Universidade de Brasília

Departamento de Estatística - EST

Lista Prática 1

Arthur Gonçalves de Souza - 221022794

Análise de Regressão Linear

Profa. Terezinha Ribeiro

Brasília

2025

Sumário

1 Introdução

Este relatório possui como objetivo resolver dois problemas envolvendo análise de regressão linear simples. O **Problema 1** envolve dados referentes à um estudo de intervenções de tomografia computadorizada (fluoroscopia) no abdômen, no qual foram coletadas 19 observações, e a análise nesse caso possui como objetivo o ajuste do melhor modelo de regressão linear simples possível para explicar a Dose total de radiação recebida, com base na variável explicativa Tempo total do procedimento.

Por sua vez, o **Problema 2** envolve dados referentes à 27 imóveis, e nesse caso, o principal interesse da análise é ajustar o melhor modelo de regressão linear possível para explicar o preço de venda (em US$ 100) desses imóveis, podendo utilizar variáveis explicativas como imposto do imóvel (em US$ 100), área do terreno (em 1000 pés quadrados), área construída (em 1000 pés quadrados), e idade da residência (em anos).

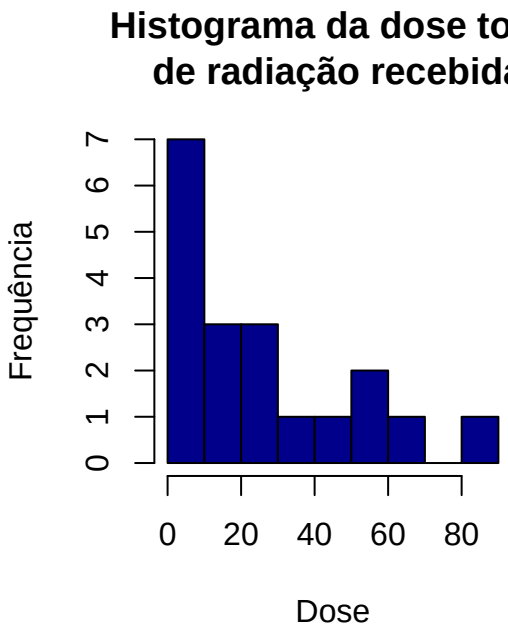
2 Análises

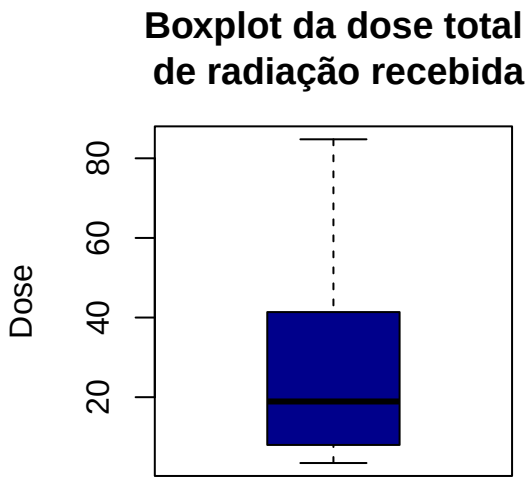
2.1 Modelo de Regressão para ajuste da dose total de radiação recebida

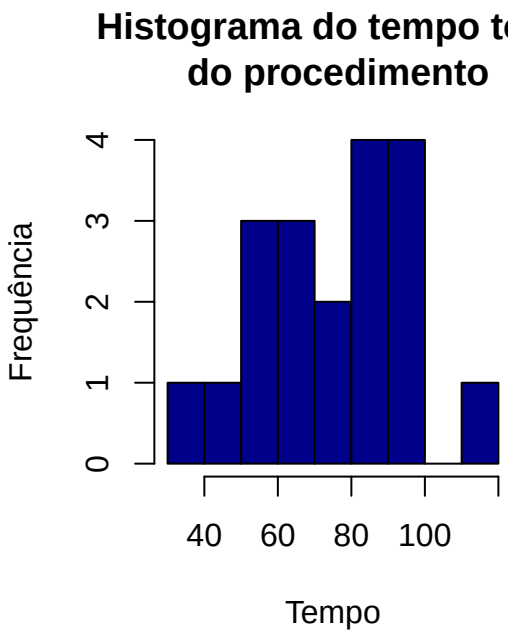
Nesta primeira seção, foi feita uma análise com o objetivo de encontrar o melhor modelo de regressão linear ajustado para explicar a Dose total de radiação recebida. Para tal, utilizou-se de análise descritiva das variáveis, ajuste dos modelos, verificação de suposições e análise de diagnóstico.

2.1.1 Análise Descritiva

Inicialmente, para compreender como as variáveis em estudo se comportam, foram elaborados histogramas e boxplots, além da análise das medidas descritivas. Após isso, verificou-se também como se dava a relação entre as variáveis.







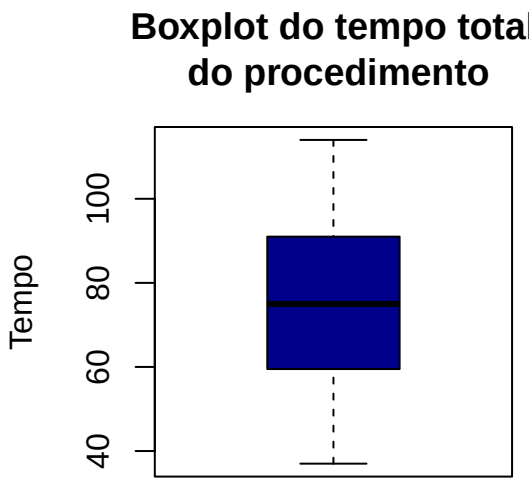
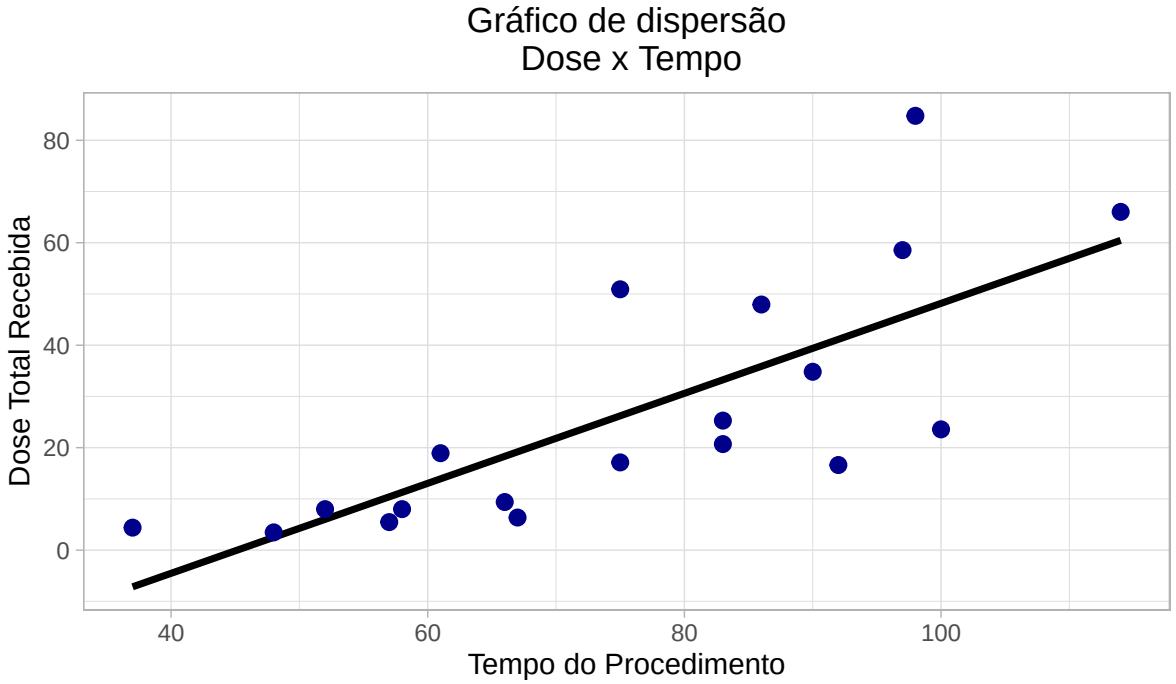


Tabela 1: Tabela de Medidas-resumo

Medidas	Tempo	Dose
Média	75,74	26,86
Desvio Padrão	20,5	23,85
Min	37	3,46
1º Quartil	59,5	8
Mediana	75	18,92
3º Quartil	91	41,38
Máximo	114	84,77
Coef. de variação	0,27	0,89

A partir dos gráficos, percebe-se uma certa assimetria à direita na distribuição dos valores das doses totais de radiação recebida, possuindo uma maior concentração de valores no intervalo de 0 à 10. Destaca-se ainda a diferença entre a média e a mediana, o que demonstra a influência que os valores mais altos podem estar exercendo nos dados. Em relação ao tempo total do procedimento, temos uma distribuição mais simétrica, evidenciada pela proximidade entre a mediana e a média



Analisando a relação entre as duas variáveis, percebe-se uma relação linear positiva, indicando que conforme o tempo aumenta, a Dose total também tende a aumentar. Vale destacar também que a variação dos dados também aparenta aumentar de acordo com o tempo do procedimento.

2.1.2 Transformações

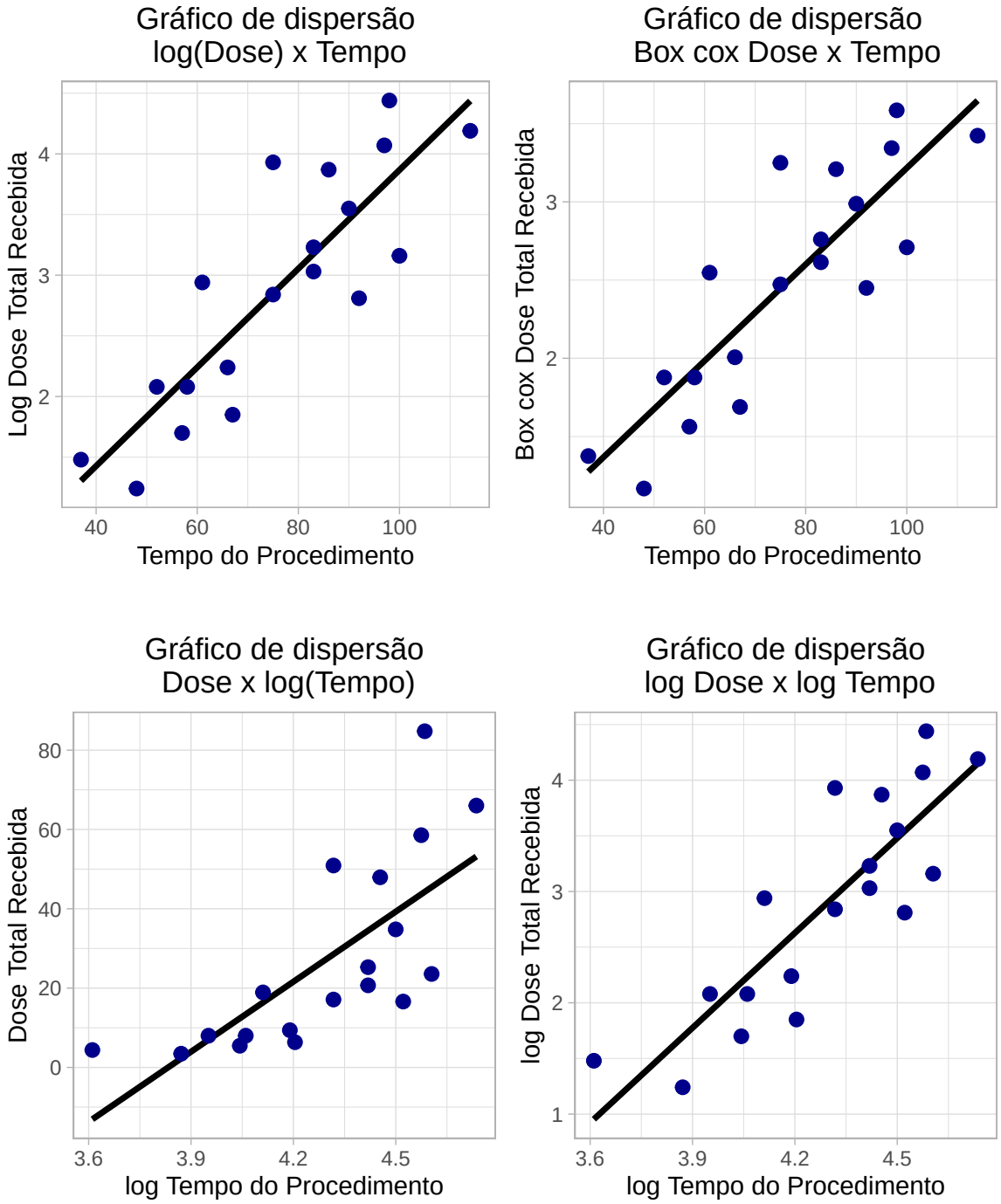
Com o objetivo de linearizar ainda mais a relação entre a Dose total recebida e o Tempo total do procedimento, foram aplicados transformações por logarítimico na variável resposta, na covariável e em ambas. Além dessas, também utitilizou-se transformação por BoxCox para a variável resposta, na qual, por meio de testes, foi encontrado o lambda ideal de -0,1.

A transformação de boxcox é uma das transformações mais conhecidas e utilizadas para normalizar a distribuição da variável resposta e estabilizar a variância. Ela é definida por:

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

onde $y > 0$. Apesar de essa transformação ser definida para todo λ real, usualmente testa-se valores de $\lambda \in [-2, 2]$.

Os gráficos de dispersão referentes à essas transformações se encontram abaixo.



Analisando os dois primeiros gráficos, é perceptível que a relação da variável resposta transformada com a covariável se tornou mais linear em ambos os casos, sendo inclusive os dois gráficos bastante similares. Isso se deve ao valor de lambda utilizado no boxcox ser bem próximo de 0, que é o lambda referente à transformação logarítimica. Tratando das relações que utilizaram transformação na covariável, o gráfico considerando a resposta original não apresentou uma relação mais linear, enquanto que a relação log na resposta e na covariável originou uma linearização semelhante aos gráficos que apresentam transformação apenas na resposta.

A partir do que foi observado, optou-se por ajustar três modelos de regressão linear, alterando a apenas a variável resposta, sendo em um a variável original, e nos outros dois a variável

transformada, descartando a relação com log em ambas variáveis, devido à sua maior complexidade por um ganho que o modelo mais “simples” já proporcionou.

Tabela 2: Modelos candidatos

Resposta	Covariável
Dose	Tempo de procedimento
log Dose	Tempo de procedimento
Boxcox Dose	Tempo de procedimento

2.1.3 Modelo Dose x Tempo

2.1.3.1 Ajuste do modelo

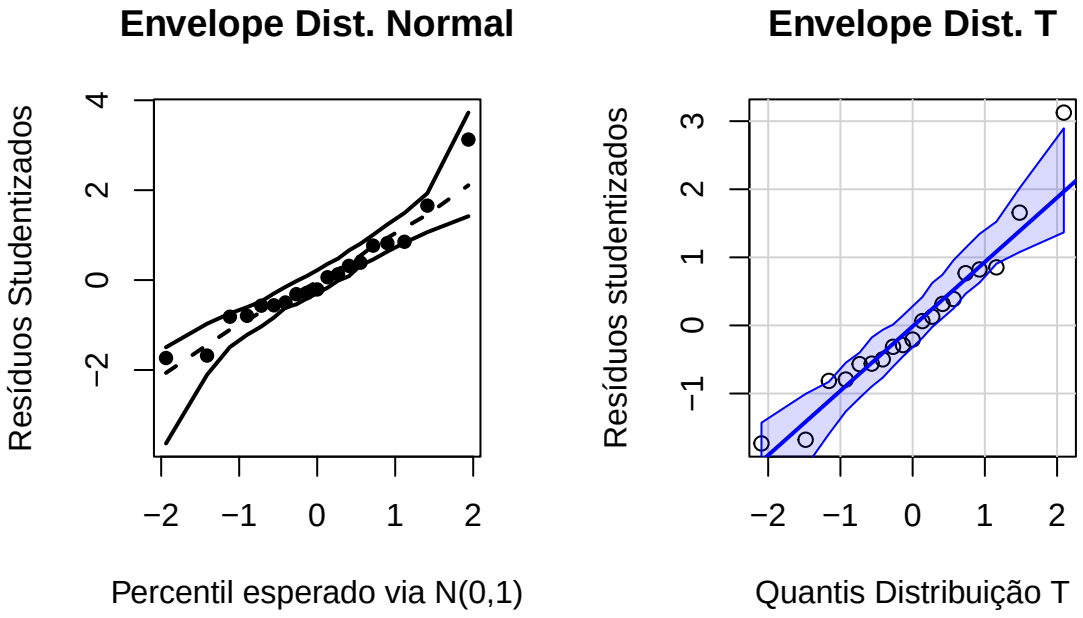
Tabela 3: Ajuste do Modelo Dose x Tempo

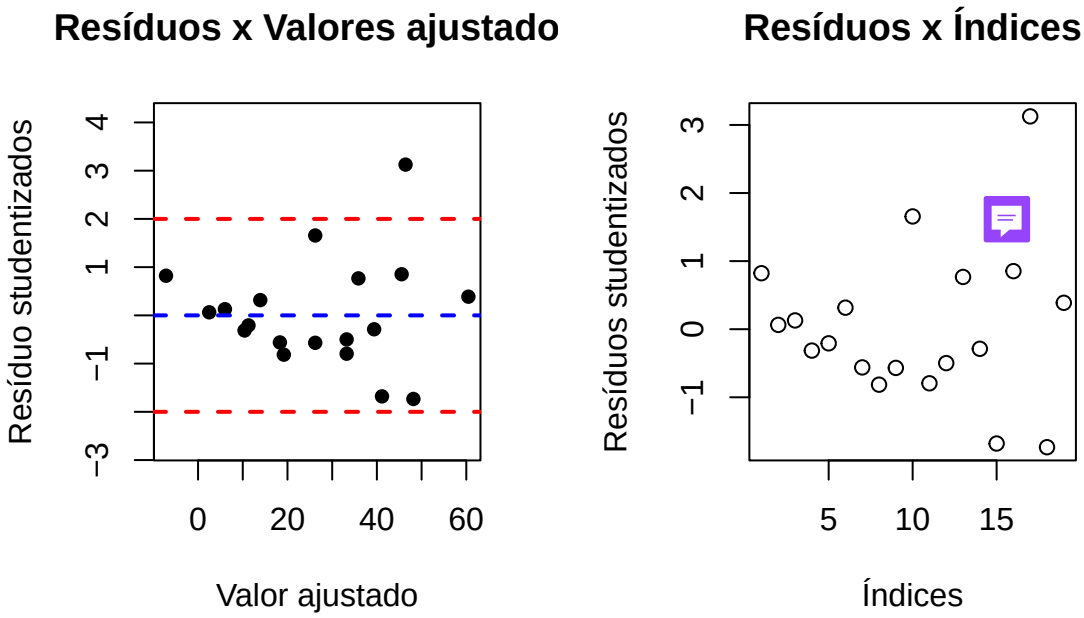
Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	-39,663799	14,486293	-2,738	0,014
Tempo	0,878313	0,184958	4,7487	0,0002
R^2	0,570200			
R^2 Ajustado	0,544900			

A partir da tabela, temos que o modelo apresentou um ajuste adequado, considerando os dois coeficientes significativos, a um nível de significância de 5%

2.1.3.2 Análise das suposições

Ao ajustar um modelo, é importante verificar se as suposições necessárias para sua inferência estão sendo cumpridas. Para isso, foram elaborados os seguintes gráficos:





Analisando os gráficos, não percebe-se um desvio em relação à suposição de normalidade dos erros, tanto observando o envelope pela distribuição Normal quanto pela distribuição exata dos resíduos, a T de Student, que é mais adequada já que não há um tamanho de amostra tão grande.

A respeito da homoscedasticidade, é perceptível pelo gráfico de Resíduos x Valores ajustados que a variabilidade está aumentando conforme o valor ajustado aumenta, o que é um ponto de atenção, e além disso, destaca-se a observação 17 como um possível aberrante, ou seja, é provável que a observação 17 tenha obtido um ajuste ruim para a variável resposta, e consequentemente, pode ter impactado numa possível distorção do intercepto.

Ademais, analisando a não-correlação dos erros, percebe-se novamente uma maior variabilidade, neste caso à medida que os dados foram coletados, mas sem apresentar nenhuma tendência clara, o que leva a não considerar como um desvio dessa suposição.

Para gerar mais evidências para tal decisão, optou-se por aplicar um teste para cada uma delas, sendo o primeiro, o teste Shapiro-Wilk, que avalia a aderência dos resíduos studentizados à distribuição Normal, o segundo, o teste Goldfield-Quandt, que analisa se as variâncias dos resíduos studentizados são iguais, e por último o teste Durbin-Watson, que avalia a presença de correlação entre os resíduos studentizados. Esses testes possuem as seguintes hipóteses nula:

Tabela 4: Testes de hipóteses para as suposições

Teste	Hipótesese nula
Shapiro-Wilk	Os resíduos Studentizados seguem distribuição normal padrão
Goldfield-Quandt	As variâncias dos resíduos Studentizados são iguais
Durbin-Watson	Os resíduos Studentizados são não-correlacionados

Tabela 5: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,1883	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,0016	Rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,3812	Não rejeita hipótese nula

A partir dos resultados obtidos pelos testes, e também pelo que foi percebido nos gráficos, concluímos que esse modelo ajustado não cumpre com a suposição de homoscedasticidade, sendo evidenciado ainda pelo p-valor = 0,0039 do teste Goldfield-Quandt, que rejeita a hipótese nula à um nível de significância de 5%.

Portanto, concluímos aqui a análise do modelo Dose x Tempo, visto que o mesmo não cumpre com as suposições necessárias para seu procedimento inferencial.

