Regressão Quantílica

Aplicações na Engenharia de Avaliações

Luiz Fernando Palin Droubi[*]
Carlos Augusto Zilli[†]
Murilo Damian Ribeiro[‡]
Norberto Hochheim[§]

11/12/2019

Resumo

A NBR 14.653-02 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, 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 seja 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, sendo apenas utilizado o valor médio, sendo que diferentes efeitos das características em imóveis de valores diferentes são 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

ZIETZ et al. (2008) mostra 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 ZIETZ et al. (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.

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

CADE; NOON (2003) mostra...

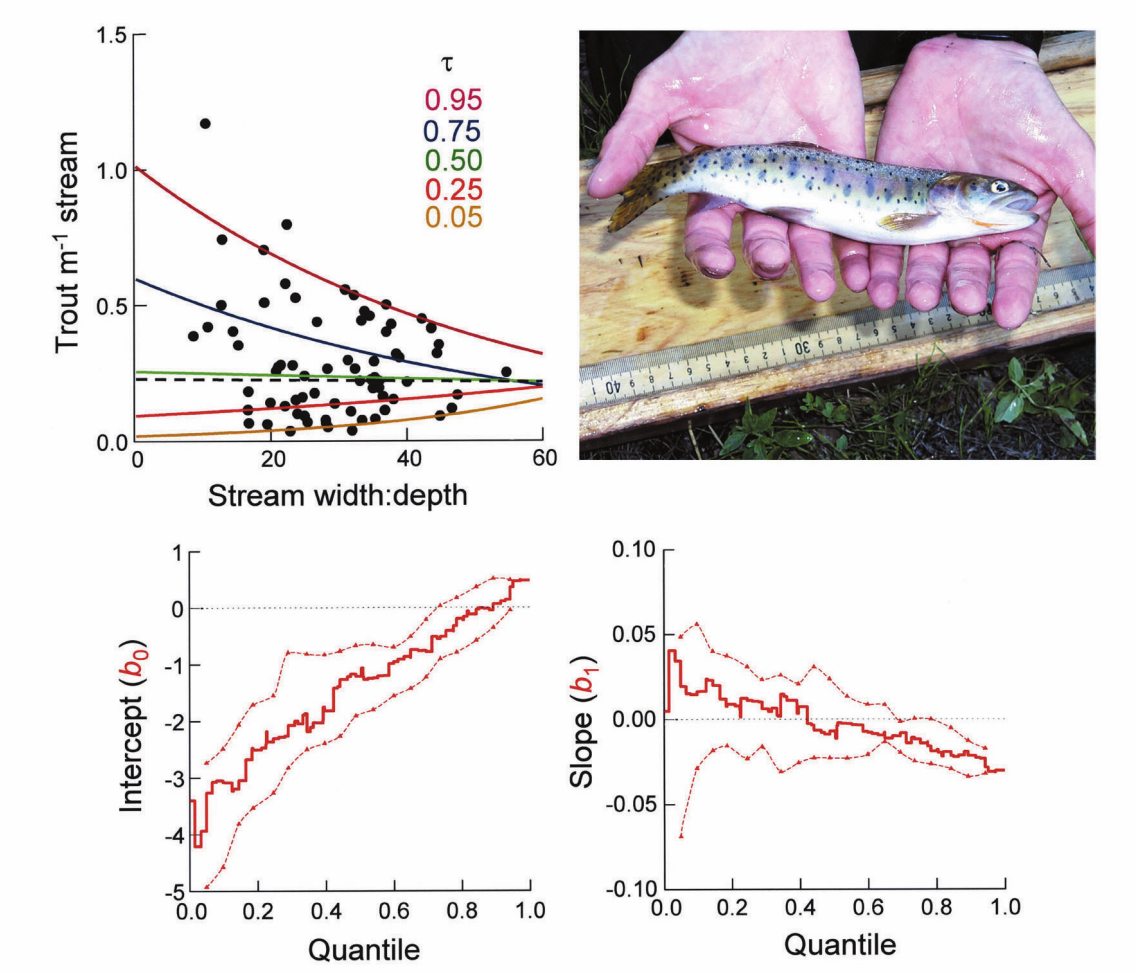

Figura 1: Mudanças na densidade de trutas (y) em função do quociente da largura sobre altura de um canal (X). Fonte: [@QReco].

