**DISTRIBUIÇÃO LOGNORMAL:PROPRIEDADES E APLICAÇÕES NA ENGENHARIA DE AVALIAÇÕES**

**Luiz Fernando Palin Droubi**

**Ministério do Planejamento, Desenvolvimento e Gestão**

MPDG/SPU/SC

Endereço

luiz.droubi@planejamento.gov.br

**Willian Zonato**

**Ministério do Planejamento, Desenvolvimento e Gestão**

MPDG/SPU/SC

Endereço

willian.zonato@planejamento.gov.br

**Norberto Hochheim**

Universidade Federal de Santa Catarina

Endereço

hochheim@gmail.com

**Resumo:**

Pretende-se com este artigo detalhar o motivo pelo qual a transformação de variável dependente pela função logaritmo é frequentemente adequada na área de avaliação de imóveis. Um procedimento muito comum nesta área é a adoção de transformações para a obtenção de um "melhor" modelo de regressão. A mais usual e preferida de muitos avaliadores é a função logaritmo, especialmente para a variável dependente. Muitas vezes esta transformação é adequada e percebe-se uma notória melhora no ajuste do modelo. Outras vezes, esta transformação pode não ser adequada. Apesar do modelo aparentar-se melhor ajustado, problemas podem ocorrer quanto às verificações das hipóteses clássicas da regressão, as quais nem sempre os avaliadores estão tão atentos quanto estão com as verificações dos intervalos de confiança e níveis de significância. No entanto, o avaliador que assim procede estará verificando intervalos de confiança e níveis de significâncias incorretos, haja vista que a hipótese da heteroscedasticidade implica na incorreção destas inferências. Entendemos que a melhor maneira para apresentar aos avaliadores a importância de criteriosas escolhas de transformações seja através da análise do histograma da variável original e transformada. Normalmente, uma boa escolha de transformação leva a uma distribuição aproximadamente normal. Quando a variável dependente apresenta distribuição lognormal, esta transformação é a transformação logaritmica. Desta maneira, demonstramos as características básicas desta distribuição, sua formulação, características além do seu relacionamento com a distribuição normal. Por fim, demonstramos as implicações da adoção da transformação da variável dependente e abordamos o problema da retransformação da variável dependente à sua escala original.​

**Palavras-chave:** distribuição lognormal, regressão-linear, heterocedasticidade, Box-Cox

**Abstract**

The purpose of this article is to detail the reason why the transformation of the dependent variable by the logarithm function is often adequate in real estate appraisals. A very common procedure in this field is the adoption of transformations to obtain a better regression model. The most usual and preferred of many transformations is the logarithm function. Often this transformation is adequate and there is a noticeable improvement in the fit of the model. At other times, this transformation may not be adequate: although the model appears to be better adjusted, problems with the verifications of the classical linear regression hypotheses may appear, which are often neglected by the appraiser. However, if he procedes without verifying heteroscedasticity, he might be verifying confidence intervals and incorrect significance levels, since this hypothesis implies the incorrectness of these inferences. We understand that the best way to present to the appraisers the importance of careful choices of transformations is by analyzing the histogram of the original and transformed variable. Typically, a good choice of transformation leads to a transformed variable distributed closely to a normal distribution. When the dependent variable presents lognormal distribution, this transformation is the logarithmic transformation. So we demonstrate the basic characteristics of this distribution, its formulation, characteristics and its relationship with the normal distribution. Finally, we demonstrate other ways to deal with heteroscedastic data, the implications of adopting the transformation of the dependent variable and address the problem of the retransformation of the dependent variable to its original scale.

**Keywords:** Lognormal distribution, linear-regression, heteroscedasticity, Box-Cox

.

1. **INTRODUÇÃO**

A transformação de variáveis é um procedimento comum na Engenharia de Avaliações. No entanto, a transformação dos dados por vezes é realizada sem uma análise profunda do comportamento das variáveis. A *Food and Drug Administration* (FDA), órgão federal dos EUA que atua no controle da comercialização de alimentos e medicamentos no país, recomenda:

“A transformação desnecessária de dados deve ser evitada. Caso tenha sido realizada transformação de dados, uma justificativa para a escolha da transformação junto com a interpretação das estimativas dos efeitos do tratamento com base nos dados transformados deve ser fornecida.” (FDA, 1988 apud KEENE (1985))

No entanto, a transformação logarítmica é especial, por uma série de aspectos, como pode ser visto em KEENE (1985). A distribuição lognormal apresenta diversas aplicações práticas. É comum, na área de avaliação de imóveis, mas não apenas, nos depararmos com dados que seguem esta distribuição. Neste artigo pretendemos demonstrar as principais características da distribuição lognormal, sua relação com a distribuição normal de Gauss, assim como debatemos a melhor maneira de se lidar com dados lognormais.

