Trabalho Final - Análise de Regressão

Thiago Tavares Lopes

11 julho 2024

Sumário

1 Seleção do modelo

Nesta seção temos a construção do primeiro modelo no qual foi ajustado com todas as variáveis para explicação da variável resposta *price*.

1.1 Variáveis do estudo

O estudo foi direcionado a dados referente a informações de imóveis e suas características o seu valor de mercado. Como mencionando anteriormente a construção do modelo visa verificar como algumas características específicas podem influenciar no preço do imóvel. O banco de dados em questão pode ser acessado no link, para fins de um melhor controle do ajuste do modelo, o banco de dados foi reduzido para 110 observações. Na tabela 1, temos a apresentação das variáveis em estudo.

Tabela 1: Variável do estudo sobre o preço de imóveis

Variável	**Descrição**
Price	Preço do imóvel
Bathrooms	Número de banheiros no imóvel
Bedrooms	Número de quartos no imóvel
Floors	Número de andares no imóvel
Sqft_living	Metragem quadrada do imóvel
Sqft_above	Metragem do imóvel (sem o porão)
Sqft_lot	Metragem quadrada do terreno
Yr_built	Ano de construção do imóvel
Waterfront	Vista para o mar

Já na figura 1, é apresentado o gráfico de correlação par as variáveis em estudo. Neste caso, podemos verificar uma correlação positiva moderada (0,443185) entre o número de banheiros o preço do imóvel, também temos uma correlação positiva (0,652976) forte para imóveis com maior metragem quadrada, o que leva a crer que área construída pode ser uma uma variável muito importante para o modelo. Por último, há correlação negativa muito fraca (-0,060581) com imóveis mais antigos tendem a ser ligeiramente menos caros.

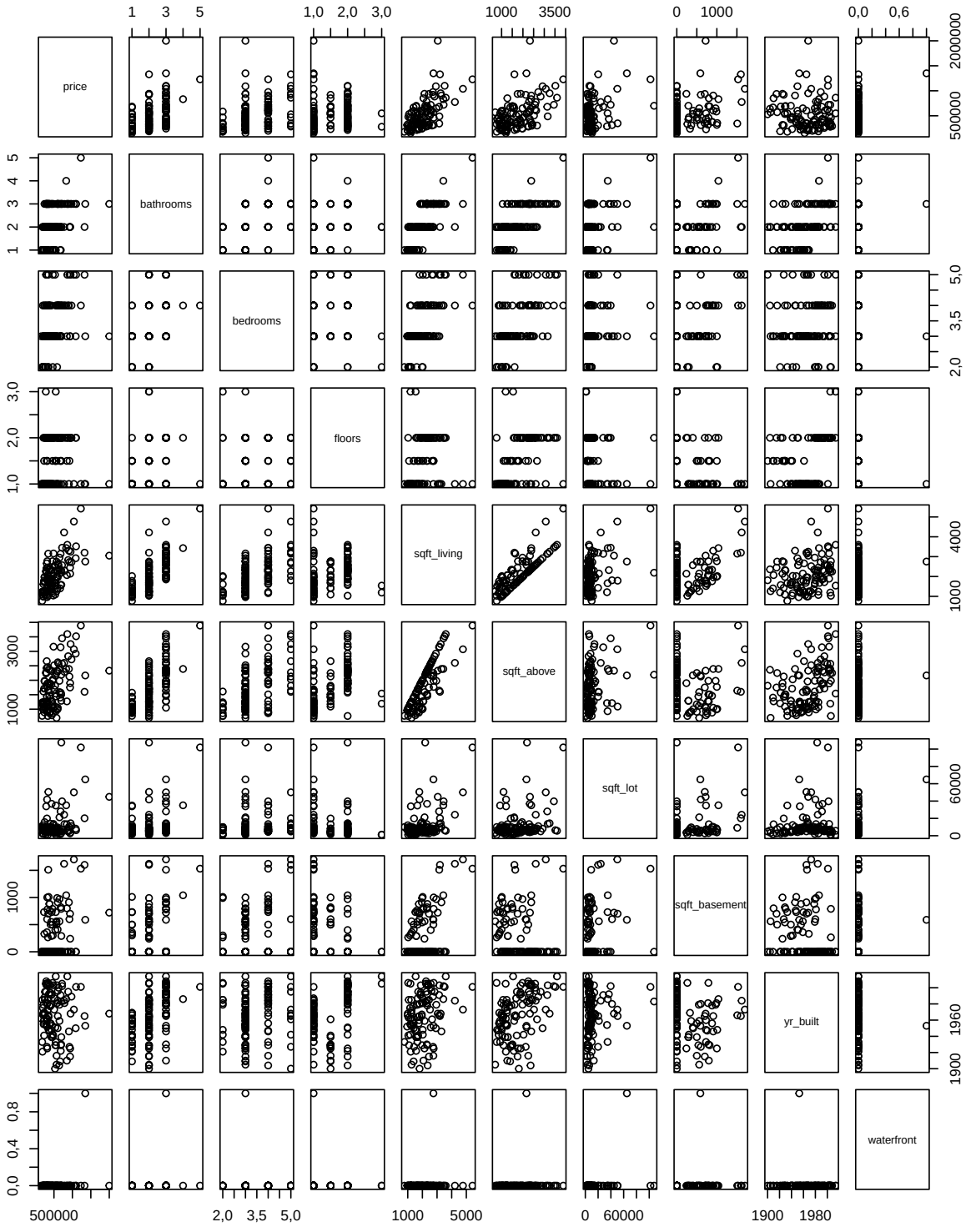


Figura 1: Gráfico de dispersão

Na tabela 2 é apresentado o resultado da análise descritiva das variáveis em estudo, como média, mediana, 1° e 3° quartil, além dos valores de máximo e mínimo. Logo, podemos indicar o primeiro modelo ajustado com todas as variáveis:

