemos ressaltar que o vetor de parâmetros $\boldsymbol{\beta}$ deve ser indexado a τ pois um dos interesses, nesse caso, é exatamente estudar se esse vetor assume diferentes valores para τ's diferentes.

Uma colocação importante deve ser feita aqui, uma vez que, diferentemente da análise de regressão usual, os modelos de regressão quantílica têm uma característica bastante peculiar, que é a quantidade de curvas a serem interpretadas. Inclusive essas diversas curvas podem

ser construídas considerando o mesmo conjunto de variáveis explicativas ou não. Por outro lado, esses modelos podem ser utilizados para concentrar a análise somente em algum ponto específico da distribuição condicional da variável resposta. Podemos dizer que os modelos de regressão quantílica ajudam a obter uma visão mais completa da relação entre as variáveis estudadas.

Sobre os resultados obtidos a partir do ajuste dos modelos, também podemos verificar algumas relações importantes. Sem perda de generalidade, vamos considerar o quantil condicional de $Y|X$ em (1.7), com apenas uma variável explicativa, ou seja,

$$Q_\tau(Y|x) = \alpha(\tau) + \beta(\tau)x.$$

Assim, não é difícil ver que se os coeficientes estimados para $\beta(\tau)$ para diferentes valores de τ são muito próximos, aparentemente variando em torno de uma constante, então podemos dizer que há evidências a favor da suposição de que os erros são independentes e identicamente distribuídos. Porém, se esses coeficientes variam em função de τ, então os erros podem estar apresentando alguma forma de heterocedasticidade. Koenker (2005) exibe alguns exemplos e usa alguns dados para comparar essas duas situações e o comportamento das curvas estimadas pela regressão quantílica. A conclusão desse fato é que modelos de regressão quantílica são capazes de incorporar uma possível heterocedasticidade, que seria detectada a partir da variação das estimativas dos coeficientes $\beta(\tau)$ para diferentes τ's.

Tendo enunciado a idéia, resta mostrar como chegar à solução $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ de (1.4) e como inferir sobre os parâmetros do modelo, mas esse tema será tratado com mais detalhe no Capítulo 2. Veremos que essa solução pode ser encontrada utilizando métodos de programação linear já implementados em diversos aplicativos estatísticos. Na sequência, estudaremos alguns exemplos para mostrar as diferenças entre a utilização da média condicional e da regressão quantílica, que estima os quantis condicionais.

1.3 Exemplos

Poluição de cidades norte-americanas

Para exemplificar a diferença entre as duas abordagens apontadas na seção anterior, vamos comparar as metodologias ajustando um modelo a dados reais. Para isso, vamos usar os dados de poluição do ar medida em 41 cidades norte-americanas entre os anos de 1969 e 1971. Os dados foram retirados de Hand et al. (1994) e encontram-se no Apêndice B.

Para verificar a relação entre duas variáveis a partir da análise de regressão, foram utilizadas as variáveis quantidade de dióxido de enxofre em miligramas por metro cúbico (SO2) e temperatura em graus Fahrenheit (Temp). Ambas medidas são valores médios observados entre os anos de 1969 e 1971. O interesse nesse exemplo é quantificar o efeito da temperatura na poluição do ar medida pela quantidade de dióxido de enxofre presente por metro cúbico.

Vamos ajustar dois modelos para estimar o efeito da temperatura na poluição do ar, um

estimando o parâmetro do modelo utilizando o método dos mínimos quadrados e outro utilizando o método da mínima soma dos erros absolutos, também conhecido como regressão L_1 e que se baseia em encontrar $\boldsymbol{\beta}$ que minimiza a soma $\sum |y_i - \boldsymbol{x_i}'\boldsymbol{\beta}|$. É importante mencionar que este modelo está estimando a mediana condicional e é um caso particular da regressão quantílica, para $\tau = 1/2$.

Podemos observar as duas retas ajustadas na Figura 1.2, em que a linha pontilhada representa a regressão da mediana, enquanto a linha cheia representa a regressão da média. Na Tabela 1.1 podem ser consultadas as estimativas para cada parâmetro do modelo. Sem entrar em detalhes sobre a inferência relativa aos parâmetros, verificou-se que todos as estimativas são significantes ao nível de 5%.

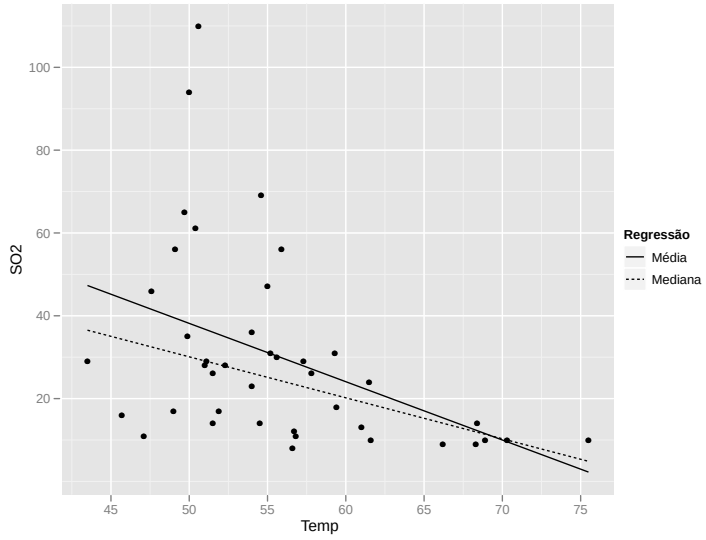

Figura 1.2: *Comparação do ajuste da regressão da média e da regressão da mediana.*

Parâmetro	Regressão da Média	Regressão da Mediana
Intercepto	108,57	79,56
Temperatura	-1,41	-0,99

Tabela 1.1: *Estimativa dos parâmetros dos modelos ajustados a partir do Modelo Linear Normal e da Regressão Quantílica*

A análise dos resultados nos permite dizer que os dois ajustes são muito próximos nesse exemplo. A diferença que se pode notar é com relação à inclinação entre as duas retas ajustadas, em que a reta da regressão da média tem um efeito negativo maior que a regressão da mediana. Em outras palavras, segundo a regressão da média, a cada 1 ºF de aumento da temperatura média das cidades americanas, estima-se que há a diminuição de 1,41 miligramas na quantidade média de dióxido de enxofre na atmosfera, enquanto que de acordo com a regressão da mediana, a diminuição seria de 0,99 miligramas na mediana da quantidade de dióxido de enxofre a cada grau aumentado na temperatura. Além disso, podemos citar que a estimativa da regressão da média parece ter sido mais influenciada pelas observações nas

cidades de Chicago e Providence, que têm temperaturas médias baixas, porém concentração de dióxido de enxofre no ar bastante alta, o que faz com que essas cidades fiquem um pouco mais afastadas da nuvem de pontos. Essas cidades têm menor influência na reta estimada pelo método da regressão quantílica, uma vez que essa reta apresentou menor inclinação. Em outras palavras, o modelo de regressão quantílica se mostrou mais robusto nesse simples exemplo.

Imunoglobulina G em crianças

Continuando a motivação inicial de analisar algumas propriedades dos modelos de regressão quantílica, vamos utilizar agora os dados de Isaacs et al. (1983) sobre a concentração de imunoglobulina G, em gramas por litro de sangue, em crianças com idade entre 6 meses e 6 anos. O interesse no problema é explicar a variação da imunoglobulina G em função da idade. Além disso, gostaríamos de mostrar como a utilização de modelos de regressão quantílica pode fornecer uma visão mais completa da distribuição condicional da variável resposta. Os dados estão disponíveis no Apêndice B.

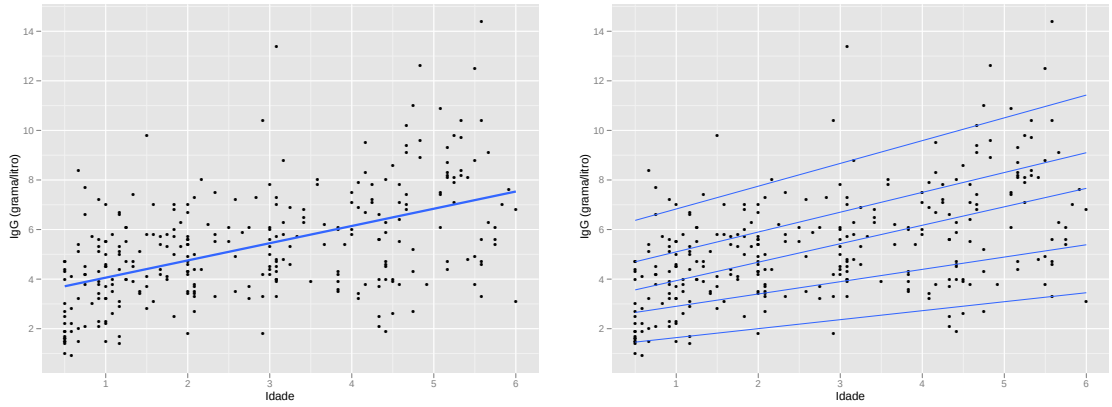

Figura 1.3: *Ajuste de um modelo de regressão linear e diversos ajustes da regressão quantílica para os valores de* $\tau = 0,05; 0,25; 0,50; 0,75; 0,95$.

Na Figura 1.3, podemos comparar a abordagem comumente utilizada em análise de regressão e uma possível abordagem da regressão quantílica. Lado a lado, podemos notar tanto a estimativa do efeito da idade na concentração média de imunoglobulina G quanto as estimativas dos efeitos da idade em diferentes quantis da distribuição condicional de imunoglobulina G. Um dos intuitos dessa comparação é discutir como a utilização de modelos de regressão quantílica possibilita analisar de forma abrangente o efeito de idade na concentração de imunoglobulina G. Isso é possível uma vez que esse tipo de análise pode traçar a relação tanto em regiões mais centrais com a mediana, a qual poderia inclusive substituir a estimativa do efeito médio, quanto nas caudas da distribuição condicional.

É interessante notar nesse exemplo que as inclinações das retas no segundo gráfico são muito parecidas, o que significa dizer que o efeito de idade é praticamente o mesmo em toda a distribuição condicional da variável resposta. Nesse caso, o apelo em usar os modelos de regressão quantílica é menor, pois não há indicativos de que a relação das duas variáveis

em estudo seja diferente para τ's diferentes. Entretanto, há diversos exemplos na literatura em que se verifica que o efeito de variáveis independentes diferem para quantis diferentes da distribuição condicional (ver, por exemplo, Buchinsky (1994)). Por outro lado, mesmo no caso de paralelismo entre as retas estimadas, ainda restaria o interesse de estimar os quantis condicionais da variável resposta em função da idade.

Curvas de crescimento do Índice de Massa Corpórea para homens

Por último, nesse capítulo, vamos considerar um exemplo de curvas de crescimento geradas a partir de modelos de regressão quantílica. Temos o interesse em estimar curvas de crescimento relacionadas ao Índice de Massa Corpórea (IMC), que é a razão entre o peso (em kg) e o quadrado da altura (em m^2), medida usualmente utilizada para definir sobrepeso e obesidade. O índice IMC é uma medida importante porque a obesidade pode estar relacionada a diversos problemas de saúde, tanto físicos como psicológicos. Nesse sentido, é importante relacionar o IMC com outras variáveis, por exemplo, idade. Com a variável idade, é possível estudar como é a variação do IMC ao longo dos anos para diversos quantis, utilizando regressão quantílica.

Para isso, vamos utilizar dados de uma amostra de um centro de estatísticas para saúde dos Estados Unidos da América. Trata-se do *National Center for Health Studies*, que conduz uma pesquisa nacional anual para examinar a saúde e a nutrição dos cidadãos norte-americanos. Para a análise, foram consideradas as pesquisas entre os anos de 1999 e 2002. As variáveis utilizadas aqui serão somente IDADE e IMC, para homens com idade entre dois e oitenta anos. No total foram selecionados 8.202 homens para esse estudo, após a retirada de observações com informações faltantes em qualquer uma das variáveis de interesse. Na Figura 1.4, podemos observar a variação do IMC dos indivíduos selecionados por idade. Utilizamos um efeito gráfico prático para análises de dispersão com muitas observações, que deixa o gráfico mais claro onde há menor concentração de pontos e mais escuro onde há maior concentração de pontos. Dessa forma, podemos notar que muitos indíviduos participantes da pesquisa têm menos de 20 anos, e também que o IMC das pessoas após os 20 anos se situa principalmente entre 20 e 30.

Para construir os modelos de regressão para diversos quantis diferentes, vamos considerar diversas potências de idade, formulando assim um modelo polinomial, como em Chen (2005). Serão consideradas as mesmas potências em todos os quantis ajustados, sendo esses 0,03, 0,05, 0,10, 0,25, 0,50, 0,75, 0,90, 0,95 e 0,97. O quantil condicional do IMC em função das potências de idade pode ser escrito da seguinte forma:

$$\begin{aligned} Q_\tau(\text{IMC}|\text{Idade}) &= \beta_0(\tau) + \beta_1(\tau)\text{Idade}^{-1} + \beta_2(\tau)\text{Idade}^{1/2} + \beta_3(\tau)\text{Idade} + \beta_4(\tau)\text{Idade}^2 \\ &\quad +\beta_5(\tau)\text{Idade}^{3/2} + \beta_6(\tau)\text{Idade}^3. \end{aligned}$$

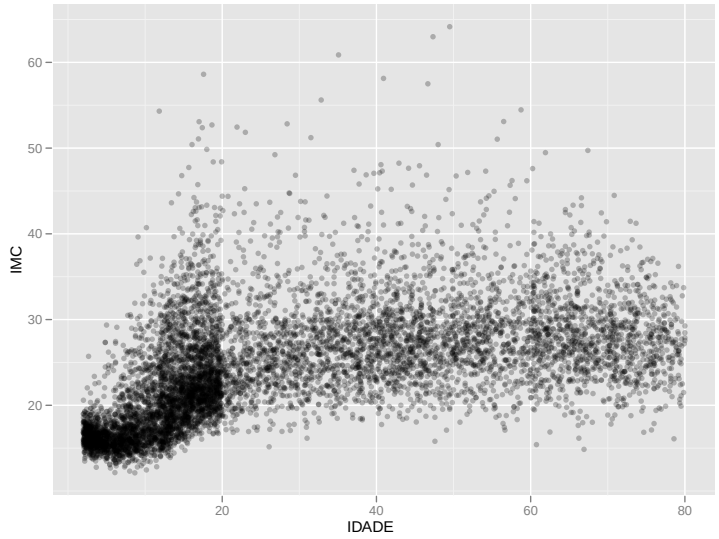

Figura 1.4: *Gráfico de dispersão entre IMC e idade.*

Os resultados do ajuste para a mediana podem ser vistos na Tabela 1.2. Novamente, não vamos entrar em detalhes sobre o cálculo da significância de cada estimativa, assunto que será tratado no Capítulo 2, porém podemos dizer que todas essas estimativas foram significantes ao nível de 5%.

Parâmetro	Estimativa
Intercepto	98,599
1 / Idade	-30,619
$\sqrt{\text{Idade}}$	-81,085
Idade	30,429
$\text{Idade}^{3/2}$	-4,968
Idade^2	0,325
Idade^3	-0,001

Tabela 1.2: *Estimativas dos parâmetros do modelo ajustado para a mediana condicional de IMC em função da Idade.*

Tendo ajustado os modelos para todos os quantis mencionados anteriormente, podemos construir as curvas de crescimento para cada quantil, a partir dos valores preditos para cada idade entre 2 e 80 anos. O resultado obtido pode ser observado na Figura 1.5.

Pode-se notar a partir do gráfico um crescimento em todas as curvas de forma bastante similar a partir dos 10 anos de idade até os 25 anos, aproximadamente. Em seguida, há um período de constância das estimativas dos quantis do IMC até praticamente os 70 anos de idade, momento em que há uma queda nos valores do índice. Um dos pontos de interesse nesse tipo de estudo é estimar parâmetros para se classificar se uma pessoa está acima do peso ou abaixo do peso normal para sua idade. Por exemplo, se considerarmos o quantil de ordem 97% como um delimitador para sobrepeso, podemos classificar se uma pessoa está com sobrepeso baseado em sua idade e IMC, a partir desse gráfico. O mesmo pode ser feito

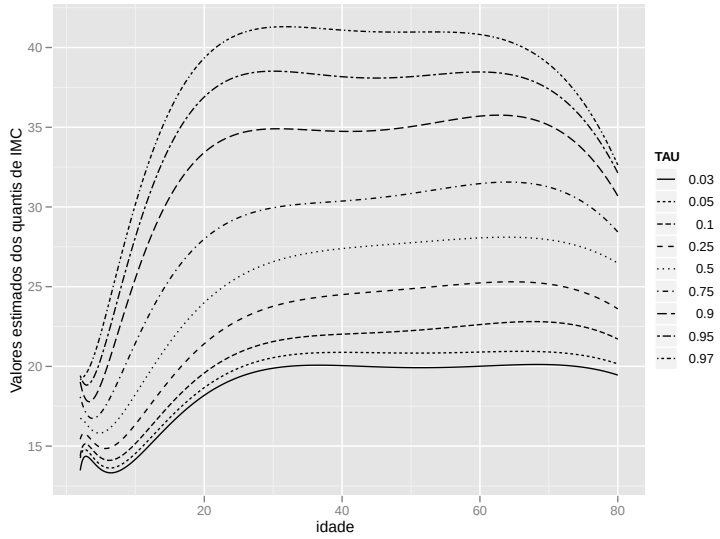

Figura 1.5: *IMC em função da idade para diversos valores de τ.*

também para a cauda inferior da distribuição condicional do índice IMC, verificando quais valores definem uma pessoa muito abaixo do peso esperado.

No próximo capítulo, vamos discutir um pouco mais sobre o processo de estimação dessas curvas, assim como da construção de intervalos de confiança e testes de hipóteses para os parâmetros do modelo.

1.4 Objetivos e organização do trabalho

Feitas essas considerações iniciais e uma introdução dos modelos de regressão quantílica, o objetivo fundamental do presente trabalho é motivar a utilização dos modelos de regressão quantílica. Com esse objetivo, apresentamos os principais métodos de inferência relacionados aos modelos de regressão quantílica. Além disso, também nos propomos a buscar na literatura os presentes métodos de análise da qualidade de ajuste para esses modelos. E por último, temos o interesse em aplicar as técnicas discutidas nesse texto em um conjunto de dados sobre renda no Brasil.

Sobre a organização do trabalho, no Capítulo 2, apresentamos os conceitos relacionados a estimação e inferência dos parâmetros, assim como algumas propriedades do modelo. No Capítulo 3, apresentaremos uma análise da qualidade do ajuste desses modelos. No Capítulo 4, aplicamos as técnicas apresentadas a um conjunto de dados reais. Finalmente, no Capítulo 5, discutimos algumas conclusões obtidas neste trabalho. Analisamos as vantagens e desvantagens dos métodos propostos e sugerimos algumas linhas de pesquisa para trabalhos futuros.

Capítulo 2

Inferência nos Modelos de Regressão Quantílica

Para os procedimentos inferenciais tratados nesse capítulo, vamos considerar a seguinte formulação. Seja $\boldsymbol{Y}$ um vetor $n \times 1$ de observações que seguem o seguinte modelo linear

$$\boldsymbol{Y} = X\boldsymbol{\beta}(\boldsymbol{\tau}) + \boldsymbol{\epsilon} \tag{2.1}$$

em que X é uma matriz de planejamento de constantes conhecidas $n \times p$, $\boldsymbol{\beta}(\boldsymbol{\tau})$ é um vetor $p \times 1$ de parâmetros desconhecidos, e $\boldsymbol{\epsilon}$ é um vetor de erros independentes e identicamente distribuídos com função de distribuição F e quantil de ordem τ igual a zero.

Tendo visto no Capítulo 1 a motivação para o presente trabalho, assim como exemplos da utilização da regressão quantílica, vamos tratar nessa parte do texto sobre outros importantes aspectos desses modelos, como a forma de estimação, construção de intervalos de confiança e testes de hipóteses relacionados aos parâmetros do modelo, além de discutirmos propriedades como equivariância e robustez que esses modelos de regressão quantílica apresentam. Finalizando, elaboramos estudos de simulação para verificar a acurácia de alguns métodos apresentados nesse capítulo.

Conforme já dito anteriormente, a estimação dos parâmetros do modelo de regressão quantílica depende de algoritmos de programação linear. O que esse texto se propõe a fazer não é discutir esses métodos em profundidade, mas apresentar a questão da programação linear envolvida nos modelos de regressão quantílica e mostrar quais as opções de uso e algumas diferenças de performance entre os métodos disponíveis nos softwares estatísticos.

Com relação à construção de intervalos de confiança para os parâmetros, indicaremos quais os diferentes procedimentos que podem ser utilizados, assim como discutiremos as principais dificuldades encontradas nessa parte da regressão quantílica. Além disso, vamos apresentar também testes de hipóteses lineares gerais para hipóteses do tipo $H_0 : C\boldsymbol{\beta}(\boldsymbol{\tau}) = \boldsymbol{c}$, em modelos de regressão quantílica como em (2.1). Em seguida, tanto para avaliar a cobertura dos intervalos de confiança quanto para analisar o poder e o nível de significância dos testes de hipóteses, foram realizados estudos de simulação.

Por último, enunciamos as propriedades de equivariância e robustez, que são alguns dos motivos pelos quais os modelos de regressão quantílica podem ser preferidos com relação à regressão normal linear, em que os parâmetros são estimados pelo método dos mínimos quadrados.

2.1 Estimação dos parâmetros

Um dos grandes atrativos dos modelos de regressão mais utilizados é a forma do estimador de mínimos quadrados para o vetor de parâmetros $\boldsymbol{\beta}$ quando a matriz X de planejamento do modelo é de posto completo e os erros são homocedásticos. Nesse caso, podemos escrever $\hat{\boldsymbol{\beta}}$, estimador de mínimos quadrados de $\boldsymbol{\beta}$, da seguinte forma

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y}, \tag{2.2}$$

em que X é a matriz de planejamento e $\boldsymbol{Y}$ é o vetor de variáveis respostas. Ainda que a estrutura de covariância dos erros seja um pouco mais complicada, por exemplo, heterocedástica, mas conhecida, o estimador do vetor de parâmetros ainda pode ser definido de forma fechada.

Infelizmente, o mesmo não pode ser dito para os modelos de regressão quantílica. Como o estimador é obtido a partir da minimização da soma de erros absolutos ponderados, conforme discutido no capítulo anterior, não é possível obter um estimador que possa ser calculado de forma direta.

Por esse motivo, os estudos baseados em modelos de regressão L_1 não obtiveram muito sucesso inicialmente, devido principalmente à complexidade computacional envolvida nesses problemas, situação que se alterou com a chegada dos algoritmos de programação linear. Somente a partir da descoberta de que o problema da minimização de erros absolutos poderia ser escrito como um problema de programação linear, é que os primeiros avanços da regressão L_1 aconteceram.

Considerando um modelo como em (2.1), devemos lembrar que para obter o estimador $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ tínhamos que minimizar a seguinte soma de erros absolutos ponderados

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i}'\boldsymbol{\beta}). \tag{2.3}$$

Koenker (2005) mostra que o estimador $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ pode ser obtido reformulando o problema anterior e transformando-o em um problema de programação linear. Inicialmente, podemos transformar a minimização de interesse em (2.3), como

$$\min_{(\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\nu})\in\mathbb{R}^p\times\mathbb{R}^{2n}_+} \left\{\tau \mathbf{1}_{\boldsymbol{n}}'\boldsymbol{\mu} + (1-\tau)\mathbf{1}_{\boldsymbol{n}}'\boldsymbol{\nu} | X\boldsymbol{\beta} + \boldsymbol{\mu} - \boldsymbol{\nu} = \boldsymbol{Y}\right\},$$

em que $\mathbf{1}_{\boldsymbol{n}}'$ denota um vetor $1 \times n$ de valores iguais a 1, $\boldsymbol{\mu}$ e $\boldsymbol{\nu}$ são vetores $n \times 1$, sendo

μ_i e ν_i seus respectivos termos. Esses valores são definidos como

$$\mu_i = \begin{cases} y_i - \hat{y}_i & \text{se } y_i - \hat{y}_i > 0, \\ 0, & \text{caso contrário;} \end{cases} \qquad \nu_i = \begin{cases} \hat{y}_i - y_i & \text{se } y_i - \hat{y}_i < 0, \\ 0, & \text{caso contrário,} \end{cases}$$

com $\hat{y}_i = \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$.

Em seguida, podemos enunciar um problema de programação linear (P) de forma usual, como,

$$\begin{array}{ll} \text{(P)} & \min\limits_{\theta} \boldsymbol{d'\theta} \\ \text{sujeito a} & B\boldsymbol{\theta} = \boldsymbol{Y} \\ & \boldsymbol{\theta} \geqslant 0, \end{array}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\phi'}, \boldsymbol{\varphi'}, \boldsymbol{\mu'}, \boldsymbol{\nu'})'$, $\boldsymbol{\phi} = [\boldsymbol{\beta}]_+$, $\boldsymbol{\varphi} = [-\boldsymbol{\beta}]_+$, $\boldsymbol{\nu}$ e $\boldsymbol{\mu}$ são os mesmos definidos anteriormente, $[\boldsymbol{z}]_+$ é a parte não negativa de um conjunto $\boldsymbol{z}$, ou seja, os termos β_{i+} de $[\boldsymbol{\beta}]_+$ podem ser definidos como

$$\beta_{i+} = \begin{cases} \beta_i & \text{se } \beta_i > 0, \\ 0, & \text{caso contrário.} \end{cases}$$

A transposta da matriz B é definida como

$$B' = \begin{bmatrix} X \\ -X \\ I_n \\ -I_n \end{bmatrix}$$

em que I_n representa a matriz identidade de ordem n. Além disso, o vetor $\boldsymbol{d}$ é definido como

$$\boldsymbol{d} = (\mathbf{0'}, \mathbf{0'}, \tau \mathbf{1_n}', (1-\tau)\mathbf{1_n'}),$$

em que $\mathbf{0}' = (0\ 0 \ldots 0)_p$. Esta formulação (P) representa um problema de programação linear padrão.

Inicialmente na utilização de programação linear em modelos de regressão L_1, podemos citar o algoritmo proposto por Barrodale e Roberts (1973), por ser um dos primeiros realmente eficientes para estimar os parâmetros do modelo. Sua implementação adapta o algoritmo *simplex* para o problema de minimização de desvios absolutos. Segundo Chen e Wei (2005), este algoritmo pode ser visto como computacionalmente exigente para bancos de dados com muitas observações, mas ainda assim razoável para conjuntos de dados com até 5000 observações e 50 variáveis. A adaptação desse algoritmo para o problema da regressão quantílica encontra-se em Koenker e d'Orey (1987).

Um procedimento mais eficiente para bancos de dados de grandes dimensões foi sugerido por Portnoy e Koenker (1997), no qual os autores utilizam um algoritmo de programação linear conhecido como ponto interior. Segundo Chen e Wei (2005), verificou-se que essa técnica

tem performance superior ao algoritmo *simplex*. Por esse motivo, esse algoritmo é preferível na presença de bancos de dados com muitas observações.

Retornando à questão da preferência pelo método dos mínimos quadrados na análise de regressão, um ponto importante em sua utilização é a facilidade computacional e rapidez do método, uma vez que, com o avanço da capacidade de processamento dos computadores, uma operação como multiplicação de matrizes e a inversão do resultado dessa multiplicação para obter uma estimativa do vetor de parâmetros pode ser considerada como uma tarefa trivial. Com relação à ordem de complexidade computacional, o algoritmo do método dos mínimos quadrados requer $\mathcal{O}(np^2)$ operações, ao passo que a ordem de complexidade do algoritmo da regressão quantílica é $\mathcal{O}(n^{5/2}p^3)$, o que coloca esse método em desvantagem se comparado ao método dos mínimos quadrados. Por esse motivo, Portnoy e Koenker (1997) propuseram uma modificação no algoritmo para a regressão quantílica, adicionando um passo de pré-processamento no algoritmo. Com essa melhoria, os autores obtiveram, em algumas situações, performances semelhantes às do método de mínimos quadrados.

Resumindo, dentre esses dois algoritmos mencionados, *simplex* e ponto interior, temos que o primeiro é o mais estável, pois sempre encontra uma solução para o problema, enquanto que o segundo pode apresentar dificuldades se existirem *outliers* nas variáveis explicativas. Todavia, o algoritmo de ponto interior é muito rápido para problemas com muitas observações, mas poucas variáveis independentes.

Para mais detalhes sobre performances dos algoritmos, assim como sugestões de melhoria nos processos computacionais de estimação e também outras formas de estimação, indicamos Chen e Wei (2005). Todas as rotinas para estimação dos parâmetros dos modelos de regressão quantílica estão implementadas nos principais aplicativos estatísticos, mas a principal referência é o pacote `quantreg` no software `R` (Koenker, 2011).

No pacote `quantreg`, para utilizar o método *simplex* de Barrodale e Roberts (1973) adaptado para modelos de regressão quantílica para estimação dos parâmetros do modelo, deve-se usar o argumento `method="br"`. Para o método de ponto interior, usa-se `method="fn"`, ou ainda `method="pfn"` se o interesse é utilizar o pré-processamento, o qual melhora consideravelmente o desempenho do algoritmo. Esses argumentos são utilizados dentro da função `rq` ou `rq.fit`.

Tendo visto como pode ser feita a estimação dos parâmetros em modelos de regressão quantílica, vamos agora discutir o problema da inferência sobre os parâmetros do modelo.

2.2 Intervalos de confiança para os parâmetros do modelo

Para a construção de intervalos de confiança para os parâmetros de modelos de regressão quantílica, apresentaremos três métodos que podem ser utilizados, baseados em resultados assintóticos, *bootstrap* e testes de escores ordinais.

a) Método baseado em resultados assintóticos

No modelo (2.1) com erros independentes, normalmente distribuídos, com média zero e variância σ^2, verifica-se que

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, (X'X)^{-1}\sigma^2\right),$$

em que $\hat{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados definido em (2.2).

