Universidade de Brasília

IE - Departamento de Estatística

Trabalho de Conclusão de Curso 2

Um Estudo sobre Regressão Quantílica

Adolfo Manoel Dias da Silva

Relatório Final

Orientador: Profº. Dr. Antônio Eduardo Gomes

Brasília

18 de dezembro de 2018

Adolfo Manoel Dias da Silva

Um Estudo sobre Regressão Quantílica

Orientador: Profº. Dr. Antônio Eduardo Gomes

Brasília

18 de dezembro de 2018

Sumário

1 INTRODUÇÃO

A teoria de Regressão originou-se em meados do século XVIII com Galton. Em um de seus trabalhos, ele estudou a relação entre as variáveis altura dos pais e dos filhos a fim de entender como a primeira variável influenciava na segunda. E nesse estudo, observou que se o pai fosse muito alto ou baixo demais, o filho teria uma altura tendendo à media. Esse fato foi chamado por ele de regressão para indicar que há uma tendência dos dados regredirem à media.

A regressão por mínimos quadrados ordinários (mqo) era bastante utilizada por fornecer boas estimativas sob certas condições e por permitir um tratamento analítico, ou seja, sem a necessidade do recurso computacional. Desse modo, ganhava destaque já que os computadores da época tinham a capacidade de processamento bem limitada, a ponto de não permitir o desenvolvimento de outros métodos estatísticos. Mas, com a evolução da tecnologia da informação, diversas técnicas começaram a emergir e ter visibilidade. Uma delas foi a regressão quantílica. Um dos pioneiros dessa técnica foi Boscovich(1760). Seus estudos eram sobre a forma elíptica da terra associados à ideia de mediana. Entretanto, os responsáveis pela maior visibilidade foram Koenker e Basset(1978).

A curva de regressão fornece um grande resumo das médias das distribuições condicionais dados os valores observados das covariáveis. Poderíamos ir além e computarmos várias curvas de regressão associadas a muitos pontos percentuais da distribuição condicional e desse modo ter uma visão mais completa do conjunto de dados. Mas, usualmente, isso não é realizado, e logo a regressão linear usual dá uma visão incompleta. De forma análoga, assim como a média dá uma visão incompleta de uma única distribuição, a curva de regressão dá uma visão incompleta de um conjunto de distribuições.Koenker and Bassett Jr [1978]

Esse problema foi solucionado pela utilização de Regressão Quantílica com a colaboração dos pesquisadores Koenker e Basset(1978). Essa técnica ganhou importância e, com o grande avanço obtido a partir da evolução dos computadores e do uso de programação linear, começou a ser aplicada, de forma mais intensiva, em várias pesquisas nas áreas de economia, finanças, ecologia e medicina. Verificaram que seria possível entender melhor a relação entre variáveis explicativas e resposta, em diversos quantis condicionais e não somente em relação à média condicional, como no método por mínimos quadrados ordinários (mqo). Outras vantagens seriam: robustez a outliers, dispensa de distribuição paramétrica e aplicabilidade na presença de heterocedasticidade.

No presente trabalho, realizaremos um estudo sobre Regressão Quantílica, com abordagens paramétrica e não paramétrica mediante a utilização do software estatístico R. Nesse, encontra-se disponível o pacote **quantreg** usado para estimar e fazer inferência sobre as funções quantílicas condicionais.

A organização será feita em três partes, além da introdução e das considerações finais. Na primeira parte, descreveremos os procedimentos metodológicos e a fonte de dados. Na segunda, apresentaremos os aspectos teóricos da técnica e, na terceira, faremos a comparação dos resultados.

2 OBJETIVOS

2.1 Objetivos Gerais

Estudar a técnica de *Regressão Quantílica*.

2.2 Objetivos Específicos

- Apresentar os aspectos teóricos da Regressão Quantílica.
- Implementar a técnica utilizando o software R.
- Comparar as regressões quantílica e clássica.
- Descrever a abordagem não paramétrica com o auxílio do estimador Kernel.
- Analisar e interpretar os resultados obtidos para diferentes quantis.

3 METODOLOGIA

3.1 Regressão Quantílica

A regressão quantílica está emergindo de forma gradual como uma abordagem abrangente para a análise estatística de modelos de resposta não linear. Complementa os métodos baseados em mínimos quadrados na estimação de funções de médias condicionais com uma técnica geral para estimar famílias de funções quantílicas condicionais e capaz de expandir enormemente a flexibilidade dos métodos de regressão paramétrica e não paramétrica.

O método baseia-se na minimização dos erros absolutos ponderados. Para entendê-lo, partiremos da comparação entre média e quantil, pois a regressão quantílica generaliza os quantis univariados para a distribuição condicional. Em diversos estudos, a comparação entre média e mediana é padrão para detectar se a distribuição e símetrica ou assimétrica. Algumas definições serão apresentadas a seguir:

Definição 3.1.1 (Média)**.** *Seja Y uma variável aleatória genérica. A média μ de Y é o ponto q de sua distribuição tal que*

$$S_\tau(q) = \sum_{i=1}^{n}(y_i - q)^2$$

assuma valor mínimo, ou seja,

$$\mu = \underset{q}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - q)^2$$

Definição 3.1.2 (Mediana)**.** *A mediana m da distribuição Y é o valor de q que minimiza a soma dos desvios absolutos:*

$$m = \underset{q}{\operatorname{argmin}}\, E|Y - q|$$

Definição 3.1.3 (Quantil)**.** *O τ-ésimo quantil da distribuição de Y é o menor valor de y tal que:*

$$F(y) = P(Y \leq y) = \tau$$

Dado um conjunto de observações, o τ-ésimo quantil é o valor y para o qual pelo menos 100τ % das observações assumem valores menores ou iguais a y e no máximo $100(1-\tau)$ % assumem valores maiores ou iguais a y.

A função quantil é definida como sendo a inversa da função de distribuição acumulada de Y:

$$Q_Y(\tau) = {F_Y}^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$$

Em problemas de minimização para o caso discreto, o τ-ésimo quantil poderá ser definido como o ponto central da distribuição de Y que minimiza a soma ponderada dos desvios:

$$q_\tau = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - c)$$

em que $\rho_\tau(.)$ é a função perda definida por:

$$\rho_\tau(u) = u(\tau - I(u \leq 0)) \qquad \text{em que } I(.) \text{ é a função indicadora.}$$

Para o caso contínuo, temos que procurar um q $\in$ R que minimiza a perda esperada $E(\rho_\tau(X - q))$. Mostraremos a seguir como encontrá-lo:

Utilizando a definição de esperança matemática, segue que:

$$\begin{aligned}
E(\rho_\tau(X-q)) &= \int_{-\infty}^{\infty} \rho_\tau(x-q)dF(x) \\
&= \int_{-\infty}^{q} \rho_\tau(x-q)dF(x) + \int_{q}^{\infty} \rho_\tau(x-q)dF(x) \\
&= \int_{-\infty}^{q} (x-q)(\tau-1)dF(x) + \int_{q}^{\infty} (x-q)\tau dF(x) \\
&= (\tau-1)\int_{-\infty}^{q} (x-q)dF(x) + \tau\int_{q}^{\infty} (x-q)dF(x) \\
&= \tau\left[\left(\int_{-\infty}^{q} xdF(x) + \int_{q}^{\infty} xdF(x)\right) - \left(\int_{-\infty}^{q} qdF(x) + \int_{q}^{\infty} qdF(x)\right)\right] \\
&- \int_{-\infty}^{q} xdF(x) + \int_{-\infty}^{q} qdF(x) \\
&= \tau\int_{-\infty}^{\infty} xdF(x) - \tau q\int_{-\infty}^{\infty} dF(x) - \int_{-\infty}^{q} xdF(x) + q\int_{-\infty}^{q} dF(x) \\
&= \tau\int_{-\infty}^{\infty} xdF(x) - \tau q - \int_{-\infty}^{q} xdF(x) + qF(q)
\end{aligned}$$

Supondo que a esperança matemática da variável aleatória X exista, encontra-se o ponto q que minimiza a função perda esperada, diferenciando a última equação obtida, em relação a q, e igualando a zero, dado que a 2ª derivada é positiva para todos valores reais de q, conforme abaixo:

$$\frac{\partial}{\partial q}\Big(E(\rho_\tau(X-q)\Big) = \frac{\partial}{\partial q}\Big(\tau\int_{-\infty}^{\infty} xdF(x) - \tau q - \int_{-\infty}^{q} xdF(x) + qF(q)\Big)$$

$$= \underbrace{\frac{\partial}{\partial q}\Big(\tau\int_{-\infty}^{\infty} xdF(x)\Big)}_{\text{zero (independe de q)}} - \underbrace{\frac{\partial}{\partial q}(\tau q)}_{\tau} - \underbrace{\frac{\partial}{\partial q}\left(\int_{-\infty}^{q} xdF(x)\right)}_{qf(q)} + \underbrace{\frac{\partial}{\partial q}\Big(qF(q)\Big)}_{F(q)+qf(q)}$$

$$\frac{\partial}{\partial q}\Big(E(\rho_\tau(X-q)\Big) = 0 \Rightarrow F(q) - \tau = 0$$

Se a função de distribuição acumulada $F(\cdot)$ for conhecida e possuir inversa, tomemos o menor valor de $q = F^{-1}(\tau)$ que satisfaz a última equação obtida. Caso a $F(\cdot)$ seja desconhecida, estimamos empiricamente.

Segue abaixo uma ilustração gráfica mostrando a contribuição de cada realização da variável aleatória U para função perda. Foram simuladas 1000 observações de uma distribuição uniforme $U = [-1, 1]$. Através dos gráficos, conseguiremos verificar o valor da função perda e a taxa de contribuição em cada ponto. Para valores não positivos de argumento, o peso será $\tau - 1$ e, para os positivos, será τ.

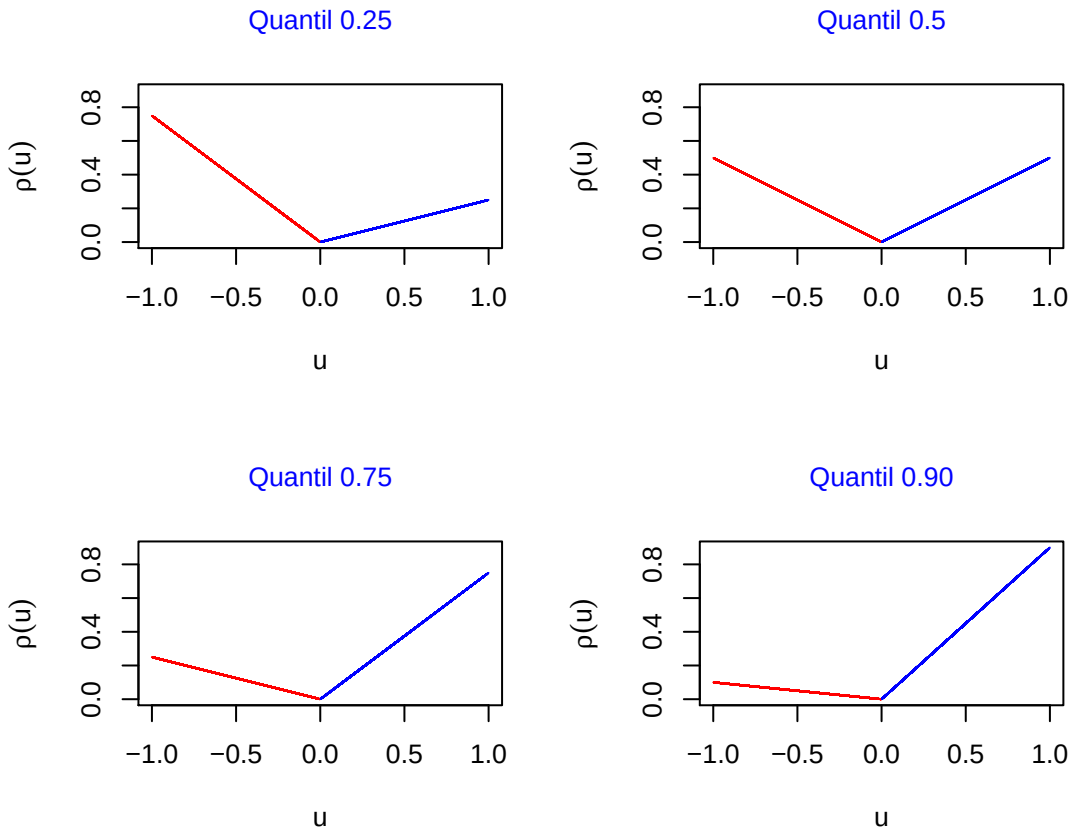


Figura 1: Gráficos da função perda para $\tau \in \{0.25, 0.5, 0.75, 0.9\}$

Exemplo 3.1.1. *Em teoria da decisão, estaremos interessados em prever um valor da variável aleatória Y com função distribuição de probabilidade acumulada F. Queremos encontrar um previsor $\hat{y}$ de Y que minimize a perda esperada:*

$$E(\rho_\tau(Y - \hat{y})) = (\tau - 1) \int_{-\infty}^{\hat{y}} (y - \hat{y}) dF(y) + \tau \int_{\hat{y}}^{\infty} (y - \hat{y}) dF(y)$$

Diferenciando com relação ao previsor $\hat{y}$, resulta:

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

Quando a solução é unica, tem-se $\hat{y} = F^{-1}(\tau)$. Caso contrário , um intervalo para os τ-ésimo quantis é obtido como solução.

O gráfico abaixo representa a soma agregada referente aos quartis. Foi construído a partir da geração de uma amostra aleatória da distribuição normal padrão. Pela sua simetria, a mediana coincide com a média. Podemos observar que, para a curva referente ao 2^o quartil, a soma $S_\tau(.)$ é minimizada quando $q = 0$,

como era esperado.

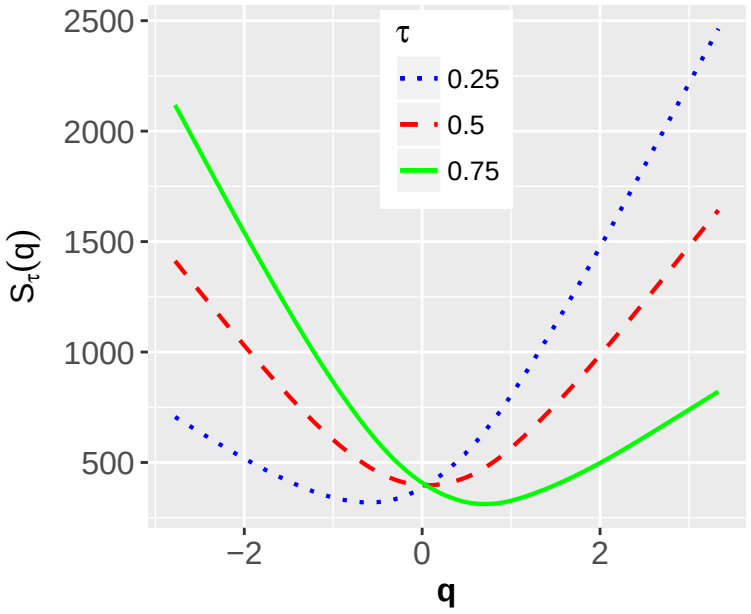


Figura 2: Exemplo da perda agregada

Para o caso em que a função de distribuição acumulada é desconhecida, podemos estimá-la através da função de distribuição empírica.