$$\hat{y} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x9$$

em que β_1 é o intercepto do modelo de regressão linear, y é a variável desfecho *price*, e o vetor das covariáveis é $(x_2, x_3, x_4, x_5, x_6, x_7, x_8)^T$=(bathrooms, bedrooms, bloors, sqft_living, sqft_above, sqft_lot, yr_built, waterfront).

Tabela 2: Análise descritiva

price	bathrooms	bedrooms	floors	sqft_living	sqft_above	sqft_lot	sqft_basement	yr_built	waterfront
Min. : 153000	Min. :1,00	Min. :2,00	Min. :1,00	Min. : 770	Min. : 700	Min. : 1044	Min. : 0	Min. :1900	Min. :0,00000
1st Qu.: 309250	1st Qu.:2,00	1st Qu.:3,00	1st Qu.:1,00	1st Qu.:1410	1st Qu.:1182	1st Qu.: 5000	1st Qu.: 0	1st Qu.:1947	1st Qu.:0,00000
Median : 444000	Median :2,00	Median :3,00	Median :1,50	Median :1970	Median :1625	Median : 7208	Median : 0	Median :1968	Median :0,00000
Mean : 523678	Mean :2,19	Mean :3,38	Mean :1,48	Mean :2073	Mean :1764	Mean : 12685	Mean : 309	Mean :1966	Mean :0,00909
3rd Qu.: 662750	3rd Qu.:3,00	3rd Qu.:4,00	3rd Qu.:2,00	3rd Qu.:2557	3rd Qu.:2305	3rd Qu.: 10189	3rd Qu.: 597	3rd Qu.:1989	3rd Qu.:0,00000
Max. :2000000	Max. :5,00	Max. :5,00	Max. :3,00	Max. :5420	Max. :3890	Max. :107593	Max. :1700	Max. :2014	Max. :1,00000

Para o primeiro modelo, apenas *sqft_living*, *yr_built*, *waterfront* foram significativas para a variável resposta. Por conseguinte, temos que o coeficiente de determinação (R^2) foi de 0,547 e o R^2 ajustado ($\bar{R^2}$)foi de 0,511, então podemos concluir que esse modelo explica, aproximadamente, 54,7% da variação de y. Em sequência, foi utilizado o método *Stepwise* que usa o critério de informação de Akaike (AIC), no qual o melhor modelo é selecionado e exclui-se as possíveis covariáveis.

Tabela 3: Resultados da Regressão Linear

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	6564399,8666	1710593,4199	3,83750	0,00022	***
bathrooms	35791,4352	42103,8017	0,85008	0,39729	
bedrooms	-13307,9357	32486,0623	-0,40965	0,68293	
floors	42259,9564	56323,6340	0,75031	0,45481	
sqft_living	218,3433	59,2017	3,68813	0,00037	***
sqft_above	5,6021	67,9407	0,08246	0,93445	
sqft_lot	2,0192	1,4506	1,39201	0,16698	
yr_built	-3372,2212	883,9489	-3,81495	0,00023	***
waterfront	517352,4755	225226,7896	2,29703	0,02368	*

[a] Nota: *** p<0.001; ** p<0.01; * p<0.05; . p<0.1

Um novo modelo foi construído baseado no resultado do critério de informação de Akaike, o qual podemos observar abaixo:

$$\hat{y} = \beta_1 + \beta_5 x_5 + \beta_8 x_8 + \beta_9 x_9$$

Temos que a variável *price* será explicada pelas variáveis $sqft\_living$, yr_built, $waterfront$. A tabela 4, apresenta os resultados referente ao modelo, temos que as três variáveis continuam significativas para o modelo, porém *waterfront* se tornou mais significativa. Por conseguinte, o coeficiente de determinação (R^2) teve uma pequena diminuição para 0,532 e o R^2 ajustado ($\bar{R^2}$) aumentou para 0,519. Por último, ao verificar as suposições do modelo em questão, as suposições $[S_0]$, $[S_3]$ e $[S_5]$ não foram atendidas. Então, foi necessário realizar a análise de diagnóstico e influência.

Tabela 4: Resultados da Regressão Linear

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	5267421,33	1421022,515	3,7068	0,00034	***
sqft_living	257,52	25,246	10,2007	0,00000	***
yr_built	-2687,41	729,990	-3,6814	0,00037	***
waterfront	622120,78	210835,128	2,9508	0,00390	**

[a] Nota: *** p<0.001; ** p<0.01; * p<0.05; . p<0.1

1.2 Análise de diagnóstico e influência

Nesta seção temos a análise diagnóstico para o segundo modelo com apenas três covariáveis. No primeiro teste de influência temos a verificação de pontos de alavancagem que busca identificar pontos atípicos no domínio das variáveis explicativas do modelo. No gráfico 2 se refere a alavancagem, no qual avaliamos possíveis ponto influentes para o modelo. Nota-se que as observações 6, 48, e 69 estão acima do limite estabelecido, porém ainda se faz necessário verificar se de fato essas observações são influentes.