**2. REVISÃO BIBLIOGRÁFICA**

**2.1 Formulação**

A formulação da distribuição lognormal para os parâmetros e pode ser vista abaixo (FARIAS)

 (1)

**2.2 Propriedades**

**2.2.1 Valor Esperado e Variância**

O valor Esperado de uma variável aleatória com distribuição lognormal é (FARIAS):

**  (2)

E sua variância é:

**  (3)

**2.2.2 Medidas de Tendência Central**

A figura 1 mostra a posição das medidas de tendência central (moda, média e mediana) para um variável aleatória de distribuição log-normal.

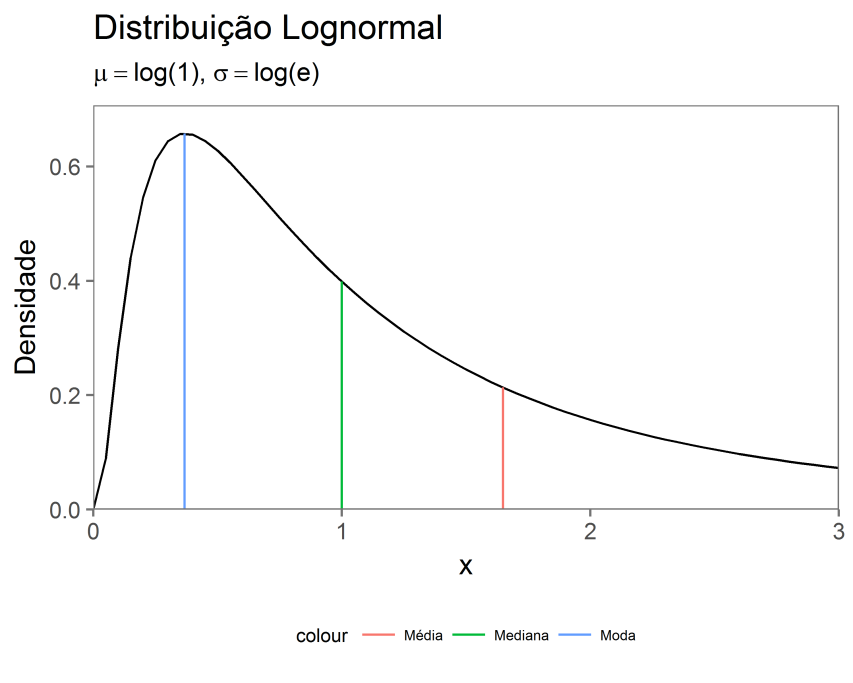


Figura 1 - Posições de medidas de tendência central numa distribuição lognormal

Fonte: Autor

**2.2.3 Efeito das variações do desvio-padrão na forma da distribuição**

Através da análise da figura 2 percebemos que, apesar das mudanças de forma decorrentes dos diversos fatores de , as características da distribuição log-normal são sempre a mesmas: tem início na origem do sistema de coordenadas, depois apresenta um pico (moda) que, considerando constante, tem o valor da sua abcissa cada vez menor, e uma cauda tendendo a zero quando o valor das abcissas tende a .

