Universidade de Brasília
Departamento de Estatística

Lista Prática 1

João Vítor Rocha Silva

Relatório apresentado para a disciplina Análise de Regressão Linear - 2024.2 - EST0038 como parte dos requisitos necessários para aprovação.

Brasília
2025

Sumário

1 Problema 1:

1.1 Análise Descritiva

Tabela 1: Resumo dos dados fluoro

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Time	37.00	59.50	75.00	75.74	91.00	114.00
Dose	3.46	8.00	18.92	26.86	41.38	84.77

A Tabela 1 apresenta as estatísticas descritivas das variáveis Time (tempo do procedimento) e Dose (dose de radiação recebida). O tempo varia entre 37 s e 114 s, com uma média próxima à mediana (75,74 s e 75 s, respectivamente), sugerindo uma distribuição relativamente simétrica.

Por outro lado, a dose de radiação apresenta maior variabilidade, variando de 3,46 rads a 84,77 rads, com média (26,86 rads) superior à mediana (18,92 rads), indicando assimetria positiva. Isso sugere a presença de valores elevados que podem influenciar as análises subsequentes.

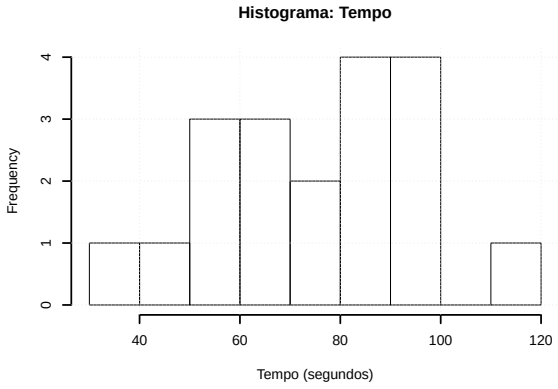


Figura 1: Histograma do tempo total do procedimento.

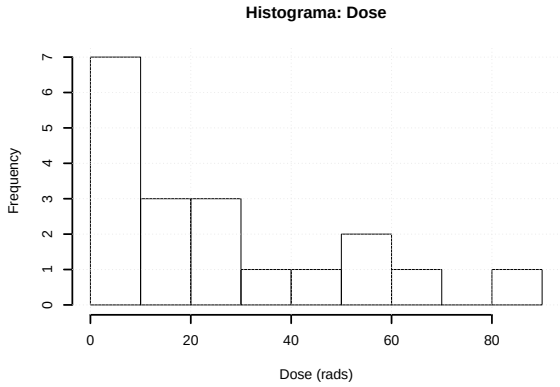


Figura 2: Histograma da dose total de radiação.

O histograma de tempo mostra a frequência do tempo total do procedimento. A distribuição do tempo tem um pico entre 80 e 100 segundos. Há uma menor frequência de procedimentos com tempos muito baixos ou muito altos, indicando uma distribuição que pode ser aproximadamente normal, mas ligeiramente inclinada à direita.

O histograma da dose mostra a frequência da dose administrada. A distribuição da dose tem um pico em torno de 0-10 unidades, mostrando que a maioria das doses está

concentrada nesse intervalo. A distribuição é inclinada à direita, com algumas doses muito altas, corroborando a análise da tabela onde vimos que a média é maior que a mediana.

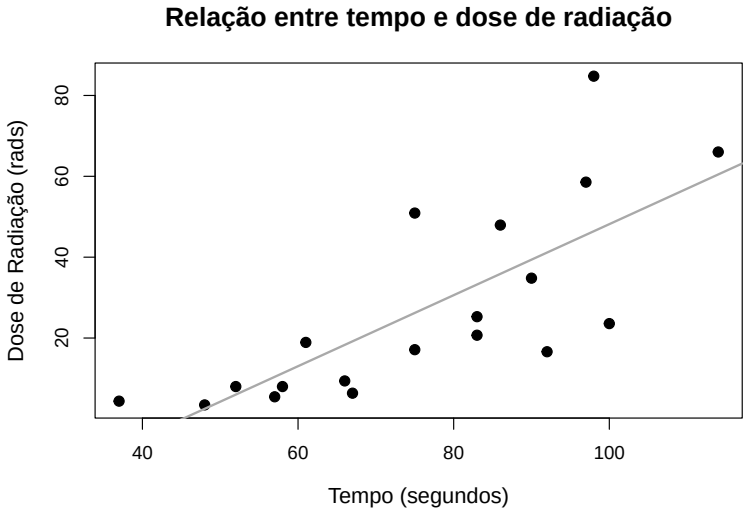


Figura 3: Relação entre tempo dose de radiação

Há uma clara tendência de que, à medida que o tempo total do procedimento aumenta, a dose total de radiação também aumenta. Isso é evidenciado pela linha de tendência que mostra uma correlação positiva entre as duas variáveis. Além disso, a maioria dos pontos de dados estão próximos à linha de tendência, o que sugere uma relação linear relativamente forte entre o tempo e a dose. Porém, vale destacar que existem alguns pontos dispersos que se afastam da linha de tendência.

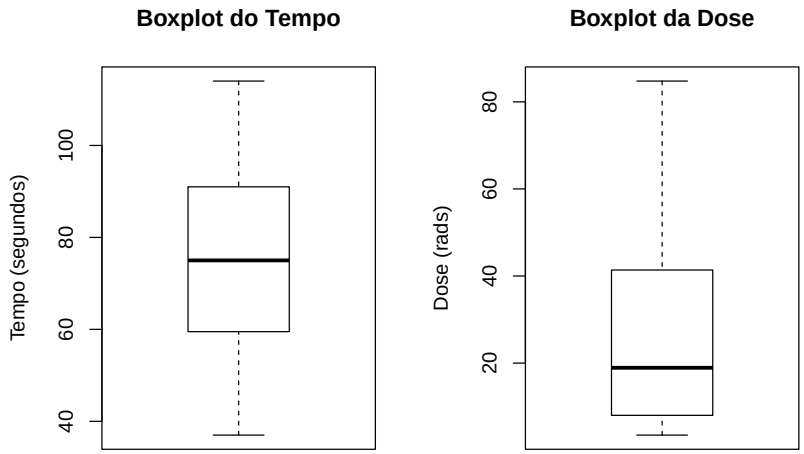
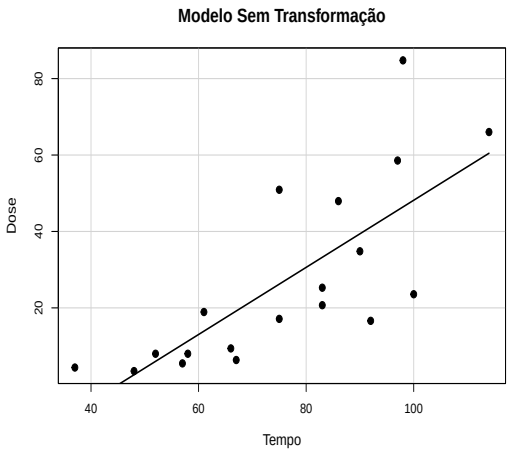


Figura 4: Boxplot do tempo e da dose de radiação

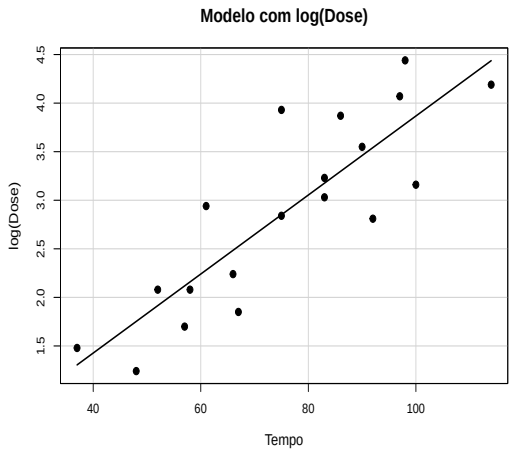
O boxplot do tempo do procedimento sugere uma distribuição relativamente simétrica, com a mediana centralizada e limites bem definidos entre o primeiro e terceiro quartil. Por outro lado, o boxplot da dose de radiação confirma a assimetria positiva observada nos histogramas. A mediana está mais próxima do primeiro quartil, e há uma cauda superior mais alongada, sugerindo a presença de valores elevados que podem ser considerados outliers.

1.2 Transformações e Ajuste de Modelos de Regressão Linear

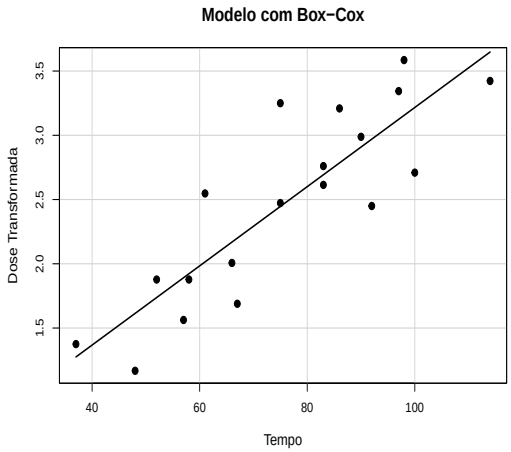
Foram realizadas três ajustes e transformações resultando em três modelos de regressão linear simples para avaliar a relação entre o tempo total do procedimento e a dose total de radiação recebida: o modelo sem transformações, o modelo com logaritmo aplicado à variável resposta (Dose), e o modelo com a transformação Box-Cox.



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)



(c) Modelo com Box-Cox

Figura 5: Gráficos de dispersão dos modelos ajustados: sem transformação, log(Dose) e Box-Cox.

A Figura 5 apresenta os gráficos de dispersão dos modelos ajustados sem transformação, com transformação logarítmica da dose e com transformação de Box-Cox.

No modelo sem transformação, observa-se que a relação entre tempo e dose é positiva, mas os pontos apresentam dispersão significativa. Já no modelo com log(Dose), a relação se torna mais linear, e os pontos estão mais alinhados em torno da reta de regressão, sugerindo um melhor ajuste em comparação ao modelo sem transformação. Por fim, no modelo com Box-Cox, observa-se um comportamento semelhante ao modelo logarítmico, com os pontos mais próximos da reta ajustada, o que indica uma melhoria na relação entre as variáveis.

As transformações logarítmica e de Box-Cox melhoram a linearidade do modelo, reduzindo a dispersão e tornando a relação entre as variáveis mais ajustada à regressão.

Tabela 2: Resumo dos Modelos Ajustados

(a) Modelo Sem Transformação

Estatística	Valor
Intercepto (β_0)	-39.66
Erro Padrão do Intercepto	14.49
Coeficiente (β_1)	0.8783
Erro Padrão do Coeficiente	0.1850
R^2	0.5702
R^2 Ajustado	0.5449
Erro Padrão Residual	16.09
p-valor do Modelo	0.000186

(b) Modelo com log(Dose)

Estatística	Valor
Intercepto (β_0)	-0.1999
Erro Padrão do Intercepto	0.4686
Coeficiente (β_1)	0.0407
Erro Padrão do Coeficiente	0.0060
R^2	0.7310
R^2 Ajustado	0.7152
Erro Padrão Residual	0.5205
p-valor do Modelo	3.11e-06

(c) Modelo com Box-Cox

Estatística	Valor
Intercepto (β_0)	0.1348
Erro Padrão do Intercepto	0.3498
Coeficiente (β_1)	0.0308
Erro Padrão do Coeficiente	0.0045
R^2	0.7369
R^2 Ajustado	0.7215
Erro Padrão Residual	0.3885
p-valor do Modelo	2.57e-06

O modelo sem transformação apresenta um ajuste moderado, com $R^2 = 0.5702$. Além disso, o erro padrão residual é relativamente alto (16.09), sugerindo maior dispersão

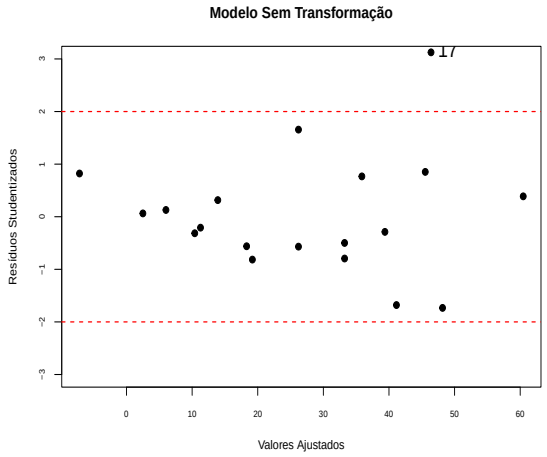
dos resíduos. Já o modelo com log(Dose) melhora consideravelmente o ajuste, com $R^2 = 0.7310$ e um erro residual reduzido para 0.5205, indicando uma relação mais linear e menor variabilidade dos resíduos. Por fim, o modelo com transformação de Box-Cox apresenta um desempenho ligeiramente superior ao modelo logarítmico, com um $R^2 = 0.7369$ e um erro padrão residual ainda menor (0.3885).

A comparação dos modelos sugere que as transformações logarítmica e de Box-Cox melhoram significativamente o ajuste da regressão, reduzindo a dispersão dos resíduos e aumentando a explicabilidade do modelo.

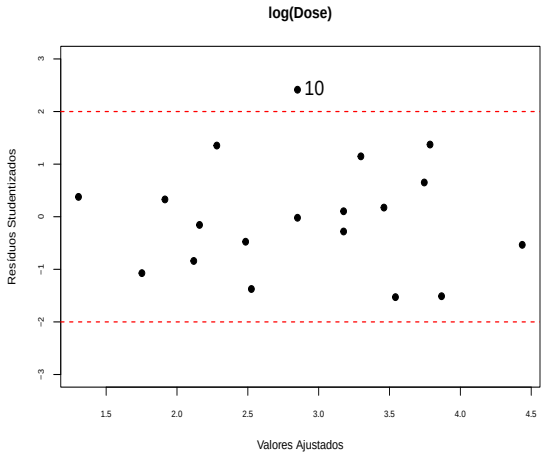
1.3 Análises de diagnóstico

1.3.1 Gráficos resíduos studentizados

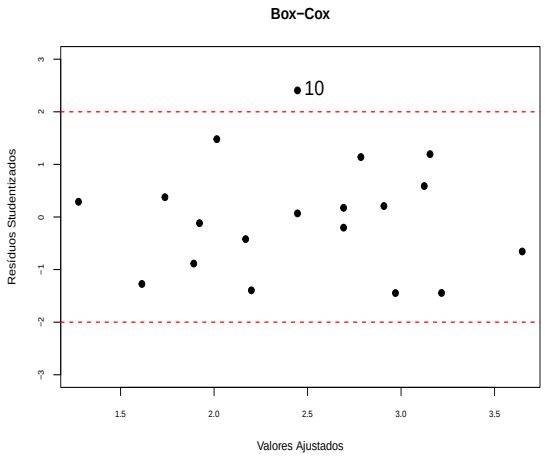
Os gráficos de resíduos studentizados foram utilizados para avaliar a adequação de cada modelo ajustado às suposições da regressão linear, eles oferecem insights sobre a adequação dos modelos e possíveis desvios das suposições da regressão linear, como homoscedasticidade e ausência de padrões nos resíduos.



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)



(c) Modelo com Box-Cox

Figura 6: Gráficos de resíduos studentizados versus valores ajustados para os modelos ajustados: sem transformação, com log(Dose) e com transformação Box-Cox.

No modelo sem transformação, observa-se a presença de um outlier significativo (ponto 17), cujos resíduos studentizados ultrapassam a faixa de ± 2.

Ao aplicar a transformação logarítmica na variável preditora, a distribuição dos resíduos aparenta estar mais homogênea, sugerindo uma melhoria na estabilidade da variância. No entanto, um novo ponto (identificado como 10) se destaca como outlier.

No modelo ajustado com a transformação Box-Cox, a distribuição dos resíduos continua dispersa ao redor de zero, e o ponto 10 mantém-se como um outlier. A amplitude da variação dos resíduos parece ligeiramente mais homogênea do que nos outros modelos, mas a presença de resíduos studentizados elevados ainda pode indicar possíveis pontos influentes.

1.3.2 Testes

Agora, analisaremos a normalidade dos resíduos dos três modelos ajustados utilizando o teste estatístico de Shapiro-Wilk.

O teste de Shapiro-Wilk avalia se os resíduos dos modelos seguem uma distribuição normal. A hipótese nula (H_0) assume que os resíduos seguem uma distribuição normal, enquanto a hipótese alternativa (H_1) assume que os resíduos não seguem uma distribuição normal. O p-valor indica se rejeitamos ou não H_0:

$p > 0.05 \Rightarrow$ Não rejeitamos $H_0 \Rightarrow$ Os resíduos podem ser considerados normais.

$p \leq 0.05 \Rightarrow$ Rejeitamos $H_0 \Rightarrow$ Os resíduos não seguem uma distribuição normal.

Tabela 3: Resultado do teste de normalidade Shapiro-Wilk

Modelo	Estatística	p-valor
Sem Transformação	0.93197	0.1883
log(Dose)	0.96508	0.6755
Box-Cox	0.95519	0.4819

Com base nos resultados dos testes, podemos concluir que nenhum dos modelos apresentou p-valor menor que 0.05, ou seja, não há evidências estatísticas para rejeitar a normalidade dos resíduos em nenhum dos casos. O modelo com log(Dose) apresentou a melhor estatística W (0.96508) e o maior p-valor (0.6755), sugerindo que seus resíduos estão mais próximos da normalidade. O modelo sem transformação apresentou o menor valor de W (0.93197), o que pode indicar uma maior tendência de desvio da normalidade em comparação com os modelos transformados.

Após a análise da normalidade dos resíduos, é necessário verificar se a suposição de homocedasticidade é atendida. Caso essa suposição seja violada, o modelo pode apresentar

heterocedasticidade, o que pode comprometer a validade das inferências estatísticas, como os intervalos de confiança e testes de significância.

Para avaliar a presença de heterocedasticidade, será aplicado o teste de Goldfeld-Quandt. Esse teste divide os dados em dois grupos com base nos valores da variável independente e compara as variâncias dos resíduos entre eles. A hipótese testada é:

H_0: Os resíduos apresentam variância constante (homocedasticidade).

H_1: Os resíduos apresentam variâncias diferentes entre os grupos (heterocedasticidade).

A interpretação do teste se baseia no p-valor, onde:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de heterocedasticidade.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, indicando heterocedasticidade significativa.

Tabela 4: Resultados dos testes Goldfeld-Quandt

Modelo	**p-valor**
Sem transformação	0.003936
Log(Dose)	0.3398
Box-Cox	0.4484

Ao observar os resultados do teste, conclui-se que o modelo sem transformação apresenta heterocedasticidade significativa, indicando que a suposição de variância constante dos erros não é válida nesse caso. Os modelos com transformações (log(Dose) e Box-Cox) não apresentam evidências de heterocedasticidade, sugerindo que essas transformações ajudaram a estabilizar a variância dos resíduos.

Após a verificação da normalidade e da homocedasticidade dos resíduos, é importante avaliar a suposição de independência dos erros no modelo de regressão. A violação dessa suposição pode indicar a presença de autocorrelação, o que pode afetar a eficiência dos estimadores e comprometer a validade das inferências estatísticas.

Para essa análise, será aplicado o teste de Durbin-Watson, que tem como objetivo detectar a autocorrelação dos resíduos da regressão.

A hipótese testada pelo teste de Durbin-Watson é:

H_0: Os resíduos são independentes (não há autocorrelação).

H_1: Os resíduos apresentam autocorrelação.

A interpretação do p-valor segue a seguinte lógica:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de autocorrelação.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, há evidências de autocorrelação nos resíduos.

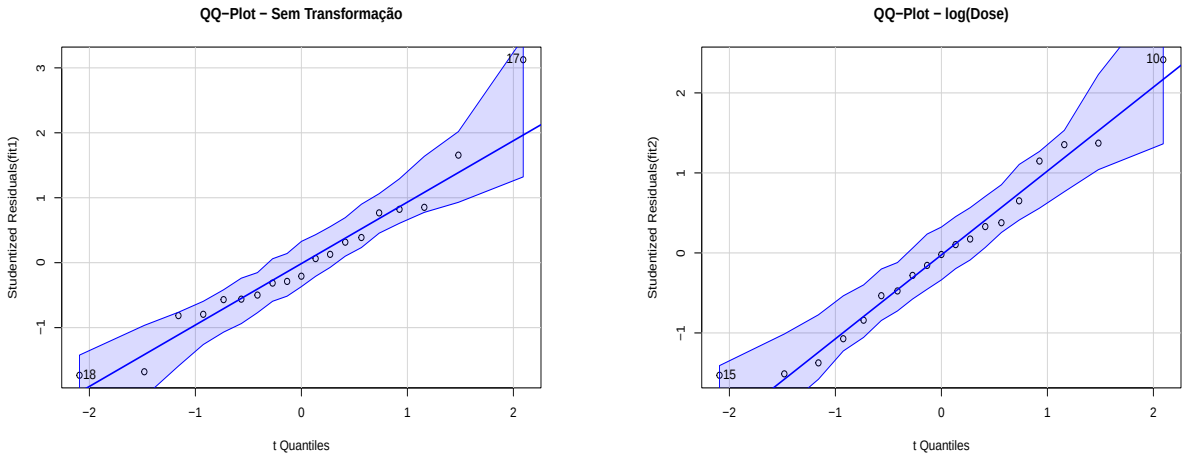
Tabela 5: Resultados dos testes Durbin-Watson

Modelo	**p-valor**
Sem transformação	0.3812
Log(Dose)	0.6887
Box-Cox	0.7926

Nenhum dos modelos apresentou p-valor menor que 0.05, ou seja, não há evidências estatísticas de autocorrelação nos resíduos. Os resultados indicam que os resíduos são independentes e não seguem um padrão sistemático ao longo das observações.

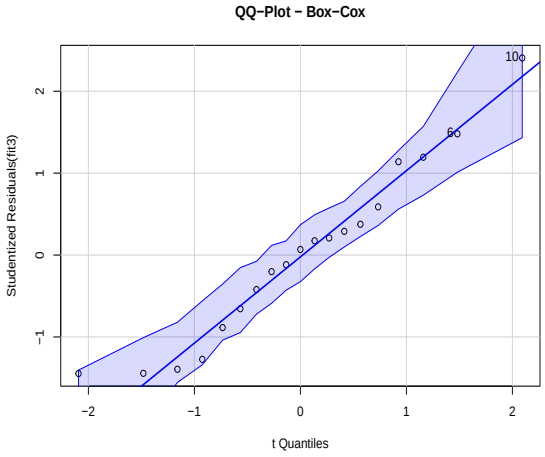
Como a autocorrelação não é um problema, não há necessidade de ajustes adicionais no modelo para lidar com essa questão.

1.3.3 Gráficos QQ-plots dos resíduos studentizados



(a) Modelo Sem Transformação

(b) Modelo com log(Dose)



(c) Modelo com Box-Cox

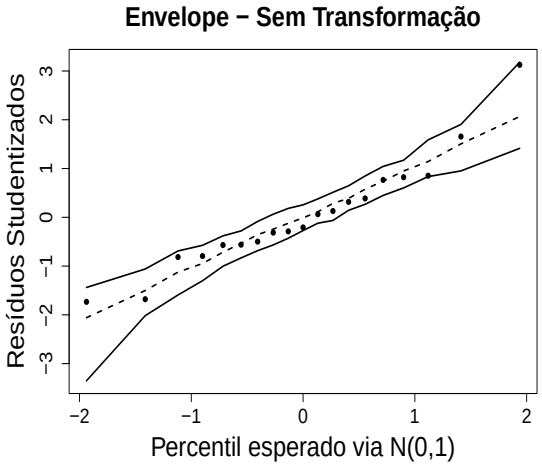
Figura 7: Avaliação gráfica da distribuição dos resíduos studentizados

No QQ-Plot - Sem Transformação (Dose vs Tempo), observa-se que os pontos apresentam certo alinhamento à reta central, mas com desvios perceptíveis, especialmente nas extremidades e alguns pontos ligeiramente ultrapassam os limites estabelecidos.