Com esse resultado, é possível construir intervalos de confiança para as componentes do vetor $\boldsymbol{\beta}$ do tipo $\hat{\boldsymbol{\beta}} \pm 1,96\sqrt{(X'X)^{-1}\hat{\sigma}^2}$, com coeficiente de confiança de 95%, em que 1,96 representa o quantil de ordem 97,5% da distribuição normal padrão.

Com o intuito de construir intervalos de confiança para os parâmetros dos modelos de regressão quantílica, utilizaremos resultados assintóticos, ao invés de resultados exatos. Além disso, a matriz de covariâncias assintótica do vetor de estimadores dos parâmetros do modelo pode ser estimada de duas formas diferentes, que serão apresentadas a seguir.

Tendo em vista os modelos de regressão L_1 inicialmente, Bassett e Koenker (1978) obtiveram a distribuição assintótica do vetor de estimadores dos parâmetros do modelo, considerando o estimador da mínima soma dos erros absolutos. No entanto, devemos lembrar que a regressão da mediana é um caso particular da regressão quantílica. Dessa forma, Koenker e Bassett (1978) generalizaram o resultado proposto no artigo anterior para vários quantis, sob as suposições do modelo (2.1). Segue o teorema provado pelos autores para o vetor de estimadores dos parâmetros do modelo de regressão quantílica, notando que o estimador $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau_i})$ é a solução do problema da minimização da soma dos erros absolutos ponderados, definida em (2.3), para $\tau = \tau_i$, dada uma amostra de n observações.

Teorema 1. *Seja* $\left\{\hat{\boldsymbol{\beta}}(\boldsymbol{\tau_1}), \hat{\boldsymbol{\beta}}(\boldsymbol{\tau_2}), \ldots, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau_m})\right\}$, *com* $0 < \tau_1 < \tau_2 < \cdots < \tau_m < 1$, *uma sequência de estimadores para os parâmetros do modelo (2.1). Seja* $\xi_i(\tau_i) = F^{-1}(\tau_i)$ *o quantil de ordem* τ_i *e assuma que*

(i) F *é contínua e tem densidade* f *contínua e positiva em* ξ_i, *para* $i = 1, 2, \ldots, m$.

(ii) A matriz X *de planejamento tem uma coluna de uns.*

(iii) $\lim_{n\to\infty} n^{-1}X'X = Q$, *matriz positiva definida.*

Nessas condições, $\sqrt{n}(\hat{\boldsymbol{\beta}}(\boldsymbol{\tau_1}) - \boldsymbol{\beta}(\boldsymbol{\tau_1}), \ldots, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau_m}) - \boldsymbol{\beta}(\boldsymbol{\tau_m})) \xrightarrow{D} N_{m\times p}(\mathbf{0}, V(\tau_1, \ldots, \tau_m))$, *em que a matriz de covariâncias,* $V(\tau_1, \ldots, \tau_m)$, *pode ser definida como*

$$\Omega(\tau_1, \ldots, \tau_m; F) \otimes Q^{-1}$$

sendo $\Omega(\tau_1, \ldots, \tau_m; F)$ *a matriz de covariâncias entre* m *quantis amostrais de amostras aleatórias com distribuição* F *e* $\otimes$ *indica o produto de Kronecker.*

Para simplificar o resultado anterior, podemos considerar o caso particular em que estamos interessados somente em um quantil específico, digamos τ. Nesse caso, segundo

Kocherginsky et al. (2005), a matriz de covariâncias assintótica de $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ para a situação em que os erros não são identicamente distribuídos é dada por

$$V(\tau) = \tau(1-\tau)(X'FX)^{-1}(X'X)(X'FX)^{-1}, \tag{2.4}$$

em que $F = \text{diag}\,(f_1(0), ..., f_n(0))$, matriz diagonal e f_j, $j = 1, \ldots, n$, é a função densidade dos erros. É importante notar que se $f_1(x) = \cdots = f_n(x) = f(x)$, ou seja, se os erros são identicamente distribuídos, então (2.4) se reduz a

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

Tendo em vista (2.4) e (2.5), foram propostos estimadores para $V(\tau)$. Para o caso (2.5), segundo Kocherginsky et al. (2005), uma estimativa de $1/f(0)$ pode ser obtida usando uma diferença de quantis empíricos dos resíduos, com

$$\frac{\hat{F}^{-1}(\tau + h_n) - \hat{F}^{-1}(\tau - h_n)}{2h_n} \tag{2.6}$$

em que $\lim_{n\to\infty} h_n = 0$. No pacote `quantreg`, para inferência sobre os parâmetros do modelo segundo esse procedimento deve-se usar o comando `se="iid"` na função `summary.rq`. Essa função fornece os valores das estimativas dos parâmetros do modelo, assim como seus erros padrão e significância de cada estimativa. O método padrão para o cálculo de h_n nesse caso é baseado no resultado de Hall e Sheather (1988), existindo outras possibilidades para esse cálculo (Koenker, 2005).

Para a estimação de $V(\tau)$ em (2.4), uma possibilidade é substituir o valor de $f_i(0)$ na matriz $X'FX$ por uma estimativa assintoticamente não viciada. Uma maneira implementada no pacote `quantreg` é substituir $f_i(0)$ por

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

Para utilizar esse método de inferência, basta tomar o comando `se="nid"` na função `summary.rq`. Para mais detalhes sobre esses resultados assintóticos, indicamos o artigo de Koenker e Machado (1999).

Com a estimação da matriz $V(\tau)$, é possível construir os intervalos de confiança para cada termo do vetor de parâmetros $\boldsymbol{\beta}(\boldsymbol{\tau})$ utilizando os resultados do Teorema 1.

b) Método *Bootstrap*

Outro método bastante utilizado para inferir sobre os parâmetros do modelo é a reamostragem. Efron e Tibshirani (1993) discutem como o método pode ser utilizado em modelos de regressão, na estimação da matriz de covariâncias do vetor de estimadores dos parâmetros do modelo. Uma das formas de utilizar o *bootstrap*, sugerida por Koenker (2005), com essa finalidade é selecionar os pares de observações $(Y_i, \boldsymbol{x_i})$ com probabilidade $1/n$, em que n

é o tamanho da amostra, de forma a construir um novo vetor $\boldsymbol{Y}^*$ com valores da variável resposta e uma nova matriz de planejamento X^*. Esse procedimento é repetido, digamos, R vezes, e em cada repetição o vetor $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\tau})$ é calculado. Com essas R estimativas para o vetor de parâmetros do modelo de interesse, estimamos o erro padrão de $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ a partir do erro padrão observado nas reamostras.

O problema desse método é a necessidade de se ajustar o modelo de regressão quantílica para cada reamostra gerada, e em casos em que tanto o número de observações quanto o número de variáveis explicativas do modelo são grandes, o método pode se tornar bastante demorado. As sugestões de tamanho de reamostragens a serem realizadas nesse processo, segundo Efron e Tibshirani (1993), variam de acordo com o uso da técnica do *bootstrap*, normalmente sendo utilizados valores como 50, 200 ou 1000 reamostras. No pacote `quantreg` do aplicativo estatístico `R`, para se utilizar esse método para inferir sobre o vetor de parâmetros do modelo, na função `summary.rq` deve-se usar o comando `se="boot"`. Para indicar o número de reamostras que devem ser utilizadas, por exemplo, se o interesse é utilizar 50 reamostras, deve-se tomar o argumento `R=50`. Dessa forma, um intervalo de confiança para $\beta_i(\tau)$, com coeficiente de confiança $\gamma = 1 - \alpha$, é

$$\hat{\beta}_i(\tau) \pm z_{\alpha/2}\widehat{\text{E.P.}(\hat{\beta}_i(\tau))}$$

em que $z_{\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição normal padrão e $\widehat{\text{E.P.}(\hat{\beta}_i(\tau))}$ é o estimador do erro padrão do estimador do parâmetro $\beta_i(\tau)$ obtido através do procedimento *bootstrap*.

Ainda com relação à inferência sobre os parâmetros do modelo de regressão quantílica, uma vez que a utilização do *bootstrap* pode ser computacionalmente exigente, He e Hu (2002) desenvolveram um novo método denominado *Markov Chain Marginal Bootstrap (MCMB)*, que foi adaptado para a regressão quantílica por Kocherginsky et al. (2005).

O algoritmo básico da adaptação do MCMB para regressão quantílica pode ser descrito da seguinte maneira. Para manter conformidade com a notação utilizada em Kocherginsky et al. (2005), vamos definir $x_{i,j}$ como o j-ésimo componente de $\boldsymbol{x_i}$, $\boldsymbol{x_{i,(j-)}}$ e $\boldsymbol{x_{i,(j+)}}$ como os vetores contendo os primeiros $j-1$ e os últimos $p-j$ componentes de $\boldsymbol{x_i}$, respectivamente, em que $\boldsymbol{x_i}$ identifica a i-ésima linha da matriz de planejamento X. Com isso, considerando o modelo (2.1) podemos escrever $\boldsymbol{x_i'\beta} = x_{i,j}\beta_j + \boldsymbol{x'_{i,(j-)}\beta_{(j-)}} + \boldsymbol{x'_{i,(j+)}\beta_{(j+)}}$, para qualquer $1 \leqslant j \leqslant p$.

Seja a derivada da função de perda $\rho_\tau(u)$

$$\psi_\tau(r) = \tau - I(r < 0). \tag{2.7}$$

Defina os resíduos como $r_i = y_i - \boldsymbol{x_i'\hat{\beta}(\tau)}$ e $z_i = \psi_\tau(r_i)x_i - \bar{z}$, em que $\bar{z} = n^{-1}\sum_{i=1}^{n}\psi_\tau(r_i)x_i$. O algoritmo iterativo têm início com as estimativas dos parâmetros do modelo de regressão quantílica $\boldsymbol{\beta^{(0)}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ no passo 0 e a atualização dos valores é feita de acordo com os passos seguintes.

1. k <- k + 1.

2. Para cada valor $j \in \{1, 2, \ldots, p\}$ de forma crescente, tome amostras com reposição de $z = \{z_1, \ldots, z_n\}$ para obter $\{z_1^{k,j}, \ldots, z_n^{k,j}\}$, e então encontre $\beta_j^{(k)}$ como raiz da equação

$$\sum_{i=1}^{n} \psi_\tau \left(y_i - \boldsymbol{x}'_{\boldsymbol{i},(\boldsymbol{j}-)} \boldsymbol{\beta}^{(\boldsymbol{k})}_{(\boldsymbol{j}-)} - x_{i,j}\beta_j^{(k)} - \boldsymbol{x}'_{\boldsymbol{i},(\boldsymbol{j}+)} \boldsymbol{\beta}^{(\boldsymbol{k-1})}_{(\boldsymbol{j}+)} \right) = \sum_{i=1}^{n} z_i^{k,j}. \tag{2.8}$$

3. Repita os passos 1 e 2 até que se complete uma pré-determinada quantidade de replicações K.

O passo 2 é necessário para extrair uma amostra independente $\{z_1^{k,j}, \ldots, z_n^{k,j}\}$ para cada j. Na equação em (2.8), estamos calculando $\beta_j^{(k)}$ usando os valores mais recentes das estimativas dos outros parâmetros. Como resultado dessa construção, obtemos a sequência $\boldsymbol{\beta^{(1)}}, \ldots, \boldsymbol{\beta^{(K)}}$, que é uma cadeia de Markov. Um resultado importante demonstrado por He e Hu (2002) é que a matriz de covariâncias amostral de $\boldsymbol{\beta^{(k)}} (k = 1, \ldots, K)$ se aproxima consistemente de $V(\tau)$ para grandes valores de n e K.

Esse novo procedimento também está implementado no pacote `quantreg`. Para utilizá-lo, basta tomar o argumento `bsmethod="mcmb"` na função `summary.rq`, além do comando `se ="boot"`.

c) Método de testes de escores ordinais

Por último, ainda com relação à inferência sobre o vetor de parâmetros dos modelos de regressão quantílica, existe na literatura um terceiro método denonimado de teste de escores ordinais, proposto inicialmente por Gutenbrunner e Jureckova (1992).

Kocherginsky et al. (2005) discutem alguns detalhes sobre esse método. Por exemplo, o método apresenta dificuldades computacionais quando utilizado em banco de dados muito grandes. Além disso, esse método não estima a matriz de variância e covariância dos estimadores dos parâmetros, uma vez que a inferência é feita a partir de intervalos de confiança construídos a partir de algoritmos de programação linear.

Koenker (2005) faz uma apresentação do método de forma mais completa fornecendo tanto a motivação para o uso bem como toda a teoria que envolve esse resultado e os passos para a construção dos intervalos de confiança para os parâmetros. De forma simplificada, essa metodologia utiliza as estatísticas de ordem condicionais para calcular a função escore e a respectiva estatística do teste de escore. Em seguida, é utilizado um algoritmo de programação linear para verificar para quais valores do vetor de estimativas dos parâmetros do modelo a hipótese nula não é rejeitada. Dessa forma, ao final do processo iterativo, é construído um intervalo de confiança para os parâmetros de interesse. O autor adianta que, ao contrário dos métodos baseados em resultados assintóticos ou *bootstrap*, o intervalo de confiança não é necessariamente simétrico em torno da estimativa dos parâmetros.

O método também está implementado no pacote estatístico `quantreg` e para sua execução o argumento `se = "rank"` deve ser fornecido dentro da função `summary.rq`.

Na Seção 2.4, elaboramos e apresentamos um pequeno estudo de simulação com o intuito de comparar a eficiência desses métodos para diferentes tamanhos de amostra e diferentes valores de τ, em modelos com erros tanto simétricos como assimétricos.

2.3 Teste da Hipótese Linear Geral

Existem algumas alternativas na literatura para testar hipóteses lineares gerais da forma

$$H_0 : C\boldsymbol{\beta}(\boldsymbol{\tau}) = \boldsymbol{c}, \tag{2.9}$$

em que C é uma matriz de constantes conhecidas, de posto completo e $\boldsymbol{c}$ é um vetor de constantes conhecidas, no modelo (2.1), quando os parâmetros são estimados pelo método da regressão quantílica. Por exemplo, se um modelo como em (1.6) fosse ajustado, então uma hipótese de interesse seria verificar se todos os $\beta_i(\tau)$, $i = 1, \ldots, p$ são iguais a zero, contra a hipótese alternativa de que pelo menos um deles seja diferente de zero, como é feito no teste da tabela de análise de variância para modelos de regressão clássica.

Koenker (2005) formula testes do tipo Wald, que podem ser utilizados para verificar a hipótese (2.9), como também hipóteses envolvendo diversos quantis e diversos parâmetros de forma simultânea. Considerando um problema em que são estimados m diferentes modelos da forma (2.1), então a hipótese linear geral sobre o vetor $\boldsymbol{\zeta} = (\boldsymbol{\beta(\tau_1)'}, \ldots, \boldsymbol{\beta(\tau_m)'})'$, em que $\boldsymbol{\beta(\tau_j)}$ é o vetor com p parâmetros para $\tau = \tau_j$ $(j = 1, \ldots, m)$, pode ser escrita da seguinte forma

$$H_0 : C\boldsymbol{\zeta} = \boldsymbol{c},$$

em que C é uma matriz de constantes conhecidas, $q \times mp$, de posto completo q e $\boldsymbol{c}$ é um vetor de constantes conhecidas, $q \times 1$.

Nestas condições, a estatística de teste é

$$T_n = n(C\hat{\boldsymbol{\zeta}} - \boldsymbol{c})'[CV_n^{-1}C']^{-1}(C\hat{\boldsymbol{\zeta}} - \boldsymbol{c}),$$

em que V_n é a matriz $mp \times mp$,

$$V_n(\tau_1, \ldots, \tau_m) = \begin{bmatrix} V_n(\tau_1, \tau_1) & V_n(\tau_1, \tau_2) & \cdots & V_n(\tau_1, \tau_m) \\ V_n(\tau_2, \tau_1) & V_n(\tau_2, \tau_2) & \cdots & V_n(\tau_2, \tau_m) \\ \vdots & \vdots & \ddots & \vdots \\ V_n(\tau_m, \tau_1) & V_n(\tau_m, \tau_2) & \cdots & V_n(\tau_m, \tau_m) \end{bmatrix}$$

e cada matriz $V_n(\tau_i, \tau_j)$, $p \times p$, é dada por

$$V_n(\tau_i, \tau_j) = [\tau_i \wedge \tau_j - \tau_i\tau_j]H_n(\tau_i)^{-1}J_nH_n(\tau_j)^{-1},$$

com $\tau_i \wedge \tau_j$ representando o mínimo entre τ_i e τ_j, $i \neq j, j = 1, 2, \ldots, m$ e J_n e $H_n(\tau)$ são definidos como

$$J_n = \frac{\sum_{i=1}^{n} \boldsymbol{x_i x_i'}}{n} \quad \text{e}$$

$$H_n(\tau) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \boldsymbol{x_i x_i'} f_i(\xi_i(\tau))}{n}.$$

O termo $f_i(\xi_i(\tau))$ denota a densidade condicional da variável resposta, y_i, avaliada no quantil de ordem τ, $\xi_i(\tau)$.

Uma hipótese de bastante interesse é a de que todos os parâmetros do modelo com $p-1$ variáveis explicativas são iguais a zero, ou seja

$$\beta_2(\tau) = \beta_3(\tau) = \cdots = \beta_p(\tau) = 0, \tag{2.10}$$

para τ definido no modelo (2.1), sendo que β_1 se refere ao intercepto do modelo. A suposição dos erros independentes e identicamente distribuídos simplifica a notação da matriz de covariâncias, que definimos sob essa suposição em (2.5).

Após alguns cálculos, verifica-se que a estatística de teste para a hipótese (2.10) é dada por

$$T_n = n \sum_{i=2}^{p} \frac{\hat{\beta}_i^2(\tau)}{\text{Var}\left(\hat{\beta}_i(\tau)\right)}.$$

Essa estatística pode ser reescrita da seguinte forma

$$T_n = n \frac{f^2(0)}{\tau(1-\tau)} \sum_{i=2}^{p} \frac{\hat{\beta}_i^2(\tau)}{v_{ii}},$$

em que v_{ii} é i-ésimo elemento da diagonal da matriz $(X'X)^{-1}$ e $f(0)$ deve ser substituido por uma estimativa, para que o valor acima possa ser considerado uma estatística. Uma opção é substituir $1/f(0)$ por (2.6) na matriz de covariâncias dos estimadores dos parâmetros.

A estatística T_n tem assintoticamente distribuição χ^2_q sob H_0, em que q é o posto da matriz C. Para a hipótese (2.10), a estatística $T_n \xrightarrow{D} \chi^2_{p-1}$. A implementação desse teste está feita no pacote `quantreg` e pode ser obtida utilizando a função `anova.rq`, com o argumento `test="Wald"`.

Além dessa possibilidade, Chen et al. (2008) desenvolveram um método que, segundo definição dos próprios autores, pode ser visto como uma análise de variância para modelos de regressão L_1. A estatística de teste que os autores propuseram, inicialmente, é a seguinte

$$M_n = \sum_{i=1}^{n} |y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}| - \sum_{i=1}^{n} |y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|$$

em que $\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ é o estimador de $\boldsymbol{\beta}$ no modelo reduzido sob H_0 e $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ é o estimador de $\boldsymbol{\beta}$ no

modelo completo. Verifica-se que essa estatística de teste coincide com a estatística de teste da razão de verossimilhança, para a hipótese em (2.10), quando os erros têm distribuição de Laplace. Os autores ainda mostram que

$$M_n \xrightarrow{D} \frac{\chi^2_q}{4f(0)},$$

em que q é o número de linhas da matriz C e $f(.)$ é a função densidade dos erros. Porém, para evitar a estimação desse valor da função densidade, os autores propuseram a seguinte transformação da estatística M_n,

$$M_n^* = \min_{\beta\in\Omega_0} \sum_{i=1}^{n} w_i|y_i - \boldsymbol{x_i'\beta}| - \min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w_i|y_i - \boldsymbol{x_i'\beta}| - (\sum_{i=1}^{n} w_i|y_i - \boldsymbol{x_i'\hat{\beta}_r}| - \sum_{i=1}^{n} w_i|y_i - \boldsymbol{x_i'\hat{\beta}_c}|),$$

em que $w_1, \ldots, w_n$ é uma sequência de variáveis aleatórias não negativas independentes e identicamente distribuídas com média 1 e variância 1 e Ω_0 é o espaço paramétrico gerado pela hipótese nula. Com relação aos pesos utilizados na definição da nova estatística de teste, w_i, a distribuição exponencial de parâmetro 1 pode ser utilizada para gerar os valores, pois tem média e variância 1. Chen et al. (2008) provam que, sob H_0

$$M_n^* \xrightarrow{D} \frac{\chi^2_q}{4f(0)}.$$

Por esse motivo, os autores defendem que, ao invés de estimar a densidade em $f(0)$, a região crítica para a estatística de teste M_n possa ser construída a partir da distribuição empírica de M_n^*. Tendo em vista esses resultados para regressão L_1, Chen et al. (2008) argumentam que os mesmos também podem ser utilizados na regressão quantílica, com a simples troca do desvio absoluto pela função de perda definida em (1.2). Nesse caso, a estatística de teste seria igual a

$$M_n = \min_{\boldsymbol{\beta}\in\Omega_0} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'\beta}) - \min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'\beta}).$$

De forma análoga, devemos reescrever a estatística M_n^* utilizando a função de perda $\rho_\tau(u)$ para a construção da região crítica da estatística de teste M_n nos modelos de regressão quantílica.

Este teste também está implementado no pacote estatístico `quantreg`. Para utilizá-lo basta fornecer o argumento `test="anowar"` dentro da função `anova.rq`. Na escolha do número de reamostras que serão utilizadas para o cálculo do nível descritivo do teste, ou simplesmente valor-p, por exemplo, deve-se tomar `R=5.000`, caso pretenda-se utilizar 5.000 reamostras.

Finalizando, Gutenbrunner et al. (1993) propõem um teste para hipóteses lineares do tipo (2.9) que é baseado em Gutenbrunner e Jureckova (1992) e os escores ordinais de re-

gressão introduzidos nesse artigo. Para mais detalhes sobre o teste, como estatística de teste, resultados assintóticos, ver Gutenbrunner et al. (1993) e Koenker (2005). Para utilização do teste, dentro da função `anova.rq`, deve-se usar o comando `test="rank"`.

Realizamos um estudo de simulação, que é apresentado na Seção 2.5, para comparar o nível de significância e o poder dos três testes discutidos nessa seção para diferentes valores de τ e diferentes formulações de modelos.

2.4 Simulações para comparação dos intervalos de confiança propostos

Com a finalidade de verificar a performance dos métodos de inferência descritos na Seção 2.2, elaboramos um estudo de simulação supondo duas situações para a distribuição dos erros do modelo: um caso com erro apresentando distribuição normal e outro com distribuição Gama. Em ambas as situações foram construídos 1000 intervalos de confiança. Diversos valores para os quantis, τ, foram fixados para avaliar a cobertura dos intervalos de confiança em diferentes partes da distribuição condicional da variável resposta.

Os valores presentes nas Tabelas 2.1 e 2.2 representam a porcentagem de vezes em que o intervalo com 90% de confiança construído com base no estimador $\hat{\beta}_1$ continha o verdadeiro valor de β_1, nas 1000 simulações feitas. Também foram utilizados tamanhos de amostras diferentes, 50, 500 e 5.000, com o intuito de comparar o desempenho dos métodos. Em todos as simulações, o verdadeiro valor de β_1 foi fixado em 1. Com relação às legendas dessas tabelas, seguem alguns esclarecimentos:

- `iid` - utiliza a estimativa da matriz de covariâncias assintótica dos estimadores dos parâmetros do modelo, supondo que os erros são independentes e identicamente distribuídos;
- `nid` - utiliza a estimativa da matriz de covariâncias assintótica dos estimadores dos parâmetros do modelo, supondo que os erros são independentes, porém não identicamente distribuídos;
- `bootXY` - utiliza o método de *bootstrap* para estimar o erro padrão do estimador do parâmetro do modelo, reamostrando os pares (y_i, x_i);
- `bootMCMB` - utiliza o método *Markov Chain Marginal Bootstrap* para estimar o erro padrão do estimador do parâmetro do modelo;
- `rankIID` - utiliza o método dos testes de escores ordinais, com suposição dos erros independentes e identicamente distribuídos;
- `rankNID` - utiliza o método dos testes de escores ordinais, com suposição dos erros independentes, porém não identicamente distribuídos.

Modelo 1: Erro simétrico

Como primeiro modelo de análise, vamos considerar um modelo de regressão linear com apenas uma variável explicativa:

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \varepsilon_i.$$

Nesse primeiro modelo, utilizamos tanto x_i quanto ε_i com distribuição normal padrão. Somente ε_i e y_i são variáveis aleatórias, enquanto que o uso de x_i com distribuição normal foi devido à necessidade de gerar valores para essa variável. Inicialmente, o interesse é verificar se os métodos apresentados para a inferência sobre os parâmetros do modelo são adequados para diferentes quantis em um modelo bastante simples, considerando também diferentes tamanhos de amostra. É possível notar alguns resultados interessantes sobre esse modelo que podem ser observados na Tabela 2.1 e que apontamos a seguir:

	Métodos inferenciais					
	`iid`	`nid`	`bootXY`	`bootMCMB`	`rankIID`	`rankNID`
τ **= 0,10**						
n = 50	0,767	0,847	0,921	0,911	0,874	0,872
n = 500	0,877	0,900	0,896	0,882	0,892	0,885
n = 5000	0,895	0,900	0,883	0,893	0,903	0,886
τ **= 0,25**						
n = 50	0,852	0,899	0,913	0,919	0,879	0,881
n = 500	0,884	0,891	0,891	0,875	0,902	0,901
n = 5000	0,886	0,895	0,881	0,891	0,913	0,880
τ **= 0,50**						
n = 50	0,872	0,896	0,902	0,904	0,882	0,871
n = 500	0,880	0,899	0,904	0,893	0,876	0,892
n = 5000	0,893	0,877	0,894	0,871	0,904	0,894
τ **= 0,75**						
n = 50	0,844	0,891	0,912	0,903	0,886	0,876
n = 500	0,899	0,904	0,907	0,893	0,899	0,898
n = 5000	0,889	0,910	0,897	0,886	0,884	0,889
τ **= 0,90**						
n = 50	0,784	0,843	0,889	0,948	0,857	0,862
n = 500	0,862	0,903	0,895	0,899	0,884	0,880
n = 5000	0,878	0,891	0,886	0,889	0,885	0,898

Tabela 2.1: *Probabilidade de cobertura dos intervalos de confiança para o Modelo 1*

(i) para pequenas amostras, de tamanho 50 nesse caso, os desempenhos de todos os métodos não são tão bons com exceção dos métodos que utilizam *bootstrap*. É importante notar que nas caudas da distribuição condicional, no caso $\tau = 0,1$ e $\tau = 0,9$, o método `iid` apresentou resultado bastante insatisfatório, com porcentagem de inclusão bem diferente do coeficiente de confiança fixado;

(ii) para grandes amostras, considerando 5000 observações nas simulações feitas, não percebemos uma diferença de performance evidente entre os diferentes métodos;

(iii) o método `iid` apresentou resultados mais consistentes principalmente quando o tamanho da amostra foi de 5000. Com o tamanho 500 e novamente nas caudas da distribuição condicional, os resultados ainda não são tão bons;

(iv) sobre os métodos de reamostragem, embora não estejamos apresentando o tempo de execução de cada procedimento, verificamos uma diferença considerável entre os dois métodos, no sentido de que, para amostras grandes, `bootMCMB` foi muito mais rápido que o `bootXY`. Com relação ao desempenho na construção dos intervalos de confiança, talvez o método `bootXY` tenha apresentado uma ligeira vantagem sobre o outro, porém, em geral, ambos métodos apresentaram resultados muito satisfatórios;

(v) os métodos baseado nos testes de escores ordinais mostraram uma boa performance na construção do intervalo de confiança longe das caudas da distribuição condicional da variável resposta. Para valores de τ iguais a 0,1 e 0,9, a cobertura do intervalo de confiança ficou um pouco menor do que o esperado.

Modelo 2: Erro assimétrico

Novamente, vamos considerar um modelo de regressão linear com apenas uma variável explicativa,

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \varepsilon_i,$$

porém com distribuições para X e ε diferentes do Modelo 1. Para esse segundo exemplo, adotamos x_i com distribuição Uniforme(0,10) e ε_i com distribuição Gama com média 1 e variância 1. Nessa simulação, desejamos analisar o desempenho dos métodos de inferência citados quando uma distribuição assimétrica para os erros é utilizada. Sobre os resultados que podem ser consultados na Tabela 2.2, destacamos os seguintes:

(i) novamente para pequenas amostras verificamos que os desempenhos de todos os métodos estão longe do esperado com exceção para regressão da mediana, em que os métodos com reamostragem e `rankNID` apresentaram bons resultados;

(ii) por outro lado, para grandes amostras, todos os métodos apresentaram cobertura dos intervalos de confiança próximos de 0,9, até mesmo para valores de τ que representam a cauda da distribuição condicional da variável resposta;

(iii) o método que não supõe independência e distribuição idêntica para os erros apresentou resultados muito bons, em geral, com problemas apenas para tamanhos de amostras menores com τ igual a 0,1 e 0,9;

	Métodos inferenciais					
	iid	nid	bootXY	bootMCMB	rankIID	rankNID
τ **= 0,10**						
n = 50	0,837	0,823	0,935	0,989	0,877	0,887
n = 500	0,879	0,872	0,894	0,892	0,892	0,902
n = 5000	0,896	0,896	0,893	0,882	0,898	0,887
τ **= 0,25**						
n = 50	0,861	0,888	0,928	0,912	0,884	0,883
n = 500	0,894	0,896	0,89	0,872	0,901	0,877
n = 5000	0,913	0,897	0,903	0,911	0,886	0,888
τ **= 0,50**						
n = 50	0,851	0,919	0,902	0,903	0,859	0,889
n = 500	0,880	0,895	0,897	0,897	0,881	0,884
n = 5000	0,910	0,899	0,910	0,894	0,897	0,901
τ **= 0,75**						
n = 50	0,802	0,945	0,882	0,931	0,877	0,880
n = 500	0,876	0,909	0,915	0,877	0,89	0,905
n = 5000	0,895	0,900	0,900	0,911	0,915	0,896
τ **= 0,90**						
n = 50	0,688	0,900	0,886	0,990	0,888	0,870
n = 500	0,856	0,913	0,909	0,890	0,898	0,884
n = 5000	0,887	0,903	0,891	0,895	0,887	0,903

Tabela 2.2: *Probabilidade de cobertura dos intervalos de confiança para o Modelo 2*

(iv) os métodos de reamostragem apresentaram resultados ótimos nessa formulação em que a distribuição dos erros é assimétrica, com única exceção ocorrendo para amostras de tamanho 50, o que, conforme já foi dito, foi uma dificuldade para todos os métodos.