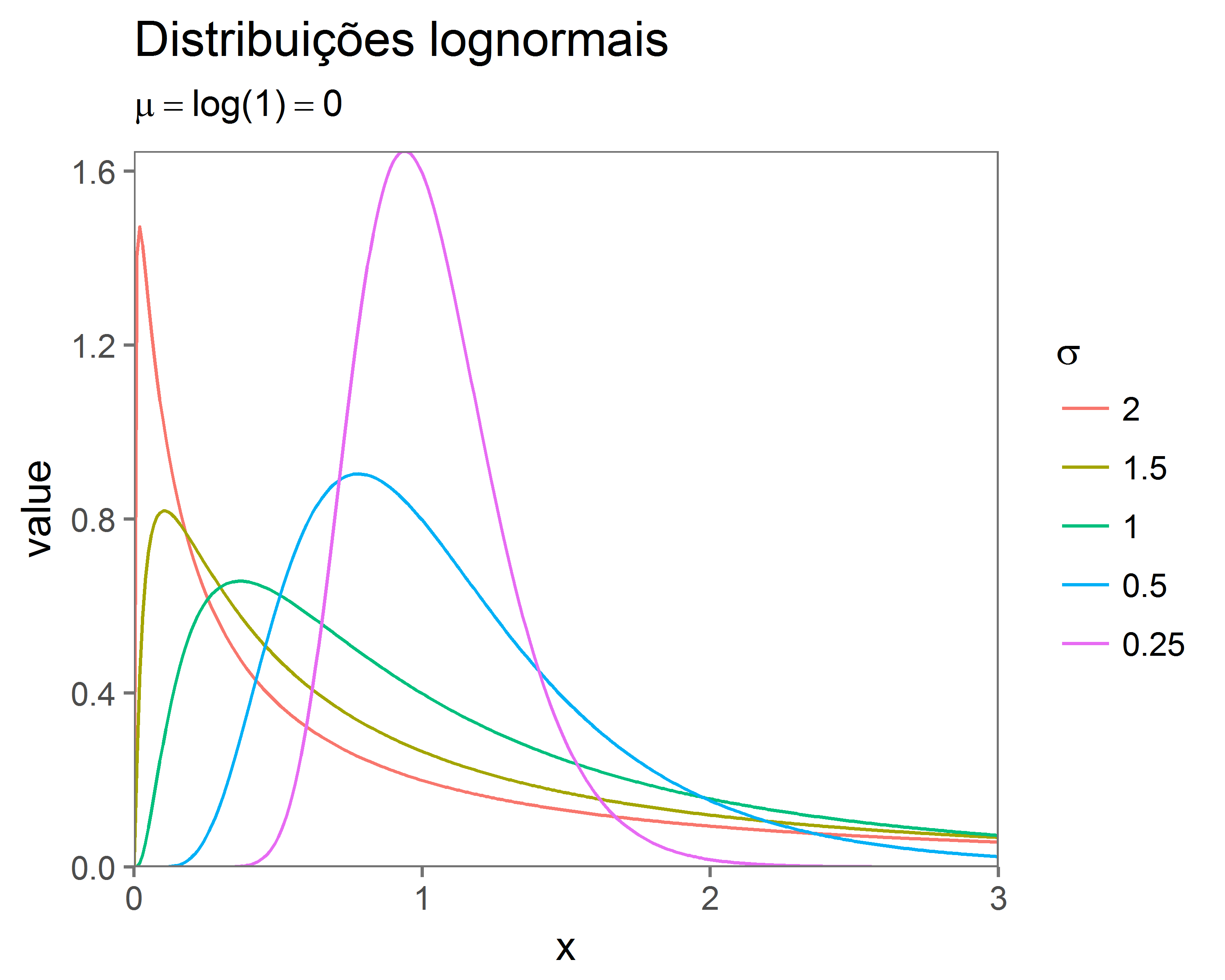


Figura 2 - Distribuição lognormal com e diversos valores de desvio-padrão

Fonte: Autor

**2.2.4 Relação com a distribuição normal**

Lembrando que a função densidade de probabilidade de uma variável aleatória com distribuição normal é dada por:

 (4)

E que para a distribuição normal-padrão () a função densidade de probabilidade torna-se:

 (5)

Seja uma variável aleatória de distribuição normal padronizada (), a função densidade de probabilidade e . Então () é igual a:

 (7)

o que equivale a:

 (8)

ou seja, a distribuição de uma variável , em que é equivalente a distribuição de uma variável lognormal com parâmetros e .

A figura 3 ilustra este fato.