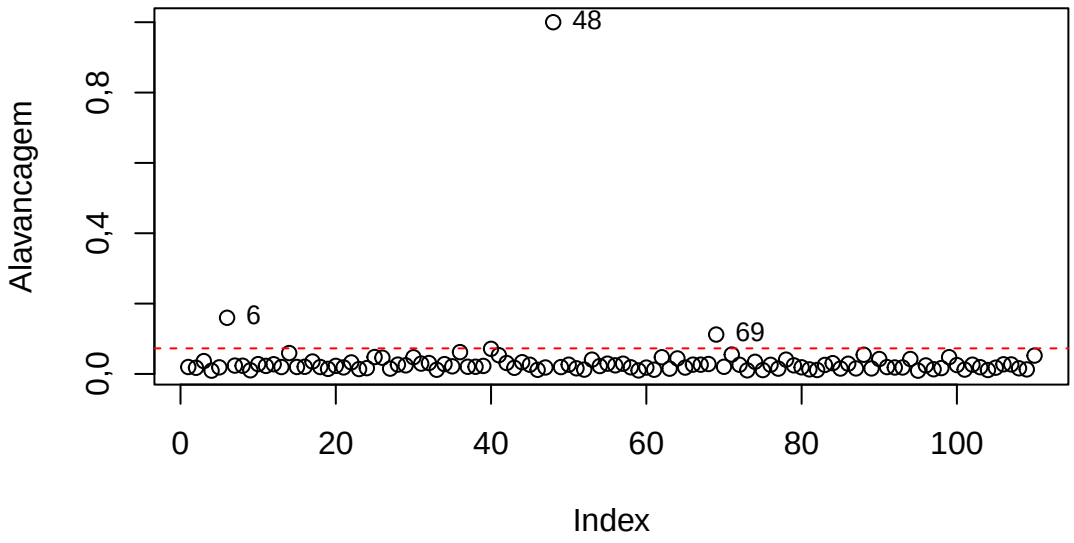


Figura 2: Gráfico de Alavancagem

Já o gráfico 3 mostra o DFFITs e leva em consideração o quando cada observação influência no valor estimado para $\hat{y}$. Podemos notar que desta vez temos as observações 20, 64, 68 estão sendo influentes para $\hat{y}$, principalmente a observação 20.

Na sequência temos os gráficos dos DFBETA para β do novo modelo ajustado. Podemos notar que a observação 20 aparece tanto $\beta_5, \beta_8, \beta_9$. O que pode ser mais um indicativo que essa observação é influente. Para verificar tal questão, mais dois testes foram feitos, são eles distância de Cook e gráfico dos resíduos.

