Pràctica 2: Modelització predictiva

Tipologia i cicle de vida de les dades

Autor: Albert Torres Rubio

Juny 2020

Contents

1 Descripció del dataset

El dataset escollit per l'anàlisi és el que conté dades relacionades amb el vi negre [1]. Concretament, és un dataset que recull dades de variants Portugueses del vi negre (*Vinho Verde*). El dataset ha estat obtingut a partir de la plataforma *Kaggle* mitjançant el següent enllanç:

- https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/download

El dataset conté dades relacionades amb la composició física i química dels diferents vins enregistrats, juntament, amb una nota que determina la qualitat d'aquests.

Des d'un punt de vista general les dades poden tenir un caràcter imporant ja que estan relacionant intrínsecament qualitats físiques/químiques amb una qualificació. Això implica que si l'estudi dona bons resultats és podría traçar una relació directe entre la qualitat d'un vi i les seves propietats físiques/químiques, que són objectives i mesurables. Aquest tipus d'anàlisi serien de gran utilitat a productors de vins, per exemple, per ajsutar els preus de les ampolles dde manera més precisa i ràpida, segons les propietats d'aquest. També, es podria "informatitzar" les qualifiacions de vins de tal manera que ja no dependrien només d'un catador si no també d'un analitzador "digital".

Consqüentment, la primera pregunta que vé a la ment quan s'observa el dataset i la seva descripció, és si hi ha relació entre les dades físiques/químiques i la qualificació dels vins. També, en el cas de que sí hagués una relació, quin tipus de relació seria. I finalment, es pot plantejar la pegunta de si es podria construir un predictor, prou eficient, que a partir de les dades físiques/químiques del vins determinés si el vi és de bona qualitat?

Per altra banda, un altre possible anàlisi del dataset, seria crear una variable nova a partir d'un un valor de tall de la qualifiacio. Aquesta variable indicaria si un vi està en el grup dels "bons" o dels "dolents". A partir d'aquí, es podrien realitzar els anàlisis convenients. També, es podrien realitzar contrastos d'hipòtesi per tal d'extreure propietats interesants de les mostres que puguin ser inferides a la població.

2 Integració i selecció de les dades d'interès a analitzar

Per tal d'analitzar les dades, s'han descarregat de l'enllaç anterior indicat i s'han desat dins la carpeta *data* del projecte amb el nom de *winequality-red.csv*. El següent troç de codi carrega le dades dins el projecte de R i mostra les seves característiques principals.

```
# LCàrrega de les dades
wineData <- read.csv('../data/winequality-red.csv')
str(wineData)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
sapply(wineData, class)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##            "numeric"             "numeric"            "numeric"
##       residual.sugar            chlorides  free.sulfur.dioxide
##            "numeric"             "numeric"            "numeric"
## total.sulfur.dioxide               density                   pH
##            "numeric"             "numeric"            "numeric"
##            sulphates               alcohol              quality
##            "numeric"             "numeric"            "integer"
```

Observant el resultat anterior i llegint la descripció del dataset proveït a *Kaggle*, es pot resumir les variables que conté el dataset i el tipus d'aquestes:

Variable	Nom	Descripció	Tipus
fixed.acidity	Acidesa (no volàtil)	Quantitat d'àcids no volàtils relacionats amb el vi	Decimal
volatile.acidity	Acidesa (volàtil)	Quantitat d'acid acètic en el vi	Decimal
citric.acid	Àcid cítric	Quantitat d'àcid cítric en el vi	Decimal
residual.sugar	Sucre residual	Quantitat de sucre restant, després de la fermentació	Decimal
chlorides	Sal	Quantitat de sal en el vi	Decimal
free.sulfur.dioxide	Diòxid de sulfur lliure	Quantitat de SO2 en forma lliure	Decimal
total.sulfur.dioxide	Diòxid de sulfur total	Quantitat total de SO2 en totes les seves formes	Decimal
density	Densitat	Densitat del vi	Decimal
pH	pH	PH del vi	Decimal
sulphates	Sulfats	Quantitat de sulfats	Decimal
alcohol	Alcohol	% d'alcohol	Decimal
quality	Qualitat	Qualificació del vi basada en dades sensorials	Enter

El dataset conté 1599 registres de vins i 12 variables diferents. Donat que totes les variables són atributs fisico-químics rellevants dels vins, s'han considerat com a d'interés, i per tant, tot el set de dades participarà a l'estudi.

3 Neteja de les dades

3.1 Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

S'analitzen les dades per veure si hi ha algun element buit:

```
# La funció de R mira si ha algun valor buit (' ') o 'N/A'
any(is.na(wineData))
```

```
## [1] FALSE
```

Com es pot observar no existeix cap variable que contingui un valor buit. Considerant la possibilitat de que el valor 0 denoti mancança de la dada, es descarta completament. Com hem vista abans, totes les dades són valors numèrics i, a través de la compressió global del dataset, el valor 0 està dins del domini de totes les variables. Es a dir, que si apareix un valor 0 en els registres, aquest s'ha de considerar com un valor vàlid.

En el cas que s'haguèssin trobat valors buits es podrien dur a terme diferents aproximacions per tal de tractar aquests valors. Uns quants exemples d'aproximacions per tractar aquests valors buits serien:

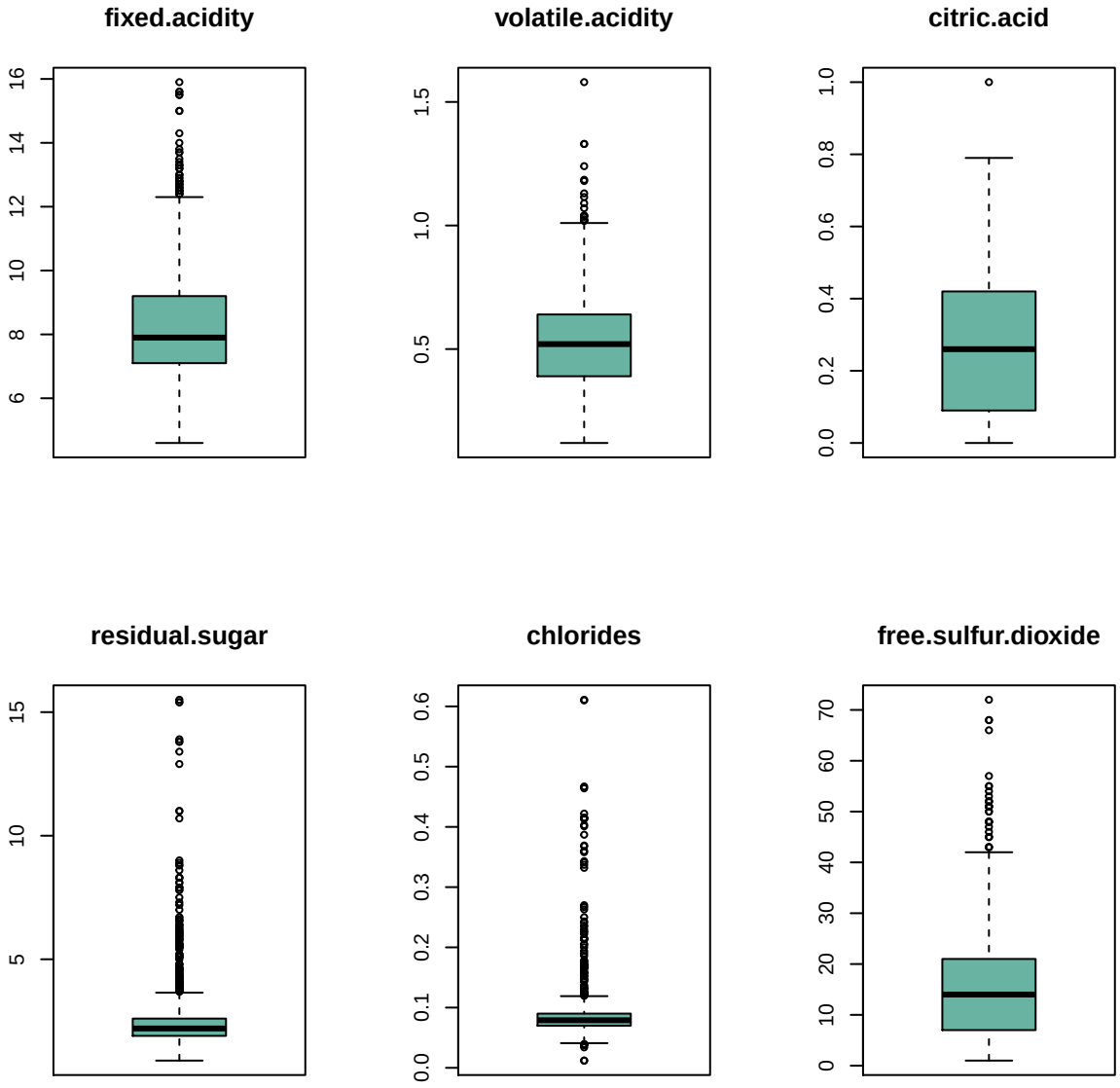
- Eliminar els registres amb valors buits.