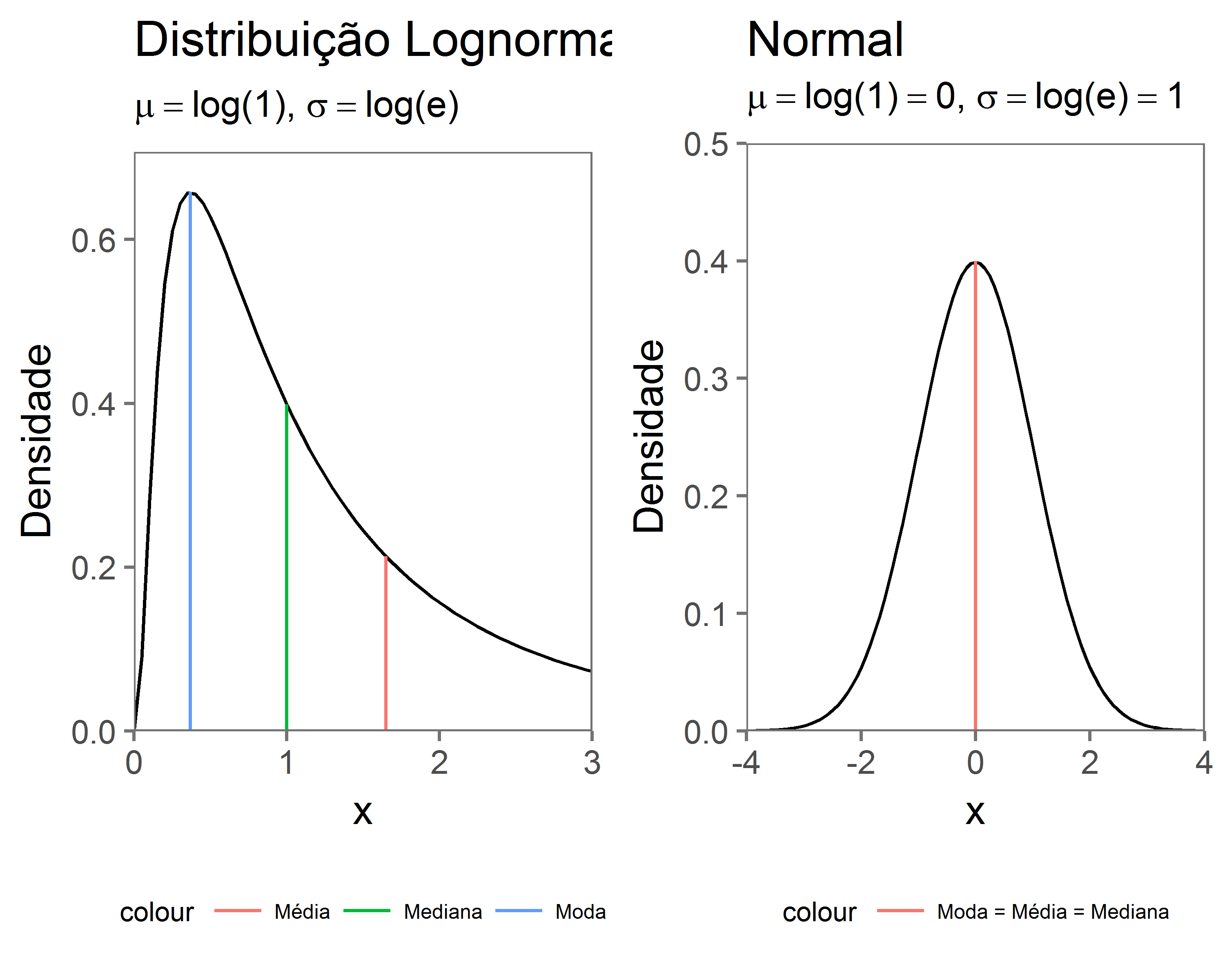


Figura 3 - Comparação entre distribuições normal e lognormal padronizadas.

Fonte: Autor

**2.2.5 Analogia com o Teorema do Limite Central**

Assim como o resultado da soma de diversas variáveis independentes com distribuições quaisquer resulta numa variável aleatória de distribuição normal (Teorema do Limite Central), o produto de diversas variáveis aleatórias resulta numa distribuição lognormal.

**2.2.6 Transformação de variável e Homoscedasticidade**

De acordo com Matloff (2017, p. 138), se uma variável aleatória W é aproximadamente normal, com baixo coeficiente de variação ( CV = σ/µ ), e g ( W ) é uma função suave, então a nova variável também será aproximadamente normal, com média g(EW) e variância:

 (9)

Assumindo que os erros de uma função de regressão sejam heteroscedásticos, seguindo uma função conhecida σ ( t ) = µ ( t ), se aplicarmos a função logaritmo natural à variável dependente, lembrando que a derivada da função logaritmo natural é  e que , segundo a equação 9, teremos:

 (10)

Ou seja, o uso da transformação logaritmo natural, para este caso em particular, conduz à homoscedasticidade do modelo.



De acordo com Matloff (2017, p. 138), ainda, se , a transformação raiz-quadrada é que traria de volta a homoscedasticidade.

**3. EXEMPLO**

**3.1 Dados**

Os dados aqui utilizados são oriundos de Hochheim (2015, pp. 21–22).

**3.2 Ajuste de distribuições aos dados**

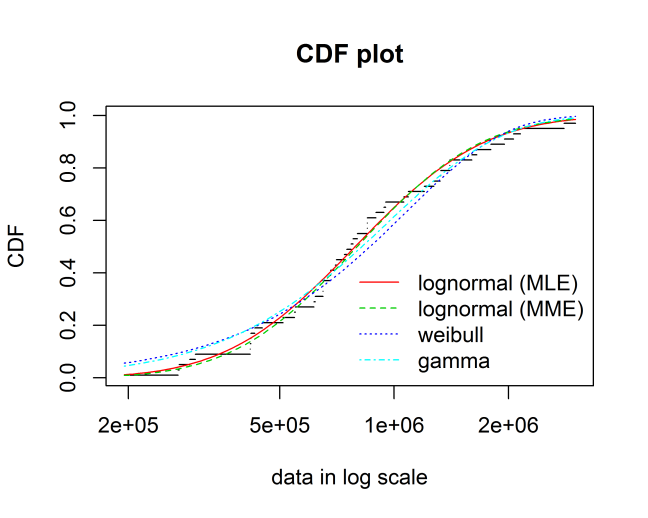


Figura 4 - Função densidade acumulada para os dados e para as distribuições testadas

