Primeira Prova de Modelos de Regressão

Carlos Eduardo Alves de Souza

04/03/2022

```
# Importando livrarias
library(robustbase)
```

```
## Warning: package 'robustbase' was built under R version 4.1.2
```

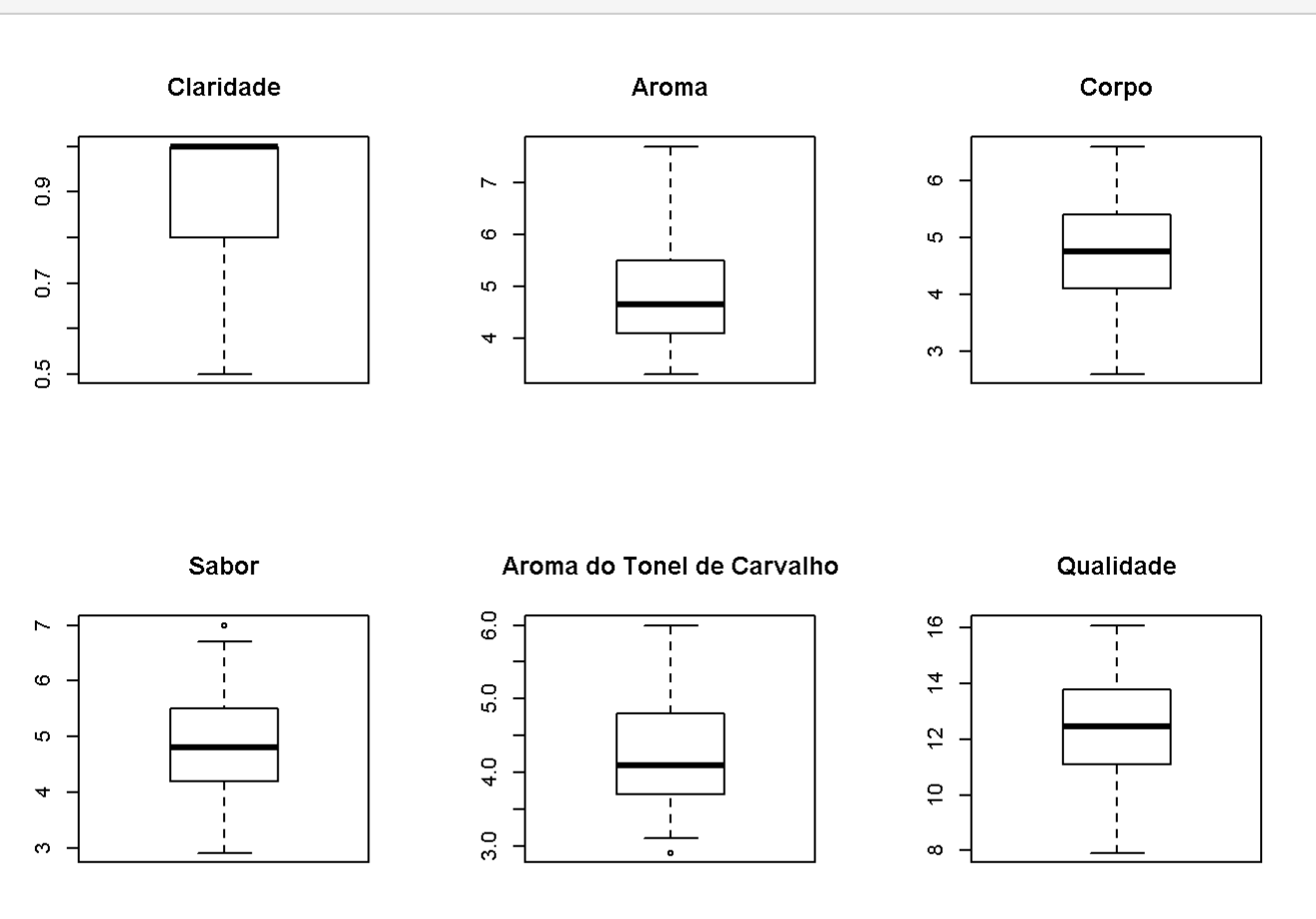
```
rm(list = ls()) # removendo variàveis já salvas...
```

```
# Importando a base de dados wine
wine = read.table("C:/Users/Carlo/Downloads/wine.txt", header = T)

# Transformação da covariável "região"
wine $ regiao <- factor(wine$regiao); wine
```

```
##    claridade aroma corpo sabor aromac qualidade regiao
## 1        1.0   3.3   2.8   3.1    4.1       9.8      1
## 2        1.0   4.4   4.9   3.5    3.9      12.6      1
## 3        1.0   3.9   5.3   4.8    4.7      11.9      1
## 4        1.0   3.9   2.6   3.1    3.6      11.1      1
## 5        1.0   5.6   5.1   5.5    5.1      13.3      1
## 6        1.0   4.6   4.7   5.0    4.1      12.8      1
## 7        1.0   4.8   4.8   4.8    3.3      12.8      1
## 8        1.0   5.3   4.5   4.3    5.2      12.0      1
## 9        1.0   4.3   4.3   3.9    2.9      13.6      3
## 10       1.0   4.3   3.9   4.7    3.9      13.9      1
## 11       1.0   5.1   4.3   4.5    3.6      14.4      3
## 12       0.5   3.3   5.4   4.3    3.6      12.3      2
## 13       0.8   5.9   5.7   7.0    4.1      16.1      3
## 14       0.7   7.7   6.6   6.7    3.7      16.1      3
## 15       1.0   7.1   4.4   5.8    4.1      15.5      3
## 16       0.9   5.5   5.6   5.6    4.4      15.5      3
## 17       1.0   6.3   5.4   4.8    4.6      13.8      3
## 18       1.0   5.0   5.5   5.5    4.1      13.8      3
## 19       1.0   4.6   4.1   4.3    3.1      11.3      1
## 20       0.9   3.4   5.0   3.4    3.4       7.9      2
## 21       0.9   6.4   5.4   6.6    4.8      15.1      3
## 22       1.0   5.5   5.3   5.3    3.8      13.5      3
## 23       0.7   4.7   4.1   5.0    3.7      10.8      2
## 24       0.7   4.1   4.0   4.1    4.0       9.5      2
## 25       1.0   6.0   5.4   5.7    4.7      12.7      3
## 26       1.0   4.3   4.6   4.7    4.9      11.6      2
## 27       1.0   3.9   4.0   5.1    5.1      11.7      1
## 28       1.0   5.1   4.9   5.0    5.1      11.9      2
## 29       1.0   3.9   4.4   5.0    4.4      10.8      2
## 30       1.0   4.5   3.7   2.9    3.9       8.5      2
## 31       1.0   5.2   4.3   5.0    6.0      10.7      2
## 32       0.8   4.2   3.8   3.0    4.7       9.1      1
## 33       1.0   3.3   3.5   4.3    4.5      12.1      1
## 34       1.0   6.8   5.0   6.0    5.2      14.9      3
## 35       0.8   5.0   5.7   5.5    4.8      13.5      1
## 36       0.8   3.5   4.7   4.2    3.3      12.2      1
## 37       0.8   4.3   5.5   3.5    5.8      10.3      1
## 38       0.8   5.2   4.8   5.7    3.5      13.2      1
```

Análise dos dados

```
# Análise descritiva, por meio do gráfico boxplot
par(mfrow = c(2,3))
adjbox(wine $ claridade, main = "Claridade")
adjbox(wine $ aroma, main = "Aroma")
adjbox(wine $ corpo, main = "Corpo")
adjbox(wine $ sabor, main = "Sabor")
adjbox(wine $ aromac, main = "Aroma do Tonel de Carvalho")
adjbox(wine $ qualidade, main = "Qualidade")
```



