Bruno de Oliveira Alves

Análise de Modelos de Regressão Quantílica e um estudo sobre suas diferenças em relação ao Modelo de Regressão Linear Clássico

Niterói - RJ, Brasil

26 de Julho de 2022

Universidade Federal Fluminense

Bruno de Oliveira Alves

Análise de Modelos de Regressão Quantílica e um estudo sobre suas diferenças em relação ao Modelo de Regressão Linear Clássico

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof[a]. Dra. Karina Yuriko Yaginuma

Niterói - RJ, Brasil

26 de Julho de 2022

Universidade Federal Fluminense

Bruno de Oliveira Alves

Análise de Modelos de Regressão Quantílica e um estudo sobre suas diferenças em relação ao Modelo de Regressão Linear Clássico

Monografia de Projeto Final de Graduação sob o título *"Análise de Modelos de Regressão Quantílica e um estudo sobre suas diferenças em relação ao Modelo de Regressão Linear Clássico"*, defendida por Bruno de Oliveira Alves e aprovada em 26 de Julho de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profª. Dra. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Profª. Dra. Patrícia Lusié Velozo da Costa
Departamento de Estatística – UFF

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Niterói, 26 de Julho de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

A474a Alves, Bruno de Oliveira
Análise de Modelos de Regressão Quantílica e um estudo sobre suas diferenças em relação ao Modelo de Regressão Linear Clássico / Bruno de Oliveira Alves ; Karina Yuriko Yaginuma, orientadora. Niterói, 2022.
72 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Regressão quantílica. 2. Regressão linear. 3. Método dos mínimos erros absolutos ponderados. 4. Inferência estatística. 5. Produção intelectual. I. Yaginuma, Karina Yuriko, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Tendo em vista que análise de regressão é uma das técnicas mais usadas para medir a relação de uma variável com uma ou um conjunto de variáveis, essa análise é sensível a presença de *outliers*, este estudo visa apresentar a regressão quantílica, que é um modelo de regressão mais robusta a presença de *outliers*, que decorre do fato de ser utilizado o método da minimização dos erros absolutos ponderados para estimar os parâmetros do modelo, assim é possível estimar retas para qualquer um dos quantis amostrais dos dados, além do fato de permitir que os resíduos não sejam normalmente distribuídos. Realizou-se então um estudo sobre como estimar os parâmetros, os intervalos de confiança para os parâmetros e medidas para analisar a qualidade do ajuste do modelo de regressão quantílica. Diante disso, foi realizado um estudo usando a renda dos brasileiros comparando a regressão linear clássica e a regressão quantílica, onde a constatação foi de que a regressão quantílica se mostra mais qualificada do que a regressão linear, para analisarmos variáveis que tem uma distribuição assimétrica das suas observações, um segundo estudo foi realizado, a partir de dados sobre as chances de admissão há programas de mestrado de estudantes indianos, onde utilizamos de análises gráficas para verificarmos como os resíduos estavam distribuídos.

Palavras-chave: Regressão quantílica. Regressão linear. Método dos mínimos erros absolutos ponderados. Inferência estatística. Medidas de qualidade de ajuste.

Agradecimentos

Eu agradeço, em primeiro lugar, aos meus pais, Euzilene Franco de Oliveira e Roberto Carlos Santana, e aos meu avós, Jomar Franco de Oliveira e Veríssimo de Oliveira, pelo apoio incondicional em todos os momentos, não só da graduação mas da minha vida.

A minha Orientadora Profª. Dra. Karina Yuriko Yaginuma, pela orientação e suporte, sem ela este trabalho jamais seria possível.

Ao Prof. Dr. Jony Arrais, por sempre incentivar a curiosidade aos seus alunos, além de ter me orientado no projeto de Monitoria da disciplina Métodos Computacionais para est. II.

Ao Prof. Dr. Rafael Santos Erbisti, por ter ministrado as disciplinas de Modelos Linares I e Modelos linares Generalizados, além disso, ter aceitado fazer parte da banca de avaliação deste projeto, assim como, a Profª. Dra. Patrícia Lusié Velozo da Costa, que também me orientou na Monitoria da disciplina Métodos Computacionais para est. II neste último período da graduação.

Aos meus colegas que de graduação, companheiros de estudo e que me ajudaram muito durante todo o processo, Ricardo Junqueira, Lucas Moura, Robson Arthur e Matheus Felipe, além da Monitora Thais de Almeida, sem a monitoria dela em Probabilidade II, Não estaria escrevendo está monografia agora.

E, por último, a coordenação e o departamento de Estatística da UFF, por toda a estrutura necessária para o aprendizado que resultou neste trabalho.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Neste capítulo serão apresentadas a motivação para essa monografia, as principais diferenças do modelo de regressão quantílica para o modelo de regressão linear clássico, os objetivos principais e a forma como este trabalho foi organizado.

1.1 Motivação

A análise de regressão é uma das técnicas amplamente utilizadas para analisar dados multifatoriais. Seu amplo apelo e utilidade resultam de um processo conceitualmente lógico de usar uma equação para expressar a relação entre uma variável de interesse e um conjunto de variáveis preditoras relacionadas (MONTGOMERY; PECK; VINING, 2021). Este tipo de análise tem grande utilidade em estudos estatísticos, por causa de sua simplicidade em interpretar os resultados obtidos, além do grande número de programas estatísticos que podem fazer a análise dos modelos de regressão nos tempos atuais.

Normalmente usamos o método da minimização dos quadrados dos erros para estimar os parâmetros do modelo de regressão, esta técnica gera uma curva de regressão que resume as médias das distribuições correspondentes as variáveis preditoras, porém neste trabalho iremos focar no método da minimização dos erros absolutos ponderados, tal técnica resulta nos modelos de regressão quantílica, o estudo desse tipo de modelo é o principal foco deste trabalho.

Ainda que o método dos mínimos quadrados seja o mais usual, ele possui algumas limitações como, a forte associação do método com a distribuição normal dos erros, ou seja, os erros tem que estar distribuídos de forma simétrica, quando isso não ocorre a performance deste método na estimação dos parâmetros não é adequada.

Outra questão relevante para o detrimento desse método é a influência de *outliers*, dados que se diferenciam drasticamente de todos os outros e que podem causar anomalias nos resultados e nas estimativas dos parâmetros do modelo. Isso exige uma avaliação

criteriosa de quanto cada um desses pontos influencia no ajuste do modelo aos dados, já que tanto *outliers* na variável resposta quanto nas preditoras podem atrapalhar na identificação correta da relação entre as variáveis de interesse, além disso, tal avaliação pode exigir um grande trabalho, fazendo com que não seja interessante usar esse método nesses casos.

Ao contrário do método dos mínimos quadrados dos erros, a minimização dos erros absolutos é robusta à presença de *outliers* na variável resposta, além disso, esse método se mostra mais eficiente quando a distribuição dos erros não é uma normal, ele também descreve melhor uma posição central ao estimar o valor da mediana da distribuição condicional da variável resposta, assim como também é possível produzir estimativas para qualquer quantil condicional da distribuição.

Apesar das vantagens já descritas acima do método da minimização dos erros absolutos sobre a minimização dos quadrados dos erros, o primeiro foi preterido por muito tempo devido ao seu alto custo computacional em relação ao dos mínimos quadrados, apenas depois do avanço dos computadores e da utilização da programação linear é que esse método começou a crescer e ser mais utilizado em análises de regressão.

Um modelo de regressão usual ajustado pelo método dos mínimos quadrados ordinários resume toda informação das variáveis independentes observadas às suas médias. A média é uma informação resumida e incompleta de uma distribuição, da mesma forma que a regressão é uma visão limitada de um conjunto de distribuições. Uma possibilidade bem mais completa seria ajustar diversas curvas de regressão referentes a diversos quantis da distribuição (KOENKER, 2005).

A principal motivação para este trabalho é que os modelos de regressão quantílica são capazes de gerar as diversas curvas de regressão referentes a diversos quantis, sendo assim, este tipo de modelo de regressão é a escolha ideal quando buscamos uma visão mais ampla da relação entre a variável resposta com suas preditoras. O principal método de estimação para os parâmetros do modelo de regressão quantílica é o método da minimização dos erros absolutos citado anteriormente, porém quando estamos interessados em estimar diversos quantis, deve-se fazer uma uma ponderação na minimização dos erros.

1.2 Objetivos

Nesta seção vamos apresentar os objetivos desta monografia. O objetivo principal é estudar o modelo de regressão quantílica e as vantagens de usar tal modelo de regressão, sendo os objetivos específicos:

- Apresentar os modelos de regressão quantílica, assim como o método para a estimação de seus parâmetros.
- Explorar as estimativas de diversos quantis amostrais, a fim de buscar uma análise de regressão mais ampla da relação das variáveis de interesse.
- Estudar as principais diferenças do modelo de regressão quantílica em relação a regressão linear clássica (quantil vs média) em determinadas situações, ou seja, estudar quando uma é melhor do que a outra.

1.3 Organização

Além do Capítulo de Introdução, esta monografia está estruturada da seguinte maneira: no Capítulo 2 serão apresentados os métodos utilizados para estimar os parâmetros do modelo de regressão quantílica, os intervalos de confiança e medidas para avaliar a qualidade do ajuste do modelo. Em seguida, no Capítulo 3 será apresentada a análise dos resultados dos dois estudos de caso realizados e por fim, no Capítulo 4 serão apresentadas as principais conclusões obtidas.

2 Materiais e Métodos

Neste capítulo vamos apresentar os conceitos necessários para estudar os modelos de regressão quantílica. Primeiro, vamos apresentar a definição de quantis e uma ideia prévia da regressão quantílica, na Seção 2.1, iremos apresentar os conceitos para estimação de parâmetros na Seção 2.2, assim como intervalos de confiança para os mesmos na seção 2.3. Em seguida, vamos discutir sobre qualidade do ajuste do modelo de regressão quantílica na Seção 2.4 e por último vamos comentar sobre as bases de dados que serão usadas nos nosso estudos práticos.

2.1 Regressão Quantílica

Os quantis de uma população são pontos estabelecidos em intervalos regulares, a partir da função de distribuição acumulada denotada aqui por $F(X)$,de uma variável aleatória X. Então utilizando a função inversa da distribuição acumulada no ponto τ, define-se que:

$$F^{-1}(\tau) = \inf\{X : F(x) \geq \tau\} \tag{2.1}$$

é chamado de τ-ésimo quantil da variável aleatória X. Em particular, temos que a mediana corresponde a: $F^{-1}(1/2)$. O primeiro quantil e o terceiro correspondem a $F^{-1}(1/4)$ e $F^{-1}(3/4)$, respectivamente.

Os quantis amostrais podem ser definidos da seguinte forma: o quantil de ordem τ de uma amostra é o valor m tal que $100\tau\%$ dos valores amostrais são inferiores a ele, com $0 < \tau < 1$ (SANTOS, 2012).

Na otimização matemática, um função de perda é uma função que mapeia um evento ou valores de uma ou mais variáveis em um número real, ela representa intuitivamente algum"custo" associado a o evento. Um problema de otimização busca minimizar uma função de perda (WALD, 1950). Na estatística, normalmente uma função de perda é usada para estimativa de parâmetro, e o evento em questão é alguma função da diferença

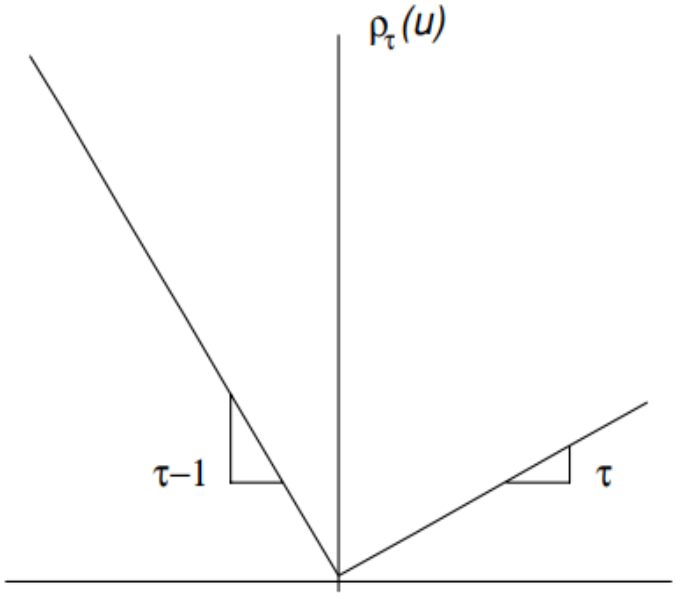
entre os valores estimados e verdadeiros.

Segundo Koenker (2005), os quantis surgem de uma problema de otimização simples que é fundamental para entendermos os modelos de regressão quantílica. Considere um problema teórico de decisão simples, onde desejamos encontrar uma estimativa pontual para uma variável aleatória com função de distribuição acumulada F. Considere a função de perda descrita pela função linear

$$\rho_\tau(u) = u(\tau - I(u < 0)), \tag{2.2}$$

para algum $\tau \in (0, 1)$, em que I é a função indicadora.

Figura 1: Função de regressão quantílica para $\rho(u)$



Fonte: Koenker (2005)

Suponha que queremos encontrar $\hat{x}$ que minimiza a perda esperada, onde $\hat{x}$ é um preditor da variável X. Então, temos que minimizar a seguinte expressão

$$E[\rho_\tau(X - \hat{x})] = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{2.3}$$

Diferenciando a Equação (2.3) em relação a $\hat{x}$ e depois igualando a zero, temos que

$$0 = (1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{2.4}$$

Como a função de distribuição acumulada F é monótona, qualquer elemento de $\{x : F(x) = \tau\}$ minimiza a perda esperada, ou seja, $\hat{x} = F^{-1}(\tau)$ quando temos o caso de

uma única solução. No caso contrário, vamos ter um intervalo τ de quantis, onde devemos escolher o menor elemento desse intervalo, para aderir a convenção de que a função quantílica empírica seja contínua à esquerda (KOENKER, 2005).

Em geral um bom estimador para a perda linear assimétrica deve nos conduzir aos quantis. No caso simétrico o valor de perda produz o quantil da mediana, ou seja, $\tau = 1/2$, nesse caso em particular temos que,

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

Assim, temos que $E[\rho_{1/2}(X - \hat{x})]$ é equivalente a minimização $E[|X - \hat{x}|]$, onde $\hat{x}$ resulta no valor da mediana, a prova desse resultado é encontrada em Hao e Naiman (2007).

2.1.1 Uma introdução à regressão quantílica

O método de estimação mais geral para os quantis de uma variável de interesse (ou resposta), pode ser obtido com base em um problema de otimização simples, onde os quantis são expressos como soluções desse problema. A partir disso pode-se fazer um paralelo com o método dos mínimos quadrados, pois essa técnica nos mostrará o caminho para o desenvolvimento dos modelos de regressão quantílica. Sabe-se que dada uma amostra com n observações de uma variável aleatória Y com média μ temos que, a média amostral é a solução para o seguinte problema de minimização

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{2.5}$$

Então, dado X uma matriz de constantes conhecidas com tamanho $n \times p$, o modelo de regressão linear que expressa a média condicional de Y dado $X = x$, como uma função linear nos parâmetros β, deve-se realizar a seguinte transformação, $\mu = x^T\beta$, assim iremos obter o estimador $\hat{\beta}$ através do método dos mínimos quadrado resolvendo a seguinte expressão

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - {x_i}^T\beta)^2, \tag{2.6}$$

onde x_i^T é a i-ésima da matriz X.

Então, a partir do que foi desenvolvido em Koenker e Bassett (1978), se tivermos interesse em especificar o quantil condicional de ordem τ de Y dado $X = x$ como uma função linear, $Q_\tau(Y|X) = {x_i}^T\beta(\tau)$, em que $\beta(\tau)$ é um vetor de parâmetros do modelo, estimado a partir do $\hat{\beta}(\tau)$ que é obtido minimizando a seguinte soma dos erros ponderados

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - {x_i}^T\beta). \tag{2.7}$$

Assim, no caso do método dos mínimos quadrados, a relação linear entre as duas variáveis pode ser escrita como

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \tag{2.8}$$

em que ε_i tem média 0, logo a média condicional de $Y|x$ pode ser escrita como

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

Então, como temos o interesse em estudar os diferentes quantis da distribuição condicional da variável resposta Y, temos que, valem as seguintes relações lineares

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_{i1} + ... + \beta_p(\tau)x_{ip} + u_i(\tau), \tag{2.9}$$

em que u_i são variáveis aleatórias $i.i.d$ (independentes e identicamente distribuídas) com quantil de ordem τ igual a zero, então pode-se escrever o quantil condicional de ordem τ de $Y|x$ como

$$Q_\tau(Y|x) = \beta_0(\tau) + \beta_1(\tau)x_1 + ... + \beta_p(\tau)x_p. \tag{2.10}$$

Deve-se destacar o fato de que o vetor de parâmetros β estar indexado a τ, isso ocorre pois, um dos interesses nesse caso é estudar se esse vetor assume valores diferentes para τ's diferentes.

Uma outra informação importante é que, diferentemente da análise de regressão linear clássica, os modelos de regressão quantílica tem uma característica bem peculiar, que é o fato de poderem ser construídas diversas curvas para interpretar um mesmo conjunto de variáveis explicativas ou não. Assim como, esse tipo de modelo pode ser usado para analisar um ponto específico da distribuição condicional da variável resposta.

Sem perder a generalidade, pode-se considerar o quantil condicional de $Y|X$ da

Equação (2.10) com apenas uma variável explicativa, ou seja

$$Q_\tau(Y|x) = \alpha(\tau) + \beta(\tau)x. \tag{2.11}$$

Assim, se os valores dos coeficientes estimados para $\beta(\tau)$ são muito próximos entre si, para valores diferentes de τ, ou seja, variando em torno de uma constante, temos evidências de que os erros são $i.i.d.$ Entretanto, se os valores dos coeficientes estimados variam em função de τ, significa os erros podem apresentar alguma forma de heterocedasticidade Koenker (2005) exibe o comportamento das curvas estimadas pela regressão quantílica, através de exemplos usando dados bivariados para comparar o comportamento dos erros nos dois casos. Então, a conclusão foi de que os modelos de regressão quantílica são capazes de incorporar uma possível forma de heterocedasticidade nos erros e que, a mesma é detectada a partir da variação das estimativas dos coeficientes $\beta(\tau)$ para diferentes $\tau's$.

Antes de seguirmos para a próxima seção, vamos fazer um pequeno exemplo para elucidar as informações apresentadas até aqui nesse capítulo, além de mostrar as diferenças entre a utilização da média condicional e da regressão quantílica.

2.1.2 Exemplos

Poluição do ar de cidades norte-americanas

Para comparar as duas metodologias, vamos utilizar os dados da poluição do ar medida em 41 cidades norte-americanas entre os anos de 1969 e 1971 para ajustar os modelos. Os dados foram retirados de Hand et al. (1993). Um estudo similar usando esses mesmos dados foi apresentado em Santos (2012).