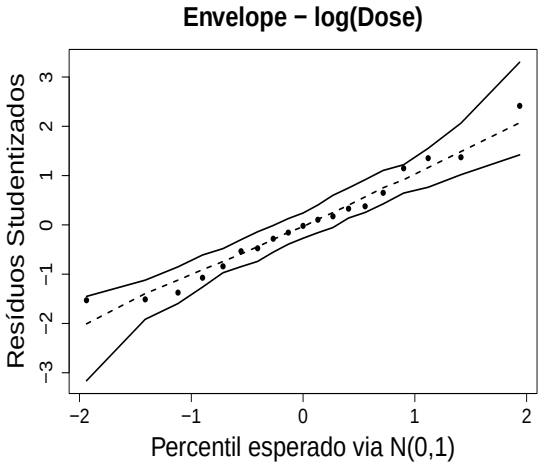
No QQ-Plot - log(Dose), os pontos estão mais alinhados à reta central, ainda há alguma dispersão ao longo da reta, especialmente nas extremidades, onde alguns pontos ligeiramente ultrapassam os limites estabelecidos.

No QQ-Plot - Box-Cox, os pontos também estão mais alinhados à reta central, sugerindo que os resíduos seguem, em grande parte, a distribuição normal. Porém, não obstante do modelo Log(Dose), ainda há pequenas discrepâncias.

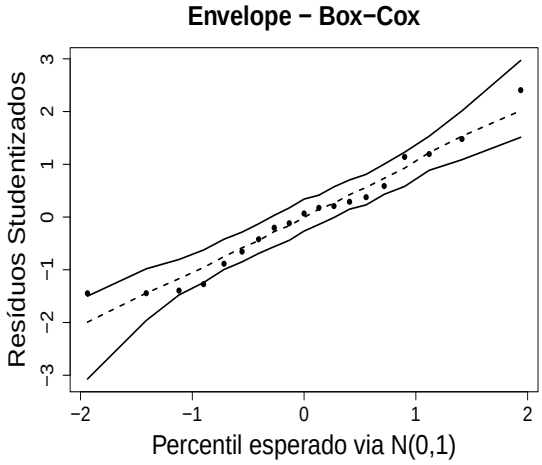
1.3.4 Gráficos de probabilidade normal



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)



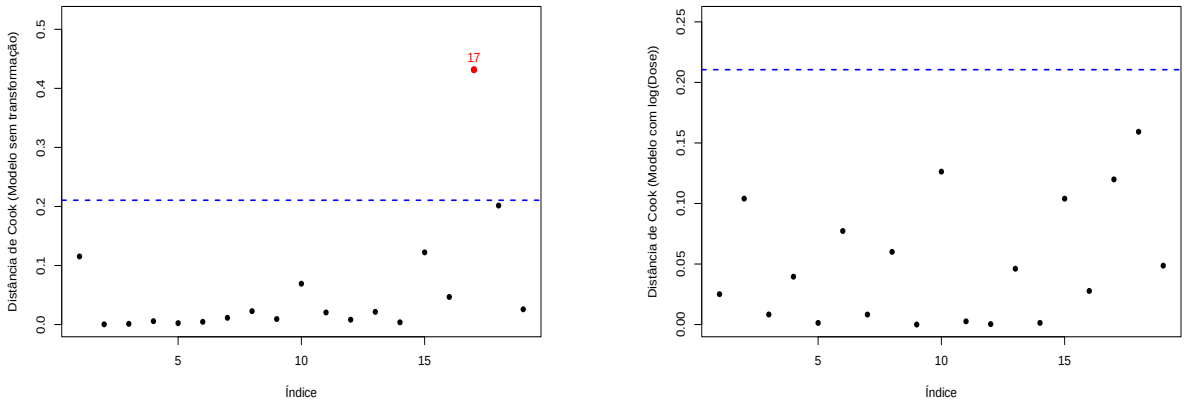
(c) Modelo com Box-Cox

Figura 8: Gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado

Complementando a análise dos gráficos QQ-Plots, os gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado na figura 8 mostram que os resíduos permanecem, na sua maior parte, dentro da faixa de confiança esperada para uma distribuição normal N(0,1), mas ainda com algumas discrepâncias nas extremidades e alguns pontos próximos do envelope.

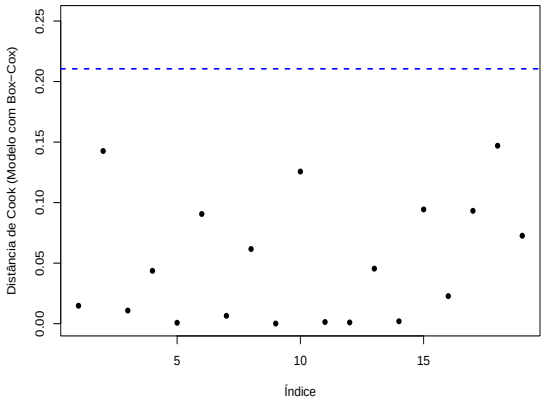
1.3.5 Medidas de Influência

Para avaliar a influência de observações individuais no ajuste do modelo, será analisada a Distância de Cook. Valores elevados de Distância de Cook indicam pontos influentes, que podem distorcer a análise e comprometer a validade dos resultados.



(a) Modelo Sem Transformação

(b) Modelo com log(Dose)



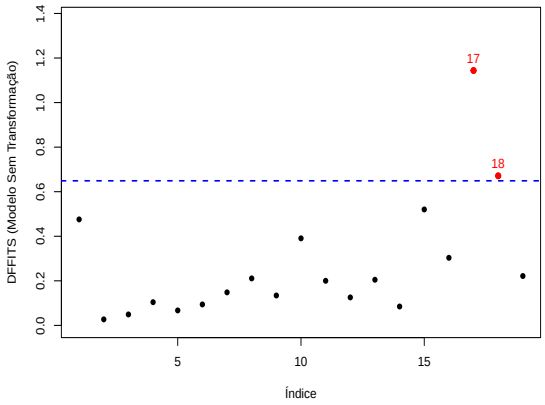
(c) Modelo com Box-Cox

Figura 9: Gráficos da Distância de Cook para os três modelos ajustados.

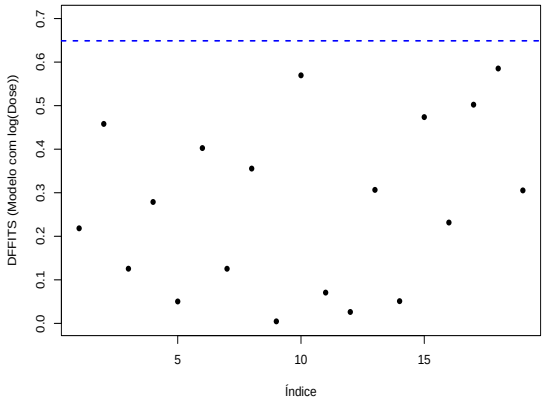
No modelo sem transformação, observa-se que o ponto 17 aparece como um possível influente. Com a transformação logarítmica, os valores da Distância de Cook tornam-se mais uniformes, sem a presença de pontos extremos muito discrepantes. Já no modelo com transformação de Box-Cox, observa-se um comportamento semelhante ao do modelo log(Dose), com uma distribuição das distâncias mais equilibrada e sem pontos possivelmentes influentes.

Para complementar a análise de influência, será examinada a métrica DFFITS.

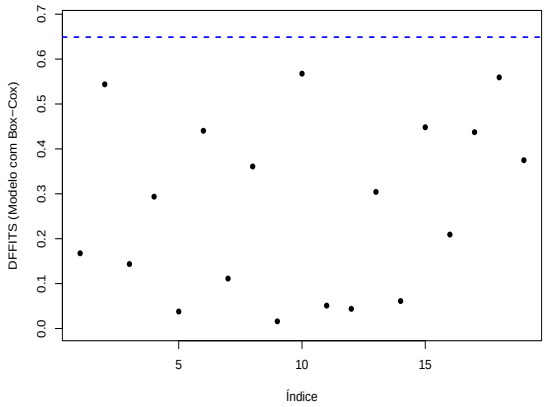
Essa medida avalia o impacto de cada observação na predição do modelo, indicando se um ponto específico tem uma grande influência sobre os valores ajustados.



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)



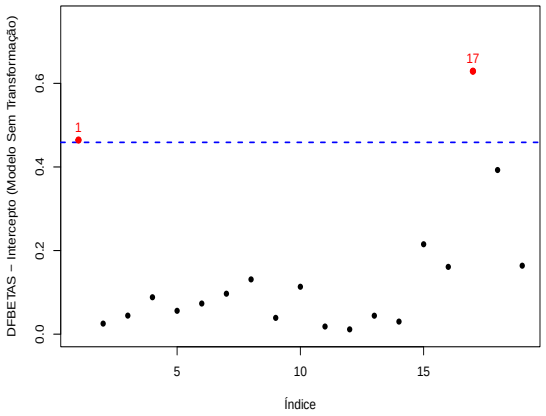
(c) Modelo com Box-Cox

Figura 10: Gráficos da métrica DFFITS para os três modelos ajustados

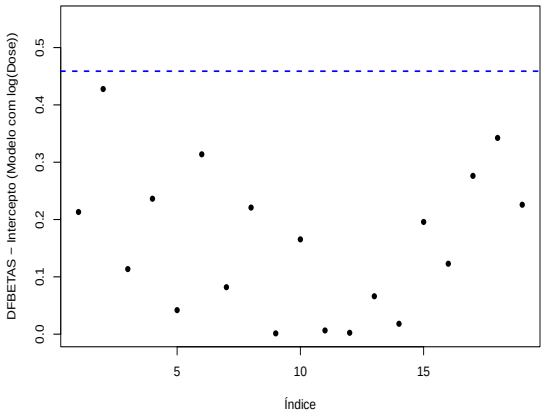
No Modelo Sem Transformação, observa-se os pontos 17 e 18 como dois possíveis pontos influentes.

Com a transformação logarítmica, os valores de DFFITS apresentam uma distribuição mais equilibrada e dispersa, sem tendência aparente. Nenhuma observação ultrapassa a linha de referência, ou seja, sem possíveis pontos influentes. Já no modelo com transformação de Box-Cox, observa-se um padrão semelhante ao do modelo log(Dose).

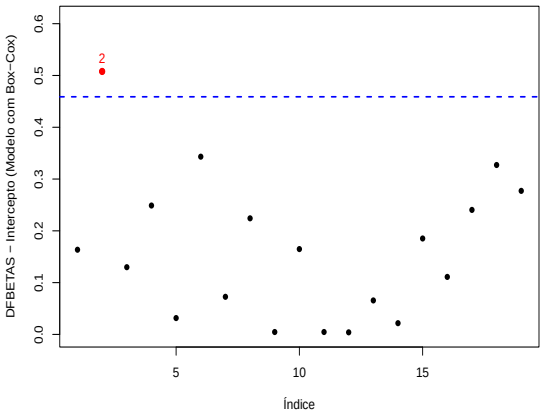
Dando continuidade à análise de influência, será examinada a métrica DFBETAS para o intercepto do modelo. Essa medida avalia o impacto que cada observação tem sobre a estimativa do coeficiente da constante.



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)

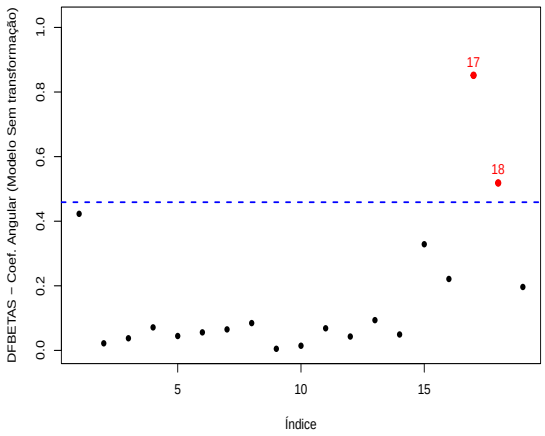


(c) Modelo com Box-Cox

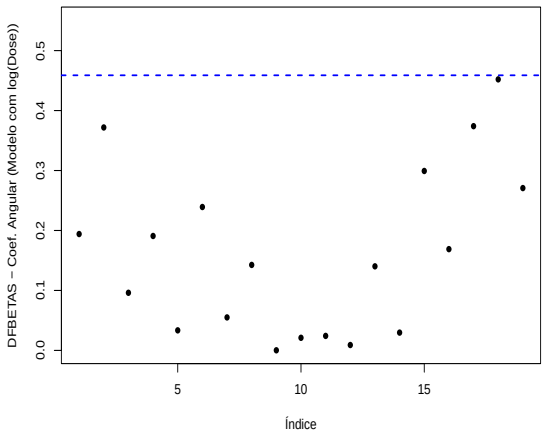
Figura 11: Gráficos da métrica DFBETAS para o intercepto nos três modelos ajustados

Os gráficos da métrica DFBETAS para o intercepto indicam que, no Modelo Sem Transformação, as observações 1 e 17 são possíveis pontos influentes. No Modelo log(Dose) os valores de DFBETAS tornam-se mais equilibrados e dispersos, sem nenhuma observação ultrapassando o limite de referência. Já no Modelo com Box-Cox, a observação 2 aparece como possível influente.

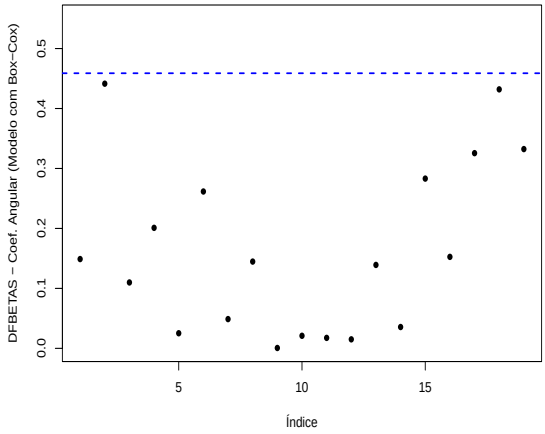
Para complementar a análise da influência de observações individuais no modelo, será examinada a métrica DFBETAS para o coeficiente angular. Essa métrica mede o impacto que cada observação tem sobre a inclinação da reta de regressão, ou seja, na estimativa do coeficiente associado à variável preditora.



(a) Modelo Sem Transformação



(b) Modelo com log(Dose)



(c) Modelo com Box-Cox

Figura 12: Gráficos da métrica DFBETAS para o coeficiente angular nos três modelos ajustados.

Além do intercepto, também é essencial verificar a influência das observações sobre o coeficiente angular, que representa a inclinação da relação entre as variáveis.

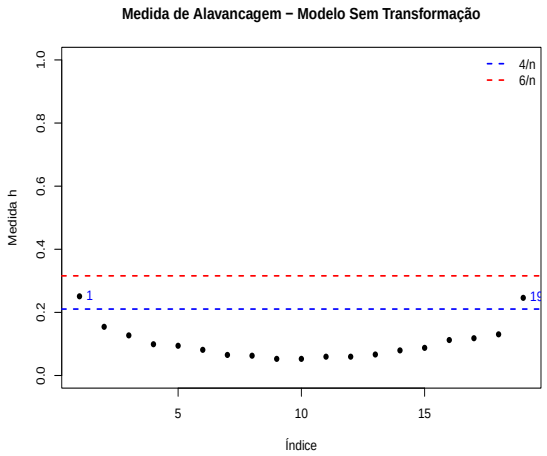
Os gráficos da métrica DFBETAS para o coeficiente angular indicam que, no Modelo Sem transformação as observações 17 e 18 aparecem como possíveis pontos influentes. Já no Modelo log(Dose) e Box-Cox, não aparecem nenhum possível ponto influente.

1.3.6 Alavancagem

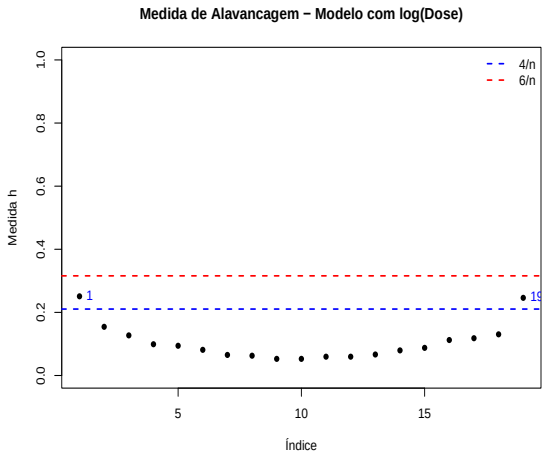
Dando continuidade à etapa de diagnóstico, serão analisados os gráficos de alavancagem. Os gráficos de alavancagem mostram a medida h de cada observação nos diferentes modelos ajustados. As linhas de referência 4/n e 6/n indicam limites convencionais para

identificar observações com alto impacto na estimativa dos coeficientes.

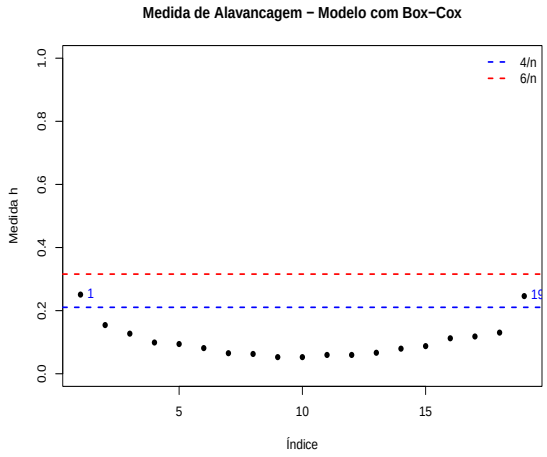
Os três modelos, Modelo sem transformação, Modelo Log(Dose) vs Imposto e Modelo Box-Cox, destacam as observações 1 e 19.



(a) Modelo sem transformação



(b) Modelo Log(Dose)



(c) Modelo Box-Cox

Figura 13: Gráficos de alavancagem para os modelos

1.4 Modelo Selecionado: Interpretação e Predição

Com base nas análises realizadas, optou-se pelo modelo com transformação logarítmica na variável resposta (`log(Dose)`) como a abordagem mais adequada para descrever a relação entre o tempo do procedimento e a dose total de radiação. Embora o modelo com transformação Box-Cox tenha apresentado resultados semelhantes, o valor estimado de λ (Figura 9) próximo de zero indica que essa transformação se aproxima do logaritmo natural. Sabe-se que a transformação de Box-Cox é definida por:

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

No caso em questão:

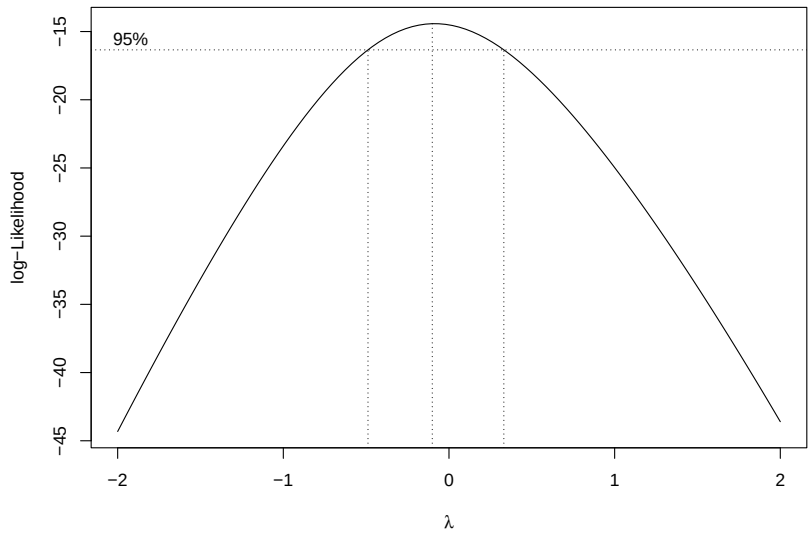


Figura 14: Estimativa do parâmetro λ para a transformação de Box-Cox com intervalo de confiança de 95%

Assim, priorizou-se o modelo logarítmico por sua simplicidade e interpretação direta.

A equação final ajustada para o modelo logarítmico é:

$$\log(\text{Dose}) = \beta_0 + \beta_1 \cdot \text{Tempo} + \epsilon$$

Substituindo os coeficientes ajustados ($\beta_0 = -0.1999$ e $\beta_1 = 0.0407$), a equação estimada fica:

$$\log(\text{Dose}) = -0.1999 + 0.0407 \cdot \text{Tempo} + \epsilon$$

Utilizando esta equação, foram realizadas predições para novos valores de tempo. As doses predizidas foram obtidas revertendo a transformação logarítmica, permitindo a apresentação dos resultados na escala original da variável resposta. A Tabela a seguir apresenta as predições para diferentes tempos de procedimento:

Tabela 6: Predições da dose recebida para novos valores de Tempo de procedimento

Tempos fixados (segundos)	**Predições (rads)**
38	3.84
44	4.90
47	5.54
59	9.02
68	13.01
69	13.55
81	22.08
82	22.99
95	39.02
96	40.63
101	49.80
108	66.20
110	71.81
112	77.90
113	81.13

Esses resultados reforçam que o modelo logarítmico é capaz de oferecer predições consistentes e robustas, com simplicidade e boa explicabilidade. Além disso, a escolha desse modelo facilita a interpretação e aplicação prática dos resultados no contexto analisado.

2 Problema 2:

2.1 Análise Descritiva:

2.1.1 Medidas-resumo

Tabela 7: Resumo estatístico das variáveis do conjunto *imoveis.txt*

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Imposto (US$ 100)	3.891	5.180	6.093	7.245	8.304	15.420
Área do terreno (1000 pés²)	2.275	4.722	5.850	6.348	7.563	12.800
Área construída (1000 pés²)	0.975	1.188	1.488	1.512	1.658	3.420
Idade da residência (anos)	3.00	30.00	40.00	36.48	47.00	62.00
Preço do imóvel (US$ 100)	25.90	29.95	36.90	38.50	40.75	84.90

A análise descritiva dos dados revela algumas características interessantes das variáveis.

O imposto (em US 100) varia entre 3.891 e 15.420 dólares, com a maior parte dos valores concentrados em torno de 6.093 dólares (mediana), sendo a média um pouco mais alta, em 7.245 dólares. Isso indica uma leve assimetria com alguns valores mais elevados.

Em relação à área do terreno (em 1000 pés²), os valores vão de 2.275 a 12.800 pés quadrados. A mediana de 5.850 pés quadrados está próxima da média de 6.348 pés quadrados, sugerindo que a maioria dos terrenos tem áreas ligeiramente acima da mediana.

A área construída (também em 1000 pés²) varia de 0.975 a 3.420 pés quadrados. A mediana de 1.488 pés quadrados e a média de 1.512 pés quadrados indicam que a distribuição é relativamente uniforme.

Quanto à idade das residências, que varia de 3 a 62 anos, a mediana é de 40 anos, ligeiramente superior à média de 36.48 anos. Isso mostra que muitas das residências analisadas são mais antigas.