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

Com o intuito de obter o menor valor para perda esperada, procuremos o previsor $\hat{y}$ que minimize a soma:

$$S_n(\hat{y}) = \sum_{i=1}^{n} \rho_\theta(y_i - \hat{y})$$

3.1.1 Modelo Quantílico Linear

Pode-se generalizar as propriedades dos quantis incondicionais para os condicionais. Vimos que a média incondicional minimiza a soma esperada segundo Davino et al. [2014]. Para o caso geral, essa soma também é minimizada pela média condicional. No modelo de regressão quantílica linear, o quantil condicional é especificado pela seguinte equação:

$$Q_{Y_i}(\tau|X = x_i) = x_i^T \beta(\tau) \tag{1}$$

Para o j-ésimo preditor, o efeito marginal é o coeficiente, para o τ-ésimo quantil, dado por:

$$\beta_j(\tau) = \frac{\partial Q_\tau}{\partial x_j}(y|x) \tag{2}$$

em que $\beta(\tau)$ é o vetor de parâmetros e x_i é o vetor de preditores.

As propriedades para o quantil incondicional são válidas para o condicional. Dada uma amostra com n observações,$\{(X_i, Y_i), i = 1, \ldots, n\}$, de um modelo linear $Q_Y(\tau|X = x) = x^T\beta(\tau)$, a τ-ésima estimativa do coeficiente de regressão quantílica é:

$$\hat{\beta}(\tau) = \underset{\beta(\tau)\in R^p}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^T\beta(\tau)\Big\} \tag{3}$$

Utilizando o conjunto de dados *Segreg* do pacote alr4 do software R, ajustamos, para exemplificar, os modelos de regressão quantílicas considerando as ordens $\tau \in \{0.10, 0.25, 0.5, 0.75, 0.90\}$, além do modelo de regressão por mqo. Desse modo, temos uma figura mais completa da distribuição dos dados.

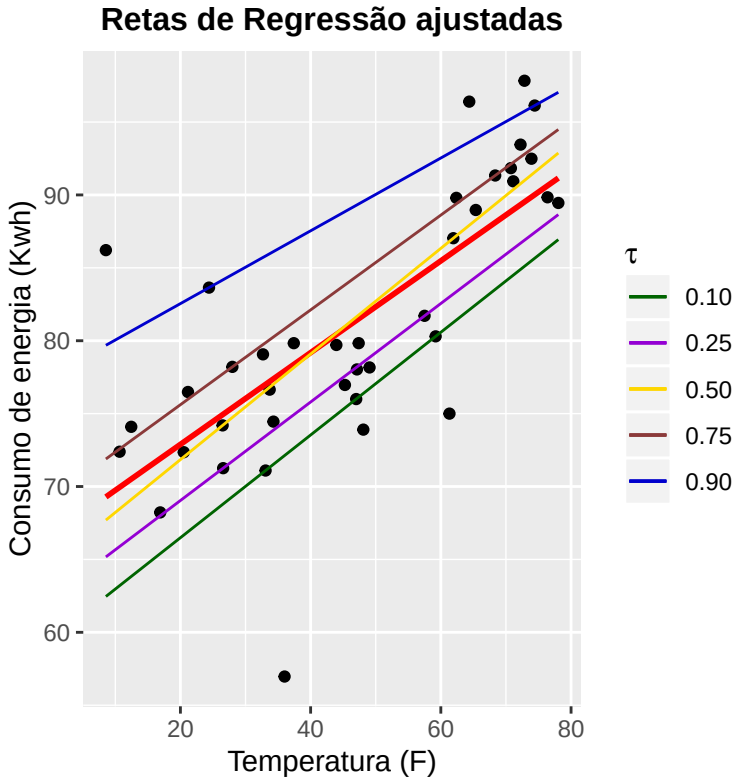


Figura 3: Regressões Quantílica para $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ e por mqo

3.1.2 Aspectos Inferenciais da Regressão Quantílica

A regressão quantílica fornece estimadores mais robustos e eficientes comparada com a regressão por mínimos quadrados ordinários e não necessita de suposições acerca da parte aleatória do modelo.

Considere o seguinte função de τ;

$$\hat{\beta}(\tau) = \underset{\beta(\tau)\in R^p}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^T \beta(\tau) \Big\} \tag{4}$$

Os estimadores da regressão quantílica são consistentes, isto é, $\mid \hat{\beta}_n(\tau) - \beta(\tau) \mid \to 0$ em probabilidade quando $n \to \infty$ e assume as seguintes condições de reguralidade:

1. As funções inversas dos quantis condicionais ${Q_\tau}^{-1}(Y|x_j)$ são absolutamente contínuas com densidade contínuas $f(Y|x_j)$.

2. Existem matrizes positivas definidas A_0 e A_1 tais que:

   - $n^{-1} \sum_{j=1}^{n} x_j {x_j}^T \to A_0$ quando $n \to \infty$
   - $n^{-1} \sum_{j=1}^{n} {f_j}^2(F^{-1}(\tau) x_j x_j^T \to A_1$ quando $n \to \infty$
   - $\mid x_{(n)} \mid / n^{1/2} \to 0$

Sob as condições mencionadas, determina-se a distribuição assintótica do estimador $\hat{\beta}(\tau)$, primeiramente, quando os erros são independentes e identicamente distribuídos:

$$\sqrt{n}\Big(\hat{\beta}_n(\tau) - \beta(\tau)\Big) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau)} A_0^{-1}\right) \tag{5}$$

E, segundo, quando temos erros independentes, porém não são identicamente distribuídos:

$$\sqrt{n}\Big(\hat{\beta}_n(\tau) - \beta(\tau)\Big) \xrightarrow{d} N\Big(0, \tau(1-\tau) A_1^{-1} A_0 A_1^{-1}\Big) \tag{6}$$

Dados os quantis τ_i e τ_j, a matriz de covariância assintótica é dada por:

$$Cov_a\left(\sqrt{n}\left(\hat{\beta}_n(\tau_i) - \beta(\tau_i\right), \sqrt{n}\left(\hat{\beta}_n(\tau_j) - \beta(\tau_j)\right)\right) = \Big(\tau_i \wedge \tau_j - \tau_i \tau_j\Big) A_1(\tau_i)^{-1} A_0 A_1(\tau_j)^{-1} \tag{7}$$

Agora, para fazer inferências com base na distribuição assintótica da regressão quantílica, devemos estimar a matriz de variância. Assumindo que os erros sejam independentes e identicamente distribuídos, a matriz de variância é dada por:

$$var\left(\sqrt{n}\hat{\beta}(\tau)\right) = \frac{\tau(1-\tau}{\hat{f}^2\Big(F^{-1}(\tau)\Big)}\hat{A}_0^{-1}, \tag{8}$$

em que $\hat{A}_0 = n^{-1}\sum_{j=1}^{n} x_j {x_j}^T$ e a função de dispersão $d(\tau) = \Big(f(F^{-1}(\tau)\Big)^{-1}$.

A estimação da função de dispersão é obtida usando a diferença das funções de distribuições empíricas:

$$\hat{d}_n(\tau) = \frac{\hat{F}_n{}^{-1}(\tau + h_n|\bar{x}) - \hat{F}_n{}^{-1}(\tau - h_n|\bar{x})}{2h_n} \tag{9}$$

em que $\hat{F}_n{}^{-1}(\tau|\bar{x})$ é a estimativa do quantil condicional de Y dado a média amostral $\bar{x}$ e h_n é p parâmetro de suavização tal que $h_n \to 0$ quando $n \to \infty$.

E, quando os erros são não iid, a matriz de covariância é dada por:

$$var\left(\sqrt{n}\hat{\beta}(\tau)\right) = \tau(1-\tau)\hat{A}_1{}^{-1}\hat{A}_0\hat{A}_1{}^{-1}, \tag{10}$$

onde $\hat{A}_1 = n^{-1}\sum_{j=1}^{n}\Big(\hat{f}_j{}^2(F^{-1}(\tau)\Big)x_j{x_j}^T$ e,

$$\hat{f}_j\Big(F^{-1}(\tau)\Big) = \frac{2h_n}{{x_j}^T\hat{\beta}(\tau + h_n) - {x_j}^T\hat{\beta}(\tau - h_n)}$$

3.1.3 Transformações da variável resposta

No modelo de regressão quantílica, podemos modificar a variável resposta por transformação monótona. Essa propriedade, chamada de equivariância, refere-se à capacidade de usar as mesmas regras de interpretação quando os dados ou modelo estão sujeitos a uma transformação.

3.1.4 Método para determinar desvio padrão, p-valor e ICs

Foi utilizado o software R para determinação dos desvios-padrão, p-valor e intervalos de confiança das tabelas da seção 4. A teoria que dá suporte aos códigos baseia-se no método de programação linear simplex que encontra-se disponível em Davino et al. [2014].

3.1.5 Estimação não paramétrica

Os métodos não paramétricos de estimação de funções de densidade de probabilidade têm-se tornado ferramentas sofisticadas e alternativas para o tratamento de dados. Podem ser aplicados sem fazer qualquer condição restritiva sobre a forma de

uma função desconhecida. Devido a essa maior flexibilidade em relação aos métodos paramétricos, têm sido muito utilizados em estudos estatísticos. Utilizaremos nesse trabalho a estimação por Kernel.

Seja $\{X_i\}_{i=1}^n$ uma amostra aleatória obtida de uma população da qual desconhecemos a forma da distribuição dos dados. Considere a seguinte função indicadora:

$$I(x - X_i) = \begin{cases} 1 & se \quad x - X_i \geq 0 \\ 0 & se \quad x - X_i < 0 \end{cases} \tag{11}$$

$I(x - X_i)$ é uma variável aleatória que segue uma distribuição de Bernoulli cuja probabilidade de sucesso é dada por $F(x)$. A função de distribuição acumulada empírica é:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x - X_i \geq 0) \tag{12}$$

O estimador $\hat{F}$ é uma função escada que apresenta descontinuidade nos pontos $\{X_i\}$, chamados pontos de salto. Substituindo a função indicadora acima por uma função de classe C^1, de modo que $\hat{F}$ tenha as propriedades de uma função de distribuição acumulada, obtemos o estimador Kernel de $F(x)$, representado por:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\frac{(x-X_i)}{h}} K(u)du \tag{13}$$

em que h é o parâmetro de suavização e a $K(\cdot)$ é uma função tal que:

$$\int_{-\infty}^{\infty} K(u)du = 1 \quad ; \tag{14}$$

$$\int_{-\infty}^{\infty} uK(u)du = 0 \quad ; \tag{15}$$

$$K(-u) = K(u) \quad \text{(função com paridade par)} \quad \text{e} \tag{16}$$

$$K(u) \geq 0 \quad \text{(não negatividade).} \tag{17}$$

3.1.6 Estimação não paramétrica da média condicional

O estimador abaixo, que se baseia na função Kernel, foi proposto por Nadaraya e Watson (1964), segundo Simonoff [1996], para estimação não paramétrica da média condicional da distribuição de Y dado x.

$$\hat{m}_{nw}(x) = \sum_{i=1}^{n} \frac{K(\frac{x-X_i}{h})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})} Y_i \tag{18}$$

em que h e $K(\cdot)$ foram definidos anteriormente.

Dada uma amostra $\{X_i, Y_i\}_{i=1}^{n}$, ao avaliar esse estimador em um conjunto de pontos, obteremos uma curva.

O parâmetro h serve para controlar a suavização dessa curva. Sua estimação é obtida por meio de validação cruzada ou mediante erro quadrático médio integrado. A estimativa por essa última técnica é obtida encontrando o valor de h que minimiza o erro quadrático médio integrado (MISE) e, quanto maior o valor de h valor, mais suave será a curva.

Na subseção, descreveremos um algoritmo geral para estimação não paramétrica dos quantis condicionais.

3.1.7 Estimação não paramétrica de densidades

Podemos também estimar as densidades. Em estatística, a estimativa via Kernel da função densidade é uma forma não paramétrica de estimação. Seguem, abaixo, os estimatidores via Kernel das funções densidade marginal de Y, densidade condicional de Y dado a realização da variável X e função de distribuição acumulada condicional.

- Estimador de Kernel para densidade marginal da variável resposta Y

$$\hat{f_Y}(y) = \frac{1}{nh_y} \sum_{i=1}^{n} K_y\Big(\frac{y - Y_i}{h_y}\Big) \tag{19}$$

- Estimador de Kernel para a função de densidade condicional da variável resposta Y dada a covariável x

$$\hat{f}_{Y|X}(y|x) = \frac{1}{h_y} \sum_{i=1}^{n} W_i(x) K_y(\frac{y - Y_i}{h_y}) \tag{20}$$

em que $K(.)$ é a função densidade da normal padrão e $W_i(x) = \frac{K_x(\frac{x-X_i}{h_x})}{\sum_{i=1}^{n} K_x(\frac{x-X_i}{h_x})}$ é o peso da i-ésima observação de X tal que $\sum_{i=1}^{n} W_i(x) = 1$.

Esse estimador foi obtido de (19) pela substituição da fração $\frac{1}{n}$ pelos pesos W_i.

- Estimador de Kernel para a função de distribuição condicional da variável resposta Y dada a covariável x

$$\hat{F}_{Y|X}(y|x) = \frac{1}{h_y}\sum_{i=1}^{n}\frac{K_x(\frac{x-X_i}{h_x})\mathbb{K}_y(\frac{y-Y_i}{h_y})}{\sum_{i=1}^{n}K_x(\frac{x-X_i}{h_x})} \tag{21}$$

em que $\mathbb{K}(.)$ é o Kernel obtido da função de distribuição acumulada da normal padrão.

Outro modo de estimação da função densidade condicional $f_{Y|X}(.|x)$ seria estimar $f_{X,Y}(x,y)$ via kernel bivariado e $f_X(x)$ via kernel univariado dada a amostra $\{X_i, Y_i\}_{i=1}^n$, obtendo, nesse caso, a seguinte estimativa:

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)}$$

3.1.8 Algoritmo da bissecção para regressão não paramétrica

Com a finalidade de estimar os quantis condicionais de uma distribuição não paramétrica, implementamos um algoritmo que se baseia no método da bissecção.

Baseia-se na estimativa via Kernel da função de distribuição acumulada condicional de Y dado x:

$$\hat{F}_{Y|X}(y|x) = \frac{1}{h_y}\sum_{i=1}^{n}\frac{K_x(\frac{x-X_i}{h_x})\mathbb{K}_y(\frac{y-Y_i}{h_y})}{\sum_{i=1}^{n}K_x(\frac{x-X_i}{h_x})}$$

Essa função é utilizada para obter o quantil condicional de Y dado x, usando o fato de que este é o inverso da função de distribuição.

