Lista 4

Guilherme Navarro - NºUSP: 8943160 E Matheus da Ponte Nicolay Nº USP: 10297477

Exercício 1

Um modelo expresso por: $y_i = \beta x_i + e_i$ com $i = 1, .., n$

Sendo y_i a variável resposta, x_i a variável independente, e_i o erro aleatório e por fim β o acréscimo (ou decréscimo) esperado na resposta y_i quando x_i é acrescido de uma unidade.

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

a)

Vamos considerar que $Q(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$ Seja uma função dos erros ao quadrado e queremos minimiza-los, então devemos calcular sua derivada e igualarmos a zero, logo:

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

Então o Estimador de mínimos quadrados para β é:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Propor um estimador não viciado para σ^2

Sabendo que:

$SQTot = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$SQReg = \sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2$

$SQRes = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$

$$\widehat{\sigma}^2 = S^2 = \frac{Q(\hat{\beta})}{n-1} = \frac{SQRes}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\widehat{e_i}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\beta} x_i)^2$$

Onde $Q(\widehat{\beta})$ é a soma dos quadrados dos resíduos.

Para verificar que S^2 é um estimador não viezado para σ^2 temos que mostrar que $\mathbb{E}(S^2) = \sigma^2$ E como SQRes = SQTot - SQReg, temos que:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\mathbb{E}(SQReg) = \frac{1}{n-1}\mathbb{E}(SQTot - SQRes) = \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2) =$$

$$= \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}((\widehat{\beta} x_i)^2 - 2\widehat{\beta} Y_i x_i + \bar{Y}^2) = \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\widehat{\beta}\sum_{i=1}^{n} x_i Y_i + \widehat{\beta}^2\sum_{i=1}^{n}(x_i)^2) =$$

$$= \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\widehat{\beta}\sum_{i=1}^{n} x_i Y_i + \widehat{\beta}^2\frac{\sum_{i=1}^{n} Y_i x_i}{\widehat{\beta}}) = \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\widehat{\beta}\sum_{i=1}^{n} x_i Y_i + \widehat{\beta}\sum_{i=1}^{n} Y_i x_i) =$$

$$= \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}(Y_i - \bar{Y})^2) - \mathbb{E}(\widehat{\beta}\sum_{i=1}^{n} x_i Y_i) = \frac{1}{n-1}(\mathbb{E}(Y_i) - n\mathbb{E}(\bar{Y}^2) - \mathbb{E}(\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}\sum_{i=1}^{n} x_i Y_i)) =$$

$$= \frac{1}{n-1}(\mathbb{E}(Y_i) - n\mathbb{E}(\bar{Y}^2) - \frac{1}{\sum_{i=1}^{n} x_i^2}\mathbb{E}(\sum_{i=1}^{n}(x_i Y_i)^2)) = \frac{1}{n-1}(\sum_{i=1}^{n}[Var(Y_i) + \mathbb{E}^2(Y_i)] - n[Var(\bar{Y}) + \mathbb{E}^2(\bar{Y})]$$

$$-\frac{1}{\sum_{i=1}^{n} x_i^2}[Var(\sum_{i=1}^{n} x_i Y_i) + \mathbb{E}^2(\sum_{i=1}^{n} x_i Y_i)) = \frac{1}{n-1}(\sum_{i=1}^{n}\sigma^2 + (\beta x_i)^2) - n(\frac{\sigma^2}{n} + (\beta x_i)^2) - \frac{1}{\sum_{i=1}^{n} x_i^2}[\sum_{i=1}^{n}\beta x_i]^2 =$$

$$= \frac{1}{n-1}(n\sigma^2 + \sum_{i=1}^{n}(\beta x_i)^2 - \sigma^2 - (\sum_{i=1}^{n}\beta x_i)^2) = \frac{1}{n-1}\sigma^2(n-1) = \sigma^2$$

Portanto S^2 é um estimador não viezado para σ^2

b)

Para o estimador $\widehat{\beta}$ temos:

$$\mathbb{E}(\widehat{\beta}) = \mathbb{E}(\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}) = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}\mathbb{E}(Y_i) = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}\beta x_i = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2}\beta = \beta$$