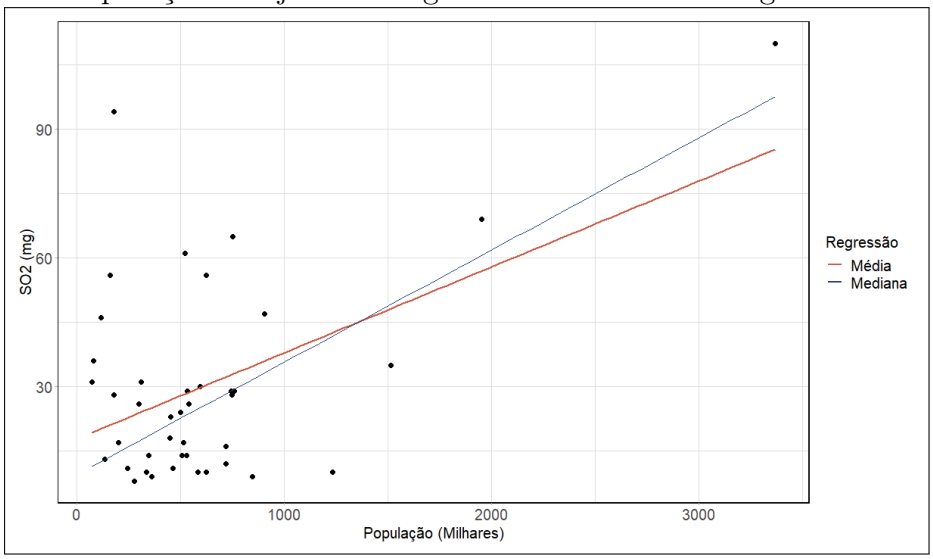
Para ajustar o modelo, vamos utilizar as variáveis quantidade de dióxido de enxofre em miligramas por metro cúbico ($SO2$) e o tamanho da população (em 1970) em milhares. Neste exemplo, estamos interessados em observar o efeito do tamanho da população na poluição do ar da cidade medida pela quantidade de dióxido de enxofre presente por metro cúbico ($SO2$), ou seja, nesse estudo $SO2$ é a nossa variável Y e a População é a nossa variável explicativa X.

Vamos ajustar dois modelos para essa análise, um utilizando a regressão linear, ou seja, estimando o parâmetro do modelo pelo o método dos mínimos quadrados e usando a media condicional das variáveis de interesse e outro modelo usando a regressão quantílica, onde o β é estimado através da Equação (2.7).

Na Figura 2 pode-se observar as duas retas estimadas, onde a reta verde representa a

estimação da média e a reta azul o valor estimado da mediana.

Figura 2: Comparação do ajuste da regressão da média e da regressão da mediana.



Na Tabela 1 observa-se que as estimativas dos parâmetros de ambos os modelos estimados. Analisando os resultados observa-se que a inclinação das duas retas difere uma da outra, em que a reta da regressão da mediana tem um efeito positivo maior que a da regressão da média, ou seja, segundo regressão da mediana que a cada aumento de mil habitantes na população das cidades, estima-se que há um aumento mediano de 26,1 miligramas de dióxido de enxofre na atmosfera, enquanto na regressão da média, o aumento é em média de 20 miligramas de dióxido de enxofre na atmosfera. Além disso, podemos citar o fato de a estimativa da média foi mais influenciada pelas observações da cidade de Providence, que tem níveis altos de dióxido de enxofre no ar, apesar de não ter uma grande população. Também foi verificado que todas as estimativas são significantes ao nível de 5%.

Tabela 1: Estimativa dos parâmetros dos modelos ajustados a partir da Regressão Linear (Média) e da Regressão Quantílica (Mediana).

Parâmetros	Média	Mediana
Intercepto	17,868	9,549
População	0,0200	0,0261

Além da estimativa para a mediana, se pode utilizar os modelos de regressão quantílica também para fornecer uma visão mais abrangente da relação entre a variável $SO2$ e a População, tal abordagem pode ser observada na Figura 3.

Nota-se que as inclinações das retas são diferentes, o que significa que o efeito do tamanho da população é diferente em alguns pontos da distribuição da variável resposta

Figura 3: Diversos ajustes da regressão quantílica para os valores de τ= 0,05; 0,25; 0,50; 0,75; 0,95.

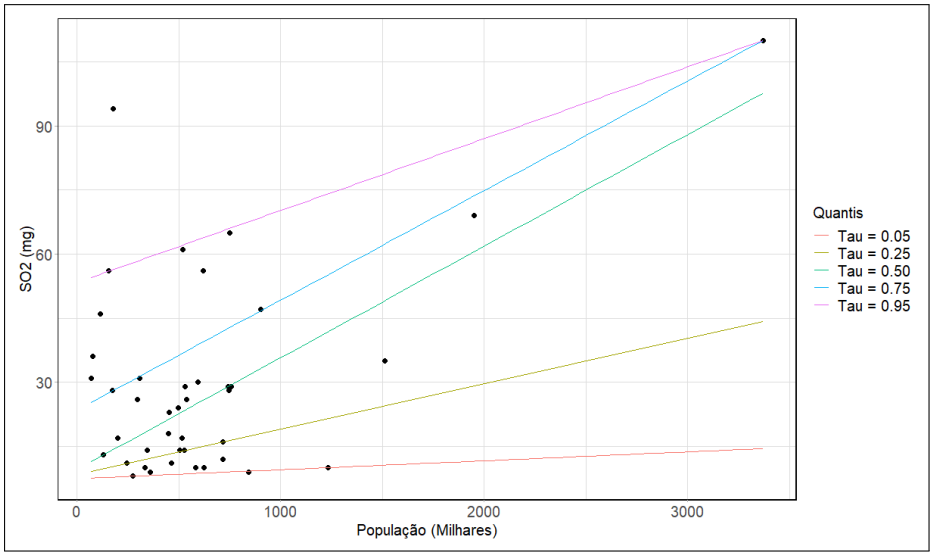


Tabela 2: Estimativa dos parâmetros do modelo de regressão quantílica para os valores de τ= 0,05; 0,25; 0,50; 0,75; 0,95.

	Quantis1(τ)				
Parâmetros	0,05	0,25	0,50	0,75	0,95
Intercepto	7,4205	8,3723	9,5490	23,4801	53,3429
População	0,0021	0,0106	0,0261	0,0257	0,0168

(SO2). Neste caso, o uso da regressão quantílica apresentou resultados interessantes, já que as retas estimadas apresentam inclinações diferentes, o que mostra que o efeito da população é diferente em pontos distintos da distribuição condicional de $SO2$.

2.2 Estimação dos parâmetros nos modelos de regressão quantílica

Antes de começarmos a comentar sobre a estimação dos parâmetros do modelo de regressão quantílica, vamos apresentar a formulação do modelo que será usada para realizar os procedimentos de estimação deste trabalho, que é o seguinte modelo linear

$$Q_\tau(Y) = X\beta(\tau) + \epsilon(\tau), \tag{2.12}$$

onde $Y = (Y_1, ..., Y_n)^T$ é o vetor de variáveis resposta com dimensão $n \times 1$, X é a matriz de planejamento ou design do modelo de dimensões $n \times p$, em que $X_i = (X_{i1}, ..., X_{ip})^T$ é o vetor com as p variáveis explicativas associadas à i-ésima observação da Matriz X, $\beta(\tau) = (\beta_1(\tau), ..., \beta_p(\tau))^T$ é um vetor com $p \times 1$ parâmetros desconhecidos e $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$

é um vetor de erros independentes e identicamente distribuídos que possuem função de distribuição F e quantil de ordem τ igual a zero.

Como mencionado no Capítulo 1, uma das grande vantagens dos modelos de regressão linear é a forma do estimador de mínimos quadrados para o vetor de parâmetros β, que quando a matriz X é de posto completo e os erros ϵ são homocedásticos então, o estimador de mínimos quadrados de β, $\hat{\beta}$ pode ser escrito como

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{2.13}$$

Apesar da estrutura da covariância dos erros, quando os mesmos são heterocedásticos, ser mais complicada, o estimador $\hat{\beta}$ para o vetor de parâmetros β ainda pode ser definido de forma fechada, pelo método dos mínimos quadrados. Infelizmente, não pode-se dizer o mesmo para os modelos de regressão quantílica. Já que seu estimador é obtido através da minimização do erros ponderados, e como foi comentado no Capítulo 1 e nas seções anteriores deste presente capítulo, não temos a possibilidade de obter um estimador calculado de forma analítica.

Esse é o motivo pelo qual os estudos baseados em modelos de regressão L_1 como o de Barrodale e Roberts (1973), onde a regressão L_1 pode ser vista como um caso particular dos modelos de regressão quantílica, pois é o caso quando o quantil de interesse τ é igual à 0,5, ou seja, quando estamos interessados em calcular a mediana, não obtiveram muito sucesso inicialmente, devido principalmente à complexidade computacional para estimar os parâmetros desses modelos. Essa situação só foi alterada com a descoberta de que o problema de minimização dos erros absolutos poderia ser escrito como um problema de programação linear, assim acontecendo os primeiros avanços nos modelos de regressão L_1.

Considerando o modelo dado pela Equação (2.12), sabe-se que para obter o estimador $\hat{\beta}(\tau)$ devemos calcular a minimização dos erros absolutos ponderados, apresentado na Equação (2.7), entretanto, o estimador $\hat{\beta}(\tau)$ não pode ser obtido de uma forma analítica, então, reformulando o problema apresentando anteriormente e transformando o mesmo em um problema de programação linear como é mostrado em Koenker (2005). No nosso caso, o problema de programação linear equivalente ao método de minimização dos erros absolutos ponderados é

$$\min_{(\beta,u,\nu)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \{\tau 1_n^T u + (1-\tau) 1_n^T \nu\}, \tag{2.14}$$

sujeito à condição $Y = X\beta + u - \nu$, onde 1_n^T é um vetor de valores iguais a 1 com dimensão

$1 \times n$, u e ν são os erros positivos e negativos, respectivamente e tem dimensão igual à $n \times 1$, sendo eles definidos como

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

A utilização de programação linear em modelos de regressão L_1 foi iniciada através do algoritmo de Barrodale e Roberts (1973), pois ele foi um dos primeiros que se mostrou realmente eficientes para estimar os parâmetros do modelo L_1. A sua implementação é uma adaptação do algoritmo *simplex* para o problema de minimização dos desvios absolutos. O algoritmo *simplex* permite encontrar valores ideais em situações em que diversos aspectos precisam ser respeitados, quando temos um problema, estabelecemos inequações que representam restrições para as variáveis. A partir daí, testa-se as possibilidades de maneira que otimizem o resultado da forma mais rápida possível. O algoritmo de Barrodale e Roberts (1973) torna-se computacionalmente exigente em banco de dados com muitas observações, mas possui performance razoável para bancos de dados com até 5000 observações e 50 variáveis Chen e Wei (2005). Este algoritmo também foi adaptado para o problema da regressão quantílica, essa adaptação é encontrada em Koenker e d'Orey (1987).

Portnoy e Koenker (1997) sugeriram um procedimento mais eficiente para bancos de dados mais extensos, onde eles utilizam o algoritmo de programação linear conhecido como *ponto interior*, este método segue um caminho através do interior do conjunto de soluções viáveis para o problema. Segundo Chen e Wei (2005), esta técnica mostrou performance superior ao algoritmo *simplex*. Assim, tornando a utilização desse algoritmo mais recomendada em caso de bancos de dados com muitas observações.

Como já comentado, a preferência pela utilização do método dos mínimos quadrados nas análises de regressão, tem como um dos ponto principais o fato do método ser computacionalmente rápido e fácil, já que devido aos avanços da capacidade de processamento dos computadores, as operações de multiplicar matrizes e a inversão do resultado dessa multiplicação, a fim de obter-se uma estimativa para o vetor de parâmetros do modelo, é possível ser considerada como uma tarefa trivial (SANTOS, 2012). Segundo Santos (2012) com relação à ordem de complexidade computacional, o algoritmo do método dos mínimos quadrados requer $O(np^2)$ operações, enquanto que, a complexidade do algoritmo da regressão quantílica é $O(n^{5/2}p^3)$, o que coloca esse método em uma desvantagem computacional se comparado ao método dos mínimos quadrados. Devido a essa complexidade maior Portnoy e Koenker (1997) propuseram uma modificação no algoritmo do *ponto*

interior para a regressão quantílica, que foi adicionar um passo de pré-processamento no algoritmo e através dessa mudança, foi obtido uma melhora na performance do algoritmo, onde em algumas situações, performances semelhantes às do método dos mínimos quadrados foram obtidas.

Comparando o algoritmo *simplex* com o algoritmo do ponto interior, pode-se dizer que o *simplex* é mais estável, já que uma das suas características é o fato de sempre se encontrar uma solução para o problema, já o segundo apresentar dificuldades se as variáveis explicativas possuírem *outliers*. Entretanto, esse algoritmo é muito mais rápido para problemas com muitas observação, mas poucas variáveis respostas.

O pacote `quantreg` do software `R`, que foi desenvolvido por Koenker et al. (2018) é a principal referência na implementação dos algoritmos para estimação dos parâmetros dos modelos de regressão quantílica. Nesse pacote, para utilizarmos o algoritmo do método *simplex* de Barrodale e Roberts (1973) adaptado para a estimação dos parâmetros do modelo de regressão quantílica, deve-se adicionar dentro da função `rq` ou `rq.fit` o argumento *method* = "br". Para o método do *ponto interior* de Portnoy e Koenker (1997), deve-se usar o argumento *method* = "fn" ou se o interesse for em utilizar o algoritmo com a etapa de pré-processamento, o que melhorar o desempenho do algoritmo, o argumento deve ser *method* = "pfn".

Depois de ver como é feita a estimação pontual dos parâmetros de um modelo de regressão quantílica, vamos discutir na próxima seção sobre os intervalos de confiança para os parâmetros, assim como, o problema da inferência sobre os parâmetros do modelo.

2.3 Intervalos de confiança para os parâmetros

Para realizar a construção dos intervalos de confiança dos parâmetros do modelo de regressão quantílica, vamos apresentar dois métodos que são utilizados normalmente, o método baseado em resultados assintóticos e o método *boostrap*. O primeiro método vai nos permitir avaliar os casos em que os erros sejam independentes e identicamente distribuídos(*iid*) ou independente e não identicamente distribuídos (*inid*), já o método *boostrap* torna-se interessante devido ao fato de não precisarmos fazer nenhuma suposição sobre os erros

2.3.1 Método baseado em resultados assintóticos

Para construir inicialmente intervalos de confiança para os parâmetros da regressão L_1, Bassett e Koenker (1978) obtiveram uma distribuição assintótica do vetor de estimadores dos parâmetros do modelo, usando o estimador da mínima soma dos erros absolutos ponderados. Como dito anteriormente, a regressão L_1 é um caso particular da regressão quantílica. Assim, Koenker e Bassett (1978) generalizaram o resultado proposto para a regressão quantílica, segue abaixo o teorema para o vetor de estimadores dos parâmetros do modelo de regressão quantílica. Note que, o estimador $\hat{\beta}(\tau_i)$ é a solução do problema de minimização da soma dos erros absolutos ponderados definida na Equação em (2.7), para $\tau = \tau_i$, dada uma amostra de n observações.

Teorema 2.1 *Seja* $\{\hat{\beta}(\tau_1), \hat{\beta}(\tau_2), \ldots, \hat{\beta}(\tau_m)\}$, *com* $0 < \tau_1 < \tau_2 < \ldots < \tau_m < 1$, *uma sequência de estimadores para os parâmetros do modelo dado em (2.12). 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* $1, 2, \ldots, m$.

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

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

Nessas condições,

$$\sqrt{n}(\hat{\beta}(\tau_1) - \beta(\tau_1), \ldots, \hat{\beta}(\tau_m) - \beta(\tau_m)) \xrightarrow{D} N_{m\times p}(0, V(\tau_1, \ldots, \tau_m)), \tag{2.15}$$

em que $V(\tau_1, \ldots, \tau_m)$ *é a matriz de covariâncias.*

No caso particular em que estamos interessados em um quantil τ específico, segundo Kocherginsky, He e Mu (2005), para o caso em que os erros são *inid* a matriz de covariâncias assintóticas de $\hat{\beta(\tau)}$ é dada por

$$V(\tau) = \tau(1-\tau)(X'FX)^{-1}(X'X)(X'FX)^{-1}, \tag{2.16}$$

em que F é igual a matriz diagonal $f_1(0), \ldots, f_n(0)$ e f_j, $j = 1, \ldots, n$ é a função de densidade dos erros.

No caso dos erros *iid* temos que $f_1(x) = \cdots = f_n(x) = f(x)$, então a matriz de covariâncias assintóticas de $\hat{\beta(\tau)}$ será

$$V(\tau) = \frac{\tau(1-\tau)}{f^2(0)}\tau(1-\tau)(X'X)^{-1}. \tag{2.17}$$

Com a estimação da matriz de covariâncias $V(\tau)$, é possível construir os intervalos de confiança para cada termo do vetor de parâmetros $\beta(\tau)$ utilizando os resultados do Teorema 2.1

2.3.2 Método *Bootstrap*

O método *Bootstrap* é muito utilizado para inferir sobre os parâmetros do modelo de regressão usando um tipo de reamostragem. Efron e Tibshirani (1994) discutem como usar esse método para estimar a matriz de covariâncias do vetor $\hat{\beta}$ do modelo. Na regressão quantílica, foi sugerida uma forma de usar o método *Bootstrap* por Koenker (2005), essa maneira consiste em selecionar pares de observações (Y_i,x_i) com probabilidade $1/n$ em que n é o tamanho da amostra, assim construímos um novo vetor com valores da variável resposta Y^* e uma nova matriz de planejamento X^*. Esse procedimento é repetido R vezes, onde é calculado um vetor $\hat{\beta}^*(\tau)$ em cada repetição feita. Então, com essas R estimativas que foram geradas para o vetor de parâmetros do modelo, estimamos o erro padrão de $\hat{\beta}(\tau)$ a partir do erro padrão observado nas R amostras.

A grande vantagem desse método está no fato de não precisarmos de nenhuma suposição sobre os erros. Entretanto, como nesse método temos a necessidade de ajustar o modelo de regressão quantílica em cada reamostragem gerada, ocorre que, nos casos em que temos muitas variáveis explicativas e muitas observações, esse método tende a se tornar demorado. Efron e Tibshirani (1994) sugere que o número de reamostragens a serem utilizadas nesse processo sejam de 50, 200 ou 1000 reamostragens.

De modo a buscar uma solução para o problema do alto custo computacional do método *Bootstrap* na inferência sobre os parâmetros do modelo de regressão quantílica. He e Hu (2002) desenvolveram um novo método denominado Markov Chain Marginal Bootstrap (MCMB), que foi adaptado para os modelos de regressão quantílica por Kocherginsky, He e Mu (2005). Santos (2012) apresenta através de simulações, que esse novo método além de construir os intervalos muito mais rápido, ele apresenta um desempenho muito próximo ao método anterior nas estimativas geradas, ou seja, ele é mais indicado para amostras grandes.

2.4 Análise da qualidade do ajuste do Modelo de Regressão Quantílica

Nos modelos de regressão linear com erros distribuídos normalmente, o coeficiente de determinação do modelo, comumente chamado de R^2, é bastante usado como uma medida de qualidade de ajuste desses modelos de regressão, já que o R^2 é possível ser interpretado como o percentual da variabilidade da variável resposta explicada pelas variáveis explicativas, entretanto, para os modelos de regressão quantílica essa estatística assim como outras que serão discutidas nesta monografia, vão ser consideradas apenas como medidas-resumo do modelo ajustado. Esse coeficiente no modelo de regressão linear pode ser obtido calculando a seguinte expressão matemática

$$R^2 = 1 - \frac{SQE}{SQT}, \tag{2.18}$$

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

Nos modelos de regressão L_1 foi inicialmente apresentado por McKean e Sievers (1987) e André et al. (2000) uma estatística da classe R^2, denominada por Santos (2012) como R. Essa estatística deve possuir algumas propriedades interessantes, propriedades essas que são apresentadas em Santos (2012) e são as seguintes:

(P1) R deve estar ligada diretamente ao critério de ajuste, uma vez que há a possibilidade dessa medida 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) $0 \leq R \leq 1$, com 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.

