PRESSUPOSTOS CLÁSSICOS DOS MODELOS DE REGRESSÃO LINEAR E SUAS IMPLICAÇÕES SOBRE AS AVALIAÇÕES EM MASSA.

Willian Zonato
Ministério do Planejamento, Desenvolvimento e Gestão
MPDG/SPU/SC
Praça XV de Novembro, 336, Centro, Florianópolis/SC, 88010-400
willian.zonato@planejamento.gov.br

Luiz Fernando Palin Droubi
Ministério do Planejamento, Desenvolvimento e Gestão
MPDG/SPU/SC
Praça XV de Novembro, 336, Centro, Florianópolis/SC, 88010-400
luiz.droubi@planejamento.gov.br

Norberto Hochheim
Universidade Federal de Santa Catarina
Campus Reitor João David Ferreira Lima, s/n - Trindade, Florianópolis/SC, 88040-900
hochheim@gmail.com

Resumo:
Este artigo visa mostrar com clareza as hipóteses clássicas da regressão linear e as implicações do não atendimento destas hipóteses no processo de inferência. Especial atenção é dada à hipótese de maior impacto sobre este processo, a saber, a hipótese da homoscedasticidade. Especialmente nos modelos de avaliação em massa, o problema da heteroscedasticidade tende a ser mais comum, haja vista a grande amplitude do domínio do modelo. Desta maneira, mostramos como surge a heteroscedasticidade, como esta pode ser detectada, e as maneiras de contorná-la. Suas implicações são detalhadas com precisão. Para o presente artigo foi realizado um estudo de caso, utilizando-se dados de imóveis comerciais na região central da cidade de Florianópolis/SC, fazendo uso de diferentes transformações para a variável dependente. Objetivando alcançar a transformação mais adequada da variável dependente, para o modelo de regressão, foi aplicado o método de Box-Cox. Por fim, foram comparadas as estimativas realizadas com os diferentes modelos, fazendo uso do método de Eicker-White para o cômputo da matriz de Covariância na presença de heteroscedasticidade. Constatou-se que aplicação do método Eicker-White, para fins de contorno da heterocedasticidade, pode ser considerada uma boa alternativa frente a pesquisa de transformações de variaveis que estabilizem a variância do modelo de regressão.
Palavras-chave: heterocedasticidade, covariância, regressão linear, avaliação em massa.

Abstract:

This article aims to show clearly the classic hypotheses of linear regression and the implications of not attending these hypotheses in the process of inference. Special attention is given to the hypothesis of greater impact on this process, namely the hypothesis of homoscedasticity. Especially in mass evaluation models, the problem of heteroskedasticity tends to be more common, due to the large breadth of the model domain. In this way, we show how heteroscedasticity arises, how it can be detected, and the ways to circumvent it. Their implications are detailed precisely. Finally, a case study with commercial real estate data is carried out in the central region of the city of Florianópolis / SC, making use of different transformations for the dependent variable. It is also detailed the Box-Cox method to search for the most appropriate transformation of the dependent variable Finally, the estimates made with the different models are compared using the Eicker-White method for the computation of the Covariance matrix in the presence of heteroskedasticity. It was verified that the application of the Eicker-White method, for contouring the heteroscedasticity, can be considered a good alternative to the research of transformations of variables that stabilize the variance of the regression model.

Keywords: heteroskedasticity, Covariance, Linear Regression, Mass Appraisal.

1. INTRODUÇÃO

A avaliação em massa de imóveis, também chamada de avaliação coletiva, é um procedimento de larga escala, com o objetivo de determinar de forma sistemática os valores destes, mantendo uma justa proporcionalidade entre eles, aplicando-se técnicas consagradas e ferramentas tecnológicas disponíveis para o tratamento dos dados amostrais pesquisados (LIPORONI, 2014).

Liporoni (2014) ainda cita que as metodologias possíveis para o tratamento destes dados amostrais são: a metodologia determinística e a probabilística. A metodologia probabilista é a que tratamos no presente artigo, pois esta se fundamenta na utilização da regressão linear múltipla para estimar os valores da avaliação em massa.

De acordo com Matloff (2015), quase todos os dados na vida real são heteroscedásticos. Apesar disto, um pressuposto básico da regressão linear é a homoscedasticidade dos resíduos. Ou seja, a regressão linear assume que a variância da variável resposta Y dado X, em outros termos, $Var(Y \mid X = t)$ é constante para todo o intervalo de variação de X, ou seja, qualquer que seja o valor de $t \in X$. Em termos estritamente matemáticos:

Se:

$$Var(Y \mid X = t) = k \qquad \forall t \in X \tag{1}$$

onde k é uma constante, então diz-se que o modelo de regressão ordinário Y ~ X é homoscedástico.