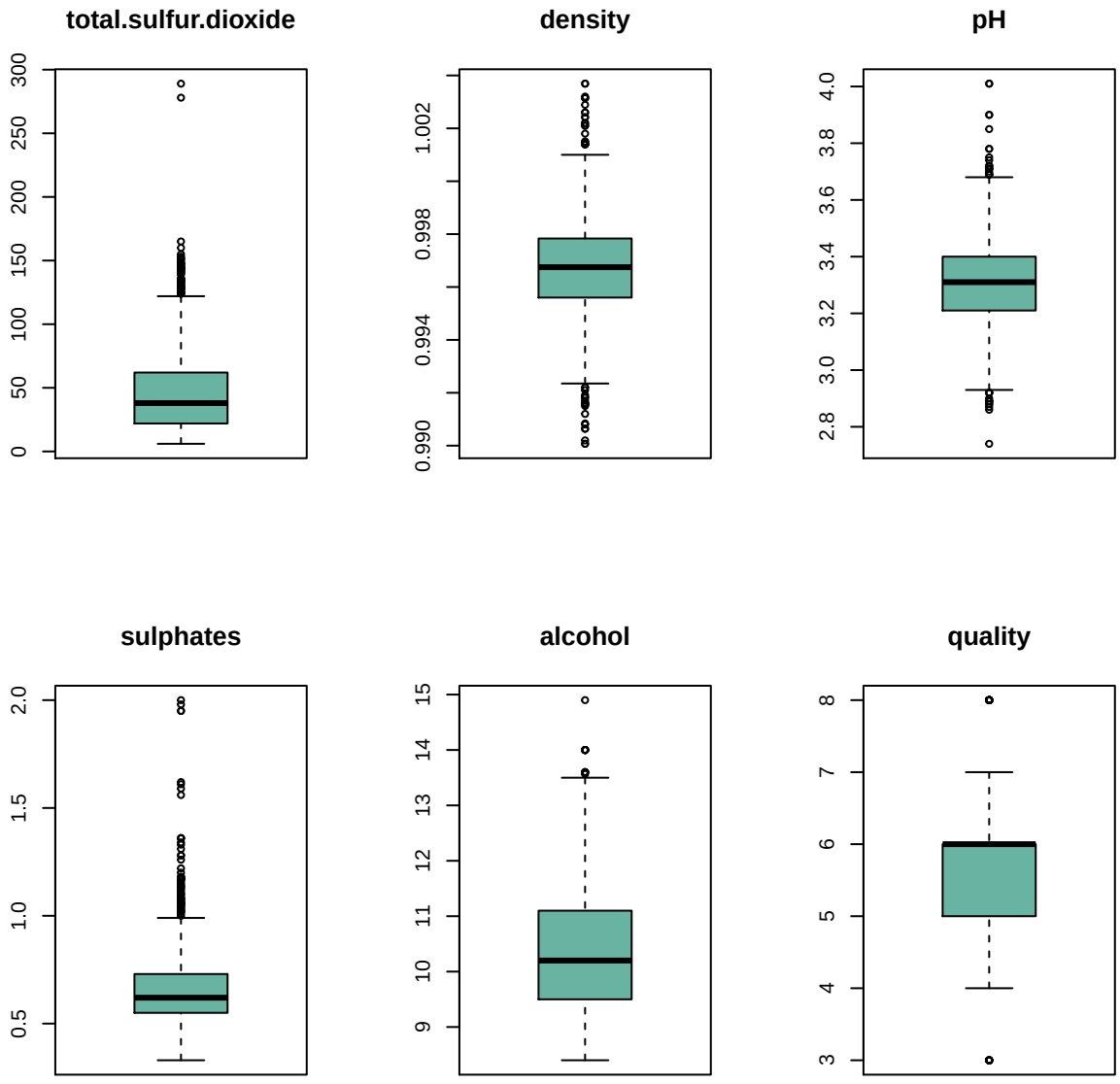
- Substituir els valors per una mateixa mesura de tendència central, com la mitjana o la mediana.
- Utilitzar mètodes probabilístics com el kNN, missForest, regressions... per predir el valor perdut i substituir-lo al dataset.
- Etc.

3.2 Identificació i tractament de valors extrems

Per tal d'analitzar si existeixen valors extrems en el conjunt de dades, utilitzarem la funció *boxplot* d'R.

```
par(mfrow=c(1,3))
for (i in 1:ncol(wineData)) {
  boxplot(wineData[,i], main=colnames(wineData)[i], col="#69b3a2")
}
```





A través dels gràfics es pot observar com R determina que existeixen valors extrems en totes les variables. Peró, a continuació, s'analitzarà si aquests valors extrems són erronis o simplement són valors distants peró petanyents al domini de les mostres. Un valor extrem és un valor que s'allunya dels valors "comuns" de la variable peró no vol dir que sigui un valor erroni en un principi.

```
outliers_table <- matrix(ncol=4)
for (i in 1:ncol(wineData)) {
  outliers <- boxplot.stats(wineData[,i])$out
  outliers_table <- rbind(outliers_table, c(col_names[i],
                                            round(mean(wineData[,i]), 3),
                                            round(min(outliers), 3),
                                            round(max(outliers), 3)))
}
kable(outliers_table[-1,],
```

```
    col.names = c("Variable", "Mitja", "Mínim valor extrem", "Màxim valor extrem"),
    booktabs = T) %>%
kable_styling(latex_options = c("striped")) %>%
kable_styling(position = "center")
```

Variable	Mitja	Mínim valor extrem	Màxim valor extrem
fixed.acidity	8.32	12.4	15.9
volatile.acidity	0.528	1.02	1.58
citric.acid	0.271	1	1
residual.sugar	2.539	3.7	15.5
chlorides	0.087	0.012	0.611
free.sulfur.dioxide	15.875	43	72
total.sulfur.dioxide	46.468	124	289
density	0.997	0.99	1.004
pH	3.311	2.74	4.01
sulphates	0.658	1	2
alcohol	10.423	13.567	14.9
quality	5.636	3	8

Amb el codi anterior i la taula creada a partir d'aquest, es poden observar els rangs que formen tots els valors extrems de les variables (mínim valor extrem i màxim valor extrem). De tal manera que no hi ha cap valor extrem fora d'aquest rang. Analitzant tots els rangs, cap rang conté cap valor totalment incoherent amb la variable que representa. Simplement aquests valors s'allunyen de la seva mitja, no de manera excesiva com per a considerar-los valors extrems incoherets o erronis. Com a resultat d'aquest anàlisi, no s'aplicarà cap tractament sobre els valors extrems ja que es consideren veritables i possibles valors que poden adquirir les propietats físiques i químiques del vi.

