Trabajo Final

Álvaro Muñoz Ruiz Almudena Luque Castro María Granero Alarcón

2 de junio de 2025

UNIVERSIDAD
DE GRANADA

Estadística Computacional

Índice

1 Introducción **3**

2 Los datos **3**

3 El modelo **4**

4 Análisis estadístico **5**
4.1 Regresión lineal múltiple 7
4.2 Regresión logística 13

5 Conclusiones **16**

6 Referencias **17**

1. Introducción

El vino es un producto emblemático tanto cultural como económico de diversas naciones, entre ellas España o Portugal, donde su elaboración cuenta con una larga tradición y una gran influencia a nivel internacional. En particular, la región norte de Portugal destaca por la producción de vinos tintos con características distintivas. Durante el proceso de elaboración y evaluación del vino, se registran diversas variables fisicoquímicas que permiten analizar su composición y determinar su calidad proporcionando una base sólida para estudios orientados a comprender los factores que inciden en la valoración final del producto.

El objetivo principal del análisis es identificar posibles relaciones significativas entre las características fisicoquímicas del vino y su calidad. Este conocimiento puede contribuir a mejorar los procesos de producción, permitiendo a los elaboradores optimizar las propiedades del vino en función de los factores que más inciden en su valoración, o, por otro lado, tratar de abaratar los costes sin reducir la calidad del producto. Además, los modelos estadísticos obtenidos podrían emplearse como herramientas predictivas para estimar la calidad de nuevas muestras a partir de sus parámetros fisicoquímicos. Para ello, se aplicarán técnicas de análisis estadístico en R aprendidas durante el curso, como modelos de regresión lineal múltiple y lineales generalizados (en concreto, se aplicará la regresión logística).

2. Los datos

En este contexto, se pretende analizar un conjunto de datos compuesto por 1599 muestras de vino tinto originarias de la región norte de Portugal. Los datos se encuentran en el fichero `winequality-red.csv`, disponible en [1]. En dicho archivo podemos observar que cada muestra se evalúa en base a once variables fisicoquímicas y una adicional denominada `quality`, que recoge la valoración de cada muestra. Continuamos la sección mostrando la información que aporta cada variable en la siguiente tabla:

Variable	**Descripción**
`fixed.acidity`	Acidez no volátil (g/L)
`volatile.acidity`	Concentración de ácidos volátiles (ácido acético) (g/L)
`citric.acid`	Cantidad de ácido cítrico (g/L)
`residual.sugar`	Cantidad de azúcar no fermentado tras la fermentación (g/L)
`chlorides`	Concentración de cloruro de sodio (g/L)
`free.sulfur.dioxide`	Concentración de moléculas de dióxido de azufre que no están unidas a otros compuestos en el vino (mg/L)
`total.sulfur.dioxide`	Cantidad total de dióxido de azufre (mg/L)
`density`	Masa por unidad de volumen del vino (g/cm^3)
`pH`	Describe cuán ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico)
`sulphates`	Concentración de sulfato de sodio (g/L)
`alcohol`	Porcentaje de alcohol por volumen en el vino
`quality`	Calidad de producto

Se observa que las once variables que componen el data frame `winequality-red` son numéri-

cas. De ellas, las variables `fixed.acidity`, `volatile.acidity`, `citric.acid`, `residual.sugar`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide`, `density`, `pH`, `sulphates` y `alcohol`, constituyen las propiedades fisicoquímicas de las muestras y la variable `quality` evaluada en una escala métrica entre 3 (muy mala) y 8 (excelente). Por último, cabe señalar que estos datos han servido como objeto de estudio en diversos análisis. Estos estudios se pueden encontrar en [1] en el apartado 'Discussion'.

3. El modelo

Para el análisis estadístico de los datos de vinos recogidos se ha decidido utilizar los modelos de regresión lineal y de regresión logística. El motivo por el cual se ha decidido usar ambos se debe a la diferente interpretación de resultados que ofrecen, cuya combinación puede resultar rica para obtener conclusiones precisas. Tanto en el caso lineal como en el logístico, los datos a analizar deben cumplir ciertas hipótesis para que el modelo pueda ser aplicado, ya que, en caso opuesto, si bien se puede llegar a errores en el cálculo de los coeficientes, sería más importante el hecho de que estos pierden cierta validez y efectividad.

Aunque las hipótesis varían dependiendo del modelo, sí que comparten algunas: (i) Independencia de las observaciones entre sí. (ii) Ausencia de multicolinealidad, es decir, que las variables medidas no dependen unas de otras. (iii) Ausencia de valores anómalos, como aquellos valores no medidos (NA) o "*outliers*" que hayan de ser eliminados. (iv) Tamaño muestral adecuado.

A los modelos de regresión lineal han de ser sumadas dos hipótesis más: Normalidad de los errores y Homocedasticidad o varianza de los errores constante.