(v) os testes de escores ordinais também apresentaram resultados interessantes e, assim como os métodos que utilizam reamostragem, tiveram maior problema somente com amostras pequenas;

Com base nas análises feitas a partir das simulações, considerando os dois modelos formulados, pudemos perceber alguns resultados importantes. Em primeiro lugar, os métodos que utilizam o *bootstrap* tiveram uma boa performance de forma geral. No entanto, para problemas com muitas observações a utilização do método `bootXY` pode se tornar bastante demorada. Por isso, sugere-se o uso do método `bootMCMB`, pois apresentou boa probabilidade de cobertura de forma geral e é muito mais rápido que o método anterior. Segundo, para amostras grandes, a escolha do método não parece influenciar no resultado, pois todos apresentaram resultados satisfatórios. Por último, para pequenas amostras, principalmente na inferência sobre os parâmetros de modelos em que o interesse está mais nas caudas da distribuição condicional da variável resposta, deve-se tomar um cuidado maior na escolha do método. De forma geral, os testes de escores ordinais e o método `bootXY` apresentaram resultados mais próximos do esperado.

2.5 Simulações para comparação dos testes propostos

Na Seção 2.3 foram apresentadas três possibilidades diferentes para testar hipóteses lineares gerais do tipo (2.9). É interessante conhecer algumas características desses diferentes testes, como tamanho e poder. Um estudo similar foi realizado por Chen et al. (2008), restrito no entanto a modelos de regressão para a mediana. Serão considerados portanto valores de τ distintos, os mesmos já utilizados nas simulações da seção anterior e um modelo da seguinte forma

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \beta_2(\tau)x_{i2} + u_i,$$

com duas situações diferentes para a distribuição de probabilidade dos erros. No primeiro modelo, assumiremos os erros com distribuição normal padrão, enquanto que no segundo será utilizada a distribuição t-Student com 1 grau de liberdade. Foram feitas essas escolhas para comparar o desempenho dos testes quando deparados com erros que podem assumir grandes valores absolutos com maior probabilidade que no caso da distribuição normal. Dessa forma, queremos verificar se essas observações são capazes de influenciar o desempenho do teste com relação ao seu nível de significância e poder para testar a hipótese

$$H_0 : \beta_2(\tau) = 0.$$

Em ambas formulações, para as variáveis X_1 e X_2 são gerados valores a partir da distribuição normal padrão. O tamanho da amostra nessa simulação foi de 400 observações.

A Tabela 2.3 apresenta a porcentagem de vezes que os testes rejeitaram a hipótese nula ao nível de 5%, tomando $\beta_2(\tau) = 0$. Com isso, estamos interessados em estimar o tamanho de cada teste. A Tabela 2.5 apresenta a porcentagem de vezes que os testes rejeitaram a hipótese nula também ao nível de 5%, quando $\beta_2(\tau) = 0,1$. Dessa forma, estaremos estimando o poder de cada teste nessas condições determinadas. Foram executas 1000 simulações em ambos os casos. Com relação às legendas de cada tabela, seguem alguns esclarecimentos:

- `Wald` - teste de Wald para modelos de regressão quantílica;
- `anowar` - teste que calcula a diferença da soma absoluta ponderada de resíduos entre o modelo reduzido e o modelo completo;
- `rank` - teste que utiliza os escores ordinais.

Como podemos verificar nos resultados apresentados na Tabela 2.3, no primeiro modelo, na formulação de testes para as estimativas dos parâmetros nas caudas da distribuição, a proporção de erros do tipo I ficaram um pouco acima do nível de significância fixado. Esse fato ocorreu com todos os métodos, com menor intensidade no método que utilizava os desvios absolutos ponderados. Por outro lado, quando considerados os erros com distribuição t-Student, a taxa de erros do tipo I dos métodos `Wald` e `anowar`, em todos os quantis

		τ				
Modelo	Método	0,10	0,25	0,50	0,75	0,90
1	`Wald`	0,062	0,063	0,049	0,043	0,065
	`anowar`	0,055	0,047	0,048	0,043	0,060
	`rank`	0,054	0,052	0,049	0,048	0,061
2	`Wald`	0,044	0,048	0,039	0,041	0,062
	`anowar`	0,037	0,044	0,039	0,045	0,043
	`rank`	0,040	0,046	0,045	0,050	0,053

Tabela 2.3: *Estimativas dos erros do tipo I para os testes propostos nos modelos formulados, para amostra de tamanho igual a 400.*

estudados, apresentaram valores muito diferentes do valor fixado de 0,05. O método `rank`, que utiza os escores ordinais, teve desempenho bom, com tamanho sempre próximo de 0,05, com exceção para $\tau = 0,10$, em que a porcentagem de erros do tipo I foi igual a 0,04.

Tendo em vista os resultados insatisfatórios dos métodos `Wald` e `anowar`, quando considerada a distribuição de t-Student para os erros, decidimos aumentar o tamanho da amostra para 4.000 e verificar se havia alguma alteração. Os resultados obtidos estão dispostos na Tabela 2.4. Ambos métodos apresentaram melhor desempenho com o aumento no número de observações. Tal comportamento pode ser uma evidência contra o uso destes testes para amostras de tamanho reduzido.

	τ				
Método	0,10	0,25	0,50	0,75	0,90
`Wald`	0,046	0,054	0,054	0,052	0,049
`anowar`	0,048	0,052	0,055	0,046	0,043

Tabela 2.4: *Estimativas dos erros do tipo I para os testes propostos com tamanho de amostra igual a 4.000, somente no modelo com erros com distribuição de t-Student.*

Com relação ao poder do teste, conforme podemos verificar na Tabela 2.5, para o primeiro modelo, o teste que considera os desvios absolutos ponderados, `anowar`, tem poder um pouco menor que os outros dois. Entre os demais, o teste de Wald foi ligeiramente superior ao teste de escores ordinais. Para o segundo modelo simulado, novamente o teste menos poderoso é o teste que utiliza o método `anowar` e, em geral, não é possível dizer qual é o mais poderoso, havendo uma equivalência entre os testes `Wald` e `rank`.

Finalizando esse capítulo, discutiremos a seguir propriedades interessantes dos modelos de regressão quantílica que se referem à robustez e à propriedade de equivariância.

		τ				
Modelo	Método	0,10	0,25	0,50	0,75	0,90
1	`Wald`	0,342	0,341	0,38	0,382	0,334
	`rank`	0,343	0,336	0,356	0,354	0,326
	`anowar`	0,317	0,322	0,329	0,334	0,313
2	`Wald`	0,287	0,311	0,289	0,327	0,295
	`rank`	0,295	0,304	0,301	0,315	0,295
	`anowar`	0,270	0,272	0,263	0,293	0,275

Tabela 2.5: *Poder dos testes propostos nos dois modelos formulados, quando* $\beta_2(\tau) = 0,1$.

2.6 Robustez e equivariância em modelos de regressão quantílica

Conforme foi discutido no Capítulo 1, a análise de regressão da média apresenta ótimas propriedades quando a distribuição dos erros é normal. No entanto, quando isso não é verificado, uma possibilidade é recorrer a transformações da variável resposta. Um exemplo bastante conhecido é a função logaritmo, que é frequentemente utilizada quando se está diante de uma distribuição assimétrica à direita, como renda. Porém, a variável resposta transformada deve ter distribuição normal para que o modelo possa usufruir das propriedades ótimas. Entretanto, há um aspecto dessa transformação que muitas vezes não é discutido após o ajuste do modelo e que pode ser enunciado da seguinte maneira.

Suponhamos que a variável de interesse Y não possui distribuição normal e é transformada na variável $W = \log Y$. Agora, com uma distribuição mais próxima da desejada, ajusta-se o seguinte modelo com as suposições usuais,

$$W_i = \alpha + \beta x_i + \varepsilon_i.$$

Dessa forma, podemos dizer que

$$E(W|x) = \alpha + \beta x.$$

No entanto, neste caso,

$$E(W|x) = E(\log Y|x) \neq \log E(Y|x).$$

Portanto, não se pode exponencializar o resultado obtido para $E(W|x)$ para obter o valor esperado da variável aleatória Y, que era a variável de interesse inicial.

Por outro lado, quantis usufruem de uma propriedade importante que pode ser denominada de equivariância a transformações monótonas. Seja $h(.)$ uma função não decrescente no conjunto $\mathbb{R}$. Então, para qualquer variável aleatória Y,

$$Q_\tau(h(Y)) = h(Q_\tau(Y)), \tag{2.11}$$

em que $Q_\tau(Y)$ representa o quantil de ordem τ da variável Y. O resultado (2.11) pode ser obtido a partir do fato elementar que, para qualquer função h monótona,

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

Com relação ao problema da presença de *outliers* na variável resposta, que se refere à grande influência que esses pontos têm nas estimativas do método de mínimos quadrados, os modelos de regressão quantílica se apresentam como uma alternativa robusta para esse problema.

Koenker e Bassett (1978) provam que os modelos de regressão quantílica apresentam uma importante propriedade de robustez, que pode ser explicada por um exemplo bastante simples. Considere uma nuvem de pontos, com um plano de regressão estimado para o quantil condicional de ordem τ, passando por entre esses pontos. Agora, selecionamos um ponto qualquer e adicionamos, ou subtraímos, unidades na variável resposta, de modo que essa transformação em seu valor não faça com que esse ponto ultrapasse o plano. Ocorre que o plano obtido não se altera mesmo após a transformação, independentemente da distância que esse ponto seja levado, sempre lembrando da condição que o ponto não ultrapasse o plano, ou seja, que o sinal do resíduo desse ponto se mantenha.

Um exemplo dessa robustez pode ser vista na Figura 2.1, em que foram ajustadas duas retas de quantis condicionais, para τ igual a 0,25 e 0,75, em dados simulados. Em seguida, para valores que ficaram acima da reta do quantil de ordem 0,75 adicionou-se 30 unidades arbitrárias na variável Y, enquanto que para aquelas que ficaram abaixo da reta do quantil 0,25, foram subtraídas 30 unidades. O resultado que se pode observar é que não há diferença nos ajustes antes e depois desse movimento dos pontos, tanto para quantil de ordem 0,25 quanto para o quantil de ordem 0,75.

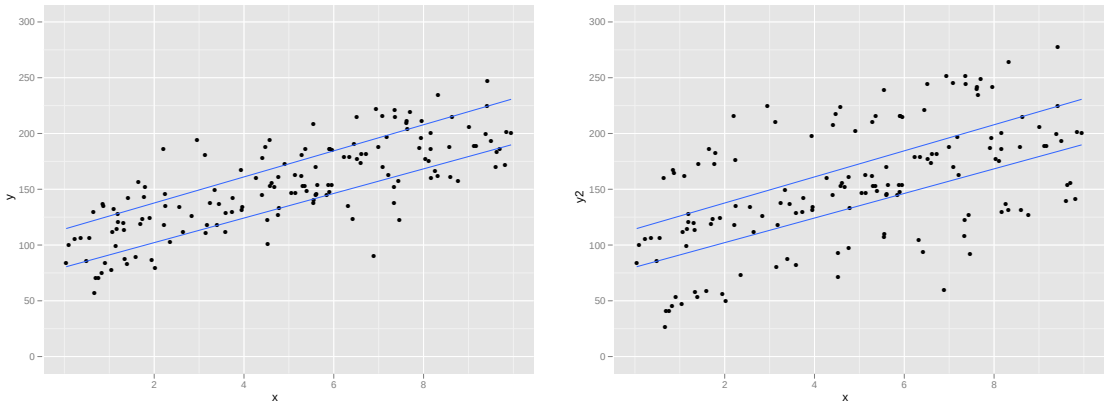

Figura 2.1: *Comparação de ajustes antes e depois de pontos terem sido deslocados no eixo y*

Ainda com relação ao estudo da influência da contaminação de pontos nos modelos de regressão quantílica, indicamos o livro de Koenker (2005) para mais informações sobre pontos

de ruptura (*breakdown point*) nesses modelos. Segundo o autor, esse valor pode atingir até 0,29, o que pode ser considerado razoável.

No presente capítulo, discutimos importantes aspectos dos modelos de regressão quantílica, como estimação dos parâmetros, construção de intervalos de confiança para os parâmetros do modelo e propriedades como robustez e equivariância. No próximo capítulo, estaremos interessados em discutir técnicas de análise de qualidade de ajuste dos modelos de regressão quantílica.

Capítulo 3

Análise da Qualidade do Ajuste do Modelo de Regressão Quantílica

Vimos nos capítulos anteriores tanto uma motivação para a utilização de modelos de regressão quantílica quanto alguns conceitos inferenciais importantes para esses modelos. No presente capítulo, estamos interessados em discutir o ajuste do modelo com relação à sua adequabilidade aos dados de interesse. Para isso, introduziremos algumas medidas-resumo e testes para avaliar a qualidade de ajuste do modelo. Por último, indicaremos algumas análises que podem ser feitas a partir de gráficos de envelope.

Inicialmente, apresentaremos um coeficiente de determinação para modelos de regressão quantílica, partindo do coeficiente de explicação R^2 já bastante utilizado em modelos de regressão clássica. Discutiremos também medidas propostas anteriormente para modelos de regressão L_1.

Em seguida, consideraremos o teste de falta de ajuste proposto por He e Zhu (2003). Primeiramente, devemos enunciar o teste formalmente e depois utilizaremos exemplos para motivar o seu uso.

Além disso, se considerarmos a ligação da distribuição Laplace Assimétrica com os modelos de regressão quantílica, então podemos utilizar técnicas de análise de qualidade de ajuste já bastante utilizadas, principalmente em modelos lineares generalizados. Nesse sentido, vamos discutir a construção e análise de gráficos de envelope para os modelos de regressão quantílica.

É importante ressaltar que implementamos diversas funções no pacote estatístico `R` para a análise dos resultados propostos nesse capítulo. Todas essas funções encontram-se disponíveis no Apêndice A.

3.1 Coeficiente de determinação em modelos de regressão quantílica

Se considerarmos o modelo linear (1.5) com erros com distribuição normal, então uma medida bastante utilizada na análise desses tipos de modelos é o coeficiente de determinação do modelo, também conhecido como R^2. Essa estatística pode ser calculada da seguinte forma:

$$R^2 = \frac{\text{SQT} - \text{SQE}}{\text{SQT}}$$

em que $\text{SQT} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ e $\text{SQE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ são denominados de soma de quadrados totais e soma de quadrados dos resíduos, respectivamente.

Esta estatística pode ser interpretada como o percentual da variabilidade da variável resposta explicada pelas variáveis explicativas, por isso é utilizada muitas vezes como uma medida de qualidade de ajuste. No entanto, como essa última colocação é bastante discutível (ver Kvalseth (1985)), vamos então considerar a estatística R^2, e outras que serão obtidas de forma semelhante na sequência do texto, somente como medidas-resumo do modelo ajustado. Para analisar a qualidade do ajuste, vamos nos referir ao teste proposto na Seção 3.2.

Inicialmente para os modelos de regressão L_1, McKean e Sievers (1987) e Andre et al. (2000) apresentam uma estatística do tipo R^2, denonimada aqui de R somente, tendo em vista algumas propriedades interessantes que essa medida deveria possuir, propriedades essas que listamos a seguir:

(P1) R deve estar ligada diretamente ao critério de ajuste, uma vez que esta medida pode ser utilizada como medida da qualidade de ajuste de um modelo;

(P2) R deve medir a melhoria no ajuste do modelo com a adição de variáveis preditoras e como tal deve manter uma relação com um teste de hipóteses com o intuito de verificar se o efeito das variáveis adicionadas é nulo;

(P3) R deve ser adimensional, e invariante sobre variações de escala e localização das variáveis resposta e preditoras;

(P4) $0 \leqslant R \leqslant 1$, 1 significando um ajuste perfeito do modelo e 0 a total falta de ajuste;

(P5) R deve aumentar com a inclusão de parâmetros adicionais;

(P6) R deve ser robusto.

McKean e Sievers (1987) apresentam a seguinte estatística como coeficiente de determinação. Essa estatística, denominada pelos autores de R_2, satisfaz às propriedades listadas anteriormente e é definida como

$$R_2 = \frac{\text{RAD}}{\text{RAD} + (n - p - 1)(\hat{\sigma}/2)},$$

em que RAD é a redução da soma absoluta dos resíduos do modelo reduzido

$$Y = \alpha + \epsilon.$$

para o modelo em questão e p é a quantidade de variáveis preditoras.

Assim, se considerarmos a soma dos erros absolutos do modelo com p variáveis preditoras como

$$\text{SAE}(\hat{\beta}) = \sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad \hat{y}_i = \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$$

e a soma dos erros absolutos do modelo reduzido, somente com o intercepto, $\hat{y}_i = \hat{\alpha}$, como

$$\text{SAE}(\hat{\alpha}) = \sum_{i=1}^{n} |y_i - \hat{\alpha}|,$$

então, RAD = SAE($\hat{\alpha}$) − SAE($\hat{\beta}$). Conforme já discutido no primeiro capítulo, $\hat{\alpha}$ é a mediana da variável resposta Y. Ainda sobre a estatística R_2 proposta, $\hat{\sigma}$ é o estimador para o parâmetro de escala σ, que na regressão L_1 é definido como

$$\hat{\sigma} = \frac{1}{2f(0)},$$

em que $f(.)$ é a função densidade dos erros.

Andre et al. (2000) argumentam, apresentando um contra-exemplo, que o coeficiente proposto não satisfaz à propriedade P5, por isso sugerem uma nova estatística. Essa nova sugestão difere da anterior somente pelo estimador $\hat{\sigma}$. O novo estimador de σ deve ser calculado como sendo a média da soma dos erros absolutos do modelo, isto é,

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|.$$

Por outro lado, para os modelos de regressão quantílica, uma primeira tentativa de definir um coeficiente de determinação foi feita por Koenker e Machado (1999). Entretanto, a discussão das propriedades dessa estatística feita pelos autores é pouco aprofundada. Dessa forma, discutiremos alguns aspectos dessa medida-resumo para modelos de regressão quantílica de uma forma um pouco diferente.

Utilizando a notação de modelos encaixados, podemos definir a estatística de uma forma mais geral. Considere um modelo linear para o quantil condicional, com p variáveis explicativas,

$$Q_\tau(Y_i|x) = \boldsymbol{x'_{i1}}\boldsymbol{\beta_1(\tau)} + \boldsymbol{x'_{i2}}\boldsymbol{\beta_2(\tau)}, \tag{3.1}$$

em que $\boldsymbol{x_i}$, i-ésima linha da matriz X de planejamento, é particionada em duas partes denominadas $\boldsymbol{x_{i1}}$ e $\boldsymbol{x_{i2}}$ de dimensões $p-q$ e q, respectivamente. Dessa forma, um particionamento semelhante deve ser considerado para o vetor de parâmetros $\boldsymbol{\beta(\tau)}$.

Seja $\hat{\boldsymbol{\beta}}\boldsymbol{(\tau)}$ o estimador que minimiza a soma $\sum \rho_\tau(y_i - \boldsymbol{x_i'\beta})$ para o modelo completo, e $\tilde{\boldsymbol{\beta}}\boldsymbol{(\tau)}$ o estimador para o modelo reduzido,

$$Q_\tau(Y_i|x) = \boldsymbol{x'_{i1}}\boldsymbol{\beta_1(\tau)}, \tag{3.2}$$

que está relacionado com a restrição q-dimensional

$$H_0 : \boldsymbol{\beta_2(\tau)} = \mathbf{0}. \tag{3.3}$$

Considerando agora a soma dos erros absolutos ponderados do modelo completo, inicialmente, da seguinte forma,

$$\hat{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\boldsymbol{(\tau)}),$$

e, em seguida, do modelo reduzido,

$$\tilde{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x'_{i1}}\tilde{\boldsymbol{\beta}}\boldsymbol{(\tau)}),$$

então, o coeficiente de determinação para a regressão quantílica do modelo (3.1) com relação ao modelo reduzido sob a hipótese (3.3) é definido da seguinte forma

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}. \tag{3.4}$$

Se considerarmos no vetor de parâmetros $\boldsymbol{\beta_2(\tau)}$ os coeficientes de regressão associados a todas as variáveis explicativas disponíveis, de forma que o modelo reduzido tenha apenas o intercepto, então $R^1(\tau)$ calculado se assemelha bastante ao coeficiente de explicação R^2 comumente utilizado na análise de regressão clássica.

Como $\tilde{\boldsymbol{\beta}}\boldsymbol{(\tau)}$ é obtido restringindo $\hat{\boldsymbol{\beta}}\boldsymbol{(\tau)}$, verifica-se que $\hat{V}(\tau) \leqslant \tilde{V}(\tau)$ e com isso $R^1(\tau)$ se encontra dentro do intervalo $[0, 1]$, satisfazendo P4. Fato similar ocorre com o coeficiente de explicação R^2.

Por outro lado, diferentemente de R^2, que mede o relativo sucesso de dois modelos para a média condicional em função de termos da variância residual, segundo Koenker e Machado (1999), $R^1(\tau)$ mede o relativo sucesso de correspondentes modelos de regressão quantílica em um específico quantil em função de uma apropriada soma de resíduos absolutos ponderados. Dessa forma, $R^1(\tau)$ constitui uma medida local de qualidade de ajuste do modelo de regressão

quantílica para um particular quantil.

Ainda sobre $R^1(\tau)$, podemos dizer que se o modelo em (3.1) é "melhor" que o modelo gerado a partir da restrição em (3.3), então $\hat{V}(\tau)$ deve ser significativamente menor que $\tilde{V}(\tau)$. Aqui, melhor deve ser entendido no sentido que o ajuste do modelo para o quantil condicional de ordem τ é alterado de forma significativa pela inclusão das covariáveis x_2 no modelo.

Para exemplicar essa estatística, utilizaremos o banco de dados da concentração de Imunoglobulina G, do Capítulo 1. O objetivo nesse exemplo era analisar a variação da concentração de Imunoglobulina G em função da idade. Obtivemos naquele capítulo estimativas das retas de regressão quantílica para alguns quantis condicionais e verificamos que estas apresentavam um comportamento de paralelismo entre elas.

Na Figura 3.1, podemos observar o valor da estatística $R^1(\tau)$ para valores de τ de 0,05 a 0,95, com intervalos de tamanho 0,05. A partir da análise do gráfico, podemos dizer que a contribuição da variável idade é muito parecida em diferentes quantis da variável resposta Imunoglobulina G.

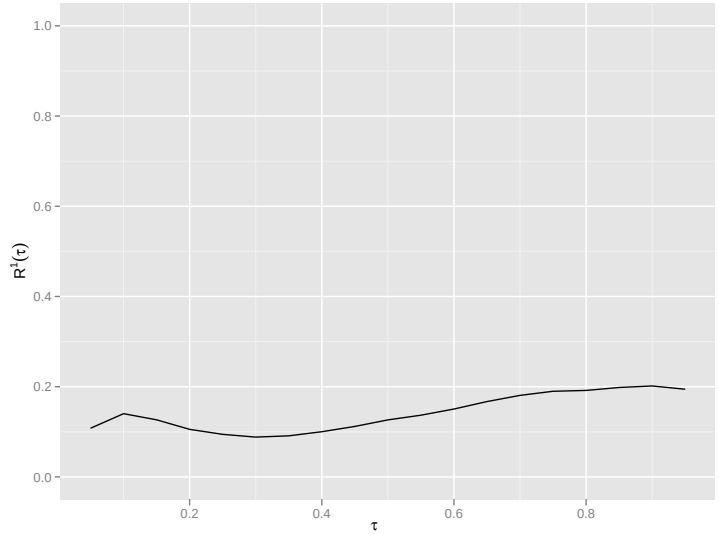

Figura 3.1: *Cálculo de $R^1(\tau)$ para o exemplo da regressão quantílica da imunoglobulina em função da idade, com crianças de 6 meses a 6 anos.*

Para complementar esse exemplo, faremos uso também de outro banco de dados citado no primeiro capítulo, sobre a poluição do ar em algumas cidades americanas. Havíamos estimado os coeficientes da relação linear da média e da mediana condicional da quantidade média de dióxido de enxofre (SO2) em função da temperatura média (Temp) das cidades. Vamos considerar agora também as variáveis número de fábricas que empregam mais de 20 trabalhadores (Fab) e também a população (Pop) de cada cidade.

Calcularemos também a estatística $R^1(\tau)$ nas regressões quantílicas do dióxido de enxofre em função das três variáveis, em três regressões com uma única variável independente e uma regressão com as três.

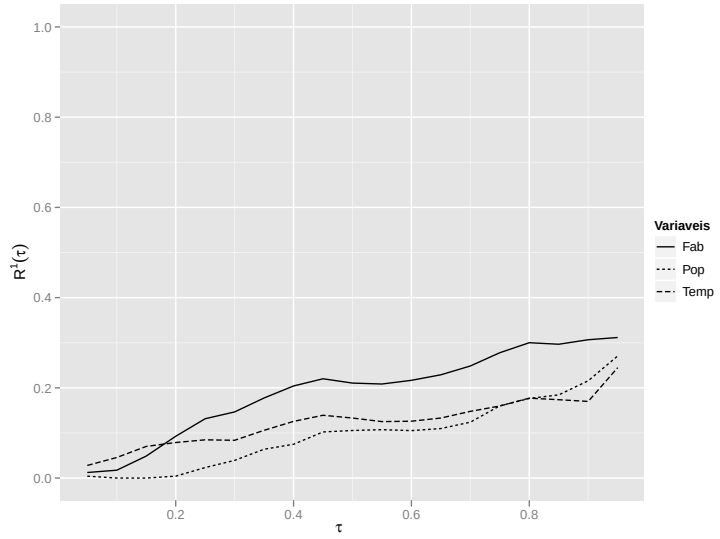

Figura 3.2: *Cálculo de $R^1(\tau)$ para o exemplo da regressão quantílica de S02 em função de temp, fab e pop, em 41 cidades americanas.*

Pode-se notar na Figura 3.2 o crescimento de $R^1(\tau)$ em função dos valores de τ, comportamento bastante semelhante nas três variáveis selecionadas. Tal fato indica que, aparentemente, a contribuição dessas variáveis independentes na distribuição condicional da variável dióxido de enxofre é maior na cauda direita da distribuição. Não estamos considerando aqui se essas contribuições são significativas, nem se o modelo está bem ajustado. Inclusive, essa verificação ajudaria a checar se essa estatística atende a propriedade P2, citada no início desta seção. No entanto, o número de observações desse exemplo é muito pequeno, e conforme já vimos no Capítulo 2, os intervalos de confiança para os parâmetros não são tão eficientes nesses casos. Outra característica importante a se notar nesse gráfico é a maior contribuição da variável que indica o número de fábricas que empregam mais de 20 trabalhadores, em quase todas as distribuições condicionais da variável resposta, com exceção das associadas aos quantis 0,05 e 0,10, de acordo com a estatística $R^1(\tau)$. Outro resultado que pode ser observado a partir dessa estatística é a contribuição conjunta dessa três variáveis, conforme se nota na Figura 3.3. Devido à própria construção da estatística $R^1(\tau)$, se verifica que o modelo com mais variáveis tem maior contribuição na redução da soma absoluta ponderada dos erros do modelo, quando passamos do modelo reduzido para um modelo mais completo.

Com relação às propriedades P1-P6 listadas no início dessa seção, podemos mostrar que a estatística em (3.4) não é robusta à presença de valores aberrantes no modelo. Conforme se verifica na Figura 3.4, a forma de todas as curvas de $R^1(\tau)$ se modificam quando alteramos uma observação na variável resposta de forma a torná-la um *outlier*. O que foi feito nesse caso foi alterar uma observação que tinha um valor igual a 69 para 690. A consequência da alteração desse único valor na amostra é a mudança nos valores e na forma da curva do coeficiente de determinação do modelo construído anteriormente e observado na Figura 3.3.

