Modelos lineales. Máster en Ciencia de Datos. ETSE. Universitat de València

Entregable 2: Modelos de regresión logística

Carlos Blom-Dahl Ramos y Daniel Lillo Plaza

El archivo `diabetes` puedes encontrar variables relacionadas con la mortalidad por diabetes `MORT`. En concreto nos interesa trabajar con las siguientes variables como predictoras de la mortalidad.

- `EDAT`: edad actual del paciente
- `BMI`: índice de masa corporal
- `EDATDIAG`: Edad al diagnóstico
- `TABAC`: Hábito tabáquico (tres categorías)
- `SBP`: Presión arterial sitólica
- `DBP`: Presión arterial diastólica
- `ECG`: Resultado del electrocardiograma (tres categorías)
- `CHD`: Antecedentes cardiacos

Ejercicio 1:

1. Antes de comenzar, como estamos interesados en aquellos pacientes que han muerto, crea una nueva variable que asigne 1 a muerto y 0 a vivo. Trabajaremos con esta nueva variable para la mortalidad.
2. Ajustar el modelo para predecir la mortalidad por diabetes en función de los predictores que consideres relevantes y que sean significativos, depurándolo al máximo.
3. Interpreta los coeficientes del modelo obtenido.
4. ¿Has conseguido un modelo explicativo? (Es decir, ¿es el modelo adecuado?, ¿ajusta bien?)
5. Con el modelo final obtenido:

   - Calcula el porcentaje de predicciones acertadas usando todos los datos. Para ello, haz la tabla de clasificación correspondiente.

   - ¿Quienes tienen más probabilidad de morir, los que tienen ECG normal o ECG frontera?

   - Calcula la probabilidad de morir para un paciente de 40 años para los distintos valores de ECG.

```
# Leemos los datos
library(readxl)
diabetes <- read_excel("./data/diabetes.xlsx")
diabetes$NUMPACIE<-NULL
diabetes$TABAC<-factor(diabetes$TABAC)
```

```
diabetes$ECG<-factor(diabetes$ECG)
diabetes$CHD<-factor(diabetes$CHD)
```

```
diabetes$MORT<-factor(diabetes$MORT, levels=c("Vivo", "Muerto"), labels=c(0,1))
str(diabetes)
```

```
## tibble [149 x 10] (S3: tbl_df/tbl/data.frame)
##  $ MORT    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 2 1 ...
##  $ TEMPSVIU: num [1:149] 0 0.3 1.1 1.1 1.5 1.8 1.9 2 2.2 2.5 ...
##  $ EDAT    : num [1:149] 53 82 35 55 61 69 41 78 74 75 ...
##  $ BMI     : num [1:149] 34.5 25.3 25.8 22.1 29.2 22.3 32 28.7 27.1 49.7 ...
##  $ EDATDIAG: num [1:149] 47 50 34 33 54 56 31 77 54 57 ...
##  $ TABAC   : Factor w/ 3 levels "ex-fumador","fumador",..: 1 3 1 1 3 3 2 3 2 2 ...
##  $ SBP     : num [1:149] 150 176 126 222 184 152 142 178 168 174 ...
##  $ DBP     : num [1:149] 88 96 82 102 802 74 90 862 84 82 ...
##  $ ECG     : Factor w/ 3 levels "Anormal","Frontera",..: 1 3 3 2 3 1 2 3 2 2 ...
##  $ CHD     : Factor w/ 2 levels "No","Si": 2 2 1 2 1 2 2 1 2 2 ...
```