2.1.4 Modelo Log(Dose) x Tempo

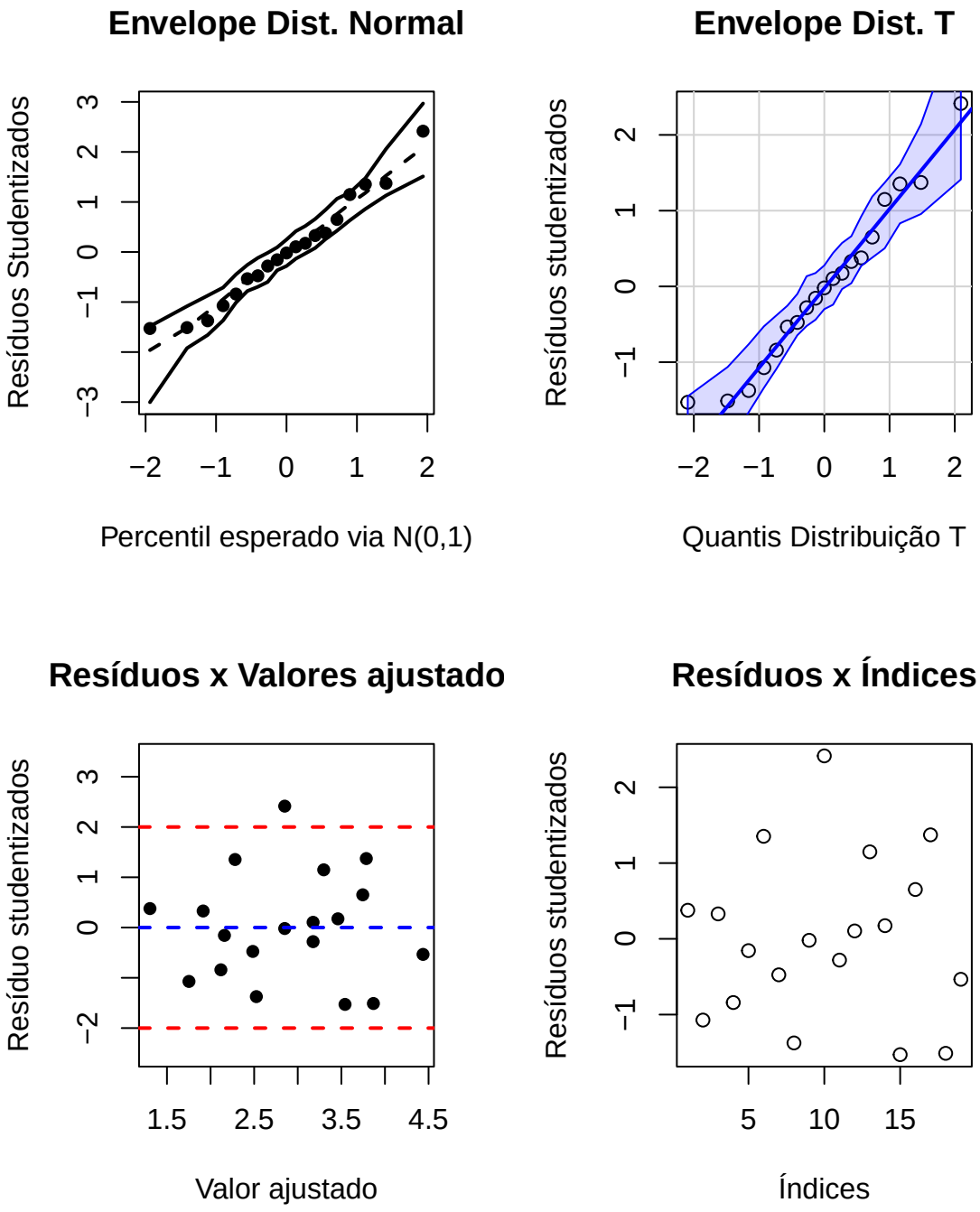
2.1.4.1 Ajuste do modelo

Tabela 6: Ajuste do modelo Log(Dose) x Tempo

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	-0,199939	0,468632	-0,4266	0,675
Tempo	0,040673	0,005983	6,7976	< 0,0001
R^2	0,731000			
R^2 Ajustado	0,715200			

A partir da tabela, temos que o modelo utilizando a resposta transformada via logarítimico apresentou um bom ajuste, apesar de considerar significativo apenas o coeficiente associado à variável explicativa, mas ainda assim manteremos o intercepto por questões inferenciais. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 70% da variabilidade da resposta.

2.1.4.2 Análise das suposições



Analisando os gráficos, não percebe-se nenhum desvio em relação à suposição de normalidade dos erros, analisando nesse caso os resíduos studentizados, tanto pelo Envelope da Distribuição Normal, quanto pelo exato da Distribuição T.

Em questão à homoscedasticidade, não é perceptível nenhum tipo de tendência em relação aos resíduos studentizados, aparentando variarem de forma aleatória, destacando-se apenas a observação 10 como um ponto a ser investigado posteriormente.

A respeito da não correlação dos erros, também não percebe-se nenhuma tendência pelo gráfico Resíduos x Índices.

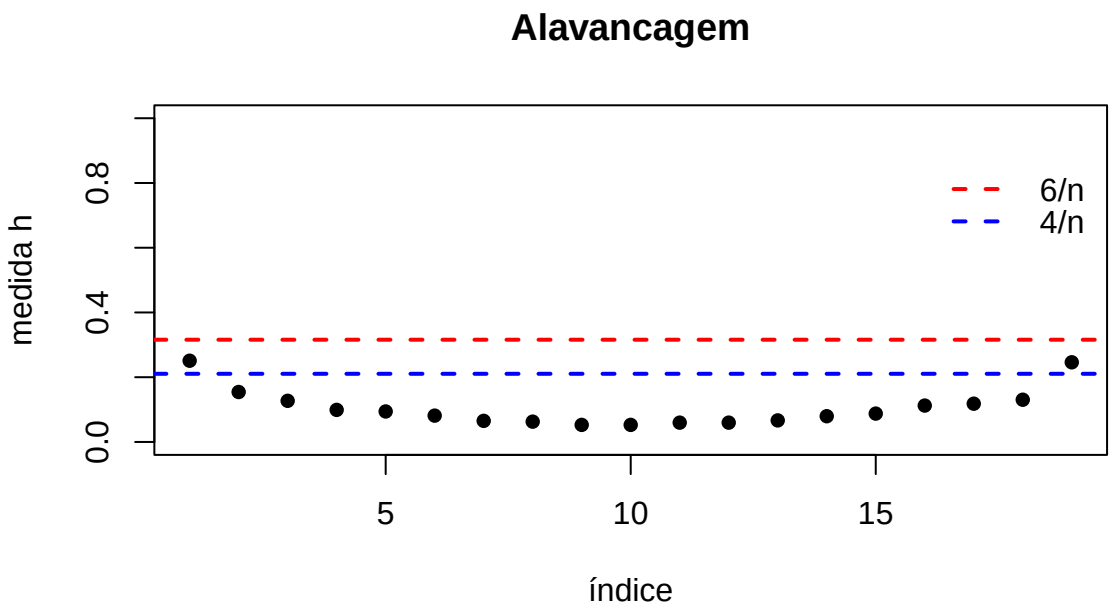
Tabela 7: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,6755	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,3251	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,6887	Não rejeita hipótese nula

A partir dos testes aplicados, reforça-se a conclusão de que o modelo não apresenta desvios para nenhuma suposição, e portanto, segue-se para a análise de alavancagem e influência.

2.1.4.3 Análise de alavancagem e influência

Pontos alavanca são pontos que possuem peso desproporcional no próprio valor ajustado, geralmente tendo essa desproporcionalidade causada por possuírem uma inconformidade em relação ao valor da covariável. Além disso, eles também podemo exercer influência atípica nas estimativas dos coeficientes da regressão. Para identificar esses pontos, analisamos suas medidas de alavancagem, como pode-se visualizar no gráfico abaixo:



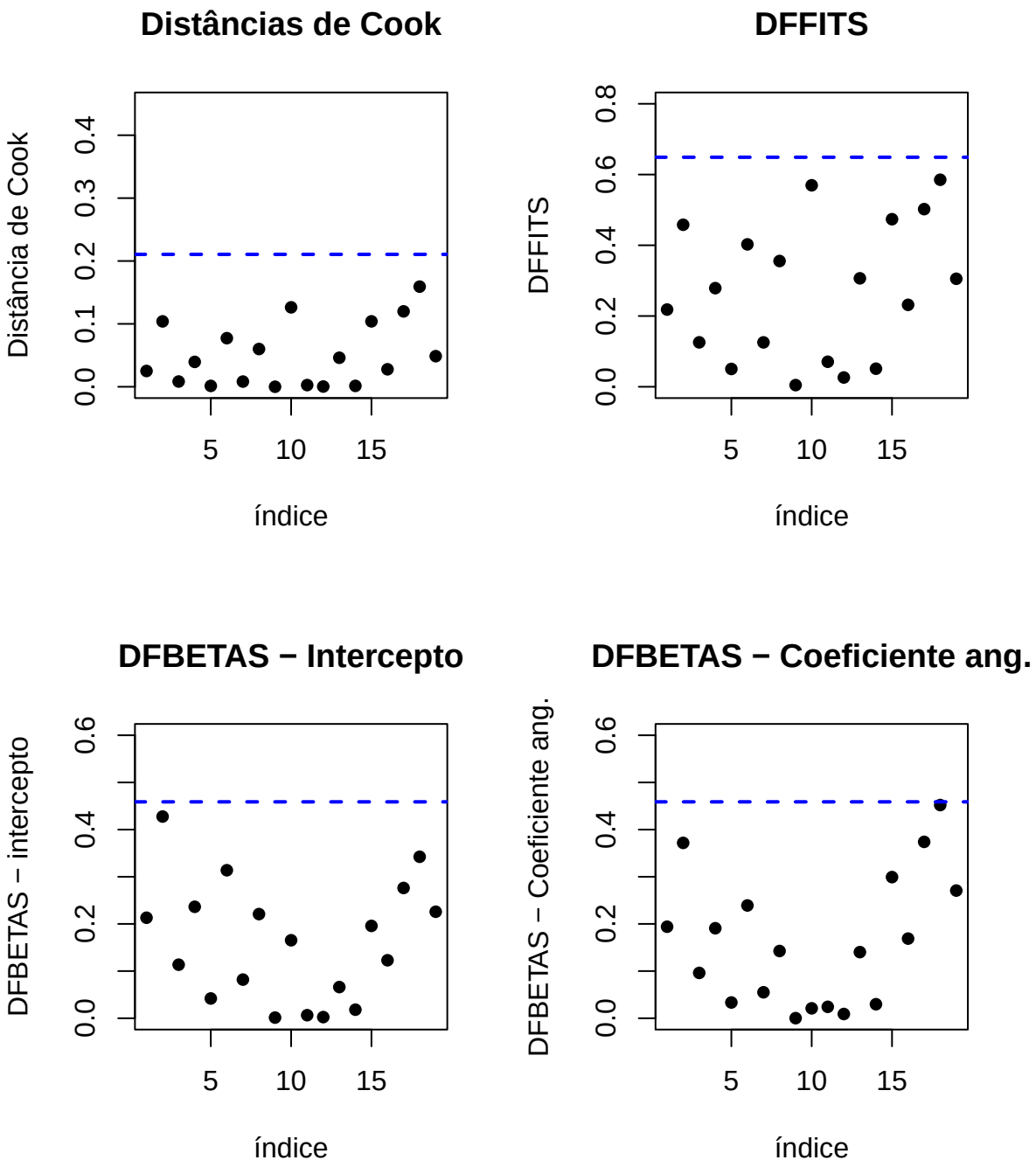
A partir do gráfico, nota-se que os pontos 1 e 19 podem ser classificados como possíveis pontos de alavanca, devendo-se então realizar uma investigação sobre os mesmos.

Já em relação à influência, são classificadas como influentes aquelas observações que exercem influência desproporcional nas estimativas dos parâmetros do modelo e podem causar mudança inferencial. Destaca-se que usualmente um ponto influente é aberrante e/ou alavanca.

Para realizar a identificação desses pontos, existem 3 medidas de influência:

- **Distância de Cook**: Mede o efeito que a exclusão da i-ésima observação causa nos ajustes da resposta para todas as observações;

- **DFFITS (Difference in fits)**: Mensura o impacto da exclusão da i-ésima observação apenas no ajuste do ponto em questão;
- **DFBETAS (Difference in betas)**: Avalia o impacto da i-ésima observação nas estimativas dos coeficientes.



Por meio dos gráficos plotados acima, não é possível observar nenhuma observação candidata à ponto de influência.

2.1.4.4 Investigação dos pontos atípicos

Para observar se os pontos considerados atípicos pelos procedimentos anteriores de fato causam algum tipo de influência ou distorção nos modelos, foram ajustados modelos retirando os pontos 1, 10 e 19, individualmente e conjuntamente, e a partir disso mensurou-se o impacto que os mesmos possuem nas estimativas do modelo, os quais são apresentados abaixo:

Tabela 8: Estimativas dos modelos retirando os pontos atípicos

Pontos	Beta 0	Mudança no Beta 0	Beta 1	Mudança no Beta 1
Com todos pontos	-0,199	0%	0,040	0%
Retirando 1	-0,302	-51,2%	0,041	2,9%
Retirando 10	-0,268	-34,2%	0,040	0,2%
Retirando 19	-0,308	-54,0%	0,042	4,0%
Retirando 1, 10 e 19	-0,544	-172,10%	0,044	9,1%

Tabela 9: P-valores dos modelos retirando os pontos atípicos

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Com todos pontos	0,675	0%	0,00000311	0%
Retirando 1	0,592	-12,3%	0,0000165	429%
Retirando 10	0,527	-21,9%	0,000000872	-71%
Retirando 19	0,562	-16,7%	0,0000133	327%
Retirando 1, 10 e 19	0,339	-49,8%	0,0000204	555%

Analisando os resultados obtidos, percebe-se claramente que a retirada dos pontos impacta na diminuição do valor associado ao β_0, enquanto não apresenta um impacto considerável no $\beta 1$. Apesar dessas alterações, a retirada desses pontos não gera mudança de significância nem de sinal com relação aos coeficientes, e nesse sentido a influência dos pontos atípicos não aparenta distorcer significativamente o ajuste do modelo.

Em resumo, a partir de tudo o que foi apresentado para este modelo, sem apresentar desvios das suposições e sem impacto considerável dos pontos atípicos, conclui-se que ele é um bom modelo apto para explicar a variável resposta, nesse caso transformada.

2.1.5 Modelo Boxcox(Dose) x Tempo

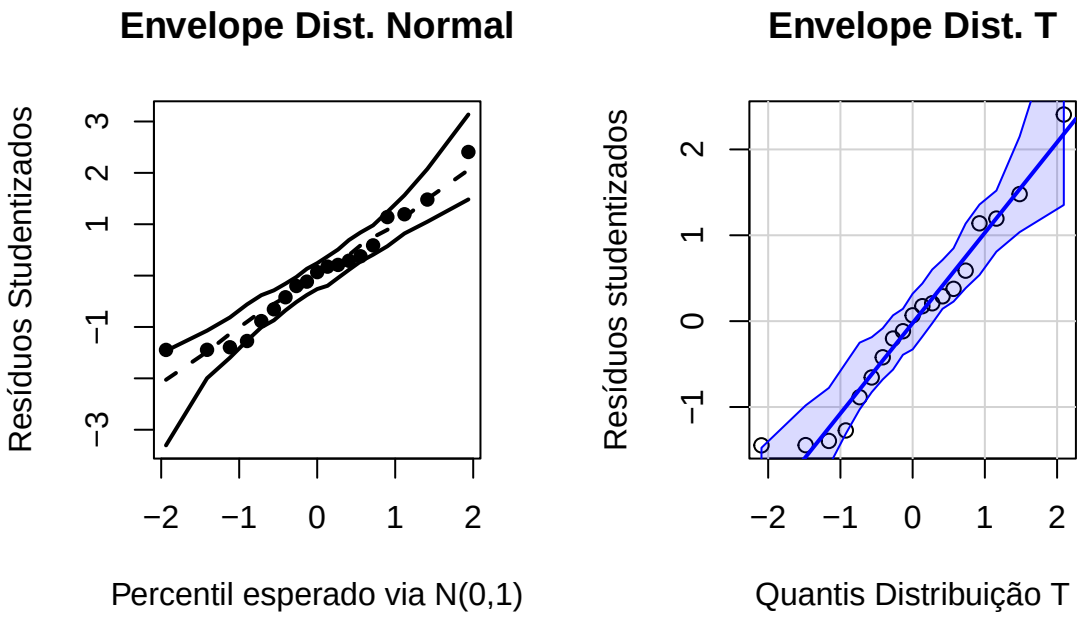
2.1.5.1 Ajuste do modelo

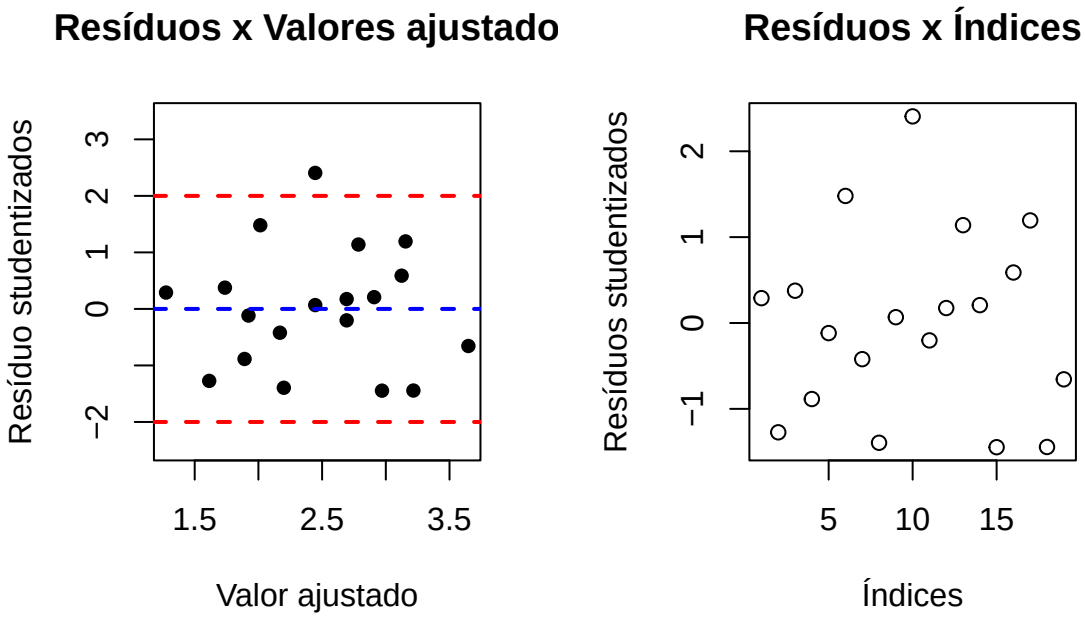
Tabela 10: Ajuste do modelo Boxcox Dose x Tempo

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	0,134845	0,34975	0,3855	0,7046
Tempo	0,030817	0,004466	6,9011	< 0,0001
R^2	0,736900			
R^2 Ajustado	0,721500			

A partir da tabela, temos que o modelo utilizando a resposta transformada via Boxcox, com $\lambda = -0,1$, apresentou um bom ajuste, apesar de considerar significativo apenas o coeficiente associado à variável explicativa, mas ainda assim manteremos o intercepto por questões inferenciais. Destaca-se também o valor de R^2, que assim como foi visto no modelo passado, indica que este modelo está conseguindo explicar mais de 70% da variabilidade da resposta.

2.1.5.2 Análise das suposições





Analisando os gráficos, novamente não percebe-se nenhum desvio em relação à suposição de normalidade dos erros, tanto pelo Envelope da Distribuição Normal, quanto pelo exato da Distribuição T.

Em questão à homoscedasticidade, não é perceptível nenhum tipo de tendência em relação aos resíduos studentizados, aparentando variarem de forma aleatória, destacando-se novamente apenas a observação 10 que novamente deve ser investigado.

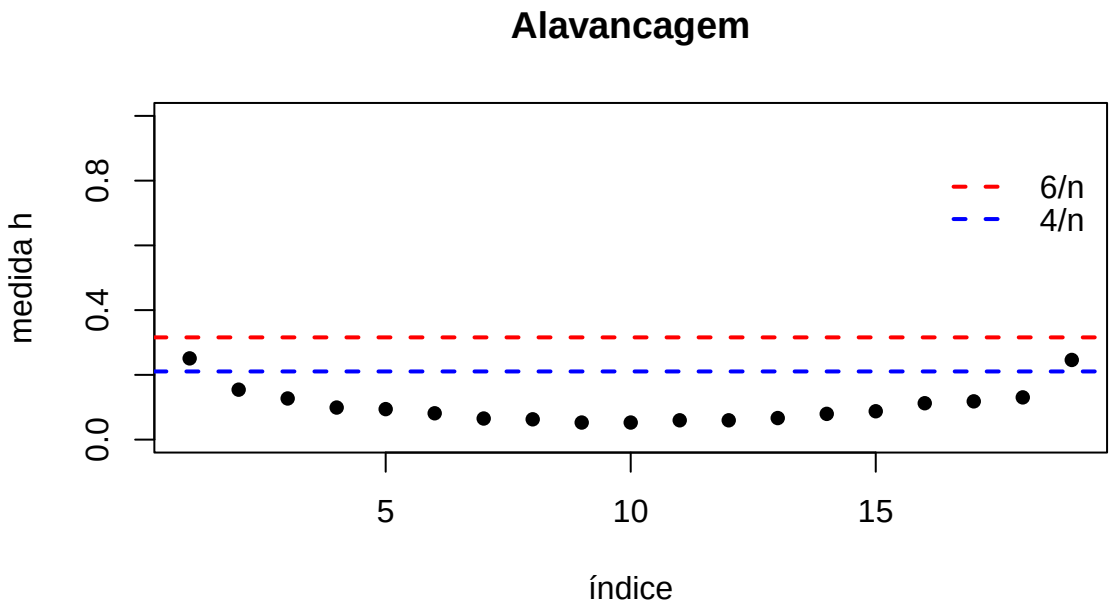
A respeito da não correlação dos resíduos, também não percebe-se nenhuma tendência pelo gráfico Resíduos x Índices, indicando que os resíduos foram obtidos de forma aleatória.

