FATORES DETERMINANTES NO PREÇO DE VENDA DE CASAS (CANADÁ, 1987)

ANA HERMÍNIA ANDRADE E SILVA
HELOISA DE MELO RODRIGUES
SADRAQUE ENEAS DE FIGUEIREDO LUCENA

RESUMO. O mercado imobiliário é uma das áreas de maior dinâmica do setor econômico e a avaliação dos bens é dificultada pela variação nas características e atributos dos imóveis. O presente trabalho tem como objetivo determinar os fatores que apresentam maior influência no preço de venda de casas (variável resposta) no Canadá em 1987 por meio de análise de regressão linear múltipla. Foi utilizada a transformação logarítmica na variável resposta para corrigir os desvios da normalidade dos dados, assim como na variável explicativa tamanho do lote do imóvel. Verificou-se que os seguintes fatores determinam o preço de venda das casas: o tamanho do lote do imóvel, se a casa tem 2, 3 ou 4 banheiros, a presença de 2, 3 ou 4 pavimentos, possuir uma entrada para garagem, ter um porão totalmente construído, usar gás para aquecimento de água, possuir ar-condicionado central e ainda, haver no imóvel 1 ou 2 garagens.

1. INTRODUÇÃO

O mercado imobiliário figura como uma das áreas de maior dinâmica do setor econômico e a avaliação dos bens é dificultada pela variação nas características e atributos dos imóveis (Steiner et al. 2008) (Bailey et al. 1963). Nesse contexto, procedimentos estatísticos utilizados de forma sistemática reduzem a avaliação subjetiva de profissionais da área, fornecendo um modelo que considere a heterogeneidade dos imóveis e suas localizações, representando o mercado com máxima precisão possível (Moreira et al. 2010).

Dentre as técnicas utilizadas para avaliação de imóveis, uma das mais disseminadas é a análise de regressão múltipla, que considera um modelo validado estatisticamente (contendo características importantes e seus respectivos pesos) que é empregado na projeção do valor de um imóvel analisado (Dantas 1998)(Isakson 2001)(Ramsland & Markham 1998).

O presente estudo tem por objetivo determinar os fatores que apresentam maior influência no preço de casas vendidas no Canadá em 1987 por meio de uma análise de regressão linear múltipla. Na seção 2 é apresentado o modelo de regressão, possíveis problemas e suas soluções ao realizar as análises. Na seção 3 é exposta uma análise descritiva do banco de dados e na seção 4 é apresentada

Key words and phrases. Preço de venda de casas, Regressão Linear Múltipla, Variáveis Dummy.

a seleção do modelo de regressão e analisado o modelo final. Na seção 5 são apresentadas as considerações finais.

2. Regressão Múltipla

Em um modelo de regressão busca-se identificar quais características da população (em geral a média) é afetada pelo comportamento de outras variáveis, chamadas variáveis explicativas, regressores ou covariadas. Podemos definir o modelo de regressão linear múltipla da seguinte forma (Draper & Smith 1998)

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

em que $\underset{\sim}{y}$ é um vetor $(T \times 1)$ contendo T observações da variável resposta; X é uma matriz $(T \times k)$ contendo as variáveis explicativas em cada coluna (a primeira coluna de X contém apenas 1 se há intercepto no modelo); $\underset{\sim}{\beta}$ é um vetor $(T \times 1)$ de parâmetros e $\underset{\sim}{\epsilon}$ é um vetor $(T \times 1)$ de erros. O vetor de parâmetros $\underset{\sim}{\beta}$ é estimado por mínimos quadrados ordinários (MQO), sendo dado por

$$\underset{\sim}{b} = X(X'X)^{-1}X'\underset{\sim}{y}.$$

Sobre o modelo de regressão são feitas as seguintes suposições:

- O modelo estimado é o modelo correto;
- Os erros têm média zero, isto é, $\mathbb{E}(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$;
- (Homoscedasticidade) A variância dos erros é constante, isto é, $Var(\underset{\sim}{\epsilon}) = \sigma^2 I$, $(0 < \sigma^2 < \infty)$ e I é a matriz identidade;
- (Não-autocorrelação) A covariância dos erros é zero, isto é, $Cov(e_t, e_s) = 0$, em que e_t e e_s são os erros correspondentes às observações t e s, $t \neq s$;
- As colunas de X são linearmente independentes, isto é, X possui posto completo (posto$(X) = k$, $k < T$);
- Os erros têm distribuição normal. Esta suposição é utilizada para estimação intervalar e testes.

Em um modelo de regressão uma ou mais variáveis explicativas não necessariamente assumem valores em uma escala contínua, ou seja, a variável aleatória é tida como um fator que pode assumir dois ou mais níveis distintos. Nestes casos utilizamos o conceito de variáveis *dummy*, isto é, variáveis que assumem apenas dois possíveis valores (usualmente 0 e 1) indicando a ausência ou presença de uma característica. Quando uma variável assume m níveis distintos, pode-se construir $m-1$ *dummies* para que não haja multicolinearidade entre a primeira coluna da matriz X e as *dummies* criadas (Draper & Smith 1998).

Durante o ajuste devemos selecionar quais regressores devem estar no modelo. Uma estratégia é utilizar o critério proposto por Schwarz (1978) denominado *Bayesian information criterion*, no qual é escolhido o modelo que apresentar menor BIC, dado por

$$BIC = \log \frac{SSE}{T} + \frac{k}{T} \log T,$$

em que SSE é a soma de quadrados dos resíduos do modelo, T é o número de observações e k é o número de parâmetros no modelo.

Após ajustar o modelo de regressão é interessante verificar a qualidade global do ajuste. Uma forma é utilizar o coeficiente de determinação

$$R^2 = \frac{SSR}{SST},$$

em que SSR é a soma de quadrados da regressão e SST é a soma de quadrados total. O R^2 é uma medida da capacidade explicativa do modelo de regressão e pertence ao intervalo (0,1). O R^2 pode ser interpretado como a proporção da variabilidade total explicada pelo modelo de regressão ajustado.

Outra medida de qualidade do ajuste é o coeficiente de determinação ajustado, denotado por $\bar{R}^2$, sendo esta uma medida que não necessariamente aumenta com o acréscimo de regressores no modelo (Draper & Smith 1998). Este coeficiente é útil para comparar ajustes e sua forma é dada por

$$\bar{R}^2 = 1 - \frac{SSE/(T-k)}{SST/(T-1)}.$$

Ao ajustar o modelo, pode-se detectar observações que destoam das demais. Quando a discrepância ocorre com respeito aos regressores, a observação é chamada de ponto de alavanca; nos casos em que a observação destoa em relação à variável reposta, esta é denominada *outlier*. Para determinar se uma observação é um ponto de alavanca, obtemos os elementos da diagonal principal da matriz "chapéu" $H = X(X'X)^{-1}X'$, em que X' é a transposta da matriz X. Caso alguma observação tenha valor maior que $3k/T$, dizemos que o ponto é de alavanca. Para verificar se o ponto t é um *outlier*, obtemos o resíduo studentizado, dado por

$$e_t^* = \frac{y_t - \underset{\sim}{x_t}' \underset{\sim}{b(t)}}{\hat{\sigma}(t)\sqrt{1 + \underset{\sim}{x_t}'[X(t)'X(t)]^{-1}\underset{\sim}{x_t}}},$$

em que y_t é a t-ésima observação da variável resposta, $\underset{\sim}{x_t}$ é um vetor contendo os valores das variáveis explicativas para a observação t, $\underset{\sim}{b(t)}$ representa a estimativa de β sem a t-ésima observação, $\hat{\sigma}(t)$ é a estimativa da variância dos

resíduos sem a observação t e $X(t)$ é a matriz X sem a t-ésima linha. Assumindo que os erros são normais, temos que e_t^* tem distribuição t_{T-k-1} e uma observação é considerada um *outlier* se e_t^* não pertence ao intervalo $[-2,2]$.