- **Descrição do Algoritmo**

1. Dada uma amostra $\{X_i, Y_i\}$ de tamanho n:
   (a) Construa os objetos limite inferior *linf* e limite superior *lsup* e associe a eles os valores mínimo e máximo da amostra $\{y_i\}$, respectivamente.
   (b) calcule a diferença *dif* entre os limites *linf* e *lsup*.
2. Atribua à estimativa do parâmetro de suavização h_X ótimo o valor prescrito por Silverman (1986):

$$h_X = 0.9An^{-\frac{1}{5}} \quad \text{em que} \quad A = min\left(s, \frac{IQR}{1.34}\right)$$

sendo s o desvio padrão amostral de X e IQR o intervalo interquartílico de $\{x_i\}$

3. Atribua à estimativa do parâmetro de suavização $h_{Y|x_j}$ ótimo a expressão, abaixo, obtida com base na estimativa da variância condicional:

$$h_{Y|x_j} = \frac{1}{5}\left\{ \sum_{i=1}^{n} \frac{K_x(\frac{x-X_i}{h_x})Y_i^2}{K_x(\frac{x-X_i}{h_x})} - \Big(\sum_{i=1}^{n} \frac{K_x(\frac{x-X_i}{h_x})Y_i}{K_x(\frac{x-X_i}{h_x})} \Big)^2 \right\}^{\frac{1}{2}} \propto \sqrt{\hat{Var}(Y|x)}$$

4. Estabeleça como limite de tolerância *tol* um $\varepsilon > 0$ a ser utilizado no critério de parada do método;

5. Para cada ordem τ do quantil fixa em $\{0.10, 0.25, 0.50, 0.75, 0.90\}$:
   (a) Obtenha o valor $lmed = (lsup + linf)/2$;
   (b) Calcule $\hat{F}_{Y|X}(y|x)$ para y igual a *linf*, *lsup* e *lmed*:
      i. Se $\hat{F}_{Y|X}(lmed|x) > \tau$, faça *lsup* igual a *lmed*;
         Se $\hat{F}_{Y|X}(lmed|x) \leq \tau$, faça *linf* igual a *lmed*;
      ii. Recalcule $dif = lsup - linf$ e $lmed = (lsup + linf)/2$;
   (c) Se $dif < \varepsilon$, pare e faça $Q_\tau(Y|x) = lmed$;
      Se $dif \geq \varepsilon$, vá para o passo 5b)

3.2 Descrição e fonte dos dados

3.2.1 Dados de estudo sobre o consumo de energia (*Segreg*)

O primeiro banco de dados a ser utilizado nesse trabalho como exemplo de aplicação refere-se ao consumo de energia e à temperatura média de um edifício no Campus da Universidade de Minnesota - EUA. Foram obtidos por fonte secundária: Pacote alr4 do livro de Weisberg [2005]. A coleta foi realizada por 39 meses, entre os anos de 1988 e 1992, a cada dia. Porém, somente a temperatura média mensal foi considerada para exploração e análise de dados.

Existem duas variáveis contínuas: temperatura média mensal em graus Fahrenheit (Temp) e consumo de eletricidade em kwh (C). Será modelado o consumo de eletricidade em função da temperatura através da técnica de regressão quantílica. Na figura 3.2.1, segue o gráfico de dispersão no qual podemos visualizar a relação entre essas duas variáveis.

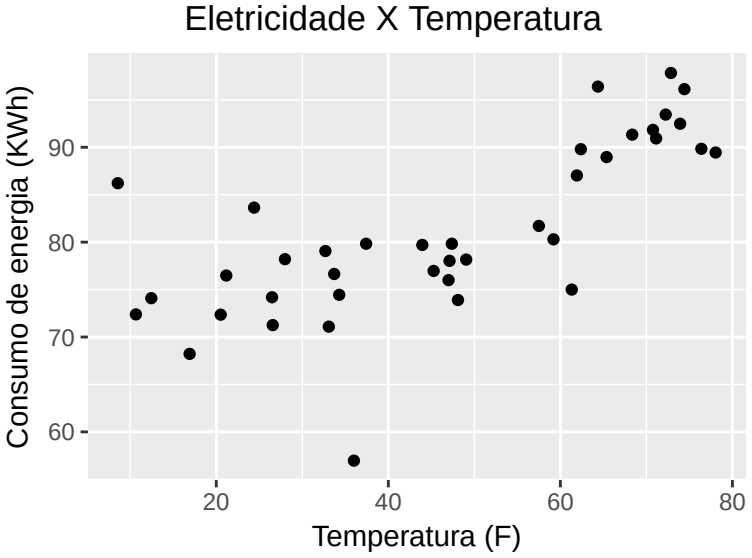


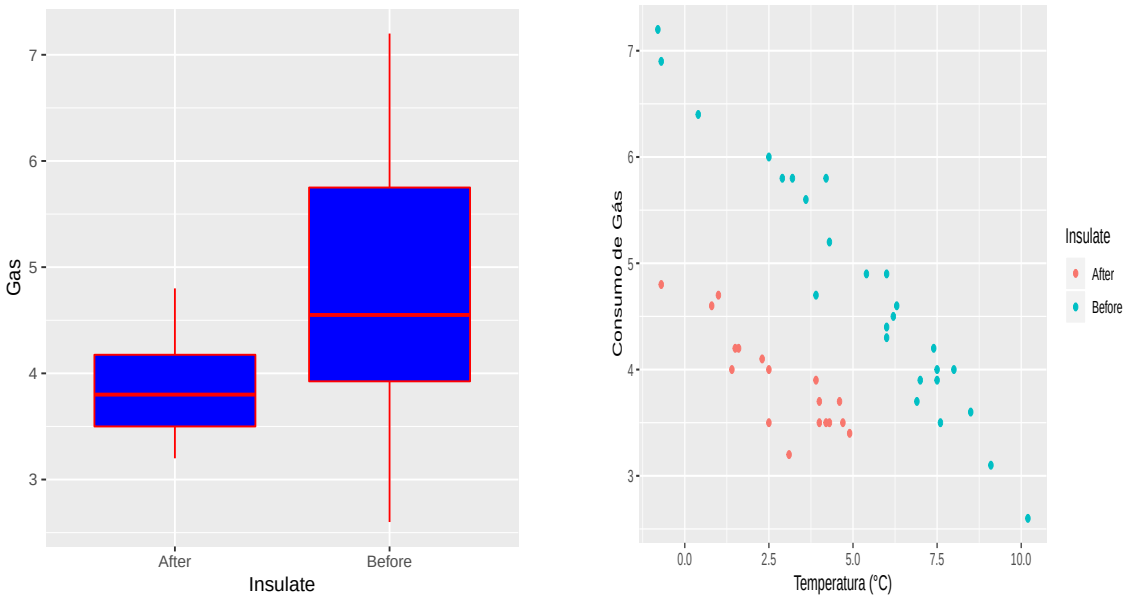
Figura 4: Relação entre Eletricidade e Temperatura

3.2.2 Dados sobre o consumo de gás em residência (*Insulgas*)

Esses dados possuem 44 observações e referem-se à utilização de gás natural em uma casa. O consumo semanal de gás (em 1000 pés cúbicos) e a temperatura média externa (em graus Celsius) foram registrados por 26 semanas antes e 30 semanas após o isolamento da parede da cavidade ter sido instalado. O termostato da casa foi definido em 20 graus Celsius, por toda parte.

As váriaveis envolvidas no estudo são:

- Insulate: covariável categorizada com dois níveis que define o momento da medição da temperatura(antes ou depois do isolamento)
- Temp: covariável contínua que representa a temperatura externa
- Gas: variável resposta consumo semanal de gás em 1000 pés cúbicos.



(a) Consumo de gás

(b) Dispersão Gas x Temp

Figura 5: Gráficos: a)Box-plot do consumo de gás pelo período de medição da temperatura b)Gráfico de dispersão do consumo de gás pela temperatura externa

3.2.3 Dados de seguro de vida (*Insurance*)

Esse conjunto de dados possui 1338 observações de 7 variáveis. Dentre estas, quatro são quantitativas (age, bmi, children e charges) e três categóricas nominais (sex, smoker e region).

- age: idade do beneficiário primário;
- bmi: índice de massa corporal;
- children: número de crianças cobertas pelo seguro;
- charges: despesas médicas individuais faturadas;
- sex: gênero do contratante do seguro (feminimo/masculino);
- smoker: indicador fumante (sim/não);
- region: área residencial do beneficiário nos EUA (nordeste, sudeste, sudoeste, noroeste).

Na figura 6, são mostradas as seguintes informações. A parte inferior esquerda da diagonal mostra matrizes de dispersão juntamente com a média bivariada denotada por um ponto vermelho; uma curva em vermeho que mostra uma relação mais flexível entre duas variáveis e uma elipse (círculo) que mostra a intensidade de correlação, em que a correlação mais alta é indicada por uma elipse mais elástica. Os diagramas em diagonal mostram o histograma de cada variável e densidade estimada. Os números no canto superior direito da diagonal são os valores de correlação entre 2 variáveis.

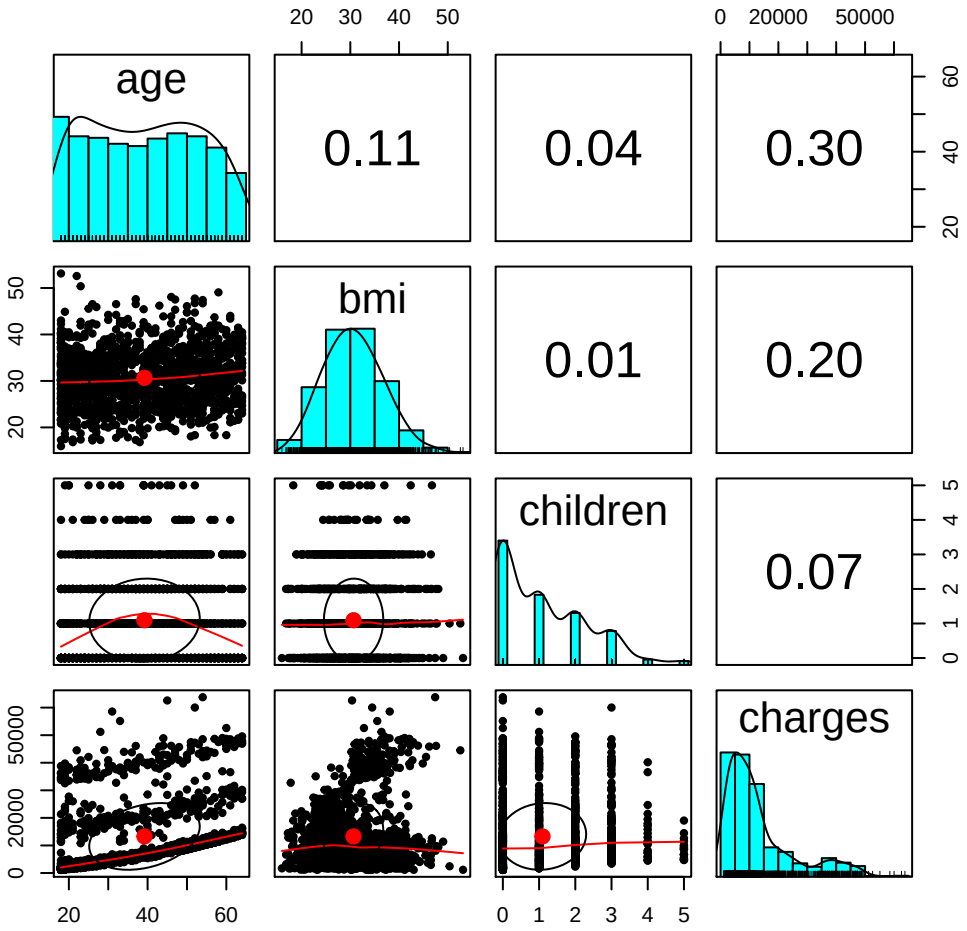
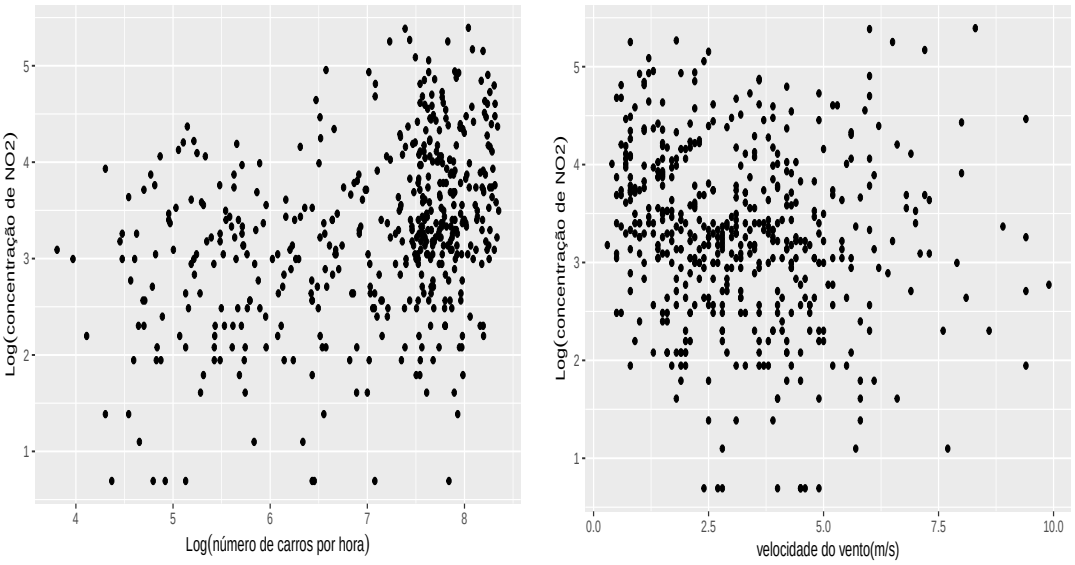


Figura 6: Gráfico com os histogramas e correlações das variáveis das variavéis age, bmi, children e charges

3.2.4 Dados de estudo sobre poluição do ar (*PMA10*)

Os dados são uma subamostra com 500 observações de um estudo sobre a relação entre poluição do ar em uma estrada da Noruega, volume de tráfego de carros e variáveis meteorológicas. Eles foram coletados pela Administração de Estradas Públicas da Noruega. A variável resposta (V1) consiste em valores horários do logaritmo da concentração de partículas NO2, medida em Alnabru em Oslo, Noruega, entre outubro de 2001 e agosto de 2003. As variáveis preditoras são:

- V2: logaritmo do número de carros por hora;
- V3: temperatura em graus Celsius ao nível 2 metros acima do solo;
- V4: velocidade do vento (m/s);
- V5: diferença de temperatura entre 2 e 25m de altitude, em graus Celsius;
- V7: hora do dia e,
- V8: número de dias da medição no período de outubro de 2001 a agosto de 2003.



(a) Dispersão V1 X V2

(b) Dispersão V1 X V4

Figura 7: Diagramas de dispersão

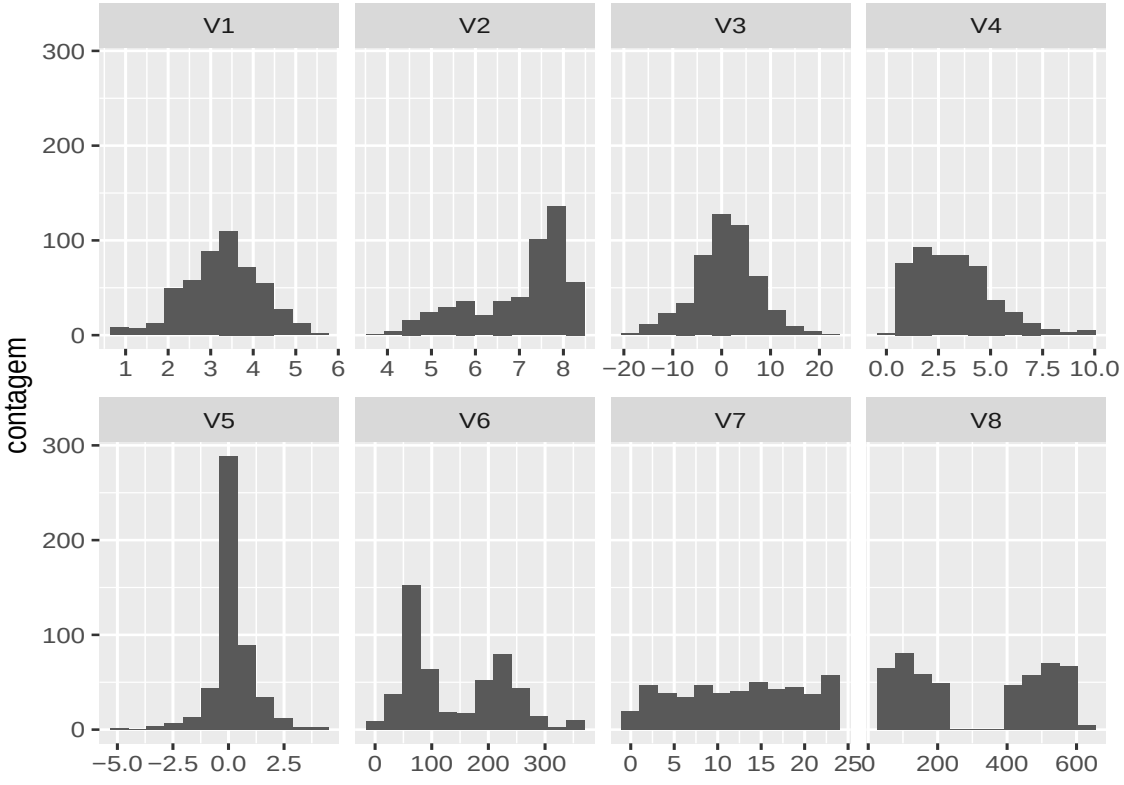


Figura 8: Histograma das variáveis: V1, V2, V3, V4, V5, V6, V7 e V8

3.2.5 Dados de uma usina geradora de eletricidade (Folds)

Esse conjunto de dados contém 9568 observações coletadas de uma usina de ciclos combinados durante 6 anos (2006-2011), período no qual o funcionamento ocorreu com carga total. As variáveis ambientais médias por hora envolvidas no estudo são:

- AT: Temperatura Ambiente;
- AP: Pressão Ambiente;
- RH: Umidade Relativa;
- V: Vácuo de Exaustão;
- PE: Produção de Energia Elétrica.