2 Regressão Quantílica

2.1 Breve Histórico

STIGLER (1986)

Segundo Koenker (2009, p. 371), o gráfico mais importante da história da estatística é o gráfico de Galton, por nós reproduzido na figura 1.

O gráfico ilustra o fenômeno, descoberto por Galton, da regressão à média, cuja importância até hoje se faz presente em diversos estudos científicos, que estabelecem grupos de controle e tratamento para isolar os efeitos do tratamento pesquisado do efeito do fenômeno da regressão à média (ou reversão à média).

Segundo Koenker (2000), a característica essencial da regressão linear clássica, derivada deste gráfico, é que o efeito do covariante na variável resposta é inteiramente capturado pela seguinte expressão:

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

enquanto a aleatoriedade remanescente de Y dado X é modela por um termo de erro aditivo independente de X.

Boscovich propôs em 1760 (PORTNOY; KOENKER, 1997, p. 281) – portanto ainda antes da introdução do método dos mínimos quadrados por Legendre em 1805[1], no seu trabalho intitulado *Nouvelles méthodes pour la determination des orbites des comètes* –, o seguinte problema:

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

Laplace resolveu o problema matematicamente em 1789 (PORTNOY; KOENKER, 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 Laplace ficou em segundo plano, até que Edgeworth, em 1887 propõe o primeiro algoritmo capaz de obter o intercepto e o coeficiente angular da reta do método dos mínimos desvios absolutos, relaxando a restrição de que a soma dos resíduos seja igual a zero ($\sum \hat{u}_i = 0$) (PORTNOY; KOENKER, 1997, p. 281).

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 do método dos mínimos desvios absolutos se deve a Arrow e Hoffenberg, em 1959.(PORTNOY; KOENKER, 1997, p. 281).

[1] Gauss ligou o método dos mínimos quadrados à distribuição normal em seu trabalho intitulado *Theoria Motus Corporum Coelestium* de 1809 mas a origem do método se deve ao trabalho pioneiro de Legendre. Houve discordâncias entre os dois na disputa pela invenção do método, e outros achados na época. Ver a este respeito STIGLER (1977) e STIGLER (1981).

