Avaliação pela Moda, Média ou Mediana?

Teoria e simulações

Luiz Fernando Palin Droubi[*]
Norberto Hochheim[†]
Willian Zonato[‡]

24/07/2018

"Eu sou o homem que com a máxima ousadia descobriu o que já fora descoberto antes." (CHESTERTON, 2008, p. 12).

1 INTRODUÇÃO

Existe na área da avaliação de imóveis uma discussão frequente e a nosso ver indesejável a respeito da adoção da estimativa de tendência central adotada para a predição de valores quando da utilização de modelos lineares log-normais, isto é, modelos em que a variável resposta aparece transformada pela função logaritmo natural.

Como veremos oportunamente, quando um modelo linear log-normal estiver razoavelmente bem-ajustado, com um baixo erro-padrão, a adoção de qualquer estimativas de tendência central, moda, média ou mediana, resultará em valores praticamente equivalentes.

No entanto, na presença de grande dispersão dos dados, os valores do erro-padrão da regressão linear pode se tornar relativamente alto, e a diferença entre as avaliação por uma ou outra medida de tendência central pode tornar-se relevante, levando a uma situação altamente indesejável: um imóvel poderá ser "corretamente" avaliado por dois avaliadores independentes com uma diferença significativa entre elas, haja vista que a NBR14.653-02 (2011) se omite a este respeito.

Pretende-se com este artigo dar a este problema uma abordagem formal, com o intuito de sugerir uma padronização das avaliações, sem no entanto especificar qual medida de tendência central é a correta, haja vista que todas elas tem seus prós e contras e nenhuma delas pode ser dita melhor do que a outra.

Como veremos adiante, no entanto, a escolha da utilização de uma ou outra medida deveria ser prévia à escolha do método, pois propiciaria assim a escolha de um método mais adequado à previsão daquela medida.

O avaliador, por exemplo, pode entender que a medida de tendência central mais adequada é a mediana, haja vista que esta é sensivelmente menos afetada pela presença de eventuais *outliers* no conjunto de dados. Sugeriríamos, neste caso, a adoção da técnica da regressão à mediana, método muito bem fundamentado, tal qual a regressão linear clássica e disponível em vários *software* estatísticos.

[*]SPU/SC, luiz.droubi@planejamento.gov.br
[†]UFSC, hochheim@gmail.com
[‡]SPU/SC, willian.zonato@planejamento.gov.br

Porém, adotado o método da regressão linear clássica, entendemos que a escolha do estimador deveria sempre ser a média, haja vista que o método de regressão linear é, por definição (como veremos oportunamente), uma regressão *à média* da variável dependente.

Ou seja, a escolha apropriada do método para a realização da avaliação poderia até ficar a cargo do avaliador (ou a cargo do contratante), mas dado o método, entendemos que caberia à NBR14.653-02 especificar o estimador adequado.

2 DESENVOLVIMENTO E FUNDAMENTAÇÃO

> Major Point 1: When we talk about the relationship of one variable to one or more others, we are referring to the regression function, which expresses the mean of the first variable as a function of the others. The key word here is *mean*! (MATLOFF, 2009, p. 386, grifo do autor)

2.1 Valor Esperado

Segundo BENNETT (2006), a **esperança matemática** ou **valor esperado** de uma variável aleatória é a soma do produto de cada probabilidade de saída da experiência pelo seu respectivo valor. Isto é, representa o valor médio 'esperado' de uma experiência se ela for repetida muitas vezes. Matematicamente, a Esperança de uma variável aleatória X é representada pelo símbolo $\mathbb{E}(X)$

Segundo Matloff (2009, p. 42), o valor esperado tem um papel central em probabilidade e estatística. A definição mais ampla de valor esperado de uma variável aleatória X, válida tanto para variáveis discretas como contínuas, é:

$$\lim_{n\to\infty} = \frac{X_1 + \ldots + X_n}{n}$$

2.1.1 Cômputo do Valor Esperado de uma variável aleatória discreta

Segundo Matloff (2009, p. 44), o valor esperado de uma variável aleatória X que assume valores definidos no conjunto A é:

$$\mathbb{E}(X) = \sum_{c\in A} cP(X = c)$$

2.1.2 Cômputo do Valor Esperado de uma variável aleatória contínua

O Valor Esperado de uma variável aleatória contínua W pode ser escrito da seguinte forma (MATLOFF, 2009, p. 128)

$$\mathbb{E}(W) = \int_{-\infty}^{\infty} t f_W(t) dt$$

onde $f_Y(x)$ é a função densidade de probabilidade de x.

2.1.3 Propriedades do Valor Esperado

Seja a um escalar e U uma variável aleatória (MATLOFF, 2017, p. 47):

$$\mathbb{E}(aU) = a\mathbb{E}(U)$$

Sejam a e b dois escalares e U e V duas variáveis aleatórias, não necessariamente independentes, então:

$$\mathbb{E}(aU + bV) = a\mathbb{E}(U) + bE(V)$$

Finalmente, sejam U e V duas variáveis aleatórias *independentes*:

$$\mathbb{E}(UV) = \mathbb{E}(U)\mathbb{E}(V)$$

Porém, se U e V não forem independentes, esta propriedade falha (covariância).

2.1.4 Lei da expectativa total

(MATLOFF, 2009, p. 339)

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|W)]$$

2.1.5 Lei da Variância total

(MATLOFF, 2009, p. 345)

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|W)] + \text{Var}[\mathbb{E}(Y|W)]$$

2.2 Desigualdade de Jensen

Segundo , seja $\varphi(x)$ uma função convexa, então:

$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\varphi(X)\right]$.

Como pode-se demonstrar, a função e^x é uma função convexa, pois possui derivada segunda sempre maior que zero ($f'' = e^x > 0$).

