

Regressão Quantílica

Quando utilizar!?

*Luiz Fernando Palin Droubi**

Carlos Augusto Zilli†

Murilo Damian Ribeiro‡

Norberto Hochheim§

02/01/2020

Resumo

A NBR 14.653-02 ([ABNT, 2011](#)) recomenda que, na Engenharia de Avaliações de imóveis urbanos, para o tratamento dos dados seja utilizada metodologia científica, mesmo no tratamento de dados por fatores, o que usualmente é feito através do método da regressão linear clássica ou ordinária, ainda que a norma também cite outros métodos, como a regressão espacial, a análise envoltória de dados e as redes neurais artificiais. No entanto, através destes métodos, o que se obtém são coeficientes ou fatores **médios** da contribuição de uma característica do imóvel na formação do valor final. Ocorre que a contribuição de uma determinada característica para a formação do valor final dos imóveis pode ser diferente para os diferentes quantis da distribuição de probabilidades. É possível até que uma determinada característica que se mostre insignificante no método da regressão linear seja significativa na regressão quantílica, já que na regressão linear o que se estima é se, **em média**, uma determinada característica tem influência na formação do valor total de um imóvel. Ocorre que uma determinada característica pode influenciar positivamente o preço de venda dos imóveis de maior valor e negativamente o preço de venda dos imóveis de menor valor (ou *vice versa*), sendo que, **em média**, o seu efeito é nulo, o que no entanto não quer dizer que aquela variável não tenha qualquer influência na formação de preço dos imóveis do mercado em análise. Em suma, esta diferente contribuição das características no valor final dos imóveis, atualmente, é ignorada (utiliza-se apenas o valor médio), fazendo com que eventuais diferenças nos efeitos das características em imóveis de valores diferentes sejam negligenciadas. A regressão quantílica é um método que permite estimar a real influência de cada característica ao longo de toda a distribuição de probabilidades dos imóveis de um mercado, o que pode se demonstrar útil na avaliação de imóveis urbanos em determinados mercados o que atualmente pode passar despercebido aos olhos do avaliador acostumado com os métodos estatísticos clássicos.

*SPU/SC, lfpdroubi@gmail.com

†IFSC, carloszilli@gmail.com

‡UFSC, murilodamianr@gmail.com

§UFSC, hochheim@gmail.com

1 Introdução

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x 's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions. — [Mosteller e Tukey \(1977\)](#) (apud [KOENKER, R., 2000](#), p. 19)

Em muitas áreas da Ciência, como na Ecologia, é possível que um pesquisador, ao analisar dados como os da figura 1 com a utilização do consagrado método da regressão linear, chegue à conclusão de que não há qualquer correlação entre duas variáveis, já que, pelo estudo das variáveis, *em média*, o efeito do regressor encontrado é nulo ($\beta = 0$), tal como mostra a reta de regressão linear (tracejada) na figura.

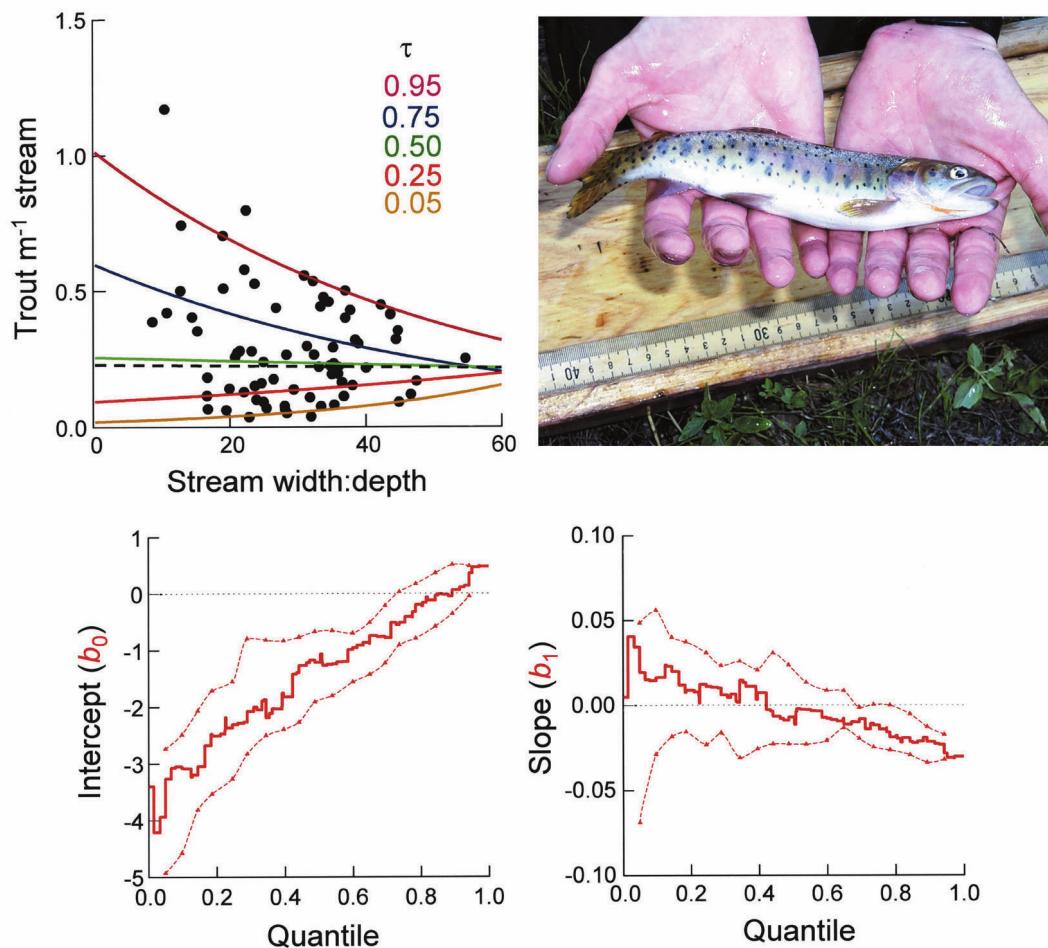


Figura 1: Mudanças na densidade de trutas (y) em função do quociente largura/profundidade de um canal (X). Fonte: [Cade e Noon \(2003, p. 413\)](#).

No entanto, como bem observaram [Cade e Noon \(2003, p. 412-413\)](#), se os autores

deste estudo tivessem se baseado *apenas* no método da regressão linear, eles teriam fatalmente chegado a esta conclusão errônea. Na realidade, porém, com o auxílio da regressão quantílica, mostrou-se que a relação existe, porém pode ser percebida apenas nos quantis superiores dos dados. Fisicamente, o que deve ser deduzido é que, com o aumento do quociente largura/profundidade de um canal, limita-se a densidade de trutas no canal. Esta limitação não é percebida nos quantis inferiores dos dados, de maneira que, *em média*, o seu efeito é nulo, mas claramente o efeito é considerável nos quantis superiores dos dados. Isto ocorre porque efeitos de fatores limitantes não são bem representados pela média das distribuições de probabilidades, onde a presença de muitos outros fatores limitantes não-medidos podem estar presentes (CADE; NOON, 2003, p. 413).

