



**Universidade de Brasília  
Departamento de Estatística**



## **Lista prática 1**

**Ana Luisa Sousa de Oliveira**

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

**Brasília  
2025**

## Sumário

<b>1 Problema 1</b>	4
1.1 Faça uma breve análise descritiva dos dados a partir de gráficos e medidas-resumo. Comente.	4
1.2 Usando transformações, verifique se é possível melhorar a relação entre resposta e covariável. Comente.	6
1.3 Ajuste modelos de regressão linear simples de acordo com as relações observadas nas subseções 1.1 e 1.2. Comente	7
1.4 Faça análises de diagnóstico completas para cada modelo de regressão simples ajustado na subseção 1.3, comentando a adequação ou desvio das suposições necessárias. Verifique se existem observações atípicas.	11
1.5 A partir das conclusões obtidas na subseção 1.4, escolha o melhor modelo de regressão simples ajustado. Para o modelo escolhido, realize as interpretações adequadas. Discuta sobre a explicabilidade do modelo. Faça previsões da dose de radiação recebida fixando novos valores para o tempo total do procedimento.	22
<b>2 Problema 2</b>	23
2.1 Faça uma análise descritiva dos dados a partir de gráficos e medidas-resumo.	23
2.2 Usando transformações, verifique se é possível melhorar as relações observadas entre resposta e covariável. Comente.	26
2.3 Ajuste modelos de regressão linear simples de acordo com as relações observadas nas subseções 2.1 e 2.2. Comente.	31
2.4 Faça análises de diagnóstico completas para cada modelo de regressão simples ajustado na subseção 2.3, comentando a adequação ou desvio das suposições necessárias. Verifique se existem observações atípicas.	34
2.5 A partir das conclusões obtidas na subseção 2.4, escolha o melhor modelo de regressão simples ajustado. Para o modelo escolhido, realize as interpretações adequadas. Discuta sobre a explicabilidade do modelo. Faça previsões do preço de imóveis fixando novos valores para a covariável considerada.	45

# 1 Problema 1

Os conceitos e procedimentos relacionados aos Modelos de Regressão Linear Simples (MRLS) presentes nesta seção foram fundamentados em (RIBEIRO, 2024a) e (RIBEIRO, 2024b).

Os dados *fluoro* da biblioteca **GLMsData** contêm 19 observações e são referentes a um estudo de intervenções de tomografia computadorizada (fluoroscopia) no abdômen. Nesse estudo mediu-se o tempo total do procedimento (em segundos) e a dose total de radiação recebida (em rads). A variável resposta é a dose total de radiação recebida.

## 1.1 Faça uma breve análise descritiva dos dados a partir de gráficos e medidas-resumo. Comente.

A análise descritiva dos dados constitui-se como etapa fundamental para modelagens e estudos robustos posteriores. Dessa forma, deseja-se nesta seção compreender as medidas-resumo de ambas as variáveis, classificadas como quantitativas, dos dados *fluoro*, além de estabelecer uma relação visual entre as duas por meio de um gráfico de dispersão.

Quadro 1: Medidas-resumo do tempo total do procedimento e da dose total de radiação recebida - Dados *fluoro*.

Estatística	Tempo	Dose
Média	75,74	26,86
Desvio padrão	20,50	23,85
Mínimo	37,00	3,46
1º Quartil	59,50	8,00
Mediana	75,00	18,92
3º Quartil	91,00	41,38
Máximo	114,00	84,77



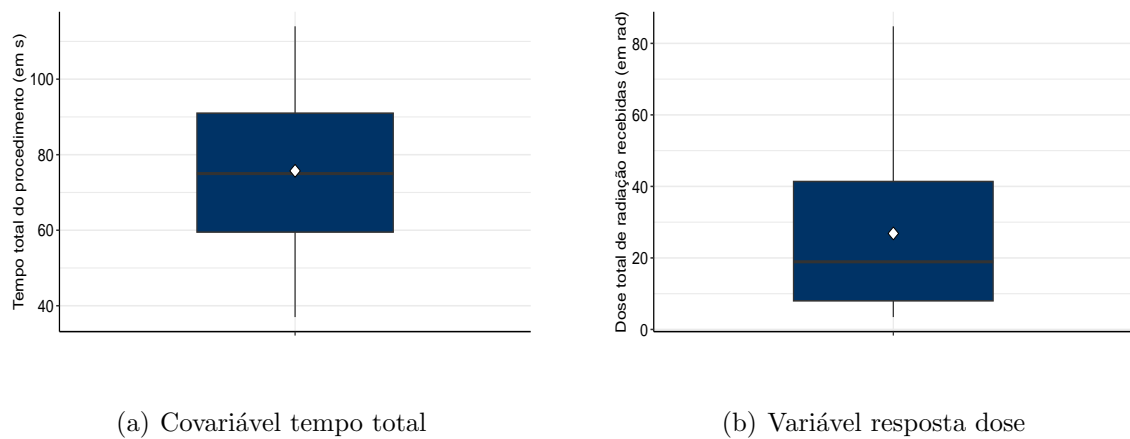


Figura 1: Gráficos box-plot das variáveis - Dados *fluoro*.

Através do Quadro 1, é possível notar que o tempo total médio do procedimento é de 75,74 segundos, sendo 37 e 114 segundos o tempo de duração mínimo e máximo, respectivamente, de uma intervenção de fluoroscopia no abdômen. Além disso, 50% dos procedimentos tiveram duração de ao menos 75 segundos, conforme indica a mediana. É perceptível que tal variável tem distribuição simétrica e não apresenta pontos atípicos, conforme a [Figura 1](#).

Quanto à variável resposta dose total de radiação recebida, emitiu-se em média 26,86 rads, com um desvio padrão de 23,85 rads. Houve uma intervenção que irradiou 84,77 rads, representando a maior quantidade registrada dentre os 19 procedimentos. Ademais, 50% das intervenções apresentaram valores entre 8 e 41,38 rads de dose total de radiação emitida. Pela Figura 1, evidencia-se a assimetria positiva desta variável, isto é, a quantidade de dose total de radiação emitida concentra-se em valores mais baixos.

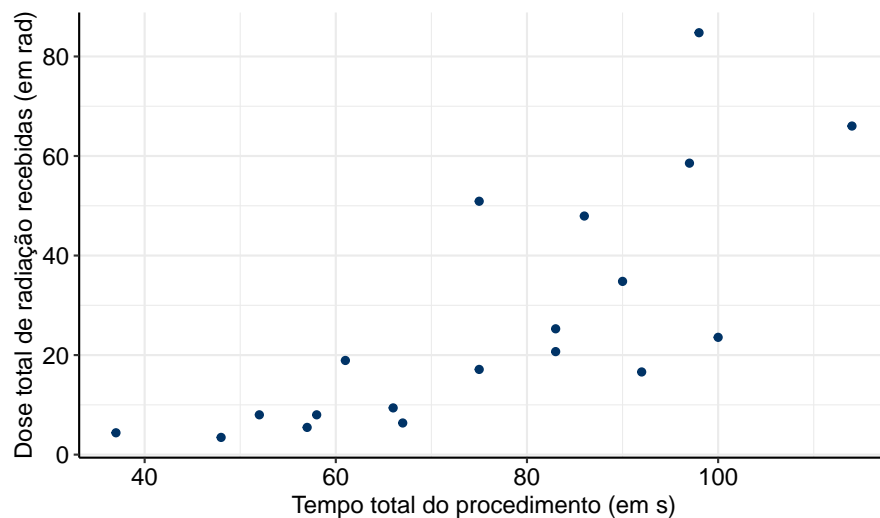
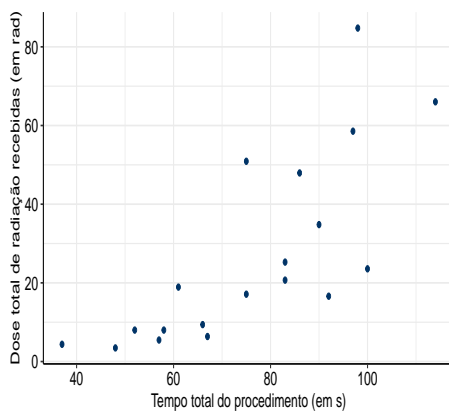


Figura 2: Gráfico de dispersão entre tempo total do procedimento e a dose total de radiação recebida - Dados *fluoro*.

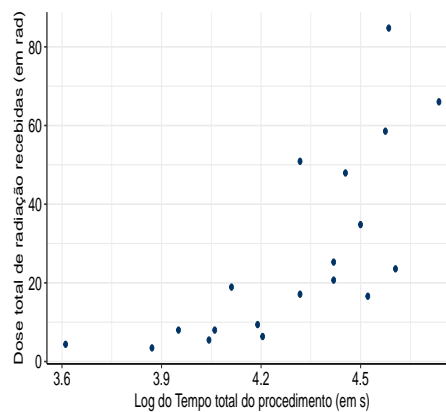
A Figura 2 permite observar a associação entre ambas as variáveis estudadas. Nota-se que provavelmente há uma relação positiva não linear, já que o desenho formado pelos pontos se assemelha a uma curva exponencial. Dessa maneira, será necessário aplicar transformações nas variáveis, a fim de linearizar a relação para que o modelo de regressão linear simples possa ser devidamente utilizado.

## 1.2 Usando transformações, verifique se é possível melhorar a relação entre resposta e covariável. Comente.

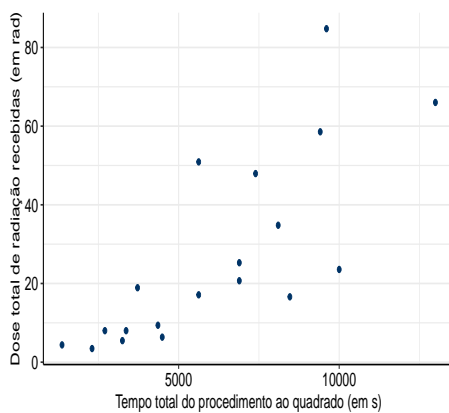
As transformações de variáveis em regressão linear são importantes porque ajudam a melhorar o desempenho do modelo, ajustando-o para que os pressupostos da regressão sejam melhor atendidos, tais como linearidade da relação entre variável resposta e covariável, homoscedasticidade e normalidade dos resíduos. Outrossim, transformações reduzem o impacto de *outliers*, além de poderem melhorar a interpretabilidade do modelo.



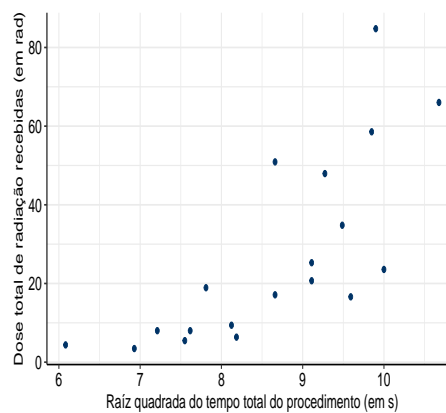
(a) Dose  $\sim$  Tempo



(b) Dose  $\sim \log(\text{Tempo})$



(c) Dose  $\sim \text{Tempo}^2$



(d) Dose  $\sim \sqrt{\text{Tempo}}$

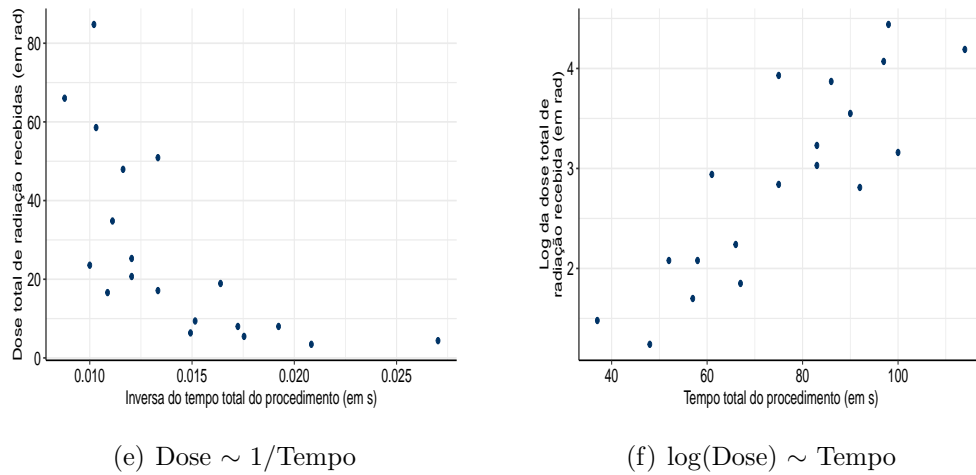


Figura 3: Gráficos de dispersão entre a dose total de radiação recebida e o tempo total do procedimento, considerando diferentes transformações da variável Tempo - Dados *fluoro*.

Foram aplicadas as transformações logarítmica, ao quadrado, raiz quadrada e inversa. A transformação de Box-Cox não foi aplicada dado o valor  $\lambda = -0,1$  encontrado ser muito próximo a zero, indicando que a transformação logarítmica poderia ser uma boa opção. As relações entre a variável resposta e a covariável com suas respectivas transformações podem ser observadas na Figura 3.

A partir desta mesma Figura, é possível observar que as três transformações que mais tornaram linear a relação entre as variáveis foram as logarítmicas e ao quadrado. Embora a transformação ao quadrado pareça ter gerado uma linearização mais evidente em comparação com a transformação logarítmica na covariável, esta última apresenta uma distribuição mais homogênea dos **resíduos**, ou seja, é mais homoscedástica. No entanto, é evidente que a transformação logarítmica na resposta foi a que mais efetivamente linearizou a relação entre resposta e covariável.

