Lista 6 - MAE0328

Guilherme Navarro NºUSP:8943160 Leonardo Noronha NºUSP:9793436

Exercício 1

Considere os dados da Tabela 1 sobre um experimento de evolução de calor na hidratação de cimento (Y), em calorias por grama, de composições distintas no que diz respeito à porcentagem de aluminato de cálcio (x_1), silicato tricálcico (x_2), aluminoferrite tetracálcico (x_3) e silicato dicálcico (x_4). Os dados encontram-se no pacote MPV do software R sob o nome cement.

Tabela 1: Dados de um experimento com cimento conhecidos como *Hald Cement Data*.

Y	x_1	x_2	x_3	x_4
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

(a) Formalize e ajuste um modelo de regressão múltipla para esses dados considerando todas as quatro covariáveis. Calcule os resíduos studentizados modificados $t_i^*,\ i \in \{1, .., 13\}$

Resolução

Considerando o modelo de regressão linear múltipla que consiste em $Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta p x_{pi} + e_i$ onde os e_i são independentes $\forall\ i\ \in \{1, .., n\}$ e $\mathbb{E}(e_i) = 0$ e $Var(e_i) = \sigma^2$.

Assim, nossa reta estimada é : $\widehat{\mu}_i = 62.41 + 1.55x_{1i} + 0.51x_{2i} + 0.10x_{3i} - 0.144x_{4i}$

Os redíduos studentizados modificados $t_i^*,\ i \in \{1, .., 13\}$ foram calculados utilizando a fórmula

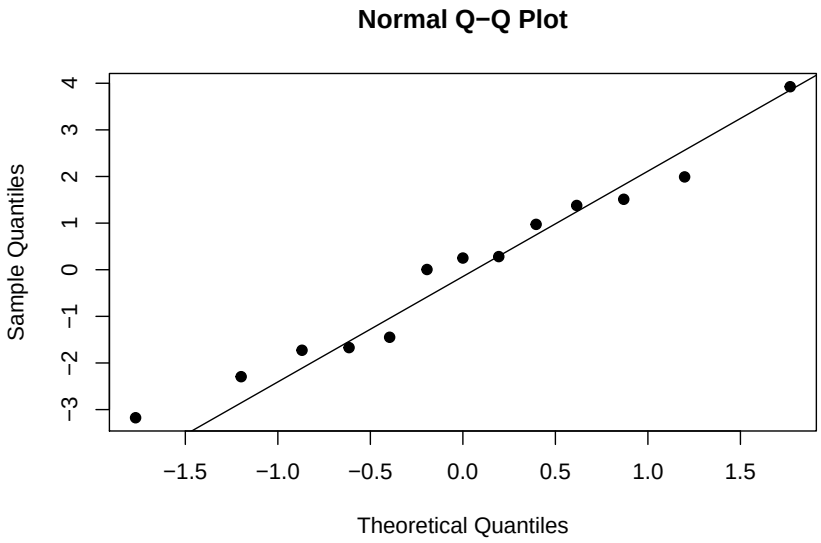
$$t_i^* = \frac{\widehat{e}_i}{\sigma_{(i)}\sqrt{1 - h_{ii}}},\ i \in \{1, .., 13\}$$

e obtimos o seguinte resultado:

```
##           1           2           3           4           5
##  0.002714706  0.734526654 -1.058093203 -0.824036397  0.119767490
##           6           7           8           9          10
##  2.017049821 -0.721820523 -1.967482994  0.645903738  0.197257449
##          11          12          13
##  1.085864774  0.439362041 -1.145888712
```

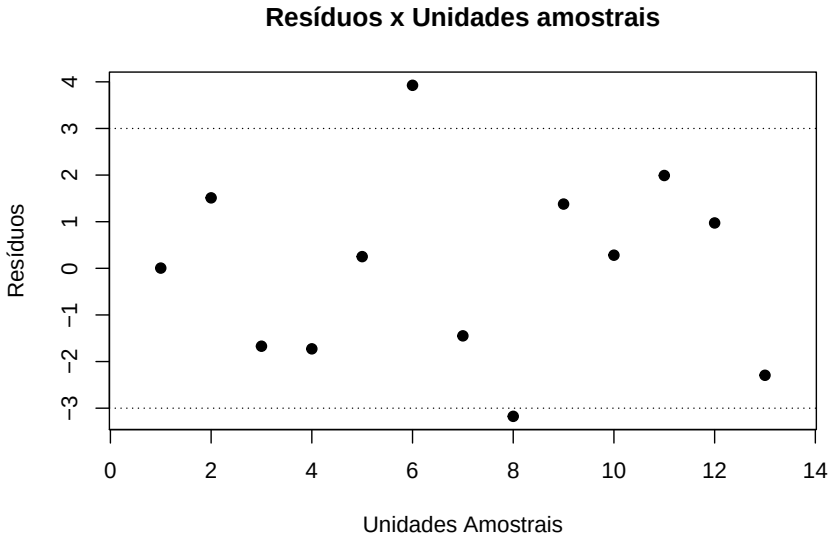
(b) Esboce e comente os seguintes gráficos:

i. de quantil dos resíduos contras os quantis esperados de uma normal padrão;