A seguinte estatística é apresentada por McKean e Sievers (1987) como coeficiente de determinação, sendo denominada por eles como R_2, essa estatística satisfaz as propriedades listadas acima e ela é definida como

$$R_2 = \frac{RAD}{RAD + (n - p - 1)(\hat{\sigma}/2)}, \tag{2.19}$$

sendo RAD igual a soma absoluta dos resíduos do Modelo Reduzido $Y = \alpha + \epsilon$ e p é a quantidade variáveis explicativas (ou preditoras) do modelo de interesse. Então, considerando a soma dos erros absolutos do modelo com p variáveis explicativas sendo igual a

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

em que, $\hat{y}_i = x_i^T \hat{\beta}$, já a soma dos erros absolutos do Modelo Reduzido, que é o modelo somente com o intercepto, $\hat{y}_i = \hat{\alpha}$ é definida como

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

onde $\hat{\alpha}$ é a mediana da variável resposta Y, então o RAD pode ser reescrito como,

$$RAD = SAE(\hat{\alpha}) - SAE(\hat{\beta}).$$

Por fim, temos que $\hat{\sigma}$ é o estimador para o parâmetro σ, que na regressão é igual a

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

onde $f(.)$ é a função de densidade dos erros.

André et al. (2000) sugeriram uma nova estatística argumentando com um contra-exemplo, que o coeficiente R_2 não satisfazia a propriedade **(P5)**, entretanto, essa nova estatística só difere da anterior somente pelo estimador de σ, onde $\hat{\sigma}$ é obtido calculando a média da soma dos erros absolutos do modelo, ou seja,

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

Para os modelos de regressão quantílica, o primeiro coeficiente de determinação foi feito por Koenker e Machado (1999). A seguir vamos mostrar como calcular essa medida de qualidade de ajuste para os modelos de regressão quantílica.

Considere um modelo de regressão quantílica, com p variáveis explicativas,

$$Q(Y_i|x) = x_{i1}^T\beta_1(\tau) + x_{i2}^T\beta_2(\tau), \tag{2.20}$$

em que $x_{i.}$, i-ésima linha da matriz X de planejamento e essa matriz é particionada em duas partes denominadas x_{i1} e x_{i2} de dimensões iguais a $p-q$ e q, respectivamente. O vetor de parâmetros $\beta(\tau)$ deve ser particionado de forma semelhante a matriz X. Esse modelo definido acima será denominado como Modelo Completo, onde o estimador que minimiza a soma dos erros absolutos ponderados deste modelo é o $\hat{\beta}(\tau)$.

Então, vamos definir o modelo de regressão quantílica reduzido como

$$Q(Y_i|x) = x_{i1}^T\beta_1(\tau) \tag{2.21}$$

onde x_{i1} e $\beta_1(\tau)$ são os mesmo já definidos acima. O estimador que minimiza a soma dos erros absolutos ponderados no Modelo Reduzido é o $\tilde{\beta}(\tau)$

Assim, considerando a soma dos erros absolutos ponderados do Modelo Completo, inicialmente, como,

$$\hat{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - {x_i}^T\hat{\beta}(\tau)), \tag{2.22}$$

e, a do Modelo Reduzido,

$$\tilde{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - {x_i}^T\tilde{\beta}(\tau)), \tag{2.23}$$

então, o coeficiente de determinação para regressão quantílica, $R^1(\tau)$, do modelo dado em (2.20) com relação ao Modelo Reduzido é definido da seguinte forma

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}. \tag{2.24}$$

Se considerarmos no vetor de parâmetros $\beta_2(\tau)$ 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 (SANTOS, 2012).

Entretanto, diferente do R^2, que mede o sucesso relativo de dois modelos de regressão linear em função da variância residual, segundo Koenker e Machado (1999), o $R^1(\tau)$ mede o relativo sucesso de correspondentes modelos de regressão quantílica em um específico quantil em função de uma apropriada soma de resíduos absolutos ponderados, ou seja, o

$R^1(\tau)$ constitui uma medida local de qualidade de ajuste do modelo de regressão quantílica para um quantil em particular.

Podemos dizer ainda, sobre o $R^1(\tau)$, que se o Modelo Completo for "melhor" que o Modelo Reduzido, então $\hat{V}(\tau)$ deve ser significativamente menor que $\tilde{V}(\tau)$. Nesse caso, "melhor" deve ser entendido no sentido que, a inclusão das covariáveis x_2 no modelo para o quantil condicional de ordem τ altera de forma significativa o ajuste do modelo.

Como $\tilde{\beta}(\tau)$ é obtido restringindo $\hat{\beta}(\tau)$, segue-se que $\hat{V}(\tau) \leq \tilde{V}(\tau)$, então o $R^1(\tau)$ se encontra dentro do intervalo $[0, 1]$, satisfazendo assim a propriedade **(P4)**. Esse fato ocorre de forma similar para o coeficiente R^2

Em relação as outras propriedade **(P1)**-**(P6)** que foram listadas no início dessa seção, temos que, a estatística $R^1(\tau)$ não é robusta à presença de *outliers* no modelo, ou seja, ela não satisfaz a propriedade **(P6)** (SANTOS, 2012), porém como essa estatística satisfaz as outras propriedades, essa estatística é a mais indicada para avaliar a qualidade do ajuste dos modelos de regressão quantílica.

Além disso, vamos utilizar uma métrica bem conhecida para comparar os modelos estimados, o Critério de Informação de Akaike (AIC), apresentado no Apêndice B.

2.5 Dados

Nesta Seção vamos comentar sobre os dados que serão utilizados nos estudos aplicados desta monografia, onde todas as análises serão realizadas no software `RStudio`[1] e considerando um nível de significância para as análises inferenciais é igual a $\alpha = 0,05$. Como os modelos de regressão linear clássica devem cumprir algumas suposições em relação aos erros, vamos utilizar testes estatísticos para verificar os pressupostos de normalidade e homocedasticidade dos erros. Para verificar a normalidade o teste usado será o de Kolmogorov-Smirnov, enquanto para testar a homocedasticidade o teste usado será Breusch-Pagan que são apresentados brevemente no Apêndice A.

2.5.1 PNAD Contínua

Inicialmente, vamos utilizar dados obtidos a partir da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADc), referentes ao terceiro trimestre de 2020. Essa pesquisa é realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em uma amostra

[1] https://www.rstudio.com

de domicílios brasileiros, essa pesquisa investiga várias características socioeconômicas da população brasileira [2] para selecionarmos os dados necessários para realizar as análises. Para o nosso estudo, foram selecionados dados sobre a renda do Brasil referentes ao ano de 2020. Assim, foram escolhidos indivíduos que possuem idade entre 14 e 80 anos e que recebiam pelo menos um terço do salário mínimo vigente naquele ano, que era igual a $R\$348,33$, com isso ficamos inicialmente com 123.568 observações. Onde as 8 variáveis da nossa base são as seguintes:

- **Renda Mensal:** Variável numérica que assume apenas valores positivos, onde temos o rendimento mensal habitual em Reais de todos os trabalhos para pessoas de 14 anos ou mais de idade; (Valores inteiros)
- **Idade:** Variável numérica que assume apenas valores inteiros, com a idade completa em anos do indivíduo (0 a 130);
- **Sexo:** Variável categórica que indica o sexo do indivíduo (1-Homem e 2-Mulher);
- **Cor ou raça:** Variável categórica que indica a cor ou raça do indivíduo (1-Branca, 2-Preta, 3-Amarela, 4-Parda, 5-Indígena e 9-Ignorado);
- **Escolaridade:** Variável categórica que indica o nível de instrução mais elevado alcançado pelo indivíduo (1-Sem instrução e menos de 1 anos de estudo, 2-Fundamental incompleto ou equivalente, 3-Fundamental completo ou equivalente, 4-Médio incompleto ou equivalente, Médio completo ou equivalente, Superior incompleto ou equivalente e Superior completo);
- **Região do domicílio:** Variável categórica que indica a região onde está localizada a região do domicílio da pessoa (1-Urbana e 2-Rural);
- **Número de pessoas no domicílio:** Variável numérica inteira, que indica o número de componentes no domicílio, excluindo as pessoas cuja condição no domicílio era pensionista, empregado doméstico ou parente do empregado doméstico (1 a 30);
- **Número de Horas trabalhadas por semana:** Variável numérica, que indica o número de horas habitualmente trabalhadas por semana em todos os trabalhos do indivíduo (1 a 120).

[2] Utilizamos o pacote para `linguagem R`, `PNADcIBGE` desenvolvido pelo próprio *IBGE*

2.5.2 Chance de admissão

Para o segundo estudo, vamos utilizar dados sobre as chances de admissão de um estudante indiano em um programa de Mestrado, esse conjunto de dados foi inspirado em bancos de dados de Graduandos da UCLA (Universidade da Califórnia em Los Angeles). Essa base foi desenvolvida para ajudar os universitários, dando à eles uma ideia sobre as chances dos mesmo ingressar em um programa de mestrado, em uma universidade de interesse. A base foi criada por Acharya, Armaan e Antony (2019), em seu trabalho, podemos encontrar mais detalhes sobre a mesma, além disso, ela pode ser adquirida no site do Kaggle.

A base de dados possui 500 observações e 9 variáveis, sendo a primeira variável o número do indivíduo, ou seja, é apenas uma variável que indica o número da observação, então, não vamos usar ela para a análise de regressão. A variável de interesse ou resposta, da nossa base é a Chance de Admissão, enquanto que as variáveis que usaremos para explicar a chance serão: o exame GRE, o Teste de inglês (TOEFL), a classificação da universidade, a declaração de objetivos (SOP), a carta de recomendação (LOR), a média de pontos das notas (GPA) e a experiência em Pesquisa. A seguir vamos explicar melhor o que cada variável significa:

- **Chance de Admissão:** Variável numérica, com as chances de cada um dos 500 estudantes conseguirem ser admitidos em algum programa de mestrado, onde esta variável pode assumir valores no intervalo $[0, 1]$;
- **GRE:** Variável numérica que contém a pontuação de cada um dos estudantes, num exame padronizado que mede as habilidades consideradas importantes para o sucesso em programas de pós-graduação: raciocínio verbal, raciocínio lógico, pensamento crítico e redação analítica (0 a 340);
- **Teste de inglês (TOEFL):** Variável numérica que contém a pontuação dos estudantes no teste de inglês, visando avaliar o nível do estudante em falar e entender o inglês em um nível acadêmico (0 a 120);
- **Classificação da Universidade :** Nível da classificação da Universidade onde de cada estudante se graduou (1 ao 5);
- **Declaração de objetivos (SOP):** Variável numérica que contém a pontuação dos estudantes, numa categoria de dissertação, onde o estudante explica ao comitê de admissão por que ele deve ser escolhido para o programa, mostrando suas qualificações

e disponibilidade para se comprometer com o campo de estudo ou de pesquisa (1 a 5);

- **Carta de recomendação (LOR):** Variável numérica onde se encontra o valor dado a carta de recomendação de cada um dos indivíduos, essa carta é fornecida por um membro do corpo docente da instituição de ensino onde o estudante se graduou, destacando as qualidades positivas do candidato e as possíveis contribuições dele (1 a 5);

- **Média de pontos das notas (GPA):** Variável numérica que contém a média de pontos das notas de cada um dos estudantes, esse valor representa uma visão geral de como foi o desempenho de cada estudante ao longo da sua graduação (0 a 10);

- **Pesquisa:** Variável categórica que indica se o aluno realizou alguma pesquisa durante a graduação (0-Não e 1-Sim).

3 Análise dos Resultados

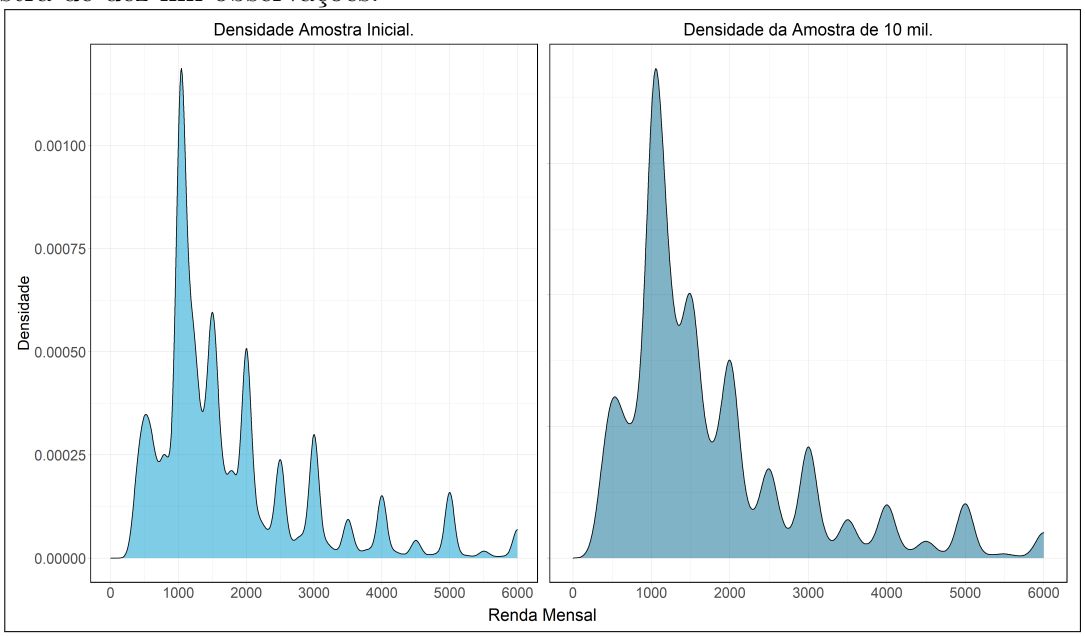
No primeiro capítulo desta presente monografia, foi apresentada uma motivação inicial para o uso dos modelos de regressão quantílica. No segundo capítulo, apresentamos uma breve definição sobre os quantis, além de uma introdução aos modelos de regressão quantílica, os métodos de estimação dos parâmetros do modelo de regressão quantílico e algumas medidas para avaliar a qualidade do ajuste desses modelos. Nesse capítulo, vamos aplicar esses resultados vistos nos capítulos anteriores em duas bases de dados, além de, comparar o modelo de regressão quantílica ao modelo de regressão linear clássico ajustados aos dados.

3.1 Dados sobre a Renda Mensal no Brasil

A utilização de modelos de regressão quantílica é comumente utilizada para explicar a relação da renda com outras variáveis socioeconômicas. Para facilitar as estimativas dos modelos, devido ao alto custo computacional do método de estimação dos modelos de regressão quantílica, resolvemos retirar uma amostra com dez mil observações da nossa amostra inicial. Assim, ficamos com uma base de dados de dez mil observações, com 8 variáveis, sendo uma delas a variável renda, que é nossa variável resposta, além das variáveis socioeconômicas já definidas na Subseção 2.5.1. Para mostrar que nossa amostra retirada representa os dados, na Figura 4, pode-se observar que a densidade da Renda é similar em ambas as amostras. Já na Tabela 5, temos dois boxplots da variável renda em cada uma das amostra , nota-se uma igualdade nos valores da variável renda nas amostras até o terceiro quartil pelo menos, a linha vermelha em ambos os gráficos representa a média da renda mensal em cada amostra, sendo a média igual a $R\$2432,00$ reais na amostra de 10 mil observações e na inicial o valor médio foi de $R\$2426,00$ reais, outro fato a se destacar, é que o número de *outliers* na amostra inicial é bem maior, do que o da amostra de 10 mil.

Além disso, observando as variáveis explicativas também notamos que elas possuem

Figura 4: Densidade da variável Renda, em reais, na amostra inicial (Completa) e na amostra de dez mil observações.



comportamentos próximos quando comparadas em ambas as amostras, como mostrado nas Figuras 6, 7 e 8.

Figura 5: Gráfico de distribuição da renda em reais, na amostra inicial (Completa) e na amostra de dez mil observações.

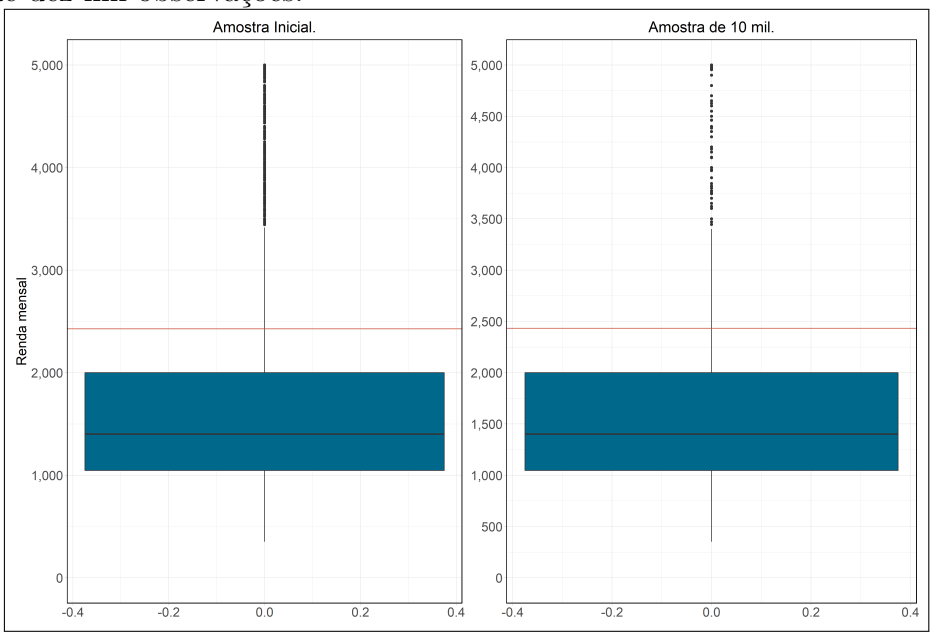
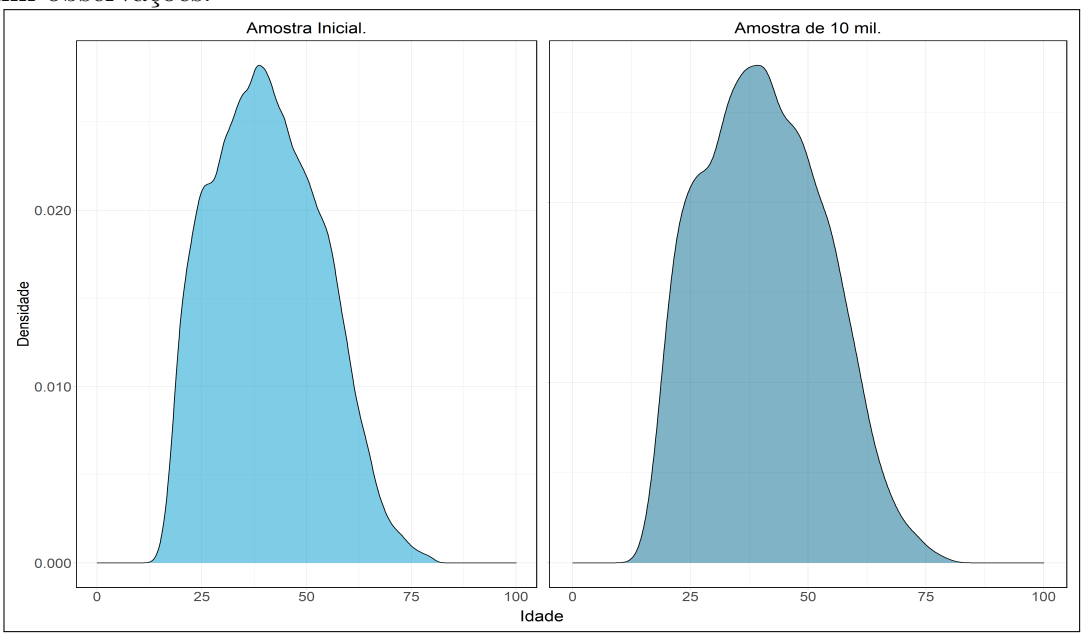


Figura 6: Densidade da variável Idade, na amostra inicial (Completa) e na amostra de dez mil observações.



Logo, com a nossa amostra de 10 mil observações, conseguimos ter uma boa representatividade dos dados totais, assim conseguindo utilizar os modelos de regressão quantílica e regressão linear para modelar a renda do Brasil, em 2020.