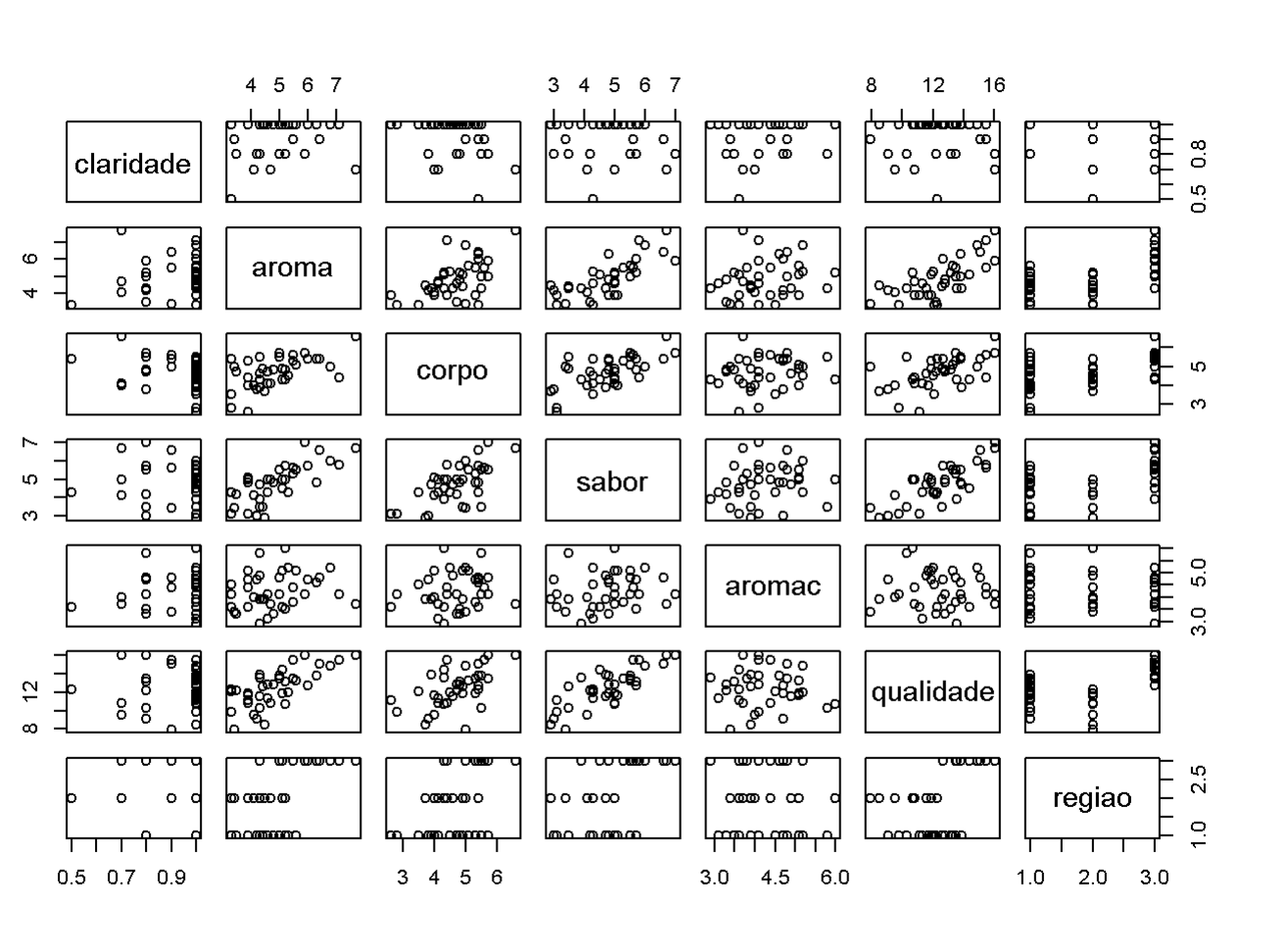
Percebemos a covariável "sabor" e "aromac" possuem outliers, sendo possível verificar se esse ponto considerado como outliers são influênte ao modelo. Já as covariáveis "sabor, qualidade e corpo" tem um comportamento simetrico, logo "aroma e aromac", exibem uma ligeira assimetria. Por fim, a covariável "claridade" tem-se um comportamento diferente das demais, não bem representado no gráfico de boxplot. Podemos averiguar isso atráves da função "summary", com a finalidade de verificar os pontos médios, a mediana e/ou quantis.

```
summary(wine)
```

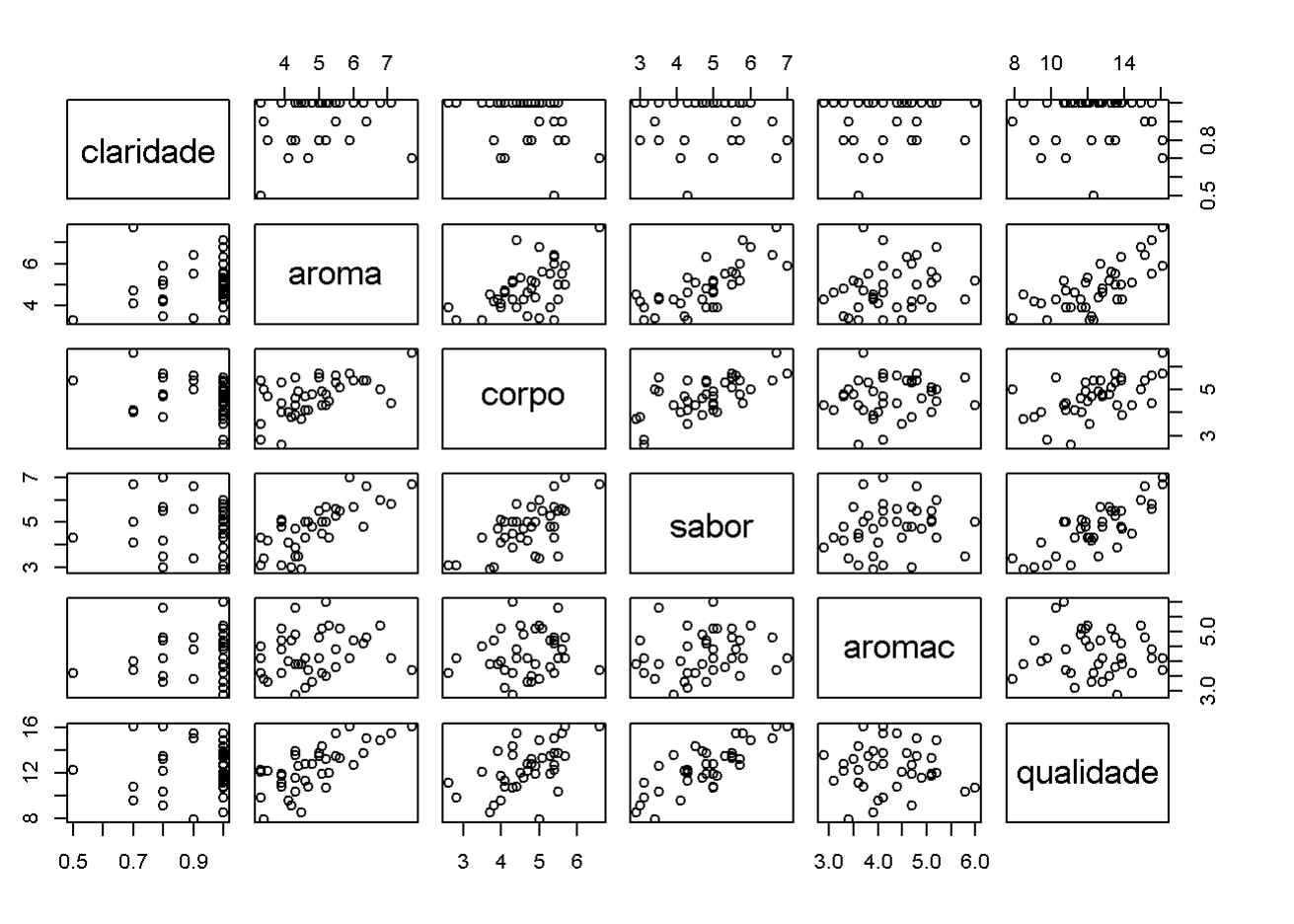
```
##    claridade          aroma           corpo           sabor      
##  Min.   :0.5000   Min.   :3.300   Min.   :2.600   Min.   :2.900  
##  1st Qu.:0.8250   1st Qu.:4.125   1st Qu.:4.150   1st Qu.:4.225  
##  Median :1.0000   Median :4.650   Median :4.750   Median :4.800  
##  Mean   :0.9237   Mean   :4.847   Mean   :4.684   Mean   :4.768  
##  3rd Qu.:1.0000   3rd Qu.:5.450   3rd Qu.:5.375   3rd Qu.:5.500  
##  Max.   :1.0000   Max.   :7.700   Max.   :6.600   Max.   :7.000  
##      aromac        qualidade     regiao
##  Min.   :2.900   Min.   : 7.90   1:17  
##  1st Qu.:3.700   1st Qu.:11.15   2: 9  
##  Median :4.100   Median :12.45   3:12  
##  Mean   :4.255   Mean   :12.44         
##  3rd Qu.:4.775   3rd Qu.:13.75         
##  Max.   :6.000   Max.   :16.10         
```