E para Variância de $\widehat{\beta}$ temos:

$$Var(\widehat{\beta}) = Var(\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}) = (\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2})^2 Var(Y_i) = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^4}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Assim, como temos que os erros tem $\mathbb{E}(e_i) = 0$ e $Var(e_i) = \sigma^2$ tem assintóticamente (TLC) distribuição aproximadamente normal, logo $e_i \sim N(0; \sigma^2)$ o que implica $Y_i \sim N(\beta x; \sigma^2)$, isso mostra que β é uma combinaçãode v.a. com distribuição aproximandamente normal, portanto:

$$\widehat{\beta} \sim N(\beta; \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$$

c)

Tomando do item anterior que $\widehat{\beta}$ tem distribuição aproximadamente Normal especificar um IC com $\gamma \in (0,1)$, utilizando o estimador não viciado para $Var(\widehat{\beta}) = \widehat{\sigma}^2$ calculado no item a), portanto pelo fato de usar uma estimador para σ^2 a distribuição passará a ser t-Student com (n-1) graus de liberdade, e como $\widehat{\beta} \sim N(\beta; \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$ Padronizando, temos:

$$\frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}} \sim N(0,1) \Rightarrow \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}}} \sim t(n-1) \therefore$$

$$\mathbb{P}(-t_{\frac{\gamma}{2}(n-1)} \leq \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}}} \leq t_{\frac{\gamma}{2}(n-1)}) = 1 - \gamma$$

$$\Rightarrow \mathbb{P}(\widehat{\beta} - t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}} \leq \widehat{\beta} - \beta \leq t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}}) = 1 - \gamma$$

$$\Rightarrow \mathbb{P}(\widehat{\beta} - t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}} \leq \beta \leq \widehat{\beta} + t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}}) = 1 - \gamma$$

$$\Rightarrow IC(\beta; \gamma) = [\widehat{\beta} - t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}} \leq \beta \leq \widehat{\beta} + t_{\frac{\gamma}{2}(n-1)}\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n} x_i^2}}]$$

Com $\gamma \in (0,1)$

Exercício 2

Volume US (cm^3)	Peso Ideal (g)	Volume US (cm^3)	Peso Ideal (g)
656	630	737	705
692	745	921	955
588	690	923	990
799	890	945	725
766	825	816	840
800	960	584	640
693	835	642	740
602	570	970	945

a)
Um modelo de regressão linear simples tem a seguinte equação:

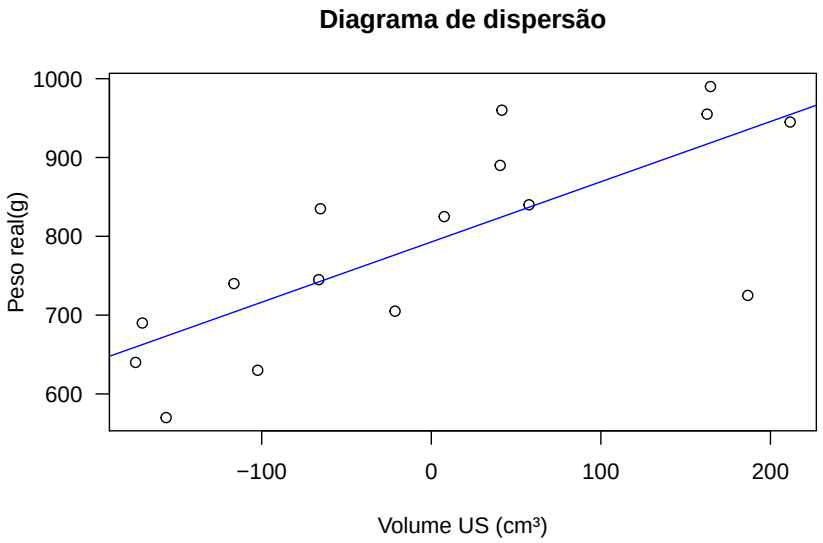
$$Y = 0.7642X + 213.28$$

Onde 0.7642 representa o acréscimo esperado no peso real (g) quando o volume previsto em cm^3 é incrementado em uma unidade, já o valor 213.28 é o valor esperado de um figado com volume 0 cm^3, que por sinal não faz sentido, logo fazendo uma correção no modelo de forma que $Peso_i = \alpha + \beta(vol_i - 758.375)$ onde 758.375 é a média dos volumes, assim o modelo ajustado terá a seguinte equação:

$$Y = 792.81 + 0.7642X$$

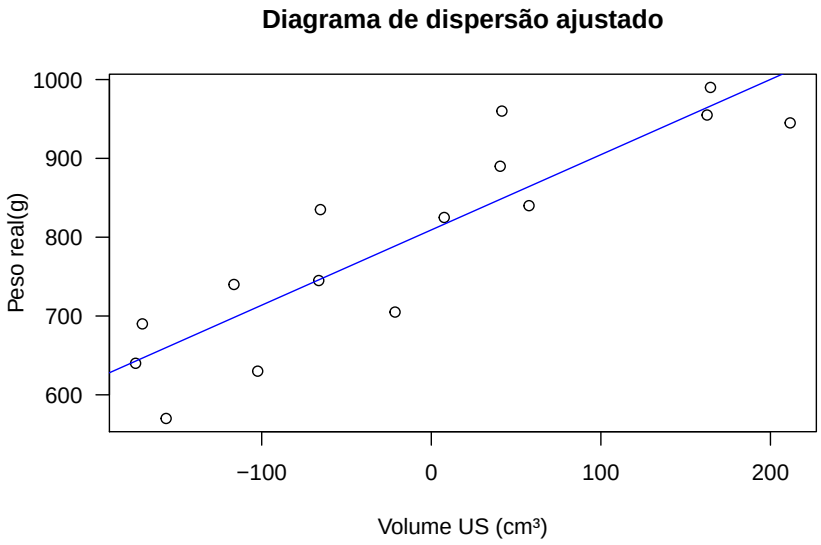
Analogamente ao exercício anterior 0.7642 tem a mesma interpretação, porém o peso de 792.81 gramas é o valor esperado de um figado com volume 758.375 cm^3.

b)



Pode-se perceber que o gráfico de dispersão acima, que quanto maior o valor do volume do fígado em cm^3 encontrado no Ultrasom maior é o valor do peso real em gramas.

c)



Ajustado o modelo obtido no item a), temos $Y = 0.9547X + 809.22$ onde 0.9547 representa o acréscimo esperado no peso real (g) quando o volume previsto em cm³ é incrementado em uma unidade, porém o peso de 809.22 gramas é o valor esperado de um figado com volume 758.375 cm³.

d)
Fazendo uma análise da qualidade do modelo através de medidas descritivas e resíduos temos:

Medidas descritivas dos resíduos				
Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
-89.914	-45.24	-0.841	41.94	111.05