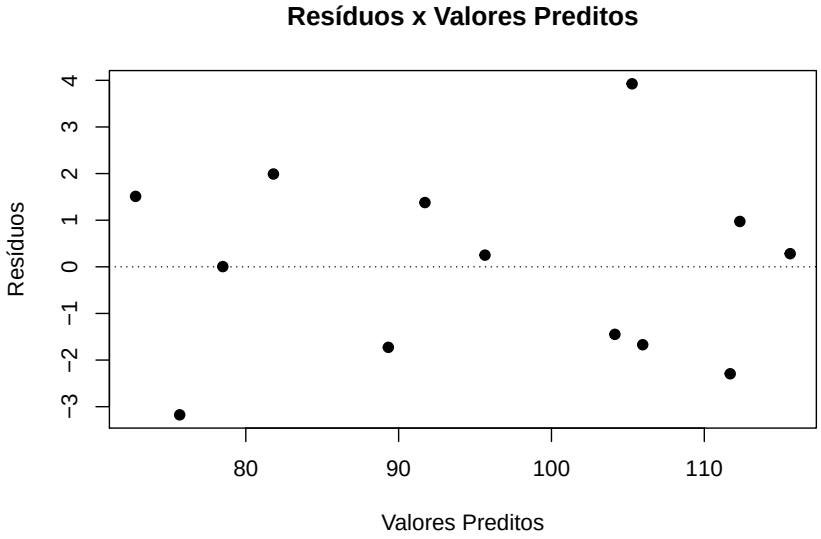
Podemos notar que os pontos estão bem ajustados a reta, logo a suposição de normalidade parece válida.

ii. de resíduos contra os índices das unidades amostrais;



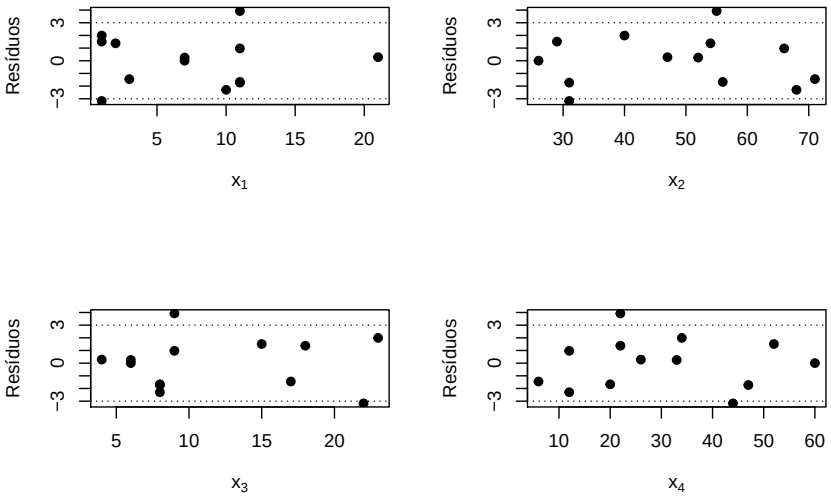
Podemos notar que a maioria das observações estão entre -3 e 3 e dispersas de forma aleatória (sem a presença de alguma linha de tendência)

iii. dos resíduos contra $\widehat{\mu}_i,\ i \in \{1, .., 13\}$;