2.2.1 Erro médio quadrático (MSE)

Seja π o valor de uma estimativa. Então o seu erro médio quadrático (MSE) é dado por:

$$\text{MSE} = \int (y - \pi) f(y) dy = \mathbb{E}[(y - \pi)^2] = \mathbb{E}(y^2) - 2\pi\mathbb{E}(y) + \pi^2$$

Para encontrar o valor mínimo do erro médio quadrático (MSE) quando π varia, fazemos:

$$\frac{d(\mathbb{E}(y^2) - 2\pi\mathbb{E}(y) + \pi^2)}{d\pi} = 0 \therefore \pi = \mathbb{E}(y)$$

Ou seja, a estimativa pelo valor esperado é a estimativa que minimiza e erro médio quadrático.

2.2.2 Valor Esperado condicional

O valor esperado de uma variável aleatória y estatísticamente relacionada com outra outra variável aleatória x é:

$$\mathbb{E}(y|x) = \int y \frac{f(x, y)}{f(x)} dy$$

onde:

- $f(x, y)$ é a função densidade da distribuição de probabilidade conjunta de x e y e
- $f(x) = \int f(x, y) dy$ é a função de distribuição de probabilidade condicional de x.

2.3 Estimadores

> Earlier, we often referred to certain estimators as being "natural." For example, if we are estimating a population mean, an obvious choice of estimator would be the sample mean. But in many applications, it is less clear what a "natural" estimate for a population quantity of interest would be. We will present general methods for estimation in this section. We will also discuss advanced methods of inference (MATLOFF, 2009, p. 303).

A definição de um *estimador* para um parâmetro ou uma variável θ é uma função $\hat{\theta}(X)$, que mapeia o espaço amostral para um conjunto de estimativas amostrais, em que X é uma variável aleatória dos dados observados. É usual denotar uma estimativa em para um determinado ponto $x \in X$ por $\hat{\theta}(X = x)$ ou, mais simplesmente, $\hat{\theta}(x)$.

2.3.1 Propriedades de um estimador

Nesta seção adotamos que $\hat{\theta}$ é um estimador da variável aleatória θ.

2.3.1.1 Erro

$$e(x) = \hat{\theta}(x) - \theta$$

2.3.1.2 Desvio

$$d(x) = \hat{\theta}(x) - \mathbb{E}(\hat{\theta}(X))$$

onde $\mathbb{E}(\hat{\theta}(X))$ é o Valor Esperado do estimador.

2.3.1.3 Variância

A variância de um estimador θ é (MATLOFF, 2009, p. 52):

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

2.3.1.4 Coeficiente de Variação

O coeficiente de variação de um estimador é uma medida admensional que compara o desvio-padrão de uma variável ou estimador θ à sua média, conforme abaixo (MATLOFF, 2009, p. 56):

$$CV = \frac{\sqrt{\text{Var}(\hat{\theta})}}{E[\hat{\theta}]}$$

2.3.1.5 Viés

O viés de um estimador $\hat{\theta}$ é (MATLOFF, 2009, p. 317):

$$\text{B}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

O viés coincide com o valor esperado do erro, pois $\mathbb{E}(\hat{\theta}) - \theta = \mathbb{E}(\hat{\theta} - \theta)$.

Numa regressão linear:

$$\text{B}[\hat{\mu}(x_0)] = \mathbb{E}[\hat{\mu}(x_0)] - \mu(x_0)$$

2.3.1.6 Erro médio quadrático

Segundo Shen e Zhu (2008, p. 553), o erro médio quadrático é uma medida comum da qualidade de um estimador na literatura estatística.

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Numa regressão linear, o erro médio quadrático pode ser descrito por:

$$\text{MSE}[\hat{\mu}(x_0)] = \mathbb{E}[\hat{\mu}(x_0) - \mu(x_0)]^2 = \text{Var}[\hat{\mu}(x_0)] + \text{B}^2[\hat{\mu}(x_0)]$$

2.3.1.7 Consistência

$$\lim_{n\to\infty} \hat{\theta} = \theta$$

2.3.2 Melhor estimador linear não-inviesado ou BLUE

Em estatística, é comum o uso da sigla BLUE (*Best Linear Unbiased Estimator*) para indicar o melhor estimador linear não-enviesado.

2.3.3 Tradeoff entre viés e variância

Um dos problemas conhecidos dos modelos de regressão linear ou outros modelos estatísticos em geral é o sobreajustamento (do inglês *overfitting*). Resumidamente, *overfitting* é o ato de ajustar um modelo tão bem ajustado aos dados amostrais, que este se torna incapaz de fazer boas previsões para outros dados que não os do modelo. Segundo Matloff (2017, p. 24), um modelo sobreajustado é um modelo tão elaborado que "capta o ruído ao invés do sinal".

Segundo Matloff (2017, pp. 24–26), pelo contrário, um modelo com menor número de variáveis explicativas estará enviesando os seus resultados (no sentido de enviesamento sistêmico, inerente à amostragem, não proposital), e o acréscimo de uma variável independente a este modelo estará assim reduzindo o seu viés.

Por outro lado, de acordo com Matloff (2017, p. 25), quanto maior for o número de variáveis do modelo – mantido o mesmo número de dados amostrais –, maior será a variabilidade coletiva dos regressores e, assim, maior será a variância dos coeficientes estimados.

Desta maneira, modelos em modelos mais simples, a redução do viés do mesmo através da adição de um novo regressor compensa o aumento na variabilidade conjunta do modelo, até que este número de regressores atinja um número ótimo, quando a diminuição adicional do viés gerada pela adição de um regressor torna-se tão pequena que não compensa a variabilidade dos coeficientes estimados. Um modelo com variáveis explicativas maior do que este número ótimo estará, portanto, sobreajustado.

Ou seja, existe um *tradeoff* entre viés e variância: para qualquer estimador estatístico (MATLOFF, 2017, p. 25), não se pode reduzir o seu viés sem aumentar a sua variância e vice-versa. Temos que conviver sempre com algum viés e temos que aceitar alguma variância.

Matematicamente, isto decorre do desenvolvimento da expressão do Erro Médio Quadrático (MSE) (MATLOFF, 2017, p. 49):

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta]^2$$

Temos que:

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[\mathbb{E}[\hat{\theta}] - \theta)^2] + \mathbb{E}[2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)]$$

como:

- o termo $\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$ é igual à variância do estimador;
- o termo $\mathbb{E}[\mathbb{E}[\hat{\theta}] - \theta)^2]$ é o quadrado do viés do estimador;
- e, finalmente, o termo $\mathbb{E}[2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)]$ é nulo, haja vista que $\mathbb{E}[\hat{\theta} - E(\hat{\theta})] = 0$.