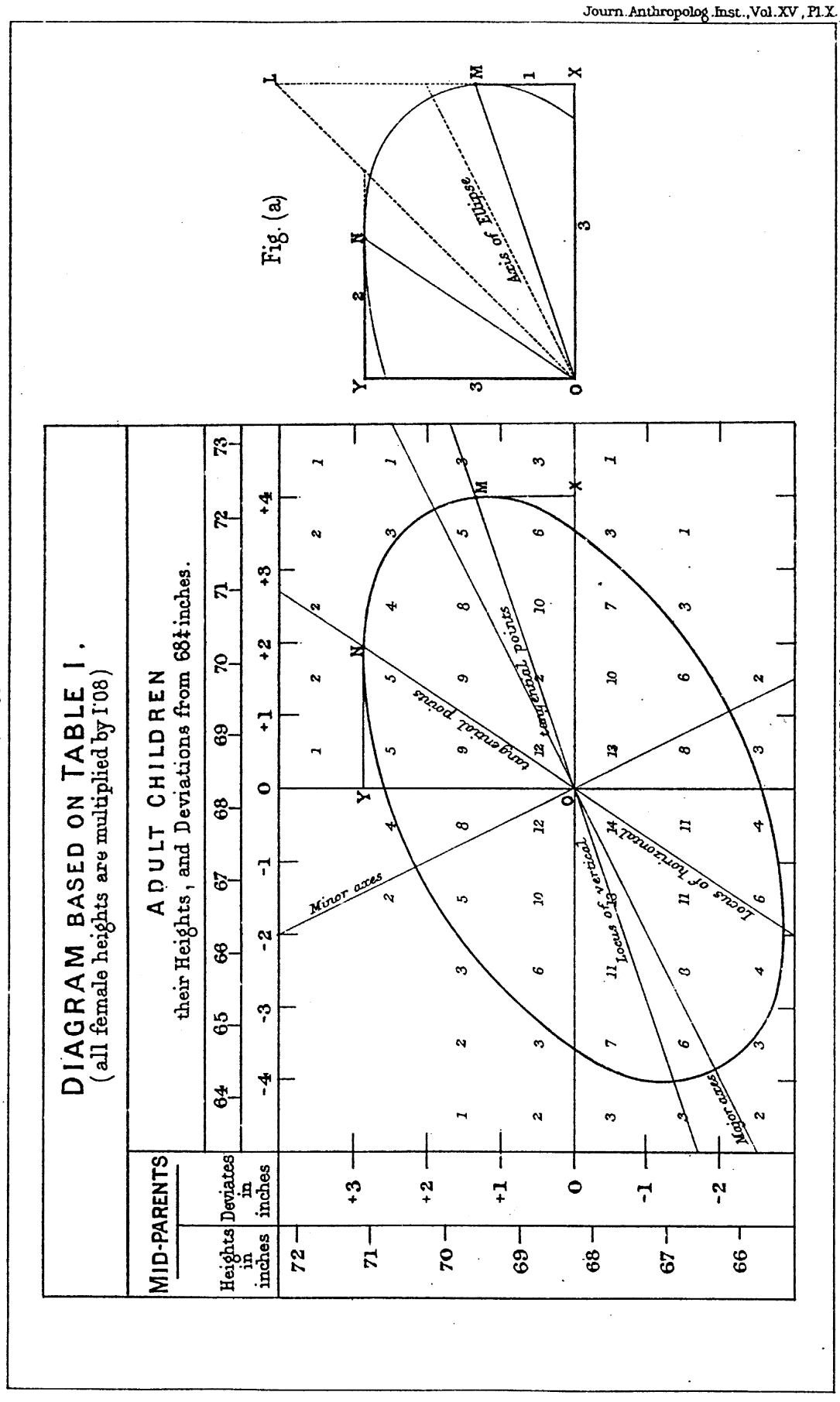

Figura 2: Regressão à média (Galton, 1885)

Koenker e Basset (1978) generalizaram o problema de minimização do erro médio absoluto, o que equivale à regressão à mediana, ao problema de encontrar os 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 Estimação de quantis

Existem diversas formas de se obter os quantis de uma amostra.

2.2.2 O problema de estimar 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:

$$\mu(Y) = \mathbb{E}Y = \arg\min_c \sum_{i=1}^{n} \frac{1}{n}(y_i - c)^2$$

$$Me = \arg\min_c \sum_{i=1}^{n} \frac{1}{n}|y_i - c|$$

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 ($\rho_\tau(.)$) em lugar da função módulo (ver figura 1):

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

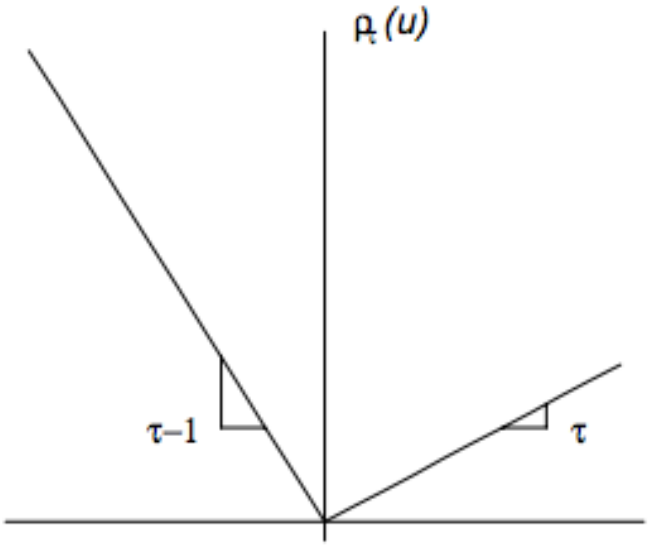

Figura 3: Função de perda ou custo.

Fonte: KOENKER; 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, pp. 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, pp. 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 e igualando-o a zero, obtendo-se assum um único solução para o cálculo do valor de β (ver MATLOFF, 2017, pp. 49–50).

O mesmo não se pode dizer da regressão à mediana a 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 precessos 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.