Antes de proceder a crear los modelos, como el objetivo es crear un clasificador, dividiremos a nuestro conjunto de datos en *train* (que emplearemos para ajustar el modelo) y *test* (que emplearemos para validarlo).

```
## 80% del tamaño de la muestra
smp_size <- floor(0.8 * nrow(diabetes))

## fijamos la semilla para hacer el análisis reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(diabetes)), size = smp_size)

train <- diabetes[train_ind, ]
test <- diabetes[-train_ind, ]
```

```
pacman::p_load(MASS)

# Creamos un modelo mínimo solo con la constante.
min.model = glm(MORT ~ 1, family=binomial, data = train)

# Creamos un modelo máximo con todas las variables.
max.model = glm(MORT~ . , family=binomial, data = train)

# Mediante el método stepwise, pudiendo ir hacia delante y hacia detrás,
# vamos a ir depurando el modelo por minimización del AIC.
step(max.model,direction="both",
scope=list(lower=min.model,upper=max.model))
```

```
## Start:  AIC=114.63
```

```
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD
##
##            Df Deviance    AIC
## - TABAC     2   91.260 111.26
## - TEMPSVIU  1   90.691 112.69
## - BMI       1   90.716 112.72
## - EDATDIAG  1   90.996 113.00
## - SBP       1   91.059 113.06
## - DBP       1   91.270 113.27
## - CHD       1   92.327 114.33
## <none>          90.627 114.63
## - ECG       2   95.849 115.85
## - EDAT      1   95.547 117.55
##
## Step:  AIC=111.26
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + SBP + DBP + ECG + CHD
##
##            Df Deviance    AIC
## - TEMPSVIU  1   91.268 109.27
## - BMI       1   91.345 109.34
## - EDATDIAG  1   91.482 109.48
## - SBP       1   91.515 109.52
## - DBP       1   91.637 109.64
## - CHD       1   92.516 110.52
## <none>          91.260 111.26
## - ECG       2   96.090 112.09
## - EDAT      1   95.563 113.56
## + TABAC     2   90.627 114.63
##
## Step:  AIC=109.27
## MORT ~ EDAT + BMI + EDATDIAG + SBP + DBP + ECG + CHD
##
##            Df Deviance    AIC
## - BMI       1   91.351 107.35
## - EDATDIAG  1   91.483 107.48
## - SBP       1   91.567 107.57
## - DBP       1   91.652 107.65
## - CHD       1   92.516 108.52
## <none>          91.268 109.27
## - ECG       2   96.524 110.52
## + TEMPSVIU  1   91.260 111.26
## - EDAT      1   96.117 112.12
## + TABAC     2   90.691 112.69
```

```
##
## Step:  AIC=107.35
## MORT ~ EDAT + EDATDIAG + SBP + DBP + ECG + CHD
##
##            Df Deviance    AIC
## - EDATDIAG  1   91.592 105.59
## - SBP       1   91.702 105.70
## - DBP       1   91.786 105.79
## - CHD       1   92.531 106.53
## <none>          91.351 107.35
## - ECG       2   96.605 108.61
## + BMI       1   91.268 109.27
## + TEMPSVIU  1   91.345 109.34
## - EDAT      1   96.415 110.42
## + TABAC     2   90.773 110.77
##
## Step:  AIC=105.59
## MORT ~ EDAT + SBP + DBP + ECG + CHD
##
##            Df Deviance    AIC
## - SBP       1   91.870 103.87
## - DBP       1   91.958 103.96
## - CHD       1   92.624 104.62
## <none>          91.592 105.59
## - ECG       2   96.924 106.92
## + EDATDIAG  1   91.351 107.35
## + BMI       1   91.483 107.48
## + TEMPSVIU  1   91.591 107.59
## + TABAC     2   91.122 109.12
## - EDAT      1  103.785 115.78
##
## Step:  AIC=103.87
## MORT ~ EDAT + DBP + ECG + CHD
##
##            Df Deviance    AIC
## - DBP       1   92.077 102.08
## - CHD       1   92.752 102.75
## <none>          91.870 103.87
## - ECG       2   97.070 105.07
## + SBP       1   91.592 105.59
## + EDATDIAG  1   91.702 105.70
## + BMI       1   91.713 105.71
## + TEMPSVIU  1   91.857 105.86
## + TABAC     2   91.584 107.58
```

```
## - EDAT      1  104.229 114.23
##
## Step:  AIC=102.08
## MORT ~ EDAT + ECG + CHD
##
##            Df Deviance    AIC
## - CHD       1   93.108 101.11
## <none>          92.077 102.08
## - ECG       2   97.239 103.24
## + DBP       1   91.870 103.87
## + BMI       1   91.895 103.89
## + EDATDIAG  1   91.938 103.94
## + SBP       1   91.958 103.96
## + TEMPSVIU  1   92.071 104.07
## + TABAC     2   91.846 105.85
## - EDAT      1  106.107 114.11
##
## Step:  AIC=101.11
## MORT ~ EDAT + ECG
##
##            Df Deviance    AIC
## <none>          93.108 101.11
## - ECG       2   97.256 101.26
## + CHD       1   92.077 102.08
## + DBP       1   92.752 102.75
## + BMI       1   93.053 103.05
## + EDATDIAG  1   93.069 103.07
## + TEMPSVIU  1   93.073 103.07
## + SBP       1   93.091 103.09
## + TABAC     2   93.052 105.05
## - EDAT      1  106.163 112.16
##
## Call:  glm(formula = MORT ~ EDAT + ECG, family = binomial, data = train)
##
## Coefficients:
## (Intercept)         EDAT  ECGFrontera    ECGNormal
##    -4.47382      0.07894     -1.67717     -1.52509
##
## Degrees of Freedom: 118 Total (i.e. Null);  115 Residual
## Null Deviance:       116.8
## Residual Deviance: 93.11     AIC: 101.1
```