Ainda sobre pontos que destoam dos demais, há os pontos de influência, os quais alteram as estimativas dos parâmetros. Uma medida para identificar tais pontos é a Distância de Cook, que é dada por

$$c_t = \frac{\hat{e}_t^2 h_t}{\hat{\sigma}^2 k(1-h_t)^2},$$

em que $\hat{e}_t$ é o erro estimado para a t-ésima observação, h_t é o t-ésimo elemento da diagonal da matriz H e $\hat{\sigma}^2$ é a variância estimada do resíduo. Assim, a observação t é atípica se $c_t > \frac{8}{T-2k}$.

2.1. **Possíveis Problemas.**

A seguir são apresentados os possíveis problemas que podem surgir no decorrer do processo de modelagem.

2.1.1. *Multicolinearidade.*

O problema de multicolinearidade ocorre quando o posto$(X) < k$, em que X é a matriz modelo e k é o número de regressores (Montgomery et al. 2006). Dizemos que há multicolinearidade exata se $\exists\ \underset{\sim}{c} = (c_1,\ldots,c_k)' \neq \underset{\sim}{0}$ tal que

$$c_1\underset{\sim}{x_1} + c_2\underset{\sim}{x_2} + \cdots + c_k\underset{\sim}{x_k} = \underset{\sim}{0},$$

em que $\underset{\sim}{x_j}$ é a j-ésima coluna de X.

Dizemos que há multicolinearidade quase exata quando a equação acima vale de forma aproximada. Quando temos multicolinearidade exata a matriz $X'X$ é singular, logo não podemos obter $\underset{\sim}{b}$ de forma única. Além disso, é impossível obter os efeitos individuais, pois não podemos variar um regressor e deixar outro constante. Já na multicolinearidade quase exata conseguimos estimar esses efeitos, mas estas estimativas são imprecisas e tem alta variância, pois $X'X$ está próxima da singularidade.

Para detectar multicolinearidade quase exata dispomos de várias técnicas. Podemos utilizar, por exemplo, o coeficiente de correlação entre os regressores 2 a 2. Se em módulo esse valor for superior a 0,8 então há multicolinearidade quase exata. Outro método, conhecido como Medida de Theil, é expresso da seguinte forma

$$H = R^2 = \sum_{j=2}^{k} (R^2 - R_j^2),$$

em que R_j^2 é o R^2 da regressão sem o regressor x_j. Se os regressores forem ortogonais entre si, $H = 0$.

Para detectar tal multicolinearidade podemos usar também os fatores de inflação da variância (VIF), em que se algum dos elementos da matriz $(X_c' X_c)^{-1} > 5$, há multicolinearidade, sendo X_c a matriz dos regressores padronizados. Temos ainda um teste baseado nos autovalores de $X'X$, em que o número de condição se dá pela raiz quadrada da razão entre o maior e o menor autovalor desta matriz. Caso este número seja maior que 30, há multicolinearidade quase exata.

2.1.2. *Não normalidade.* Outro problema que pode acontecer ao utilizamos o método de mínimos quadrados ordinários é a violação de normalidade dos erros, necessária para estimação intervalar e testes de hipóteses. Para verificar tal pressuposto podemos utilizar um teste de normalidade. O teste de Bera-Jarque (Thadewald & Büning 2004) tem a seguinte estatística de teste

$$BJ = T\left(\frac{\hat{a}^2}{6} + \frac{(\hat{c}-3)^2}{24}\right),$$

em que $\hat{a}$ é o estimador da assimetria, $\hat{c}$ é o estimador da curtose e T é o tamanho da amostra. Sob a hipótese nula, esta estatística de teste tem distribuição aproximadamente χ^2_2.

2.1.3. *Heteroscedasticidade.* A mais comum das violações é a correspondente à suposição de homoscedasticidade dos erros. Para utilizar o método de mínimos quadrados ordinários utiliza-se a suposição de que os erros têm variância constante. Para testar tal suposição podemos usar o teste de Breush-Pagan-Godfrey (Breusch & Pagan 1979), mais conhecido como Teste de Breush-Pagan. Suponha que

$$\sigma_t^2 = h(\alpha_1 + \alpha_2 z_{t2} + \cdots + \alpha_s z_{ts}),$$

em que h é a função cedástica, sendo ela contínua e duas vezes diferenciável. Os z são variáveis que afetam as variâncias. Sendo

$$m_t = \hat{e}_t^2 - \tilde{\sigma}^2,$$

em que $\hat{e}_t = y_t - \underbrace{x_t b}$ e $\tilde{\sigma} = \frac{\hat{e}'\hat{e}}{T}$. Regressando $\underbrace{m}$ sobre Z, sendo SSR_a definido como a soma dos quadrados desta regressão auxiliar, temos a seguinte estatística de teste

$$LM_{BP} = \frac{SSR_A}{2},$$

em que, sob a hipótese de homoscedasticidade LM_{BP} distribui-se assintoticamente como χ^2_{s-1}. Uma limitação deste teste é a necessidade da suposição de normalidade, pois neste caso é utilizado o estimador para a variância. Para tentar resolver este problema Koenker (1981) propôs uma modificação para o teste de Breush-Pagan. Tal teste, que ficou conhecido como teste de Koenker, baseia-se em um estimador para variância que não depende de normalidade, tornado-o

poderoso com ou sem esta suposição. A estatística de teste é dada por

$$LM_k = TR_A^2,$$

em que R_A^2 é o R^2 da regressão auxiliar. Sob a hipótese de homoscedasticidade, LM_k distribui-se assintoticamente como uma χ^2_{s-1}.

2.1.4. *Especificação incorreta do modelo.* Ao estimarmos um modelo linear, partimos da pressuposição de que o modelo proposto está corretamente especificado. Para testar tal suposição podemos fazer uso do teste `Reset` (J.B.Ramsey 1969). Sendo a hipótese de nulidade de que o modelo linear está corretamente especificado, a ideia deste foi inserir regressores não lineares, de forma que a inclusão de tais regressores é testada. Caso o regressor não linear seja significativo para o modelo, há indícios de não linearidade, rejeitando assim a hipótese nula. Sendo $\hat{y} = X\underbrace{b}$, em geral podemos incluir como regressores não lineares: $\hat{y}^2$, $\hat{y}^3$, $\hat{y}^4$, x_j^2, $\widetilde{x_j^3}$, x_j^4, em que $j = 2, \ldots, k$. Para a exclusão destes regessores podemos usar o Teste F.

2.2. **Possíveis soluções.** Para tentar solucionar ou minimizar problemas de não normalidade e multicolinearidade quase exata podemos aplicar transformações de variáveis. Opcionalmente, no caso de multicolinearidade, podemos fazer uso da Regressão Ridge, em que substituímos a matriz $X'X$ por $X'X + K$, sendo K chamado de parâmetro de encolhimento. Essa substituição tem como objetivo resolver o problema da proximidade de $X'X$ da singularidade. No caso em que não utilizamos a regressão Ridge podemos aplicar algumas transformações nos regressores, como a transformação logarítmica, quadrática, cúbica, etc. Em hipótese alguma devemos excluir a variável que está causando a multicolinearidade quase exata. Pode-se ainda, caso faça sentido na prática, combinar tais variáveis.