Tabela 11: Testes de hipóteses aplicados

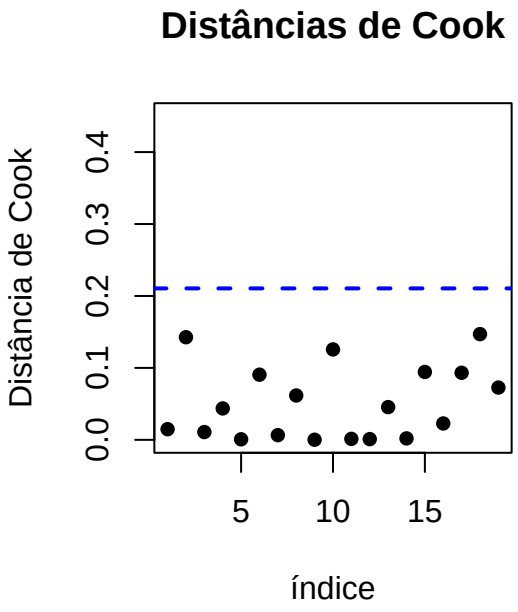
Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,4819	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,4444	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,7926	Não rejeita hipótese nula

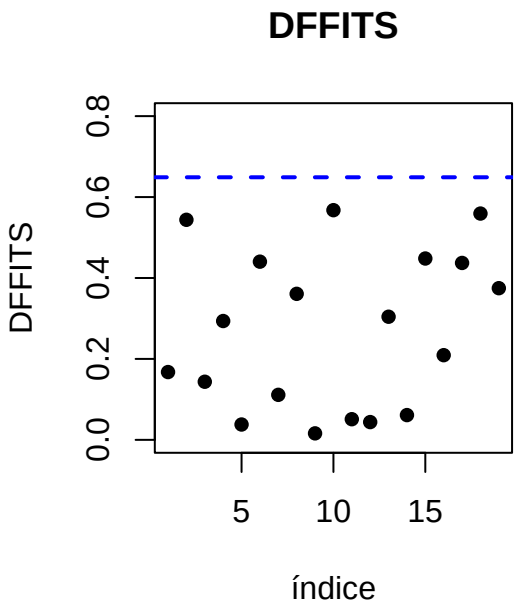
A partir dos testes aplicados, reforça-se a conclusão de que o modelo não apresenta desvios para nenhuma suposição, e portanto, segue-se para a análise de alavancagem e influência.

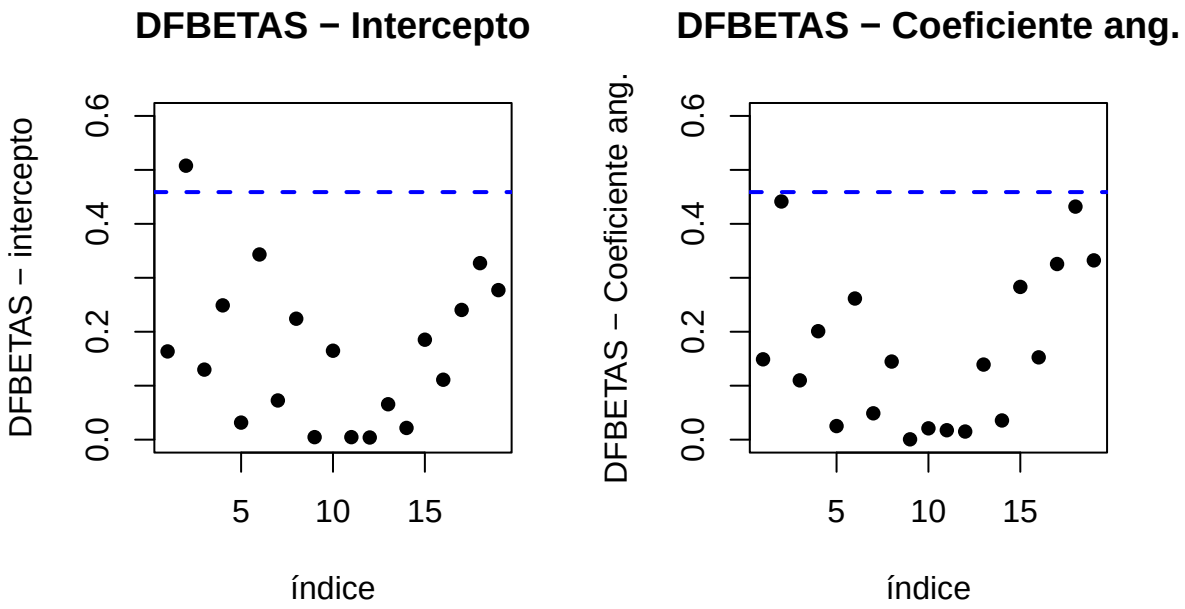
2.1.5.3 Análise de influência e alavancagem



A partir do gráfico, nota-se, assim como no modelo anterior, que os pontos 1 e 19 podem ser classificados como possíveis pontos de alavanca, devendo-se então realizar uma investigação sobre os mesmos.







Em relação à influência, desta vez foi possível observar, por meio dos DFBETAS para o intercepto, o ponto 2 como um possível ponto influente, e que nesse caso, aparente causar influência desproporcional no intercepto.

2.1.5.4 Investigação dos pontos atípicos

A partir das técnicas aplicadas anteriormente, foram observados os pontos 1, 2, 10 e 19, como possíveis pontos atípicos, e portanto, estes serão investigados por meio do ajuste de modelos retirando os mesmos.

Tabela 12: Estimativas dos modelos retirando os pontos atípicos

Pontos	Beta 0	Mudança no Beta 0	Beta 1	Mudança no Beta 1
Com todos pontos	0,13	0%	0,030	0%
Retirando 1	0,07	-43%	0,031	2,21%
Retirando 2	0,30	129%	0,028	-6,28%
Retirando 10	0,08	-37%	0,030	0,26%
Retirando 19	0,03	-73%	0,032	4,89%
Retirando 1, 2, 10 e 19	0,09	-28%	0,031	0,97%

Tabela 13: P-valores dos modelos retirando os pontos atípicos

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Com todos pontos	0,705	0%	0,00000257	0%
Retirando 1	0,856	21,5%	0,0000153	496%
Retirando 2	0,415	-41,0%	0,0000122	376%
Retirando 10	0,790	12,0%	0,000000727	-71%
Retirando 19	0,926	31,4%	0,00000966	418%

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Retirando 1, 2, 10 e 19	0,83793	18,9%	0,00013	4.962%

A partir dos resultados obtidos na investigação, nota-se que o ponto 2 é o único que aumenta o valor associado ao β_0 quando é retirado, impactando em uma mudança relativa considerável de 129% do mesmo. A retirada dos pontos de modo geral não trouxe uma variação considerável para o β_1. Dessa forma, apesar da grande mudança que o ponto 2 traz intercepto, não foi percebida nenhuma mudança de sinal nem de significância na investigação, o que também torna o modelo apto para escolha.

2.1.6 Conclusão

Por meio dos resultados obtidos, acredita-se que o melhor modelo de regressão simples ajustado seja o **Modelo log(Dose) x Tempo**, pois apresentou maior adequabilidade às suposições e possui menos pontos atípicos que o modelo boxcox(Dose) x Tempo, além de também não possuir uma interpretação tão dificultada quanto o mesmo.

Tabela 14: Ajuste do modelo escolhido - Log(Dose) x Tempo

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	-0,199939	0,468632	-0,4266	0,675
Tempo	0,040673	0,005983	6,7976	< 0,0001
R^2	0,731000			
R^2 Ajustado	0,715200			

Analisando mais especificamente o ajuste do modelo, tem-se $\beta_0 = -0,19$, indicando a média do log da Dose total quando fixamos o valor da covariável Tempo à zero, o que não faz sentido no contexto do estudo. Além disso, tem-se $\beta_1 = 0,04$, que representa a variação na média do log da Dose total quando acrescenta-se uma unidade na covariável Tempo.

Como utilizamos uma transformação na variável resposta para ajustar o modelo, os coeficientes são obtidos considerando a escala transformada. Felizmente, para o caso da transformação logarítimica, é possível facilmente encontrar as estimativas dos coeficientes para a escala da resposta original, apenas aplicando o exponencial nas estimativas do modelo com base na resposta transformada. Dessa forma, obtém-se que $e^{\beta_0} = 0,81$ e $e^{\beta_1} = 1,04$, ou seja, 0,81 é a média da resposta Dose total quando fixamos a covariável tempo à zero (não faz sentido nesse caso), e 1,04 é a variação na média da resposta Dose total quando acrescenta-se um segundo no tempo total do procedimento.

Além disso, foi calculado o intervalo de confiança para os coeficientes, que segue na tabela abaixo:

Tabela 15: Intervalo de confiança para os coeficientes

Coeficientes	Limite Inferior	Limite Superior
Intercepto	-1,189	0,789
Tempo total	0,028	0,053

Ademais, vale destacar que modelo apresenta $R^2 = 0,731$, representando que mais de 73% da variabilidade da Dose total recebida está sendo explicada pela modelo, o que é um indicativo de que o modelo consegue explicar bem a variável resposta.

2.1.7 Predições

A partir dos coeficientes estimados pelo modelo ajustado, podemos realizar predições para a variável resposta com base em novos valores fixados para a covariável Tempo total, desde que estes estejam dentro do intervalo de valores do banco de dados original. Como nesse caso queremos predizer a resposta original, utilizou-se o exponencial dos coeficientes obtidos no modelo ajustado para a transformação logarítimica. As predições obtidas encontram-se na tabela abaixo:

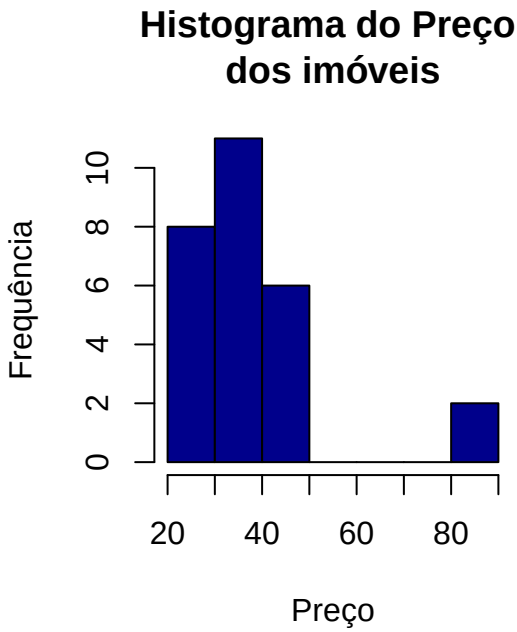
Tabela 16: Predições da dose recebido para novos valores de Tempo de procedimento

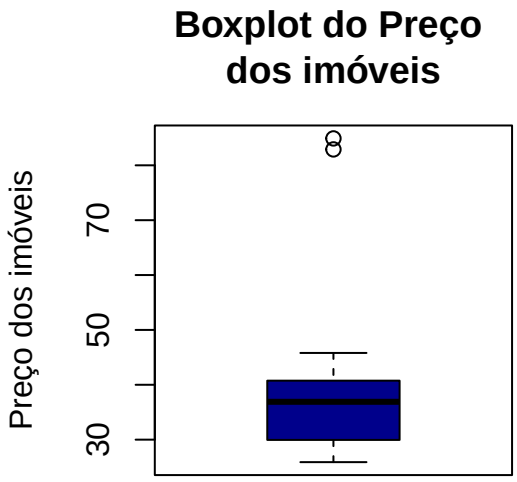
Tempos fixados	Predições
38	3,84
41	4,34
53	7,07
59	9,02
63	10,62
65	11,52
78	19,54
79	20,35
85	25,98
89	30,57
91	33,16
95	39,02
102	51,87
103	54,02
109	68,95

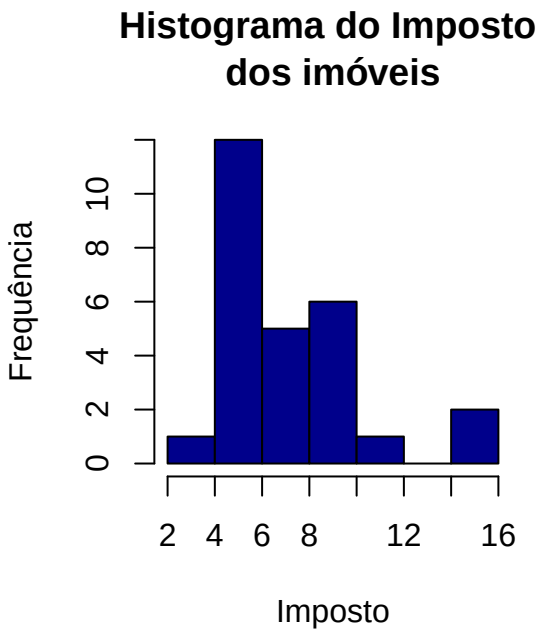
2.2 Modelo de Regressão para ajuste do preço de imóveis

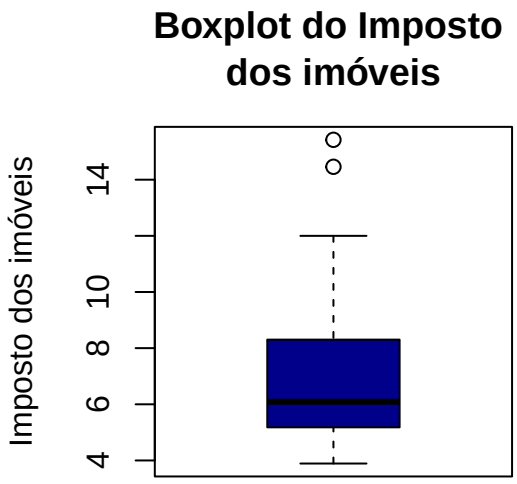
Na segunda seção, foi feita uma análise com o objetivo de encontrar o melhor modelo de regressão linear ajustado para explicar o preço de imóveis (em US$ 100). Para tal, utilizou-se de análise descritiva das variáveis, ajuste dos modelos, verificação de suposições e análise de diagnóstico.

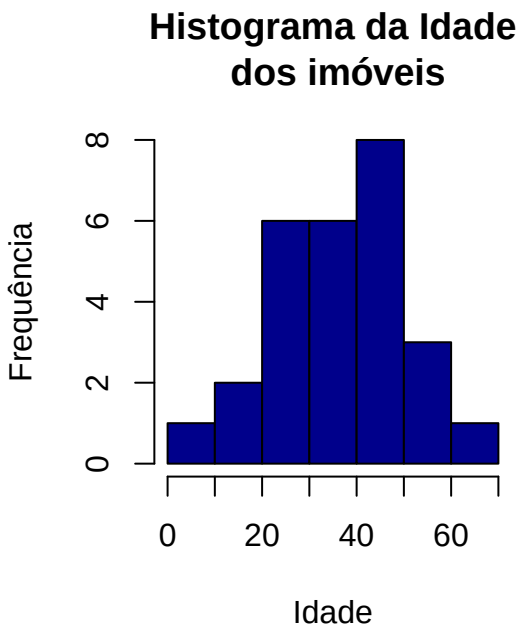
2.2.1 Análise Descritiva

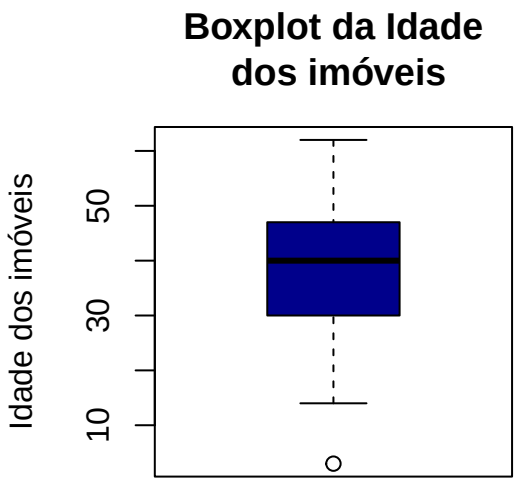


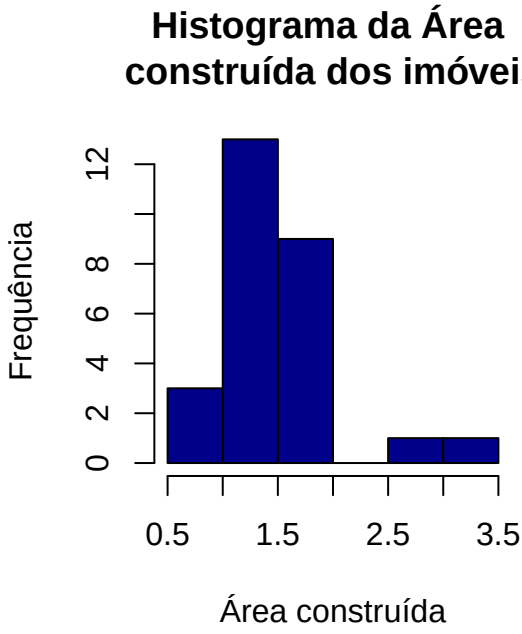


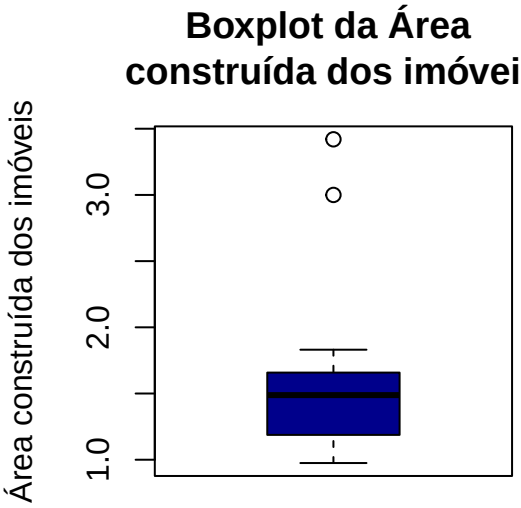


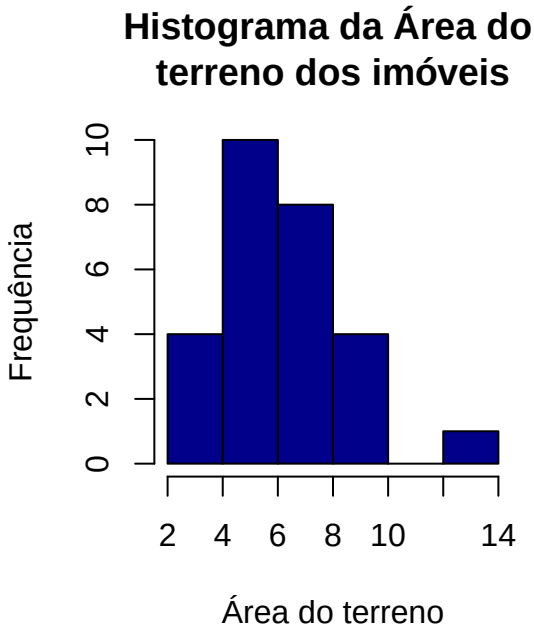



Histograma da Idade dos imóveis
Frequência
0
2
4
6
8
0
20
40
60
Idade


Boxplot da Idade dos imóveis
Idade dos imóveis
10
30
50


Histograma da Área construída dos imóveis
Frequência
0
2
4
6
8
12
0.5
1.5
2.5
3.5
Área construída


Boxplot da Área construída dos imóveis
Área construída dos imóveis
1.0
2.0
3.0



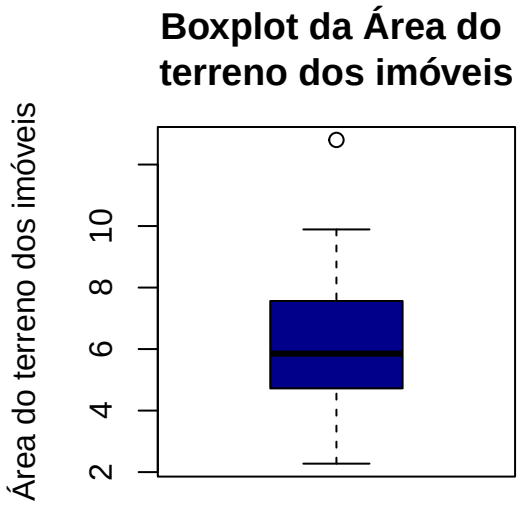


Tabela 17: Medidas Resumo

Medida	Imposto	Área do terreno	Área construída	Idade	Preço
Média	7,24	6,35	1,51	36,48	38,5
Desvio Padrão	2,88	2,4	0,56	14,05	14,31
Min	3,89	2,28	0,98	3	25,9
1º Quartil	5,18	4,72	1,19	30	29,95
Mediana	6,09	5,85	1,49	40	36,9
3º Quartil	8,3	7,56	1,66	47	40,75
Máximo	15,42	12,8	3,42	62	84,9
Coef. de variação	0,4	0,38	0,37	0,39	0,37

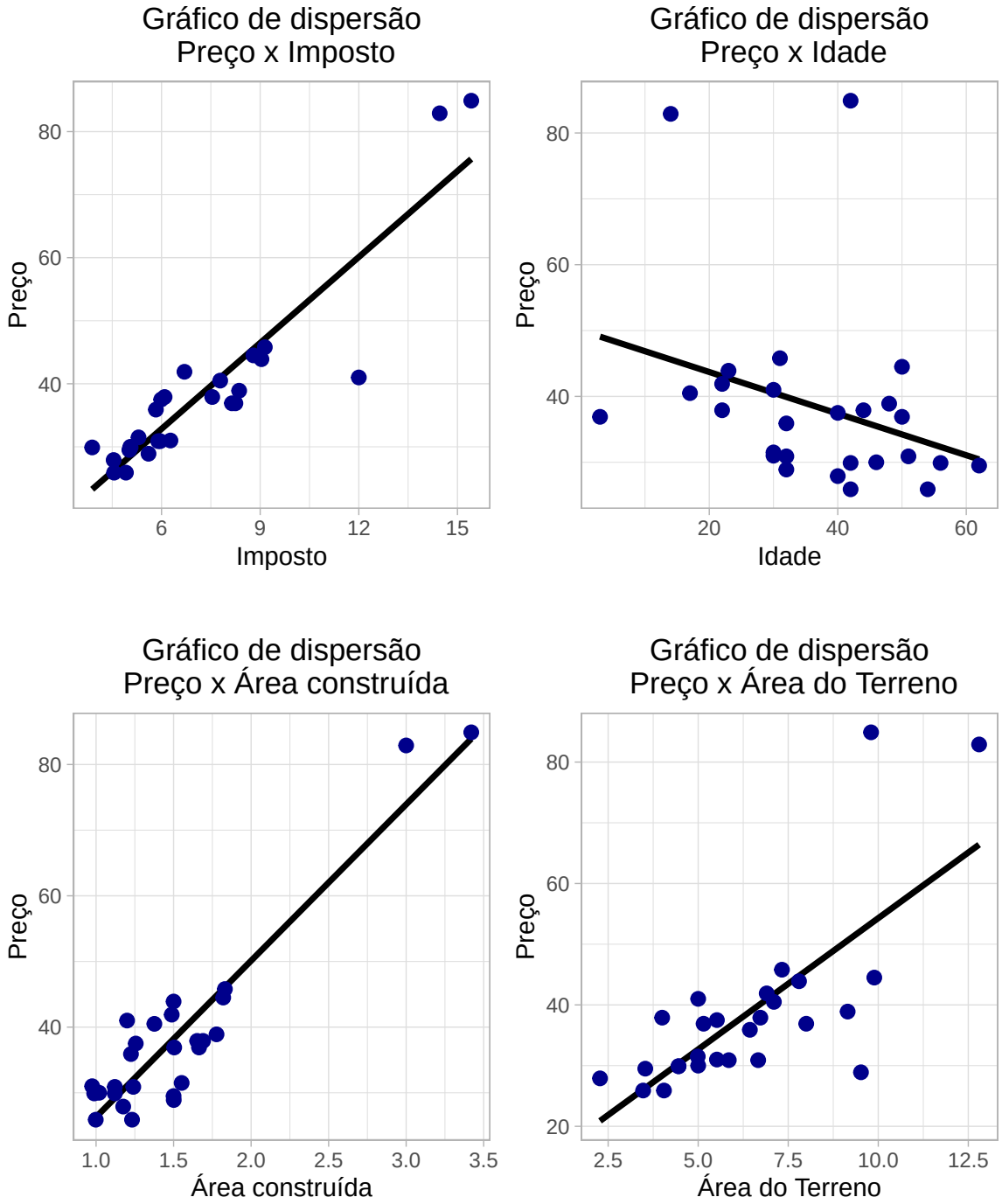
Analisando o gráfico e as medidas-resumo, é possível perceber que a variável resposta, preço dos imóveis, apresenta uma distribuição assimétrica à direita, sendo isso ainda evidenciado pela presença de outliers superiores, ou valores discrepantes, no boxplot, além de apresentar também grande concentração dos imóveis com preço de até 5 mil dólares.