A homoscedasticidade, porém, não é necessária para o simples intuito da previsão de valores através do modelo de regressão. As estimativas obtidas na presença ou na ausência da homoscedasticidade são as mesmas, haja vista que os valores dos coeficientes de regressão são também os mesmos. O que muda são as inferências feitas quanto aos erros do modelo, tais como os intervalos de confiança, por exemplo.

Satisfeita a hipótese da homoscedasticidade, são calculados com precisão os intervalos de confiança dos regressores, os p-valores de seus testes de hipótese e os intervalos de confiança das previsões.

Uma das maneiras de contornar a presença de heteroscedasticidade é a utilização de transformações adequadas das variáveis do modelo, em especial a transformação da variável dependente.

No entanto, a utilização de transformações de variáveis em modelos de regressão linear devem ser estabelecidas de forma criteriosa, sob pena de induzir a resultados incoerentes e interpretação equivocada dos modelos.

De fato, é sabido que a *Food and Drug Administration* (FDA), não recomenda o uso de transformações de variáveis, exigindo a elaboração de explicações para o modelo transformado e a fundamentação da aplicação das transformações.

Mesmo a NBR 14653-2 (2011) faz restrições ao uso das transformações, recomendando a elaboração de gráficos entre as variáveis para a melhor compreensão do funcionamento das variáveis, assim como métodos estatísticos como o método de Box-Cox.

Considerando-se a existência de outras maneiras para se contornar o problema da heteroscedasticidade, cabe discutir se a transformação de variáveis seria a forma mais adequada de contorno, em que pese ser o procedimento usualmente adotado na prática da engenharia de avaliações.

O referencial teórico considerado no presente trabalho aponta que a hipótese da normalidade não se mostra tão relevante quanto a hipótese da heteroscedasticidade. A linearidade do modelo deve ser verificada sempre.

A aplicação do método de Eicker-White para o cômputo da matriz de covariância permite calcular erros robustos para a regressão linear. Seu cômputo é computacionalmente leve e a adoção deste método ao invés da transformação de variáveis é vantajoso, haja vista que não introduz qualquer deformação nos dados originais.

2. REGRESSÃO LINEAR

2.1. Pressupostos clássicos de um modelo de regressão linear

De acordo com Matloff (2009, pp. 419–429), a análise clássica pressupõe três hipóteses para a regressão linear:

- Linearidade do Modelo
- A distribuição condicional de Y dado X é normal.
- Homogeneidade da variância (Homocedasticidade) da variável resposta Y dado o conjunto de variáveis independentes X.

Satisfeitas as três hipóteses acima, pode-se concluir pela validade do modelo, assim como de seus erros e intervalos de confiança. Neste passo, pretende-se demonstrar no presente artigo como verificar estas hipóteses com confiabilidade, e quais as implicações destas hipóteses não serem atendidas.

2.1.1 A hipótese da linearidade do modelo

A hipótese da linearidade do modelo de regressão linear clássico pode ser resumido matematicamente como abaixo:

$$E(Y \mid X = t) = t'\beta \tag{2}$$

Ou seja, existe um vetor β tal que a equação acima é válida para todo *t*.

Esta hipótese é fundamental, mas não é de grande interesse, pois se esta hipótese não for aceita, simplesmente isto significa que não temos uma regressão linear, ou seja, significa que não podemos prever Y linearmente em função de X, pois Y é alguma função não-linear de X. Normalmente, no entanto, após uma adequada transformação de variáveis, a equação 2 será aproximadamente válida.

2.1.2. A hipótese da distribuição normal de $Y \mid X_i$

Segundo Matloff (2015, p. 55) a hipótese da normalidade não é tão importante, dado que os valores dos coeficientes de regressão β são normais mesmo que a hipótese da normalidade não seja satisfeita. Aliás, mesmo a hipótese da homoscedasticidade não é necessária para o cálculo dos β_i, que são não-enviesados, consistentes e assintoticamente normais mesmo sem que estas hipóteses sejam atendidas. O problema começa a ter relevância apenas no cálculo dos erros e intervalos de confiança. A hipótese da normalidade, no entanto, é uma aproximação que nunca vai ocorrer no mundo real:

We must live with approximations one way or the other, and the end result is that the normality assumption is not very important. (MATLOFF, 2015, p. 84)

2.1.3. A hipótese da homoscedasticidade

A hipótese da homoscedasticidade pode ser resumida na equação 3.

$$\sigma^2(t) = \mathrm{Var}(Y \mid X = t) = k \tag{3}$$

Ou seja, a hipótese da homoscedasticidade é a hipótese de que a variância da variável dependente é constante qualquer que seja o valor da(s) variável(eis) independente(s).