No caso da não normalidade podemos tentar transformar a variável resposta, pois esta tem uma relação linear com os erros. Caso esta seja normal, os erros também serão. Uma transformação bastante utilizada em variáveis positivas é a transformação de Box-Cox (Box & Cox 1964). Ela é dada por

$$y^* = \begin{cases} \frac{y^{\lambda-1}}{\lambda} & , \text{se } \lambda \neq 0 \\ \log y & , \text{se } \lambda = 0 \end{cases},$$

em que o estimador de λ é obtido pelo método de máxima verossimilhança. Esta transformação reduz desvios de normalidade e heteroscedasticidade e é fácil de ser utilizada. Porém, além de só poder ser utilizada para variáveis positivas, para alguns valores de λ a variável obtida é limitada, além disso, dependendo também da escolha do parâmetro, a interpretação pode não ser trivial.

Para resolver o problema de heteroscedasticidade podemos fazer uso do estimador de mínimos quadrados generalizado viável. Porém, para tanto, precisamos estimar a matriz de covariâncias do estimador de forma consistente tanto sob homoscedasticidade como sob heteroscedasticidade, necessitando assim a estimação das T variâncias desconhecidas. O estimador da covariância de $\underset{\sim}{b}$ é expresso por $(X'X)^{-1}X'\hat{\Phi}X(X'X)^{-1}$, em que $\hat{\Phi}$ é a matriz de covariância estimada dos erros. Para tentar solucionar este problema, White (1980) teve a grande ideia de visualizar este estimador de forma diferente, em que precisaria estimar $X'\hat{\Phi}X$ ao invés de estimar apenas $\hat{\Phi}$. Dessa forma, não temos mais que estimar T variâncias, pois $X'\hat{\Phi}X$ tem dimensão k x k. O estimador proposto por White ficou conhecido como HC0, expresso da seguinte forma

$$HC0 = (X'X)^{-1}X'\hat{\Phi}X(X'X)^{-1},$$

em que $\hat{\Phi} = \text{diag}\{\hat{e}_1^2, \ldots, \hat{e}_T^2\}$.

Este estimador é consistente tanto sob homoscedasticidade como sob heteroscedasticidade, porém, em amostras finitas é viesado, tendendo a subestimar as variâncias verdadeiras. Surgiram então alguns estimadores baseados na grande ideia de White, visando melhorar esse problema. Assim surgiu o HC1 (MacKinnon & White 1985), em que a ideia foi multiplicar o estimador HC0 por $\frac{T}{T-k}$, porém esse estimador ainda tinha problemas, pois é atribuído o mesmo peso para todas as observações, sem levar em consideração a alavancagem. Outro estimador proposto foi o HC2 (MacKinnon & White 1985): $HC2 = (X'X)^{-1}X'\hat{\Phi}_2X(X'X)^{-1}$, de forma que

$$\hat{\Phi}_2 = \text{diag}\left\{\frac{\hat{e}_1^2}{1-h_1}, \ldots, \frac{\hat{e}_T^2}{1-h_T}\right\},$$

em que h_i é o i-ésimo elemento da matriz "chapéu", expressa por $X'(X'X)^{-1}X$. Davidson & MacKinnon (1993) propuseram outro estimador, conhecido como HC3: $HC3 = (X'X)^{-1}X'\hat{\Phi}_3X(X'X)^{-1}$, sendo

$$\hat{\Phi}_3 = \text{diag}\left\{\frac{\hat{e}_1^2}{(1-h_1)^2}, \ldots, \frac{\hat{e}_T^2}{(1-h_T)^2}\right\}.$$

Levando em consideração os altos pontos de alavancagem, Cribari-Neto (2004) propôs o HC4: $HC4 = (X'X)^{-1}X'\hat{\Phi}_4X(X'X)^{-1}$, de forma que

$$\hat{\Phi}_4 = \text{diag}\left\{\frac{\hat{e}_1^2}{(1-h_1)^{\delta_1}}, \ldots, \frac{\hat{e}_T^2}{(1-h_T)^{\delta_T}}\right\},$$

em que $\delta_t = \min\left\{\frac{h_t}{\bar{h}}, 4\right\}$.

Além disso, Cribari-Neto et al. (2007) observaram que os pontos de alta alavancagem afetam a variância dos demais, propondo assim HC5: $HC5 = (X'X)^{-1} X'\hat{\Phi}_5 X(X'X)^{-1}$, de forma que

$$\hat{\Phi}_5 = \text{diag}\left\{ \frac{\hat{e}_1^2}{(1-h_1)^{\delta_1}}, \ldots, \frac{\hat{e}_T^2}{(1-h_T)^{\delta_T}} \right\},$$

em que $\delta_t = \min\left\{\frac{h_t}{\bar{h}}, \max\left\{4, \frac{0.7T\text{hmáx}}{k}\right\}\right\}$.

Para proceder testes de hipóteses utilizamos os testes conhecidos como *quasi-t* ou *quasi-z*, em que substituímos o estimador consistente da covariância de $\underset{\sim}{b}$ sob homoscedasticidade, por um dos estimadores consistentes tanto sob homoscedasticidade quanto heteroscedasticidade propostos. Estudos de simulação em Cribari-Neto (2004) mostraram que, sob heteroscedasticidade, os estimadores HC3 e HC4 têm bom desempenho. Porém, na presença de pontos de alta alavancagem, HC4 é superior, além de apresentar bom desempenho em testes *quasi-t*, mesmo subestimando as variâncias e sendo viesado.

3. ANÁLISE DESCRITIVA

Com a finalidade de identificar as características de interesse das variáveis de estudo, foi utilizado um conjunto de técnicas para descrever e resumir os dados, estes correspondentes a preços de casas vendidas no Canadá, em 1987. Ao todo, foram analisadas 546 observações ($T = 546$), em que as variáveis em estudo estão dispostas na Tabela 1.

TABELA 1. Variáveis do estudo

Variável	Descrição
price	Preço de venda da casa
lotsize	Tamanho do lote do imóvel
bedrooms	Número de quartos
bathrms	Número de banheiros
stories	Número de pavimentos, excluindo o porão
driveway	Há entrada para carros
recroom	Há um quarto de recreação
fullbase	Há um porão totalmente construído
gashw	Uso de gás para aquecimento de água
airco	Há ar-condicionado central
garagepl	Número de vagas na garagem
prefarea	É localizada no bairro preferido da cidade

A variável preço de venda da casa corresponde à variável resposta e as demais são as variáveis explicativas do presente estudo. A Tabela 2 apresenta algumas

medidas descritivas com relação às variáveis preço de venda da casa e tamanho do lote do imóvel. Assim, é possível notar que o preço médio de venda das casas é de 68120 dólares canadenses, com desvio-padrão de 26702,67. Ainda observa-se um preço mediano de 62000 dólares canadenses. Quanto ao tamanho do lote, verifica-se que o tamanho médio foi de $463,50m^2$, apresentando desvio-padrão de $195,13m^2$ e tamanho mediano de $414m^2$. Vale ressaltar que originalmente os dados para o tamanho do lote estavam na unidade de pés quadrados, sendo posteriormente transformados em metros quadrados por meio de conversão adequada (1 pé quadrado = 0,09 metros quadrados).

