Práctica 2: Limpieza y validación de los datos

Tipología y ciclo de vida de los datos

Jose María Bernet Fernández

7 de January, 2019

Contents

1 Descripción del dataset

El dataset ha sido elegido desde la plataforma Kaggle (enlace) corresponde a una colección de 1599 (registros) vinos portugueses donde se describen 12 (columnas) caracteristicas distribuidos en fisicoquímicas (entradas) y sensorial (salida). Las columnas del dataset son las siguientes:

- **fixed.acidity** : Nivel de acidez del vino.
- **volatile.acidity** : Nivel de acido acético, mucha cantidad de este ácido provoca un sabor amargo.
- **critic.acid** : Nivel de cítricos.
- **residual.sugar** : Cantidad de azucar que queda en el vino una vez termina la fermentación.
- **chlorides** : Cantidad de sal que contiene el vino.
- **free.sulfur.dioxide** : Cantidad de dióxido de azufre, previene el crecimiento microbiano y la oxidación del vino.
- **total.sulfur.dioxide** : Nivel de acidez del vino.
- **density** : Densidad.
- **ph** : Nivel de PH.
- **sulphates** : Nivel de sulfatos.
- **alcohol** : Porcentaje de alcohol.
- **quality** : Calidad del vino, basado en datos sensoriales.

Gracias a este dataset, podemos deducir que tipo de vino podría tener mejor calidad que otro, tomando cómo muestra los datos fisicoquimicos de un vino cualquiera que analizemos y en base a los datos sensoriales de calidad que tenemos.

De este modo, con la información disponible podemos poner a disposición de las diferentes empresas que elaboran vino las cantidades y porcentajes exactos en la elaboración para intentar conseguir unos valores de calidad superiores, en función de los analisis que tenemos.

2 Integración y selección de los datos de interés a analizar

El dataset se encuentra en formato CSV, este fichero lo cargaremos y volcaremos sobre un dataframe, en principio las 12 columnas que disponemos van a ser utilizadas ya que no existen columnas de identificación ni valores que sobren, más adelante veremos si podemos prescindir de alguna de las columnas por que no sea lo suficientemente relevante.

```
# Cargamos el csv a un dataframe que denominaremos wine

wine <- read.csv(file="winequality-red.csv", header=TRUE, sep=",",
                         strip.white=TRUE)
```

3 Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Antes de comprobar si existen valores NA, haremos un summary para ver los tipos de datos que disponemos y ver si existe algun valor extraño, en principio, los valores son todos númericos y estan acotados entre dos intervalos por los que no deberíamos de efectuar ninguna operación.

```
# Hacemos el summary de los datos

kable(summary(wine)) %>% kable_styling(latex_options="scale_down")
```

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
	Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900	Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901	Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
	1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900	1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
	Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200	Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968	Median :3.310	Median :0.6200	Median :10.20	Median :6.000
	Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539	Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967	Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
	3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600	3rd Qu.:0.09000	3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9978	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
	Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500	Max. :0.61100	Max. :72.00	Max. :289.00	Max. :1.0037	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

```
# Consultamos el tipo de variable

kable(sapply(wine, function(x) class(x)))
```

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

```
# Miramos los valores mínimos y máximos de cada variable

kable(t(sapply(wine, function(x){ c(min(x),max(x))})))
```

fixed.acidity	4.60000	15.90000
volatile.acidity	0.12000	1.58000
citric.acid	0.00000	1.00000
residual.sugar	0.90000	15.50000
chlorides	0.01200	0.61100
free.sulfur.dioxide	1.00000	72.00000
total.sulfur.dioxide	6.00000	289.00000
density	0.99007	1.00369
pH	2.74000	4.01000
sulphates	0.33000	2.00000
alcohol	8.40000	14.90000
quality	3.00000	8.00000

Podemos comprobar que todas las variables son de tipo numerico menos la calidad que es un entero, además los valores mínimos y máximos no nos hacen sospechar de que hubiese datos corruptos. Comprobamos ahora si existiese algún valor NA, además miraremos si existiese algún cero, en el resumen de máximos y mínimos ya vimos que solo podía haber una columna con ceros que es el citric.acid pero esto es normal, ya que el vino puede tener un contenido cítrico o no, depende del tipo de vino.

```
# Comprobamos valores vacios
kable(sapply(wine, function(x) sum(is.na(x))))
```

	x
fixed.acidity	0
volatile.acidity	0
citric.acid	0
residual.sugar	0
chlorides	0
free.sulfur.dioxide	0
total.sulfur.dioxide	0
density	0
pH	0
sulphates	0
alcohol	0
quality	0

Podemos ver que tamopco tenemos ningún valor NA. En el caso de haber algún valor vació o algún cero en una columna donde no debiese de estar, tendremos primero que ver la importancía de la columna en sí para el resultado final, si es una columna relevante lo mejor será eliminar la muestra para no falsear los datos, esto es posible por que tenemos un dataset bastante grande (1599) en un conjunto más pequeño sería una perdida importante, si por el contrario la columna no es relevante para nuestro estudio podríamos dejar el registro, todo depende del grado de error que estemos dispuesto a asumir.

3.2 Identificación y tratamiento de valores extremos.

El único dato que se sale un poco de la normalidad es el total.sulfur.dioxide,de todos modos comprobaremos ahora mediante boxplot si los datos contienen valores extremos (outliers). Haremos uso de la función boxplot.stats dentro de sapply para ver de un vistazo todos los valores que se consideran extremos en primera instancia.

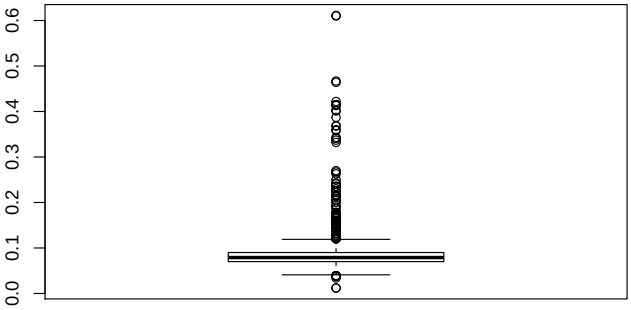
```
# Ejecutamos el boxplot.stats para todas las columnas
sapply(wine, function(x) boxplot.stats(x)$out)
```

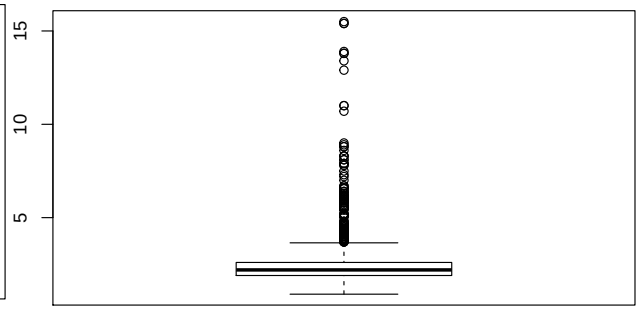
```
$fixed.acidity
 [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8
[15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4
[29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2
[43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6

$volatile.acidity
 [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
[12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040

$citric.acid
[1] 1

$residual.sugar
  [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
 [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
 [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
 [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
 [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
 [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
 [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
 [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
 [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
[100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
[111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
[122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
[133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
[144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
[155]  7.80

$chlorides
  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178
 [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343
 [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122
 [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171
 [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157
 [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126
 [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120
 [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136
 [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415
[100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235
[111] 0.230 0.038

$free.sulfur.dioxide
 [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51
[24] 51 52 55 55 48 48 66

$total.sulfur.dioxide
 [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127
[18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148
[35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147
[52] 147 131 131 131

$density
```

