Parcial 1: Modelos lineales generalizados y no paramétricos

Angie Rodríguez Duque

Octubre 23 de 2020

Introducción

El presente trabajo tiene como finalidad ajustar un Modelo Logístico, para un conjunto de datos de vino tinto que consta de 1599 observaciones y 12 variables, 11 de las cuales son sustancias químicas.

```
dim(Datos)
```

```
## [1] 1599   12
```

Las variables son:

1. **Acidez fija:** La mayoría de los ácidos implicados en el vino son fijos o no volátiles (no se evaporan fácilmente).
2. **Acidez volátil:** La cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.
3. **Ácido cítrico:** Encontrado en pequeñas cantidades, el ácido cítrico puede agregar "frescura" y sabor a los vinos.
4. **Azúcar residual:** Es la cantidad de azúcar que queda después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y los vinos con más de 45 gramos / litro se consideran dulces.
5. **Cloruros:** Es la cantidad de sal del vino.
6. **Dióxido de azufre libre:** La forma libre de SO_2 existe en equilibrio entre el SO_2 molecular (como gas disuelto) y el ion bisulfito; Previene el crecimiento microbiano y la oxidación del vino.
7. **Dióxido de azufre total:** Es la cantidad de formas libres y unidas de SO_2; en concentraciones bajas, el SO_2 es mayormente indetectable en el vino, pero en concentraciones de SO_2 libre superiores a 50 ppm, el SO_2 se hace evidente en la nariz y el sabor del vino.
8. **Densidad:** La densidad es cercana a la del agua dependiendo del porcentaje de alcohol y contenido de azúcar.
9. **pH:** Describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH.
10. **Sulfatos:** Aditivo del vino que puede contribuir a los niveles de dióxido de azufre (SO_2), que actúa como antimicrobiano y antioxidante.
11. **Alcohol:** El porcentaje de contenido de alcohol del vino.
12. **Calidad:** Variable de respuesta (basada en datos sensoriales, puntuación entre 0 y 10).