4 Anàlisi de les dades

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar)

Es realitzaran 3 anàlisis:

1. El primer anàlisi consistirà en realitzar un contrast d'hipòtesi del valor mitjà de la qualitat del vi amb el valor (qualificació) de 5. Aquest anàlisi vol determinar si la qualitat general del vi negre *Vinho Verde* és igual o superior a 5. Es a dir, es pretén respondre la següent pregunta: ¿Es pot afirmar que la qualitat mitjana d'un vi negre *Vinho Verde*, és acceptable? Considerant que "acceptable" correspon a una qualificació de 5 o superior.

2. El segon anàlisi serà crear un model de regressió lineal que sigui predictor de la qualitat del vi a través de les propietats fisico-químiques d'aquest. Es crearan dos models lineals, en el primer s'utilitzarà una matriu de correlació per determinar quines variables s'inclouran en el model i en el segon es durà a terme un anàlisi previ de components principals (PCA). A través d'aquest estudi es voldrà determinar, per una banda quines són les propietats del vi que creen un millor model de regressió lineal i per l'altra, si es millora la predicció d'aquest model fent un anàlisi previ de components principals.

3. Finalment, l'últim anàlisi consistirà en crear un model de regressió logística que intenti determinar si un vi es "bo" o "dolent", entenent com a "bo" una qualitat superior o igual a 5 i "dolent" una qualitat

inferior a 5. Per aquest anàlisi es crearà una variable dicotòmica que classificarà els vins en "bons" i "dolents" segons la qualifiació d'aquests. Tindrà el valor 1 en cas de que el vi sigui "bo" i 0 en el cas contrari:

```
wineData$goods <- ifelse(wineData$quality >= 5, 1, 0)
```

4.2 Comprovació de la normalitat i homogeneïtat de la variància

Es comprovarà tots si les dades segueixen una distribució normal i la seva homocedasticitat respecte la variable qualitat. Per realitzar el test de normailitat utilitzarem els tests *Shapiro-Wilk* i *Kolmogorov-Smirnov (Lilliefors)*.

```
library(nortest)

norm_table <- matrix(ncol=4)
for (i in 1:12) {
  p_val_1 = shapiro.test(wineData[,i])$p.value
  p_val_2 = lillie.test(wineData[,i])$p.value
  if (p_val_1 >= 0.05 && p_val_2 >= 0.05) {
    normal <- "Sí"
  } else {
    normal <- "No"
  }

  norm_table <- rbind(norm_table, c(col_names[i], p_val_1, p_val_2, normal))
}

kable(norm_table[-1,],
      col.names = c("Variable", "Shapiro-Wilk", "Kolmogorov-Smirnov (Lilliefors)",
                    "Normalitat"),
      booktabs = T) %>%
  add_header_above(c(" ", "p-valor" = 2), italic = T) %>%
  kable_styling(latex_options = c("striped")) %>%
  kable_styling(position = "center")
```

	p-valor		
Variable	Shapiro-Wilk	Kolmogorov-Smirnov (Lilliefors)	Normalitat
fixed.acidity	1.52501179295091e-24	6.98245550337254e-53	No
volatile.acidity	2.69293489456032e-16	4.48908406781828e-12	No
citric.acid	1.02193162131975e-21	9.85942912473966e-30	No
residual.sugar	1.02016171149076e-52	3.98171188614972e-309	No
chlorides	1.17905575371677e-55	1.26010678750114e-306	No
free.sulfur.dioxide	7.69459692029225e-31	1.28359886564504e-53	No
total.sulfur.dioxide	3.57345139578549e-34	7.94099617249671e-64	No
density	1.93605282884883e-08	6.25170665491422e-08	No
pH	1.71223728301906e-06	2.24404752955602e-06	No
sulphates	5.82314039765996e-38	4.60248846310498e-68	No
alcohol	6.64405672007326e-27	2.39150143471842e-64	No
quality	9.51508538364454e-36	1.95145497407858e-283	No

Com es pot observar, cap *p-valor* supera el valor de significació 0.05, de tal manera que cap variable segueix una distribució normal. Per tant, s'utilitzarà el test *Fligner-Killeen* per avaluar la homocedasticitat de les variables, ja que és un test orientat a variables que no segueixen un distribució normal.

```
hom_table <- matrix(ncol=3)
for (i in 1:11) {
  p_val = fligner.test(wineData$quality, wineData[, i])$p.value
  if (p_val >= 0.05) {
    flag <- "Sí"
  } else {
    flag <- "No"
  }
  hom_table <- rbind(hom_table, c(col_names[i], p_val, flag))
}

kable(hom_table[-1,],
      col.names = c("Variable", "p-valor (Fligner-Killeen*)", "Homocedasticitat"),
      booktabs = T) %>%
  kable_styling(latex_options = c("striped")) %>%
  kable_styling(position = "center")
```

Variable	p-valor (Fligner-Killeen*)	Homocedasticitat
fixed.acidity	0.981773123740242	Sí
volatile.acidity	0.362141747253417	Sí
citric.acid	0.236197626264641	Sí
residual.sugar	0.603307502614504	Sí
chlorides	0.564220853343594	Sí
free.sulfur.dioxide	0.695479131023802	Sí
total.sulfur.dioxide	0.0183192310199995	No
density	0.993806366292141	Sí
pH	0.523517118615673	Sí
sulphates	0.0413823548768257	No
alcohol	4.15745166691387e-07	No

Com es pot observar, les variables que guarden homogeneïtat de la variança respecte la qualitat són les que obtenen un *p_valor* més elevat que 0.05 (nivell de signifiació).

Com a punt final d'aquest apartat, com s'ha vist que cap variable es distribuex de manera normal, s'aplicarà el teorema del límit central (TLC). S'obtindran els resultats dels anàlisis considerant que les variables s'aproximen a una distribució normal, ja que tenim un gran nombre de registres (1599). Els resultats obtinguts seràn aproximacions prou acurades.

El TLC estableix que quan el tamany de la mostra es suficientment gran (N>30), la distribució de les mitjanes segueix aproximadament una distribució normal. El que implica, entre altres factors, que es poden aplicar procediments estadístics comuns, que requereixen que les dades siguin aproximadament normals. Permet aplicar aquests procediments útils a poblacions que són considerablement no normalas, com ocorre en el cas d'aquest estudi.