Fonte: Autor

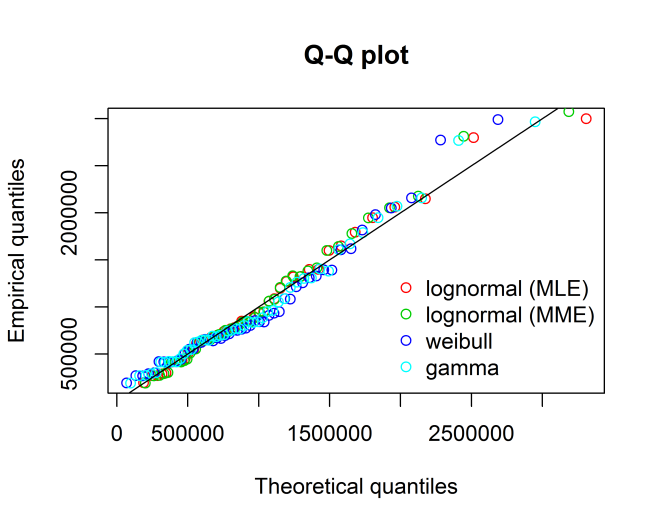


Figura 5 - Gráfico QQ para as diversas distribuições ajustadas

Fonte: Autor

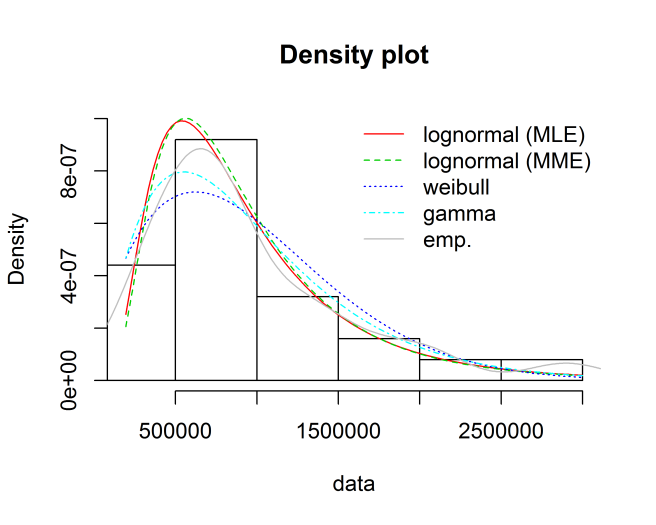


Figura 6 - Ajuste da distribuição empírica para funções de probabilidade conhecidas.

Fonte: Autor

Percebe-se pela análise das figuras 4 a 6 que o melhor ajuste se deu para a distribuição lognormal ajustada seja pelo método dos momentos (MME) ou pelo método da verossimilhança (MLE), haja vista que as outras distribuições inicialmente crescem mais rapidamente e tem pico mais achatado que os dados empíricos e as distribuições log-normais ajustadas.

**3.3 Gráficos**

As figuras 7 a 10 demonstram que os valores observados para a variável do conjunto de dados mencionados acima (HOCHHEIM, 2015, pp. 21–22) apresentam distribuição aproximadamente lognormal, com parâmetros

1. Densidade

A figura 7 mostra o gráfico da função densidade de probabilidade (FDP) construídos com os parâmetros e  obtidos da variável *valor*.

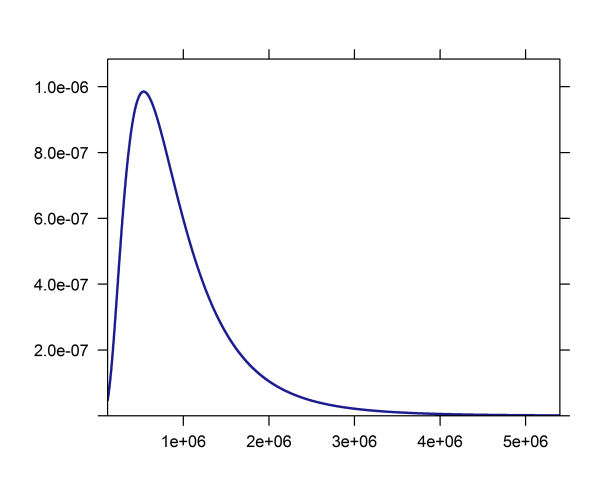


Figura 7 -FDP com parâmetros obtidos dos dados da variável *valor*

Fonte: Autor

1. Histograma com densidade superposta

A figura 8 mostra o histograma dos dados da variável valor, superposto com a curva da função densidade de probabilidade (FDP) da figura 7.

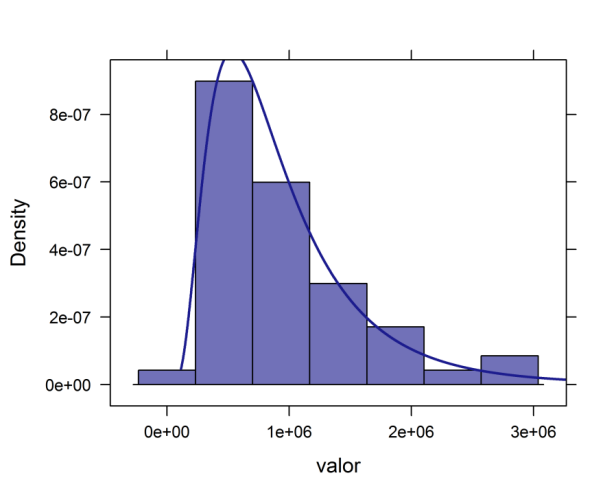


Figura 8 - Histograma das variável *valor* com FDP superposta

Fonte: Autor

1. Cumulativa

A figura 9 ilustra o gráfico da função cumulativa de probabilidade (FCP) para a variável *valor*.