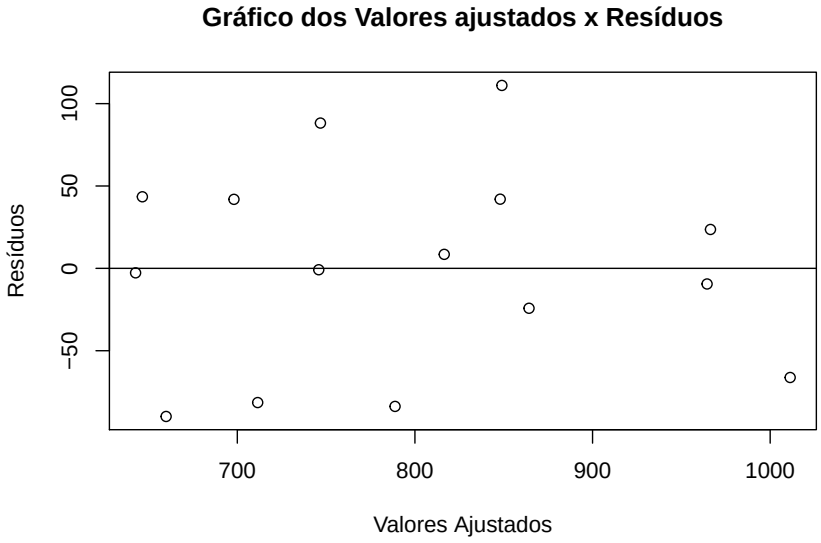
O coeficiente $R^2 = 0.791$

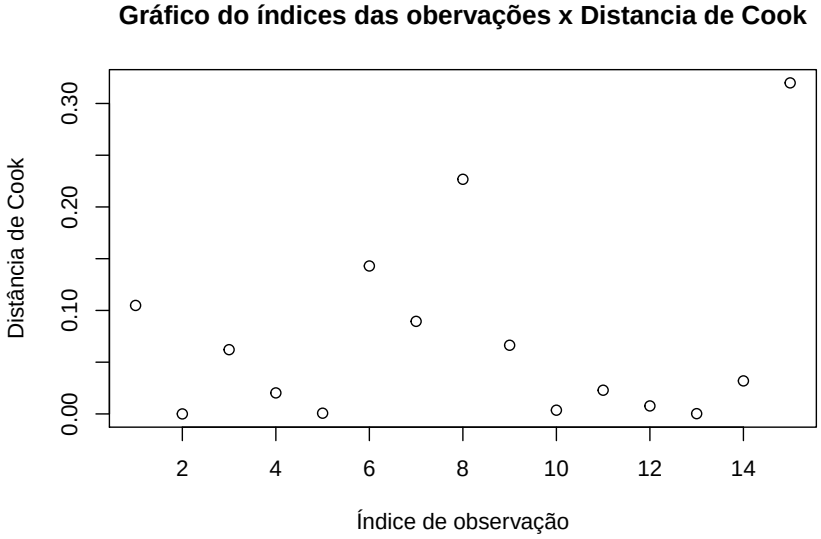
Correlção linear de Pearson = 0.8893361

$$SQTot = \sum_{i=n}^{n} (y_i - \bar{y})^2 = 2.5334333 \times 10^5$$

$$SQReg = \sum_{i=n}^{n} (\widehat{y_i} - \bar{y})^2 = 2.0035625 \times 10^5$$

$$SQRes = \sum_{i=n}^{n} (y_i - \widehat{y_i})^2 = 5.2987084 \times 10^4$$





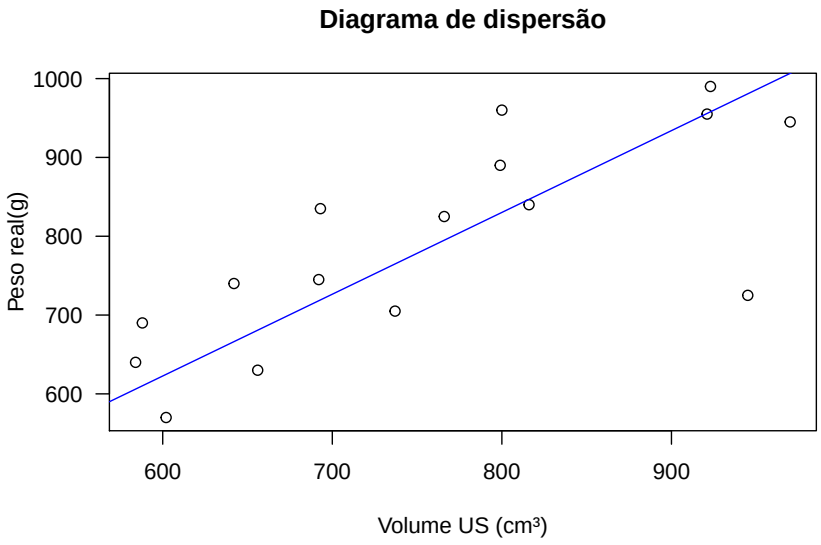
Assim podemos concluir que foi feito um bom ajuste no modelo, pois temos uma correlação e um coeficiente de determinação altos e também uma baixa distância de Cook, com apenas a remoção de um dado muito discrepante.

e)
a') Um modelo de regressão linear simples sem intercepto tem a seguinte equação:

$$Y = 1.038X$$

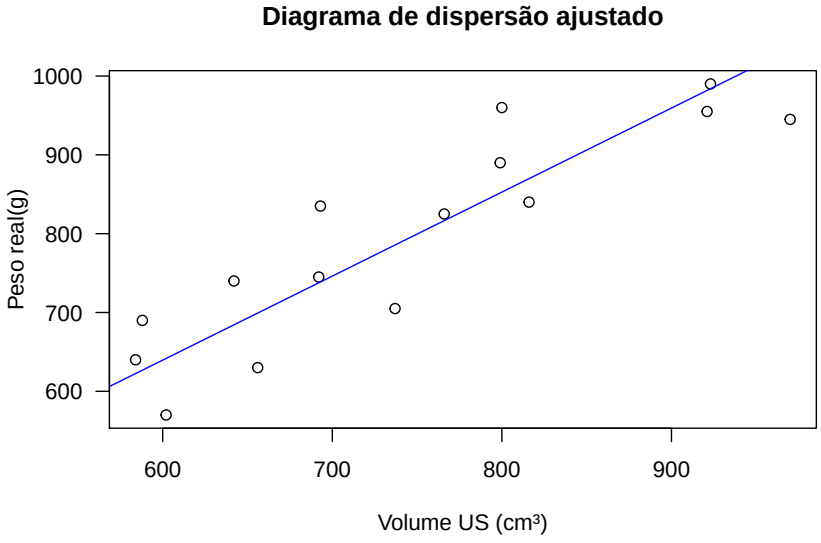
Onde 1.038 representa o acréscimo esperado no peso real (g) quando o volume previsto em cm^3 é incrementado em uma unidade.