Por fim, o preço dos imóveis (em US 100) varia de 25.90 a 84.90 dólares. A mediana é de 36.90 dólares, com a média em 38.50 dólares, sugerindo uma distribuição equilibrada dos preços.

2.1.2 Imposto

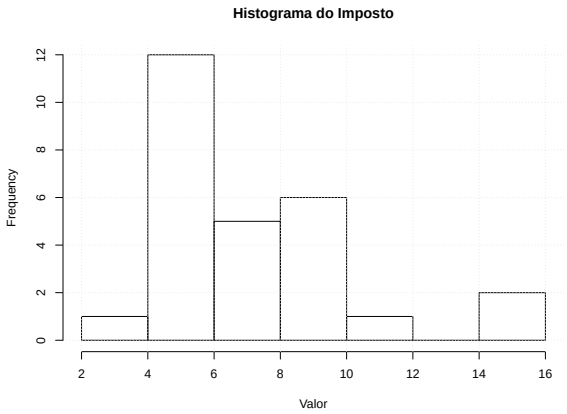


Figura 15: Histograma do imposto sobre os imóveis.

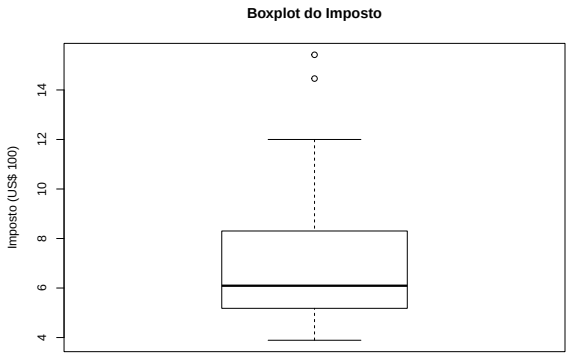


Figura 16: Boxplot do imposto sobre os imóveis (US$ 100).

A Figura 14 apresenta o histograma do imposto sobre imóveis, evidenciando uma concentração maior em valores baixos e uma assimetria à direita, indicando a presença de alguns imóveis com impostos elevados.

A Figura 15, com o boxplot, reforça essa assimetria, mostrando que a mediana está mais próxima do quartil inferior e que há outliers, representando imóveis com impostos significativamente acima da maioria. Isso sugere uma distribuição desigual, possivelmente influenciada por características como localização e tamanho dos imóveis.

2.1.3 Área Total (areaT)

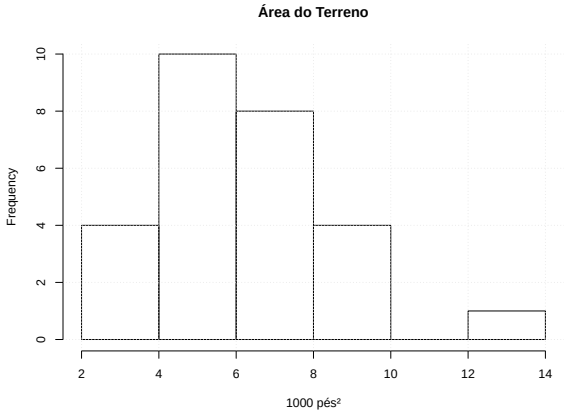


Figura 17: Histograma da área total dos terrenos (1000 pés^2).

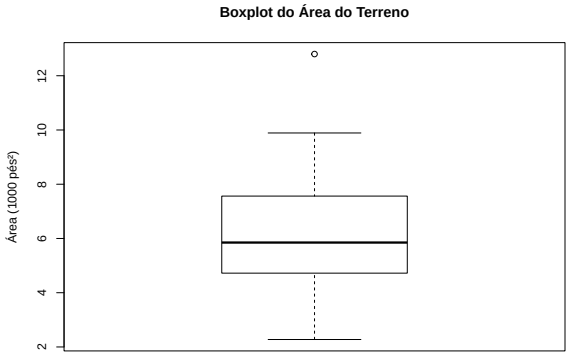


Figura 18: Boxplot da área total dos terrenos (1000 pés^2).

A Figura 16 apresenta o histograma da área total dos terrenos, indicando que a maioria dos terrenos possui áreas menores, enquanto alguns apresentam áreas significativamente maiores, evidenciando uma distribuição assimétrica à direita.

A Figura 17, com o boxplot, confirma essa assimetria, mostrando a mediana deslocada para o quartil inferior e a presença de outliers, que representam terrenos com áreas consideravelmente maiores que a média. Isso sugere uma variação significativa no tamanho dos terrenos analisados.

2.1.4 Área Construída (areaC)

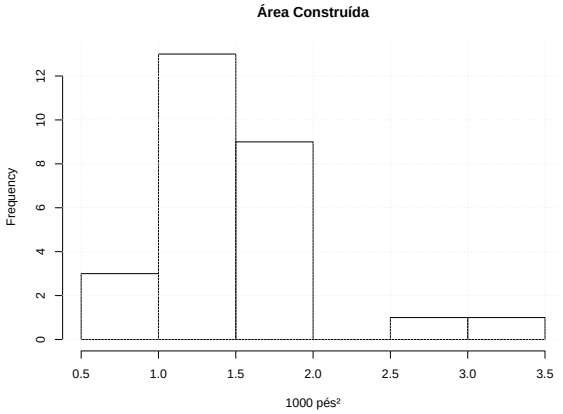


Figura 19: Histograma da área construída dos imóveis (1000 pés^2).

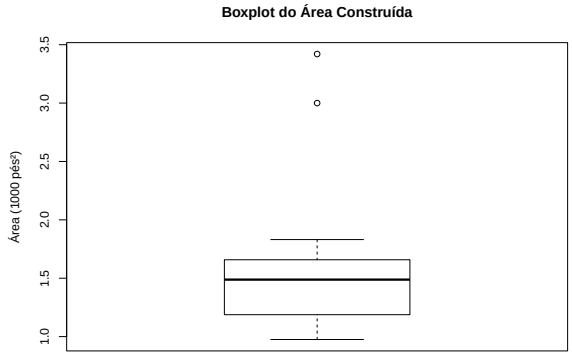


Figura 20: Boxplot da área construída dos imóveis (1000 pés^2).

A Figura 18 exibe o histograma da área construída dos imóveis, evidenciando que a maioria das construções possui áreas menores, com poucos casos de imóveis significativamente maiores, indicando uma distribuição assimétrica à direita.

A Figura 19, com o boxplot, reforça essa tendência, mostrando a mediana deslocada para o quartil inferior e a presença de outliers, representando imóveis com áreas construídas muito superiores à média. Isso sugere uma variação expressiva no tamanho das construções analisadas.

2.1.5 Idade

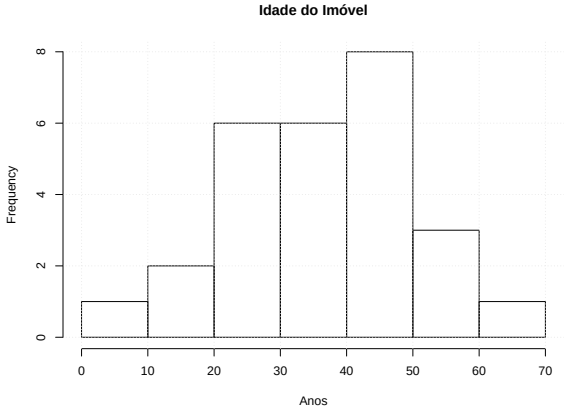


Figura 21: Histograma da idade dos imóveis (anos).

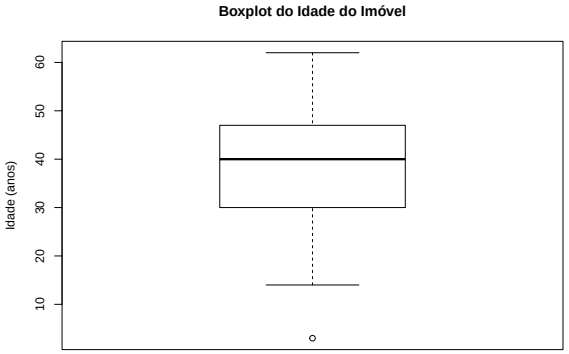


Figura 22: Boxplot da idade dos imóveis (anos).

A Figura 20 apresenta o histograma da idade dos imóveis, mostrando uma distribuição aproximadamente simétrica, com maior concentração em faixas intermediárias de idade.

A Figura 21, com o boxplot, confirma essa distribuição, com a mediana centralizada e poucos outliers. Isso sugere que a maioria dos imóveis possui idades semelhantes, sem grande variação extrema.

2.1.6 Preço

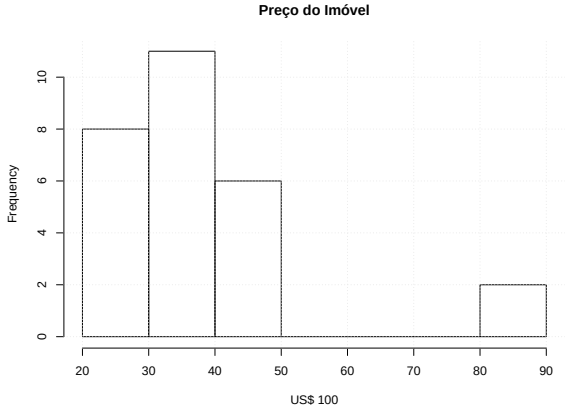


Figura 23: Histograma do preço dos imóveis (US$ 100).

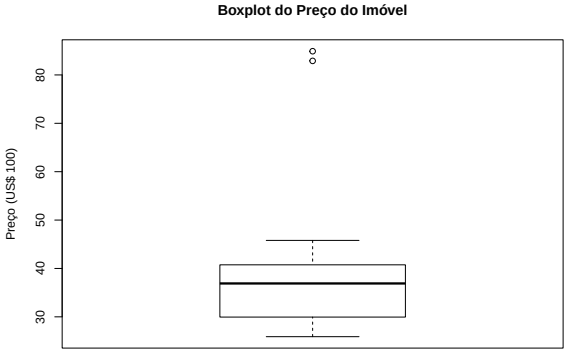


Figura 24: Boxplot do preço dos imóveis (US$ 100).

A Figura 22 apresenta o histograma do preço dos imóveis, evidenciando uma concentração maior em valores baixos, com alguns casos de preços elevados, indicando uma distribuição assimétrica à direita.

A Figura 23, com o boxplot, confirma essa assimetria, mostrando a mediana próxima ao quartil inferior e a presença de outliers, representando imóveis com preços significativamente acima da maioria. Isso sugere uma grande variação nos valores dos imóveis analisados.

2.1.7 Gráficos de Disperções

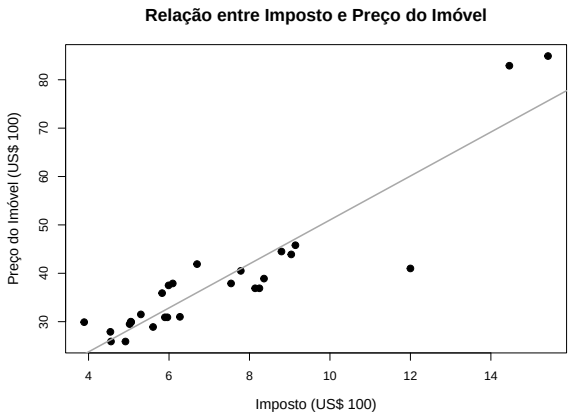


Figura 25: Relação entre imposto e preço do imóvel.

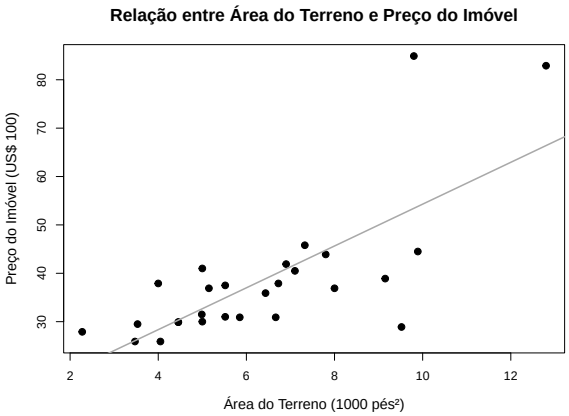


Figura 26: Relação entre área total do terreno e preço do imóvel.

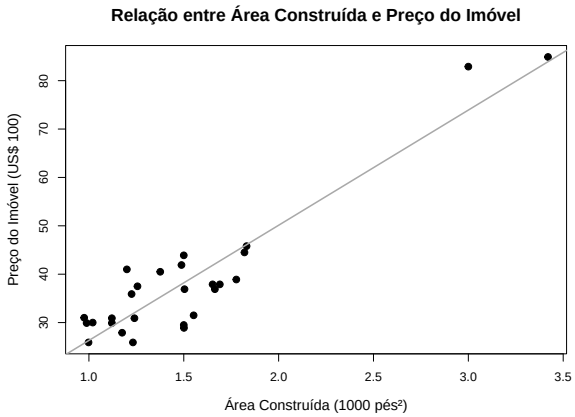


Figura 27: Relação entre área construída e preço do imóvel.

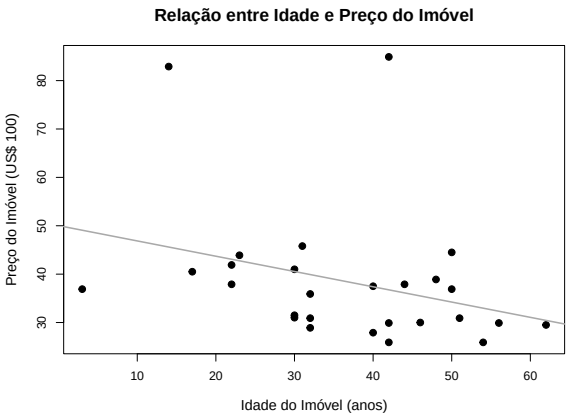


Figura 28: Relação entre idade e preço do imóvel.

Relação entre Imposto e Preço do Imóvel (Figura 24): O gráfico indica uma relação positiva entre o imposto e o preço do imóvel, o que era esperado, já que impostos geralmente acompanham o valor da propriedade. No entanto, há uma dispersão

considerável, sugerindo que outros fatores influenciam essa relação.

Relação entre Área Total do Terreno e Preço do Imóvel (Figura 25): A relação entre a área do terreno e o preço do imóvel também é positiva, mas apresenta maior dispersão. Isso indica que, embora terrenos maiores tendam a ser mais caros, essa variável sozinha não determina o preço, pois fatores como localização e infraestrutura podem ser determinantes.

Relação entre Área Construída e Preço do Imóvel (Figura 26): Essa relação apresenta a correlação mais forte entre as variáveis analisadas. Imóveis com maior área construída tendem a ser mais valorizados, o que é esperado, pois representam maior espaço útil e maior investimento em construção.

Relação entre Idade e Preço do Imóvel (Figura 27): Diferente das outras variáveis, a idade do imóvel mostra uma tendência levemente negativa em relação ao preço. Isso sugere que imóveis mais antigos tendem a valer menos, possivelmente devido a desgaste ou necessidade de reformas, mas a grande dispersão indica que essa relação não é uniforme.

2.2 Transformações e Ajuste de Modelos de Regressão Linear

Em todas as análises realizadas para cada grupo de variáveis, foram considerados quatro modelos de regressão linear, visando identificar a melhor relação entre a variável resposta e a variável preditora: (i) um modelo sem transformação, (ii) um modelo com transformação logarítmica na variável resposta, (iii) um modelo utilizando a transformação Box-Cox, e (iv) um modelo com transformação logarítmica na variável preditora.

A escolha dos modelos que avançarão para análise em termos de explicação da variabilidade dos dados e potencial de atendimento às suposições da regressão linear foi baseada na comparação dos coeficientes de determinação ajustados (R^2 ajustado), nos erros padrão dos coeficientes e na significância global do modelo (p-valor). Dessa forma, foi possível descartar modelos menos adequados.

2.2.1 Grupo 1 - Preço vs Área Total

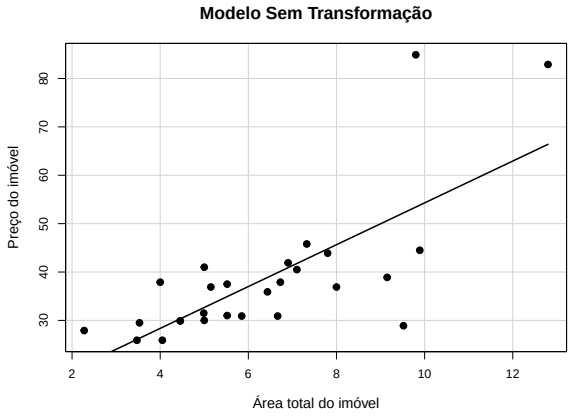


Figura 29: Sem transformação.

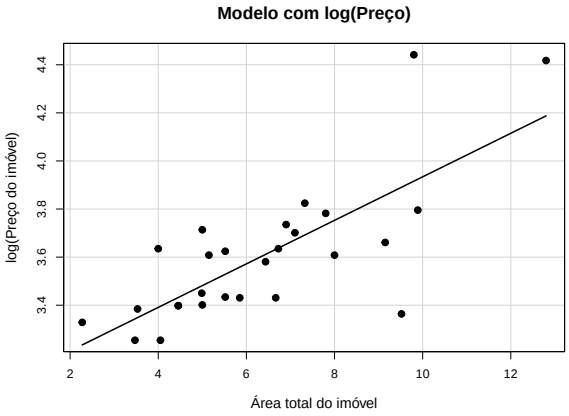


Figura 30: Com transformação logarítmica.

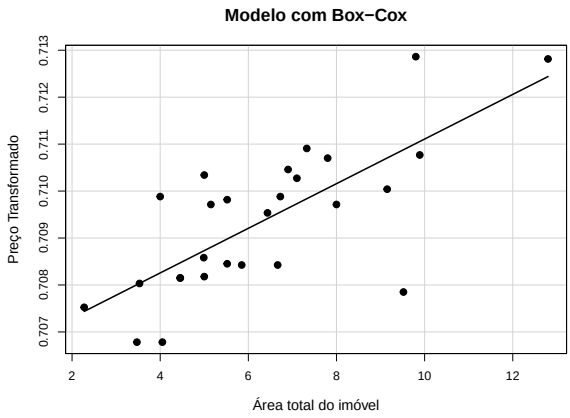


Figura 31: Com transformação Box-Cox.

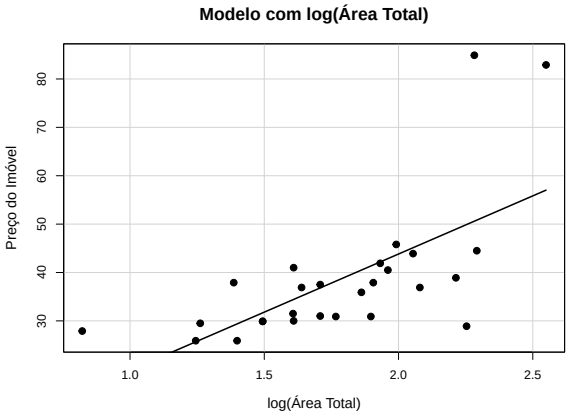


Figura 32: Modelo com log(Área Total).

Na Figura 28, sem transformação, observa-se uma relação positiva entre o preço e a área total do imóvel, há uma dispesão significativa dos pontos para maiores valores de área total. Na Figura 29, a transformação logarítmica foi aplicada ao preço. Esse ajuste parece ter melhorado a linearidade do modelo, além parece ter ocorrido uma possível estabilização da variância. A Figura 30 apresenta a transformação Box-Cox. Comparado ao modelo sem transformação, percebe-se que os pontos estão mais distribuídos ao longo da reta e que a dispersão foi reduzida. Já na Figura 31, a transformação logarítmica foi aplicada à área total, nota-se que alguns pontos ainda apresentam dispersão, o que pode indicar que essa transformação não foi tão eficaz.

Tabela 8: Resumo dos Modelos Ajustados para Preço vs. Área Total

(a) Modelo Sem Transformação

Estatística	Valor
Intercepto (β_0)	11.0570
Erro Padrão do Intercepto	5.5409
Coeficiente (β_1)	4.3234
Erro Padrão do Coeficiente	0.8183
R^2	0.5276
R^2 Ajustado	0.5087
Erro Padrão Residual	10.03
p-valor do Modelo	1.79e-05

(b) Modelo com log(Preço)

Estatística	Valor
Intercepto (β_0)	3.02887
Erro Padrão do Intercepto	0.10825
Coeficiente (β_1)	0.09051
Erro Padrão do Coeficiente	0.01599
R^2	0.5618
R^2 Ajustado	0.5443
Erro Padrão Residual	0.1959
p-valor do Modelo	6.8e-06

(c) Modelo com Box-Cox

Estatística	Valor
Intercepto (β_0)	0.7064
Erro Padrão do Intercepto	0.0005966
Coeficiente (β_1)	0.0004753
Erro Padrão do Coeficiente	0.0000881
R^2	0.5379
R^2 Ajustado	0.5194
Erro Padrão Residual	0.00108
p-valor do Modelo	1.35e-05

(d) Modelo com log(Área Total)

Estatística	Valor
Intercepto (β_0)	-4.237
Erro Padrão do Intercepto	10.262
Coeficiente (β_1)	24.028
Erro Padrão do Coeficiente	5.643
R^2	0.4204
R^2 Ajustado	0.3972
Erro Padrão Residual	11.11
p-valor do Modelo	0.0002548

O modelo sem transformação apresentou um R^2 de 0.5276 e erro padrão residual de 10.03, indicando um ajuste insatisfatório, apesar de ser estatisticamente significativo. O modelo com log(Preço) demonstrou um desempenho superior, com um R^2 de 0.5618 e erro residual muito mais baixo (0.1959), além de ser estatisticamente significativo. O modelo com Box-Cox também teve um bom ajuste, com erro padrão residual de 0.00108, mas o R^2 ajustado foi ligeiramente inferior ao do modelo com log(Preço). Por fim, o modelo com log(Área Total) foi descartado, pois apresentou um R^2 de 0.4204 e erro residual de 11.11, indicando um desempenho inferior aos demais.

Dessa forma, optou-se por seguir para as próximas etapas de diagnóstico com os modelos Box-Cox e com log(Preço), garantindo a verificação das suas suposições antes da interpretação final dos resultados.

2.2.2 Grupo 2 - Preço vs Área Construída

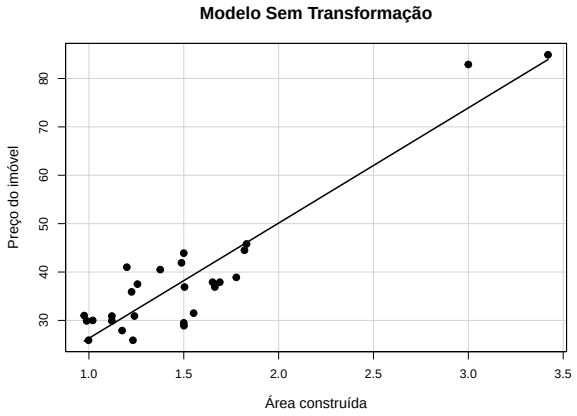


Figura 33: Sem transformação.

