Lista 2 MLG

Davi Wentrick Feijó

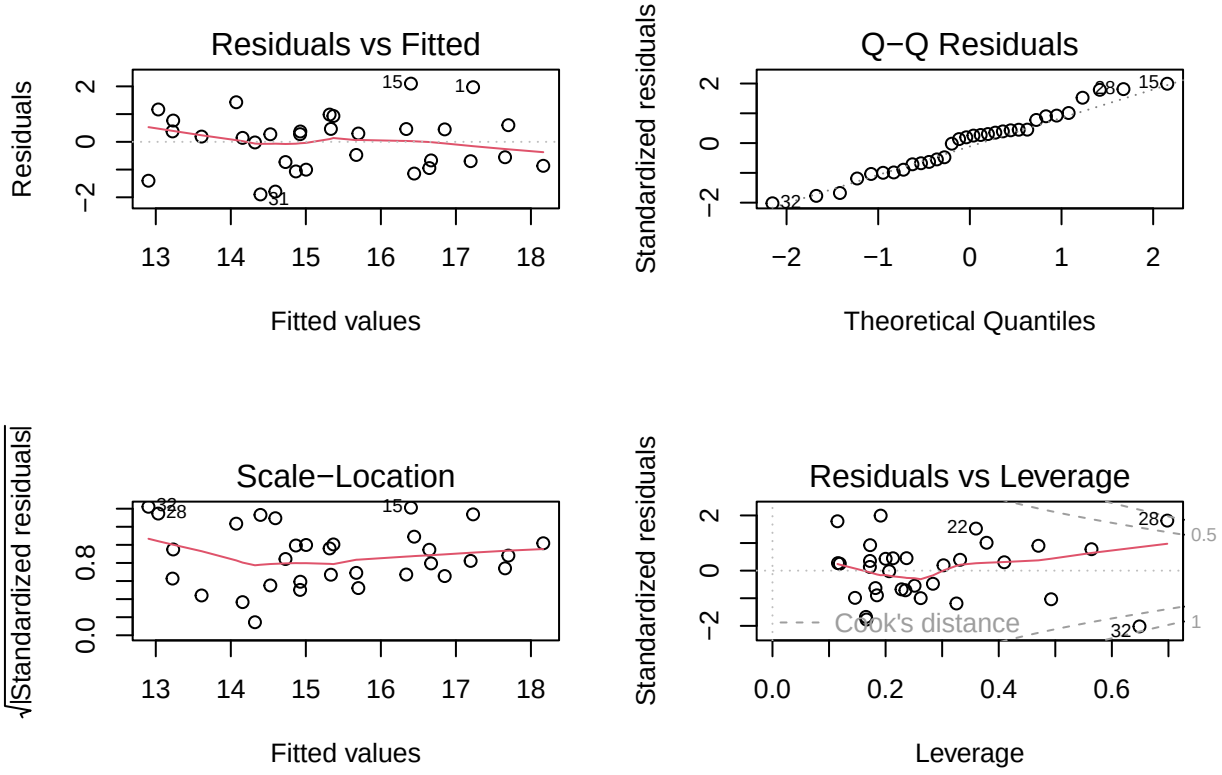
2023-10-02

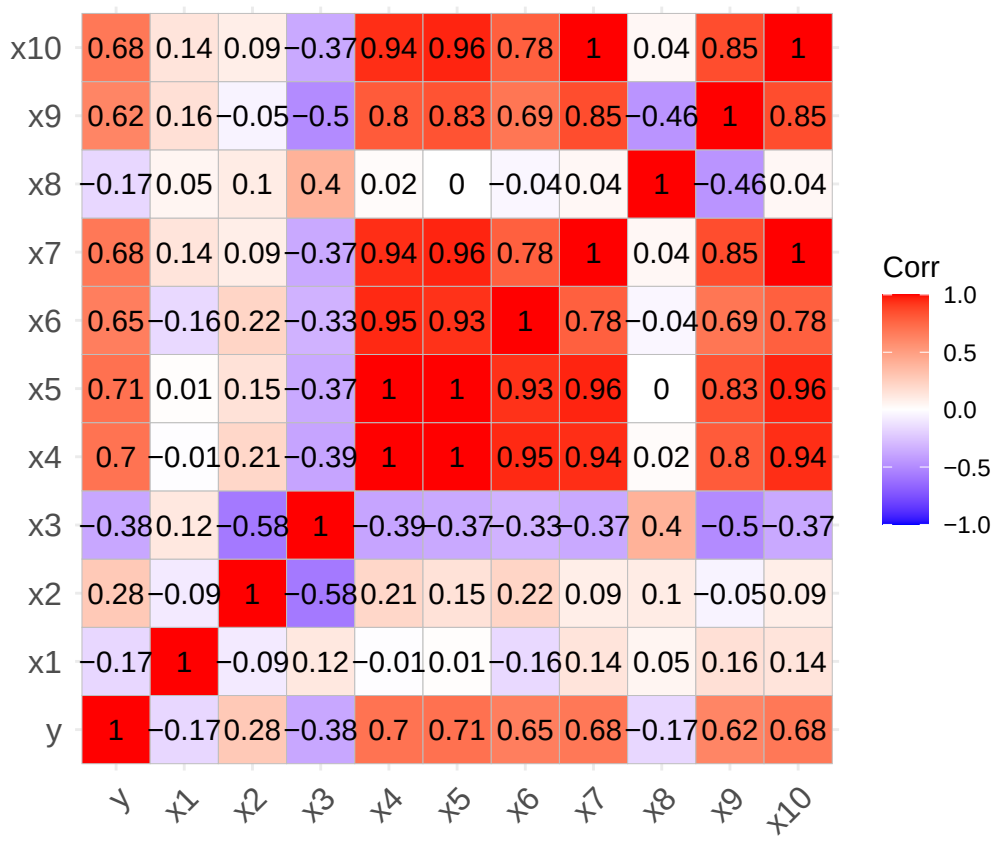
Q1. Considere os dados sobre a qualidade do vinho tinto, apresentados no ficheiro `Q01-data.txt`. Ajuste o modelo de regressao linear multipla, e faca uma analise completa desses dados. Que conclusoes voce tira dessa analise? (use 5% de significancia durantes as analises)

Vamos fazer uma analise do modelo usando todo o banco para ver os resultados gerais.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10, data = Q01_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8952 -0.7626  0.2315  0.4999  2.0991
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.20843   14.61153  -0.836   0.4120
## x1           -0.84577    0.58596  -1.443   0.1624
## x2            7.41839    3.51235   2.112   0.0457 *
## x3            0.01046    0.00857   1.220   0.2347
## x4           -1.94732    2.22110  -0.877   0.3897
## x5            4.89518    3.21850   1.521   0.1419
## x6           -1.43382    1.81263  -0.791   0.4370
## x7                 NA         NA      NA       NA
## x8          -11.42517    7.88120  -1.450   0.1606
## x9           -0.10802    0.22040  -0.490   0.6287
## x10                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.171 on 23 degrees of freedom
## Multiple R-squared:  0.6753, Adjusted R-squared:  0.5624
## F-statistic:  5.98 on 8 and 23 DF,  p-value: 0.0003399
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1  2.805   2.805  2.0444   0.16621
## x2         1  6.745   6.745  4.9159   0.03677 *
## x3         1  6.223   6.223  4.5352   0.04413 *
## x4         1 36.140  36.140 26.3393 3.363e-05 ***
## x5         1  7.928   7.928  5.7784   0.02468 *
## x6         1  0.279   0.279  0.2037   0.65598
## x8         1  5.192   5.192  3.7842   0.06406 .
## x9         1  0.330   0.330  0.2402   0.62868
## Residuals 23 31.558   1.372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