b')



Sem o intercepto, pode-se perceber que o gráfico de dispersão acima, que quanto maior o valor do volume do fígado em cm³ encontrado no Ultrasom maior é o valor do peso real em gramas.

c')



Ajustado o modelo obtido no item a'), temos $Y = 1.066X$ onde 1.066 representa o acréscimo esperado no peso real (g) quando o volume previsto em cm³ é incrementado em uma unidade.

d')
Fazendo uma análise da qualidade do modelo através de medidas descritivas e resíduos temos:

Medidas descritivas dos resíduos				
Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
-88.998	-49.559	7.344	46.963	107.218

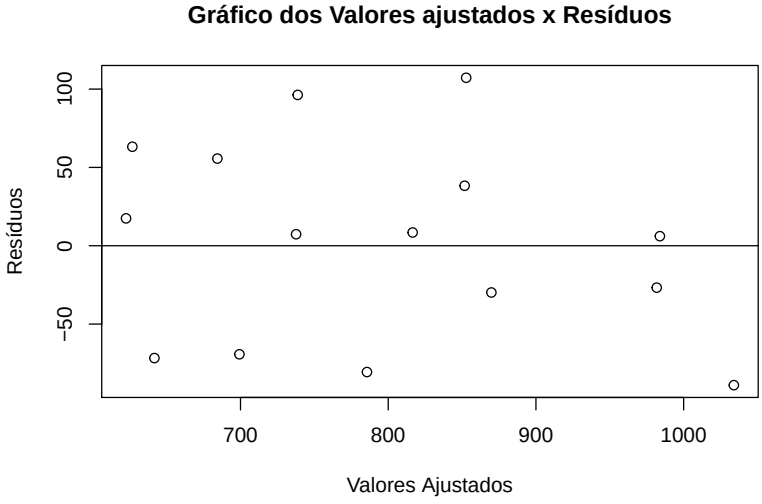
O coeficiente $R^2 = 0.99$

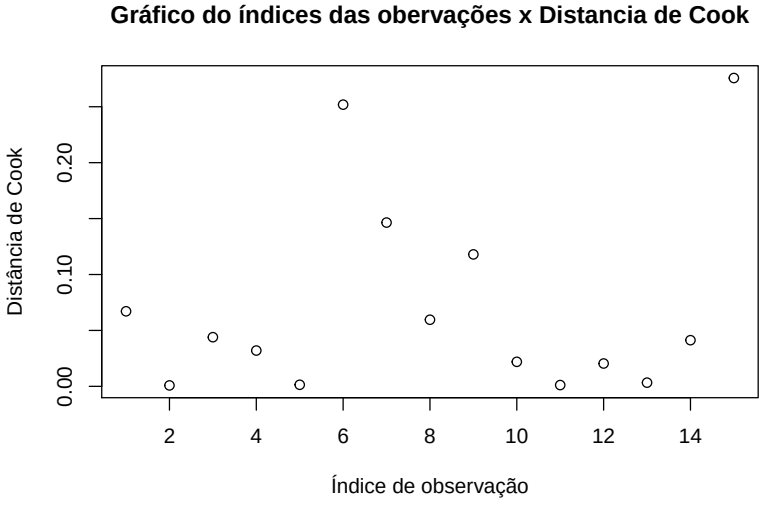
Correlção linear de Pearson = 0.8893361

$$SQTot = \sum_{i=n}^{n} (y_i - \bar{y})^2 = 2.5334333 \times 10^5$$

$$SQReg = \sum_{i=n}^{n} (\widehat{y_i} - \bar{y})^2 = 2.4986537 \times 10^5$$

$$SQRes = \sum_{i=n}^{n} (y_i - \widehat{y_i})^2 = 3477.967436$$



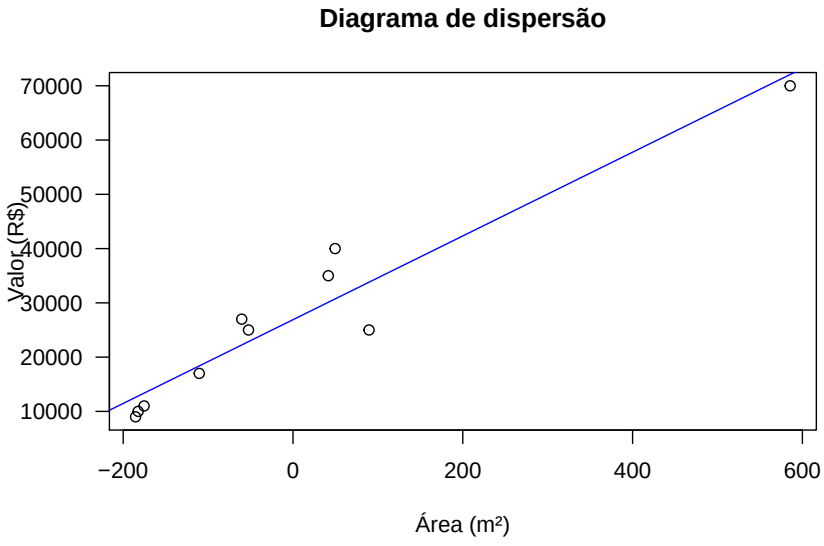