A continuación y una vez nombradas las hipótesis a seguir, se define el modelo de regresión lineal. Dada n observaciones independientes de una variable aleatoria Y, sean $\{Y_1, \dots, Y_n\}$. Estas observaciones siguen entonces un modelo lineal si:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \epsilon_i \quad i = 1, \dots, n \tag{1}$$

donde $\beta_1, \dots, \beta_i$ son parámetros desconocidos, x_{ij} valores medidos de la respectiva variable j dada la observación i y ϵ_i errores aleatorios. De este modo se puede realizar una formulación matricial del modelo tal que:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{2}$$

Esta ecuación matricial puede escribirse tal que $Y = X\beta + \epsilon$, y la matriz X se conoce como matriz del modelo. Este modelo generalizado se denomina "*modelo lineal de Gauss-Markov*" al verificar la condición de error normalizado previamente mencionada: $\epsilon \sim N(0, I_n)$.

Para el caso particular de este estudio, se utilizará el modelo de regresión lineal múltiple, cuya matriz X es de la forma:

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \tag{3}$$

Por lo que el modelo se define entonces como:

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \epsilon_i \tag{4}$$

donde los parámetros $\beta_1, \ldots, \beta_i$ son denominados coeficientes de regresión. De ellos pueden surgir interpretaciones sobre la fuerza que tiene cada variable explicativa sobre la respuesta en el modelo.

Por otra parte, los modelos de regresión logística son diferentes, ya que requieren como hipótesis principal que la variable respuesta sea binaria, ya sea nominal u ordinal. Es por ello que en multitud de ocasiones los datos han de ser transformados para satisfacer esta condición. En este modelo se tiene entonces que la variable respuesta sigue una distribución de Bernouilli tal que $Y_i \sim B(p_i, n_i)$. Para simplificar el desarrollo, supondremos que los niveles de esta variable son $\{0, 1\}$. Es por ello que podemos asumir que, si la variable respuesta depende de las explicativas, se tiene $P(Y = 1) = p(x)$, con $x = (x_1, \ldots, x_n)$.

El modelo de regresión logística múltiple se define entonces como:

$$\text{logit}[p(\boldsymbol{x})] = \log\left[\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right] = \sum_{i=0}^{p} \beta_i x_i = \boldsymbol{\beta' x} \tag{5}$$

De forma que, despejando para $p(x)$:

$$p(\boldsymbol{x}) = \frac{\exp(\sum_{i=0}^{p} \beta_i x_i)}{1 + \exp\left(\sum_{i=0}^{p} \beta_i x_i\right)} = \frac{e^{\boldsymbol{\beta' x}}}{1 + e^{\boldsymbol{\beta' x}}} \tag{6}$$

Este paso es de elevada importancia, ya que observamos que los términos que acompañan a las variables explicativas son de la forma e^{β_i}. Se conocen como "*odds-ratio*", y representan que, dado el incremento en una unidad de la variable explicativa x_i, la probabilidad de éxito ($P(Y = 1)$) aumenta en e^{β_i} frente a la de fracaso (esto se conoce como razón de probabilidades). De nuevo, estos factores son relevantes de cara a la interpretación del modelo.

4. Análisis estadístico

En primer lugar, realizamos un análisis exploratorio de las variables. Representamos el diagrama de puntos de algunas variables (que posteriormente veremos que son de interés), y hacemos un resumen estadístico de estos. En el diagrama de puntos, visto en 1 no se aprecian codependencias entre las variables explicativas ni dependencia de ellas con la variable respuesta.

```
summary(vinos$quality)
```

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

```
#Dado que hay muchos datos, las nubes de puntos apenas son legibles
#Por ello, se toma una muestra suficientemente grande como para
#apreciar comportamiento pero a la vez pueda ser visible.
```