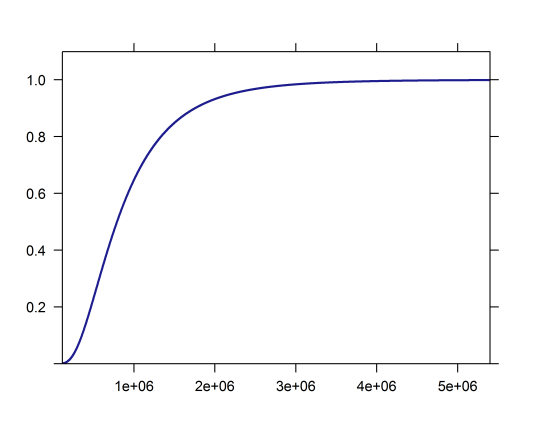


Figura 9 - FCP com parâmetros obtidos dos dados da variável *valor*

Fonte: Autor

1. Distribuição da variável ln(*valor*)

A figura 10 mostra a distribuição da variável ln(*valor*). Pode-se notar que, conforme esperado, já que a distribuição da variável *valor* é aproximadamente lognormal, seu logaritmo tem distribuição aproximadamente normal.

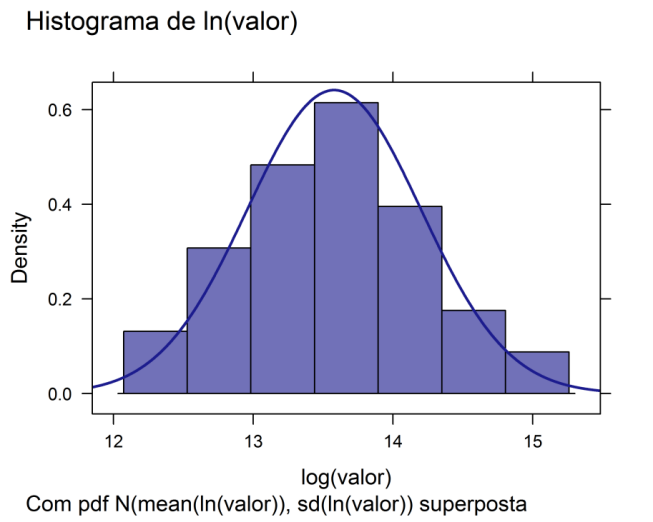


Figura 10 - Histograma com função densidade de probabilidade normal superposta

Fonte: Autor

**3.4 - Modelos**

Detectando-se a presença de variável resposta com distribuição lognormal, pode-se proceder da seguinte maneira:

* Proceder com a transformação da variável resposta pela função logarítmica;
* Proceder com a variável na escala original, corrigindo posteriormente a heteroscedasticidade com o método de Eickert-White;
* Proceder com o método dos mínimos quadrados ponderados.

É fácil mostrar que o modelo linear com a variável resposta logaritmizada, ou seja, com distribuição normal, é melhor ajustado que o modelo linear de uma variável resposta lognormal. Na tabela 1, no entanto, mostra-se que, para o presente caso, esta melhora de ajuste é modesta, próxima a 4,5%.

No entanto, o modelo sem transformação, segundo o teste de Breusch-Pagan, apresenta heteroscedasticidade, enquanto no modelo transformado pela função logaritmo natural, esta heteroscedasticidade desaparece.

A função máxima verosimilhança de Box-Cox também vai apresentar como transformação ótima a transformação logarítmica, como demonstra a figura 11

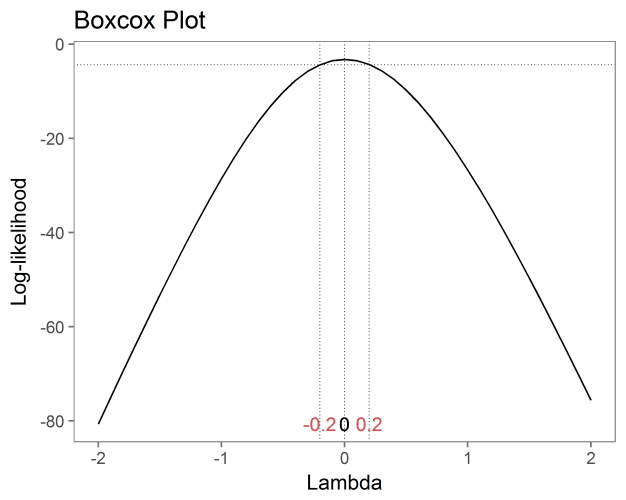


Figura 11 - Gráfico da função verossimilhança de Box-Cox

Fonte: Autor

Tabela 1- Comparação entre modelos com e sem transformação da variável resposta.