```r
# Guardamos el modelo obtenido en fit1.
fit1<-glm(formula = MORT ~ EDAT + ECG, family = binomial, data = train)
```

```r
fit1.s<-summary(fit1)

# Vamos a considerar ahora otro modelo máximo, en el cual tendremos en cuenta
# todas las interacciones 2 a 2 entre las variables del modelo.
max.model2=glm(MORT~ (.)^2 , family=binomial, data = train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Mediante stepwise de nuevo, depuramos el modelo nuevamente.
step(max.model,direction="both",
scope=list(lower=min.model,upper=max.model2))
```

```
## Start:  AIC=114.63
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD
##
##                      Df Deviance    AIC
## + SBP:DBP             1   77.918 103.92
## + BMI:TABAC           2   76.450 104.45
## + TEMPSVIU:DBP        1   79.758 105.76
## - TABAC               2   91.260 111.26
## + DBP:CHD             1   86.594 112.59
## - TEMPSVIU            1   90.691 112.69
## + DBP:ECG             2   84.703 112.70
## - BMI                 1   90.716 112.72
## - EDATDIAG            1   90.996 113.00
## - SBP                 1   91.059 113.06
## + TEMPSVIU:EDATDIAG   1   87.238 113.24
## - DBP                 1   91.270 113.27
## + EDATDIAG:CHD        1   87.445 113.44
## + EDATDIAG:ECG        2   85.650 113.65
## + TABAC:DBP           2   86.300 114.30
## - CHD                 1   92.327 114.33
## <none>                    90.627 114.63
## + EDAT:EDATDIAG       1   88.921 114.92
## + EDAT:CHD            1   88.935 114.94
## + TEMPSVIU:ECG        2   87.410 115.41
## + EDAT:BMI            1   89.537 115.54
## + EDAT:DBP            1   89.846 115.85
## - ECG                 2   95.849 115.85
## + TEMPSVIU:SBP        1   90.074 116.07
## + EDATDIAG:DBP        1   90.114 116.11
## + BMI:SBP             1   90.235 116.23
## + EDAT:ECG            2   88.277 116.28
```

```
## + EDAT:SBP           1   90.403 116.40
## + TEMPSVIU:EDAT      1   90.424 116.42
## + TABAC:CHD          2   88.449 116.45
## + BMI:CHD            1   90.569 116.57
## + TEMPSVIU:BMI       1   90.599 116.60
## + BMI:DBP            1   90.607 116.61
## + BMI:EDATDIAG       1   90.610 116.61
## + SBP:CHD            1   90.623 116.62
## + TEMPSVIU:CHD       1   90.625 116.62
## + EDATDIAG:SBP       1   90.625 116.62
## + EDAT:TABAC         2   89.077 117.08
## + BMI:ECG            2   89.267 117.27
## - EDAT               1   95.547 117.55
## + TABAC:SBP          2   89.660 117.66
## + SBP:ECG            2   89.832 117.83
## + EDATDIAG:TABAC     2   89.847 117.85
## + TEMPSVIU:TABAC     2   89.898 117.90
## + TABAC:ECG          4   88.353 120.35
##
## Step:  AIC=103.92
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD + SBP:DBP

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                     Df Deviance     AIC
## + BMI:TABAC          2   68.285  98.285
## - TABAC              2   78.834 100.834
## - TEMPSVIU           1   77.955 101.955
## - BMI                1   78.007 102.007
## - EDATDIAG           1   78.416 102.416
## - CHD                1   78.637 102.637
## + TEMPSVIU:ECG       2   73.032 103.032
## <none>                   77.918 103.918
## + EDATDIAG:CHD       1   76.067 104.067
## + EDAT:SBP           1   76.625 104.625
## + EDAT:CHD           1   76.720 104.720
## + EDAT:EDATDIAG      1   76.800 104.800
## - ECG                2   82.829 104.829
## + EDAT:BMI           1   76.921 104.921
## + EDAT:DBP           1   76.959 104.959
## + TEMPSVIU:EDATDIAG  1   77.020 105.020
## + BMI:SBP            1   77.107 105.107
## + TEMPSVIU:SBP       1   77.200 105.200
## + DBP:ECG            2   75.223 105.223
```

```