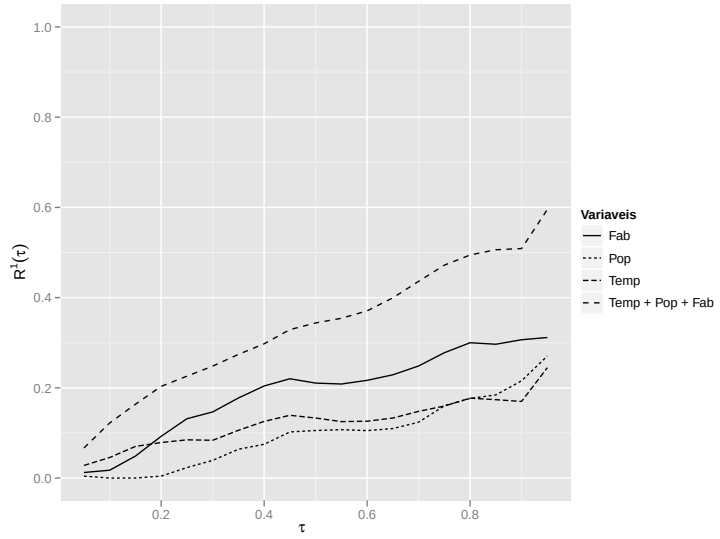

Figura 3.3: *Cálculo de $R^1(\tau)$ para o exemplo da regressão quantílica de SO2 em função da Temp, Man e Pop, separada e conjuntamente, em 41 cidades americanas.*

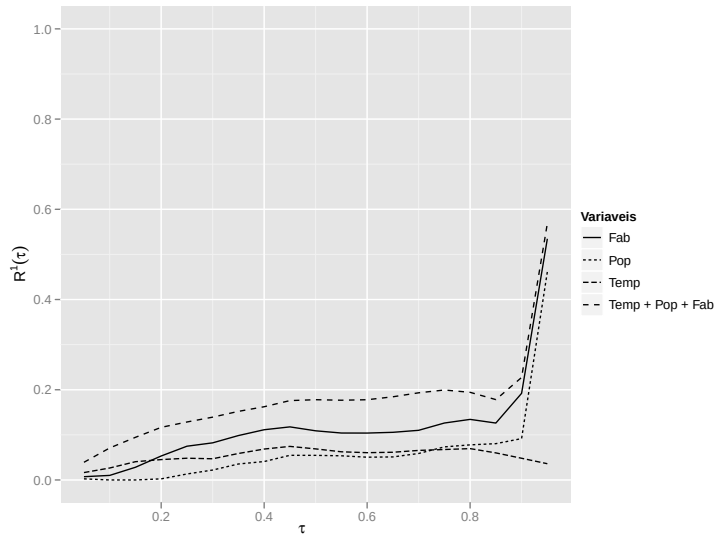

Figura 3.4: *Cálculo de $R^1(\tau)$ para o exemplo da regressão quantílica de SO2 em função da Temp, Man e Pop, separada e conjuntamente, em 41 cidades americanas, com uma observação aberrante.*

3.2 Teste da falta de ajuste em modelos de regressão quantílica

Para uma introdução sobre testes de falta de ajuste, tendo em vista inicialmente a análise de modelos de regressão clássica, indicamos o livro de Draper e Smith (1981). De forma resumida, a estatística apresentada pelos autores testa a hipótese de linearidade do modelo, $H_0 : E(Y|x) = \beta_0 + \beta_1 x$ contra a hipótese de não linearidade do modelo. Em outras palavras, a hipótese de interesse pode ser enunciada como $H_0 : E(\varepsilon_1) = E(\varepsilon_2) = \cdots = E(\varepsilon_n) = 0$, no modelo $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon}$.

Para modelos de regressão quantílica, He e Zhu (2003) apresentam também uma estatís-

tica com o intuito de testar a hipótese de linearidade do modelo. A diferença na formulação desses autores, com relação ao teste em Draper e Smith (1981), se refere ao método utilizado para construir o teste. Para chegar a essa estatística de teste para verificar a falta de ajuste em modelos de regressão quantílica, os autores se basearam no trabalho de Stute (1997), que propôs decompor os resíduos em componentes principais e a partir dessa análise verificar se há afastamento do modelo ajustado com relação a um modelo hipótetico. Essa sugestão de decomposição utiliza um processo de somas acumuladas, ao passo que, no caso da regressão quantílica, essa decomposição utiliza um processo de somas acumuladas ponderadas.

Em seu artigo, He e Zhu (2003) propuseram duas abordagens para o teste, considerando erros homocedásticos e também heterocedásticos. Trataremos aqui somente do primeiro caso, deixando indicado o artigo para o entendimento do segundo. Nesse sentido, vamos supor o modelo

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta(\tau)} + e_i \tag{3.5}$$

em que os erros são independentes e identicamente distribuídos com quantil de ordem τ igual a zero. Seja $\psi_\tau(u)$ a derivada da função de perda definida em (2.7), $\boldsymbol{\hat{\beta}(\tau)}$ a estimativa do parâmetro $\boldsymbol{\beta(\tau)}$ do modelo em (3.5) para o quantil condicional de ordem τ e defina os resíduos como

$$r_i = y_i - \boldsymbol{x_i'}\boldsymbol{\hat{\beta}(\tau)}.$$

Nessas condições, os autores definem inicialmente o bloco principal para o teste de falta de ajuste em modelos de regressão quantílica como

$$\boldsymbol{R_n(t)} = n^{-1/2}\sum_{j=1}^{n}\psi(r_j)\boldsymbol{x_j}I(\boldsymbol{x_j} \leqslant \boldsymbol{t}),$$

em que I representa a função indicadora, sendo que $I(\boldsymbol{x_j} \leqslant \boldsymbol{t}) = 1$ somente quando todos os termos de $\boldsymbol{x_j}$ são menores ou iguais aos correspondentes componentes de t.

A estatística de teste é definida como

$$T_n = \max_{\|\boldsymbol{a}\|=1} n^{-1}\sum_{i=1}^{n}\left(\boldsymbol{a'R_n(x_i)}\right)^2. \tag{3.6}$$

Como T_n pode ser escrita na forma

$$T_n = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a'}\left[n^{-1}\sum_{i=1}^{n}\boldsymbol{R_n(x_i)R_n'(x_i)}\right]\boldsymbol{a}$$

então, o seu valor coincide com o maior autovalor da matriz $n^{-1}\sum_i \boldsymbol{R_n(x_i)R_n'(x_i)}$.

O cálculo do nível descritivo do teste envolve simulações para decidir se a estatística T_n excede o quantil superior de ordem α de T. Para melhor entendimento da teoria que envolve esse teste, indicamos Stute (1997). He e Zhu (2003) mostram que, sob a hipótese

de que o modelo é da forma (3.5) e para n grande, T_n converge em distribuição para a variável aleatória T, em que T é o maior autovalor de $\int \boldsymbol{R(t)R'(t)w(t)dt}$ e w é a função de distribuição de X.

Segundo os autores, como a distribuição limite do teste é independente da distribuição dos erros, então para calcular o nível descritivo do teste, várias reamostras de $(Y_1^*, \ldots, Y_n^*)$ são geradas, em que os Y_j^* possuem qualquer distribuição contínua com o quantil de ordem τ igual a zero. Em cada reamostra gerada, a estatística de teste é calculada a partir dos resíduos r_j^*, $r_j^* = y_j^* - \boldsymbol{x_j'}\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$, da seguinte forma,

$$\boldsymbol{R_n^*(t)} = n^{-1/2} \sum_{j=1}^{n} \psi(r_j^*)\boldsymbol{x_j} I(\boldsymbol{x_j} \leqslant \boldsymbol{t}).$$

Tendo essa sugestão em vista, executamos um estudo de simulação utilizando a distribuição Laplace assimétrica (Yu e Zhang, 2005) definida com três parâmetros, sendo μ parâmetro de localização, σ parâmetro de escala e τ parâmetro de assimetria, para gerar as observações Y_j^*. No entanto, observamos que somente a suposição de que variáveis Y_j^* sejam contínuas com o quantil de ordem τ igual a zero não foi suficiente para produzir os resultados esperados do teste. No caso da distribuição Laplace assimétrica, foi necessário substituir σ pelo seu estimador de máxima verossimilhança, que pode ser calculado como

$$\hat{\sigma} = n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i). \tag{3.7}$$

Consideramos para esse estudo de simulação as mesmas condições da Seção 4.1 de He e Zhu (2003), sob a hipótese nula. Dessa forma, fixando o nível de significância do teste em 10%, verificamos a proporção de rejeição do teste em 1000 simulações nas duas abordagens, isto é, utilizando o estimador de máxima verossimilhança de σ para gerar os valores Y_j^*, que chamaremos de Modelo 1, e não considerando esse estimador, que chamaremos de Modelo 2. Os resultados obtidos estão dispostos na Tabela 3.1.

Tamanho da amostra	Modelo 1	Modelo 2
20	0,103	0,062
50	0,104	0,059
100	0,100	0,052

Tabela 3.1: *Proporção de rejeições para o teste de falta de ajuste, considerando ou não o estimador de máxima verossimilhança de σ no cálculo do p-valor do teste.*

A partir dessa simulação, podemos concluir que é necessária a utilização do estimador de máxima verossimilhança de σ quando utilizada a distribuição Laplace Assimétrica para gerar os Y_j^* no cálculo do p-valor do teste. Verificamos que, quando não foi utilizado o estimador, o nível de significância observado do teste ficou abaixo do valor fixado para os três tamanhos de amostra. Essa queda é corrigida quando geramos Y_j^* levando em consideração $\hat{\sigma}$. Embora

a estimativa da probabilidade de erro do tipo I seja inferior à fixada, esse fato pode gerar um aumento da probabilidade do erro do tipo II.

Para diminuir o custo computacional no esquema de reamostragem para obter o nível descritivo da estatística T_n, He e Zhu (2003) sugerem utilizar os seguintes passos. Seja K um número inteiro modesto, 10 ou 20 por exemplo, e seja T^* o maior autovalor de $n^{-1}\sum_i \boldsymbol{R_n^*}(\boldsymbol{x_i})\boldsymbol{R_n^{*'}}(\boldsymbol{x_i})$ de cada realização de $\boldsymbol{R_n^*}$, então os passos para estimar o p-valor do teste são os seguintes:

1. Gere K cópias independentes de T^*. Seja $\hat{p}$ a proporção de vezes que $T^* > T_n$. Faça $M = K$.

2. Seja $p_1 = \max[\hat{p}; 0, 1]$, $p_L = \hat{p} - 3(p_1(1-p_1)/M)^{1/2}$ e $p_U = \hat{p} + 3(p_1(1-p_1)/M)^{1/2}$, em que M é igual ao número de cópias de T^* utilizadas.

3. Se $p_U < \alpha$, então rejeite H_0. Se $p_L > \alpha$, então não rejeite H_0. Em caso contrário, continue o processo, em primeiro lugar, gerando K cópias indepentes de T^*, segundo, atualizando $M \leftarrow M + K$ e, por último, atualizando $\hat{p}$ usando o total de cópias de M cópias de T^*; vá então ao Passo 2.

O processo iterativo entre os passos 2 e 3 termina quando M excede um valor prédeterminado, como 1000 iterações, por exemplo. A idéia básica desse processo sequencial é simples. Os valores p_L e p_U são utilizados como um intervalo de confiança de 99% para o p-valor de T_n. O processo iterativo é terminado de forma mais rápida quando as evidências contra a hipótese nula são fracas. Por exemplo, se o valor-p é igual a 0,4, então a probabilidade que $\hat{p} \leqslant 0,1$ com $M = 20$ é igual a 0,0036, ou seja, a probabilidade é muito pequena. Por outro lado, se a evidência contra a hipótese nula é forte, de forma que o valor-p é pequeno, então o uso de p_1 ao invés de $\hat{p}$ ao construir o intervalo de confiança no Passo 2 garante um mínimo de 80 reamostras utilizadas para $\alpha = 0,1$. Quanto mais perto o valor do nível descritivo do teste é de α, maior o número de reamostras que serão necessárias.

Para entender melhor o comportamento desse teste de falta de ajuste, propusemos algumas situações com o intuito de analisar o valor da estatística de teste, quando estimamos a mediana condicional de $Y|x$, com o modelo $\boldsymbol{Y} = \beta_0(0,5) + \beta_1(0,5)\boldsymbol{x} + \boldsymbol{\varepsilon}$. As situações propostas foram simuladas da seguinte maneira

(a) $\boldsymbol{Y} = \boldsymbol{\varepsilon}$,

(b) Utilizando $\boldsymbol{Y}$ do item anterior, calculamos $\boldsymbol{Y}^* = \boldsymbol{Y} + a$,

(c) $\boldsymbol{Y} = -0,2\boldsymbol{x}^2 + \boldsymbol{x} + \varepsilon$,

(d) $\boldsymbol{Y} = +0,2\boldsymbol{x}^3 - 0,2\boldsymbol{x}^2 + \boldsymbol{x} + \varepsilon$,

em que $X \sim U(0, 10)$, $\varepsilon \sim N(0, 1)$, a assume valores 2 ou -2 e foi proposto da mesma maneira que na Seção 2.6, ou seja, quando y_i está acima da reta estimada para a mediana,

então tomamos $a = 2$, enquanto que quando y_i está abaixo da reta estimada para a mediana, foi adotado $a = -2$. Vimos naquela seção que a reta estimada não é influenciada por essa alteração devido à robustez do estimador. O motivo para analisar essa situação é ressaltar que a estatística de teste para falta de ajuste do modelo também não sofre alteração devido ao fato de não levar em conta no seu cálculo o valor absoluto do resíduo, mas somente o seu sinal. As quatro situações formuladas estão apresentadas na Figura 3.5.

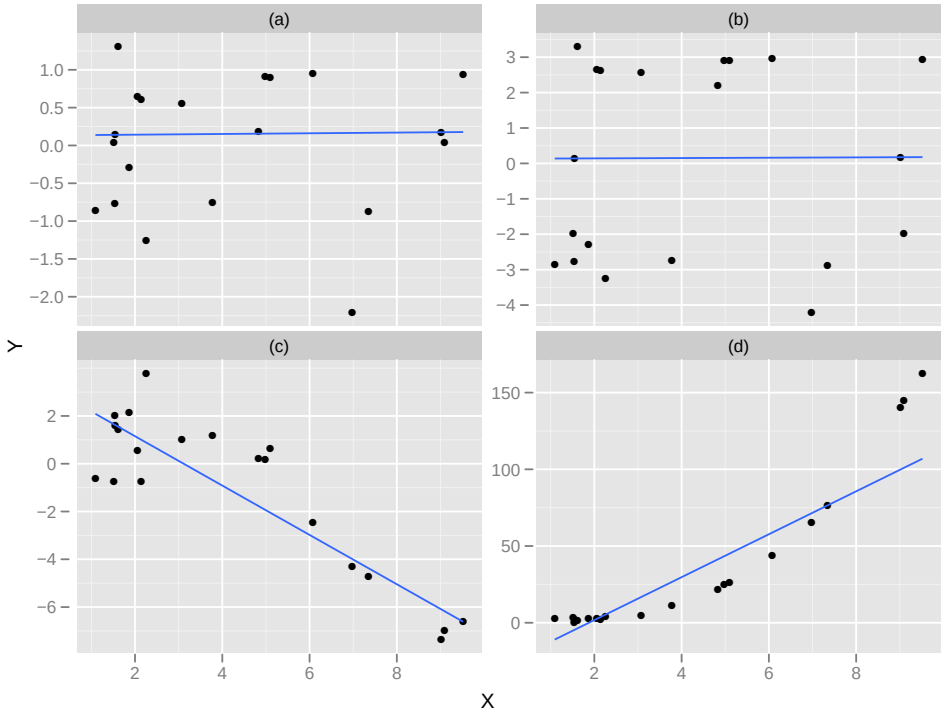

Figura 3.5: *Diferentes situações para o cálculo da estatística de falta de ajuste.*

A estimativa para a estatística de teste, T_n, e o seu respectivo p-valor em cada situação podem ser observados na Tabela 3.2.

Situação	T_n	p-valor
(a)	0,831	0,162
(b)	0,831	0,210
(c)	1,389	0,036
(d)	2,247	0,006

Tabela 3.2: *Cálculo de T_n e seu respectivo p-valor nas quatro situações propostas.*

É importante observar alguns pontos com relação ao cálculo de T_n e as conclusões que podem ser tiradas nessas situações propostas. Em primeiro lugar, o teste de falta de ajuste rejeitou a relação linear $\boldsymbol{Y} = \beta_0(0,5)+\beta_1(0,5)\boldsymbol{x}+\boldsymbol{\varepsilon}$ nas situações (c) e (d), o que era esperado, uma vez que a variável aleatória Y, nesses casos, não foi gerada a partir de uma relação do tipo $\beta_0 + \beta_1\boldsymbol{x}$. Em segundo lugar, como já havíamos adiantado, a estatística de teste, T_n, não leva em consideração o valor absoluto do resíduo, somente o seu sinal, logo os casos (a)

e (b) apresentam o mesmo valor para essa estatística. O fato dos p-valores não serem os mesmos também se refere à maneira como estes são calculados, que ocorre por simulação. E por último, a não rejeição da hipótese de linearidade nas duas primeiras situações não deve ser confundida com a hipótese do coeficiente de X no modelo $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon}$ ser igual a zero. Nas situações (c) e (d), devemos concluir que não há evidências para dizer que $\beta_1 = 0$, porém rejeitamos também a hipótese de linearidade do modelo.

Além disso, é importante destacar que o valor da estatística T_n cresce ao passarmos das situações (a) e (b) para (c) e depois para (d). Isso é interessante do ponto de vista do teste de falta de ajuste, porque está associado a um aumento no afastamento da hipótese de linearidade, principalmente quando passamos de (c) para (d).

Para a execução do teste ao longo dessa seção, implementamos algumas funções no aplicativo estatístico R, que estão disponibilizadas no Apêndice A.

3.3 Análise Gráfica

Finalizando o capítulo, gostaríamos de propor a verificação da qualidade do ajuste do modelo por meio de análises de gráficos dos resíduos. Esse tipo de análise está ligado à suposição de alguma distribuição para a variável resposta. Nos modelos de regressão quantílica, como o estimador para os parâmetros do modelo é o estimador de máxima verossimilhança quando os erros têm distribuição de Laplace Assimétrica (ver Apêndice C), então consideraremos essa distribuição para os erros do modelo nessa seção.

Antes de qualquer discussão sobre os possíveis gráficos de análise dos resíduos, devemos definir exatamento qual o resíduo que pode ser utilizado com o intuito de verificar a qualidade do ajuste do modelo. Dunn e Smyth (1996) propõem, em um contexto geral, utilizar os resíduos quantílicos, os quais, apesar do nome, não têm relação com os modelos de regressão quantílica. Na verdade, o nome desses resíduos está vinculado ao método como esses são calculados.

Antes de definir esses resíduos, devemos enunciar alguns resultados para a distribuição Laplace Assimétrica. Assim, se $Y \sim LA(\mu, \sigma, \tau)$, então sua função de distribuição acumulada é da seguinte forma

$$F(y;\mu,\sigma,\tau) = \begin{cases} \tau \exp\left(\dfrac{1-\tau}{\sigma}(y-\mu)\right), & \text{se } y \leqslant \mu, \\ 1-(1-\tau)\exp\left(-\dfrac{\tau}{\sigma}(y-\mu)\right), & \text{se } y > \mu. \end{cases}$$

Tendo em vista essa definição, podemos apresentar os resíduos quantílicos. Como a função $F(y;\mu,\sigma,\tau)$ é contínua, então, pelo Teorema da Transformação Integral, é também uma variável aleatória uniformemente distribuída no intervalo (0,1). Nesse caso, os resíduos quantílicos são definidos da seguinte maneira

$$r_{q,i} = \Phi^{-1}\left\{F(y_i, \hat{\mu}_i, \hat{\sigma}, \tau)\right\} \tag{3.8}$$

em que $\Phi(.)$ é a função de distribuição acumulada da distribuição normal padrão, $\hat{\mu} = \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$, $\hat{\sigma}$ é o estimador de máxima verossimilhança de σ como em (3.7) e τ é o parâmetro fixado para o ajuste do modelo de regressão quantílica. Segundo os autores, a menos da variabilidade amostral em $\hat{\mu}, \hat{\sigma}$, os resíduos $r_{q,i}$ são exatamente normais padrão. Isto implica que a distribuição de $r_{q,i}$ converge para normal padrão se β e σ são consistemente estimados. Ainda segundo os autores, esses resíduos são um caso particular dos resíduos brutos[1] propostos por Cox e Snell (1968).

Tendo sido realizada a definição dos resíduos quantílicos propostos por Dunn e Smyth (1996), devemos notar algumas relações importantes. Conforme indicado no Apêndice C, o parâmetro μ define o quantil de ordem τ da distribuição Laplace Assimétrica. Logo, quando supomos uma relação linear para o quantil condicional da variável resposta Y dado as variáveis explicativas X, da seguinte forma

$$Q_\tau(Y|x) = \boldsymbol{x_i'}\boldsymbol{\beta}(\boldsymbol{\tau})$$

e considerando que os erros do modelo, como em (3.5), têm distribuição Laplace Assimétrica, então o estimador da mínima soma dos erros absolutos ponderados, $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$, coincide com o estimador de máxima verossimilhança. Dessa forma, obtemos estimadores consistentes de μ e σ para substituir na expressão de $r_{q,i}$.

Para exemplificar a praticidade desses resíduos, vamos considerar inicialmente um problema univariado para uma amostra de tamanho 1000 com distribuição Laplace Assimétrica, com parâmetros μ, σ e τ iguais a 0, 1 e 0,50, respectivamente. Em seguida, calculemos os resíduos $r_{q,i}$, para $\tau = 0,5$, porém com um pequeno erro nos estimadores de μ e σ. Utilizaremos $\hat{\mu} = 2$ e $\hat{\sigma} = 2$. Na Figura 3.6, temos o histograma dos resíduos quantílicos calculados para esses valores, mas com um erro nos estimadores.

Com isso, podemos observar que, caso os estimadores sejam imprecisos, então a distribuição resultante dos $r_{q,i}$ não vai ser normal padrão. É notável que a distribuição resultante dos resíduos quantílicos, nesse exemplo, está um pouco deslocada à esquerda, em comparação com a normal padrão, ou seja, possui uma assimetria à direita.

Agora, com relação ao uso desses resíduos em modelos de regressão quantílica, também podemos exemplificar suas qualidades em detectar falta de ajuste dos modelos postulados. Retomemos os dados da seção anterior, quando descrevemos quatro situações diferentes para aplicar o teste de ajuste, em especial, as situações (c) e (d) em que verificamos que havia uma falta de ajuste do modelo para a mediana condicional. Construindo o gráfico dos resíduos quantílicos em função dos valores preditos pelo modelo, obtemos a Figura 3.7.

É possível notar pelo padrão dos resíduos apresentados nos gráficos que há alguma relação não linear entre a variável resposta e a variável explicativa que o modelo não está sendo capaz de explicar. Nesse exemplo, sabemos que a variável resposta tem relação polinomial de grau 3 com a variável explicativa, por isso, os resíduos quantílicos apresentaram esse

[1] Traduzido do termo *crude residuals*

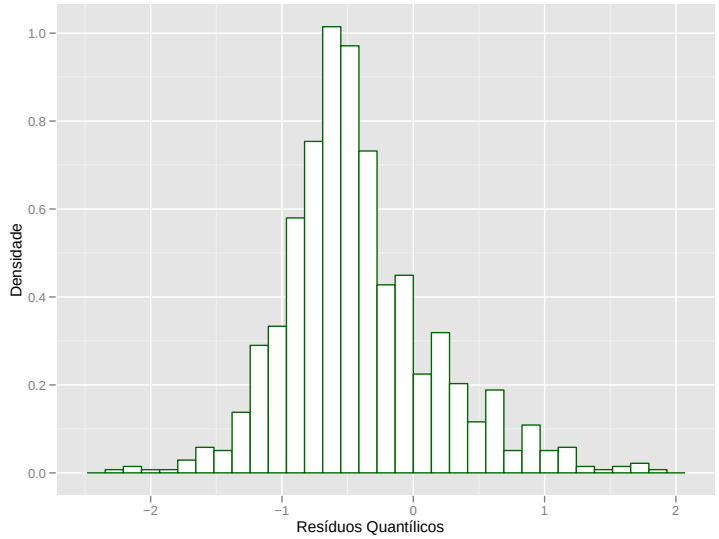

Figura 3.6: *Histograma dos resíduos quantílicos para os dados gerados, com erro nos estimadores dos parâmetros da distribuição desses dados.*

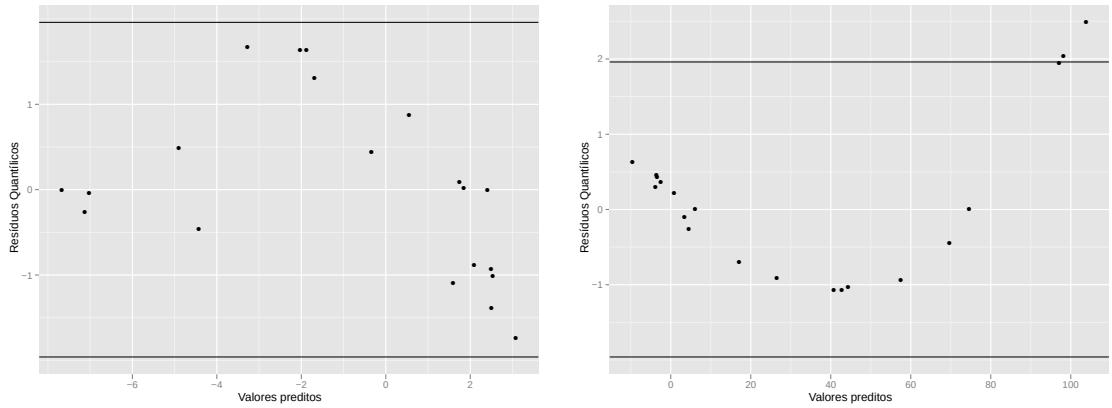

Figura 3.7: *Gráfico dos resíduos quantílicos em função dos valores preditos para as situações (c) e (d), respectivamente.*

comportamento.

Com os resíduos quantílicos postulados, e tendo em vista que espera-se que estes tenham distribuição normal se os parâmetros são consistentemente estimados, então outros gráficos podem ser utilizados para a verificação da qualidade do ajuste, além daquele apresentado anteriormente com os resíduos quantílicos em função dos valores preditos. Gráficos como *qq-plot* e histogramas também são interessantes para analisar se a distribuição dos resíduos está próxima da distribuição normal. Além disso, consideraremos também gráficos de envelope, que são bastante utilizados na análise de modelos lineares generalizados, ver Atkinson (1981).

Para motivar o uso do gráfico de envelope em modelos de regressão quantílica, utilizaremos novamente os dados de poluição do ar em cidades norte-americanas.

Seja o modelo para a mediana condicional da quantidade de enxofre em miligramas por metro cúbico em função das variáveis temperatura em graus Fahrenheit, número de fábricas que empregam mais de 20 homens e população de cada cidade. Tal modelo foi

utilizado para exemplificar o cálculo do coeficiente de determinação $R^1(\tau)$. Para esse modelo ajustado, obtivemos o gráfico de envelope para os resíduos apresentado na Figura 3.8. A partir do gráfico de envelope construído, concluímos que a distribuição Laplace Assimétrica com parâmetro $\tau = 0,5$ é adequada para explicar a distribuição condicional da variável resposta.

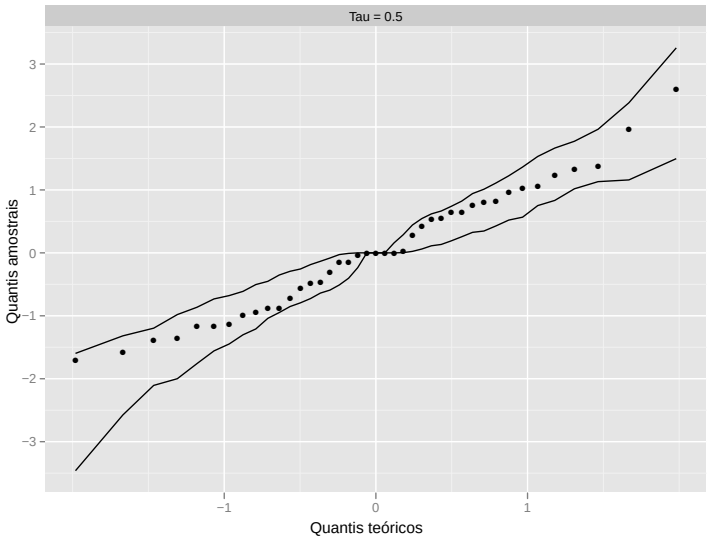

Figura 3.8: *Gráfico de envelope para a mediana condicional de SO2 em função de MAN, FAB e TEMP.*

Para finalizar essa análise gráfica dos modelos de regressão quantílica, gostaríamos de propor uma possibilidade de análise inicial que pode ser feita quando utilizados os modelos de regressão quantílica. Entendemos que a retirada e a adição de variáveis explicativas mudam os resultados relacionados ao modelo considerado, isto é, mudam a relação estimada entre as variáveis resposta e explicativas. De maneira similar, com a regressão quantílica é possível variar também o quantil condicional com o intuito de encontrar aquele quantil para o qual essa relação é mais interessante.

Com esse objetivo, retomando o banco de dados sobre concentração de imunoglobulina G em crianças, podemos gerar o gráfico de envelope, assim como os outros gráficos sugeridos nessa seção, para diversos valores dos quantis condicionais da concentração de imunoglobulina G dada a variável Idade. O resultado pode ser observado na Figura 3.9.

Com os diversos gráficos de envelope apresentados, podemos analisar inicialmente para quais faixas de quantis a relação entre imunoglobulina e idade é mais adequada. Nesse caso, diríamos que possivelmente a distribuição condicional da variável resposta em função da idade não tem assimetria à esquerda, pois os modelos estimados para os quantis acima de 0,50 não parecem estar bem ajustados, de acordo com a Figura 3.9. Além disso, os gráficos de envelope sugerem que a relação entre concentração de imunoglobulina e idade é mais forte para os quantis entre 0,2 e 0,5.