Em relação à covariável Imposto, nota-se que esta também apresenta outliers superiores, apesar de não apresentar uma distribuição tão assimétrica quanto a Resposta. Destaca-se também uma grande concentração de imóveis com Imposto entre 400 e 600 dólares, e que dentre as variáveis do banco, Imposto é a que apresenta maior coeficiente de variação.

Por outro lado, a variável idade apresenta uma distribuição simétrica, por sua vez contendo um valor discrepante inferior, sendo a variável que mais aparenta seguir uma distribuição normal, o que faz sentido dado sua natureza.

Tratando da covariável Área construída, esta é a que apresenta distribuição mais semelhante à da variável resposta, sendo assimétrica à direita e também contendo dois outliers superiores. Destaca-se ainda uma considerável concentração de imóveis com área construída entre 1000 e 1500 pés quadrados.

Por último, a covariável Área do Terreno apresenta uma distribuição bem menos assimétrica que a covariável Área construída, sendo perceptível a presença de um valor discrepante com Área do Terreno entre 12 e 14 mil pés quadrados.



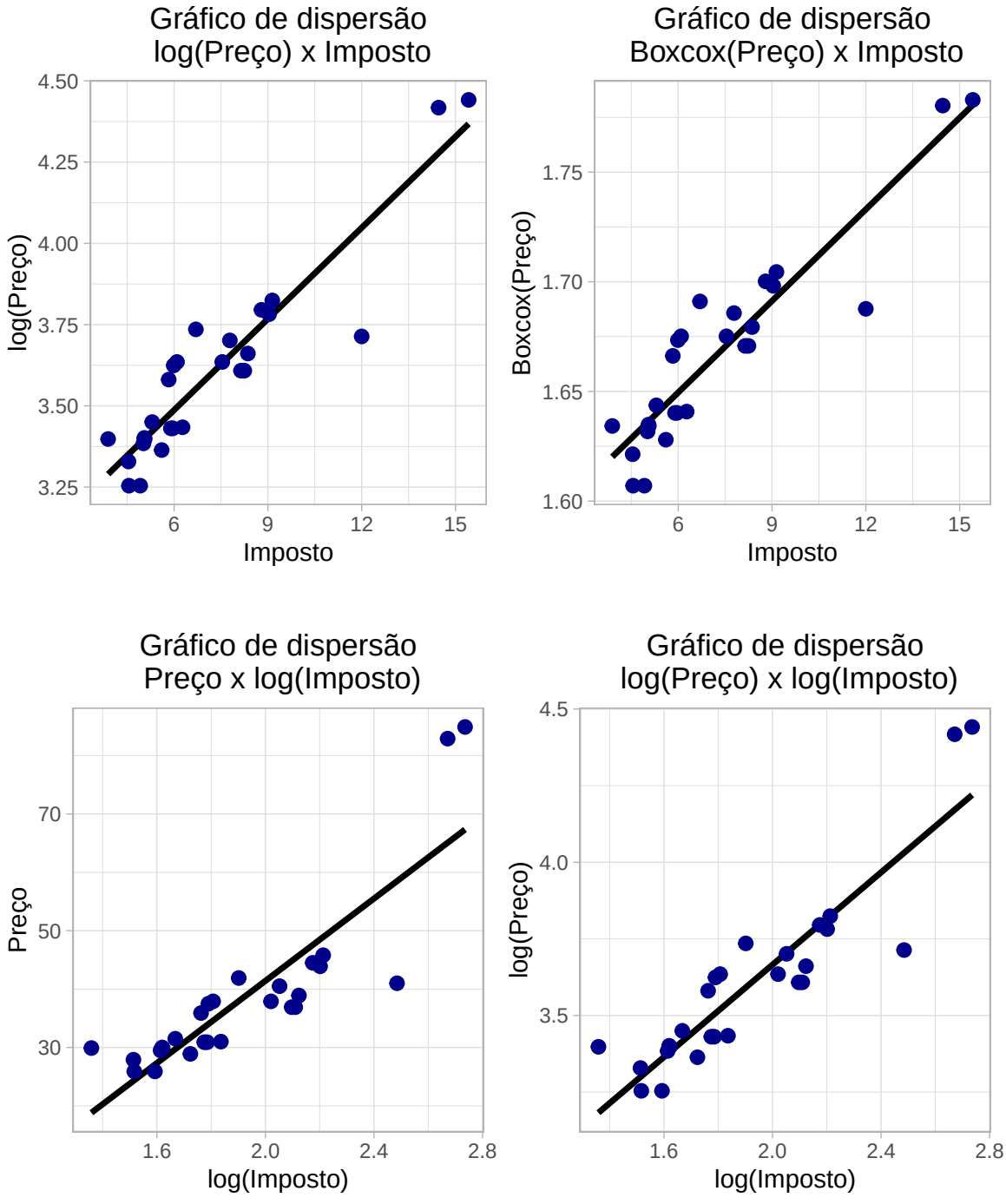
Analisando a relação entre a variável resposta Preço do Imóvel e as covariáveis, destacam-se Imposto e Área construída, que apresentam relações positivas consideravelmente lineares com a variável resposta. No caso das duas, é perceptível a presença dos dois outliers para ambas variáveis, que provavelmente serão classificados como pontos atípicos mais a frente.

A relação entre o preço dos imóveis e a idade não possui uma relação tão linear, sendo novamente perceptível a presença e o impacto dos dois outliers. Posteriormente a resposta será transformada, para tentar linearizar mais essa relação.

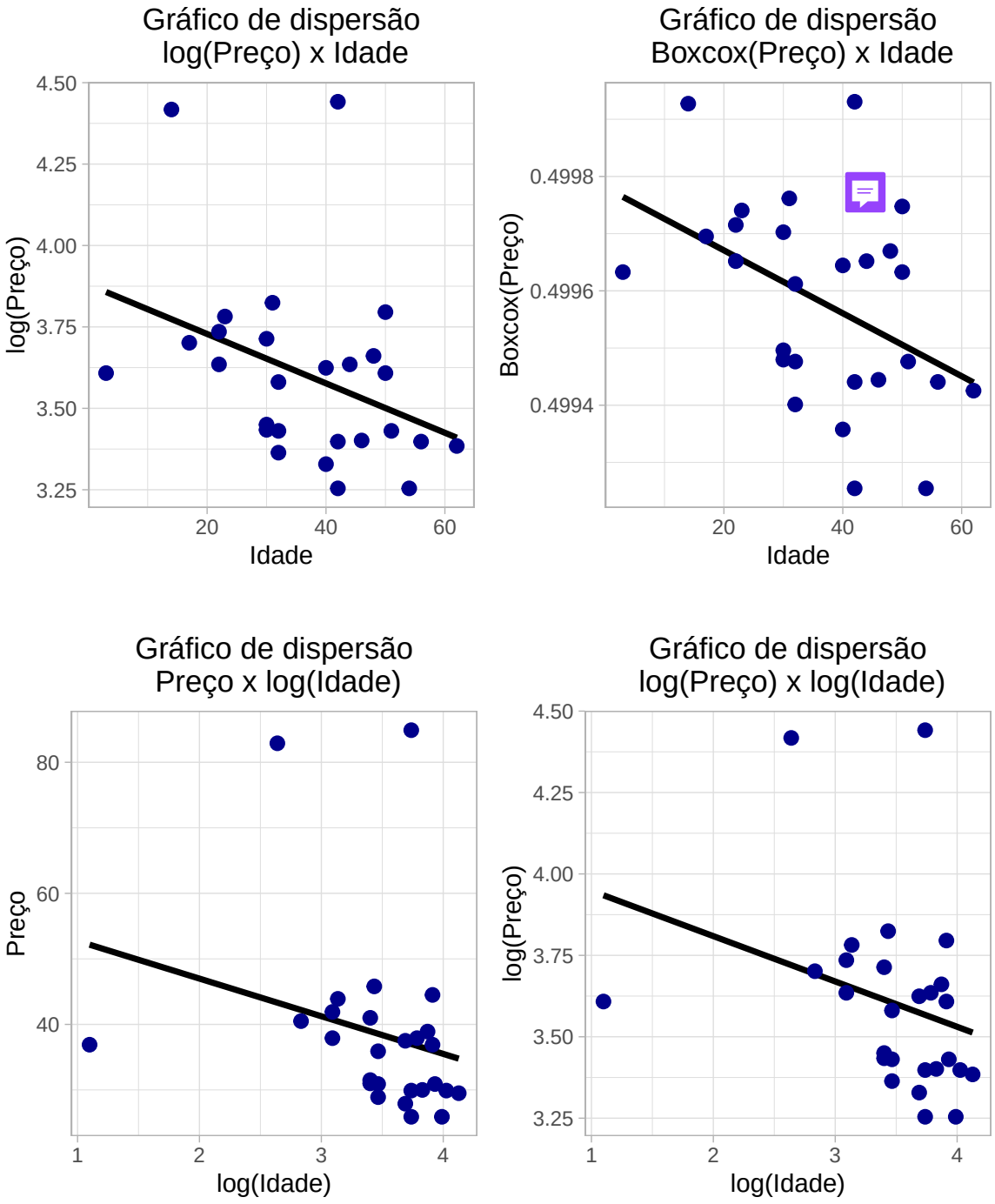
A variável Área do Terreno também apresenta uma relação linear positiva com o Preço dos imóveis, mas também não é tão forte quando a do Imposto ou da Área construída.

2.2.2 Transformações

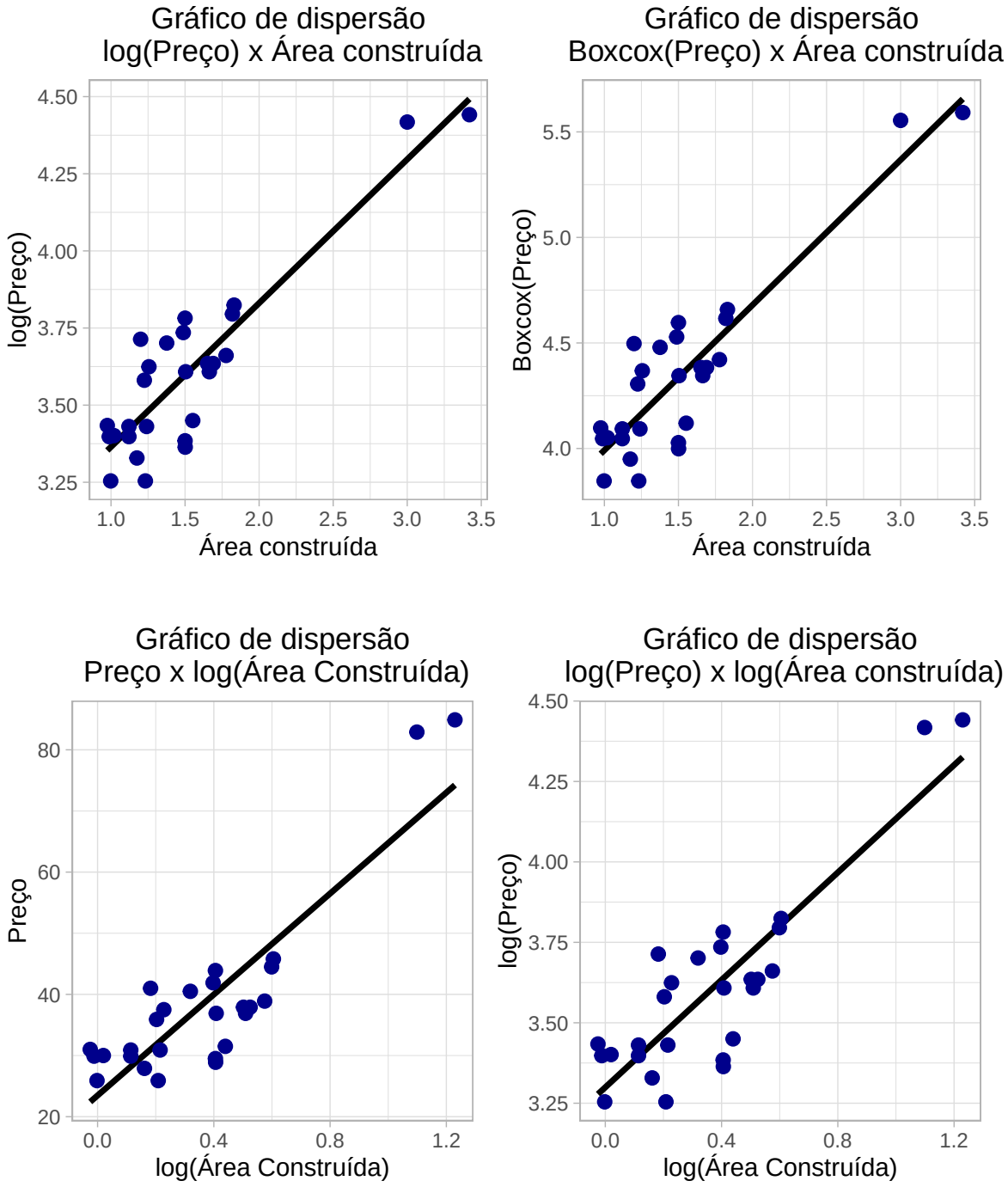
Com o objetivo de linearizar ainda mais a relação entre o Preço dos imóveis e as outras variáveis, aplicaram-se duas transformações na variável resposta, sendo uma a transformação logarítimica, e a outra uma transformação por Boxcox, na qual, por meio de simulação, foram encontrados o lambda ideal para cada relação. Além dessas, também testou-se a transformação logarítimica na covariável e a transformação logarítimica em ambos.



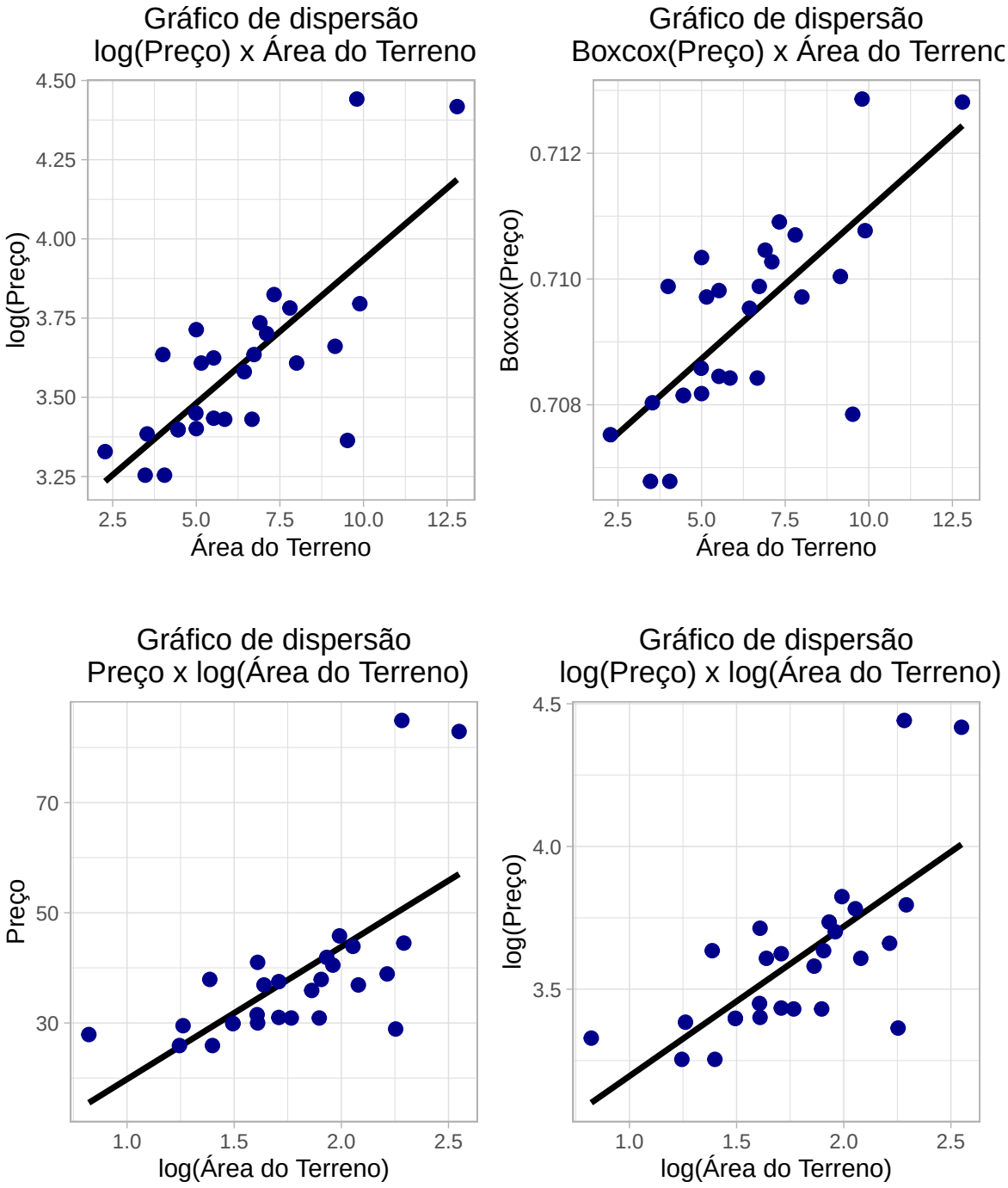
As transformações para a Resposta nesse caso conseguiriam linearizar ainda mais a relação entre Preço e o Imposto, com a transformação de Boxcox para $\lambda = -0,5$ apresentando uma relação aparentemente ainda melhor que a por log. Por outro lado, as relações com base no log do Imposto não demonstraram uma melhora significante. Assim, serão ajustados modelos para ambas transformações da resposta com a covariável original, além do ajuste de um modelo considerando a relação original.



Nesse caso, a transformação logarítimica não conseguiu linearizar tão bem a relação, tanto aplicada na resposta, na covariável ou em ambas, enquanto que a transformação por Boxcox, com $\lambda = -2$, obteve um resultado melhor. Nesse sentido, utilizando a covariável Idade, será ajustado apenas o modelo transformado por Boxcox.



A respeito da relação da resposta com a covariável Área construída, temos que a relação original já estava bem linearizada, e as transformações apenas na resposta mantiveram praticamente com a mesma relação, sendo utilizado $\lambda = 0,1$ para a transformação Boxcox. Nos casos em que se utilizou transformação logarítimica na covariável, não obteve-se uma melhora tão boa quanto as anteriores, mas ainda assim melhoraram a relação original. Dessa forma, serão ajustados modelos para a relação original e também para estas quatro transformações, considerando possíveis complicações relacionadas às suposições que podem surgir no modelo baseado na relação original.



Sobre a relação da variável Resposta com a covariável Área do Terreno, as duas transformações aplicadas somente à resposta obtiveram sucesso em linearizar mais a relação original, enquanto que as relações considerando transformação na covariável não obtiveram o mesmo sucesso. Nesse contexto, temos que a transformação da resposta via boxcox com $\lambda = -1,4$ conseguiu uma relação bem mais linear que a transformada por log(preço). Assim, utiliando a covariável Área do Terreno, será ajustado apenas o modelo transformado por Boxcox.