## + TEMPSVIU:DBP       1   77.445 105.445
## + EDATDIAG:SBP       1   77.502 105.502
## + BMI:ECG            2   75.521 105.521
## + SBP:CHD            1   77.635 105.635
## + DBP:CHD            1   77.638 105.638
## + EDAT:TABAC         2   75.657 105.657
## + EDATDIAG:ECG       2   75.703 105.703
## + TEMPSVIU:CHD       1   77.773 105.773
## + EDATDIAG:DBP       1   77.791 105.791
## + BMI:DBP            1   77.834 105.834
## + TEMPSVIU:BMI       1   77.868 105.868
## + TEMPSVIU:EDAT      1   77.894 105.894
## + BMI:CHD            1   77.901 105.901
## + BMI:EDATDIAG       1   77.911 105.911
## + SBP:ECG            2   76.400 106.400
## + EDAT:ECG           2   76.454 106.454
## + TABAC:CHD          2   76.561 106.561
## + EDATDIAG:TABAC     2   76.572 106.572
## + TABAC:SBP          2   76.735 106.735
## + TEMPSVIU:TABAC     2   77.345 107.345
## + TABAC:DBP          2   77.687 107.687
## - EDAT               1   86.276 110.276
## + TABAC:ECG          4   76.940 110.940
## - SBP:DBP            1   90.627 114.627
##
## Step:  AIC=98.29
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD + SBP:DBP + BMI:TABAC

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                     Df Deviance     AIC
## + BMI:SBP            1   63.950  95.950
## - TEMPSVIU           1   68.335  96.335
## + EDAT:CHD           1   64.653  96.653
## - EDATDIAG           1   68.756  96.756
## - CHD                1   68.941  96.941
## + EDAT:ECG           2   63.410  97.410
## <none>                   68.285  98.285
## + EDAT:BMI           1   66.401  98.401
## + TEMPSVIU:ECG       2   64.645  98.645
## + EDAT:SBP           1   66.733  98.733
## + EDATDIAG:CHD       1   66.743  98.743
## + BMI:DBP            1   66.782  98.782
## - ECG                2   72.782  98.782
```

```
## + EDAT:TABAC         2   65.205  99.205
## + TEMPSVIU:EDATDIAG  1   67.395  99.395
## + TEMPSVIU:BMI       1   67.499  99.499
## + EDATDIAG:ECG       2   65.525  99.525
## + EDATDIAG:SBP       1   67.558  99.558
## + TEMPSVIU:DBP       1   67.597  99.597
## + TEMPSVIU:SBP       1   67.656  99.656
## + EDAT:DBP           1   67.719  99.719
## + SBP:ECG            2   65.756  99.756
## + EDAT:EDATDIAG      1   67.883  99.883
## + BMI:EDATDIAG       1   67.954  99.954
## + TABAC:CHD          2   66.089 100.089
## + EDATDIAG:DBP       1   68.207 100.207
## + TEMPSVIU:CHD       1   68.213 100.213
## + DBP:CHD            1   68.216 100.216
## + BMI:CHD            1   68.236 100.236
## + SBP:CHD            1   68.259 100.259
## + TABAC:SBP          2   66.277 100.277
## + TEMPSVIU:EDAT      1   68.282 100.282
## + DBP:ECG            2   66.420 100.420
## + BMI:ECG            2   66.976 100.976
## + EDATDIAG:TABAC     2   67.078 101.078
## + TABAC:DBP          2   67.909 101.909
## + TEMPSVIU:TABAC     2   68.237 102.237
## + TABAC:ECG          4   65.239 103.239
## - EDAT               1   75.748 103.748
## - BMI:TABAC          2   77.918 103.918
## - SBP:DBP            1   76.450 104.450
##
## Step:  AIC=95.95
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD + SBP:DBP + BMI:TABAC + BMI:SBP

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                     Df Deviance     AIC
## + EDAT:CHD           1   58.938  92.938
## - EDATDIAG           1   63.996  93.996
## - TEMPSVIU           1   64.121  94.121
## - CHD                1   64.403  94.403
## + EDAT:ECG           2   58.640  94.640
## + EDATDIAG:CHD       1   61.252  95.252
## <none>                   63.950  95.950
## + TEMPSVIU:EDATDIAG  1   62.088  96.088
## + TEMPSVIU:ECG       2   60.397  96.397
```

```
## + EDAT:TABAC         2   60.485  96.485
## + EDATDIAG:ECG       2   60.601  96.601
## + SBP:ECG            2   60.834  96.834
## + EDAT:DBP           1   63.241  97.241
## + EDAT:SBP           1   63.283  97.283
## + EDATDIAG:SBP       1   63.295  97.295
## - ECG                2   69.424  97.424
## + EDAT:EDATDIAG      1   63.467  97.467
## + DBP:CHD            1   63.594  97.594
## + TABAC:CHD          2   61.693  97.693
## + BMI:CHD            1   63.726  97.726
## + SBP:CHD            1   63.761  97.761
## + TEMPSVIU:SBP       1   63.776  97.776
## + TEMPSVIU:DBP       1   63.781  97.781
## + BMI:DBP            1   63.812  97.812
## + TEMPSVIU:EDAT      1   63.823  97.823
## + EDATDIAG:DBP       1   63.849  97.849
## + EDAT:BMI           1   63.868  97.868
## + BMI:EDATDIAG       