# Parcial 1: Modelos lineales generales y no paramétricos

Cesar A. Saavedra Vanegas

10/23/2020

## Parcial 1

Este conjunto de datos de vino tinto consta de 1599 observaciones y 12 variables, 11 de las cuales son sustancias químicas. 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).

### Base de datos vinos

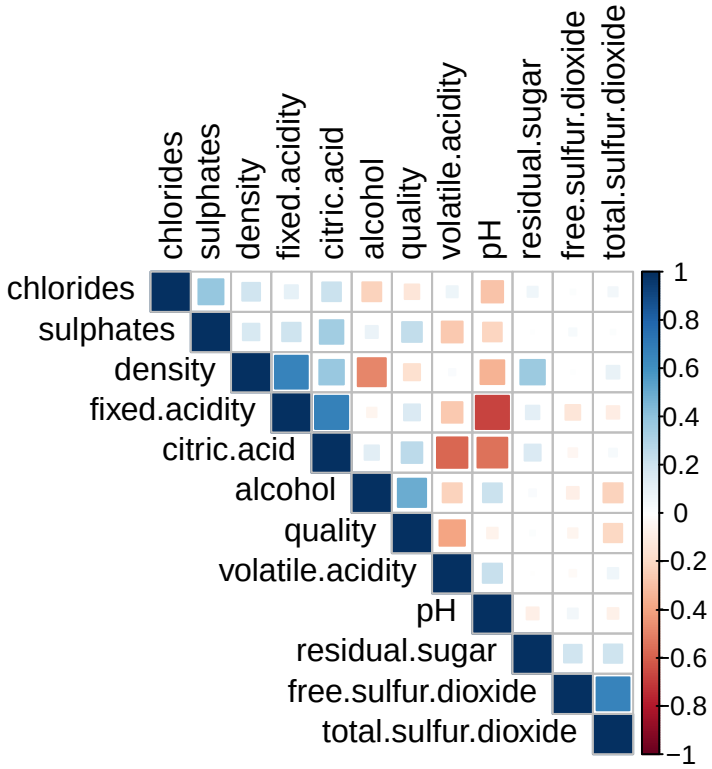
```
Datos <- read.table("Datos.txt",header=T,sep = ",")
```

Se selecciona un tamaño de muestra de 1200 vinos para realizar el modelo, tal y como se muestra a continuación:

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

Como primera medida, se realiza un gráfico de correlación para observar las posibles relaciones existentes entre las variables de la base de datos, tal y como se muestra a continuación:

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




Del gráfico es posible observar que las variables más fuertemente correlacionadas con la calidad son la acidez volátil y el alcohol, para nuestro modelo esto es importante dado que la variable alcohol sera una de las predictoras en nuestro modelo.

## Variables indicadoras

Se procede con la conversión de las variables "alcohol" en una variable indicadora de dos niveles: "Bajo" y "Alto" y de la variable "quality" como dicotómica con sus niveles: "0" y "1", estas nuevas variables se denominan "alcoholAB" y "calidadAB" respectivamente. Este procedimiento se realiza de la siguiente forma:

```
# Variable indicadora alcoholAB
alcoholAB <- vector()
alcoholAB[muestra$alcohol < 12] <- "Bajo"
alcoholAB[muestra$alcohol >= 12] <- "Alto"
alcoholAB <- as.factor(alcoholAB)
table(alcoholAB)
```

```
## alcoholAB
## Alto Bajo
##  127 1073
```

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

```
## calidadAB
##    0    1
## 1037  163
```