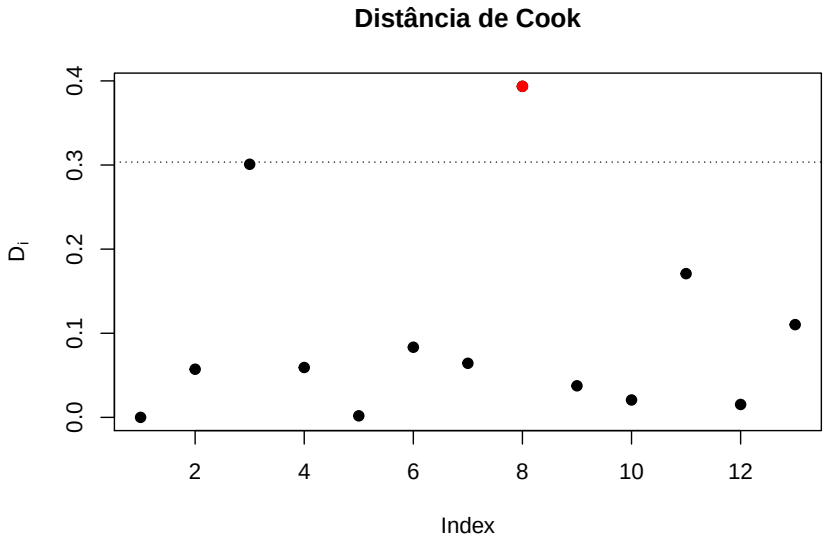
Podemos notar que a distribuição dos pontos é aleatória logo não temos problemas com heterocedasticidade.

iv. de resíduos contra cada uma das covariáveis;



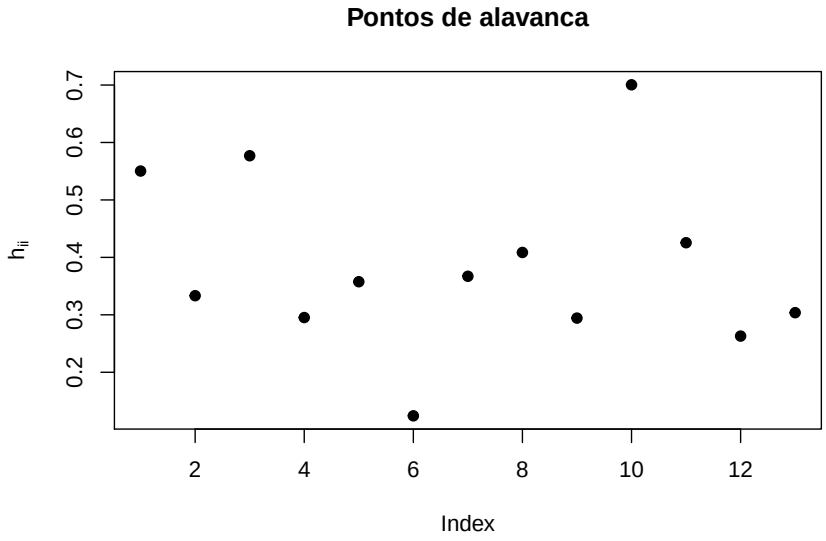
Podemos notar que há uma evidência de relação equivocada ou falta de alguma covariável para explicar o modelo.

v. das distâncias de Cook;



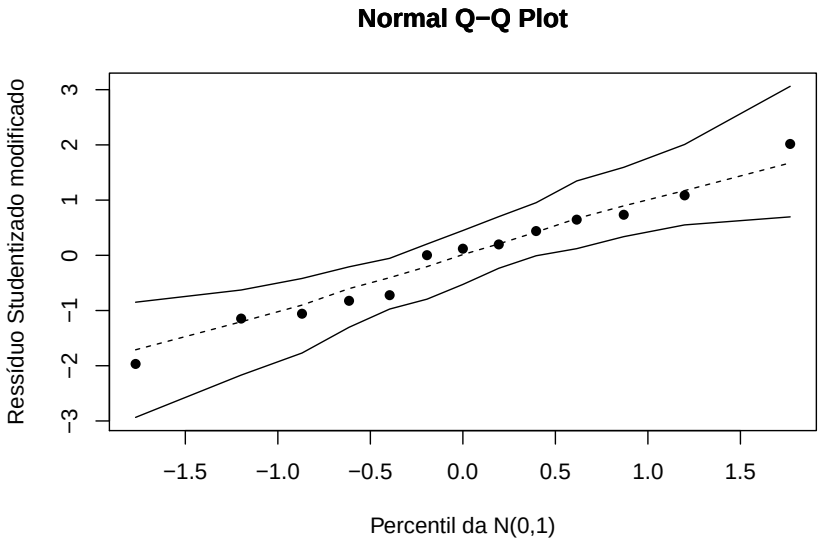
Com o gráfico da distância de cook, podemos notar que temos somente um ponto que está acima do limite $3 * \bar{D}$.

vi. de alavancagem;



Podemos notar que não há pontos de alavacagem, pois o maior ponto de alavancagem é 0.7 e o limite que define pontos de alavancagem é de 0.769.

vii. de envelope simulado;



Podemos notar que com o envelope simulado, que os dados que estão no qqplot do item i. confirmam a hipótese de normalidade para os erros do modelo.

