

Lista 2 MLG

Davi Wentrick Feijó

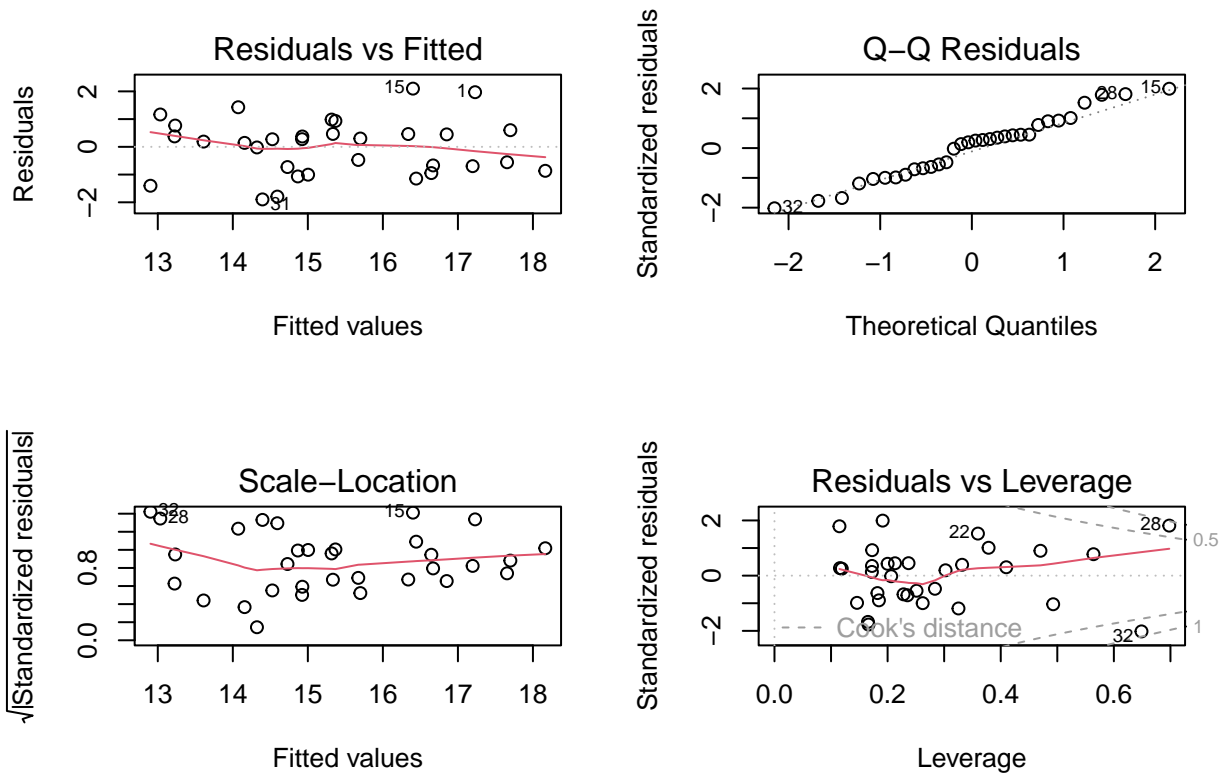
2023-10-02

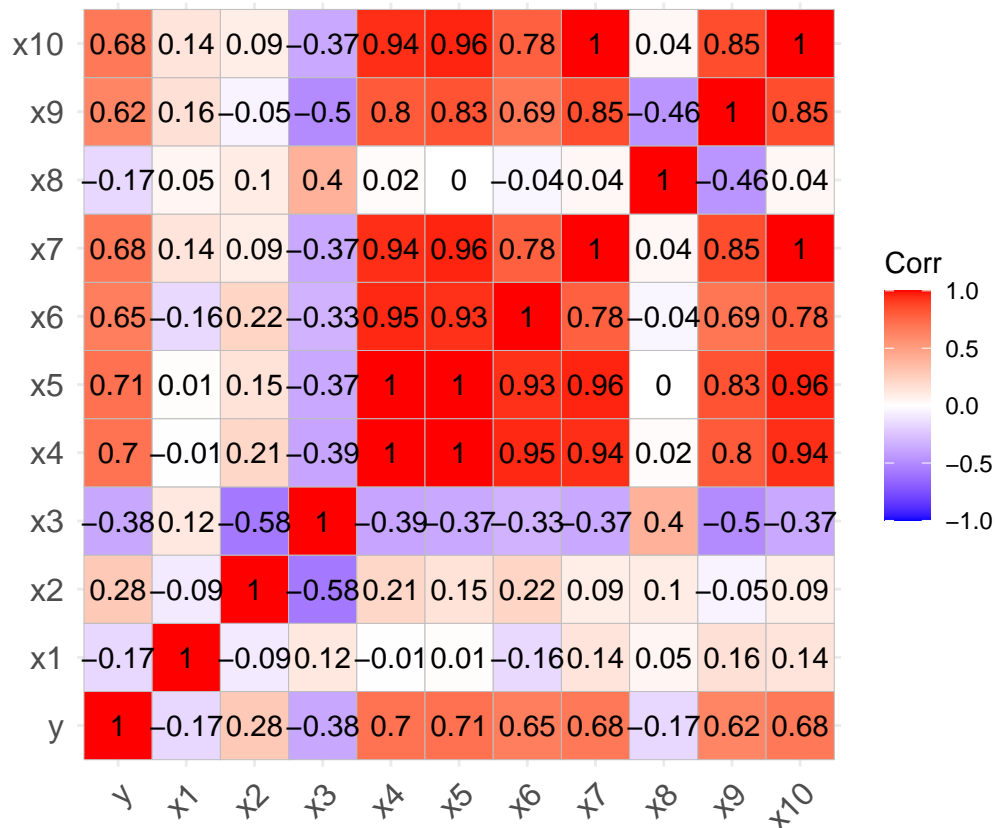
Q1. Considere os dados sobre a qualidade do vinho tinto, apresentados no ficheiro Q01-data.txt. Ajuste o modelo de regressao linear multipla, e faca uma analise completa desses dados. Que conclusoes voce tira dessa analise? (use 5% de significancia durante as analises)

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 possível observar que algumas covariáveis estão fortemente correlacionadas. Assim, pede-se:

(a) **Proponha algum método para resolver o problema da multicolinearidade no conjunto de dados.** Podemos perceber que temos um problema incomum de algumas variáveis terem correlação perfeita, isso definitivamente atrapalha o modelo e dificulta qualquer método de seleção de variáveis, especialmente dentro do R. Logo é necessário removermos essas variáveis problemáticas. Em seguida podemos rodar um modelo de seleção de variáveis. Nesse caso vamos retirar as variáveis x4, x5 e x10 que possuem múltiplas correlações altas com outras variáveis