Estadísticas descriptivas

```
summary(Datos)
```

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

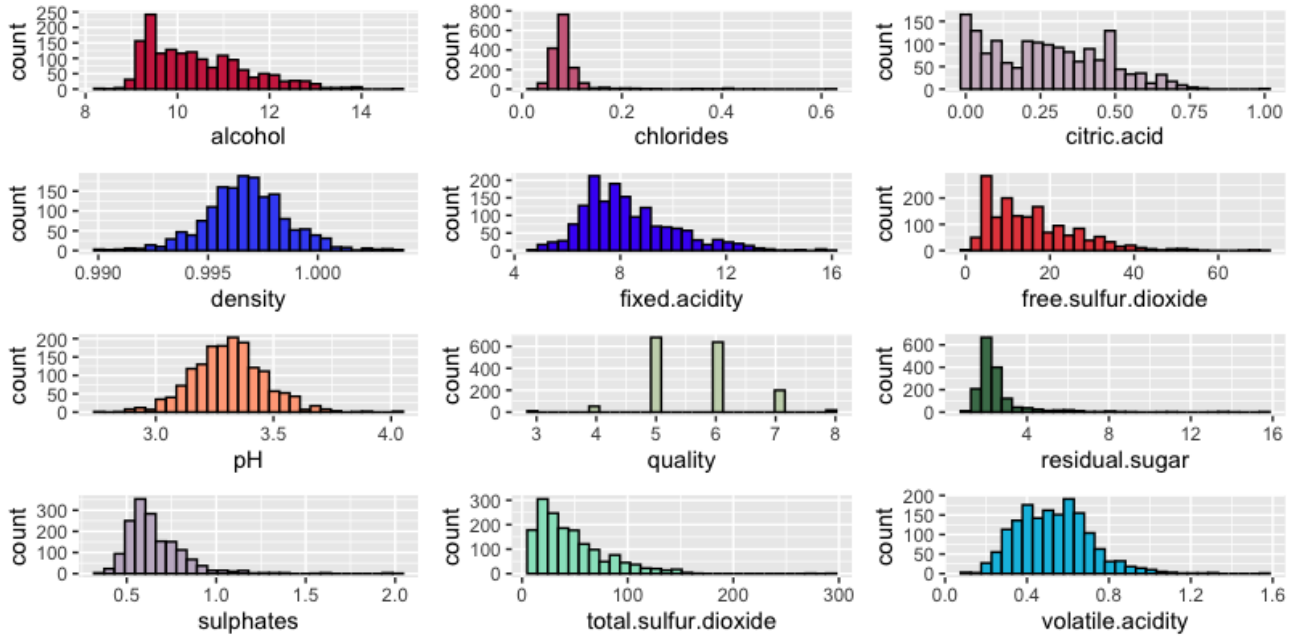


Figure 1: Distribución de las variables

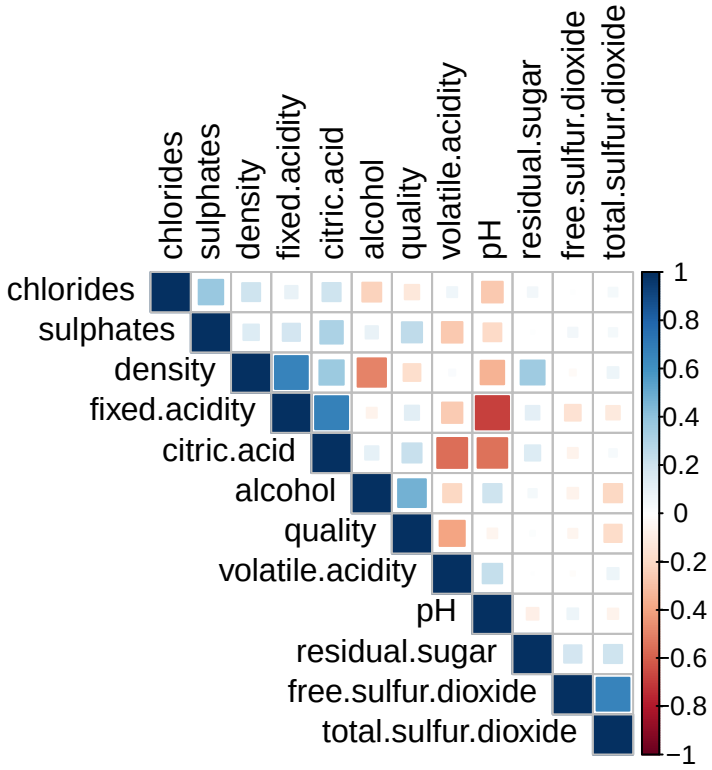
Observaciones

- Algunas de las variables tienen distribuciones normales (densidad, acidez fija, pH, acidez volátil).
- Algunas variables están un poco sesgadas hacia el extremo inferior de los valores (cloruros, ácido cítrico, azúcar residual, dióxido de azufre total).

- La variable calidad tiene solo 6 valores discretos.

Correlación

```
corrplot(cor(Datos), method="square", type="upper", order="hclust", tl.col="black")
```



- La densidad tiene una correlación muy fuerte con la acidez fija.
- Las variables más fuertemente correlacionadas con la calidad son la acidez volátil y el alcohol.
- El alcohol tiene una correlación negativa con la densidad. Esto es evidente por el hecho de que la densidad del agua es mayor que la densidad del alcohol.
- Es posible observar que las variables pH y acidez fija presentan una correlación negativamente fuerte, lo cual nos indica que a mayor pH menor será la acidez, y viceversa, a menor pH mayor acidez. Lo cual se ve reflejado en la calidad final del vino.

Muestra

Se elige una muestra de mil doscientos (1200) vinos de esta base de datos y trabaje, tal como sigue:

```
#Tamaño de la muestra
n <- 1200
```

Variable indicadora: alcoholAB

Se convierte la variable "alcohol" en una variable indicadora con dos niveles: "bajo" y "alto", esta nueva variable se denomina: "alcoholAB".

```
# Variable indicadora alcoholAB
alcoholAB <- vector()
alcoholAB[muestra$alcohol < 12] <- "Bajo"
```