(c) Calcule os fatores de inflação da variância FIV_j para cada covariável e analise se existe um problema de multicolinearidade

Resolução

Utilizando a fórmula

$$FIV_j = \frac{1}{1 - R_j^2}$$

Obtivemos:

$FIV_1 = 38.49$

$FIV_2 = 254.423$

$FIV_3 = 46.868$

$FIV_4 = 282.512$

Geralmente, o VIF é indicativo de problemas de multicolinearidade se $VIF > 10$ como todos os valores estão maiores que 10 é recomendável remover as covárias que apresentam maior VIF.

(d) Remova a(s) covariável(is) que apresentaram multicolinearidade, ajuste novamente o modelo e refaça o item (a). Verifique se o ajuste melhorou, justificando sua resposta.

Resolução

Removendo a covariável que apresentou o maior VIF (no caso x_4) temos que o novo modelo é do tipo $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta 3 x_{3i} + e_i$ onde os e_i são independentes $\forall\ i\ \in \{1, .., 13\}$ e $\mathbb{E}(e_i) = 0$ e $Var(e_i) = \sigma^2$.

Assim, nossa reta estimada é : $\widehat{\mu}_i = 48.19 + 1.7x_{1i} + 0.66x_{2i} + 0.25x_{3i}$

Recalculando o VIF_j para no novo modelo obtivemos:

$FIV_1 = 3.25$

$FIV_2 = 1.06$

$FIV_3 = 3.14$

Como todos os valores estão abaixo de 10, o problema de multicolinearidade foi resolvido.

Os resíduos para o novo modelo são:

```
##           1           2           3           4           5           6
## -0.08034362  0.81280089 -0.59406662 -0.78418247  0.08775509  2.17070343
##           7           8           9          10          11          12
## -0.78164896 -2.10668141  0.71919406  0.15805792  1.03793694  0.38806688
##          13
## -1.23387005
```

(e) Considerando novamente todas as covariáveis da Tabela 1, apresente uma implementação computacional do algoritmo de seleção backward. Faça a seleção de um modelo utilizando seu algoritmo com $q_s = 0.05$. Compare esse com o modelo obtido no item (d).

Resolução

Exercício 2

Um experimento para avaliar o crescimento de bactérias *E. coli* foi conduzido da seguinte maneira: 1) foram utilizados 30 recipicientes com o mesmo número de bactérias da mesma espécie; 2) o i-ésimo recipiente foi exposto à temperatura de 70ºF pelo tempo de x_i minutos; 3) ao final do tempo x_i, o número de bactérias resultante foi contado (Y_i). A Tabela 2 apresenta o número de bactérias e os respectivos tempos de exposição.

Tabela 2: Número de bactérias (Y) e tempo (x), em minutos

i	Y_i	x_i	i	Y_i	x_i
1	20	1	16	150	26
2	60	3	17	80	28
3	20	4	18	130	30
4	60	6	19	60	31
5	30	8	20	30	33
6	30	9	21	150	35
7	10	11	22	80	36
8	60	13	23	100	38
9	50	15	24	200	40
10	80	16	25	120	42
11	40	18	26	100	43
12	90	20	27	110	45
13	140	21	28	80	47
14	40	23	29	160	48
15	100	25	30	150	50

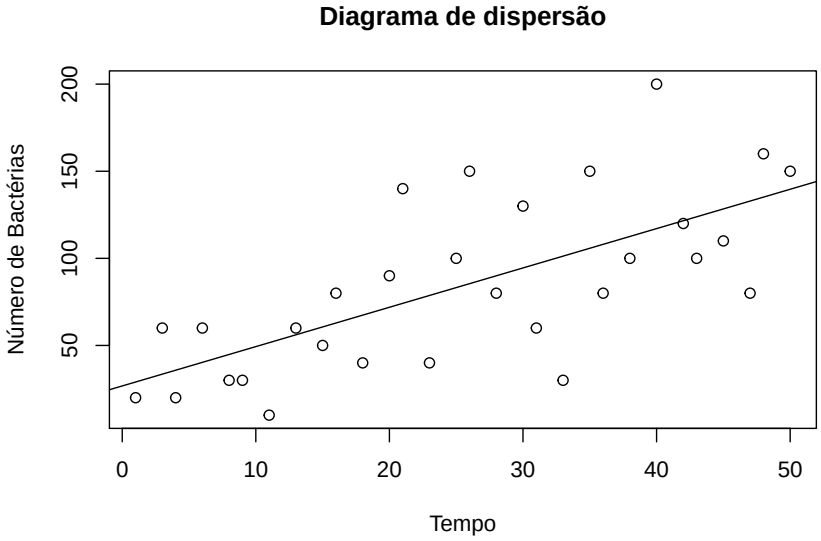
(a) Ajuste o modelo linear $Y_i = \beta_0 + \beta_1 x_i + e_i$, em que $e_i \overset{iid}{\sim} N(0, \sigma^2),\ i \in \{1, ..., 30\}$