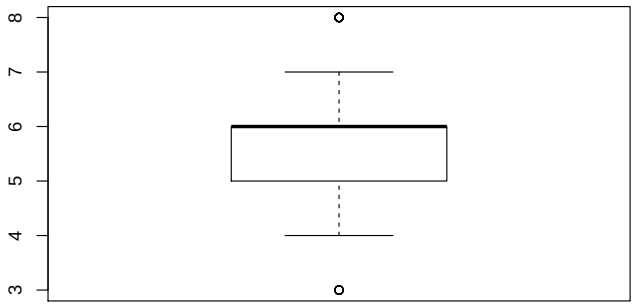
```
 [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220
 [9] 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140
[17] 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260
[25] 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007
[33] 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369
[41] 1.00369 1.00242 0.99182 1.00242 0.99182

$pH
 [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87
[15] 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78
[29] 4.01 2.90 4.01 3.71 2.88 3.72 3.72

$sulphates
 [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08
[15] 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13
[29] 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17
[43] 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03
[57] 1.17 1.10 1.01

$alcohol
 [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
 [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000

$quality
 [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

Descubrimos que hay muchisimos valores considerados extremos, pero esto no quiere decir que todos los valores sean erroneos, sino que la mayoría de muestras del dataset se comprenden en un rango definido de valores y hay otras tantas que se salen de la media, no por ello son errores, miramos por ejemplo un par de boxplots que tengan bastantes balores extremos y hacemos también summary para comprobar los cuantiles.

```r
# Dibujamos boxplot
boxplot(wine$chlorides)
boxplot(wine$residual.sugar)
boxplot(wine$quality)
```








```
# Mostramos summary
summary(wine$chlorides)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

```
summary(wine$residual.sugar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.900   1.900   2.200   2.539   2.600  15.500
```

```
summary(wine$quality)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   6.000   5.636   6.000   8.000
```

Podemos ver mirando los cúantiles y las gráficas cómo la mayoría de muestras se engloban en pequeñas cantidades, por ejemplo, el primer cuantíl y el tercero de los chlorides son 0.07 y 0.09 respectivamente, sin embargo su valor máximo son 0.611. Dado que no tenemos los conocmientos químicos para asegurar que los datos extremos son erroneos, debemos de suponer que existen tipos de vinos que tienen caracteristicas diferentes a los otros, por lo que en este caso nos quedaremos con los datos tal y como estan.

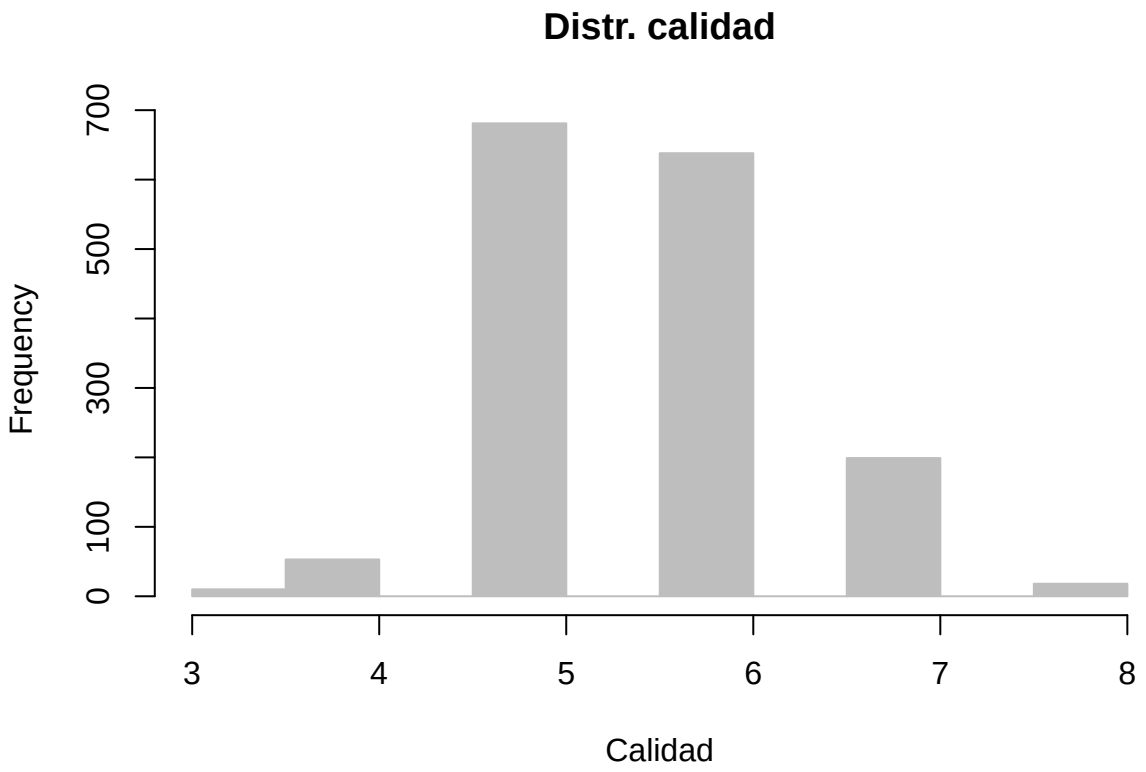
---

4 Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para entender un poco más los datos que queremos usar en nuestro analisis vamos a mostrar una seríe de gráficas para ver cómo se distribuyen algunos datos.

```
hist(wine$quality,col='gray',border='gray',main='Distr. calidad',xlab='Calidad')
```