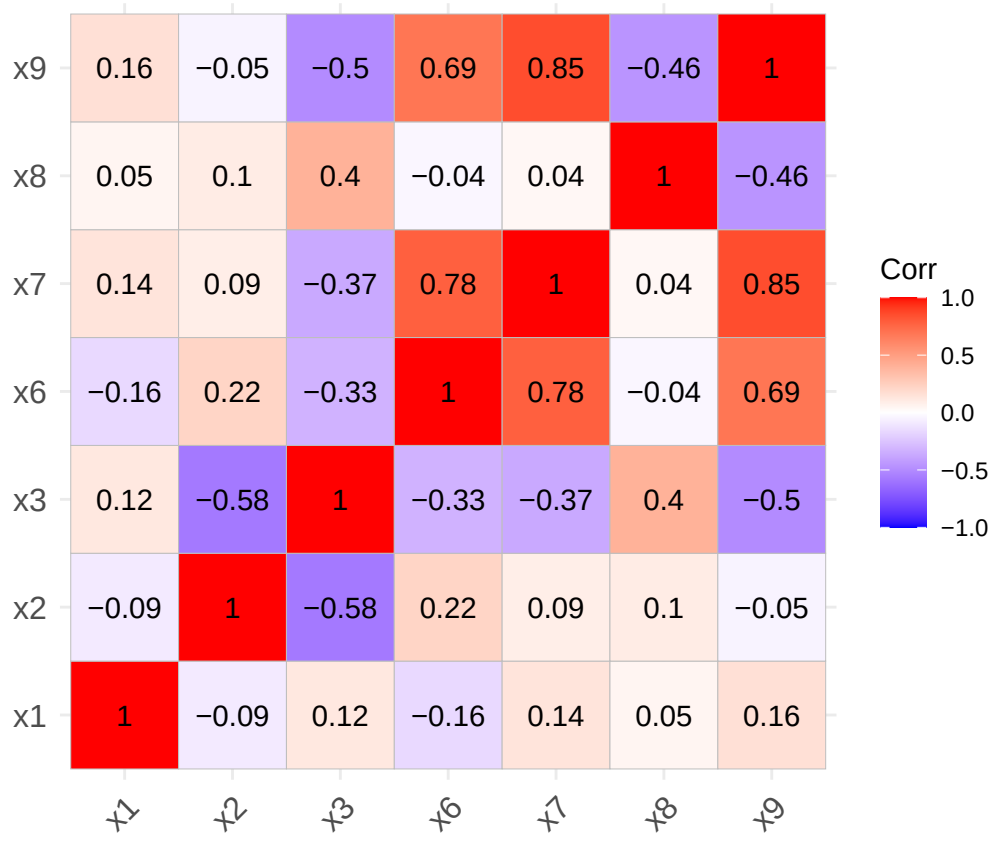
Com base no conjunto de dados, e possıvel observar que algumas covariaveis estao fortemente correlacionadas. Assim, pede-se:

(a) Proponha algum metodo para resolver o problema da multicolinearidade no conjunto de dados. Podemos perceber que temos um problema incomum de algumas variaveis terem correlaçao perfeita, isso definitivamente atrapalha o modelo e dificuldat qualquer metodo de selecao de variaveis, especialemnte dentro do R. Logo é necessario removermos essas variaveis problematicas. Em seguida podemos rodar um modelo de selecao de variaveis. Nesse caso vamos retirar as variaves x4, x5 e x10 que possuem multiplas correlaçoes altas com outras variaveis

```
modelo1_completo = lm(y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9 , data = Q01_data)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1  2.805   2.805  2.0444   0.16621
## x2         1  6.745   6.745  4.9159   0.03677 *
## x3         1  6.223   6.223  4.5352   0.04413 *
## x4         1 36.140  36.140 26.3393 3.363e-05 ***
## x6         1  1.674   1.674  1.2199   0.28081
## x7         1  6.534   6.534  4.7622   0.03956 *
## x8         1  5.192   5.192  3.7842   0.06406 .
## x9         1  0.330   0.330  0.2402   0.62868
## Residuals 23 31.558   1.372
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



(b) Usando algum metodo de selecao de variaveis, obtenha o modelo final para o conjunto de dados.

	Intercepto	X1	X2	X3	X4	X6	X7	X8	X9
Forward	-12.2084	-0.8458	7.4184	0.0105	-1.9473	3.4614	4.8952	-11.4252	-0.108
Backward	-10.1055	-0.9628	6.2506	0.0123	NA	NA	1.6429	-8.7781	NA
Both	-10.1055	-0.9628	6.2506	0.0123	NA	NA	1.6429	-8.7781	NA

Podemos ver que o metodo backward e both deram os mesmos resultados, contudo se escolhermos o modelo com base no R^2 ajustado veremos que o modelo do forward tem um resultado um pouco menor, logo vamos com utilizar o modelo encontrado pelos outros metodos.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x7 + x8, data = Q01_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84035 -0.81723  0.06688  0.55012  2.32578
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -10.105468   9.318344  -1.084   0.2881
## x1           -0.962779   0.414668  -2.322   0.0283 *
## x2            6.250592   2.419073   2.584   0.0157 *
## x3            0.012292   0.006793   1.810   0.0819 .
## x7            1.642902   0.268359   6.122  1.8e-06 ***
## x8           -8.778121   3.504817  -2.505   0.0189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.131 on 26 degrees of freedom
## Multiple R-squared:  0.658,  Adjusted R-squared:  0.5923
## F-statistic: 10.01 on 5 and 26 DF,  p-value: 1.994e-05
```

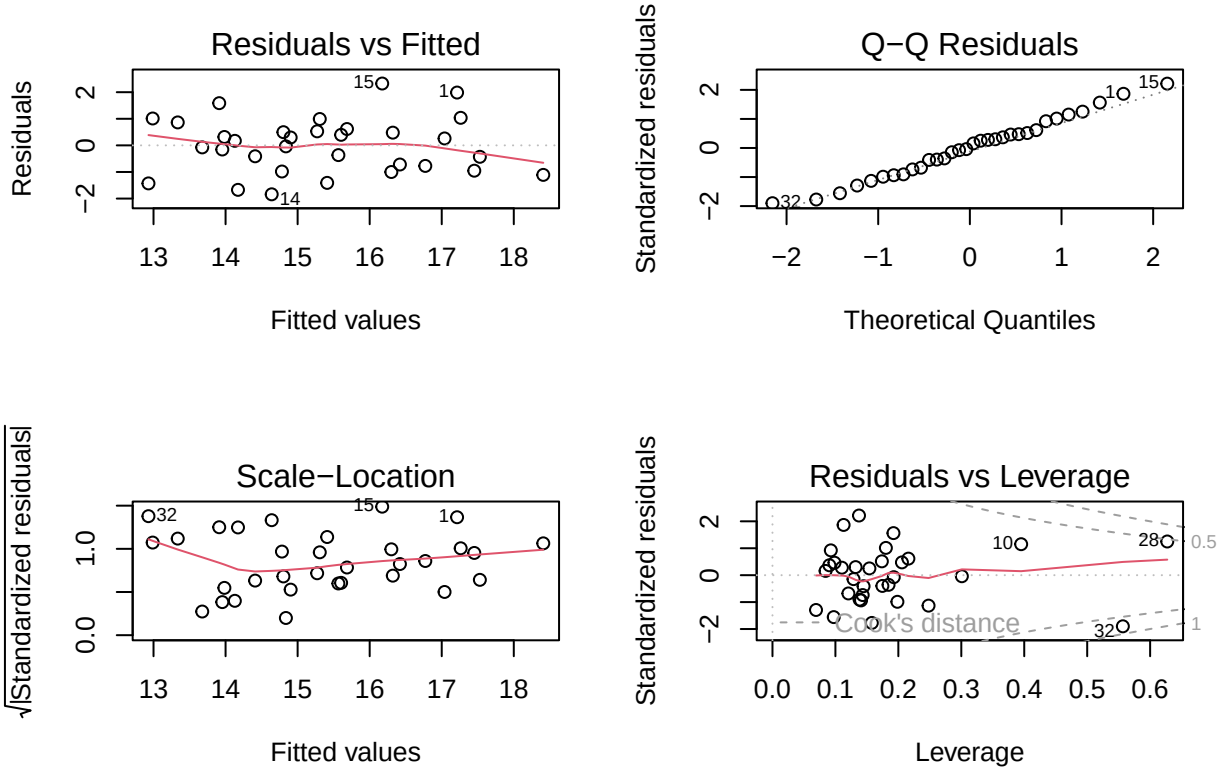
(c) Apresente a tabela da Analise de Variancia para testar a significancia global dos coeficientes do modelo final. Apresente as hip´otese de teste, e conclua.