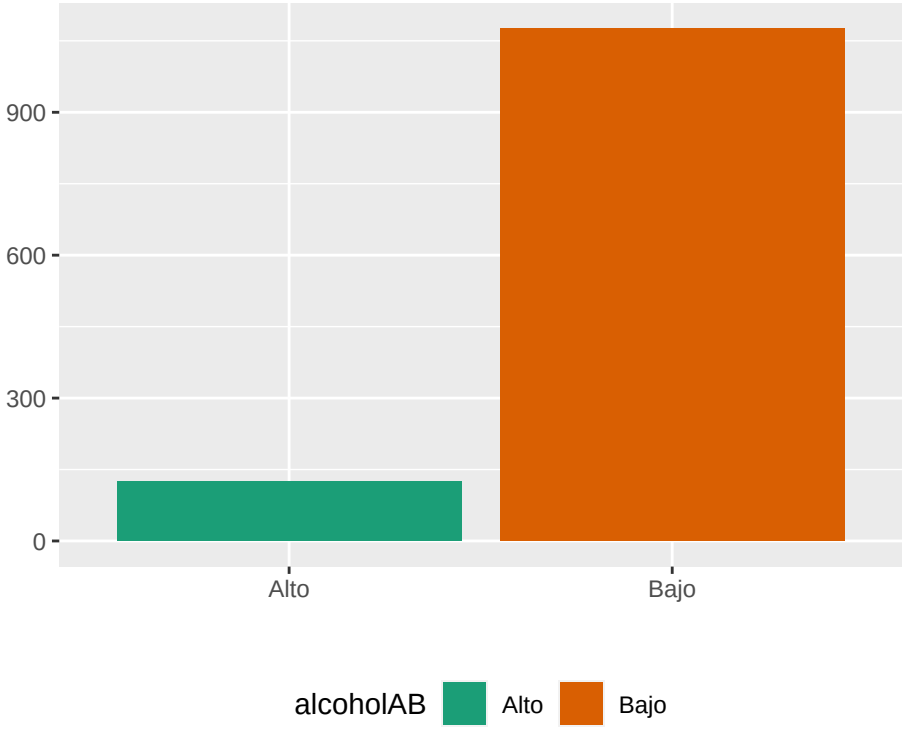
```r
alcoholAB[muestra$alcohol >= 12] <- "Alto"
alcoholAB <- as.factor(alcoholAB)
table(alcoholAB)
```

```
## alcoholAB
## Alto Bajo
##  124 1076
```

```r
G2 <- ggplot(data = muestra, aes(x=alcoholAB, fill=alcoholAB)) +
  geom_bar(position="dodge") + ylab("") + xlab(" ") +
  scale_fill_discrete(name = "alcohol:") + scale_fill_brewer(palette="Dark2") +
  theme(legend.position="bottom")
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

```r
G2
```



Variables dicotomica: calidadAB

Se convierte la variable "calidad" (quality) en una variable dicotómica y se denomina "calidadAB". Se forma un grupo con los vinos que tienen calidades 7 y 8 y otro con los demás vinos.

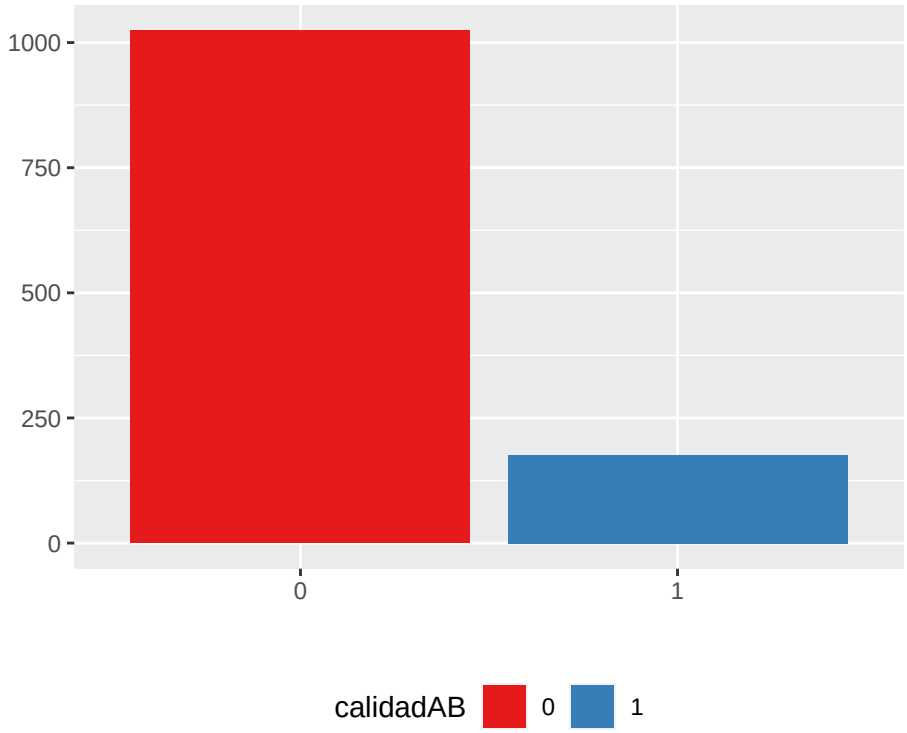
```r
# Variable indicadora quality
calidadAB <- vector()
calidadAB[muestra$quality <= 6] <- "0"
calidadAB[muestra$quality > 6] <- "1"
calidadAB <- as.factor(calidadAB)
table(calidadAB)
```