TABELA 2. Principais medidas descritivas das variáveis relacionadas ao Preço de venda da casa e ao Tamanho do lote do imóvel

Medidas Descritivas	Preço de venda	Tamanho do lote
Mínimo	25000	148,50
1° quartil	49120	324,00
Mediana	62000	414,00
Média	68120	463,50
3° quartil	82000	572,40
Máximo	190000	1458,00
Desvio-padrão	26702,67	195,13

Para visualizar melhor os valores acima com relação ao preço de venda das casas, foi construído um box-plot e um histograma, ambos apresentados na Figura 1. A partir da análise desses gráficos, é possível perceber uma assimetria positiva nas observações, indicando que a curva da distribuição da variável resposta não é simétrica e possivelmente não segue uma distribuição normal. Para verificar tal informação, foi realizado o teste de Bera-Jarque a fim de verificar se os dados, com relação à variável preço de venda das casas, seguem uma distribuição normal. O referido teste apresentou valor de p menor do que 0,001, indicando a rejeição da hipótese nula de normalidade. A solução para a não-normalidade da variável resposta será dada logo mais, na Seção 4. Por meio do box-plot, ainda pode-se notar algumas observações mais destoantes das demais, caracterizando-as como *outliers*.

Na Tabela 3 são apresentadas as frequências absolutas e relativas das variáveis qualitativas que compõem o banco de dados. Assim, é possível inferir que 55,13% das casas tem 3 quartos, 73,63% possui 1 banheiro, em 85,90% há entrada para carros, 82,23% não tem quarto de recreação e 65,02% das casas não possui porão. Não há uso de gás para aquecimento de água em 95,42% das casas, nem ar-condicionado central em 68,32%. Também nota-se que em 54,94%

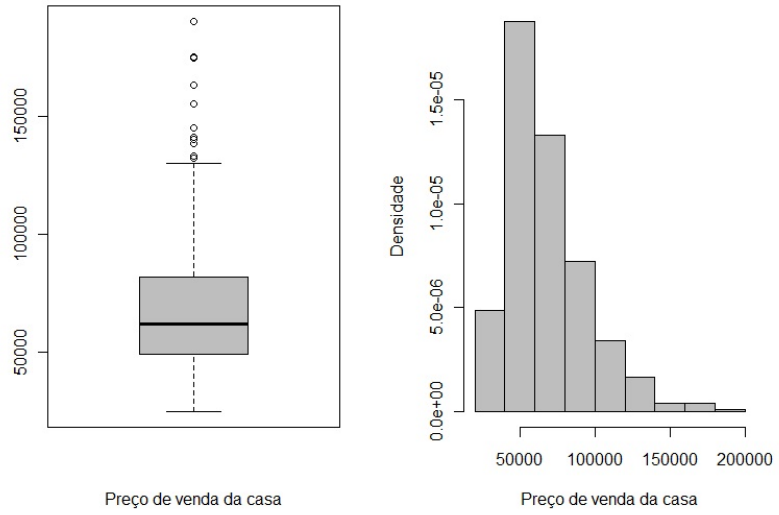

FIGURA 1. Box-plot e histograma da variável resposta Preço de venda da casa

das casas não há garagem e ainda 76,56% não estão localizadas no bairro preferido da cidade.

TABELA 3. Frequência das variáveis qualitativas

Variáveis	Frequência absoluta (N)	Frequência relativa (%)
Número de quartos		
1	2	0,37
2	136	24,90
3	301	55,13
4	95	17,40
5	10	1,83
6	2	0,37
Número de banheiros		
1	402	73,63
2	133	24,36
3	10	1,83
4	1	0,18
Número de pavimentos		
1	227	41,57
2	238	43,59
3	40	7,33
4	41	7,51
Há entrada para carros		
Sim	469	85,90
Não	77	14,10
Há um quarto de recreação		
Sim	97	17,77
Não	449	82,23
Há um porão		
Sim	191	34,98
Não	355	65,02
Uso de gás para aquecimento de água		
Sim	25	4,58
Não	521	95,42
Há ar-condicionado central		
Sim	173	31,68
Não	373	68,32
Número de vagas na garagem		
0	300	54,94
1	126	23,08
2	108	19,78
3	12	2,20
É localizada no bairro preferido da cidade		
Sim	128	23,44
Não	418	76,56

4. ANÁLISE DE REGRESSÃO

Esta seção apresenta uma análise de regressão, em que vários modelos foram testados a fim de encontrar um modelo final que melhor representasse a relação entre o preço de venda da casa e as demais covariadas do estudo. No decorrer desta etapa, foram realizados testes de normalidade, de homoscedasticidade, de especificação correta do modelo, além de análises de resíduo e diagnóstico, com a finalidade de detectar problemas com o ajuste proposto. Ainda fez-se necessário verificar a existência ou não de multicolinearidade entre as variáveis explicativas.

Inicialmente, foi feito um diagrama de dispersão entre a variável resposta preço de venda da casa e a variável explicativa quantitativa contínua tamanho do lote para detectar se há relação linear entre essas duas variáveis. Na Figura 2 é possível observar a relação linear positiva entre o preço de venda da casa e o tamanho do lote da mesma. Assim, à medida que o tamanho do lote aumenta, o preço de venda da casa tende a aumentar também.

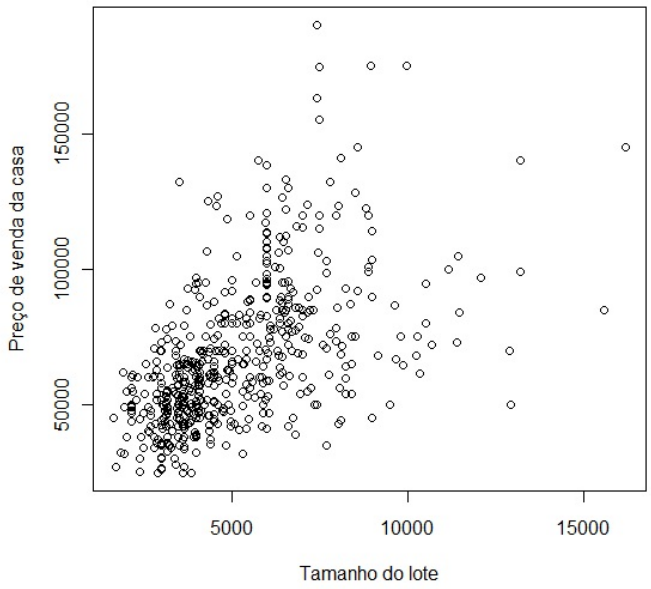

FIGURA 2. Diagrama de dispersão entre o preço de venda da casa e o tamanho do lote do imóvel

Em seguida, as variáveis número de quartos, números de banheiros, número de pavimentos e número de vagas na garagem foram transformadas em variáveis *dummy*. Vale ressaltar que categorias de algumas variáveis, por apresentar baixa frequência, foram agrupadas. Isso aconteceu com o número de quartos, que passou a ficar com três *dummies*, como também com a variável número de banheiros, esta com duas *dummies*.

Um primeiro problema que percebemos, ainda na Seção 3, foi que a variável resposta preço de venda da casa não segue uma distribuição normal. Com isso, nesta etapa da análise, fez-se necessário uma transformação com o objetivo de alcançar a normalidade. Para tal, foi aplicada a transformação de Box-Cox, considerando o modelo completo. O gráfico que contém o valor de $\hat{\lambda}$, este estimado pela maximização da função de verossimilhança, mostra que $\hat{\lambda} = 0$ está contido no intervalo com 95% de confiança. Assim, foi aplicada a transformação logarítmica na variável resposta, em que foi possível constatar a normalidade da variável transformada por meio do teste de normalidade de Jarque-Bera ($p = 0,3318$) e do histograma apresentado na Figura 3.