Após de executar a função "summary", é notório que cada ponto emitido esta em torno da média, indicando que uma certa simétria nos dados, para cada covariável presente, mesmo apresentando os gráficos boxplot em que duas destas colunas exibiram valores como outliers. Como mencionado a variável claridade em que a mediana 3º quartil e o valor maximo são pontos iguais, ou seja, realmente esta variável pode ser considerada não simetrica.

```
# Gráfico de dispersão
pairs(wine) # Gráfico de dispersão para o banco de dados inteiro
```

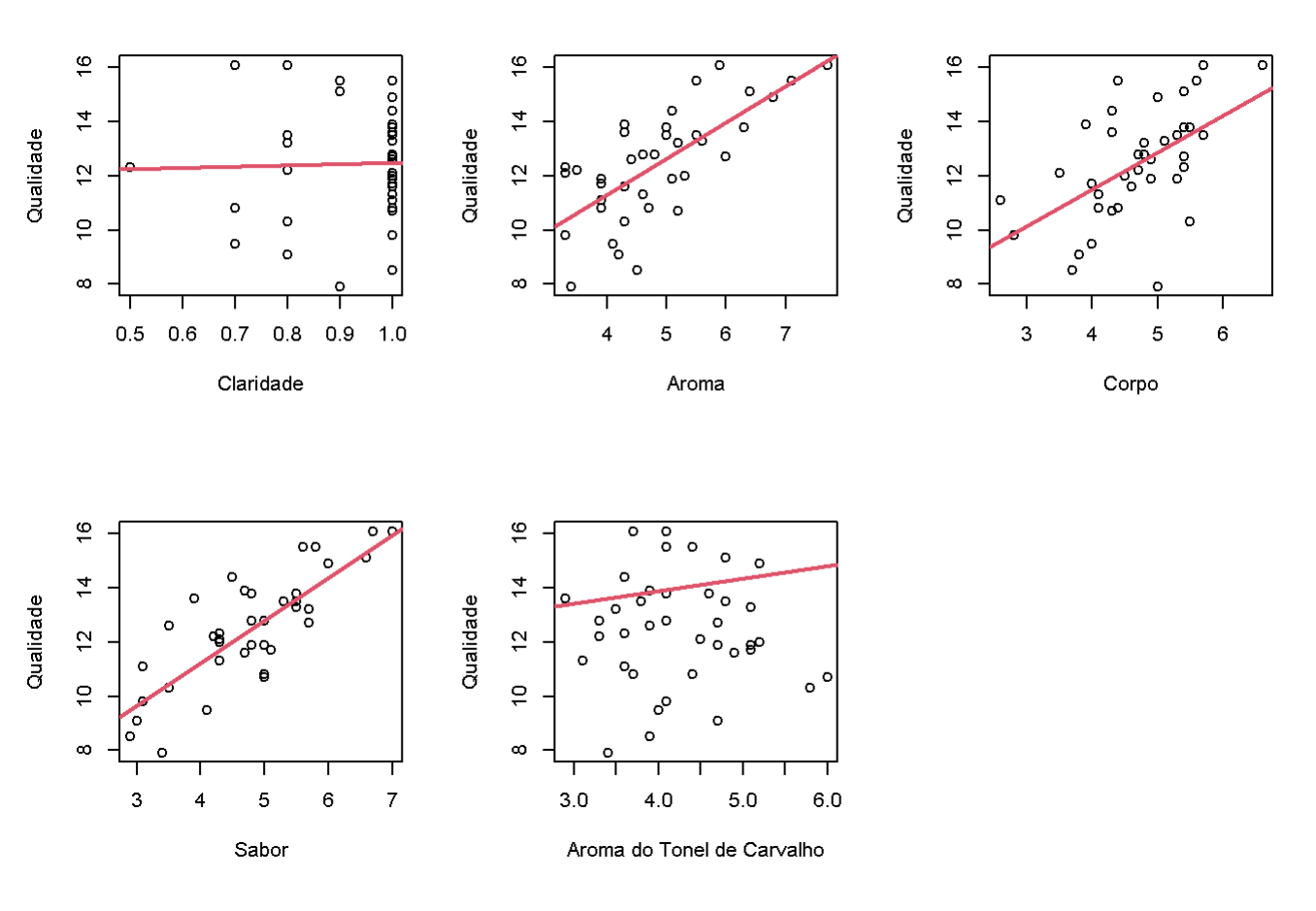



```
pairs(wine [, -7]) # Gráfico de dispersão retirando a coluna região
```



Apresentando os dois gráficos, podemos observar que a covariável região nao tem tanto influência nas demais colunas mencionadas devido ela ser uma variável contegorica com 3 posições, a coluna Claridade também mostra o mesmo comportamento da variável região. Após remoção da região, podemos ver que as variáveis Sabor e Qualidade exibem uma relação crescente com a qualidade do vinho, outrora as covariáveis não relatadas apresentam uma variabilidade maior que não permite facilmente identificar alguma dependência.

```
# Diagrama de dispersão para cada covariável
par(mfrow = c(2,3))

plot(qualidade ~ claridade, xlab = "Claridade", ylab = "Qualidade", data = wine)
abline(lm(qualidade ~ claridade, data = wine), col = 2, lwd = 2)

plot(qualidade ~ aroma, xlab = "Aroma", ylab = "Qualidade", data = wine)
abline(lm(qualidade ~ aroma, data = wine), col = 2, lwd = 2)

plot(qualidade ~ corpo, xlab = "Corpo", ylab = "Qualidade", data = wine)
abline(lm(qualidade ~ corpo, data = wine), col = 2, lwd = 2)

plot(qualidade ~ sabor, xlab = "Sabor", ylab = "Qualidade", data = wine)
abline(lm(qualidade ~ sabor, data = wine), col = 2, lwd = 2)

plot(qualidade ~ aromac, xlab = "Aroma do Tonel de Carvalho", ylab = "Qualidade", data = wine)
abline(lm(qualidade ~ claridade, data = wine), col = 2, lwd = 2)
```