Assim podemos concluir que com o ajuste do modelo e a utilização do mesmo sem intercepto é melhor neste caso, pois temos uma correlação e um coeficiente de determinação altos e também uma baixa distância de Cook, com apenas a remoção de um dado muito discrepante.

Exercício 3

Imóvel	Área (m²)	Valor (R$)
1	128,00	10.000,00
2	125,00	9.000,00
3	200,00	17.000,00
4	4.000,00	200.000,00
5	258,00	25.000,00
6	360,00	40.000,00
7	896,00	70.000,00
8	400,00	25.000,00
9	352,00	35.000,00
10	250,00	27.000,00
11	135,00	11.000,00
12	6.492,00	120.000,00
13	1.040,00	35.000,00
14	3.000,00	300.000,00

a)
Notamos que os imóvel 4,12,13,14, tinham dados de área e valor inconsitentes, para o modelo linear, portanto removemos das observações, assim o Diagrama de Dispersão ajustado:



b)
Um modelo de regressão linear simples ajustado tem a seguinte equação:

$$Y = 77.16X + 26900$$

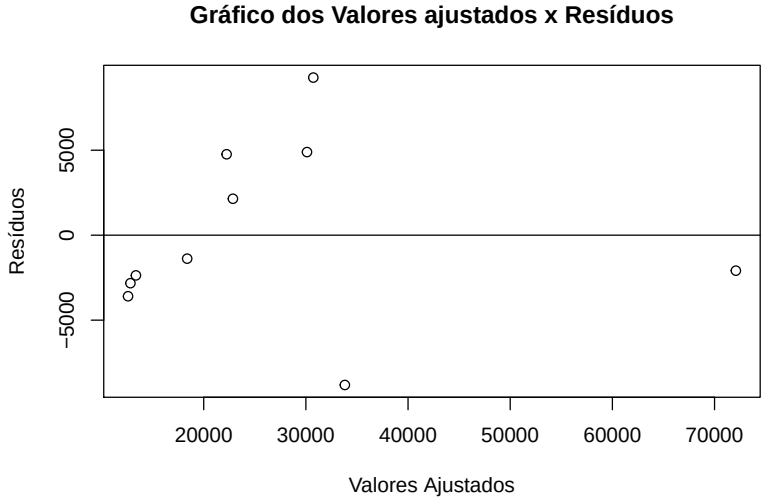
Fazendo uma avaliação da qualidade do modelo, temos:

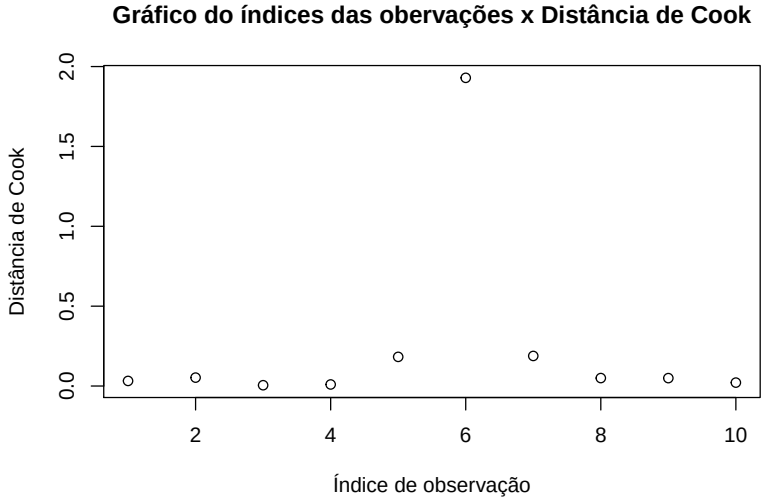
$\widehat{\beta} = 77.16$ Com erro padrão = 8.096