### 1.3 Ajuste modelos de regressão linear simples de acordo com as relações observadas nas subseções 1.1 e 1.2. Comente

A reta de regressão linear simples ajustada é definida como

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

em que

- $Y_i$  é a  $i$ -ésima amostra da variável resposta  $Y$ ;
- $\beta_0$  e  $\beta_1$  são os parâmetros que representam o intercepto e coeficiente angular, res-


pectivamente, da reta de regressão;

- $x_i$  é o  $i$ -ésimo valor fixado da covariável  $x$ ;
- $\epsilon_i$  é o  $i$ -ésimo erro aleatório.

Com base nas transformações discutidas na subseção 1.2, os seguintes Modelos de Regressão Linear Simples (MRLS) foram ajustados. A Tabela 1 a seguir apresenta as estimativas dos coeficientes, os erros padrão, a estatística  $t$  e os respectivos valores de  $p$ . Nesses modelos,  $Y$  representa a variável resposta (dose total recebida) e  $x$  corresponde à variável independente (tempo total do procedimento). Ademais, para posterior análise considera-se o nível de significância  $\alpha = 5\%$ .

- Modelo 1:  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$
- Modelo 2:  $Y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i, \quad i = 1, 2, \dots, n$
- Modelo 3:  $Y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i, \quad i = 1, 2, \dots, n$
- Modelo 4:  $\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$

Tabela 1: Estimativas, erros padrão, estatística  $t$  e  $p$ -valor sob os MRLS ajustados - Dados *fluoro*.

	Modelo 1: Dose $\sim$ Tempo				Modelo 2: Dose $\sim \log(\text{Tempo})$			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
 <i>parâmetro</i>								
$\beta_0$	-39,663	14,486	-2,738	< 0,001	-225,64	58,800	-3,837	< 0,001
$\beta_1$	0,878	0,185	4,749	< 0,001	58,870	13,680	4,303	< 0,001

	Modelo 3: Dose $\sim \text{Tempo}^2$				Modelo 4: $\log(\text{Dose}) \sim \text{Tempo}$			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>								
$\beta_0$	-9,253	8,151	-1,135	0,272	-0,199	0,468	-0,427	0,675
$\beta_1$	0,005	0,001	4,940	< 0,001	0,040	0,005	6,798	< 0,001

Tabela 2: Intervalos de confiança sob os MRLS ajustados - Dados *fluoro*.

Parâmetro	Modelo 1: Dose $\sim$ Tempo		Modelo 2: Dose $\sim \log(\text{Tempo})$	
	Limite Inferior	Limite Superior	Limite Inferior	Limite Superior
$\beta_0$	-70,227	-9,100	-349,702	-101,567
$\beta_1$	0,488	1,268	30,006	87,732

Parâmetro	Modelo 3: Dose $\sim \text{Tempo}^2$		Modelo 4: $\log(\text{Dose}) \sim \text{Tempo}$	
	Limite Inferior	Limite Superior	Limite Inferior	Limite Superior
$\beta_0$	-26,450	7,943	-1,188	0,788
$\beta_1$	0,003	0,008	0,028	0,053

O Modelo 1 assume uma relação linear simples entre tempo e dose, com um coeficiente angular positivo e significativo, indicando que um aumento de uma unidade no tempo do procedimento está associado a um acréscimo médio de 0,878 unidades na dose total recebida.

Ademais, o coeficiente angular estimado para  $\log(x_i)$ , presente no Modelo 2, também é significativo, mostrando que essa transformação ainda captura bem a relação, embora sua interpretação seja menos intuitiva.

Já o Modelo 3 introduz um termo quadrático para o tempo, permitindo capturar possíveis padrões curvos na relação entre as variáveis. Apesar de o coeficiente quadrático ser estatisticamente significativo, seu efeito prático pode ser pequeno. Além disso, o intercepto não é estatisticamente significativo, o que pode indicar que ele não desempenha um papel essencial no ajuste do modelo.

Por fim, o Modelo 4 aplica  $\log$  à variável resposta. O  $\beta_1$  obtido continua significativo. Entretanto, assim como no Modelo 3, o intercepto não é estatisticamente significativo. Similar ao Modelo 2, a interpretação do modelo também é menos intuitiva.

Pela Tabela 2 confirmam-se as significâncias dos coeficientes a um nível de confiança de 95%. Para os modelos 3 e 4,  $\beta_0$  não é estatisticamente significativo, haja vista que possuem o valor 0 dentro de seus respectivos intervalos. Já o parâmetro  $\beta_1$  é significativo em todos os modelos apresentados.

A seguir são apresentadas as tabelas ANOVA e testes F para cada um dos modelos indicados. Note que o teste F, através da estatística  $F_0 = \text{QMReg} / \text{QMRes}$ , avalia a



importância da covariável  $x$  para explicar  $Y$ , por meio das hipóteses

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0. \end{cases}$$

Tabela 3: Tabelas ANOVA e testes F sob os MRLS ajustados - Dados *fluoro*.

	Modelo 1: Dose $\sim$ Tempo				Modelo 2: Dose $\sim \log(\text{Tempo})$			
	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F
<i>fonte de variação</i>								
Regressão	1	5.838	5.838	22,55	1	5.338,2	5.338,2	18,517
Resíduo	17	4.401	258		17	4.900,8	288,3	
Total	18	10.239			18	10.239		

	Modelo 3: Dose $\sim \text{Tempo}^2$				Modelo 4: $\log(\text{Dose}) \sim \text{Tempo}$			
	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F
<i>fonte de variação</i>								
Regressão	1	6.035,3	6.035,3	24,407	1	12,518	12,518	46,207
Resíduo	17	4.203,7	247,3		17	4,605	0,270	
Total	18	10.239			18	17,123		

Segundo  $H_0$ , a estatística  $F_0 \sim F_{1,17}$ . A partir dos valores observados, obtêm-se p-valores menores que 0,001 para todos os modelos apresentados. Assim, rejeita-se  $H_0$  a um nível de significância de 5% para todos os modelos, indicando que há evidências estatísticas para afirmar que  $\beta_1 \neq 0$ . Em outras palavras, o teste F indica que todos os modelos apresentados são estatisticamente significativos, ou seja, o tempo total do procedimento (variável explicativa) está ajudando a explicar a variação da dose total de radiação recebida (variável resposta).

**1.4 Faça análises de diagnóstico completas para cada modelo de regressão simples ajustado na subseção 1.3, comentando a adequação ou desvio das suposições necessárias. Verifique se existem observações atípicas.**

As técnicas de diagnóstico para modelos de regressão linear simples (MRLS) são fundamentais para garantir que as suposições do modelo sejam atendidas e para avaliar sua adequação. O diagnóstico adequado ajuda a melhorar a qualidade da modelagem, identificar problemas potenciais e refinar as previsões. Os MRLS assumem

- (1) Linearidade entre as variáveis dependente e independente;
- (2) Normalidade dos erros, ou seja, os erros devem seguir uma distribuição normal;
- (3) Independência dos erros;
- (4) Homoscedasticidade, isto é, a variância dos erros deve ser constante.

A seguir verifica-se a normalidade dos **resíduos**, conforme o pressuposto (2) dos modelos, através do gráfico de probabilidade normal dos resíduos studentizados e teste de hipóteses Shapiro-Wilk, a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

$$\begin{cases} H_0 : \text{Os resíduos seguem distribuição normal} \\ H_1 : \text{Os resíduos seguem outro modelo.} \end{cases}$$

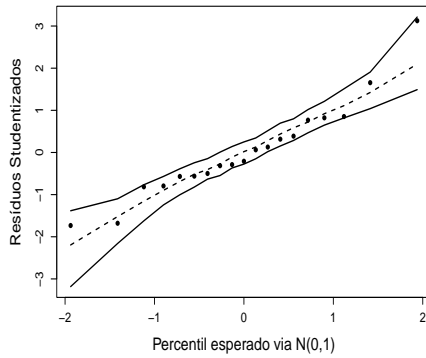
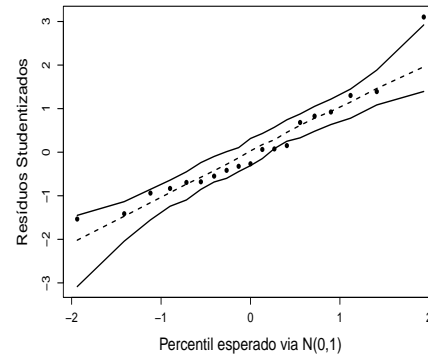
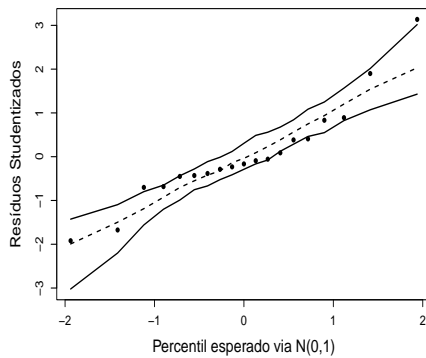
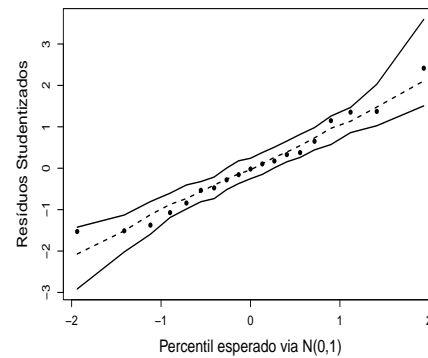
(a) Modelo 1: Dose  $\sim$  Tempo(b) Modelo 2: Dose  $\sim \log(\text{Tempo})$ (c) Modelo 3: Dose  $\sim \text{Tempo}^2$ (d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$ 

Figura 4: Gráficos de probabilidade normal dos resíduos studentizados com envelopes simulados sob os MRLS ajustados - Dados *fluoro*.

Quadro 2:  $p$ -valor do teste de Shapiro-Wilk sob os MRLS ajustados - Dados *fluoro*

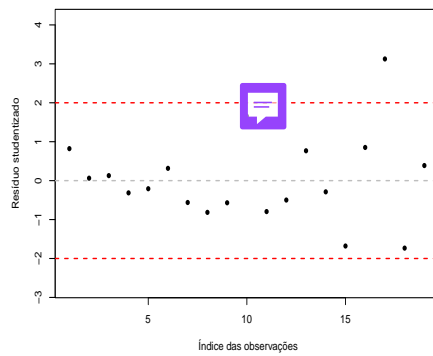
Modelo	$p$ -valor	Decisão do teste
1	0,188	Não rejeita $H_0$
2	0,169	Não rejeita $H_0$
3	0,072	Não rejeita $H_0$
4	0,675	Não rejeita $H_0$

Com base na Figura 4 e no Quadro 2, pode-se afirmar que há evidências estatísticas de que os resíduos de todos os modelos seguem uma distribuição normal. Entretanto, observa-se que nos Modelos 2 e 3 houve observações fora das bandas de confiança.

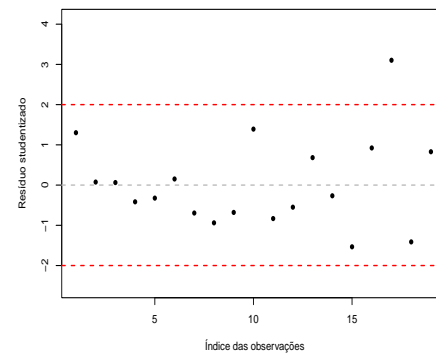
A seguir investiga-se a independência dos **resíduos**, conforme o pressuposto (3), por meio do gráfico resíduos studentizados vs índice e teste de hipóteses Durbin-Watson,

a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

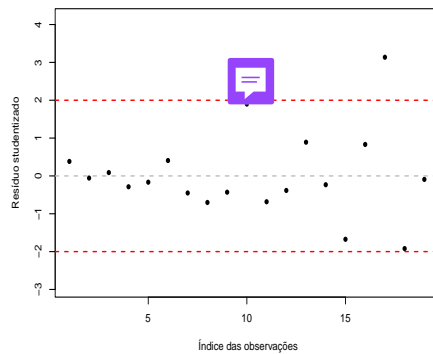
$$\begin{cases} H_0 : \text{Os resíduos são independentes} \\ H_1 : \text{Os resíduos são autocorrelacionados.} \end{cases}$$



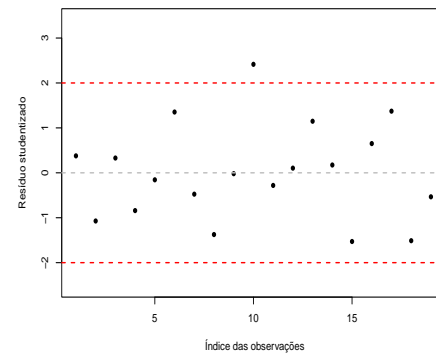
(a) Modelo 1: Dose  $\sim$  Tempo



(b) Modelo 2: Dose  $\sim \log(\text{Tempo})$



(c) Modelo 3: Dose  $\sim \text{Tempo}^2$



(d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$

Figura 5: Gráficos dos resíduos studentizados pelo índice das observações sob os MRLS ajustados - Dados *fluoro*.

Quadro 3:  $p$ -valor do teste de Durbin-Watson sob os MRLS ajustados - Dados *fluoro*

Modelo	$p$ -valor	Decisão do teste
1	0,381	Não rejeita $H_0$
2	0,630	Não rejeita $H_0$
3	0,339	Não rejeita $H_0$
4	0,688	Não rejeita $H_0$

Em todos os modelos, os pontos não **aparentam seguir alguma tendência**, distribuindo-se de forma aleatória pelo gráfico, o que sugere a não correlação dos resíduos.