Novamente aqui, implementamos funções no `R` para o cálculo das diversas medidas, estatísticas de teste e gráficos sugeridos nesse capítulo. Todas essas funções estão disponibilizadas

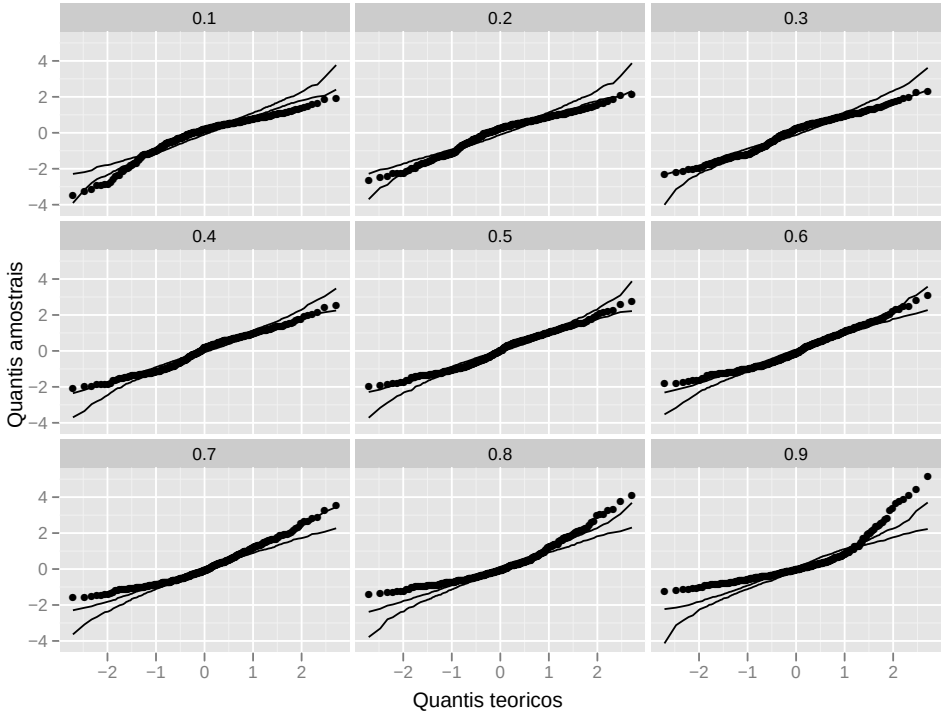

Figura 3.9: *Gráficos de envelope para modelos de regressão quantílica que estimam o efeito de idade na concentração de imunoglobulina em crianças.*

no Apêndice A para consulta e uso.

No presente capítulo, apresentamos algumas possibilidades para se avaliar determinados aspectos relacionados à qualidade de ajuste do modelo. Terminamos apresentando análises gráficas dos modelos, considerando a distribuição Laplace Assimétrica. No próximo capítulo, iremos consolidar o uso das técnicas aqui apresentadas em aplicações com dados reais.

Capítulo 4

Aplicações

No primeiro capítulo desse texto, apresentamos uma motivação inicial para o uso de modelos de regressão quantílica. Em seguida, discutimos alguns resultados inferenciais relacionados a esses modelos, como estimação, intervalos de confiança e teste de hipóteses para os parâmetros do modelo. No terceiro capítulo, vimos alguns procedimentos propostos na literatura para avaliar a qualidade de ajuste de modelos de regressão quantílica e propusemos uma análise por meio de gráficos considerando a distribuição Laplace Assimétrica. Nesse capítulo, vamos aplicar todos esses resultados em dois bancos de dados, com o intuito de consolidar a utilização de modelos de regressão quantílica.

Inicialmente, reconsideramos o banco de dados sobre poluição em cidades situadas nos Estados Unidos da América, apresentado no primeiro capítulo e retomado no terceiro. Discutiremos como as relações lineares entre as variáveis, construídas para estimação da média da variável resposta a partir da análise de regressão usual, podem não ser válidas para diferentes quantis da distribuição condicional da variável resposta.

Na sequência, utilizaremos dados obtidos a partir da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2009. Essa pesquisa apresenta informações sobre renda no Brasil, além de diversas variáveis socio-econômicas, que permitem a construção de modelos de regressão quantílica. Com esse dados, iremos comparar as estimativas dos parâmetros dos modelos para diferentes quantis.

4.1 Dados de poluição de cidades norte-americanas

Conforme já dito no Capítulo 1, esses dados foram retirados de Hand et al. (1994) e se referem a dados de poluição do ar medida em 41 cidades norte-americanas entre os anos de 1969 e 1971. Os dados estão disponibilizados no Apêndice B.

No presente capítulo, utilizaremos todas as variáveis desse banco de dados. São elas:

- SO2: concentração de dióxido de sulfato no ar, em miligramas por metro cúbico;
- TEMP: temperatura média anual, em graus Fahrenheit;

- FAB: número de fábricas que empregam mais de 20 pessoas;
- POP: população, em milhares de habitantes;
- VENTO: velocidade média anual dos ventos, em milhas por hora;
- CHUVA: volume médio anual de chuvas, em polegadas;
- DIASCHUVA: número médio de dias com chuva na cidade.

Na Figura 4.1, podemos observar o diagrama de dispersão da variável SO2 em função das outras variáveis incluídas nesse estudo. Entretanto, não é possível observar nenhuma relação clara entre essas variáveis.

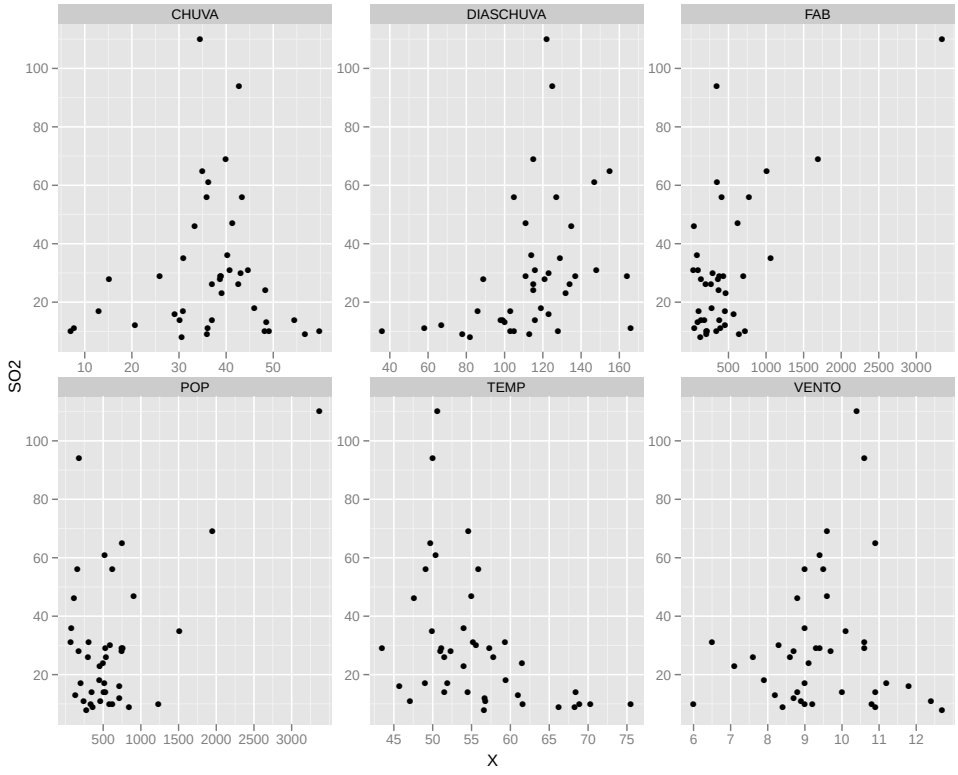

Figura 4.1: *Gráficos de dispersão da variável SO2 em função das outras variáveis incluídas no estudo.*

Em uma primeira análise, poderíamos ajustar um modelo de regressão usual, para quantificar o valor esperado para SO2 em função das outras variáveis:

$$\text{SO2} = \beta_0 + \beta_1 \text{TEMP} + \beta_2 \text{FAB} + \beta_3 \text{POP} + \beta_4 \text{VENTO} + \beta_5 \text{CHUVA} + \beta_6 \text{DIASCHUVA} + \boldsymbol{\varepsilon}, \quad (4.1)$$

em que os erros $\boldsymbol{\varepsilon}$ são normalmente distribuídos com média zero e variância constante.

As estimativas para os parâmetros desse modelo, bem como as demais medidas associadas à inferência sobre eles, estão na Tabela 4.1.

Parâmetro	Estimativa	Erro Padrão	Valor-t	P-valor
β_0	111,73	47,32	2,36	0,024
β_1	-1,27	0,62	-2,04	0,049
β_2	0,06	0,02	4,12	$< 0{,}001$
β_3	-0,04	0,02	-2,60	0,013
β_4	-3,18	1,82	-1,75	0,089
β_5	0,51	0,36	1,41	0,167
β_6	-0,05	0,16	-0,32	0,750

Tabela 4.1: *Estimativas para os parâmetros do modelo (4.1).*

Retirando a variável DIASCHUVA do modelo, a variável POP deixa de apresentar contribuição significante ao nível de 5%, o que nos deixa com o seguinte modelo final, com as mesmas suposições anteriores,

$$\text{SO2} = \beta_0 + \beta_1 \text{TEMP} + \beta_2 \text{FAB} + \beta_4 \text{VENTO} + \beta_5 \text{CHUVA} + \varepsilon. \tag{4.2}$$

Um resumo da análise inferencial para esse novo modelo encontra-se na Tabela 4.2.

Parâmetro	Estimativa	Erro Padrão	Valor-t	P-valor
β_0	123,12	31,29	3,93	$< 0{,}001$
β_1	-1,61	0,40	-4,01	$< 0{,}001$
β_2	0,03	0,00	5,62	$< 0{,}001$
β_4	-3,63	1,89	-1,92	0,063
β_5	0,52	0,23	2,29	0,028

Tabela 4.2: *Estimativas para os parâmetros do modelo (4.2).*

Segundo esse último modelo ajustado, estima-se que que o aumento em um grau Fahrenheit na temperatura média, mantidas as demais variáveis explicativas fixas, diminui a concentração média de dióxido de sulfato em 1,61 miligrama por metro cúbico. Além disso, o aumento de uma unidade no número de fábricas com mais de 20 empregados aumenta a concentração média de SO_2 em 0,03 miligramas. E ainda, o aumento em uma unidade na velocidade média do vento diminui a concentração média em 3,63, ao passo que o aumento de uma unidade no volume médio de chuva aumenta a concentração média de dióxido de sulfato em 0,53 miligramas por metro cúbico.

Por outro lado, essa análise de regressão pode ser feita utilizando os modelos de regressão quantílica, estudando se esses efeitos são os mesmos em diferentes quantis da distribuição condicional da variável resposta. Se considerarmos inicialmente somente a regressão da mediana, assim como no primeiro capítulo, temos as seguintes estimativas para os parâmetros, conforme Tabela 4.3, já desconsiderando a variável DIASCHUVA, que também não se mostrou significativa nesse modelo.

Nessa tabela, observamos as estimativas dos parâmetros e os respectivos intervalos de

Parâmetros	Estimativas	Limite inferior	Limite superior
β_0	96,46	81,95	144,57
β_1	-0,86	-1,87	-0,72
β_2	0,06	0,04	0,08
β_3	-0,03	-0,06	-0,01
β_4	-3,79	-6,84	-1,88
β_5	0,17	0,08	0,84

Tabela 4.3: *Estimativas dos parâmetros para a regressão da mediana.*

confiança com coeficiente de confiança igual a 95% considerando o método de escores ordinais, que apresentou boa performance nos estudos de simulação construídos na Seção 2.4. Com relação à interpretação dos resultados, estimamos que o aumento de uma unidade da variável POP, que corresponde a 1.000 pessoas, mantidas as outras variáveis fixas, diminui a concentração mediana de dióxido de sulfato no ar em 0,03 miligramas por metro cúbico. As outras estimativas têm interpretação similar à do modelo da média, com diferença somente na ordem de grandeza das estimativas e também que a interpretação dos efeitos das variáveis se referem à concentração mediana e não concentração média de dióxido de sulfato no ar.

Entretanto, conforme já vimos ao longo desse texto, os modelos de regressão quantílica permitem avaliar a relação das variáveis envolvidas no estudo além de uma posição central, que corresponderia à estimação da média e da mediana. Nesse sentido, podemos estimar o efeito dessas variáveis na concentração de dióxido de sulfato em diferentes pontos da distribuição condicional dessa variável, como na cauda inferior, com o quantil condicional de ordem 10% e também na cauda superior, com o quantil condicional de ordem 90%, por exemplo.

Com esse intuito, foram construídas as Figuras 4.2, 4.3 e 4.4 de modo a ilustrar essa relação em diferentes pontos da distribuição condicional da variável resposta. Nessas figuras, podemos observar as estimativas dos coeficientes de regressão de cada variável explicativa nos modelos de regressão quantílica, para os quantis 0,1 até 0,9, com diferença de 0,1 entre eles. Além dos coeficientes estimados, as figuras fornecem também um intervalo de confiança com coeficiente de confiança 0,95 para os parâmetros, considerando o métodos dos escores ordinais. É importante notar que esses intervalos de confiança não são necessariamente simétricos em torno da estimativa do parâmetro.

Com relação aos valores estimados para os parâmetros dos diferentes modelos de regressão quantílica, não há uma variação muito grande se considerarmos os diferentes quantis propostos. Se notarmos os intervalos de confiança construídos, podemos destacar que a variável FAB tem coeficientes positivos e estatisticamente significantes para quantis iguais ou superiores a 0,4. Além disso, é importante notar que a variável CHUVA, no quantil 0,5, tem o intervalo de confiança contendo o valor 0, diferentemente do indicado na Tabela 4.3. Essa diferença é explicada pela presença da variável DIASCHUVA, que é considerada para estimar os diferentes quantis observadas nos gráficos, ao passo que na regressão da mediana

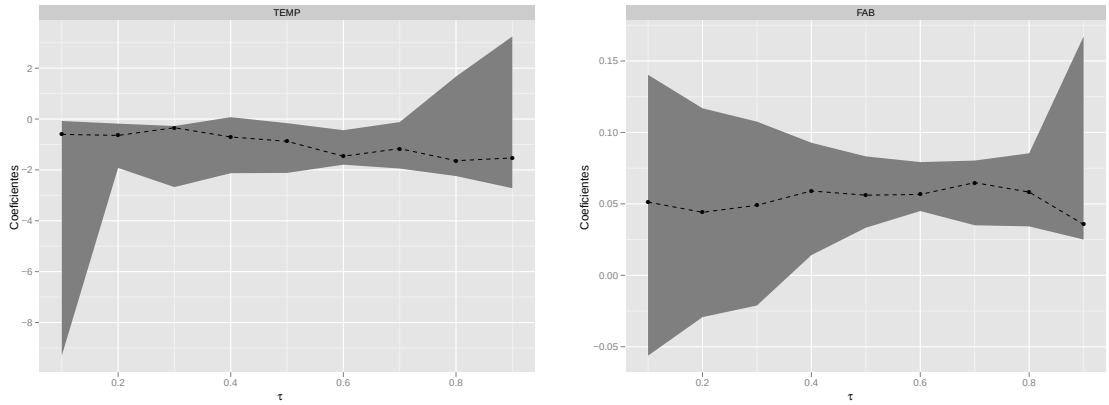

Figura 4.2: *Estimativas dos coeficientes de regressão para as variáveis TEMP e FAB em diferentes modelos de regressão quantílica com quantis iguais a* $0,1; 0,2; \ldots; 0,9$ *e variável resposta SO2.*

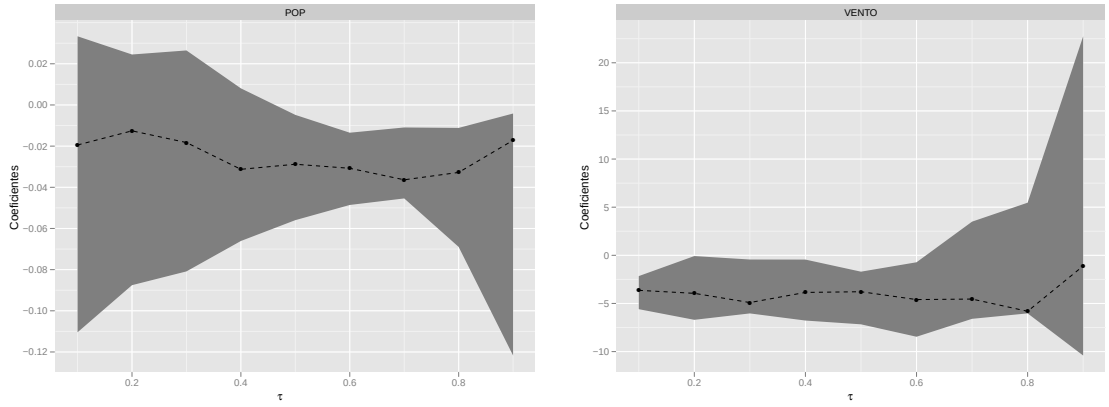

Figura 4.3: *Estimativas dos coeficientes de regressão para as variáveis POP e VENTO em diferentes modelos de regressão quantílica com quantis iguais a* $0,1; 0,2; \ldots; 0,9$ *e variável resposta SO2.*

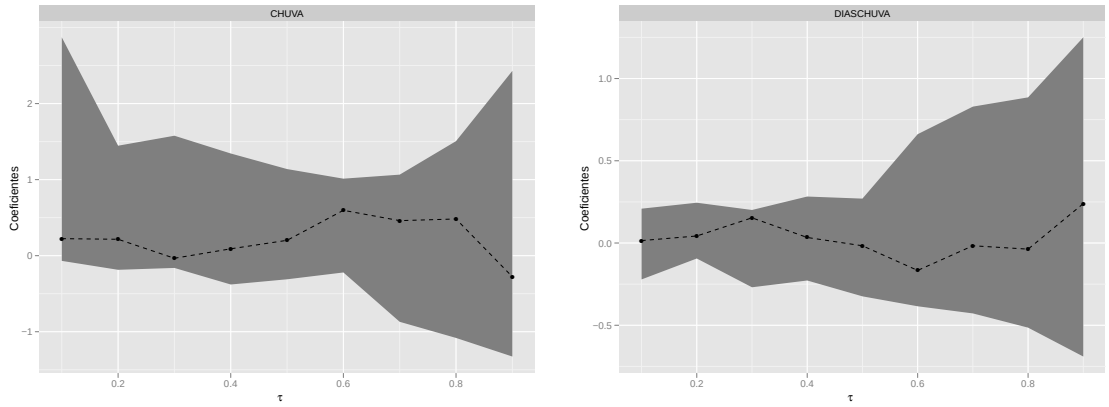

Figura 4.4: *Estimativas dos coeficientes de regressão para as variáveis CHUVA e DIASCHUVA em diferentes modelos de regressão quantílica com quantis iguais a* $0,1; 0,2; \ldots; 0,9$ *e variável resposta SO2.*

essa variável não foi utilizada. Ainda com relação à variável DIASCHUVA, a mesma não se mostrou significante em nenhum dos modelos de regressão quantílica ajustados considerando os intervalos de confiança construídos, enquanto que as demais variáveis se mostraram

significantes em alguns quantis específicos.

A Tabela 4.4 apresenta as estimativas dos coeficientes de regressão quantílica para os quantis 0,1, 0,25, 0,5, 0,75, 0,9, somente para as variáveis explicativas cuja contribuição foi significante. Além das estimativas, são apresentados também os intervalos de confiança para os parâmetros e, para comparação, as estimativas dos respectivos parâmetros no modelo de regressão para a média da variável dióxido de sulfato.

		Quantis					
Variável	Parâmetros	0,10	0,25	0,50	0,75	0,90	Média
Intercepto	β_0	95,32	96,96	96,46	94,33	124,87	123,12
		[39,22;155,67]	[60,34;143,52]	[81,95;144,57]	[24,28;116,21]	[52,39;165,29]	
TEMP	β_1	-1,02	-1,04	-0,86	-1,11	-1,54	-1,61
		[-1,77;-0,30]	[-1,78;-0,58]	[-1,87;-0,72]	[-1,65;-0,05]	[-1,99;0,87]	
FAB	β_2	-	-	0,06	0,05	0,04	0,03
				[0,04;0,08]	[0,04;0,07]	[0,04;0,15]	
POP	β_3	-	-	-0,03	-0,03	-0,02	-
				[-0,06;-0,01]	[-0,04;-0,01]	[-0,12;-0,02]	
VENTO	β_4	-3,05	-3,17	-3,79	-	-	-3,63
		[-5,09;-1,32]	[-4,81;-2,00]	[-6,84;-1,88]			
CHUVA	β_5	0,12	0,17	0,17	-	-	0,52
		[0,09;3,91]	[0,09;0,94]	[0,08;0,84]			

Tabela 4.4: *Estimativas para os diversos modelos de regressão quantilíca.*

Duas conclusões interessantes podem ser observadas da Tabela 4.4. Em primeiro lugar, as variáveis VENTO e CHUVA apenas apresentam efeitos significativos na regressão da mediana e dos menores quantis, 0,1 e 0,25, dado que as demais variáveis se encontram no modelo. Por outro lado, as variáveis FAB e POP têm um comportamento oposto, com efeitos significativos nos maiores quantis, 0,75 e 0,9, e também na mediana. Todos as variáveis incluídas nesse estudo, com exceção de DIASCHUVA, apresentaram efeito significativo ao nível de 5% para a regressão da mediana.

Considerando esses modelos ajustados, podemos utilizar os testes de hipóteses propostos na Seção 2.3 para testar

$$H_0 : \beta_1(\tau) = \beta_2(\tau) = \beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = 0, \tag{4.3}$$

contra a hipótese de que pelo menos um desses parâmetros é diferente de zero. No exemplo, para usar esses testes consideraremos somente aqueles parâmetros que foram selecionados, de acordo com a Tabela 4.4. Os níveis descritivos desses testes encontram-se na Tabela 4.5.

Para esse exemplo, não foi possível calcular o valor da estatística do teste de hipótese segundo o método `Wald` proposto na Seção 2.3 devido a problemas numéricos. Todavia, de acordo com o nível de significância apresentado na Tabela 4.5, rejeitamos a hipótese em

	Quantis				
Método	0,10	0,25	0,50	0,75	0,90
`anowar`	0,001	0,001	$< 0{,}001$	$< 0{,}001$	0,002
`rank`	$< 0{,}001$	$< 0{,}001$	0,001	0,001	0,001

Tabela 4.5: *Nível descritivo dos testes de hipóteses (4.3)*

(4.3), ou seja, os dados sugerem dependência dos quantis populacionais da variável SO2 em pelo menos uma das cinco variáveis explicativas utilizadas no ajuste do modelo.

Para avaliar a qualidade do ajuste, vamos considerar o teste de falta de ajuste descrito na Seção 3.2. De forma análoga ao que foi feito nos testes de hipóteses, utilizaremos os modelos selecionados para cada quantil com os coeficientes apresentados na Tabela 4.4. Sendo assim, o p-valor do teste para os quantis propostos pode ser observado na Tabela 4.6.

	Estatística de teste	p-valor
0,10	31,37	0,608
0,25	110,80	0,104
0,50	7.238,73	0,738
0,75	11.330,20	0,376
0,90	5.675,69	0,450

Tabela 4.6: *Nível descritivo para o teste de falta de ajuste para cada modelo de regressão quantílica ajustado.*

Com relação à conclusão que obtemos a partir dos testes de falta de ajuste, de acordo com os níveis descritivos obtidos, não rejeitamos a hipótese de linearidade dos modelos ajustados. Dessa forma, não rejeitamos os modelos propostos para explicar os quantis condicionais da concentração de dióxido de sulfato no ar.

Na próxima seção, utilizaremos um conjunto de dados de renda da população brasileira, para ilustrar as técnicas de regressão quantílica apresentadas. Nessa aplicação, faremos também uso da análise gráfica proposta no capítulo anterior para exemplificação.

4.2 Dados de renda no Brasil

A utilização de modelos de regressão quantílica para explicar a relação entre renda e outras variáveis explicativas é largamente abordada na literatura. Yu et al. (2005) se baseiam em dados de renda de trabalhadores do sexo masculino na Grã-Bretanha para incentivar o uso de modelos de regressão quantílica bayesianos. Buchinsky (1994) estuda a transformação da estrutura de renda da década de 80 nos Estados Unidos a partir de modelos de regressão quantílica, verificando o efeito de experiência e anos de estudo em diferentes quantis da distribuição condicional da renda. Melly (2005) analisa as diferenças salariais entre cargos no setores público e privado na Alemanha através de modelos de regressão quantílica.

Tendo esses resultados em vista, iremos nos basear nos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2009, que possui informações de renda, além de outras variáveis socio-econômicas, como idade, sexo, anos de ensino, as quais utilizaremos para construir modelos de regressão quantílica adequados à realidade brasileira.

Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), responsável por efetuar a pesquisa, a PNAD tem o caráter de investigar anualmente características gerais da população, como educação, trabalho, rendimento e habitação. A pesquisa só não ocorre quando há o CENSO da população, como no ano de 2010. Por esse motivo, os dados mais recentes disponíveis para utilização se referem às informações de 2009.

Para a nossa aplicação, vamos selecionar indíviduos que possibilitem o estudo da relação de renda com outras variáveis. Dessa forma, selecionamos inicialmente somente pessoas com idade entre 18 e 80 anos, que trabalharam ao menos 40 horas por semana e que recebiam ao menos um terço do salário mínimo vigente no ano de 2009, que era igual a R$465,00. Com isso, obtemos uma amostra de 122.727 observações.

As variáveis consideradas nesse primeiro momento são a renda em reais, sexo, idade, estado civil, etnia, anos de estudo e Unidade Federativa (UF). Como o tamanho da amostra obtida é grande e isso pode dificultar alguns resultados de interesse, como o teste de falta de ajuste, utilizaremos a variável UF para diminuir o número de observações. Já que a variável resposta é a renda, analisando a distribuição incondicional dessa variável, caracterizada pela forte assimetria à direita, observamos que o estado de Rondônia possui características muito parecidas com os dados do Brasil, conforme podemos verificar pela Figura 4.5 e pela tabela de estatísticas descritivas da variável renda.

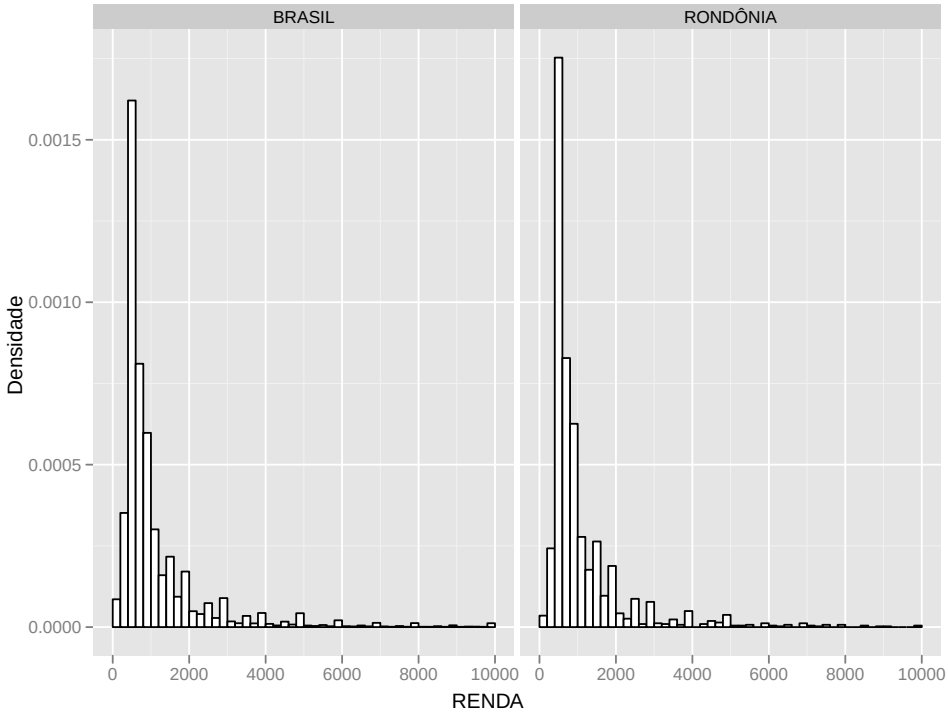

Figura 4.5: *Histograma da variável Renda, em reais, no Brasil e em Rondônia.*

	Minímo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Brasil	155	500	735	1.288	1.250	350.000
Rondônia	156	500	730	1.213	1.215	50.000

Tabela 4.7: *Estatísticas descritivas da Renda, em reais, no Brasil e em Rondônia.*

A Tabela 4.7 denota uma acentuada coincidência nas estatísticas descritivas da renda no Brasil e no Estado de Rondônia. Utilizando mais quantis além desses apresentados nessa tabela, notamos que essa coincidência acontece até o quantil 0,95.

Além disso, as variáveis explicativas também são muito próximas quando comparadas, vide Figuras 4.6 e 4.7, e as Tabelas 4.8, 4.9 e 4.10, com exceção da variável Etnia, uma vez que a proporção de pardos na composição do Estado de Rondônia é maior que no Brasil.