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

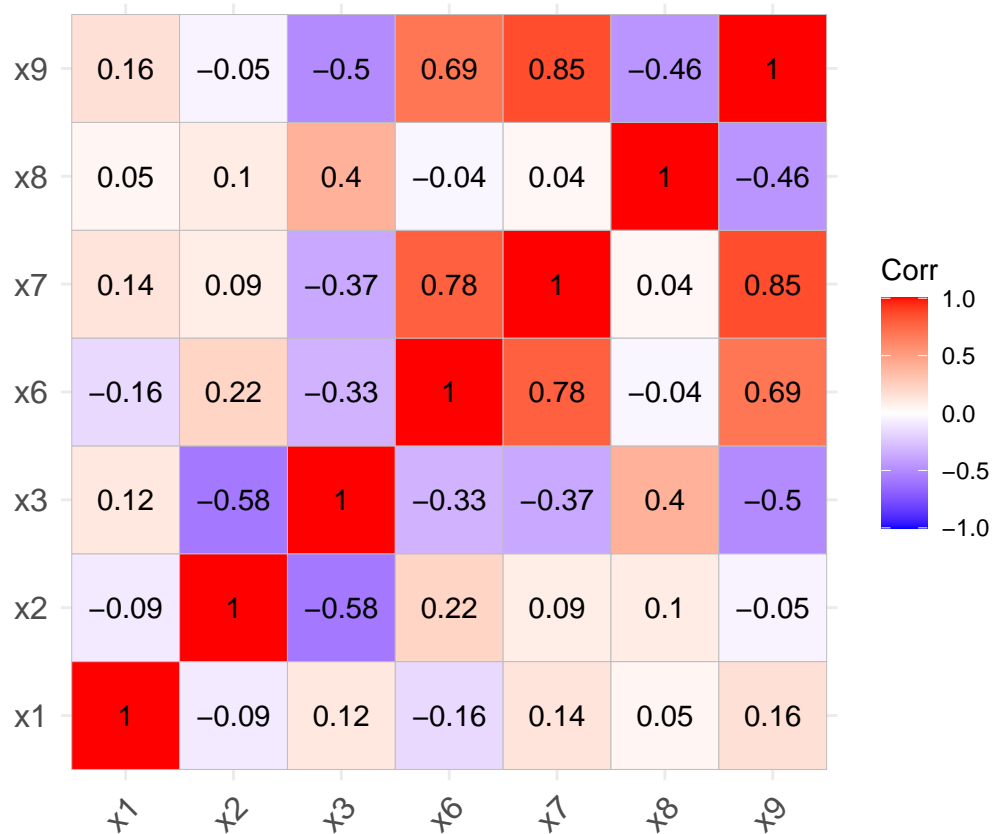
```
## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1  2.805    2.805   2.0444  0.16621
## x2         1  6.745    6.745   4.9159  0.03677 *
## x3         1  6.223    6.223   4.5352  0.04413 *
## x4         1 36.140   36.140  26.3393 3.363e-05 ***
## 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
```

```
matriz_cor = cor(Q01_data %>% dplyr::select(-c(1,5,6,11)))
```

```
# Converter a matriz de correlação em um data frame
df_cor <- as.data.frame(matriz_cor)
```

```
# Criar o gráfico de correlação (corr plot)
ggcorrplot(df_cor,lab = TRUE)
```



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

```
##          X.Intercept.          x1          x2          x3          x4          x6          x7
## Forward    -12.20843 -0.8457729 7.418389 0.01045843 -1.94732 3.461351 4.895176
## Backward   -10.10547 -0.9627786 6.250592 0.01229231      NA      NA 1.642902
## Both       -10.10547 -0.9627786 6.250592 0.01229231      NA      NA 1.642902
##
##           x8           x9
## Forward -11.425166 -0.1080232
## Backward  -8.778121      NA
## Both      -8.778121      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.