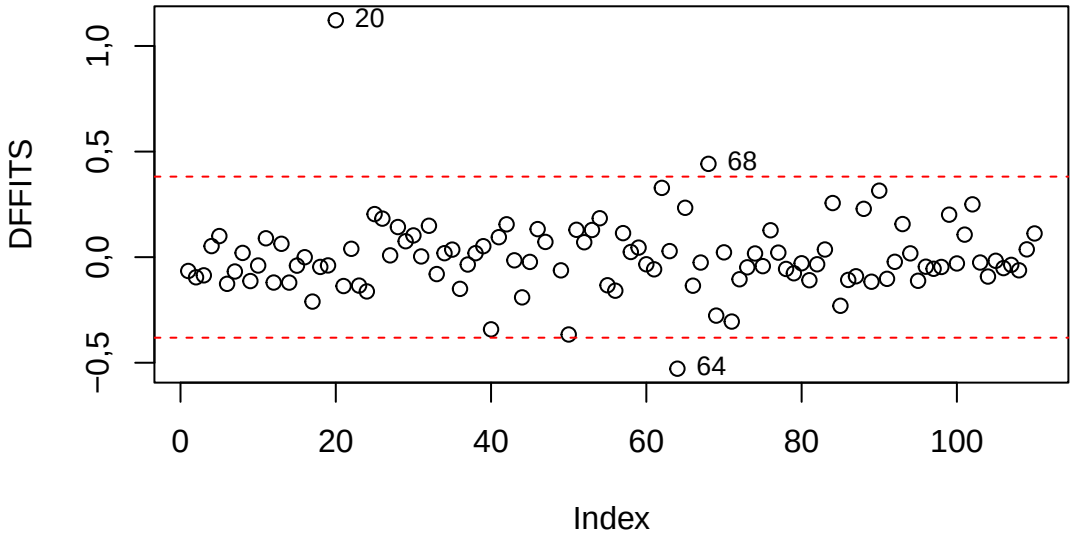


Figura 3: Gráfico DFFIT

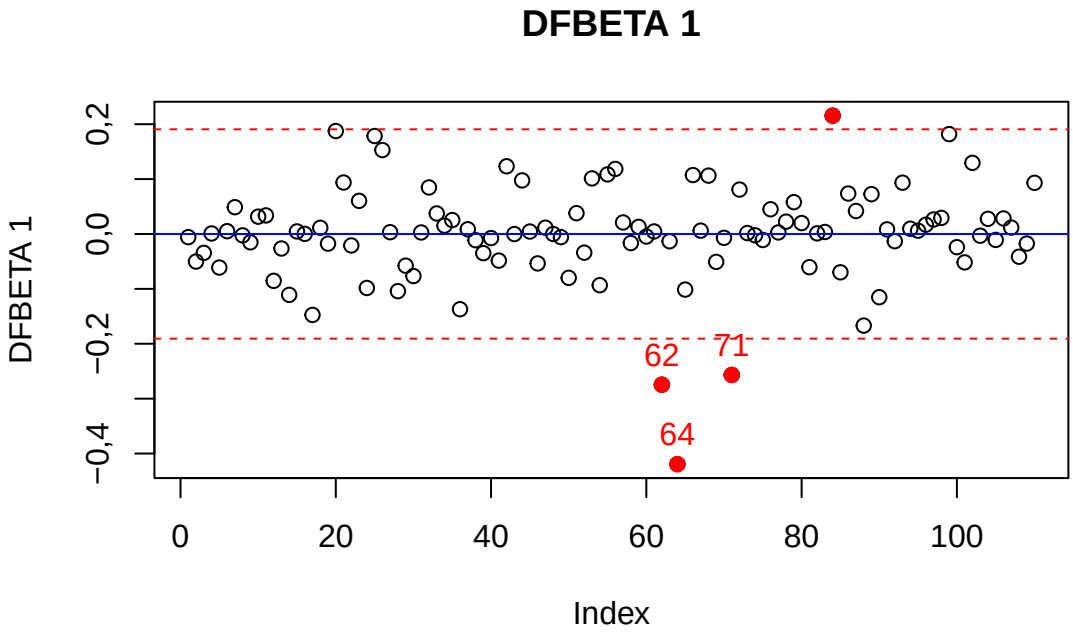


Figura 4: DFBETA 1

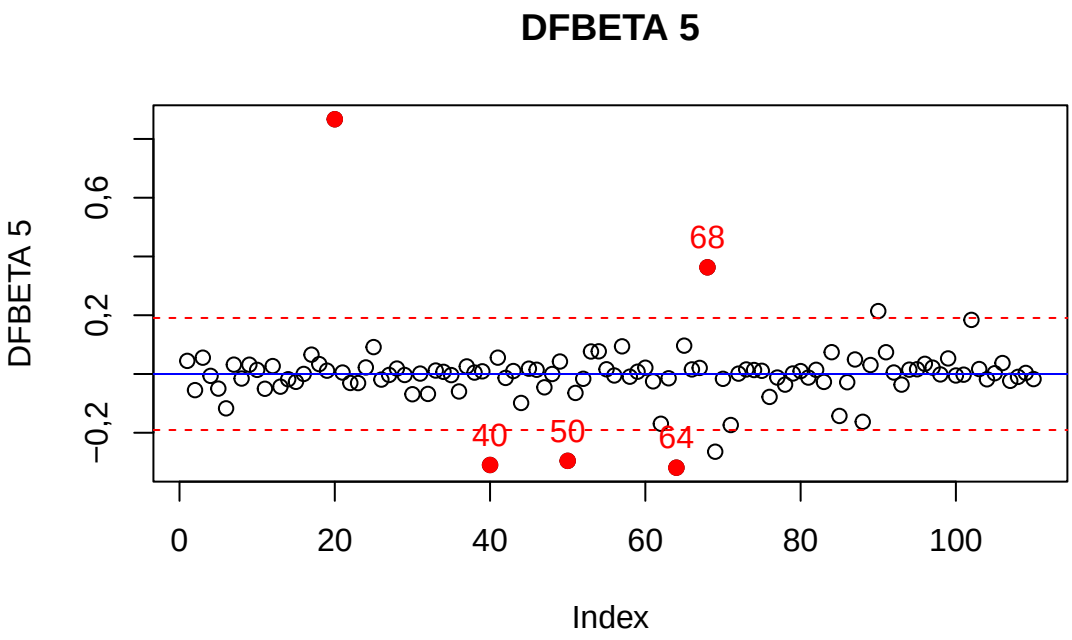


Figura 5: DFBETA 5

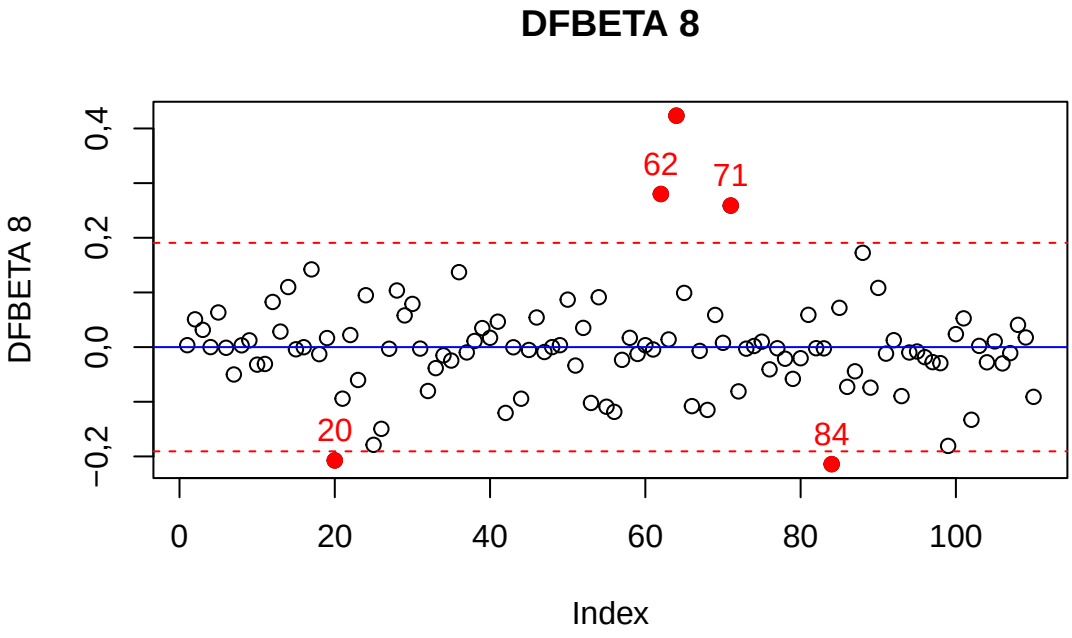


Figura 6: DFBETA 8

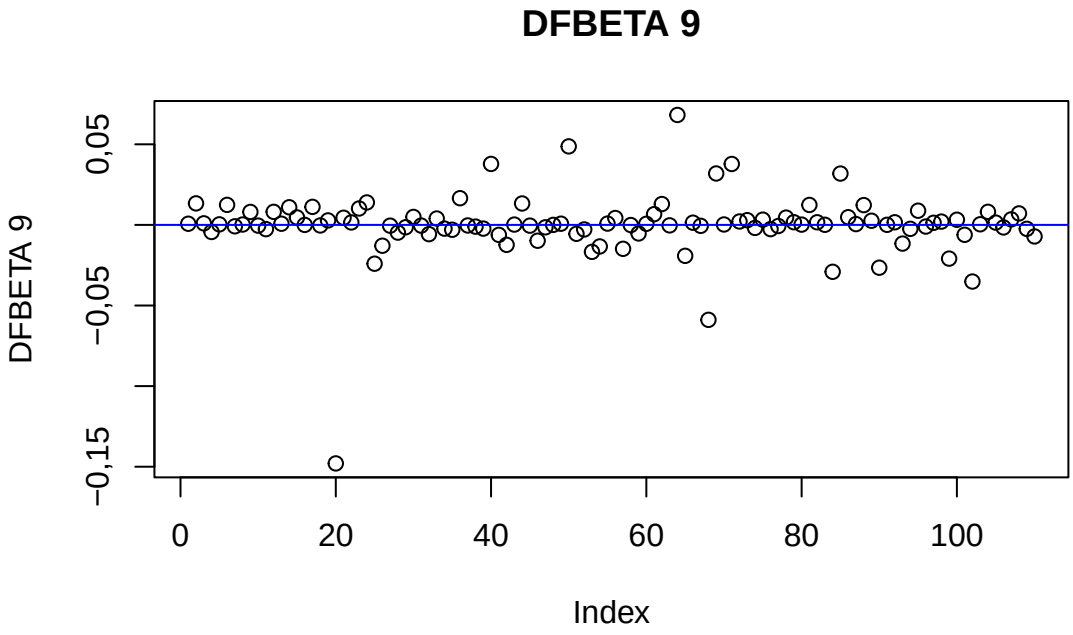


Figura 7: DFBETA 9

No gráfico 8 temos a distância de Cook, e novamente a observação 20 aparece como discrepante e fora do limite estabelecido.

O mesmo ocorre no gráfico de resíduos 8, temos a observação 20 sendo apontada como distante dos limites estabelecidos, ou seja, podemos considerar que temos um ponto de influência no modelo. Portanto, torna-se necessário a remoção desta observação do banco de dados.

No gráfico 9 temos o histograma dos resíduos, nota-se que o mesmo está bem diferente do que se espera de uma distribuição normal. Já no gráfico **??** temos o envelope simulado e podemos notar que os resíduos estão fora das bandas.

```
## Gaussian model (lm object)
```