O mesmo comportamento pode ou poderia ser encontrado na Engenharia de Avaliações. Imagine-se que ao analisar um mercado de lotes urbanos o Engenheiro de Avaliações se depare com os dados mostrados na figura 2. Ao analisar os dados através da regressão linear à média (reta tracejada), o Engenheiro poderia, erroneamente concluir que a área não teria nenhum efeito sobre o valor unitário dos lotes, quando na realidade o que ocorre é que o efeito da variável área é o de aumentar suavemente o valor unitário dos lotes de menor valor, e diminuir de uma forma um pouco mais agressiva o valor unitário dos lotes dos quantis superiores.

Em suma, globalmente, o efeito médio é zero, o que não significa que a variável não possua qualquer significância na formação de preços dos imóveis.

A regressão quantílica permite que a influência de uma característica qualquer de um imóvel tenha efeitos diferentes para diferentes faixas de valores de imóveis.

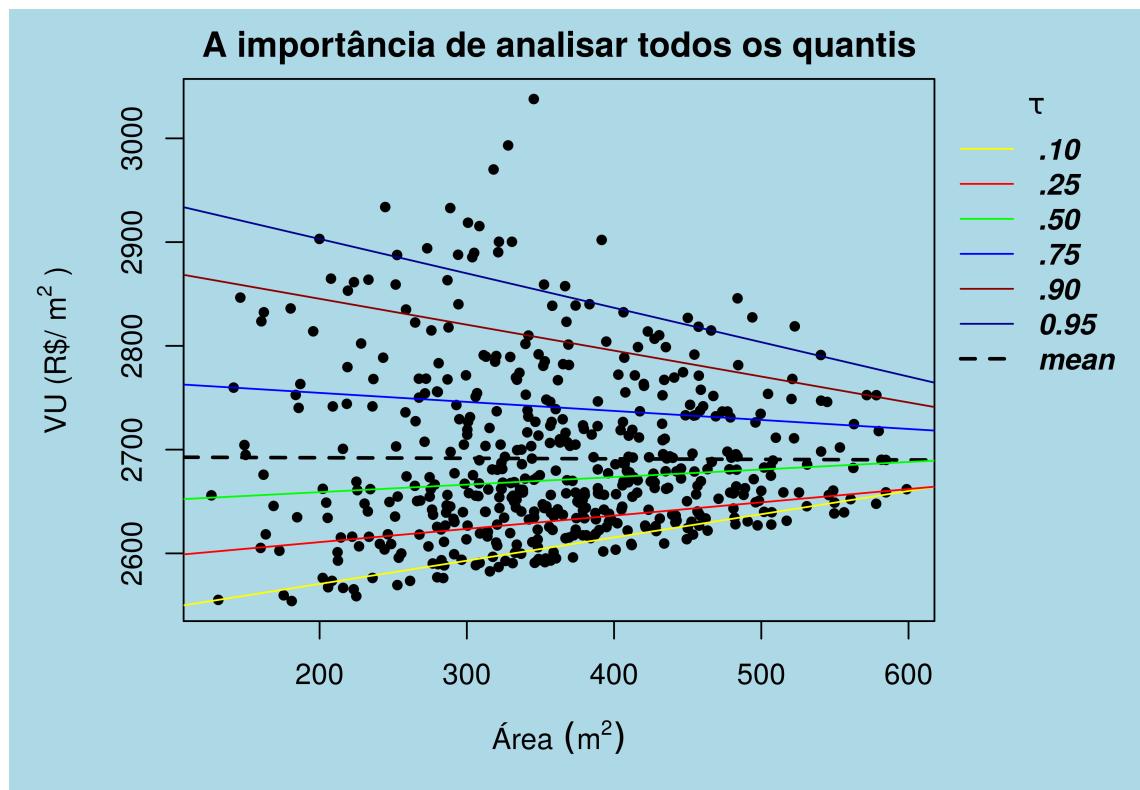


Figura 2: Dados gerados randomicamente para ilustrar a importância da regressão quantílica na Engenharia de Avaliações. Fonte: do Autor.

Joachim Zietz, Emily Norman Zietz e Sirmans (2008) mostram que os conflitos a respeito da contribuição de uma determinada característica na formação dos preços de venda de imóveis residenciais podem ser esclarecidos através da regressão quantílica. Diferentes valores para os coeficientes de regressão linear para algumas características podem ser encontrados ao longo da distribuição de preços de venda de imóveis. Ou seja, algumas características dos imóveis residenciais podem ser mais valorizados por compradores de imóveis de mais alto valor do que por compradores de imóveis de menor valor.

Segundo Joachim Zietz, Emily Norman Zietz e Sirmans (2008), variáveis como área construída, área do lote e número de banheiros tem um impacto maior nos imóveis de maior valor de venda, enquanto outras variáveis parecem ter um comportamento constante para todos os preços de venda de imóveis, como garagens e distância ao centro, entre outras.

2 Revisão Bibliográfica

2.1 Breve Histórico

Segundo Roger Koenker (2000, p. 347), o gráfico mais importante da história da estatística é o gráfico de Galton, reproduzido na figura 3.

O gráfico ilustra o fenômeno, descoberto por Galton, da regressão à média, cuja importância até hoje se faz presente em diversos estudos científicos. Em estudos clínicos para a aferição do real efeito de um novo medicamento, por exemplo, há necessidade de se estabelecerem dois grupos de pacientes (chamados de grupos de controle e de grupo de tratamento) para isolar os efeitos do tratamento pesquisado do efeito da regressão à média (ou reversão à média) (ver JAMES, 1973). Nestes estudos, apenas as pessoas do grupo de tratamento realmente são tratadas com o medicamento em teste. Desta maneira, a diferença das médias entre os grupos é isolada da variação biológica natural e da variação devido aos erros de aferição, que estão sujeitos à regressão à média.

Segundo Roger Koenker (2000, p. 349), a característica essencial da regressão linear clássica, derivada deste gráfico, é que o efeito do covariante na variável resposta é inteiramente capturado pela expressão abaixo, uma simples “mudança de local”, enquanto a aleatoriedade remanescente de Y dado X é modelada por um termo de erro aditivo independente de X .