2.1.3.1. Heteroscedasticidade

A heteroscedasticidade é exatamente o inverso da homoscedasticidade. Ou seja, dados heteroscedásticos são aqueles em que:

$$\mathrm{Var}(Y \mid X) = f(X) \tag{4}$$

Graficamente, a heteroscedasticidade pode ser notada através da análise do gráfico de dispersão da variável dependentes *versus* variável independente, como pode ser observado na figura 1.

2.1.3.2. A ocorrência da heterocedasticidade

A heteroscedasticidade ocorre, na prática, por diversos motivos. Primeiramente, os dados observados para a variável resposta podem ter realmente variância inconstante ao longo do intervalo da(s) variável(eis) dependente(s) e não há nada que possamos fazer quanto aos dados reais. Uma opção seria a transformação da variável dependente ou da(s) variável(eis) independente(s) mas este problema de como contornar a heteroscedasticidade será visto na secção 2.1.2.5.

Outros motivos para a ocorrência de modelos heteroscedásticos incluem: má especificação do modelo, má escolha das transformações das variáveis ou ainda a falta de alguma(s) variável(eis) explicativa(s) no modelo.

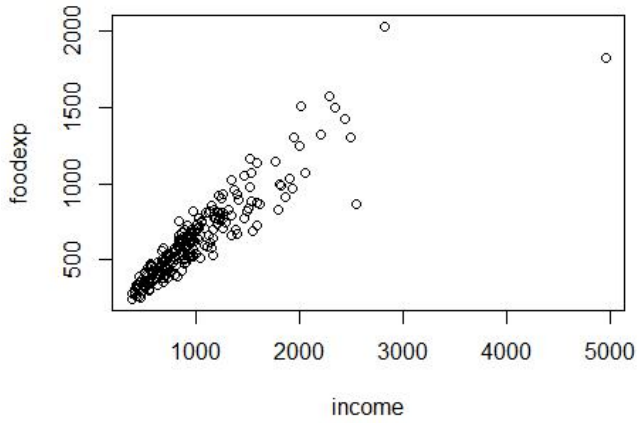

Figura 1 - Gastos com alimentação *versus* renda
Fonte: KOENKER; BASSETT (1982)

2.1.3.3. A detecção da heterocedasticidade

A detecção da heteroscedasticidade pode ser feita através de diversos testes já sacramentados estatisticamente. O teste de Breusch-Pagan, por exemplo, detecta forma lineares de heteroscedasticidade, mas não detecta as suas formas não-lineares. Para formas não-lineares de heteroscedasticidade pode-se aplicar o teste de White, um caso particular do teste de Breusch-Pagan. Outra maneira é através da análise gráfica dos resíduos vs. valores ajustados (ε *vs.* $\hat{Y}$).

Para o estudo em tela foi aplicado o teste de Breusch-Pagan, aos dados de Engel, no software estatístico R version 3.4.3 (2017-11-30), através da função bptest do pacote lmtest (ZEILEIS; HOTHORN, 2002).

2.1.3.4. Implicações da presença da heteroscedasticidade

A simples presença da heteroscedasticidade não anula o modelo de regressão como um todo e todo o processo de inferência. Apenas algumas partes da análise de inferência usual estarão comprometidas: os erros, testes de significância dos regressores e os intervalos de confiança, tanto dos regressores, como das estimativas porventura obtidas com o modelo.

2.1.3.4.2 Intervalo de confiança dos regressores

Os intervalos de confiança para cada regressor, supondo a homoscedasticidade do modelo, são calculados pela seguinte expressão (FARAWAY, 2004, p. 38):

$$\hat{\beta} \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{(X^T X)_{ii}^{-1}} \quad (5)$$

Tabela 2- Resultados do teste de Breusch-Pagan, através do software R 3.4.3.