Assim, deve-se levar em conta que 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

2.2.3.3 Transformação e retransformação

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

2.2.3.4 Eficiência computacional

2.2.3.5 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, 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.

2.2.3.6 Média

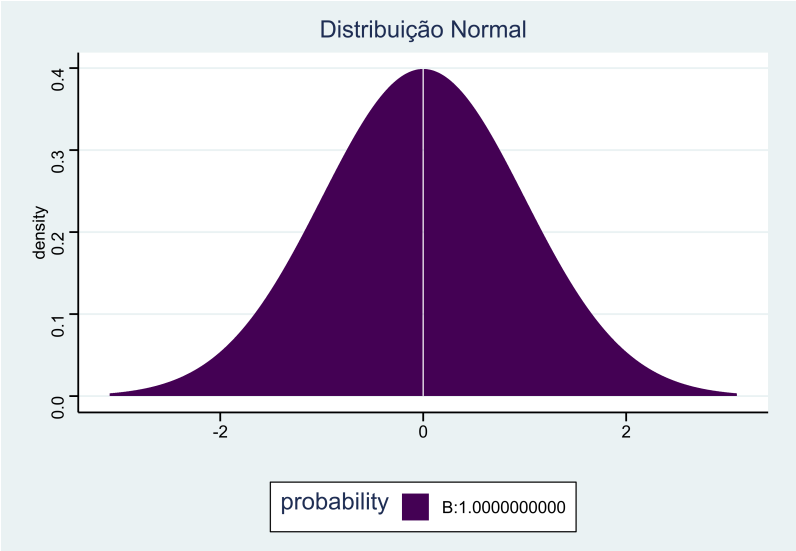

Figura 4: Distribuição Normal.

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

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

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

2.2.3.7 Mediana

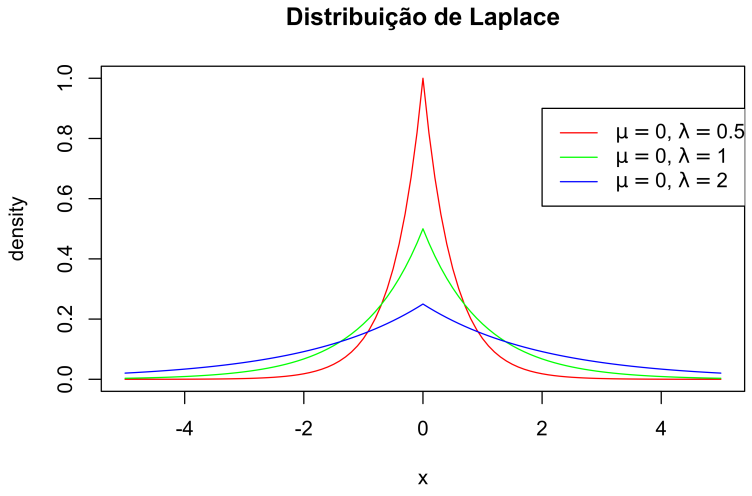

Figura 5: Distribuição de Laplace.

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

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

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

2.2.4 Inferência

2.3 Aplicações da regressão quantílica

2.3.1 Na Engenharia de Avaliações

ZIETZ et al. (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 aa 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 6 ilustra a aplicação da regressão quantílica e da regressão linear para este caso. Na figura 6, 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.