Portanto temos, matematicamente, que:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{B}^2(\hat{\theta})$$

2.4 A avaliação pela média

2.4.1 Regressão Linear

2.4.1.1 Definição precisa

Sejam Y e X duas variáveis e $m_{Y;X}(t)$ uma função tal que:

$$m_{Y;X}(t) = \mathbb{E}(Y|X = t)$$

Chamamos $m_{Y;X}$ de **função de regressão de** Y **dado** X (MATLOFF, 2009, p. 386, grifo do autor). Em geral, $m_{Y;X}(t)$ é a **média** de Y para todas as unidades da população para as quais $X = t$ (MATLOFF, 2009, p. 386, grifo nosso).

> The word “regression” is an allusion to the famous comment of Sir Francis Galton in the late 1800s regarding “regression toward the mean.” This referred to the fact that tall parents tend to have children who are less tall closer to the mean – with a similar statement for short parents. The predictor variable here might be, say, the father’s height F, with the response variable being, say, the son’s height S. Galton was saying that $\mathbb{E}(S|F) < F$.

Segundo Matloff (2009, p. 386, grifo do autor), ainda, a função $m_{Y;X}(t)$ é uma função da **população**, ou seja, apenas **estimamos** uma equação de regressão ($\hat{m}_{Y;X}(t)$) à partir de uma amostra da população.

> The function $m_{Y;X}(t)$ is a population entity, so we must estimate it from our sample data. To do this, we have a choice of either assuming that $m_{Y;X}(t)$ takes on some parametric form, or making no such assumption. If we opt for a parametric approach, the most common model is linear […] (MATLOFF, 2009, p. 389).

Segundo Matloff (2009, pp. 394–397), as proposições acima sobre a função $m_{Y;X}$ pode ser generalizada para outras quantidades de regressores em X e seus termos de interação, tal que:

$$m_{Y;X}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 + \beta_4 t_1^2$$

Notando que o termo **regressão linear** não necessariamente significa que o gráfico da função de regressão seja uma linha reta ou um plano, mas que se refere a função de regressão ser linear em relação aos seus parâmetros (β_i).

2.4.2 Estimação em modelos de regressão paramétricos

Segundo Matloff (2009, p. 389), é possível demonstrar que o mínimo valor da quantidade[1] $\mathbb{E}[(Y - g(X))^2]$ é obtido, entre todas as outras funções, para $g(X) = m_{Y;X}(X)$. Porém, "se pretendemos minimizar o erro médio absoluto de predição, $\mathbb{E}(|Y - g(X)|)$, a melhor função seria a mediana $g(Y) = mediana(Y|X)$." (MATLOFF, 2009, p. 389).

Matloff (2009) aqui está se referindo à um outro tipo de regressão, chamada de regressão quantílica, mais especificamente, à regressão à mediana, ou seja, ao quantil de 50%.

2.4.3 A equação de regressão linear

Como veremos nesta seção, a equação de regressão linear $\mu(t)$ é uma *função da população*, que geralmente não nos está acessível, pois temos acesso apenas a uma parte (amostra) desta população em estudo. O que fazemos, então é *estimar* uma equação de regressão $\hat{\mu}(t)$ para que possamos prever os valores reais da variável em análise.

Tem que se levar em conta que a equação de regressão linear não é uma equação determinística, mas probabilística. No dia-a-dia da prática de engenharia de avaliações, assim como em outras áreas, no entanto, a equação de regressão é usualmente escrita simplificadamente, sem o termo de erro ϵ, ou seja, a equação de regressão é escrita como uma equação determinística, da forma $Y = \alpha + X\beta$ ou, em termos de variáveis de avaliação de imóveis, $VU = \alpha + A\beta$, onde VU representa o valor unitário dos imóveis e A a sua área.

No entanto, estas equações são uma simplificação da equação de regressão. Na verdade, a equação de regressão $\mu(t)$ é uma função da *população* e pode ser escrita formalmente como abaixo (MATLOFF, 2017, p. 66):

$$\mu(t) = \beta_0 + \beta_1 t_1 + ... + \beta_p t_p$$

Como o termo de erro da equação, ou seja, o erro que cometeríamos ao prever Y se nós efetivamente conhecessemos a equação de regressão da população, é (MATLOFF, 2017, p. 67):

$$\epsilon = Y - \mu(t)$$

Então podemos escrever a equação de regressão de outra maneira, como abaixo (MATLOFF, 2017, p. 67):

$$Y = \beta_0 + \beta_1 t_1 + ... + \beta_p t_p + \epsilon$$

Onde ϵ é uma variável aleatória supostamente tal que $\mathbb{E}(\epsilon) = 0$ e $\text{Var}(\epsilon) = \sigma^2$, ou simplesmente $\epsilon \sim N(0, \sigma^2)$.

[1] Erro médio quadrático de predição

Num modelo onde não há a adoção de qualquer transformação para a variável dependente, verificada a hipótese da normalidade, esta equação de regressão é também a equação de estimação da variável Y, ou seja, para uma equação de regressão sem transformação de variáveis, pode-se escrever:

$$\mathbb{E}[Y|X] = \mathbb{E}[\alpha + X\beta] + \mathbb{E}[\epsilon] = \alpha + X\beta$$

Haja vista que o valor esperado para o termo de erro ϵ é zero.