1   63.884  97.884
## + TEMPSVIU:BMI       1   63.940  97.940
## + TEMPSVIU:CHD       1   63.946  97.946
## + EDATDIAG:TABAC     2   62.163  98.163
## - BMI:SBP            1   68.285  98.285
## + DBP:ECG            2   62.339  98.339
## + TABAC:SBP          2   62.618  98.618
## + BMI:ECG            2   62.771  98.771
## - EDAT               1   69.188  99.188
## + TABAC:DBP          2   63.423  99.423
## + TEMPSVIU:TABAC     2   63.676  99.676
## + TABAC:ECG          4   60.899 100.899
## - SBP:DBP            1   72.885 102.885
## - BMI:TABAC          2   77.107 105.107
##
## Step:  AIC=92.94
## MORT ~ TEMPSVIU + EDAT + BMI + EDATDIAG + TABAC + SBP + DBP +
##     ECG + CHD + SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                     Df Deviance     AIC
## - EDATDIAG           1   58.963  90.963
## - TEMPSVIU           1   59.502  91.502
## <none>                   58.938  92.938
## + DBP:ECG            2   55.492  93.492
## + TABAC:SBP          2   55.551  93.551
```

```
## + EDAT:SBP           1   57.584  93.584
## + DBP:CHD            1   57.796  93.796
## + EDATDIAG:SBP       1   57.899  93.899
## + TEMPSVIU:CHD       1   58.097  94.097
## + TEMPSVIU:ECG       2   56.146  94.146
## + TEMPSVIU:SBP       1   58.177  94.177
## + EDAT:BMI           1   58.281  94.281
## - ECG                2   64.328  94.328
## + TEMPSVIU:BMI       1   58.381  94.381
## + EDAT:TABAC         2   56.431  94.431
## + EDAT:DBP           1   58.466  94.466
## + TEMPSVIU:EDAT      1   58.486  94.486
## + SBP:CHD            1   58.587  94.587
## + EDAT:EDATDIAG      1   58.598  94.598
## + TEMPSVIU:EDATDIAG  1   58.710  94.710
## + TEMPSVIU:DBP       1   58.756  94.756
## + BMI:CHD            1   58.884  94.884
## + EDATDIAG:CHD       1   58.893  94.893
## + BMI:DBP            1   58.913  94.913
## + EDATDIAG:DBP       1   58.917  94.917
## + BMI:EDATDIAG       1   58.937  94.937
## + EDATDIAG:TABAC     2   57.383  95.383
## + SBP:ECG            2   57.530  95.530
## - EDAT:CHD           1   63.950  95.950
## + BMI:ECG            2   58.108  96.108
## + EDAT:ECG           2   58.112  96.112
## + EDATDIAG:ECG       2   58.257  96.257
## + TEMPSVIU:TABAC     2   58.496  96.496
## + TABAC:CHD          2   58.612  96.612
## + TABAC:DBP          2   58.647  96.647
## - BMI:SBP            1   64.653  96.653
## - SBP:DBP            1   66.864  98.864
## + TABAC:ECG          4   57.523  99.523
## - BMI:TABAC          2   76.154 106.154
##
## Step:  AIC=90.96
## MORT ~ TEMPSVIU + EDAT + BMI + TABAC + SBP + DBP + ECG + CHD +
##     SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                  Df Deviance     AIC
## - TEMPSVIU        1   59.613  89.613
## <none>               58.963  90.963
## + TABAC:SBP       2   55.590  91.590
```

```
## + DBP:ECG         2   55.622  91.622
## + EDAT:SBP        1   57.647  91.647
## + DBP:CHD         1   57.901  91.901
## + TEMPSVIU:ECG    2   56.161  92.161
## + TEMPSVIU:CHD    1   58.162  92.162
## + TEMPSVIU:SBP    1   58.235  92.235
## + EDAT:BMI        1   58.316  92.316
## + TEMPSVIU:BMI    1   58.430  92.430
## + EDAT:TABAC      2   56.452  92.452
## - ECG             2   64.541  92.541
## + SBP:CHD         1   58.587  92.587
## + TEMPSVIU:EDAT   1   58.619  92.619
## + EDAT:DBP        1   58.622  92.622
## + TEMPSVIU:DBP    1   58.757  92.757
## + BMI:CHD         1   58.896  92.896
## + EDATDIAG        1   58.938  92.938
## + BMI:DBP         1   58.942  92.942
## + SBP:ECG         2   57.533  93.533
## - EDAT:CHD        1   63.996  93.996
## + BMI:ECG         2   58.120  94.120
## + EDAT:ECG        2   58.132  94.132
## + TEMPSVIU:TABAC  2   58.581  94.581
## + TABAC:CHD       2   58.624  94.624
## + TABAC:DBP       2   58.649  94.649
## - BMI:SBP         1   65.212  95.212
## - SBP:DBP         1   66.894  96.894
## + TABAC:ECG       4   57.550  97.550
## - BMI:TABAC       2   76.391 104.391
##
## Step:  AIC=89.61
## MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD + SBP:DBP +
##     BMI:TABAC + BMI:SBP + EDAT:CHD
##
##              Df Deviance     AIC
## <none>            59.613  89.613
## + DBP:ECG     2   55.915  89.915
## + TABAC:SBP   2   56.302  90.302
## + DBP:CHD     1   58.403  90.403
## + EDAT:SBP    1   58.568  90.568
## + EDAT:BMI    1   58.743  90.743
## + TEMPSVIU    1   58.963  90.963
## + EDAT:TABAC  2   57.073  91.073
## + SBP:CHD     1   59.207  91.207
## + EDAT:DBP    1   59.337  91.337
```