4.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents

4.3.1 Contrast del valor mitjà de la qualitat del vi amb una qualificació de 5

- Hipòtesi nul · la:

$$H_0 : \mu = 5$$

- Hipòtesi alternativa:

$$H_1 : \mu > 5$$

On μ és la qualificació mitjana i la hipòtesi alternativa és unilateral.

```
# Utilitzem la funció t.test per realitzar el contrast d'hipòtesi
t_greater <- t.test(wineData$quality, mu = 5, alternative = "greater")
print(t_greater)
```

```
##
##  One Sample t-test
##
## data:  wineData$quality
## t = 31.493, df = 1598, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 5
## 95 percent confidence interval:
##  5.602785      Inf
## sample estimates:
## mean of x
##  5.636023
```

Com es pot observar el *p_valor* del test ha sorit inferior al nivel de significació establert (0.05), per tant, es rebutja la hipòtesi nul · la en favor a l'alternativa. D'aquesta manera es pot determinar que la qualitat mitjana dels vins negres (*Vinho Verde*) és superior al 5 amb una confiança del 95%.

Seguidament, es calcula el test bilateral per obtenir l'interval de confiança:

```
t_sided <- t.test(wineData$quality)
print(t_sided)
```

```
##
##  One Sample t-test
##
## data:  wineData$quality
## t = 279.07, df = 1598, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.596410 5.675635
## sample estimates:
## mean of x
##  5.636023
```

```
max_ci <- t_sided$conf.int[1]
min_ci <- t_sided$conf.int[2]
```

L'interval de confiança del valor mitjà de la qualitat del vi és:

$$\mu \in [5.59641, 5.6756351]$$

Com es pot observar, el valor 5 no està comprès dintre de l'interval. L'interval de confiança ens indica que de cada 100 estudis independents amb diferents conjunts de mostres cada cop, en 95 (com a mínim) dels 100 casos tindrem la veritable qualitat mitjana del vi compresa dintre dels intervals calculats per cada estudi. Podem afirmar que el valor de la qualitat mitjana dels vins negres (*Vinho Verde*) estarà compresa dins l'interval anterior amb una confiança del 95%.

4.3.2 Models de regressió lineal

Per crear el model de regressió lineal s'esculliran les variables més correlacionades amb la qualitat del vi:

```
c_mat <- cor(wineData, method="spearman")
print(c_mat)
```

```
##                       fixed.acidity volatile.acidity  citric.acid residual.sugar
## fixed.acidity           1.00000000      -0.27828222  0.661708417    0.220700864
## volatile.acidity       -0.27828222       1.00000000 -0.610259467    0.032385603
## citric.acid             0.66170842      -0.61025947  1.000000000    0.176417306
## residual.sugar          0.22070086       0.03238560  0.176417306    1.000000000
## chlorides               0.25090411       0.15877025  0.112576508    0.212959242
## free.sulfur.dioxide    -0.17513656       0.02116264 -0.076451575    0.074617864
## total.sulfur.dioxide   -0.08841741       0.09411014  0.009399602    0.145375058
## density                 0.62307076       0.02501412  0.352285261    0.422265863
## pH                     -0.70667359       0.23357152 -0.548026276   -0.089970954
## sulphates               0.21265375      -0.32558398  0.331074404    0.038332000
## alcohol                -0.06657566      -0.22493168  0.096455544    0.116548131
## quality                 0.11408367      -0.38064651  0.213480914    0.032048168
## goods                   0.05610154      -0.16720587  0.112192373    0.007605807
##                          chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.2509041064       -0.1751365613        -0.0884174083
## volatile.acidity      0.1587702548        0.0211626414         0.0941101376
## citric.acid           0.1125765077       -0.0764515753         0.0093996024
## residual.sugar        0.2129592419        0.0746178640         0.1453750584
## chlorides             1.0000000000        0.0008051686         0.1300333418
## free.sulfur.dioxide   0.0008051686        1.0000000000         0.7896978767
## total.sulfur.dioxide  0.1300333418        0.7896978767         1.0000000000
## density               0.4113896972       -0.0411776800         0.1293321018
## pH                   -0.2343612736        0.1156791779        -0.0098414382
## sulphates             0.0208254792        0.0458623500        -0.0005038194
## alcohol              -0.2845039422       -0.0813673063        -0.2578060251
## quality              -0.1899223356       -0.0569006455        -0.1967350754
## goods                -0.0063041128        0.0889217360         0.0878704610
##                           density           pH     sulphates     alcohol
## fixed.acidity          0.62307076 -0.706673595  0.2126537506 -0.06657566
## volatile.acidity       0.02501412  0.233571519 -0.3255839818 -0.22493168
## citric.acid            0.35228526 -0.548026276  0.3310744040  0.09645554
## residual.sugar         0.42226586 -0.089970954  0.0383320002  0.11654813
## chlorides              0.41138970 -0.234361274  0.0208254792 -0.28450394
## free.sulfur.dioxide   -0.04117768  0.115679178  0.0458623500 -0.08136731
## total.sulfur.dioxide   0.12933210 -0.009841438 -0.0005038194 -0.25780603
## density                1.00000000 -0.312055078  0.1614782344 -0.46244458
```

```
## pH                   -0.31205508  1.000000000 -0.0803060380  0.17993243
## sulphates             0.16147823 -0.080306038  1.0000000000  0.20732955
## alcohol              -0.46244458  0.179932427  0.2073295535  1.00000000
## quality              -0.17707407 -0.043671935  0.3770601991  0.47853169
## goods                 0.01236563 -0.096778152  0.1315709604  0.02944211
##                          quality        goods
## fixed.acidity         0.11408367  0.056101544
## volatile.acidity     -0.38064651 -0.167205868
## citric.acid           0.21348091  0.112192373
## residual.sugar        0.03204817  0.007605807
## chlorides            -0.18992234 -0.006304113
## free.sulfur.dioxide  -0.05690065  0.088921736
## total.sulfur.dioxide -0.19673508  0.087870461
## density              -0.17707407  0.012365628
## pH                   -0.04367193 -0.096778152
## sulphates             0.37706020  0.131570960
## alcohol               0.47853169  0.029442115
## quality               1.00000000  0.363932452
## goods                 0.36393245  1.000000000
```

Com es pot observar a la columna "quality" no hi ha cap variable altament correlacionada amb aquesta propietat. Si més no, les variables que tenen un valor absolut més alt de correlació són: l'alcohol amb una correlació de 0.4785317 i l'acidesa (volàtil) amb un valor de -0.3806465.