```
## calidadAB
##    0    1
## 1024  176
```

```
G3 <- ggplot(data = muestra, aes(x=calidadAB, fill=calidadAB)) +
  geom_bar(position="dodge") + ylab("") + xlab(" ") +
  scale_fill_discrete(name = "Calidad:") + scale_fill_brewer(palette="Set1") +
  theme(legend.position="bottom")
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

```
G3
```



Modelo logístico con variable indicadora

Para ajustar este modelo se hace uso de la función glm() para modelos lineales generalizados, una clase de modelos en los que se incluye el modelo logístico. En nuestro caso, se ajusta un modelo lineal generalizado usando como respuesta la variable calidadAB y como variables de predicción las variables "acidez fja" (fixed acidity) y "alcoholAB". Además, como la variable de respuesta "calidadAB" es una variable dicotómica especificamos el argumento family = binomial.

```
Modelo<- glm(calidadAB ~ fixed.acidity + alcoholAB + fixed.acidity*alcoholAB,
             data = muestra, family = "binomial")
```

```
summary(Modelo)
```

```
##
## Call:
## glm(formula = calidadAB ~ fixed.acidity + alcoholAB + fixed.acidity *
##     alcoholAB, family = "binomial", data = muestra)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1668  -0.5090  -0.4309  -0.3938   2.3446
```

```
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.23257    0.67871  -0.343   0.7319
## fixed.acidity                0.01305    0.08264   0.158   0.8745
## alcoholABBajo               -3.88183    0.82727  -4.692  2.7e-06 ***
## fixed.acidity:alcoholABBajo  0.22169    0.09768   2.269   0.0232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1000.52  on 1199  degrees of freedom
## Residual deviance:  896.28  on 1196  degrees of freedom
## AIC: 904.28
##
## Number of Fisher Scoring iterations: 5
```

El modelo obtenido es el siguiente:

$$calidadAB = -0.23257 + 0.01305 * Acidezfija + -3.88183 * AlcoholABBajo + 0.22169 * Acidezfija * AlcoholABBajo$$

A partir de los anteriores resultados y teniendo en cuenta que la interpretación de los p-valores es similar a la del modelo lineal. Es posible evidenciar que la variable alcoholAB-Bajo es altamente significativa ($2.7e-06$). De igual forma, la interacción entre ambas varias explicativas acidez fija y alcoholABBajo es significativa (0.0232)

Respecto a la interpretación de los coeficientes del modelo logit, estos se interpretan como el logaritmo del odds ratio. Así, si nos fijamos en el coeficiente de la variable acidez fija (0.01305), está positivamente relacionada con el logaritmo del odds ratio de la calidadAB del vino, la cual aumentaría 0.01305 unidades por cada unidad que aumenta la acidez fija. Por otro lado, la variable alcoholABBajo se encuentra negativamente relacionada con el logaritmo del odds ratio de la calidadAB del vino, el cual disminuiría -3.88183 unidades por cada unidad que aumenta el alcoholABBajo.

Bibliografía

- Fox, J. (2015),Applied regression analysis and generalized linear models, SagePublications.
- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.