```
## + BMI:CHD     1   59.489  91.489
## + EDATDIAG    1   59.502  91.502
## + BMI:DBP     1   59.609  91.609
## + SBP:ECG     2   57.994  91.994
## - EDAT:CHD    1   64.232  92.232
## + EDAT:ECG    2   58.697  92.697
## + BMI:ECG     2   58.741  92.741
## - ECG         2   66.989  92.989
## + TABAC:CHD   2   59.222  93.222
## + TABAC:DBP   2   59.321  93.321
## - BMI:SBP     1   65.610  93.610
## - SBP:DBP     1   67.024  95.024
## + TABAC:ECG   4   57.755  95.755
## - BMI:TABAC   2   76.471 102.471

##
## Call:  glm(formula = MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD +
##     SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD, family = binomial,
##     data = train)
##
## Coefficients:
##         (Intercept)                EDAT                 BMI
##           20.613225            0.106216           -1.582008
##        TABACfumador     TABACNo fumador                 SBP
##            8.103436          -12.597125           -0.191725
##                DBP         ECGFrontera           ECGNormal
##            0.422857           -1.053956           -4.766013
##               CHDSi             SBP:DBP    BMI:TABACfumador
##          -12.957766           -0.002287           -0.341148
## BMI:TABACNo fumador              BMI:SBP          EDAT:CHDSi
##            0.368325            0.010251            0.184504
##
## Degrees of Freedom: 118 Total (i.e. Null);  104 Residual
## Null Deviance:       116.8
## Residual Deviance: 59.61     AIC: 89.61
```

```r
# Guardamos el resultado obtenido en fit2.
fit2<-glm(formula = MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD +
    SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD, family = binomial,
    data = train)
fit2.s<-summary(fit2)
fit2.s
```

```
##
## Call:
```

```
## glm(formula = MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD +
##     SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD, family = binomial,
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.62436  -0.39906  -0.21061  -0.02845   2.62173
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          20.613225  22.747639   0.906   0.3648
## EDAT                  0.106216   0.047077   2.256   0.0241 *
## BMI                  -1.582008   0.663079  -2.386   0.0170 *
## TABACfumador          8.103436   6.678911   1.213   0.2250
## TABACNo fumador     -12.597125   5.583870  -2.256   0.0241 *
## SBP                  -0.191725   0.143910  -1.332   0.1828
## DBP                   0.422857   0.185003   2.286   0.0223 *
## ECGFrontera          -1.053956   1.392603  -0.757   0.4492
## ECGNormal            -4.766013   2.174078  -2.192   0.0284 *
## CHDSi               -12.957766   7.127146  -1.818   0.0691 .
## SBP:DBP              -0.002287   0.001012  -2.259   0.0239 *
## BMI:TABACfumador     -0.341148   0.235010  -1.452   0.1466
## BMI:TABACNo fumador   0.368325   0.177589   2.074   0.0381 *
## BMI:SBP               0.010251   0.004628   2.215   0.0268 *
## EDAT:CHDSi            0.184504   0.100879   1.829   0.0674 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.844  on 118  degrees of freedom
## Residual deviance:  59.613  on 104  degrees of freedom
## AIC: 89.613
##
## Number of Fisher Scoring iterations: 7
```

```
anova(fit1, fit2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: MORT ~ EDAT + ECG
## Model 2: MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD + SBP:DBP +
##     BMI:TABAC + BMI:SBP + EDAT:CHD
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
## 1       115     93.108
## 2       104     59.613 11   33.496 0.0004371 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Sí que hay diferencia estadística significativa entre fit1 y fit2, por ello
# tenemos que decantarnos por fit2 pese a que tenga más variables.

# Aunque tenemos variables que no son significativas en fit2, como por ejemplo
# ECGFrontera, no podemos eliminar las que son de ese estilo ya que la otra
# caategoría que hemos obtenido sí es significativa y consideramos que la
# ECG en su conjunto es una variable importante. Nos decantamos por no
# fusionar las categorías.
```