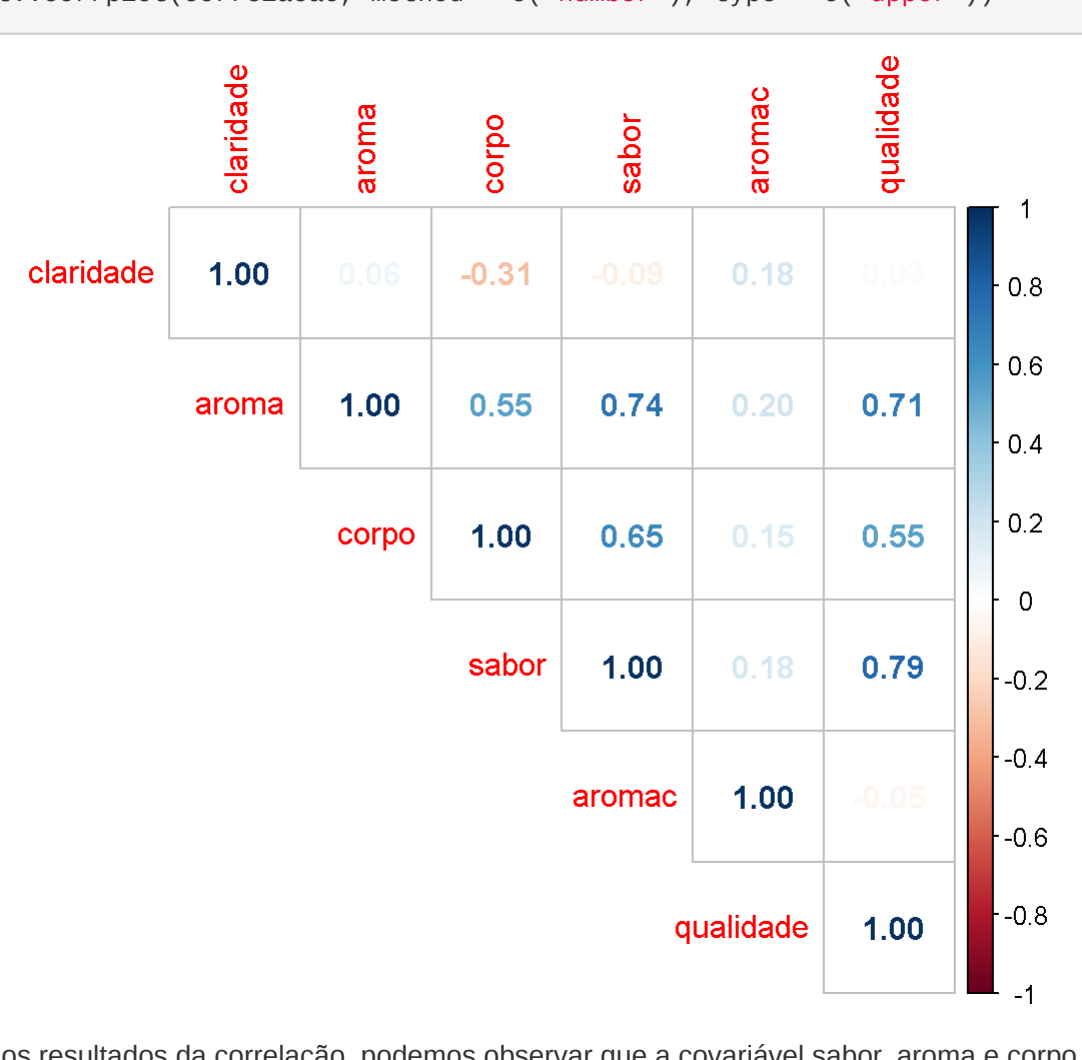


"Claridade" tem uma tendência levimente positiva, em que o R^2 se aproxima de zero, ou seja, não uma tendência significativa por serem uma variável categorica. A "Aromac", tem os seus valores aleatóriamente distribuídos, mas a sua linha de tendência não mostra uma relação linear. As demais covariáveis apresentadas (Aroma, Corpo e sabor) existem uma correlação positiva forte

```
# Exibindo os valores da correlação entre as variáveis
correlacao <- cor(wine[,-7], method = c("pearson")); correlacao
```

```
##             claridade     aroma      corpo       sabor      aromac   qualidade
## claridade  1.00000000 0.0619021 -0.3083783 -0.08515993  0.18321471  0.02844131
## aroma      0.06190210 1.0000000  0.5489102  0.73656121  0.20164445  0.70732432
## corpo     -0.30837826 0.5489102  1.0000000  0.64665917  0.15210591  0.54870219
## sabor     -0.08515993 0.7365612  0.6466592  1.00000000  0.17976051  0.79004713
## aromac     0.18321471 0.2016444  0.1521059  0.17976051  1.00000000 -0.04704047
## qualidade  0.02844131 0.7073243  0.5487022  0.79004713 -0.04704047  1.00000000
```

```
corrplot::corrplot(correlacao, method = c("number"), type = c("upper"))
```



Por meio dos resultados da correlação, podemos observar que a covariável sabor, aroma e corpo possuem uma correlação forte positiva com a covariável qualidade. A coluna aromac tem uma posição negativa fraca, indicando que não uma relacao com a variável qualidade o mesmo comportamento acontece com a variável claridade.

Aplicação do Modelo de Regressão Múltipla

```
# Ajuste do modelo incluido todas as covariáveis
ajuste_1 = lm(qualidade ~ claridade + aroma + corpo + sabor + aromac + regiao, data = wine); ajuste_1
```

```
## 
## Call:
## lm(formula = qualidade ~ claridade + aroma + corpo + sabor + 
##     aromac + regiao, data = wine)
## 
## Coefficients:
## (Intercept)    claridade        aroma        corpo        sabor       aromac  
##     7.81437      0.01705      0.08901      0.07967      1.11723     -0.34644  
##     regiao2      regiao3  
##    -1.51285      0.97259  
```

```
# Resumo dos resultados do modelo múltiplo
summary(ajuste_1)
```

```
## 
## Call:
## lm(formula = qualidade ~ claridade + aroma + corpo + sabor + 
##     aromac + regiao, data = wine)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.80824 -0.58413 -0.02081  0.48627  1.70909 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.81437    1.96944   3.968 0.000417 ***
## claridade    0.01705    1.45627   0.012 0.990736    
## aroma        0.08901    0.25250   0.353 0.726908    
## corpo        0.07967    0.26772   0.298 0.768062    
## sabor        1.11723    0.24026   4.650 6.25e-05 ***
## aromac      -0.34644    0.23301  -1.487 0.147503    
## regiao2     -1.51285    0.39227  -3.857 0.000565 ***
## regiao3      0.97259    0.51017   1.906 0.066218 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9154 on 30 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.7997 
## F-statistic:  22.1 on 7 and 30 DF,  p-value: 3.295e-10
```