No entanto, quando a variável dependente Y é transformada, este termo de erro desprezado na equação de regressão acima é de suma importância para o computo do valor esperado da variável original, como veremos neste artigo, pois ele determina a equação de estimação da variável original. Por exemplo, no caso que aqui nos interessa, que é o da transformação logarítmica da variável dependente, temos:

$$\begin{gathered}
\ln(Y) = \alpha + X\beta + \epsilon \Leftrightarrow \\
Y = \exp(\alpha + X\beta)\exp(\epsilon) \Leftrightarrow \\
\mathbb{E}[Y|X] = \mathbb{E}[\exp(\alpha + X\beta)]\mathbb{E}[\exp(\epsilon)|X] \Leftrightarrow \\
\mathbb{E}[Y|X] = \exp(\alpha + X\beta)\mathbb{E}[\exp(\epsilon)|X]
\end{gathered}$$

O fundamental a se perceber aqui é que, quando há transformação da variável dependente, para voltarmos à variável original, temos que levar em conta o termo de erro, haja vista que uma propriedade do valor esperado é a de que $\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$, como veremos a seguir. Mais precisamente, para funções convexas, pela desigualdade de Jensen, $f(\mathbb{E}[X]) \leq \mathbb{E}[(f(x)]$. Isto implica que o valor esperado da exponencial do termo de erro que precisamos estimar é maior do que a exponencial do valor esperado do erro, ou seja, $\mathbb{E}[\exp(\epsilon)|X] \geq \exp(\mathbb{E}[\epsilon|X]) = 1$.

Consideramos equivocado, portanto, a afirmação abaixo:

> Ao adotar o valor proposto pela equação de regressão linear, o perito, como acima referido, estará informando o Juiz a quem se dirige que o valor pelo qual avaliou o bem é dado por Yc; adicionalmente, há um componente aleatório, de caráter aditivo ou subtrativo, com determinado desvio-padrão, cujo resultado, porém, tanto excederá o valor Yc, como lhe ficará aquém, com a mesma probabilidade de 50%. Decorre isto do princípio dos "eventos comparáveis"; o perito avaliou, na realidade, o logaritmo neperiano de Yc; os resíduos aleatórios são medidos como Ln(Y)-Ln(Yc), onde Ln(Yc) é o valor central de uma distribuição normal e, portanto, sua mediana. **Conseqüentemente, seu homólogo, antilogaritmo, necessariamente, marcará, na distribuição lognormal de Y/Yc, também a mediana. No caso, como o antilogaritmo de 0 é a unidade, a mediana de Y/Yc terá valor 1,0** (GIANNAKOS; LEÃO, 1996, p. 13, grifo nosso).

Ou seja, consideramos equivocado a consideração de que os erros aleatórios e com distribuição normal na equação de regressão logaritmizada podem ser diretamente retransformados por um fator de erro multiplicativo igual a 1, já que isto viola a desigualdade de Jensen.

Desta maneira, no nosso ponto de vista seria errôneo afirmar que, ao utilizar a avaliação pela média, se "introduz, na regressão linear, como fator de decisão, as características da função dita 'originária', não-linear, transformada em logarítmica precisamente para alcançar linearidade;

viola os pressupostos do método de mínimos quadrados, fundamento da regressão, ou, alternativamente, equivale a adulterar a amostra original, multiplicando, no caso presente, todos os seus valores…" (1996, p. 5).

GIANNAKOS; LEÃO (1996) faz uma crítica à avaliação pela moda da distribuição lognormal, crítica esta muito bem elaborada e da qual não discordamos no todo. Porém, o mesmo trabalho faz também uma defesa a nosso ver injustificada da utilização da estimativa pela mediana desta distribuição. Concordamos com GIANNAKOS; LEÃO (1996) que a moda não é o valor mais provável, contudo, a nosso ver, pelo motivo que **o valor mais provável é o Valor Esperado** da variável, ou seja, o seu valor médio, como veremos.

Mesmo em GIANNAKOS; LEÃO (1996), encontra-se que "a média aritmética é o 'valor esperado' da variável".

Na verdade, o que poderia ser afirmado é que, ao avaliar pela média, o avaliador estaria se aproximando melhor da equação de regressão do que ao avaliar pela moda ou pela mediana, haja vista que faz parte da equação de regressão o termo de erro multiplicativo, de valor sabidamente maior do que 1 (pela desigualdade de Jensen), a que se refere Giannakos e Leão (1996).

2.4.4 O problema da retransformação das variáveis

Segundo (SHEN; ZHU, 2008, p. 552), modelos lineares lognormais tem muitas aplicações e muitas vezes é de interesse prever a variável resposta ou estimar a média da variável resposta na escala original para um novo conjunto de covariantes.

Segundo Shen e Zhu(2008, p. 552), se $Z = (Z_1, \cdots, Z_n)^T$ é o vetor variável resposta de distribuição lognormal e $x_i = (1, x_{i1}, \cdots, x_{ip})^T$ é o vetor dos covariantes para a observação i, um modelo linear lognormal assume a seguinte forma:

$$Y = \ln(Z) = X\beta + \epsilon$$

onde $X = (x_1, \cdots, x_n)^T$, $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$, e $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$ com $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.(*identically independently distributed*) (SHEN; ZHU, 2008, pp. 552–553).

Em muitos casos, para um novo conjunto de covariantes x_0, pode-se estar interessado em prever a variável resposta em sua escala original:

$$Z_0 = e^{x_o^T\beta + \epsilon_0}$$

ou estimar a média condicional da variável resposta:

$$\mu(x_0) = \mathbb{E}[Z_0|x_o] = e^{x_o^T\beta + \frac{1}{2}\sigma^2}$$

De acordo com Shen e Zhu(2008, p. 553), se β e σ^2 são ambos conhecidos, então é fácil demonstrar que o melhor estimador de Z_0 é de fato $\mu(x_0)$. Contudo, na prática, ambos β e σ^2 são desconhecidos e precisam ser estimados para a obtenção de $\mu(x_0)$.

Segundo Shen e Zhu (2008, p. 552), existem na literatura diversos estimadores baseados em diversos métodos inferenciais, como **ML** (*Maximum Likelihood Estimator*), **REML** (*Restricted*

ML Estimator), **UMVU** (*Uniformly Minimum Variance Unbiased Estimator*), além de um estimador **REML** com viés corrigido.

Na prática, estes estimadores pertencem a uma classe de estimadores definida na expressão abaixo:

$$\left\{\hat{\mu_c}(x_0) : \hat{\mu_c}(x_0) = \exp(x_0^T\hat{\beta} + cRSS/2), c = \frac{1}{n-a}, a < n\right\}$$

Shen e Zhu(2008) então propõem dois novos estimadores baseados na minimização do erro médio quadrático assintótico (MM) e do viés assintótico (MB).