Figura 7: Densidade da variável Horas trabalhadas por semana, na amostra inicial (Completa) e na amostra de dez mil observações.

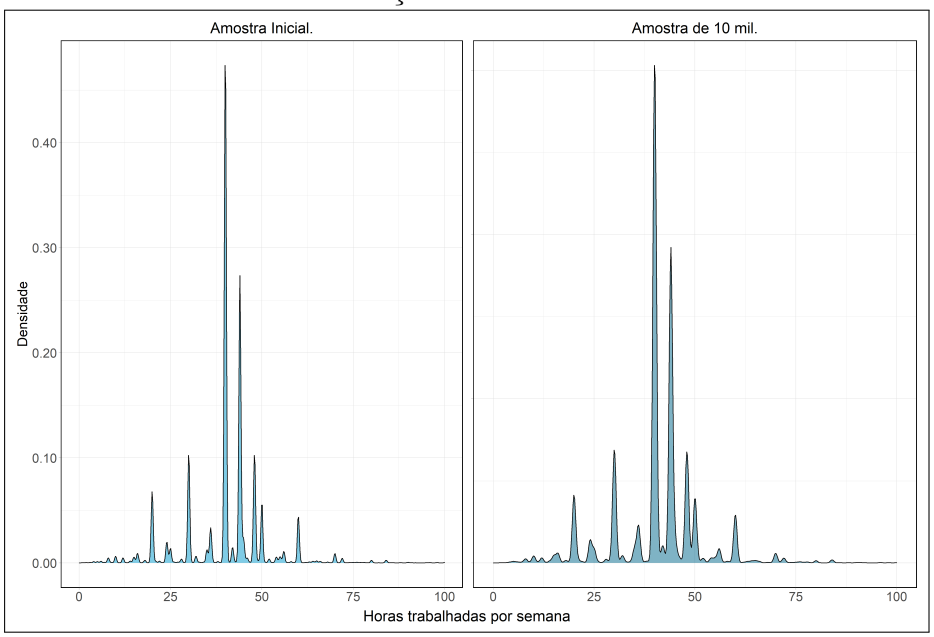


Figura 8: Densidade da variável Número de pessoas no domicílio, na amostra inicial (Completa) e na amostra de dez mil observações.

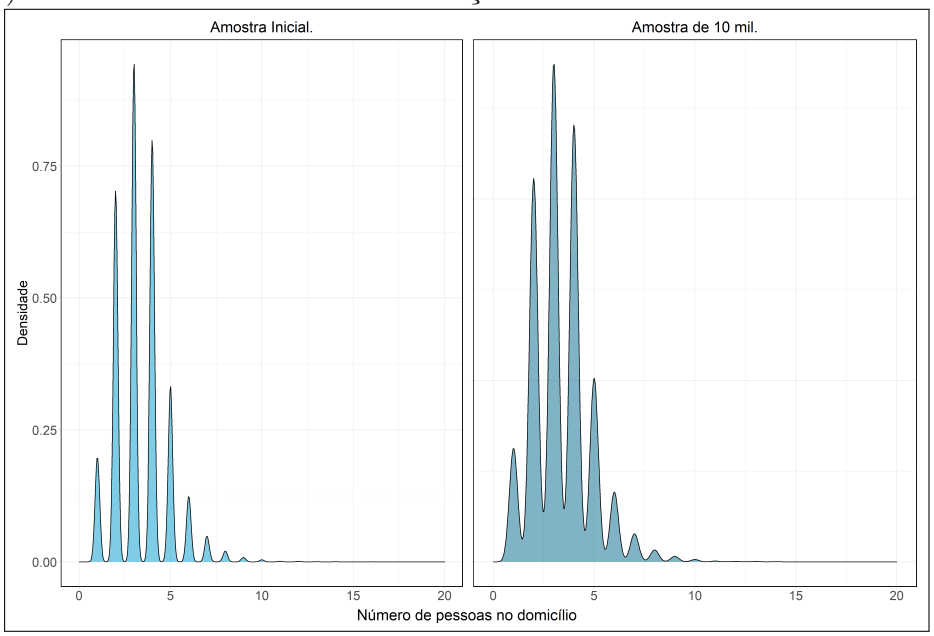


Figura 9: Densidade da variável Número de pessoas no domicílio, na amostra inicial (Completa) e na amostra de dez mil observações.

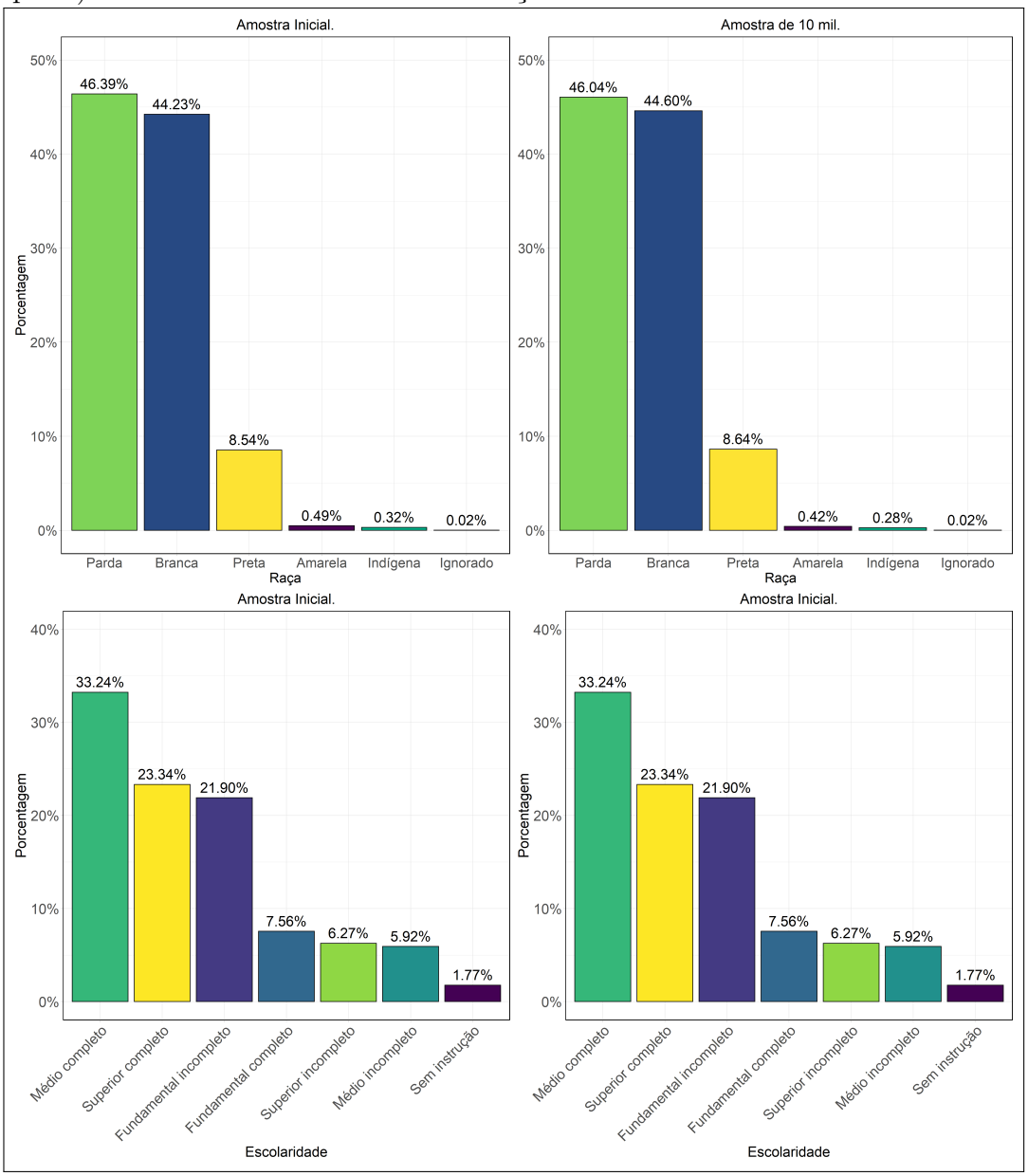
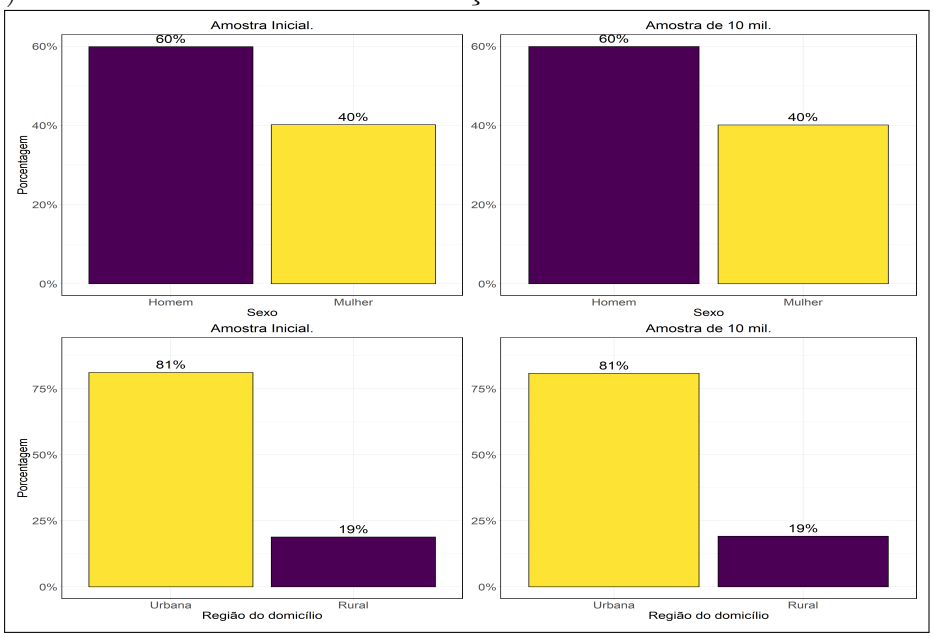


Figura 10: Densidade da variável Número de pessoas no domicílio, na amostra inicial (Completa) e na amostra de dez mil observações.



Na regressão linear vamos considerar o seguinte modelo para nossa amostra

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \epsilon_i. \tag{3.1}$$

onde:

- Y_i é a renda mensal do indivíduo da amostra;
- X_{i1} é idade do indivíduo;
- X_{i2} é a variável categórica que indica a raça do indivíduo;
- X_{i3} é a variável categórica que indica o nível de escolaridade do indivíduo;
- X_{i4} é a variável indicadora do sexo do indivíduo;
- X_{i5} é a variável indicadora da região do domicílio do indivíduo;
- X_{i6} é o número de pessoas que moram na mesma residência que o indivíduo;
- X_{i7} Número de horas trabalhadas por semana pelo indivíduo.
- ϵ_i é o termo de erro aleatório, onde esses erro são iid com distribuição Normal, média igual a zero e variância constante.

Já na regressão quantílica o modelo será

$$\begin{aligned} Y_i = \beta_0(\tau) + \beta_1(\tau)X_{i1} + \beta_2(\tau)X_{i2} + \beta_3(\tau)X_{i3} + \beta_4(\tau)X_{i4} \\ +\beta_5(\tau)X_{i5} + \beta_6(\tau)X_{i6} + \beta_7(\tau)X_{i7} + \epsilon_i(\tau). \end{aligned} \tag{3.2}$$

onde as variáveis explicativas e resposta são as mesmas do modelo dado pela Equação em (3.1). Os coeficientes $\beta(\tau)$ podem ser calculados para qualquer τ um dos quantis de interesse, além do $\epsilon_i(\tau)$ pode ser *iid*, *inid* e *bootstrap*, como mostrado na Seção 2.3. Para verificarmos qual a melhor suposição para o erro padrão dos coeficientes do modelo, testamos qual dos dois métodos se sai melhor na nossa base de dados, o Método assintótico (*iid* e *inid*) ou o Método *bootstrap*(Sem suposição sobre os erros). Foi estimado os valores da regressão quantílica para o quantil da mediana ($\tau = 0,5$). Observando a Tabela 3 notamos que não há uma grande diferença nos valores dos erros padrão para cada estimativa, comparando os diferentes métodos, exceto na estimativa da Raça ignorado, onde o método assintótico mostraram valores de bem superiores ao método *bootstrap*. Logo, decidimos utilizar o método *bootstrap* para estimar o erro padrão dos parâmetros do nosso modelo de regressão quantílica, pois ele mostrou uma boa eficiência nas estimativas, além de não necessitar de suposições sobre os erros do modelo. Já para estimar os coeficientes do modelo em si, o método do *ponto interior* foi o escolhido, pois como comentado da Seção 2.2 esse método é mais eficiente em base de dados com muitas observações, o que é o caso da nossa base de estudo.

Nas Figuras 12 e 11 temos uma comparação das variáveis explicativas numéricas e categóricas do modelo, vale destacar que a variável número de pessoas que moram na residência foi transformada em uma variável categórica, para que pudéssemos obter uma melhor visualização da mesma em relação a renda. Podemos notar que os indivíduos do sexo masculino tem uma renda maior que os do sexo feminino, isso também é observado na Tabela 3, quanto ao número de pessoas na residência, a categoria 12 apresentou uma variabilidade muito grande nos valores, isso se deve ao fato de que só possuem três observações nessa categoria, já que a mediana dessa categoria é igual a $R\$1.045$, que é inferior a mediana total da amostra que é de $R\$1.500$, ou seja essa variabilidade é influenciada por um indivíduo que possui renda igual a $R\$10.000$. Em relação a raça, os indivíduos autodeclarados brancos, parecem ter rendas maiores do que os indivíduos de outras raças, só equiparados aos indivíduos de que tiveram a raça ignorada, porém só tem duas observações dessa categoria na base, logo não é possível comparar as categorias de forma real. Os indivíduos que moram na região urbana e possuem ensino superior completo também possuem indicativos de rendas maiores do que indivíduos que moram

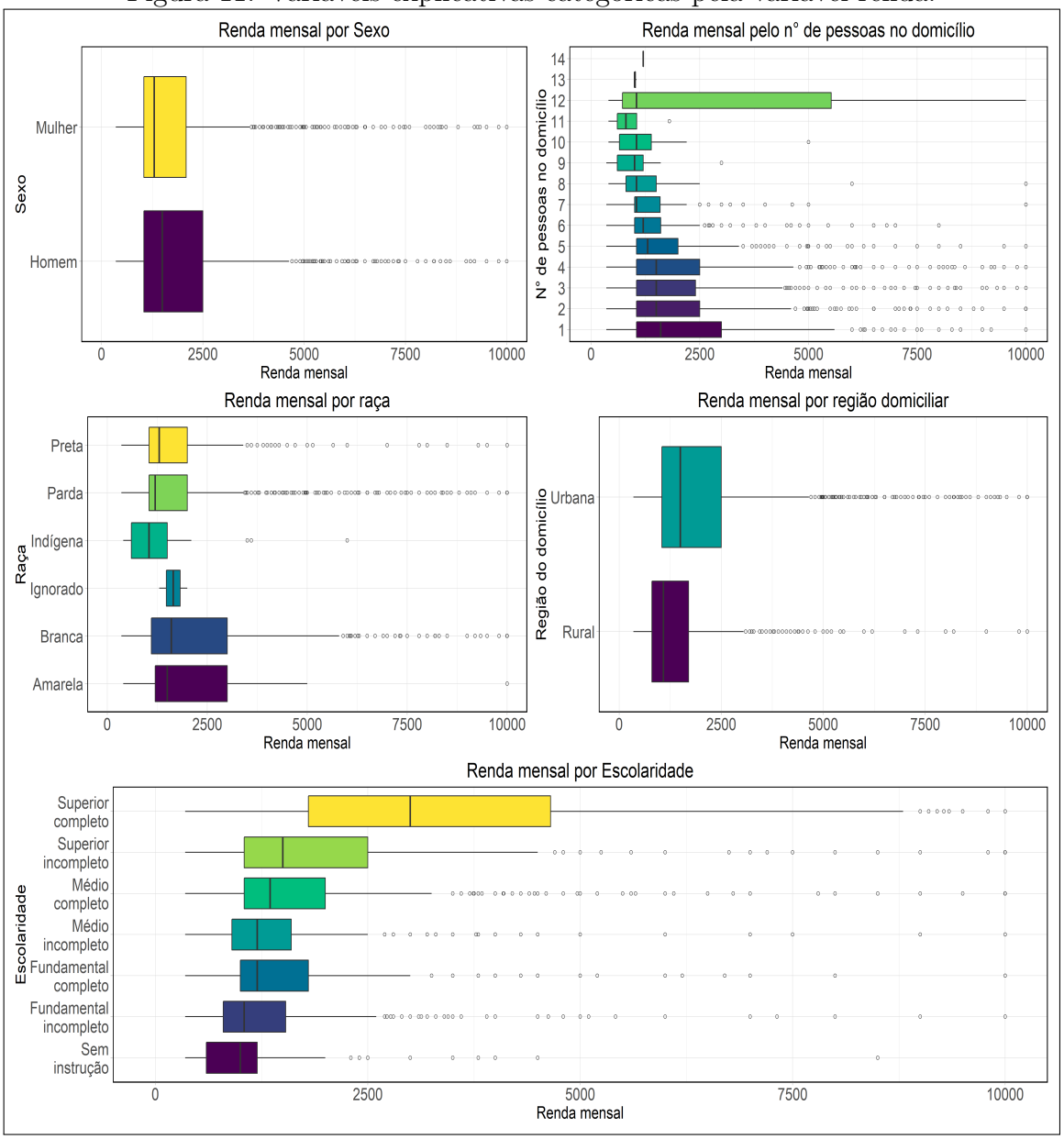
Tabela 3: Valores dos erros padrão para métodos inferenciais diferentes.

		Erro Padrão		
Parâmetros	Estimativas	iid	inid	boot
Intercepto	- 82,06	126,17	107,09	133,70
Idade	16,68	0,64	0,62	0,79
Raça branca	199,29	116,14	100,13	122,84
Raça indígena	- 26,10	182,93	123,47	171,96
Raça parda	- 6,81	116,23	99,51	122,52
Raça preta	- 4,01	118,45	100,91	123,19
Raça ignorado	25,85	542,09	2325,98	285,52
Escolaridade fund. incom.	25,14	26,14	16,71	21,49
Escolaridade fund. com.	50,30	22,93	22,09	24,33
Escolaridade med. incom.	242,43	28,35	25,47	26,94
Escolaridade med. com.	579,47	31,12	29,18	32,81
Escolaridade sup. incom.	654,04	37,90	33,50	40,40
Escolaridade sup. com.	1564,38	39,71	36,52	41,95
Sexo feminino	- 286,68	16,07	14,44	17,75
Região urbana	184,65	20,14	15,72	19,08
Horas trabalhadas	22,92	0,71	0,55	0,92
N° de pessoas	- 19,13	5,32	4,13	5,87

na região rural e que tem níveis de escolaridades menores. Vale ressaltar a presença de valores atípicos em quase todas categorias da variáveis estudadas.

Em relação as variáveis quantitativas, a Figura 12 mostra que é difícil afirmar que existe uma relação linear crescente entre as variáveis e a renda, o que faz sentido em relação a idade, já que a maioria das pessoas mais velhas são aposentadas, ou seja, recebem em sua grande maioria um salário mínimo, essas suposições vão ser verificadas na análise de regressão linear e quantílica.

Figura 11: Variáveis explicativas categóricas pela variável renda.