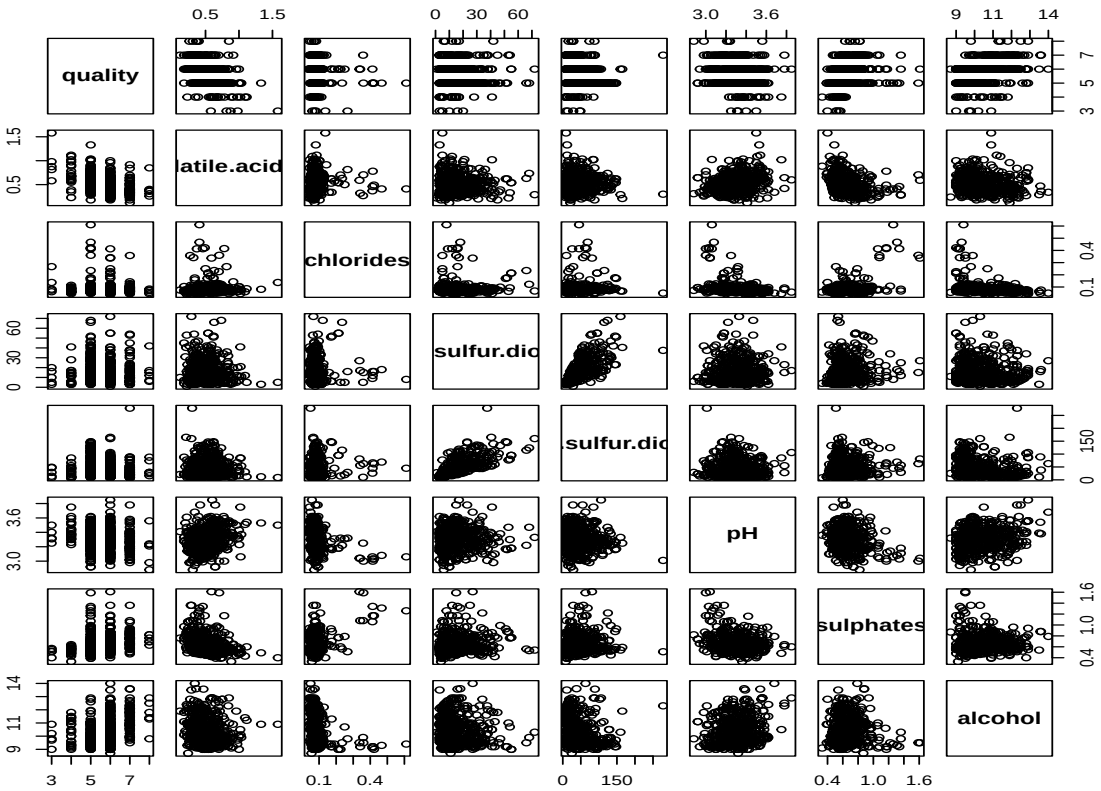

Figura 1: Diagrama de puntos de algunas variables explicativas del problema.

```
x <- sample(1:nrow(vinos), 500)
pairs(vinos[x, variables],
      labels = variables,
      cex.labels = 1.2,
      font.labels = 2)
```

Variable	Min	Mediana	Media	Max
fixed.acidity	4.60	7.9	8.32	15.90
volatile.acidity	0.1200	0.5200	0.5278	1.5800
citric.acid	0.000	0.260	0.271	1.000
residual.sugar	0.900	2.200	2.539	15.500
chlorides	0.01200	0.07900	0.08747	0.61100
free.sulfur.dioxide	1.00	14.00	15.87	72.00
total.sulfur.dioxide	6.00	38.00	46.47	289.00
density	0.9901	0.9968	0.9967	1.0037
pH	2.740	3.310	3.311	4.010
sulphates	0.3300	0.6200	0.6581	2.0000
alcohol	8.40	10.20	10.42	14.90
quality	3.000	6.000	5.636	8.000

Tabla 1: Resumen estadístico de las variables

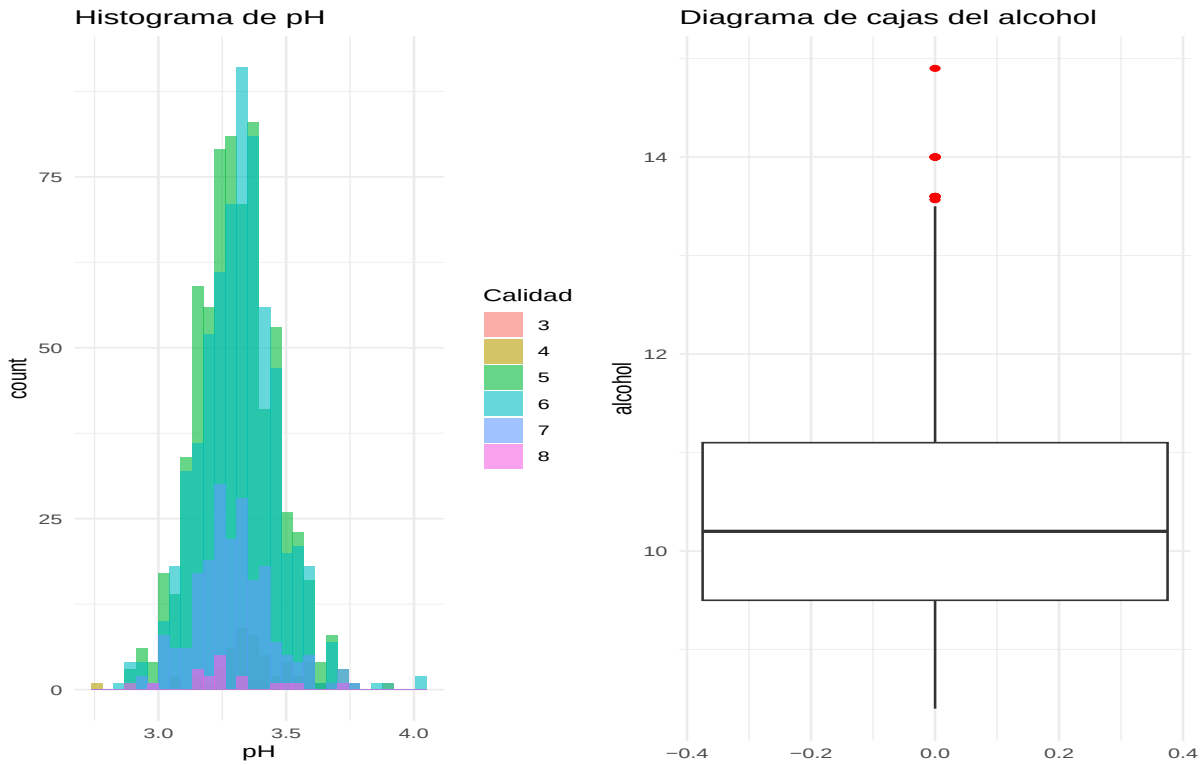

Figura 2: Histograma de pH y boxplot de alcohol.