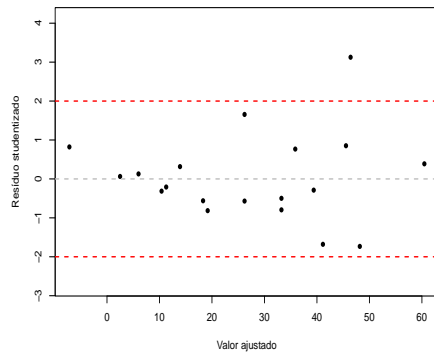
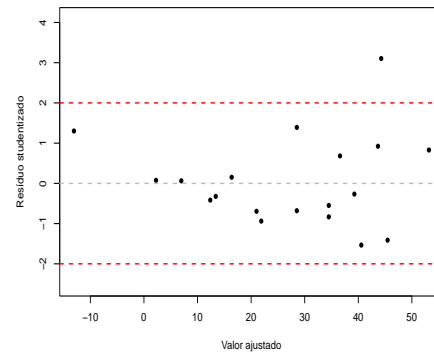
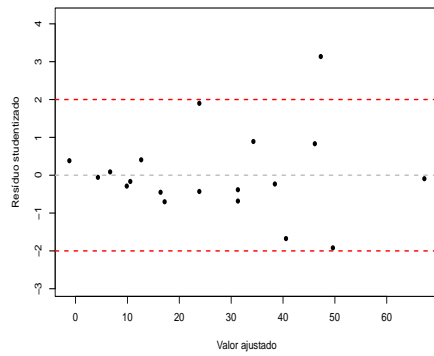
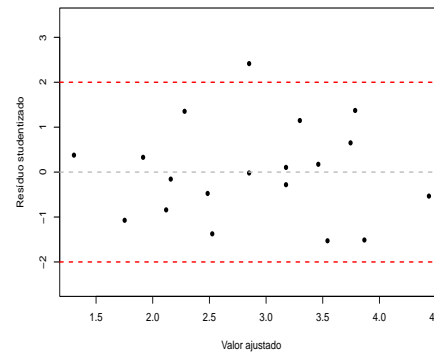
Além disso, pelo Quadro 3, todos os modelos apresentam resíduos independentes. É importante destacar que o teste usado é desenvolvido sob a suposição de que os erros da regressão são normalmente distribuídos, e como isso é verdade para todos os modelos, o teste é mais confiável.

A seguir estuda-se a homoscedasticidade do modelo, conforme o pressuposto (4), por meio do gráfico resíduos studentizados vs valor ajustado e teste de hipóteses Goldfeld-Quandt, a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

$$\begin{cases} H_0 : \text{Não há heteroscedasticidade} \\ H_1 : \text{Há heteroscedasticidade.} \end{cases}$$

Para tornar o teste mais poderoso, omitiu-se um grupo central de  $r = n/3$  observações, resultando em  $r \approx 6,33$ .



(a) Modelo 1: Dose  $\sim$  Tempo(b) Modelo 2: Dose  $\sim \log(\text{Tempo})$ (c) Modelo 3: Dose  $\sim \text{Tempo}^2$ (d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$ Figura 6: Gráficos dos resíduos studentizados pelo valor ajustado sob os MRLS ajustados - Dados *fluoro*.Quadro 4:  $p$ -valor do teste de Goldfeld-Quandt sob os MRLS ajustados - Dados *fluoro*

Modelo	$p$ -valor	Decisão do teste
1	0,003	Rejeita $H_0$
2	0,004	Rejeita $H_0$
3	0,003	Rejeita $H_0$
4	0,339	Não rejeita $H_0$

Na Figura 6, observa-se que, nos Modelos 1, 2 e 3, a variância é menor para os baixos valores ajustados, quando comparada aos valores mais altos. No entanto, no Modelo 4, a variância dos resíduos é constante para todos os valores ajustados. Conforme mostrado no Quadro 4, apenas o Modelo 4 apresenta homocedasticidade.

A seguir avaliam-se as observações atípicas dos MRLS ajustados, que são classificadas em

- (a) Pontos aberrantes: são aqueles que apresentam ajustes inadequados para a variável

resposta ( $Y$ ), resultando em resíduos discrepantes. Esses pontos geralmente distorcem o intercepto do modelo de regressão ajustado. Normalmente, considera-se como pontos aberrantes aqueles cujos resíduos studentizados absolutos são maiores que 3, ou seja,  $|t_i^*| > 3$ .



- (b) Pontos alavanca: são aqueles que têm um peso desproporcional no valor ajustado de  $\hat{Y}$ . Esses pontos geralmente apresentam um perfil distinto em relação aos outros, especialmente no que diz respeito ao valor da variável explicativa  $x$ . Além disso, o ponto de alavanca "puxa" a reta de regressão ajustada em sua direção, pois geralmente afeta o coeficiente angular estimado da reta.
- (c) Pontos influentes: são aqueles que têm uma influência desproporcional nas estimativas dos parâmetros do modelo e podem alterar as conclusões inferenciais. Estas observações podem causar mudanças relativas consideráveis nas estimativas  $\beta_0$ ,  $\beta_1$  e  $\sigma^2$ , além poderem mudar o sinal da estimativa e a significância dos coeficientes  $\beta_0$  e  $\beta_1$ .

A partir das Figura 5 e 6 é possível observar que todos os modelos possuem um ponto aberrante, com exceção do Modelo 4, já que  $|t_i^*| > 3$ .

Com relação aos pontos alavanca, define-se a medida de alavancagem da  $i$ -ésima observação como

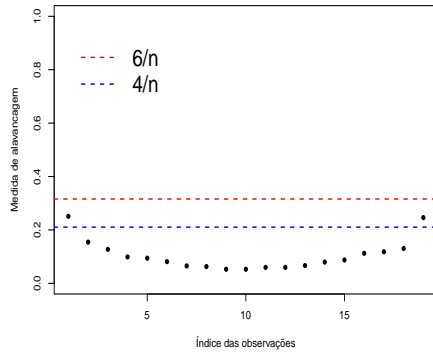
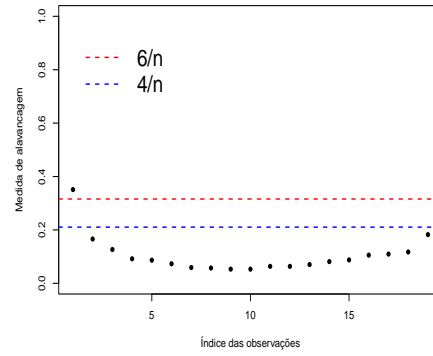
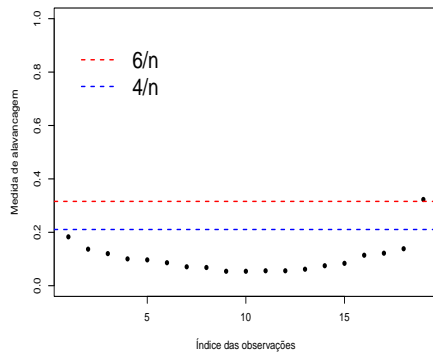
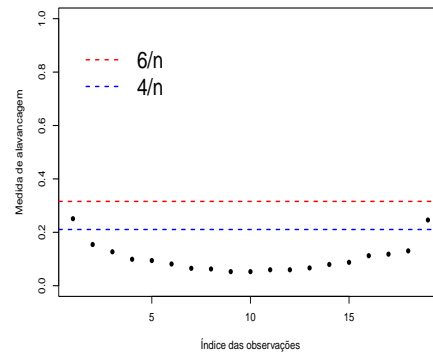
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}},$$

em que



- $n$  é o número de observações da amostra;
- $x_i$  é o  $i$ -ésimo valor fixado da covariável  $x$ ;
- $\bar{x}$  é a média da covariável  $x$ ;
- $S_{xx} = \sum_{i=1}^n x_i(x_i - \bar{x})$ .

Como a média amostral de  $h_{ii}$  é  $2/n$ , pode-se investigar as observações que sejam maiores que o dobro ou que o triplo da média amostral, isto é,  $4/n$  e  $6/n$  respectivamente.

(a) Modelo 1: Dose  $\sim$  Tempo(b) Modelo 2: Dose  $\sim \log(\text{Tempo})$ (c) Modelo 3: Dose  $\sim \text{Tempo}^2$ (d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$ Figura 7: Gráficos da medida de alavancagem sob os MRLS ajustados - Dados *fluoro*.

Com base na Figura 7, todos os modelos apresentaram ao menos 1 ponto de alavancagem, sendo este a primeira ou a última observação. Apenas os Modelos 2 e 3 possuem uma observação com  $h_{ii} > 6/n$ .

Com relação aos pontos de influência, realiza-se a identificação desses pontos influentes no ajuste do modelo de regressão linear por meio de três medidas de influência: DFFITS, DFBETAS e Distância de Cook.

A medida DFFITS (Difference in Fits) pode ser vista como a variação padronizada no  $i$ -ésimo valor ajustado de  $Y$  quando o  $i$ -ésimo ponto é excluído. Tal medida para a  $i$ -ésima observação é definida por

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_i^{(i)}}{\sqrt{\sigma_{(i)}^2 h_{ii}}},$$




em que  $\sigma_{(i)}^2$  é a variância da resposta  $\sigma^2$  estimada excluindo a  $i$ -ésima observação. Para identificar pontos de influência a partir desta medida, basta avaliar as observações tais que

$$|DFFITs_i| > 2\sqrt{\frac{2}{n}}.$$

Para avaliar o impacto da  $i$ -ésima observação nas estimativas dos coeficientes  $\beta_0$  e  $\beta_1$ , pode-se utilizar a medida DFBETAS, definida por

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_i^{(i)}}{\sqrt{\text{Var}(\hat{\beta}_j^{(i)})}}$$

em que  $\text{Var}(\hat{\beta}_j^{(i)})$  é a variância de  $\hat{\beta}_j$  avaliada na estimativa de  $\sigma^2$  obtida excluindo a  $i$ -ésima observação. Uma regra geral é investigar os pontos tais que 

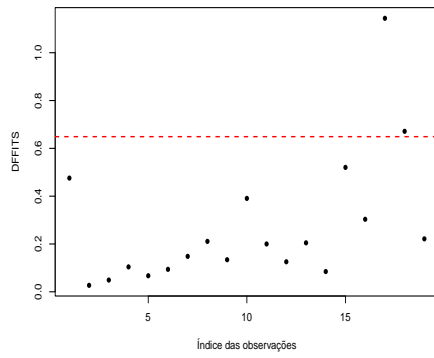
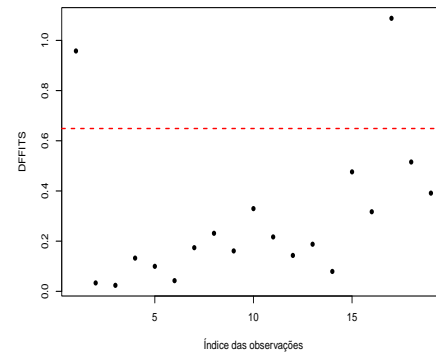
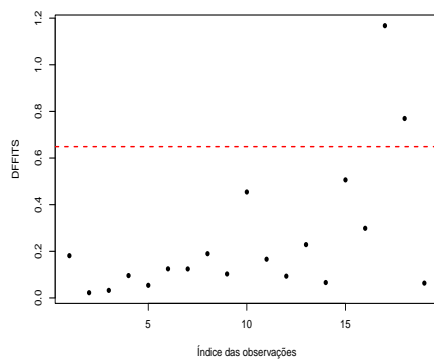
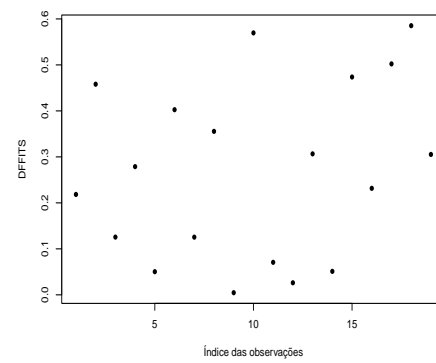
$$|DFBETAS_{ji}| > \frac{2}{\sqrt{n}}.$$

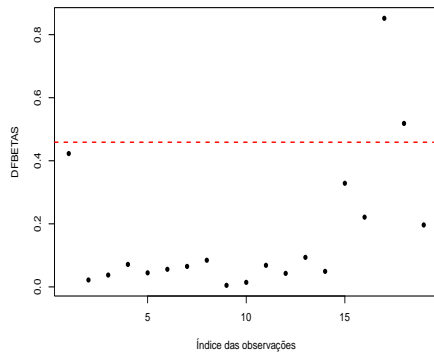
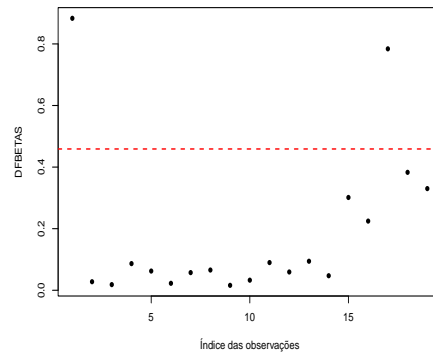
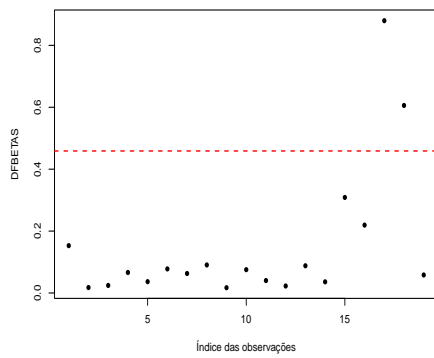
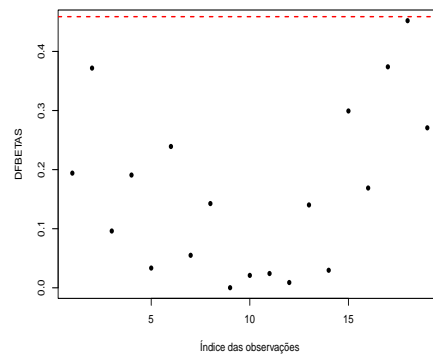
Por fim, a medida da Distância de Cook para a  $i$ -ésima observação é definida por