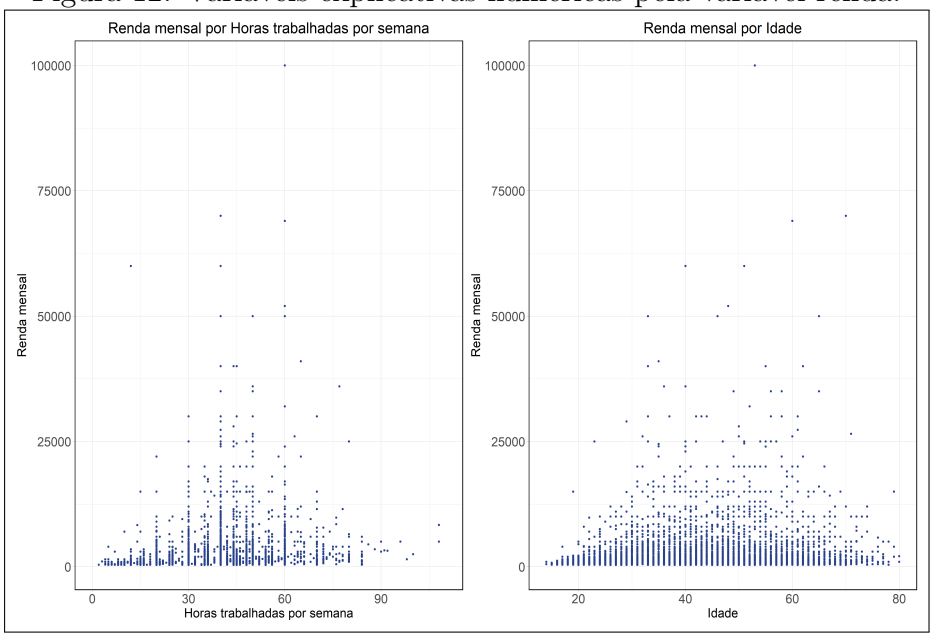


3.1.1 Regressão linear clássica

O primeiro modelo estimado foi o da regressão linear clássica, o modelo estimado é o mesmo apresentado na Equação em (3.1). Na Tabela 4 temos as estimativas dos coeficientes desse modelo de regressão, que foi denotado como Modelo Completo, observamos que as variáveis número de pessoas e raça não parecem ser muito significativas para a explicação da renda, assim como os níveis de escolaridade fundamental incompleto e fundamental completo. Logo, foi estimado um segundo modelo sem essas variáveis e sem essas categorias, a fim de buscar um modelo melhor ajustado aos dados.

Na Tabela 5 temos os coeficientes estimados para esse novo modelo, que foi chamado de

Figura 12: Variáveis explicativas numéricas pela variável renda.



Modelo Reduzido, é possível notar que apenas os indivíduos com ensino médio incompleto não são explicativos para explicar a distribuição da renda na nossa amostra, entretanto se adicionarmos essa categoria a categoria base da variável escolaridade, que é a sem instrução as categorias de ensino médio completo não será significativa para explicar a renda dos indivíduos, por isso vamos mantê-la no nosso modelo. O coeficiente R^2 apresentou o mesmo valor em ambos os modelos, porém o valor em ambos foi baixo, só 21% da variabilidade dos dados é explicada por ambos modelos. Além disso o AIC dos Modelos 1 e 2 foram respectivamente $189879,7$ e $135289,1$. Logo, o Modelo Reduzido foi o escolhido entre os dois, pois além de ter um AIC menor, ele tem um R^2 igual ao do Modelo Completo só que com menos variáveis, ou seja, ele é o modelo mais parcimonioso.

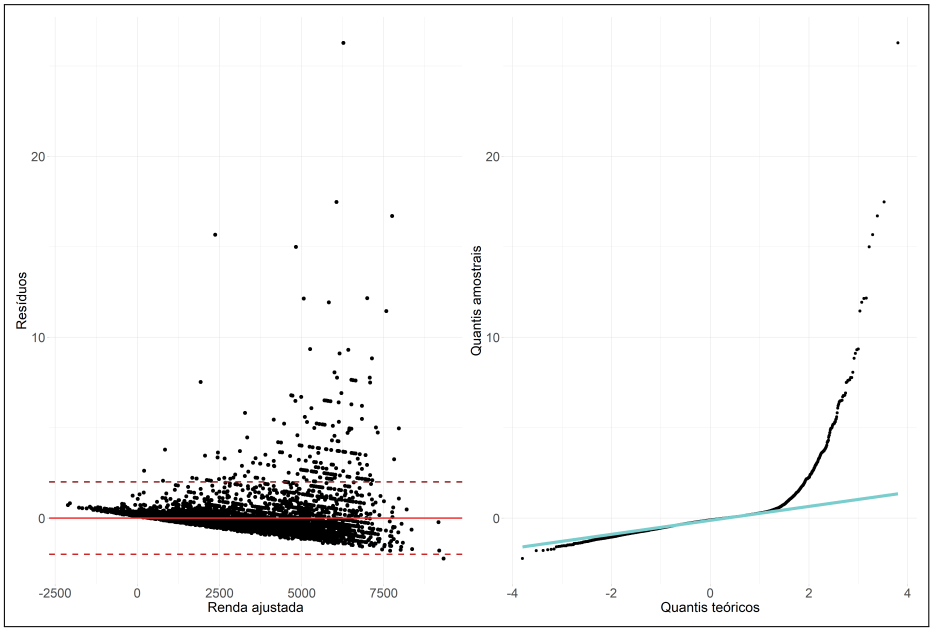
Tabela 4: Estimativas, erro padrão e p-valor do modelo de regressão clássica (Modelo Completo).

Parêmetro	Estimativa	Erro Padrão	p-valor
Intercepto	-572,86	541,15	0,29
Idade	45,95	2,76	0,00
Raça branca	-324,03	498,12	0,52
Raça indígena	-1007,48	784,60	0,20
Raça parda	-837,81	498,52	0,09
Raça preta	-912,23	508,05	0,07
Raça ignorado	-1391,46	2325,09	0,55
Escolaridade fund. incom.	-35,43	112,12	0,75
Escolaridade fund. com.	13,48	98,36	0,89
Escolaridade med. incom.	428,34	121,58	0,00
Escolaridade med. com.	933,66	133,47	0,00
Escolaridade sup. incom.	1159,58	162,54	0,00
Escolaridade sup. com.	3153,34	170,33	0,00
Sexo feminino	-875,21	68,92	0,00
Região urbana	327,54	86,40	0,00
Horas trabalhadas	38,83	3,03	0,00
N° de pessoas	-28,02	22,82	0,22
		R^2	0,21

Tabela 5: Estimativas, erro padrão e p-valor do modelo de regressão clássica sem as variáveis número de pessoas e a raça (Modelo Reduzido).

Parêmetro	Estimativa	Erro Padrão	p-valor
Intercepto	-1976,83	264,07	0,00
Idade	63,15	3,83	0,00
Escolaridade med. incom.	428,34	120,48	0,14
Escolaridade med. com.	592,83	183,73	0,00
Escolaridade sup. incom.	898,54	166,21	0,00
Escolaridade sup. com.	3335,16	214,63	0,00
Sexo feminino	-1038,48	92,56	0,00
Região urbana	299,92	139,01	0,03
Horas trabalhadas	49,17	4,34	0,00
		R^2	0,21

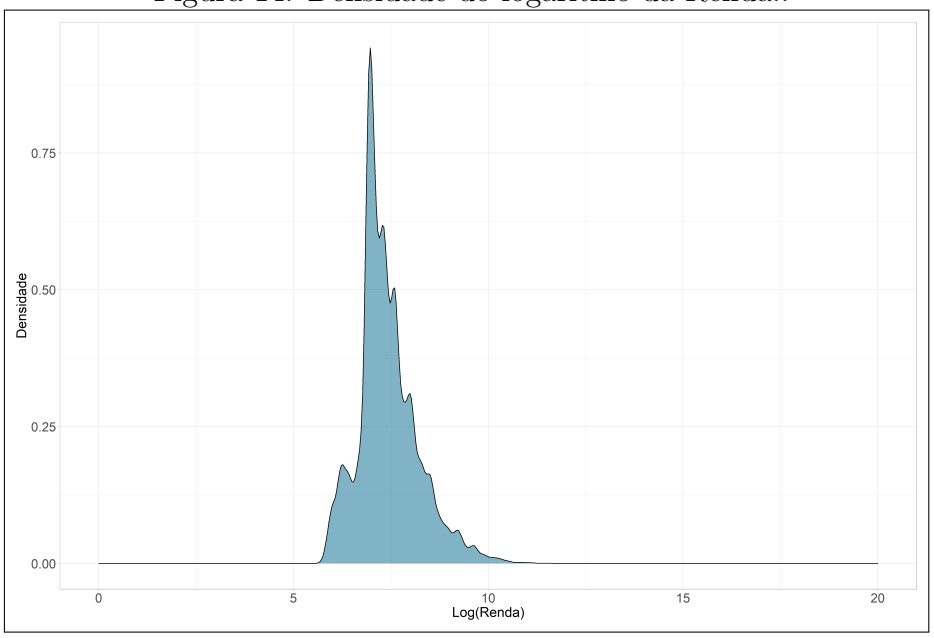
Figura 13: Análise dos resíduos versus a Renda ajustada pelo modelo reduzido e dos quantis teóricos versus os amostrais.



Como apresentado nos capítulos anteriores dessa monografia o modelo de regressão linear clássica deve cumprir algumas suposições em relação aos resíduos do modelo, que são eles serem normais e homocedásticos, para verificar essas suposições vamos utilizar análises gráficas e de testes estatísticos já comentados no início desse capítulo. Os testes de Kolmogorov-Smirnov e Breusch-Pagan apresentaram um $p-valor < 0,05$, além disso, analisando a Figura 13 observamos que a cauda superior dos resíduos foge muito dos quantis da normal, enquanto a comparação dos resíduos versus os valores da renda ajustados pelo modelo, evidência que muitos pontos estão fora do intervalo $[-2, 2]$, além de indicar a existência de um possível padrão na variabilidade dos erros, ou seja, as análises gráficas corroboram com os resultados dos testes realizados, por isso devemos realizar uma transformação na variável resposta.

A transformação escolhida foi a logarítmica, ou seja, foi estimado um novo modelo de regressão linear, onde a variável resposta agora é o logaritmo da renda. Na Figura 14 podemos visualizar que a densidade do logaritmo da renda é mais simétrico do que densidade da variável renda sem a transformação, mostrada na Figura 4. Então, foi estimado um novo modelo, o Modelo Log. Suas estimativas podem ser observadas na Tabela 6. Vale ressaltar o fato de que primeiro foi estimado um modelo com todas as variáveis, porém assim como no Modelo Completo, a variável Raça não foi significativa para explicação do logaritmo da renda, além de algumas categorias da escolaridade. Assim, o modelo apresentado na Tabela 6 é aquele que contém só as variáveis explicativas que foram signi-

Figura 14: Densidade do logaritmo da Renda..

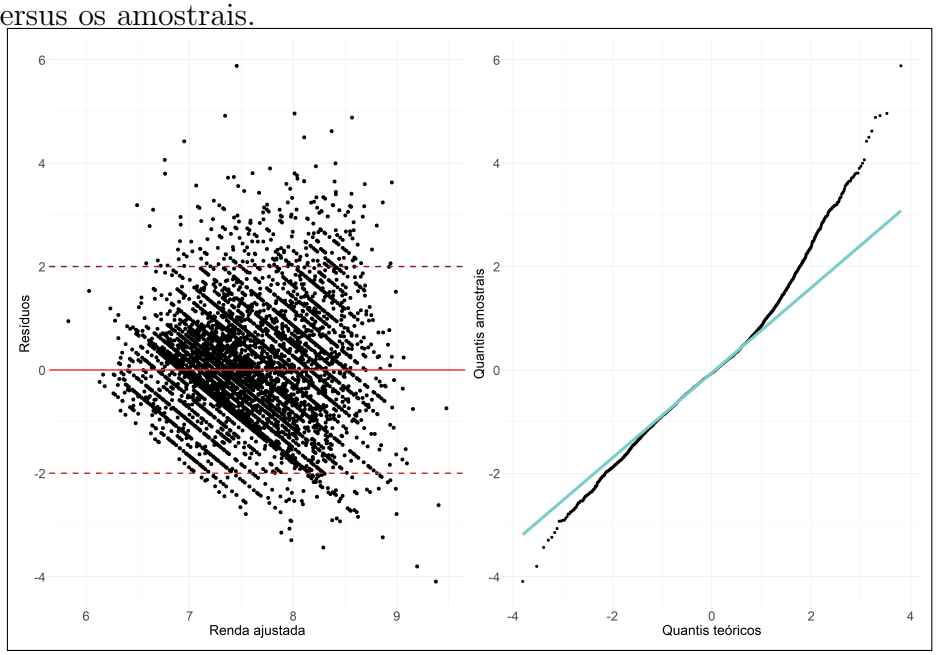


ficantes para a explicação do logaritmo da renda, assim como no Modelo Reduzido, uma categoria da escolaridade não foi significativa, nesse modelo foi o nível médio completo.

Quanto a adequabilidade do modelo, a Tabela 7 mostra que o Modelo Log foi o modelo que melhor se ajustou aos dados, porém assim como os outros modelos estimados, ele também teve as suposições de normalidade e homocedasticidade dos erros rejeitada pelos teste de Kolmogorov-Smirnov e Breusch-Pagan, respectivamente. Analisando os resíduos do Modelo Log, pela Figura 15, tem-sea uma melhora na variabilidade dos resíduos, porém as caudas do gráfico dos quantis, mostra que as caudas da amostra destoam muita da distribuição normal. Como foi observado anteriormente na Figura 5, nossa variável resposta tem muitos valores atípicos e como é sabido o modelo de regressão linear clássico é sensível a presença de desses valores, por isso, utilizamos duas medidas comuns para essas verificar a influência desses pontos no ajuste do modelo de regressão linear, que é a distância de cook e os pontos de alavanca (Laverage).

Na Figura 16 podemos observar que temos muitas observações acima da linha de limite, ou seja, temos muitas observações que influenciam no ajuste. A alternativa muito usada nesses casos é a retirada dessas observações depois estimar outro modelo sem elas, porém como nesse caso em específico temos muitas observações atípicas, para aceitar os pressupostos do modelo de regressão linear, seria necessário retirar cerca de 52% das observações, isso não será feito neste caso, já que a retirada de tantas observações pode acarretar na perda do poder preditivo do nosso modelo em relação aos dados. Logo, podemos concluir que a regressão linear clássica não é ideal para analisar nossa base de

Figura 15: Análise dos resíduos versus a Renda ajustada pelo modelo log e dos quantis teóricos versus os amostrais.



dados em questão.

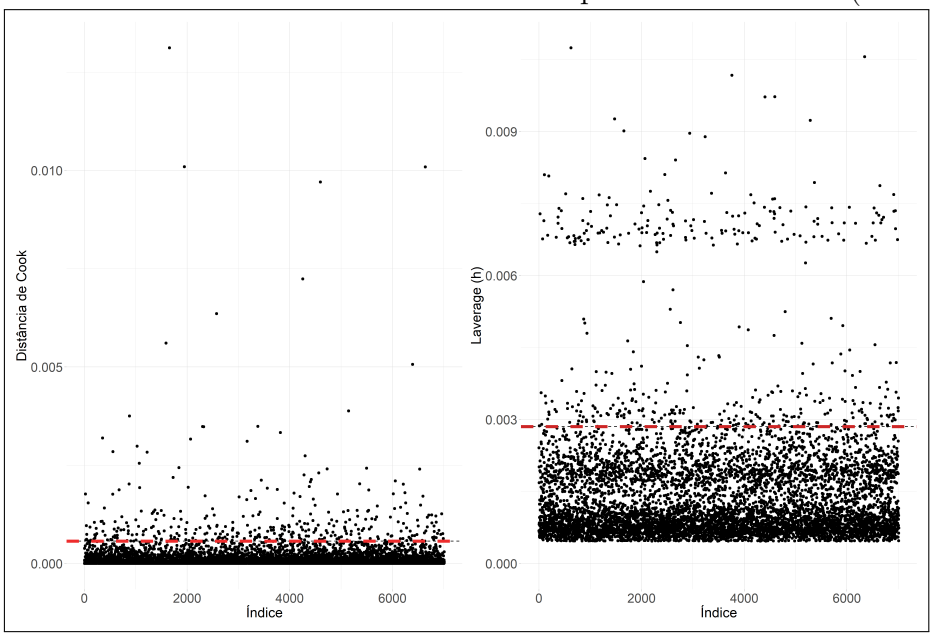
Tabela 6: Estimativas, erro padrão e p-valor do modelo de regressão considerando o logaritmo da renda.

Parêmetro	Estimativa	Erro Padrão	p-valor
Intercepto	6,09	0,05	0,00
Idade	0,02	0,00	0,00
Escolaridade med. incom.	-0,05	0,02	0,01
Escolaridade med. com.	0,04	0,03	0,19
Escolaridade sup. incom.	0,22	0,03	0,00
Escolaridade sup. com.	1,05	0,03	0,00
Sexo feminino	-0,27	0,01	0,00
Região urbana	0,15	0,02	0,03
Horas trabalhadas	0,02	0,00	0,00
N° de pessoas	-0,02	0,01	0,00
		R^2	0,42

Tabela 7: Comparação da adequabilidade dos modelos estimados , através do AIC e do R^2

Modelos	AIC	R^2
Modelo Completo	189879,7	0,21
Modelo Reduzido	135289,1	0,21
Modelo Log	12855,08	0,42

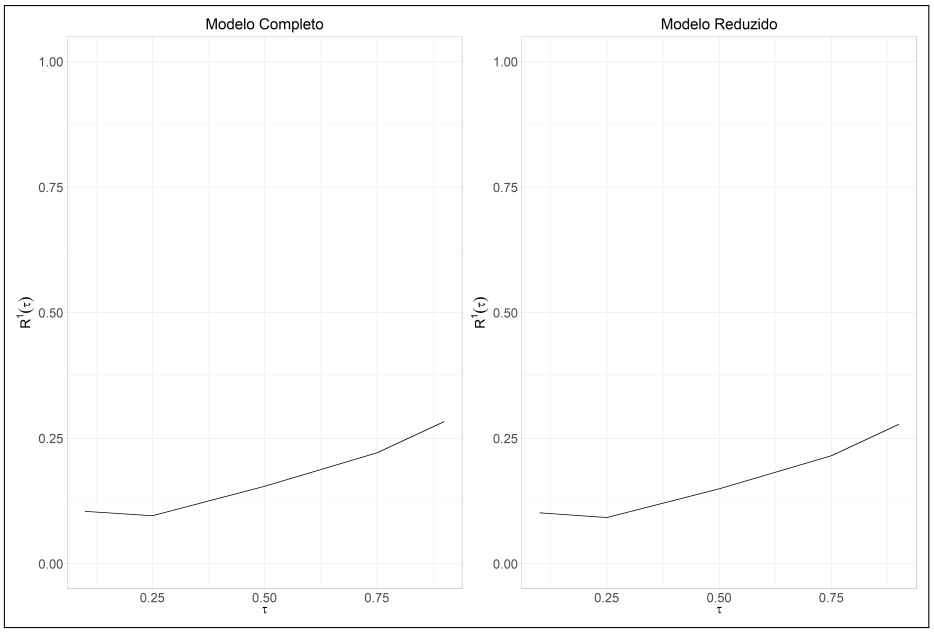
Figura 16: Análise da Distância de Cook o dos pontos de alavanca (Laverage) .