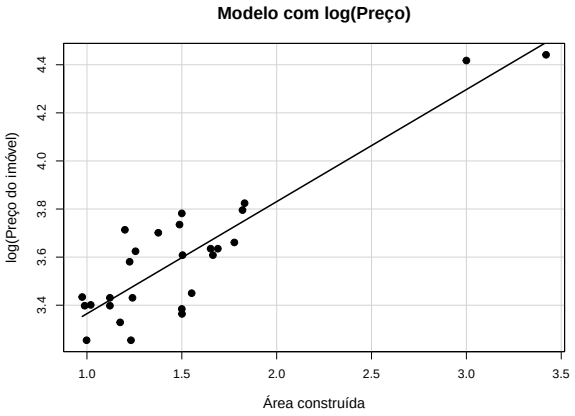


Figura 34: Com transformação logarítmica.

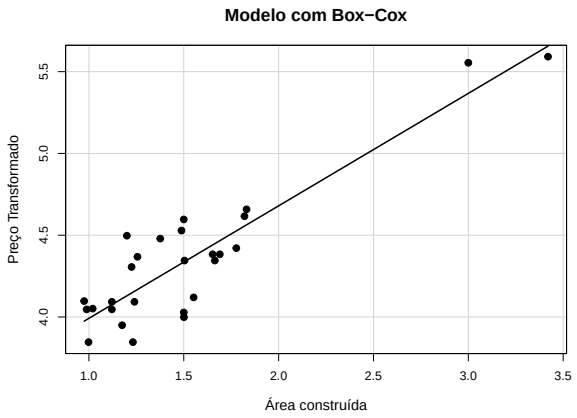


Figura 35: Com transformação Box-Cox.

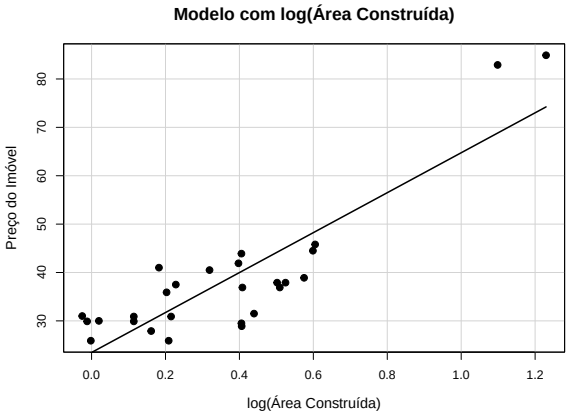


Figura 36: Modelo com log(Área Construída).

Na Figura 32, sem transformação, a relação entre o preço e a área construída é positiva e relativamente linear. A Figura 33 apresenta a transformação logarítmica aplicada ao preço. Essa modificação parece melhorar a linearidade do modelo, pois os pontos estão mais alinhados à reta de regressão. Além disso, a dispersão dos valores altos de área construída parece ter sido reduzida. Na Figura 34, a transformação Box-Cox parece corrigir melhor a dispersão dos valores altos de área construída. Por fim, a Figura 35 exibe o modelo com transformação logarítmica aplicada à área construída, percebe-se que a dispersão dos pontos ainda está presente em alguns trechos da reta de regressão.

Tabela 9: Resumo dos Modelos Ajustados para Preço vs. Área Construída

(a) Modelo Sem Transformação

Estatística	Valor
Intercepto (β_0)	2.506
Erro Padrão do Intercepto	3.054
Coeficiente (β_1)	23.804
Erro Padrão do Coeficiente	1.899
R^2	0.8627
R^2 Ajustado	0.8572
Erro Padrão Residual	5.406
p-valor do Modelo	2.81e-12

(b) Modelo com log(Preço)

Estatística	Valor
Intercepto (β_0)	2.89870
Erro Padrão do Intercepto	0.07412
Coeficiente (β_1)	0.46603
Erro Padrão do Coeficiente	0.04609
R^2	0.8035
R^2 Ajustado	0.7957
Erro Padrão Residual	0.1312
p-valor do Modelo	2.56e-10

(c) Modelo com Box-Cox

Estatística	Valor
Intercepto (β_0)	3.30423
Erro Padrão do Intercepto	0.10625
Coeficiente (β_1)	0.68767
Erro Padrão do Coeficiente	0.06606
R^2	0.8125
R^2 Ajustado	0.8050
Erro Padrão Residual	0.1881
p-valor do Modelo	1.42e-10

(d) Modelo com log(Área Construída)

Estatística	Valor
Intercepto (β_0)	23.470
Erro Padrão do Intercepto	2.196
Coeficiente (β_1)	41.279
Erro Padrão do Coeficiente	4.681
R^2	0.7567
R^2 Ajustado	0.7470
Erro Padrão Residual	7.196
p-valor do Modelo	3.803e-09

O modelo sem transformação apresentou o maior $\mathbf{R^2}$ (0.8627) e um erro padrão residual de 5.406, indicando um excelente ajuste. Seu p-valor extremamente baixo (2.81e-12) confirma a significância estatística. O modelo com log(Preço) também teve um bom desempenho, com $\mathbf{R^2}$ de 0.8035 e erro residual muito menor (0.1312), o que pode indicar melhor distribuição dos resíduos e maior robustez estatística, sendo uma alternativa viável para lidar com heterocedasticidade. O modelo com Box-Cox apresentou $\mathbf{R^2}$ de 0.8125 e erro padrão residual de 0.1881, demonstrando um ajuste adequado. Sua leve vantagem sobre o modelo log(Preço) sugere que a transformação pode ajudar na estabilização da variância dos resíduos. Já o modelo com log(Área Construída) teve o menor $\mathbf{R^2}$ (0.7567) e o maior erro residual (7.196), demonstrando desempenho inferior, sendo assim descartado.

Diante disso, os modelos sem transformação, com log(Preço) e com Box-Cox seguirão para a etapa de diagnóstico para verificação das suposições antes da interpretação final.

2.2.3 Grupo 3 - Preço vs Idade

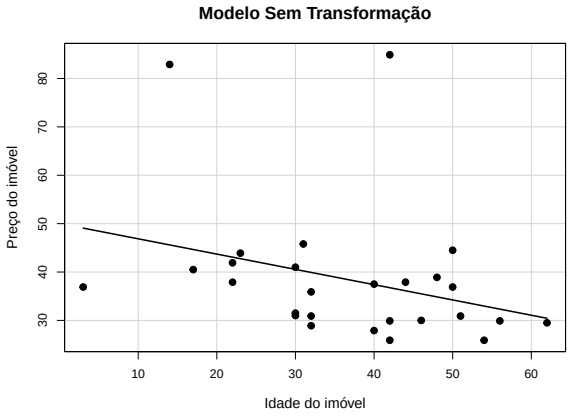


Figura 37: Sem transformação.

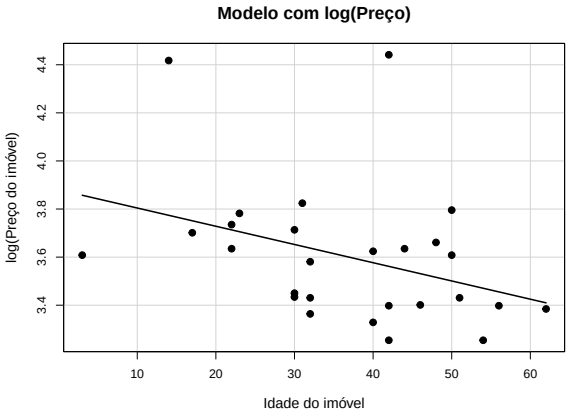


Figura 38: Com transformação logarítmica.

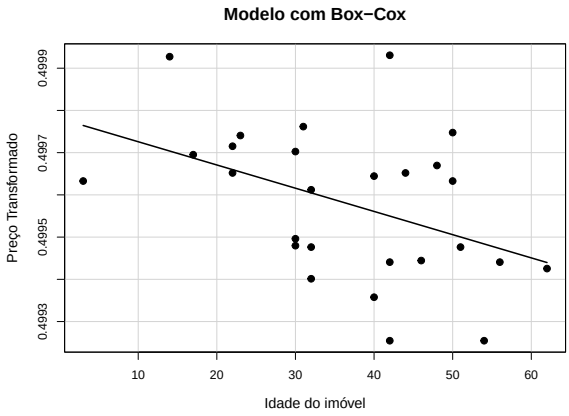


Figura 39: Com transformação Box-Cox.

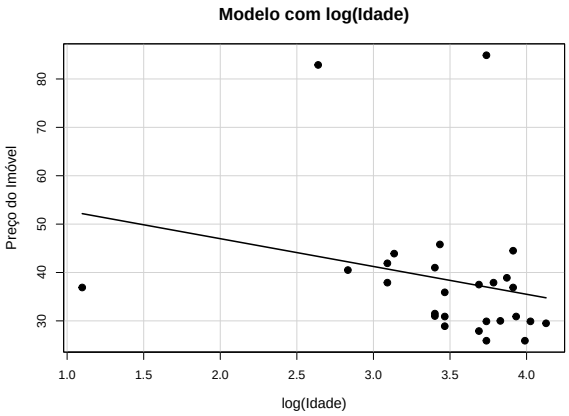


Figura 40: Com transformação log(Idade).

A Figura 36, sem transformação, apresenta uma relação negativa entre o preço e a idade do imóvel, mas com uma dispersão considerável dos pontos, o que pode indicar heterocedasticidade. Na Figura 37, a transformação logarítmica foi aplicada ao preço. Com essa modificação, a dispersão dos pontos parece ter sido reduzida, especialmente para valores mais altos de idade. No entanto, ainda há certa variabilidade na parte inferior do gráfico. Já a Figura 38 adota a transformação Box-Cox, esse modelo parece ter suavizado ainda mais a dispersão dos pontos ao longo da reta. Por fim, a Figura 39 aplica a transformação logarítmica à idade do imóvel, percebe-se que ainda há bastante dispersão ao longo da reta de regressão, especialmente em imóveis mais antigos.

Tabela 10: Resumo dos Modelos Ajustados para Preço vs. Idade do Imóvel

(a) Modelo Sem Transformação

Estatística	Valor
Intercepto (β_0)	50.0248
Erro Padrão do Intercepto	7.5494
Coeficiente (β_1)	-0.3159
Erro Padrão do Coeficiente	0.1936
R^2	0.0963
R^2 Ajustado	0.0601
Erro Padrão Residual	13.87
p-valor do Modelo	0.1152

(b) Modelo com log(Preço)

Estatística	Valor
Intercepto (β_0)	3.88008
Erro Padrão do Intercepto	0.14984
Coeficiente (β_1)	-0.007585
Erro Padrão do Coeficiente	0.003842
R^2	0.1349
R^2 Ajustado	0.1003
Erro Padrão Residual	0.2753
p-valor do Modelo	0.05951

(c) Modelo com Box-Cox

Estatística	Valor
Intercepto (β_0)	0.4998
Erro Padrão do Intercepto	0.00008866
Coeficiente (β_1)	-0.000005501
Erro Padrão do Coeficiente	0.000002273
R^2	0.1898
R^2 Ajustado	0.1574
Erro Padrão Residual	0.0001629
p-valor do Modelo	0.02312

(d) Modelo com log(Idade)

Estatística	Valor
Intercepto (β_0)	58.477
Erro Padrão do Intercepto	16.250
Coeficiente (β_1)	-5.744
Erro Padrão do Coeficiente	4.606
R^2	0.05856
R^2 Ajustado	0.0209
Erro Padrão Residual	14.16
p-valor do Modelo	0.2239

O modelo sem transformação apresentou um $\mathbf{R^2}$ de 0.0963 e um erro padrão residual de 13.87, indicando um ajuste fraco. Além disso, seu p-valor (0.1152) mostra que o modelo não é estatisticamente significativo ao nível de 5%. O modelo com log(Preço) teve um desempenho um pouco melhor, com $\mathbf{R^2}$ de 0.1349 e erro residual reduzido (0.2753). Apesar de sua leve melhora na distribuição dos resíduos, seu p-valor (0.05951) indica que a relação entre as variáveis ainda é fraca. O modelo com Box-Cox obteve o melhor ajuste, apresentando um $\mathbf{R^2}$ de 0.1898 e o menor erro padrão residual (0.0001629). Foi o único modelo estatisticamente significativo ao nível de 5% (p-valor = 0.02312), sugerindo um ajuste ligeiramente mais adequado. Por outro lado, o modelo com log(Idade) demonstrou o pior desempenho, com o menor $\mathbf{R^2}$ (0.05856), o maior erro residual (14.16) e seu p-valor (0.2239) indica que a relação entre as variáveis não é estatisticamente significativa.

Diante disso, apenas o modelo com Box-Cox será levado para a etapa de diagnóstico para verificação das suposições antes da interpretação final.

2.2.4 Grupo 4 - Preço vs Imposto

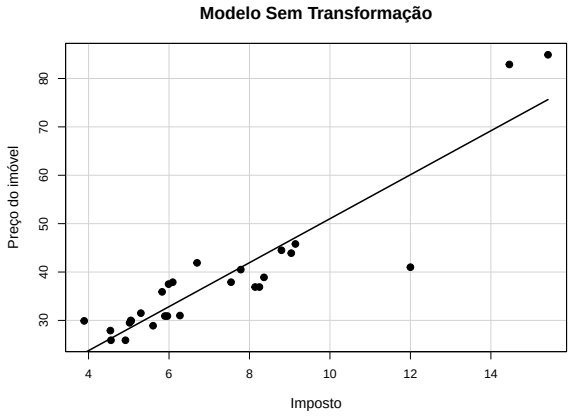


Figura 41: Sem transformação.

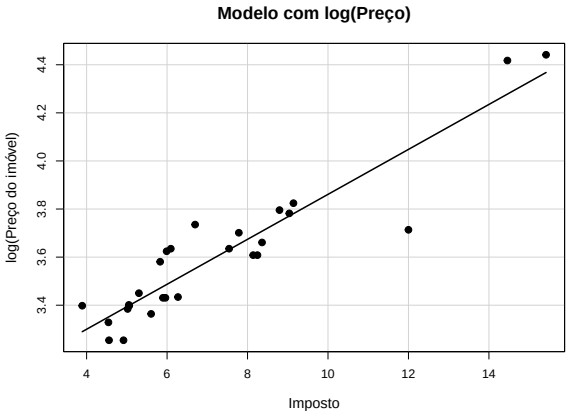


Figura 42: Com transformação logarítmica.

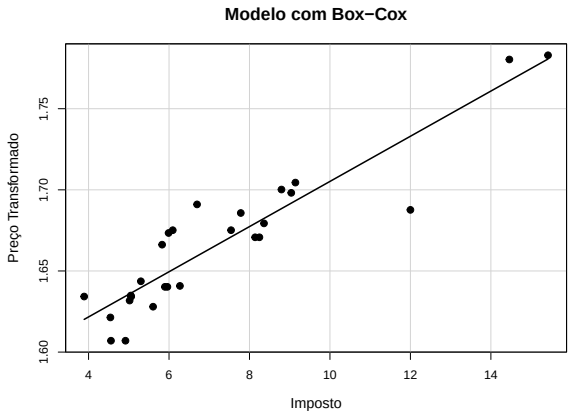


Figura 43: Com transformação Box-Cox.

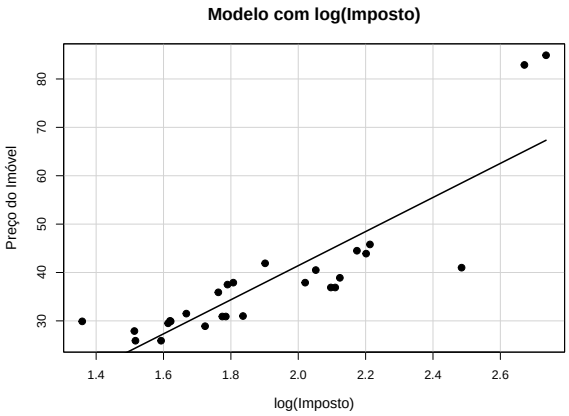


Figura 44: Com transformação log(Imposto).

A Figura 40, sem transformação, mostre que a relação entre o preço e o imposto é positiva, observa-se também que a dispersão dos pontos aumenta conforme o imposto cresce. Na Figura 41, a transformação logarítmica foi aplicada ao preço. Com essa modificação, a dispersão dos pontos parece ter sido reduzida, e a relação entre as variáveis aparenta ser mais linear do que no modelo sem transformação. A Figura 42 apresenta a transformação Box-Cox, a distribuição dos pontos parece estar mais alinhada com a reta de regressão, reduzindo as discrêpancias do modelo original. Por fim, a Figura 43 exibe o modelo em que a transformação logarítmica foi aplicada ao imposto, percebe-se que os pontos continuam apresentando certa dispersão para valores mais altos de imposto.

Tabela 11: Resumo dos Modelos Ajustados para Preço vs. Imposto

(a) Modelo Sem Transformação

Estatística	Valor
Intercepto (β_0)	5.583
Erro Padrão do Intercepto	3.111
Coeficiente (β_1)	4.543
Erro Padrão do Coeficiente	0.400
R^2	0.8377
R^2 Ajustado	0.8312
Erro Padrão Residual	5.878
p-valor do Modelo	2.31e-11

(b) Modelo com log(Preço)

Estatística	Valor
Intercepto (β_0)	2.925999
Erro Padrão do Intercepto	0.058180
Coeficiente (β_1)	0.093499
Erro Padrão do Coeficiente	0.007481
R^2	0.8620
R^2 Ajustado	0.8565
Erro Padrão Residual	0.1099
p-valor do Modelo	2.99e-12

(c) Modelo com Box-Cox

Estatística	Valor
Intercepto (β_0)	1.565994
Erro Padrão do Intercepto	0.009179
Coeficiente (β_1)	0.013919
Erro Padrão do Coeficiente	0.001180
R^2	0.8476
R^2 Ajustado	0.8415
Erro Padrão Residual	0.01734
p-valor do Modelo	1.04e-11

(d) Modelo com log(Imposto)

Estatística	Valor
Intercepto (β_0)	-29.051
Erro Padrão do Intercepto	8.289
Coeficiente (β_1)	35.232
Erro Padrão do Coeficiente	4.256
R^2	0.7327
R^2 Ajustado	0.7220
Erro Padrão Residual	7.544
p-valor do Modelo	1.25e-08

O modelo sem transformação apresentou um R^2 de 0.8377 e um erro padrão residual de 5.878, indicando um ajuste satisfatório e estatisticamente significativo (p-valor = 2.31e-11). O modelo com log(Preço) demonstrou um desempenho superior, com um R^2 de 0.8620 e um erro residual muito mais baixo (0.1099), além de ser estatisticamente significativo, tornando-se a melhor opção para prosseguir com a análise. O modelo com Box-Cox também apresentou um bom ajuste, com um R^2 de 0.8476 e um erro padrão residual de 0.01734. Seu desempenho próximo ao modelo log(Preço) sugere que ambas as transformações podem ser úteis para a estabilização da variância. Por outro lado, o modelo que aplicou logaritmo na variável explicativa (log(Imposto)) apresentou o menor R^2 (0.7327) e o maior erro residual (7.544), demonstrando um ajuste inferior e sendo descartado.

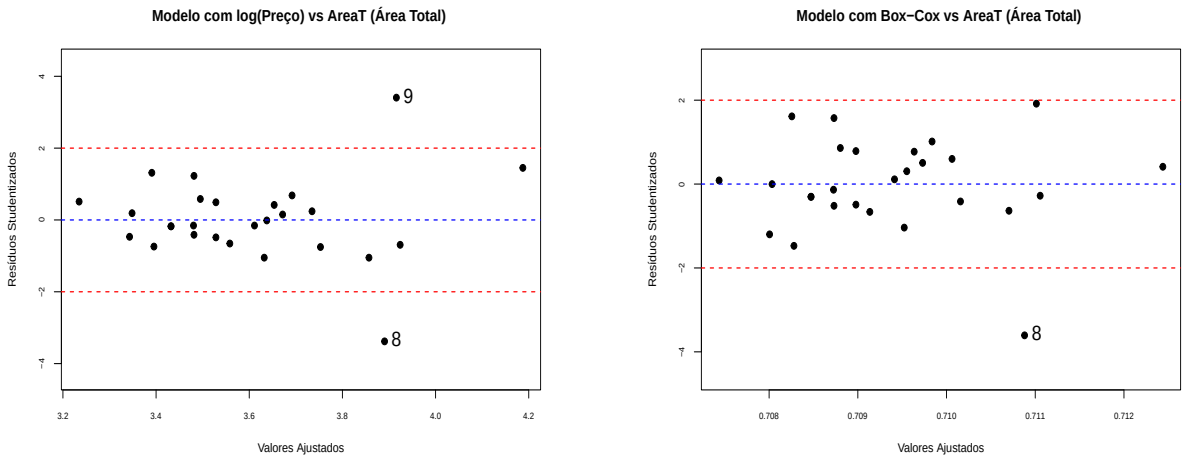
Diante disso, os modelos sem transformação, com log(Preço) e com Box-Cox seguirão para a etapa de diagnóstico para verificação das suposições antes da interpretação final.

2.3 Análises de diagnóstico

2.3.1 Diagnóstico covariável - Área Total

Os gráficos de resíduos studentizados foram utilizados para avaliar a adequação de cada modelo ajustado às suposições da regressão linear, eles oferecem insights sobre a adequação dos modelos e possíveis desvios das suposições da regressão linear, como homoscedasticidade e ausência de padrões nos resíduos.

No modelo sem transformação, os pontos 8 e 9 se destacam como outliers. Ao aplicar a transformação logarítmica, somente o ponto 8 mantém-se como um outlier.



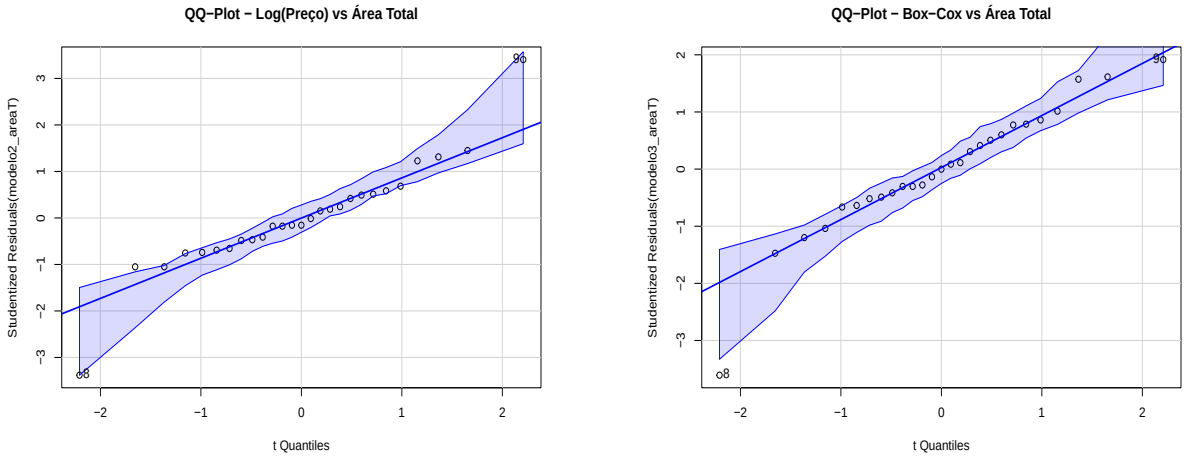
(a) Modelo com log(Preço)

(b) Modelo com Box-Cox

Figura 45: Gráficos de resíduos studentizados versus valores ajustados para os modelos ajustados: log(Preço) e com transformação Box-Cox.

No QQ-Plot - Log(Preço) vs Área Total a maioria dos pontos segue a linha azul central, isso indica que os resíduos estão aproximadamente distribuídos normalmente. Porém, é visível que alguns pontos estão ligeiramente fora desse limite.