Results	
BP	12,934
df	6
p-valor	0,04409

Fonte: Autor

Tabela 3- Estimativas resultantes do modelo heteroscedástico com intervalo de confiança de 80%

Results	
Predicted value	2330,78
CI lower	2265,13
CI upper	2396,42
Amplitude	0,056

Fonte: Autor

É fácil notar pela análise da equação 5, portanto, que os IC dos regressores calculados estarão comprometidos quando $\hat{\sigma} \neq cte$, e não um escalar, como diz a hipótese da homoscedasticidade.

Tabela 3- Coeficientes do modelo linear sobre os dados de Engel, através do R 3.4.3.

	Dependent variable
	foodexp
Income	0.485
	(0.014)
	t = 33.772
	p = 0.000***
Constant	147.475
	(15.957)
	t = 9.242
	p = 0.000***
Observations	235
R^2	0.830
Adjusted R^2	0.830
Residual Standart Error	114.108 (df = 233)
F Statistic	114.108 (df = 1; 233)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fonte: Autor

2.1.3.4.3 Na realização de estimativas e avaliações.

Um modelo heteroscedástico ainda é capaz de fazer previsões para a estimativa central sem comprometimento algum, mas é incapaz de calcular os intervalos de confiança ou de predição, pois estes são dependentes dos erros, assim como os IC para os regressores.

O intervalo de confiança para a resposta média, por exemplo, é calculado de acordo com a expressão 6 (FARAWAY, 2004, p. 42):

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0} \tag{6}$$

Já o intervalo de predição é calculado de acordo com a equação 7 (FARAWAY, 2004, p. 42):

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \tag{7}$$

Donde é fácil notar, também, porque em um modelo heteroscedástico os IC para as estimativas estão comprometidos ($\hat{\sigma} \neq cte$).

2.1.3.5. Possibilidade de contorno quanto a heterocedasticidade

A heteroscedasticidade é um problema prático que pode ser contornado de diversas maneiras, conforme cada caso.

Na Engenharia de Avaliações, a tendência é procurar transformar as variáveis de maneira que ela desapareça. Porém, nem sempre o avaliador está tão atento à heteroscedasticidade quanto ao grau de ajuste do modelo, ou à presença de outliers na amostra. Algumas vezes os avaliadores encontram modelos que estão de acordo com o que eles esperam quanto ao grau de ajuste, normalidade dos resíduos, e outras verificações, mas não se lembram de verificar a homoscedasticidade. Adicionalmente, infelizmente, nem sempre os modelos mais ajustados encontrados com os atuais *software* de avaliações comerciais disponíveis na praça são os melhores modelos.

Felizmente, existem outras maneiras de contornar a heteroscedasticidade. Segundo Matloff (2015), porém, existe pouca literatura onde estes métodos são explicitados, como por exemplo a teoria de Eicker-White, que desenvolve uma inferência assintótica válida para dados heteroscedásticos (MATLOFF, 2015).

Com o método de Eicker-White é possível calcular erros robustos para o modelo, o que torna possível o cálculo dos intervalos de confiança na presença de heteroscedasticidade.

3. ESTUDO DE CASO

À guisa de exemplo, a seção 3.1 apresenta a aplicação do método de Eicker-White aos dados de Engels. Já a seção 3.2 apresenta dois modelos de regressão, com transformações da variável dependente, resultantes de dados de imóveis comerciais no centro de Florianópolis.

3.1 Aplicação do Método de Eicker-White

Foi aplicado o método de Eicker-White aos dados de Engel, vistos na figura 1. Os erros calculados de acordo com este método são ditos erros robustos, ou seja, os erros são calculados corretamente, mesmo na presença de heteroscedasticidade. Os resultados podem ser vistos na tabela 4.

Tabela 4- Resultados da aplicação do método de Eicker-White

	Dependent variable	
	foodexp	
	Default (1)	*Robust (2)*
Income	0.485***	0.485***
	(0.014)	(0.0052)
Constant	147.475	147.475
	(15.957)	(46.449)
Observations	235	235
R^2	0.830	0.830
Adjusted R^2	0.830	0.830
Residual Standart Error	114.108 (df = 233)	114.108 (df = 233)
F Statistic	114.108 (df = 1; 233)	114.108 (df = 1; 233)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Fonte: Autor

Destaca-se que, conforme esperado, os coeficientes tem valores idênticos aos calculados anteriormente. Porém os erros-padrão (e portanto o IC dos coeficientes e os p-valores) tem valores drasticamente diferentes.

Abaixo pode ser visto o efeito da correção da heteroscedasticidade sobre as estimativas feitas com o modelo, para renda (*income*) igual a $ 4.500,00.

