Lista 4 - Estatística descritiva

Guilherme Marthe, nusp: 8661992

1. Considere o modelo

$$y_i = \beta x_i + e_i, \quad i = 1, ..., n$$

Em que $E(e_i) = 0$ e $Var(e_i) = \sigma^2$, erros aleatórios e não correlacionados.

a) Obtenha o estimador de mínimos quadrados de β e proponha um estimador não viciado para σ^2.

Para chegar no estimador de mínimos quadrados, basta rearranjar os termos da equação acima da seguinte forma:

$$Q(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta \cdot x_i)^2$$

Então, através da condição de primeira ordem para otimizar $Q(\beta)$ temos:

$$\frac{dQ}{d\beta} = 0$$

E derivando $Q(\beta)$ temos:

$$\frac{dQ}{d\beta} = \sum_{i=1}^{n} 2 \cdot (y_i - \beta \cdot x_i) \cdot (-x_i) = 0 \implies \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Para estimarmos σ^2 podemos partir do pressuposto que quando o calculamos em uma regressão com intercpeto, a correção para acabar com o viés leva em conta a soma dos quadrados dos resíduos e o número de parâmetros estimados. Como aqui no caso da regressão simples sem intercepto estimamos apenas uma parâmetro, propomos o seguinte estimador:

$$\widehat{Var(e_i)} = \widehat{\sigma}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\mathbf{SQR}}{n-1}$$

(b) Que distribuição (aproximada) você proporia para o estimador de β obtido em (a)? Justifique sua resposta.

Sob a hipótese de n grande o suficiente, podemos depender do teorema do limite central para propor uma distribuição para $\hat{\beta}$. Uma vez que $E(\hat{\beta}) = \beta$ e $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$ podemos propor que:

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{Var(x_i)})$$

(c) Com base nos itens anteriore, especifique um intervalo de confiança aproximado para β, com coeficiente de confiança $\gamma \in (0, 1)$.

Assim, podemos propor o seguinte intervalo de confiânça para $\hat{\beta}$:

$$\beta \in [\hat{\beta} \pm \widehat{ep(\hat{\beta})} \cdot z_\gamma]$$

Onde z_γ o quantil da normal padrão sob $\gamma\%$ de confiança.

2. Num estudo realizado na Faculdade de Medicina da Universidade de São Paulo foram colhidos dados de 16 pacientes submetidos a transplante intervivos e, em cada um deles, obtiveram-se medidas tanto do peso (g) real do lobo direito do fígado quanto de seu volume (em cm3) previsto pré-operatoriamente por métodos ultrassonográficos. O objetivo é estimar o peso real por meio do volume previsto. Os dados estão dispostos na tabela a seguir.

volume	peso
656	630
692	745
588	690
799	890
766	825
800	960
693	835
602	570
737	705
921	955
923	990
945	725
816	840
584	640
642	740
970	945

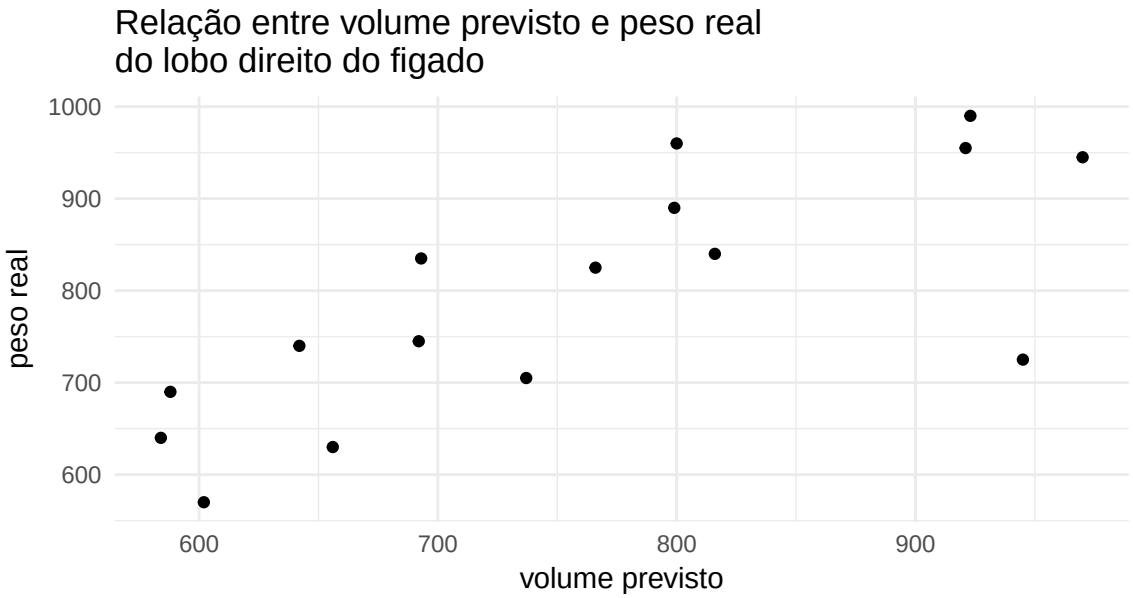
(a) Proponha um modelo de regressão linear simples com intercepto para analisar os dados e interprete seus parâmetros.