3.1.2 Regressão linear quantílica

Agora iremos estimar o modelo de regressão quantílica mostrado na Equação em (3.2). Esse modelo estimado é apresentado na Tabela 9, podemos observar que temos estimativas para os parâmetros nos quantis de $\tau = 0{,}10$; 0,25; 0,50; 0,75 e 0,90, nota-se que a variável Raça não foi significativa em nenhum dos quantis, enquanto ao nível de escolaridade, os indivíduos com nível até o ensino fundamental incompleto só foi significativo no quantil $\tau = 0,25$, ou seja, ela só se mostrou influente em salários mais baixos, o que é de se esperar visto que esse indivíduos tem rendas menores em relação às outras escolaridades, como já havia sido observado na Figura 11, já o nível fundamental completo, foi significativo no quantis $\tau = 0{,}50$; 0,75 e 0,90, isso talvez decorra do fato dessa categoria possuir uma forte assimetria positiva, além de um valor maior do primeiro quartil do que as categorias de escolaridade ensino médio incompleto e sem instrução, essas análises podem ser observadas na Figura 11. Então vamos estimar um segundo modelo sem a variável raça e sem a categoria da escolaridade até o fundamental incompleto. As estimativas do Modelo Reduzido são apresentadas na Tabela 10, podemos observar que o nível de escolaridade fundamental completo foi não significante para explicação da renda do indivíduo nos quantis $\tau = 0{,}10$ e 0,75, porém quando retiramos essa categoria da análise a categoria escolaridade ensino médio incompleto fica não significativa, para explicar a renda, então, como ela ainda se mostrou significativa em três quantis vamos manter essa categoria na análise. O número de pessoas também se mostrou não significativa em dois quantis, porém como ela foi significativa nos outros três quantis analisados, vamos manter

Figura 17: Valores da estatística $R^1(\tau)$ para todos os quantis de interesse em ambos os Modelos.



essa variável na análise.

Em relação a qualidade do ajuste dos Modelos estimados, completo e reduzido, na Tabela 8 temos os valores do AIC para os diferentes quantis de ambos os modelos, verificamos que os valores do AIC do Modelo Completo são maiores nos quantis $\tau = 0{,}10$; 0,25; 0,50 e 0,75 do que os valores do AIC do Modelo Reduzido, ao observamos os valores da estatística $R^1(\tau)$ na Figura 17 é possível notar que não existe diferença na variabilidade explicada pelos modelos, ou seja, mesmo com uma variável a menos o Modelo Reduzido consegue explicar a renda de um indivíduo tão bem quanto o Modelo Completo, por isso, vamos escolher o Modelo Reduzido, além de ter valores menores do AIC na maioria dos quantis analisados.

Tabela 8: Valores do AIC por quantil, em cada um dos modelos estimados.

	AIC por Quantil				
Modelos estimados	0,10	0,25	0,50	0,75	0,90
Modelo Completo	171221,8	172146,5	176191,7	183962,5	193799,2
Modelo Reduzido	171144,3	172084,3	176157,8	183958,9	193803,8

Na Figura 18 podemos observar o comportamento das estimativas dos coeficientes para as variáveis idade e horas trabalhadas por semana, em todos os quantis, nessas variáveis é possível notar que quanto maior o quantil de interesse maior é o valor do coeficiente, porém esse aumento ocorre de forma diferente em cada uma das variáveis. Na Figura 19 temos o comportamento dos coeficientes da variável sexo feminino em comparação

Tabela 9: Estimativas para os parâmetros do Modelo Completo nos diferentes quantis (p-valor).

	Quantis				
Parâmetro	0,10	0,25	0,50	0,75	0,90
Intercepto	-26,53	18,76	-82,06	-2,80	2488,07
	(0,88)	(0,91)	(0,61)	(1,00)	(0,30)
Idade	6,32	8,33	16,68	29,05	51,82
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Raça branca	64,10	130,56	199,29	292,94	-1674,05
	(0.69)	(0,43)	(0,19)	(0,78)	(0,49)
Raça ignorado	441,68	186,11	25,85	-406,65	-3849,42
	(0,23)	(0,58)	(0,93)	(0,70)	(0,14)
Raça indígena	-183,65	-163,25	-26,10	-287,75	-2152,61
	(0,29)	(0,46)	(0,89)	(0,79)	(0,38)
Raça Parda	-38,03	7,78	-6,81	-38,90	-2258,35
	(0,81)	(0,96)	(0,96)	(0,97)	(0,35)
Raça Preta	3,36	35,57	-4,01	-67,68	-2219,26
	(0,98)	(0,83)	(0,98)	(0,95)	(0,36)
Escolaridade sup. com.	606,39	848,40	1564,38	2806,38	5438,78
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade sup. incom.	129,48	293,96	654,04	1309,77	3098,91
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade med. com.	83,89	255,35	579,47	1091,48	2330,12
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade med. incom.	47,36	109,91	242,43	439,78	1025,44
	(0,03)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade fund. com.	24,10	27,03	50,30	108,04	336,16
	(0,26)	(0,09)	(0,04)	(0,02)	(0,00)
Escolaridade fund. incom	2,44	27,73	25,14	19,18	83,19
	(0,92)	(0,07)	(0,23)	(0,59)	(0,27)
Sexo feminino	-103,92	-141,11	-286,68	-544,34	-863,88
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Horas trabalhadas	12,74	15,97	22,92	29,08	34,30
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Região urbana	183,51	175,19	184,65	249,92	197,93
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
N° de pessoas	-8,52	-10,00	-19,13	-32,40	-19,83
	(0,09)	(0,01)	(0,00)	(0,00)	(0,30)

Tabela 10: Estimativas para os parâmetros do Modelo Reduzido nos diferentes quantis (p-valor).

	Quantis				
Parâmetro	0,10	0,25	0,50	0,75	0,90
Intercepto	32,93	135,40	112,11	285,55	779,13
	(0,49)	(0,00)	(0,06)	(0,00)	(0,00)
Idade	6,42	8,24	16,36	29,93	53,68
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade sup. com.	549,66	766,41	1433,79	2631,28	5321,82
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade sup. incom.	116,50	333,97	731,06	1427,96	3346,18
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade med. com.	89,15	241,00	475,80	918,87	2026,66
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade med. incom.	30,86	70,84	130,84	237,44	711,40
	(0,19)	(0,00)	(0,00)	(0,00)	(0,00)
Escolaridade fund. com.	18,83	42,91	47,76	59,65	228,04
	(0,31)	(0,00)	(0,04)	(0,13)	(0,01)
Sexo feminino	-102,13	-138,99	-258,10	-547,11	-894,39
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Horas trabalhadas	12,86	16,59	23,16	29,91	35,34
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Região urbana	169,02	169,21	181,86	213,54	230,18
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
N° de pessoas	-9,46	-13,86	-23,49	-35,76	-31,18
	(0,06)	(0,01)	(0,00)	(0,00)	(0,12)

a categoria base, que é o sexo masculino e a variável região urbana em comparação a região rural, que é a categoria base dessa variável. Em relação ao sexo, podemos notar que a renda das mulheres é menor em comparação a dos homens em todos os quantis, quanto maior o quantil menos as mulheres tem de renda em comparação aos homens. Quanto a região do domicílio, notamos que quanto maior o quantil maior é a renda de quem reside em áreas urbanas. Na Figura 20 temos a variável número de pessoas que moram na residência, podemos observar um relação negativa entre essa variável e a renda em todos os quantis, assim como o valor da estimativa vai diminuindo em relação que o valor do quantil aumenta, exceto, em comparação ao quantil 0,90 apresenta um valor maior da estimativa em relação ao quantil 0,75, entretanto, a relação da variável com a renda ainda é mesma dos outros quantis, outra observação interessante é o fato de nos quantis 0,10 e 0,90 o intervalo de confiança do coeficiente contém o valor 0, ou seja, como o 0 está dentro do intervalo a estimativa do coeficiente foi considerada não significativa nesses quantis.

Nas Figura 20 temos também, o gráfico das estimativas do coeficiente de uma das categorias da variável escolaridade, a escolaridade fundamental completa, enquanto nas Figuras 21 e 22 temos as outras categorias dessa variável, é interessante notar que, assim como no caso da variável número de pessoas, os quantis em que essa categoria não se mostrou significativa, são quantis onde o intervalo de confiança da estimativa do coeficiente tem o 0 contido nele. Em relação as categorias da variável escolaridade, vale ressaltar o fato de que a categoria base dessa variável é composta pelas categorias, sem instrução e fundamental incompleto. Podemos notar que os indivíduos com a escolaridade até o nível fundamental completo, ganham um renda mensal maior que os de um nível inferior durante todos os quantis da amostra, mas com uma diferença ainda maior observando apenas o quantil 0,90, já a escolaridade ensino médio incompleto, que foi não significativa no quantil 0,10 pelo mesmo motivo. Olhando a relação da escolaridade com a renda, podemos observar que quanto maior a escolaridade do indivíduo maior será sua renda conforme o quantil de interesse aumenta, em especial no quantil 0,90, temos um diferença maior das rendas, isso decorre do fato da reta estimada desse quantil, explicar os maiores valores da nossa amostra.

Em relação ao $R^1(\tau)$ apresentado na Figura 17, é importante notar que essa estatística, como explicado anteriormente na Seção 2.4 é uma estimativa local da qualidade do ajuste do modelo de regressão quantílica, ou seja, é o valor que cada reta estimada representa da variabilidade da renda explicada pelas variáveis explicativas, porém não é correto somar esses valores e dizer que o nosso modelo explica 84% da variabilidade da renda, pois a

Figura 18: Estimativas dos coeficientes e intervalo de confiança das variáveis Idade e Horas trabalhadas.

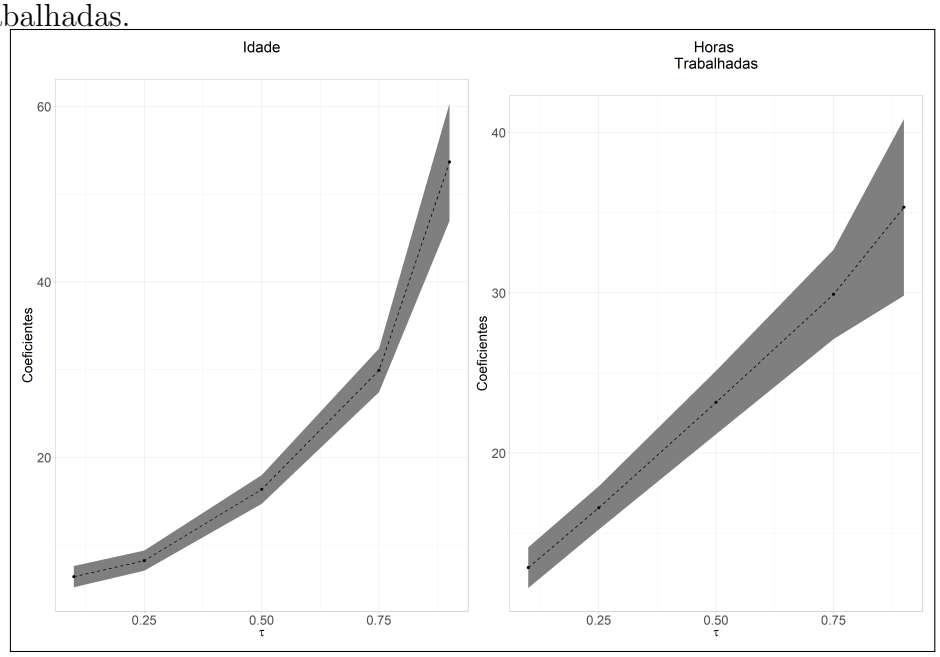


Figura 19: Estimativas dos coeficientes e intervalo de confiança das variáveis Sexo e Região de domicílio.

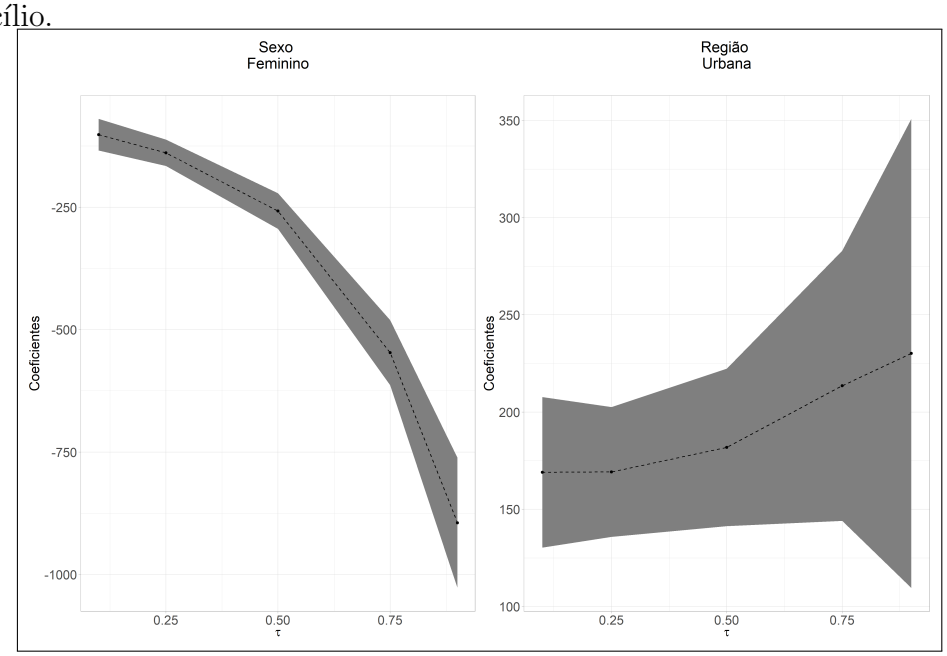


Figura 20: Estimativas dos coeficientes e intervalo de confiança das variáveis para a Escolaridade até o Fundamental completo e N° de pessoas.

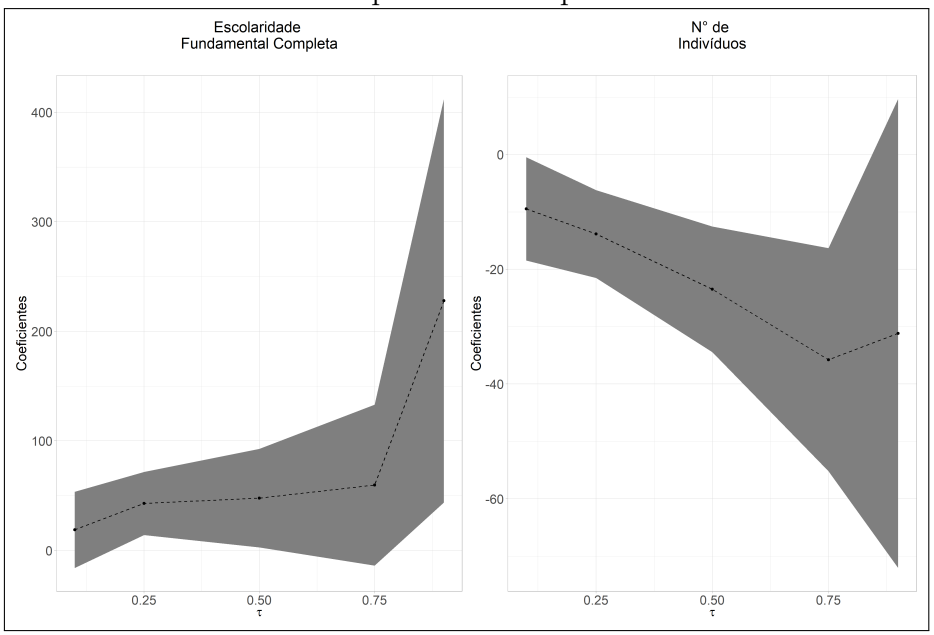


Figura 21: Estimativas dos coeficientes e intervalo de confiança para a Escolaridade até o Ensino Médio completo e incompleto.

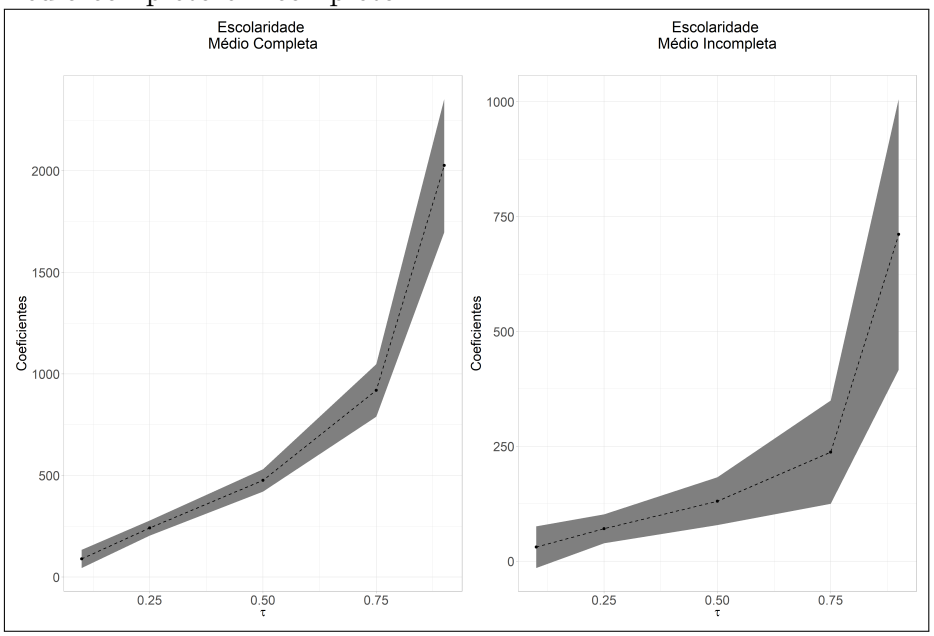
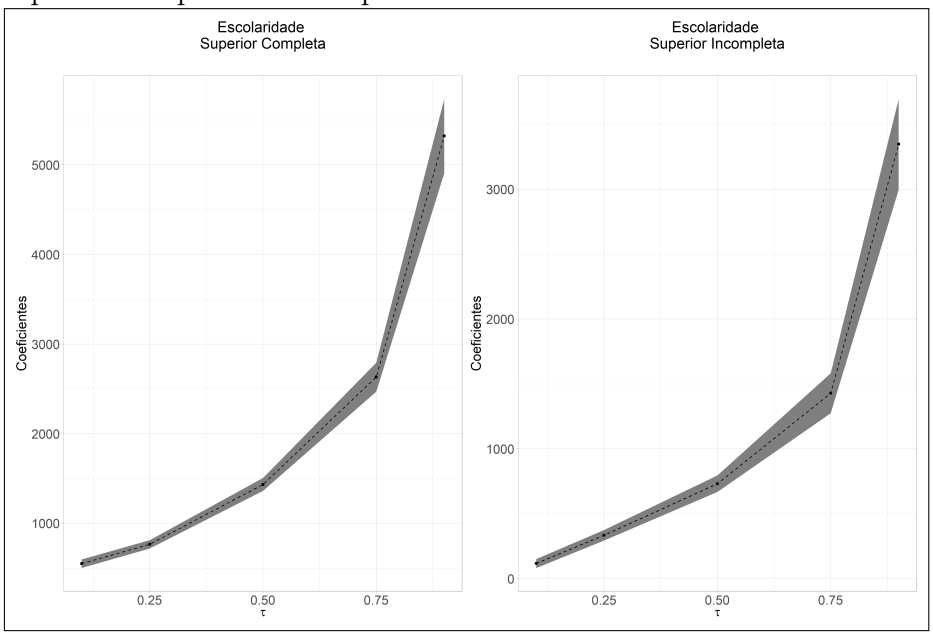


Figura 22: Estimativas dos coeficientes e intervalo de confiança para a Escolaridade até o Ensino Superior completo e incompleto.



observações que são explicadas por mais de um reta estimada, mas é compreensível dizer que os valores explicados pela reta 0.10 são diferentes dos do quantil 0,90, que é exatamente um dos pontos fortes da regressão quantílica, ou seja, nosso modelo estimado consegue ter uma visão mais abrangente da relação da renda com nossas variáveis explicativas. Importante ressaltar que as retas que melhor explicaram a variabilidade da renda foram as retas estimadas referentes aos quantis 0,05, 0,75, 0,90

Sobre os resíduos do modelo, como usamos o método *bootstrap* para estimar os erros do modelo, como comentado na Seção 2.3, não é necessário nenhum tipo de suposição sobre os resíduos do nosso modelo, ou seja, não precisamos verificar se os resíduos são normais, ou seguem qualquer distribuição específica, que foi um dos motivos que impediu a interpretação do modelo de regressão linear clássico. Logo, para a base de dados utilizada nessa análise, os modelos de regressão quantílica se mostraram mais adequados para explicar a renda mensal dos brasileiros.