Tabela 5- Estimativas do modelo, através do software R 3.4.3

Results	
Predicted value	2330,78
CI lower	1962,76
CI upper	2698,79
Amplitude	0,316

Fonte: Autor

3.2 Problemas decorrentes da má escolha de transformações para a variável dependente

Para exemplificar o que ocorre com os modelos em que há presença de heteroscedasticidade foram elaborados dois modelos à partir de dados de imóveis comerciais no centro de Florianópolis em 2017. O resumo dos modelos pode ser visto na tabela 6.

Como percebe-se, para estes dados, o melhor modelo é o modelo (1) onde a variável resposta é transformada pela função raiz-quadrada, como pode-se observar na figura 2 ($\hat{\lambda} \approx 0{,}5$).

Tabela 6: Comparação de modelos com diferentes transformações de variáveis.

	Dependent variable	
	model (1)	*model (2)*
	sqrt (VU)	*log (VU)*
Andar	0.701 (-0.310, 1.711)	-
	t = 1.359	-
	p = 0.179	-
Vagas	3.726 (1.950, 5.502)	0.082 (0.042, 0.122)
	t = 4.113	t = 3.995
	p = 0.0002**	p = 0.0002***
Padrão Alto	10.270 (1.603, 18.938)	0.281 (0.096, 0.465)
	t = 2.322	t = 2.985
	p = 0.024**	p = 0.004**
Idade	-0.695(-1.062,-0.327)	-0.016(-0.024,-0.008)
	t = -3.705	t = -3.849
	p = 0.0005**	p = 0.0003**
AreaPrivativa:Loja	-0.242(-0.369,-0.116)	-0.007(-0.009,-0.004)
	t = -3.764	t =-5.378
	p =0.0004**	p = 0.0000**
AreaPrivativa:Sala	-0.135(-0.178,-0.092)	-0.003(-0.004,-0.002)
	t = -6.126	t = -6.421
	p = 0.0000**	p = 0.0000**
Constant	108.230(96.187, 120.273)	9.446(9.229, 9.662)
	t = 17.614	t = 85.438
	p = 0.0000**	p = 0.0000**
Observations	75	75
R^2	0.646	0.635
Adjusted R^2	0.614	0.608
Residual Standart Error	13.413 (df = 68)	0.304 (df = 69)
F Statistic	20.639*** (df = 6; 68)	23.968*** (df = 5; 69)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Fonte: Autor

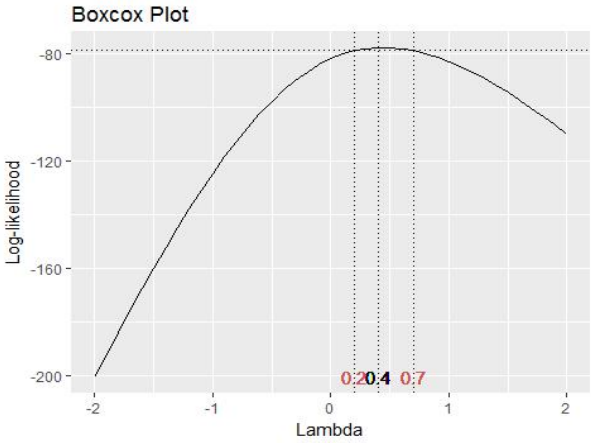

Figura 2 - Diagrama de máxima verossimilhança de Box-Cox.
Fonte: Autor

Em ambos os modelos foram retirados os dados considerados *outliers*, quais sejam, os dados(pontos) nº 2, 54, 56, 70 e 79. Também foi desconsiderada a variável *Conservacao*, por não se mostrar significante.

A única diferença entre os dois modelos, então, é a transformação aplicada à variável dependente: enquanto no primeiro modelo utilizamos a raiz-quadrada, no segundo modelo utilizamos a função *log* natural.

Pode-se notar que, embora os modelos sejam parecidos, o primeiro possui resíduos homoscedásticos, enquanto no segundo eles não o são, como pode ser observado nas figuras 3 e 4, respectivamente.

No primeiro modelo, com a transformação adequada, os resíduos são praticamente constantes ao longo de todo o intervalo de valores previstos. Já no segundo, os resíduos variam (quase) linearmente com a escala, reduzindo (em módulo) conforme aumenta o valor ajustado.