Proponho o seguinte modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, ..., n$$

Onde y_i representa peso real do lobo direito do fígado, x_i representa o volume previsto. Neste modelo, β_1 representa a mudança esperada do peso real(y_i) para a mudança em uma unidade de volume previsto (x_i). Além disso, da maneira que está construido o modelo, β_0 fica sem uma interpretação direta, pois um volume previsto de 0 não faz sentido.

(b) Construa um gráfico de dispersão apropriado. Interprete.



O gráfico de dispersão anterior indica alguns pontos:

- o modelo linear pode ser o mais indicado para esse conjunto de dados;
- porém, pode-se haver uma tendência logarítimica ao se checar uma dispersão menor o redor dos pontos no canto superior direito;
- o ponto no canto inferior direito pode se mostrar discrepante com relação aos outros.

(c) Ajuste o modelo e interprete os valores obtidos para os parâmetros

	Model 1
(Intercept)	213.28
	(133.33)
volume	0.76***
	(0.17)
R^2	0.58
Adj. R^2	0.55
Num. obs.	16
RMSE	87.91

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

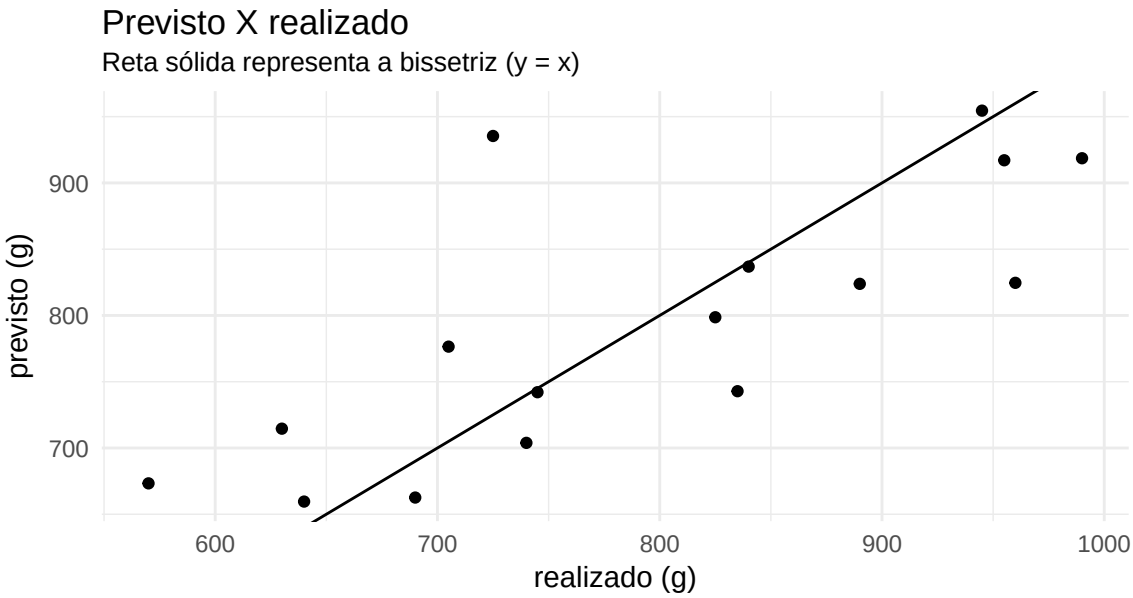
Table 1: Statistical models

O modelo ajustado acima tem a seguinte interpretação:

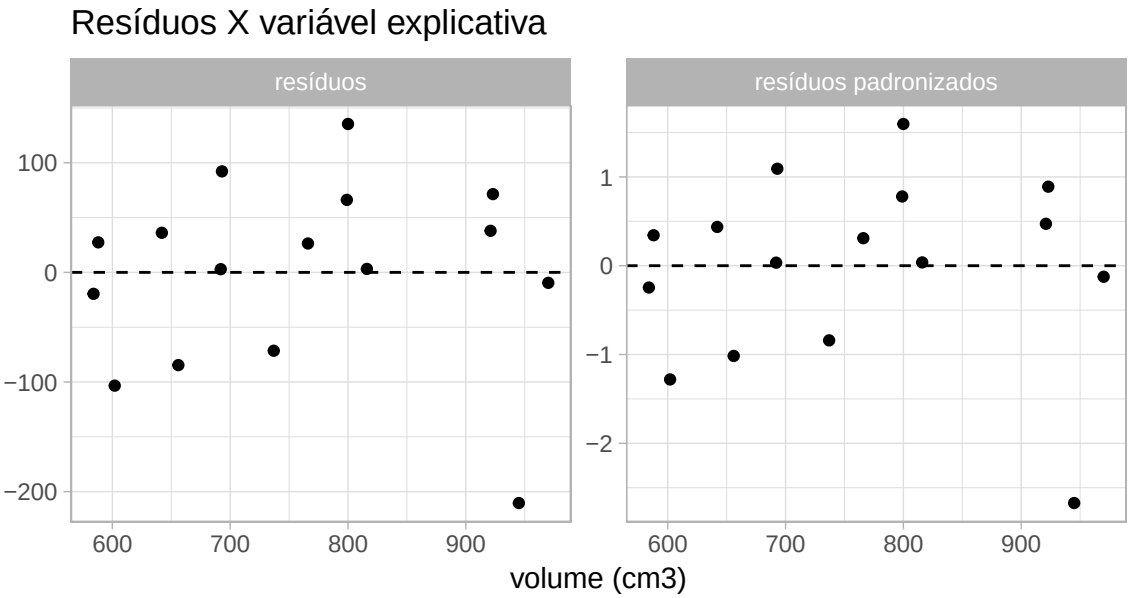
- a cada mudança de uma unidade (cm3) de volume previsto é esperado um aumento de 0.7642 gramas do peso real do lobo direito do fígado.
- é possível notar que devido ao alto erro padrão da estimativa do intercepto, talvez não faça sentido mantê-lo na forma funcional do modelo.

(d) Avalie o ajuste do modelo por meio de medidas descritivas e resíduos.

De acordo com a tabela do item anterior, as medidas de R2 e R2 ajustado indicam um que a variável volume estimado representam 58% das variações do peso do rim.

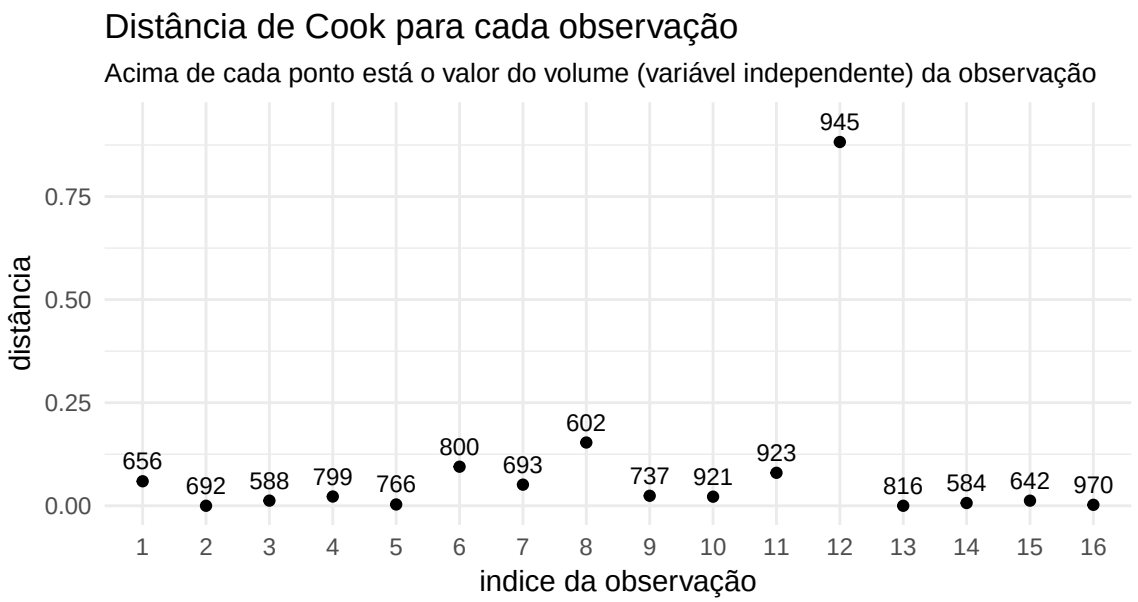


A dispersão dos pontos previstos pela regressão e realizados (acima) uma dispersão possívelmente aleatória o suficiente ao longo da reta para indicar um bom ajuste.