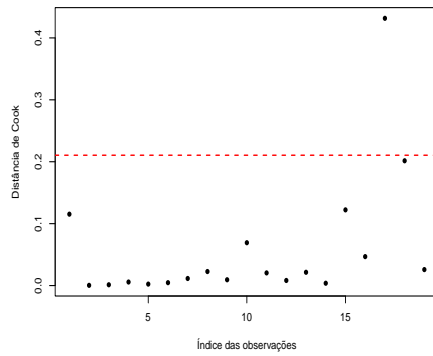
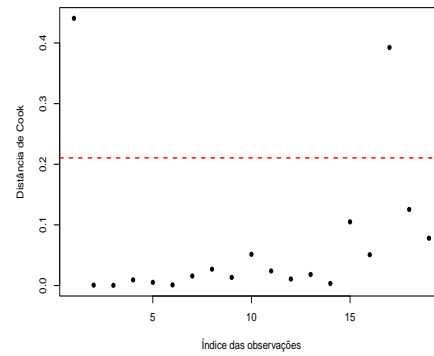
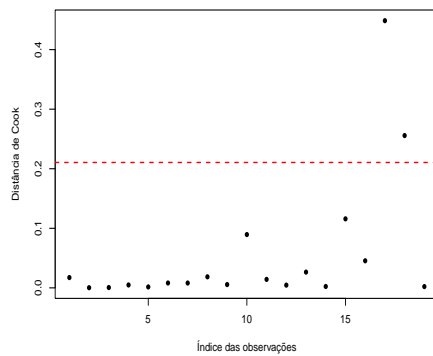
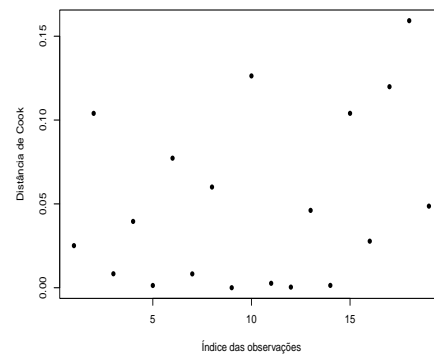
$$D_i = \frac{1}{2\hat{\sigma}^2} \sum_{j=1}^p (\hat{Y}_j - Y_j^{(i)})^2.$$

Para  $n$  grande, uma regra é identificar o ponto como possível influente se

$$D_i > 4/n.$$

(a) Modelo 1: Dose  $\sim$  Tempo(b) Modelo 2: Dose  $\sim \log(\text{Tempo})$ (c) Modelo 3: Dose  $\sim \text{Tempo}^2$ (d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$ Figura 8: Gráficos da medida DFFITS sob os MRLS ajustados - Dados *fluoro*.

(a) Modelo 1:  $\text{Dose} \sim \text{Tempo}$ (b) Modelo 2:  $\text{Dose} \sim \log(\text{Tempo})$ (c) Modelo 3:  $\text{Dose} \sim \text{Tempo}^2$ (d) Modelo 4:  $\log(\text{Dose}) \sim \text{Tempo}$ Figura 9: Gráficos da medida DFBETAS sob os MRLS ajustados - Dados *fluoro*.

(a) Modelo 1:  $Dose \sim Tempo$ (b) Modelo 2:  $Dose \sim \log(Tempo)$ (c) Modelo 3:  $Dose \sim Tempo^2$ (d) Modelo 4:  $\log(Dose) \sim Tempo$ Figura 10: Gráficos da distância de Cook sob os MRLS ajustados - Dados *fluoro*.

Com base nas Figuras 8, 9 e 10, observa-se a presença de pontos influentes segundo as três medidas analisadas. Entre os quatro modelos, apenas o Modelo 4 não apresentou esses pontos. Além disso, todas as medidas indicaram que os demais modelos possuem dois pontos influentes cada.

A seguir são avaliados os modelos propostos retirando-se os pontos atípicos, isto é, as observações aberrantes, de alavanca e influentes.

Tabela 4: Estimativas, erros padrão, estatística  $t$  e  $p$ -valor sob os MRLS ajustados retiradas as observações atípicas - Dados *fluoro*.

	Modelo 1: Dose $\sim$ Tempo				Modelo 2: Dose $\sim \log(\text{Tempo})$			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>								
$\beta_0$	-43,938	13,557	-3,241	0,005	-234,110	59,010	-3.967	0,001
$\beta_1$	0,914	0,175	5,211	< 0,001	60,040	13,670	4,393	< 0,001

	Modelo 3: Dose $\sim \text{Tempo}^2$				Modelo 4: $\log(\text{Dose}) \sim \text{Tempo}$			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>								
$\beta_0$	-32,052	13,428	-2,387	0,031	-0,391	0,541	-0,723	0,481
$\beta_1$	0,753	0,185	4,062	0,001	0,043	0,007	5,844	< 0,001

Dada a Tabela 4, observa-se que os sinais dos parâmetros se mantiveram para todos os modelos. Além disso, todas as estimativas de  $\beta_0$  diminuíram e de  $\beta_1$  aumentaram com menores erros padrão. Quanto à significância dos parâmetros, apenas o Modelo 3 teve mudança -  $\beta_0$  passou a ser significativo.

**1.5 A partir das conclusões obtidas na subseção 1.4, escolha o melhor modelo de regressão simples ajustado. Para o modelo escolhido, realize as interpretações adequadas. Discuta sobre a explicabilidade do modelo. Faça previsões da dose de radiação recebida fixando novos valores para o tempo total do procedimento.**

A partir da análise dos pressupostos, apenas o Modelo 4 atendeu às suposições necessárias - normalidade, independência e homoscedasticidade dos erros. Além disso, o Modelo tem bom ajuste com  $R^2 = 0,731$ .

Dessa forma, a reta ajustada pelo modelo escolhido é definida como

$$\log(Y_i) = -0,199 + 0,04x_i.$$

Teoricamente, o intercepto  $\beta_0$  representa o valor esperado de  $\log(Y_i)$  quando  $x_i = 0$ . Contudo, como a covariável  $x$  assume apenas valores maiores que 0, essa interpretação

não é válida. Já  $\beta_1$  indica a taxa de variação de  $\log(Y_i)$  para cada aumento unitário em  $x$ .

Fixando-se um valor arbitrário para  $x$  no intervalo entre o valor mínimo (37s) e máximo (114s) observado na amostra que ajustou o modelo, pode-se prever o valor de  $Y$ . Sendo  $x_i = 57$ , por exemplo, tem-se que  $Y_i = 8,01$ . Em outras palavras, pela previsão do modelo, um procedimento de 57 segundos de duração total emitiria 8,01 rads de dose total.

## 2 Problema 2

Os conceitos e procedimentos relacionados aos Modelos de Regressão Linear Simples (MRLS) presentes nesta seção foram fundamentados em (RIBEIRO, 2024a) e (RIBEIRO, 2024b).

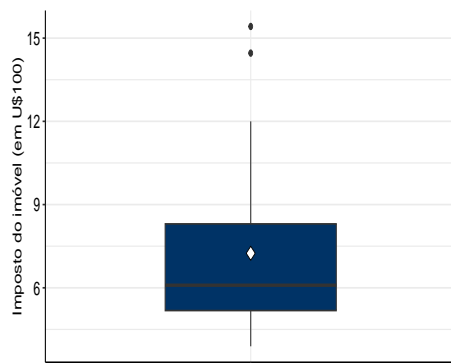
Os dados no arquivo *imoveis.txt* são informações referentes a 27 imóveis. Há interesse em explicar o preço de venda (em US\$ 100) destes imóveis. As variáveis que podem ser utilizadas para explicar o preço dos imóveis são: imposto do imóvel (em US\$ 100), área do terreno (em 1000 pés quadrados), área construída (em 1000 pés quadrados), e idade da residência (em anos).

### 2.1 Faça uma análise descritiva dos dados a partir de gráficos e medidas-resumo.

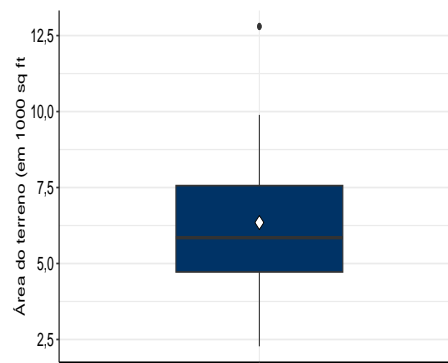
A análise descritiva dos dados constitui-se como etapa fundamental para modelagens e estudos robustos posteriores. Dessa forma, deseja-se nesta seção compreender as medidas-resumo das cinco variáveis, classificadas como quantitativas, dos dados *imoveis*, além de estabelecer uma relação visual entre a variável resposta preço de venda com cada uma das demais covariáveis por meio de gráficos de dispersão.

Quadro 5: Medidas-resumo do imposto do imóvel, área do terreno, área construída, idade da residência e preço de venda do imóvel - Dados *imoveis*

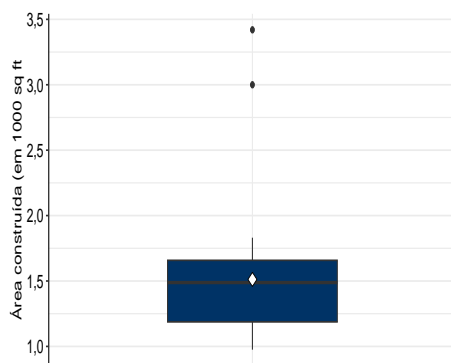
Estatística	Imposto	Área do Terreno	Área Contruída	Idade	Preço de Venda
Média	7,24	6,35	1,51	36,48	38,50
Desvio padrão	2,88	2,40	0,56	14,05	14,31
Mínimo	3,89	2,28	0,98	3,00	25,90
1º Quartil	5,18	4,72	1,19	30,00	29,95
Mediana	6,09	5,85	1,49	40,00	36,90
3º Quartil	8,30	7,56	1,66	47,00	40,75
Máximo	15,42	12,80	3,42	62,00	84,90



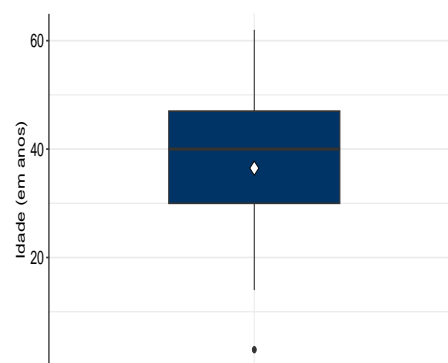
(a) Covariável imposto



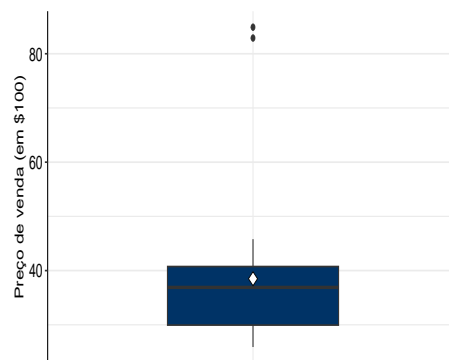
(b) Covariável área do terreno



(c) Covariável área construída



(d) Covariável idade

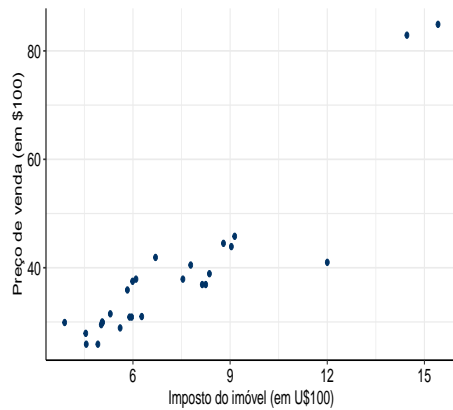


(e) Variável resposta preço de venda

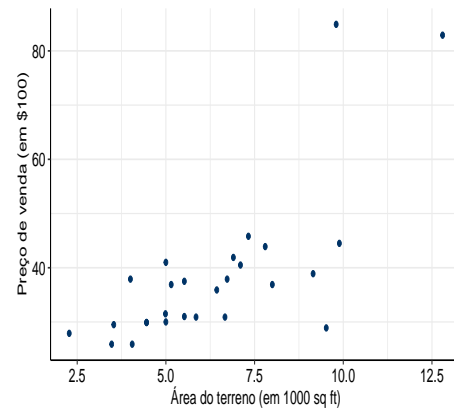
Figura 11: Gráficos box-plot das variáveis - Dados *imoveis*

O Quadro 6 apresenta medidas-resumo de todas as variáveis do conjunto de dados. Quanto à variável resposta preço de venda, tem-se que sua média é R\$3.850. Além disso, os preços mais baixo e alto registrados entre estes imóveis é de R\$2.590 e R\$8.490 respectivamente e 25% das propriedades custam acima de R\$4.075. Pela Figura 11, identificam-se dois imóveis *outliers* com preços de venda muito maiores do que os demais.