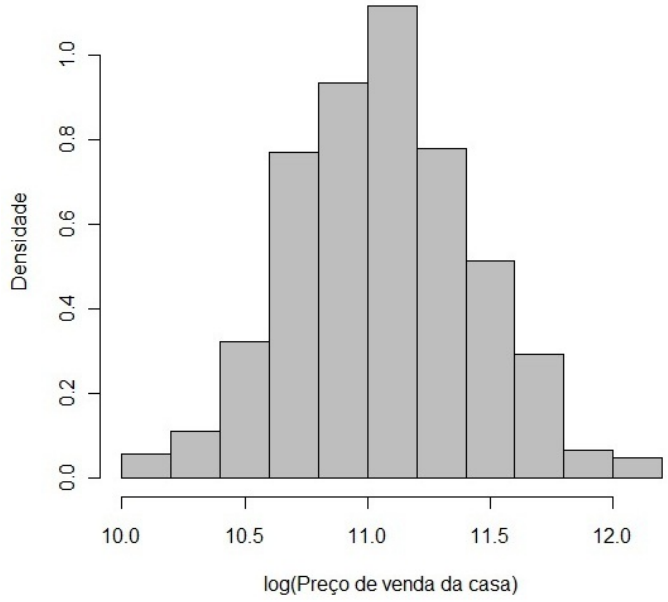

FIGURA 3. Histograma da variável resposta Preço de venda da casa transformada

Em seguida, foram ajustados dois modelos com todas as variáveis, o primeiro considerando a variável resposta transformada e o segundo, além da transformação no preço de venda da casa, ainda apresenta a variável explicativa tamanho do lote do imóvel transformada por meio da função logarítmica. Assim, considerando esses dois modelos, a fim de verificar a presença de multicolinearidade entre as variáveis regressoras, foi calculado os Fatores de Inflação da Variância (VIF) como forma de detectar a multicolinearidade. Em ambos os modelos, os valores de VIF foram menores do que 5, indicando que não indícios de multicolinearidade entre as variáveis regressoras. Em seguida, foi escolhido um dos dois modelos por meio do R^2 ajustado. O modelo completo 1, o qual considera somente a variável resposta transformada e o modelo completo 2, com as variáveis resposta preço de venda da casa e explicativa tamanho do lote do

imóvel transformadas, obtiveram R^2 ajustado no valor de 67,2% e 68,2%, respectivamente. Assim, optou-se por escolher o modelo completo 2 para dar continuidade à análise de regressão.

4.1. **Seleção do modelo final.** Para saber se interações dois a dois entre as variáveis explicativas deveriam estar presentes no modelo, foi utilizado o método BIC. O modelo proposto por este método apresentou resíduos não normais (valor de p obtido com o teste de Jarque-Bera: 0,036) e heteroscedásticos, este último confirmado por meio do valor de p do teste de Koenker dado por 0,005. Como este modelo ainda apresentou pontos de alta alavancagem (Figura 4), foram retiradas variáveis não significativas utilizando testes *quasi-t* com o estimador HC4. Deve-se ressaltar que o teste de normalidade aqui realizado foi desenvolvido sob a suposição de homoscedasticidade. Vale observar também que o método BIC apresentou duas interações entre as covariadas no seu modelo. Porém, estas não foram significativas ao nível de 5% na realização do teste.

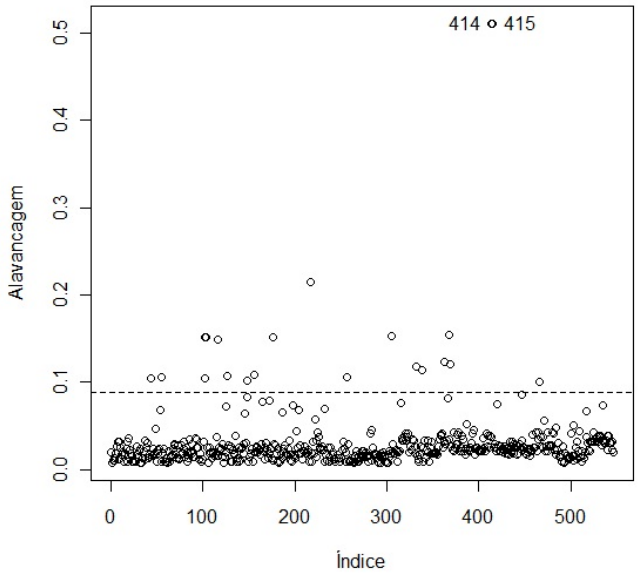

FIGURA 4. Pontos de alavanca do modelo proposto pelo método BIC

4.2. **Modelo Selecionado.** Após a retirada de variáveis não significativas por meio do teste *quasi-t*, foi escolhido o modelo dado por

$$\begin{aligned}\log(price)_t = {} & \beta_1 + \beta_2 \log(lotsize)_t + \beta_3 bathrms2_t + \beta_4 bathrms34_t + \beta_5 stories2_t + \beta_6 stories3_t \\ & + \beta_7 stories4_t + \beta_8 driveway_t + \beta_9 fullbase_t + \beta_{10} gashw_t + \beta_{11} airco_t \\ & + \beta_{12} garagepl1_t + \beta_{13} garagepl2_t + \epsilon_t,\end{aligned}$$

em que $t = 1, \ldots, 546$.

Para detectar a presença de possíveis *outliers*, estes caracterizados por destoar das demais observações com relação à variável resposta, foi utilizado o teste de Bonferroni, o qual assume em sua hipótese nula a não existência de *outliers* entre as observações. Com um valor de p de $0,321$, pode-se dizer que não há indícios da presença de *outliers* entre as observações.

Ainda foi de interesse verificar se o modelo escolhido estaria corretamente especificado. Para tal, utilizou-se o teste RESET de Ramsey, em que a hipótese de nulidade é dada pela afirmação de que o modelo está corretamente especificado. Com base no valor de p de $0,713$, é possível afirmar que o modelo selecionado está corretamente especificado.

As estimativas dos coeficientes do modelo de regressão final é apresentado na Tabela 4, assim como os respectivos valores de p, obtidos por meio de testes *quasi-t*, utilizando o estimador HC4.

TABELA 4. Estimativas dos coeficientes de regressão e respectivos valores de p (testes *quasi-t*)

Efeito	Estimativa	Valor de p
Intercepto	$8,601$	$< 0,001$
loglotsize	$0,340$	$< 0,001$
bathrms2	$0,182$	$< 0,001$
bathrms34	$0,357$	$< 0,001$
stories2	$0,105$	$< 0,001$
stories3	$0,265$	$< 0,001$
stories4	$0,298$	$< 0,001$
driveway	$0,121$	$< 0,001$
fullbase	$0,155$	$< 0,001$
gashw	$0,149$	$0,007$
airco	$0,170$	$< 0,001$
garagepl1	$0,077$	$0,002$
garagepl2	$0,122$	$< 0,001$

Com isso, percebe-se que quando o tamanho do lote aumenta em 1%, o preço médio da casa aumenta em $0,34\%$. Ainda, quando a casa tem 2 banheiros, a taxa de variação média no preço de venda é de $18,2\%$, em que esta aumenta com o acréscimo do número de banheiros na casa, ou seja, há $35,7\%$ de variação no preço de venda do imóvel quando este possui 3 ou 4 banheiros. Quanto ao número de pavimentos, uma casa que tem apenas 2 pavimentos ocasiona uma variação média de $10,5\%$ no preço, enquanto que em um imóvel que possui 3 e 4 pavimentos, essa taxa de variação aumenta para $26,5\%$ e $29,8\%$, respectivamente. Uma casa que possui entrada para carros faz o preço de venda variar, em média, $12,1\%$, enquanto o imóvel que possui o porão totalmente construído faz

esse preço variar em 15,5%. Casas que utilizam gás para aquecimento de água e possuem ar-condicionado central acometem uma variação média no preço de venda de 15,5% e 17,0%, respectivamente. Por fim, em relação ao número de garagens, tem-se que quando um imóvel tem 2 garagens, a taxa de variação média no preço é de 7,7%, aumentando essa taxa para 12,2% quando a casa possui 3 vagas de garagem.