Vamos a pasar ahora a explicar los coeficientes obtenidos en el modelo `fit2`. Recordemos cómo era este modelo.

```
fit2.s
```

```
##
## Call:
## glm(formula = MORT ~ EDAT + BMI + TABAC + SBP + DBP + ECG + CHD +
##     SBP:DBP + BMI:TABAC + BMI:SBP + EDAT:CHD, family = binomial,
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.62436  -0.39906  -0.21061  -0.02845   2.62173
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           20.613225  22.747639   0.906   0.3648
## EDAT                   0.106216   0.047077   2.256   0.0241 *
## BMI                   -1.582008   0.663079  -2.386   0.0170 *
## TABACfumador           8.103436   6.678911   1.213   0.2250
## TABACNo fumador      -12.597125   5.583870  -2.256   0.0241 *
## SBP                   -0.191725   0.143910  -1.332   0.1828
## DBP                    0.422857   0.185003   2.286   0.0223 *
## ECGFrontera           -1.053956   1.392603  -0.757   0.4492
## ECGNormal             -4.766013   2.174078  -2.192   0.0284 *
## CHDSi                -12.957766   7.127146  -1.818   0.0691 .
## SBP:DBP               -0.002287   0.001012  -2.259   0.0239 *
## BMI:TABACfumador      -0.341148   0.235010  -1.452   0.1466
## BMI:TABACNo fumador    0.368325   0.177589   2.074   0.0381 *
## BMI:SBP                0.010251   0.004628   2.215   0.0268 *
## EDAT:CHDSi             0.184504   0.100879   1.829   0.0674 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.844  on 118  degrees of freedom
## Residual deviance:  59.613  on 104  degrees of freedom
## AIC: 89.613
##
## Number of Fisher Scoring iterations: 7
```