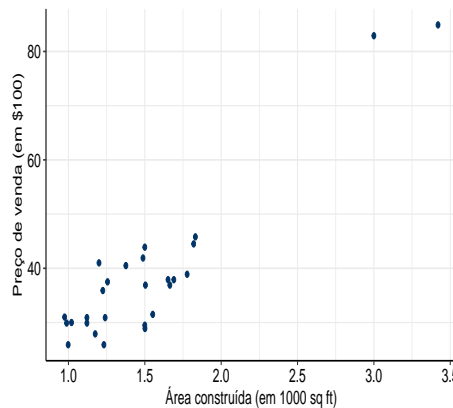
Outrossim, todas as covariáveis apresentam pontos atípicos e algum nível de assimetria. Os imóveis *outliers* com preços significativamente superiores são aqueles que possuem maiores áreas construídas e impostos mais elevados. Todas as propriedades à venda possuem, em média, uma idade avançada desde sua construção (36,48 anos), além de 50% dos imóveis terem ao menos 25,47% de seus terrenos com área construída.



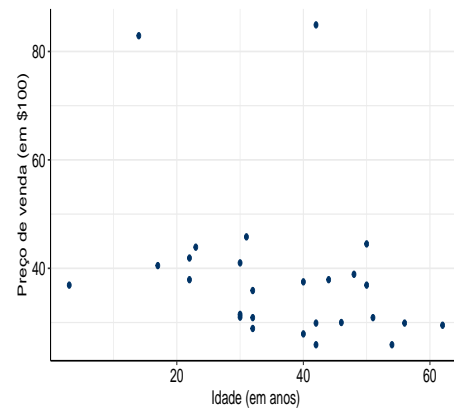
(a) Relação preço de venda e imposto do imóvel



(b) Relação preço de venda e área do terreno



(c) Relação preço de venda e área construída



(d) Relação preço de venda e idade do imóvel

Figura 12: Gráficos de dispersão da variável resposta preço de venda com as demais covariáveis - Dados *imoveis*

A Figura 12 ilustra visualmente a relação entre a variável resposta e as demais covariáveis. Em todos os casos observam-se dois pontos aberrantes que devem ser tratados para evitar distorções no modelo de regressão.

Com respeito às covariáveis imposto e área construída, ambas aparentam ter uma relação linear positiva com o preço de venda dos imóveis. Contudo, as variáveis explicativas área do terreno e idade do imóvel aparentam ter uma relação exponencial positiva e negativa respectivamente com a variável dependente. Dessa forma será necessário apli-

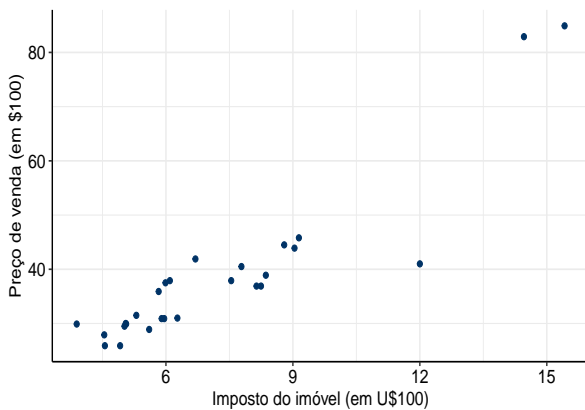


car transformações nas variáveis, a fim de linearizar relações e tratar *outliers* para que o modelo de regressão linear simples possa ser devidamente utilizado.

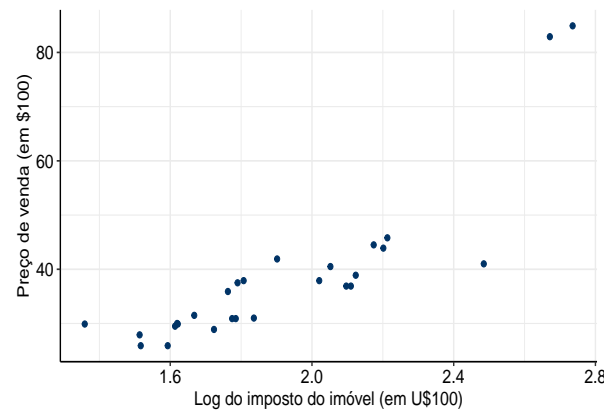
## 2.2 Usando transformações, verifique se é possível melhorar as relações observadas entre resposta e covariável. Comente.

As transformações de variáveis em regressão linear são importantes porque ajudam a melhorar o desempenho do modelo, ajustando-o para que os pressupostos da regressão sejam melhor atendidos, tais como linearidade da relação entre variável resposta e covariável, homoscedasticidade e normalidade dos resíduos. Outrossim, transformações reduzem o impacto de *outliers*, além de poderem melhorar a interpretabilidade do modelo.

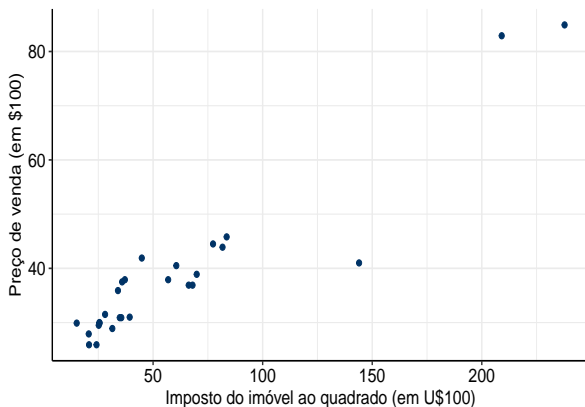
A seguir foram aplicadas as transformações logarítmica, ao quadrado, raiz quadrada, inversa e Box-Cox (indicada por \*) a todas as covariáveis separadamente.



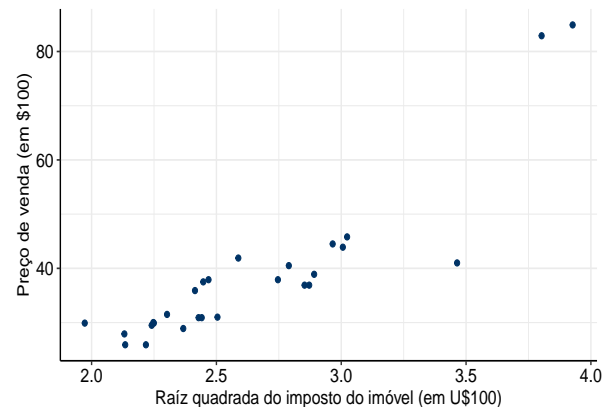
(a) Preço  $\sim$  Imposto



(b) Preço  $\sim \log(\text{Imposto})$



(c) Preço  $\sim \text{Imposto}^2$



(d) Preço  $\sim \sqrt{\text{Imposto}}$

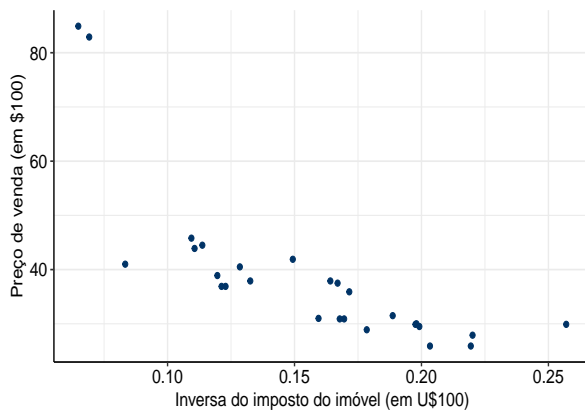
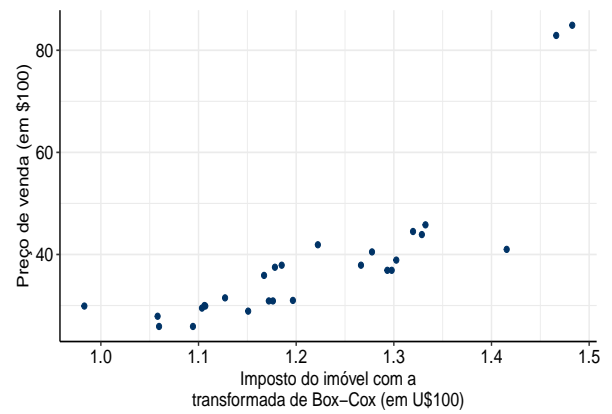
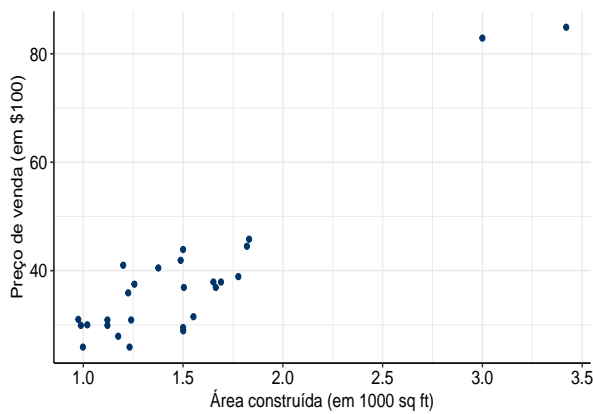
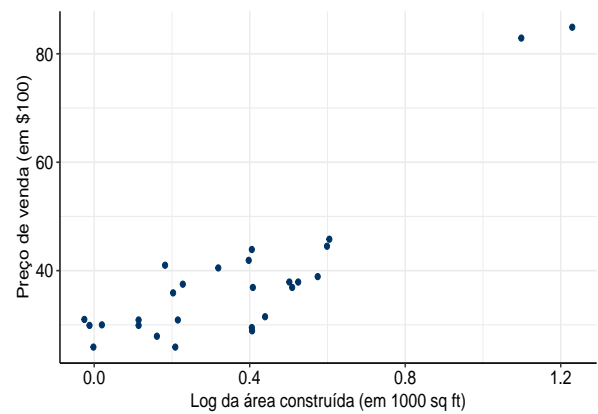
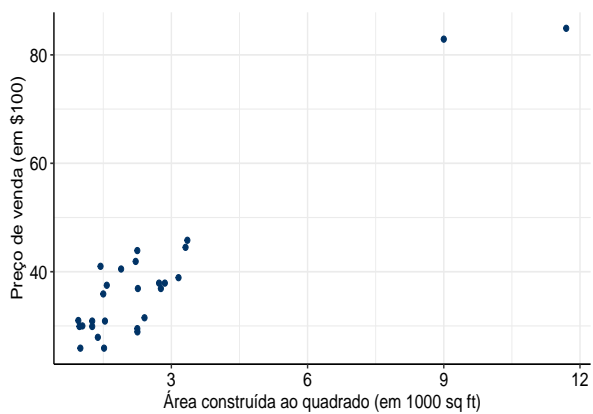
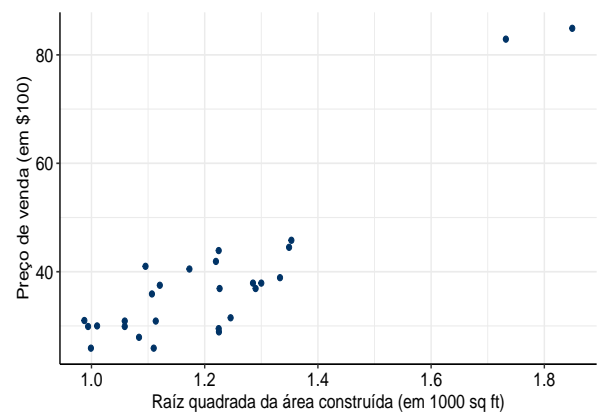
(e)  $\text{Preço} \sim 1/\text{Imposto}$ (f)  $\text{Preço} \sim \text{Imposto}^*$ 

Figura 13: Gráficos de dispersão entre o preço de venda e o imposto incidente no imóvel com transformações aplicadas - Dados *imoveis*

(a)  $\text{Preço} \sim \text{Área construída}$ (b)  $\text{Preço} \sim \log(\text{Área construída})$ (c)  $\text{Preço} \sim \text{Área construída}^2$ (d)  $\text{Preço} \sim \sqrt{\text{Área construída}}$