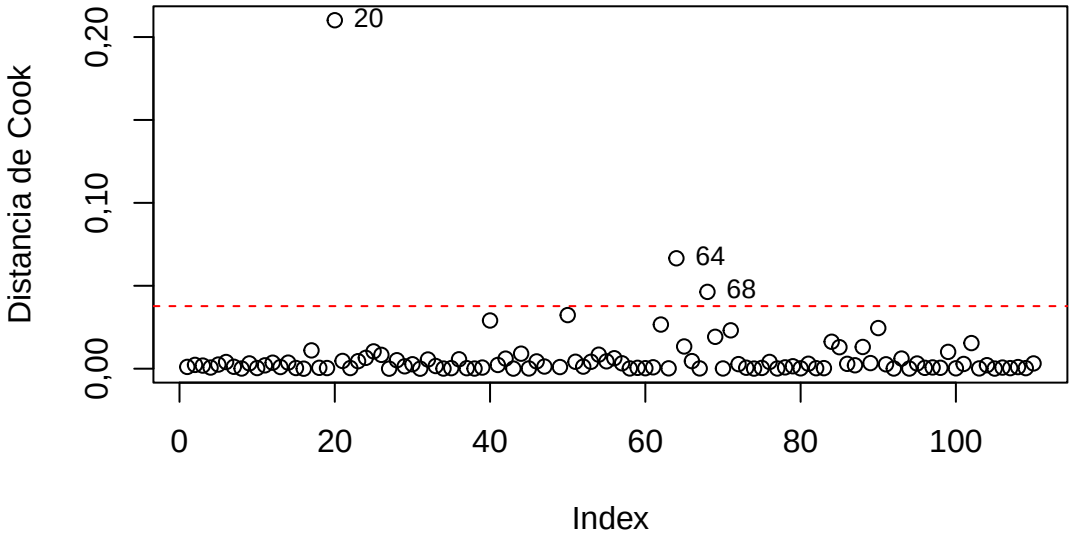


Figura 8: Distância de Cook

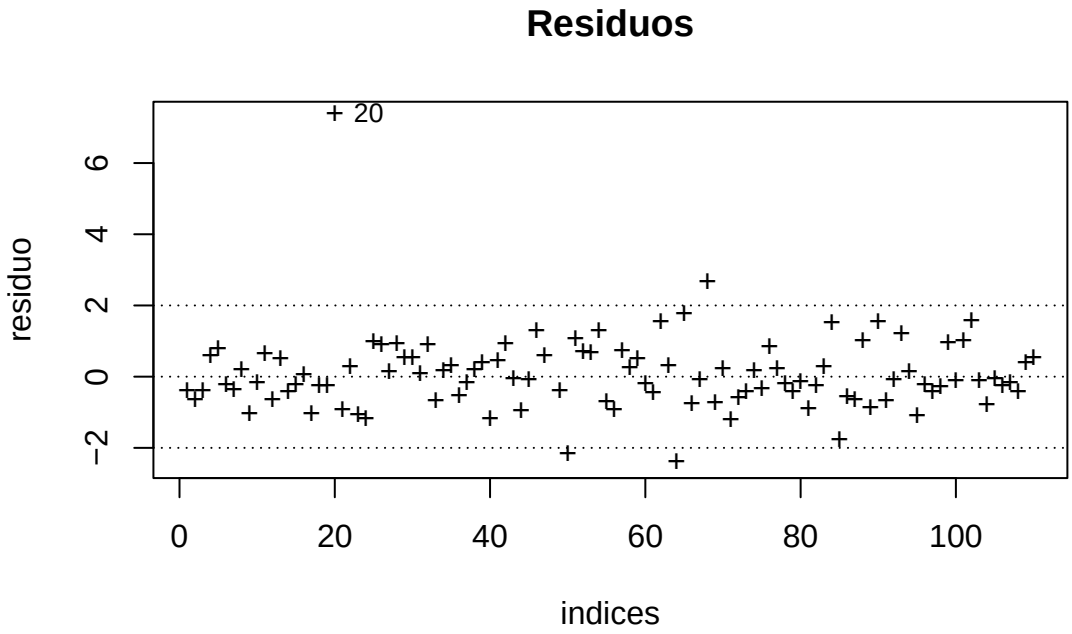


Figura 9: Gráfico dos Resíduos

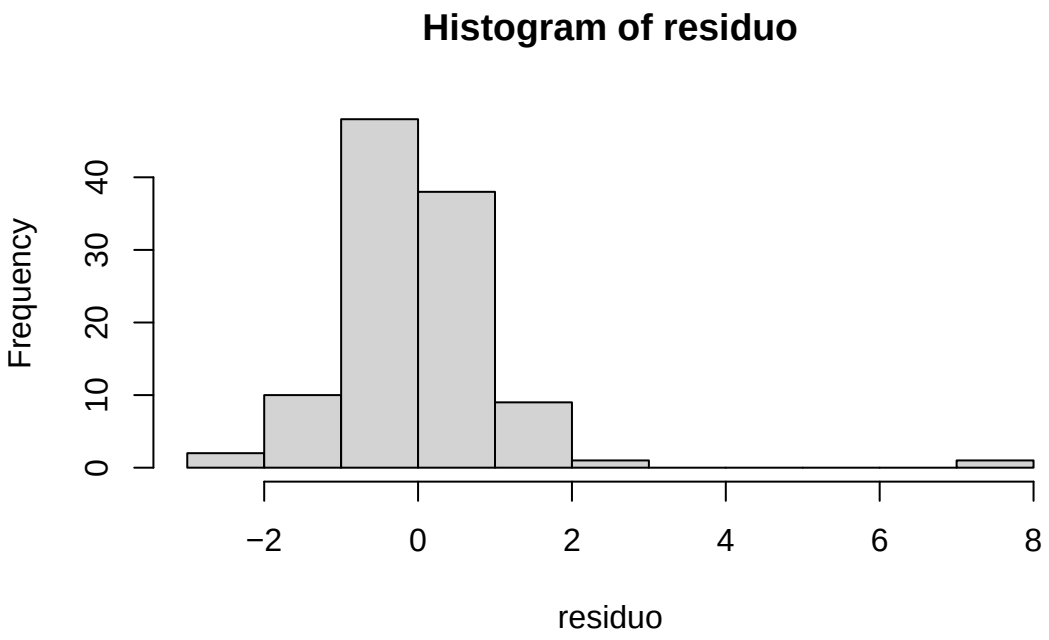


Figura 10: Histograma dos resíduos

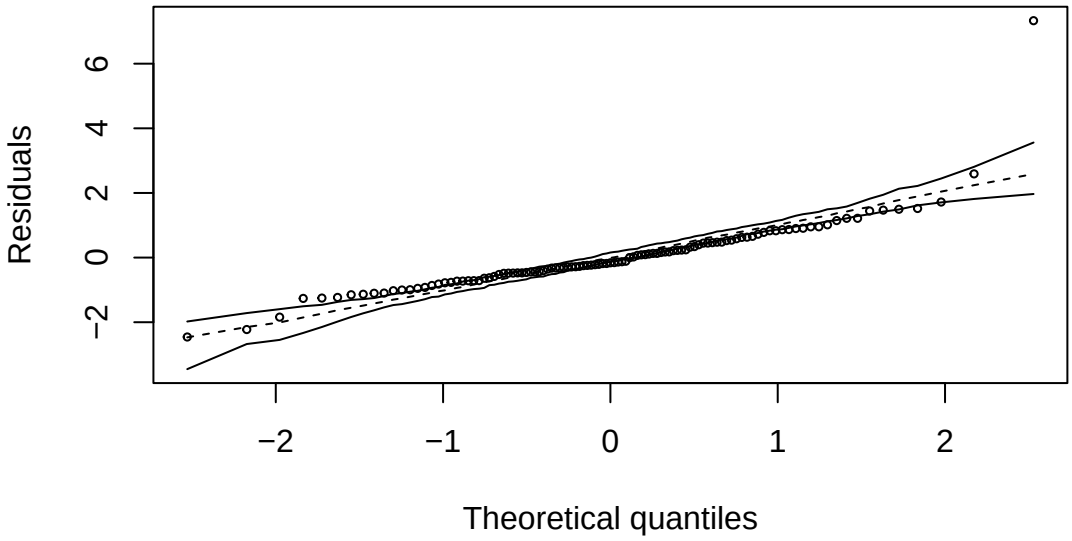


Figura 11: Histograma dos resíduos

No primeiro momento, foi feita a remoção da observação 20 e foi refeita a análise diagnóstico. Por conseguinte, novas observações foram identificadas como influentes para para o modelo e causou a rejeição de ao menos duas das suposições obrigatórios para validação do modelo de regressão linear. Então, após mais alguns ajustes, no total foram removidas 7 observações, são elas: $20, 61, 63, 64, 66, 67, 68$ novas análises diagnósticos foram feitas.

1.3 Validação dos pressupostos

Por se tratar de uma regressão linear múltipla, é necessário que o modelo ajusto siga algumas suposições, são elas:

S_0 : O modelo está corretamente especificado;

S_1 : A média dos erros é igual a zero;

S_2 : Homoscedasticidade dos erros;

S_3 : Não haver autocorrelação;

S_4 : Ausência de Multicolinearidade;

S_5 : Os erros seguem uma distribuição normal.