Resolução

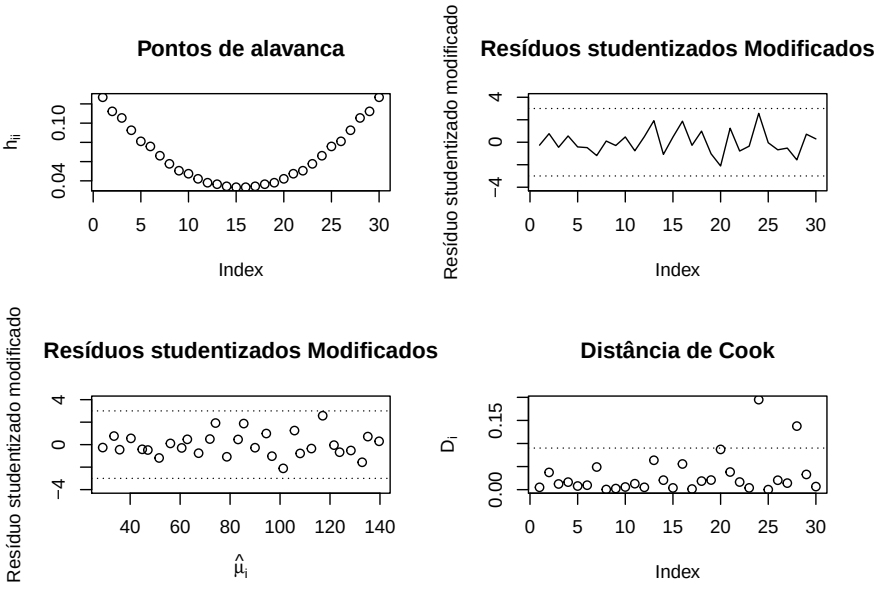
O modelo ajustado tem equação: $\widehat{\mu}_i = 26.76 + 2.26 x_i$

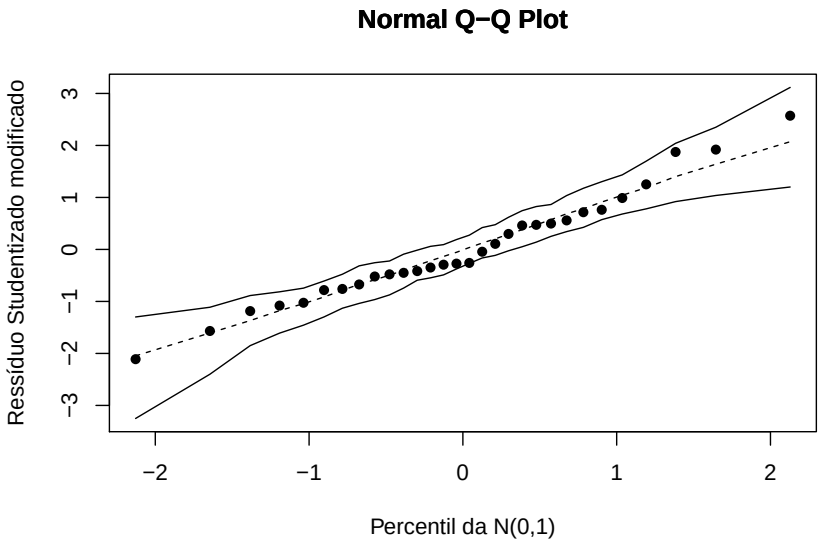
(b) Apresente o gráfico de dispersão com a reta estimada. Faça uma análise de resíduos e verifique se há indícios de desvio das suposições do modelo.