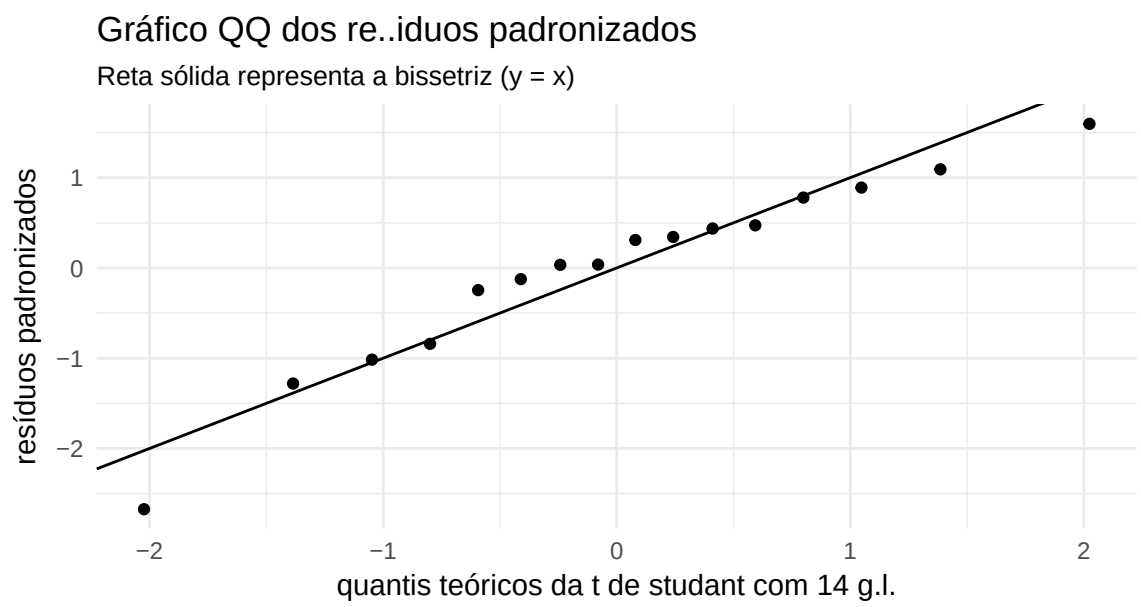


Ao analizar gráfico de resíduos e gráfico de resíduos padronizados ao longo da variável explicativa volume previsto, podemos ressaltar alguns pontos:

- existe um padrão relativamente ascendente (pelo menos não aleatório) dos resíduos ao longo das faixas de 600g e 800g do volume previsto;
- a medida encontrada no canto inferior esquerdo de ambos os gráficos pode ser considerada um ponto anormal em termos do resíduo resultante do modelo escolhido;



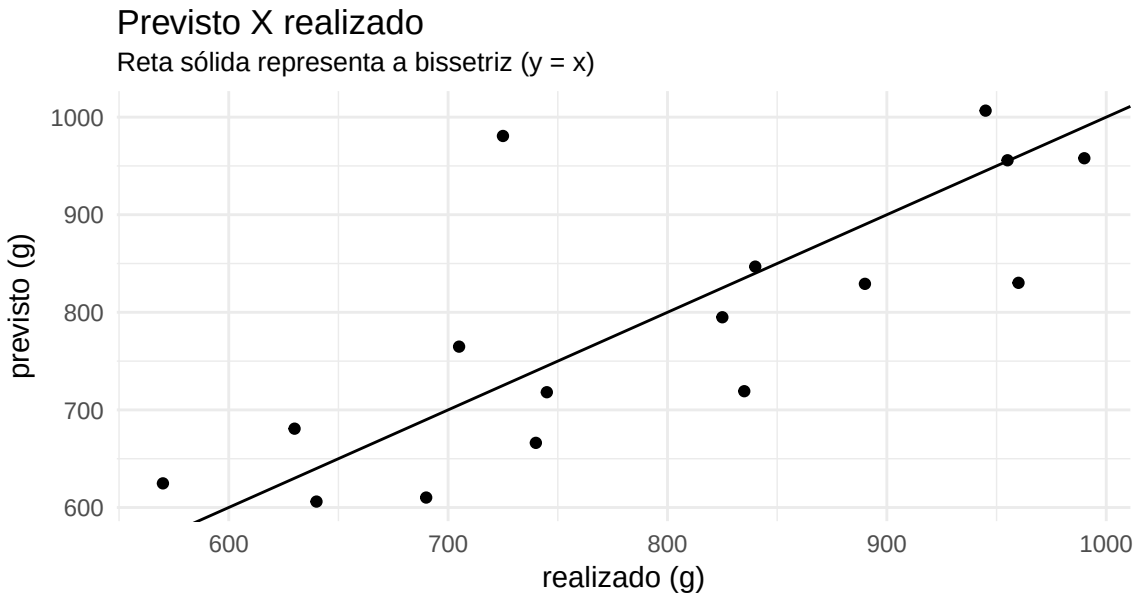
Analisando o gráfico da distância de cook para cada observação, podemos ver que a observação 12 (com um valor de volume previsto igual a 945) possui uma distância de Cook maior que todas as outras observções, indicando um possível outlier.