3.2 Dados sobre a admissão de estudantes a Programas de Mestrado

Como comentado na Subseção 2.5.2, nossa base possui 9 variáveis e 500 observações, onde a variável a nossa variável de interesse é a chance de admissão, que é uma variável contida no intervalo $[0, 1]$, então, vamos definir que a nossa variável resposta assume os

valores 0 e 1, com as seguintes probabilidades:$P(Y = 0) = 0$ e $P(Y = 1) = 1$. Como nenhuma das variáveis explicativas tem valor 0, o valor do intercepto não possui um real importância para a nossa análise, por isso ele não será apresentado em nossos resultados.

Assim, tendo comentado brevemente sobre cada uma das variáveis, o modelo de regressão à ser estimado será:

$$\begin{aligned} Y_i = \beta_1(\tau)X_{i1} + \beta_2(\tau)X_{i2} + \beta_3(\tau)X_{i3} + \beta_4(\tau)X_{i4} \\ +\beta_5(\tau)X_{i5} + \beta_6(\tau)X_{i6} + \beta_7(\tau)X_{i7} + \epsilon_i(\tau). \end{aligned} \tag{3.3}$$

onde:

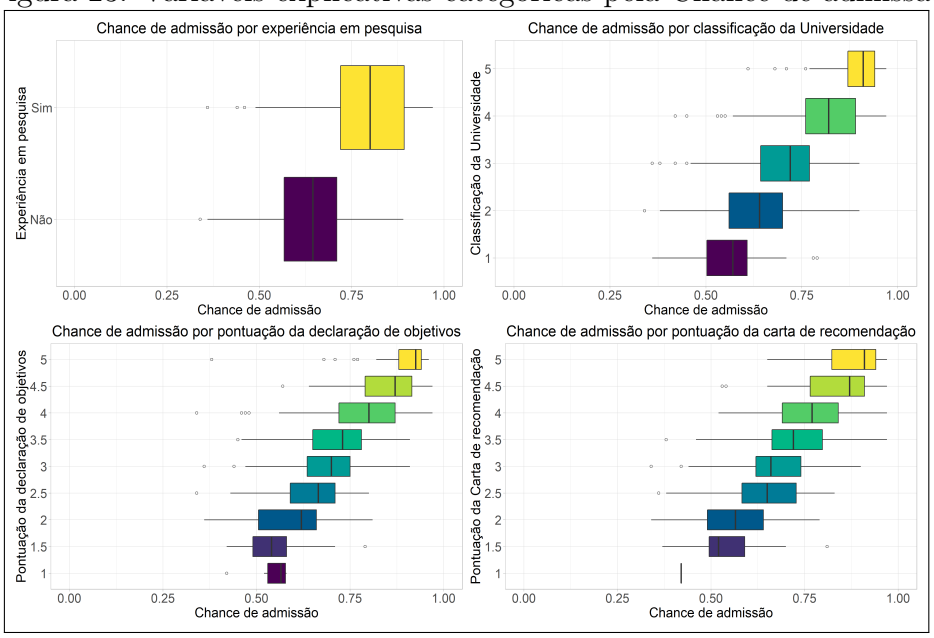
- Y_i é a chance do i-ésimo estudante ingressar num programa de mestrado;
- X_{i1} é a pontuação do i-ésimo estudante no exame GRE;
- X_{i2} é a pontuação do i-ésimo estudante no TOEFL ;
- X_{i3} é a classificação da universidade onde o i-ésimo estudante se graduou;
- X_{i4} é a pontuação do i-ésimo estudante na SOP;
- X_{i5} é a pontuação do i-ésimo estudante na LOR;
- X_{i6} é o valor do GPA do i-ésimo estudante;
- X_{i7} Variável categoria que indica se o aluno realizou alguma pesquisa.
- ϵ_i é o termo de erro aleatório, do modelo de regressão quantílica.

A Figura 23 apresenta as variáveis categóricas da base de dados pela chance de admissão, além delas, a pontuação obtida pelos estudantes nas variáveis declaração de objetivos (SOP) e carta de recomendação (LOR) também podem ser visualizadas, apesar de numéricas essas duas variáveis foram colocadas nessa figura, apenas para obter uma análise melhor das mesmas em relação a chance de admissão do estudante. Nota-se que as os estudantes que realizaram pesquisa tem mais chances de conseguir serem admitidos num programa de mestrado, do que aqueles que não fizeram alguma, quanto a classificação da universidade, observamos que os estudantes que se formaram em universidades de classificações superiores também parecem ter mias chances de serem aceitos. Em relação a pontuação da carta de recomendação e da declaração de abjetivos, é fácil observar que quanto maior a pontuação maiores as chances do estudante ser aceito em um programa de mestrado. Em todas as variáveis, podemos observar alguns valores atípicos ou fora

do padrão, além disso, vale ressaltar que a maioria das categorias das variáveis estão de distribuídas de forma assimétrica.

Já na 24 temo as variáveis quantitativas que são as pontuações do GRE e do teste de inglês (TOEFL), além da média de pontos das notas (GPA), em relação à chance de admissão, todas as variáveis apresentaram uma relação linear crescente. Logo, parece que a análise de regressão linear quantílica conseguirá explicar bem a relação dessas variáveis com a variável resposta.

Figura 23: Variáveis explicativas categóricas pela Chance de admissão.

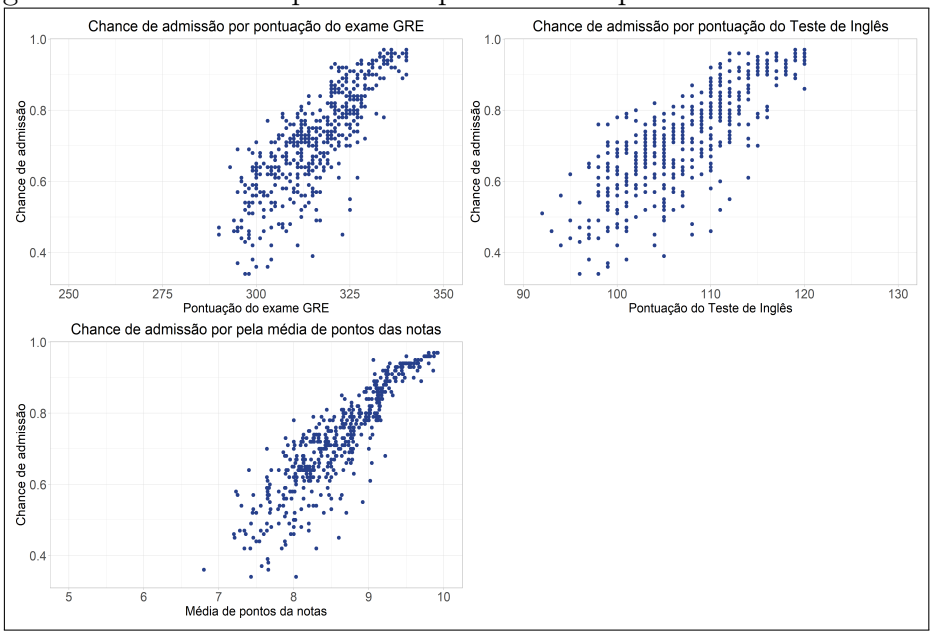


O Modelo 3.3 vai ter seus coeficientes estimados usando o método *simplex* de Barrodale-Roberts, que como já tínhamos comentado na Seção 2.2 tem uma melhor performance que o do *ponto interior* em bases de dados com menos observações, em relação ao erro padrão dos coeficientes, vamos utilizar o método *Markov Chain Marginal Bootstrap (MCMB)*, assim como no outro método *bootstrap*, ele não nos obriga a fazer suposições sobre os resíduos do modelo, entretanto, neste estudo em específico, iremos fazer uma breve análise sobre os resíduos do modelo quantílico estimado.

3.2.1 Regressão linear quantílica

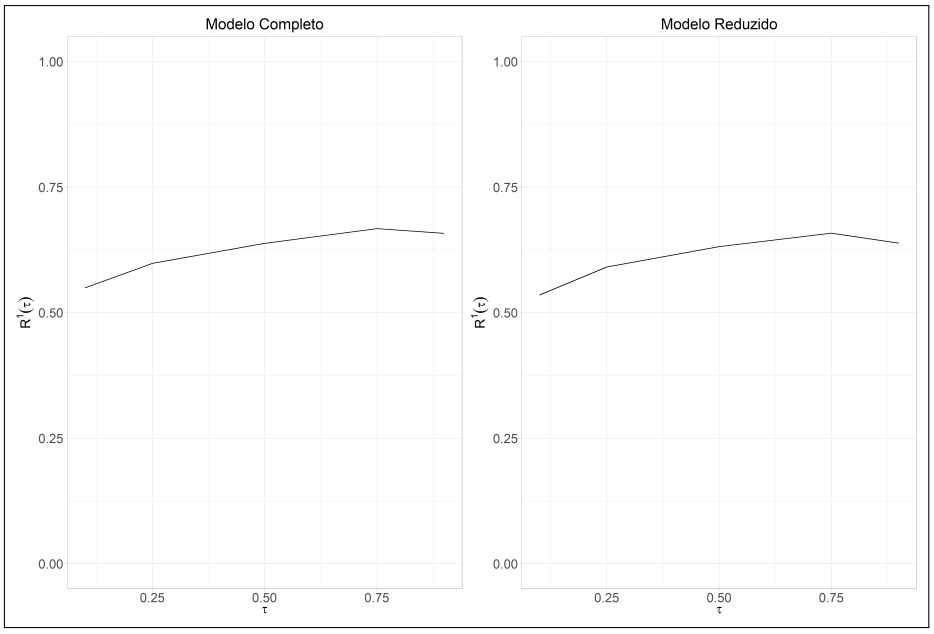
O Modelo Completo mostrado na Equação (3.3) foi estimado para os quantis de $\tau = 0{,}10;\ 0{,}25;\ 0{,}50;\ 0{,}75$ e $0{,}90$ e as estimativas dos seus parâmetros, assim como, os p-valores dos coeficientes são apresentados na Tabela 12, nota-se que a variável categórica que representa a Classificação da Universidade quanto ao seu nível, se apresentou não

Figura 24: Variáveis explicativas quantitativas pela Chance de admissão.



significativa em suas categorias na maioria dos quantis de interesse, a única categoria que foi as universidades de classificação nível 2, entretanto, se estimarmos um modelo apenas com essa categoria e a categoria base da variável, que as universidades classificadas como nível 1, as universidade de nível 2 acabam se tornando não significativas para explicar a chance de admissão do estudante em um programa de mestrado. Outra variável que se mostrou não significativa no nosso modelo foi a variável SOP, que só foi significativa no quantil $\tau = 0,90$. A partir dessas análises, foi estimado um Modelo Reduzido, sem essas variáveis não significativas, as estimativas dos parâmetros deste modelo, bem como os seus p-valores pode ser observado na Tabela 13, podemos notar que os coeficientes deste modelo se apresentaram significativos em quase todos os quantis, exceto pelas variáveis pesquisa no quantil $\tau = 0,25$, GRE no quantil $\tau = 0,90$ e a TOEFL no quantil $\tau = 0,10$. Na Tabela 11 temos os valores do AIC para diferentes quantis em ambos os modelos, nota-se que com exceção do quantil $\tau = 0,90$, o Modelo Reduzido teve valores menores do AIC em comparação ao Modelo Completo, além disso na Figura 25 temos os valores do coeficiente $R^1(\tau)$ em cada um dos quantis para ambos os modelos, podemos verificar que não há diferenças significativas do coeficiente $R^1(\tau)$ entre os modelos, assim, escolhemos o Modelo Reduzido em comparação ao Modelo Completo, pois apresentou uma qualidade de ajuste similar ao outro modelo, mas com menos variáveis, ou seja, mesmo tendo uma simplicidade maior, ele conseguiu apresentar uma qualidade similar ao outro modelo menos simples e de qualidade de ajuste similar. Além disso vale ressaltar o fato de que o quantil $\tau = 0,75$ apresentou um valor de $R^1(0,75) = 0.658$, que foi o maior valor observado dessa métrica

Figura 25: Valores da estatística $R^1(\tau)$ para todos os quantis de interesse em ambos os Modelos.



entre os quantis estudados, ou seja, foi a reta estimada que melhor explicou a variabilidade da chance de admissão pelas variáveis explicativas do Modelo Reduzido.

Tabela 11: Valores do AIC por quantil, em cada um dos modelos estimados.

	AIC por Quantil				
Modelos estimados	0,10	0,25	0,50	0,75	0,90
Modelo Completo	-960,66	-1233,42	-1426,90	-1468,83	-1416,30
Modelo Reduzido	-960,98	-1246,39	-1440,52	-1472,24	-1391,94

Na Figura 26 podemos analisar o comportamento das variáveis GRE e TOEFL, podemos observar que elas tem comportamentos diferentes, enquanto a pontuação do GRE começou com o coeficiente crescendo a medida que o valor do quantil aumentava, porém a partir do quantil $\tau = 0,50$ a estimativa desse coeficiente foi diminuindo, até ser não significativa no quantil $\tau = 0,90$, vale ressaltar que essa variável tem um valor mínimo de 290, ou seja, o mínimo que um estudante tirou foi 290, o que significa que um se observarmos o quantil $\tau = 0,25$ um estudante que tirou 290 pontos no GRE teve a chance de admissão aumentada em 0,87. Já na variável TOEFL, teve o mesmo valor estimado nos quantis $\tau = 0{,}10$; 0,25 e 0,50 e depois teve um comportamento crescente nos quantis $\tau = 0{,}75$ e 0,90, lembrando que se um estudante tirou 100 pontos no teste, olhando para o quantil $\tau = 0,90$, o estudante aumentou a sua chance admissão em um programa de mestrado em 0,5. Assim, temos que ressaltar o fato de que o GRE tem uma capacidade de influenciar maior na chance de admissão do que a TOEFL, devido ao fato dos coeficientes estimados dessas variáveis terem valores próximos e o GRE ter pontuações máximas

Tabela 12: Estimativas para os parâmetros do Modelo Completo nos diferentes quantis (p-valor).

	Quantis				
Parâmetro	0,10	0,25	0,50	0,75	0,90
Pesquisa sim	0,019	0,019	0,028	0,025	0,026
	(0,314)	(0,018)	(0,000)	(0,000)	(0,000)
GRE	0,003	0,003	0,002	0,001	0,001
	(0,050)	(0,000)	(0,000)	(0,021)	(0,028)
TOEFL	0,004	0,002	0,003	0,003	0,004
	(0,042)	(0,004)	(0,000)	(0,000)	(0,000)
Universidade nível 2	-0,080	-0,035	0,007	0,018	0,026
	(0,013)	(0,013)	(0,494)	(0,140)	(0,029)
Universidade nível 3	-0,061	-0,019	0,014	0,030	0,019
	(0,041)	(0,179)	(0,164)	(0,013)	(0,108)
Universidade nível 4	-0,049	-0.020	0,016	0,032	0,020
	(0,095)	(0,197)	(0,225)	(0,026)	(0,182)
Universidade nível 5	-0,040	-0,008	0,032	0,045	0,031
	(0,218)	(0,669)	(0,013)	(0,002)	(0,054)
SOP	-0,007	0,004	0,008	0,008	0,013
	(0,560)	(0,370)	(0,052)	(0,087)	(0,005)
LOR	0,042	0,019	0,007	0,009	0,006
	(0,000)	(0,000)	(0,033)	(0,010)	(0,103)
GPA	0,122	0,120	0,110	0,100	0,084
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)

Tabela 13: Estimativas para os parâmetros do Modelo Reduzido nos diferentes quantis (p-valor).

	Quantis				
Parâmetro	0,10	0,25	0,50	0,75	0,90
Pesquisa sim	0,032	0,017	0,025	0,029	0,021
	(0,040)	(0,059)	(0,001)	(0,000)	(0,005)
GRE	0,002	0,003	0,002	0,001	0,001
	(0,020)	(0,000)	(0,001)	(0,023)	(0,092)
TOEFL	0,003	0,003	0,003	0,004	0,005
	(0,102)	(0,004)	(0,000)	(0,000)	(0,000)
LOR	0,035	0,023	0,016	0,017	0,015
	(0,000)	(0,000)	(0,000)	(0,000)	(0,001)
GPA	0,144	0,125	0,123	0,106	0,094
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)

Figura 26: Estimativas dos coeficientes e intervalo de confiança para as pontuações do GRE e do TOEFL.

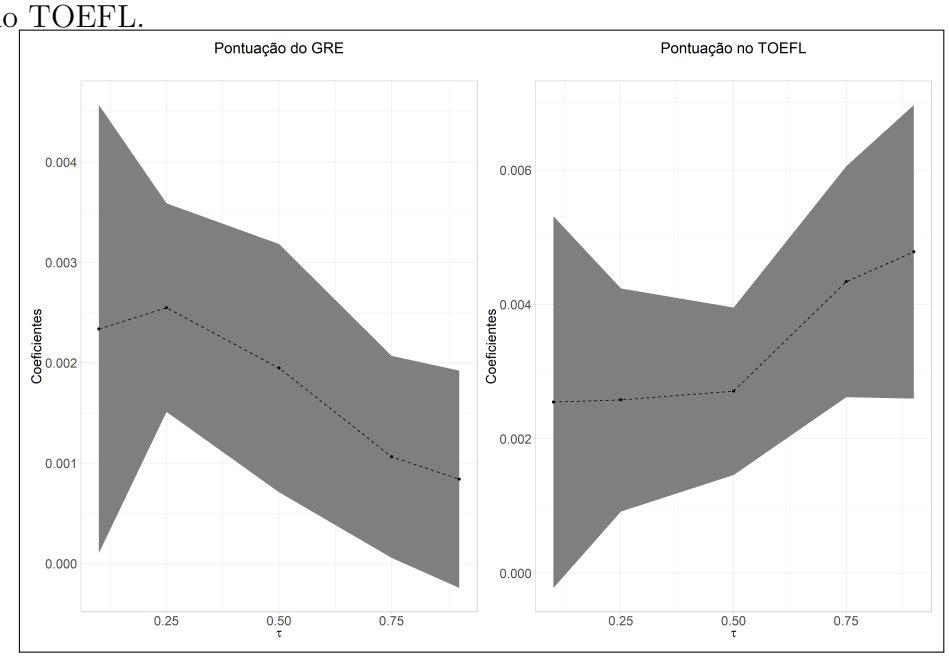


Figura 27: Estimativas dos coeficientes e intervalo de confiança para as pontuações da Carta de recomendação e do GPA.

