Avaliação pela Moda, Média ou Mediana?

Teoria e simulações

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

27/04/2018

1 INTRODUÇÃO

Existe na área da avaliação de imóveis uma discussão frequente e 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.[1]

Pretende-se com este artigo dar a este problema de uma abordagem formal. 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.

> 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*! [matloff2009, 386, grifo do autor]

2 REVISÃO BIBLIOGRÁFICA

> 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).

2.1 Estimadores

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, $\hat{\theta}(x)$.

2.1.1 Propriedades de um estimador

2.1.1.1 Erro

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

2.1.1.2 Erro médio quadrático

$$MSE = E[\hat{\theta}(X) - \theta]$$

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

[1]Neste artigo esta função é representada por *log*.

2.1.1.3 Desvio

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

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

2.1.1.4 Variância

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

2.1.1.5 Viés

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

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

2.1.1.6 Consistência

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

2.2 Melhor estimador linear não-inviesado ou BLUE (Best Linear Unbiased Estimator)

Em estatística, é comum o uso da sigla BLUE para indicar o melhor estimador linear não-enviesado.

2.3 Regressão linear

2.3.1 Definição precisa

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

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

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** da de Y para todas as unidades da população para as quais $X = t$ (MATLOFF, 2009, p. 386, grifo nosso).

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 Estimação em modelos de regressão paramétricos

Segundo Matloff (2009, p. 389), é possível demonstrar que o mínimo valor da quantidade $E[(Y - g(X))^2]$[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, $E(|Y - g(X)|)$, a melhor função seria a mediana $g(Y) = mediana(Y|X)$." (MATLOFF, 2009, p. 389).

2.5 Esperança matemática ou Valor Esperado

Segundo WIKIPEDIA (2018), a "**esperança matemática** 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 $E[X]$, de tal forma que, pela definição dada acima, no caso de uma variável aleatória discreta:

$$E[X] = \sum_{i=1}^{\infty} x_i p(x_i)$$

Já para uma variável aleatória contínua, o valor esperado torna-se:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

2.6 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 log-normal assume a seguinte forma:

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

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

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.

[2] Erro médio quadrático de predição

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

2.6.1 Regressão Linear

De acordo com Duan (1983, p. 606), o Valor Esperado 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):

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

Numa regressão log-linear, 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 **??**.

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):

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

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 $E[e^{\epsilon}] = 1$, como vimos na seção anterior. 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 $e^{-\sigma^2/2}$:

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

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 $E[Y]$, de acordo com DUAN (1983), assume a forma vista na equação abaixo(DUAN, 1983, p. 606; MANNING; MULLAHY, 1999, p. 2 e 6):

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

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

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

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

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

2.6.2 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 $E[e^{\epsilon}] = f(x)$, de modo que:

$$log(E[Y|X]) = X\beta + log(f(x))$$

É desnecessário dizer que, para estes modelos o estimador para a média é diferente de $E[Y] = e^{\beta_0+\beta_1 X} \cdot e^{\frac{1}{2}\sigma^2}$, 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[3].

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

2.7 Modelo linear generalizado (*GLM*)

De acordo com (MANNING; MULLAHY, 1999, pp. 3–4), um modelo linear generalizado com uma função de ligação logarítmica estimam $log(E[Y|X])$ diretamente, de tal maneira que:

$$log(E[Y|X]) = X\beta$$

ou

$$E[Y|X] = e^{X\beta}$$

2.8 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 *root mean squared prediction error*), conforme expressão a seguir (SHEN; ZHU, 2008, p. 564):

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

3 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.

[3] ver link

3.1 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).

3.2 Cálculo do RMSPE

3.2.1 Regressão linear

Os valores ajustados com os estimadores da moda, média e mediana podem ser vistos na tabela abaixo:

	Y	Média	Mediana	Moda
AP_1	1.060.000	1.029.765	1.020.713	1.002.846
AP_2	510.000	628.132	622.610	611.712
AP_3	780.000	855.052	847.535	832.700
AP_4	550.000	736.956	730.478	717.691
AP_5	850.000	1.011.300	1.002.410	984.863
AP_6	300.000	358.594	355.441	349.220
AP_7	750.000	724.106	717.741	705.177
AP_8	650.000	657.475	651.695	640.288
AP_9	620.000	658.389	652.601	641.177
AP_10	740.000	662.002	656.182	644.696
AP_11	770.000	818.933	811.734	797.525
AP_12	680.000	702.573	696.397	684.207
AP_13	850.000	681.544	675.553	663.728
AP_14	420.000	551.781	546.931	537.357
AP_15	547.000	673.810	667.887	656.196
AP_16	1.600.000	1.413.047	1.400.625	1.376.108
AP_17	1.320.000	1.115.664	1.105.857	1.086.499
AP_18	615.000	645.338	639.665	628.468
AP_19	705.000	722.736	716.383	703.843
AP_20	418.000	435.824	431.993	424.431
AP_21	270.000	243.440	241.300	237.077
AP_22	418.000	485.426	481.159	472.736
AP_23	650.000	630.016	624.478	613.547
AP_24	700.000	774.614	767.805	754.365
AP_25	680.000	729.864	723.448	710.784
AP_26	420.000	350.336	347.256	341.178
AP_27	195.000	229.411	227.394	223.414
AP_28	290.000	279.686	277.228	272.375
AP_29	272.000	246.194	244.030	239.758
AP_30	430.000	399.634	396.121	389.187
AP_31	895.000	615.032	609.625	598.954
AP_32	450.000	454.828	450.830	442.938
AP_33	1.950.000	1.474.903	1.461.938	1.436.347
AP_34	2.150.000	2.597.848	2.575.011	2.529.937
AP_35	940.000	969.142	960.623	943.808
AP_36	1.400.000	1.334.839	1.323.105	1.299.945
AP_37	1.090.000	1.002.811	993.996	976.596
AP_38	1.272.000	999.341	990.556	973.217
AP_39	2.800.000	1.921.706	1.904.812	1.871.470
AP_40	1.796.000	2.075.621	2.057.374	2.021.361
AP_41	1.400.000	1.398.114	1.385.824	1.361.566
AP_42	3.000.000	3.306.637	3.277.569	3.220.197
AP_43	1.200.000	1.062.442	1.053.103	1.034.669
AP_44	800.000	646.536	640.853	629.635
AP_45	950.000	668.014	662.142	650.551

	Y	Média	Mediana	Moda
AP_46	2.061.000	2.267.978	2.248.041	2.208.690
AP_47	1.326.000	1.575.944	1.562.090	1.534.746
AP_48	850.000	776.375	769.550	756.079
AP_49	1.650.000	1.509.488	1.496.218	1.470.028
AP_50	650.000	834.750	827.412	812.929

Os valores encontrados para o erro de predição médio quadrático para cada estimador foram: 203.939,11 para a média, 204.006,84 para a mediana e 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 média possui viés de $-1,5\sigma^2$.

3.2.2 Modelo linear generalizado

4 CONCLUSÃO

Como vimos na seção **??**, 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. Considerando que são satisfeitas as hipóteses que

REFERÊNCIAS

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

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>..

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

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

MATLOFF, N. S. **From algorithms to z-scores: Probabilistic and statistical modeling in computer science**. Davis, California: Orange Grove Books, 2009.

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>..

WIKIPEDIA. Valor esperado — Wikipedia, the free encyclopedia., 2018. Disponível em: <https://pt.wikipedia.org/wiki/Valor_esperado>..