Dessa forma, temos 10 modelos candidatos para serem ajustados, os quais são apresentados na tabela abaixo:

Tabela 18: Modelos candidatos

Número	Resposta	Covariável
1	Preço	Imposto
2	log(Preço)	Imposto
3	Boxcox(Preço)	Imposto
4	Boxcox(Preço)	Idade
5	Preço	Área construída
6	log(Preço)	Área construída
7	Boxcox(Preço)	Área construída
8	Preço	log(Área construída)
9	log(Preço)	log(Área construída)
10	Boxcox(Preço)	Área do Terreno

2.2.3 Modelo Preço x Imposto

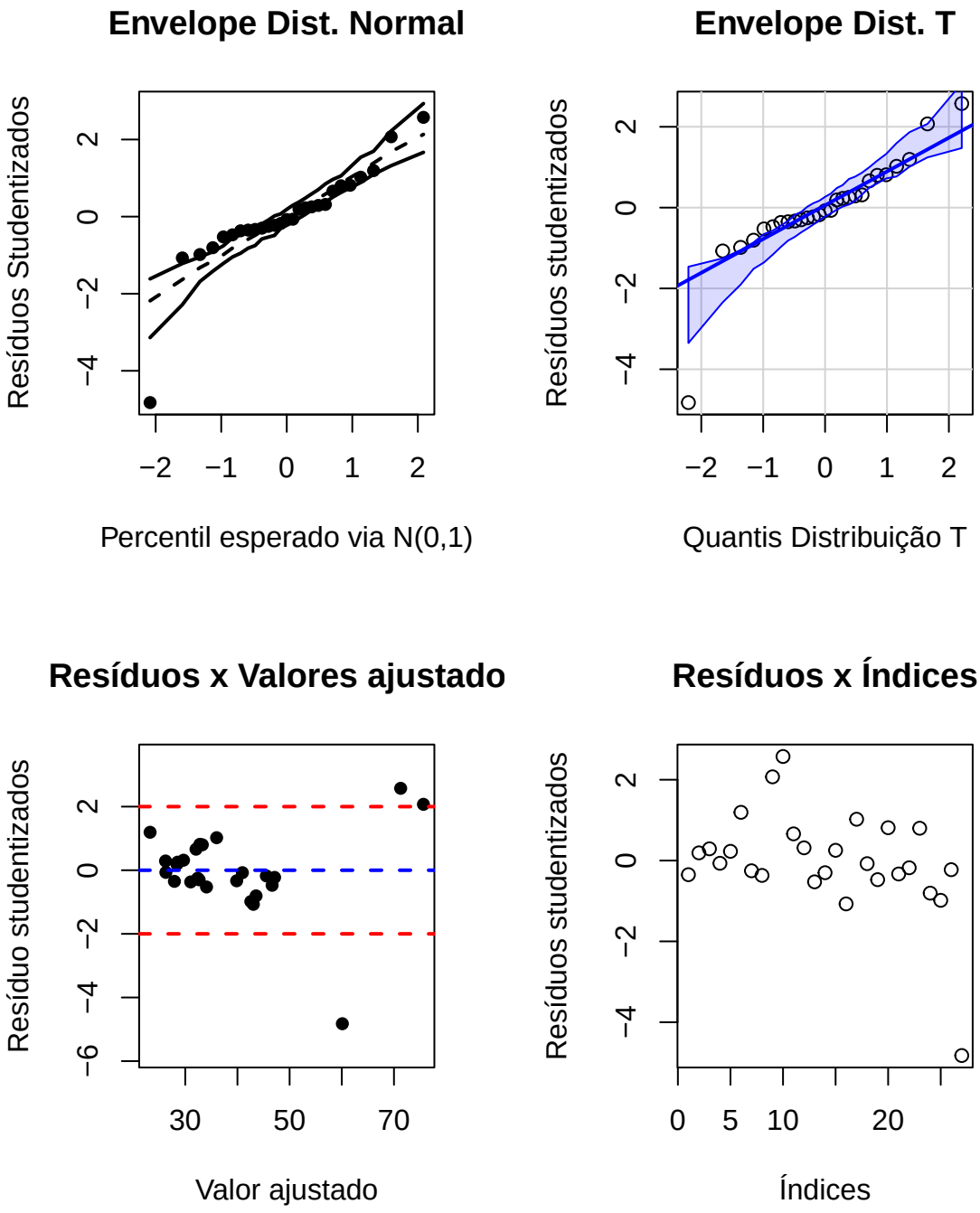
2.2.3.1 Ajuste do modelo

Tabela 19: Ajuste do modelo Preço x Imposto

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	5,583305	3,110721	1,7949	0,0848
imposto	4,543530	0,399978	11,3595	< 0,0001
R^2	0,837700			
R^2 Ajustado	0,831200			

A partir da tabela, temos que o modelo utilizando a resposta original e a covariável Imposto apresentou um bom ajuste, apesar de considerar significativo apenas o coeficiente associado à variável explicativa, mas ainda assim manteremos o intercepto por questões inferenciais. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 83% da variabilidade da resposta, o que é um ótimo resultado.

2.2.3.2 Análise das suposições



Tratando das suposições sob o modelo, é perceptível pelos gráficos de envelope que há uma considerável parcela de observações fora do envelope simulado, indicando que provavelmente há um desvio dessa suposição.

Em relação à homoscedasticidade, percebe-se uma tendência à maior variação dos dados à medida que o valor ajustado aumenta, indicando que provavelmente a variância dos resíduos não seja a mesma. Destaca-se ainda que os pontos 9, 10 e 27 são considerados como possíveis aberrantes, e provavelmente obtiveram um ajuste da resposta ruim.

A respeito da não correlação, não é perceptível uma grande tendência em relação à coleta dos dados.

Tabela 20: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,0003	Rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,0029	Rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,0580	Não rejeita hipótese nula

A partir da aplicação dos testes, temos que, sob um nível de significância de 5%, há evidências estatísticas suficientes para rejeitar as hipóteses nula tanto relacionadas com a normalidade dos resíduos quando com sua homoscedasticidade. Nesse sentido, e também pelo que pôde ser visto nos gráficos, o modelo em questão não cumpre com suposições necessárias para sua inferência, e portanto, sua análise encerra nesse ponto.

2.2.4 Modelo log(Preço) x Imposto

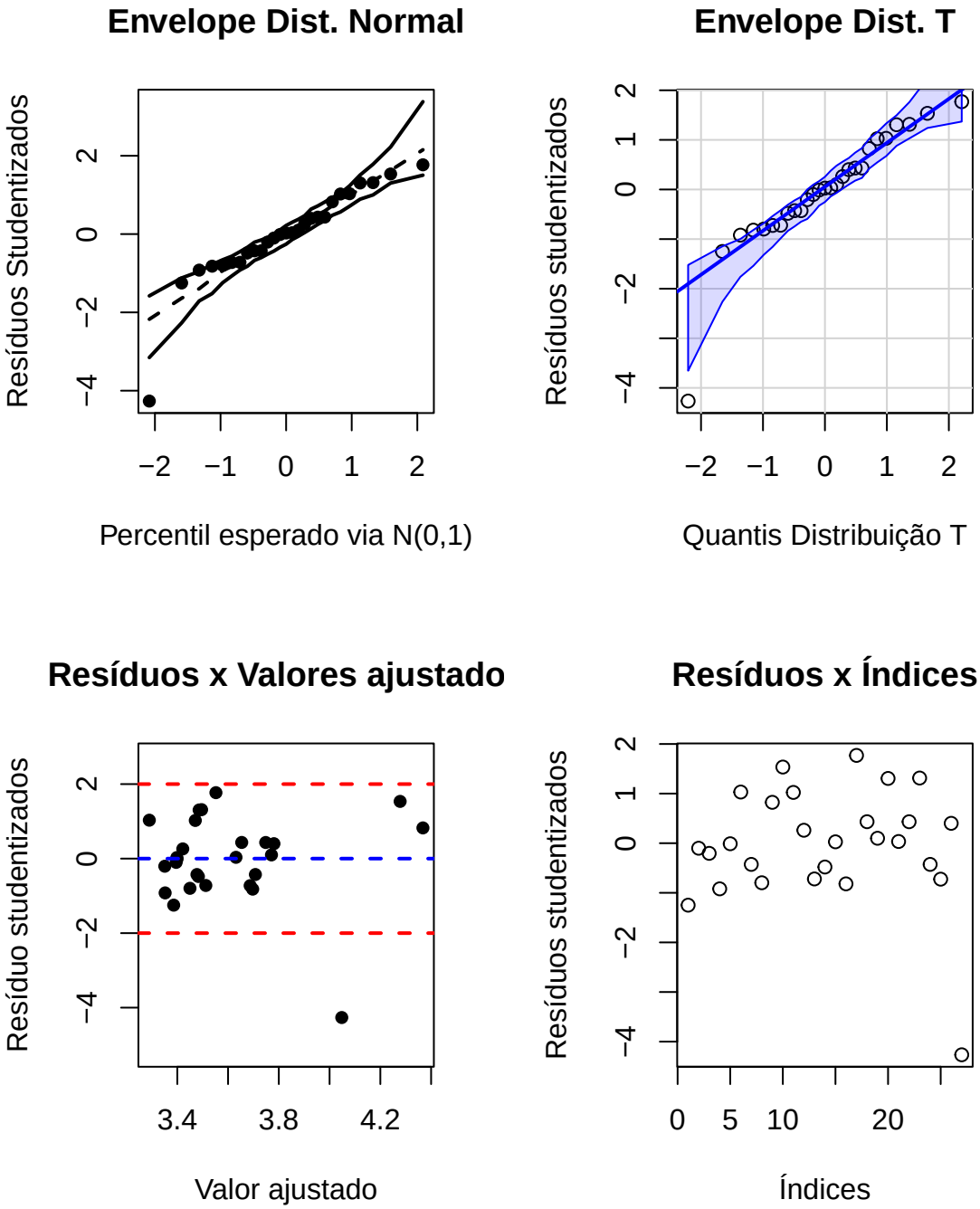
2.2.4.1 Ajuste do modelo

Tabela 21: Ajuste do Modelo log(Preço) x Imposto

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	2,925999	0,05818	50,2918	< 0,0001
imposto	0,093499	0,007481	12,4984	< 0,0001
R^2	0,862000			
R^2 Ajustado	0,856500			

A partir da tabela, temos que o modelo utilizando a resposta transformada via logarítimico apresentou um ótimo ajuste, considerando os dois coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 86% da variabilidade da resposta, o que é um resultado ainda melhor que o obtido pela resposta original e a mesma covariável.

2.2.4.2 Análise das suposições



A respeito das suposições, tratando-se da normalidade dos erros, novamente nota-se algumas observações fora do envelope, menos que do que foi percebido no modelo anterior, mas ainda assim preocupante para a adequação à suposição de normalidade.

Em relação à homoscedasticidade, é perceptível a mesma tendência de aumento da variação dos resíduos à medida que o valor ajustado aumenta, indicando uma possível diferença de variâncias. Destaca-se neste modelo apenas a observação 27 como provável aberrante, indicando que houve uma melhora no ajuste dos pontos 9 e 10 em comparação com o modelo anterior.

Tratando da não correlação dos resíduos, não percebe-se nenhuma tendência em relação ordem de coleta dos dados.

Tabela 22: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,0027	Rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,1235	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,1708	Não rejeita hipótese nula

A partir dos testes aplicados, novamente são apresentadas evidências estatísticas suficientes para acreditar que o modelo não está adequado para as suposições que devem feitas sobre ele, e portanto, sua análise se encerra nesse ponto.

2.2.5 Modelo Boxcox(Preço) x Imposto

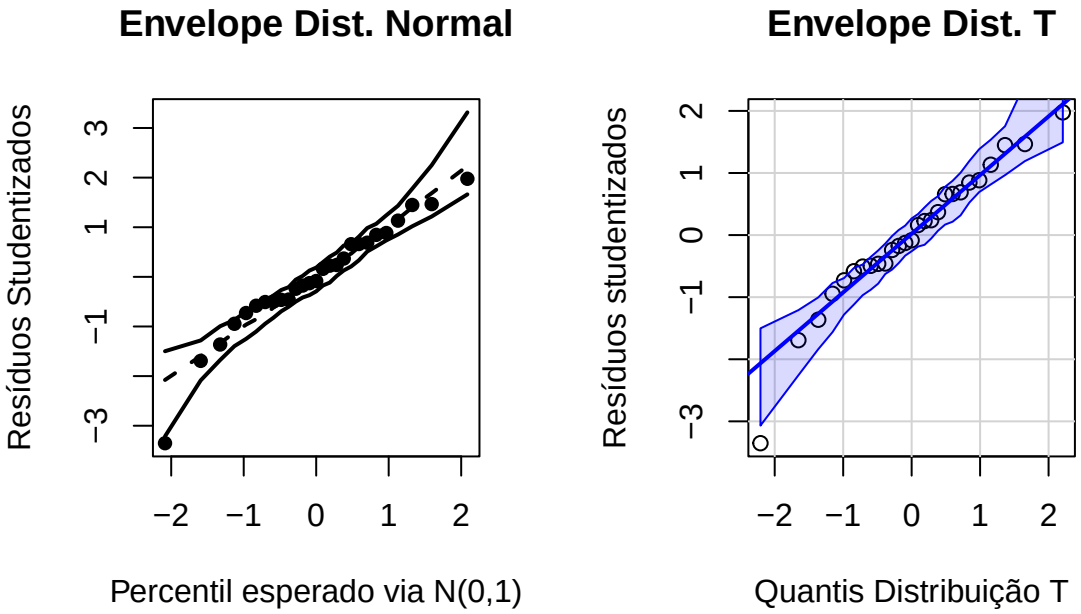
2.2.5.1 Ajuste do modelo

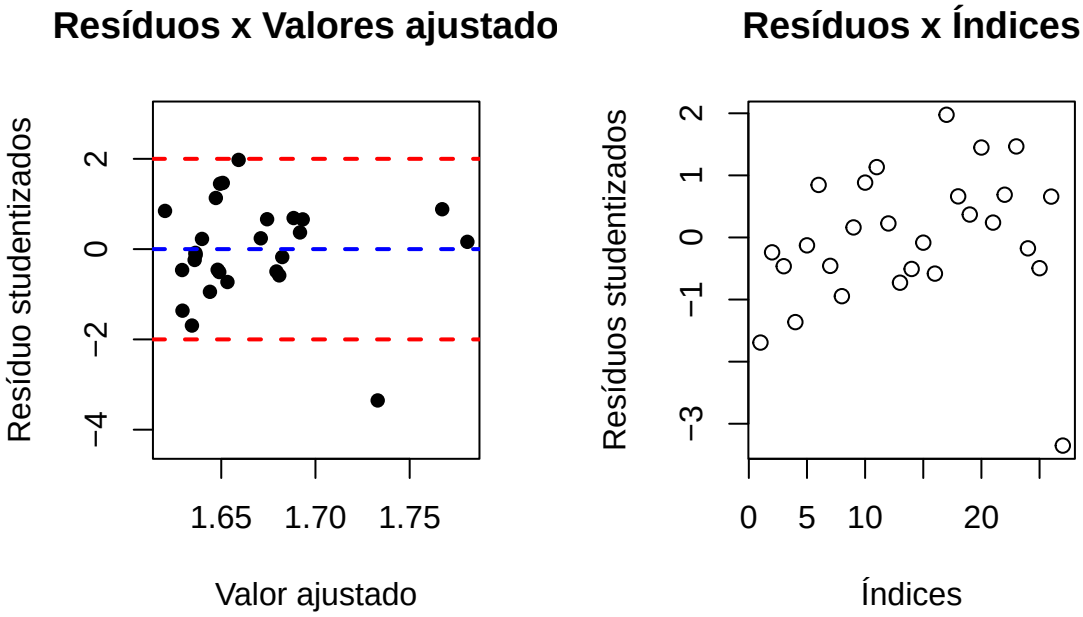
Tabela 23: Ajuste do modelo Boxcox(Preço) x Imposto

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	1,565994	0,009179	170,6065	< 0,0001
imposto	0,013919	0,00118	11,7935	< 0,0001
R^2	0,847600			
R^2 Ajustado	0,841500			

A partir da tabela, temos que o modelo utilizando a resposta transformada via Boxcox com $\lambda = -0,5$ apresentou um ótimo ajuste, considerando ambos os coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 84% da variabilidade da resposta, que é um ótimo indicador, sendo próximo do atingido pelo modelo do log da resposta.

2.2.5.2 Análise das suposições





Em relação à suposição de normalidade dos erros, este modelo obteve um resultado melhor que os dois modelos anteriores, apresentando apenas uma observação realmente fora do envelope em ambos gráficos, o que pode ser considerado aceitável.

Tratando da suposição de homoscedasticidade, percebe-se um aumento da variabilidade dos resíduos à medida que o valor ajustado aumenta, apresentando o ponto 27 como possível aberrante, comportamento que já pôde ser notado nos modelos anteriores.

Sobre a não correlação dos resíduos, agora é possível observar uma certa tendência crescente no valor dos resíduos studentizados, indicando um indício contra a suposição de não-correlação dos resíduos.

Tabela 24: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,2248	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,3679	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,0933	Não rejeita hipótese nula

Apesar de todos os testes aplicados não rejeitarem a hipótese nula, opta-se por descartar o modelo em questão devido ao desvio da suposição de não correlação apresentado no gráfico.

2.2.6 Modelo Boxcox(Preço) x Idade

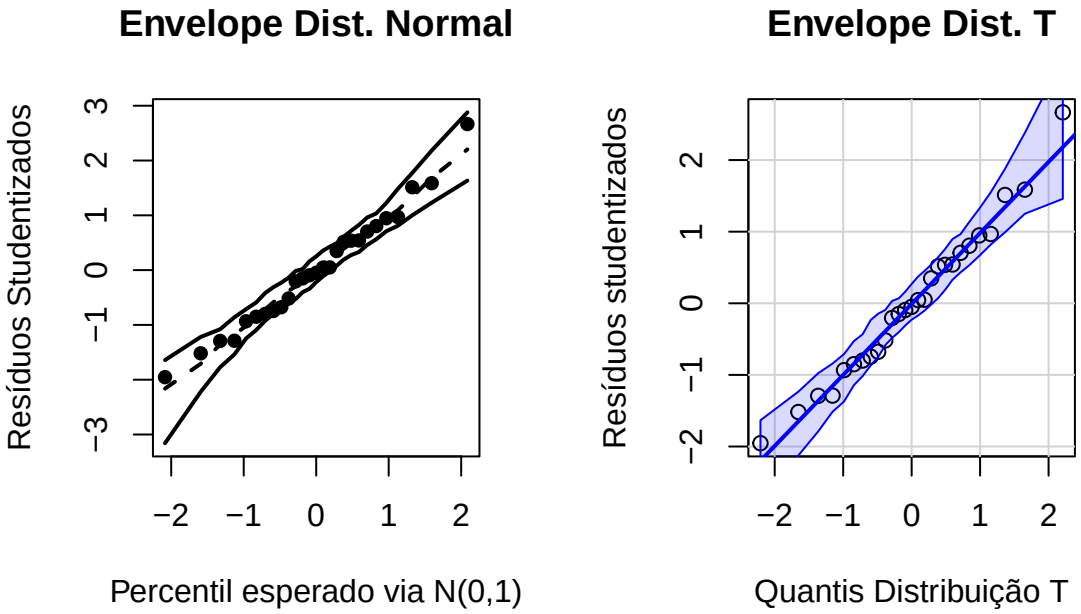
2.2.6.1 Ajuste do modelo

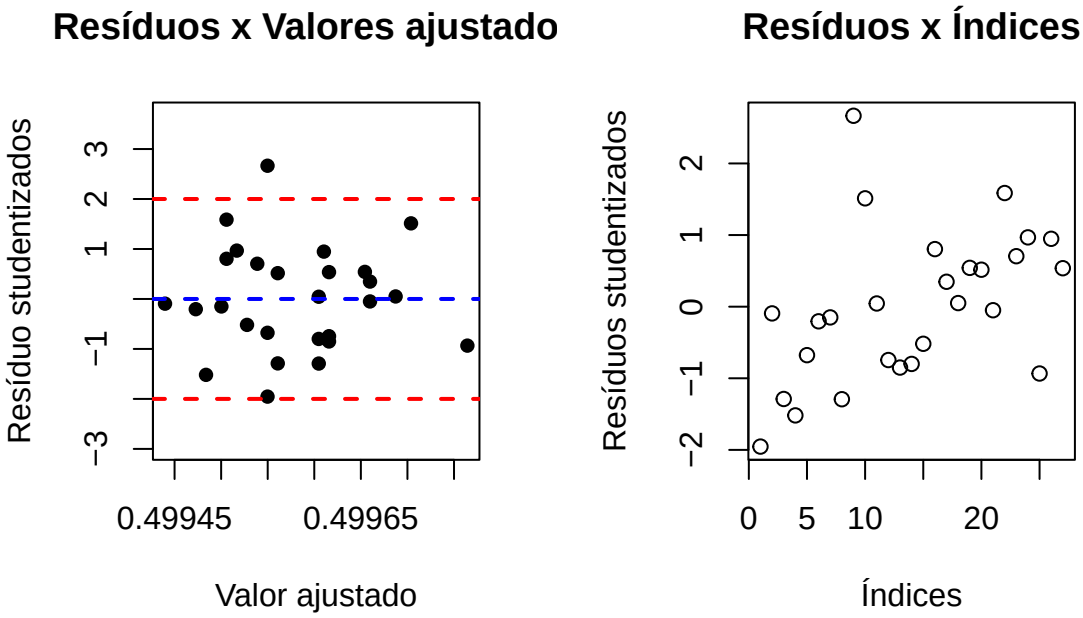
Tabela 25: Ajuste do modelo Boxcox(preço) x Idade

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	0,499781	0,000089	5637,1501	< 0,0001
idade	-0,000006	0,000002	-2,42	0,0231
R^2	0,189800			
R^2 Ajustado	0,157400			

A partir da tabela, temos que o modelo utlizando a variável resposta transformada via Boxcox com $\lambda = -2$ apresentou um ajuste não tão bom, apesar de considerar os dois coeficientes significativos, a um nível de significância de 5%. O modelo apresente um valor baixo para R^2, indicando que este só consegue explicar menos de 20% da variabilidade da resposta.