De maneira que a diferença entre os estimadores supra-citados pode ser resumida ao parâmetro a:

$$a_{ML} = 0$$

$$a_{REML} = p + 1$$

$$a_{MM} = p - 1 - 3nv_0 - 3RSS/(2m)$$

$$a_{MB} = p + 1 - nv_0 - RSS/(2m)$$

2.4.4.1 Estimadores não-paramétricos

De acordo com Duan (1983, p. 606), o Valor Esperado $\mathbb{E}$ de uma variável resposta Y que tenha sido transformada em valores η durante a regressão linear por uma função $g(Y)$ **não-linear** não é igual ao valor da simples retransformação da variável transforma pela sua função inversa $h(\eta) = g^{-1}(Y)$. Em outros termos (DUAN, 1983, p. 606):

$$\mathbb{E}[Y_0] = \mathbb{E}[h(x_0\beta + \epsilon)] \neq h(x_o\beta)$$

Reparar que o termo de erro faz parte da composição do valor esperado da variável de regressão. Em uma regressão linear clássica, sem transformação, $\mathbb{E}[\epsilon] = 0$, então $\mathbb{E}[Y_0] = \mathbb{E}[x_0\beta]$.

Numa regressão linear logaritmizada, ou seja, uma regressão linear com o logaritmo da variável dependente ($h(\eta) = g^{-1}(\eta) = \exp(\eta)$), para efetuar apropriadamente a retransformação das estimativas de volta a sua escala original, precisa-se ter em conta a desigualdade mencionada na seção 2.1.

Segundo (MANNING; MULLAHY, 1999), quando ajustamos o logaritmo natural de uma variável Y contra outra variável X através da seguinte equação de regressão:

$$ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

Se o erro ϵ é normalmente distribuído, com média zero e desvio padrão σ^2, ou seja, se $\epsilon \sim N(0, \sigma^2)$, então (DUAN, 1983, p. 606; MANNING; MULLAHY, 1999, p. 6):

$$\mathbb{E}[Y|X] = e^{\beta_0+\beta_1 X} \cdot \mathbb{E}[e^{\epsilon}] \neq e^{\beta_0+\beta_1 X}$$

Embora o valor esperado dos resíduos ϵ seja igual a zero, ele está submetido a uma transformação não linear, de maneira que não podemos afirmar que $\mathbb{E}[e^{\epsilon}] = 1$ (como vimos na seção 2.2, $\mathbb{E}[\exp(x)] > \exp[\mathbb{E}(x)]$). Desta maneira, o estimador abaixo, chamado em (SHEN; ZHU, 2008, p. 554) de *naive back-transform estimator*, ou simplesmente **BT** não é consistente e é enviesado, tendo viés multiplicativo de valor assintótico igual a $e^{-\sigma^2/2}$:

$$\mathbb{E}[Y|X] = e^{\beta_0+\beta_1 X}$$

Segundo (SHEN; ZHU, 2008, p. 554), ainda, o valor de $e^{-\sigma^2/2}$ é sempre menor do que 1(SHEN; ZHU, 2008, p. 554).

> As a result, the BT estimator underestimates $\mu(x_0)$, and the bias is large when σ^2 is large. In our study, it appears that the BT estimator performs much worse than the other estimators[...]Actually, the BT estimator is more suitable for estimating the median of Z0, which is $\exp(x_0^T\beta)$ in this case.

Porém se o termo de erro ϵ é normalmente distribuído $N(0, \sigma^2)$, então um estimador não-enviesado para o valor esperado $\mathbb{E}[Y]$, de acordo com DUAN (1983), assume a forma vista na equação abaixo(DUAN, 1983, p. 606; MANNING; MULLAHY, 1999, p. 2 e 6):

$$\mathbb{E}[Y] = e^{\beta_0+\beta_1 X} \cdot e^{\frac{1}{2}\sigma^2}$$

Cabe salientar, segundo (MANNING; MULLAHY, 1999, p. 6), que se o termo de erro não for i.i.d (independente e identicamente distribuído), mas for homoscedástico, então:

$$\mathbb{E}[Y|X] = s \times e^{X_0\beta}$$

onde $s = \mathbb{E}[e^{\epsilon}]$.

De qualquer maneira, o valor esperado de Y é proporcional à exponencial da previsão na escala log.

DUAN (1983) apresenta então um estimador não-paramétrico (*smearing estimate*), independente da função de transformação $h(\eta)$ e da distribuição dos erros $F(\epsilon)$, tal que:

$$\hat{\mathbb{E}}[Y_0] = \int h(x_0\hat{\beta} + \epsilon)d\hat{F}_n(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} h(x_0\hat{\beta} + \hat{\epsilon}_i)$$

2.4.4.2 Modelos Heteroscedásticos

Modelos heteroscedásticos não são raros, especialmente no caso de variáveis envolvendo valores em moeda, sendo muito comum em modelos econométricos. Em sua essência, são heteroscedásticos aqueles modelos lineares cujo termo de erro não pode ser considerado totalmente independente, ou seja, existe alguma função (linear ou não), tal que $\mathbb{E}[e^{\epsilon}] = f(x)$, de modo que:

$$\ln(\mathbb{E}[Y|X]) = X\beta + \ln(f(x))$$

É desnecessário dizer que, para estes modelos o estimador para a média é diferente de $\mathbb{E}[Y] = e^{\beta_0+\beta_1 X} \cdot e^{\frac{1}{2}\sigma^2}$, haja vista que σ^2 não é mais um escalar, mas uma função.

Existem diversas maneiras de se contornar este problema. Por exemplo, através da eliminação do viés através da utilização de uma função que modele a variância $\sigma^2(X)$, ou através do estimador sanduíche[2].

Cabe ainda salientar que, para os modelos heteroscedásticos, não apenas os erros estão comprometidos, mas também os intervalos de confiança.

2.5 A avaliação pela mediana

A avaliação pela mediana através de modelos de regressão linear clássicos, como prentendemos demonstrar neste artigo, não é uma boa opção. Muito melhor, para este caso, seria fazer uso de um método consagrado e automatizado praticamente da mesma maneira que a regressão linear clássica à média: a regressão à mediana, ou, regressão quantílica à mediana.