Assim, entre os fatores encontrados que determinam o preço de venda de casas, percebe-se que aqueles de maior influência sob o preço são, nesta ordem: a presença de 3 ou 4 banheiros no imóvel, o tamanho do lote, ter 3 pavimentos e possuir 4 pavimentos excluindo o porão.

4.3. **Análise de Resíduos e Diagnóstico.** Os resíduos são interpretados como o desvio entre os valores observados e ajustados, sendo uma medida de variabilidade da variável resposta não explicada pelo modelo de regressão. Assim, a análise de resíduos é uma maneira eficaz de identificar vários tipos de inadequacidades do modelo. Entre os objetivos principais da análise de resíduos e diagnóstico estão verificar se há afastamentos sérios das suposições feitas para o modelo e detectar observações atípicas que destoam do conjunto (outliers, pontos de alavanca e influentes).

A suposição de normalidade dos erros foi inicialmente verificada por meio do teste de Jarque-Bera, em que foi rejeitada a hipótese nula de que os resíduos seguem uma distribuição normal ($p = 0,005$). Porém, de acordo com a Figura 5, na qual se pode observar o gráfico normal de probabilidades com envelope do modelo ajustado, não há indícios de afastamento da suposição de normalidade dos erros.

A partir da Figura 6, nota-se alguns pontos de alavanca e outliers, sendo ainda possível observar um comportamento aleatório dos resíduos (gráfico resíduos studentizados *versus* valores ajustados), indicando que não há indícios de rejeição da hipótese de homoscedasticidade, verificada também por meio do teste de Koenker (valor de $p = 0,083$).

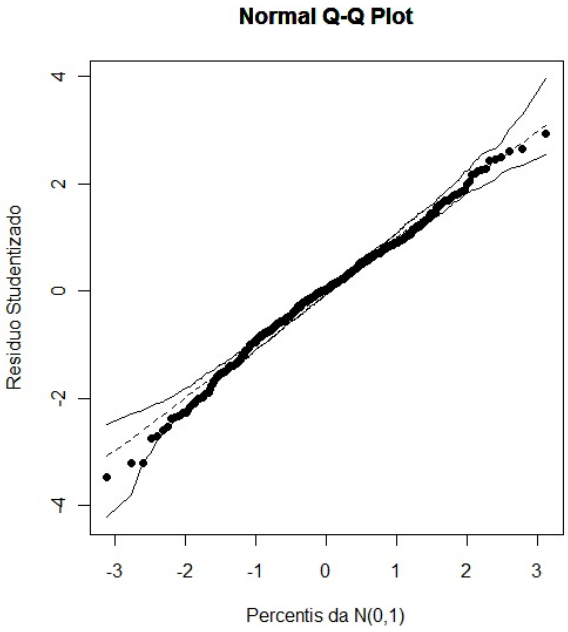

FIGURA 5. Gráfico normal de probabilidades com envelope do modelo ajustado sob erros normais

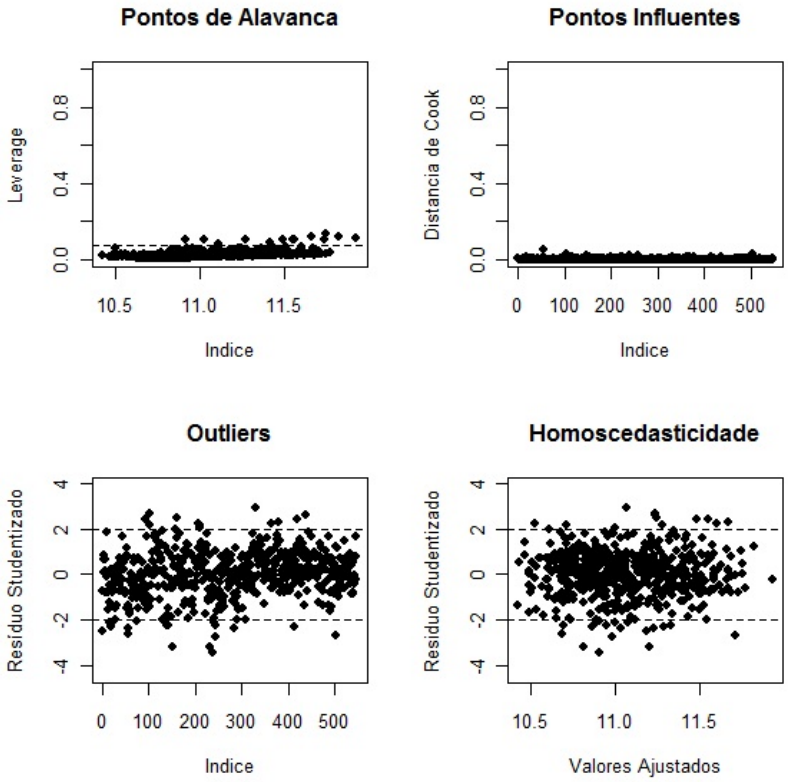

FIGURA 6. Gráficos para detecção de pontos de alavanca, influentes e outliers, assim como a suposição de homoscedasticidade dos erros

5. Considerações Finais

No presente trabalho, foi de interesse verificar quais fatores (variáveis explicativas/regressoras) poderiam influenciar no preço de venda das casas (variável resposta). Foi observada uma relação linear entre o preço de venda das casas e o tamanho do lote. Verificou-se ainda que a variável resposta não seguia uma distribuição normal, sendo então proposta primeiramente a transformação de Box-Cox para tentar corrigir tal problema. Assim, foi observado que apenas a transformação logarítmica na variável resposta foi suficiente para atender à normalidade dos dados.

Foi verificado que não há multicolinearidade entre as variáveis regressoras, fato este confirmado por meio dos Fatores de Inflação de Variância (VIF). Em seguida, foram utilizados alguns métodos de seleção de modelos, como R^2 ajustado, BIC, testes $quasi-t$ (usando o estimador HC4), teste RESET de especificação correta do modelo, entre outros. Foram excluídos alguns fatores do modelo completo por estes não mostrarem relevância. Vale observar que foram testadas interações dois a dois entre as covariáveis do modelo.

Assim, os resultados indicaram que os seguintes fatores determinam o preço de venda das casas: o tamanho do lote do imóvel, se a casa tem 2, 3 ou 4 banheiros, a presença de 2, 3 ou 4 pavimentos, possuir uma entrada para garagem, ter um porão totalmente construído, usar gás para aquecimento de água, possuir ar-condicionado central e ainda, haver no imóvel 1 ou 2 garagens.

Na análise de resíduos e diagnóstico, foi verificado que, embora, de acordo com o teste de normalidade dos resíduos, a suposição de normalidade dos erros tenha sido rejeitada, o gráfico normal de probabilidades com envelope do modelo ajustado não apresentou indícios contra a suposição de normalidade. Observou-se ainda a presença de pontos de alavanca e outliers e também a suposição de erros homoscedásticos atendida.