A Central de Ciclo Combinado (CCPP) é composta por turbinas a gás (GT), turbinas a vapor (VT) e geradores de vapor de recuperação de calor. A eletricidade é gerada por turbinas a gás e a vapor, que são combinadas em um ciclo e transferidas de uma para outra. Enquanto o vácuo é coletado e tem efeito sobre a turbina a vapor, outras três variáveis do ambiente afetam o desempenho do GT. As medidas são obtidas de vários sensores localizados ao redor da fábrica que registram as variáveis de ambiente a cada segundo. O objetivo é prever a produção de energia elétrica a partir das observações das covariáveis ambientais mencionadas acima.

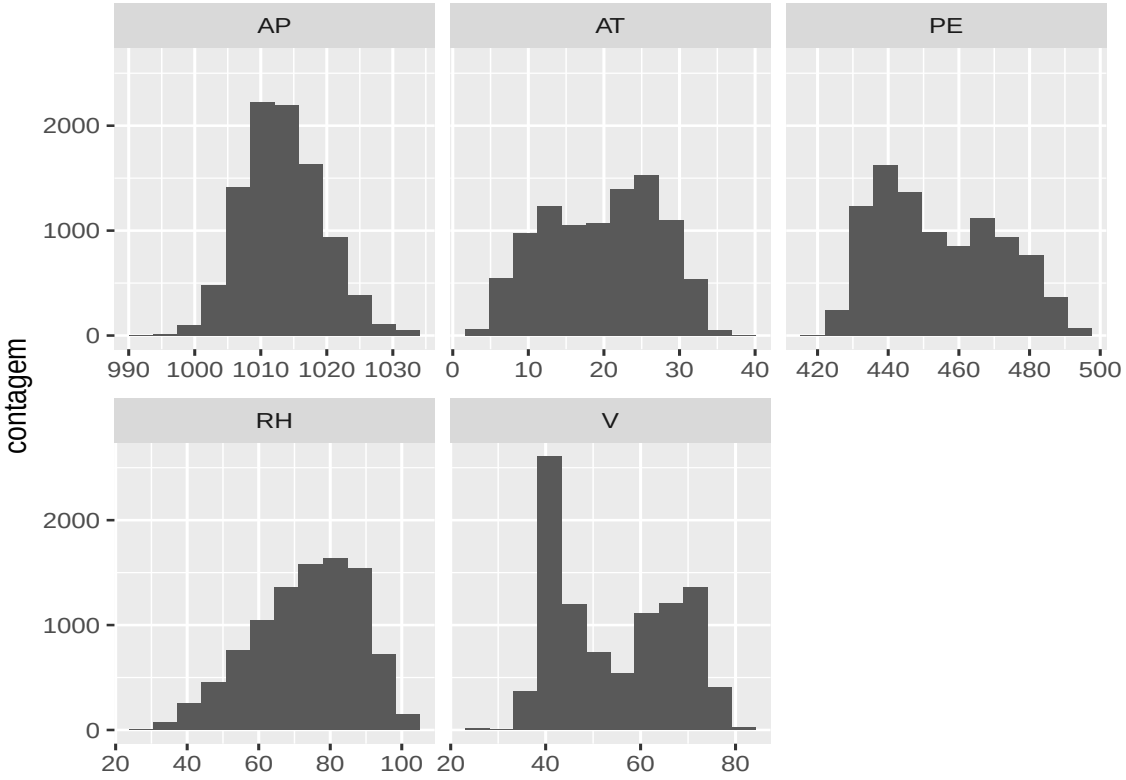


Figura 9: Histograma das variáveis: AT- Temperatura Ambiente, AP- Pressão Ambiente, RH- Umidade Relativa, V- Vácuo e PE- Produção de Energia

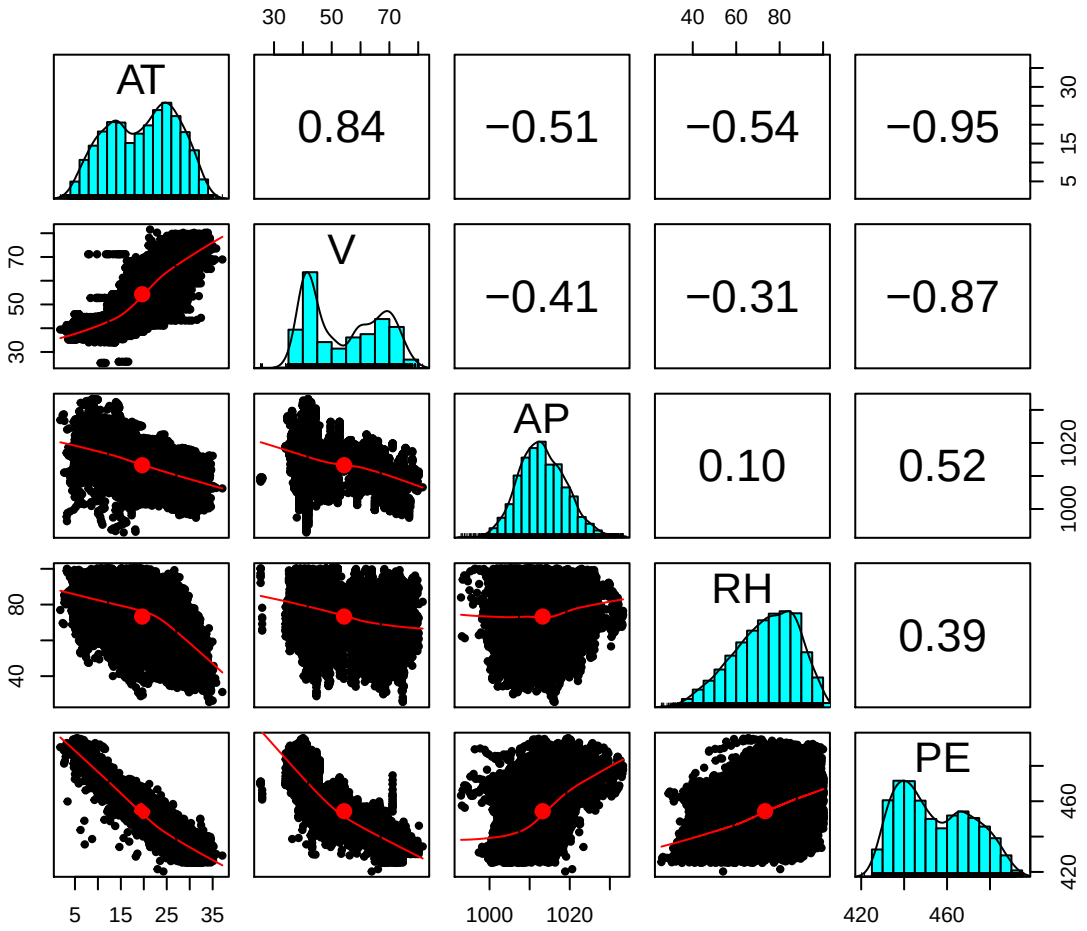


Figura 10: Gráfico com os histogramas e correlações das variáveis: AT- Temperatura Ambiente, AP- Pressão Ambiente, RH- Umidade Relativa, V- Vácuo e PE- Produção de Energia

3.2.6 Dados sobre imóveis em New York (housing-price)

Esse conjunto de dados refere-se a preços e propriedades de casas colocadas à venda em Nova York. Essas informações foram obtids de uma amostra aleatória de 1057 casas retiradas do Saratoga Housing Data (De Veaux), com as seguintes variáveis:

- **Price**: Preço avaliado da casa;
- **Living.Area**: Tamanho da sala de estar;

- **Age**: Idade do imóvel;
- **Bathrooms**: Quantidade de banheiros;
- **Bedrooms**: Quantidade de dormitórios;
- **Fireplace**: Indicador de presença de Lareira;
- **Lot.Size**: Tamanho do lote.

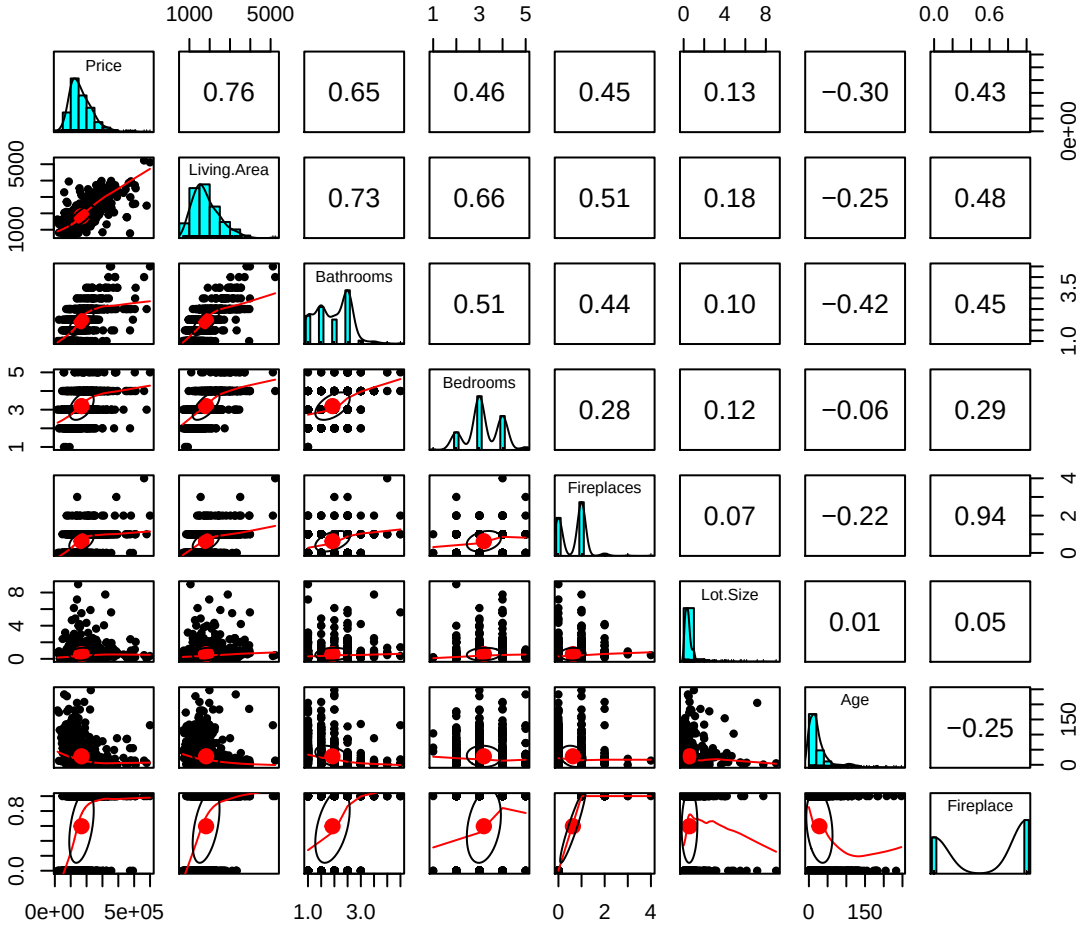


Figura 11: Gráfico com os histogramas e correlações das variáveis: Price, Living.Area, Age, Bathrooms, Bedrooms, fireplace e Lot.Size

4 APLICAÇÃO E ANÁLISE DOS RESULTADOS

4.1 Dados sobre consumo de energia (*Segreg*)

Conforme descrito anteriormente, esse conjunto de dados contém observações de consumo de energia e temperatura média mensal.

O objetivo é modelar o consumo de energia analisando os efeitos da covariável temperatura média mensal nas respostas quantílica e média (consumo médio), utilizando um modelo linear.

- Modelo de Regressão Quantílica

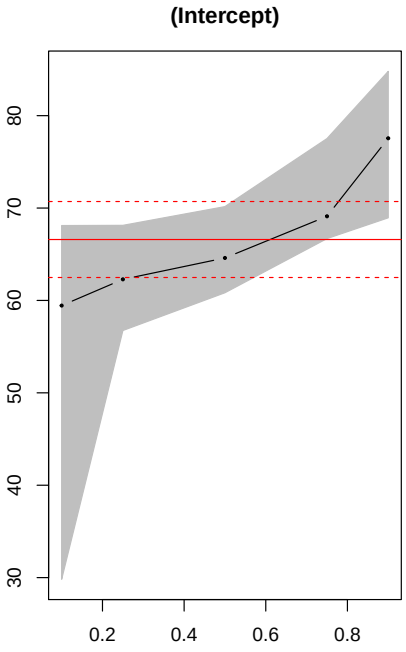
$$Q_C(\tau|Temp) = \hat{\beta}_0 + \hat{\beta}_1 Temp \tag{22}$$

- Modelo de Regressão por mqo

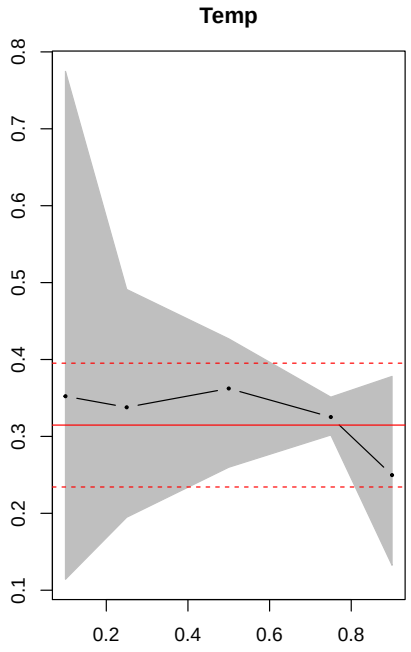
$$E(C|Temp) = \hat{\beta}_0 + \hat{\beta}_1 Temp \tag{23}$$

As estimativas dos coeficientes, $\hat{\beta}_0$ e $\hat{\beta}_1$, podem ser obtidas solucionando um problema de programação linear de minimização da função objetivo perda, conforme descrito na seção 3.1.1. Em geral, utiliza-se o algoritmo simplex para o qual existem rotinas implementadas em softwares estatísticos. Nesse trabalho, foi utilizado o software R. Os códigos encontram-se disponíveis no apêndice.