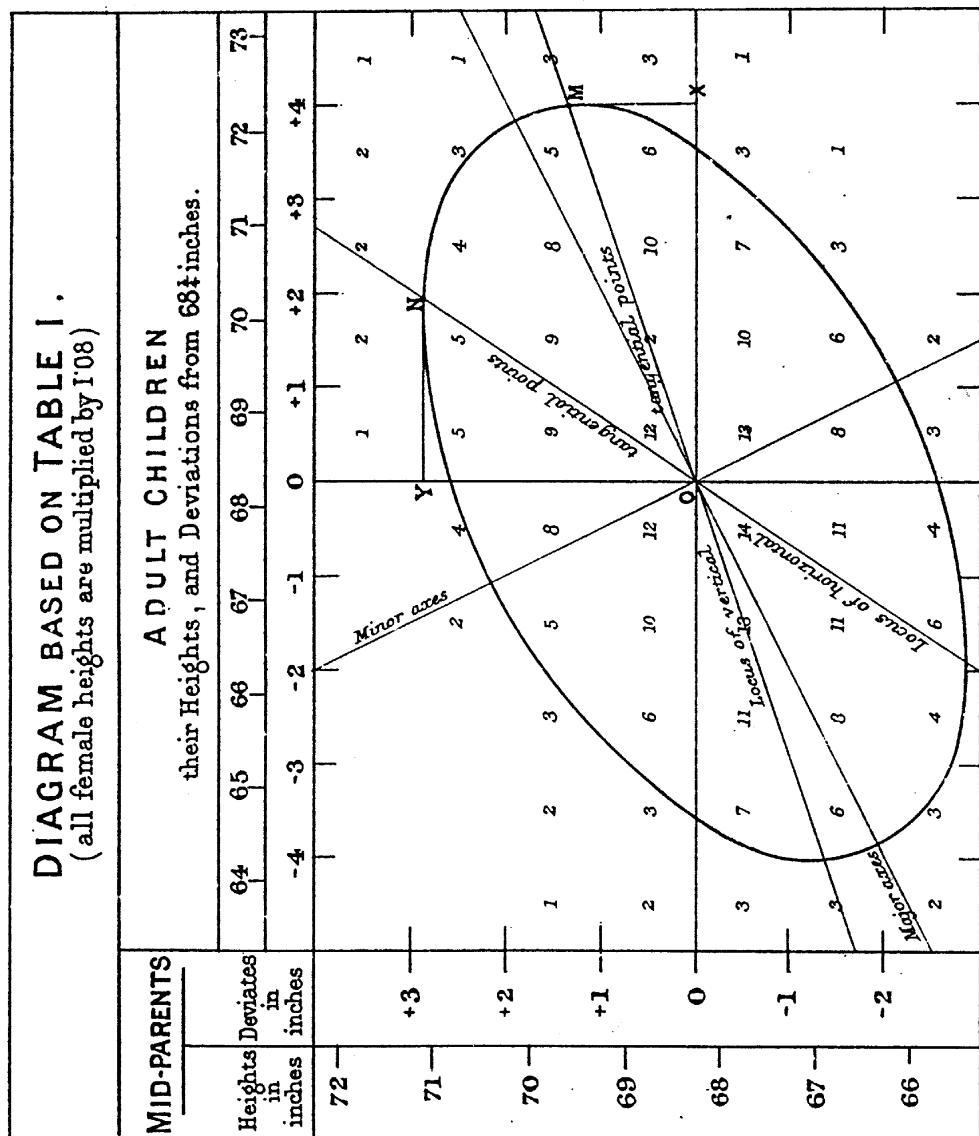
$$\mathbb{E}(Y|X = x) = x'\beta$$

Talvez prevendo que o seu método fizesse os seus colegas estatísticos se aterem apenas ao estudo das médias, Galton (1889, p. 62) (*apud* KOENKER, R., 2000, p. 350) repreendeu os seus colegas que:

limited their inquiries to Averages, and do not seem to revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of a native of one of our flat English counties, whose retrospect of Switzerland was

Plate X.

DIAGRAM BASED ON TABLE I.
 (all female heights are multiplied by 108)



J.P. & W.W. Holmes, M.D.

Figura 3: Regressão à média (Galton, 1885). Fonte: Roger Koenker (2000, p. 348)

that, if the mountains could be thrown into its lakes, two nuisances would be got rid of at once.

Ironicamente, contudo, apesar da repreensão de Galton, muito provavelmente foi o seu método o principal responsável por restringir o escopo das investigações estatísticas à comparação de médias por décadas (KOENKER, R., 2000, p. 350).

Muito anterior ao descobrimento por Galton (1889) do fenômeno da regressão à média e ainda anterior ao descobrimento do método dos mínimos quadrados por Legendre (1805)¹, Boscovich propôs, em 1760, o seguinte problema, em busca de estimar se a hipótese elipsoidal da forma da terra, prevista na obra-prima de Newton (1687), estaria correta, em função de cinco observações do comprimento de um grau de latitude obtidas em diferentes longitudes (KOENKER, R., 2000, p. 353; PORTNOY; KOENKER, R., 1997, p. 281; STIGLER, 1986, p. 40):

Encontrar $\hat{\alpha}$ e $\hat{\beta}$ tais que:

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

com $\sum \hat{u}_i = 0$ e $\sum |\hat{u}_i| = \min!$.

Boscovich (1770) propôs uma solução através de um algoritmo geométrico. Alguns anos depois, em 1789, Laplace (1793) resolveu o problema matematicamente, no que pode ser considerada a primeira análise de regressão (PORTNOY; KOENKER, R., 1997, p. 281).

Posteriormente, com a chegada do *méthode des moindres carrés*, ou seja, do método dos mínimos quadrados ordinários, o método dos mínimos erros absolutos de Boscovich/Laplace ficou em segundo plano. Posteriormente, Edgeworth (1887) argumentou que, quando os erros não seguem a lei de Gauss, a regressão à mediana poderia efetuar melhores estimativas. Então Edgeworth (1888), relaxando a restrição de que a soma dos resíduos seja igual a zero ($\sum \hat{u}_i = 0$) (PORTNOY; KOENKER, R., 1997, p. 281), propõe o primeiro o protótipo do que viria se tornar, na década de 1940, o algoritmo simplex, capaz de obter, de forma iterativa, o intercepto e o coeficiente angular da reta do método dos mínimos desvios absolutos.

Na década de 40 surgiram os primeiros algoritmos simplex destinados à otimização, algoritmos estes que se ajustam às necessidades do métodos dos mínimos desvios absolutos, que não possui solução analítica, mas iterativa. A primeira aplicação dos algoritmos simplex para a resolução do método dos mínimos desvios absolutos se deve a Charnes, Cooper e Ferguson (1955) (ver PORTNOY; KOENKER, R., 1997, p. 281; KOENKER, R., 2018, p. 4).

Barrodale e Roberts (1974) propuseram a forma moderna do algoritmo simplex que, por muitos anos e até hoje é utilizado para a minimização do erro médio absoluto.

Roger W Koenker e Bassett (1978) generalizaram o problema de minimização do erro médio absoluto, o que equivale à regressão à mediana, ao problema de encontrar os

¹Gauss (1809) ligou o método dos mínimos quadrados à distribuição normal mas a origem do método se deve ao trabalho pioneiro de Legendre (1805). Houve discordâncias entre os dois na disputa pela invenção do método, entre outros achados na época. Ver a este respeito Stigler (1977); Stigler (1981) e Stigler (1986).

diversos quantis de distribuição através da aplicação de uma função de perda assimétrica, correspondente àquele quantil, chegando-se assim à regressão quantílica.