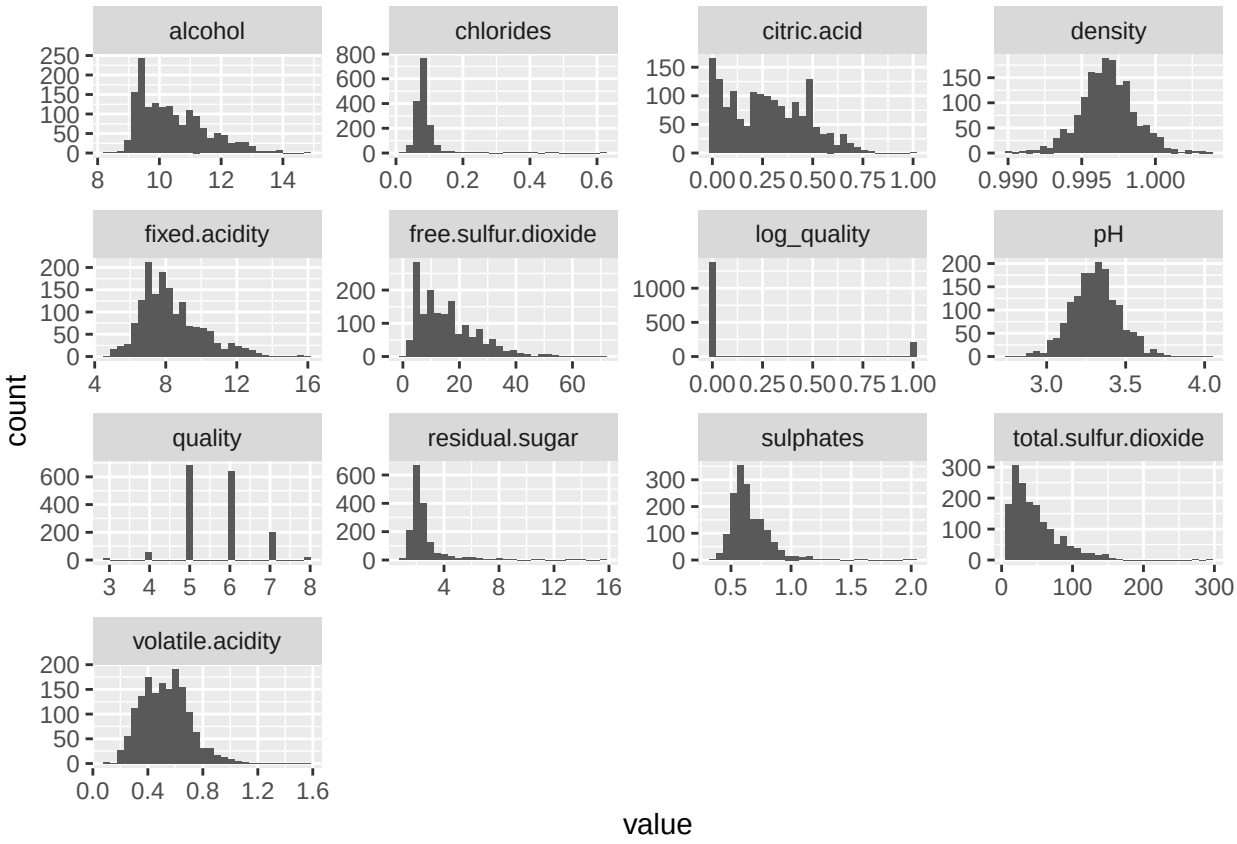


Lo primero que vemos es el histográma de la calidad, con este gráfico podemos ver la distribución de las diferentes calificaciones que han obtenido los vinos, si nos fijamos, practicamente la mitad se puede encontrar entre el 5 y el 6 (si consultamos el summary que hicimos anteriormente vemos cómo la media es 5.636), por lo que para realizar nuestros análisis podemos decir que un vino con puntuación menor o igual que 6 es considerado de menor calidad y un vino con puntuación superior o igual a 6 será considerado de una calidad mejor.

Con la idea de usarlo posteriormente en un modelo de regresión logística, prepararemos una nueva columna del dataframe con 0 y 1, indicando si el vino es de peor, o mejor calidad, usaremos como punto nota de refencia el 6.

```
wine$log_quality <- ifelse(wine$quality <=6 ,0, 1)
```

```
# Vemos el histograma de todas las columnas
wine %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```



Al mostrar el histograma de todas las columnas se pueden preveer que algunas tienen muchas posibilidades de seguir una distribución normal.

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Para comprobar si nuestros atributos del dataset siguen una distribución normal haremos uso del test de Shapiro-Wilks. Existen diversas gráficas que sirven también para comprobarlo pero aunque parezcan datos seguros, siguen dejando lugar a la interpretación, con el test de Shapiro-Wilks simplemente miraremos nos devuelva un p-valor mayor o igual que 0.5 para poder decir que sigue una distribución normal.

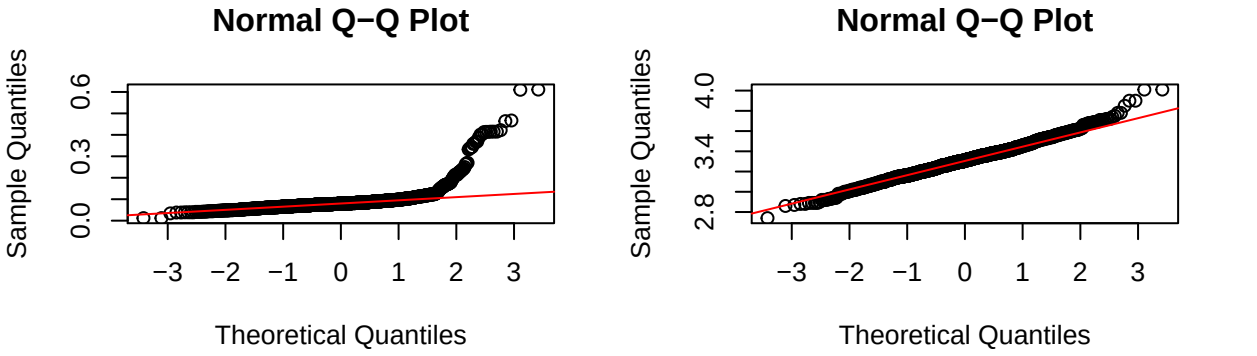
```
# Usamos Sapply para realizar el test de saphiro en todas las columnas del dataset.
sapply(wine, function(x){ (shapiro.test(x))$p.value})
```

```
       fixed.acidity     volatile.acidity          citric.acid
        1.525012e-24         2.692935e-16         1.021932e-21
      residual.sugar            chlorides  free.sulfur.dioxide
        1.020162e-52         1.179056e-55         7.694597e-31
total.sulfur.dioxide              density                   pH
        3.573451e-34         1.936053e-08         1.712237e-06
           sulphates              alcohol              quality
        5.823140e-38         6.644057e-27         9.515085e-36
         log_quality
        3.726209e-58
```

Podemos ver cómo todos los atributos tienen un p-valor muy inferior al 0.5, por lo que consideramos que se alejan bastante de una distribción normal. Podemos ver ahora cómo se podría hacer esta comprobación

de manera gráfica con algunos atributos por ejemplo chlorides y pH, que son los que tienen el p-valor más distante, para ver las diferencias.

```
par(mfrow=c(2,2))

qqnorm(wine$chlorides)
qqline(wine$chlorides, col = "red")

qqnorm(wine$pH)
qqline(wine$pH, col = "red")
```



Podemos ver cómo el pH, al ser el que tiene un p-valor más grande, muestra un parecido mayor a la normal (los puntos se aproximan más a la linea roja) sin embargo en la otra gráfica, la de chlorides, vemos cómo se aleja enormemente de la linea roja.

Para comprobar ahora la homogeneidad de la varianza tenemos varías funciones disponibles, haremos uso del test de Fligner-Killeen que es el más recomendado cuando no tenemos datos con distribución normal, para realizar el test necesitaremos comparar dos conjuntos de datos, enfrentaremos por tanto aquellos registros que