A continuació, es provaran diverses combinacions de variables per intetar trobar el millor model de regressió lineal:

```
# Model lineal amt totes les variables
lm_all <- lm(quality ~ ., data=wineData[,1:12])

# Model lineal amb alcohol i acidesa volàtil
lm_al_va <- lm(quality ~ alcohol+volatile.acidity, data=wineData[,1:12])

# Model lineal amb les variables  amb el codi de signifiació més alt
# segons el primer model
lm_custom <- lm(quality ~ alcohol+volatile.acidity+chlorides+
                  total.sulfur.dioxide+sulphates+free.sulfur.dioxide
                +pH, data=wineData[,1:12])

coef <- lm_custom$coefficients
r2_custom <- summary(lm_custom)$adj.r.squared

summary(lm_all)
```

```
##
## Call:
## lm(formula = quality ~ ., data = wineData[, 1:12])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
```

```
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
summary(lm_al_va)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity, data = wineData[,
##     1:12])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59342 -0.40416 -0.07426  0.46539  2.25809
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.09547    0.18450   16.78   <2e-16 ***
## alcohol            0.31381    0.01601   19.60   <2e-16 ***
## volatile.acidity -1.38364    0.09527  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6678 on 1596 degrees of freedom
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.3161
## F-statistic: 370.4 on 2 and 1596 DF,  p-value: < 2.2e-16
```

```
summary(lm_custom)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + chlorides +
##     total.sulfur.dioxide + sulphates + free.sulfur.dioxide +
##     pH, data = wineData[, 1:12])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
```

```
## alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
## volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
## free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
## pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

A partir dels models lineals creats es pot observar com la bondat de l'ajust d'aquests no és massa significativa, ja que el coeficient de determinació ajustat és molt baix en cada cas. Això implica que no serien bons predictors de la qualitat del vi. Si més no, el millor model de regressió lienal obtingut ha estat el següent:

$$quality = \beta_0 + \beta_{alcohol} * AL + \beta_{volatile.acidity} * VA + \beta_{chlorides} * CHL +$$
$$+\beta_{total.sulfur.dioxide} * TSD + \beta_{sulphates} * SUL + \beta_{free.sulfur.dioxide} * FSD + \beta_{pH} * PH$$

$$quality = 4.4300987 + 0.2893028 * AL - 1.0127527 * VA - 2.0178138 * CHL$$
$$-0.0034822 * TSD + 0.8826651 * SUL + 0.0050774 * FSD - 0.4826614 * PH$$

Amb un coeficient de determinació ajustat de 0.3566527.

Seguidament s'estudiarà l'existència o no de multicolinealitat entre les covariables del model anterior. Es calcularan els factors d'inflació respectius (VIF). Aquest factor determina l'efecte de la colinealitat de la variància sobre un model de regressió.

```
library(faraway)
vif_lm_custom <- faraway::vif(lm_custom)
print(vif_lm_custom)
```

```
##              alcohol     volatile.acidity            chlorides
##             1.220157             1.241819             1.333333
## total.sulfur.dioxide            sulphates  free.sulfur.dioxide
##             1.943920             1.321931             1.882706
##                   pH
##             1.254570
```

S'observa com els VIFs corresponents a les variables explicatives no superen el valor 5 [2][3]. Això implica que no existeix una colinealitat forta entre aquestes variables i per tant, no són dependents entre sí. D'aquesta manera es confirma que es poden utilitzar com a variables explicatives independents en el model de regressió.

El següent pas en aquest estudi es calcular un nou model de regressió lineal aplicant prèviament un anàlisi de components principals (ACP), per veure si es milloren els resultats. L'anàlisi de components principals és una tècnica utilitzada per a descriure un conjunt de dades en termes de noves variables ("components") no correlacionades. L'objectiu és trobar les components que expliquen la màxima variança de les dades originals, major sigui la variància que expliquen de les dades originals major és la informació que contenen les components. L'ACP bucsa la projecció segons la que les dades queden millor respresanteds entermes de mínims quadrats. Per aplicar aquesta tècnica s'utilitzarà la funció *prcomp()* de R:

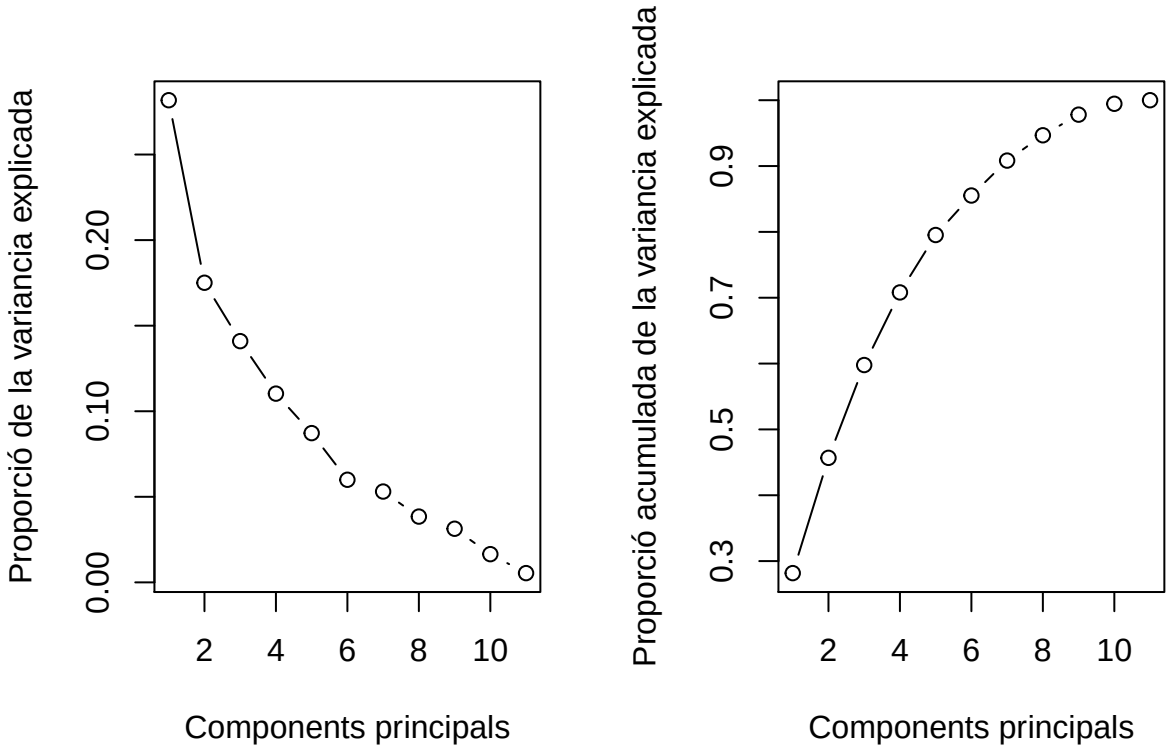
```r
# S'aplica la funció prcomp a les variables del dataset
# sense incluir la columna "quality", ni "goods"
prin_comp <- prcomp(wineData[,1:11], scale = T, center = T)
std_dev <- prin_comp$sdev
pr_var <- std_dev^2
prop_varex <- pr_var/sum(pr_var)

# Suma acumulativa del percentatge de la variancia explicada
# per les components principals
cumsum(prop_varex)
```

```
##  [1] 0.2817393 0.4568220 0.5977805 0.7080744 0.7952827 0.8552471 0.9083191
##  [8] 0.9467697 0.9781008 0.9945856 1.0000000
```

```r
# Gràfics representatius
par(mfrow=c(1,2))
plot(prop_varex, xlab = "Components principals",
     ylab = "Proporció de la variancia explicada",
     type = "b")

plot(cumsum(prop_varex), xlab = "Components principals",
 ylab = "Proporció acumulada de la variancia explicada",
 type = "b")
```



Com es pot observar a partir dels resultats anteriors, amb les 7 primeres components principals ja s'explica el 90% de la variancia de les dades originals del dataset. Per tant, es crearà el nou model de regressió amb aquestes 7 components. Per dur a terme aquest anàlisi s'haurà de crear un nou dataset on estiguin aquestes

components principals, juntament amb la qualificació dels vins del dataset original.