2.2 Referencial teórico

2.2.1 Interpretação Geométrica

2.2.1.1 O gráfico dual

edgeworth criou o gráfico dual, ferramenta essencial para o desenvolvimento do seu procedimento de minimização dos resíduos absolutos. O gráfico dual nada mais é do que o gráfico, em duas dimensões, dos parâmetros α e β a serem estimados, com o parâmetro α nas ordenadas e o parâmetro β no eixo das abscissas.

2.2.1.2 Algoritmo inicial

[Edgeworth \(1888\)](#) mostrou que, para a regressão à mediana bivariada, pontos no espaço amostral correspondem a retas no espaço paramétrico $(x_i, y_i) \mapsto \{(\alpha, \beta) : \alpha = y_i - x_i\beta\}$ e retas através de pares de pontos no espaço amostral correspondem a pontos (interseção das duas retas geradas pelos pontos) no espaço paramétrico. Todos os pares de observações assim gerados produzem $\binom{n}{2}$ pontos no espaço paramétrico.

[Edgeworth \(1888\)](#) então propôs um algoritmo geométrico, iniciando em algum ponto do espaço paramétrico, iterativamente, seguindo o caminho mais íngreme, minimizando os resíduos, encontrando assim uma solução ótima. Este procedimento pode ser generalizado para n dimensões, porém

2.2.2 O problema de estimação de quantis como um problema de minimização

Pode-se demonstrar que, assim como a média aritmética μ de uma variável aleatória tem a propriedade de minimizar a soma dos desvios quadráticos de cada observação em relação a ela ([MATLOFF, 2017](#), p. 50), a mediana tem a propriedade de minimizar a soma dos desvios médios absolutos de cada observação ([MATLOFF, 2017](#), p. 260). Ou seja:

$$\begin{aligned}\mu(Y) &= \mathbb{E}Y = \arg \min_c \sum_{i=1}^n \frac{1}{n} (y_i - c)^2 \\ Me &= \arg \min_c \sum_{i=1}^n \frac{1}{n} |y_i - c|\end{aligned}$$

Sabe-se que a mediana de uma variável equivale ao quantil de 50%. Assim, outros quantis podem ser obtidos com formulação análoga à formulação acima, porém com a aplicação de uma função de perda assimétrica ($\rho_\tau(\cdot)$) em lugar da função módulo (ver figura 4):

$$Q_\tau(Y) = \arg \min_c \sum_{i=1}^n \rho_\tau(y_i - c)$$

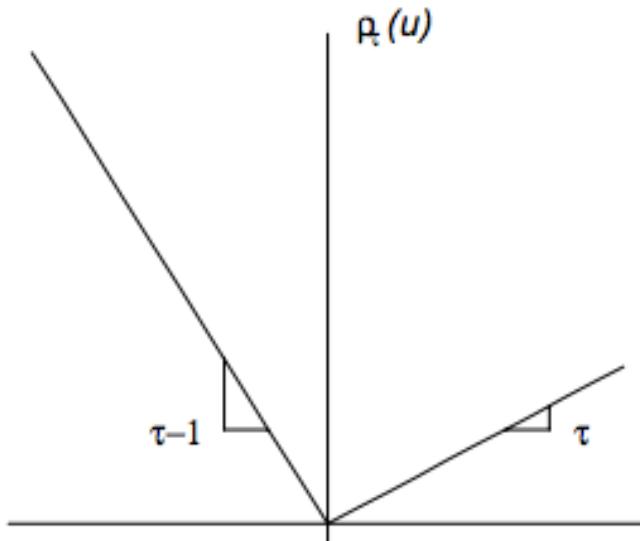


Figura 4: Função de perda ou custo.

Fonte: Roger Koenker e Hallock (2001).

2.2.3 Regressão linear e quantílica

A regressão linear pode ser vista como uma forma de minimização, assim como a média de uma população pode ser visto como o problema de minimização descrito acima.

A diferença é que no caso da regressão linear, ao invés de minimizar em relação a um escalar, desta vez se minimiza o erro em prever uma variável Y em relação a uma função de outra variável X, $f(X)$. Pode-se demonstrar que entre todas as funções $f(X)$, a que minimiza o erro médio quadrático de Y dado X ($\mathbb{E}[(Y - f(X))^2]$) é a função de regressão $\mu(t) = \mathbb{E}(Y|X = t)$ (MATLOFF, 2017, p. 49-50).

Analogamente, pode-se demonstrar que a mediana condicional é a função que minimiza o erro médio absoluto de Y dado X ($\mathbb{E}(|Y - f(X)|)$) (MATLOFF, 2017, p. 260-261).

2.2.3.1 Unicidade da solução

Pode-se demonstrar que a regressão linear, ou seja, a minimização de $\mathbb{E}[(Y - X\beta)^2]$ possui uma única solução e esta solução pode ser encontrada analiticamente, bastando para isso efetuar a derivação parcial deste termo em relação a β e igualando-o a zero, obtendo-se assim uma única solução para β a qual usualmente designa-se $\hat{\beta}$ (ver MATLOFF, 2017, p. 49-50).

O mesmo não se pode dizer da regressão à mediana e, mais genericamente, da regressão quantílica. Nesta abordagem, há múltiplas soluções possíveis, assim como numa amostra de tamanho par existem duas medianas possíveis. Ainda, as soluções do problema de minimização da regressão quantílica não podem ser encontradas analiticamente, sendo necessária a utilização de processos iterativos para a obtenção do(s) mínimo(s).

Contudo, deve-se ter em mente que, em ambos os processos de minimização, seja para a regressão linear ou para a regressão quantílica, trabalha-se com apenas uma amostra da população estudada. Desta forma, os valores de $\hat{\beta}$ encontrados são apenas estimativas dos valores reais de β , ou seja, os valores da população.

A diferença entre as múltiplas soluções da regressão quantílica é da ordem de $1/n$, enquanto a amplitude da precisão da estimativa é de tamanho $1/\sqrt{n}$. Assim, presume-se que as múltiplas soluções possíveis, para os casos práticos estão dentro da margem de erro para a primeira estimativa encontrada pelo algoritmo.

2.2.3.2 Robustez da solução

Uma das principais vantagens da regressão à mediana consiste na sua robustez à presença de *outliers*. Enquanto a solução da regressão linear é altamente sensível à presença de eventuais *outliers* na amostra, a solução de regressão quantílica é

2.2.3.3 Transformação e retransformação

Na regressão linear, com a aplicação do método à uma variável produto da transformação da variável dependente original, não é suficiente a simples retransformação dessa variável à escala original para a obtenção da solução, já que $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$, devido à desigualdade de Jensen (ver [DROUBI; HOCHHEIM; ZONATO, 2019](#), p. 207).