REFERÊNCIAS

Bailey, M. J., Muth, R. F. & Nourse, H. O. (1963), 'A regression method for real estate price index construction', *Journal of the American Statistical Association* **58**(304), 933–942.

Box, G. & Cox, D. (1964), 'An analysis of tranformations', *Journal of the Royal Statistical Society* **26**, 211–252.

Breusch, T. & Pagan, A. (1979), 'Simple test for heteroscedasticity and random coefficient variation', *Econometrica* **47**, 1287–1294.

Cribari-Neto, F. (2004), 'Asymptotic inference under heteroscedasticity of unknown form', *Computational Statistics & Data Analysis* **45**, 215–233.

Cribari-Neto, F., Souza, T. & Vasconcellos, K. (2007), 'Inference under heteroscedasticity and leveraged data', *Comunications in Statistics, Theory and Methods* **36**, 1877–1888.

Dantas, R. A. (1998), *Engenharia de avaliações*, Pini, São Paulo.

Davidson, R. & MacKinnon, J. (1993), 'Estimation and inference in econometrics', *New York: Oxford University Press* .

Draper, N. R. & Smith, H. (1998), *Applied regression analysis*, Vol. 3.

Isakson, H. R. (2001), 'Using multiple regression analysis in real estate appraisal', *Appraisal Journal* **69**(4), 424–430.

J.B.Ramsey (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society* **31**, 350–371.

Koenker, R. (1981), 'A note on studentizing a test got heteroscedasticity.', *Jornal of Econometrics* **17**, 107–112.

MacKinnon, J. & White, H. (1985), 'Some heteroscedasticity-consistent covariance matrix estimators with improved finite-sample properties', *Jor* **29**, 305–325.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2006), *Introduction to linear regression analysis*, 4th edn.

Moreira, D. S., dos Santos Silva, R. & da Rocha Fernandes, A. M. (2010), 'Engenharia de avaliações de imóveis apoiada em técnicas de análise multicritério e redes neurais artificiais', *Revista de Sistemas de Informação da FSMA* **6**, 49–58.

Ramsland, M. O. & Markham, D. E. (1998), 'Market-supported adjustments using multiple regression analysis', *Appraisal Journal* **66**(2), 181–191.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Steiner, M. T. A., Neto, A. C., Braulio, S. N. & Alves, V. (2008), 'Métodos estatísticos multivariados aplicados à engenharia de avaliações', *Gest. Prod.* **15**(1), 23–32.

Thadewald, T. & Büning, H. (2004), 'Jarque-bera test and its competitors for testing normality - a power comparison', *Institute for Statistics and Econometrics* .

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48 (4)**, 817–838.