Tras representar el histograma de las once variables que tenemos, observamos que algunas parecen seguir una distribución normal, como por ejemplo, el pH. También respresentamos los diagramas de cajas de todas las variables y observamos bastantes valores atípicos, pero no merece la pena eliminar sus entradas debido a que no se desvían tanto del resto de datos. Mostramos por ejemplo el histograma del pH y el diagrama del alcohol en 2.

4.1. Regresión lineal múltiple

Nuestro objetivo es predecir la calidad del vino según sus características fisicoquímicas. Para ello proponemos un modelo de regresión lineal múltiple.

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \epsilon_i \quad i = 1, \ldots, n \tag{7}$$

donde la variable de respuesta Y es la calidad del vino (quality), y como variables explicativas, o covariables, se consideran las 11 caracteríticas medidas: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates y alcohol.

Ajustamos ahora el modelo de regresión lineal múltiple a las n=1599 observaciones que tenemos de las variables, usando la función lm.

```
mod1<-lm(quality~ .,vinos) ; mod1

##
## Call:
```

```
## lm(formula = quality ~ ., data = vinos)
##
## Coefficients:
##          (Intercept)         fixed.acidity      volatile.acidity
##            21.965208              0.024991             -1.083590
##          citric.acid        residual.sugar             chlorides
##            -0.182564              0.016331             -1.874225
##  free.sulfur.dioxide  total.sulfur.dioxide               density
##             0.004361             -0.003265            -17.881164
##                   pH            sulphates              alcohol
##            -0.413653              0.916334              0.276198
```

El resultado nos muestra los coeficientes estimados, a partir de los cuales se obtienen los valores estimados de la calidad de cada vino mediante la siguiente expresión lineal:

$$\begin{aligned}\hat{Y}_i =& 0.025x_{i1} - 1.084x_{i2} - 0.183x_{i3} + 0.016x_{i4} - 1.874x_{i5} + 0.004x_{i6} - 0.003x_{i7} \\ & - 17.881x_{i8} - 0.413x_{i9} + 0.916x_{i10} + 0.2762x_{i11} + 21.965\end{aligned}$$

Los coeficientes estimados de las covariables no son directamente comparables entre sí, ya que no todas están expresadas en las mismas unidades de medida. Sin embargo, si nos enfocamos en aquellas variables que se miden en g/L —x_1, ..., x_5 y x_{10}—, podemos afirmar que la covariable con mayor influencia en la calidad del vino es x_5 (`chlorides`), que representa la concentración de cloruro de sodio, ya que presenta el coeficiente de mayor valor absoluto.

Por el contrario, la variable con menor influencia dentro de este grupo sería x_1, correspondiente a la acidez no volátil.

Las variables con coeficientes negativos disminuyen la calidad del vino, como es el caso de `chlorides`, mientras que aquellas con coeficientes positivos la aumentan, como `alcohol`.

Estudiamos ahora en qué medida las once percepciones de forma conjunta consiguen describir la calidad del vino midiendo la bondad del ajuste. Obtengamos entonces el coeficiente de determinación, R^2.

```
summary(mod1)
```

```
##
## Call:
## lm(formula = quality ~ ., data = vinos)
##
## 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
```

El coeficiente de regresión multiple es 0.3606, por lo que el ajuste no parece ser demasiado bueno. Sospechamos que esto puede deberse a la naturaleza de los datos, ya que `quality`, nuestra variable de respuesta, es ordinal discreta. Justificaremos esto mejor más adelante.

Inferencia sobre el modelo

Contraste de regresión: significación conjunta de las covariables: El estadístico de contraste obtenido es 81.35 y el p-valor asociado es aproximadamente 0. Se rechaza entonces la hipótesis nula y se concluye que el modelo de regresión ajustado es útil y tiene sentido, ya que las covariables consideradas de modo conjunto permiten explicar la calidad de los vinos. Por lo tanto, el modelo tiene suficiente significación a nivel poblacional.

Contrastes de significación individual de las variables:
A continuación, se analiza la influencia individual que cada una de las once percepciones consideradas ejerce sobre la calidad del vino. Se plantean once contrastes de hipótesis, uno por cada variable, en los cuales rechazar la hipótesis nula implica concluir que la percepción correspondiente tiene una influencia significativa en la calidad. Por el contrario, si no se rechaza la hipótesis nula, se concluye que dicha covariable no aporta información relevante y, por tanto, el modelo podría simplificarse eliminándola.

Observando el *summary* vemos que hay 5 variables (`volatile.acidity`, `chlorides`, `free.sulfur.dioxide`, `sulphates` y `alcohol`) que tienen una significación a nivel poblacional del $0\,\%$. Las variables `free.sulfur.dioxide` y `pH` tienen un nivel del $1\,\%$, por lo que las seguimos aceptando. Por el contrario rechazamos el resto de variables, entre ellas el término independiente.

Diagnóstico del modelo

De momento, se llega a la conclusión de que aunque los datos no tienen una fuerte relación lineal entre sí, el modelo parece ser útil, y hay variables que se pueden eliminar para simplificar el modelo.