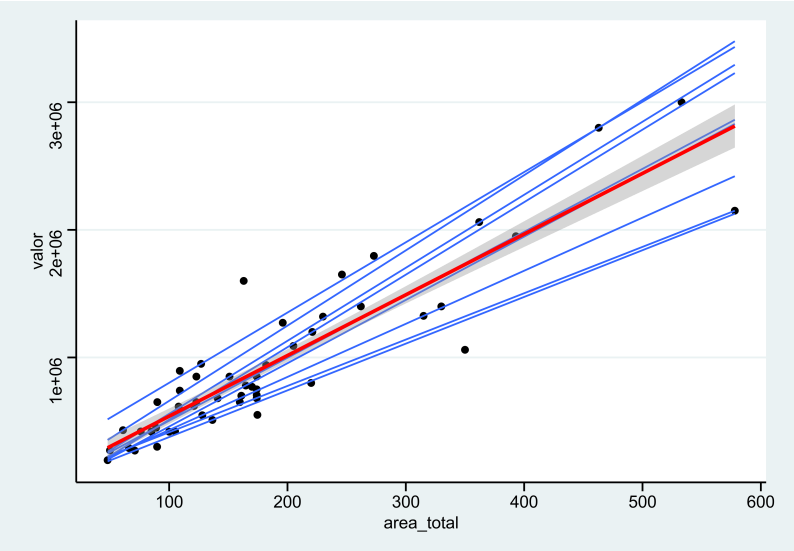

Figura 6: Regressão Linear e Quantílica para dados heteroscedásticos.

Assim como na regressão linear, uma conveniente transformação das variáveis pode ser aplicada para a obtenção da homoscedasticidade. Isto pode ser visto na figura 7, onde as retas para os diferentes quantis obtidas pela regressão quantílica agora são praticamente paralelas entre si, indicando que a heteroscedasticidade foi removida.

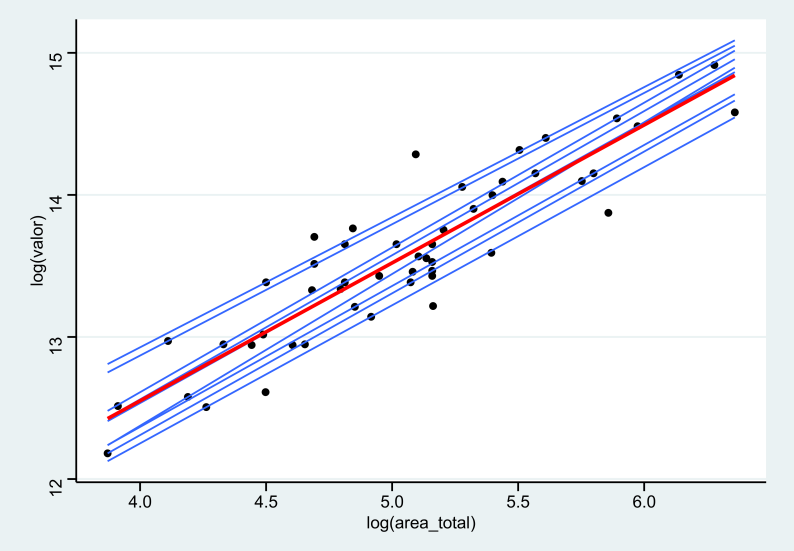

Figura 7: Regressão Linear e Quantílica com dados transformados.

Os coeficientes das retas de regressão quantílica podem ser plotados como na figura 8. 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 estimação 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 significamente diferentes, estatísticamente, do coeficiente da regreessão linear.

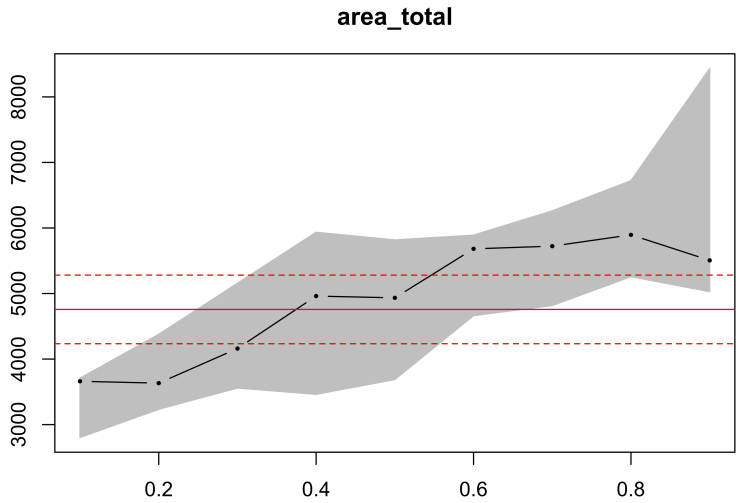

Figura 8: Variação dos coeficientes de regressão quantílica (variáveis originais).

Já para os dados transformados, pode-se notar na figura 9 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).