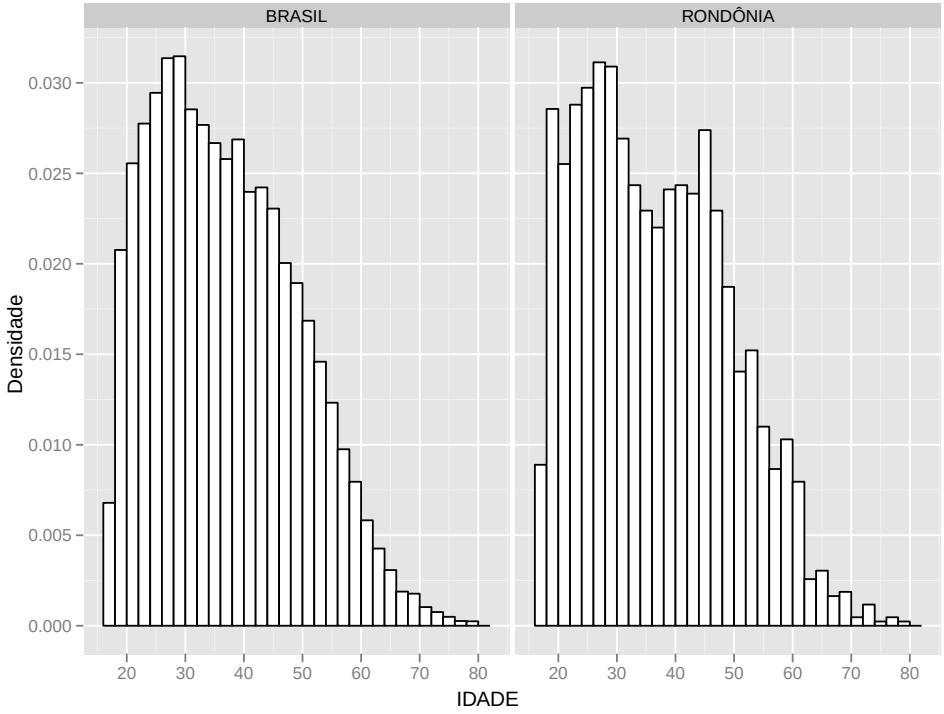

Figura 4.6: *Histograma da variável Idade, no Brasil e em Rondônia.*

Estado Civil	Brasil	Rondônia
Solteiro	42,82	46,58
Casado	49,30	45,69
Separado	2,76	1,97
Divorciado	3,46	4,45
Viúvo	1,65	1,31

Tabela 4.8: *Distribuição da variável Estado Civil no Brasil e em Rondônia, em porcentagem.*

Utilizando os dados do estado de Rondônia, obtemos uma amostra de 2.136 observações. Ainda que não possamos dizer que os resultados estimados com base nessa amostra sejam

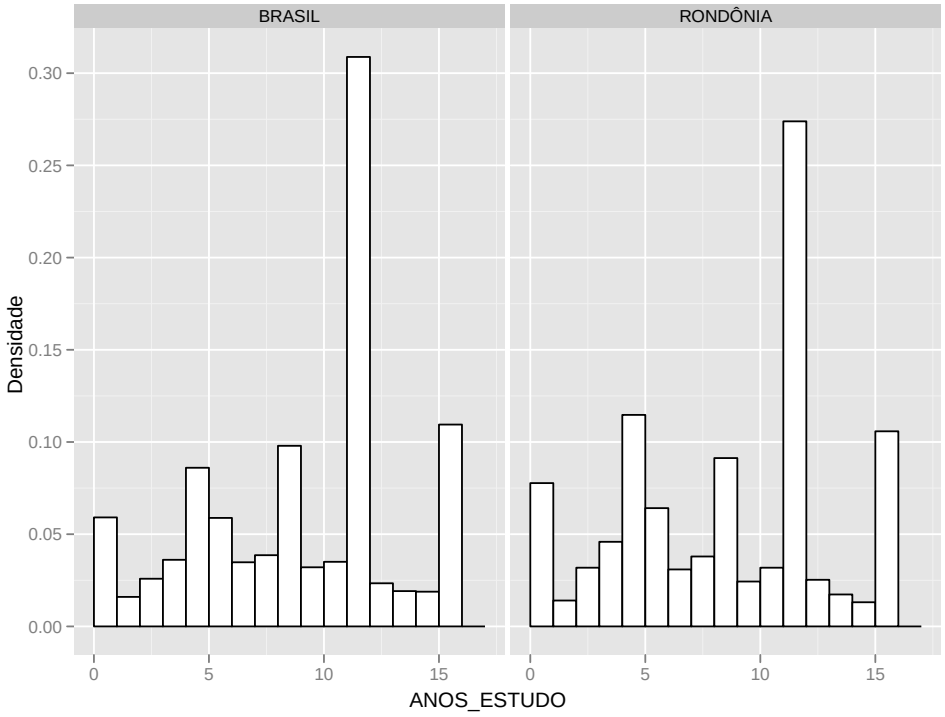

Figura 4.7: *Histograma da variável Anos de Estudo, no Brasil e em Rondônia.*

Sexo	Brasil	Rondônia
Feminino	34,94	31,69
Masculino	65,06	68,31

Tabela 4.9: *Distribuição da variável Sexo no Brasil e em Rondônia, em porcentagem.*

Etnia	Brasil	Rondônia
Branca	47,33	34,41
Preta	8,25	8,99
Parda	43,81	55,57
Amarela	0,37	0,47
Indígena	0,24	0,56

Tabela 4.10: *Distribuição da variável Etnia no Brasil e em Rondônia, em porcentagem.*

similares para o Brasil, entendemos que com esse conjunto de dados é possível ter uma boa demonstração da utilização dos modelos de regressão quantílica com dados brasileiros.

Dada a distribuição da variável Estado Civil, com o intuito de simplificar os modelos a serem estimados, vamos considerar as categorias "Separado", "Divorciado" e "Viúvo" como uma categoria única, chamada aqui de "Outros". De forma análoga, para a variável Etnia, as etnias diferentes de "Branca" formarão uma categoria única denonimada de "Outras". Dessa maneira, o número de parâmetros nos modelos é menor e a interpretação das estimativas é facilitada.

Consideraremos o seguinte modelo de interesse

$$\begin{aligned} y_i &= \beta_0(\tau) + \beta_1(\tau)\text{Idade}_i + \beta_2(\tau)\text{Sexo}_i + \beta_3(\tau)\text{Casado}_i + \beta_4(\tau)\text{Solteiro}_i \\ &\quad + \beta_5(\tau)\text{Etnia}_i + \beta_6(\tau)\text{AnosEstudo}_i + u_i, \end{aligned} \tag{4.4}$$

em que y_i é a renda do i-ésimo indivíduo presente na amostra; Idade_i é idade do i-ésimo indivíduo presente na amostra; AnosEstudo_i é a quantidade de anos de estudo do i-ésimo indivíduo presente na amostra; e as variáveis categóricas são definidas da seguinte forma

$$\text{Sexo}_i = \begin{cases} 1 & \text{se se o i-ésimo indivíduo presente na amostra é homem,} \\ 0, & \text{caso contrário;} \end{cases}$$

$$\text{Casado}_i = \begin{cases} 1 & \text{se se o i-ésimo indivíduo presente na amostra é casado,} \\ 0, & \text{caso contrário;} \end{cases}$$

$$\text{Solteiro}_i = \begin{cases} 1 & \text{se se o i-ésimo indivíduo presente na amostra é solteiro,} \\ 0, & \text{caso contrário.} \end{cases}$$

Vamos supor também que o quantil de ordem τ de $\boldsymbol{u}$, erro do modelo, é igual a zero.

Para a estimação dos coeficientes do modelo, conforme foi discutido na Seção 2.1, como o banco de dados tem muitas observações, optou-se pelo método de ponto interior. Verificamos que esse método realmente é mais rápido que o método *simplex* principalmente quando o número de observações é superior a 10.000 registros. Entretanto, para o banco de dados com as observações do estado de Rondônia, não houve diferença no tempo para estimação dos parâmetros. Além disso, como já era esperado, as estimativas utilizando os diferentes métodos são exatamente as mesmas.

Com relação ao método de utilizado para construção do intervalo de confiança para os parâmetros do modelo, de acordo com o que apresentamos na Seção 2.2, os diferentes métodos podem ser utilizados nesse exemplo, uma vez que o número de observações é grande. Para efeito de comparação, podemos observar na Tabela 4.11 a diferença nos erros-padrão para cada estimativa, utilizando o método em que consideramos que os erros são independentes e identicamente distribuídos (`iid`), o método de *bootstrap* com a utilização do algoritmo MCMB (`bootMCMB`) e também o método sem suposição de mesma distribuição para os erros (`nid`). Os valores foram estimados para a regressão da mediana, ou seja, com $\tau = 0,5$.

Podemos notar que não há uma diferença muito grande no valor do erro-padrão das estimativas, com exceção para o parâmetro $\beta_4(0,5)$, em que o valor do erro-padrão segundo o método `nid` é cerca de 50% maior que o valor do método `iid`. Por esse motivo, utilizaremos na continuação dessa seção o método `bootMCMB` para construção de intervalos de confiança para os parâmetros. Esse método, conforme vimos nas Seções 2.2 e 2.4, não necessita de suposições sobre os erros do modelo e também apresentou boa performance na construção de intervalos de confiança para os parâmetros.

		Erro Padrão		
Parâmetro	Estimativa	`iid`	`bootMCMB`	`nid`
$\beta_0(0,5)$	-282,06	66,22	69,40	83,53
$\beta_1(0,5)$	14,78	0,94	0,93	1,20
$\beta_2(0,5)$	251,24	22,28	20,65	18,45
$\beta_3(0,5)$	62,07	21,15	22,26	21,39
$\beta_4(0,5)$	-121,94	41,07	54,11	59,94
$\beta_5(0,5)$	-23,41	39,22	56,71	60,38
$\beta_6(0,5)$	50,25	2,45	3,02	2,74

Tabela 4.11: *Valores dos erros-padrão para diferentes métodos inferenciais.*

De forma semelhante ao que foi feito na seção anterior, iremos considerar gráficos como nas Figuras 4.2 a 4.4 para identificar em quais quantis as variáveis explicativas são significativas. O resultado pode ser observado nas Figuras 4.8 a 4.10.

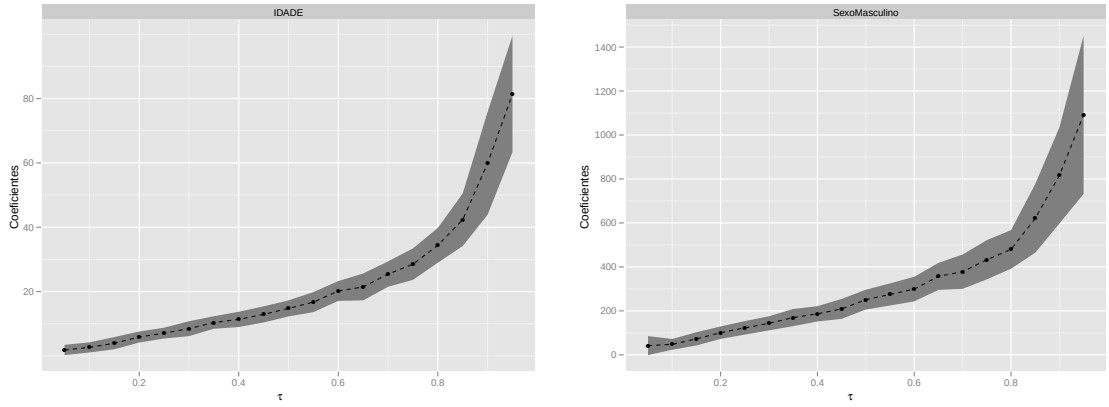

Figura 4.8: *Estimativas dos coeficientes e intervalo de confiança das variáveis Idade e Sexo.*

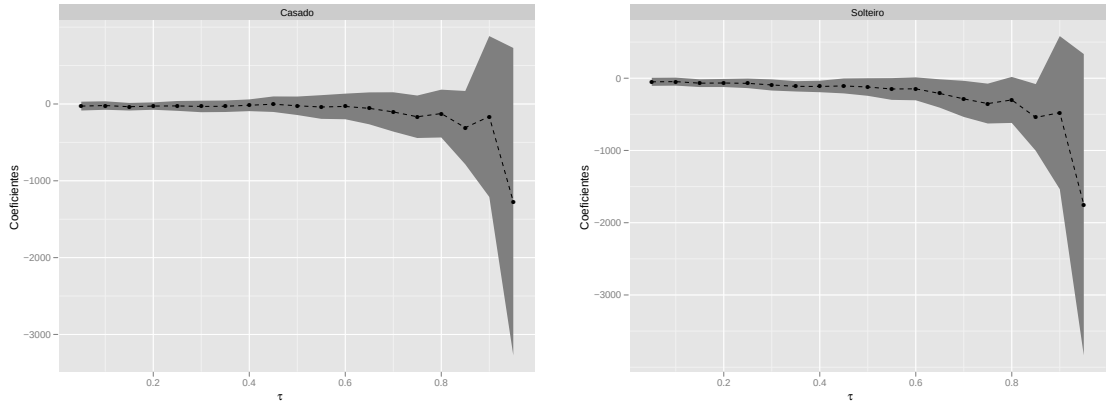

Figura 4.9: *Estimativas dos coeficientes e intervalo de confiança das variáveis Casado e Solteiro.*

Alguns resultados interessantes podem ser visualizados nesses gráficos construídos com os intervalos de confiança para os parâmetros dos modelos em diferentes quantis. Em primeiro lugar, o comportamento das estimativas dos coeficientes de regressão ao longo dos

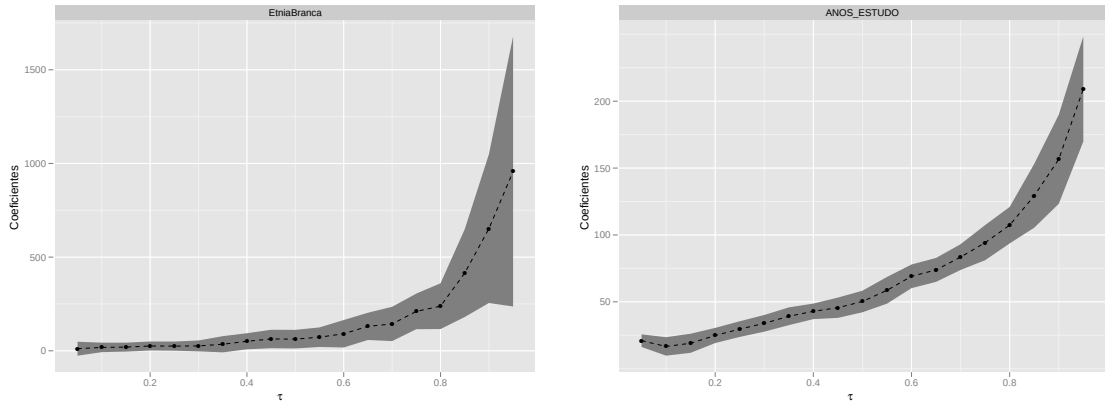

Figura 4.10: *Estimativas dos coeficientes e intervalo de confiança para modelos de regressão quantílica para variável Etnia e Anos de Estudo diferentes quantis de interesse.*

quantis para as variáveis Idade, Sexo, Etnia e Anos de Estudo são muito parecidos, isto é, nessas variáveis, conforme se aumenta o quantil de interesse, maior é o valor do coeficiente relacionado a essas variáveis. Entretanto, esse aumento ocorre de forma diferente para cada parâmetro estimado. Por exemplo, para a variável Etnia, de acordo com os intervalos de confiança construídos para o parâmetro $\beta_5(\tau)$, até o quantil $0,35$ não há evidências para dizer que o efeito dessa variável é diferente de zero, enquanto que para as outras variáveis isso não ocorre. Em segundo lugar, a variável Casado não parece ser significativa em nenhum quantil. E ainda, os coeficientes da variável Solteiro apresentam um comportamento muito próximo da variável Casado, porém em alguns quantis, o intervalo de confiança para o parâmetro não contém o zero.

Retirando a variável Casado do ajuste para esses quantis, a variável Solteiro sofre uma alteração grande nos seus coeficientes, conforme podemos verificar na Figura 4.11. A retirada dessa variável Casado não altera os resultados para as outras variáveis.

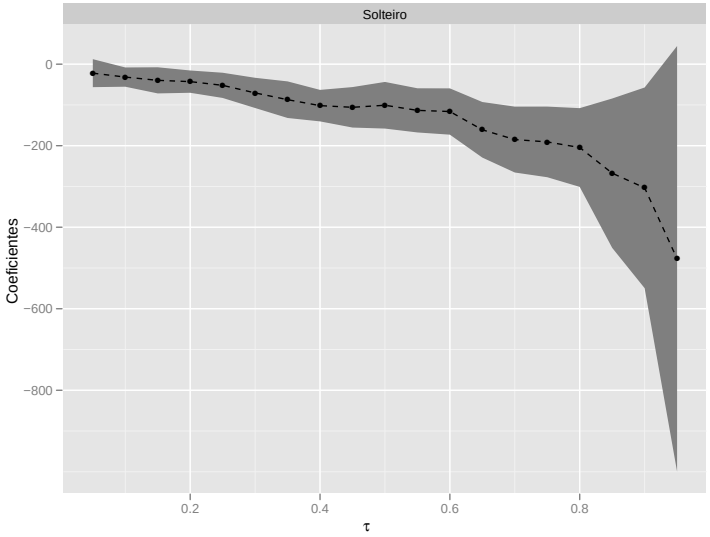

Figura 4.11: *Estimativas dos coeficientes e intervalo de confiança para modelos de regressão quantílica para variável Solteiro diferentes quantis de interesse.*

Com a retirada da variável Casado, de forma que esse estado civil agora é incluído na categoria de referência, os coeficientes da variável Solteiro são sempre negativos e do quantil 0,10 até 0,90 são significativos, de acordo com o intervalos de confiança construídos. Podemos dizer então, a partir dos coeficientes estimados, que em grande parte da distribuição condicional da renda, pessoas solteiras ganham menos que casados, divorciados, separados e viúvo, quando mantemos as outras variáveis constantes. De forma análoga, pessoas de etnia branca recebem mais que as pessoas de outras etnias e homens recebem mais que mulheres. É importante lembrar, entretanto, que no caso da variável Etnia, essa diferença só é realmente notada a partir do quantil 0,40. Em outras palavras, na cauda inferior da distribuição condicional da variável Renda, não parece haver diferenças entre pessoas de etnia branca e outras etnias.

Ainda sobre esses modelos ajustados devemos testar a hipótese

$$H_0 : \beta_1(\tau) = \beta_2(\tau) = \beta_3(\tau) = \beta_4(\tau) = \beta_5(\tau) = \beta_6(\tau) = 0, \tag{4.5}$$

contra a hipótese de que pelo menos um dos parâmetros é diferente de zero, de acordo com os testes de hipóteses propostos na Seção 2.3. Fazendo isso, verificamos que utilizando qualquer um dos métodos discutidos, a conclusão para o teste em (4.5) é a rejeição da hipótese nula em todos os quantis propostos.

Se utilizarmos o coeficiente de determinação discutido na Seção 3.1, vide Figura 4.12, para estudar o ajuste desses modelos, vemos que o ganho em utilizar essas variáveis, como Sexo, Idade, Anos de Estudo, Estado Civil e Etnia, é maior nos quantis superiores da distribuição condicional da renda. Isso talvez possa ser devido ao fato da característica de assimetria à direita da distribuição da renda, de forma que nos menores quantis, devido à maior concentração dos valores, não há uma diminuição considerável da soma ponderada dos resíduos absoluta quando passamos de um modelo somente com intercepto para um modelo com essas variáveis.

Numa etapa seguinte, construímos cinco modelos com apenas uma variável explicativa de cada vez. As mesmas conclusões são obtidas considerando essa abordagem, conforme Figura 4.13, com as variáveis Idade e Anos de Estudo obtendo maiores valores para $R^1(\tau)$ conforme se aumenta o valor de τ. A Figura 4.14 é o mesmo gráfico, porém em uma escala maior para melhorar a visualização da contribuição de cada variável de forma separada.

De modo geral, os coeficientes de explicação foram baixos e isso pode ser devido ao grande tamanho da amostra. No nosso exemplo, n=2136.

Selecionando menos quantis para análise, de maneira similar ao que foi feito na seção anterior, podemos ter uma idéia mais precisa da diferença entre as estimativas nas caudas inferior e superior, além de uma posição central como a mediana, além da média. Para isso, vamos selecionar as variáveis que melhor se ajustam nos quantis 0,10, 0,25, 0,50, 0,75 e 0,90. As estimativas de cada variável com seu respectivo erro-padrão podem ser observadas na Tabela 4.12.

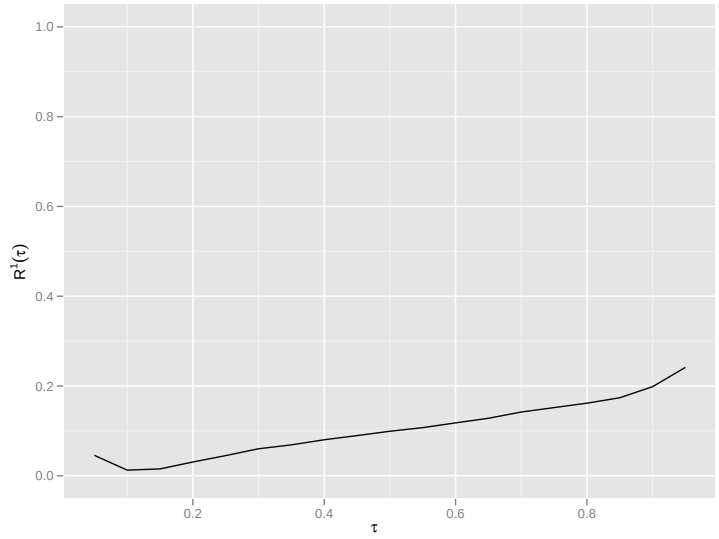

Figura 4.12: *Coeficiente de determinação para os modelos de regressão quantílica ajustados.*

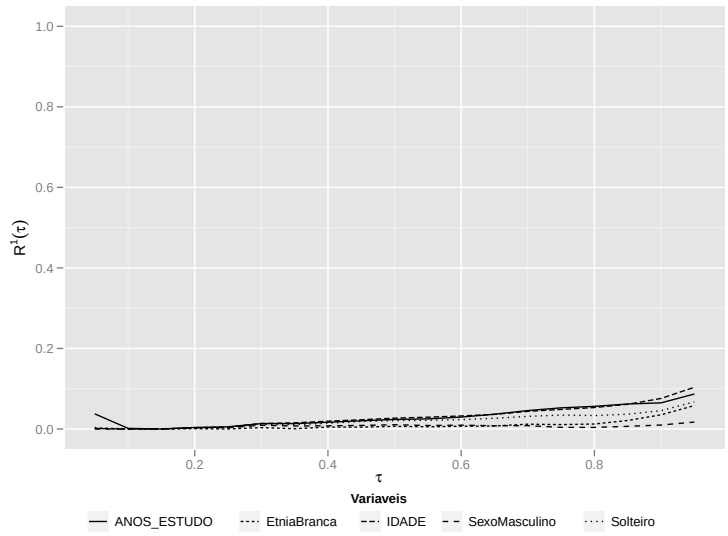

Figura 4.13: *Coeficiente de determinação para os modelos de regressão quantílica ajustados somente com uma variável explicativa.*

Assim como já havíamos discutido, a variável Etnia não se mostrou significante para os quantis 0,10 e 0,25. Além disso, o intercepto do modelo também não apresentou significância para o quantil 0,25. Para a estimativa do efeito dessas variáveis, ajustamos um modelo linear generalizado com distribuição gama e função de ligação identidade, para que a ordem das estimativas pudesse ter alguma comparação com as estimativas dos modelos de regressão quantílica. A escolha dessa distribuição se deve à assimetria da variável resposta, renda.

Uma característica interessante das estimativas dos efeitos das variáveis explicativas na média da variável resposta é a proximidade dos valores com diferentes quantis dos modelos de regressão quantílica. Por exemplo, o intercepto do modelo para a média não se mostrou significante assim como na regressão do quantil 0,25. Por outro lado, a estimativa do efeito da variável Solteiro no quantil 0,90 foi igual a -303,40, valor muito próximo da estimativa

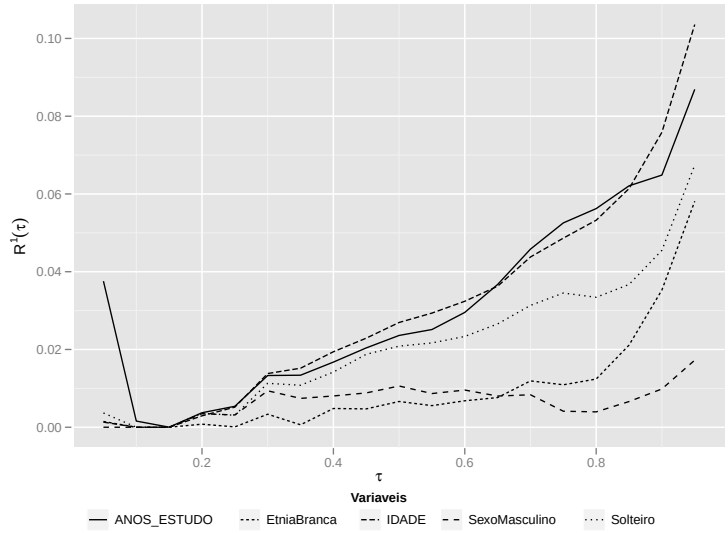

Figura 4.14: *Coeficiente de determinação para os modelos de regressão quantílica ajustados somente com uma variável explicativa, com a escala alterada.*

Variável	0,10	0,25	0,50	0,75	0,90	Média
Intercepto	180,59	-	-306,40	-849,29	-1967,47	-
	(59,19)		(81,72)	(117,59)	(328,58)	
Idade	2,79	7,03	14,84	28,57	60,74	15,27
	(0,76)	(0,25)	(1,39)	(2,60)	(7,35)	(1,21)
Sexo	48,23	120,68	250,52	422,86	807,38	230,63
	(12,50)	(10,61)	(23,05)	(42,58)	(119,66)	(43,64)
Etnia	-	-	62,16	214,29	702,75	218,20
			(24,83)	(53,76)	(195,07)	(54,53)
Solteiro	-27,95	-44,86	-100,80	-190,71	-303,40	-301,20
	(11,96)	(13,82)	(25,71)	(45,46)	(126,79)	(42,17)
Anos Estudo	17,18	28,96	50,40	94,29	155,52	54,63
	(3,28)	(1,34)	(4,23)	(6,29)	(16,45)	(4,50)

Tabela 4.12: *Estimativas para os parâmetros nos diferentes modelos de regressão quantilíca e seus respectivos erros-padrão.*

do coeficiente dessa variável na regressão para a média, -301,20.

Com relação a essa proximidade da estimativa da variável Solteiro para o quantil 0,90 e para a média, a explicação pode estar na influência que alguns pontos podem ter na estimativa para a média, enquanto que isso não acontece nos modelos de regressão quantílica, resultado discutido na Seção 2.6.

Para ilustrar isso, retiramos alguns pontos e analisamos a diferença que esses pontos têm na estimativa. Assim, se retirarmos os cinco maiores valores de renda dessa amostra de mais de 2.000 observações, os quais não são solteiros, então a estimativa do correspondente parâmetro na regressão da média passa de -301,20 para -267,50, isto é, uma variação de mais de 10% no coeficiente causada pela retirada de menos de 1% da amostra. Enquanto isso, nos modelos de regressão quantílica, essa retirada de pontos provoca uma variação nas

estimativas dos coeficientes bem menor.

Outro resultado importante que deve ser notado da Tabela 4.12 se refere às estimativas tanto da regressão quantílica quanto da regressão para a média para as variáveis Idade, Sexo e Anos de Estudo. Conforme verificamos anteriormente, para essas variáveis, a medida que aumenta o quantil condicional, maior o valor para o coeficiente da regressão quantílica. Também para essas variáveis, a estimativa do coeficiente de regressão nas regressões da média e mediana estão bem próximos. E ainda, as estimativas são todas positivas, então não há diferença nas conclusões sobre a influência das variáveis, isto é, homens recebem mais que mulheres, a cada ano que passa, o salário de uma pessoa tende a aumentar, assim como mais anos de estudos costumam significar melhor remuneração.

O uso de modelos de regressão quantílica se torna mais interessante nesse caso devido à possibilidade de calcular esses efeitos para diferentes quantis da distribuição condicional de renda. No exemplo brasileiro, a diferença estimada no quantil de ordem 0,1 da renda entre homens e mulheres, mantidas as demais variáveis fixas, é igual a R$ 48,23, enquanto que no quantil 0,90, essa diferença é estimada em R$ 807,38.

Essa análise pode ser interessante quando se deseja medir a desigualdade de renda em diferentes países, por exemplo. Nesse sentido, Koenker (2005) faz uma adaptação do coeficiente de Gini utilizando modelos de regressão quantílica, que possibilita a avaliação de alterações no valor do coeficiente provocadas por mudanças na distribuição das variáveis explicativas, e não somente na distribuição da variável resposta.

Finalizando o estudo, utilizaremos a seguir a análise gráfica proposta na Seção 3.3. Conforme foi discutido nessa seção, esse tipo de proposta está relacionado à distribuição Laplace Assimétrica. De forma simplificada, podemos dizer que devemos verificar se alguma distribuição Laplace Assimétrica, a partir dos parâmetros estimados pelo modelo de regressão quantílica, é adequada para explicar a distribuição de renda. Dado que temos a informação que a distribuição de renda é assimétrica à direita, então podemos avaliar distribuições Laplace Assimétrica com parâmetros de assimetria, τ, menores que 0,5. Dessa maneira, iremos considerar somente modelos sem a variável Etnia.

Essa análise gráfica leva em consideração os resíduos definidos em (3.8). Nesse caso, como esses resíduos devem ter distribuição normal quando o modelo está bem ajustado, podemos analisar os gráficos dos resíduos quantílicos em função dos valores ajustados, assim como o histograma dos resíduos, para identificar em qual quantil a normalidade é mais provável. Os resultados podem ser observados nas Figuras 4.15 e 4.16.

Os resultados obtidos sugerem que o modelo de regressão quantílica no quantil 0,05 parece ser mais adequado para explicar a variação de renda em função das variáveis Idade, Sexo, Anos de Estudo e Estado Civil. Porém, essa escolha pode ser discutida, uma vez que ainda existem muitos pontos fora da banda de confiança na Figura 4.15. Por esse motivo, podemos utilizar uma transformação na variável resposta, por exemplo, a transformação logaritmo. Esse tipo de alteração não traz problemas quando utilizamos os modelos de regressão quantílica, conforme apresentado na Seção 2.6.