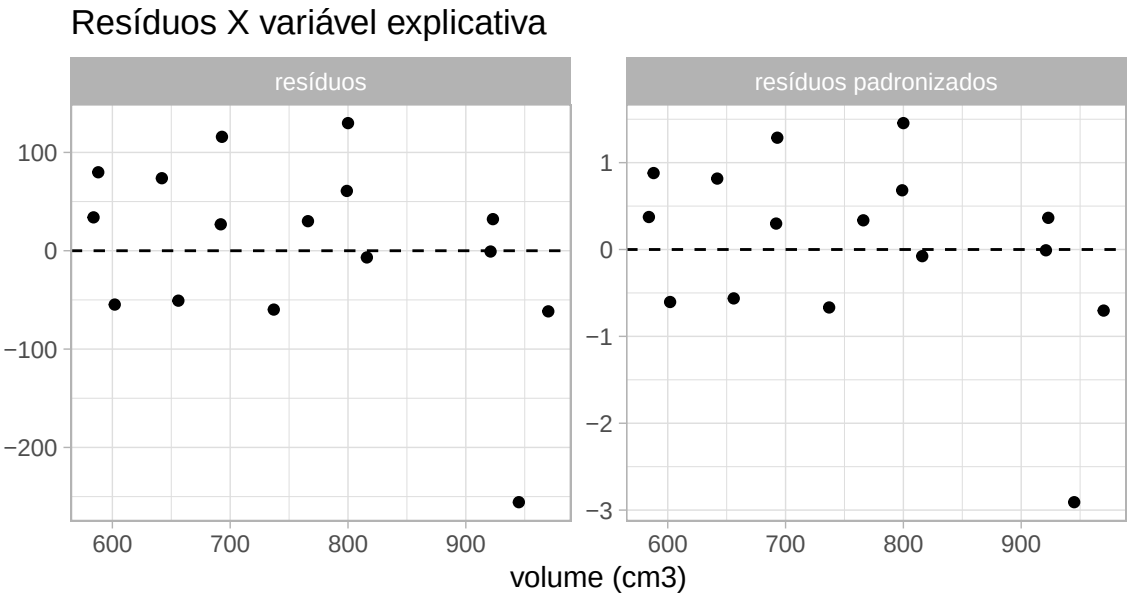


Por fim, um gráfico QQ dos reśiduos padronizados com relação a distribuição t de student com 14 graus de liberdade indica que talvez o ajuste do modelo seja adquado uma vez que a distribuição estão relativamente dispersa ao longo da bissetriz.

(e) Repita os itens anteriores considerando um modelo linear simples sem intercepto.

Primeiro reproduziremos os gráficos e tebelas do item anterior, e em seguida analisaremos brevemente as diferenças.