Já na regressão quantílica, com a aplicação de **transformações monotônicas**, pode-se afirmar que o quantil da variável transformada corresponde ao mesmo quantil da variável original. Assim, seja $f(\cdot)$ uma transformação monotônica qualquer em \mathbb{R} , e $Q(\cdot)$ a função quantil, para uma variável aleatória qualquer Y pode-se escrever:

$$Q_{f(Y)}(\tau) = f(Q_Y(\tau))$$

[Roger W Koenker e Bassett \(1978](#), p. 39-40) elencam ainda quatro propriedades de equivariância para a regressão quantílica, reproduzidas abaixo:

$$\begin{aligned} \hat{\beta}(\tau; \lambda y, X) &= \lambda \hat{\beta}(\tau; y, X), & \lambda \in [0, \infty), \\ \hat{\beta}(1 - \tau; \lambda y, X) &= \lambda \hat{\beta}(\tau; y, X) & \lambda \in (-\infty, 0], \\ \hat{\beta}(\tau; y + X\gamma, X) &= \hat{\beta}(\tau; y, X) + \gamma & \gamma \in \mathbb{R}^k, \\ \hat{\beta}(\tau; y, XA) &= A^{-1} \hat{\beta}(\tau; y, X) & A_{K \times K} \text{ não-singular.} \end{aligned}$$

2.2.3.4 Eficiência computacional

Segundo [Roger Koenker \(2018](#), p. 3), para amostras de tamanho menor do que alguns milhares de dados, com algumas dúzias de parâmetros a serem estimados, o que é hoje considerado um problema modesto pela ciência de dados, o algoritmo de [Barrodale e Roberts \(1974\)](#) é considerado extremamente eficiente.

De fato, segundo [Roger Koenker \(2018](#), p. 6), algoritmos do tipo simplex como o de **barrolade**, ditos *exterior point algorithms*, são mais eficientes para problemas de número moderado de dados e parâmetros.

Já para a solução de problemas com maior número de dados, procedimentos do tipo *interior point algorithms*, como são substancialmente mais rápidos e precisos (KOENKER, R., 2018, p. 6).

2.2.3.5 Eficiência estatística

Pode-se demonstrar que, se por um lado a média amostral tem eficiência estatística assintótica igual a 1 quando a distribuição dos dados é normal ou gaussiana, ela tem menos da metade da eficiência da mediana quando a distribuição for a distribuição de Laplace e tem eficiência zero para a distribuição de Cauchy (KOENKER, R. W.; BASSETT, 1978, p.36). Desta maneira, quando há incerteza quanto à real distribuição dos dados, a média, no caso escalar, ou a média condicional (regressão linear), considerada o estimador ótimo para o caso gaussiano, pode ser preferível à mediana, ou outros estimadores, ditos “ineficientes” (KOENKER, R. W.; BASSETT, 1978, p.36)

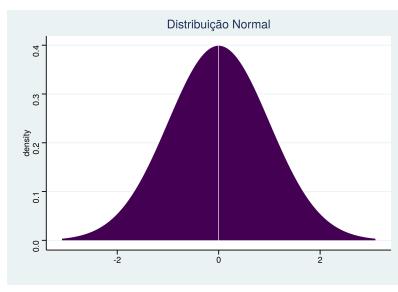
2.2.3.6 Estimador de máxima verossimilhança

Pode-se demonstrar que, quando a distribuição é normal o estimador de máxima verossimilhança para o parâmetro μ da distribuição é a média amostral.

Analogamente, se a distribuição dos dados for a distribuição de Laplace, o estimador de máxima verossimilhança para o parâmetro é a mediana.

Isto implica que, se a distribuição dos dados é normal, são necessários $\pi/2$ (1,57) vezes mais dados para que a estimativa de μ através da mediana seja tão eficiente quanto a estimativa através da média. Isto implica, por sua vez, que intervalos de confiança obtidos para a regressão quantílica são 25% mais largos do que os intervalos de confiança para a regressão linear (KOENKER, R., 2000, p. 354; DASGUPTA, 2008, p. 92).

No entanto, se a distribuição dos dados for a distribuição de Laplace, pode-se demonstrar que são necessários duas vezes mais dados para que a média estime μ com a mesma precisão da mediana.



$$\hat{\mu} = \frac{1}{n} \sum x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2}$$

$$f(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2/\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$$

Figura 5: Distribuição Normal.

$$\hat{\mu} = \arg \min_c \sum |x_i - c|$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n |x_i - \hat{\mu}|$$

$$f(x|\mu, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|x-\mu|}{\lambda}\right)$$

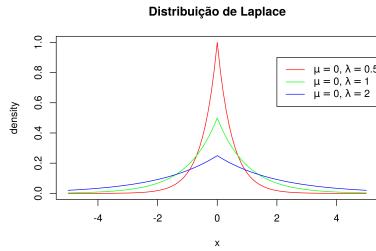


Figura 6: Distribuição de Laplace.

2.2.4 Inferência

2.3 Aplicações da regressão quantílica

2.3.1 Na Engenharia de Avaliações

[Joachim Zietz, Emily Norman Zietz e Sirmans \(2008\)](#) mostra...

3 Estudos de Caso

Para os estudos de caso foram utilizados os dados disponíveis em [Hochheim \(2015\)](#).

3.1 Duas dimensões

Assim como na regressão linear, é mais fácil a compreensão da regressão quantílica através de exemplos em duas dimensões, e depois generalizar para n dimensões.

Seja primeiramente o caso de dados heteroscedásticos. A figura 7 ilustra a aplicação da regressão quantílica e da regressão linear para este caso. Na figura 7, a reta vermelha é a reta de regressão linear entre as variáveis. A área sombreada em cinza é o intervalo de confiança para a regressão linear @80%. As retas azuis são as retas de regressão quantílica para os quantis 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 e 0,9.

A regressão quantílica neste caso pode ser usada para demonstrar a não validade dos intervalos de confiança (IC) e predição (IP) para a regressão linear para este tipo de dados: como a variância da população não é constante, mas aumenta com o aumento da área, as retas da regressão quantílica se abrem. Como os intervalos de confiança e predição na inferência clássica são calculados considerando-se que a variância da população é constante, este efeito não se observa no formato do IC.

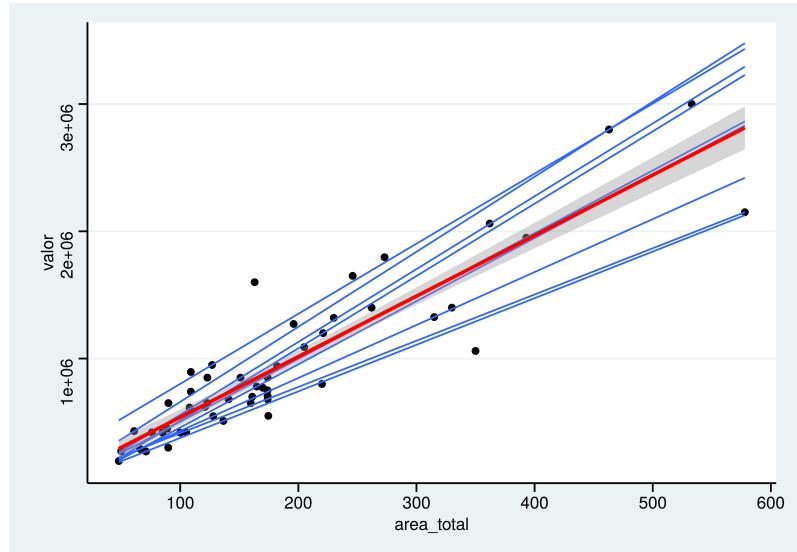


Figura 7: Regressão Linear e Quantílica para dados heteroscedásticos.