```
# Es crea el nou dataset a paritr de les components principals
winePrinComp <- as.data.frame(prin_comp$x[,1:7])
winePrinComp <- cbind(winePrinComp, wineData$quality)
colnames(winePrinComp)[8] <- "quality"

# Es crea el nou model de regressió lineal
lm_prcomp <- lm(quality ~ ., data=winePrinComp)
r2_prcomp <- summary(lm_prcomp)$adj.r.squared
summary(lm_prcomp)
```

```
##
## Call:
## lm(formula = quality ~ ., data = winePrinComp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70114 -0.37034 -0.06334  0.49300  1.94724
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.636023   0.016415 343.343  < 2e-16 ***
## PC1          0.050621   0.009327   5.427 6.61e-08 ***
## PC2         -0.225087   0.011832 -19.023  < 2e-16 ***
## PC3         -0.258946   0.013187 -19.637  < 2e-16 ***
## PC4         -0.032376   0.014908  -2.172    0.030 *
## PC5         -0.083721   0.016765  -4.994 6.57e-07 ***
## PC6          0.025114   0.020218   1.242    0.214
## PC7          0.095257   0.021491   4.432 9.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6564 on 1591 degrees of freedom
## Multiple R-squared:  0.3422, Adjusted R-squared:  0.3393
## F-statistic: 118.3 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Com es pot observar no es millora el model de regressió lineal previ i la bondat de l'ajust continua sent baixa, ja que el coeficient de determinació ajustat és, en aquest cas, 0.3393393.

4.3.3 Model de regressió logística

En aquest apartat, es realitzarà el model de regressió logística que intentarà predir la bona qualitat del vi. Es a dir si el vi tindrà una qualitat superior o inferior a 5. Es provaran diverses combinacions de variables per intetar trobar el millor model:

```
# Nou dataset sense la variable "quality""
# ja que analitzem la variable "goods"
goodsData <- wineData
goodsData$quality <- NULL

# Model logístic amt totes les variables
glm_all <- glm(goods ~ ., data=goodsData, family = "binomial")

# Model logísitc amb alcohol i acidesa volàtil
```

```r
glm_al_va <- glm(goods ~ alcohol+volatile.acidity, data=goodsData, family = "binomial")

# Model lineal amb les variables  amb el codi de signifiació més alt
# segons el primer model
glm_custom <- glm(goods ~ volatile.acidity+residual.sugar+pH+alcohol,
                  data=goodsData, family = "binomial")

glm_coef <- glm_all$coefficients
aic <- glm_all$aic

summary(glm_all)
```

```
##
## Call:
## glm(formula = goods ~ ., family = "binomial", data = goodsData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4093   0.1260   0.1871   0.2792   1.7153
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -388.41209  191.52107  -2.028 0.042556 *
## fixed.acidity          -0.48230    0.24609  -1.960 0.050010 .
## volatile.acidity       -4.58409    0.83210  -5.509 3.61e-08 ***
## citric.acid            -0.96437    1.24516  -0.774 0.438639
## residual.sugar         -0.33217    0.11037  -3.010 0.002616 **
## chlorides              -6.29348    3.08127  -2.042 0.041102 *
## free.sulfur.dioxide     0.01967    0.02299   0.856 0.392248
## total.sulfur.dioxide    0.01469    0.00829   1.771 0.076499 .
## density               411.36832  195.24622   2.107 0.035124 *
## pH                     -5.71974    1.65474  -3.457 0.000547 ***
## sulphates               1.12978    1.28177   0.881 0.378091
## alcohol                 0.72499    0.24964   2.904 0.003683 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.97  on 1598  degrees of freedom
## Residual deviance: 431.29  on 1587  degrees of freedom
## AIC: 455.29
##
## Number of Fisher Scoring iterations: 7
```

```r
summary(glm_al_va)
```

```
##
## Call:
## glm(formula = goods ~ alcohol + volatile.acidity, family = "binomial",
##     data = goodsData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.1753   0.1518   0.2169   0.2972   1.3593
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        5.1904     1.5487   3.351 0.000804 ***
## alcohol            0.1098     0.1439   0.763 0.445244
## volatile.acidity  -5.1169     0.6493  -7.880 3.27e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.97  on 1598  degrees of freedom
## Residual deviance: 464.24  on 1596  degrees of freedom
## AIC: 470.24
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm_custom)
```

```
##
## Call:
## glm(formula = goods ~ volatile.acidity + residual.sugar + pH +
##     alcohol, family = "binomial", data = goodsData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1861   0.1396   0.2071   0.2981   1.3565
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      11.36159    2.94677   3.856 0.000115 ***
## volatile.acidity -4.74674    0.67527  -7.029 2.07e-12 ***
## residual.sugar   -0.10543    0.08094  -1.303 0.192706
## pH               -2.25082    0.94926  -2.371 0.017734 *
## alcohol           0.24748    0.15739   1.572 0.115863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.97  on 1598  degrees of freedom
## Residual deviance: 457.53  on 1594  degrees of freedom
## AIC: 467.53
##
## Number of Fisher Scoring iterations: 7
```

Segons els diferents models creats es pot observar que el que presenta un menor AIC (*Akaike Information Criterion*) és el primer model, el que conté totes les variables del dataset. L'AIC és una mesura de la qualitat relativa d'un model estadístic. El menor AIC dels models creats serà, en relació als altres, el millor model d'aquest últim anàlisi. En conseqüència, el millor model obtingut segons l'AIC és el següent:

$$logit(goods) = \beta_0 + \beta_{fixed.acidity} * FA + \beta_{volatile.acidity} * VA + \beta_{citric.acid} * CA +$$
$$+\beta_{residual.sugar} * RS + \beta_{chlorides} * CHL + \beta_{free.sulfur.dioxide} * FSD + \beta_{total.sufur.dioxide} * TSD +$$
$$\beta_{density} * D + \beta_{pH} * PH + \beta_{sulphates} * SUL + \beta_{alcohol} * AL$$

$$logit(goods) = -388.4120898 + -0.4822986 * FA - 4.5840949 * VA - 0.9643682 * CA$$
$$-0.3321741 * RS + -6.2934795 * CHL + 0.019671 * FSD$$
$$0.0146852 * TSD + 411.3683162 * D - 5.7197375 * SUL + 1.1297783 * AL$$

Amb un AIC de 455.29.

L'AIC és un paràmetre que pondera entre la bondat de l'ajust i la complexitat del model. En el següent apartat es veurà mitjançant certs gràfics la bondat de l'ajust d'aquest millor model de regressió logísitca per a determinar si és un bon model predictor.

Com a punt final de l'anàlisi, s'aplicarà el test de Hosmer-Lemeshow que avalúa paramètricament la bondat de l'ajust d'un model logístic. Aquest test es basa en dividir la mostra d'acord a les variables predites de probabilitat per a cada una de les observacions. Les observacions són dividides en g grups d'acord amb la probabilitat predita de cada observació. Comunment es pren $g = 10$. Aleshores, es tracta de contar interval per interval l'observat i el predit per a cadascún dels resultats de la variable dependent dicotòmica. Si l'ajust és bò, indica que la p predita s'associa amb el resultat 1 de la varaible binomial. Mitjançant un test chi quadrat es comparen les freqüències esperades de cada interval amb les observacions reals. Finalment, un valor elevat del *p-valor* del test chi quadrat indicarà que l'ajust es bò.

```
library(ResourceSelection)
hl <- hoslem.test(glm_all$y, glm_all$fitted.values, g=10)
print(hl)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glm_all$y, glm_all$fitted.values
## X-squared = 7.1065, df = 8, p-value = 0.5252
```

Com es pot observar el p-valor (0.5251845) és superior al nivell de significació del 0.05, per tant, l'ajust del model logísitic és bo, segons el test H-L. Un *p_valor* superior al nivell de significancia del test H-L indica que el model s'ajsuta a la realitat, per tant, implica que l'observat s'ajusta suficientment al esperat sota el model.