Segundo os valores obtidos do "ajuste_1", somente as covariáveis "sabor" e "região" foram significativas, em que o p_valor<0.10. As demais covariáveis não tiveram uma representatividade ao modelo.

```
# Aplicando um novo modelo de ajuste sem a covariável região
ajuste_2 <- lm(qualidade ~ claridade + aroma + corpo + sabor + aromac, data = wine); ajuste_2
```

```
## 
## Call:
## lm(formula = qualidade ~ claridade + aroma + corpo + sabor + 
##     aromac, data = wine)
## 
## Coefficients:
## (Intercept)    claridade        aroma        corpo        sabor       aromac  
##      3.9969       2.3395       0.4826       0.2732       1.1683      -0.6840  
```

```
# Resumo dos resultados do novo modelo múltiplo
summary(ajuste_2)
```

```
## 
## Call:
## lm(formula = qualidade ~ claridade + aroma + corpo + sabor + 
##     aromac, data = wine)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.85552 -0.57448 -0.07092  0.67275  1.68093 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.9969     2.2318   1.791 0.082775 .  
## claridade     2.3395     1.7348   1.349 0.186958    
## aroma         0.4826     0.2724   1.771 0.086058 .  
## corpo         0.2732     0.3326   0.821 0.417503    
## sabor         1.1683     0.3045   3.837 0.000552 ***
## aromac       -0.6840     0.2712  -2.522 0.016833 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.163 on 32 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.6769 
## F-statistic: 16.51 on 5 and 32 DF,  p-value: 4.703e-08
```

Obs.: Se levarmos em consideração a exclusão da covariável "região" como indicado na parte em análise descritiva, segundo os resultados obtidos, podemos ver que a não inclusão dela (região) no ajuste_2, muda totalmente os valores das estimativas (coeficientes - betas) e muda drasticamente os valores do p_valor, indicando que somente as covariáveis: corpo e claridade não é significativo ao modelo onde o p_valor destas é > 0.10, conforme a referência na prova. É notório observar que o R-squared e o R ajustado para o modelo ajuste_2 e menor o que se encontra no modelo ajuste_1. Por fim, atráves do resultado do R-quadrado entre os dois modelos apresentos podemos incluir a covariável região.

```
# Aplicando um novo modelo de ajuste segundo os resultado obtidados do ajuste_1, excluindo as covariáveis não sig
nificativas
ajuste_3 <- lm(qualidade ~ sabor + regiao, data = wine); ajuste_3
```

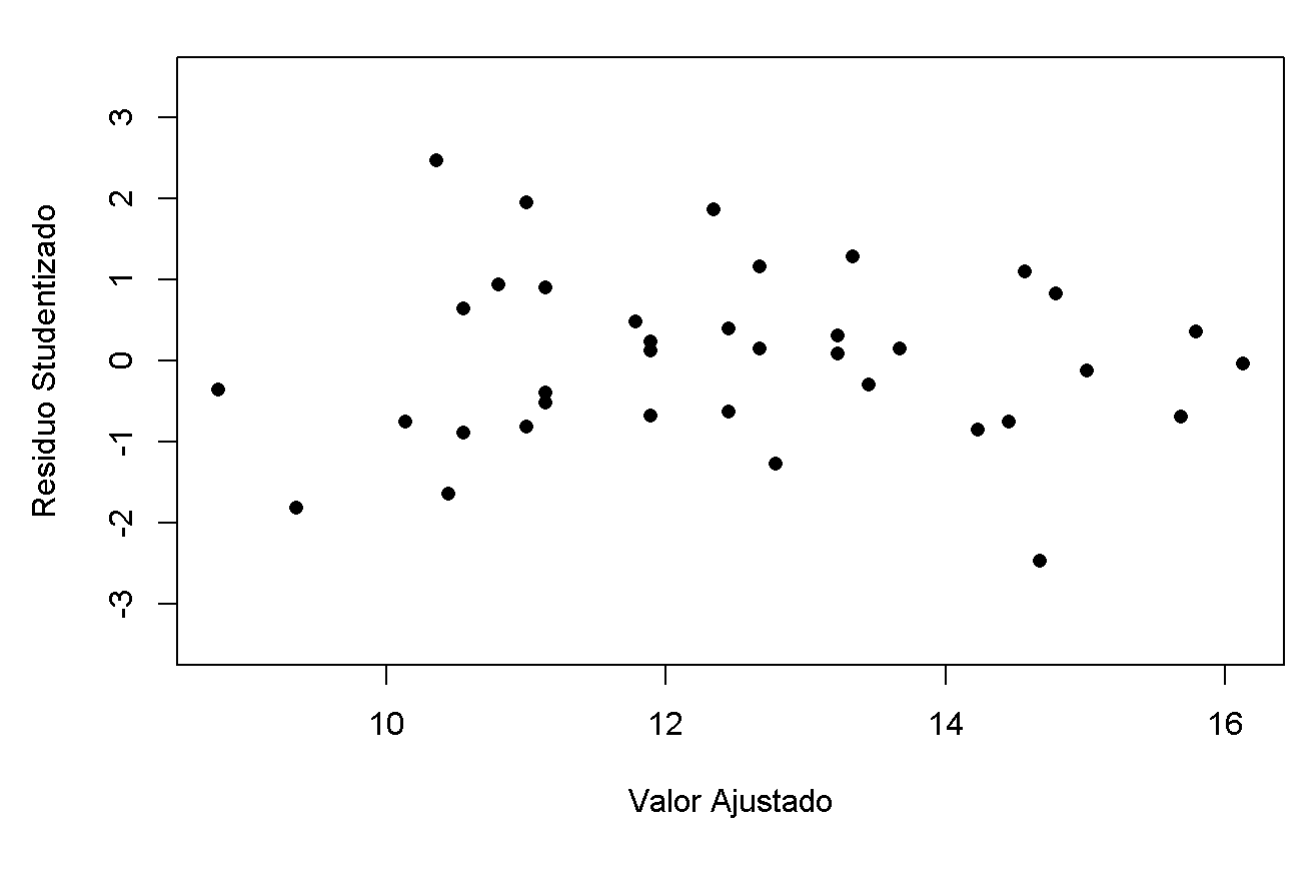
```
## 
## Call:
## lm(formula = qualidade ~ sabor + regiao, data = wine)
## 
## Coefficients:
## (Intercept)        sabor      regiao2      regiao3  
##       7.094        1.116       -1.533        1.223  
```

```
summary(ajuste_3)
```