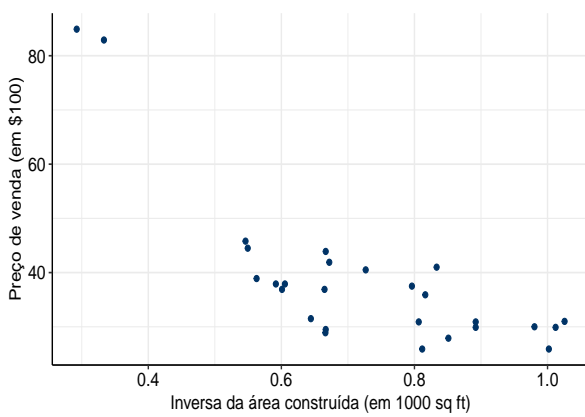
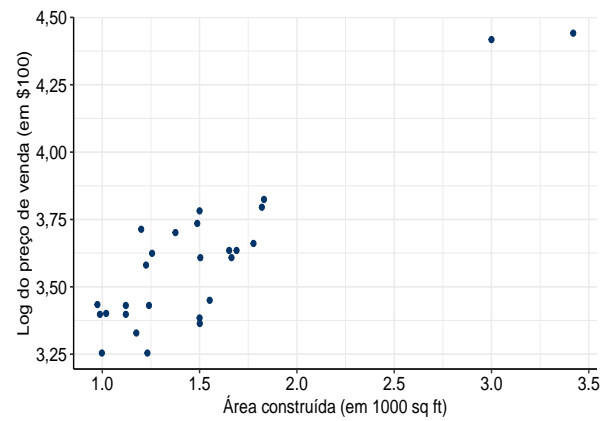
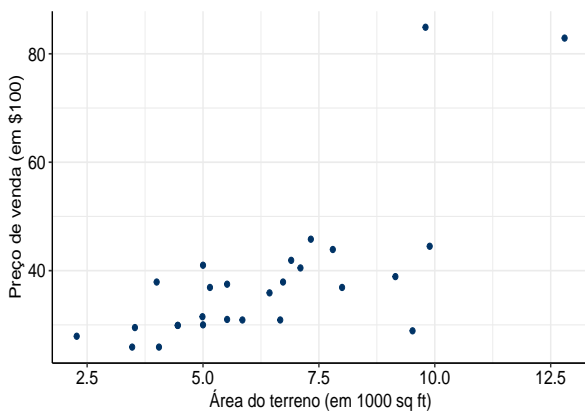
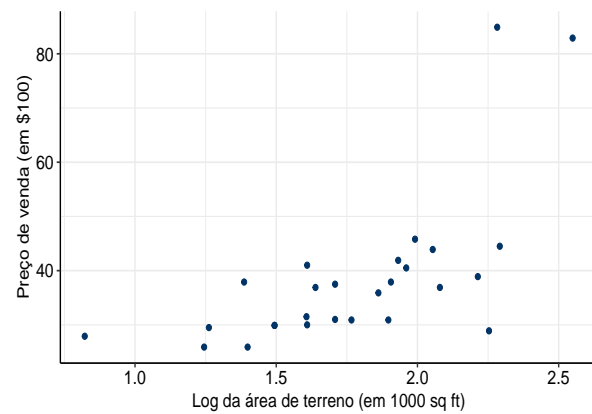
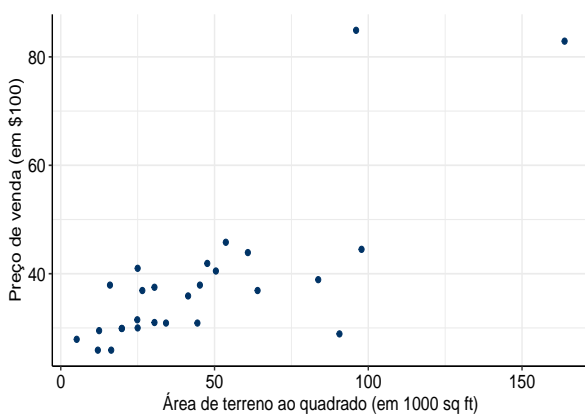
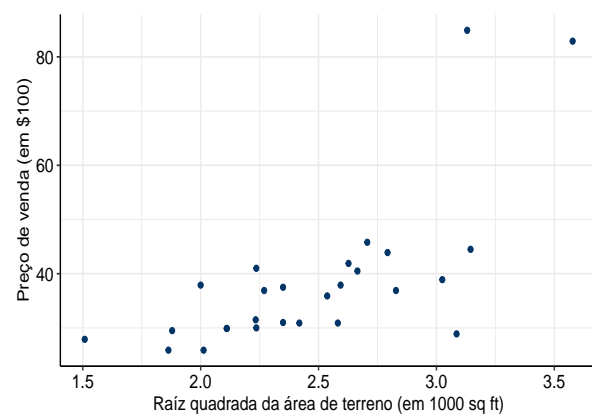
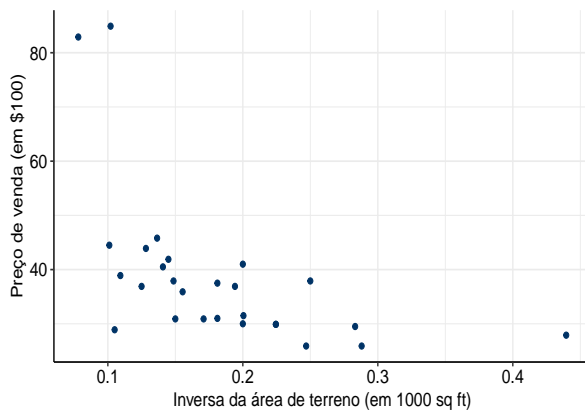
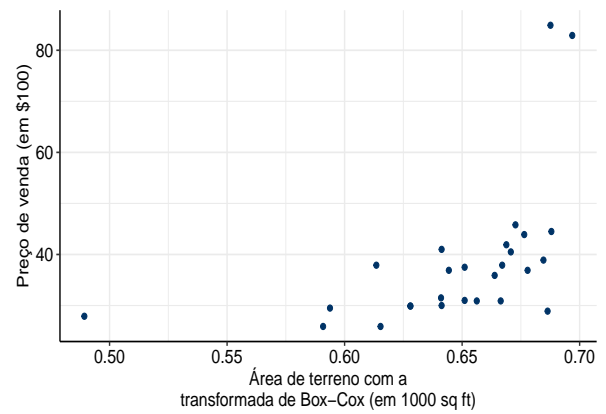
(e)  $\text{Preço} \sim 1/\text{Área construída}$ (f)  $\log(\text{Preço}) \sim \text{Área construída}$ 

Figura 14: Gráficos de dispersão entre o preço de venda e a área construída com transformações aplicadas - Dados *imoveis*

(a)  $\text{Preço} \sim \text{Área do terreno}$ (b)  $\text{Preço} \sim \log(\text{Área do terreno})$ (c)  $\text{Preço} \sim \text{Área do terreno}^2$ (d)  $\text{Preço} \sim \sqrt{\text{Área do terreno}}$

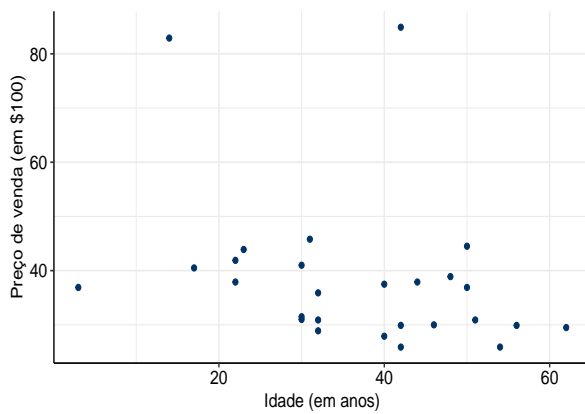


(e)  $\text{Preço} \sim 1/\text{Área do terreno}$

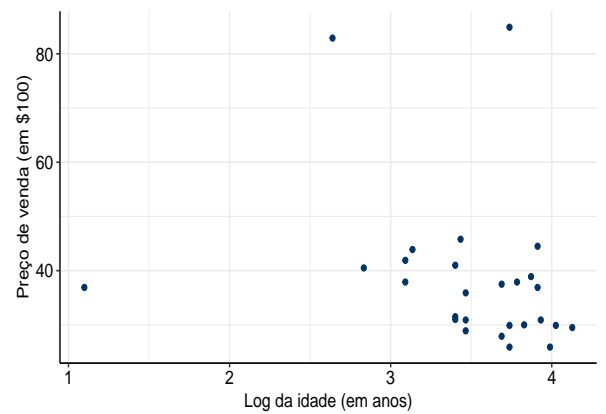


(f)  $\text{Preço} \sim \text{Área do terreno}^*$

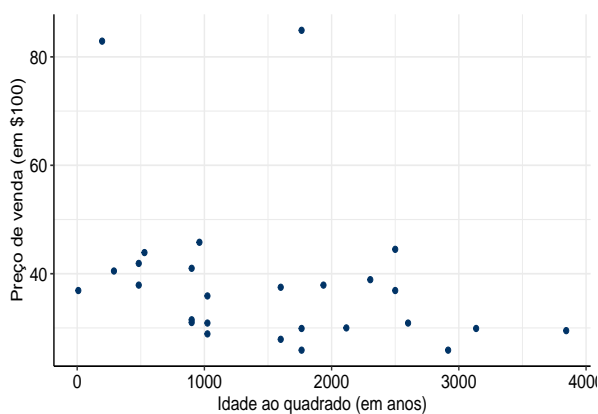
Figura 15: Gráficos de dispersão da variável resposta preço de venda e a área do terreno com transformações aplicadas - Dados *imoveis*



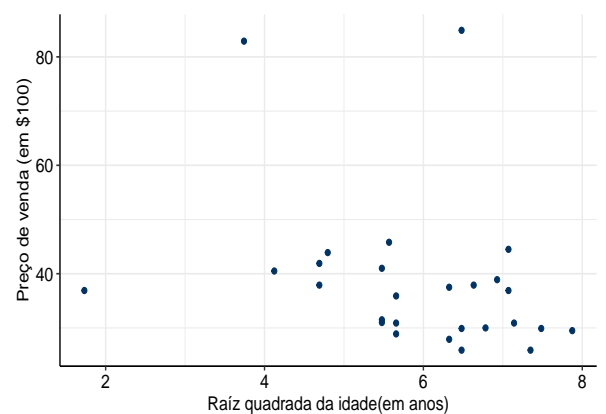
(a)  $\text{Preço} \sim \text{Idade}$



(b)  $\text{Preço} \sim \log(\text{Idade})$



(c)  $\text{Preço} \sim \text{Idade}^2$



(d)  $\text{Preço} \sim \sqrt{\text{Idade}}$

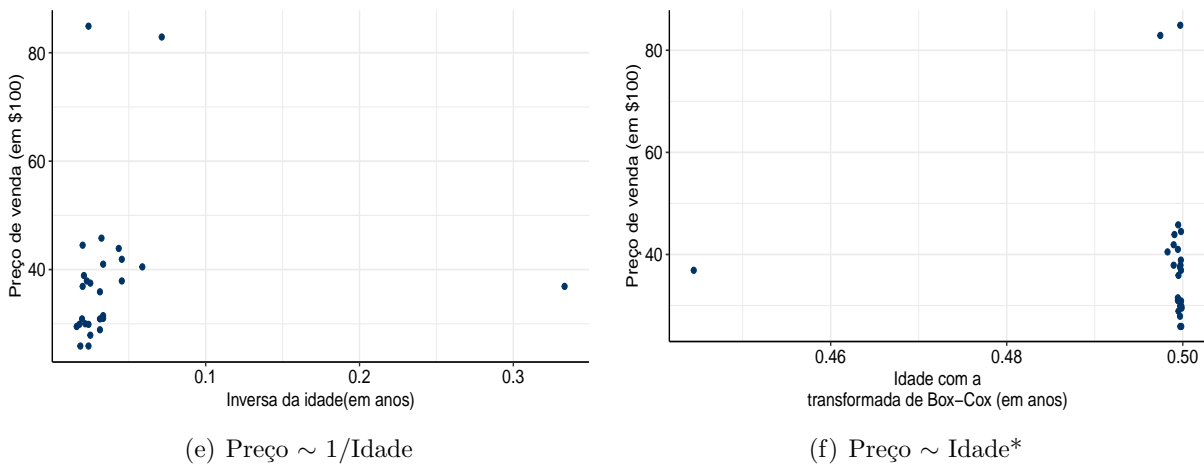


Figura 16: Gráficos de dispersão da variável resposta preço de venda e a idade do imóvel com transformações aplicadas - Dados *imoveis*

A partir das Figuras 13, 14, 15 e 16, é possível observar como se estabelecem as relações entre a variável resposta e as covariáveis transformadas separadamente. Naturalmente, cada transformação possui características próprias que podem impactar o atendimento dos pressupostos da regressão linear simples. Por exemplo, algumas transformações tornam o conjunto de pontos mais homoscedástico, mas não são tão eficazes na linearização dos dados, enquanto outras melhoram a linearidade, mas não garantem a mesma homocedasticidade.

A transformação de Box-Cox não foi aplicada à variável resposta no modelo  $\text{Preço} \sim \text{Área}$  construída, haja vista o valor de  $\lambda = 0,06$  ser próximo a zero, indicando que tal transformação teria um desempenho similar ao modelo  $\log(\text{Preço}) \sim \text{Área}$  construída, entretanto com uma pior interpretabilidade. Para as demais covariáveis imposto, área do terreno e idade, no entanto, se manteve a transformação de Box-Cox com os valores de  $\lambda$  iguais a -0,5, -1,39 e -2 respectivamente.

Apesar das transformações aplicadas para linearizar as relações entre resposta e covariáveis, não houve mudanças relevantes. Portanto, opta-se pelos modelos mais simples, devido à sua maior interpretabilidade. As variáveis imposto, área construída e área do terreno apresentam uma relação linear com o preço de venda. No entanto, a covariável idade não demonstra correlação aparente com a variável resposta, já que não há tendência de crescimento ou decrescimento entre as variáveis.

### 2.3 Ajuste modelos de regressão linear simples de acordo com as relações observadas nas subseções 2.1 e 2.2. Comente.

A reta de regressão linear simples ajustada é definida como

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

em que



- $Y_i$  é a  $i$ -ésima amostra da variável resposta  $Y$ ;
- $\beta_0$  e  $\beta_1$  são os parâmetros que representam o intercepto e coeficiente angular, respectivamente, da reta de regressão;
- $x_i$  é o  $i$ -ésimo valor fixado da covariável  $x$ ;
- $\epsilon_i$  é o  $i$ -ésimo erro aleatório.

Com base nas transformações discutidas na subseção 2.2, os seguintes Modelos de Regressão Linear Simples (MRLS) foram ajustados. A Tabela 8 a seguir apresenta as estimativas dos coeficientes, os erros padrão, a estatística  $t$  e os respectivos valores de  $p$ . Nesses modelos,  $Y$  representa a variável resposta (preço de venda) e  $x$  corresponde, para cada modelo, às covariáveis imposto, área construída e área do terreno respectivamente. Ademais, para posterior análise considera-se o nível de significância  $\alpha = 5\%$ .

- Modelo 5:  $Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad i = 1, 2, \dots, n$
- Modelo 6:  $Y_i = \beta_0 + \beta_1 x_{2i} + \epsilon_i, \quad i = 1, 2, \dots, n$
- Modelo 7:  $Y_i = \beta_0 + \beta_1 x_{3i} + \epsilon_i, \quad i = 1, 2, \dots, n$

Tabela 5: Estimativas, erros padrão, estatística  $t$  e  $p$ -valor sob os MRLS ajustados - Dados *imoveis*.