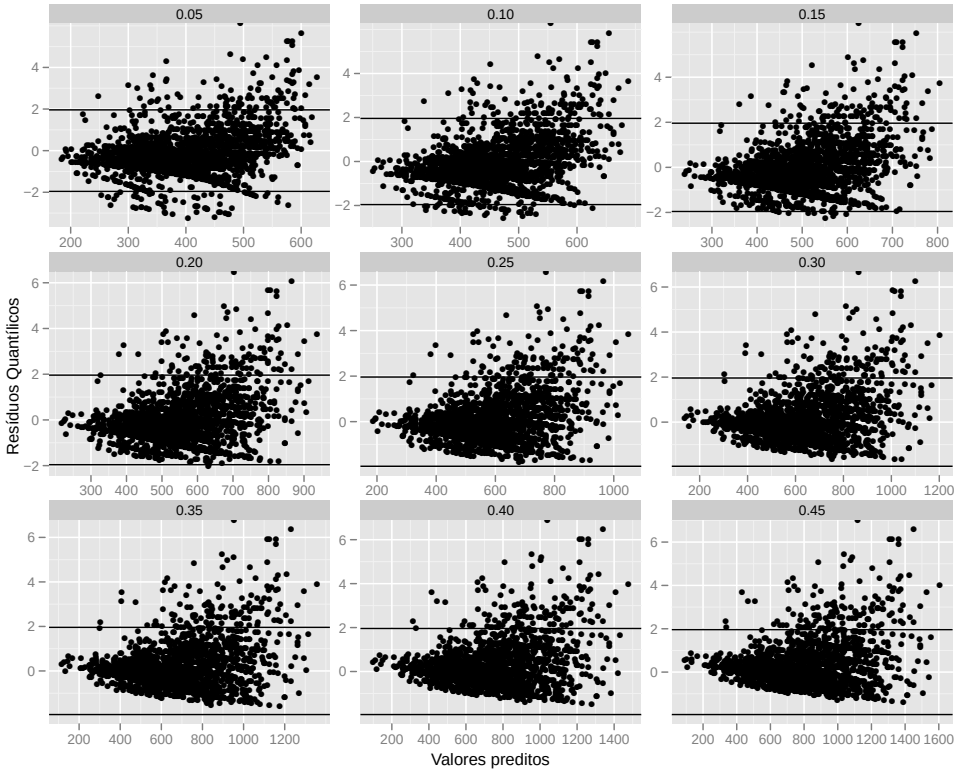

Figura 4.15: *Gráficos dos resíduos quantílicos em função do valor ajustado para os modelos de regressão quantílica ajustados.*

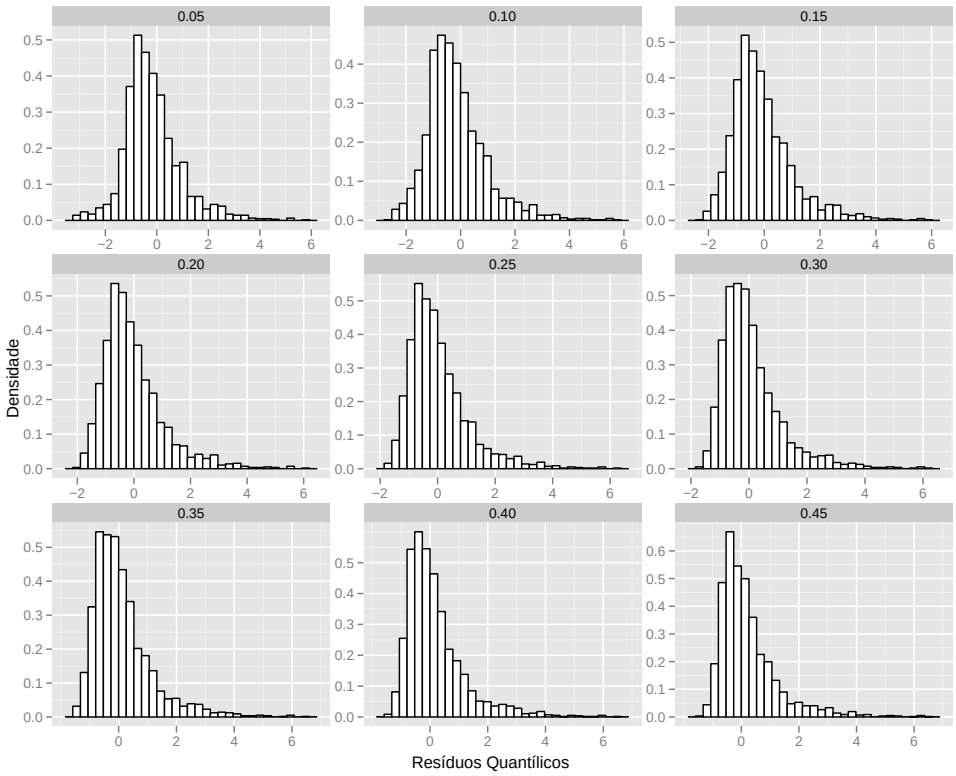

Figura 4.16: *Histograma dos resíduos quantílicos para os modelos de regressão quantílica ajustados.*

Nesse caso, iremos estimar os parâmetros para o modelo

$$\begin{aligned}\log y_i &= \beta_0(\tau) + \beta_1(\tau)\text{Idade}_i + \beta_2(\tau)\text{Sexo}_i + \beta_4(\tau)\text{Solteiro}_i \\ &\quad +\beta_5(\tau)\text{Etnia}_i + \beta_6(\tau)\text{AnosEstudo}_i + u_i.\end{aligned}$$

Esse modelo é muito parecido com o em (4.4), porém sem a variável *dummy* Casado. A suposição sobre os u_i também vale para esse modelo.

O gráfico dos resíduos quantílicos em função dos valores ajustados para o modelo com logaritmo da renda como variável resposta, assim como o envelope, estão nas Figuras 4.17 e 4.18. Foram considerados os quantis 0,1 até 0,9 com 0,1 de diferença entre eles. Após a transformação logaritmo, os quantis mais adequados de acordo com os gráficos gerados, estão entre os valores 0,3 e 0,6. Para uma melhor análise, podemos gerar os gráficos de envelope somente com quantis próximos desses valores. O resultado se encontra na Figura 4.19.

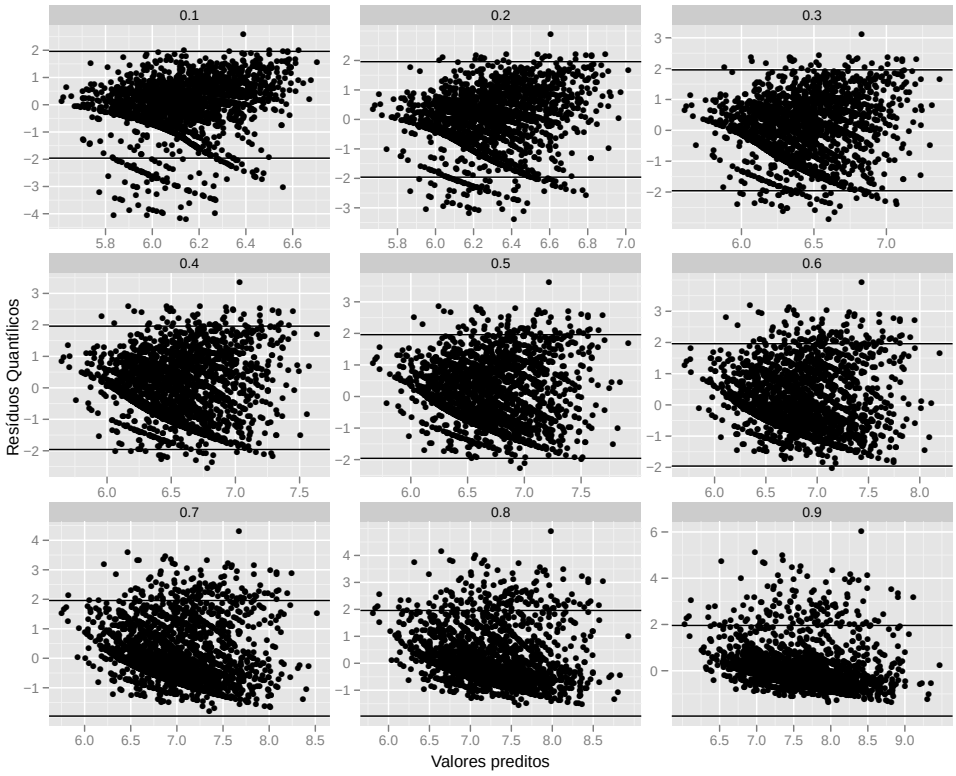

Figura 4.17: *Gráficos dos resíduos quantílicos em função do valor ajustado nos modelos de regressão quantílica com o logaritmo da renda como variável resposta.*

A partir desse novo gráfico, podemos identificar o quantil 0,35 como parâmetro de assimetria da distribuição Laplace Assimétrica que mais se adequa para explicar a distribuição condicional do logaritmo da renda. As estimativas para os parâmetros desse modelo estão apresentadas na Tabela 4.13

Utilizando a propriedade de equivariância dos modelos de regressão quantílica, podemos utilizar a regressão para o logaritmo da renda para estimar o quantil condicional da renda.

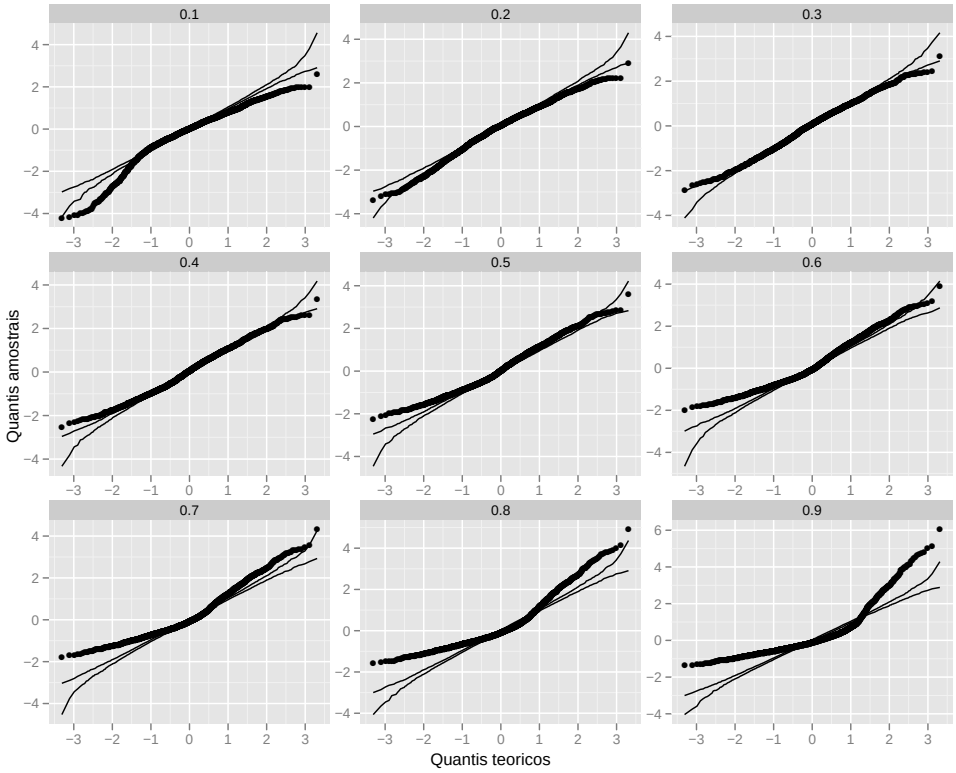

Figura 4.18: *Envelope para os resíduos nos modelos de regressão quantílica com o logaritmo da renda como variável resposta.*

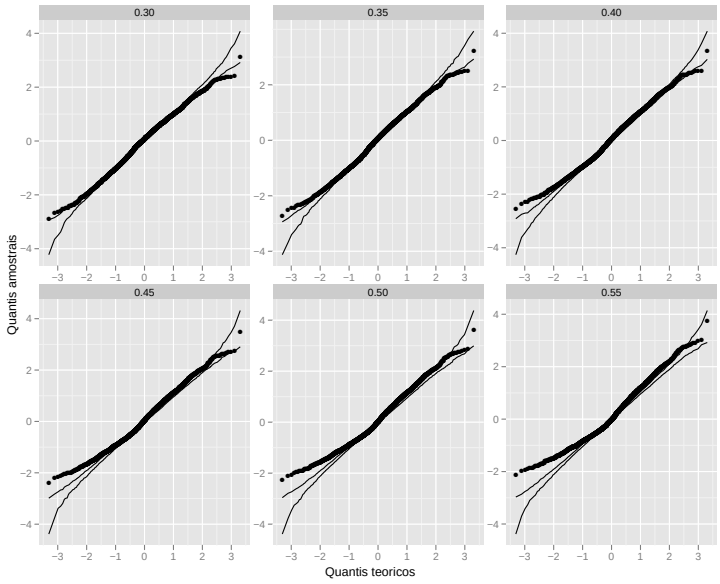

Figura 4.19: *Envelope para os resíduos nos modelos de regressão quantílica com o logaritmo da renda como variável resposta.*

Dessa forma, para pessoas com 30 anos de idade, do sexo masculino, solteiras, de etnia branca e com 10 anos de estudo, de acordo com o modelo ajustado, estimamos o quantil de ordem 0,35 do logaritmo da renda como

	Estimativa	Erro-Padrão	valor-t	valor-p
Intercepto	5,24	0,08	62,08	0,00
Idade	0,02	0,00	10,48	0,00
Sexo	0,27	0,03	8,77	0,00
Solteiro	-0,14	0,03	-4,30	0,00
Etnia	0,07	0,03	2,11	0,04
Anos de Estudo	0,06	0,00	14,83	0,00

Tabela 4.13: *Estimativas para ajuste do modelo de regressão quantílica com $\tau = 0,35$ para o logaritmo da renda como variável resposta.*

$$\begin{aligned} \widehat{Q_{0,35}(\log y_i|x)} &= 5,24 + 0,02 \times 30 + 0,27 - 0,14 + 0,07 + 0,06 \times 10 \\ &= 6,64. \end{aligned}$$

Mas pela propriedade da equivariância, temos que

$$Q_{0,35}(y_i|x) = \exp(Q_{0,35}(\log y_i|x)).$$

Logo, temos que a estimativa do quantil condicional de ordem 0,35 da renda de homens solteiros, de etnia branca, com 10 anos de estudo e 30 anos de idade é

$$\mathrm{e}^{6,64} = 765,10.$$

Finalizando esse capítulo, aplicamos o teste de falta de ajuste apresentado na Seção 3.2. Calculamos as estatísticas de testes e os níveis descritivos para os modelos ajustados do quantil condicional 0,05 da renda e do quantil 0,35 para o logaritmo da renda. Ambos os testes rejeitaram a hipótese nula de linearidade do modelo. Por esse motivo, consideramos que mais variáveis devem ser utilizadas para explicar o quantil condicional da renda. Não foi possível acrescentar diferentes combinações ou mesmo verificar a linearidade para outros quantis devido ao alto tempo de processamento necessário para a execução desse teste para esse conjunto de dados.

Capítulo 5

Conclusões

Nesta dissertação foram estudados os modelos de regressão quantílica. Inicialmente foi feita uma introdução contextualizando a utilização desses modelos em relação à análise de regressão usual. Em seguida, apresentamos as formas de estimação, construção de intervalos de confiança e os testes de hipóteses para os parâmetros. Comparamos os diferentes métodos através de estudos de simulação. Além disso, discutimos propriedades dos modelos de regressão quantílica, como equivariância e robustez que esses modelos apresentam. No terceiro capítulo, formalizamos o coeficiente de explicação para modelos de regressão quantílica enunciado por Koenker e Machado (1999). Na sequência, o teste de falta de ajuste proposto por He e Zhu (2003) foi discutido. E por último no capítulo, a utilização de gráficos para avaliar a qualidade do ajuste em modelos de regressão quantílica foi proposta considerando a distribuição Laplace Assimétrica. No capítulo de aplicação das técnicas discutidas ao longo da dissertação, utilizamos os dados de poluição de cidades norte-americanas e também informações sobre renda e outras variáveis considerando o contexto brasileiro para exemplificar o uso dos modelos de regressão quantílica.

Entendemos que entre as principais contribuições desse trabalho estejam os estudos de simulação para o teste de hipóteses proposto por Chen et al. (2008), pois o autor do artigo sugere a utilização do teste para modelos de regressão quantílica, porém não apresenta os resultados relativos ao uso deste nessa situação. Além disso, ainda não há na literatura a sugestão da utilização de gráficos como os da Seção 3.3 para modelos de regressão quantílica. Outra contribuição desse trabalho que devemos listar é a criação dos programas disponibilizados no Apêndice A, que foram essenciais na discussão das técnicas apresentadas no Capítulo 3.

5.1 Considerações Finais

A análise de regressão é uma das principais técnicas estatísticas utilizadas na análise de dados. Por esse motivo, entendemos que é importante que novas técnicas sejam sempre difundidas para que a análise da relação entre a variável resposta e suas variáveis expli-

cativas seja feita da melhor forma possível. Dessa maneira, tendo identificado a falta de material em português sobre o tema, entendemos que essa dissertação possa ser importante na disseminação do conteúdo relacionado aos modelos de regressão quantílica.

Por fim, acreditamos que seja interessante a discussão e comparação entre os modelos de regressão quantílica com os modelos de regressão usual, que estimam os efeitos das variáveis explicativas na média condicional da variável resposta, por percebermos o potencial dos modelos de regressão quantílica segundo a sua proposta de análise, identificando relações que não seriam possíveis utilizando a abordagem para a média somente. Com isso, sugerimos a utilização dos modelos de regressão quantílica, tendo em vista o objetivo da análise de regressão em questão. Entretanto, não podemos deixar de reconhecer o avanço da análise de regressão usual e a importância dessa em diversos estudos.

5.2 Sugestões para Pesquisas Futuras

Para temas de pesquisas futuras, sugerimos o estudo de novos testes de falta de ajuste para modelos de regressão quantílica e medidas de ajuste do tipo $R^1(\tau)$.

Outras sugestões de estudo são a combinações de estimativas de diferentes quantis para previsão de novos valores, além de metódos de seleção de modelos e de variáveis na análise de regressão quantílica. Outra área que poderia gerar futuros estudos são os modelos de regressão quantílicos bayesianos.

Apêndice A

Programas

Pacotes necessárias para o bom funcionamento das funções criadas

```
library(quantreg)
library(ggplot2)
library(tcltk)
```

Funções relacionadas à distribuição de Laplace assimétrica

```
## Função Densidade da Distribuição Laplace Assimétrica
dalap <- function(x, mu=0, sigma=1, tau=0.5)
{
 saida = vector(length=length(x))
 for (k in 1:length(x))
 {
   parte1 = (tau*(1-tau))/sigma
   parte2 = exp(-((x[k]-mu)/sigma)*(tau-I(x[k]<mu)))
   saida[k] = parte1*parte2
 }
 return(saida)
}
```

```
## Função de distribuição acumulada para a Distribuição de Laplace
    Assimétrica
palap <- function(q, mu=0, sigma=1, tau=0.5)
{
 saida <- vector(length=max(length(q), length(mu)))

 if (length(q)!=length(mu))
```

```
 {
  for (k in 1:length(q))
  {
    if (!q[k]>mu) saida[k]<-tau*exp((1/sigma)*(1-tau)*(q[k]-mu))
    else saida[k]<-1-(1-tau)*exp(-(tau/sigma)*(q[k]-mu))
  }
 }
 else
 {
  for (k in 1:length(q))
  {
    if (!q[k]>mu[k]) saida[k]<-tau*exp((1/sigma)*(1-tau)*(q[k]-mu[k]))
    else saida[k]<-1-(1-tau)*exp(-(tau/sigma)*(q[k]-mu[k]))
  }
 }
 return(saida)
}
```

```
## Função quantílica para a distribuição de Laplace Assimétrica
qalap <- function(p, mu=0, sigma=1, tau=0.5)
{
  if (p > 1) stop("p deve ser menor que 1")
  saida <- vector(length=length(p))
  for (k in 1:length(p))
  {
    if (0<p[k] && p[k]<tau) saida[k] <- mu + (sigma/(1-tau))*log(p[k]/tau)
    else saida[k] <- mu - (sigma/tau)*log((1-p[k])/(1-tau))
  }
  return(saida)
}
```

```
## Função geradora de números aleatórios com distribuição Laplace
    Assimétrica
ralap <- function(n, mu=0, sigma=1, tau=0.5)
{
  saida = vector(length=n)
  if (length(mu)==1) mu = rep(mu, n)
  if (length(mu)!=n) stop("Mu e n têm dimensões diferentes")
  for (k in 1:n)
  {
    u1 = rexp(1)
    u2 = rexp(1)
    saida.padrao <- u1/tau - u2/(1-tau)
    saida[k] = mu[k] + sigma*saida.padrao
  }
  return(saida)
}
```

Funções criadas para calcular o coeficiente de determinação dos modelos

```
## Gráfico que plota o valor de R^1(tau) para um modelo com diversos taus
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x, tau=1:9/10)
> grafR1(modelo)

grafR1 <- function(model.rqs, trueScale=T)
{
  if (class(model.rqs)!="rqs")
  {
    stop("Você deve usar essa função com objetos do tipo rqs")
  }
  taus = model.rqs$tau
  rho.c = model.rqs$rho
  methods = model.rqs$method
  y <- model.rqs$y

  rho.r <- rq(y~1, tau=taus, method=methods)$rho

  R1 <- 1 - rho.c/rho.r

  data.graph <- data.frame(taus, R1)

  saida = list(values=data.graph, variable=paste(model.rqs$call$formula
      [3], "", sep=""))

  if (trueScale) graph <- ggplot(data.graph, aes(x=taus, y=R1)) + ylim(c
      (0,1))
  else graph <- ggplot(data.graph, aes(x=taus, y=R1))
  graph <- graph + geom_line()+ ylab(expression(R^{1}*(tau))) + xlab(
      expression(tau))

  saida.final = list(data=saida, graph=graph)
  return(saida.final)
}
```

```
## Função que deve ser utilizada para comparar a contribuição de
## variáveis em um grupo de quantis.
## Para utilizar, deve fazer (exemplo)
> g1 <- grafR1(modelo1)
> g2 <- grafR1(modelo2)
> graf2R1(c(g1, g2))

graf2R1 <- function(objects, trueScale=T)
{
```

```
  quant.var <- length(objects)
  if (quant.var==1)
  {
    stop("Você precisa de mais variáveis para utilizar essa função")
  }

  objects <- lapply(objects, grafR1)

  taus = unlist(lapply(1:length(objects), function(x) objects[[x]]$data$
      values$taus))
  R1 = unlist(lapply(1:length(objects), function(x) objects[[x]]$data$
      values$R1))
  Variaveis = unlist(lapply(1:length(objects), function(x) rep(objects[[x
      ]]$data$variable, length(objects[[x]]$data$values$taus))))

  dataToUse <- data.frame(taus, R1, Variaveis)

  if (trueScale) graph <- ggplot(dataToUse, aes(x=taus, y=R1, group=
      Variaveis)) + ylim(c(0,1))
  else graph <- ggplot(dataToUse, aes(x=taus, y=R1, group=Variaveis))
  graph + geom_line(aes(linetype=Variaveis)) + ylab(expression(R^{1}*(tau)
      )) + xlab(expression(tau)) + opts(legend.position="bottom", legend.
      direction="horizontal")
}
```

Funções adaptadas para o teste de falta de ajuste

```
## Função auxiliar no cálculo do teste de falta de ajuste
psi.rq <- function(u,tau)  tau - I(u < 0)
```

```
## Função auxiliar no cálculo do teste de falta de ajuste
Rn <- function(residuals, tt, X, tau)
{
  n <- length(residuals)
  valor.final = 0
  for (k in 1:nrow(X))
  {
    valor.final = valor.final + psi.rq(residuals[k], tau) * X[k,] * all(X[
        k,] <= tt)
  }
  valor.final = valor.final/n^0.5

  return(valor.final)
}
```

```
## Função que calcula o valor da estatística de teste para o teste de
    falta de ajuste
## Retirar \ de dentro do loop antes de rodar essa função.
tn.rq <- function(model)
{
  res <- model$res
  z <- model$x
  n <- length(res)
  tau <- model$tau

  A <- matrix(0,ncol(z),ncol(z))
  for (k in 1:nrow(z))
  {
    A <- A + Rn(res,z[k,],z,tau)  \% * \% t(Rn(res, z[k,],z,tau))
  }
  A <- A/n
  tn <- eigen(A)$value[1]

  return(tn)
}
```

```
## Função que calcula o p-valor do teste da estatística de falta de ajuste
    .
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> tn.rqStar(modelo)

tn.rqStar <- function(model, msize=1000)
{
  predictors <- model$call$formula[3]
  tn <- tn.rq(model)

  n <- length(model$res)
  tau <- model$tau
  method <- model$method

  tn.star.values <- vector(length=msize)

  total <- msize
  # create progress bar
  pb <- tkProgressBar(title = "Barra de progresso", min = 0, max = total,
      width = 300)

  for (k in 1:msize)
  {
    setTkProgressBar(pb,k,label=paste(round(k/total*100,3),"%"))
    random.y <- ralap(n, tau=tau)
    model.star <- rq(as.formula(paste("random.y ~ ", model$call$formula
```

```
        [3])), tau=tau, method=method)

    tn.star.values[k] <- tn.rq(model.star)
  }

  p.value <- sum(tn.star.values>=tn)/msize
  list(p.value=p.value, values.tnStar <- tn.star.values)
}
```

```
## Função que calcula o p-valor da estatística de falta considerando o
    processo iterativo proposto por He e Zhu(2003).
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> tn.rqStarIT(modelo)

tn.rqStarIT <- function(model, mTot=500, step=20, alpha=0.05)
{
  result = "Accept Null"

  predictors <- model$call$formula[3]
  tn <- tn.rq(model)

  n <- length(model$res)
  tau <- model$tau
  method <- model$method

  tn2 <- vector()

  for (k in 1:step)
  {
    random.y <- ralap(n, tau=tau)
    model.star <- rq(as.formula(paste("random.y~", model$call$formula[3]))
        , tau=tau, method=method)
    tn2[k] <- tn.rq(model.star)
  }

  p <- sum(tn2>=tn)/length(tn2)

  p1=max(p, .1)
  ep.p = 3*((p1*(1-p1))/length(tn2))^0.5
  check=!any(p+ep.p<alpha, p-ep.p>alpha)

  while (check == T && length(tn2)<=mTot)
  {
    tn.star.values = vector()
    for (k in 1:step)
    {
      random.y <- ralap(n, tau=tau)
```

```
      model.star <- rq(as.formula(paste("random.y ~ ", model$call$formula
          [3])), tau=tau, method=method)

      tn.star.values[k] <- tn.rq(model.star)
    }
    tn2 <- c(tn2, tn.star.values)

    p <- sum(tn2>=tn)/length(tn2)

    p1=max(p, .1)
    ep.p = 3*((p1*(1-p1))/length(tn2))^0.5
    check=!any(p+ep.p<alpha, p-ep.p>alpha)
  }

  if (p < alpha) result="reject null"

  list(p.value=p,decision=result,countIterations=length(tn2))
}
```

```
# Função auxiliar para a utilização de computação paralela
# no cálculo da estatística de teste de falta de ajuste.
simula <- function(x, model)
{
  predictors <- model$call$formula[3]
  n <- length(model$res)
  tau <- model$tau
  method <- model$method

  random.y <- ralap(n, tau=tau)
  model.star <- rq(random.y ~ model$x - 1, tau=tau, method=method)
  return(tn.rq(model.star)$tn)
}
```

```
# Função criada para o cálculo da estatística de teste de falta de ajuste,
    utilizando computação paralela.
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> tn.rqStar.Par(modelo)

tn.rqStar.Par <- function(model, msize=1000, alpha=0.05)
{
  result = "Accept Null"

  predictors <- model$call$formula[3]
  tn <- tn.rq(model)$tn

  tn.star.values <- mclapply(1:msize,
                             simula,
```

```
                                model=model,
                                mc.preschedule = FALSE,
                                mc.set.seed = TRUE,
                                mc.cores = getOption('cores'))

  tn.star.values <- unlist(tn.star.values)

  p.value <- sum(tn.star.values>=tn)/msize

  if (p.value < alpha) result="reject null"

  print(paste("O p-valor para esse teste é igual a ", p.value, ",
      utilizando ", msize, " simulações. "))

  return(list(p.value=p.value, tn=tn, decision=result))
}
```