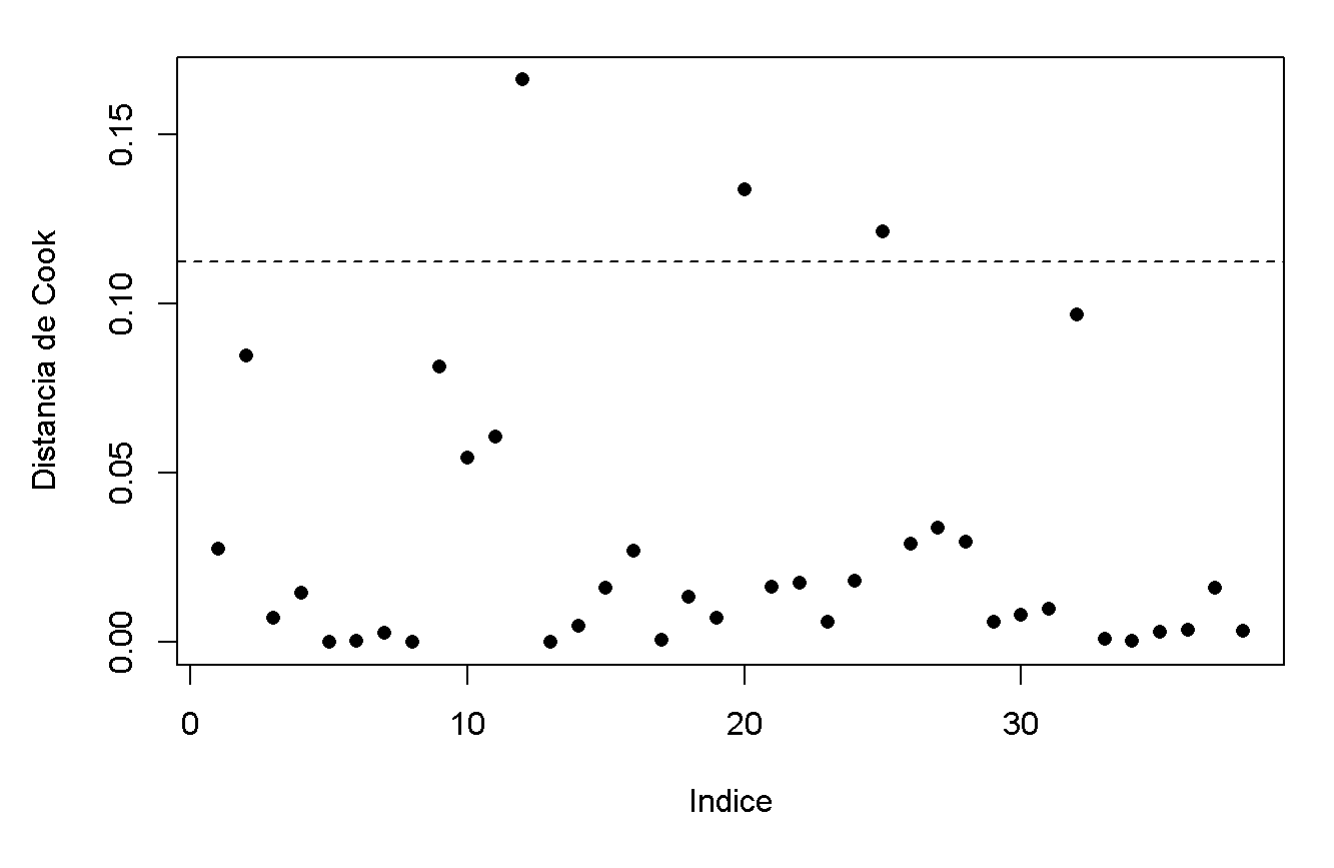
```
## 
## Call:
## lm(formula = qualidade ~ sabor + regiao, data = wine)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.97630 -0.58844  0.02184  0.51572  1.94232 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   7.0943     0.7912   8.967 1.76e-10 ***
## sabor         1.1155     0.1738   6.417 2.49e-07 ***
## regiao2      -1.5335     0.3688  -4.158 0.000205 ***
## regiao3       1.2234     0.4003   3.056 0.004346 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8946 on 34 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8087 
## F-statistic: 53.13 on 3 and 34 DF,  p-value: 6.358e-13
```

Segundo os resultos do modelo ajuste_3, todas as covariáveis foram significativas onde os resultado do respectivo p_valor é <0.10

```
# plotar os gráficos de resíduos e cook
fit.model <- ajuste_3
source("C:/Users/Carlo/Downloads/ResiduoNormal.txt")
```




```
source("C:/Users/Carlo/Downloads/CookNormal.txt")
```



Ao executar análise de resíduos e a distância de cook, observamos que os resíduos são normalmente distribuídos no intervalo [-3, 3] e a distância de cook idêntificou 3 pontos fora da margem.

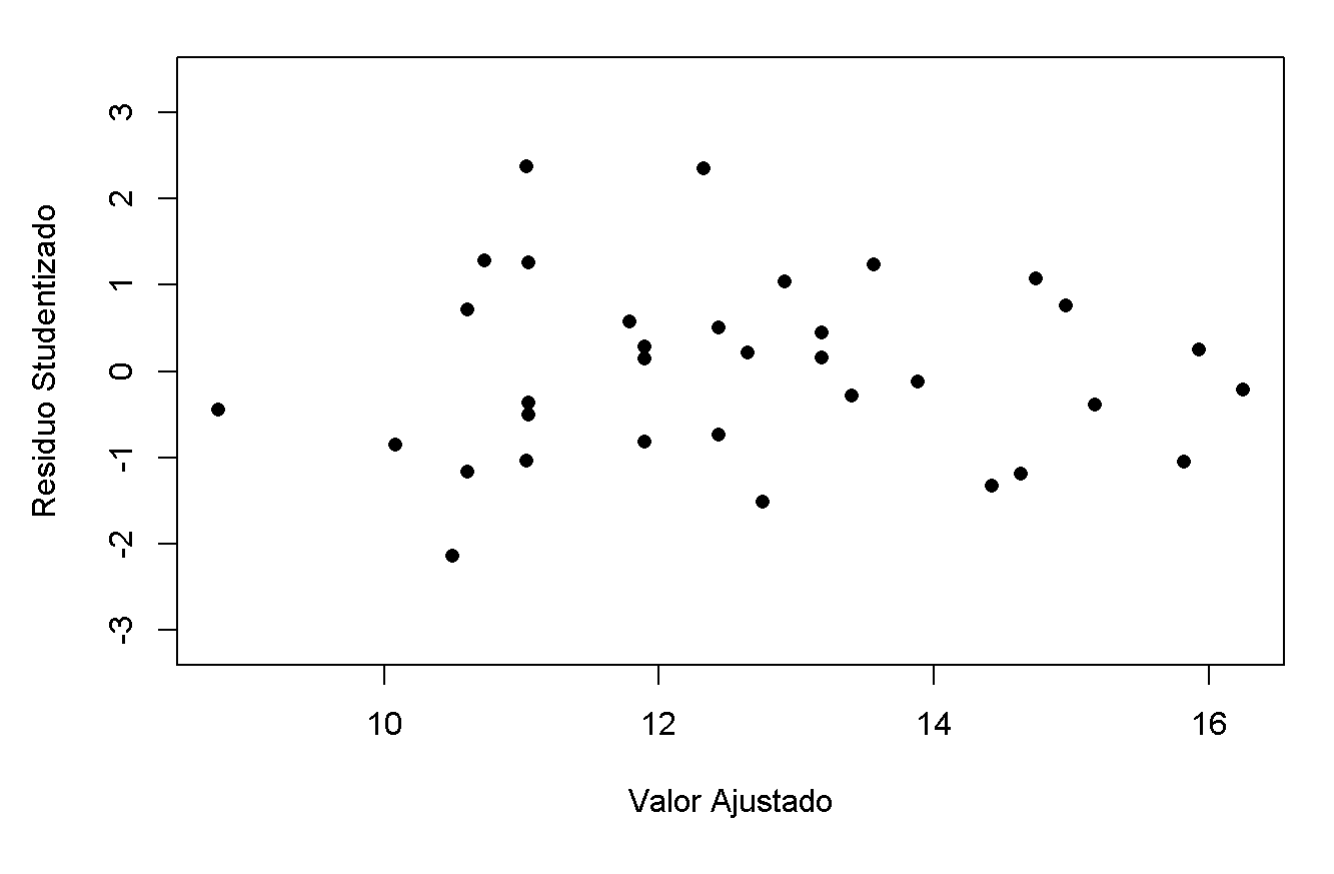
```
# idêntificando os pontos discrepantes
distance_cook <-  which(di>0.10)
wine[distance_cook,]
```

```
##    claridade aroma corpo sabor aromac qualidade regiao
## 12       0.5   3.3   5.4   4.3    3.6      12.3      2
## 20       0.9   3.4   5.0   3.4    3.4       7.9      2
## 25       1.0   6.0   5.4   5.7    4.7      12.7      3
```