Assim como na regressão linear, uma conveniente transformação das variáveis pode ser aplicada para a obtenção da homoscedasticidade. Isto pode ser visto na figura 8, onde as retas para os diferentes quantis obtidas pela regressão quantílica agora são praticamente paralelas entre si, indicando que a heteroscedasticidade foi removida.

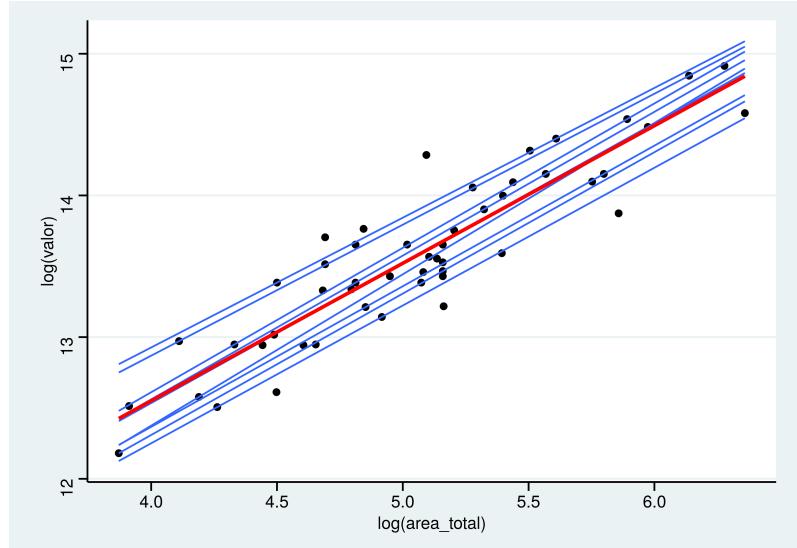


Figura 8: Regressão Linear e Quantílica com dados transformados.

Os coeficientes das retas de regressão quantílica podem ser plotados como na figura 9. Nesta figura, a reta cheia vermelha representa o coeficiente do modelo de regressão linear, enquanto a reta preta pontilhada representa os vários coeficientes da regressão quantílica. As retas vermelhas tracejadas representam o intervalo de confiança de estimativa do coeficiente de regressão linear. A área sombreada em cinza representa os intervalos de confiança para os coeficientes da regressão quantílica. Deve-se notar que, entre os quantis aproximados de 0,3 e 0,55, os coeficientes da regressão quantílica não são significantemente diferentes, estatisticamente, do coeficiente da regressão linear.

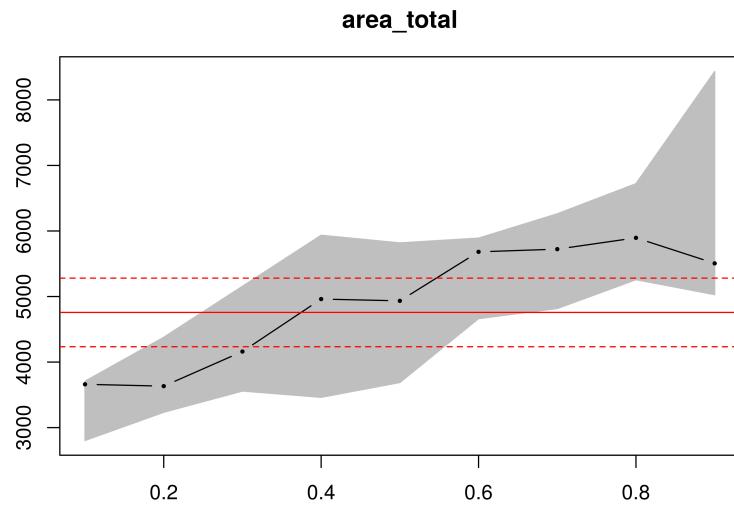


Figura 9: Variação dos coeficientes de regressão quantílica (variáveis originais).

Já para os dados transformados, pode-se notar na figura 10 que para todos os quantis, os coeficientes da regressão quantílica não podem ser considerados estatisticamente diferentes do coeficiente da regressão linear. Também se pode notar nesta figura como o estimador de regressão linear, para uma variável normalmente distribuída e na ausência de heteroscedasticidade, é mais eficiente do que o estimador da regressão quantílica, como a teoria já prevê (ver [Matloff \(2017\)](#), 238).

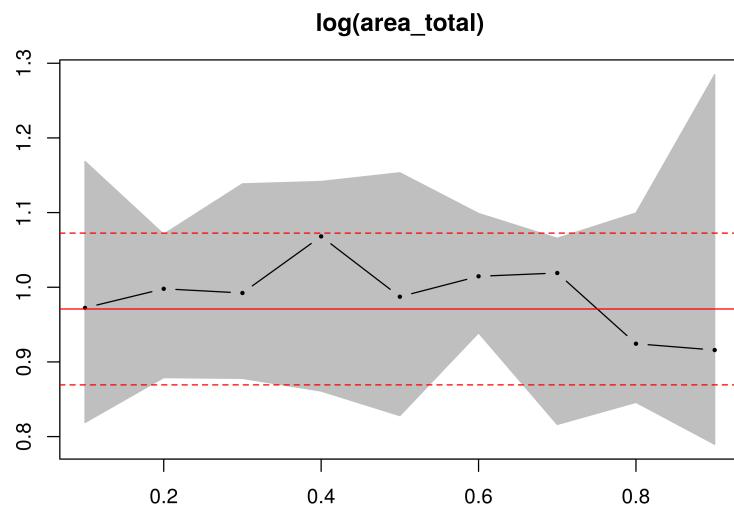


Figura 10: Variação dos coeficientes de regressão quantílica (variáveis transformadas).

3.2 Análise Multivariada

Para os dados obtidos de Hochheim ([2015](#), p. 22-23) foram ajustados dois modelos, um de regressão linear, com os dados saneados, e outro de regressão quantílica, utilizando-se a totalidade dos dados, para os quantis 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 e 0,9. Na figura 11 podem ser vistos os valores dos coeficientes de cada variável para os diferentes quantis.

Pode-se perceber, mais uma vez, que o valor dos coeficientes da regressão quantílica não diferem significantemente dos coeficientes da regressão linear (exceção para alguns quantis superiores nas variáveis `area_total` e `padrao`).