```
fligner.test(pH ~ log_quality, data = wine)
```

```
    Fligner-Killeen test of homogeneity of variances

data:  pH by log_quality
Fligner-Killeen:med chi-squared = 0.085167, df = 1, p-value =
0.7704
```

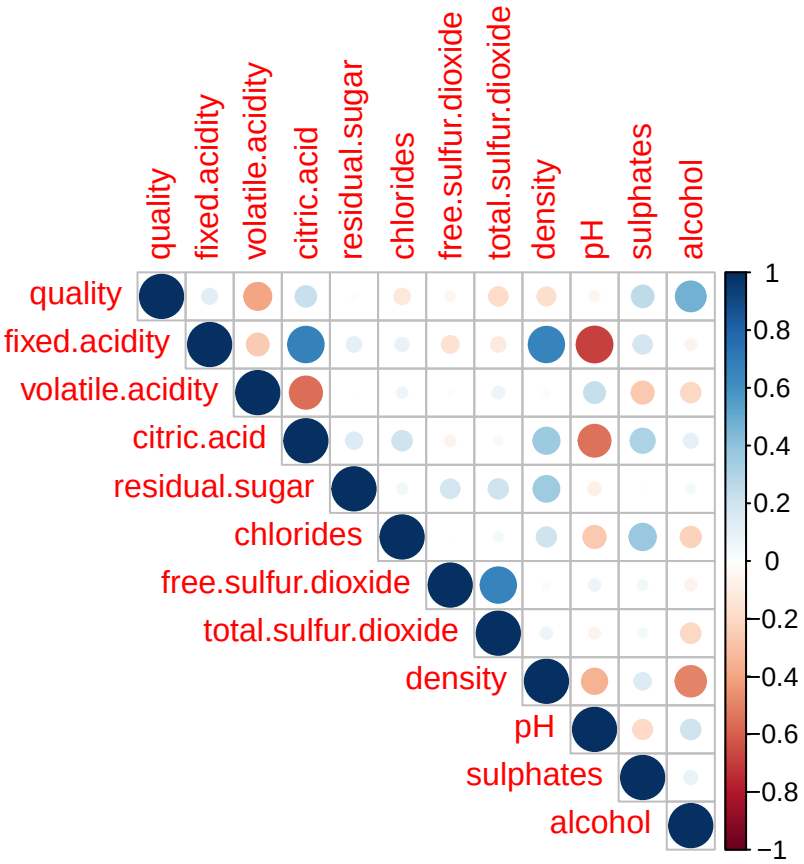
4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

4.3.1 Correlación

Cómo nuestro objetivo del estudio se basa principalmente en identificar qúe componentes fisicoquimicos afectán más directamente a la calidad del vino empezaremos el análisis mostrando una matriz de correlación de las variables respecto la calidad.

```
wine_1 <- wine[c('quality','fixed.acidity','volatile.acidity','citric.acid',
                 'residual.sugar','chlorides','free.sulfur.dioxide',
                 'total.sulfur.dioxide','density','pH','sulphates','alcohol')]
```

```
o=corrplot(cor(wine_1), method='circle', type='upper')
```



Vemos como el corrplot nos muestran en la primera fila los grados de correlación entre las distintas variables y la calidad del vino, el tamaño del circulo indica el nivel de significación y el color si es una relación inversa o directa, es decir, si a mayor cantidad de un componente la calidad subira o bajará. Aunque no haya ninguna correlación extremadamente fuerte, podemos ver cómo hay 4 elementos que tienen un circulo mayor, estos son alcohol, sulphates, citric,acid y volatile.acidity.

4.3.2 Contraste de Hipótesis