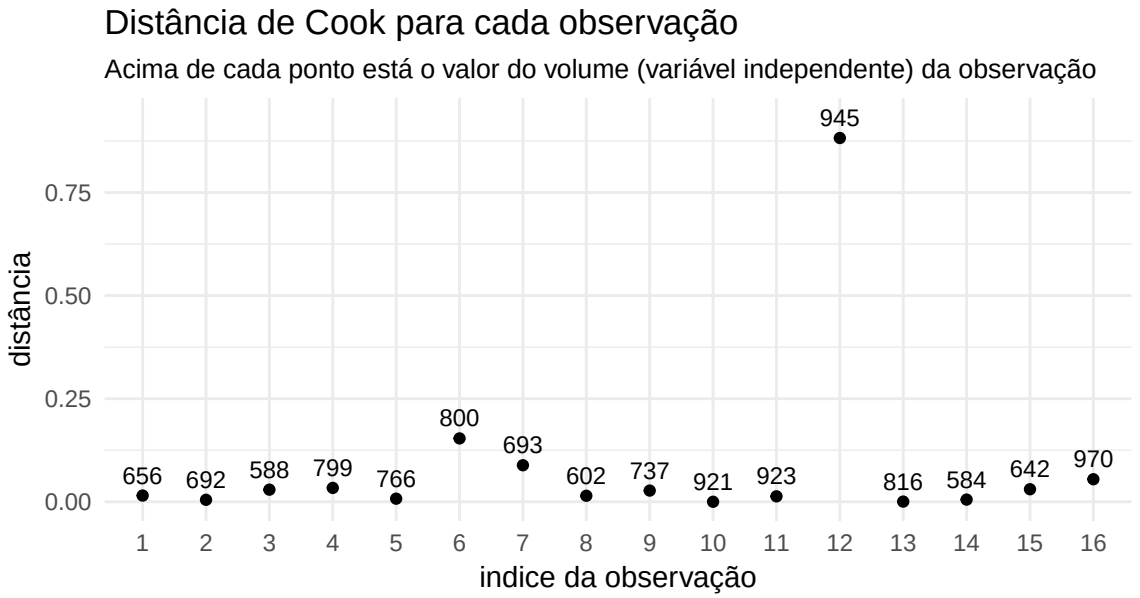


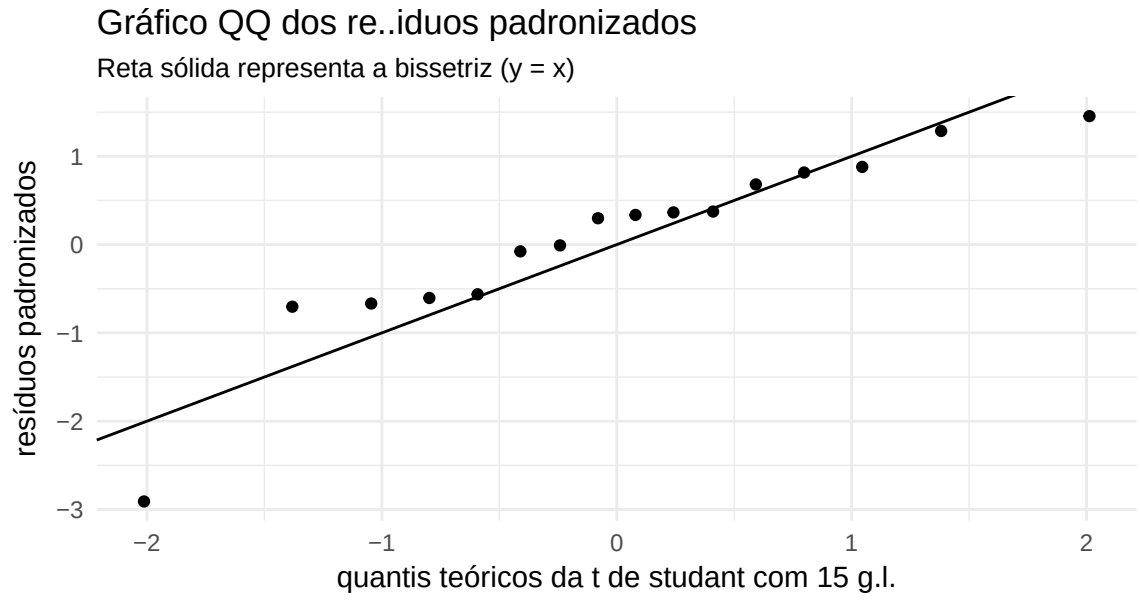


	Model 1
volume	1.04***
	(0.03)
R^2	0.99
Adj. R^2	0.99
Num. obs.	16
RMSE	92.36

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Statistical models





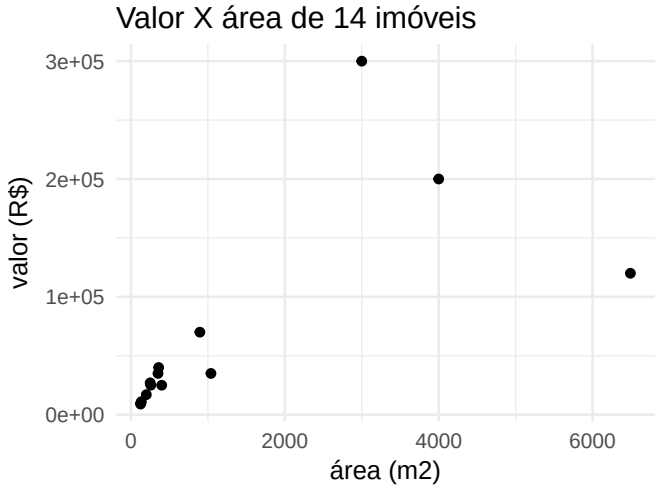
As diferenças mais relevantes entre os ajustes de cada modelo são:

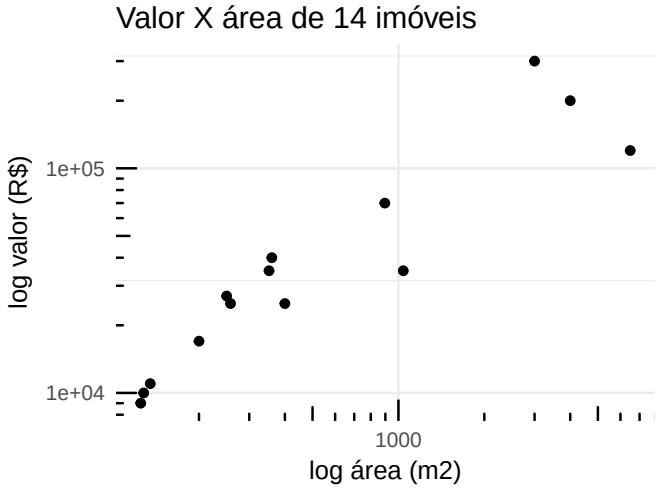
- o aumento expressivo dos coeficientes de determinação R2 e R2 ajustado indicam que 98% das variações do peso são explicadas pelo volume previsto do rim;
- um maior indicativo de que a observação 12 é de fato incomum para os dados, como evidênciado pelo aumento do resíduo padronizado delas (~3);
- o impacto na análise da distribuição do resíduos padronizados indicando ainda mais que algun pontos impedem a boa aderência dos resíduos com o modelo teórico adequado.