Já no QQ-Plot - Box-Cox vs Área Total, os pontos estão mais próximos da linha azul. Nota-se que a transformação Box-Cox parece ter melhorado a normalidade dos resíduos, porém há discrêpancias nas extremidades.

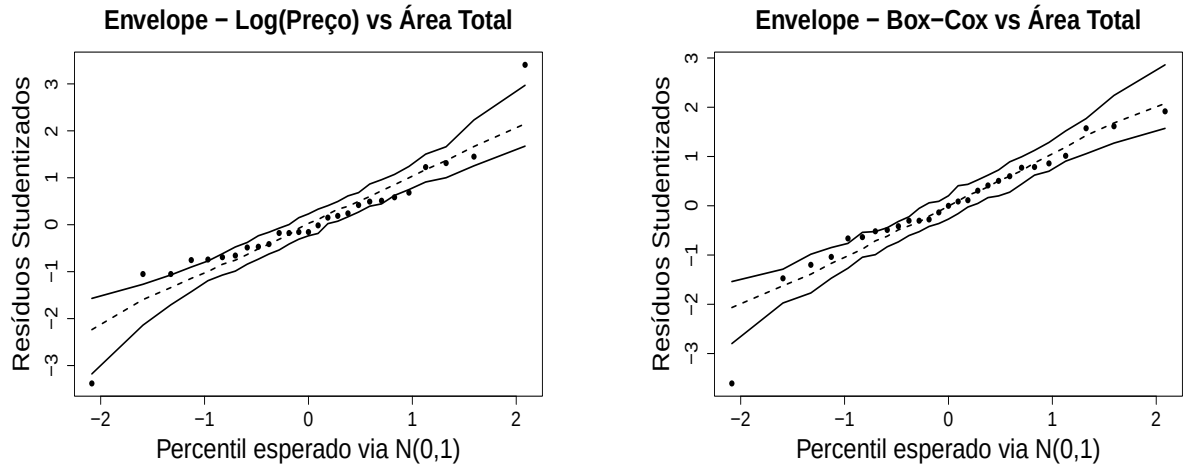


(a) Modelo com log(Preço) (b) Modelo com Box-Cox (Área Total)

Figura 46: Gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado

Complementando, os gráficos de envelope indicam que os resíduos seguem aproximadamente uma normal N(0,1), com a maioria dos pontos dentro da faixa de confiança. Pequenos desvios são observados nas extremidades.

No Modelo com Box-Cox (Área Total), os resíduos também estão distribuídos conforme o esperado para uma normal N(0,1), com um bom alinhamento na região central e leves discrepâncias nas caudas.



(a) Modelo com log(Preço) (b) Modelo com Box-Cox (Área Total)

Figura 47: Gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado

Porém, para uma análise mais concreta optou-se por aplicar o teste de Shapiro-

Wilk em ambos os modelos, onde o p-valor indica se rejeitamos ou não H_0:

$p > 0.05 \Rightarrow$ Não rejeitamos $H_0 \Rightarrow$ Os resíduos podem ser considerados normais.

$p \leq 0.05 \Rightarrow$ Rejeitamos $H_0 \Rightarrow$ Os resíduos não seguem uma distribuição normal.

Para o modelo com log(Área Total), o p-valor de 0.01638 indica a rejeição da hipótese de normalidade a um nível de 5%, sugerindo que os resíduos não seguem uma distribuição normal.

Já no modelo com Box-Cox (Área Total), o p-valor de 0.07285 não leva à rejeição da normalidade, indicando um melhor ajuste.

Modelo	**p-valor**
Log(Preço)	0.01638
Box-Cox	0.07285

Tabela 12: Resultados do teste de normalidade Shapiro-Wilk para os modelos com a covariável Área Total

Prosseguindo com o diagnóstico, foi realizado o teste de Goldfeld-Quandt para verificar a presença de heretocedasticidade nos modelos.

A interpretação do teste se baseia no p-valor, onde:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de heterocedasticidade.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, indicando heterocedasticidade significativa.

Para o modelo com transformação logarítmica da variável 'Preço', o valor do p-valor de 0.00886, o que sugere a presença de heterocedasticidade no modelo.

Por outro lado, para o modelo com transformação Box-Cox, o valor do p-valor foi de 0.134, sugerindo que o modelo Box-Cox não apresenta heterocedasticidade.

Modelo	**p-valor**
Log(Preço)	0.00886
Box-Cox	0.134

Tabela 13: Resultados do teste Goldfeld-Quandt para os modelos com a covariável Área Total

Agora, para verificar a presença de autocorrelação dos resíduos foi aplicado o teste de Durbin-Watson.

A interpretação do p-valor do teste de Durbin-Watson segue a seguinte lógica:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de autocorrelação.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, há evidências de autocorrelação nos resíduos.

No Modelo Log(Preço) apresentou um p-valor igual a 0.8535, ou seja, não há evidências suficientes para rejeitar a hipótese nula de que os resíduos não estão autocorrelacionados.

Assim como no modelo logarítmico, o Box-Cox apresentou um p-valor maior que 0.05. Esses resultados indicam que, em ambos os modelos, não há evidências fortes de autocorrelação nos resíduos.

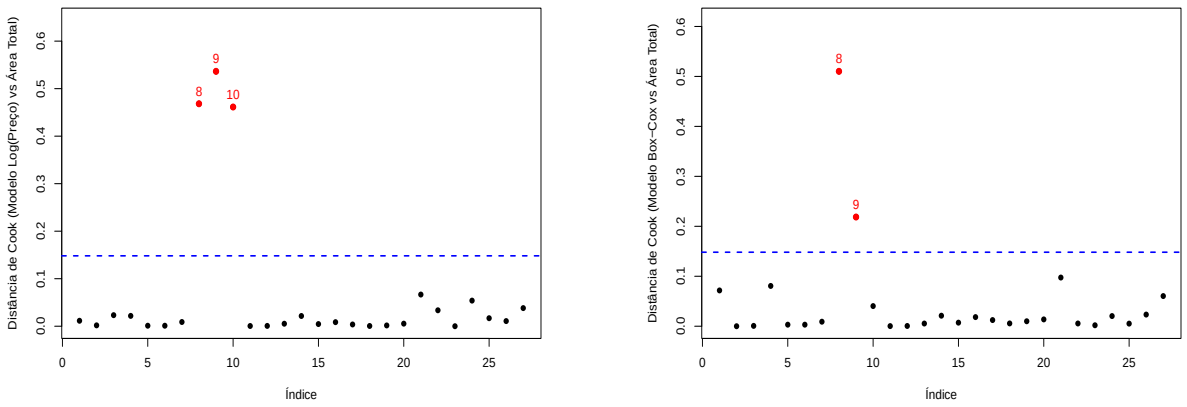
Modelo	**p-valor**
Log(Preço)	0.8535
Box-Cox	0.3082

Tabela 14: Resultados do teste Durbin-Watson para os modelos com a covariável Área Total

Avançando mais com o diagnótico, analisaremos algumas medidas de influência.

Para analisar o impacto global de cada observação no modelo ajustado, utilizamos a métrica da Distância de Cook. Essa medida permite identificar pontos que influenciam significativamente os coeficientes da regressão, podendo indicar observações que afetam de maneira relevante os ajustes realizados.

A partir dos gráficos, nota-se que o Modelo com log(Preço) indica as observações 8, 9 e 10 como possíveis pontos influentes. Já no Modelo com Box-Cox, apenas as observações 8 e 9 se destacam, sugerindo que essa transformação pode ter reduzido a influência de alguns pontos.



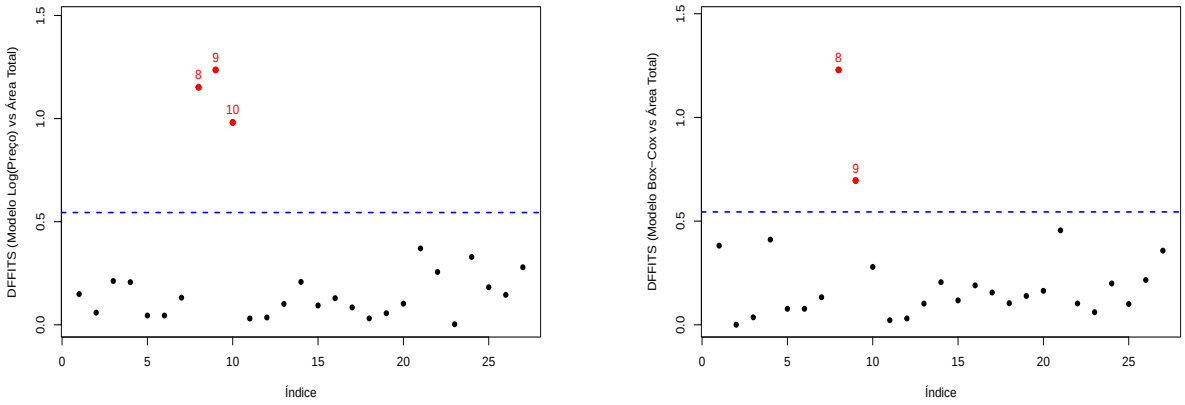
(a) Modelo com log(Preço)

(b) Modelo com Box-Cox

Figura 48: Gráficos da Distância de Cook para os dois modelos ajustados para a covariável Área Total

Embora a Distância de Cook forneça uma visão geral da influência das observações, a métrica DFFITS permite avaliar especificamente o impacto de cada ponto na predição do modelo.

Os gráficos da métrica DFFITS mostram que, no Modelo com log(Preço), as observações 8, 9 e 10 também aparecem como possíveis pontos influentes. Igualmente no Modelo Box-Cox, onde a tranformação reduziu a influência somente da observação 10.



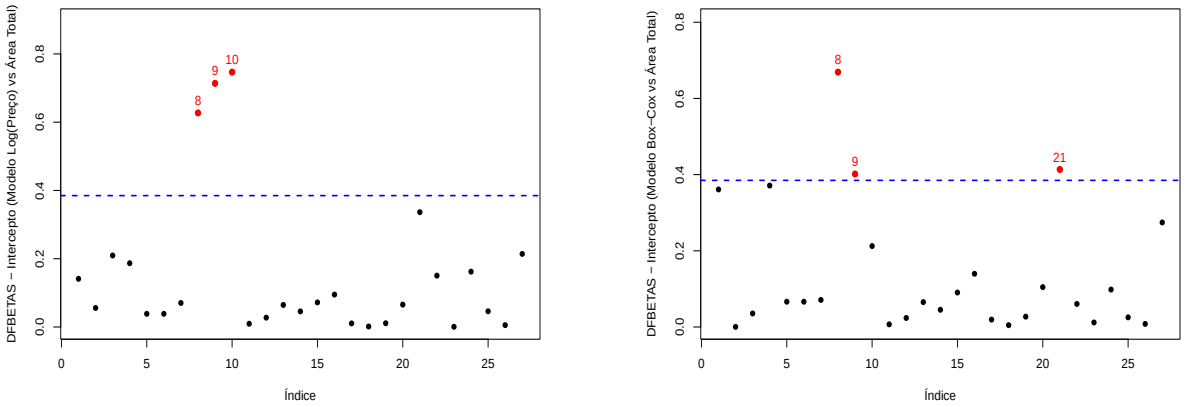
(a) Modelo com log(Preço)

(b) Modelo com Box-Cox

Figura 49: Gráficos da métrica DFFITS para os dois modelos ajustados para a covariável Área Total

Além de analisar a influência sobre as predições, é importante avaliar como cada observação afeta individualmente os coeficientes do modelo. Para isso, utilizamos a métrica DFBETAS, que mede a variação nos coeficientes ao remover uma determinada observação. Primeiramente, analisamos o impacto sobre o intercepto.

Os gráficos da métrica DFBETAS para o intercepto indicam que, no Modelo com log(Preço), as observações 8, 9 e 10 são possíveis pontos influentes. No Modelo com Box-Cox, as observações 8, 9 e 21 aparecem como possíveis influentes.

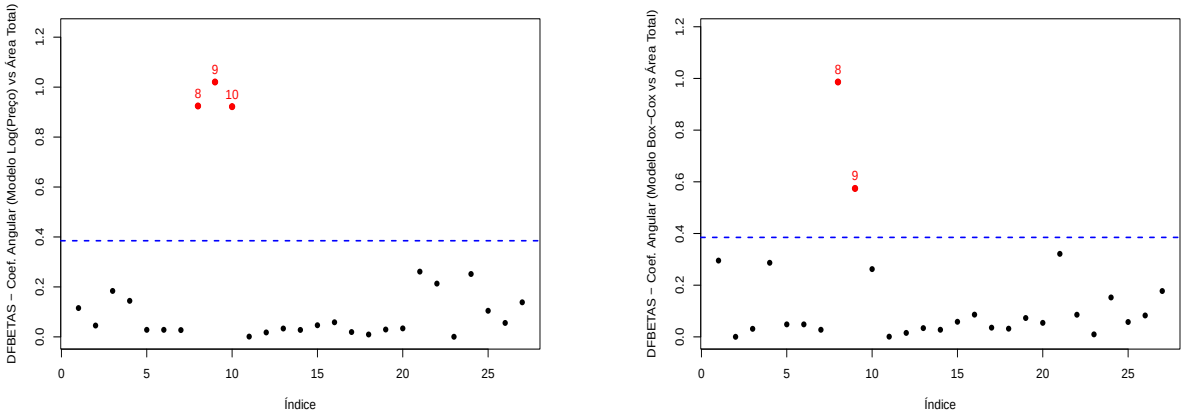


(a) Modelo com log(Preço) (b) Modelo com Box-Cox

Figura 50: Gráficos da métrica DFBETAS para o intercepto nos dois modelos ajustados para a covariável Área Total

Além do intercepto, também é essencial verificar a influência das observações sobre o coeficiente angular, que representa a inclinação da relação entre as variáveis.

No Modelo com log(Preço), as observações 8, 9 e 10 se destacam como possíveis pontos influentes. No Modelo com Box-Cox, somente as observações 8 e 9 continuaram.



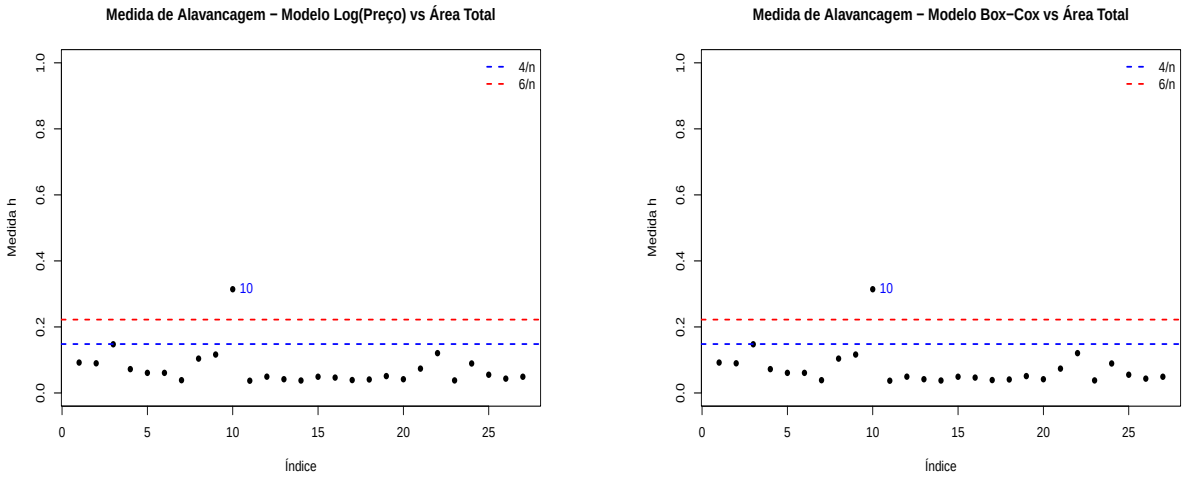
(a) Modelo com log(Preço) (b) Modelo com Box-Cox

Figura 51: Gráficos da métrica DFBETAS para o coeficiente angular nos dois modelos ajustados para a covariável Área Total

Os gráficos de alavancagem mostram a medida h de cada observação nos diferentes modelos ajustados. As linhas de referência 4/n e 6/n indicam limites convencionais para identificar observações com alto impacto na estimativa dos coeficientes.

Os dois modelos, Modelo Log(Preço) vs Área Total e Modelo Box-Cox vs Área

Total, destacam a observação 10 como um ponto com alavancagem elevada.



(a) Modelo Log(Preço) vs Área Total

(b) Modelo Box-Cox vs Área Total

Figura 52: Gráficos de alavancagem para os modelos baseados na Área Total

Finalizando a etapa de diagnóstico, analisaremos o impacto dos pontos influentes nos coeficientes dos modelos. A tabela abaixo apresenta os coeficientes da regressão dos modelos, além da remoção de possíveis pontos influentes identificados nos gráficos de alavancagem (observações 8, 9, 10 e 21).

No modelo logarítmico, a remoção dos pontos teve um grande impacto sobre β_1, especialmente ao remover todos os quatro pontos. Isso indica que esses pontos eram influentes na relação entre preço e áreaT.

No modelo Box-Cox, as mudanças foram mínimas, sugerindo que este modelo é mais robusto a essas observações.

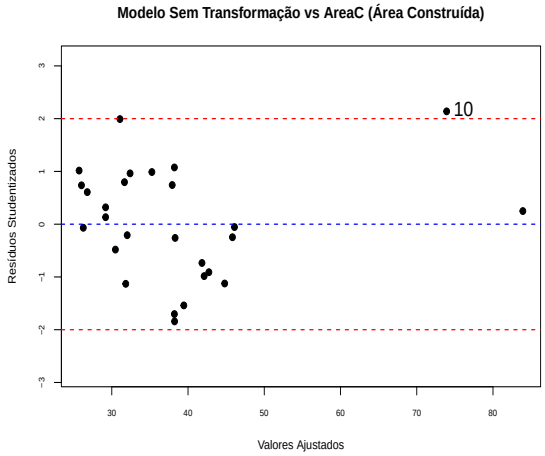
Tabela 15: Impacto dos Pontos Influentes nos Coeficientes da Regressão para Área Total e Área Construída

Modelo	β_0	β_1	p-valor
$y = \log(preco) \sim areaT$	3.02887	0.09051	6.8×10^{-6}
Sem ponto 8	2.97184	0.10292	1.24×10^{-7}
Sem ponto 9	3.09362	0.07684	1.2×10^{-5}
Sem ponto 10	3.10798	0.07609	4.05×10^{-4}
Sem ponto 21	2.99297	0.09463	4.48×10^{-6}
Sem pontos 8, 9, 10, 21	2.9426	0.4406	2.56×10^{-4}
$y = boxcox(preco) \sim areaT$	0.7064	0.0004753	1.35×10^{-5}
Sem ponto 8	0.7060	0.0005467	1.6×10^{-7}
Sem ponto 9	0.7066	0.0004272	5.55×10^{-5}
Sem ponto 10	0.7065	0.0004518	2.76×10^{-4}
Sem ponto 21	0.7061	0.0005027	5.97×10^{-6}
Sem pontos 8, 9, 10, 21	0.7060	0.0005310	9.82×10^{-6}

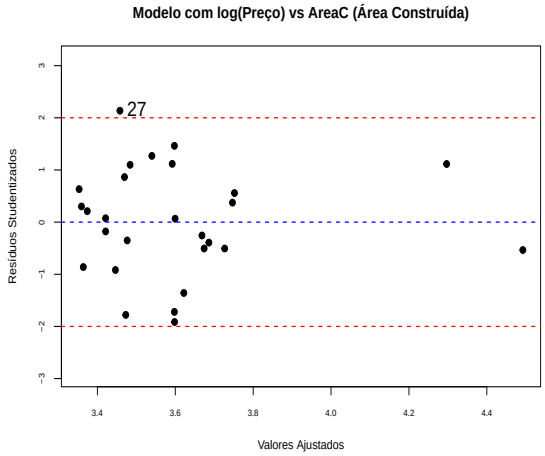
2.3.2 Diagnóstico covariável - Área Construída

Os gráficos de resíduos studentizados foram utilizados para avaliar a adequação de cada modelo ajustado às suposições da regressão linear, eles oferecem insights sobre a adequação dos modelos e possíveis desvios das suposições da regressão linear, como homoscedasticidade e ausência de padrões nos resíduos.

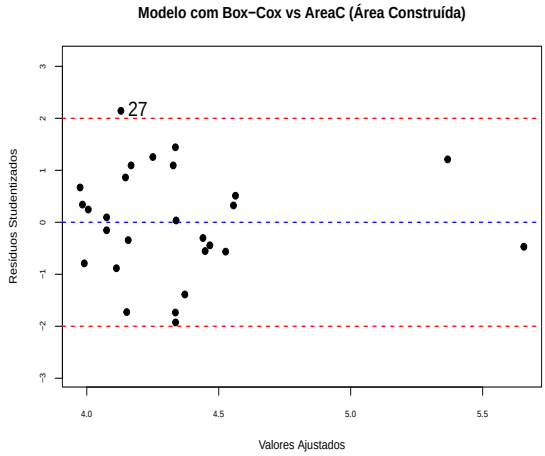
No modelo sem transformação, o ponto 10 é identificado como outlier, com resíduos studentizados elevados. Ao aplicar a transformação logarítmica, a distribuição dos resíduos se torna mais homogênea, mas o ponto 27 se destaca como outlier. No modelo ajustado com a transformação Box-Cox, o ponto 27 ainda se mantém como outlier.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)



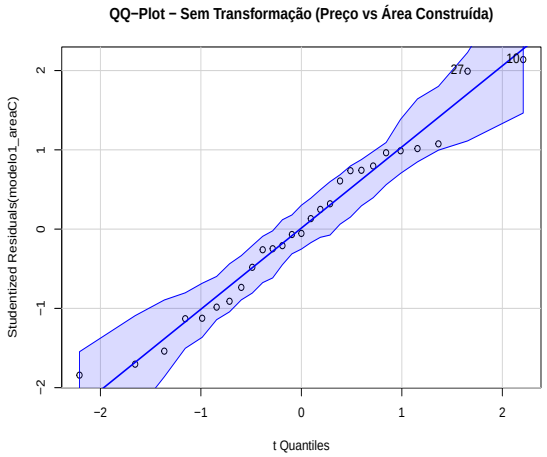
(c) Modelo com Box-Cox

Figura 53: Gráficos de resíduos studentizados versus valores ajustados para os modelos ajustados: sem transformação, com log(Preço) e com transformação Box-Cox.

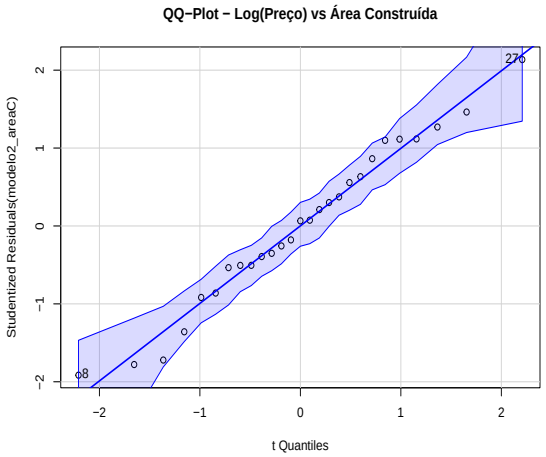
No QQ-Plot - Sem Transformação (Preço vs Área Construída), os resíduos seguem a linha central na maior parte da distribuição, mas há desvios significativos nas extremidades.

No QQ-Plot - Log(Preço) vs Área Construída, os pontos estão mais alinhados à reta central, mostrando uma melhora na distribuição dos resíduos. No entanto, ainda há alguma dispersão nas extremidades.