Para testar S_0, é utilizado o teste RESET com hipótese nula (H_0) de que o modelo está corretamente especificado. O p-valor encontrado foi de $0,4 > 0,05$, logo não rejeitamos H_0 e concluímos que o modelo está corretamente especificado.

Para testar S_1, é utilizado o teste t para as médias dos erros com hipótese nula (H_0) de que as médias dos erros é igual a zero. O p-valor encontrado foi de $1 > 0,05$, logo não rejeitamos H_0 e concluímos que a médias dos erros é igual a zero.

Para testar S_2, é usado o teste de Bressch-Pagan de Heteroscedasticidadee com hipótese nula (H_0) de que os erros são homoscedasticos. O p-valor encontrado foi de $0,083 > 0,05$, logo não rejeitamos H_0 e concluímos que os erros são homoscedasticos.

Para testar S_3, é usado o teste de Durbin-Watson de autocorrelação com hipótese nula (H_0) de que não há autocorrelação. O p-valor encontrado foi de $0,44 > 0,05$, logo não rejeitamos H_0 e concluímos que não há autocorrelação.

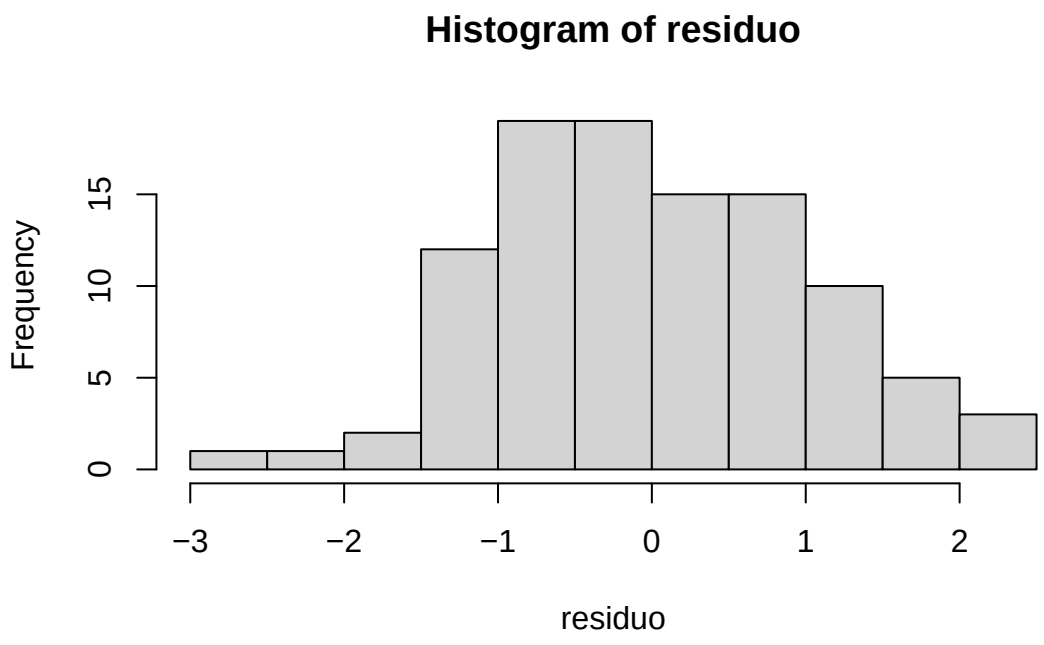
Para testar S_4, é utilizado o Fator de Inflação de Variância (VIF) para detectar multicolinearidade. O ideal é que os valores do VIFs seja igual a 1 e principalmente menores que 10. Na tabela **??**, temos os valores dos VIFs para cada covariável e todos os valores se encontram próximo de 1 o que é um indicativo de não multicolinearidade.