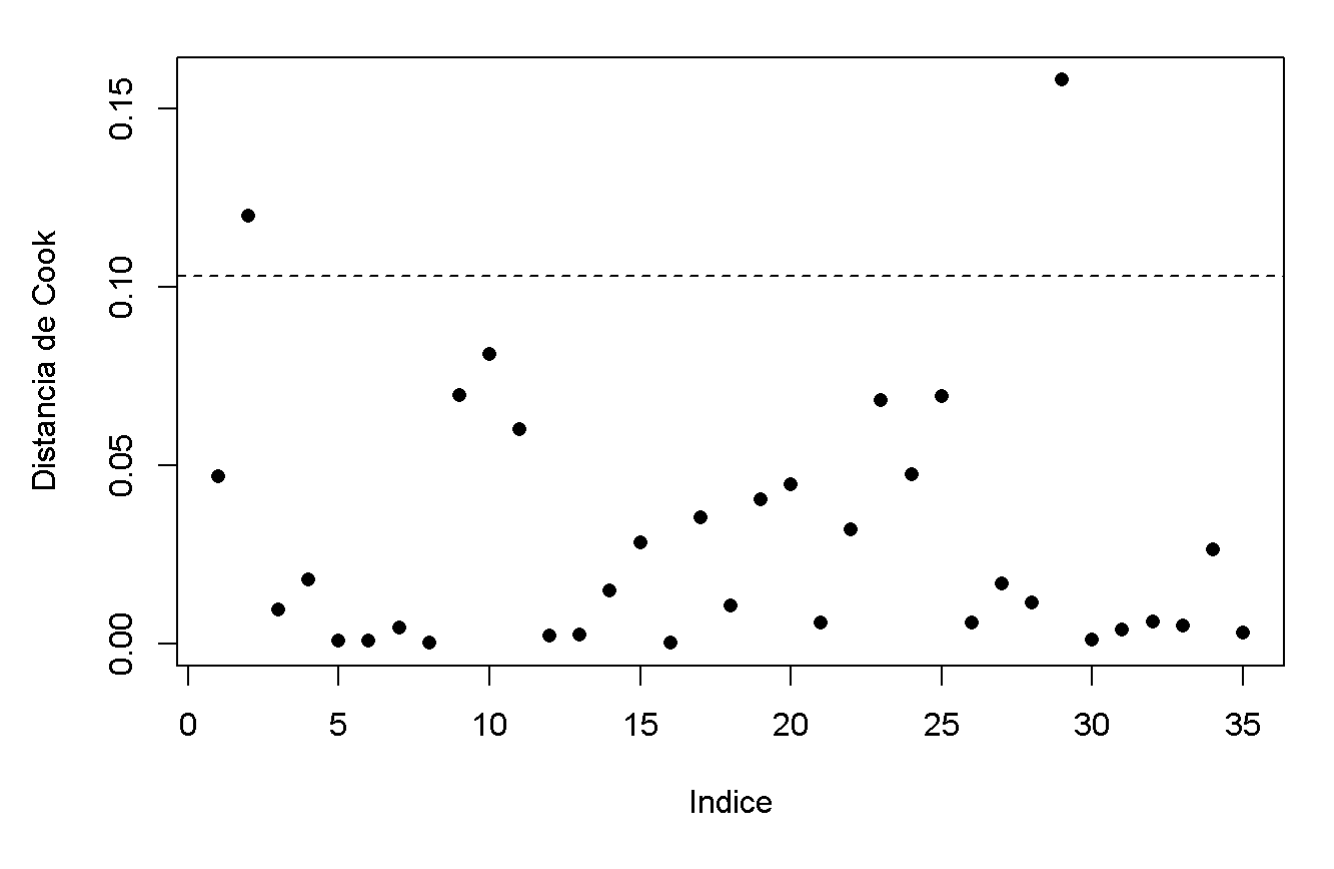
```
# Executando um novo modelo, excluindo os pontos discrepantes
ajuste_4 <- lm(qualidade ~ sabor + regiao, data = wine, subset = -c(12, 20, 25))

# Resumo do ajuste_4
summary(ajuste_4)
```

```
## 
## Call:
## lm(formula = qualidade ~ sabor + regiao, data = wine, subset = -c(12, 
##     20, 25))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.39616 -0.55695 -0.08442  0.45480  1.57559 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   7.2699     0.6681  10.882 4.10e-12 ***
## sabor         1.0754     0.1470   7.314 3.12e-08 ***
## regiao2      -1.5972     0.3332  -4.793 3.88e-05 ***
## regiao3       1.4525     0.3390   4.285 0.000165 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7403 on 31 degrees of freedom
## Multiple R-squared:  0.8728, Adjusted R-squared:  0.8605 
## F-statistic: 70.89 on 3 and 31 DF,  p-value: 5.634e-14
```

```
fit.model <- ajuste_4
source("C:/Users/Carlo/Downloads/ResiduoNormal.txt")
```



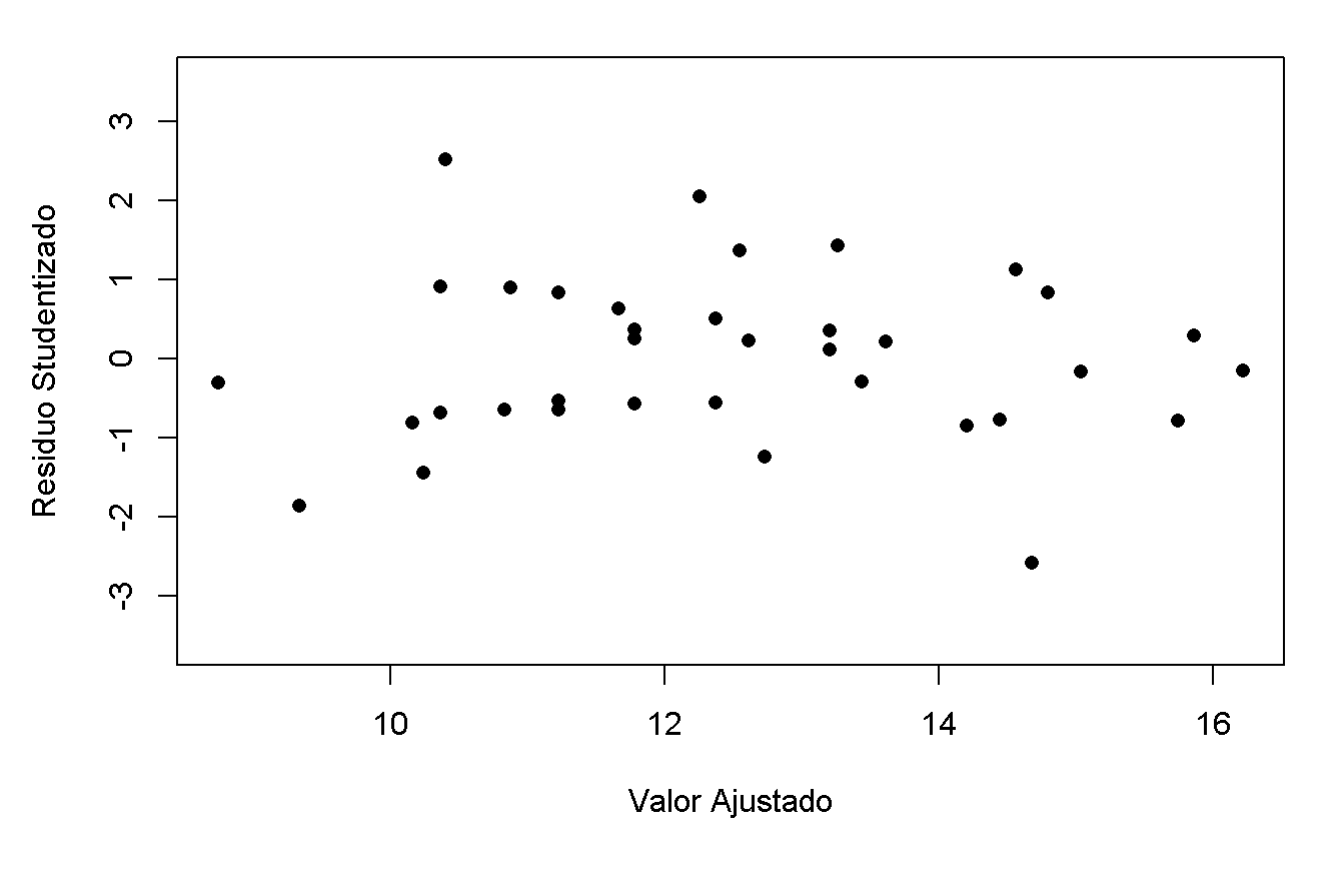

```
source("C:/Users/Carlo/Downloads/CookNormal.txt")
```