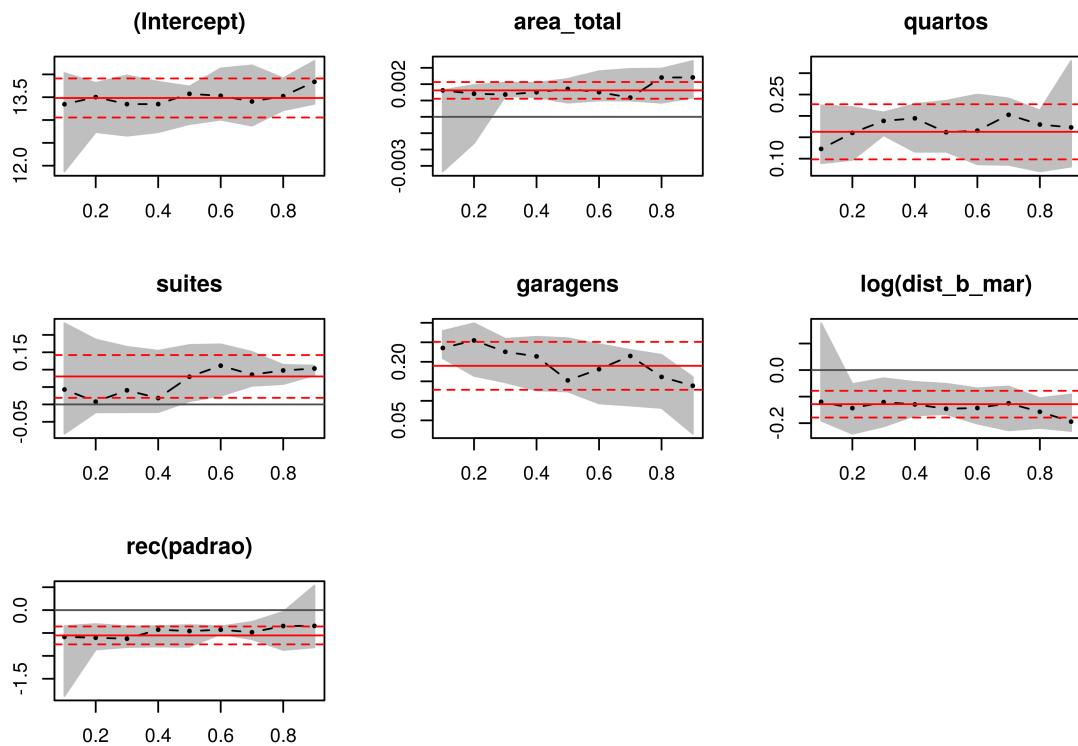


Figura 11: Coeficientes de regressão linear e quantílica. Análise multivariada.

Na tabela 1 podem ser vistos os coeficientes e estatísticas básicas dos modelos de regressão linear e de regressão à mediana (quantil 0,5).

3.2.1 Estimativas

É interessante comparar as estimativas obtidas com os modelos de regressão linear, com dados saneados, e o modelo de regressão à mediana, com a totalidade dos dados. Por um lado, o modelo de regressão linear tende a ser mais preciso para a estimativa da média, como prevê a teoria. Por outro lado, com mais dados, o modelo de regressão à mediana pode tornar-se mais eficiente.

Deve-se levar em conta que as estimativas com o modelo de regressão linear aqui apresentadas são para a mediana da distribuição lognormal.

Pelo modelo de regressão linear, o valor da estimativa central encontrado foi de R\$961.660,64, com intervalo de confiança entre R\$ r `brformat(exp(p[, "lwr"]))` e R\$ 1.000.024,94. A amplitude do intervalo de confiança foi de 7.83%.

Já pelo modelo de regressão quantílica, o valor da estimativa central encontrado foi de R\$946.467,87, com intervalo de confiança entre R\$ 886.472,34 e R\$ 1.010.523,85. A amplitude do intervalo de confiança foi de r `porcento(amplitude(exp(p1))/100)`.

Tabela 1: Comparação entre os modelos de regressão linear e regressão à mediana.

	<i>Dependent variable:</i>	
	<i>OLS</i>	<i>quantile regression</i>
	(1)	(2)
area_total	0.001 (0.001, 0.002) t = 5.113 p = 0.00001***	0.002 (0.001, 0.003) t = 2.300 p = 0.027***
quartos	0.164 (0.118, 0.209) t = 4.626 p = 0.00004***	0.162 (0.107, 0.217) t = 3.788 p = 0.0005***
suites	0.061 (0.018, 0.104) t = 1.810 p = 0.078***	0.080 (0.020, 0.139) t = 1.712 p = 0.095***
garagens	0.209 (0.166, 0.252) t = 6.247 p = 0.00000***	0.152 (0.075, 0.230) t = 2.520 p = 0.016***
log(dist_b_mar)	-0.141 (-0.176, -0.106) t = -5.174 p = 0.00001***	-0.146 (-0.210, -0.081) t = -2.904 p = 0.006***
rec(padrao)	-0.563 (-0.697, -0.428) t = -5.360 p = 0.00001***	-0.459 (-0.650, -0.267) t = -3.070 p = 0.004***
Constant	13.564 (13.268, 13.859) t = 58.847 p = 0.000***	13.574 (13.100, 14.047) t = 36.732 p = 0.000***
Observations	48	50
R ²	0.956	
Adjusted R ²	0.950	
Residual Std. Error	0.136 (df = 41)	
F Statistic	148.921*** (df = 6; 41)	

Note:

*p<0.3; **p<0.2; ***p<0.1

15 / 18

O modelo de regressão linear mostrou-se, portanto, mais eficiente do que o modelo de regressão a mediana, apesar no menor número de dados. Os limites inferior e superior do intervalo de predição @80% para o modelo de regressão linear são, respectivamente: R\$ 802.017,63 e R\$ 1.153.080,88.

Para o modelo de regressão quantílica, o intervalo de predição não faz qualquer sentido. No entanto, é possível estimar os valores diretamente para os quantis 0,1 e 0,9 da população. Nesta caso, os valores encontrados foram, respectivamente: R\$ 810.629,32 e R\$ 1.186.954,14.

Podem ainda ser calculados os intervalos de confiança @80% para as estimativas dos quantis 0,1 e 0,9.

Os limites inferior e superior do IC para o quantil 0,1 são, respectivamente: R\$ 781.253,06 e R\$ 841.110,17.

Os limites inferior e superior do IC para o quantil 0,9 são, respectivamente: R\$ 1.116.547,53 e R\$ 1.261.800,41.

Referências

ABNT. **NBR 14653-2:** Avaliação de Bens – Parte 2: Imóveis Urbanos. Rio de Janeiro, fev. 2011. p. 4.