3. Os dados abaixo são provenientes de uma pesquisa para cujo objetivo é propor um modelo para a relação entre a área construída de um determinado tipo de imóvel e o seu valor de mercado.

imovel	area	valor
1	128	10000
2	125	9000
3	200	17000
4	4000	200000
5	258	25000
6	360	40000
7	896	70000
8	400	25000
9	352	35000
10	250	27000
11	135	11000
12	6492	120000
13	1040	35000
14	3000	300000

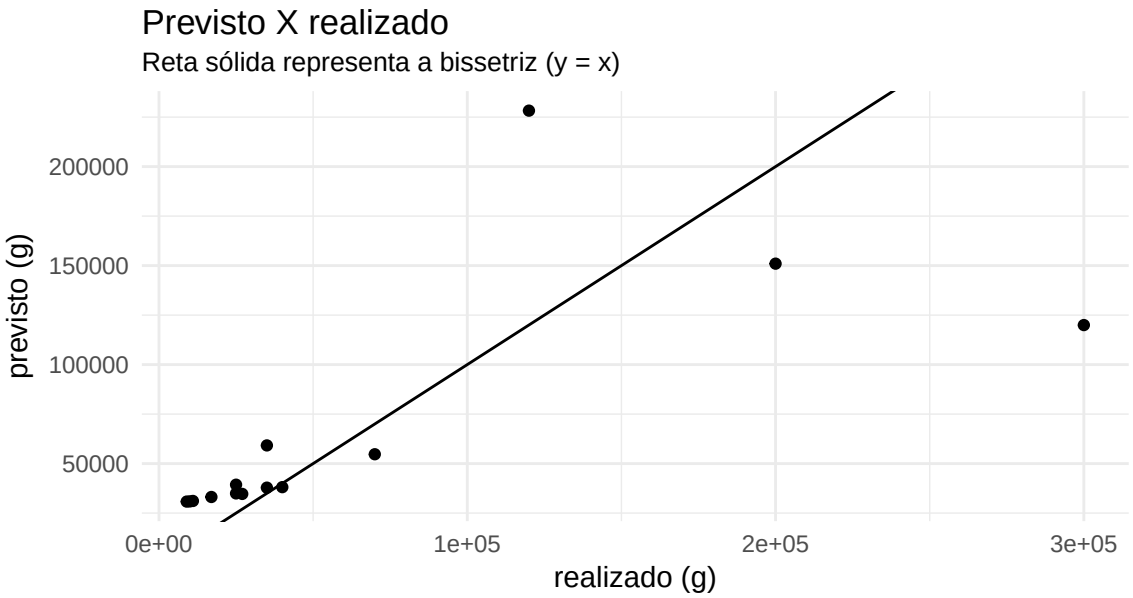
(a) Construa um gráfico de dispersão apropriado para esses dados.



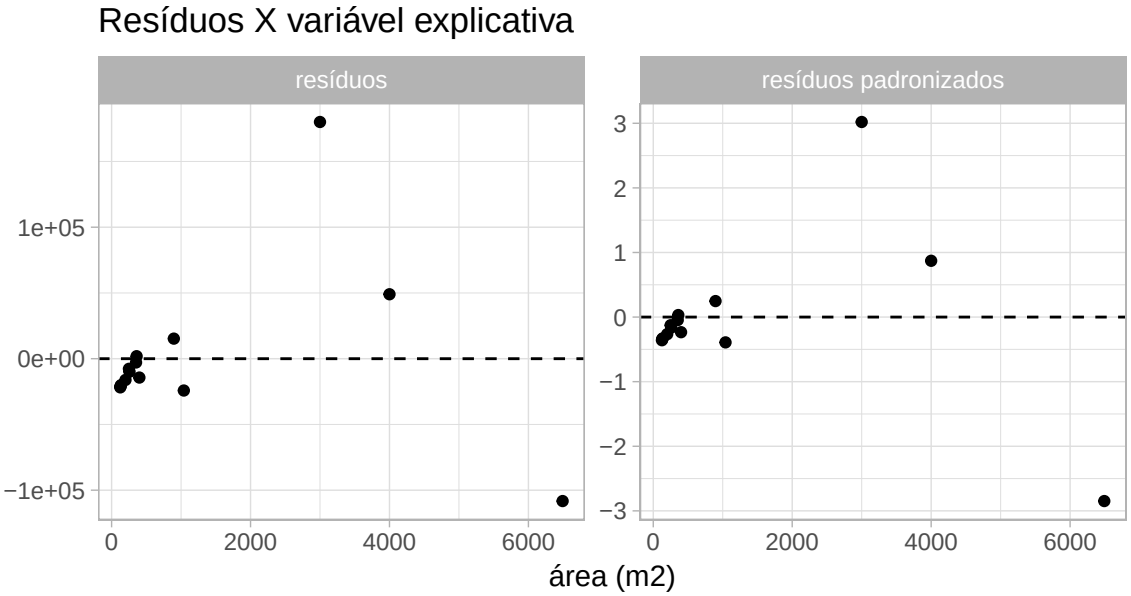


(b) Ajuste um modelo de regressão linear simples e avalie a qualidade do ajuste (obtenha estimativas dos parâmetros e de seus erros padrões, calcule o coeficiente de determinação e construa gráficos de resíduos e um gráfico tipo QQ.