*Dependent variable:*

|  |  |  |
| --- | --- | --- |
|  | valor | log(valor) |
| (1) | (2) |
| area\_total | 2,893.178 | 0.002 |
|  | (2,065.405, 3,720.951) | (0.001, 0.002) |
|  | t = 6.850 | t = 4.886 |
|  | p = 0.00000\*\*\* | p = 0.00002\*\*\* |
| quartos | 73,524.375 | 0.169 |
|  | (−34,814.143, 181,862.894) | (0.084, 0.255) |
| t = 1.330 | t = 3.870 |
|  | p = 0.191 | p = 0.0004∗∗∗ |
| suites | 111,000.591 | 0.088 |
|  | (8,045.131, 213,956.052) | (0.007, 0.170) |
|  | t = 2.113 | t = 2.121 |
|  | p = 0.041\*\* | p = 0.040\*\* |
| garagens | 148,427.448 | 0.175 |
|  | (49,657.102, 247,197.795) | (0.097, 0.253) |
|  | t = 2.945 | t = 4.394 |
|  | p = 0.006\*\*\* | p = 0.0001\*\*\* |
| dist\_b\_mar | −223.217 | −0.0003 |
| (−434.862, −11.571) | (−0.0004, −0.0001) |
| t = −2.067 | t = −3.215 |
| p = 0.045\*\* | p = 0.003\*\*\* |
| padraomedio | −146,549.393 | 0.268 |
|  | (−354,850.457, 61,751.672) | (0.103, 0.433) |
|  | t = −1.379 | t = 3.190 |
|  | p = 0.176 | p = 0.003∗∗∗ |
| padraoalto | −56,064.550 | 0.334 |
|  | (−264,003.525, 151,874.426) | (0.169, 0.498) |
| t = −0.528 | t = 3.975 |
| p = 0.600 | p = 0.0003∗∗∗ |
| Constant | 33,953.788 | 12.315 |
|  | (−267,469.800, 335,377.375) | (12.076, 12.553) |
| t = 0.221 | t = 101.170 |
|  | p = 0.827 | p = 0.000\*\*\* |
| Observations | 50 | 50 |
| R2 | 0.906 | 0.940 |
| Adjusted R2 | 0.890 | 0.930 |
| Akaike Inf. Crit. | 1,375.659 | −29.275 |
| Residual Std. Error (df = 42) | 207,903.003 | 0.165 |
| F Statistic (df = 7; 42) | 57.731\*\*\* | 94.063\*\*\* |

*Note: \**p*<*0.1; \*\*p*<*0.05; \*\*\*p*<*0.01

Fonte:Autor

**3.4.1 Modelo linear com a variável resposta transformada**

**3.4.1.1 Retransformação de variáveis**

O problema da transformação da variável resposta no logaritmo da variável resposta original, é que devemos estudar como proceder na retransformação da variável, para efetuar a avaliação do imóvel.

**3.4.1.2 A desigualdade de Jensen**

Segundo Matloff (2017, p. 142), a desigualdade de Jensen (aplicada à estatística) se traduz na seguinte expressão, válida para funções convexas:

**  (9)

Isto aplicado no caso da transformação logarítimica, que é uma função côncava, se reduz à expressão abaixo (MATLOFF, 2017, p. 142):

 (10)

Para Matloff, então, como a igualdade só irá acontecer em poucos casos especiais, a função de regressão de será quase sempre menor do que o logaritmo natural da função de regressão de , de tal forma que a suposição que dado uma variável aleatória tal que assumimos que , não podemos concluir de imediato que um modelo linear razoável seria da forma , pois, pela desigualdade de Jensen, se temos dados significantemente heteroscedásticos da variável original (), a discrepância entre os dois lados da desigualdade acima poderia variar bastante com , potencialmente produzindo uma grande distorção à forma da curva de regressão (MATLOFF, 2017, p. 143).

Segundo Becker (2012, p. 4), a desigualdade de Jensen pode ser transformada numa igualdade do tipo:

 (11)

E, de acordo com o mesmo (2012, p. 5), o valor de , chamado de Defeito de Holder, é proporcional à variância da variável aleatória , tal que se é duas vezes continuamente diferenciável e existem limites finitos e tais que para todo , então existe um valor para o qual a fórmula abaixo é válida:

**  (12)

Em suma, o valor de é proporcional à variância de ().

Deste modo, existem na literatura diversos estudos sobre qual seria o “melhor” estimador – paramétrico ou não-paramétrico – para a variável resposta original, quando da ocorrência da transformação da variável pela função logaritmo natural, como pode ser visto em DUAN (1983), MEULENBERG (1965) e SHEN; ZHU (2008).

Entendemos que, na precisão necessária para a área de avaliação de imóveis, é suficiente a adoção do estimador teórico, apesar do funcionamento dos estimadores não-paramétricos ter sido demonstrado mais eficiente do que ele.