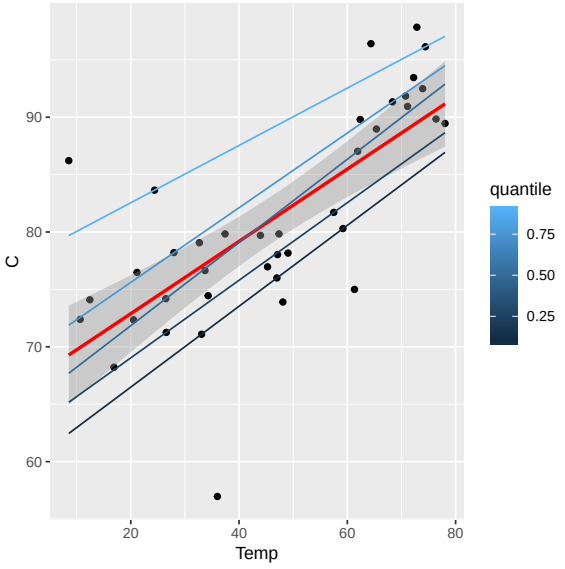
4.1.1 Gráficos e Tabelas



(a) Estimativa do Intercepto β_0



(b) Estimativa do coeficiente β_1



(c) Ajuste das retas de regressão

Figura 12: Ajustes das retas de regressão e variação dos coeficientes

Tabela 1: Modelo de Regressão por mqo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.5942	2.4998	26.64	0.0000
Temp	0.3147	0.0489	6.43	0.0000
	R^2=0.53	Adj. R^2= 0.51	Num. obs.=39	RMSE= 6.33

Tabela 2: Modelo de Regressão Quantílica.

τ	**Parâmetro**	**Estimativa**	**Erro Padrão**	**p-valor**	**ICs**	
					L.I.	L.S.
0,10	β_0	59,44	0,95	$<$ 0,001	29,8	68,10
	β_1	0,35	0,02	$<$ 0,001	0,11	0,77
0,25	β_0	62,28	2,46	$<$ 0,001	56,74	68,12
	β_1	0,34	0,05	$<$ 0,001	0,19	0,49
0,50	β_0	64,60	2,98	$<$ 0,001	60.83	70.16
	β_1	0,36	0,06	$<$ 0,001	0.25	0.43
0,75	β_0	69,11	1,83	$<$ 0,001	66.69	77.52
	β_1	0,33	0,04	$<$ 0,001	0.30	0.35
0,90	β_0	77,55	8,86	$<$ 0,001	68.94	84.79
	β_1	0,25	0,17	$<$ 0,02	0.13	0.38

4.1.2 Análise e Interpretação dos Resultados

As estimativas dos coeficientes β_1 da regressão quantílica variam, de modo insignificante, em torno da estimativa do coeficiente de regressão por mqo. Isso indica que a distribuição dos erros possui variância constante e o efeito provocado no consumo de energia pela variação de uma unidade na temperatura média mensal (Temp), para os quantis condicionais de ordens $\in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ é próximo ao efeito provocado no consumo médio pela mesma variação.

Os efeitos da covariável temperatura média mensal (Temp) na variável resposta para os quantis de ordem $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ são significantes, pois as estimativas dos coeficientes de regressão que medem a intensidade do efeito são estatisticamente significantes, quanto à nulidade, ou seja, rejeita-se a hipótese nula de igualdade a zero das estimativas de cada um dos coeficientes, conforme apresentado na tabela 2.

A tabela abaixo apresenta a análise de deviance para regressão quantílica. Estamos testando a igualdade dos coeficientes de regressão β_1 para as diferentes ordens dos quantis mencionadas. Como podemos observar, o p-valor é maior que os níveis descritivos usuais de 1%,5% e 10%. Isso implica que não há evidências suficientes para rejeição da hipótese nula de igualdade dos coeficientes.

	Df	Df residual	F-valor	Pr(≥F)
Temp	4	191	0.2336	0.9192

Tabela 3: Teste de igualdade dos coeficientes: $\tau \in \{0.10,0.25,0.5,0.75,0.9\}$

Nesse exemplo de aplicação, os dados não violam as condições para aplicação da regressão por mqo. Mas, em muitos casos, elas são violadas, havendo portanto a necessidade de aplicação da técnica de regressão quantílica a fim de obter boas estimativas para os coeficientes.

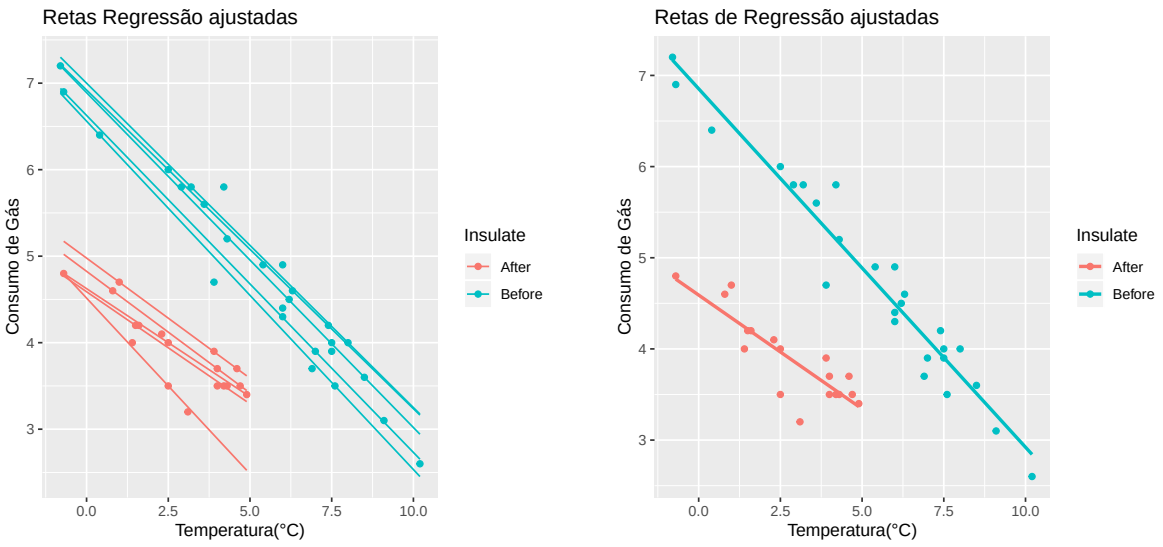
Através das retas de regressão ajustadas, detectamos homocedasticidade dos dados e, ainda, é possível avaliar se a distribuição é simétrica ou assimétrica. Para tanto, verifica se a reta para o quantil condicional de ordem 0.50 aproxima-se da reta para média condicional, e compara-se também as referentes aos quantis condicionais de ordens 0.10 e 0.25 com as referentes aos quantis de ordens 0.75 e 0.90.

Esse modelo que apresenta as estimativas dos coeficientes de regressão β_1 bem próximas é denominado de modelo de locação.

4.2 Dados sobre consumo de gás em residência (*Insulgas*)

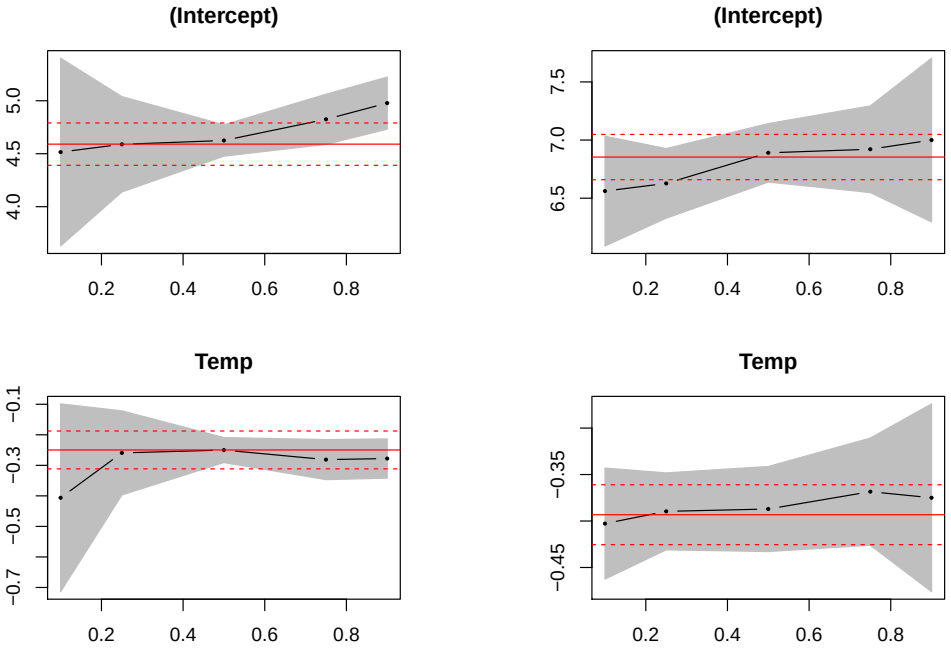
Nosso objetivo é aplicar a técnica de regressão quantílica nesse conjunto de dados reais que foi descrito anteriormente. Para isso, ajustamos o modelo de regressão para os quantis de ordem $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. Seguem abaixo os gráficos e tabelas referentes aos ajustes dos modelos de regressão por MQO e de regressão quantílica, nos quais temos a resposta consumo de gás (Gas) e as covariáveis temperatura (Temp) para cada nível da variável categorizada Isolamento (Insulate).

4.2.1 Gráficos e Tabelas



(a) Regressão Quantílica (b) Regressão por MQO

Figura 13: Ajustes de regressão da média condicional e de regressão quantílica para os quantis de ordem $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$



(a) Após o isolamento (b) Antes do isolamento

Figura 14: Variação das estimativas dos coeficientes quantílicos

Tabela 4: Ajuste do modelo antes do isolamento

τ	**Parâmetro**	**Estimativa**	**Erro Padrão**	**p-valor**	**ICs**	
					L.I.	L.S.
0,10	β_0	6,561	0,149	$<$ 0,001	-1,797.10^3	6,583
	β_1	-0,403	0,0247	$<$ 0,001	-0,417	-0,268
0,25	β_0	6,627	0,247	$<$ 0,001	6,552	6,845
	β_1	-0,389	0,040	$<$ 0,001	-0,421	-0,378
0,50	β_0	6,890	0,109	$<$ 0,001	6,716	7,156
	β_1	-0,387	0,018	$<$ 0,001	-0,436	-0,362
0,75	β_0	6,921	0,080	$<$ 0,001	6,887	7,972
	β_1	-0,368	0,013	$<$ 0,001	-0,504	-0,356
0,90	β_0	7,000	0,038	$<$ 0,001	6,952	1,797.10^3
	β_1	-0,375	0,006	$<$ 0,001	-0,545	-0,333

Tabela 5: Ajuste do modelo após o isolamento

τ	**Parâmetro**	**Estimativa**	**Erro Padrão**	**p-valor**.	**ICs**	
					L.I.	L.S.
0,10	β_0	4,51	0,300	$<$ 0,001	-1,797	4,522
	β_1	-0,41	0,093	$<$ 0,001	-0,429	-0,182
0,25	β_0	4,59	0,024	$<$ 0,001	2,605	4,611
	β_1	-0.26	0,073	$<$ 0,001	-0,278	-0,063
0,50	β_0	4,62	0,110	$<$ 0,001	4,527	4,800
	β_1	-0,25	0,034	$<$ 0,001	-0,305	-0,211
0,75	β_0	4,82	0,083	$<$ 0,001	4,636	5,013
	β_1	-0,28	0,026	$<$ 0,001	-0,306	-0,211
0,90	β_0	4,97	0,019	$<$ 0,001	4,775	1,79.10^3
	β_1	-0,27	0,006	$<$ 0,001	-1,654	-0.219

4.2.2 Interpretação dos gráficos e Análise dos Resultados

Analisaremos os ajustes separadamente por níveis (antes e depois do isolamento) da covariável categorizada.

A tabela 4 nos mostra que todas as estimativas dos coeficientes de regressão e do intercepto foram significativas ao nível descritivo usual de 1%. Observa-se pouca

variação nas estimativas dos coeficientes de regressão. Isso indica que os dados não são heterocedásticos. Podemos, portanto assumir a condição de homocedasticidade.

O gráfico da variação das estimativas dos coeficientes regressão da varíavel temperatura (Temp), antes do isolamento, confirma a suposição mencionada, pois as estimativas dos coeficientes para regressão quantílica e por MQO são bem próximas e estão dentro do intervalo de confiança ao nível 0.95, $IC(\beta_1, 95\%)$.

Considerando o modelo de regressão quantílica ajustado antes do isolamento, observa-se que, para o quantil condicional de ordem 0.10, há uma maior taxa de decréscimo. A cada aumento de 1 unidade na temperatura, há uma diminuição de aproximadamente 0,4 no consumo de gás.

Após o isolamento, observa-se que a estimativa do coeficiente de regressão β_1 não encontra-se dentro do intervalo de confiança $IC(\beta_1, 95\%)$ para ordem 0.10. Para $\tau \in \{0.25, 0.50, 0.75, 0.90\}$, houve uma subestimação dos coeficientes em relação ao coeficiente de regressão β_1 da média condicional. Se tivessemos interesse em estudar a relação entre o consumo de gás e a temperatura, na cauda inferior da distribuição condicional, seria prudente adotar a regressão quantílica.

4.3 Dados de estudo sobre poluição do ar (*PMA10*)

Selecionamos as covariáveis logaritmo do número de carros por hora (V2) e velocidade do vento (V4) pelo método stepwise tanto para regressão quantílica quanto para a regressão por mqo, mediante à utilização do software R. De acordo com o método, V2 e V4 têm maiores capacidades de contribuição para um possível modelo ajustado cuja variável resposta é a concentração de NO2 (V1).