2.2.6.2 Análise das suposições





Tratando da normalidade dos erros, não há nenhum ponto que foge consideravelmente para fora do envelope, evidenciando nenhum tipo de possível desvio à essa suposição.

Em relação à homoscedasticidade, percebe-se os resíduos distribuídos de forma bem mais aleatória, sem apresentar nenhuma tendência clara, destacando-se apenas um ponto como possível aberrante, a observação 9.

Sobre a não correlação dos resíduos, é perceptível uma tendência crescente dos resíduos studentizados com relação à ordem de coleta dos dados, sendo portanto um forte indício de desvio da suposição de não-correlação.

Tabela 26: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,9219	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,2263	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,0800	Não rejeita hipótese nula

Apesar dos testes aplicados não rejeitarem as suas respectivas hipóteses nulas, opta-se por descartar o modelo em questão devido ao forte indício de desvio da suposição de não correlação dos resíduos.

2.2.7 Modelo Preço x Área Construída

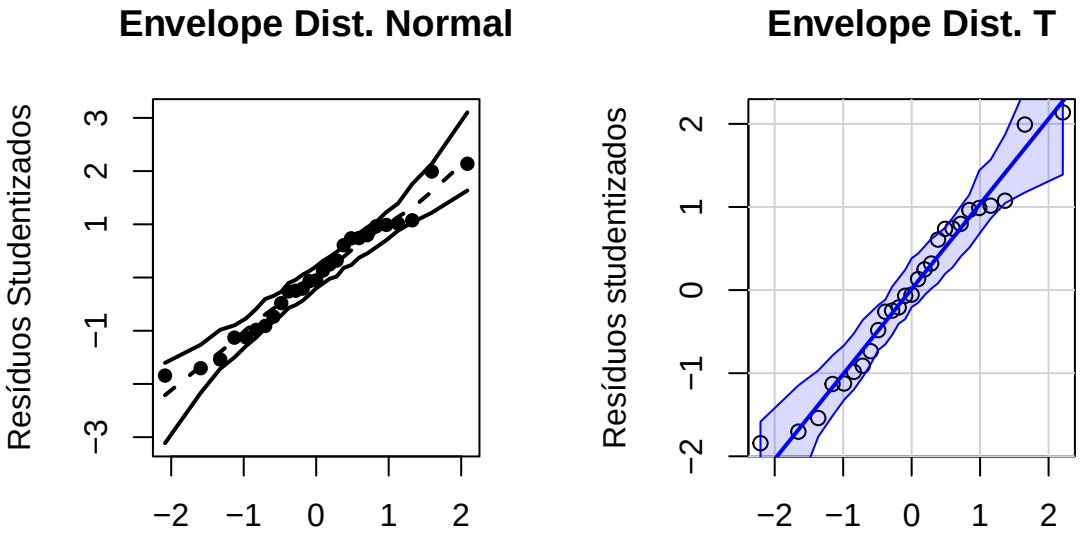
2.2.7.1 Ajuste do modelo

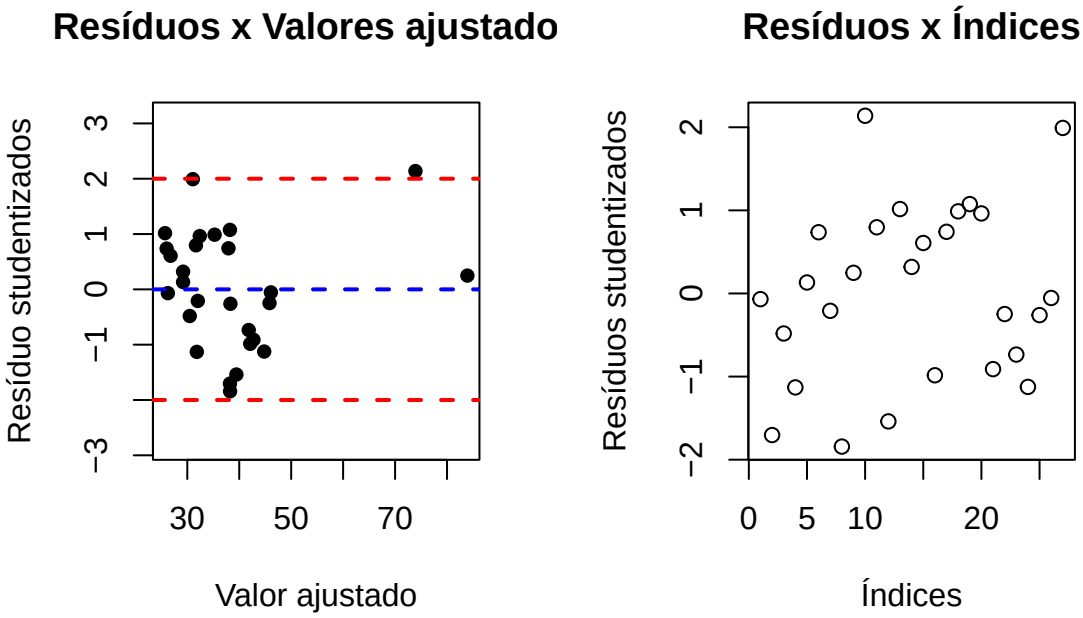
Tabela 27: Ajuste do modelo Preço x Área construída

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	2,50592	3,054311	0,8205	0,4197
areaC	23,80444	1,899146	12,5343	< 0,0001
R^2	0,86270			
R^2 Ajustado	0,85720			

A partir da tabela, temos que o modelo utilizando a resposta original e a covariável Área construída apresentou um ótimo ajuste, apesar de considerar significativo apenas o coeficiente associado à variável explicativa, mas ainda assim manteremos o intercepto por questões inferenciais. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 86% da variabilidade da resposta, o que é um ótimo resultado, sendo o modelo que melhor conseguiu explicar a resposta até o momento.

2.2.7.2 Análise das suposições





Tratando-se sobre a normalidade dos erros, não percebe-se nenhum ponto muito distante ou fora do envelope, indicando que não há indícios de um desvio dessa suposição, ainda mais considerando a distribuição exata T.

Em relação à homoscedasticidade, percebe-se uma maior homogeneidade dos resíduos e alguns valores que destoam com relação ao aumento do valor ajustado, além de apresentar um possível ponto aberrante, a observação 10.

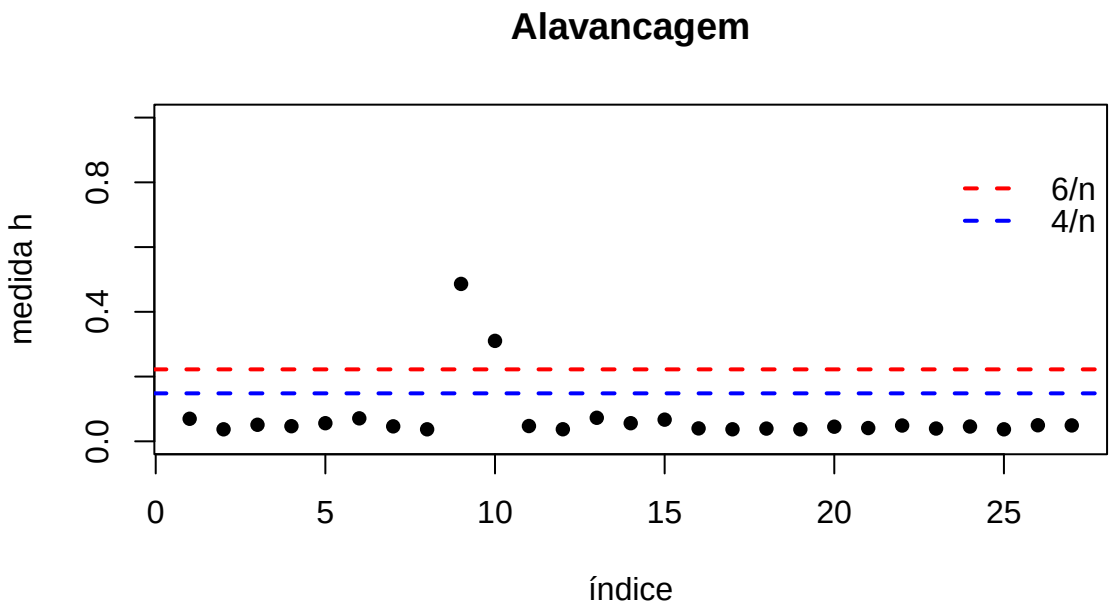
A respeito da não correlação dos erros, percebe-se uma certa tendência crescente, mas não é tão acentuada ao ponto considerar como um desvio à essa suposição.

Tabela 28: Testes de hipóteses aplicados

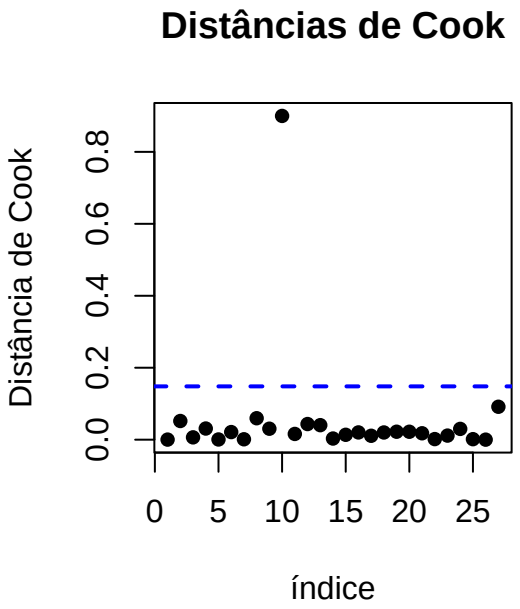
Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,7105	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,7617	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,3045	Não rejeita hipótese nula

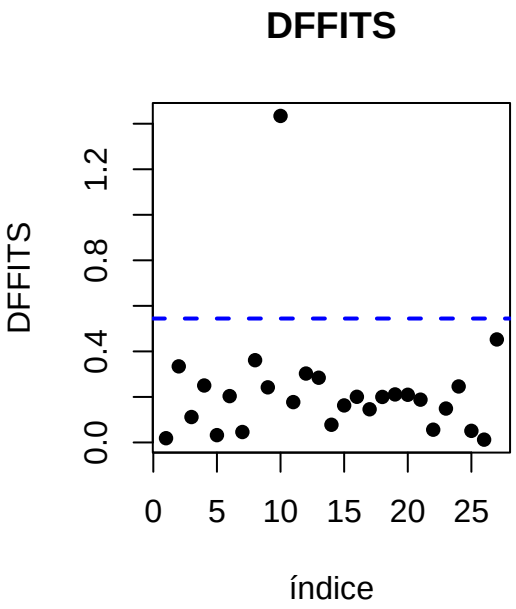
Pelos resultados apresentados pelos testes aplicados, aparentemente não há nenhuma evidência estatística que leve a crer que o modelo ajustado não se adeque às suposições feitas para seu procedimento inferencial. Nesse sentido, segue-se para a análise de alavangem e influência.

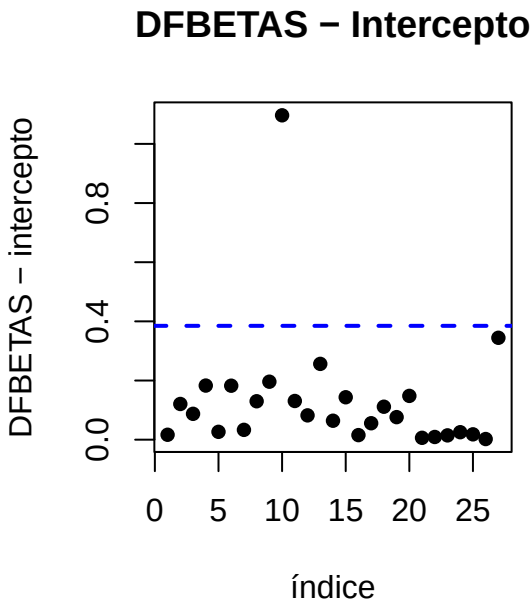
2.2.7.3 Análise de alavancagem e influência

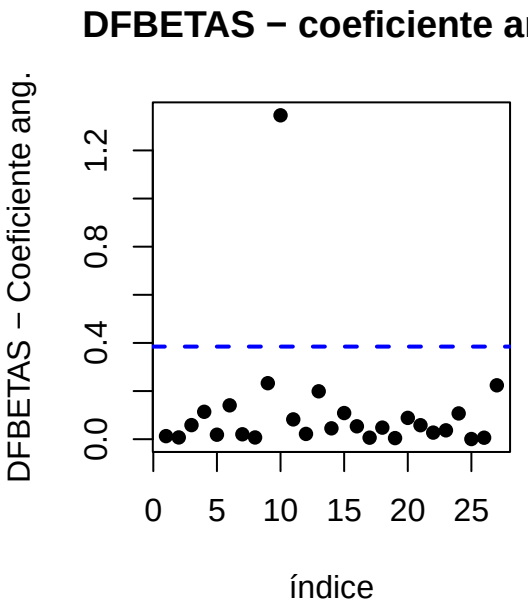


A partir do gráfico acima, que plota as medidas de alavangem dos resíduos, tem-se que os pontos 9 e 10 podem ser considerados como possíveis pontos de alavanca, ou seja, pontos que possuem um peso desproporcional em relação ao seu valor ajustado, muito provavelmente por serem considerados valores discrepantes com relação à área construída dos imóveis.









Em relação à pontos influentes, destaca-se em todos os gráficos apenas a observação 10, que curiosamente, também se destacou como um possível aberrante e também como possível ponto de alavanca, indicando claramente que tal observação não obteve um bom ajuste e também pode causando uma certa influência nas estimativas dos coeficientes.

2.2.7.4 Investigação dos pontos atípicos

A partir das técnicas aplicadas anteriormente, destacaram-se os pontos 9 e 10 como pontos atípicos, que então serão investigados por meio do ajuste de modelos retirando esses pontos.

Tabela 29: Estimativas dos modelos retirando os pontos atípicos

Pontos	Beta 0	Mudança no Beta 0	Beta 1	Mudança no Beta 1
Com todos pontos	2,506	0%	23,804	0%
Retirando 9	3,116	24,3%	23,354	-1,89%
Retirando 10	5,639	125,0%	21,414	-10,0%
Retirando 9 e 10	14,879	493,7%	14,524	-38,9%

Tabela 30: P-valores dos modelos retirando os pontos atípicos

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Com todos pontos	0,419	0%	<0,000000001	0%
Retirando 9	0,439	4,68%	0,0000000055	194.835%
Retirando 10	0,091	-78,12%	0,0000000003	11.735%
Retirando 9 e 10	0,005	-98,71%	0,00034	12.101.785.022%

Ao analisar os resultados obtidos pela investigação dos pontos atípicos, observa-se que o valor de β_0 aumenta com a retirada dos pontos, enquanto β_1 tende a diminuir. Essa dinâmica percebida foi tão considerável, que chegou a mudar a significância do p-valor associado ao β_0, indicando nesse sentido que os pontos 9 e 10 de fato exercem uma influência desproporcional na estimativa do intercepto.

Nesse sentido, apesar do modelo ter apresentado um ótimo ajuste e adequação às suposições, ele aparenta ser sensível aos pontos atípicos, e portanto, não possui robustez para explicar a resposta.

2.2.8 Modelo log(Preço) x Área Construída

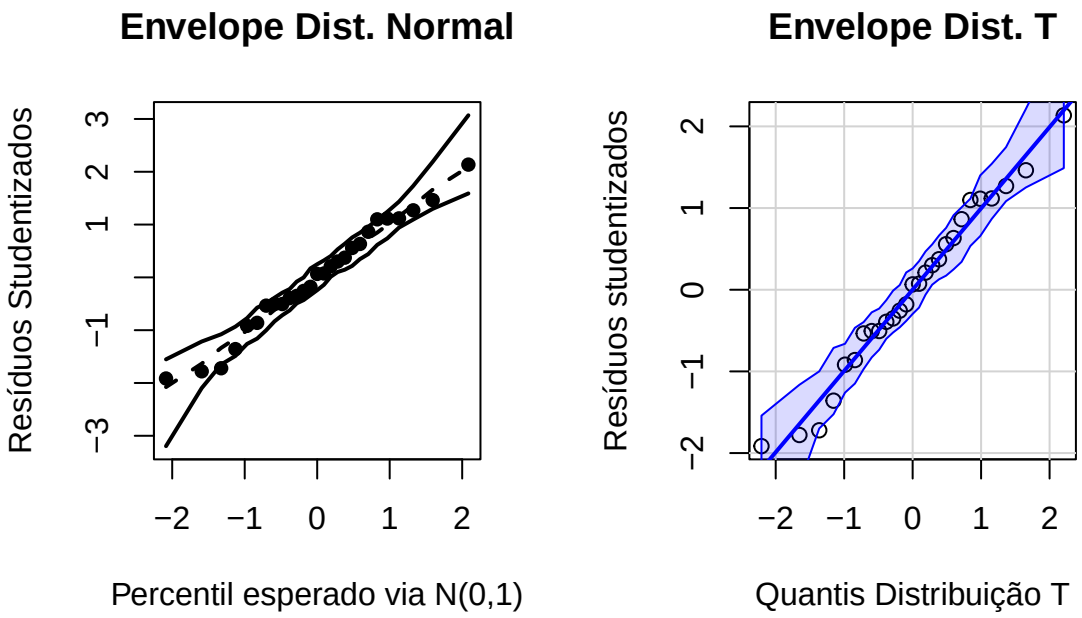
2.2.8.1 Ajuste do modelo

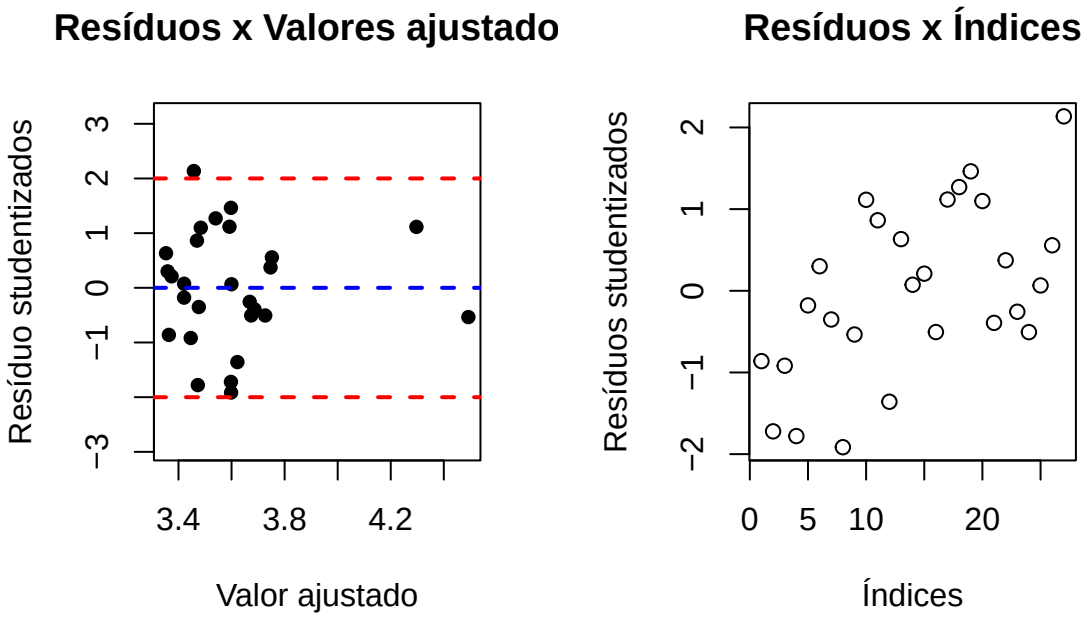
Tabela 31: Ajuste do modelo log(Preço) x Área construída

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	2,898698	0,074124	39,106	< 0,0001
areaC	0,466033	0,04609	10,1114	< 0,0001
R^2	0,803500			
R^2 Ajustado	0,795700			

A partir da tabela, temos que o modelo utilizando a resposta transformada via logarítimico apresentou um ótimo ajuste, considerando os dois coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 80% da variabilidade da resposta, o que é um bom resultado, mas apresenta menos 6% de explicabilidade com relação ao modelo anterior, ajustado considerando a resposta original.

2.2.8.2 Análise das suposições





Em relação à normalidade dos erros, novamente não há nenhum forte indício de desvio dessa suposição, apesar de ser possível visualizar dois pontos que estão posicionados mais na fronteira do envelope.

Sobre a homoscedasticidade dos erros, não nota-se nenhum tipo de comportamento heterogêneo pelos resíduos studentizados, destacando-se apenas os pontos 8 e 27 como possíveis aberrantes.

Tratando sobre a não correlação dos erros, percebe-se claramente uma relação crescente dos resíduos studentizados à medida que foram coletados, indicando um forte indício de que há um desvio contra tal suposição.

Tabela 32: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,8491	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,8424	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,0117	Rejeita hipótese nula

A partir dos resultados obtidos pelos testes aplicados, de fato há evidências estatísticas suficentes para rejeitar a hipótese nula de não correlação dos erros, indicando assim que o modelo ajustado não está adequado com relação às suposições realizadas sob o ajuste do mesmo, e portanto, encerra-se a análise do modelo neste ponto.