	Modelo 5: Preço $\sim$ Imposto				Modelo 6: Preço $\sim$ Área construída			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>								
$\beta_0$	5,583	3,111	1,795	0,084	2,506	3,054	0,820	0,420
$\beta_1$	4,543	0,400	11,359	< 0,001	23,804	1,899	12,53	< 0,001

Modelo 7: Preço $\sim$ Área do terreno					
	Estimativa	Erro padrão	$t$	$p$ -valor	
<i>parâmetro</i>					
$\beta_0$	11,057	5,540	1,996	0.057	
$\beta_1$	4.323	0.818	5.284	< 0,001	

Tabela 6: Intervalos de confiança sob os MRLS ajustados - Dados *imoveis*.

<i>Parâmetro</i>	Modelo 5: Preço $\sim$ Imposto		Modelo 6: Preço $\sim$ Área construída	
	Limite Inferior	Limite Superior	Limite Inferior	Limite Superior
$\beta_0$	-0,823	11,989	-3,784	8,796
$\beta_1$	3,719	5,367	19,893	27,715

Modelo 7: Preço $\sim$ Área do terreno		
<i>Parâmetro</i>	Limite Inferior	Limite Superior
$\beta_0$	-0,354	22,468
$\beta_1$	2,638	6,008

A partir da Tabela 8 é possível observar que em todos os modelos a estimativa de  $\beta_1$  é significativa, com valores iguais a 4,543, 23,804 e 4,323 respectivamente. Ademais, o intercepto  $\beta_0$  é estatisticamente não significativo em todos os ajustes.

Os intervalos de confiança apresentados pela Tabela 6 reforçam, a um nível de confiança de 95%, as significâncias notadas anteriormente. Como todos os  $\beta_0$  possuem o valor 0 em seus intervalos de confiança, esses parâmetros não são significativos em seus respectivos modelos. Em contrapartida, por essa mesma justificativa todos os coeficientes  $\beta_1$  são significativos.

A seguir são apresentadas as tabelas ANOVA e testes F para cada um dos modelos indicados. Note que o teste F, através da estatística  $F_0 = \text{QMReg} / \text{QMRes}$ , avalia a importância da covariável  $x$  para explicar  $Y$ , por meio das hipóteses

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0. \end{cases}$$

Tabela 7: Tabelas ANOVA e testes F sob os MRLS ajustados - Dados *imoveis*.

	Modelo 5: Preço $\sim$ Imposto				Modelo 6: Preço $\sim$ Área construída			
	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F
<i>fonte de variação</i>								
Regressão	1	4.458,5	4.458,5	129,04	1	4.591,7	4.591,7	157,11
Resíduo	25	863,8	34,6		25	730,7	29,2	
Total	26	5.322,3			26	5.322,3		

	Modelo 7: Preço $\sim$ Área do terreno			
	Graus de liberdade	Soma de quadrados	Quadrados médios	Valor F
<i>fonte de variação</i>				
Regressão	1	2.807,8	2.807,85	27,917
Resíduo	25	2.514,5	100,58	
Total	26	5.322,3		

Segundo  $H_0$ , a estatística  $F_0 \sim F_{1,25}$ . A partir dos valores observados, obtêm-se p-valores menores que 0,001 para todos os modelos apresentados. Assim, não se aceita  $H_0$  a um nível de significância de 5% para nenhum modelo, indicando que há evidências estatísticas para afirmar que  $\beta_1 \neq 0$ . Em outras palavras, o teste F indica que todos os modelos apresentados são estatisticamente significativos, ou seja, a respectiva variável explicativa está ajudando a explicar o preço de venda do imóvel (variável resposta).



## 2.4 Faça análises de diagnóstico completas para cada modelo de regressão simples ajustado na subseção 2.3, comentando a adequação ou desvio das suposições necessárias. Verifique se existem observações atípicas.

As técnicas de diagnóstico para modelos de regressão linear simples (MRLS) são fundamentais para garantir que as suposições do modelo sejam atendidas e para avaliar sua adequação. O diagnóstico adequado ajuda a melhorar a qualidade da modelagem, identificar problemas potenciais e refinar as previsões. Os MRLS assumem

- (1) Linearidade entre as variáveis dependente e independente;
- (2) Normalidade dos erros, ou seja, os erros devem seguir uma distribuição normal;
- (3) Independência dos resíduos;
- (4) Homoscedasticidade, isto é, a variância dos erros deve ser constante.



A seguir verifica-se a normalidade dos resíduos, conforme o pressuposto (2) dos modelos, através do gráfico de probabilidade normal dos resíduos studentizados e teste de hipóteses Shapiro-Wilk, a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

$$\begin{cases} H_0 : \text{Os resíduos seguem distribuição normal} \\ H_1 : \text{Os resíduos seguem outro modelo.} \end{cases}$$

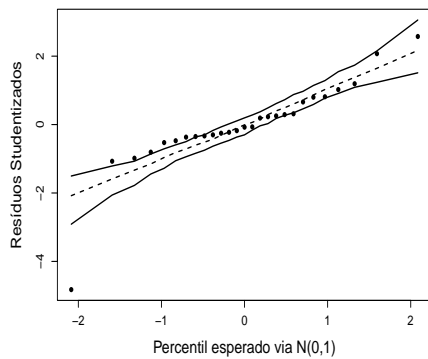
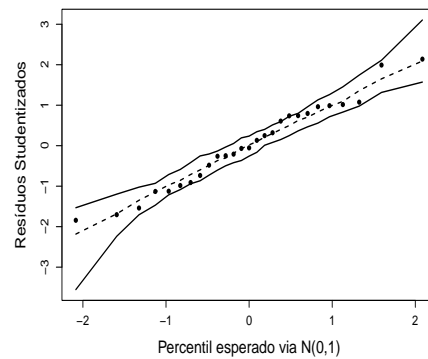
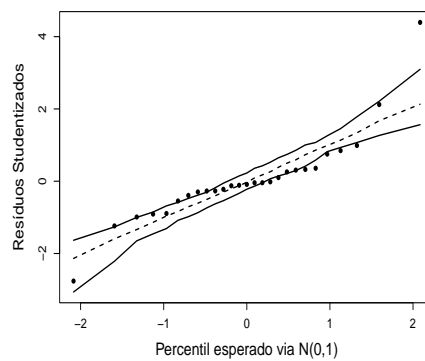
(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terreno

Figura 17: Gráficos de probabilidade normal dos resíduos studentizados com envelopes simulados sob os MRLS ajustados - Dados *imoveis*.

Quadro 6:  $p$ -valor do teste de Shapiro-Wilk sob os MRLS ajustados - Dados *imoveis*

Modelo	$p$ -valor	Decisão do teste
5	$<0,001$	Rejeita $H_0$
6	0,710	Não rejeita $H_0$
7	$<0,001$	Rejeita $H_0$

A partir da Figura 17 e do Quadro 6, pode-se afirmar que há evidências estatísticas de que apenas os resíduos do Modelo 6 seguem uma distribuição normal. Isso porque mais de 95% dos pontos estão dentro das bandas de confiança, e o teste de hipótese não rejeitou a hipótese nula  $H_0$  neste caso. Nos Modelos 5 e 7, percebe-se graficamente que muitos pontos saíram das bandas de confiança.

A seguir investiga-se a independência dos **resíduos**, conforme o pressuposto (3), por meio do gráfico resíduos studentizados vs índice e teste de hipóteses Durbin-Watson, a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

$$\begin{cases} H_0 : \text{Os resíduos são independentes} \\ H_1 : \text{Os resíduos são autocorrelacionados.} \end{cases}$$

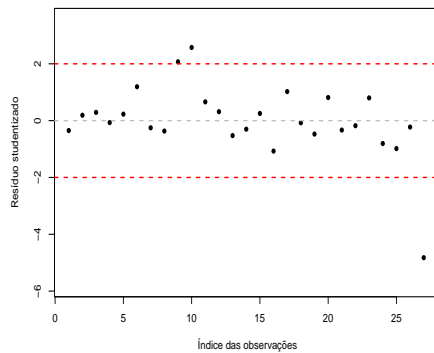
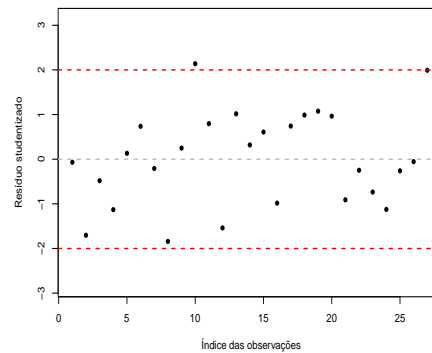
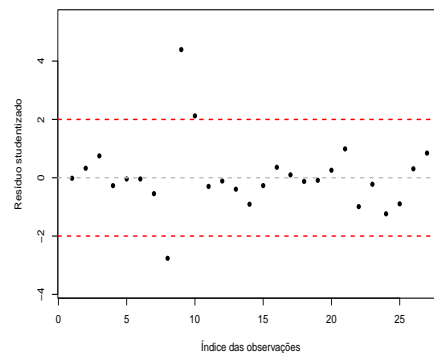
(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terreno

Figura 18: Gráficos dos resíduos studentizados pelo índice das observações sob os MRLS ajustados - Dados *imoveis*.

Quadro 7:  $p$ -valor do teste de Durbin-Watson sob os MRLS ajustados - Dados *imoveis*

Modelo	$p$ -valor	Decisão do teste
5	0,058	Não rejeita $H_0$
6	0,304	Não rejeita $H_0$
7	0,831	Não rejeita $H_0$

Em todos os modelos, os pontos não aparentam seguir alguma tendência, distribuindo-se de forma aleatória pelo gráfico, o que sugere a não correlação dos resíduos.

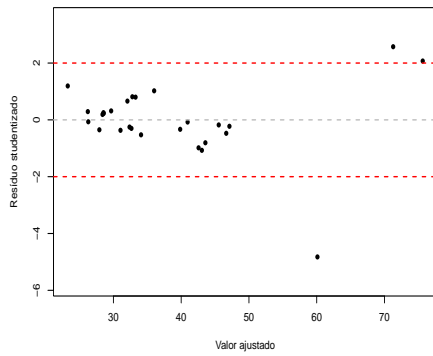
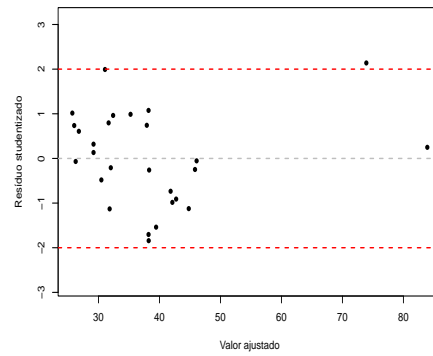
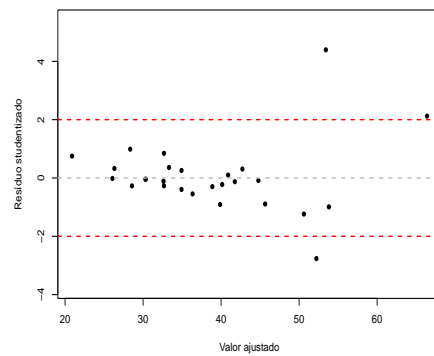
Embora o Quadro 7 mostre que o teste de Durbin-Watson não rejeitou  $H_0$  para

os Modelo 5 e 7, essa conclusão não é totalmente confiável. O teste usado é desenvolvido sob a suposição de que os erros da regressão são normalmente distribuídos. Entretanto, como apresentado pela Figura 17 e pelo Quadro 2, os resíduos desses modelos não seguem distribuição normal, o que implica em uma decisão de teste que pode não estar correta. Já para o Modelo 6 há evidências de independência dos resíduos.

A seguir estuda-se a homoscedasticidade do modelo, conforme o pressuposto (4), por meio do gráfico resíduos studentizados vs valor ajustado e teste de hipóteses Goldfeld-Quandt, a um nível de significância de 5%. Para o teste citado, as hipóteses testadas são

$$\begin{cases} H_0 : \text{Não há heteroscedasticidade} \\ H_1 : \text{Há heteroscedasticidade.} \end{cases}$$

Para tornar o teste mais poderoso, omitiu-se um grupo central de  $r = n/3$  observações, resultando em  $r = 9$ .

(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terrenoFigura 19: Gráficos dos resíduos studentizados pelo valor ajustado sob os MRLS ajustados - Dados *imoveis*.Quadro 8:  $p$ -valor do teste de Goldfeld-Quandt sob os MRLS ajustados - Dados *imoveis*

Modelo	$p$ -valor	Decisão do teste
5	$<0,001$	Rejeita $H_0$
6	0,607	Não rejeita $H_0$
7	$<0,001$	Rejeita $H_0$

Apenas o Modelo 6 parece estar distribuído aleatoriamente ao redor da linha horizontal (zero), a partir da Figura 19, sugerindo homoscedasticidade. Essa constatação é reforçada pelo Quadro 8, que indica que apenas esse modelo é homoscedástico, diferentemente dos demais.

- (a) Pontos aberrantes: são aqueles que apresentam ajustes inadequados para a variável resposta (Y), resultando em resíduos discrepantes. Esses pontos geralmente distorcem o intercepto do modelo de regressão ajustado. Normalmente, considera-se como pontos aberrantes aqueles cujos resíduos studentizados absolutos são maiores que 3, ou seja,  $|t_i^*| > 3$ .

- (b) Pontos alavanca: são aqueles que têm um peso desproporcional no valor ajustado de  $\hat{Y}$ . Esses pontos geralmente apresentam um perfil distinto em relação aos outros, especialmente no que diz respeito ao valor da variável explicativa  $x$ . Além disso, o ponto de alavanca "puxa" a reta de regressão ajustada em sua direção, pois geralmente afeta o coeficiente angular estimado da reta.
- (c) Pontos influentes: são aqueles que têm uma influência desproporcional nas estimativas dos parâmetros do modelo e podem alterar as conclusões inferenciais. Estas observações podem causar mudanças relativas consideráveis nas estimativas  $\beta_0$ ,  $\beta_1$  e  $\sigma^2$ , além poderem mudar o sinal da estimativa e a significância dos coeficientes  $\beta_0$  e  $\beta_1$ .