Resolução



Fazendo uma análise de resíduos, temos:

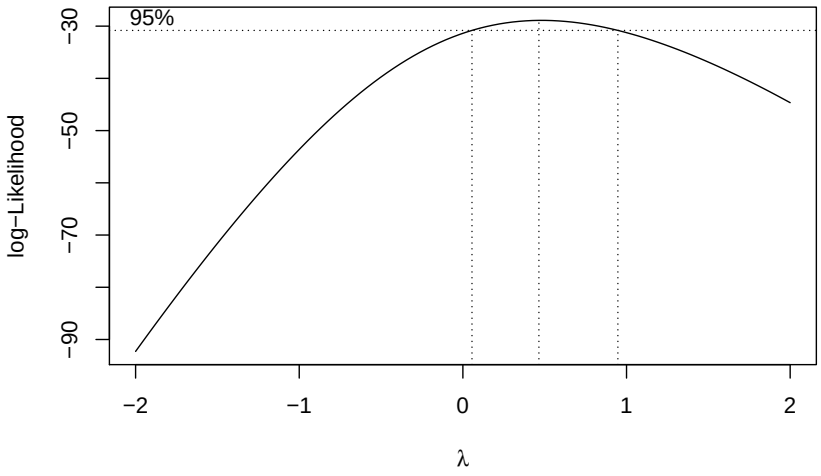




Podemos notar que há uma pequena tendência de a variância aumentar conforme o valor predito aumenta, sendo possível então uma transformação de Box-Cox para aumentar a eficiência do modelo.

(c) Aplique a transformação de Box-Cox, reajuste o modelo e refaça a análise de resíduos. Verifique se a transformação raiz quadrada pode ser utilizada.

Resolução



Oberservando o intervalo de confiança obtido pelo método de Box-Cox, podemos concluir que o melhor $\lambda = 0.5$, agora refazendo o diagrama de dispersão e a análise de resíduos novamente com os dados transformados, temos:

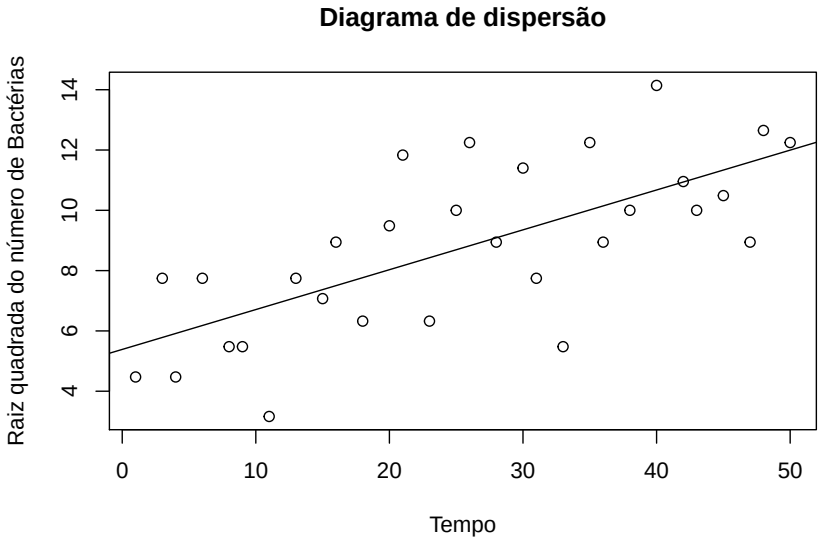
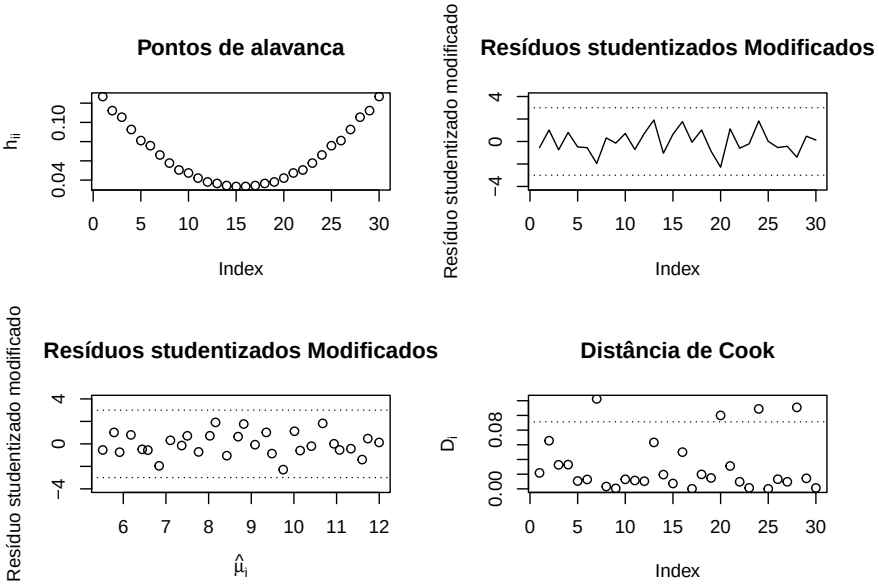
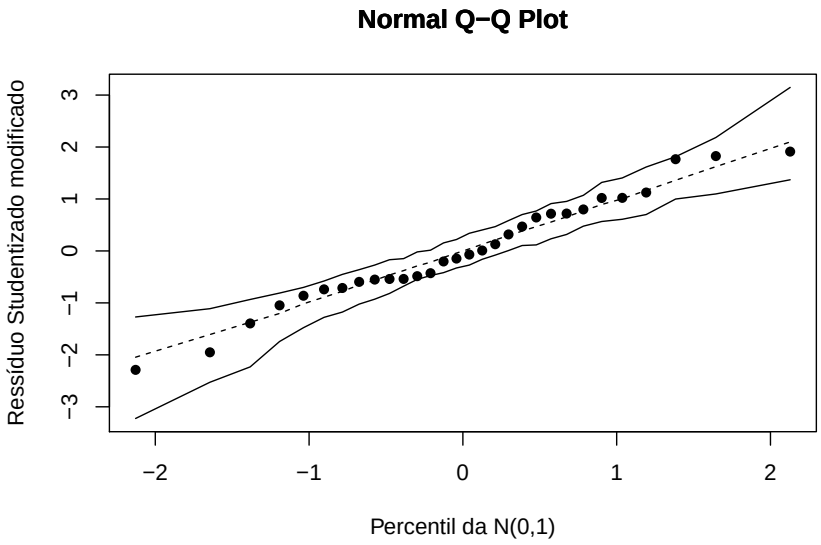