5 Representació dels resultats a partir de taules i gràfiques

Per una banda, com hem vist al tercer apartat, hem utilitzat les gràfiques *boxplot* per determinar si existien valors extrems. Per una altra, en aquest apartat analitzarem la bondat dels ajustos dels models de regressió creats a l'apartat anterior. Concretament s'utilitzarà la corba ROC i el paràmetre AUC (*Area Under the Curver*) per a determinar la precissió dels ajustos.

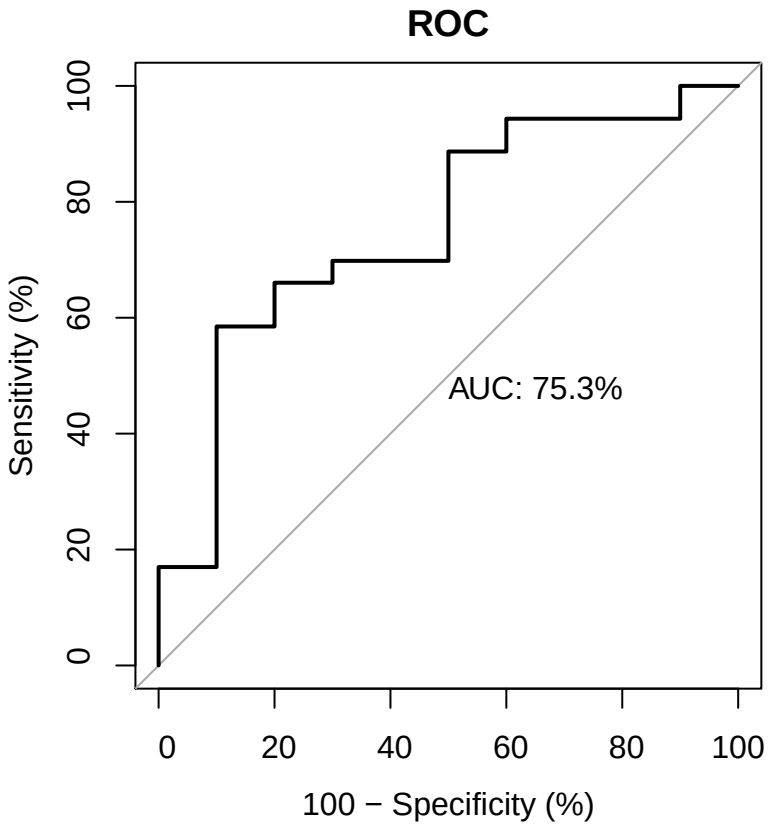
Utilitzant la llibreria pROC es pot dibuixar la corba ROC dels models i calcular l'àrea sota la corba. Si la corba ROC té una desviació important cap a l'esquerra i cap a la part superior del gràfic, està indicant que el model de regessió és un bon predictor. La corba ROC ens indica el balanç entre els veritables positius i els falsos negatius en un model de regressió.

Per altra banda, quan més elevat sigui el paràmetre AUC indicarà un millor ajust del model. El millor model correspon al model que té l'àrea sota la corba ROC més elevada.

5.1 Model de regressió lineal

Utilitzant el millor model de regressió lineal obtingut, indicat a l'apartat anterior, s'analitza a seva corba ROC:

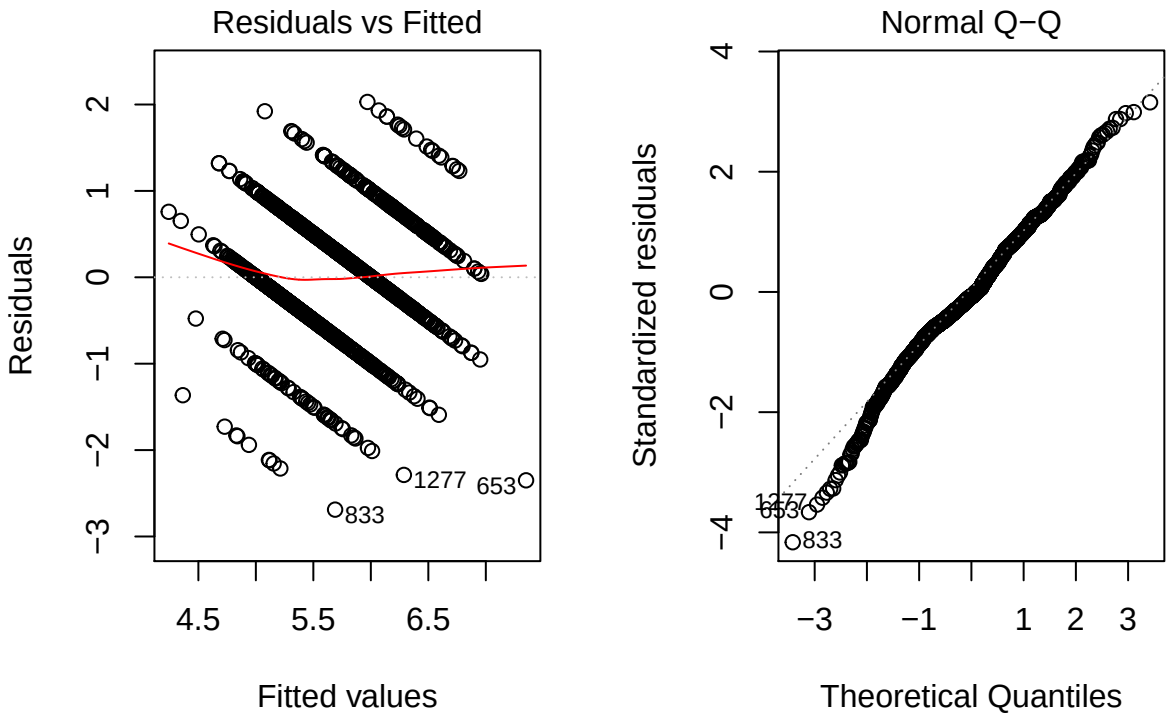
```
library(pROC)
par(pty="s")
roc_lm <- roc(wineData$quality, lm_custom$fitted.values, plot=TRUE,
legacy.axes=TRUE, print.auc = TRUE, percent = TRUE,
main = "ROC")
```



Com es pot observar, la corba no té una desviació prou significativa cap a l'esquerra ni cap a dalt, per tant, no es pot dir que la bondat de l'ajust sigui bona. Això ja quedava reflectit amb el coeficient de determinació ajustat obtingut (0.3566527), corroborant el baix ajust. A més, l'àrea sota la corba tampoc té un valor massa elevat, encara que és més d'un 70%.