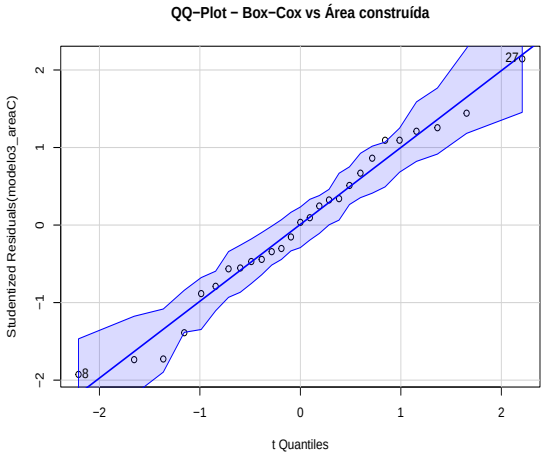
Já no QQ-Plot - Box-Cox vs Área Construída, os pontos estão bem ajustados à linha central, porém pequenas discrepâncias ainda são observadas nas extremidades e pontos ligeiramente fora do limite.



(a) Modelo sem transformação



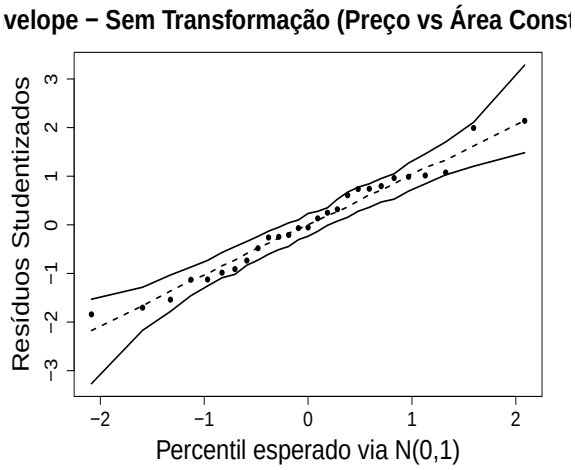
(b) Modelo com log(Preço) VS Área Construída



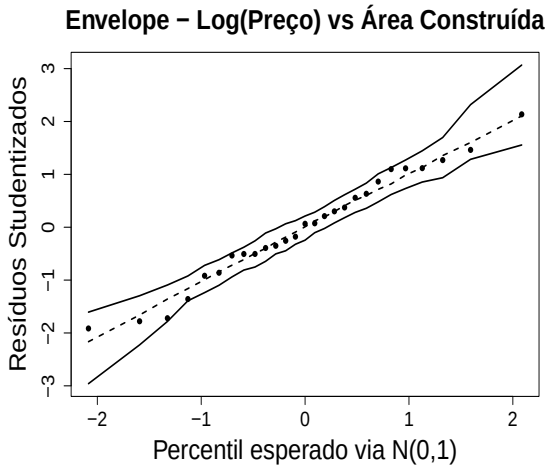
(c) Modelo com Box-Cox (Área Construída)

Figura 54: Gráficos de probabilidade normal dos resíduos Studentizados para modelos ajustados à Área Construída

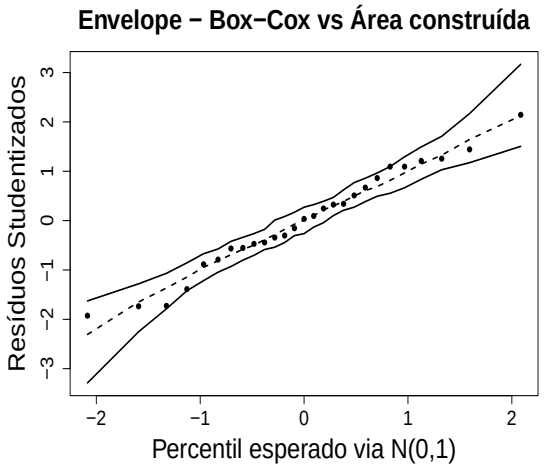
Complementando a análise dos gráficos QQ-Plots, os gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado na figura 51 mostram que os resíduos permanecem, na sua maior parte, dentro da faixa de confiança esperada para uma distribuição normal N(0,1), o que confirma que a suposição de normalidade é provavelmente atendida nos três modelos.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço) VS Área Construída



(c) Modelo com Box-Cox VS Área Construída

Figura 55: Gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado

Porém, para uma análise mais concreta optou-se por aplicar o teste de Shapiro-Wilk em ambos os modelos, onde o p-valor indica se rejeitamos ou não H_0:

$p > 0.05 \Rightarrow$ Não rejeitamos $H_0 \Rightarrow$ Os resíduos podem ser considerados normais.

$p \leq 0.05 \Rightarrow$ Rejeitamos $H_0 \Rightarrow$ Os resíduos não seguem uma distribuição normal.

Em todos os casos, os p-valores foram maiores que 0.05. Conclui-se que a hipótese nula de normalidade dos resíduos não é rejeitada. A suposição de normalidade é razoavelmente atendida para todos os modelos.

Modelo	**p-valor**
Sem transformação	0.7105
Log(Preço)	0.8491
Box-Cox	0.8481

Tabela 16: Resultados do teste de normalidade Shapiro-Wilk para os modelos com a covariável Área Construída

Prosseguindo com o diagnóstico, foi realizado o teste de Goldfeld-Quandt para verificar a presença de heretocedasticidade nos nos resíduos dos três modelos de regressão.

A interpretação do teste se baseia no p-valor, onde:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de heterocedasticidade.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, indicando heterocedasticidade significativa.

Em todos os testes, os p-valores foram superiores a 0.05, o que permite manter a hipótese nula de homocedasticidade. Isso sugere que os resíduos dos modelos sem transformação, logarítmico e Box-Cox são homogêneos, e a suposição de variâncias constantes é válida para os três modelos.

Modelo	**p-valor**
Sem transformação	0.6076
Log(Preço)	0.7648
Box-Cox	0.7545

Tabela 17: Resultados do teste Goldfeld-Quandt para os modelos com a covariável Área Construída

Agora, para verificar a presença de autocorrelação dos resíduos foi aplicado o teste de Durbin-Watson.

A interpretação do p-valor do teste de Durbin-Watson segue a seguinte lógica:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de autocorrelação.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, há evidências de autocorrelação nos resíduos.

Os resultados do teste de Durbin-Watson indicam que o modelo sem transformação não apresenta autocorrelação significativa nos resíduos, com um p-valor de 0.3045. Por outro lado, os modelos com transformação logarítmica e Box-Cox apresentam autocorrelação significativa, conforme indicam os p-valores de 0.01173 e 0.01847.

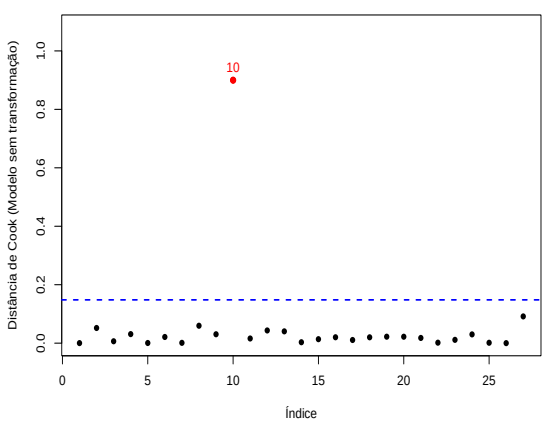
Modelo	**p-valor**
Sem transformação	0.3045
Log(Preço)	0.01173
Box-Cox	0.01847

Tabela 18: Resultados do teste Durbin-Watson para os modelos com a covariável Área Construída

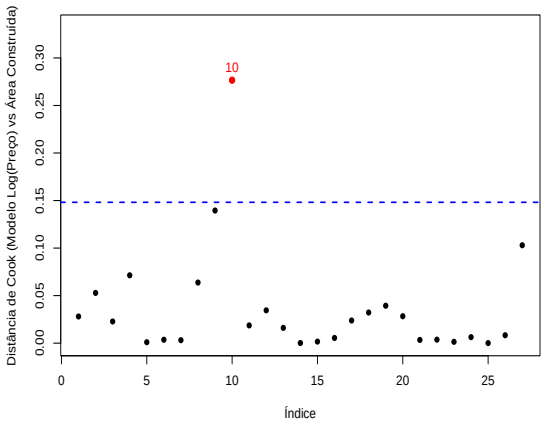
Avançando mais com o diagnótico, analisaremos algumas medidas de influência.

Para analisar o impacto global de cada observação no modelo ajustado, utilizamos a métrica da Distância de Cook. Essa medida permite identificar pontos que influenciam significativamente os coeficientes da regressão, podendo indicar observações que afetam de maneira relevante os ajustes realizados.

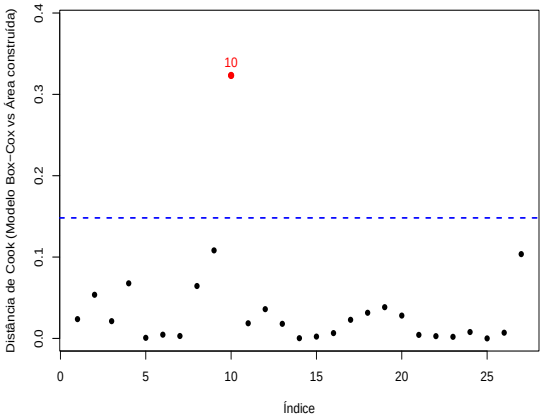
A partir dos gráficos, nota-se que no Modelo Sem Transformação, no Modelo com Log(Preço) e no Modelo Box-Cox a observação 10 se destaca, indicando esse ponto como um possível influente.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

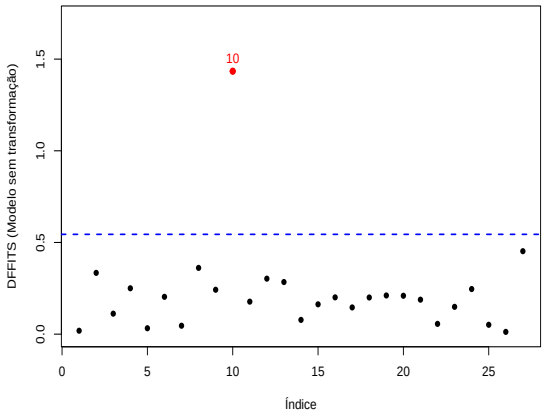


(c) Modelo com Box-Cox

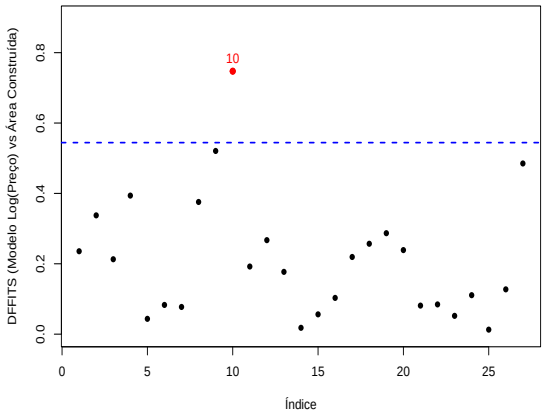
Figura 56: Gráficos da Distância de Cook para os três modelos ajustados para a covariável Área Construída

Embora a Distância de Cook forneça uma visão geral da influência das observações, a métrica DFFITS permite avaliar especificamente o impacto de cada ponto na predição do modelo.

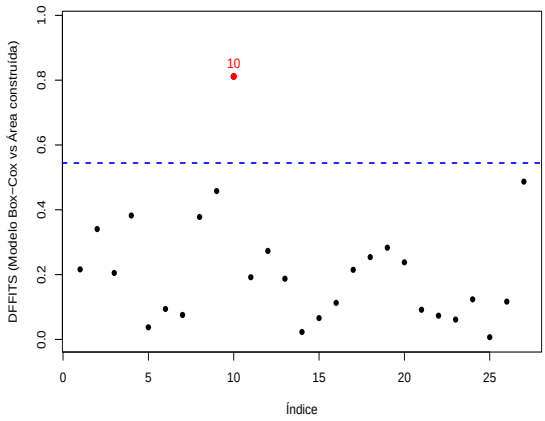
Os gráficos da métrica DFFITS mostram também que, nos três modelos a observação 10 aparece como possível ponto influente.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

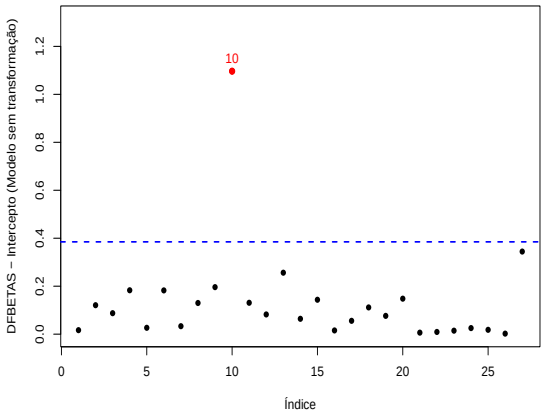


(c) Modelo com Box-Cox

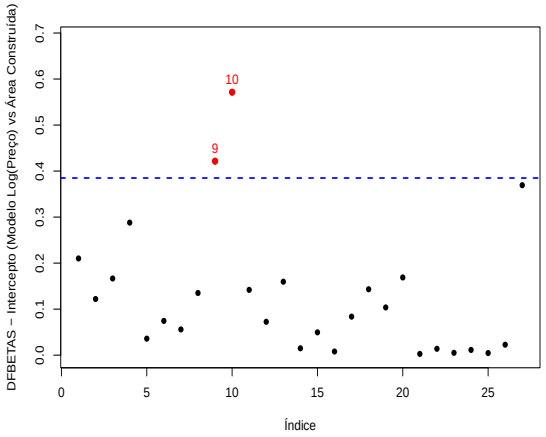
Figura 57: Gráficos da métrica DFFITS para os três modelos ajustados para a covariável Área Construída

Além de analisar a influência sobre as predições, é importante avaliar como cada observação afeta individualmente os coeficientes do modelo. Para isso, utilizamos a métrica DFBETAS, que mede a variação nos coeficientes ao remover uma determinada observação. Primeiramente, analisamos o impacto sobre o intercepto.

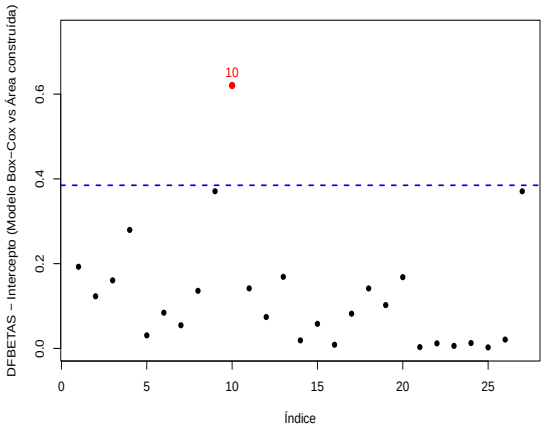
Os gráficos da métrica DFBETAS para o intercepto indicam que, no Modelo Sem Transformação e no Modelo com Box-Cox, a observação 10 aparece como possível ponto influente. Já no Modelo log(Preço), os pontos 9 e 10 aparecem como possíveis influentes.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

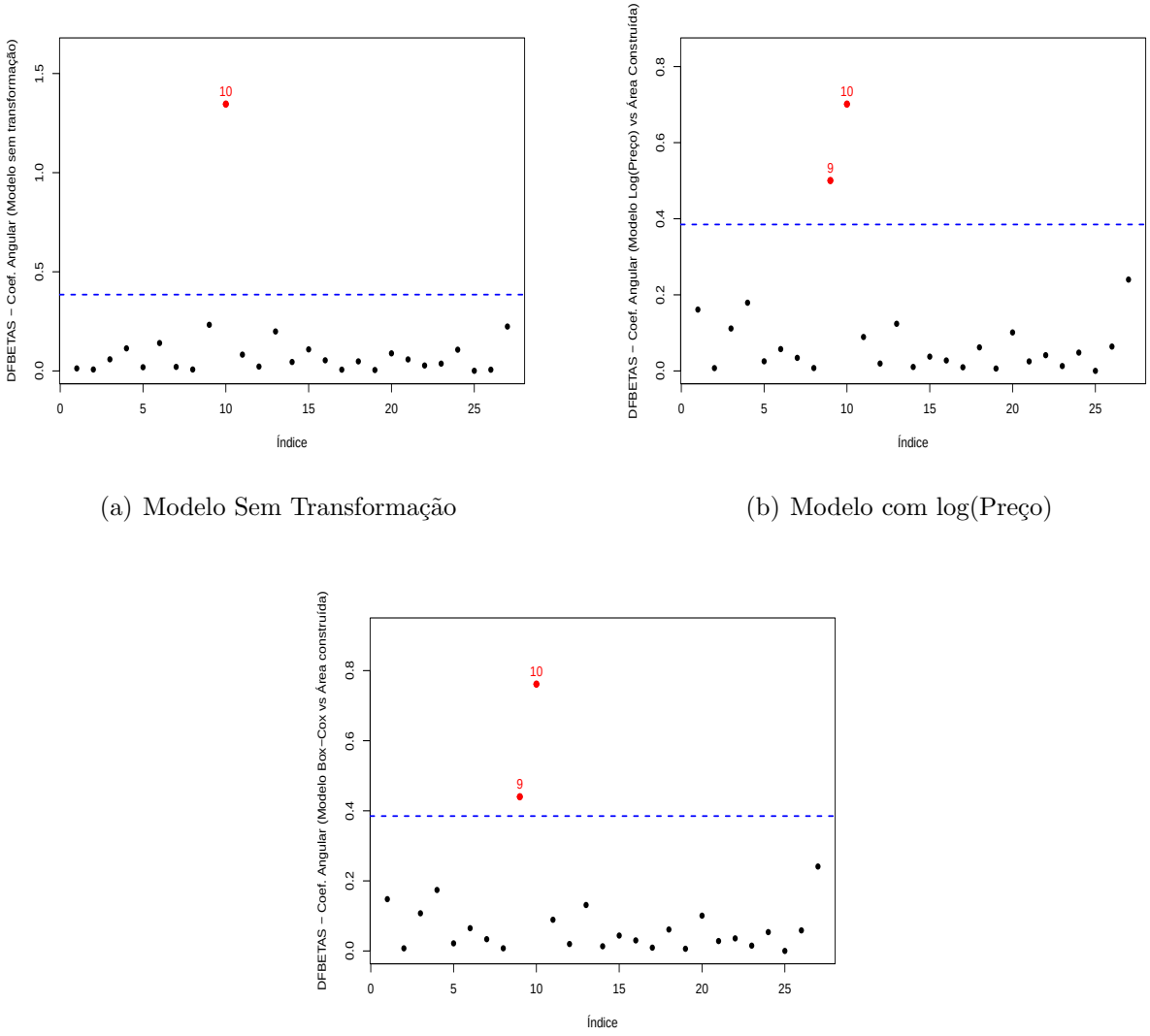


(c) Modelo com Box-Cox

Figura 58: Gráficos da métrica DFBETAS para o intercepto nos três modelos ajustados para a covariável Área Construída

Além do intercepto, também é essencial verificar a influência das observações sobre o coeficiente angular, que representa a inclinação da relação entre as variáveis.

Os gráficos da métrica DFBETAS para o coeficiente angular indicam que, no Modelo log(Preço) e no Modelo com Box-Cox, as observações 9 e 10 aparecem como possíveis pontos influentes. Já no Modelo Sem Transformação, o ponto 10 aparece como possível influente.



(a) Modelo Sem Transformação

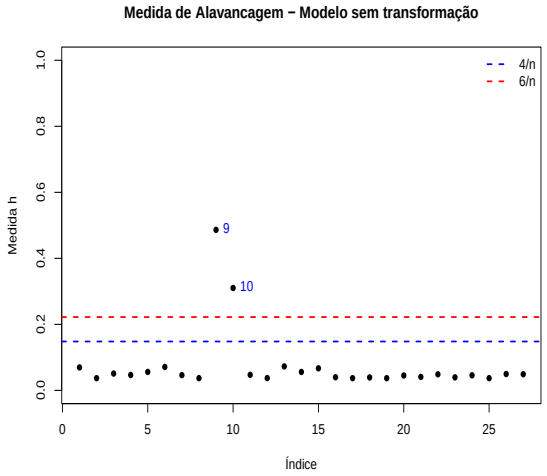
(b) Modelo com log(Preço)

(c) Modelo com Box-Cox

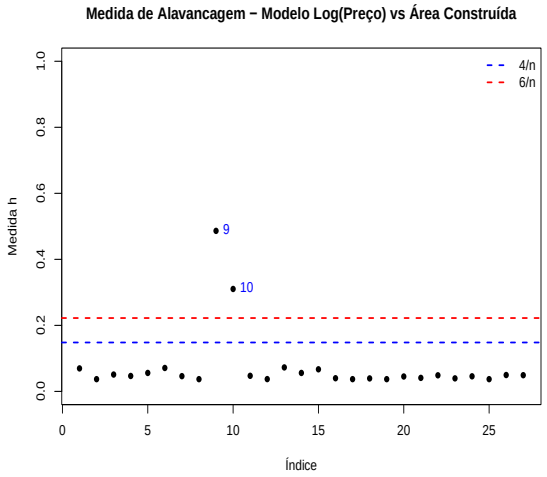
Figura 59: Gráficos da métrica DFBETAS para o coeficiente angular nos três modelos ajustados para a covariável Área Construída

Os gráficos de alavancagem mostram a medida h de cada observação nos diferentes modelos ajustados. As linhas de referência 4/n e 6/n indicam limites convencionais para identificar observações com alto impacto na estimativa dos coeficientes.

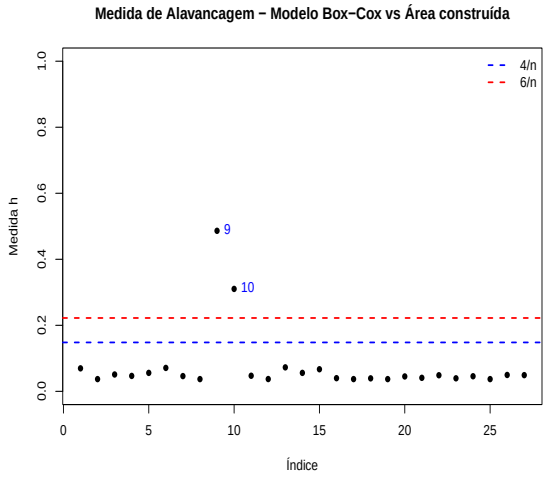
Modelo sem transformação, Modelo Log(Preço) vs Área Construída e Modelo Box-Cox vs Área Construída destacam as observações 9 e 10 como pontos com alta alavancagem.



(a) Modelo sem transformação



(b) Modelo Log(Preço) vs Área Construída



(c) Modelo Box-Cox vs Área Construída

Figura 60: Gráficos de alavancagem para os modelos baseados na Área Construída

Finalizando a etapa de diagnóstico, analisaremos o impacto dos pontos influentes nos coeficientes dos modelos. A tabela abaixo apresenta os coeficientes da regressão dos modelos, além da remoção de possíveis pontos influentes identificados nos gráficos de alavancagem (observações 9, 10 e 27).

No Modelo Sem Transformação, os pontos influentes parecem estar superestimando a relação entre área construída e preço.

No Modelo log(Preço), as mudanças nos coeficientes são menores, indicando que a transformação reduz o impacto desses pontos.