2.5.1 Regressão quantílica

Segundo CRISTINA DAVINO; VISTOCCO (2014), enquanto a média é a medida que minimiza o erro médio quadrático:

$$\mu = \underset{c}{argmin} E(Y-c)^2$$

A mediana é o valor que minimiza o erro médio absoluto:

$$Me = \underset{c}{argmin} E|Y-c|$$

Por estas simples equações, percebe-se que a média tem a propriedade de ser mais suscetível à presença de *outliers*, haja vista que os erros maiores serão mais impactantes no modelo do que os erros menores, por estarmos minimizando erro médio quadrático. Já na regressão à mediana, minimiza-se o erro médio absoluto, de maneira que a presença de um eventual *outlier* tem pouco impacto sobre a equação de regressão (à mediana).

2.5.1.1 Exemplo com duas variáveis

O gráfico da figura 1 foi reproduzido de Koenker e Hallock (2001, p. 147). Pode-se perceber que a reta de regressão linear é bastante afetada pela presença dos dois pontos com maior renda, o que faz com que a equação de regressão linear superestime os valores para os extratos de mais baixa renda, enquanto a reta de regressão à mediana apresenta maior equilíbrio, não sendo tão afetada pela presença destes pontos.

[2] ver link

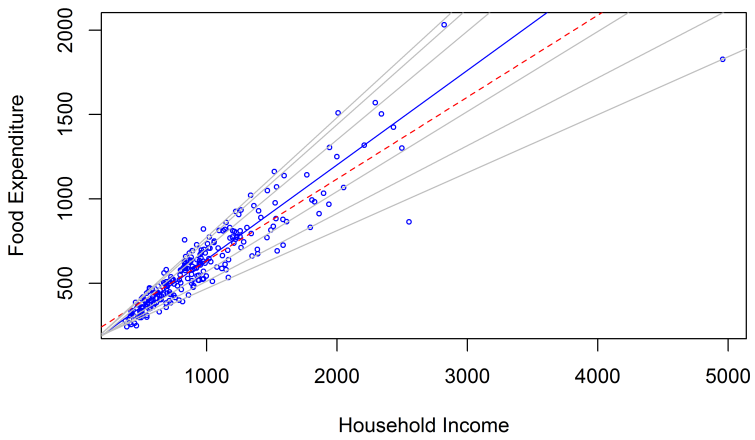

Figura 1: Comparação de modelos de regressão para média (em vermelho) e para a mediana (em azul).

Na figura 1 as retas cinzas são as regressões para os quantis de 5%, 10%, 25%, 75%, 90% e 95%.

Também é possível a transformação de variáveis nos modelos de regressão quantílica, assim como fazemos nos modelos de regressão à média. O modelo de regressão linear para a média apresentado é heteroscedástico, como o próprio gráfico da figura 1 demonstra. Nestes casos, é usual proceder com a transformação dos dados. Desta maneira, foi elaborada a figura 2, reproduzida da vinheta (2018a, p. 11) do pacote quantreg (2018b) do software estatístico R (2018), que nos mostra o modelo das variáveis em escala $\ln$.

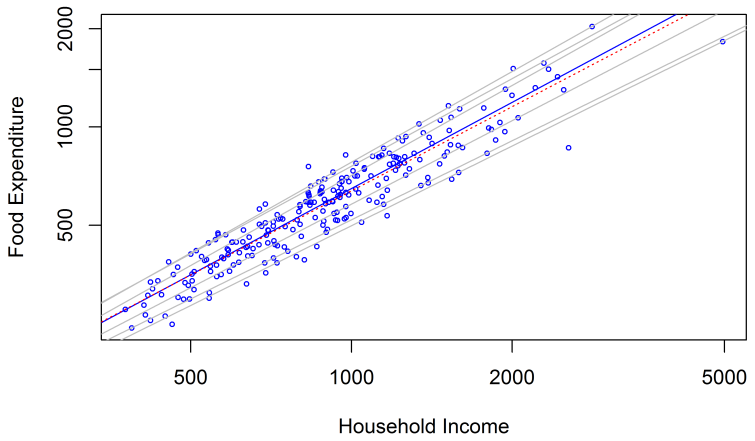

Figura 2: Comparação de modelos de regressão para média (em vermelho) e para a mediana (em azul) em escala transformada (ln).

Como esperado, a heteroscedasticidade do modelo praticamente desapareceu com a transformação dos dados.

2.6 A avaliação pela moda

Os defensores da avaliação pela moda da variável lognormal normalmente argumentam que a escolha desta é pelo motivo da moda ser o valor mais provável da variável lognormal. Ledo engano, poisinstal o valor mais provável é o valor esperado da mesma.

2.6.1 Regressão à moda

Segundo Chen *et al.* (2016, p. 1), ao contrário da regressão convencional que busca estimar a média condicional de Y dado $X = x$, a regressão à moda busca estimar a moda condicional de Y dado $X = x$.

Segundo Chen *et al.*, a regressão à moda é utilizada para buscar estruturas da distribuição de probabilidade dos dados que são perdidas quando se utiliza a clássica regressão à média.

A regressão modal clássica, segundo Chen *et al.* (2016, p. 4) pode ser resumida por um modelo de uma moda condicional tal que:

$$\text{Mode}(Y|X = x) = \beta_0 + \beta^T x$$

onde $Mode(Y|X = x)$ é a moda **global** de Y dado $X = x$.

Já o procedimento não-paramétrico proposto em Chen *et al.* (2016, p. 4), fora do escopo deste artigo, permite múltiplas modas (locais) da variável resposta.

Segundo Oelker *et al.*(2015, p. 2):

- a moda é de longe a característica mais proeminente de uma função densidade de probabilidade;
- a moda é extremamente robusta à *outliers*;
- the mode provides a location measure that is easily communicated to practitioners such that mode regression will be of high interest in applied regression situations, there may be situations where the dependence of the mode on covariates may be quite different from the dependence of the median and/or the mean,
- a regressão à moda permite lidar com variáveis dependentes truncadas.