Continuando con los analisis estadisticos, pasamos ahora a el contraste de hipotesis, queremos saber si las bebidas que tienen más concentración de alcohol son propensas a tener una mayor calidad, tal y como se ha podido preveer viendo la matriz de correlación, suponemos entonces que:

$H_0 : \mu_1 - \mu_2 > 0$ $H_1 : \mu_1 - \mu_2 =< 0$

```
wine.better.alcohol <- wine[wine$quality >= 6,]$alcohol
wine.worse.alcohol <-wine[wine$quality < 6,]$alcohol
```

De este modo tenemos que μ_1 corresponde a la media de alcohol de bebidas con una calidad inferior y μ_2 corresponde a la media de alcohol de bebidas con una calidad superior, tomamos $\alpha = 0.05$.

```
t.test(wine.worse.alcohol, wine.better.alcohol , alternative = "less")
```

```
    Welch Two Sample t-test

data:  wine.worse.alcohol and wine.better.alcohol
t = -19.782, df = 1516.8, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.8512962
sample estimates:
mean of x mean of y
 9.926478 10.855029
```

El p-valor resultado es 0.00000000000000022 muy inferior al valor de significaciín 0.05, por lo que tenemos que rechazar la hipotesis nula a favor de la hipotesis alternativa, concluyendo que las bebidas con una mayor concentración de alcohol son bebidas de mayor calidad.

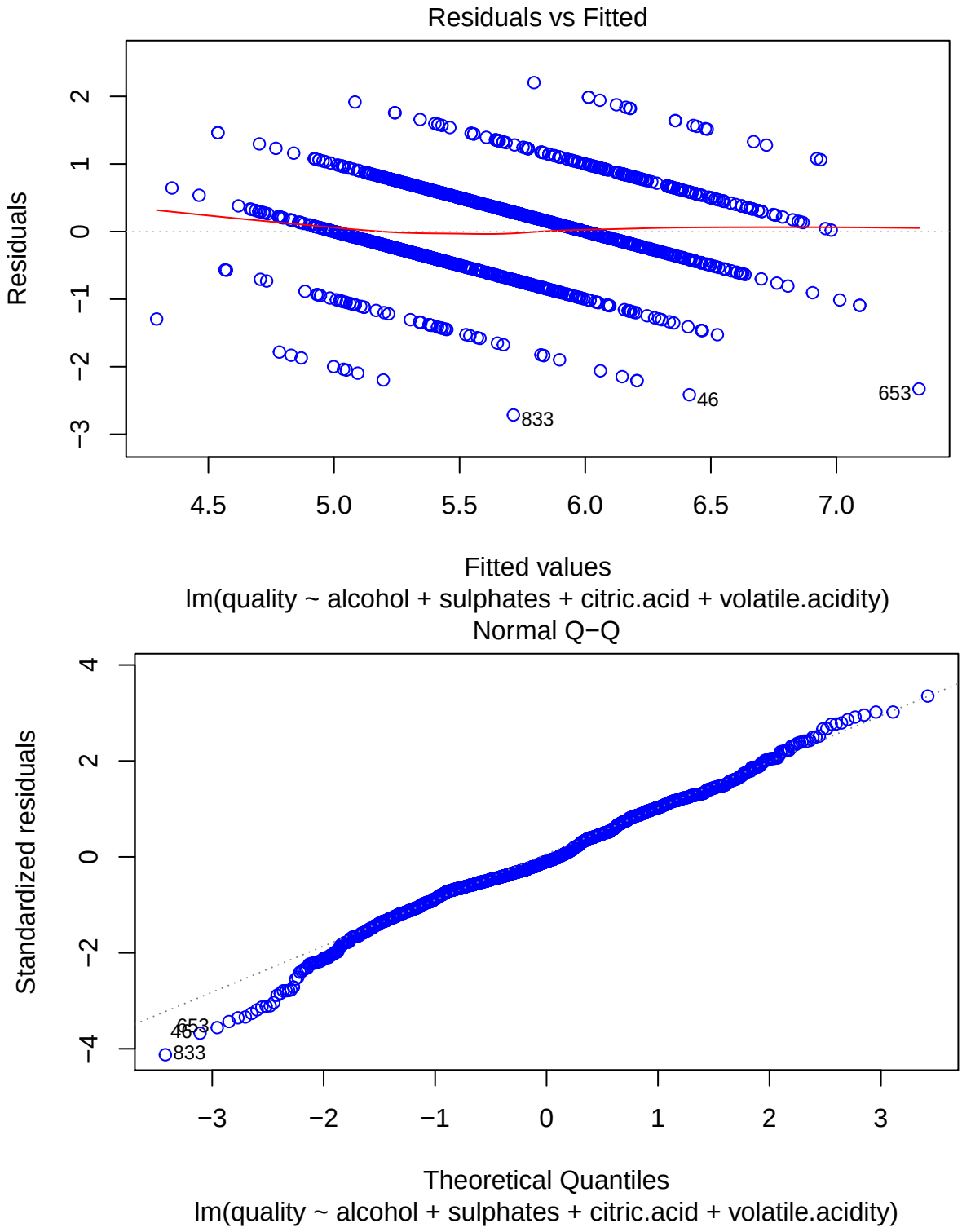
4.3.3 Regresión

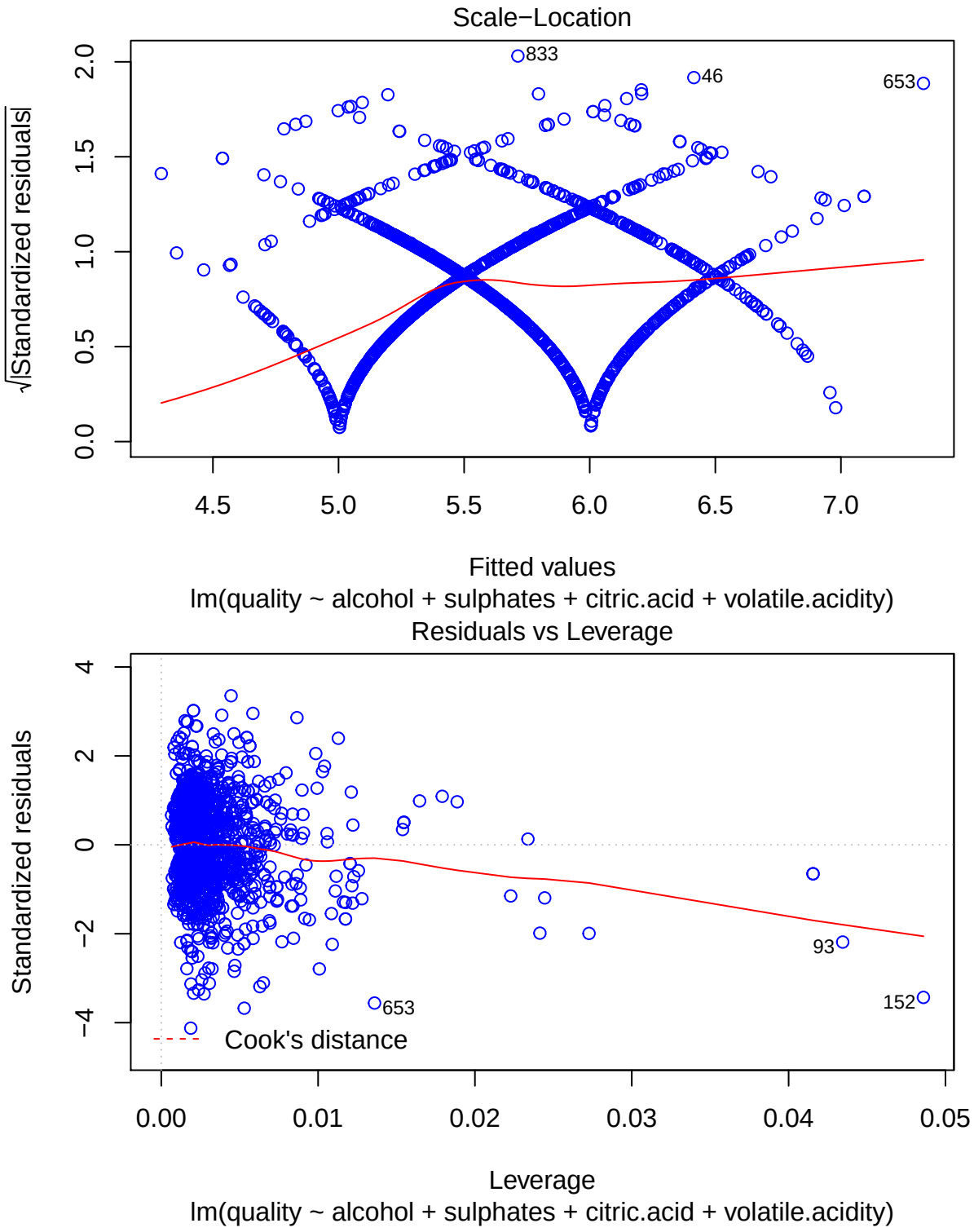
Con el fin de predecir la calidad del vino vamos a realizar una serie de modelos de regresión lineal probando los diferentes atributos que más afectaban a la calidad, como vimos en la matriz de correlación, usaremos por tanto: alcohol, sulphates, citric.acid y volatile.acidity.