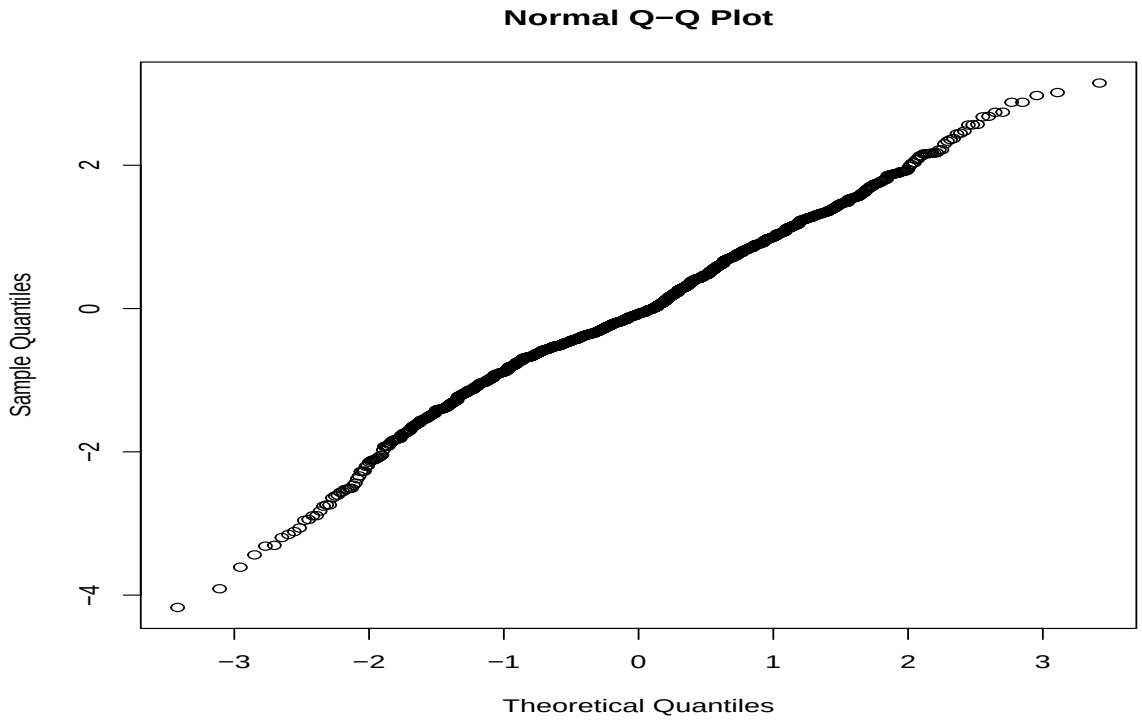

Figura 3: Gráfico qq-plot de los datos.

Pero estas conclusiones se han obtenido apoyándose en contrastes que se han desarrollado bajo fuertes hipótesis. Antes de continuar debemos realizar un análisis diagnóstico que nos permita validar o al menos descartar posibles violaciones de tales hipótesis:

Homocedasticidad: varianza constante

La hipótesis de homocedasticidad implica que los errores del modelo tienen varianza constante.

```
res<-rstandard(mod1)

qqnorm(res)
```

```
plot(mod1$fitted.values,res)
```

La gráfica 3 se acerca a un comportamiento lineal, luego podemos aceptar la hipótesis de normalidad. Claramente, observando 4, se aprecia una estructura de bandas decrecientes en el gráfico de los residuos. Hay una disminución de la dispersión de los residuos a medida que aumentan los valores ajustados. Esto se debe a que la variable de respuesta no es continua, sino ordinal discreta. Como los valores posibles de la calidad del vino son enteros, el modelo lineal intenta ajustar algo que es inherentemente escalonado, por eso los residuos se alinean en franjas. Todo esto nos indica que no se cumple la hipótesis de homocedasticidad, y por lo tanto hace que un modelo de regresión lineal no sea el más adecuado.