Função criadas para a análise gráfica dos resíduos

```
## Função utilizada para fazer o gráfico de resíduos quantílicos em função
     dos valores ajustados
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> grafResiduosRQ(modelo)

grafResiduosRQ <- function(model, scales="fixed")
{
  tau <- model$tau
  n <- ifelse(length(tau)==1, length(residuals(model)), nrow(residuals(
      model)))
  preditos <- fitted(model)
  rho.hat <- model$rho/n

  if (length(tau)>1)
  {
    rho.hat <- model$rho/n
    residuos <- list()

    for (k in 1:length(tau))
    {
      residuos[[k]] <- qnorm(palap(q=as.numeric(model$y), mu=preditos[,k],
          sigma=rho.hat[k], tau=tau[k]))
    }

    residuos <- unlist(residuos)
    preditos <- as.vector(fitted(model))
```

```
    tau <- rep(tau, each=n)

    dados <- data.frame(preditos, residuos, tau)

    g <- ggplot(dados, aes(x=preditos,y=residuos, group=tau)) + geom_point
        () + facet_wrap(~tau, ncol=3, scales=scales)
    g + geom_hline(aes(yintercept=qnorm(c(0.025, 0.975)))) + xlab("Valores
        preditos") + ylab("Resíduos Quantílicos")

  }
  else
  {
    residuos <- qnorm(palap(as.numeric(model$y), preditos, rho.hat, tau))

    dados <- data.frame(preditos, residuos)
    g <- ggplot(dados, aes(x=preditos,y=residuos)) + geom_point()
    g + geom_hline(aes(yintercept=qnorm(c(0.025, 0.975)))) + xlab("Valores
        preditos") + ylab("Resíduos Quantílicos")
  }
}
```

```
## Função utilizada para fazer o gráfico Q.Q. dos resíduos quantílicos.
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> qq.ResiduosRQ(modelo)

qq.ResiduosRQ <- function(model, scales, ncolunas)
{
  tau <- model$tau
  rho <- model$rho

  if (length(tau)== 1)
  {
    n <- length(residuals(model))

    sigmahat <- model$rho/n
    predicted <- fitted(model)
    res.quant = qnorm(palap(q=as.numeric(model$y),mu=predicted, sigma=
        model$rho/n, tau=tau))

    theoretical.quant <- qnorm(1:n/(n+1))
    sample.quant <- sort(res.quant)
    db <- data.frame(theoretical.quant, sample.quant, Tau=paste("Tau = ",
        tau, sep=""))

    g <- ggplot(db, aes(x=theoretical.quant, y=sample.quant)) + geom_point
        () + xlab("Quantis teóricos") + ylab("Quantis amostrais")
    graph <- g + facet_wrap(~Tau) + geom_abline(intercept = 0, slope = 1)
```

```
  }

  else
  {
    n <- nrow(residuals(model))

    residuos <- list()
    for(k in 1:length(tau))
    {
      residuos[[k]] <- qnorm(palap(q=as.numeric(model$y), mu=fitted(model)
          [,k], sigma=model$rho[k]/n, tau=tau[k]))
    }

    residuos <- lapply(residuos, sort)
    residuos <- unlist(residuos)
    predicted <- as.vector(fitted(model))

    tau.total <- rep(tau, each=n)

    quantis.teoricos <- vector(length=n*length(tau))
    quantis.teoricos <- qnorm(1:n/(n+1))

    for (j in 2:length(tau))
    {
      quantis.teoricos <- c(quantis.teoricos, qnorm(1:n/(n+1)))
    }

    dados <- data.frame(residuos, quantis.teoricos, tau=tau.total)

    g <- ggplot(dados, aes(x=quantis.teoricos, y=residuos, group=tau)) +
        geom_point() + facet_wrap(~tau, scales=scales, ncol=ncolunas)
    graph <- g + geom_abline(intercept = 0, slope = 1) + xlab("Quantis
        teoricos") + ylab("Quantis amostrais")
  }
  return(graph)
}
```

```
## Função utilizada para fazer o histograma dos resíduos quantílicos do
    modelo de regressão quantílica ajustado
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x)
> hist.ResiduosRQ(modelo)

hist.ResiduosRQ <- function(model, scales, ncolunas)
{
  tau <- model$tau
  rho <- model$rho
```

```
  if (length(tau)== 1)
  {
    n <- length(residuals(model))

    sigmahat <- model$rho/n
    predicted <- fitted(model)
    res.quant = qnorm(palap(q=as.numeric(model$y), mu=predicted, sigma=
        model$rho/n, tau=tau))

    sample.quant <- sort(res.quant)
    db <- data.frame(sample.quant, Tau=paste("Tau = ", tau, sep=""))

    g <- ggplot(db, aes(x=sample.quant, y=..density..)) + geom_histogram()
        + xlab("Resíduos Quantílicos") + ylab("Densidade")
    graph <- g + facet_wrap(~Tau) + geom_histogram(colour = "black", fill
        = "white")
  }

  else
  {
    n <- nrow(residuals(model))

    residuos <- list()
    for(k in 1:length(tau))
    {
      residuos[[k]] <- qnorm(palap(q=as.numeric(model$y), mu=fitted(model)
          [,k], sigma=model$rho[k]/n, tau=tau[k]))
    }

    residuos <- lapply(residuos, sort)
    residuos <- unlist(residuos)
    predicted <- as.vector(fitted(model))

    tau.total <- rep(tau, each=n)

    dados <- data.frame(residuos, tau=tau.total)

    g <- ggplot(dados, aes(x=residuos, y=..density.., group=tau)) + geom_
        histogram()  + facet_wrap(~tau, scales=scales, ncol=ncolunas)
    graph <- g + geom_histogram(colour = "black", fill = "white") + ylab("
        Densidade") + xlab("Resíduos Quantílicos")
  }
  return(graph)
}
```

Função criada para gerar o envelope da regressão de Laplace assimétrica

```
## Função utilizada para gerar o gráfico de envelope para os resíduos
    quantílicos do modelo de regressão quantílica ajustado
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x, data=dados)
> envel.rq(modelo, dados)

envel.rq <- function(model, data, ncolunas=1, scales="fixed")
{
  tau <- model$tau
  rho <- model$rho

  if (length(tau)== 1)
  {
    n <- length(residuals(model))

    sigmahat <- model$rho/n
    predicted <- fitted(model)
    res.quant = qnorm(palap(q=as.numeric(model$y), mu=predicted, sigma=
        model$rho/n, tau=tau))

    e <- matrix(0,n,100)
    e1 <- numeric(n)
    e2 <- numeric(n)
    #
    for(i in 1:100)
    {
       e[,i] <- ralap(n,mu=predicted,sigma=sigmahat, tau=tau)
       sim.model <- rq(as.formula(paste("e[,i] ~ ", model$formula[3])),
           data, tau=tau)

       e[,i] <- qnorm(palap(q=as.numeric(sim.model$y), mu=fitted(sim.model
           ), sigma=sim.model$rho/n, tau=tau))
       e[,i] <- sort(e[,i])
    }
    #
    for(i in 1:n)
    {
       eo <- sort(e[i,])
       e1[i] <- (eo[2]+eo[3])/2
       e2[i] <- (eo[97]+eo[98])/2
    }

    theoretical.quant <- qnorm(1:n/(n+1))
    sample.quant <- sort(res.quant)
```

```
  db <- data.frame(theoretical.quant, sample.quant, Tau=paste("Tau = ",
      tau, sep=""))
  db1 <- data.frame(e1=sort(e1), theoretical.quant)
  db2 <- data.frame(e2=sort(e2), theoretical.quant)

  g <- ggplot(db, aes(x=theoretical.quant, y=sample.quant)) + geom_point
      () + xlab("Quantis teóricos") + ylab("Quantis amostrais")
  graph <- g + geom_line(aes(y=e1), db1) + geom_line(aes(y=e2), db2) +
      facet_wrap(~Tau)
}

else
{
  n <- nrow(residuals(model))

  residuos <- list()
  for(k in 1:length(tau))
  {
    residuos[[k]] <- qnorm(palap(q=as.numeric(model$y), mu=fitted(model)
        [,k], sigma=model$rho[k]/n, tau=tau[k]))
  }

  residuos <- lapply(residuos, sort)
  residuos <- unlist(residuos)
  predicted <- as.vector(fitted(model))

  tau.total <- rep(tau, each=n)

  e <- list()
  e1 <- list(n)
  e2 <- list(n)

  for(k in 1:length(tau))
  {
    e[[k]] <- matrix(0,n,100)
    e1[[k]] <- numeric(n)
    e2[[k]] <- numeric(n)

    for(i in 1:100)
    {
      e[[k]][,i] <- ralap(n,mu=fitted(model)[,k], sigma=model$rho[k]/n,
          tau=tau[k])
      sim.model <- rq(as.formula(paste("e[[k]][,i] ~ ", model$formula
          [3])), data, tau=tau[k])

      e[[k]][,i] <- qnorm(palap(q=as.numeric(sim.model$y), mu=fitted(
          sim.model), sigma=sim.model$rho/n, tau=tau[k]))
      e[[k]][,i] <- sort(e[[k]][,i])
```

```
      }

      for(i in 1:n)
      {
         eo <- sort(e[[k]][i,])
         e1[[k]][i] <- (eo[2]+eo[3])/2
         e2[[k]][i] <- (eo[97]+eo[98])/2
      }
    }

    e1 <- as.numeric(unlist(lapply(e1,sort)))
    e2 <- as.numeric(unlist(lapply(e2,sort)))

    quantis.teoricos <- vector(length=n*length(tau))
    quantis.teoricos <- qnorm(1:n/(n+1))

    for (j in 2:length(tau))
    {
      quantis.teoricos <- c(quantis.teoricos, qnorm(1:n/(n+1)))
    }

    dados <- data.frame(residuos, quantis.teoricos, tau=tau.total, e1, e2)

    g <- ggplot(dados, aes(x=quantis.teoricos,y=residuos, group=tau)) +
        geom_point() + facet_wrap(~tau, scales=scales, ncol=ncolunas)
    graph <- g + geom_line(aes(y=e1, group=tau),dados) + geom_line(aes(y=
        e2, group=tau),dados) + xlab("Quantis teoricos") + ylab("Quantis
        amostrais")
  }
  return(graph)
}
```

Funções criadas para gerar os gráficos com os coeficientes dos modelos de regressão quantílica.

```
## Função criada para fazer o gráfico com os valores dos coeficientes para
     os quantis ajustados com os modelos de regressão quantílica.
## Para utilizar, deve fazer (exemplo)
> modelo <- rq(y ~ x, tau=1:9/10)
> graficoCoeficientes(modelo, se="boot")

graficoCoeficientes <- function(model, level=0.95, se)
{
  if(is.null(se)) se="nid"
```

```
  tau <- model$tau
  if (se=="boot") info<-summary(model, se=se, method="mcmb")
  else info<-summary(model, se=se)

  zalpha <- qnorm(1 - (1 - level)/2)

  if (se!="rank"){
    cf <- lapply(info, coef)
    for (i in 1:length(cf)) {
            cfi <- cf[[i]]
            cfi <- cbind(cfi[, 1], cfi[, 1] - cfi[, 2] * zalpha, cfi[, 1]
               + cfi[, 2] * zalpha)
            colnames(cfi) <- c("coefficients", "lower bd", "upper bd")
            cf[[i]] <- cfi
    }
  }

  else {
    cf <- lapply(info, coef)
    for (i in 1:length(cf)) {
            cfi <- cf[[i]]
            cfi <- cbind(cfi[, 1], cfi[, 2], cfi[, 3])
            colnames(cfi) <- c("coefficients", "lower bd", "upper bd")
            cf[[i]] <- cfi
    }
  }

  lim.inf <- as.numeric(unlist(lapply(cf, function(x) x[,2])))
  lim.sup <- as.numeric(unlist(lapply(cf, function(x) x[,3])))
  est.coef <- as.numeric(unlist(lapply(cf, function(x) x[,1])))

  variaveis <- rownames(coef(info[[1]]))

  dados <- data.frame(Tau=rep(tau,each=length(variaveis)),
                      Variaveis=rep(variaveis, length(tau)),
                      est.coef, lim.sup, lim.inf)

  lapply(variaveis, function(x) {
  graph <- ggplot(dados[dados$Variaveis==x,], aes(x=Tau, y=est.coef)) +
     facet_wrap(~Variaveis, scales="free")
  graph <- graph + xlab(expression(tau)) + ylab("Coeficientes")
  graph + geom_ribbon(aes(ymin=lim.inf, ymax=lim.sup), fill="grey50") +
     geom_point() + geom_line(linetype=2)
         })
}
```

Apêndice B

Dados utilizados na dissertação

Dados de poluição de cidades norte-americanas

Obs	SO2	TEMP	FAB	POP	VENTO	CHUVA	DIASCHUVA
1	10	70,30	213	582	6,00	7,05	36
2	13	61,00	91	132	8,20	48,52	100
3	12	56,70	453	716	8,70	20,66	67
4	17	51,90	454	515	9,00	12,95	86
5	56	49,10	412	158	9,00	43,37	127
6	36	54,00	80	80	9,00	40,25	114
7	29	57,30	434	757	9,30	38,89	111
8	14	68,40	136	529	8,80	54,47	116
9	10	75,50	207	335	9,00	59,80	128
10	24	61,50	368	497	9,10	48,34	115
11	110	50,60	3344	3369	10,40	34,44	122
12	28	52,30	361	746	9,70	38,74	121
13	17	49,00	104	201	11,20	30,85	103
14	8	56,60	125	277	12,70	30,58	82
15	30	55,60	291	593	8,30	43,11	123
16	9	68,30	204	361	8,40	56,77	113
17	47	55,00	625	905	9,60	41,31	111
18	35	49,90	1064	1513	10,10	30,96	129
19	29	43,50	699	744	10,60	25,94	137
20	14	54,50	381	507	10,00	37,00	99
21	56	55,90	775	622	9,50	35,89	105

Tabela B.1: *Dados do primeiro exemplo do Capítulo 1*

Obs	SO2	TEMP	FAB	POP	VENTO	CHUVA	DIASCHUVA
22	14	51,50	181	347	10,90	30,18	98
23	11	56,80	46	244	8,90	7,77	58
24	46	47,60	44	116	8,80	33,36	135
25	11	47,10	391	463	12,40	36,11	166
26	23	54,00	462	453	7,10	39,04	132
27	65	49,70	1007	751	10,90	34,99	155
28	26	51,50	266	540	8,60	37,01	134
29	69	54,60	1692	1950	9,60	39,93	115
30	61	50,40	347	520	9,40	36,22	147
31	94	50,00	343	179	10,60	42,75	125
32	10	61,60	337	624	9,20	49,10	105
33	18	59,40	275	448	7,90	46,00	119
34	9	66,20	641	844	10,90	35,94	78
35	10	68,90	721	1233	10,80	48,19	103
36	28	51,00	137	176	8,70	15,17	89
37	31	59,30	96	308	10,60	44,68	116
38	26	57,80	197	299	7,60	42,59	115
39	29	51,10	379	531	9,40	38,79	164
40	31	55,20	35	71	6,50	40,75	148
41	16	45,70	569	717	11,80	29,07	123

Tabela B.2: *Continuação dos dados do primeiro exemplo do Capítulo 1.*

Dados de Imunoglobulina G em crianças

IgG	Idade	IgG	Idade	IgG	Idade
1,50	0,50	2,10	0,75	2,60	1,08
2,70	0,50	4,20	0,75	5,10	1,17
1,90	0,50	3,80	0,75	4,40	1,17
4,00	0,50	5,70	0,83	3,10	1,17
1,90	0,50	3,00	0,83	5,00	1,17
4,40	0,50	3,20	0,92	1,40	1,17
1,50	0,50	5,10	0,92	6,70	1,17
2,20	0,50	2,10	0,92	5,30	1,17
1,60	0,50	2,30	0,92	1,70	1,17
4,70	0,50	3,40	0,92	6,60	1,17
1,60	0,50	3,90	0,92	2,90	1,17
1,40	0,50	4,30	0,92	6,10	1,25
3,00	0,50	5,30	0,92	4,00	1,25
2,50	0,50	7,20	0,92	5,50	1,25
1,00	0,50	3,80	0,92	4,70	1,25
4,30	0,50	5,60	0,92	6,10	1,25
4,70	0,50	1,50	1,00	4,00	1,25
1,70	0,50	7,00	1,00	7,40	1,33
1,90	0,58	4,60	1,00	4,70	1,33
0,90	0,58	3,70	1,00	3,90	1,33
4,10	0,58	4,50	1,00	4,50	1,33
2,80	0,58	4,50	1,00	5,10	1,42
2,20	0,58	5,00	1,00	3,40	1,42
5,40	0,67	5,50	1,00	3,50	1,42
8,40	0,67	5,50	1,00	3,70	1,50
2,00	0,67	3,20	1,00	5,80	1,50
5,10	0,67	3,20	1,00	4,10	1,50
1,50	0,67	2,20	1,00	9,80	1,50
3,20	0,67	2,30	1,00	2,80	1,50
7,70	0,75	3,80	1,08	5,80	1,58
4,50	0,75	3,50	1,08	7,00	1,58
6,60	0,75	5,80	1,08	3,10	1,58
4,20	0,75	4,00	1,08	4,20	1,67

Tabela B.3: *Dados do segundo exemplo do Capítulo 1.*

IgG	Idade	IgG	Idade	IgG	Idade
5,40	1,67	3,40	2,08	5,30	3,00
5,70	1,67	4,40	2,08	5,60	3,00
4,40	1,67	3,70	2,08	4,50	3,00
5,80	1,75	3,30	2,08	6,00	3,00
4,10	1,75	3,50	2,08	7,30	3,08
4,00	1,75	5,00	2,08	4,50	3,08
5,30	1,75	2,70	2,08	4,70	3,08
5,00	1,83	4,40	2,17	3,90	3,08
6,00	1,83	8,00	2,17	4,00	3,08
5,70	1,83	6,20	2,25	4,80	3,08
7,00	1,83	3,30	2,33	3,30	3,08
2,50	1,83	5,80	2,33	5,70	3,08
6,80	1,83	7,50	2,33	4,30	3,08
6,30	1,92	5,50	2,33	7,00	3,08
5,30	1,92	5,50	2,50	13,40	3,08
4,70	2,00	6,10	2,50	4,00	3,08
7,00	2,00	4,90	2,58	5,80	3,17
4,20	2,00	7,20	2,58	6,30	3,17
5,70	2,00	3,50	2,58	8,80	3,17
3,40	2,00	5,90	2,67	4,80	3,17
6,70	2,00	3,20	2,75	5,30	3,25
4,60	2,00	6,10	2,75	4,60	3,25
5,60	2,00	3,70	2,75	6,90	3,25
1,80	2,00	7,30	2,83	5,70	3,33
3,50	2,00	3,30	2,92	6,50	3,42
4,30	2,00	1,80	2,92	6,30	3,42
4,30	2,00	10,40	2,92	6,80	3,42
5,40	2,00	4,20	2,92	3,90	3,50
4,90	2,00	4,20	3,00	7,80	3,58
5,40	2,00	6,10	3,00	8,00	3,58
5,60	2,00	7,80	3,00	5,40	3,67
3,80	2,08	4,40	3,00	6,20	3,67
7,30	2,08	5,00	3,00	6,10	3,83

Tabela B.4: *Continuação dos dados do segundo exemplo do Capítulo 1.*

IgG	Idade	IgG	Idade	IgG	Idade
3,90	3,83	4,60	4,42	9,30	5,17
6,00	3,83	4,00	4,42	4,40	5,17
3,50	3,83	1,90	4,42	8,70	5,17
4,20	3,83	4,00	4,50	9,80	5,25
3,60	3,83	8,60	4,50	7,10	5,25
4,30	3,83	2,60	4,50	8,10	5,25
5,40	3,92	3,90	4,50	7,90	5,25
5,80	4,00	6,40	4,58	8,40	5,33
7,50	4,00	7,80	4,58	8,20	5,33
7,10	4,00	3,80	4,58	10,40	5,33
6,00	4,00	5,50	4,58	9,70	5,33
3,20	4,08	7,10	4,58	8,10	5,42
6,90	4,08	10,20	4,67	4,80	5,42
7,90	4,08	7,00	4,67	4,90	5,50
3,40	4,08	7,40	4,67	12,50	5,50
9,50	4,17	9,40	4,67	3,80	5,50
3,80	4,17	6,80	4,67	8,80	5,50
8,30	4,17	9,10	4,67	10,40	5,58
6,70	4,17	5,20	4,75	4,70	5,58
7,10	4,25	4,30	4,75	3,30	5,58
7,80	4,25	2,70	4,75	5,60	5,58
7,20	4,25	11,00	4,75	4,60	5,58
6,60	4,33	9,60	4,83	14,40	5,58
2,50	4,33	12,60	4,83	9,10	5,67
2,10	4,33	8,90	4,83	6,30	5,67
4,00	4,33	3,80	4,92	6,10	5,75
3,70	4,33	6,10	5,08	5,60	5,75
5,60	4,33	7,50	5,08	5,40	5,75
5,60	4,33	7,40	5,08	7,00	5,83
4,50	4,42	10,90	5,08	7,60	5,92
5,90	4,42	8,30	5,17	3,10	6,00
6,30	4,42	8,20	5,17	6,80	6,00
4,70	4,42	4,70	5,17	0,00	0,00
8,00	4,42	8,10	5,17	0,00	0,00

Tabela B.5: *Continuação dos dados do segundo exemplo do Capítulo 1.*

Apêndice C

Distribuição Laplace Assimétrica

Seguindo a formulação de Yu e Zhang (2005), dizemos que X tem distribuição Laplace assimétrica se sua função densidade de probabilidade puder ser escrita da seguinte forma:

$$f(x, \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}(\tau - I(x \leq \mu))\right),$$

em que $-\infty < \mu < \infty$ é o parâmetro de localização, $\sigma > 0$ é o parâmetro de escala e $0 < \tau < 1$ é o parâmetro de assimetria. Verifica-se que para $\tau = 0,5$, essa distribuição se reduz à distribuição Laplace, ou exponencial dupla como também é conhecida. Além disso, verifica-se também que para $\tau < 0,5$, a distribuição é assimétrica à direita e para $\tau > 0,5$, a distribuição é assimétrica à esquerda. Na Figura C.1, podemos observar a densidade dessa distribuição para três valores diferentes de τ, quando $\mu = 0$ e $\sigma = 1$.

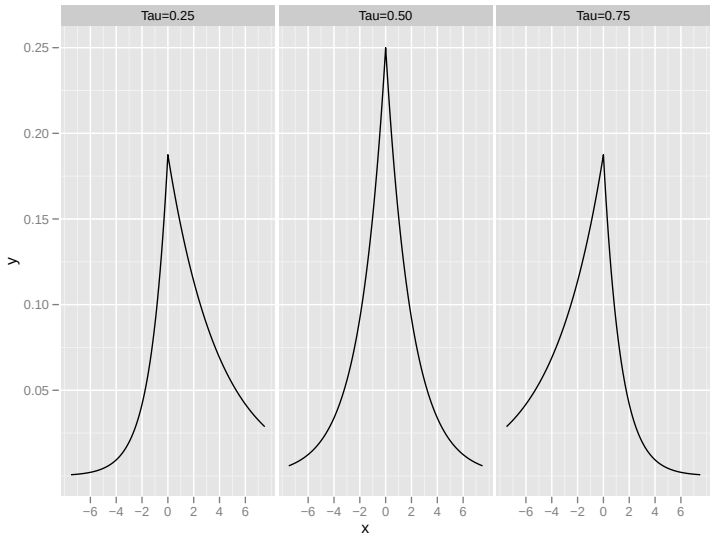

Figura C.1: *Densidade da distribuição Laplace Assimétrica* $\tau = 0,25, 0,50 e 0,75$*,* $\mu = 0$ *e* $\sigma = 1$

Utilizando a notação para a distribuição Laplace Assimétrica como $\text{LA}(\mu, \sigma, \tau)$, é possível mostrar que se $X \sim \text{LA}(0, 1, \tau)$, que é também chamada de distribuição Laplace assimétrica padrão, então $Y = \mu + \sigma X$, tem distribuição $\text{LA}(\mu, \sigma, \tau)$, de forma similar ao que ocorre na distribuição normal.

Outro resultado interessante dessa distribuição, que inclusive remete aos problemas de regressão quantílica, é que se $X \sim \text{LA}(\mu, \sigma, \tau)$, então $P(X < \mu) = \tau$ e, por conseguinte, $P(X > \mu) = 1 - \tau$, ou seja, na distribuição Laplace assimétrica, o parâmetro μ é o quantil de ordem τ da distribuição, assim como a moda da distribuição. Além disso, a esperança e a variância de X ficam definidos da seguinte forma:

$$E(X) = \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)},$$

$$\text{Var}(X) = \frac{\sigma^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}.$$

Para terminar essa pequena introdução da distribuição Laplace Assimétrica, vamos discutir um ponto relacionado diretamente à regressão quantílica. Um resultado conhecido bastante utilizado em análise de regressão é que o estimador de mínimos quadrados do vetor de parâmetros $\boldsymbol{\beta}$, discutido no Capítulo 1, é igual ao estimador de máxima verossimilhança de $\boldsymbol{\beta}$ quando a distribuição dos erros é normal com média 0 e variância constante. Da mesma forma, podemos mostrar que se os erros do modelo têm distribuição Laplace assimétrica, então o estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$ coincide com o estimador da mínima soma dos erros absolutos ponderados da regressão quantílica, da equação (1.4). Basta notar que, supondo o modelo linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{C.1}$$

sendo que ε_i tem distribuição Laplace assimétrica, então a função de verossimilhança para $\boldsymbol{\beta}$, $\boldsymbol{L(\beta)}$, é tal que

$$\boldsymbol{L(\beta)} \propto \exp\left(-\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i'\beta})\right),$$

em que $\rho_\tau(u)$ é a função de perda descrita em (1.2). Como o expoente é negativo, maximizar o valor da função de verossimilhança com relação a $\boldsymbol{\beta}$ é equivalente a minimizar a soma dentro do expoente, que conforme demonstrado em (1.4), gera o estimador de $\boldsymbol{\beta}$ na regressão quantílica.

Referências Bibliográficas

Andre, C., Elian, S., Narula, S., e Tavares, R. (2000), "Coefficients of Determination for variable selection in MSAE regression," *Communications in Statistics - Theory and Methods*, 29, 623–642. Citado na pág. 34, 35

Atkinson, A. (1981), "Two graphical displays for outlying and influential observations in regression," *Biometrika*, 68, 13–20. Citado na pág. 46

Barrodale, I. e Roberts, F. (1973), "An improved Algorithm for Discrete l_1 Linear Approximation," *SIAM Journal on Numerical Analysis*, 10, 839–848. Citado na pág. 15, 16

Bassett, G. e Koenker, R. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, 73, 618–622. Citado na pág. 17

Box, G. e Cox, D. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society. Series B*, 26, 211–252. Citado na pág. 2

Buchinsky, M. (1994), "Changes in US Wage Structure 1963-87: An Application of Quantile Regression," *Econometrica*, 62, 405–458. Citado na pág. 10, 55

Chen, C. (2005), "Growth Charts of Body Mass Index (BMI) with Quantile Regression," in *Proceedings of International Conference on Algorithmic Mathematics and Computer Science*. Citado na pág. 10

Chen, C. e Wei, Y. (2005), "Computational Issues for Quantile Regression," *Sankhiā: Indian Journal of Statistics*, 67, 399–417. Citado na pág. 15, 16

Chen, K., Ying, Z., Zhang, H., e Zhao, L. (2008), "Analysis of Least Absolute Deviation," *Biometrika*, 95, 107–122. Citado na pág. 22, 23, 28, 70

Cox, D. e Snell, E. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society. Series B*, 30, 248–275. Citado na pág. 45

Draper, N. e Smith, H. (1981), *Applied regression analysis*. Citado na pág. 39, 40

Dunn, P. e Smyth, G. (1996), "Randomized Quantile Residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244. Citado na pág. 44, 45

Efron, B. e Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall/CRC. Citado na pág. 18, 19

Gutenbrunner, C. e Jureckova, J. (1992), "Regression Rank Scores and Regression Quantiles," *The Annals of Statistics*, 20, 305–330. Citado na pág. 20, 23

Gutenbrunner, C., Jureckova, J., Koenker, R., e Portnoy, S. (1993), "Tests of Linear Hypotheses Based on Regression Rank Scores," *Journal of Nonparametric Statistics*, 2, 307–331. Citado na pág. 23, 24

Hall, P. e Sheather, S. (1988), "On the Distribution of a Studentized Quantile," *Journal of the Royal Statistical Society. Series B (Methodological)*, 50, 381–391. Citado na pág. 18

Hand, D., Lunn, A., K.J., M., e Ostrowski, E. (1994), *A Handbook of Small Data Sets*, Chapman and Hall. Citado na pág. 7, 49

Hao, L. e Naiman, D. (2007), *Quantile Regression*, Sage Publications. Citado na pág. 3

He, X. e Hu, F. (2002), "Markov Chain Marginal Bootstrap," *Journal of the American Statistical Association*, 97, 783–795. Citado na pág. 19, 20

He, X. e Zhu, L. (2003), "A Lack-of-Fit Test for Quantile Regression," *Journal of the American Statistical Association*, 98, 1013–1022. Citado na pág. 33, 39, 40, 41, 42, 70

Isaacs, D., Altman, D., Tidmarsh, C., Valman, H., e Webster, A. (1983), "Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for IgG, IgA, IgM," *Journal of Clinical Pathology*, 36, 1193–1196. Citado na pág. 9

Kocherginsky, M., He, X., e Mu, Y. (2005), "Practical Confidence Intervals for Regression Quantiles," *Journal of Computational and Graphical Statistics*, 14, 41–55. Citado na pág. 18, 19, 20

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press. Citado na pág. 1, 7, 14, 18, 20, 21, 24, 31, 65

— (2011), *quantreg: Quantile Regression*, r package version 4.62. Citado na pág. 16

Koenker, R. e Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. Citado na pág. 6, 17, 31

Koenker, R. e d'Orey, V. (1987), "Algorithm AS 229: Computing Regression Quantiles," *Journal of the Royal Statistical Society. Series C*, 36, 383–393. Citado na pág. 15

Koenker, R. e Machado, J. (1999), "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310. Citado na pág. 18, 35, 36, 70

Kvalseth, T. (1985), "Cautionary Note about R^2," *The American Statistician*, 39, 279–285. Citado na pág. 34

McKean, J. e Sievers, G. (1987), "Coefficients of Determination for Least Absolute Deviation Analysis," *Statistics and Probability Letters*, 5, 49–54. Citado na pág. 34

Melly, B. (2005), "Public-private wage differentials in Germany: evidence from quantile regression," *Empirical Economics*, 30, 505–520. Citado na pág. 55

Montgomery, D., Peck, E., e Vining, C. (2001), *Introduction to Linear Regression Analysis*, Wiley. Citado na pág. 1

Mosteller, F. e Tukey, J. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley. Citado na pág. 1

Nelder, J. e Wedderburn, W. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A*, 135, 370–384. Citado na pág. 2

Portnoy, S. e Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-error vs Absolute-error Estimators," *Statistical Science*, 12, 279–296. Citado na pág. 15, 16

Rao, C. (1973), *Linear Statistical Inference and its Applications*, Wiley. Citado na pág. 1

Searle, S. (1971), *Linear Models*, Wiley. Citado na pág. 1

Stigler, G. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge University Press. Citado na pág. 2

Stute, W. (1997), "Nonparametric Model Checks for Regression," *The Annals of Statistics*, 25, 613–641. Citado na pág. 40

Yu, K., van Kerm, P., e Zhang, J. (2005), "Bayesian Quantile Regression: An Application to the Wage Distribution in 1990s Britain," *Sankhiā: Indian Journal of Statistics*, 67, 359–377. Citado na pág. 55

Yu, K. e Zhang, J. (2005), "A Three-Parameter Asymmetric Laplace Distribution and Its Extension," *Communications in Statistics - Theory and Methods*, 34, 1867–1879. Citado na pág. 41, 92