BARRODALE, I.; ROBERTS, F. D. K. Solution of an Overdetermined System of Equations in the L1 Norm [F4]. **Commun. ACM**, ACM, New York, NY, USA, v. 17, n. 6, p. 319–320, jun. 1974. ISSN 0001-0782. DOI: [10.1145/355616.361024](https://doi.acm.org/10.1145/355616.361024). Disponível em: <<http://doi.acm.org/10.1145/355616.361024>>.

BOSCOVICH, Roger Joseph. **Voyage astronomique et geographique, dans l'état de l'église**. Paris: N. M. Tilliard., 1770.

CADE, Brian S.; NOON, Barry R. A Gentle Introduction to Quantile Regression for Ecologists. **Frontiers in Ecology and the Environment**, Ecological Society of America, v. 1, n. 8, p. 412–420, 2003. ISSN 15409295. Disponível em: <<http://www.jstor.org/stable/3868138>>.

CHARNES, A.; COOPER, W. W.; FERGUSON, R. O. Optimal Estimation of Executive Compensation by Linear Programming. **Management Science**, v. 1, n. 2, p. 138–151, 1955. DOI: [10.1287/mnsc.1.2.138](https://doi.org/10.1287/mnsc.1.2.138). eprint: <https://doi.org/10.1287/mnsc.1.2.138>. Disponível em: <<https://doi.org/10.1287/mnsc.1.2.138>>.

DASGUPTA, Anirban. **Asymtotic Theory of Statistics and Probability**. [S.l.]: Springer, 2008. DOI: [10.1007/978-0-387-75971-5](https://doi.org/10.1007/978-0-387-75971-5).

DROUBI, Luiz Fernando Palin; HOCHHEIM, Norberto; ZONATO, Willian. Estudos Interdisciplinares nas Ciências Exatas e da Terra e Engenharias. In: Florianópolis: Atena Editora, set. 2019. Avaliação pela moda, média ou mediana? DOI: [10.22533/at.ed.218191109](https://doi.org/10.22533/at.ed.218191109). Disponível em: <<http://droubi.me/sobrearea2018>>.

EDGEWORTH, Francis Ysidro. On observations relating to several quantities. **Hermathena**, Trinity College Dublin, v. 6, n. 13, p. 279–285, 1887. ISSN 00180750. Disponível em: <<http://www.jstor.org/stable/23036355>>.

- EDGEWORTH, Francis Ysidro. XXII. On a new method of reducing observations relating to several quantities. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 25, n. 154, p. 184–191, 1888. DOI: [10.1080/14786448808628170](https://doi.org/10.1080/14786448808628170). eprint: <https://doi.org/10.1080/14786448808628170>. Disponível em: <<https://doi.org/10.1080/14786448808628170>>.
- GALTON, F. **Natural Inheritance**. London: MacMillan, 1889.
- GAUSS, Carl Friedrich. **Theoria motus corporum coelestium in sectionibus conicis solem ambientium**. [S.l.: s.n.], 1809.
- HOCHHEIM, Norberto. **Engenharia de Avaliações - Módulo Básico**. Florianópolis: IBAPE - SC, 2015.
- JAMES, Kenneth E. Regression toward the Mean in Uncontrolled Clinical Studies. **Biometrics**, [Wiley, International Biometric Society], v. 29, n. 1, p. 121–130, 1973. ISSN 0006341X, 15410420. Disponível em: <<http://www.jstor.org/stable/2529681>>.
- KOENKER, Roger. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. **Journal of Econometrics**, v. 95, n. 2, p. 347–374, 2000. ISSN 0304-4076. DOI: [https://doi.org/10.1016/S0304-4076\(99\)00043-3](https://doi.org/10.1016/S0304-4076(99)00043-3). Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0304407699000433>>.
- _____. Handbook of Quantile Regression. In: New York: Chapman e Hall/CRC, 2018. Computational Methods for Quantile Regression, p. 13. DOI: [10.1201/9781315120256](https://doi.org/10.1201/9781315120256). Disponível em: <<http://www.econ.uiuc.edu/~roger/research/conopt/computation.pdf>>.
- KOENKER, Roger W; BASSETT, Gilbert. Regression Quantiles. **Econometrica**, v. 46, n. 1, p. 33–50, 1978. Disponível em: <<https://EconPapers.repec.org/RePEc:ecm:emetrp:v:46:y:1978:i:1:p:33-50>>.
- KOENKER, Roger; HALLOCK, Kevin F. Quantile Regression. **Journal of Economic Perspectives**, v. 15, n. 4, p. 143–156, dez. 2001. DOI: [10.1257/jep.15.4.143](https://doi.org/10.1257/jep.15.4.143). Disponível em: <<http://www.aeaweb.org/articles?id=10.1257/jep.15.4.143>>.
- LAPLACE, Pierre Simon. Sur quelques points du système du monde. **Mémoires de l'Academie Royale des Sciences de Paris**, p. 1–87, 1793. Reprinted in Laplace, 1878–1912, vol. 11, pp. 477–558.
- LEGENDRE, Adrien Marie. **Nouvelles méthodes pour la détermination des orbites des comètes**. [S.l.]: F. Didot, 1805.
- MATLOFF, Norman. **From Linear Models to Machine Learning**: Regression and Classification, with R examples. [S.l.]: Chapman & Hall, 2017. Disponível em: <<http://heather.cs.ucdavis.edu/draftregclass.pdf>>.
- MOSTELLER, Frederick; TUKEY, John W. **Data Analysis and Regression: a Second Course in Statistics**. [S.l.: s.n.], 1977. p. xvii + 588.
- NEWTON, Isaac. **Philosophiae Naturalis Principia Mathematica**. [S.l.: s.n.], 1687.
- PORTNOY, Stephen; KOENKER, Roger. 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. 279–296, 1997. ISSN 08834237. Disponível em: <<http://www.jstor.org/stable/2246216>>.

STIGLER, Stephen M. An attack on Gauss, published by Legendre in 1820. **Historia Mathematica**, v. 4, n. 1, p. 31–35, 1977. ISSN 0315-0860. DOI: [https://doi.org/10.1016/0315-0860\(77\)90032-5](https://doi.org/10.1016/0315-0860(77)90032-5). Disponível em: <<http://www.sciencedirect.com/science/article/pii/0315086077900325>>.

_____. Gauss and the Invention of Least Squares. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 9, n. 3, p. 465–474, mai. 1981. DOI: [10.1214/aos/1176345451](https://doi.org/10.1214/aos/1176345451). Disponível em: <<https://doi.org/10.1214/aos/1176345451>>.

_____. **The History of Statistics**: The Measurement of Uncertainty before 1900. Cambridge, Mass., & London, England: The Belknap Press of Harvard University Press, 1986. p. 410. ISBN 0674403401.

ZIETZ, Joachim; ZIETZ, Emily Norman; SIRMANS, G. Stacy. Determinants of House Prices: A Quantile Regression Approach. **The Journal of Real Estate Finance and Economics**, v. 37, n. 4, p. 317–333, 2008. Disponível em: <<https://EconPapers.repec.org/RePEc:kap:jrefec:v:37:y:2008:i:4:p:317-333>>.