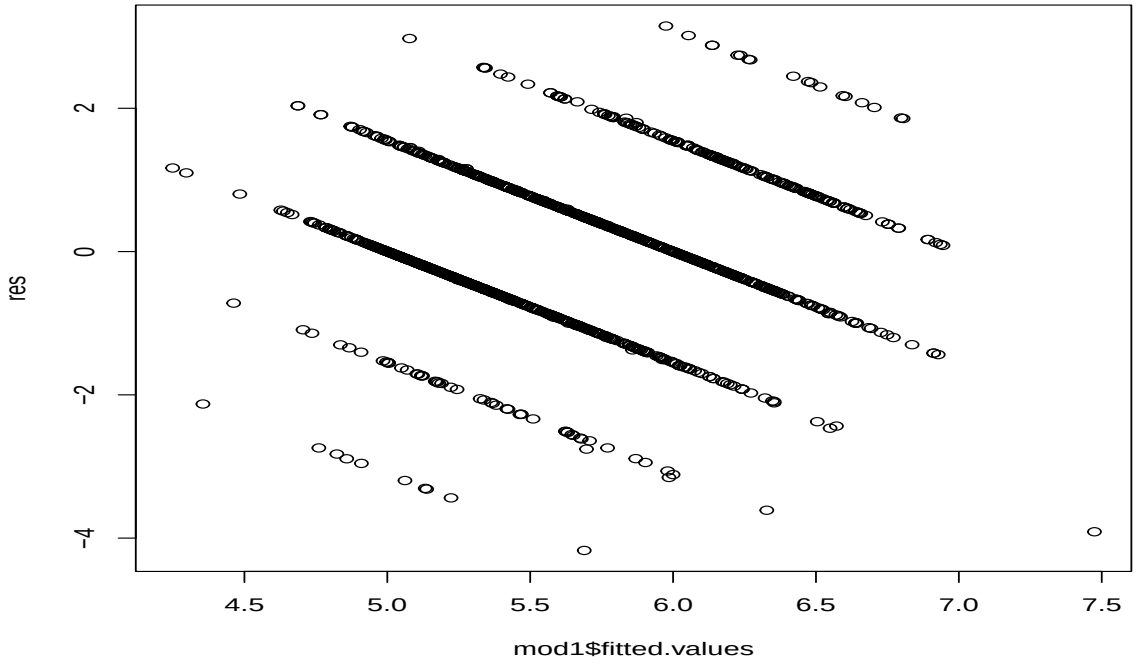

Figura 4: Gráfico de los residuos ajustados.

Aún así el modelo es eficiente, y para comprobarlo lo estudiamos usando una matriz de confusión utilizando una partición del $75\,\%$ de los datos para entrenar el modelo y el $25\,\%$ restante para probarlo.

```
#Primero creamos la partición de los datos en train y test.
set.seed(1)
train <- vinos[sample(1:1599,1200),]
indices_train <- sample(1:1599, 1200)
test <- vinos[setdiff(1:1599,indices_train),]


lm_train <- lm(quality ~ .,train[,variables])
predicciones <- predict(lm_train, newdata = test[,variables])
pred_class <- round(predicciones)
#Porque toma valores enteros, vamos a redondear los decimales
confusionMatrix(as.factor(pred_class), as.factor(test$quality))

Confusion Matrix and Statistics

          Reference
Prediction   3   4   5   6   7   8
         3   0   0   0   0   0   0
         4   0   0   1   0   0   0
         5   3  11 117  40   1   0
         6   0   3  38 112  44   2
```

```
         7   0   0   0   9  14   4
         8   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.609
                 95% CI : (0.5592, 0.6572)
    No Information Rate : 0.4035
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3695

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Sensitivity          0.000000 0.000000   0.7500   0.6957  0.23729  0.00000
Specificity          1.000000 0.997403   0.7737   0.6345  0.96176  1.00000
Pos Pred Value            NaN 0.000000   0.6802   0.5628  0.51852      NaN
Neg Pred Value       0.992481 0.964824   0.8282   0.7550  0.87903  0.98496
Prevalence           0.007519 0.035088   0.3910   0.4035  0.14787  0.01504
Detection Rate       0.000000 0.000000   0.2932   0.2807  0.03509  0.00000
Detection Prevalence 0.000000 0.002506   0.4311   0.4987  0.06767  0.00000
Balanced Accuracy    0.500000 0.498701   0.7618   0.6651  0.59953  0.50000
```

Tras calcular la matriz de confusión, vemos que no es capaz de predecir vino de calidad 3 y 8 ya que hay pocos de estos. Sin embargo, en el 5 acierta 117 y se quivoca 52 y para las calidades 6 y 7 también acierta más de lo que falla. Entonces el modelo sí es bueno en cierto modo para discernir vinos de calidad media. El coeficiente de determinación lineal es tan bajo porque los errores respecto del 3, 4 y del 8 son muy grandes.

Simplificación del modelo

Ahora se buscan posibles simplificaciones del modelo considerando solo las variables que realmente suponen una contribución significativa a la hora de describir la variable de respuesta.

```
mod2<-step(mod1)
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = vinos)
```

```
##
## 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 ***
## volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
## free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
## sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
## alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## 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
```

Este modelo se queda con las variables que suponen una constribución significativa, y como se observa, el coeficiente de determinación apenas varía con respecto al del modelo anterior. Esta vez es 0.3595 y la anterior 0.36. Las variables usadas en el primer gráfico son con las que se queda este modelo.

Podemos observar que la variable más influyente negativamente es `chlorides`, ya que cada unidad de esta disminuye la calidad del vino en -2.0178. Y la que más aumenta la calidad es `sulphates` con 0.88.