(Zilli, não sei se tu pesquisou isso na revisão bibliográfica, mas acho que se não, era bom colocar! Colocar algo do tipo: as vantagens e desvantagens da regressão quantílica. Apesar da regressão quantílica ser robusta à presença de *outliers*, ela é menos eficiente do que a regressao linear, caso a distribuição da variável estudada seja normal, claro.)

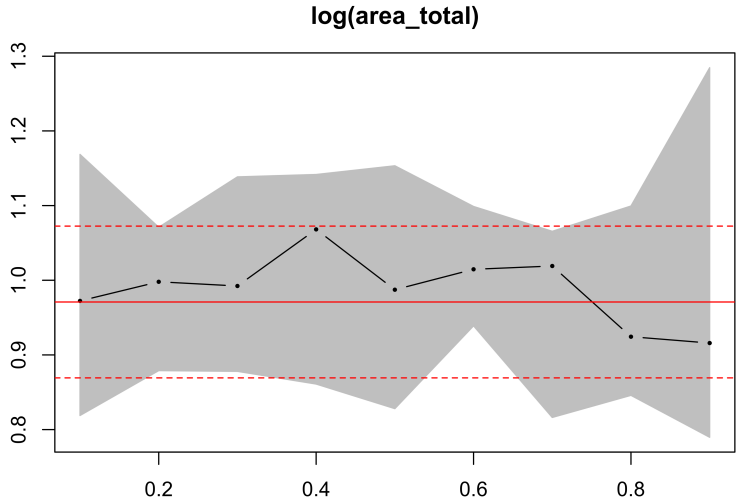

Figura 9: Variação dos coeficientes de regressão quantílica (variáveis transformadas).

3.2 Análise Multivariada

Para os dados obtidos de Hochheim (2015, pp. 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 10 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`).

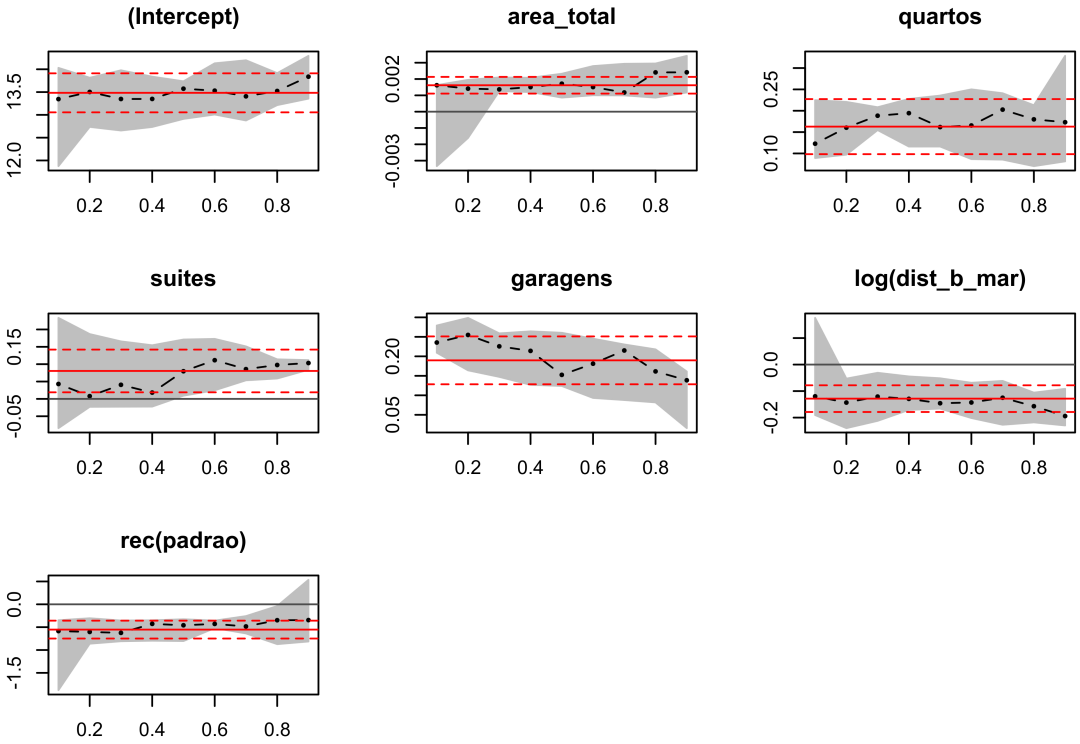

Figura 10: Coeficientes de regressão linear e quantílica. Análise multivariada.