Observa-se por meio das Figuras 18 e 19 que os Modelos 5 e 7 possuem três pontos que devem ser investigados na análise. Em especial, possuem um e dois pontos aberrantes, respectivamente, pois são pontos que  $|t_i^*| > 3|$ . Já o Modelo 6 possui apenas um ponto  $|t_i^*| > 2|$  que deve ser analisado.

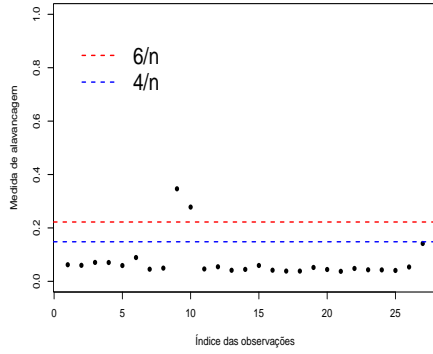
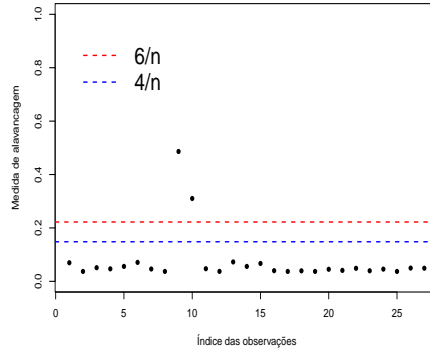
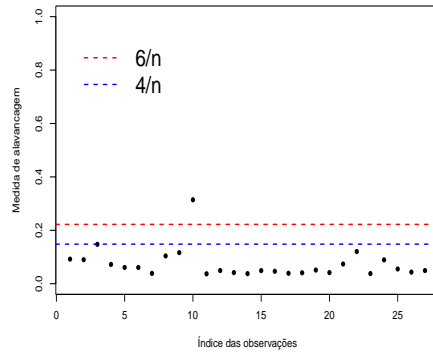
Com relação aos pontos alavanca, define-se a medida de alavancagem da  $i$ -ésima observação como

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}},$$

em que

- $n$  é o número de observações da amostra;
- $x_i$  é o  $i$ -ésimo valor fixado da covariável  $x$ ;
- $\bar{x}$  é a média da covariável  $x$ ;
- $S_{xx} = \sum_{i=1}^n x_i(x_i - \bar{x})$ .

Como a média amostral de  $h_{ii}$  é  $2/n$ , pode-se investigar as observações que sejam maiores que o dobro ou que o triplo da média amostral, isto é,  $4/n$  e  $6/n$  respectivamente.

(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terrenoFigura 20: Gráficos da medida de alavancagem sob os MRLS ajustados - Dados *imoveis*.

Com base na Figura 20, nota-se que as observações nos Modelos 5 e 6 formam uma linha horizontal abaixo da faixa  $4/n$ . Entretanto, os mesmos modelos apresentam dois pontos com  $h_{ii} > 6/n$ . Já o Modelo 7 apresenta pontos mais dispersos abaixo da linha tracejada azul, com apenas uma observação alavanca.

Com relação aos pontos de influência, realiza-se a identificação desses pontos influentes no ajuste do modelo de regressão linear por meio de três medidas de influência: DFFITS, DFBETAS e Distância de Cook.

A medida DFFITS (Difference in Fits) pode ser vista como a variação padronizada no  $i$ -ésimo valor ajustado de  $Y$  quando o  $i$ -ésimo ponto é excluído. Tal medida para a  $i$ -ésima observação é definida por

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_i^{(i)}}{\sqrt{\sigma_{(i)}^2 h_{ii}}},$$

em que  $\sigma_{(i)}^2$  é a variância da resposta  $\sigma^2$  estimada excluindo a  $i$ -ésima observação. Para identificar pontos de influência a partir desta medida, basta avaliar as observações tais que

$$|DFFITs_i| > 2\sqrt{\frac{2}{n}}.$$

Para avaliar o impacto da  $i$ -ésima observação nas estimativas dos coeficientes  $\beta_0$  e  $\beta_1$ , pode-se utilizar a medida DFBETAS, definida por

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_i^{(i)}}{\sqrt{\text{Var}(\hat{\beta}_j^{(i)})}}$$

em que  $\text{Var}(\hat{\beta}_j^{(i)})$  é a variância de  $\hat{\beta}_j$  avaliada na estimativa de  $\sigma^2$  obtida excluindo a  $i$ -ésima observação. Uma regra geral é investigar os pontos tais que

$$|DFBETAS_{ji}| > \frac{2}{\sqrt{n}}.$$

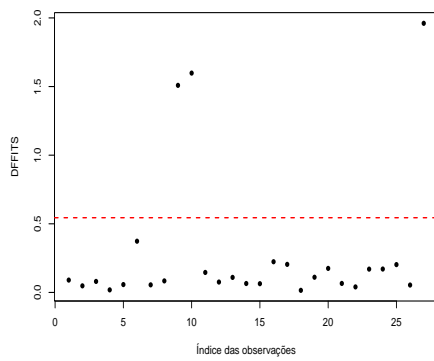
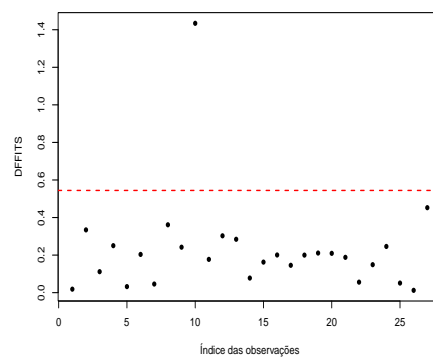
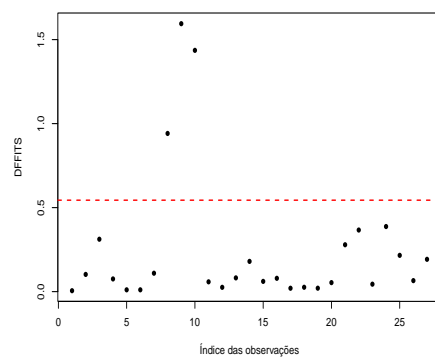
Por fim, a medida da Distância de Cook para a  $i$ -ésima observação é definida por

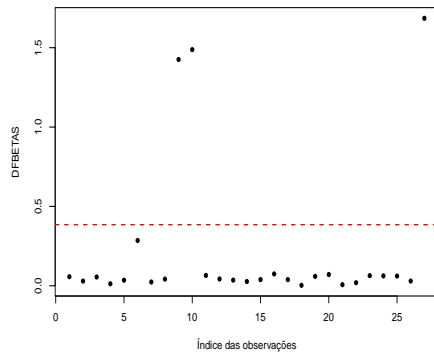
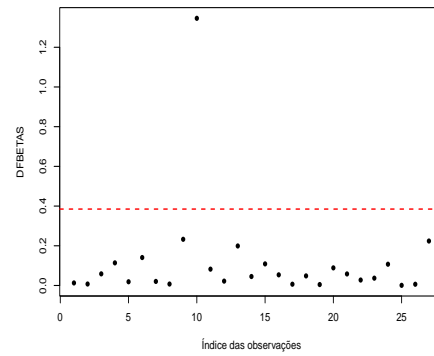
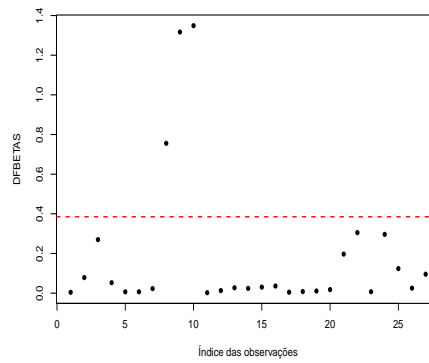
$$D_i = \frac{1}{2\hat{\sigma}^2} \sum_{j=1}^p (\hat{Y}_j - \hat{Y}_j^{(i)})^2.$$

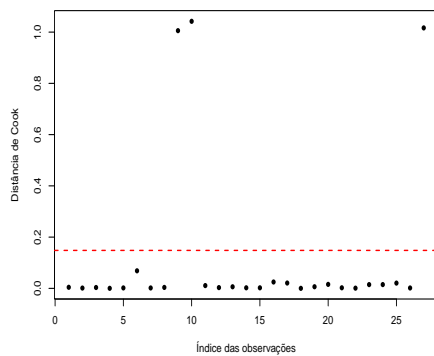
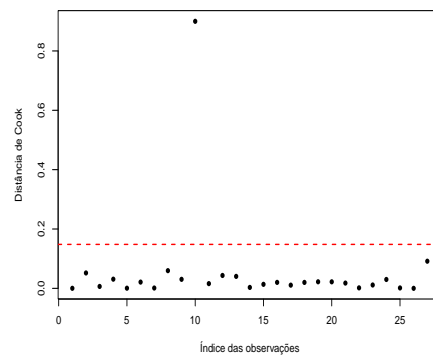
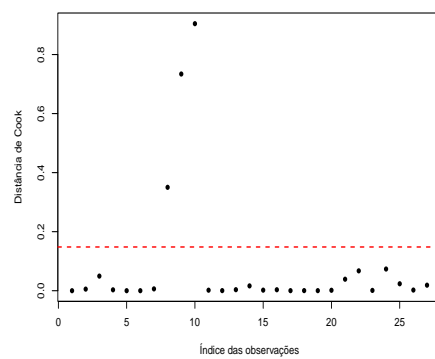
Para  $n$  grande, uma regra é identificar o ponto como possível influente se

$$D_i > 4/n.$$



(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terrenoFigura 21: Gráficos da medida DFFITS sob os MRLS ajustados - Dados *imoveis*.

(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terrenoFigura 22: Gráficos da medida DFBETAS sob os MRLS ajustados - Dados *imoveis*.

(a) Modelo 5: Preço  $\sim$  Imposto(b) Modelo 6: Preço  $\sim$  Área construída(c) Modelo 7: Preço  $\sim$  Área do terrenoFigura 23: Gráficos da distância de Cook sob os MRLS ajustados - Dados *imoveis*.

Com base nas Figuras 21, 22 e 23, observa-se a **presença de pontos** influentes segundo as três medidas analisadas. Tais medidas apontam três pontos de influência para os Modelos 5 e 7. Para o Modelo 6, DFBETAS e a distância de Cook indicaram duas observações e DFFITS indicou apenas uma.

A seguir são avaliados os modelos propostos retirando-se os pontos atípicos, isto é, as observações aberrantes, de alavanca e influentes.

Tabela 8: Estimativas, erros padrão, estatística  $t$  e  $p$ -valor sob os MRLS ajustados retiradas as observações atípicas - Dados *imoveis*.

	Modelo 5: Preço $\sim$ Imposto				Modelo 6: Preço $\sim$ Área construída			
	Estimativa	Erro padrão	$t$	$p$ -valor	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>								
$\beta_0$	13,313	2,570	5,180	< 0,001	14,879	4,845	3,071	0,005
$\beta_1$	3,325	0,390	8,526	< 0,001	14,524	3,456	4,203	< 0,001



	Modelo 7: Preço $\sim$ Área do terreno			
	Estimativa	Erro padrão	$t$	$p$ -valor
<i>parâmetro</i>				
$\beta_0$	21,318	2,841	7,503	< 0,001
$\beta_1$	2,377	0,467	5,089	< 0,001

Houve variações esperadas nas estimativas e nos erros padrão dos parâmetros, mas os sinais permaneceram os mesmos. No entanto, ocorreram mudanças na significância de  $\beta_0$  em todos os modelos. Em outras palavras, alguns  $\beta_0$  que eram significativos antes da remoção dos pontos deixaram de ser, enquanto outros que não eram passaram a ser significativos.

**2.5 A partir das conclusões obtidas na subseção 2.4, escolha o melhor modelo de regressão simples ajustado. Para o modelo escolhido, realize as interpretações adequadas. Discuta sobre a explicabilidade do modelo. Faça previsões do preço de imóveis fixando novos valores para a covariável considerada.**

A partir da análise dos pressupostos, apenas o Modelo 6 atendeu às suposições necessárias - normalidade, independência e homoscedasticidade dos erros. Além disso, o Modelo tem bom ajuste com  $R^2 = 0,862$ .

Dessa forma, a reta ajustada pelo modelo escolhido é definida como

$$\hat{Y}_i = 2,506 + 23,804x_i.$$

Teoricamente, o intercepto  $\beta_0$  representa o valor esperado de  $Y_i$  quando  $x_i = 0$ . Contudo, como a covariável  $x$  (área construída) assume apenas valores maiores que 0, essa interpretação não é válida. Já  $\beta_1$  indica que um aumento de uma unidade na área construída está associado a um acréscimo médio de 23,804 no preço de venda.

Fixando-se um valor arbitrário para  $x$  no intervalo entre o valor mínimo (0,975) e máximo (3,42) observado na amostra que ajustou o modelo, pode-se prever o valor de  $Y$ . Sendo  $x_i = 1,121$ , por exemplo, tem-se que  $Y_i = 29,19$ . Em outras palavras, pela previsão do modelo, uma propriedade de 1,121 pés quadrados teria um preço de venda de 29,19 (em U\$100).

## Referências

RIBEIRO, T. K. A. *Análise de Diagnóstico sob MRLS*. 2024. Slide ministrado em sala de aula.

RIBEIRO, T. K. A. *Regressão Linear Simples*. 2024. Slide ministrado em sala de aula.