2.2.9 Modelo Boxcox(Preço) x Área Construída

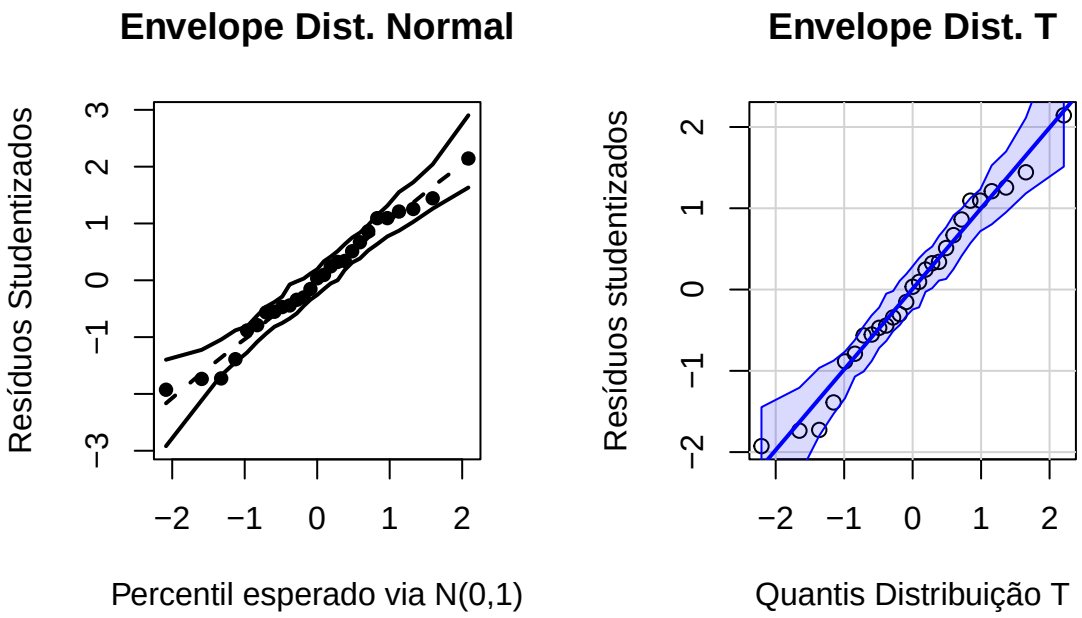
2.2.9.1 Ajuste do modelo

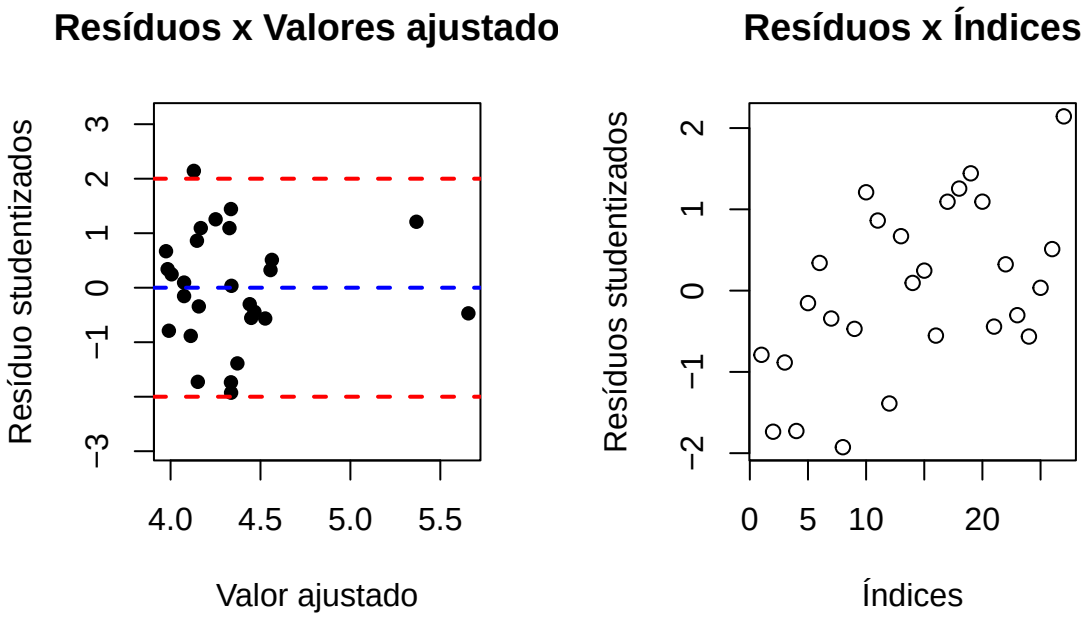
Tabela 33: Ajuste do modelo Boxcox(Preço) x Área construída

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	3,304233	0,106247	31,0995	< 0,0001
areaC	0,687666	0,066064	10,4092	< 0,0001
R^2	0,812500			
R^2 Ajustado	0,805000			

A partir da tabela, temos que o modelo utilizando a resposta transformada via Boxcox com $\lambda = 0,1$ apresentou um ótimo ajuste, considerando os dois coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 81% da variabilidade da resposta, o que é um bom resultado, mas apresenta menos 5% de explicabilidade com relação ao modelo ajustado considerando a resposta original.

2.2.9.2 Análise das suposições





Tratando-se da normalidade dos erros, temos que ambos gráficos de envelope não apresentam nenhum indício de desvio dessa suposição, com todos os pontos consideravelmente inseridos dentro do envelope.

Com relação à homoscedasticidade, novamente não observa-se nenhum tipo de tendência clara de comportamento dos resíduos studentizados, apenas destacando-se as observações 8 e 27 como possíveis aberrantes.

Sobre a não correlação, assim como para o modelo anterior, novamente é perceptível uma tendência crescente dos resíduos studentizados com relação à ordem de coleta dos dados, indicando assim um considerável indício de desvio para essa suposição no modelo.

Tabela 34: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,8481	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,8386	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,0185	Rejeita hipótese nula

Analisando os resultados obtidos pelos testes aplicados, apenas reforça-se o que já havia sido percebido nos gráficos, de que há evidências estatísticas suficientes para rejeitar a hipótese nula de que os erros são não correlacionados, e portanto, temos que modelo não está cumprindo com uma suposição necessária para a sua inferência, e portanto, encerra-se sua análise neste ponto.

2.2.10 Modelo Preço x log(Área Construída)

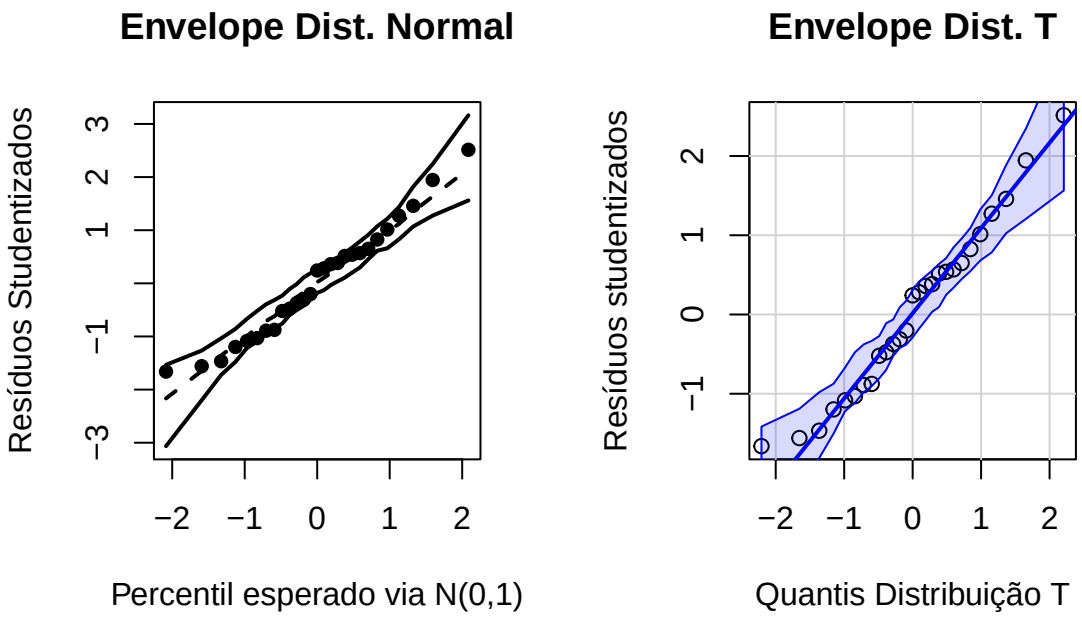
2.2.10.1 Ajuste do modelo

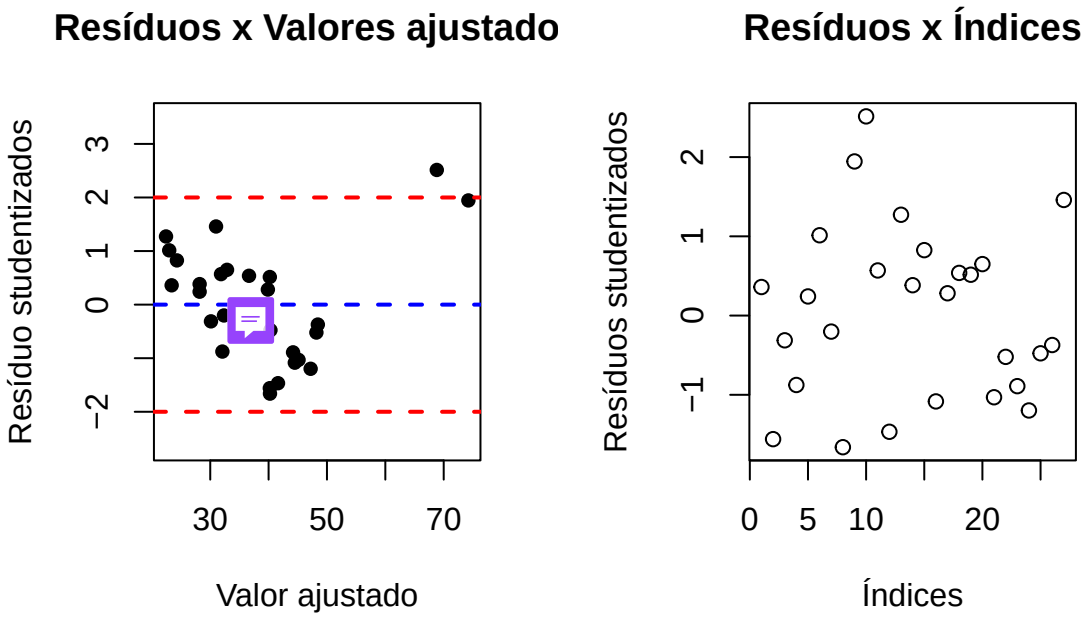
Tabela 35: Ajuste do modelo Preço x log(Área construída)

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	23,46952	2,196152	10,6867	< 0,0001
log_ac	41,27896	4,680865	8,8187	< 0,0001
R^2	0,75670			
R^2 Ajustado	0,74700			

Analisando a tabela acima, temos que o modelo utilizando a resposta original e a covariável transformada via logarítimico apresentou um ajuste muito bom, considerando os ambos coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 75% da variabilidade da resposta, o que é um bom resultado, mesmo sendo um valor abaixo do que os modelos anteriores ajustados para estas variáveis apresentaram.

2.2.10.2 Análise das suposições





Em relação à normalidade dos erros, novamente não há nenhum forte indício de desvio dessa suposição, sem apresentar pontos fora do envelope.

Sobre a homoscedasticidade dos erros, destaca-se apenas o ponto 10 com resíduo studentizado maior que 2, sendo portanto um ponto a ser investigado posteriormente. Percebe-se também um comportamento heterogêneo dos resíduos de forma geral, muito provavelmente devido à presença de outliers para as duas variáveis.

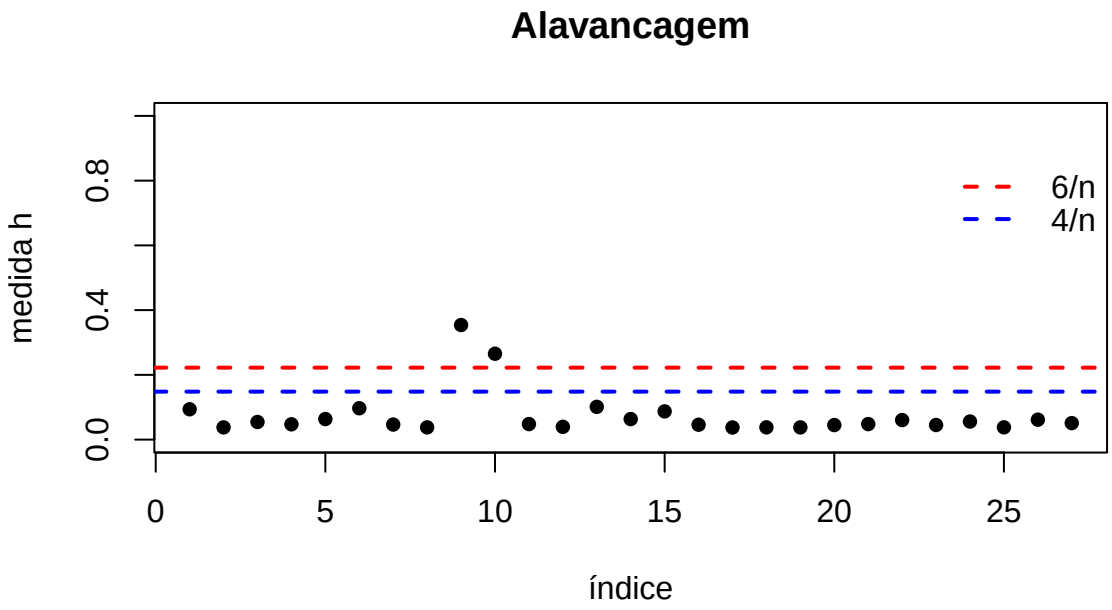
Tratando sobre a não correlação dos erros, percebe-se uma leve tendência crescente dos resíduos studentizados à medida que foram coletados, não sendo esta suficiente para interpretar-se como um desvio à essa suposição.

Tabela 36: Testes de hipóteses aplicados

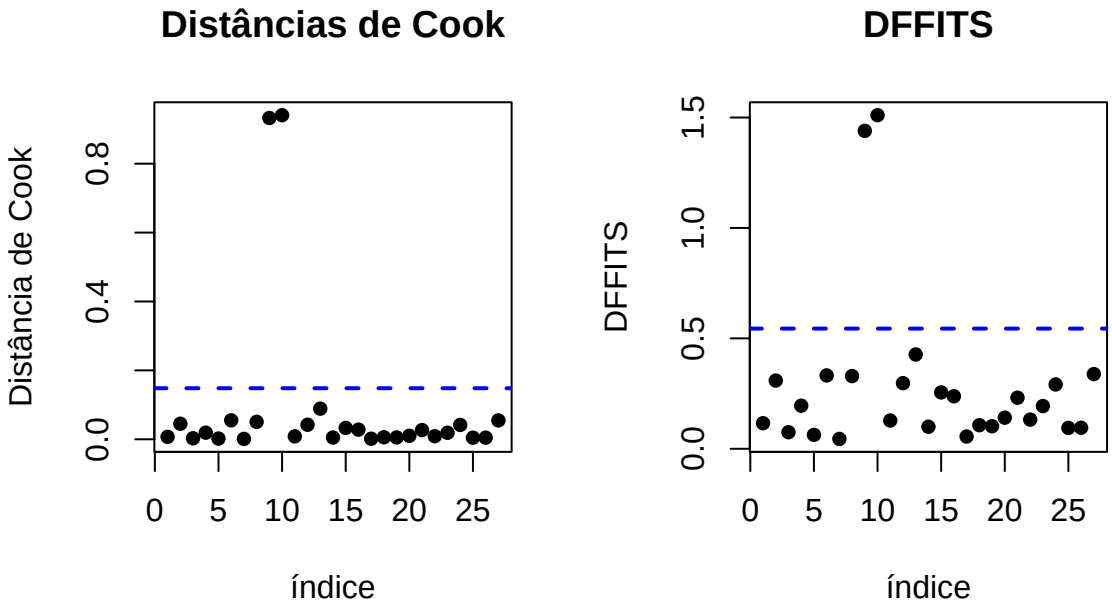
Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,6236	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,8732	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,5382	Não rejeita hipótese nula

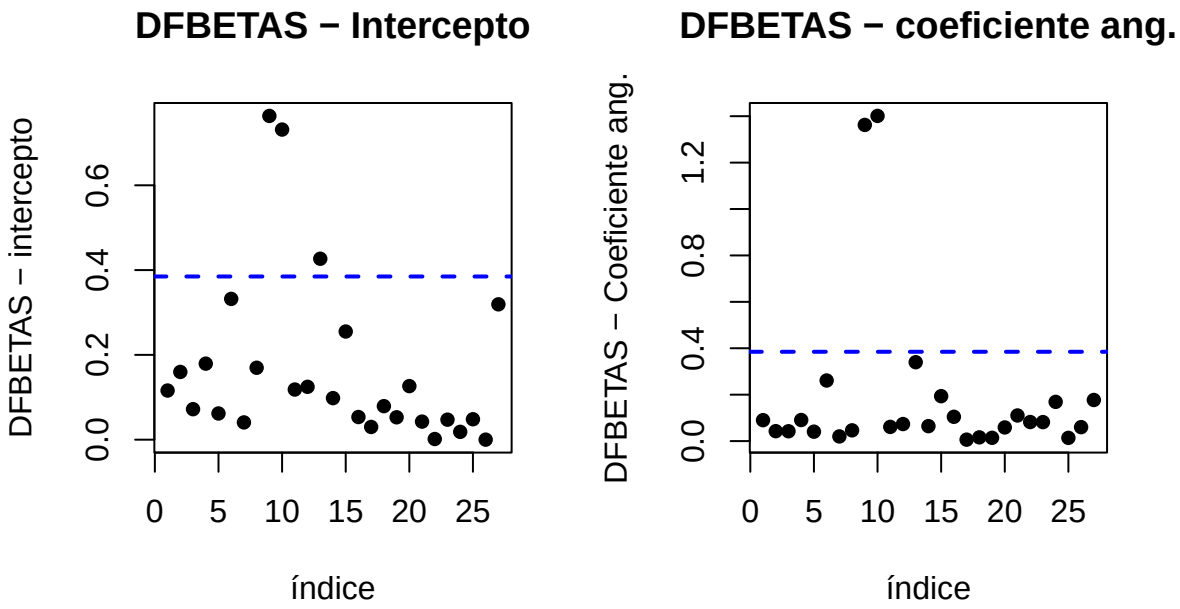
Pelos resultados apresentados pelos testes aplicados, não existe nenhuma evidência estatística indicando que o modelo ajustado não se adeque às suposições feitas para seu procedimento inferencial. Nesse sentido, segue-se para a análise de alavangem e influência.

2.2.10.3 Análise de alavancagem e influência



A partir do gráfico acima, que apresenta as medidas de alavangem dos resíduos, tem-se que os pontos 9 e 10 podem ser considerados como possíveis pontos de alavanca, ou seja, pontos que possuem um peso desproporcional em relação ao seu valor ajustado, muito provavelmente por serem considerados valores discrepantes com relação à área construída dos imóveis, e serão investigados posteriormente.





Em relação à pontos influentes, destaca-se em todos os gráficos as observações 9 e 10, além de também ser possível notar a observação 13 como um possível ponto influente com relação ao intercepto.

2.2.10.4 Investigação

A partir das técnicas aplicadas anteriormente, destacaram-se os pontos 9, 10 e 13 como pontos atípicos, que então serão investigados por meio do ajuste de modelos retirando esses pontos individual e conjuntamente.

Tabela 37: Estimativas dos modelos retirando os pontos atípicos

Pontos	Beta 0	Mudança no Beta 0	Beta 1	Mudança no Beta 1
Com todos pontos	23,4	0%	41,2	0%
Retirando 9	25,0	6,7%	35,2	-14,6%
Retirando 10	24,9	6,2%	35,3	-14,4%
Retirando 13	22,5	-3,9%	42,8	3,8%
Retirando 9, 10 e 13	28,6	21,8%	20,4	-50,3%

Tabela 38: P-valores dos modelos retirando os pontos atípicos

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Com todos pontos	<0,0000000010%		<0,0000000010%	
Retirando 9	<0,000000001	-37,4%	0,000001012	26.536%
Retirando 10	<0,000000001	-84,9%	0,000000168	4.332%
Retirando 13	<0,0000000016	95,8%	0,000000004	7%

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Retirando 9, 10 e 13	<0,000000001	-99,6%	0,000558	14.672.692%

Obsevando os resultados obtidos pela investigação dos pontos, percebe-se que as observações 9 e 10 aumentam o valor do β_0 e diminuem o valor do β_1 quando são retirados, enquanto que a observação 13 apresenta o efeito contrário. Apesar dessa dinâmica, não foi apresentada nenhuma mudança de significância com relação aos coeficientes, indicando que tais pontos não geram uma alteração muito significante nas estimativas do modelo. Dessa forma, conclui-se que o modelo em questão é um forte candidato para explicar a variável resposta, apresentando a robustez necessária.