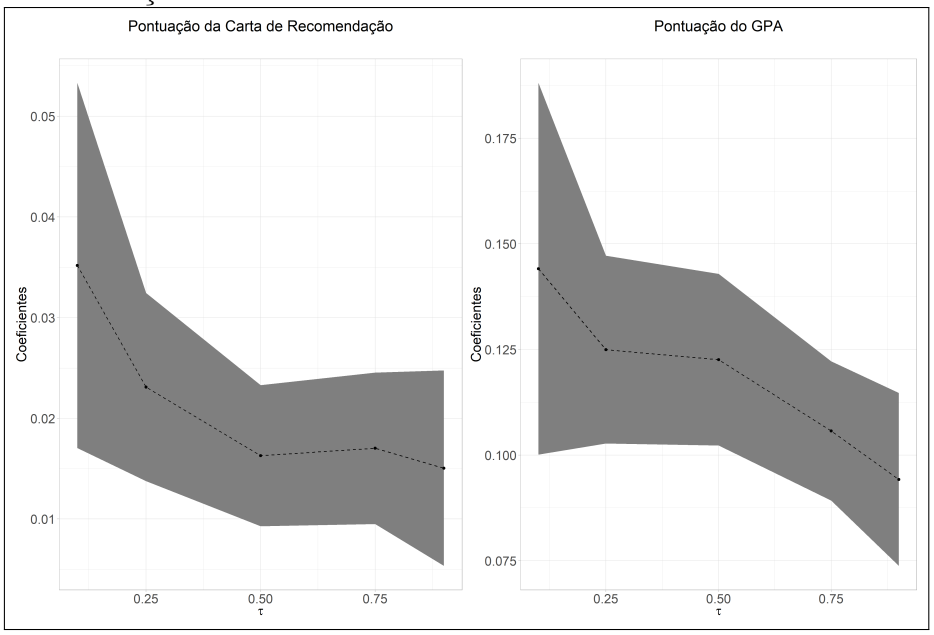
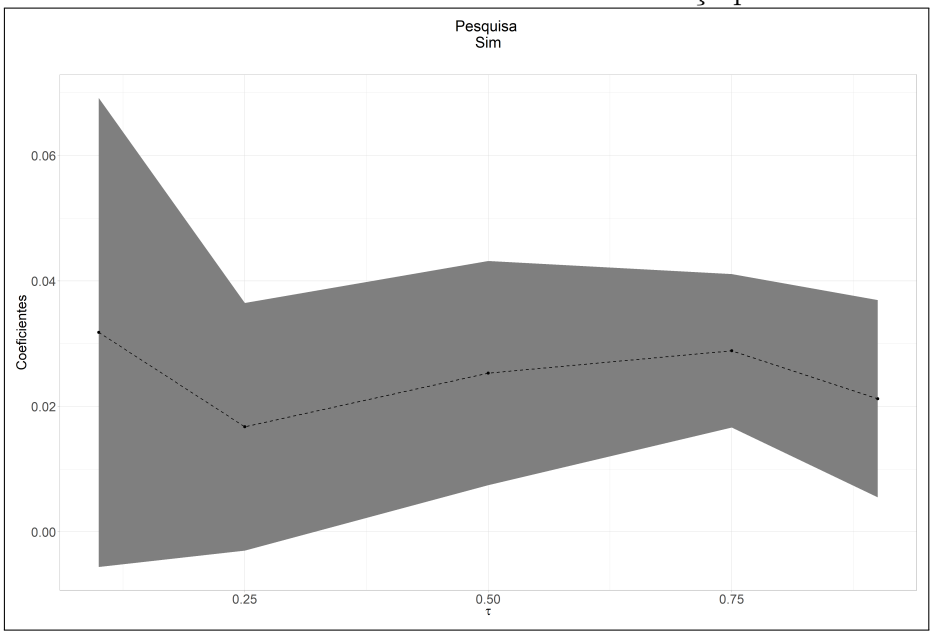


Figura 28: Estimativa dos coeficientes e intervalo de confiança para a variável Pesquisa.



iguais à 340 pontos, enquanto o máximo do TOEFL é 120 pontos.

Na Figura 27 temos as estimativas dos coeficientes das variáveis carta de recomendação (LOR) e a média de pontos das notas (GPA) dos estudantes, ambas as variáveis tiveram um comportamento decrescente em relação aos quantis de interesse, observando a variável LOR no quantil $\tau = 0,10$, temos que podemos aumentar no máximo 0,175 as chances de admissão do estudante. Já a variável GPA é aquela que tem as maiores estimativas do coeficientes, assim como a maior possibilidade de aumentar as chances do estudante ser aceito, que é de 1,43 aproximadamente, isso se olharmos para o quantil $\tau = 0,10$.

Na Figura 28 o comportamento do coeficiente da variável pesquisa nos quantis estudados, vale ressaltar que a categoria base dessa variável é não, ou seja, que o estudante não fez pesquisa, assim, podemos notar que os estudantes que fizeram pesquisa durante a graduação tem mais chances de serem admitidos do que os estudantes que não fizeram pesquisa. Esse coeficiente tem um comportamento que varia de quantil para quantil, mas as maiores chances de quem fez pesquisa em relação à quem não fez é no quantil $\tau = 0,10$, enquanto a menor é no quantil $\tau = 0,25$.

Como comentado anteriormente, vamos fazer uma breve análise sobre os resíduos do nosso modelo quantílico reduzido. Na Figura 29 temos os resíduos quantílicos em função do valor ajustado da chance de admissão para os quantis $\tau =$ 0,10; 0,25; 0,50; 0,75 e 0,90, podemos observar que a dispersão dos resíduos diminui conforme o valor ajustado aumenta, configurando uma heterogeneidade na variância dos erros em todos os quantis

Figura 29: Gráficos dos resíduos quantílicos em função do valor ajustado no Modelo Reduzido nos diferentes quantis.

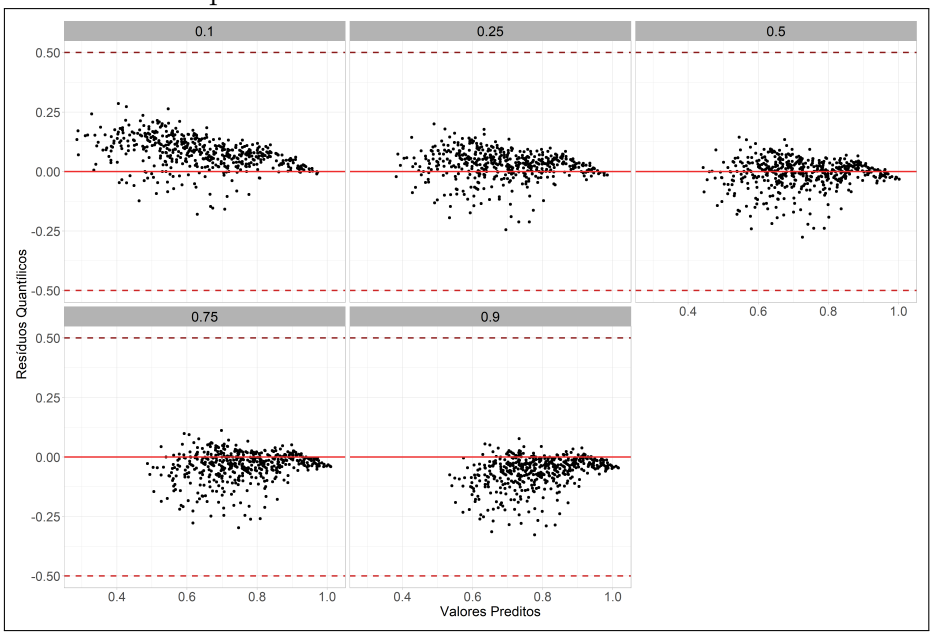
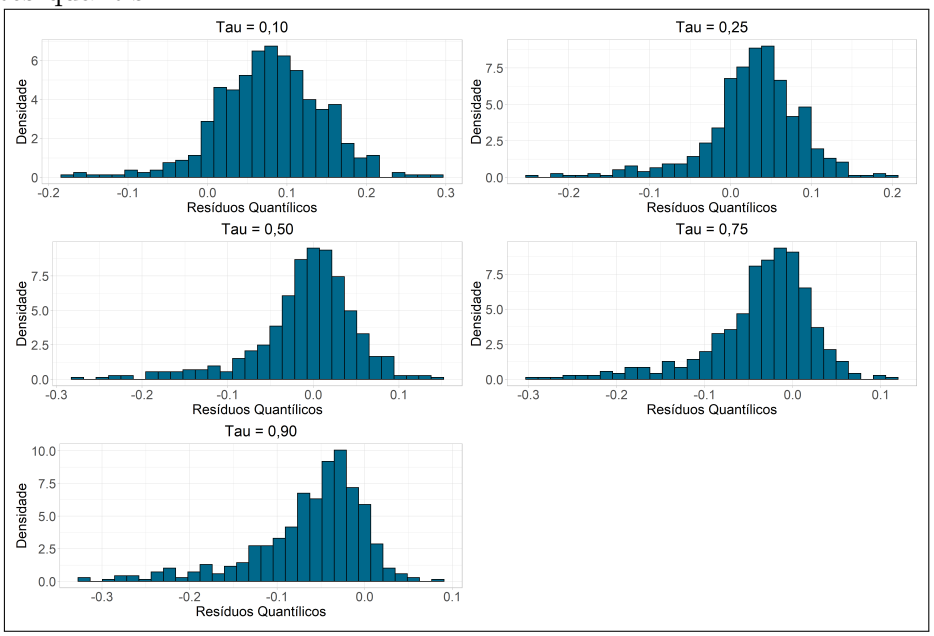


Figura 30: Histograma dos resíduos quantílicos para o Modelo Reduzido ajustado para os diferentes quantis.



observados. Já na Figura 30 podemos ver um histograma para os resíduos quantílicos em cada um dos quantis de interesse, é possível notar que a distribuição dos resíduos é assimétrica, tendo um padrão de assimetria parecido em todos os quantis, com exceção do quantil $\tau = 0,10$, que parece ter um comportamento um pouco mais simétrico do que os outros, ou seja, talvez seja possível utilizar uma mesma distribuição de probabilidade para os resíduos, mas ela não será a priori a distribuição normal.

4 Conclusões

A análise de regressão linear é uma técnica amplamente utilizada para expressar a relação de uma variável de interesse (resposta) e um conjunto de variáveis preditoras relacionada, além de nos permitir realizar previsões da variável de interesse. A análise de regressão mais comum utilizada é a regressão linear clássica, tal expressão é representada por uma equação de primeiro grau, ou seja, uma reta. Os parâmetros deste modelo de regressão são estimados utilizando o método dos mínimos quadrados, entretanto, este método tem algumas limitações, que é fato de possuir uma forte associação com a distribuição normal e a influência que observações atípicas podem causar na estimativa dos parâmetros do modelo de regressão linear.

Devido a essas limitações, foram desenvolvidos durante os anos várias maneiras de se contornar esses problemas, a regressão quantílica surge como uma dessas formas, como a mesma utiliza o método da minimização dos erros absolutos ponderados para estimar os parâmetros do seu modelo, onde tal método é mais robusto a observações atípicas e também não possui forte ligação com a distribuição normal, pelo contrário, ele se ajusta bem a distribuições assimétricas. Além disso, na regressão quantílica podemos estimar mais de uma reta para os dados, já que ao invés da regressão linear que estima a reta em relação a média da distribuição condicional da variável de interesse por suas preditoras, a regressão quantílica estima retas para qualquer um dos quantis da distribuição condicional, assim podemos estimar mais de uma reta para um mesmo conjunto de dados, obtendo uma visão mais ampla e completa da relação da variável de interesse com suas preditoras.

Assim, o foco desta monografia foi explorar os modelos de regressão quantílica, estimando retas para os diferentes quantis, além de utilizar variáveis respostas com distribuição assimétrica, para mostrar a vantagem em se utilizar esses modelos em relação ao modelo de regressão linear clássico. Desta maneira, buscando desenvolver um maior debate sobre a regressão quantílica, devido não só a escassez de estudos sobre o tema, como também a grande parte dos estudos e matérias de referência estarem em inglês como o livro de Koenker (2005), entendemos que esse trabalho possa ajudar na disseminação e

na motivação para novos estudos envolvendo os modelos de regressão quantílica, como o trabalho de Santos (2012).

No primeiro estudo aplicado realizado nesta monografia, utilizamos dados sobre a renda dos brasileiros, que foram obtidos através da PNAD contínua de 2020, além de informações socioeconômicas da população brasileira. Nesse estudo comparamos o desempenho de ambos os modelos, o linear e o quantílico, onde o modelo de regressão linear apresentou dificuldade em cumprir os pressupostos necessários para utilizarmos a análise em si, em contrapartida a regressão quantílica conseguiu modelar os dados, já que não necessita supor nem normalidade e nem homocedasticidade dos resíduos do modelo, ou seja, como nossa a variável renda apresentava muitos valores atípicos além de possuir uma forte assimetria à esquerda, a regressão quantílica se mostrou mais interessante para expressar a relação da Renda com as variáveis socioeconômicas observadas. Também observamos essa relação para diferentes quantis de interesse $\tau = 0{,}10$; 0,25; 0,50; 0,75 e 0,90, onde foi possível notar que através da medida de qualidade $R^1(\tau)$ que as retas para cada um dos quantis tinha uma capacidade do modelo de explicar a relação crescente a medida que o valor do quantil aumentava.

No segundo estudo, analisamos as chances de admissão de um estudante indiano em um programa de Mestrado e quais características influenciavam para aumentar essas chances. Construímos retas para diferentes quantis, depois realizando uma breve análise sobre os resíduos de cada uma dessas retas estimadas, observando que os mesmos não era simétricos, além de nem todas as retas serem identicamente distribuídas.

O custo computacional para gerar esses modelos é um possível crítica a esse método, já que devido ao fato do método da minimização dos erros absolutos ponderados não nos permitir obter os estimadores dos parâmetros do modelo de forma analítica, mesmo usando métodos de programação linear o tempo para a estimação dos parâmetros é consideravelmente longo, se comparado a regressão linear clássica.

Uma possível transformação na variável resposta no modelo de regressão quantílica pode melhorar ainda mais o ajuste do modelo, apesar de não apresentada nesta monografia, um modelo usando o logaritmo da Renda como variável reposta foi estimado e mostrou resultados interessantes, todavia, vale ressaltar que a interpretação do resultado no modelo de regressão quantílica difere um pouco nesses casos da interpretação do modelo de regressão linear.

Durante o desenvolvimento desta monografia foi observado que existem ainda muitos temas para possíveis trabalhos relacionados a regressão quantílica, como modelos de

regressão quantílica não-lineares, bayesianos e localmente polinomial.

Por fim, o modelo de regressão quantílica se mostrou melhor que o de regressão linear para variáveis assimétricas, além de conseguir obter uma visão mais ampla da relação da variável de interesse com suas preditoras. Logo, considera-se que foi possível explorar todos os objetivos propostos no início deste trabalho.

APÊNDICE A

A.1 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov (KS) é um teste que busca checar se os dados se aderem a alguma distribuição de probabilidade, no caso do nosso estudo a distribuição normal, ou seja, queremos verificar se os dados a serem estudados são provenientes de uma determinada distribuição de probabilidade. Este teste se baseia na função de distribuição empírica, isto é, o teste compara a função de distribuição empírica com a função de distribuição acumulada de uma determinada distribuição de probabilidade.

Dado uma função de distribuição empírica $S(x)$ e $F(x)$ a função de distribuição acumulada de alguma distribuição de probabilidade, a estatística de teste para o teste de Kolmogorov-Smirnov será,

$$D_n = sup_x(|F(x) - S(x)|),$$

onde as hipóteses do teste KS são:

$$H0 : F(x) = S(x);$$
$$H1 : F(x) \neq S(x),$$

isto é, a hipótese nula (H0) é de que a distribuição empírica é igual a função de distribuição acumulada de alguma distribuição de probabilidade, no caso deste trabalho, a normal, ou seja, os dados são provenientes de uma normal. Já a hipótese alternativa (H1) será a de que a função empírica é diferentes da função de distribuição acumulada, ou seja, os dados não são provenientes de uma normal. Para mais informações, consultar o trabalho Miot (2017).

A.2 Teste de Breusch-Pagan

O teste de Breusch-Pagan é utilizado para testar a hipótese de homocedasticidade dos erros de um modelo de regressão linear. Se a variância dos erro é constante, então temos homocedasticidade; por outro lado, se ela variar, temos heterocedasticidade. Então as Hipóteses do teste são,

$$H0 : \text{Variância dos erros é constante;}$$
$$H1 : \text{Variância dos erros não é constante,}$$

se a estatística $\chi^2 \geq \chi_{(1-\alpha;N-1)}$, então rejeitaremos H0. Para mais informações, consultar o estudo de Kutner et al. (2005).

APÊNDICE B

B.1 O critério de informação de Akaike (AIC)

O critério de informação de Akaike (AIC) é uma métrica muito popular que mensura a qualidade do ajuste de um modelo estatístico, e que também visa a simplicidade do modelo, isto é, uma medida de comparação e seleção de modelos, em que quanto menor o valor do *AIC*, maior será a qualidade do ajuste e o mais simplificado será o modelo.

O *AIC* é fundamentado na teoria da informação, que demonstra a relação entre a verossimilhança e a quantidade de informação que se perde quando aproximamos os dados com um modelo. Quanto menos informações o modelo perde, menor será o *AIC* do modelo, ou seja, o melhor modelo é o que aproxima os dados com a menor perda de informação. O *AIC*, como comentado acima é calculado baseado na verossimilhança, com penalização baseada no número de parâmetros do modelo estimado, sendo ela calculada através da seguinte expressão,

$$AIC = -2l(\hat{\beta}; y) + 2p,$$

onde p é o número de parâmetros estimados pelo modelo e $l(\hat{\beta}; y)$ é a função de log-verossimilhança dos parâmetros do modelo.

Referências

ACHARYA, M. S.; ARMAAN, A.; ANTONY, A. S. A comparison of regression models for prediction of graduate admissions. In: IEEE. *2019 international conference on computational intelligence in data science (ICCIDS)*. [S.l.], 2019. p. 1–5.

ANDRÉ, C. D. et al. Coefficients of determinations for variable selection in the msae regression. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 29, n. 3, p. 623–642, 2000.

BARRODALE, I.; ROBERTS, F. D. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, SIAM, v. 10, n. 5, p. 839–848, 1973.

BASSETT, G.; KOENKER, R. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 73, n. 363, p. 618–622, 1978.

CHEN, C.; WEI, Y. Computational issues for quantile regression. *Sankhyā: The Indian Journal of Statistics*, JSTOR, p. 399–417, 2005.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2018.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994.

HAND, D. J. et al. *A handbook of small data sets*. [S.l.]: cRc Press, 1993.

HAO, L.; NAIMAN, D. Q. *Quantile regression*. [S.l.]: Sage, 2007.

HE, X.; HU, F. Markov chain marginal bootstrap. *Journal of the American Statistical Association*, Taylor & Francis, v. 97, n. 459, p. 783–795, 2002.

KOCHERGINSKY, M.; HE, X.; MU, Y. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 14, n. 1, p. 41–55, 2005.

KOENKER, R. *Quantile Regression*. [S.l.]: Cambridge University Press, 2005. (Econometric Society monographs 38).

KOENKER, R.; BASSETT, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, JSTOR, p. 33–50, 1978.

KOENKER, R.; MACHADO, J. A. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, Taylor & Francis, v. 94, n. 448, p. 1296–1310, 1999.

KOENKER, R. et al. Package 'quantreg'. *Cran R-project. org*, 2018.

KOENKER, R. W.; D'OREY, V. Algorithm as 229: Computing regression quantiles. *Applied statistics*, JSTOR, p. 383–393, 1987.

KUTNER, M. et al. *Applied linear statistical models.,(McGraw-Hill Irwin: New York)*. [S.l.]: NY, 2005.

MCKEAN, J. W.; SIEVERS, G. L. Coefficients of determination for least absolute deviation analysis. *Statistics & Probability Letters*, Elsevier, v. 5, n. 1, p. 49–54, 1987.

MIOT, H. A. *Avaliação da normalidade dos dados em estudos clínicos e experimentais*. [S.l.]: SciELO Brasil, 2017. 88–91 p.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021.

PORTNOY, S.; KOENKER, R. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, Institute of Mathematical Statistics, v. 12, n. 4, p. 279–300, 1997.

SANTOS, B. R. d. *Modelos de regressão quantílica*. Tese (Doutorado) — Universidade de São Paulo, 2012.

WALD, A. Statistical decision functions. Wiley, 1950.