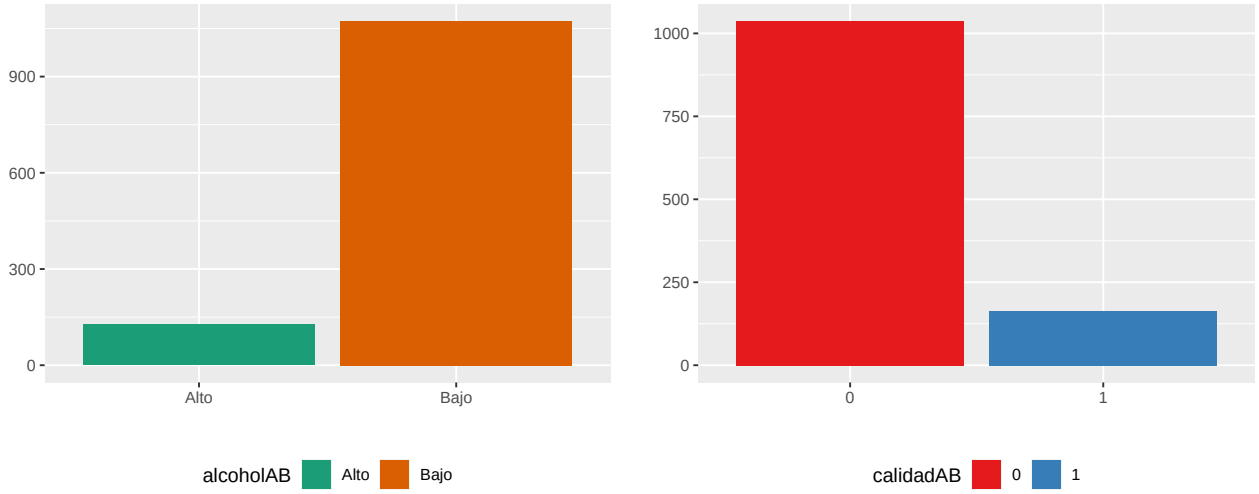
```
G2 <- ggplot(data = muestra, aes(x=alcoholAB, fill=alcoholAB)) +
  geom_bar(position="dodge") + ylab("") + xlab(" ") +
  scale_fill_discrete(name = "alcohol:") + scale_fill_brewer(palette="Dark2") +
  theme(legend.position="bottom")
```

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

```
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.
```

```
grid.arrange(G2,G3, ncol= 2)
```




A partir de los graficos anteriores se observa como en la base de datos vinos se encuentra una mayor concentración en alcoholes bajos, esto es vinos con un nivel de alcohol inferior a 12 y resultados de calidad inferior a 6.

## Modelo con variable indicadora y dicotómica

En esta sección, se procede a generar un modelo logístico con variable de respuesta ordinal (CalidadAB), ya que la variable de respuesta "calidadAB" ha sido dicotomisada, esto es, una puntuación entre 0 y 1, donde 0 representa una mala calidad y 1 una buena calidad del vino.

```
modelo.logit <- glm(calidadAB ~ fixed.acidity + alcoholAB + fixed.acidity*alcoholAB, data = muestra, fa
```

```
summary(modelo.logit)
```

```
##
## Call:
## glm(formula = calidadAB ~ fixed.acidity + alcoholAB + fixed.acidity *
##     alcoholAB, family = "binomial", data = muestra)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3805  -0.4848  -0.3846  -0.3340   2.4894
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -0.9547     0.7119  -1.341   0.1799
## fixed.acidity                  0.1029     0.0872   1.180   0.2379
## alcoholABBajo                 -4.1037     0.8793  -4.667 3.05e-06 ***
## fixed.acidity:alcoholABBajo    0.2206     0.1033   2.135   0.0327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 953.58  on 1199  degrees of freedom
## Residual deviance: 824.25  on 1196  degrees of freedom
## AIC: 832.25
##
## Number of Fisher Scoring iterations: 5
```

De acuerdo a los resultados obtenidos y teniendo en cuenta que la interpretación de los p-valores es similar a la del modelo lineal. Podemos ver que las variables alcoholABBajo y la interacción entre las variables predictoras son significativas, esto es valores-p de <3.05e-06 y 0.0327 respectivamente.

En cuanto a los coeficientes del modelo logit, estos se interpretan como el logaritmo del odds ratio. De esta manera, si nos fijamos en el coeficiente de la variable acidez fija (0.1029), está positivamente relacionada con el logaritmo del odds ratio de la CalidadAB, la cual aumentaría la calidad en 0.1029 unidades por cada unidad que aumenta la puntuación en el acidez fija.