2.2.11 Modelo log(Preço) x log(Área Construída)

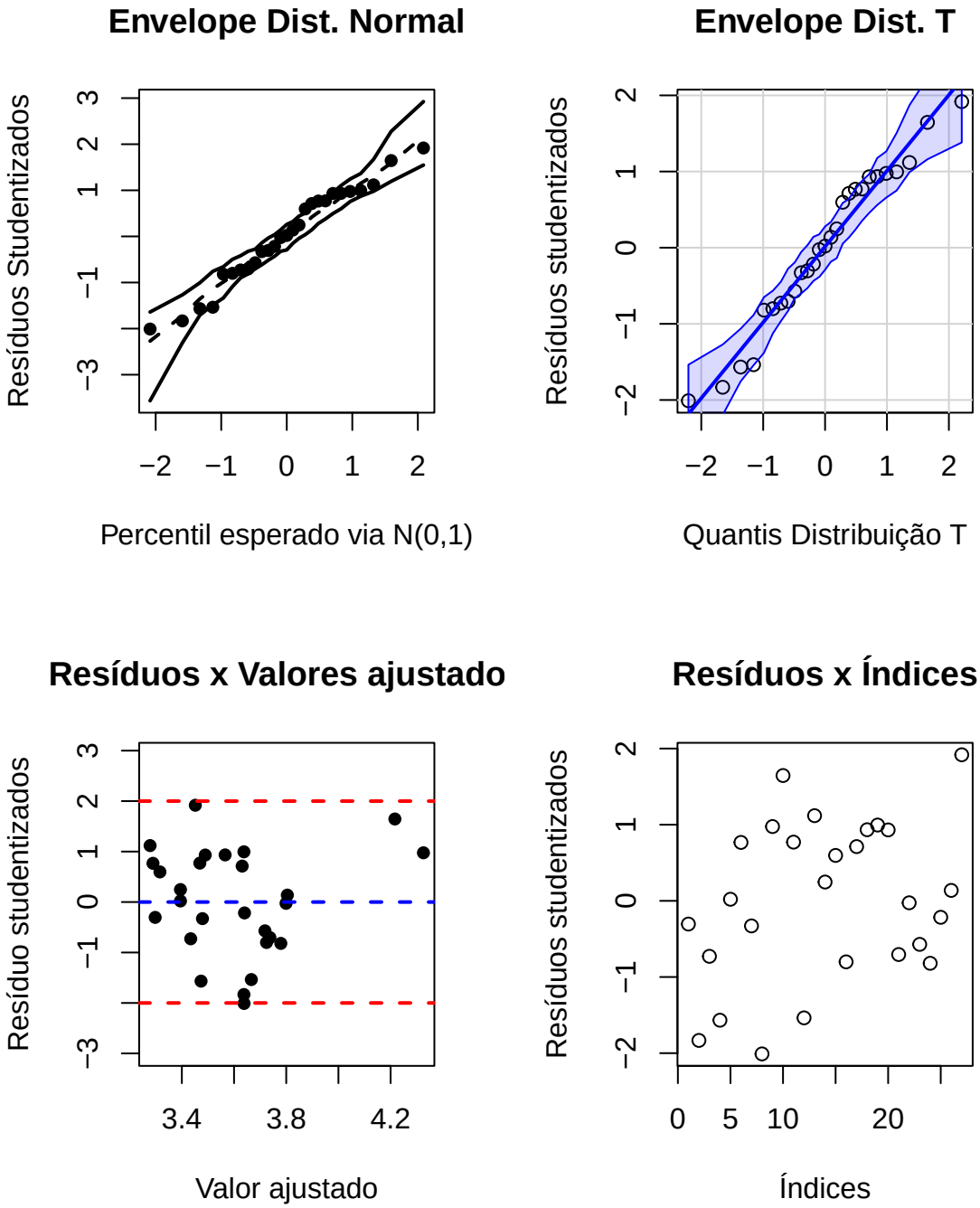
2.2.11.1 Ajuste do modelo

Tabela 39: Ajuste do modelo log(Preço) x log(Área construída)

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	3,299495	0,045015	73,2976	< 0,0001
log(areaC)	0,834557	0,095945	8,6983	< 0,0001
R^2	0,751600			
R^2 Ajustado	0,741700			

Analisando a tabela acima, temos que o modelo utilizando transformação logarítimica tanto na resposta quanto na covariável apresentou um ajuste muito bom, considerando, novamente ambos coeficientes significativos. Destaca-se também o valor de R^2, indicando que o modelo consegue explicar mais de 75% da variabilidade da resposta, o que é um bom resultado, semelhante ao obtido pelo modelo ajustdo anteriormente.

2.2.11.2 Análise das suposições



Em relação à normalidade dos erros, novamente não há nenhum forte indício de desvio dessa suposição, apresentando apenas dois pontos localizados na fronteira do envelope.

Sobre a homoscedasticidade dos erros, destaca-se apenas o ponto 8 com resíduo studentizado maior que 2, sendo portanto um ponto a ser investigado posteriormente. Percebe-se também um comportamento mais aleatório dos resíduos de forma geral, com praticamente todos os resíduos no intervalo de -2 a 2.

Tratando sobre a não correlação dos erros, percebe-se uma tendência crescente mais forte dos resíduos studentizados à medida que foram coletados, sendo esta suficiente para interpretar-se como um considerável indício de desvio à essa suposição.

Tabela 40: Testes de hipóteses aplicados

Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,5231	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,9399	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,2667	Não rejeita hipótese nula

Apesar dos testes aplicados não rejeitarem as suas respectivas hipóteses nulas, opta-se por descartar o modelo em questão devido ao forte indício de desvio da suposição de não correlação dos erros.

2.2.12 Modelo Boxcox(Preço) x Área do Terreno

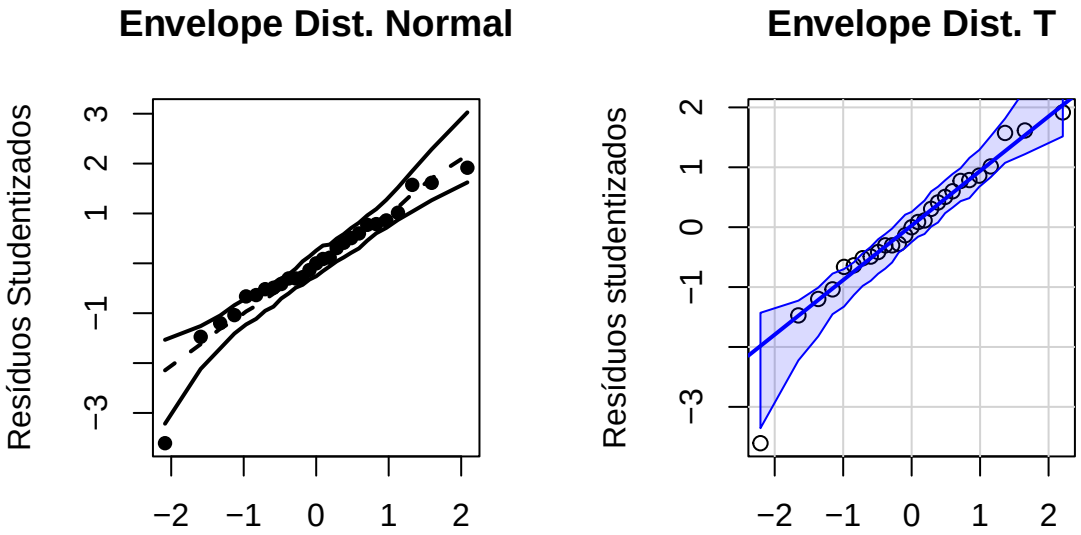
2.2.12.1 Ajuste do modelo

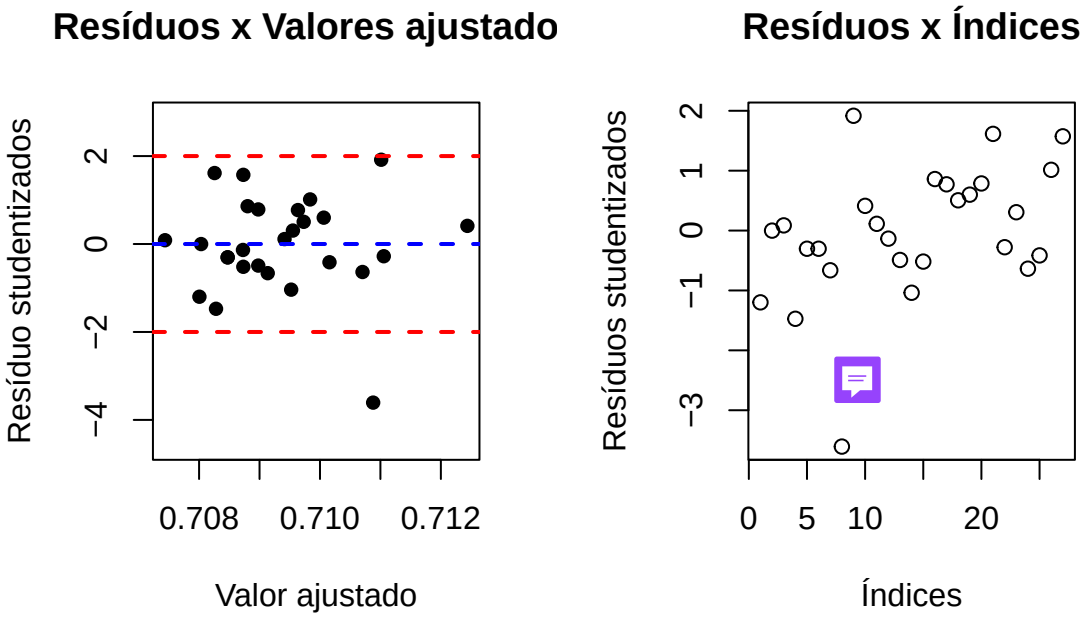
Tabela 41: Ajuste do modelo Boxcox(Preço) x Área do Terreno

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	0,706355	0,000597	1183,9297	< 0,0001
areaT	0,000475	0,000088	5,3943	< 0,0001
R^2	0,537900			
R^2 Ajustado	0,519400			

A partir da tabela, temos que o modelo utlizando a variável resposta transformada via Boxcox com $\lambda = -2$ apresentou um ajuste adequado, considerando ambos coeficientes significativos, a um nível de significância de 5%. O modelo apresenta um valor não tão alto para R^2, indicando que este consegue explicar 50% da variabilidade da resposta, o que está abaixo do que pôde-se observar nos outros modelos ajustados.

2.2.12.2 Análise das suposições





Em relação à normalidade dos resíduos, percebe-se apenas um ponto consideravelmente fora do envelope, em ambos gráficos, mas ainda assim podemos considerar sem um grande indício de desvio de normalidade.

Sobre a homoscedasticidade, percebe-se uma certa homogeneidade distribuição dos resíduos studentizados com relação aos valores ajustados, destacando-se apenas a observação 8 como provável aberrante.

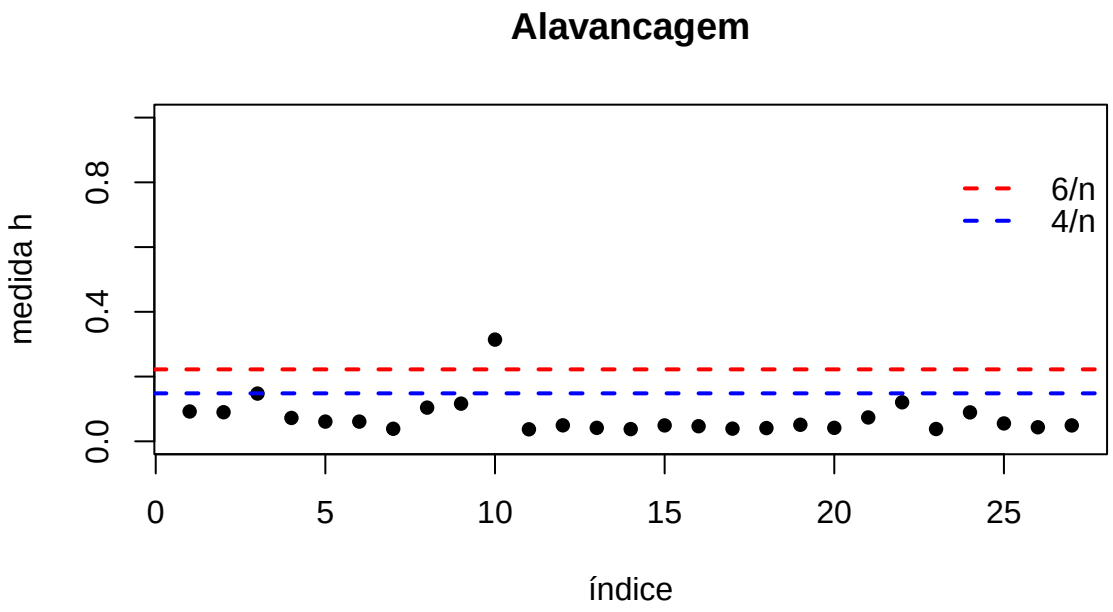
Tratando-se da suposição de não-correlação dos resíduos, não é perceptível uma tendência tão clara de comportamento dos resíduos com base na ordem de coleta da amostra.

Tabela 42: Testes de hipóteses aplicados

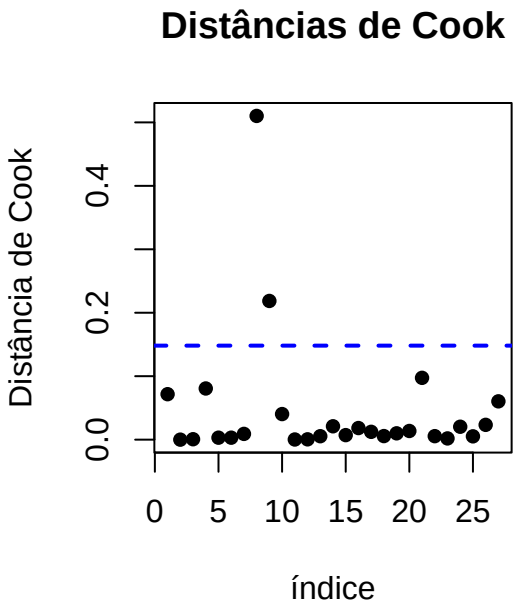
Teste	Suposições	P-valor	Decisão
Shapiro-Wilk	Normalidade	0,0729	Não rejeita hipótese nula
Goldfield-Quandt	Homoscedasticidade	0,2962	Não rejeita hipótese nula
Durbin-Watson	Não-Correlação	0,3082	Não rejeita hipótese nula

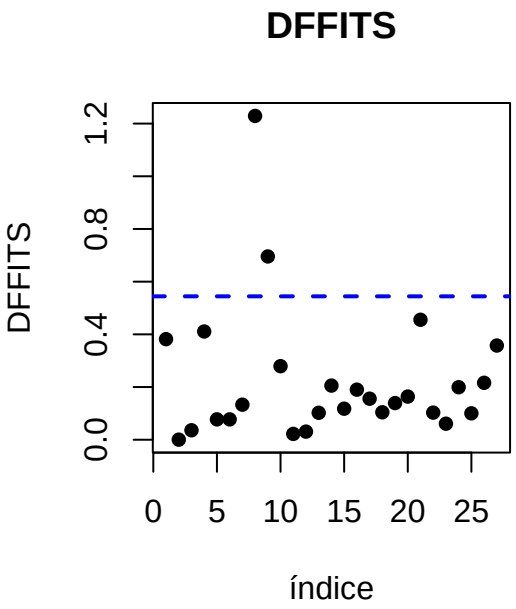
Por meio dos testes aplicados, sob um nível de significância de 5%, não há evidências estatísticas suficientes para acreditar que o modelo ajustado esteja adequado com relação às suposições necessárias para seu procedimento inferencial. Nesse sentido, segue-se para a análise de alavancagem e influência.

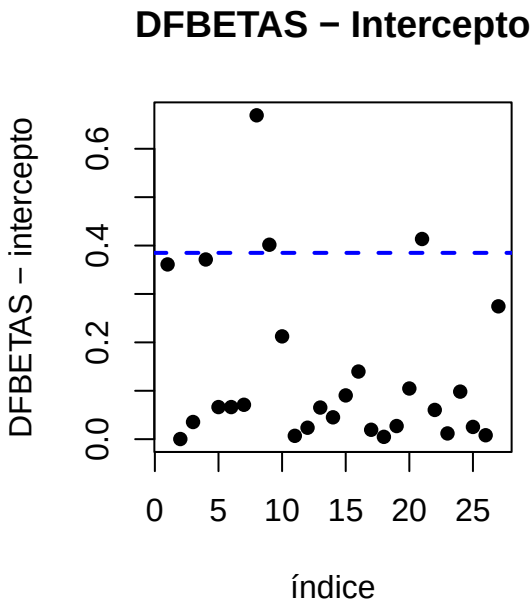
2.2.12.3 Análise de alavancagem e influência

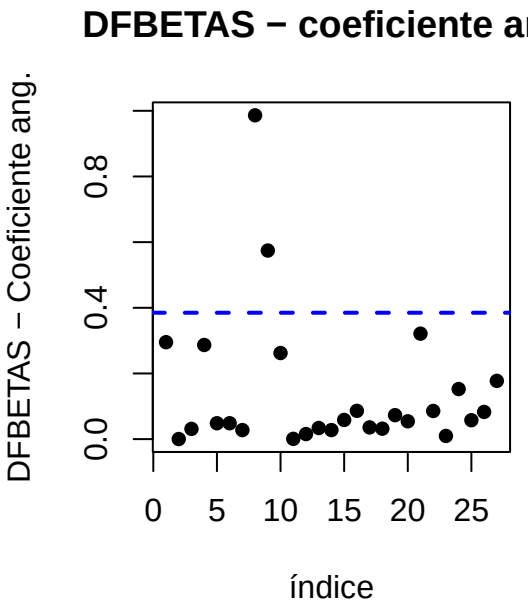


Analisando o gráfico acima, que apresenta as medidas de alavancagem pela ordem de coleta da amostra, nota-se a observação 10 como um possível ponto de alavanca, o que é justificável dado ao seu comportamento atípico com relação à covariável Área do terreno, sendo considerada um outlier.









Em relação aos pontos de influência, pelos gráficos destacam-se as observações 8, 9 e 21 como possíveis pontos de influência, sendo que o ponto 21 se destacou apenas no gráfico DFFITS - Intercepto, demonstrando assim que tal ponto talvez esteja influenciando na estimativa do intercepto. Nota-se que o ponto 8 apresenta a maior medida de influência em todos os gráficos, além de também ser considerado um aberrante.

2.2.12.4 Investigação dos pontos atípicos

A partir das técnicas aplicadas anteriormente, destacaram-se os pontos 8, 9, 10 e 21 como pontos atípicos, que então serão investigados por meio do ajuste de modelos retirando esses pontos individual e conjuntamente.

Tabela 43: Estimativas dos modelos retirando os pontos atípicos

Pontos	Beta 0	Mudança no Beta 0	Beta 1	Mudança no Beta 1
Com todos pontos	0,70	0%	0,0004	0%
Retirando 8	0,70	-0,04%	0,0005	15,0%
Retirando 9	0,70	0,03%	0,0004	-10,1%
Retirando 10	0,70	0,01%	0,0004	-4,9%
Retirando 21	0,70	-0,03%	0,0005	5,7%
Retirando 8, 9, 10 e 21	0,70	-0,05%	0,0005	11,7%

Tabela 44: P-valores dos modelos retirando os pontos atípicos

Pontos	P-valor Beta 0	Mudança no P-valor Beta 0	P-valor Beta 1	Mudança no P-valor Beta 1
Com todos pontos	<0,00000001	0%	0,0000135	0%
Retirando 8	<0,00000001	102%	0,00000016	-98%
Retirando 9	<0,00000001	7.167%	0,0000555	311%
Retirando 10	<0,00000001	371.528%	0,000276	1.948%
Retirando 21	<0,00000001	15.062%	0,00000597	-55%
Retirando 8, 9, 10 e 21	<0,00000001	18.253.242.336%	0,00000982	-27%

A partir dos resultados obtidos pela investigação dos pontos atípicos, nota-se claramente que a retirada dos pontos não impacta consideravelmente nas estimivativas dos coeficientes, sendo a maior mudança observado igual à apenas 15%, indicando portanto que tais pontos não exercem influência desproporcional em tais estimativas. Nesse sentido, o modelo também está apto para ser escolhido para explicar a variável resposta.

2.2.13 Conclusão

Tabela 45: Modelos propostos

N	Resposta	Covariável	Suposições	Pontos atípicos
1	Preço	Imposto	Desvio de normalidade e homoscedasticidade	
2	log(Preço)	Imposto	Desvio de normalidade	
3	Box-cox(Preço)	Imposto	Desvio de não correlação	
4	Box-cox(Preço)	Idade	Desvio de não correlação	
5	Preço	Área construída	Adequação das suposições	9 e 10
6	log(Preço)	Área construída	Desvio de não correlação	
7	Box-cox(Preço)	Área construída	Desvio de não correlação	
8	Preço	log(Área construída)	Adequação das suposições	9, 10 e 13
9	log(Preço)	log(Área construída)	Desvio de não correlação	
10	Box-cox(Preço)	Área do Terreno	Adequação das suposições	8, 9, 10 e 21

De acordo com a análise diagnóstica, acredito que o melhor modelo de regressão simples ajustado é o **Modelo 8, Preço x log(Área construída)**, pois este apresenta maior adequabilidade às suposições e não aparenta ser sensível aos pontos atípicos, demonstrando uma certa robustez na estimação dos parâmetros, além de explicar melhor a variabilidade dos dados, apresentando $R^2 = 0,75670$, ou seja, o modelo está explicando mais de 75% da variabilidade da resposta.

Tabela 46: Ajuste do modelo Preço x log(Área construída)

Coeficientes	Estimativa	Erro Padrão	Estatística T	P-valor
Intercepto	23,46952	2,196152	10,6867	< 0,0001
log_ac	41,27896	4,680865	8,8187	< 0,0001
R^2	0,75670			
R^2 Ajustado	0,74700			

Analisando mais especificamente o ajuste do modelo, tem-se $\beta_0 = 23,46952$, indicando a média do Preço dos imóveis quando fixamos o log da Área construída à zero, ou consequentemente, quando fixamos o valor da Área construída à 1 (equivalente à mil metros quadrados, na escala da variável). Além disso, tem-se $\beta_1 = 41,27896$, que representa a variação na média do Preço dos imóveis quando aumenta-se uma unidade no log da Área construída do imóvel. Além disso, segue abaixo o intervalo de confiança para os coeficientes:

Tabela 47: Intervalo de confiança para os coeficientes

Coeficientes	Limite Inferior	Limite Superior
Intercepto	18,946	27,993
Tempo total	31,639	50,919

2.2.14 Predições

A partir dos coeficientes estimados pelo modelo ajustado, podemos realizar predições para a variável resposta com base em novos valores fixados para o log da Área construída, desde que estes estejam dentro do intervalo de valores do banco de dados. As predições obtidas encontram-se na tabela abaixo:

Tabela 48: Predições do Preço dos imóveis (em US$ 100) para novos valores do log da Área construída (em 1000 pés quadrados)

Áreas fixadas	Predições
0,08	26,92
0,09	27,01
0,14	29,11
0,2	31,92
0,41	40,46
0,46	42,4
0,52	44,96
0,55	46,08
0,58	47,48
0,71	52,97
0,79	56,07
0,8	56,44
0,83	57,55
0,87	59,53
0,91	60,94
0,92	61,56
0,98	64,07
1	64,94
1,01	64,98
1,13	69,98