També es poden analitzar els gràfics "Residus vs. Valors Ajustats" i el "Q-Q gràfic":

```
par(mfrow=c(1,2))
plot(lm_custom, which = c(1,2))
```

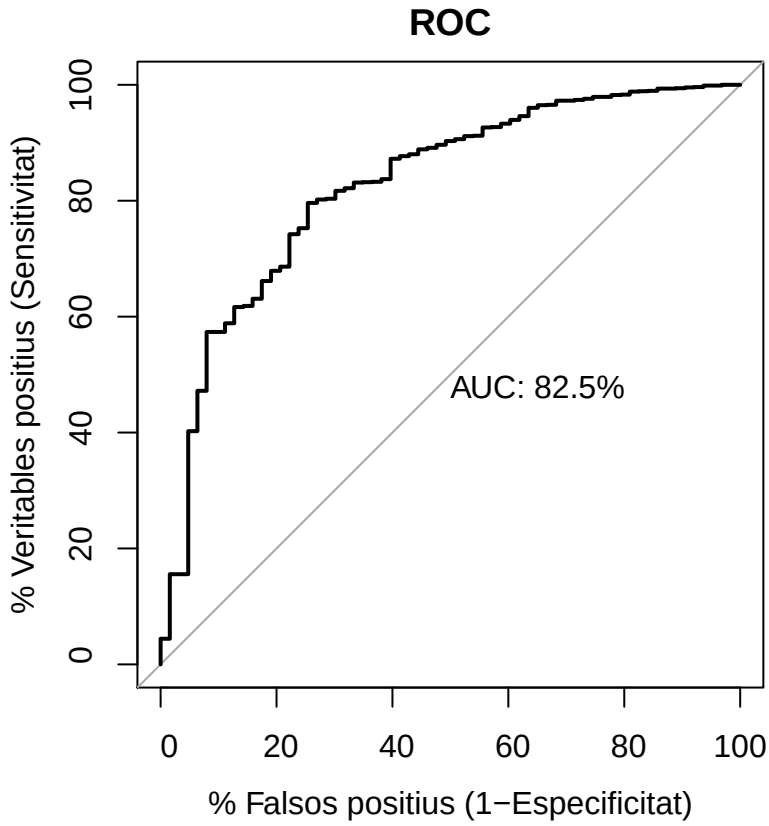


Pel que fa al primer gràfic es pot observar com hi ha una diferencia sistemàtica entre els valors ajustats i els reals, els punts no són aleatories sobre la linea horitzontal del gràfic. Pêro, si es consideren les linees diagonals que es formen, el valors si que mostren certa aleatorietat, i aquestes línees diagonals són degudes als valors enters exactes que agafa la variable "quality". En el segon gràfic es pot observar, com els els residus guarden una distribució normal segons els quantils teòrics, ja que s'ajusten prou bé a la diagonal.

5.2 Model de regressió logística

Emprant el millor model de regressió logística trobat, el que conté totes les variables del dataset, s'obté una la següent corba ROC:

```
par(pty="s")
roc_glm <- roc(glm_all$y, glm_all$fitted.values, plot=TRUE,
legacy.axes=TRUE, print.auc = TRUE, percent = TRUE,
xlab="% Falsos positius (1-Especificitat)",
ylab="% Veritables positius (Sensitivitat)",
main = "ROC")
```



Com es pot observar, en aquest cas la corba ROC presenta una accentuació més forta cap a l'esquerra i cap a dalt que en el cas anterior. Es pot afirmar que s'ha trobat una primera aproximació d'un bon model predictiu. Aquest permet distinguir quins vins són de bona qualitat (qualitat >= 5) i quins no, donades les seves propietats fisico-químiques, ja que el seu AUC és superior al 80%.

6 Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

En aquest estudi s'han vist 3 anàlisis diferents. En primer lloc, s'ha dut a terme un constrast d'hipòtesi amb la qualitat mitjana del vi. S'ha vist com la qualitat mitjana del vi negre *Vinho Verde* es superior a 5. Concretament, el valor mitjà de la qualitat d'aquest vi està comprés dins de l'interval [5.59641, 5.6756351] amb una confiança del 95%. Per tant, la resposta a la pregunta plantejada sobre si la qualitat mitjana del vi és acceptable és afirmativa.

A continuació s'han plantejat dos problemes relacionats amb una regressió lineal i una regressió logística. En el primer cas, s'ha intentat crear un model predictiu que a través de les qualitats fisico-químiques del vi determinés la qualitat d'aquest. El millor model lineal obtingut en aquest estudi ha estat el següent:

$$quality = 4.4300987 + 0.2893028 * AL - 1.0127527 * VA - 2.0178138 * CHL$$
$$-0.0034822 * TSD + 0.8826651 * SUL + 0.0050774 * FSD - 0.4826614 * PH$$

Si més no, aquest model té un coeficient de determinació ajustat de 0.36, el qual és molt baix. Aquest paràmetre indica que la bondat de l'ajust és baixa i per tant, el model no serveix com a model predictor.

A continuació, s'ha intentat veure si es millorava aquest model lineal mitjançat un anàlisi de components principals previ. S'ha creat un model amb les 7 primeres components, ja que expliquen més del 90% de la variancia de la mostra. En aquest cas s'ha obtingut un coeficient de determinació ajustat de 0.34, el qual és més baix que en el cas anterior, per tant, no s'ha millorat el model. Cal comentar que realitzant un anàlisi de correlacions entre les variables s'ha trobat que aquestes no están altament correlacionaes, per tant, és possible que altres mètodes de predicció siguin més eficients.

Per últim, s'ha creat un model de regressió logística que intenta predir si un vi és de bona o mala qualitat, es a dir, si la seva qualificació és inferior o és igual o superior a 5. El millor model logístic trobat en aquest estudi ha estat el següent:

$$\begin{gathered} logit(goods) = -388.4120898 + -0.4822986 * FA - 4.5840949 * VA - 0.9643682 * CA \\ -0.3321741 * RS + -6.2934795 * CHL + 0.019671 * FSD \\ 0.0146852 * TSD + 411.3683162 * D - 5.7197375 * SUL + 1.1297783 * AL \end{gathered}$$

Mitjançant el gràfic de la corba ROC i el paràmtere AUC (82.51%) s'ha vist com és un model acceptable, ja que l'AUC supera el 80%. Aquest model podrà pedrir si un vi negre *Vinho Verde* té una qualitat superior o inferior a 5, es a dir, si està considerat bo o dolent, a partir de les propietats fisico-químiques del vi.

Com a última conclusió, a través de l'observació de tots els resultats obtinguts en aquest estudi, és raonable considerar la possible aplicació d'anàlisis posteriors utilitzant eines més potents de *machine learning*, per tal de millorar aquests resultats. Per exemple, es podria realitzar un estudi més avançat aplicant diferents algoritmes de classifiació com el *RandomForest*, *Stochastic Gradient Descent Classifier*, *Support Vector Classifier (SVC)*, etc.

7 Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi es pot trobar al següent enllaç de Github:

- https://github.com/albert-torres/redwine-quality-data-analyser

Per altra banda, els arxius que contenen les dades originals i les dades processades (encara que no s'han aplicat gaires modificacions) es poden tobar en d'aquest projecte dins el subdirectori */data*. Els arxius han estat anomenats *winequality-red.csv* i *winequality-red-processed.csv*, respectivament.

```
write.csv(wineData, "../data/winequality-red-processed.csv")
```

Referències

[1] A. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009." [Online]. Available: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

[2] D. Dalpiaz, "Applied Statistics with R." [Online]. Available: https://daviddalpiaz.github.io/appliedstats/collinearity.html.

[3] A. Rodríguez and C. García, "El Factor de Inflacion de la Varianza en R." [Online]. Available: http://r-es.org/9jornadasR/pdf/9JUR_paper_31.pdf.