```
summary(both_md1)
```

```
##
## 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
```

```
anova(both_md1)
```

(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.

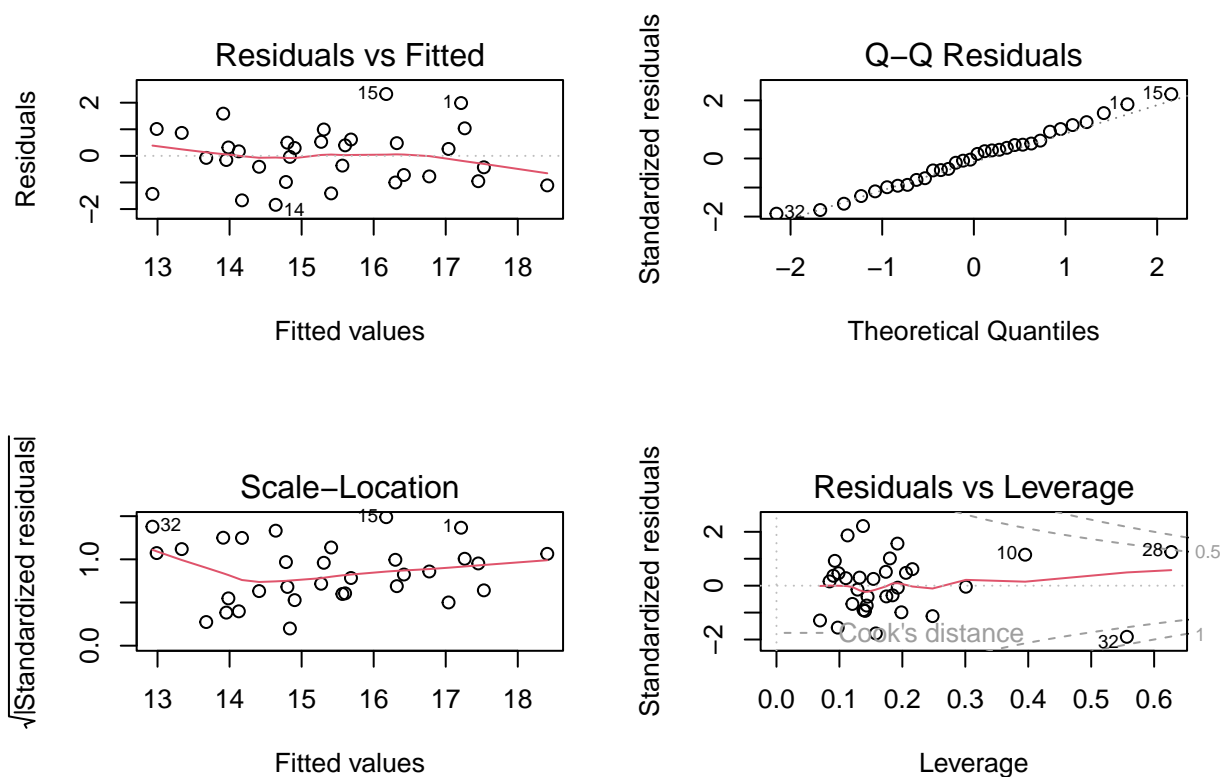
```
## Analysis of Variance Table
##
## Response: y
##      Df Sum Sq Mean Sq F value    Pr(>F)
## x1     1  2.805    2.805    2.1942  0.15055
## x2     1  6.745    6.745    5.2762  0.02992 *
## x3     1  6.223    6.223    4.8676  0.03640 *
## x7     1 40.170   40.170   31.4225 6.851e-06 ***
## x8     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
```

```
resultado_anova = aov(both_md1)
```

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 residuos e conclua.



```
shapiro.test(residuo)
```

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

```
bptest(both_md1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  both_md1
## BP = 2.0947, df = 5, p-value = 0.8359
```

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

Q02. Uma equipe de pesquisadores de saúde mental deseja comparar três métodos de tratamento da depressão grave (A, B e C=referência). Eles também gostariam de estudar a relação entre idade e eficácia do tratamento, bem como a interação (se houver) entre idade e tratamento. Cada elemento da amostra aleatória simples de 36 pacientes, foi selecionado aleatoriamente para receber o tratamento A, B ou C. Os dados obtidos podem ser encontrados no arquivo Q02-data.txt. A variável dependente y é a eficácia do tratamento; as variáveis independentes são: a idade do paciente no aniversário mais próximo, e o tipo de tratamento administrado (use 1% de significância durante as análises).

(a) Ajuste o modelo de regressão 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 interação entre as variáveis independentes, i.e., assumamos que o modelo a ser ajustado tem a seguinte formulação: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \epsilon_i$, com a suposição de que $\epsilon_i \sim N(0, \sigma^2)$. Com base no modelo anterior,

- (i) Liste todos os possíveis submodelos que podem ser obtidos usando o modelo apresentado anteriormente.
- (ii) Interprete os coeficientes de regressão associados aos fatores de interação.
- (iii) Apresente a tabela anova para testar as seguintes hipóteses, $H_0 : \beta_1 = \beta_4 = \beta_5 = 0$ contra $H_1 : \exists \beta_j \neq 0$, com $j = 1, 4, 5$.
- (iv) Faça uma análise completa dos resíduos do modelo.