4.2. Regresión logística

Debido a los resultados obtenidos al no cumplir las hipótesis adecuadas para el modelo de regresión lineal, se decide hacer un modelo de regresión logística. Este modelo es más robusto frente a las hipótesis no cumplidas, por lo que habrá de ser más adecuado. Por tanto, realizamos un tratamiento previo de los datos para aplicar el modelo logit, ya que este funciona exclusivamente en casos en los que la variable respuesta es binaria. Por ello, creamos una nueva variable sobre la calidad asignándole el valor de 1 si el vino tenía calidad mayor o igual a 6, y 0 en caso opuesto. De esta forma, este modelo nos ajusta si el vino será bueno o malo, pero no diciendo exactamente qué nivel de calidad tendrá.

```
logit <- glm(vinos$quality_binario ~ ., vinos[,-c(12)],
             family = binomial(link = "logit"))
summary(logit)


##
```

```
## Call:
## glm(formula = vinos$quality_binario ~ ., family = binomial(link = "logit"),
##     data = vinos[, -c(12)])
##
## Coefficients:
##                        Estimate Std. Error z value             Pr(>|z|)
## (Intercept)           42.949948  79.473979   0.540              0.58890
## fixed.acidity          0.135980   0.098483   1.381              0.16736
## volatile.acidity      -3.281694   0.488214  -6.722      0.0000000000179 ***
## citric.acid           -1.274347   0.562730  -2.265              0.02354 *
## residual.sugar         0.055326   0.053770   1.029              0.30351
## chlorides             -3.915713   1.569298  -2.495              0.01259 *
## free.sulfur.dioxide    0.022220   0.008236   2.698              0.00698 **
## total.sulfur.dioxide  -0.016394   0.002882  -5.688      0.0000000128736 ***
## density              -50.932385  81.148745  -0.628              0.53024
## pH                    -0.380608   0.720203  -0.528              0.59717
## sulphates              2.795107   0.452184   6.181      0.0000000006356 ***
## alcohol                0.866822   0.104190   8.320 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2209.0  on 1598  degrees of freedom
## Residual deviance: 1655.6  on 1587  degrees of freedom
## AIC: 1679.6
##
## Number of Fisher Scoring iterations: 4
```

Podemos ver que el modelo final siguiendo la regresión logística tendría la forma siguiente:

$$P(Y=1) = p(x) = e^{0.14x_{i1}} e^{-3.3x_{i2}} e^{-1.3x_{i3}} e^{0.06x_{i4}} e^{-3.9x_{i5}} e^{0.02x_{i6}} e^{-0.02x_{i7}} e^{-50.9x_{i8}} e^{-0.38x_{i9}} e^{2.8x_{i10}} e^{0.87x_{i11}} e^{42.95}$$

En cuanto a la inferencia sobre el modelo, se observa que las variables `volatile.acidity`, `total.sulfur.dioxide`, `sulphates` y `alcohol` tienen validez al $0\,\%$, `free.sulfur.dioxide` al $0.1\,\%$, y `citric.acid` y `chlorides` al $1\,\%$. Todas ellas entran en el margen de validez del $5\,\%$, por lo que tienen representación del modelo a nivel poblacional. El resto de variables no cumple esto, por tanto, no son representativas.

Hagamos un "*step* y veamos que estas son las variables con las que se queda el modelo final.

```
logit_2 <- step(logit)
```

```
summary(logit_2)
```

```
##
## Call:
## glm(formula = vinos$quality_binario ~ fixed.acidity + volatile.acidity +
##     citric.acid + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     sulphates + alcohol, family = binomial(link = "logit"), data = vinos[,
##     -c(12)])
##
## Coefficients:
##                       Estimate Std. Error z value             Pr(>|z|)
## (Intercept)          -9.216919   0.949966  -9.702 < 0.0000000000000002 ***
## fixed.acidity         0.127271   0.051081   2.492              0.01272 *
## volatile.acidity     -3.379881   0.477983  -7.071     0.00000000000154 ***
## citric.acid          -1.260357   0.560972  -2.247              0.02466 *
## chlorides            -3.529121   1.509122  -2.339              0.01936 *
## free.sulfur.dioxide   0.022082   0.008184   2.698              0.00697 **
## total.sulfur.dioxide -0.015645   0.002811  -5.565     0.00000002616223 ***
## sulphates             2.686254   0.432624   6.209     0.00000000053249 ***
## alcohol               0.905412   0.073423  12.331 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2209.0  on 1598  degrees of freedom
## Residual deviance: 1657.8  on 1590  degrees of freedom
## AIC: 1675.8
##
## Number of Fisher Scoring iterations: 4
```

Sorprendentemente, al hacer el modelo mediante step, aparece la variable `fixed.acidity` y como representativa con validez del $1\,\%$, sumada a las nombradas previamente, mientras que desaparece la variable `pH`, que tenía significación en el modelo lineal pero no en el logístico.

Debido a esto, podemos observar que la variable más importante a la hora de mejorar la calidad del vino es `sulphates`, con una odds-ratio de $e^{2.7} = 15$, por lo que, en cada incremento unitario de esta variable, la razón de probabilidades (*odds*) de que el vino sea de buena calidad es 15 veces mayor. Le sigue cerca el alcohol, con $e^{0.91} = 2.48$. Por otra parte, la variable que afecta de manera negativa con mayor intensidad es `chlorides`, con la odds-ratio de $e^{-3.5} = 0.03$, reduciendo las *odds* de que el vino sea de buena calidad en 0.03 veces.