Para testar S_5, é utilizado o teste Jarque-Bera com hipótese nula H_0 de que os erros seguem uma distribuição normal. O p-valor encontrado foi de $0,74 > 0,05$, logo não rejeitamos H_0 e concluímos que os erros possuem uma distribuição normal.

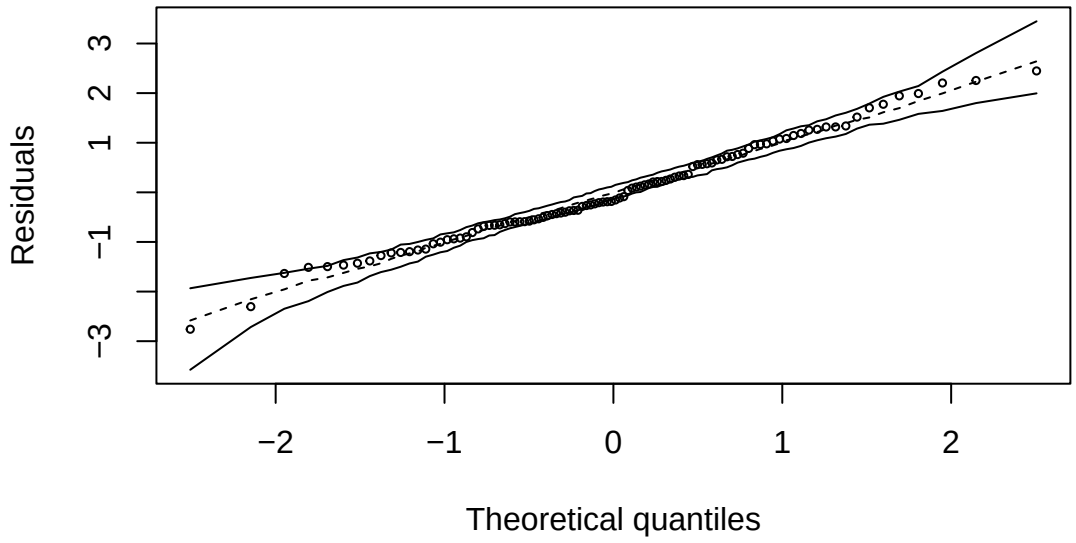
Tabela 5: Valores do VIFs para as covariáveis do modelo

Variável	sqft_living	yr_built	waterfront
VIF	1,1192	1,1135	1,0125

Para validar também as suposições, podemos verificar o novo histograma dos resíduos (gráfico **??**)e o envelopo simulado (gráfico **??**). Nota-se que agora temos diferenças nítidas nos dois gráficos quando comparados com os gráficos feitos sem a remoção das observações.




```
## Gaussian model (lm object)
```



Por fim, temos o modelo ajustado como sendo:

$$\hat{y} = \beta_1 + \beta_5 x_5 + \beta_8 x_8 + \beta_9 x_9$$

Sendo $\beta 1$ o intercepto e as covariáveis sqft_living e yr_built e waterfront. Com o novo modelo ajustado temos um R^2 de 0,633, ou seja, aproximadamente 63% da variabilidade de y é explicado pelo modelo em questão. e seu R^2 ajustado foi de 0,622

substituindo os valores dos betas estimados da tabela 4 temos:

$$y = 5267421,33 + 257,52x_5 - 2687,41x_8 + 622120,7x_9$$

Nota-se que a variável yr_built influencia negativamente no preço final do imóvel, o que faz sentido, visto que se a casa for antiga, há desvalorização da mesma no mercado imobiliário. Já a vista para o mar foi a variável que mais influenciou o preço final do imóvel, o que também faz muito sentido levando em consideração que imóveis próximos de praias, por exemplo, são mais valorizados.

2 Aplicação do modelo

Na tabela 6 abaixo temos cenários para aplicações de diferentes valores para variáveis. Como já esperado imóveis com vista para o mar são mais caros e imóveis mais novos também são mais caros quando comparados com imóveis mais antigos.

Tabela 6: Cenários de Preço dos Imóveis

Cenário	sqft_living	yr_built	waterfront	Preço Estimado (R$)
1	2000	1990	0	437.575,43
2	3000	2000	1	636.281,03
3	1500	1980	0	336.569,53
4	2500	2010	1	637.551,03
5	1800	1995	0	364.822,38

3 Referências

Linear Regression of Home Prices

GUJARATI, Damodar N. Econometria Básica. 5ª ed. Porto Alegre: AMGH Editora, 2011.