Diagrama de dispersão
Raiz quadrada do número de Bactérias
Tempo


Pontos de alavanca
Resíduos studentizados Modificados
Resíduo studentizado modificado
Index
Resíduos studentizados Modificados
Distância de Cook



Podemos notar que a curva no diagrama de disperção e o envelope simulado são mais adequados aos dados que os da análise inicial.

Exercício 5

Para os dados da Tabela 3, suponha um modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 x_i + e_i$, em que $e_i \overset{iid}{\sim} N(0, \sigma^2),\ i \in \{1, ..., 20\}$

Tabela 3: Variável resposta (Y_i) e covariável x_i.

i	Y_i	x_i	i	Y_i	x_i
1	28,29	6	11	39,45	8
2	45,58	10	12	36,90	8
3	33,49	7	13	42,07	8
4	40,27	7	14	31,83	8
5	49,56	9	15	55,11	11
6	40,00	15	16	48,29	10
7	21,83	5	17	20,58	5
8	33,96	7	18	43,49	9
9	57,07	9	19	60,27	11
10	49,90	9	20	34,56	6

(a) Apresente as estimativas de máxima verossimilhança dos parâmetros β_0, β_1 e σ^2 e, ao nível de 5% de significância, teste a hipótese $H_0 : \beta_1 = 5$ contra $H_1 : \beta_1 \neq 5$

Resolução

As estimativas de máxima verossimilhança para os parâmetros β_0, β_1 e σ^2 são $\begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{pmatrix} = \begin{pmatrix} 14.315 \\ 3.132 \end{pmatrix}$ e a estimativa de máxima verossimilhança não viciada para σ^2 é $\widehat{\sigma}^2 = 68.331$

Agora o teste, $H_0 : \beta_1 = 5$ contra $H_1 : \beta_1 \neq 5$ tomando a estatística $F_{\widehat{\beta},C,d} = \frac{(C\widehat{\beta}-d)^T[C(X^TX)^{-1}C^T]^{-1}(C\widehat{\beta}-d)}{k\widehat{\sigma^2}}$

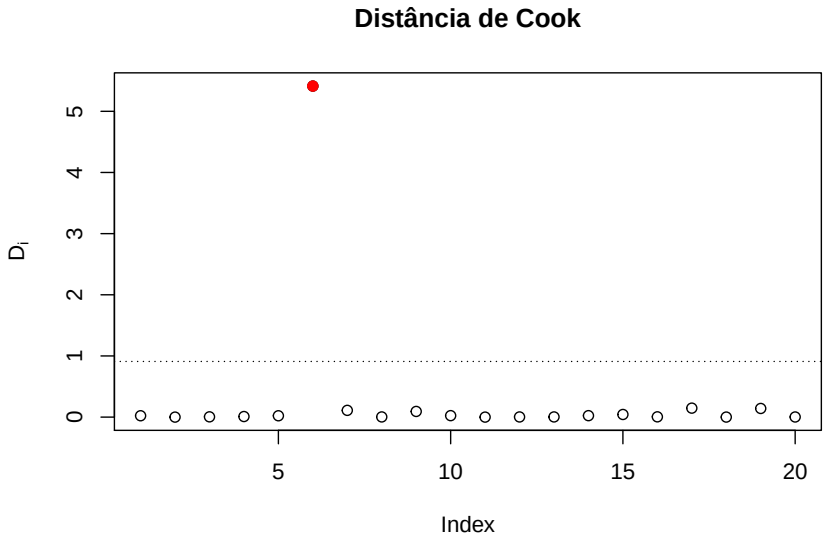
Tomando a matriz $C = (0\ 1)$ e $d = 5$ e $k = 1$, obtendo