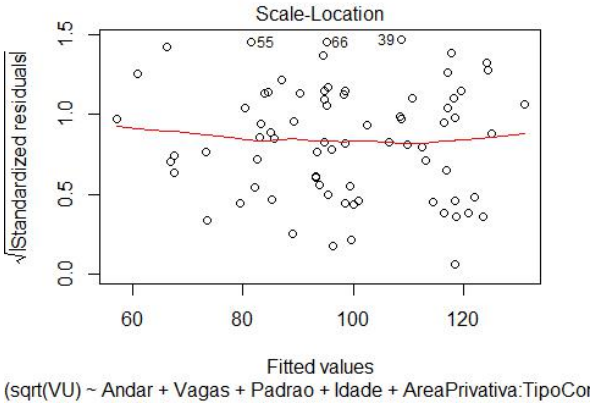

Figura 3 – Resíduos com a transformação raiz quadrada da variável dependente.
Fonte: Autor

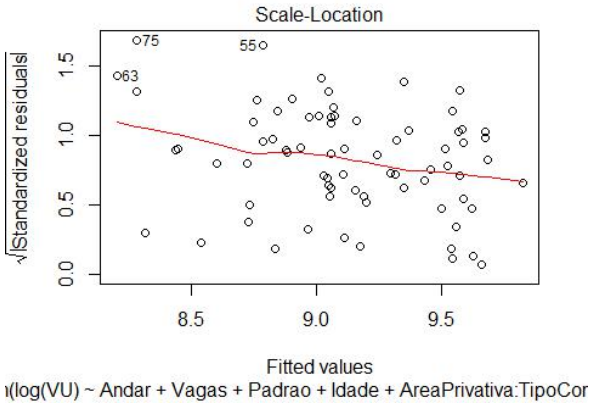

Figura 4 – Resíduos com a transformação logarítmica da variável dependente.
Fonte: Autor

O primeiro modelo, que apresenta transformação adequada (raiz-quadrada) para a variável dependente, passa no teste de Breusch-Pagan:

Tabela 6- Resultados do teste estatístico através do software R 3.4.3

Results	
BP	10,346
df	6
p-valor	0,1108

Fonte:Autor

Com a escolha inadequada da transformação da variável dependente, o modelo não passa no teste estatístico de Breusch-Pagan, ou seja, comprova-se a existência da heterocedasticidade pelo teste formal.

Tabela 7- Resultados do teste estatístico através do software R 3.4.3

Results	
BP	12,934
df	6
p-valor	0,04409

Fonte:Autor

4. CONSIDERAÇÕES FINAIS

O método de Eicker-White pode ser encontrado em diversos *software* estatísticos comerciais ou livres, como o R, e seu uso é simples e eficaz para contornar o problema da heteroscedasticidade. Haja vista as considerações de Matloff (2009), a normalidade não é tão impactante no processo de inferência e a transformação dos dados apenas para chegar à normalidade não parece ser sempre uma boa idéia, haja vista a distorção que isto pode provocar na correlação das variáveis. O método ainda é computacionalmente leve e pode ser implementado facilmente nos *software* comercias de avaliação de imóveis.

O uso do método de Eicker-White pode ser ainda mais útil na avaliação em massa de imóveis, dado que os modelos de avaliação em massa geralmente precisam ter validade num domínio mais amplo do que os modelos de avaliações específicas de um imóvel em particular e nestas grandes amplitudes de domínio é que o problema da heteroscedasticidade parece ser mais comum.

Referências Bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14.653 – **Avaliação de bens: Parte 1 – Procedimentos Gerais**. Rio de Janeiro, 2001.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14.653 – **Avaliação de bens: Parte 2 – Imóveis Urbanos**. Rio de Janeiro, 2011

FARAWAY, J. **Linear models with r**. Taylor & Francis, 2004.

KOENKER, R.; BASSETT, G. **Robust tests of heteroscedasticity based on regression quantiles**. **Econometrica**, v. 50, p. 43–61, 1982.

LIPORONI, A. S. **Avaliações em massa com ênfase em planta de valores** .Engenharia de avaliações **IBAPE/SP**, v. 2, p.644, Editora Leud , São Paulo. 2014.

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

MATLOFF, N. S. **Can you say "heteroscedasticity" 3 times fast**?, 2015. Wordpress. Disponível em: <https://matloff.wordpress.com/2015/09/18/can-you-say-heteroscedasticity-3-times-fast/>..

ZEILEIS, A.; HOTHORN, T. **Diagnostic checking in regression relationships. R News**, v. 2, n. 3, p. 7–10, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>..