Em relação à moda como estimativa de medida central, consideramos que esta se trata mais de uma curiosidade do que uma estimativa de fato: o que significa a moda de uma população de apartamentos em uma determinada cidade? A moda encontraria-se, provavelmente, nos valores dos apartamentos de 2 e 3 quartos, com uma ou duas vagas de garagem. Mas qual a utilidade disto quando o que se pretende avaliar, por exemplo, é o valor de um apartamento de 4 ou 5 quartos e 4 ou 5 vagas de garagem, ou ainda de se avaliar um apartamento com um quarto e sem vaga de garagem? Assim como os apartamento citados estão “fora de moda”, também estarão os seu valores. Contudo, estes estarão em consonância com a média ou com a mediana do mercado, dados as suas características, a depender da configuração deste.

Em outras palavras, um modelo de regressão linear é uma média *condicional* da variável resposta (ver Regressão Linear). Ou seja, pretende-se saber o valor médio de um imóvel *dado que* ele possui as seguintes características…E estas características podem estar *na moda* ou fora dela.

2.7 Validação Cruzada

Validação Cruzada ou *cross-validation* é uma técnica estatística que pode ser utilizada de diversas maneiras e consistem em dividir um conjunto de dados em duas partições distintas, chamados de partição de treino (*training set*) e partição de teste(*test set*), utilizadas para o ajuste do modelo e para a previsão da variável dependente, respectivamente. Os dados previstos na partição de teste são então comparados aos valores observados.

Neste artigo efetuaremos a validação-cruzada utilizando o procedimento chamado de *delete-one procedure*, em que se retira apenas um dado do conjunto de dados, ajusta-se um modelo e então utiliza-se este modelo para prever o valor da variável dependente para o dado retirado (SHEN; ZHU, 2008, p. 564).

Para cada observação então calcula-se o seu erro quadrático ($(Y_i - \hat{Y}_i)^2$), utilizado para o cálculo da estatística RMSPE (erro de previsão médio quadrático, ou *root mean squared prediction error*), conforme expressão a seguir (SHEN; ZHU, 2008, p. 564):

$$RMSPE = (\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2)^{1/2}$$

2.8 Estudo de Caso

Com o fim de averiguar qual estimador melhor se adequa ao procedimento de retransformação de variáveis, aplicar-se-á um comparativo entre os estimadores média, moda e mediana, através do uso da estatística RMSPE.

2.8.1 Dados

Neste estudo comparamos a precisão de diversos tipos de modelos estatísticos (regressão linear, regressão não-linear e modelo linear generalizado) sobre os dados disponíveis em Hochheim (2015, pp. 21–22).

Os coeficientes do modelo utilizado em HOCHHEIM (2015), assim como suas estatísticas básicas podem ser visualizados na tabela 1.

2.8.2 Cálculo do RMSPE

2.8.2.1 Regressão linear ordinária

Para o cálculo do RMSPE foi utilizado como referência o modelo proposto por Hochheim (2015, p. 29), ou seja, foram utilizadas as mesmas transformações de variáveis utilizadas no modelo proposto. Os valores dos β_i são calculados a cada passo.

Os valores encontrados para o erro de predição médio quadrático para cada estimador foram: **R$203.939,11** para a média, **R$204.006,84** para a mediana e **R$205.537,36** para a moda.

Como esperado, o RMSPE foi menor para a média, e maior para a moda. O que comprova a teoria, já que o *naive estimator* é enviesado com viés conhecido de $-\sigma^2/2$, logo a moda possui viés de $-1,5\sigma^2$.

Os valores ajustados com os estimadores da moda, média e mediana podem ser vistos na tabela em anexo.[3]

2.8.3 Cálculo do erro médio absoluto

Assim como a regressão linear é uma minimização do erro médio quadrático, a regressão à mediana leva a minimização do erro médio absoluto.

Para verificarmos isto, com um modelo de regressão à mediana, calcularemos o RMAPE (*root mean absolute prediction error*) e o RMSPE (*root mean squared prediction error*) para as estimativas obtidas com este modelo.

2.8.3.1 Regressão quantílica à mediana

O modelo de regressão quantílica com quantil $\tau = 0.5$, ou seja, o modelo de regressão à mediana, para os mesmos dados supra-mencionados está resumido na tabela 1.

De posse do modelo para a regressão quantílica, fazemos a previsão para a mediana da variável `valor` na escala original da mesma maneira que a fizemos para a regressão linear, ou seja, apenas aplicamos a função inversa à variável transformada ($valor = \exp(log(\hat{Y}))$). Os valores podem ser vistos na tabela em anexo.[4]

O valor de RMAPE para a regressão à mediana é de R$131.842,83, enquanto o valor do RMSPE é de R$208.063,86.

É fácil demonstrar que estes valores são bem diferentes dos obtidos pelas estimativas da regressão linear clássica (regressão à média). Para a estimativa pela mediana na regressão linear, o erro médio absoluto seria de R$ 133.234,00, bem superior ao erro médio absoluto obtido pela regressão à mediana.

Já para o RMSPE, o valor obtido na regressão linear é menor, qualquer que seja a estimativa, pela moda, média ou mediana.

Ou seja, o modelo de regressão linear minimizou o RMSPE e o modelo de regressão quantílica minimizou o RMAPE, conforme esperado.

3 CONCLUSÕES E RECOMENDAÇÕES

Entendemos que a norma brasileira (ABNT, 2011) deveria tratar este assunto de maneira clara, especificando qual estimador deveria ser utilizado para a formação de valores, ou ainda, qual seria o estimador dependendo do método utilizado pelo avaliador, se a regressão linear clássica (*i.e.*, à média), a regressão à mediana ou a regressão modal, haja visto que os três métodos são cientificamente válidos.