Probamos de nuevo a ver la matriz de confusión que obtenemos, con el mismo modelo de entrenamiento y prueba previo.

```
predicciones <- predict(logit_2, newdata = test[,variables])
#El logit devuelve una probabilidad, por lo que imponemos un umbral
pred_class <- ifelse(predicciones >= 0.5, 1, 0)
confusionMatrix(as.factor(pred_class), as.factor(test$quality_binario))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 150  71
##          1  23 155
##
##                Accuracy : 0.7644
##                  95% CI : (0.7196, 0.8052)
##     No Information Rate : 0.5664
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.5355
##
##  Mcnemar's Test P-Value : 0.000001249
##
##             Sensitivity : 0.8671
##             Specificity : 0.6858
##          Pos Pred Value : 0.6787
##          Neg Pred Value : 0.8708
##              Prevalence : 0.4336
##          Detection Rate : 0.3759
##    Detection Prevalence : 0.5539
##       Balanced Accuracy : 0.7764
##
##        'Positive' Class : 0
##
```

De esta matriz de confusión se obtienen muy buenos resultados. Por una parte, clasifica de manera correcta a la clase de vino de buena calidad el $86\,\%$ de las veces, y dada la especificidad del modelo, podremos estar un $68\,\%$ seguros de que no ha cometido errores. Viendo la matriz, es evidente que el modelo prefiere indicar que un vino es malo cuando no es así, a que un vino malo es bueno. Por lo tanto, con este modelo, si se indica un vino como buena calidad, lo será con un $86\,\%$ de seguridad.

5. Conclusiones

En primer lugar, se comentó el problema planteado referente a la calidad cuantificada del vino, así como el posible interés en un estudio estadístico de este problema, con intención de hallar aquellas posibles variables que afecten tanto de formas positivas como negativas sobre la misma. Para ello, se comenzó explicando los modelos a utilizar, el modelo de regresión lineal y el modelo de regresión logística o "*logit*", así como las hipótesis que han de cumplir para su adecuada aplicación.

Posteriormente, se realizó un exhaustivo análisis exploratorio sobre las variables, para comprobar sus escalas, sus posibles distribuciones e incluso la presencia de codependencias. No se observó ninguna codependencia entre variables, y se concluyó mediante técnicas visuales, como

pueden ser los histogramas o el "*qq-plot*", que las variables explicativas siguen distribuciones normales univariantes. Además, se realizaron diagramas de cajas así como un breve estudio de los posibles "*outliers*", pero se decidió por no eliminarlos de los datos, ya que no se consideró necesario dada la cantidad y los valores de sus entradas.

Es necesario destacar que, tras la realización del primer modelo, se trató de comprobar la hipótesis de homocedasticidad, y esta no se cumple debido al carácter cuasi-categórico de la variable respuesta, ya que su intervalo consiste en los enteros entre 3 y 8, con 3 una baja calidad y 8 una alta. Es por ello que el modelo lineal no llega a ser un buen modelo, al no cumplir una de las hipótesis requeridas para su funcionamiento.

Finalmente, se aplicaron los modelos, comenzando por el de regresión lineal. Se observó que no todas las variables eran significativas a nivel poblacional, por lo que se decidió por simplificar dicho modelo utilizando el método "*stepwise*", tomando siete de las once variables explicativas, todas ellas con un nivel de significación por debajo del $5\,\%$, por lo que son aceptadas. El resultado de este modelo, como ya se predijo, no es especialmente bueno: su coeficiente de correlación es de 0.36. Aún así, dado que tiene valor a nivel poblacional, nos permitirá conocer aquellas variables que influyen más en la calidad del vino. Se concluye entonces que estas variables son `chlorides`, influyendo de manera negativa en la calidad del vino, y `sulphates`, de manera positiva.

Dado el resultado de este modelo, se decidió aplicar el modelo de regresión logística, robusto frente al no cumplimiento de hipótesis como la homocedasticidad. De él se observó, realizando a su vez una simplificación "`stepwise`", que no todas las variables son representativas, escogiendo ocho de las once. De estas ocho variables, seis de ellas aparecen a su vez en el modelo de regresión lineal simplificado. Como resultado, este modelo presentó una sensibilidad del $86\,\%$ y una especificidad del $68\,\%$, por lo que es relativamente bueno. Como conclusión, se notó que las variables influyentes coincidían con las del modelo lineal: `chlorides` de forma negativa sobre la calidad y tanto `sulphates` como `alcohol` de forma positiva.

Por último, se concluye que, como modelo, el lineal no representa una buena aplicación para estos datos dada la no homocedasticidad, pero el logístico sí que sería considerado adecuado. Además, se obtienen los mismos resultados en ambos modelos, y es que las variables más influyentes en la calidad final del vino son `chlorides`, de forma negativa, y `sulphates`, seguida de `alcohol`, ambas de manera positiva.

6. Referencias

Referencias

[1] Kaggle, "Kaggle datasets - wine quality," 2025. `https://archive.ics.uci.edu/ml/datasets/wine+quality`. Accessed: 2025-05-26.

[2] Kaggle, "Kaggle datasets," 2025. Accessed: 2025-05-26.

[3] A. J. Izenman, *Introduction and Preview*, ch. 8. Springer texts in statistics, New York, NY: Springer New York, 2013.

[4] A. Agresti, *An introduction to categorical data analysis*, ch. 4-6. Wiley Series in Probability and Statistics, Nashville, TN: John Wiley & Sons, 2 ed., Mar. 2007.