APÊNDICE

Script do *software* R:

```
#############################################################################
# Universidade Federal de Pernambuco                                        #
# Centro de Ciências Exatas e da Natureza                                   #
# Mestrado e Estatística                                                    #
# Estatística Aplicada                                                      #
#                                                                           #
# Equipe:                                                                   #
#                                                                           #
# Ana Hermínia Andrade e Silva                                              #
# Heloisa de Melo Rodrigues                                                 #
# Sadraque Eneas de Figueiredo Lucena                                       #
#############################################################################

dados = read.table("dados.txt", header=TRUE)
dados

names(dados)
attach(dados)



################################################
BIBLIOTECAS
################################################
library(tseries)
library(MASS)
library(lmtest)
library(car)
library(sandwich)



################################################
    ANÁLISE DESCRITIVA
################################################
# Descritiva da variável resposta
summary(price)
sd(price)
par(mfrow=c(1,2))
boxplot(price, col="gray", xlab="Preço de venda da casa")
hist(price, freq=FALSE, col="gray", xlab="Preço de venda da casa",
ylab="Densidade", main="")
jarque.bera.test(price) # a variável resposta não é normal

# Descritiva da variável explicativa quantitativa
# Transformando a variável lotsize de pés quadrados para metros quadrados
```

```
lotsize = 0.09*lotsize
summary(lotsize)
sd(lotsize)

# Tabela de frequência das outras variáveis explicativas
table(bedrooms)
table(bathrms)
table(stories)
table(driveway)
table(recroom)
table(fullbase)
table(gashw)
table(airco)
table(garagepl)
table(prefarea)

# Gráfico de dispersão entre a variável resposta e a variável contínua 'lotsize'
plot(lotsize, price, xlab="Tamanho do lote", ylab="Preço de venda da casa")

# Construindo as variáveis dummy
# Como a variável bedrooms apresenta poucas observações nas categorias 1 e 6,
# agrupamos as categorias 1 e 2, assim como as categorias 5 e 6
bedrooms12 = ifelse(bedrooms==1 | bedrooms==2, 1, 0)
bedrooms3 = ifelse(bedrooms==3, 1, 0)
bedrooms4 = ifelse(bedrooms==4, 1, 0)
bedrooms56 = ifelse(bedrooms==5 | bedrooms==6, 1, 0)
bathrms2 = ifelse(bathrms==2, 1, 0)
# Como a categoria bathrms = 4 só tem uma observação, agrupamos bathrms = 3 e = 4
bathrms34 = ifelse(bathrms==3 | bathrms==4, 1, 0)
stories2 = ifelse(stories==2, 1, 0)
stories3 = ifelse(stories==3, 1, 0)
stories4 = ifelse(stories==4, 1, 0)
garagepl1 = ifelse(garagepl==1, 1, 0)
garagepl2 = ifelse(garagepl==2, 1, 0)
garagepl3 = ifelse(garagepl==3, 1, 0)

bedrooms12 = factor(bedrooms12, labels=c("Não", "Sim"))
bedrooms3 = factor(bedrooms3, labels=c("Não", "Sim"))
bedrooms4 = factor(bedrooms4, labels=c("Não", "Sim"))
bedrooms56 = factor(bedrooms56, labels=c("Não", "Sim"))
bathrms2 = factor(bathrms2, labels=c("Não", "Sim"))
bathrms34 = factor(bathrms34, labels=c("Não", "Sim"))
stories2 = factor(stories2, labels=c("Não", "Sim"))
stories3 = factor(stories3, labels=c("Não", "Sim"))
stories4 = factor(stories4, labels=c("Não", "Sim"))
garagepl1 = factor(garagepl1, labels=c("Não", "Sim"))
garagepl2 = factor(garagepl2, labels=c("Não", "Sim"))
garagepl3 = factor(garagepl3, labels=c("Não", "Sim"))
```

```
################################################
    AJUSTE DO MODELO
################################################
# Problema: a variável resposta não é normal
# Possível solução: Transformação de Box-Cox
BC = boxcox(price~lotsize + bedrooms3 + bedrooms4 + bedrooms56 + bathrms2 +
bathrms34 + stories2 + stories3 + stories4 + driveway + recroom + fullbase + gashw +
airco + garagepl1 + garagepl2 + garagepl3 + prefarea)
lambda = BC$x[BC$y==max(BC$y)] ##dá o lambda##
price.bc = (price^(lambda)-1)/(lambda)
modelo.bc  =  lm(price.bc~bedrooms3 + bedrooms4 + bedrooms56 + bathrms2 + bathrms34 +
stories2 + stories3 + stories4 + driveway + recroom + fullbase + gashw + airco +
garagepl1 + garagepl2 + garagepl3 + prefarea)
summary(modelo.bc)
# como IC 95% de lambda contem 0, usaremos a transformação log na variável resposta

# Aplicando a transformação log na variável resposta
logprice = log(price)
hist(logprice, freq=FALSE, col="gray", xlab="log(Preço de venda da casa)",
ylab="Densidade", main="")
jarque.bera.test(logprice) # com o log, a var resposta ficou normal


# Verificando se há multicolinearidade
# Matriz contendo apenas os regressores
modelo_completo1 = lm(logprice~lotsize + bedrooms3 + bedrooms4 + bedrooms56 +
bathrms2 + bathrms34 + stories2 + stories3 + stories4 + driveway + recroom +
fullbase + gashw + airco + garagepl1 + garagepl2 + garagepl3 + prefarea)
# R^2 ajustado
R2_1 = summary(modelo_completo1)$adj.r.squared;R2_1
# VIF
vif(modelo_completo1)

# Aplicando a transformação log na variável explicativa contínua lotsize
loglotsize = log(lotsize)
modelo_completo2 = lm(logprice~loglotsize + bedrooms3 + bedrooms4 + bedrooms56 +
bathrms2 + bathrms34 + stories2 + stories3 + stories4 + driveway + recroom + fullbase
+ gashw + airco + garagepl1 + garagepl2 + garagepl3 + prefarea)
# R^2 ajustado
R2_2 = summary(modelo_completo2)$adj.r.squared;R2_2
# VIF
vif(modelo_completo2)

# modelo selecionado até agora: modelo_completo2, por possuir maior R^2 ajustado
```

```
###############################################################################################
# Testando modelo completo com todas as interações
modelo.BIC = stepAIC(modelo_completo2,scope=list(lower = ~1,upper = ~.^2),k=log(546))

# Teste de heteroscedasticidade - teste de Koenker
bptest(modelo.BIC)

jarque.bera.test(modelo.BIC$resid)

# Identificando se há pontos de alavanca
chapeu=hatvalues(modelo.BIC)
plot(chapeu, xlab="Índice", ylab="Alavancagem", main="")
X = model.matrix(modelo.BIC)
cut = 3*(ncol(X))/(nrow(X))
abline(cut,0,lty=2)
identify(chapeu, n=2)

influence.measures(modelo.BIC) ##Medidas: DFBETA, DFFITS, D. de Cook e cov.ratio.
summary(influence.measures(modelo.BIC))

# Como o modelo.BIC é heteroscedástico e ainda há pontos de alavanca, usaremos o
# estimador HC4 para testar a significância dos parâmetros sob heteroscedasticidade
coeftest(modelo.BIC, vcov=vcovHC (modelo.BIC, type="HC4"))

# Seleção do modelo final
# tirando prefarea e interação de prefarea com driveway
modelo1 = lm(logprice~loglotsize + bathrms2 + bathrms34 + stories2 +stories3 +
stories4 + driveway + fullbase + gashw + airco + garagepl1 + garagepl2 + gashw*garagepl2)
coeftest(modelo1, vcov=vcovHC (modelo1, type="HC4"))

# tirando a interação de gashw com garagepl2
modelo2 = lm(logprice~loglotsize + bathrms2 + bathrms34 + stories2 + stories3 +
stories4 + driveway + fullbase + gashw + airco + garagepl1 + garagepl2)
coeftest(modelo2, vcov=vcovHC (modelo2, type="HC4"))



# Modelo final: modelo2
# Teste de heteroscedasticidade - Teste de Koenker
bptest(modelo2)

# Normalidade dos residuos do modelo
# Teste de normalidade dos resíduos
# H0: normalidade dos resíduos
jarque.bera.test(modelo2$resid)
par(mfrow=c(1,2))
hist(modelo2$resid, freq=FALSE, col="gray", xlab="Resíduos do modelo",
ylab="Densidade", main="")
```

```
qqnorm(modelo2$resid);qqline(modelo2$resid, col=2)

# Identificando se há pontos de alavanca
chapeu2=hatvalues(modelo2)
plot(chapeu2, xlab="Índice", ylab="Alavancagem", main="")
X2 = model.matrix(modelo2)
cut2 = 3*(ncol(X2))/(nrow(X2))
abline(cut2,0,lty=2)

influence.measures(modelo2) ##Medidas: DFBETA, DFFITS, D. de Cook e cov.ratio.
summary(influence.measures(modelo2))

# Teste de presença de outliers
#H0: não há outliers
outlier.test(modelo2)

# Teste de especificidade do modelo
# H0: o modelo está corretamente especificado
resettest(modelo2)



##### FUNÇÃO DIAG #####
fit.model = modelo2
lms = summary(fit.model)
X = model.matrix(fit.model)
n = nrow(X)
p = ncol(X)
H = X%*%solve(t(X)%*%X)%*%t(X)
h = diag(H)
lms = summary(fit.model)
s = lms$sigma
r = resid(lms)
ts = r/(s*sqrt(1-h))
di = (1/p)*(h/(1-h))*(ts^2)
si = lm.influence(fit.model)$sigma
tsi = r/(si*sqrt(1-h))
a = max(tsi)
b = min(tsi)
par(mfrow=c(2,2))
#PONTOS DE ALAVANCA
plot(fitted(fit.model),h,xlab="Indice", ylab="Leverage", pch=16,
main="Pontos de Alavanca", ylim=c(0,1))
cut = 3*p/n
abline(cut,0,lty=2)
#identify(fitted(fit.model),h, n=2)
#PONTOS INFLUENTES
plot(di,xlab="Indice", ylab="Distancia de Cook", pch=16,
main="Pontos Influentes", ylim=c(0,1))
```

```
#identify(di, n=2)
#OUTLIERS
plot(tsi,xlab="Indice", ylab="Resíduo Studentizado",
ylim=c(b-1,a+1), pch=16, main="Outliers")
abline(2,0,lty=2)
abline(-2,0,lty=2)
#identify(tsi, n=3)
#par(mfrow=c(1,1))
#Homoscedasticidade
plot(fitted(fit.model),tsi,xlab="Valores Ajustados",
ylab="Residuo Studentizado", ylim=c(b-1,a+1), pch=16,
main="Homoscedasticidade")
abline(2,0,lty=2)
abline(-2,0,lty=2)
#identify(fitted(fit.model),tsi, n=3)

##### ENVELOPE DA NORMAL #####
par(mfrow=c(1,1))
X = model.matrix(fit.model)
n = nrow(X)
p = ncol(X)
H = X%*%solve(t(X)%*%X)%*%t(X)
h = diag(H)
si = lm.influence(fit.model)$sigma
r = resid(fit.model)
tsi = r/(si*sqrt(1-h))
#
ident = diag(n)
epsilon = matrix(0,n,100)
e = matrix(0,n,100)
e1 = numeric(n)
e2 = numeric(n)
#
for(i in 1:100){
     epsilon[,i] = rnorm(n,0,1)
     e[,i] = (ident - H)%*%epsilon[,i]
     u = diag(ident - H)
     e[,i] = e[,i]/sqrt(u)
     e[,i] = sort(e[,i]) }
#
for(i in 1:n){
     eo = sort(e[i,])
     e1[i] = (eo[2]+eo[3])/2
     e2[i] = (eo[97]+eo[98])/2 }
#
med = apply(e,1,mean)
faixa = range(tsi,e1,e2)
#
```

```
par(pty="s")
qqnorm(tsi,xlab="Percentis da N(0,1)",
ylab="Residuo Studentizado", ylim=faixa, pch=16)
par(new=T)
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(med,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=2)
```