```r
modelo_1 <- lm(quality ~  alcohol + sulphates + citric.acid + volatile.acidity, data = wine)
summary(modelo_1)
```

```
Call:
lm(formula = quality ~ alcohol + sulphates + citric.acid + volatile.acidity,
    data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.71408 -0.38590 -0.06402  0.46657  2.20393

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)       2.64592    0.20106  13.160 < 0.0000000000000002 ***
alcohol           0.30908    0.01581  19.553 < 0.0000000000000002 ***
sulphates         0.69552    0.10311   6.746      0.0000000000212 ***
citric.acid      -0.07913    0.10381  -0.762                0.446
volatile.acidity -1.26506    0.11266 -11.229 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6588 on 1594 degrees of freedom
Multiple R-squared:  0.3361,    Adjusted R-squared:  0.3345
F-statistic: 201.8 on 4 and 1594 DF,  p-value: < 0.00000000000000022
```

```r
plot(modelo_1, col='blue')
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
4.5
5.0
5.5
6.0
6.5
7.0
46
653
833
Fitted values
lm(quality ~ alcohol + sulphates + citric.acid + volatile.acidity)
Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
653
46
833
Theoretical Quantiles
lm(quality ~ alcohol + sulphates + citric.acid + volatile.acidity)

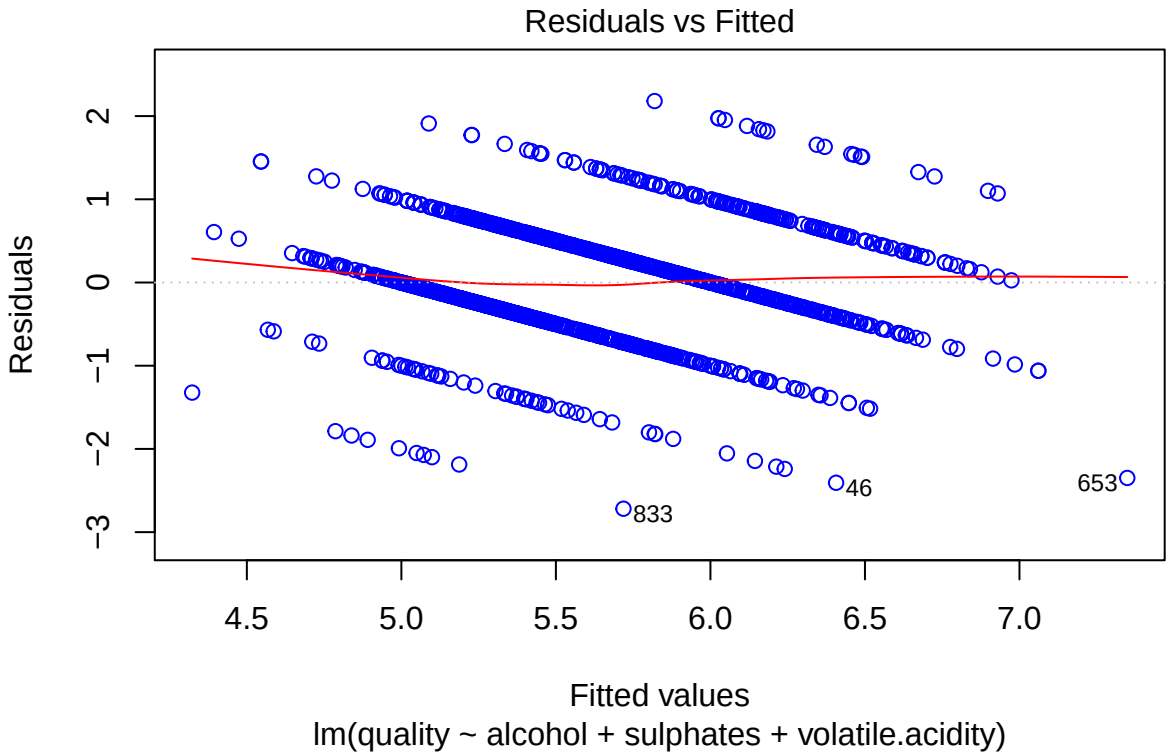


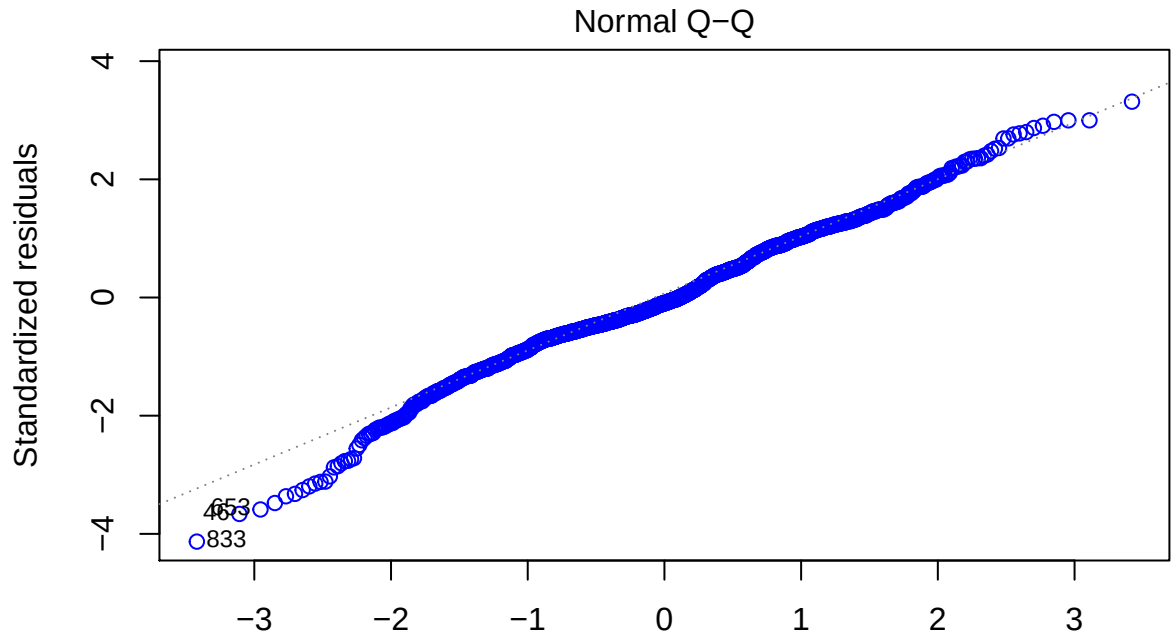
Haciendo summary del modelo podemos ver que la calidad R2 del ajuste es 0.3361, un valor bastante distante del 1, por lo que la calidad del modelo no es muy buena, nos fijamos ahora en los P-valores de las diferentes variables y descubrimos que todas excepto citric.acid tienen una influencia significativa con un p-valor inferior a un ajuste de 0.05, realizamos el modelo sacando la variable citric.acid.

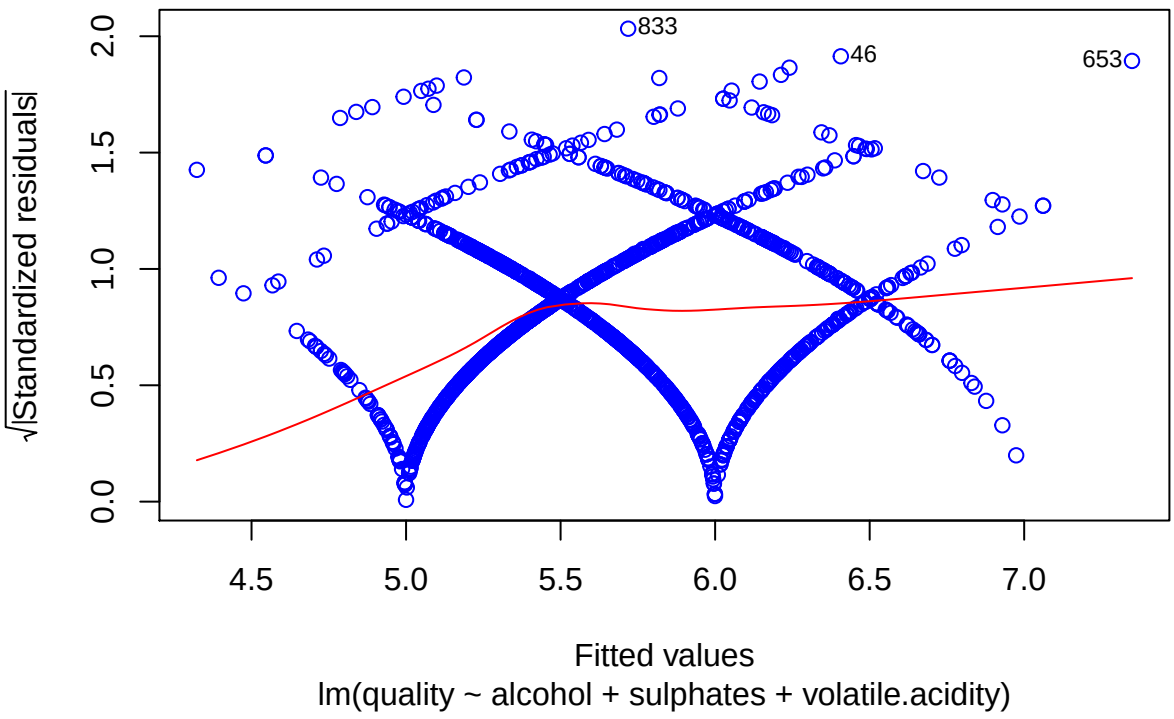
```r
modelo_2 <- lm(quality ~  alcohol + sulphates + volatile.acidity, data = wine)
summary(modelo_2)
```