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 estimação 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.

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

	Dependent variable:	
	log(valor)	
	OLS	*quantile regression*
	(1)	(2)
area_total	0.001	0.002
	(0.001, 0.002)	(0.001, 0.003)
	t = 5.113	t = 2.300
	p = 0.00001***	p = 0.027***
quartos	0.164	0.162
	(0.118, 0.209)	(0.107, 0.217)
	t = 4.626	t = 3.788
	p = 0.00004***	p = 0.0005***
suites	0.061	0.080
	(0.018, 0.104)	(0.020, 0.139)
	t = 1.810	t = 1.712
	p = 0.078***	p = 0.095***
garagens	0.209	0.152
	(0.166, 0.252)	(0.075, 0.230)
	t = 6.247	t = 2.520
	p = 0.00000***	p = 0.016***
log(dist_b_mar)	−0.141	−0.146
	(−0.176, −0.106)	(−0.210, −0.081)
	t = −5.174	t = −2.904
	p = 0.00001***	p = 0.006***
rec(padrao)	−0.563	−0.459
	(−0.697, −0.428)	(−0.650, −0.267)
	t = −5.360	t = −3.070
	p = 0.00001***	p = 0.004***
Constant	13.564	13.574
	(13.268, 13.859)	(13.100, 14.047)
	t = 58.847	t = 36.732
	p = 0.000***	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

Pelo modelo de regressão linear, o valor da estimativa central encontrado foi de R$961.660,64, com intervalo de confiança entre R$ 924.768,13 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 13.1%.

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

CADE, B. S.; NOON, B. R. A gentle introduction to quantile regression for ecologists. **Frontiers in Ecology and the Environment**, v. 1, n. 8, p. 412–420, 2003. Ecological Society of America. Disponível em: <http://www.jstor.org/stable/3868138>..

DASGUPTA, A. **Asymtotic theory of statistics and probability**. Springer, 2008.

HOCHHEIM, N. **Engenharia de avaliações - módulo básico**. Florianópolis: IBAPE - SC, 2015.

KOENKER, R. Galton, edgeworth, frisch, and prospects for quantile regression in econometrics. **Journal of Econometrics**, v. 95, n. 2, p. 347–374, 2000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304407699000433>..

KOENKER, R. The median is the message: Wilson and hilfertys experiments on the law of errors. **The American Statistician**, v. 63, n. 1, p. 20–25, 2009. Taylor & Francis. Disponível em: <https://doi.org/10.1198/tast.2009.0004>..

KOENKER, R.; HALLOCK, K. F. Quantile regression. **Journal of Economic Perspectives**, v. 15, n. 4, p. 143–156, 2001. Disponível em: <http://www.aeaweb.org/articles?id=10.1257/jep.15.4.143>..

KOENKER, R. W.; BASSETT, G. 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>..

MATLOFF, N. **From linear models to machine learning: Regression and classification, with R examples**. Chapman & Hall, 2017.

PORTNOY, S.; KOENKER, R. The Gaussian Hare and the Laplacian Tortoise: Computability of squared- error versus absolute-error estimators. **Statistical Science**, v. 12, n. 4, p. 279–296, 1997. Institute of Mathematical Statistics. Disponível em: <http://www.jstor.org/stable/2246216>..

STIGLER, S. M. An attack on Gauss, published by Legendre in 1820. **Historia Mathematica**, v. 4, n. 1, p. 31–35, 1977. Disponível em: <http://www.sciencedirect.com/science/article/pii/0315086077900325>..

STIGLER, S. M. Gauss and the invention of least squares. **Ann. Statist.**, v. 9, n. 3, p. 465–474, 1981. The Institute of Mathematical Statistics. Disponível em: <https://doi.org/10.1214/aos/1176345451>..

STIGLER, S. M. **The history of statistics: The measurement of uncertainty before 1900**. Cambridge, Mass., & London, England: The Belknap Press of Harvard University Press, 1986.

ZIETZ, J.; ZIETZ, E.; SIRMANS, G. 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>..