$$F^{obs}_{\widehat{\beta},C,d} = 5.35$$

e seu valor-p= $\mathbb{P}(F_{k,n-2} > F^{obs}_{\widehat{\beta},C,d}) = \mathbb{P}(F_{1,18} > F^{obs}_{\widehat{\beta},C,c}) = 0.032$. Logo temos evidencias estatísticas para rejeitar a hipótese nula a um nível de significância de 5%.

(b) Calcule a distância de Cook para cada observação. Usando critério apropriado, discuta se há observações possivelmente influentes nas estimativas do modelo de regressão.

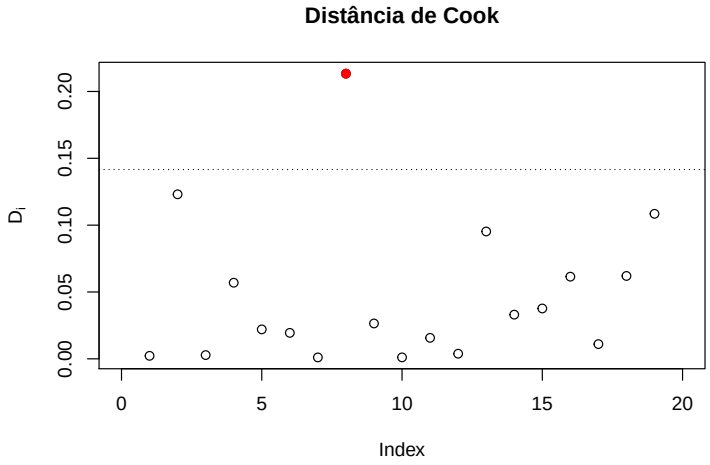
Resolução



Podemos observar pelo gráfico, pelo critério de que um ponto é influente de $D_i > 3 * \bar{D}$ onde $\bar{D} = 0.3$, logo temos apenas um ponto influente é o ponto de índice 6 onde $D_6 = 5.41$.

(c) Removendo as observações potencialmente influentes, refaça o item (a). Suas conclusões inferencias foram diferentes? Argumente contra ou a favor do modelo pressuposto para esse conjunto de dados.

Resolução



Removendo a observação potencialmente influente e refazendo o teste do item a, as estimativas de máxima verossimilhança para os parâmetros β_0, β_1 e σ^2 são $\begin{pmatrix}\widehat{\beta_0}\\ \widehat{\beta_1}\end{pmatrix} = \begin{pmatrix}-4.776\\ 5.642\end{pmatrix}$ e a estimativa de máxima verossimilhança não viciada para σ^2 é $\widehat{\sigma}^2 = 22.419$

Agora o teste, $H_0 : \beta_1 = 5$ contra $H_1 : \beta_1 \neq 5$ tomando a estatística $F_{\widehat{\beta},C,d} = \frac{(C\widehat{\beta}-d)^T[C(X^TX)^{-1}C^T]^{-1}(C\widehat{\beta}-d)}{k\widehat{\sigma^2}}$

Tomando a matriz $C = (0\ 1)$ e $d = 5$ e $k = 1$, obtendo

$$F^{obs}_{\widehat{\beta},C,d} = 1.08$$

e seu valor-p= $\mathbb{P}(F_{k,n-2} > F^{obs}_{\widehat{\beta},C,d}) = \mathbb{P}(F_{1,17} > F^{obs}_{\widehat{\beta},C,c}) = 0.312$. Logo temos evidencias estatísticas para não rejeitar a hipótese nula a um nível de significância de 5%, podemos então utilizar o valor 5 para o β_1. Podemos realizar o gráfico de dispersão com a curva do modelo após a remoção da observação potencialmente influente, em preto, e a curva sob H_0, com $\beta_1 = 5$, em vermelho. Desse modo podemos ver que a reta preta se ajusta melhor aos dados que a utilizada sob H_0, portanto ainda assim seria melhor utilizar o modelo original.