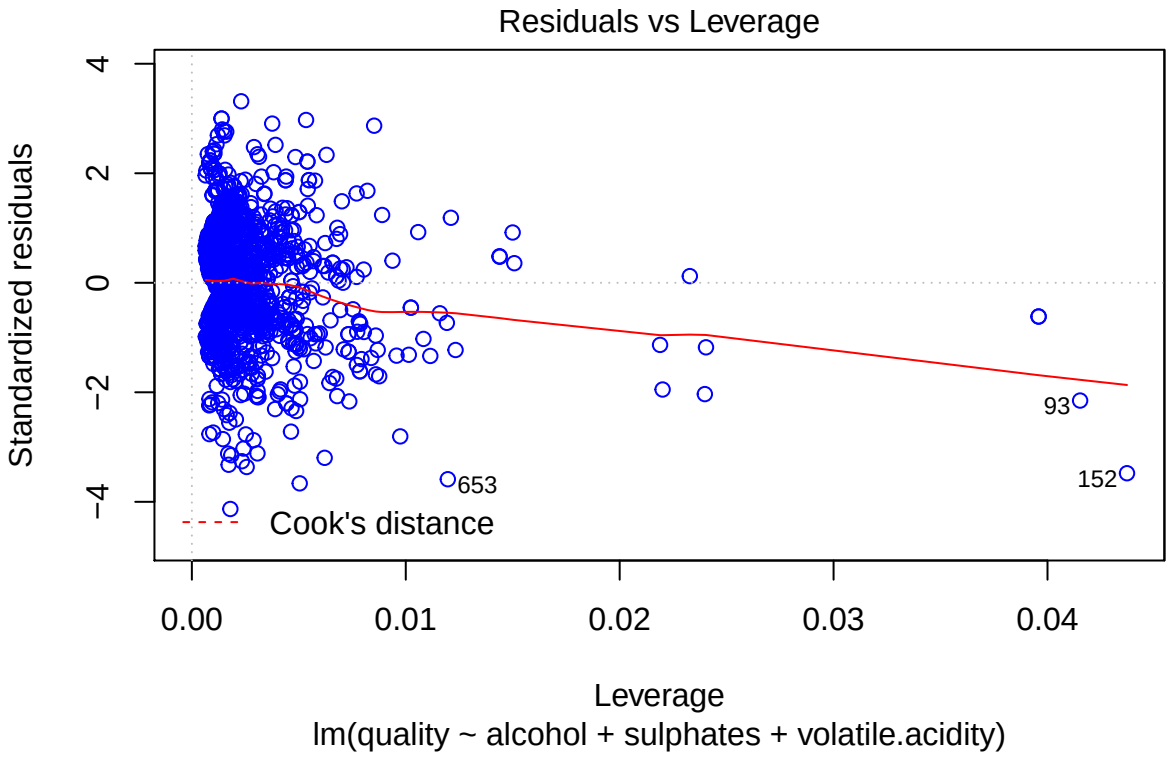
```
Call:
lm(formula = quality ~ alcohol + sulphates + volatile.acidity,
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7186 -0.3820 -0.0641  0.4746  2.1807

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)      2.61083    0.19569  13.342 < 0.0000000000000002 ***
alcohol          0.30922    0.01580  19.566 < 0.0000000000000002 ***
sulphates        0.67903    0.10080   6.737      0.0000000000226 ***
volatile.acidity -1.22140    0.09701 -12.591 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6587 on 1595 degrees of freedom
Multiple R-squared:  0.3359,    Adjusted R-squared:  0.3346
F-statistic: 268.9 on 3 and 1595 DF,  p-value: < 0.00000000000000022
```

```r
plot(modelo_2, col='blue')
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−4
653
46
833
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(quality ~ alcohol + sulphates + volatile.acidity)


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
833
46
653
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(quality ~ alcohol + sulphates + volatile.acidity)



Cómo vemos ninguno de los modelos tiene una buena calidad del ajuste, vamos a probar ahora un modelo de regresión logistica para adivinar si el vino es de calidad o no, basado en nuestro indicador log_quality creado anteriormente.

```
modelo_3 <- glm(log_quality ~  alcohol + sulphates + citric.acid + volatile.acidity,data=wine,family=bi
summary(modelo_3)
```

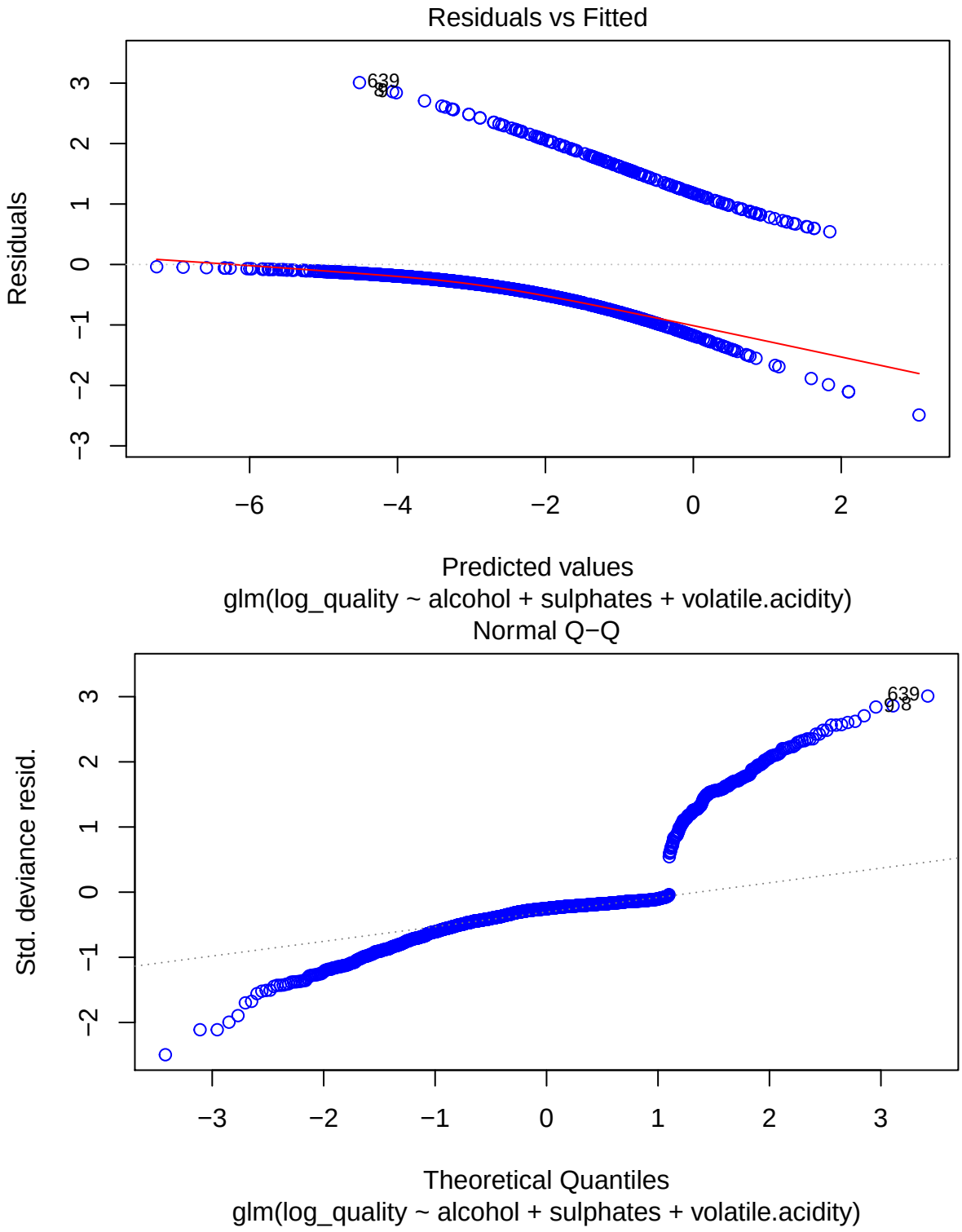
```
Call:
glm(formula = log_quality ~ alcohol + sulphates + citric.acid +
    volatile.acidity, family = binomial(), data = wine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5765  -0.4606  -0.2440  -0.1530   2.9777