Seguem abaixo os ajustes, considerando V2 e V4 como covariáveis, dos modelos de regressão mediana (caso particular de regressão quantílica) e de regressão por mqo.

$$Q_{V1}(0.5|V2, V4) = \hat{\beta}_0(0.5) + \hat{\beta}_2(0.5)V2 + \hat{\beta}_4(0.5)V4$$

$$\mathbb{E}(V1|V2, V4) = \hat{\beta}_0 + \hat{\beta}_2 V2 + \hat{\beta}_4 V4$$

4.3.1 Gráficos e Tabelas

	Regressão Mediana	Regressão por mqo
β_0	1.82^{***}	1.35^{***}
	(0.25)	(0.23)
β_2	0.26^{***}	0.32^{***}
	(0.04)	(0.03)
β_4	-0.11^{***}	-0.10^{***}
	(0.02)	(0.019)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Tabela 6: Estimativas dos coeficientes

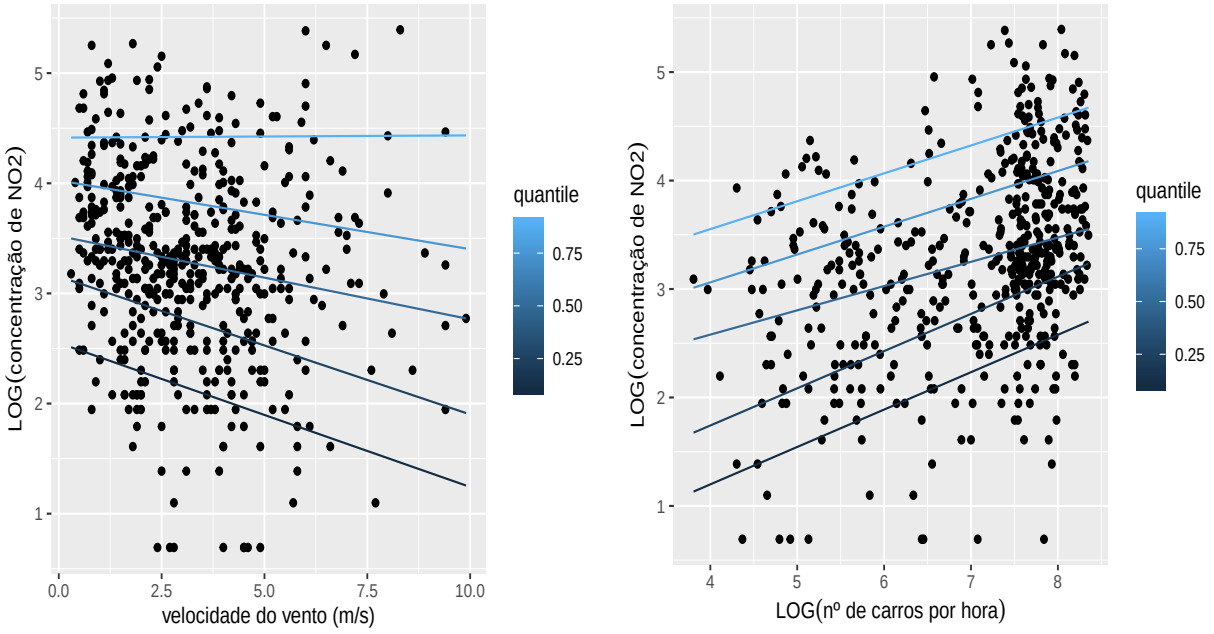
4.3.2 Análise dos Resultados

Observa-se, pela tabela 6 da subseção 4.3.1, que tanto os coeficientes da Regressão Mediana quanto os da Regressão por mínimos quadrados ordinários (mqo) são estatisticamente significantes ao nível descritivo de 1%.

Nessa mesma tabela, com base nas estimativas de β_4, nota-se que os efeitos provocados na variável resposta V1 são bem próximos. A cada aumento de uma unidade no preditor V4, há uma diminuição de 0.10 na resposta mediana de V1 enquanto, na resposta média, ocorre uma diminuição de 0.11. Agora, considerando o preditor V2, podemos observar uma diferença maior quanto aos efeitos. Nesse caso, para um aumento de 1(uma) unidade no preditor V2, fixando o valor de V4, há um aumento de 0.26 na resposta média e de 0.32 na resposta mediana. Para visualizar o ajuste desse mesmo modelo de regressão quantílica para outros quantis de ordem $\tau \in \{0.10, 0.25, 0.75, 0.9\}$, segue a tabela abaixo:

	Estimativa	Erro Padrão	t-valor	Pr(>|t|)
$\beta_0(0.10)$	-0.24	0.65	-0.36	0.72
$\beta_2(0.10)$	0.38	0.09	4.07	0.00
$\beta_4(0.10)$	-0.16	0.04	-3.93	0.00
$\beta_0(0.25)$	0.62	0.37	1.69	0.09
$\beta_2(0.25)$	0.36	0.05	7.70	0.00
$\beta_4(0.25)$	-0.11	0.03	-3.19	0.00
$\beta_0(0.50)$	1.82	0.29	6.27	0.00
$\beta_2(0.50)$	0.26	0.04	6.43	0.00
$\beta_4(0.50)$	-0.11	0.02	-4.52	0.00
$\beta_0(0.75)$	2.15	0.34	6.34	0.00
$\beta_2(0.75)$	0.28	0.05	5.63	0.00
$\beta_4(0.75)$	-0.10	0.03	-3.20	0.00
$\beta_0(0.90)$	2.49	0.29	8.54	0.00
$\beta_2(0.90)$	0.33	0.05	6.80	0.00
$\beta_4(0.90)$	-0.08	0.04	-1.99	0.05

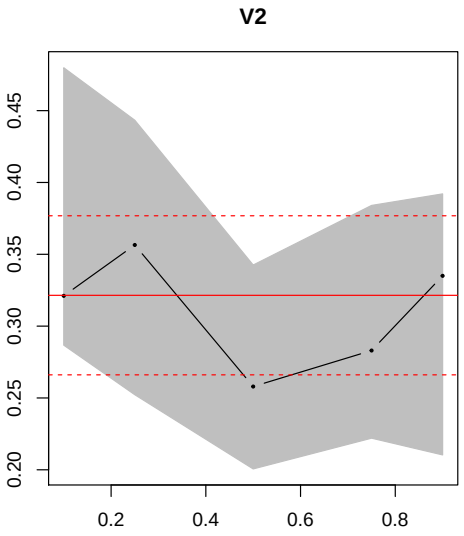
Tabela 7: Estimativas dos coeficientes de regressão



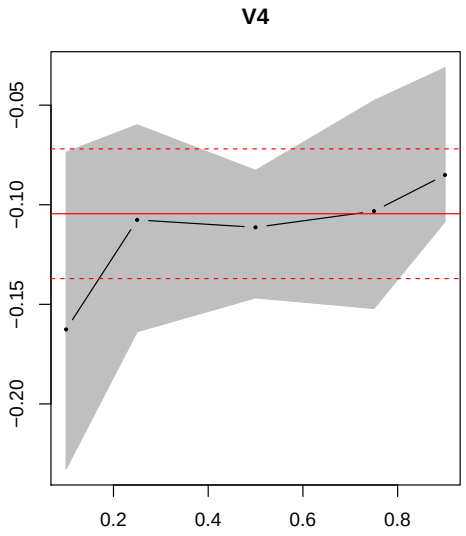
(a) Retas de regressão ajustadas para V2

(b) Retas de regressão ajustadas para V4

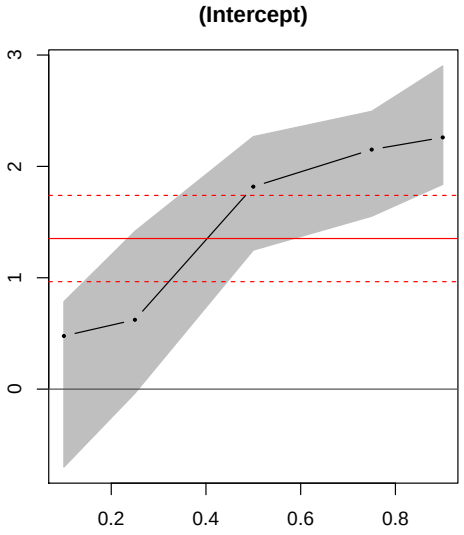
Figura 15: Modelos de Regressão Quantílica ajustados



(a) Coeficiente de regressão de V2



(b) Coeficiente de regressão de V4



(c) Intercepto

Figura 16: Variação dos coeficientes de regressão quantílica

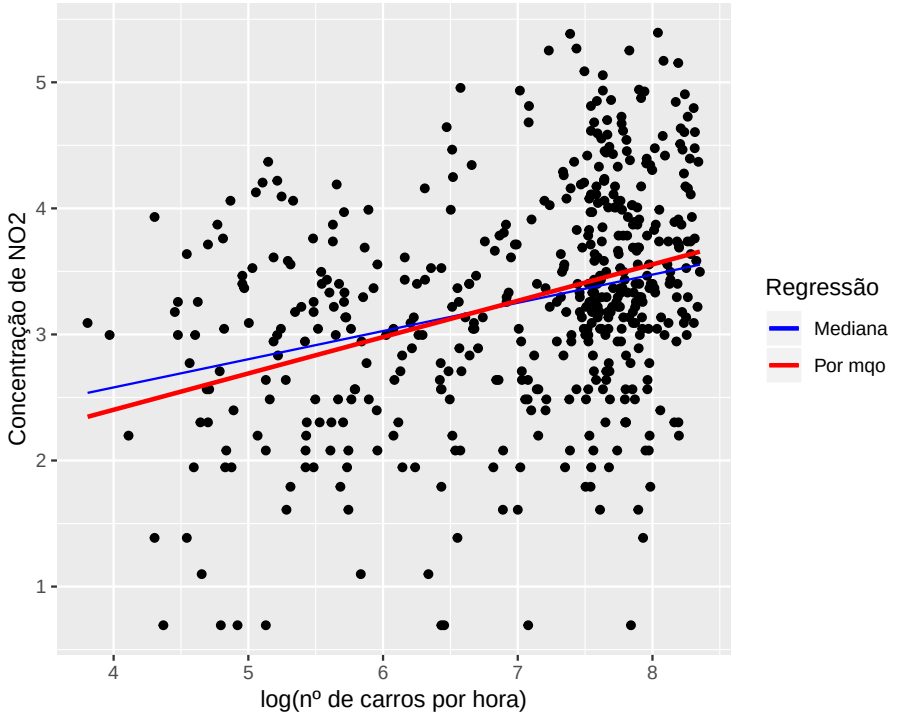


Figura 17: Gráfico de dispersão e ajustes de retas para os modelos de Regressão Mediana e da Média Condicional

Parâmetro	Estimativa	Desvio Padrão	p-valor
β_0	1.25	0.24	0.0000
β_2	0.28	0.03	0.0000
$\beta_0(0.50)$	1.68	0.36	0.0000
$\beta_2(0.50)$	0.22	0.05	0.0000

Tabela 8: Ajustes de regressão da média condicional e regressão mediana

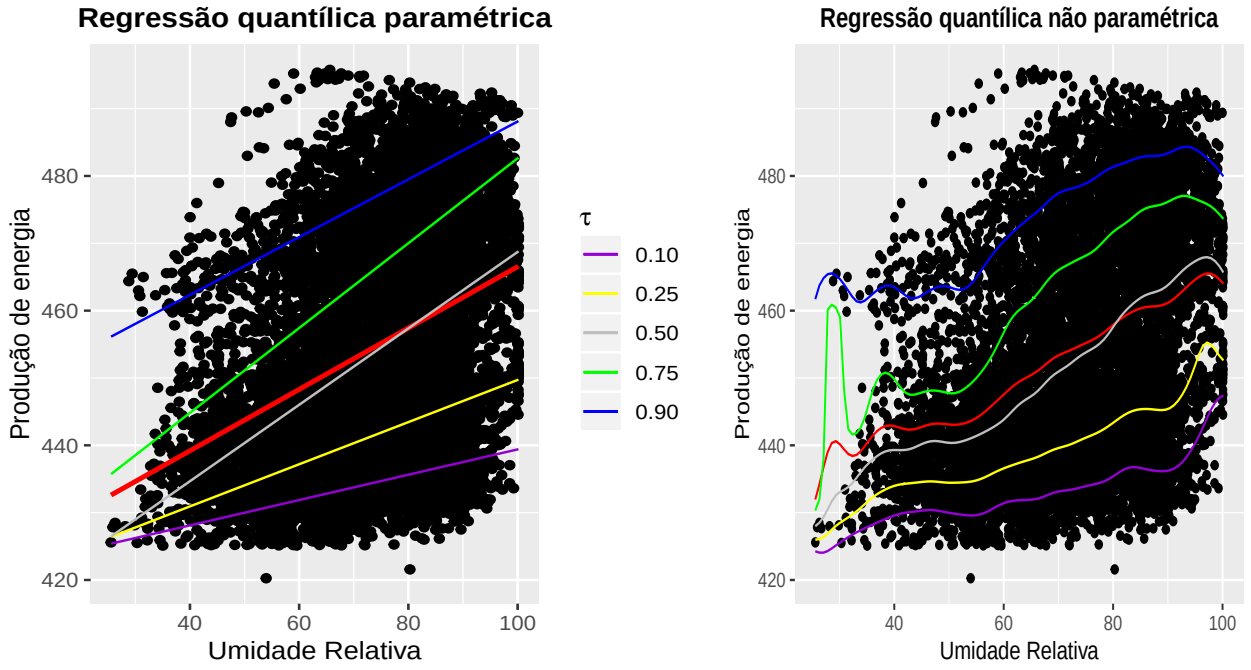
Na Tabela 8, temos o resultado do ajuste de regressão da Concentração de NO2 sobre a covariável log(número de carros por hora) (V2). Analisando os resultados, podemos dizer que os ajustes são todos significativos.

Conforme o gráfico, existe uma diferença não significativa nas inclinações das retas ajustadas.

Na regressão por mqo, a presença de grandes concentrações de NO2 para uma média do número de carros alta afetou o coeficiente de regressão. Já na regressão mediana, a estimativa do coeficiente foi menos afetada. Isso mostra a robustez do modelo de regressão quantílica.

4.4 Dados de uma usina geradora de eletricidade (Folds)

Com o objetivo de aplicar a técnica não paramétrica descrita anteriormente, ajustamos o modelo de regressão quantílica não paramétrica para examinar a relação entre a variável resposta produção de energia (PE) e a covariável umidade relativa (RH) e comparar com o modelo paramétrico.



(a) Regressão paramétrica

(b) Regressão não paramétrica

Figura 18: Ajustes das curvas de regressão quantílica

	Estimativa	Erro padrão	t-valor	Pr(>|t|)
$\beta_0(0.10)$	420.39	0.66	635.69	0.00
$\beta_1(0.10)$	0.15	0.01	16.15	0.00
$\beta_0(0.25)$	418.41	0.73	572.90	0.00
$\beta_1(0.25)$	0.31	0.01	28.75	0.00
$\beta_0(0.50)$	411.97	0.95	434.70	0.00
$\beta_1(0.50)$	0.57	0.01	40.71	0.00
$\beta_0(0.75)$	419.62	1.54	272.47	0.00
$\beta_1(0.75)$	0.63	0.02	32.24	0.00
$\beta_0(0.90)$	452.41	1.33	340.39	0.00
$\beta_1(0.90)$	0.39	0.02	22.84	0.00

Tabela 9: Ajustes de modelos de regressão quantílica

A reta e a curva ajustadas em vermelho referem-se à estimativa da média condicional para o caso paramétrico e não paramétrico, respectivamente. A curva em verde representa a regressão mediana. Observa-se que ela se aproxima da curva em vermelho. A aproximação torna-se mais evidente em distribuições simétricas nas quais a mediana coincide com a média. Nota-se pela comparação dos gráficos que a regressão não paramétrica fornece boas estimativas para os quantis condicionais. É uma boa alternativa ao método paramétrico. Logo, teríamos mais uma possibilidade de aplicação da regressão quantílica caso a forma paramétrica da relação entre x e y fosse desconhecida. A Tabela 10 mostra as estimativas dos coeficientes da regressão quantílica para $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ com os respectivos erros padrões e estatísticas t. Verifica-se que todos os coeficientes são significativos ao nível descritivo de 1%.

4.5 Dados de seguro de vida (Insurance)

Escolhemos esse conjunto de dados com objetivo de propor modelos de regressão quantílica para modelar as despesas médicas, mediante a utilização das seguintes variáveis preditoras: indicador de condição de fumante (smoker), índice de massa corporal (bmi), idade do beneficiário (age) e o número de crianças cobertas na família (children).

Ajustaremos os dois modelos abaixo, a fim de compreender como as covariáveis influeciam nas despesas médicas (charges).

1. **Modelo de Regressão Quantílica múltiplo**

$$Q_{charges}(\tau|V) = \hat{\beta}_0 + \hat{\beta}_1 smoker + \hat{\beta}_2 bmi + \hat{\beta}_3 age + \hat{\beta}_4 children$$

em que V=$\{smoker, bmi, age, children\}$ é o vetor de preditores.