Já o modelo Box-Cox se comporta de forma parecida ao logarítmico.

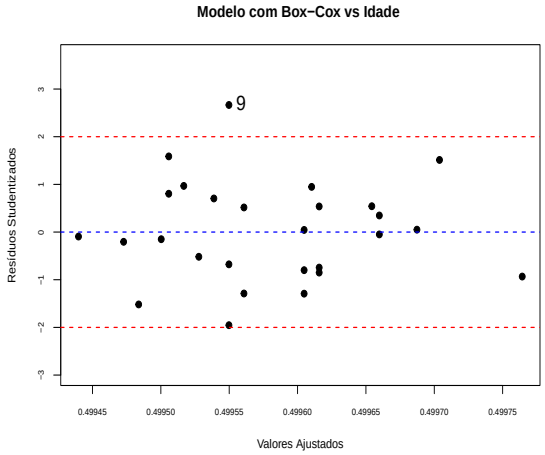
Tabela 19: Impacto dos Pontos Influentes nos Coeficientes da Regressão para Área Construída

Modelo	β_0 (Intercepto)	β_1 (Área Construída)	p-valor (β_1)
$y = preco \sim areaC$	2.506	23.804	2.81×10^{-12}
Sem ponto 9	3.116	23.354	5.47×10^{-9}
Sem ponto 10	5.639	21.414	3.32×10^{-10}
Sem ponto 27	1.511	24.207	1.24×10^{-12}
Sem pontos 9, 10, 27	13.260	15.433	1.01×10^{-4}
$y = \log(preco) \sim areaC$	2.8987	0.4660	2.56×10^{-10}
Sem ponto 9	2.8670	0.4894	6.98×10^{-8}
Sem ponto 10	2.9409	0.4339	3.12×10^{-8}
Sem ponto 27	2.8731	0.4764	7.70×10^{-11}
Sem pontos 9, 10, 27	2.9202	0.4407	1.14×10^{-4}
$y = \text{Box-Cox}(preco) \sim areaC$	3.3042	0.6877	1.42×10^{-10}
Sem ponto 9	3.2642	0.7172	4.87×10^{-8}
Sem ponto 10	3.3696	0.6378	1.84×10^{-8}
Sem ponto 27	3.2674	0.7026	4.27×10^{-11}
Sem pontos 9, 10, 27	3.3657	0.6285	1.12×10^{-4}

2.3.3 Diagnóstico covariável - Idade

O gráfico de resíduos studentizados foi utilizado para avaliar a adequação do modelo ajustado às suposições da regressão linear, ele oferece insights sobre a adequação do modelo e possíveis desvios das suposições da regressão linear, como homoscedasticidade e ausência de padrões nos resíduos.

No modelo ajustado com a transformação Box-Cox, o ponto 9 é identificado como outlier.



(a) Modelo com Box-Cox

Figura 61: Gráfico de resíduos studentizados versus valores ajustados para o modelo ajustado: transformação Box-Cox.

No QQ-Plot - Box-Cox VS Idade, os resíduos seguem a linha central na maior parte da distribuição, mas há desvios significativos nas extremidades.

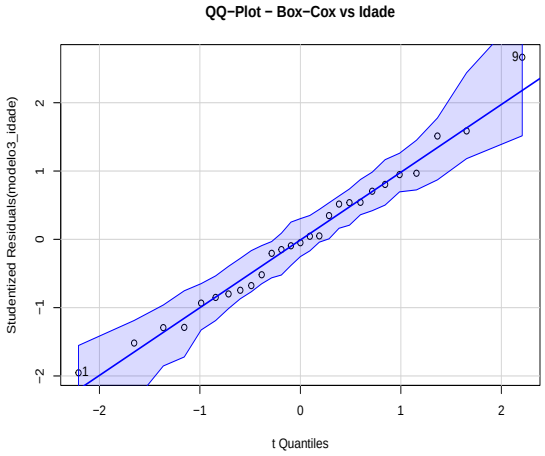


Figura 62: Gráfico de probabilidade normal dos resíduos Studentizados para o modelo com Box-Cox (Idade)

Complementando a análise do gráfico QQ-Plot, o gráfico de probabilidade normal dos resíduos Studentizados com envelope simulado na figura 57 mostram que os resíduos permanecem, na sua maior parte, dentro da faixa de confiança esperada para uma distribuição normal N(0,1), o que confirma que a suposição de normalidade é provavelmente atendida nos modelo.

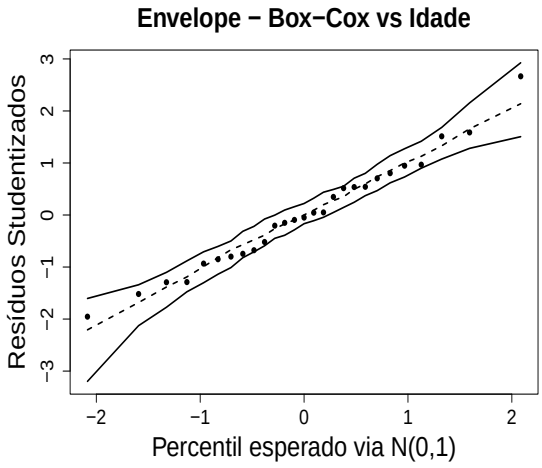


Figura 63: Gráfico de probabilidade normal dos resíduos Studentizados para o modelo com Box-Cox (Idade) com envelope simulado

Porém, para uma análise mais concreta optou-se por aplicar o teste de Shapiro-Wilk em ambos os modelos, onde o p-valor indica se rejeitamos ou não H_0:

$p > 0.05 \Rightarrow$ Não rejeitamos $H_0 \Rightarrow$ Os resíduos podem ser considerados normais.

$p \leq 0.05 \Rightarrow$ Rejeitamos $H_0 \Rightarrow$ Os resíduos não seguem uma distribuição normal.

O modelo Box-Cox apresentou um p-valor (0.9219) muito maior que 0.05, ou seja, não há evidências para rejeitar a hipótese nula de normalidade. Isso sugere que os resíduos do modelo Box-Cox seguem uma distribuição normal.

Modelo	**p-valor**
Box-Cox	0.9219

Tabela 20: Resultado do teste de normalidade Shapiro-Wilk para o modelo com a covariável Idade

Prosseguindo com o diagnóstico, foi realizado o teste de Goldfeld-Quandt para verificar a presença de heretocedasticidade nos nos resíduos dos três modelos de regressão.

A interpretação do teste se baseia no p-valor, onde:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de heterocedasticidade.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, indicando heterocedasticidade significativa.

Os resíduos do modelo Box-Cox não apresentam heterocedasticidade, já que o teste de Goldfeld-Quandt não rejeitou a hipótese nula de homocedasticidade, conforme o p-valor alto (0.6015). Isso indica que o modelo está em conformidade com a suposição de variâncias constantes.

Modelo	**p-valor**
Box-Cox	0.6015

Tabela 21: Resultados do teste Goldfeld-Quandt para os modelos com a covariável Idade

Agora, para verificar a presença de autocorrelação dos resíduos foi aplicado o teste de Durbin-Watson.

A interpretação do p-valor do teste de Durbin-Watson segue a seguinte lógica:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de autocorrelação.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, há evidências de autocorrelação nos resíduos.

O modelo Box-Cox apresentou um p-valor igual a 0.07999, o que indica que há uma autocorrelação positiva significativa. Sugere que a suposição de independência dos erros não é totalmente atendida.

Modelo	**DW**	**p-valor**
Box-Cox	1.3732	0.07999

Tabela 22: Resultados do teste Durbin-Watson para os modelos com a covariável Idade

Avançando mais com o diagnótico, analisaremos algumas medidas de influência.

Para analisar o impacto global de cada observação no modelo ajustado, utilizamos a métrica da Distância de Cook. Essa medida permite identificar pontos que influenciam significativamente os coeficientes da regressão, podendo indicar observações que afetam de maneira relevante os ajustes realizados.

A partir do gráfico, nota-se que no Modelo Box-Cox as observações 10 e 25 se destacam, indicando esses pontos como possíveis pontos influentes.

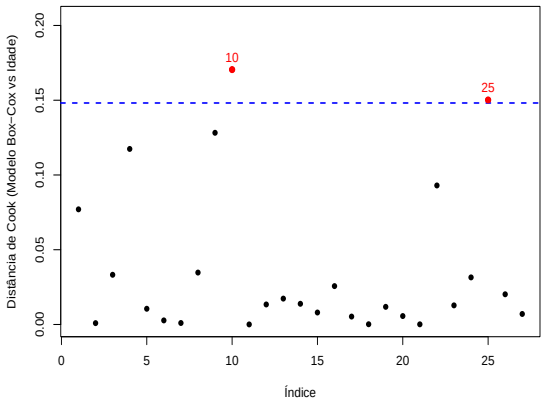


Figura 64: Gráficos da Distância de Cook para o modelo Box-Cox para a covariável Idade

Embora a Distância de Cook forneça uma visão geral da influência das observações, a métrica DFFITS permite avaliar especificamente o impacto de cada ponto na predição do modelo.

Os gráficos da métrica DFFITS mostram que, no modelo as observações 9, 10 e 25 aparecem como possíveis influentes.

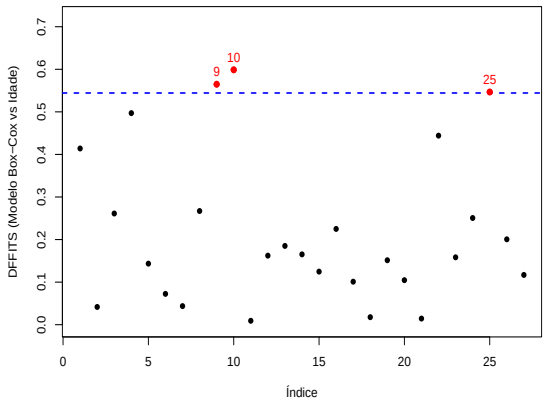


Figura 65: Gráficos da métrica DFFITS para o modelo Box-Cox para a covariável Idade

Além de analisar a influência sobre as predições, é importante avaliar como cada observação afeta individualmente os coeficientes do modelo. Para isso, utilizamos a métrica DFBETAS, que mede a variação nos coeficientes ao remover uma determinada observação. Primeiramente, analisamos o impacto sobre o intercepto.

Os gráfico da métrica DFBETAS para o intercepto indica que, no Modelo com Box-Cox, as observações 10 e 25 aparecem como possíveis pontos influentes.

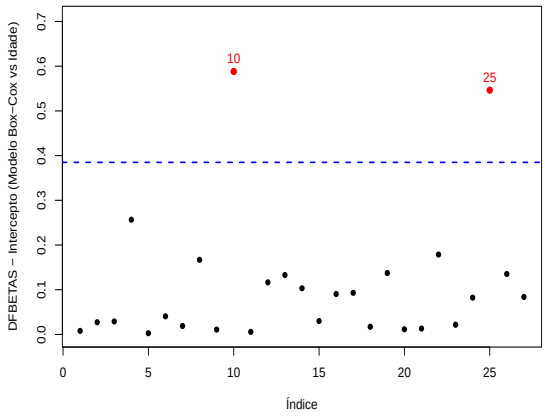


Figura 66: Gráficos da métrica DFBETAS para o intercepto para o modelo Box-Cox para a covariável Área Idade

Além do intercepto, também é essencial verificar a influência das observações sobre o coeficiente angular, que representa a inclinação da relação entre as variáveis.

O gráfico da métrica DFBETAS para o coeficiente angular indica que, no Modelo com Box-Cox, as observações 4, 10 e 25 aparecem como possíveis pontos influentes.

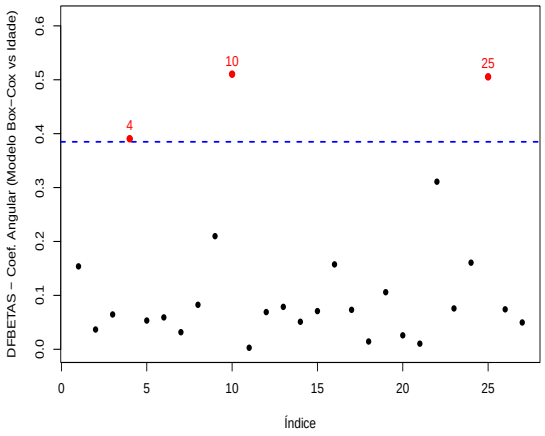
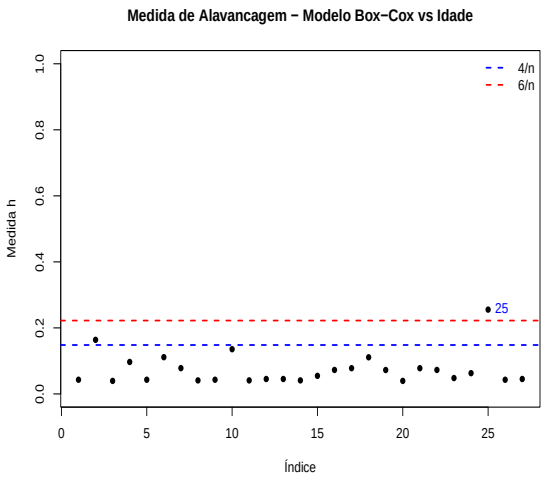


Figura 67: Gráficos da métrica DFBETAS para o coeficiente angular no modelo Box-Cox para a covariável Idade

O gráfico de alavancagem mostram a medida h de cada observação nos diferentes modelos ajustados. As linhas de referência 4/n e 6/n indicam limites convencionais para identificar observações com alto impacto na estimativa dos coeficientes.

O Modelo Box-Cox vs Idade destaca a observação 25 como um ponto influente.



(a) Modelo Box-Cox vs Idade

Figura 68: Gráfico de alavancagem para o modelo baseado na Idade

Finalizando a etapa de diagnóstico, analisaremos o impacto dos pontos influentes nos coeficientes do modelo. A tabela abaixo apresenta os coeficientes da regressão do modelo, além da remoção de possíveis pontos influentes identificados nos gráficos de alavancagem (observações 4, 9, 10 e 25).

No modelo em questão, foi observado que a relação entre idade e preço pode estar sendo impulsionada por poucas observações atípicas, e que, sem elas, a evidência estatística da associação se enfraquece.

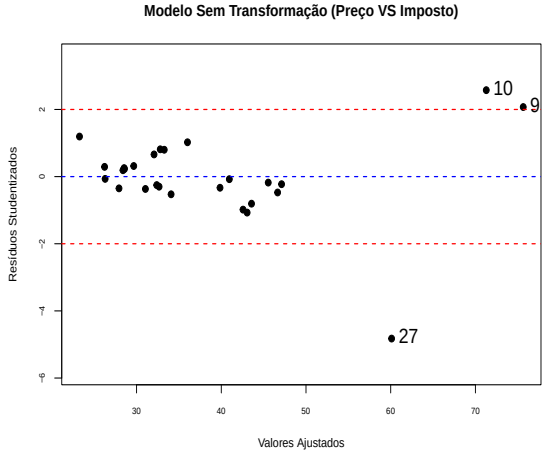
Tabela 23: Impacto dos Pontos Influentes nos Coeficientes da Regressão para Idade

Modelo	β_0 (Intercepto)	β_1 (Idade)	p-valor (β_1)
$y = \text{Box-Cox}(preco) \sim idade$	0.4998	-5.501e-06	2.31×10^{-2}
Sem ponto 4	0.4998	-4.636e-06	5.40×10^{-2}
Sem ponto 9	0.4998	-5.929e-06	7.85×10^{-3}
Sem ponto 10	0.4997	-4.370e-06	7.41×10^{-2}
Sem ponto 25	0.4998	-6.653e-06	1.69×10^{-2}
Sem pontos 4, 9, 10, 25	0.4997	-4.580e-06	7.55×10^{-2}

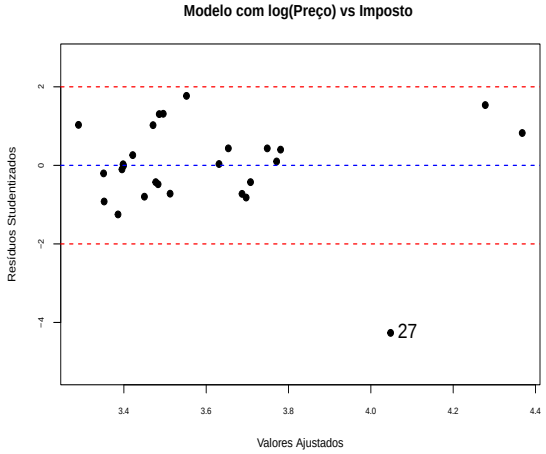
2.3.4 Diagnóstico covariável - Imposto

Os gráficos de resíduos studentizados foram utilizados para avaliar a adequação de cada modelo ajustado às suposições da regressão linear, eles oferecem insights sobre a adequação dos modelos e possíveis desvios das suposições da regressão linear, como homoscedasticidade e ausência de padrões nos resíduos.

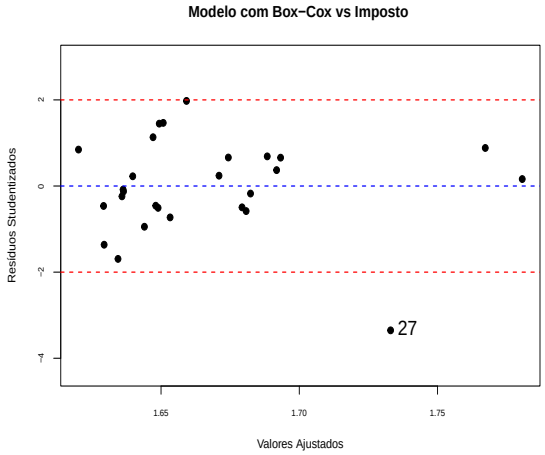
No modelo sem transformação, observa-se a presença de três outliers significativos (pontos 9, 10 e 27), cujos resíduos studentizados ultrapassam a faixa de ± 2. Ao aplicar a transformação logarítmica na variável preditora, a distribuição dos resíduos aparenta estar mais homogênea, mas o ponto 27 ainda se destaca como outlier. No modelo ajustado com a transformação Box-Cox, a dispersão dos resíduos continua ao redor de zero, mas o ponto 27 mantém-se como um outlier.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)



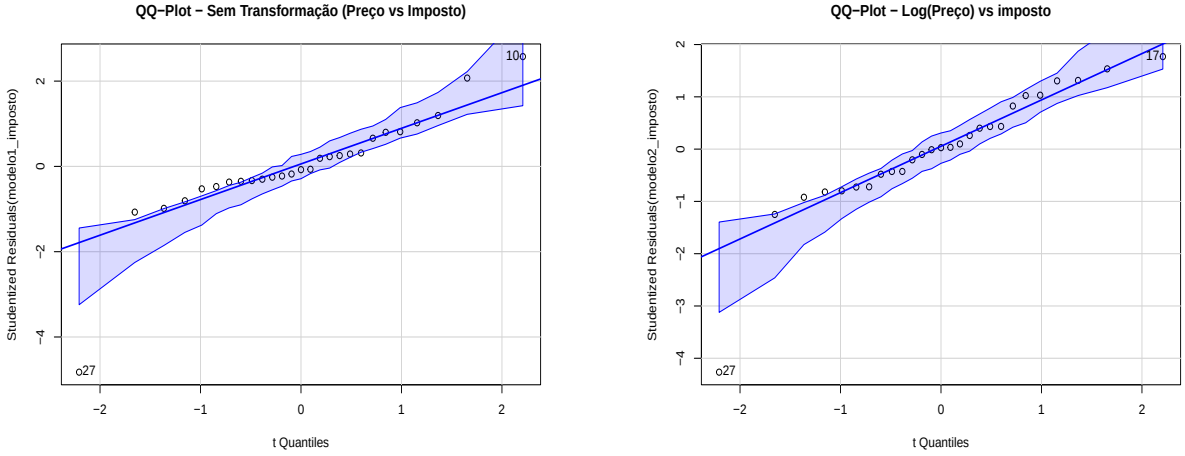
(c) Modelo com Box-Cox

Figura 69: Gráficos de resíduos studentizados versus valores ajustados para os modelos ajustados: sem transformação, com log(Preço) e com transformação Box-Cox.

No QQ-Plot - Sem Transformação (Preço vs Imposto), alguns resíduos aparecem fora do limite estabelecido.

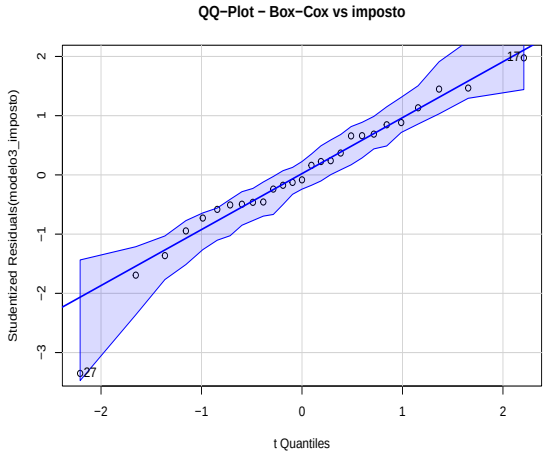
No QQ-Plot - Log(Preço) vs Imposto, os pontos estão mais alinhados à reta central. No entanto, ainda há alguma dispersão durante a reta e alguns pontos fora do limite estabelecido.

Já no QQ-Plot - Box-Cox vs Imposto, os pontos estão mais ajustados à linha central, porém pequenas discrepâncias ainda são observadas nas extremidades.



(a) Modelo sem transformação

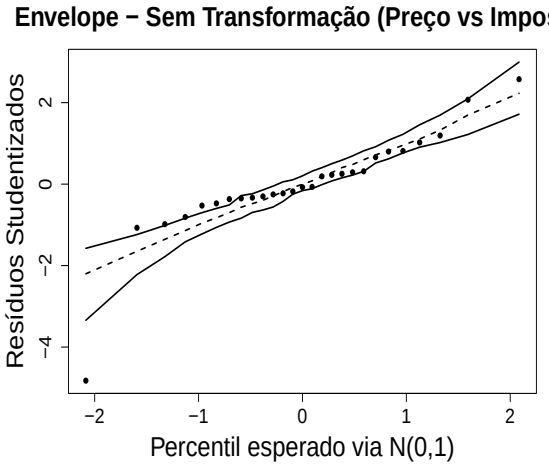
(b) Modelo com log(Preço)



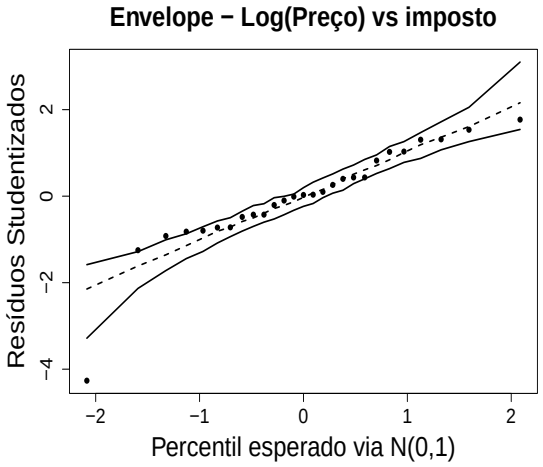
(c) Modelo com Box-Cox

Figura 70: Gráficos de probabilidade normal dos resíduos Studentizados para modelos ajustados ao Imposto

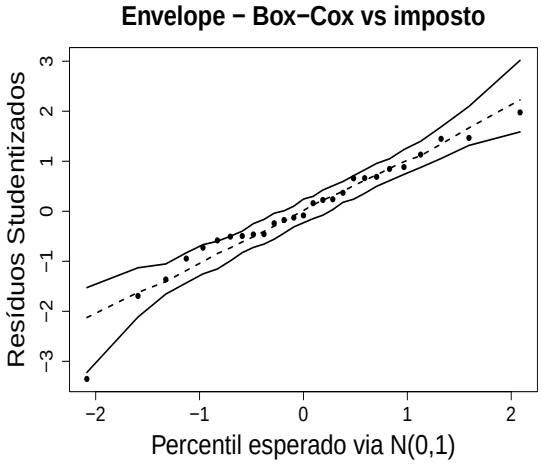
Complementando a análise do gráfico QQ-Plot, o gráfico de probabilidade normal dos resíduos Studentizados com envelope simulado na figura 63 mostram que alguns resíduos no Modelo Sem tranformação se encontram fora do envelope, igualmente ao Modelo Log(Preço). Diferentemente, o Modelo Box-Cox mostra os resíduos dentro da faixa de confiança esperada para uma distribuição normal N(0,1), o que confirma que a suposição de normalidade é provavelmente atendida neste modelo.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)



(c) Modelo com Box-Cox

Figura 71: Gráficos de probabilidade normal dos resíduos Studentizados com envelope simulado

Porém, para uma análise mais concreta optou-se por aplicar o teste de Shapiro-Wilk em ambos os modelos, onde o p-valor indica se rejeitamos ou não H_0:

$p > 0.05 \Rightarrow$ Não rejeitamos $H_0 \Rightarrow$ Os resíduos podem ser considerados normais.