Coefficients:
                Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -13.1586     1.1252 -11.694 < 0.0000000000000002 ***
alcohol            1.0114     0.0814  12.424 < 0.0000000000000002 ***
sulphates          2.4424     0.4488   5.442          0.0000000527 ***
citric.acid        0.9378     0.5270   1.780                0.0751 .
volatile.acidity  -3.5839     0.6854  -5.229          0.0000001704 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1269.92  on 1598  degrees of freedom
```

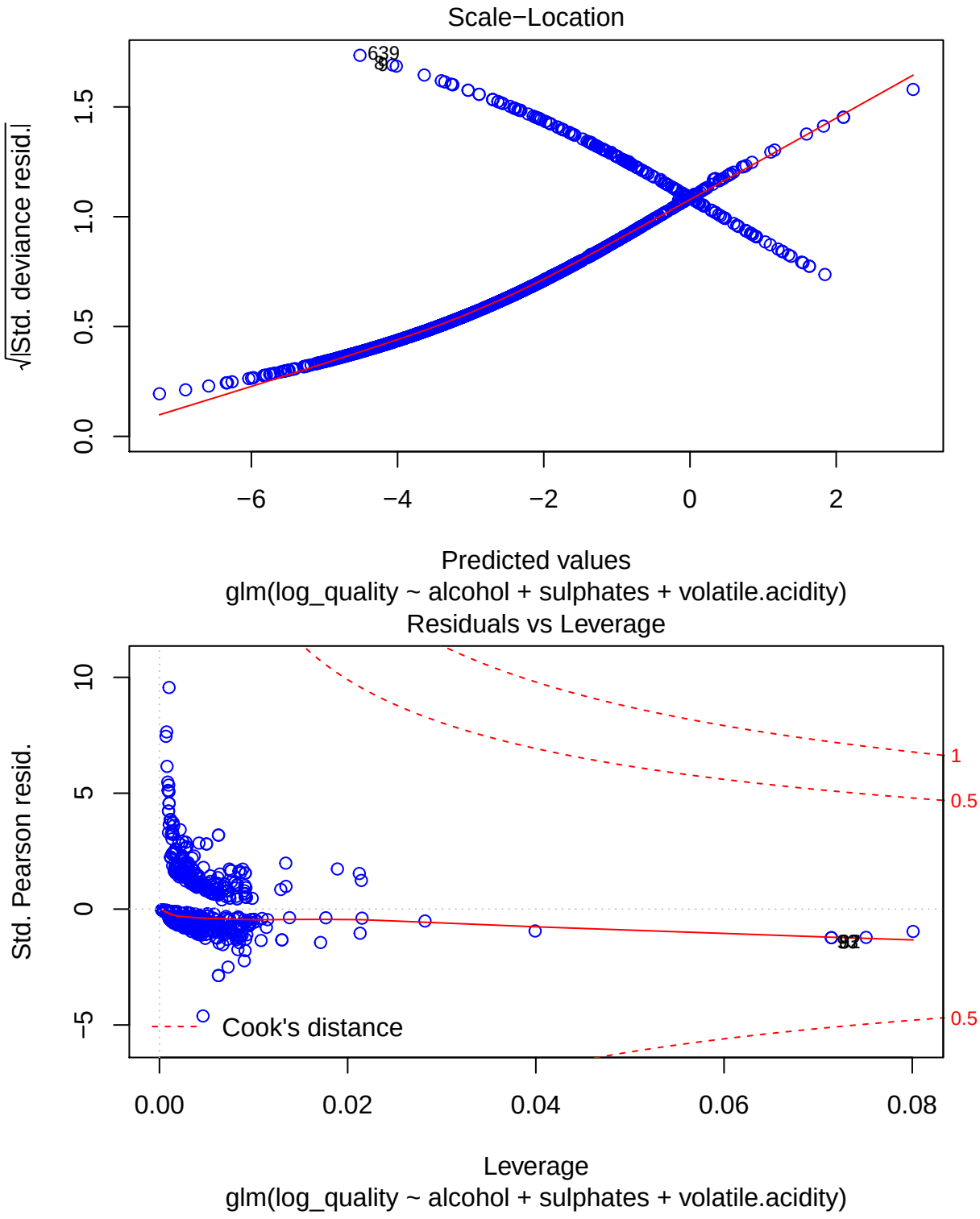
```
Residual deviance:  914.11  on 1594  degrees of freedom
AIC: 924.11

Number of Fisher Scoring iterations: 6
```

```
modelo_4 <- glm(log_quality ~  alcohol + sulphates  + volatile.acidity,data=wine,family=binomial())
summary(modelo_4)
```

```
Call:
glm(formula = log_quality ~ alcohol + sulphates + volatile.acidity,
    family = binomial(), data = wine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4893  -0.4563  -0.2519  -0.1543   3.0084

Coefficients:
                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)      -12.62864    1.07668 -11.729 < 0.0000000000000002 ***
alcohol            1.00788    0.08132  12.394 < 0.0000000000000002 ***
sulphates          2.59428    0.44229   5.866     0.00000000447751 ***
volatile.acidity  -4.21024    0.59450  -7.082     0.00000000000142 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1269.92  on 1598  degrees of freedom
Residual deviance:  917.26  on 1595  degrees of freedom
AIC: 925.26

Number of Fisher Scoring iterations: 6
```

```
plot(modelo_4, col='blue')
```


Residuals vs Fitted
Residuals
Predicted values
glm(log_quality ~ alcohol + sulphates + volatile.acidity)
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(log_quality ~ alcohol + sulphates + volatile.acidity)



Vemos ahora la matriz de confusión para analizar la calidad del modelo, usamos nuestro propio dataset como parametros de referencia mediante la columna log_quality y predecimos usando el modelo_4 que nos ha dado un AIC superior, aunque apenas hay diferencia. Usaremos cómo umbral discriminatorio un 70%, es decir sólo cuando haya más de un 0.7 en la predicción, tomaremos el valor cómo bueno.

```r
wine$qualityPredictedRaw <- predict(modelo_4, newdata=wine,type="response")
wine$qualityPredicted <- as.factor(ifelse(predict(modelo_4, newdata=wine) > 0.7, 1, 0))
wine$log_quality <- as.factor(wine$log_quality)


confusionMatrix(wine$qualityPredicted,wine$log_quality)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1371  192
         1   11   25

               Accuracy : 0.873
                 95% CI : (0.8557, 0.889)
    No Information Rate : 0.8643
    P-Value [Acc > NIR] : 0.1622

                  Kappa : 0.1654
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.9920
            Specificity : 0.1152
         Pos Pred Value : 0.8772
         Neg Pred Value : 0.6944
             Prevalence : 0.8643
         Detection Rate : 0.8574
   Detection Prevalence : 0.9775
      Balanced Accuracy : 0.5536

       'Positive' Class : 0
```

Podemos ver que tenemos 11 falsos positivos y 192 falsos negativos y una perfección del 87.3%.

5 Conclusiones

Después de realizar el estudio, para el cúal se han desarrollado diversos modelos de regresión logistica, regresión lineal y matriz de correlaciones, hemos podido identificar que existen caracteristicas fisicoquímicas que hacen aumentar la calidad del vino, estos son el alcohol, los sulfatos, el ácido cítrico y la acidez volátil.

El mejor modelo de regresión que hemos encontrado nos ha dado unas predicciones con un acierto del 87,3% lo cual, aún siendo un buen resultado, estariamos hablando de asumir un error del 12,7%.

El problema principal al que nos hemos enfrentado es la poca variedad que hay en la columna de calidad, donde la mayoria de registros se encontraban en los valores 5 y 6, si hubiesemos tenido una distribución más regular podríamos a ver conseguido mejores resultados.

En definitiva, aunque no hayamos dado con un mejor modelo ni hayamos encontrado la receta del vino perfecto, hemos encontrado las caracteristicas que a grosso modo, hacen que el vino sea mejor.