Como vimos na seção 2.4.1, o método clássico de regressão linear é uma minimização do erro médio quadrático de predição e a função de regressão $\hat{m}_{Y;X}$ é uma equação para a *média* da população Y dado X, seja ela uma função de outra variável ou não. Considerando que são

[3] https://github.com/lfpdroubi/moda-media-mediana/blob/master/tabela.xls

[4] https://github.com/lfpdroubi/moda-media-mediana/blob/master/tabela.xls

Tabela 1: Comparação entre os coeficientes de regressão linear e regressão quantílica

	Dependent variable:	
	log(valor)	
	OLS	*quantile regression*
	(1)	(2)
area_total	0.001 (0.001, 0.002)	0.002 (0.0003, 0.003)
	t = 5.113	t = 2.300
	p = 0.00001***	p = 0.027**
quartos	0.164 (0.094, 0.233)	0.162 (0.078, 0.245)
	t = 4.626	t = 3.788
	p = 0.00004***	p = 0.0005***
suites	0.061 (−0.005, 0.127)	0.080 (−0.012, 0.171)
	t = 1.810	t = 1.712
	p = 0.078*	p = 0.095*
garagens	0.209 (0.143, 0.274)	0.152 (0.034, 0.271)
	t = 6.247	t = 2.520
	p = 0.00000***	p = 0.016**
log(dist_b_mar)	−0.141 (−0.194, −0.087)	−0.146 (−0.244, −0.047)
	t = −5.174	t = −2.904
	p = 0.00001***	p = 0.006***
I(padrao^-1)	−0.563 (−0.769, −0.357)	−0.459 (−0.751, −0.166)
	t = −5.360	t = −3.070
	p = 0.00001***	p = 0.004***
Constant	13.564 (13.112, 14.016)	13.574 (12.850, 14.298)
	t = 58.847	t = 36.732
	p = 0.000***	p = 0.000***
Observations	48	50
R^2	0.956	
Adjusted R^2	0.950	
Akaike Inf. Crit.	−46.813	−38.299
Residual Std. Error	0.136 (df = 41)	
F Statistic	148.921*** (df = 6; 41)	
Note:		*p<0.1; **p<0.05; ***p<0.01

satisfeitas as hipóteses da regressão linear clássica, o melhor estimador para o valor será o da avaliação pela média, haja vista que, por definição, a regressão linear é uma função para a média.

Ora, claro está, de acordo com todos os trabalhos citados, inclusive GIANNAKOS; LEÃO (1996), que o valor esperado da variável é a média. A regressão linear com o método dos mínimos quadrados é uma regressão para a média. Isto posto, como então avaliar o valor da variável original? Porque na área de avaliações não temos interesse na previsão da variável $W = \ln(Y)$, mas sim na variável Y, ou seja, estamos interessados nos valores da variável original, não nos valores da variável transformada. Está claro que deve-se proceder a retransformação da variável W na variável original, mas para isso é preciso utilizar o estimador correto.

Esperamos ter demonstrado com este artigo que a retransformação adequada da variável de regressão linear é a estimativa pela média, que é o seu Valor Esperado, que pode ser calculado através de qualquer dos estimadores supra-citados, sem com isso adulterar a equação de regressão, muito pelo contrário, reafirmando-a.

Não pretendemos, com isto, impor quer seja a média ou a mediana a melhor estimativa. Em vários campos, a mediana tem sido adotada como melhor estimativa, por sua propriedade de estar menos vulnerável a presença de *outliers*, como ocorre com a média.

No entanto, se pretende-se efetuar uma avaliação pela mediana, entendemos que a melhor opção seria a utilização da regressão quantílica, para o quantil de 50% (obviamente), e não a utilização da retransformação inadequada da equação de regressão linear, que destina-se a estimar a média.

REFERÊNCIAS

ABNT. **NBR 14653-2: Avaliação de bens – parte 2: Imóveis urbanos**. Rio de Janeiro: Associação Brasileira de Normas Técnicas, 2011.

BENNETT, H. Lecture note 4: Expectations (moments)., 2006. MIT. Disponível em: <https://tinyurl.com/yayljdpq>..

CHEN, Y.-C.; GENOVESE, C. R.; TIBSHIRANI, R. J.; WASSERMAN, L. Nonparametric modal regression. **Ann. Statist.**, v. 44, n. 2, p. 489–514, 2016. The Institute of Mathematical Statistics. Disponível em: <https://doi.org/10.1214/15-AOS1373>..

CHESTERTON, G. K. **Ortodoxia**. São Paulo: Mundo Cristão, 2008.

CRISTINA DAVINO, M. F.; VISTOCCO, D. **Quantile regression: Theory and applications**. UK: Wiley, 2014.

DUAN, N. Smearing estimate: A nonparametric retransformation method. **Journal of the American Statistical Association**, v. 78, n. 383, p. 605–610, 1983. Taylor & Francis. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478017>..

GIANNAKOS, I. B. D. S.; LEÃO, M. L. Crítica à avaliação pela moda da distribuição log-normal. In: VIII Congresso Brasileiro de Avaliações e Perícias. **Anais...**. p.267–278, 1996. Florianópolis: COBREAP.

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

2015.

KOENKER, R. Quantile regression in R: A vignette., 2018a. Disponível em: <https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf>..

KOENKER, R. **Quantreg: Quantile regression**. 2018b.

KOENKER, R.; HALLOCK, K. F. Quantile regression. **Journal of Economic Perspectives**, v. 15, n. 4, p. 143–156, 2001.

MANNING, W. G.; MULLAHY, J. **Estimating log models: To transform or not to transform?** Working Paper, National Bureau of Economic Research, 1999.

MATLOFF, N. **Statistical regression and classification: From linear models to machine learning**. Boca Raton, Florida: Chapman & Hall, 2017.

MATLOFF, N. S. **From Algorithms to Z-Scores: Probabilistic and statistical modeling in computer science**. Davis, California: Orange Grove Books, 2009.

OELKER, M.-R.; SOBOTKAZ, F.; KLEINZ, N.; KNEIBZ, T. Semiparametric mode regression., 2015. Disponível em: <https://www.uni-goettingen.de/de/oelker_04_2015/512327.html>..

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018.

SHEN, H.; ZHU, Z. Efficient mean estimation in log-normal linear models. **Journal of Statistical Planning and Inference**, v. 138, p. 552–567, 2008. Elsevier. Disponível em: <https://www.unc.edu/~haipeng/publication/emplnM1.pdf>..