 (13)

**3.4.1 Modelo linear com erros robustos**

Uma vez detectada a heteroscedasticidade, uma outra possibilidade é permanecer com os valores da variável original e proceder com o cálculo de erros robustos pelo método de Eickert-White.Computados os erros robustos, podem ser calculados os intervalos de confiança para os regressores e os novos p-valores dos testes de significância. A tabela 2 mostra estes valores.

**3.4.2 Mínimos quadrados ponderados**

De acordo com Matloff (2017, p.139), em princípio, o Método dos Mínimos Quadrados ponderados (MQP) fornece melhores estimativas para os coeficientes e inferência estatística correta mesmo na presença de heteroscedasticidade.

Tal método consiste, analogamente ao Método dos Mínimos Quadrados Ordinários (MQO), em uma minimização. No caso dos MQO, minimiza-se a quantidade abaixo (MATLOFF, p. 69):

 (14)

Enquanto o MQP minimiza (MATLOFF, p. 133):

 (14)

Onde .

Tabela 2: Comparação entre modelos com e sem erros robustos

*Dependent variable:*

valor

default robust

(1) (2)

area\_total 2.893,178 2.893,178

(2.065,405, 3.720,951) (1.712,055, 4.074,302)

t = 6,850 t = 4,801

p = 0,00000\*\*\* p = 0,00001\*\*\*

quartos 73.524,375 73.524,375

(−34.814,143, 181.862,894) (−31.897,755, 178.946,506)

t = 1,330 t = 1,367

p = 0,191 p = 0,172

suites 111.000,591 111.000,591

(8.045,131, 213.956,052) (18.671,249, 203.329,934)

t = 2,113 t = 2,356

p = 0,041\*\* p = 0,019\*\*

garagens 148.427,448 148.427,448

(49.657,102, 247.197,795) (73.906,905, 222.947,992)

t = 2,945 t = 3,904

p = 0,006\*\*\* p = 0,0001\*\*\*

dist\_b\_mar −223,217 −223,217

(−434,862, −11,571) (−406,588, −39,845)

t = −2,067 t = −2,386

p = 0,045\*\* p = 0,018\*\*

padraomedio −146.549,393 −146.549,393

(−354.850,457, 61.751,672) (−322.293,226, 29.194,441)

t = −1,379 t = −1,634

p = 0,176 p = 0,103

padraoalto −56.064,550 −56.064,550

(−264.003,525, 151.874,426) (−197.646,893, 85.517,794)

t = −0,528 t = −0,776

p = 0,600 p = 0,438

Constant 33.953,788 33.953,788

(−267.469,800, 335.377,375) (−191.823,586, 259.731,161)

t = 0,221 t = 0,295

p = 0,827 p = 0,769

|  |  |  |
| --- | --- | --- |
| Observations | 50 | 50 |
| R2 | 0,906 | 0,906 |
| Adjusted R2 | 0,890 | 0,890 |
| Akaike Inf. Crit. | 1.375,659 | 1.375,659 |
| Residual Std. Error (df = 42) | 207.903,003 | 207.903,003 |
| F Statistic (df = 7; 42) | 57,731\*\*\* | 57,731\*\*\* |
| *Note:* |  | \*p*<*0,1; \*\*p*<*0,05; \*\*\*p*<*0,01 |

Fonte: Autor

**4 CONCLUSÃO**

Foi possível demonstrar de maneira gráfica que os dados da variável *valor* apresentados se ajustam bem a uma distribuição lognormal equivalente. Por definição, então, o logaritmo da variável possui distribuição normal.

Consideramos que o valor mais provável para a variável resposta é o seu Valor Esperado. Logo, a retransformação da variável deve ser feita para a média da variável log-normal.

Uma alternativa seria a aplicação da regressão linear ponderada (ou mínimos quadrados ponderados). No entanto, o método de Eickert-White parece ser adequado e sua aplicação menos complexa do que o método de regressão ponderada.

**Referências Bibliográficas**

BECKER, R. A. The variance drain and Jensen’s inequality. **CAEPR Working Paper**, 2012. Disponível em: <http://dx.doi.org/10.2139/ssrn.2027471>..

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>..

FARIAS, A. M. L. DE. **Métodos estatísticos aplicados à economia II: Variáveis aleatórias contínuas**. Universidade Federal Fluminense,.

FDA. **Guideline for the format and content of the clinical and statistical sections of new drug applications**. Food and Drug Administration, Public Health Service, US Department of Health and Human Services, 1988.

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

KEENE, O. N. The log transformation is special. **Statistics in Medicine**, v. 14, p. 811–819, 1985.

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

MEULENBERG, M. T. G. On the estimation of an exponential function. **Econometrica**, v. 33, n. 4, p. 863–868, 1965. [Wiley, Econometric Society]. Disponível em: <http://www.jstor.org/stable/1910362>..

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>..