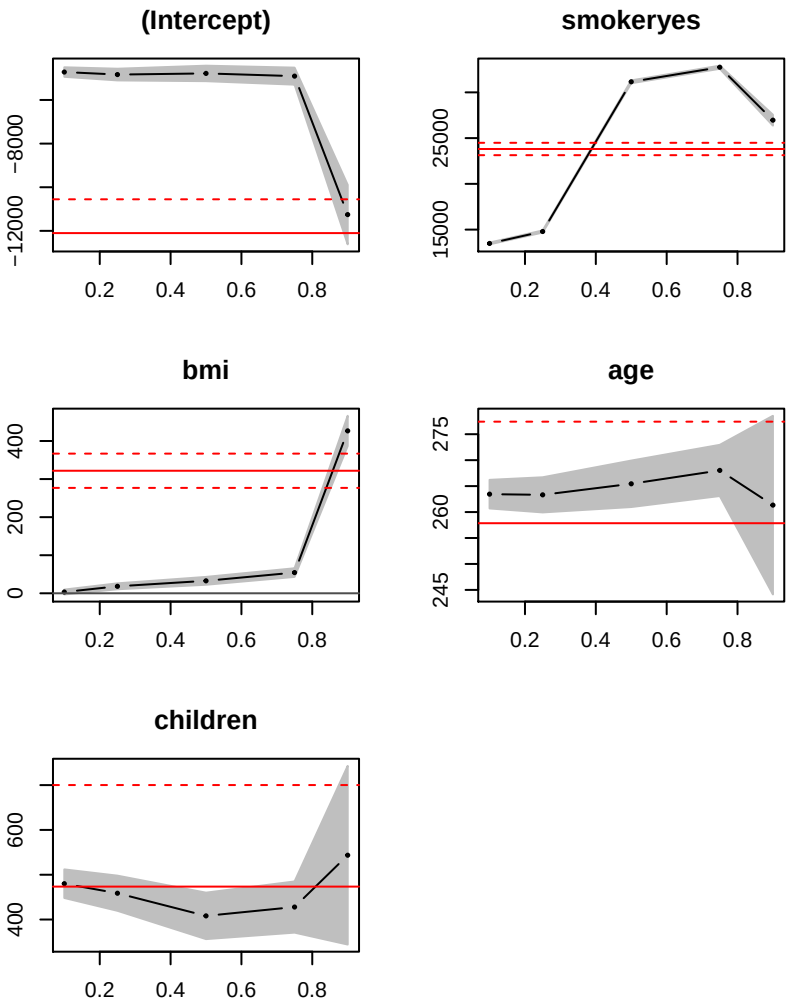


Figura 19: Variação das estimativas dos coeficientes

De acordo com as variações das estimativas dos coeficientes, conclui-se que o indicador de fumante (preditor smoker) tem um efeito positivo no aumento das despesas médicas. Observa-se que, a partir do quantil de ordem 0.4, o efeito torna-se consideravelmente alto.

A covariável idade (age) tem efeito uniforme ao longo de toda distribuição provocando um aumento nas despesas médicas, em torno de 264, para diversos quantis.

Já a quantidade de crianças cobertas pelo seguro (children) tem um efeito positivo que decai até o quantil 0.50 e, a partir desse, aumenta, porém sempre situando abaixo do aumento médio, exceto para o quantil de ordem 0.90.

2. **Modelo de Regressão Quantílica simples**

$$Q_{charges}(\tau|bmi) = \hat{\alpha}_0(\tau) + \hat{\alpha}_1(\tau)bmi \quad ; \quad \tau \in \{0.10, 0.25, 0.50, 0.75, 0.9\}$$

$$\mathbb{E}(charges|bmi) = \hat{\delta}_0 + \hat{\delta}_1 bmi$$

	tau= 0.10	tau= 0.25	tau= 0.50	tau= 0.75	tau= 0.90
(Intercept)	1418.84	2118.21	5624.70	5161.99	-10446.38
bmi	31.02	86.14	123.52	424.19	1412.55

Tabela 10: Estimativas dos coeficientes de regressão

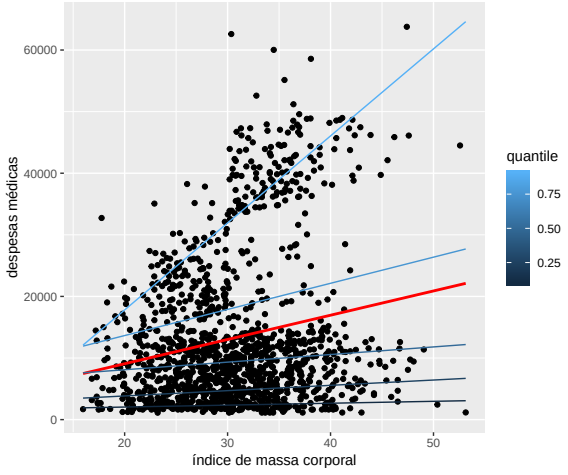


Figura 20: Modelo de regressão quantílica ajustado

DF	DF Residual	Estatística	p-valor
1.00	2675	1.084	0.297

Tabela 11: Teste de igualdade das inclinações para $\tau \in \{0.25, 0.50\}$

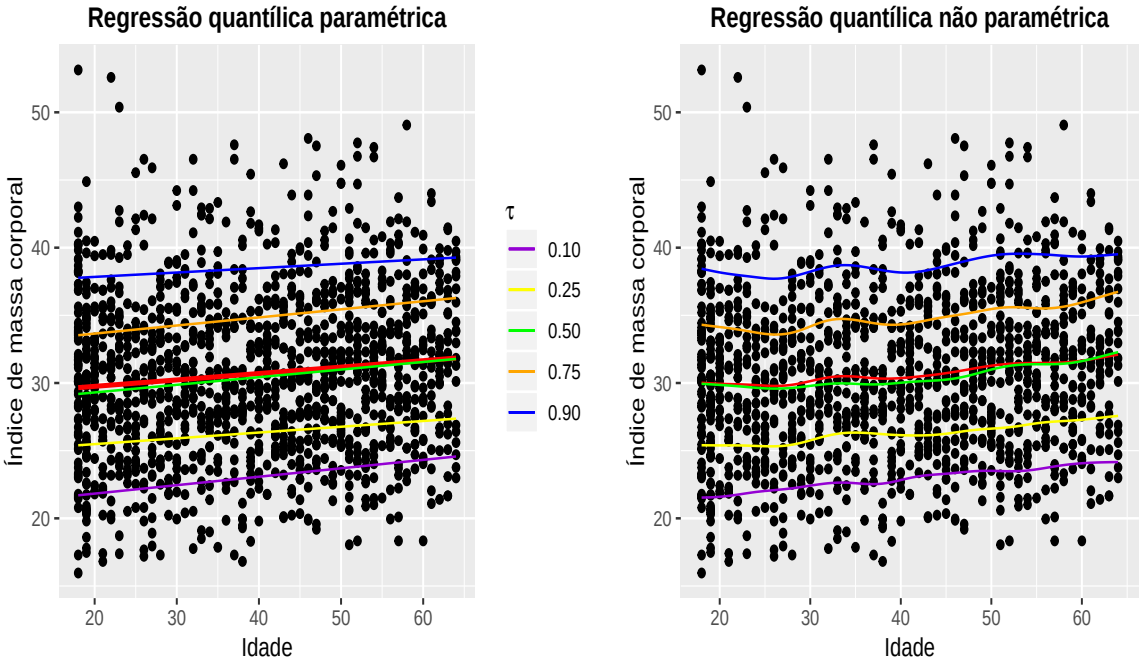
Observa-se que há variação tanto nos interceptos como nas inclinações das retas ajustadas. Logo, conclui-se que os efeitos provocados pelo índice de massa corporal aumentam dos quantis de ordem baixa para os de ordem alta, demonstrando que trata-se de um modelo escala-locação, pois apresenta alteração na tendência central e variabilidade nas despesas médicas. Isso significa que indivíduos com maior renda tende a gastar mais, em relação aos que têm menores rendas, à medida que o índice de massa corporal aumenta.

A mudança no indíce de massa corporal provoca uma maior variação nas despesas médicas no quantil de ordem 0.90 (cauda superior da distribuição). Isso indica que entre indivíduos que têm mais gastos com despesas médicas, o índice de massa corporal é um fator relevante.

Podemos notar que modelo de regressão quantílica fornece uma visão mais completa da distribuição condicional do que a fornecida pelo modelo por mqo, pois é possível obter a relação entre despesas médicas e índice de massa corporal em diversos quantis.

Existe um paralelismo entre as retas ajustadas para os quantis condicionais de ordens 0.25 e 0.50. Isso indica que há indícios de igualdade nas intensidades do efeito da variável preditora índice de massa corporal na variável resposta despesas médicas. Para comprovar, realizamos uma análise de variância para verificar a igualdade entre as inclinações. Os resultados encontram-se na Tabela 12. Verificamos que, ao nível descritivo de 1%, não há evidências suficientes para rejeição da hipótese nula de igualdade das inclinações, pois o p-valor é maior do que o referido nível.

A reta ajustada para o quantil de ordem 0.90 tem maior inclinação quando comparada à reta de regressão por mqo.Isso indica que, se adotássemos o método por minímos quadrados ordinários para examinar o efeito provocado pelo índice de massa corporal nas despesas médicas, não enxergaríamos o efeito referente aos indivíduos com mais gastos com despesa médica.



(a) regressão paramétrica

(b) regressão não paramétrica

Figura 21: Ajuste de regressão quantílica paramétrica e não paramétrica para $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$

Conforme mostrado nos gráficos 21(a) e 21(b), a curva da regressão quantílica não paramétrica é uma boa técnica alternativa para a paramétrica.

A distribuição condicional da varíavel índice de massa corporal (bmi) dada idade (age) tem uma forte simetria, logo a média se aproxima da mediana. A aproximação pode ser também ilustrada pelo gráfico 21(b) no qual, em verde, temos a representação da mediana condicional e, em vermelho, da média condicional.

Existe um paralelismo entre as retas indicando que poderíamos adotar a regressão por mqo, pois, nessa situação, os efeitos de idade nos diversos quantis condicionais e na média condicional são bem próximos.

4.6 Dados sobre imóveis em New York (housing-price)

Com esse conjunto de dados, queremos visualizar graficamente e, por meio de informações contidas nas tabelas abaixo, que as variavéis preditoras selecionadas pelo método stepwise contribuem para explicar a variabilidade dos preços dos imovéis (Price).

Desejamos também identificar casas que são muito caras ou baratas demais. Para isso, ajustamos os seguintes modelos de regressão, dado o conjunto de covariáveis $I = \{Living.Area, Bathrooms, Age, Fireplaces, Bedrooms\}$:

1. **Modelo para cauda inferior (quantil 0.10)**

$$\begin{aligned} Q_{Price}(0.10|I) &= \hat{\alpha}_0(0.10) + \hat{\alpha}_1(0.10)Living.Area + \hat{\alpha}_2(0.10)Bathrooms \\ &+ \hat{\alpha}_3(0.10)Age + \hat{\alpha}_4(0.10)Fireplaces + \hat{\alpha}_5(0.10)Bedrooms \end{aligned}$$

	Estimativa	Desvio Padrão	Estatística t	Pr(>|t|)
β_0	7616.24	7406.98	1.03	0.30
β_1	38.49	4.85	7.94	0.00
β_2	11812.51	3949.87	2.99	0.00
β_3	-651.16	37.90	-17.18	0.00
β_4	10662.12	3385.11	3.15	0.00
β_5	10606.49	2811.09	3.77	0.00

Tabela 12: Ajuste de modelo de regressão quantílica para $\tau = 0.10$

2. **Modelo para mediana condicional (quantil 0.50)**

$$\begin{aligned} Q_{Price}(0.50|I) &= \hat{\beta}_0(0.50) + \hat{\beta}_1(0.50)Living.Area + \hat{\beta}_2(0.50)Bathrooms \\ &+ \hat{\beta}_3(0.50)Age + \hat{\beta}_4(0.50)Fireplaces + \hat{\beta}_5(0.50)Bedrooms \end{aligned}$$

	Estimativa	Desvio Padrão	Estatística t	Pr(>|t|)
β_0	12290.74	5256.86	2.34	0.02
β_1	73.09	3.42	21.39	0.00
β_2	11077.22	2599.21	4.26	0.00
β_3	-239.18	53.25	-4.49	0.00
β_4	6941.96	2230.40	3.11	0.00
β_5	-692.94	1966.88	-0.35	0.72

Tabela 13: Ajuste do modelo de regressão quantílica para $\tau = 0.50$

3. **Modelo para média condicional**

$$\begin{aligned}\mathbb{E}(Price|I) &= \hat{\delta}_0 + \hat{\delta}_1 Living.Area + \hat{\delta}_2 Bathrooms \\ &+ \hat{\delta}_3 Age + \hat{\delta}_4 Fireplaces + \hat{\delta}_5 Bedrooms\end{aligned}$$

	Estimativa	Desvio Padrão	Estatística t	Pr(>|t|)
β_0	15519.5501	7322.3686	2.12	0.0343
β_1	73.6153	4.0057	18.38	0.0000
β_2	18788.9636	3674.2152	5.11	0.0000
β_3	-151.7555	48.1627	-3.15	0.0017
β_4	9209.3888	3190.2864	2.89	0.0040
β_5	-6107.7787	2756.5234	-2.22	0.0269

Tabela 14: Ajuste de modelo de regressão por mqo

4. **Modelo para cauda superior (quantil 0.90)**

$$\begin{aligned}\mathbb{Q}_{Price}(0.90|I) &= \hat{\gamma}_0(0.90) + \hat{\gamma}_1(0.90) Living.Area + \hat{\gamma}_2(0.90) Bathrooms \\ &+ \hat{\gamma}_3(0.90) Age + \hat{\gamma}_4(0.90) Fireplaces + \hat{\gamma}_5(0.90) Bedrooms\end{aligned}$$

	Estimativa	Desvio Padrão	Estatística t	Pr(>|t|)
β_0	32180.80	15524.34	2.07	0.04
β_1	101.83	9.74	10.45	0.00
β_2	16581.22	7767.98	2.13	0.03
β_3	335.98	122.41	2.74	0.01
β_4	8057.73	6980.38	1.15	0.25
β_5	-15543.91	5541.77	-2.80	0.01

Tabela 15: Ajuste de modelo de regressão quantílica para $\tau = 0.90$

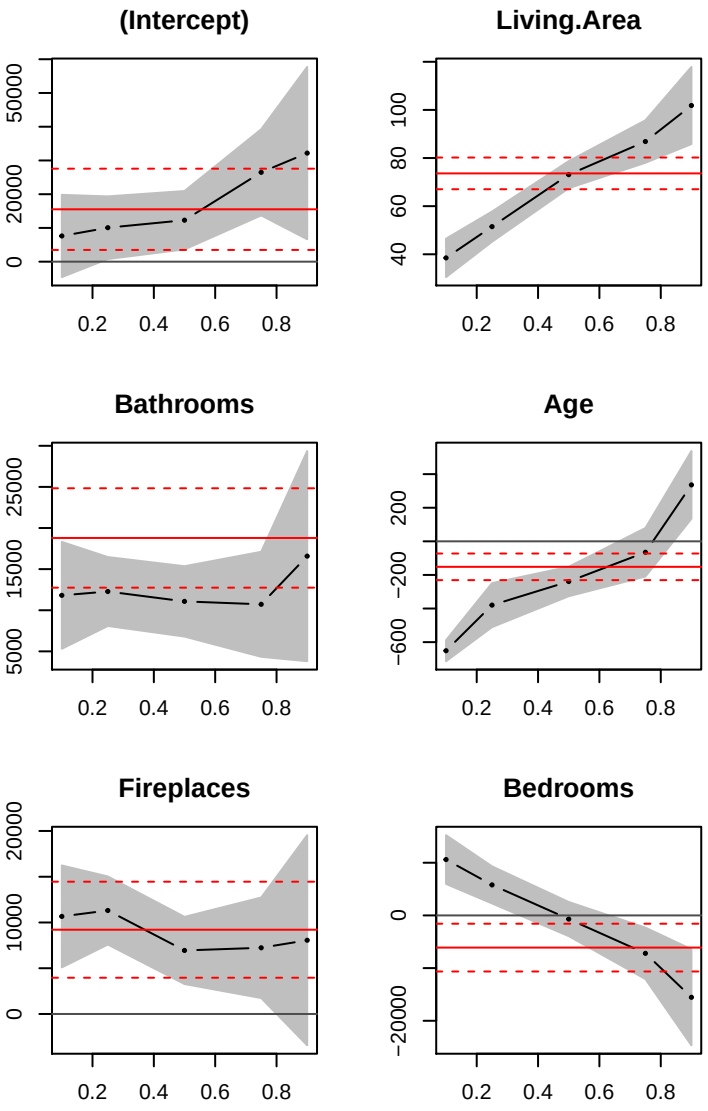


Figura 22: Variação das estimativas dos coeficientes

4.6.1 Interpretação dos gráficos e Análise dos Resultados

Baseado no gráfico da variação dos coeficientes (Figura 19 da subseção 4.6), notamos que o tamanho da sala de estar tem maior efeito na variação dos preços para casas mais caras (acima do quantil de ordem 0.50) do que para casas mais baratas (abaixo do quantil de ordem 0.50).