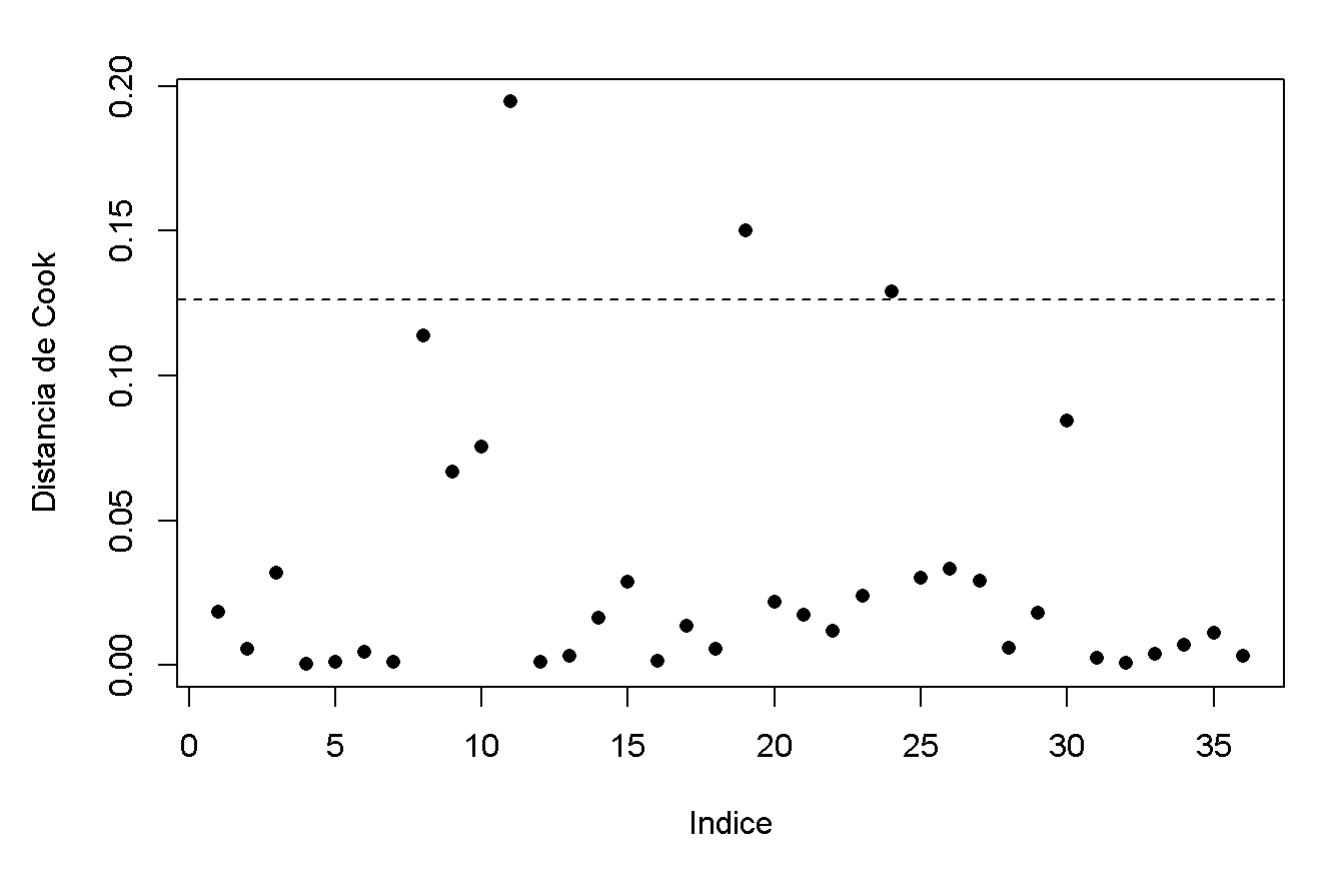

```
# Aplicando um novo modelo e removendo os pontos discrepantes identificados
ajuste_5 <- lm(qualidade ~ sabor + regiao, data = wine, subset = -c(2, 29))

# Resumo do ajuste_5
summary(ajuste_5)
```

```
## 
## Call:
## lm(formula = qualidade ~ sabor + regiao, data = wine, subset = -c(2, 
##     29))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.9820 -0.5415 -0.0121  0.5700  1.9000 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   6.6907     0.7981   8.384 1.40e-09 ***
## sabor         1.1840     0.1733   6.833 9.96e-08 ***
## regiao2      -1.3821     0.3776  -3.661 0.000899 ***
## regiao3       1.2423     0.3907   3.179 0.003268 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8704 on 32 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8256 
## F-statistic: 56.23 on 3 and 32 DF,  p-value: 7.44e-13
```

```
fit.model <- ajuste_5
source("C:/Users/Carlo/Downloads/ResiduoNormal.txt")
```




```
source("C:/Users/Carlo/Downloads/CookNormal.txt")
```



Executando os modelos: Ajuste_3, Ajuste_4 e Ajuste_5, foi identificando pontos discrepantes, ao identifica-los no ajuste_3 os mesmos foram removidos, então executamos o modelo ajuste_4, obtivemos o mesmo resultados em que, na análise de resíduos eles são normalmente distribuidos no intervalo [-3 e 3] e ao plotar a distância de cook ainda mostraram pontos discrepantes no modelo. Por fim, identificamos esses pontos discrepantes realizamos um novo ajuste (ajuste_5) removendo os referidos pontos(2 e 29), ao executar novamente a distância de cook e resíduos identificamos mais pontos discrepantes. Por tanto, não há difereça entre os modelos executados se removemos tais pontos discrepantes encontrados. Portanto, realizando a leitura do R-squared do ajuste_3 e ajuste_4, foi apontado um aumento após a remoção deste pontos (12, 20, 25), ou seja, ajuste_3 < ajuste_4. Ao executar um novo modelo ajuste_5, resultou em uma diminuição dos resultados do R-squared entre o ajuste_4 e ajuste_5, após a remoção destes pontos (2 e 29), logo ajuste_4 > ajuste_5 e o ajuste_3 < ajuste_5. Contudo, entre o modelo ajuste_3 e 4, a diferença entre eles, considerando o modelo ajuste_4 como modelo final definido.

```
# Tabela anova para ajuste_4
anova_ <- anova(ajuste_4); anova_
```

```
## Analysis of Variance Table
## 
## Response: qualidade
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## sabor      1 83.242  83.242 151.870 1.751e-13 ***
## regiao     2 33.328  16.664  30.402 4.917e-08 ***
## Residuals 31 16.992   0.548                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O modelo final (ajuste_4) teve 35 observações presentes, onde o seu p-valor é sgnificativo.