Sobre o ajuste do modelo, podemos dizer que o erro padrão do intercepto é alto, indicando que talvez não precise estar no modelo. Além disso, podemos ver pelo R2 e R2 ajustado que 48% das variações do valor de um imóvel é explicada pela sua área. É esperado que para cara metro qudrado a mais, o aumento em seu valor seja em 30 reais (aproximadamente).



O gráfico previsto X realizado indica um boa ajuste para valores realizados baixos, porém para valores altos o erro é grande.



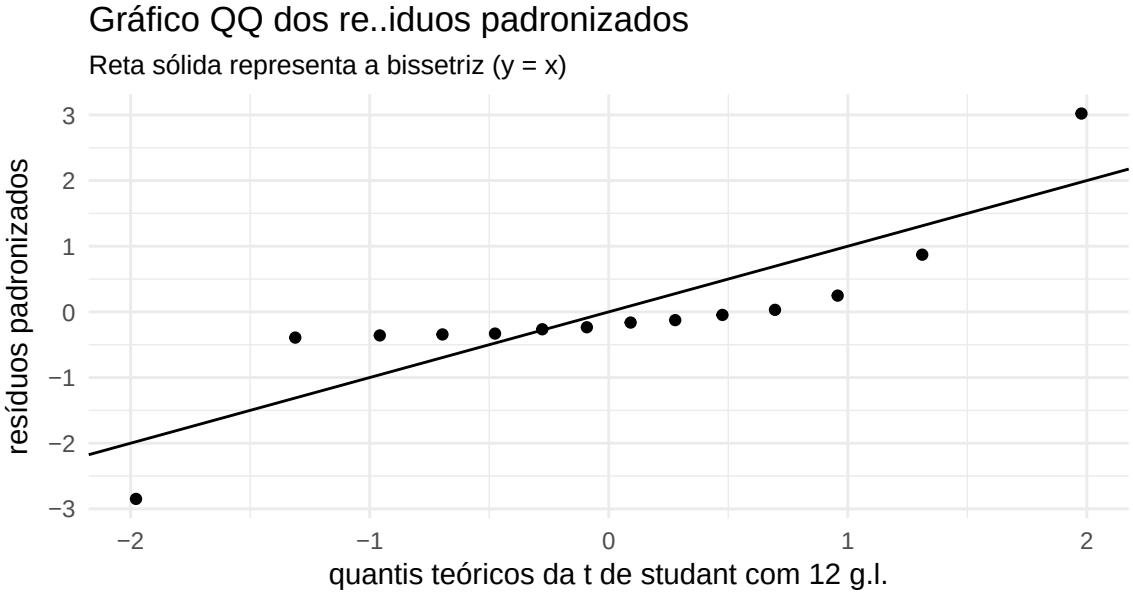
Ao analisar os resíduos e resíduos padronizados podemos ver que o mesmo padrão ocorre. Os resíduos discrepantes

	Model 1
(Intercept)	26934.57
	(20758.38)
area	31.01**
	(9.31)
R^2	0.48
Adj. R^2	0.44
Num. obs.	14
RMSE	64100.24

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Statistical models

ocorrem para valores altos de área, aqueles geralmente associados à valores de imóveis altos em que o modelo atual "erra" bastante.



O gráfico QQ para os resíduos padronizados indicam que o modelo proposto é inadequado, uma vez que os ponto sem dúvidas não ficam sobre a bissetriz.

(c) Ajuste agora um modelo linearizável

$$y = \beta x^{\gamma} e$$

em que y representa o preço e x representa a área.Avalie a qualidade do ajuste comparativamente ao modelo linear ajustado no item anterior; construa um gráfico de dispersão com os dados transformados.

Com o modelo proposto no enunciado podemos ter a seguinte versão linear dele:

$$log(y) = log(\beta) + \gamma \cdot log(x) + log(e)$$

Que é estimável usualmente por mínimos quadrados oridnários após transformar as variáveis.

	Model 1
(Intercept)	5.72***
	(0.55)
log(area)	0.77***
	(0.09)
R^2	0.87
Adj. R^2	0.85
Num. obs.	14
RMSE	0.41

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Statistical models

Claramente o ajuste é melhor, uma vez que ouve um salto grande no coeficiente de determinação e todos os parâmetros estimados têm um erro padrão baixo.

(d) Discuta as vantagens e desvantagens de uso de cada um dos modelos.

O primeiro modelo tem a vantagem de ter um parâmetro associado à variável independente facilmente interpretável. Além disso, com pequenos ajustes, o modelo também pode ter um intercepto interpretável. Todavia o segundo modelo ajuda a explicar as variações da variável resposta (valor do imóvel) melhor, ás custas de uma interpretação mais complexa do modelo.