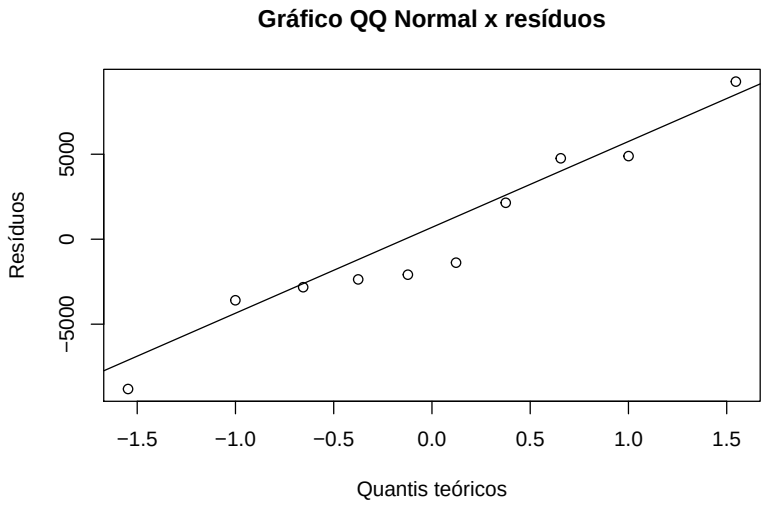
$\widehat{\alpha} = 26900$ Com erro padrão $= 1759.241$

O coeficiente $R^2 = 0.919$

Correlção linear de Pearson $= 0.958675$







Logo podemos concluir que com a remoção dos imóveis 4,12,13,14, o modelo teve uma melhora significativa com um coeficiente de determinação de 0.91, e também vemos que a suposição de normalidade para os erros parece ser adequada.

c)
Fazendo um ajuste agora um modelo linearizável do tipo $Y = \beta x^{\gamma} e$ temos:

Aplicando o ln(logaritmo neperiano) dos dois lados:

$$ln(Y_i) = ln(\beta x_i^{\gamma} e) \Rightarrow ln(Y_i) = ln(\beta) + \gamma ln(x_i) + ln(e_i)$$

Chamando: $ln(Y_i) = Y_i$

$ln(\beta) = \alpha$

$ln(x_i) = x_i$

$ln(e_i) = e_i$

Temos: $Y_i = \alpha + \gamma x_i + e_i$ assim seus estimadores de mínimos quadrados com seus respectivos erros padões são:

$\widehat{\gamma} = 0.7654$ Com erro padrão = 10.306

$\widehat{\alpha} = 5.7158$ Com erro padrão = 8.803

O coeficiente $R^2 = 0.865$

Diagrama de dispersão com dados transformados

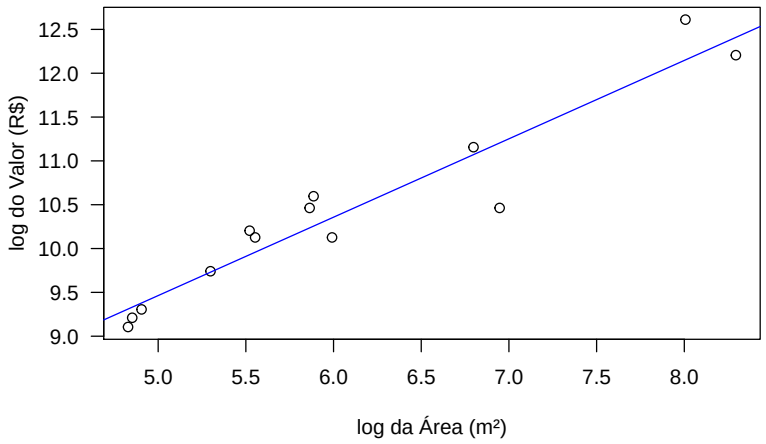


Gráfico dos Valores ajustados x Resíduos

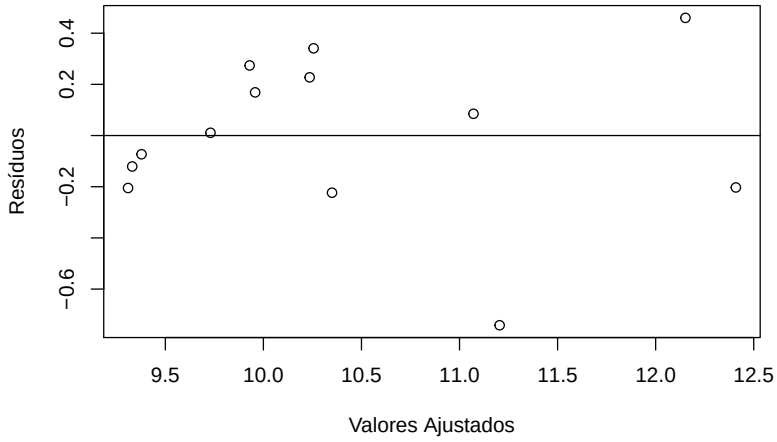
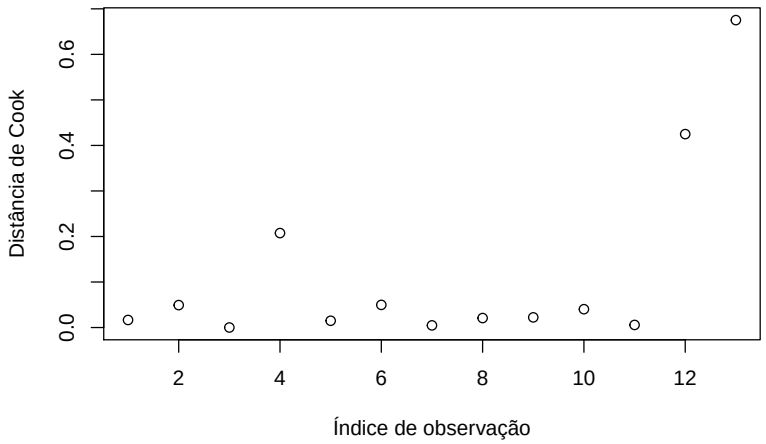
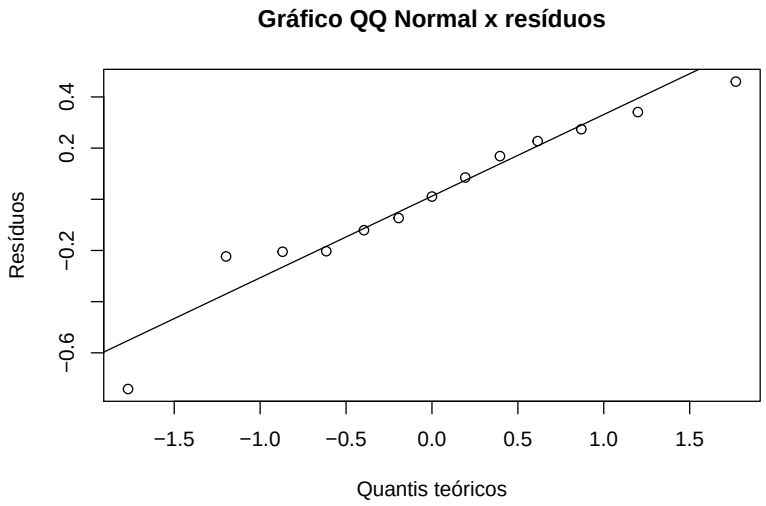


Gráfico do índices das obervações x Distancia de Cook





Comparado ao item anterior, para o modelo ficar com um bom ajuste só foi removido uma única obervação, que no caso foi o item 12, mesmo com um coeficiente de determinação um pouco menor do que o modelo do item a), percebemos que este modelo mais se adequa aos dados, pelo fato de facilitar a visualização sem precisar fazer remoções, e também vemos que a suposição de normalidade para os erros parece ser adequada.

d)
Vantagens do modelo linear: Interpretação intuitiva; adequado para dados com dependência linear.

Desvantagens do modelo linear: Para obervações com diferentes escalas pode deixar o modelo com difícil ajuste.

Vantagens do modelo linearizavél: Oberservações com escalas muito difrentes ficam próximas; ajuste facilitado.

Desvantagens do modelo linearizável: interpretatar o ln das váriaveis; não é intuitivo.