O efeito médio provocado nos preços das casas pelo aumento de uma unidade do preditor tamanho da sala-estar praticamente coincide com o efeito mediano.

Para os imóveis demasiadamente caros, a cada acréscimo de uma unidade no tamanho da sala, há um aumento de aproximadamente US$100 no preço.

Observa-se que o efeito de ter ou não lareira dentro da casa é uniforme e varia de forma insignificante em torno do efeito médio. Ou seja, desde casas mais baratas até as mais caras, o efeito é muito próximo.

A quantidade de dormitórios (Bedrooms) tem efeito positivo que decai da cauda inferior (preços muito baixos) até a cauda superior (preços bem elevados) da distribuição.

Um fato interessante a perceber é que, para casas com preços elevados (acima do quantil 0.50), o aumento na quantidade de dormitórios provoca uma desvalorização do imóvel, enquanto, para casas com preços reduzidos, há uma valorização.

A variação no preço do imóvel provocada pelo aumento na quantidade de banheiro está abaixo do efeito positivo médio, para os referidos quantis.

A idade do imóvel também influencia no preço. Para uma casa extremamente barata, um aumento de uma unidade na idade do imóvel provoca um decréscimo de aproximadamente US$600 no preço e, um acréscimo por volta de US$250, para imóveis com valores muito altos.

5 Conclusão

Nesse trabalho, vimos que a regressão quantílica pode caracterizar, por completo, a distribuição condicional da variável resposta dada as covariáveis, fornecendo uma modelagem estatística mais abrangente. Seus modelos mostraram-se úteis para detectar efeitos heterogêneos dos preditores.

Quando o interesse não é exclusicamente no comportamente médio e mas também em regiões próximas das caudas, a utilização da técnica de regressão quantílica torna-se muito importante.

As estimativas são mais consistentes e capazes de modelar a natureza heterocedástica dos dados, pois a variação das estimativas em função da ordem τ do quantil acompanha a variabilidade dos dados.

Nos exemplos de aplicação apresentados, nota-se que essa técnica fornece uma estrutura mais flexível quando há mudança da relação entre variáveis resposta e explicativas conforme o quantil que esteja sendo modelado.

A regressão por mqo não fornece boas estimativas quando suposições de homecedasticidade e normalidade são violadas. Desse modo, devemos adotar o método quantílico, principalmente, quando há interesse em estudar o comportamento entre variáveis resposta e preditoras em relação aos quantis extremos da distribuição condicional. Isso nos garante uma melhor modelagem.

Através da regressão mediana (caso particular de regressão quantílica), verifica-se, na presença de dados discrepantes, que a propriedade da robustez é satisfeita, ou seja, as estimativas dos coeficientes de regressão não são tão afetadas quanto na regressão por mqo.

Por fim, apresentamos uma noção sobre o método não paramétrico de estimação, que é uma alternativa para regressão quantílica assim como o método não paramétrico de Nadaraya-Watson é para regressão da média condicional.

6 Propostas para estudos futuros

Para estudos futuros, propomos as seguintes abordagens não exploradadas nesse trabalho:

1. Regressão quantílica paramétrica não linear;

2. Aspectos inferênciais para amostras finitas;

3. Inferência assintótica para regressão quantílica;

4. Método não paramétrico por meio do kernel bivariado;

7 Apêndice

7.1 Programação R

```
INSTALAÇÃO DE PACOTES UTILIZADOS NO TRABALHO

1)Regressão quantílica: install.packages("quantreg")

2) Outros: install.packages("texreg"); install.packages("ggplot2")
install.packages("tydiverse") ; install.packages("xtable")
install.packages("corrplot");install.packages("openxlsx")




############################################################
Gráfico da contribuição das covariáveis pra função perda
############################################################

rho<-function(u,tau){
       res1<-numeric()
       res2<-numeric()
     for (i in 1:length(u)){
       if (u[i]<0) {
       res1[i]<-u[i]*(tau-ifelse(u[i]<0,1,0))
        } else {
       res2[i]<-u[i]*(tau-ifelse(u[i]<0,1,0))
         }
     }
       return(list(neg=res1[!(is.na(res1))],
       pos=res2[!(is.na(res2))]))
}
dev.off()
par(mfrow=c(2,2))
set.seed(5849)
u<-runif(1000,-1,1)
tau=c(.25,0.5,0.75,0.9)
titulo<-c("Quantil 0.25", "Quantil 0.5","Quantil 0.75",
"Quantil 0.90")
```

```
for (i in 1:length(tau)){
perda<-rho(u,tau=tau[i])
plot(u[u<0],perda[[1]],xlim=c(-1,1),ylim=c(0,0.9),
   type="l",xlab=expression(u),ylab=expression(rho(u)),
      col="red")
lines(u[!(u<0)],perda[[2]],col="blue")
title(main = titulo[i], cex.main = 1,font.main= 1,
col.main= "blue")
}

#################################################
Gráfico da função perda pra cada ordem de quantil
#################################################
set.seed(12345)
amostra<-rnorm(1000)
taus=c(0.25,0.5,0.75)
indicadora=numeric();dif=numeric();res=numeric();sq=numeric()
s<-matrix(0,nrow=length(amostra),ncol=length(taus))
for (k in 1:length(taus)){
 for (i in 1:length(amostra)){
  for (j in 1:length(amostra)){
    dif[j]<-amostra[j] - amostra[i]
    indicadora[j]<-ifelse(dif[j]< 0,1,0)
    res[j]<-(dif[j])*(taus[k]-indicadora[j])
  }
  sq[i]<-sum(res)
  s[i,k]<-sq[i]
 }
}
colnames(s)<-c("SQ_tau1","SQ_tau2","SQ_tau3")
dt<-data.frame(amostra,s)
install.packages("ggplot2")
install.packages("reshape2")
require(ggplot2)
require(reshape2)
ggplot(data = dt, aes(x = amostra, y = SQ_tau1)) +
  geom_line(size=0.75, aes(x = amostra, y = SQ_tau1,color="SQ_tau1"),
```

```
  linetype="dotted") +
  geom_line(size=0.75, aes(y = SQ_tau2,color="SQ_tau2"),
  linetype="dashed")+
  geom_line(size=0.75, aes(y = SQ_tau3,color="SQ_tau3"),
  linetype="solid")+
  labs(title ="Soma da perda", x="c",
  y=expression(S[tau](c)),  colour=expression(tau)) +
  scale_color_manual(labels = c("0.25", "0.5","0.75"),
  values = c("blue", "red","green")) +
  theme_gray() +
  theme(legend.title = element_text(size=16, face="bold"),
  legend.direction = "vertical",
        legend.position=c(0.5, 0.8), text =
        element_text(size=16)) +
guides(color = guide_legend(override.aes = list(linetype =
c("dotted", "dashed","solid"))))

#############################################################
# REGRESSÃO  NÃO PARAMÉTRICA                                #
#############################################################

#############################
### funcao "fda"
### Calcula o valor estimado via kernel da f.d.a.
condicional de Y dado x
#############################
fda <- function(py,px,h) {
sum(dnorm((x-px)/hx,0,1)*pnorm((py-y)/h,0,1))/
sum(dnorm((x-px)/hx,0,1))
}

set.seed(63247)

# tamanho da amostra
n <- 1000
dev.off()
# covariavel
```

```
#x <- runif(n,1,10)
x<-x
# variavel resposta
#y <- max(x)+rnorm(n,x,abs(x))
y<-x
# ordem dos quantis
p <- c(0.1,0.25,0.5,0.75,0.9)

# numero de valores da covariável para os quais se calcula os
quantis

n.pontos <- 100

# valores da covariavel para os quais sao calculados os quantis
x.pontos <- seq(min(x),max(x),length.out=n.pontos)

# matrix com os quantis estimados para cada valor de x.pontos
mat.quantis <- matrix(0,n.pontos,length(p))

# parametro de suavizacao para o kernel da covariavel
#hx <- (max(x)-min(x))/5
s<-sd(x)
I<-IQR(x)/1.34
A<-min(I,s)
hx<-0.9*A*(length(x))^(-0.2) # h prescrito por Silverman(1992)
para o caso gaussiano
#hx <- sd(x)/8

# parametro de suavizacao para o kernel da variavel resposta
condicionado ao valor da covariavel
hy <- rep(0,n.pontos)
for (i in 1:n.pontos)
hy[i]<-sqrt(sum((y^2)*dnorm((x-x.pontos[i])/hx,0,1))/
sum(dnorm((x-x.pontos[i])/hx,0,1))-(sum(y*dnorm((x-x.pontos[i])/
hx,0,1))/sum(dnorm((x-x.pontos[i])/hx,0,1)))^2)
#hy[i]<-sqrt(sum((y^2)*dnorm((x-x.pontos[i])/hx,0,1)/
sum(dnorm((x-x.pontos[i])/hx,0,1)))-(sum(y*dnorm((x-x.pontos[i])/
```

```
hx,0,1)/sum(dnorm((x-x.pontos[i])/hx,0,1)))^2)

#hy[i]<-(1/5)*sqrt(sum((y^2)*dnorm((x-x.pontos[i])/hx,0,1)/
sum(dnorm((x-x.pontos[i])/hx,0,1)))-(sum(y*dnorm((x-x.pontos[i])/
hx,0,1)/sum(dnorm((x-x.pontos[i])/hx,0,1)))^2)

tol <- 1e-3  # Critério de parada do laço while

for (iq in 1:length(p)) {
  for (ix in 1:n.pontos) {
    linf <- min(y)
    lsup <- max(y)
    dif <- lsup-linf
    while (dif > tol) {
      lmed <- (linf+lsup)/2
      fda.linf <- fda(linf,x.pontos[ix],hy[ix])
      fda.lsup <- fda(lsup,x.pontos[ix],hy[ix])
      fda.lmed <- fda(lmed,x.pontos[ix],hy[ix])
      if (fda.lmed > p[iq]) lsup <- lmed
      else linf <- lmed
      dif <- dif/2
    }
    mat.quantis[ix,iq] <- lmed
  }
}



############ ESTIMADOR DE NADARAYA-WATSON #########

mNW <- function(x, X, Y, h) {

  # Argumentos da função mNW
  # x: pontos do grid de x
  # X: vetor amostral com n preditores
  # Y: vetor amostral de n variáveis respostas
  # h: parametro de suavização
```

```
  # K: kernel

  K<-function(x){(1/sqrt(2*pi))*exp(-(1/2)*x^2)}

   # Matrix com os valores de Kernel Kx
  Kx <- sapply(X, function(Xi) K((x - Xi) / h) / h)

  # Pesos
  W <- Kx / rowSums(Kx)

    drop(W %*% Y)
}
dev.off()
#dev.new()
col.plot <- c("black", "orange", "green", "blue","blueviolet")
plot(x,y,ylim=c(min(y),max(y)),xlab="",ylab="")
lines(x.pontos, mNW(x = x.pontos, X = x, Y = y, h = hx),
col = 2)
for (i in 1:length(p))
  lines(x.pontos,mat.quantis[,i],col=col.plot[i])

### FUNÇÃO PARA AJUSTAR MODELOS DE RQ
PARA DIVERSOS QUANTIS ##

install.packages("quantreg")
require(quantreg)
resumo.modelosRQ<-function(taus){
  modelos<-list()
  resumo.modelos<-list()
  for (i in 1:length(taus)){
    modelos[[i]]<-rq(C~Temp,data=dados,tau=taus[i])
    resumo.modelos[[i]]<-summary(modelos[[i]],se="iid")
  }
return(resumo.modelos)
}
resumo.modelosRQ(c(0.2,0.5,0.75,0.90))
```

```
### RETAS DE REGRESSÃO MEDIANA E POR MQO

ggplot(data=dadosPMA10,aes(x=V2,y=V1))+
      geom_point()+
      stat_quantile(aes(color="blue"),quantiles=0.5)+
      geom_smooth(aes(color="red"),method="lm",se=FALSE)+
      labs(x="n° de carros por hora",y="Concentração de NO2")+
      ggtitle("")+
      theme(plot.title=element_text(hjust=0.5))+
      scale_colour_manual(name="Regressão",
                      values=c("blue"="blue","red"="red"),
                      labels=c("Mediana","Por MQO"))

#### AJUSTE DE MODELOS DE REGRESSÃO QUANTÍLICA
PARA DIFERENTES GRUPOS ####

dados %>% group_by(Insulate)%>% ggplot(aes(x=Temp,y=Gas,color=Insulate)) +
geom_point()+
stat_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75,0.9))+
labs(title="Retas de Regressão Ajustadas",x="Temperatura(C)",y="Consumo de Gás")

###FUNCAO PARA CONSTRUÇÃO DAS CURVAS DE
REGRESSAO NÃO LINEAR###

require(quantreg)

Y<- function(x){
  v<-c(0.10,0.25,0.50,0.75,0.90)
  #x<-seq(min(distance),max(distance),1)
  coef<-matrix(0,length(v),2)
  vet<-matrix(0,nrow=length(x),ncol=length(v))
  rq.modelo<-list()
  rq.resumo<-list()
  for (j in 1:length(v)){
   for (i in 1:length(x)){
  rq.modelo[[j]]<-rq(consume~distance+I(distance^2)
  +I(distance^3),data=dados2_car,tau=v[j])
```

```
  rq.resumo[[j]]<-summary(rq.modelo[[j]])
  coef[j,1]<-rq.modelo[[j]]$coefficients[1]
  coef[j,2]<-rq.modelo[[j]]$coefficients[2]
  vet[i,j]<-coef[j,1]+coef[j,2]*x[i]^(-1/2)
   }
  }
  return(vet)
}
```

8 REFERÊNCIAS BIBLIOGRÁFICAS

Referências

Davino, C., Furno, M., & Vistocco, D. (2014). *Quantile regression*. Wiley Online Library.

Koenker, R. Quantile regression. 2005.

Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

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

Koenker, R. & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Koenker, R. & Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis / B.W. Silverman*. Chapman and Hall London ; New York.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer Series in Statistics. Springer.

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.