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1  2.805   2.805  2.1942   0.15055
## x2         1  6.745   6.745  5.2762   0.02992 *
## x3         1  6.223   6.223  4.8676   0.03640 *
## x7         1 40.170  40.170 31.4225 6.851e-06 ***
## x8         1  8.019   8.019  6.2730   0.01886 *
## Residuals 26 33.238   1.278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fonte	SSQ	GL	QM	F	P
Regressao	63.96203	5	12.792405	10.0067	1.99e-05
Erro	33.23797	26	1.278384	NA	NA
Total	97.20000	31	NA	NA	NA

O resultado indica que é signigficante e que todos os coeficientes sao diferentes de 0.

(d) Com base no modelo obtido no item anterior, faca uma analise de resıduos e conclua.




```
##
##  Shapiro-Wilk normality test
##
## data:  residuo
## W = 0.9805, p-value = 0.8139

##
##  studentized Breusch-Pagan test
##
## data:  both_mdl
## BP = 2.0947, df = 5, p-value = 0.8359
```

Podemos perceber que o modelo passa nos testes de normalidade e variancida dos residuos.

Q02. Uma equipe de pesquisadores de saude mental deseja comparar tres metodos de tratamento da depressao grave (A, B e C=referencia). Eles tambem gostariam de estudar a relacao entre idade e eficacia do tratamento, bem como a interacao (se houver) entre idade e tratamento. Cada elemento da amostra aleatoria simples de 36 pacientes, foi selecionado aleatoriamente para receber o tratamento A, B ou C. Os dados obtidos podem ser encontrados no ficheiro `Q02-data.txt`. A variavel dependente y e a eficacia do tratamento; as variaveis independentes sao: a idade do paciente no aniversario mais proximo, e o tipo de tratamento administrado (use 1% de significancia durantes as analises).

(a) Ajuste o modelo de regressao linear e interprete os resultados obtidos.

(b) Obtenha a tabela ANOVA para o modelo obtido no item (a) e interprete os resultados.

(c) Considere a possibilidade de incluir a interacao entre as varaveis independentes, i.e., assuma que o modelo a ser ajustado tem a seguinte formulacao: $yi = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \epsilon_i$ **, com a suposicao de que** $\epsilon_i \sim N(0, \sigma^2)$**. Com base no modelo anterior,**

- (i) Liste todos os possıveis submodelos que podem ser obtidos usando o modelo apresentado anteriormente.
- (ii) Interprete os coeficientes de regressao associados aos fatores de interacao.
- (iii) Apresente a tabela anova para testar as seguintes hipoteses, $H0 : \beta_1 = \beta_4 = \beta_5 = 0$ contra $H1 : \exists \beta_j 6 = 0$, com $j = 1, 4, 5$.
- (iv) Faca uma analise completa dos res´ıduos do modelo.