$p \leq 0.05 \Rightarrow$ Rejeitamos $H_0 \Rightarrow$ Os resíduos não seguem uma distribuição normal.

No Modelo Sem Transformação, o p-valor é muito baixo (0.0002936), o que indica que a hipótese nula de normalidade dos resíduos deve ser rejeitada, ou seja, não seguem uma distribuição normal. Não obstante, no Modelo Log(Preço) o p-valor (0.002702) também é baixo. Portanto, não seguem uma distribuição normal. Já no Modelo Box-Cox, o p-valor foi de 0.2248, maior que 0.05, indicando que não há evidências suficientes para rejeitar a hipótese nula, sugerindo que os resíduos do modelo Box-Cox seguem uma

distribuição normal.

Modelo	**p-valor**
Sem transformação	0.0002936
Log(Preço)	0.002702
Box-Cox	0.2248

Tabela 24: Resultados do teste de normalidade Shapiro-Wilk para os modelos com a covariável Imposto

Prosseguindo com o diagnóstico, foi realizado o teste de Goldfeld-Quandt para verificar a presença de heretocedasticidade nos nos resíduos dos três modelos de regressão.

A interpretação do teste se baseia no p-valor, onde:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de heterocedasticidade.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, indicando heterocedasticidade significativa.

No Modelo Sem Transformação, o p-valor muito baixo (0.0006811) indica que há evidências para rejeitar a hipótese nula de homocedasticidade, ou seja, as variâncias dos erros não são constantes. No Modelo Log(Preço), o p-valor (0.04073) ainda é relativamente baixo, indicando que também há heterocedasticidade nos resíduos. Já no Modelo Box-Cox, o p-valor (0.1649) é maior que 0.05, isso indica que a suposição de variâncias constantes é atendida para este modelo.

Modelo	**p-valor**
Sem transformação	0.0006811
Log(Preço)	0.04073
Box-Cox	0.1649

Tabela 25: Resultados do teste Goldfeld-Quandt para os modelos com a covariável Imposto

Agora, para verificar a presença de autocorrelação dos resíduos foi aplicado o teste de Durbin-Watson.

A interpretação do p-valor do teste de Durbin-Watson segue a seguinte lógica:

$p > 0.05 \Rightarrow$ Não rejeitamos H_0, ou seja, não há evidências de autocorrelação.

$p \leq 0.05 \Rightarrow$ Rejeitamos H_0, há evidências de autocorrelação nos resíduos.

Os resultados do teste de Durbin-Watson indicam que, para os três modelos, não há autocorrelação significativa nos resíduos (já que os p-valores são todos maiores que 0.05).

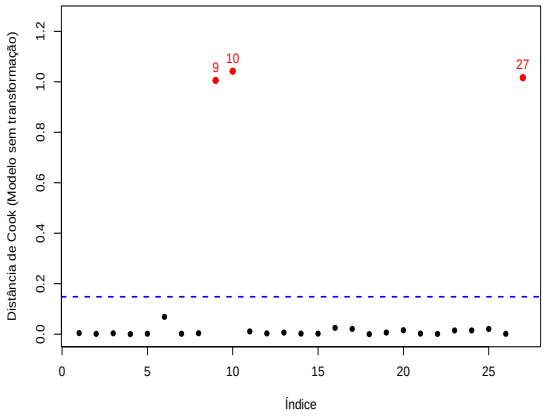
Modelo	**p-valor**
Sem transformação	0.05801
Log(Preço)	0.1708
Box-Cox	0.09331

Tabela 26: Resultados do teste Durbin-Watson para os modelos com a covariável Imposto

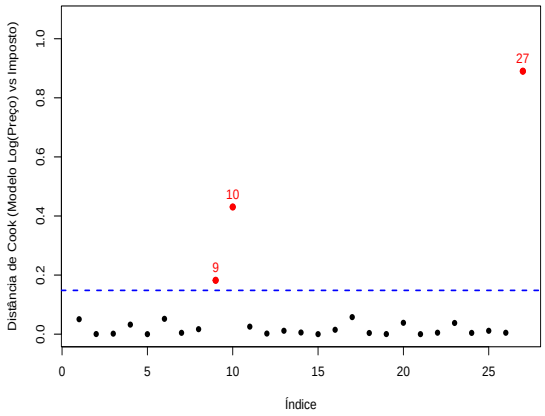
Avançando mais com o diagnótico, analisaremos algumas medidas de influência.

Para analisar o impacto global de cada observação no modelo ajustado, utilizamos a métrica da Distância de Cook. Essa medida permite identificar pontos que influenciam significativamente os coeficientes da regressão, podendo indicar observações que afetam de maneira relevante os ajustes realizados.

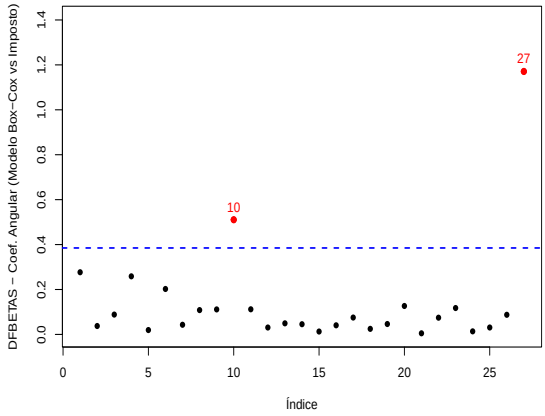
A partir dos gráficos, nota-se que no Modelo Sem Transformação e no Modelo com Log(Preço), os pontos 9, 10 e 27 aparecem como possíveis pontos influentes. E no Modelo Box-Cox as observações 10 e 27 se destacam.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

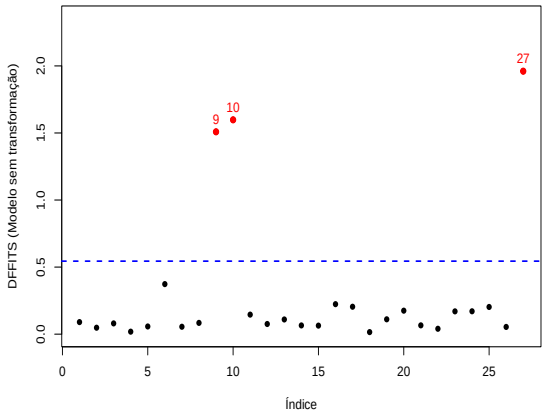


(c) Modelo com Box-Cox

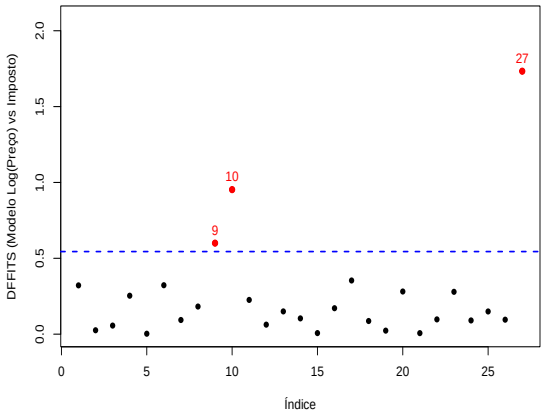
Figura 72: Gráficos da Distância de Cook para os três modelos ajustados para a covariável Imposto

Embora a Distância de Cook forneça uma visão geral da influência das observações, a métrica DFFITS permite avaliar especificamente o impacto de cada ponto na predição do modelo.

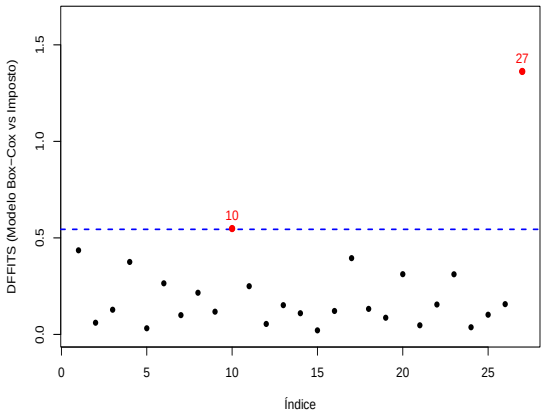
Igualmente aos gráficos da Distância de Cook, no Modelo Sem Transformação e no Modelo com Log(Preço), os pontos 9, 10 e 27 aparecem como possíveis pontos influentes. Já no Modelo Box-Cox as observações 10 e 27 se destacam.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

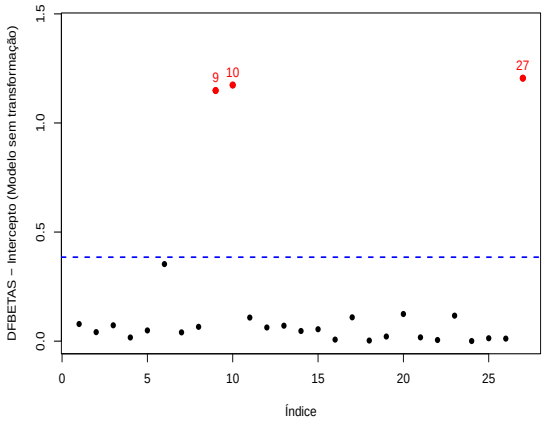


(c) Modelo com Box-Cox

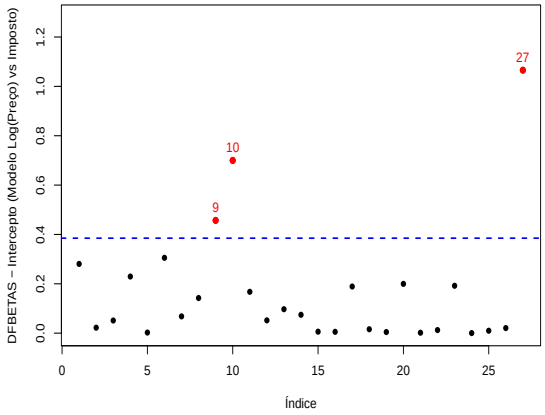
Figura 73: Gráficos da métrica DFFITS para os três modelos ajustados para a covariável Imposto

Além de analisar a influência sobre as predições, é importante avaliar como cada observação afeta individualmente os coeficientes do modelo. Para isso, utilizamos a métrica DFBETAS, que mede a variação nos coeficientes ao remover uma determinada observação. Primeiramente, analisamos o impacto sobre o intercepto.

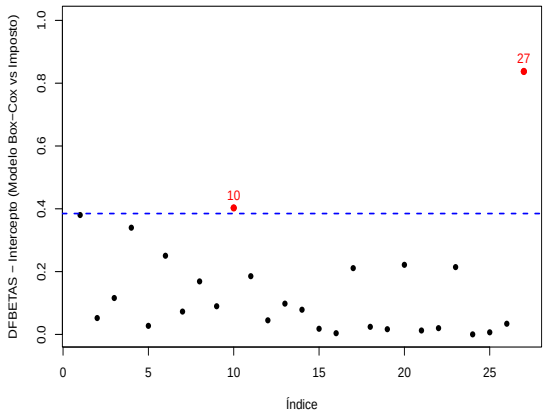
Os gráficos da métrica DFBETAS para o intercepto indicam que, no Modelo Sem Transformação e no Modelo log(Preço), os pontos 9, 10 e 27 aparecem como possíveis influentes. Já no Modelo Box-Cox, os pontos 10 e 27 se destacam.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

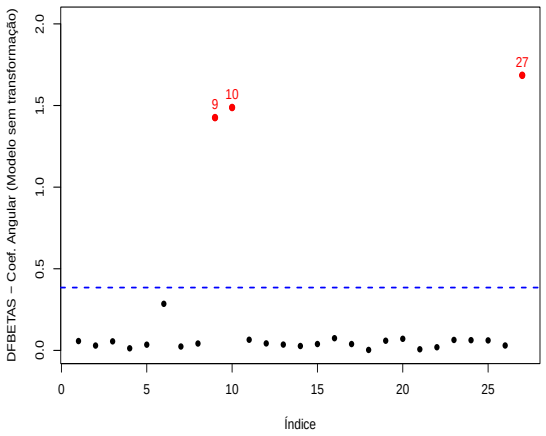


(c) Modelo com Box-Cox

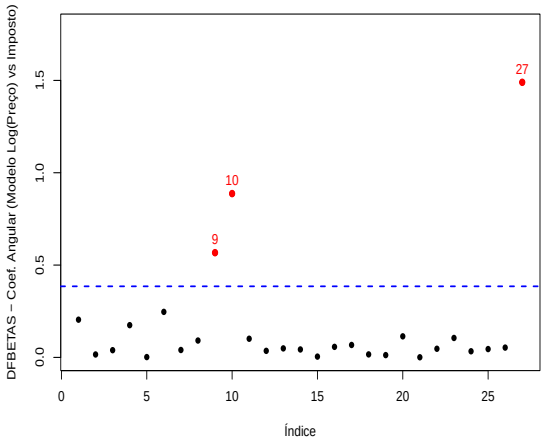
Figura 74: Gráficos da métrica DFBETAS para o intercepto nos três modelos ajustados para a covariável Imposto

Além do intercepto, também é essencial verificar a influência das observações sobre o coeficiente angular, que representa a inclinação da relação entre as variáveis.

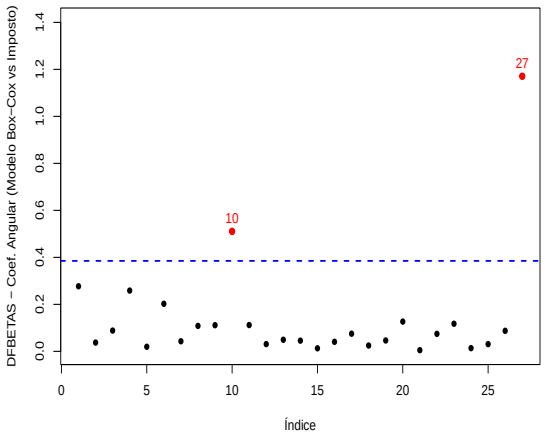
Os gráficos da métrica DFBETAS para o coeficiente angular indicam que, no Modelo Sem transformação e no Modelo log(Preço) as observações 9, 10 e 27 aparecem como possíveis pontos influentes. Já no Modelo Box-Cox, os pontos 10 e 27 aparecem como possíveis influentes.



(a) Modelo Sem Transformação



(b) Modelo com log(Preço)

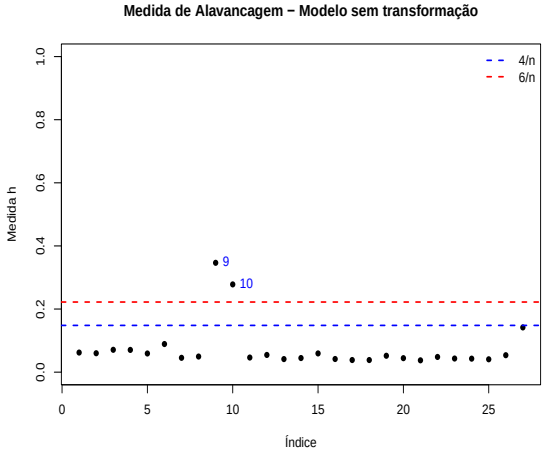


(c) Modelo com Box-Cox

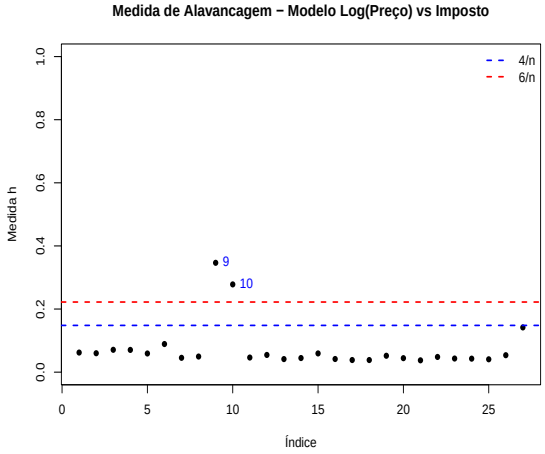
Figura 75: Gráficos da métrica DFBETAS para o coeficiente angular nos três modelos ajustados para a covariável Imposto

Dando continuidade à etapa de diagnóstico, serão analisados os gráficos de alavancagem. Os gráficos de alavancagem mostram a medida h de cada observação nos diferentes modelos ajustados. As linhas de referência 4/n e 6/n indicam limites convencionais para identificar observações com alto impacto na estimativa dos coeficientes.

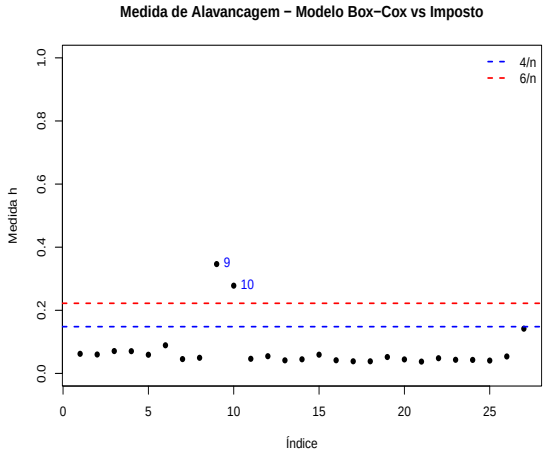
Os três modelos, Modelo sem transformação, Modelo Log(Preço) vs Imposto e Modelo Box-Cox vs Imposto, destacam as observações 9 e 10.



(a) Modelo sem transformação



(b) Modelo Log(Preço) vs Imposto



(c) Modelo Box-Cox vs Imposto

Figura 76: Gráficos de alavancagem para os modelos baseados no Imposto

Finalizando a etapa de diagnóstico, analisaremos o impacto dos pontos influentes nos coeficientes dos modelos. A tabela abaixo apresenta os coeficientes da regressão dos modelos, além da remoção de possíveis pontos influentes identificados nos gráficos de alavancagem (observações 9, 10 e 27).

No Modelo Sem Transformação, a presença dos pontos influentes superestima o impacto do imposto sobre o preço.

No Modelo log(Preço) Os pontos influentes têm menor impacto quando o modelo é transformado para escala logarítmica. A relação entre imposto e preço se mantém forte, com pouca variação no coeficiente β_1 mesmo após a remoção dos pontos influentes, mas a remoção do ponto 27 tem um impacto notável.

No Modelo Box-Cox, assim como no modelo logarítmico, a relação entre imposto

e preço se mantém robusta, mas a remoção do ponto 27 tem um impacto notável, aumentando a inclinação da reta ajustada.

Tabela 27: Impacto dos Pontos Influentes nos Coeficientes da Regressão

Modelo	β_0 (Intercepto)	β_1 (Imposto)	p-valor
$y = preco \sim imposto$	5.583	4.543	2.31×10^{-11}
Sem ponto 9	8.9425	4.0075	5.85×10^{-9}
Sem ponto 10	8.8822	4.0059	1.09×10^{-9}
Sem ponto 27	2.8569	5.0337	1.66×10^{-14}
Sem pontos 9, 10, 27	13.313	3.325	2.02×10^{-8}
$y = \log(preco) \sim imposto$	2.926	0.0935	2.99×10^{-12}
Sem ponto 9	2.9528	0.0892	7.84×10^{-10}
Sem ponto 10	2.9657	0.0870	2.53×10^{-10}
Sem ponto 27	2.8783	0.1021	9.78×10^{-15}
Sem pontos 9, 10, 27	2.9220	0.0949	3.04×10^{-8}
$y = \text{Box-Cox}(preco) \sim imposto$	1.5660	0.0139	1.04×10^{-11}
Sem ponto 9	1.5668	0.0138	1.52×10^{-9}
Sem ponto 10	1.5697	0.0133	8.46×10^{-10}
Sem ponto 27	1.5595	0.0151	2.95×10^{-13}
Sem pontos 9, 10, 27	1.5530	0.0161	4.34×10^{-8}

2.4 Modelo Selecionado: Interpretação e Previsões

Com base nas análises realizadas, optou-se pelo Modelo Sem Transformação da covariável "AreaC" (Preço $\sim$ Área Construída) como a abordagem mais adequada para descrever o problema. A principal justificativa para essa escolha está no fato de que o modelo apresentou um alto poder explicativo, com um valor de $R^2 = 0.8627$, o que indica que ele é capaz de explicar 86,27% da variabilidade do preço dos imóveis.

Além disso, os testes de normalidade, heterocedasticidade e auto-correlação indicaram que o modelo está bem ajustado aos dados, sem sinais de violação das suposições clássicas da regressão linear.

A robustez do modelo também foi confirmada pela análise dos pontos influentes, onde a remoção de pontos não resultou em alterações significativas nos parâmetros estimados. Isso reforça a confiança na estabilidade e na generalização do modelo para outras observações.

Portanto, a escolha pelo modelo sem transformação foi motivada pela sua simplicidade, boa capacidade explicativa e pela ausência de problemas graves nos testes de

diagnóstico. A equação final ajustada para o modelo logarítmico é:

$$\text{Preço} = \beta_0 + \beta_1 \cdot \text{Área Construída} + \epsilon$$

Substituindo os coeficientes ajustados ($\beta_0 = -0.1999$ e $\beta_1 = 0.0407$), a equação estimada fica:

$$\text{Preço} = 2.506 + 23.804 \cdot \text{Área Construída} + \epsilon$$

Utilizando esta equação, foram realizadas predições para preços dos imóveis. A Tabela a seguir apresenta as predições para diferentes tempos de procedimento:

Tabela 28: Predições de preço para novos valores de Área Construída

Área Construída Fixada	**Previsões**
0.998	26.26
1.121	29.19
1.200	31.07
1.232	31.83
1.240	32.02
1.256	32.40
1.376	35.26
1.488	37.93
1.500	38.21
1.652	41.83
1.664	42.12
1.777	44.81
1.831	46.09
3.000	73.92
3.420	83.92

Esses resultados reforçam que o modelo é capaz de oferecer predições consistentes e robustas, com simplicidade e boa explicabilidade. Além disso, a escolha desse modelo facilita a interpretação e aplicação prática dos resultados no contexto analisado.