Estamos en el siguiente caso:

$$(Y|\mathbf{X}) \sim Ber(p(\mathbf{X}))$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

$$logit(p(\mathbf{X})) = \ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Es decir, nuestra situación esf:

$$\frac{P(MORT=1)}{1-P(MORT=1)} = exp(20{,}61+0{,}1\cdot EDAT-1{,}58\cdot BMI+8{,}1\cdot TABAC_F-12{,}59\cdot TABAC_{NF}-0{,}19\cdot SBP+$$
$$+0{,}42\cdot DBP-1{,}05\cdot ECG_F-4{,}76\cdot ECG_N-12{,}96\cdot CHD_{Si}-0{,}002\cdot SBP\cdot DBP-0{,}34\cdot BMI\cdot TABAC_F+$$
$$+0{,}36\cdot BMI\cdot TABAC_{NF}+0{,}01\cdot BMI\cdot SBP+0{,}18\cdot EDAT\cdot CHD_{Si})$$

Y por último antes de comenzar, aclaremos qué es lo que estamos analizando. Vamos a estudiar cómo varía ese cociente, el cociente de los odds, al aumentar en una unidad una determinada variable.

```
coef<-coefficients(fit2)
################## EDAT #################
# EDAT interactúa de forma significativa con la variable CHD, por lo que
# para analizar cómo afecta a la variable respuesta tenemos que estudiar
# también el valor de CHD (Sí/No en este caso)

# Si estamos en el grupo CHD_No, entonces:
EDAT_exp<-exp(coef["EDAT"])
EDAT_exp
```

```
##     EDAT
## 1.112062
```

```r
# Cuando aumentamos en una unidad esta variable, el cociente de los odds,
# p(X)/(1-p(X)), se modifica en 1.112062.

# Si estamos en el grupo CHD_Si, entonces:
EDAT_CDH_SI_exp<-exp(coef["EDAT"]+coef["EDAT:CHDSi"])
EDAT_CDH_SI_exp
```

```
##    EDAT
## 1.33739
```

```r
# Cuando aumentamos en una unidad esta variable, el cociente de los odds,
# p(X)/(1-p(X)), se modifica en 1.33739.

# Es decir, tanto si estamos en CHD como si no, al aumentar la edad el cociente
# de los odds aumenta, es decir, la proporción de muertes aumenta más que la de
# vivos bajo las mismas condiciones. Sin embargo, este aumento es más acusado
# si además estamos en el grupo CHD_Si.

################## BMI  ##################
# BMI interactúa de forma significativa con la variable TABAC y con la variable SBP,
# por lo que para analizar cómo afecta a la variable respuesta tenemos que estudiar
# también el valor de TABAC (ex-fumador, fumador y no_fumador en este caso) y SBP
# (variable numérica).

# Si estamos en el grupo ex-fumador, entonces la variación del cociente de los
# odds viene dada por:
# exp(-1.582008+0.01025135 *SBP)=exp(-1.582008)*exp(0.01025135 *SBP)

# Si estamos en el grupo fumador, entonces la variación del cociente de los
# odds viene dada por:
# exp(-1.582008-0.3411482+0.01025135*SBP)=exp(-1.923156)*exp(0.01025135*SBP)

# Si estamos en el grupo no-fumador, entonces la variación del cociente de los
# odds viene dada por:
# exp(-1.582008+0.3683248+0.01025135*SBP)=exp(-1.213683)*exp(0.01025135*SBP)

# Por lo tanto, las diferencias entre cómo varía el cociente de los odds al
# aumentar en una unidad la variable BMI según el grupo de TABAC vienen dadas
# por exp(-1.582008)=0.2055619 para el grupo de ex-fumadores,
# exp(-1.923156)=0.146145 para el de los fumadores y
# exp(-1.213683)=0.297101 en el de los no-fumadores. Por lo tanto,
# aumentar en una unidad el BMI en cualquiera de estos grupos causará una reducción
# en el cociente de los odds (una reducción en la proporción de muertos frente a
# vivos) pero esta reducción será mayor (se reducirá más) en el grupo de los
```

```
# fumadores, luego en el de los ex-fumadores y luego en el de los no-fumadores.

# Aunque por supuesto a esto hay que añadir el efecto de SBP, que como podemos ver
# es exp(0.01025135 *SBP)=1.010304^SBP. Por lo tanto, al aumentar SBP estaremos
# aumentando el cociente de los odds, es decir, la proporción de muertos frente
# a vivos.

################## TABAC  ##################
# Ya hemos comentado lo que ocurre si en los diferentes grupos de TABAC
# aumentamos en una unidad la variable BMI. Ahora vamos a estudiar cómo
# varía el cociente de los odds si pasamos de un grupo de TABAC a otro.

# Supongamos que estamos en el grupo de los fumadores y pasamos al de los
# ex-fumadores. La diferencia entre las expresiones en los cocientes de los odds,
# es decir, si cogemos y dividimos el cociente de los odds de los ex-fumadores entre
# el cociente de los odds de los fumadores, viene dada por:
# 1/exp(8.103436-0.3411482*BMI)=exp(0.3411482*BMI)/exp(8.103436)=1.406562^BMI/3294.468.
# Es decir, si BMI>23.7433 entonces 1.406562^BMI/3294.468>1 y por tanto la
# proporción de muertos será mayor que la de vivos. Si por el contrario BMI<23.7433
# entonces la proporción de muertos será menor que la de vivos al cambiar de grupo.

# Supongamos ahora que estamos en el de los no-fumadores y pasamos al de los
# fumadores. Análogamente a antes, la diferencia entre las expresiones en los
# cocientes de los odds vendrá dada por:
# exp(8.103436-0.3411482*BMI)/exp(-12.59713 +0.3683248*BMI)=
# =exp(8.103436+12.59713)/exp((0.3411482+0.3683248)*BMI)=
# =exp(20.70057)/exp(0.709473*BMI)=977559776/(2.03292^BMI).
# Tenemos que 977559776/(2.03292^BMI)>=1, es decir, la proporción de muertos
# será mayor frente a la de vivos, si BMI<29.1774. En caso contrario,
# la proporción de muertos se verá reducida frente a la de vivos al cambiar de grupo.

################## SBP  ##################
# La variable SBP interactúa significativamente con la variable DBP y con BMI.
# Veamos qué ocurre al aumentar en una unidad SBP si hacemos la división de los
# cocientes de los odds sin el +1 (denominador) y con el +1 (numerador):
# exp(-0.1917253-0.002286931*DBP+0.01025135*BMI)
# Por lo tanto, dado un DBP, para que exp(-0.1917253-0.002286931*DBP+0.01025135*BMI)>1
# necesitamos BMI>9.75481×10^-8 (2.28693×10^6 DBP + 1.91725×10^8). Si se cumple
# esa condición, entonces la proporción de muertes frente a la de vivos habrá
# aumentado al aumentar en una unidad SBP. En caso contrario, habrá disminuidp.
# Por ejemplificarlo un poco, si DBP es 22 entonces necesitamos BMI>23.61
# para que esto se cumpla.
```

```
################## DBP  ##################
# La variable DBP interacúa significativamente con la variable SBP. Estudiemos
# el cociente como en los casos anteriores:
# exp(0.4228566-0.002286931*SBP)
# Por lo tanto, como exp(0.4228566-0.002286931*SBP)>1 si SBP<184.901, en esos casos
# se producirá un aumento en la proporción de muertes frente a la de vivos al
# aumentar en una unidad DBP.

################## ECG  ##################
# Esta variable no presenta interacciones significativas en nuestro modelo, por lo
# que estudiaremos el efecto individual de pasar de una categoría a otra.

# Supongamos que estamos en ECG anormal y pasamos a normal. Estudiemos cómo varía
# el cociente:
exp(-4.766013)
```

```
## [1] 0.008514259
```

```
# Como es menor que 1, la proporción de muertos frente a vivos disminuye al
# hacer el cambio por un factor de 117.45.

# Supongamos ahora que estamos en ECG normal y pasamos a frontera.
exp(-1.053956)/exp(-4.766013)
```

```
## [1] 40.93793
```

```
# Como es mayor que 1, la proporción de muertos frente a vivos aumenta al
# hacer el cambio por un factor de 40.93793.

################## CHD  ##################
# La variable CHD presenta interacción con la variable EDAT.
# Veamos cómo varía el cociente de los odds si pasamos de una categoría de CHS (Si)
# a la otra (No).
# 1/exp(-12.95777+0.1845043 *EDAT)
# Tenemos que es mayor que 1 si EDAT<70.2302, por lo tanto en ese caso al cambiar
# del grupo Si al grupo No obtenemos que la proporción de muertos aumenta frente
# a la de vivos.
```

Veamos si el modelo es explicativo. Para ello usaremos nuestro grupo test del principio, también para comprobar si el modelo está sobreajustado.

```
prob <- predict(fit2, newdata = train, type="response")
prediction<-rep(1, length(prob))
prediction[prob<0.5]<-0
sum((prediction-(as.numeric(train$MORT)-1))==0)/length(prob)
```

```
## [1] 0.8907563
```

```
# Veamos cuántas acertamos en el grupo test
prob <- predict(fit2, newdata = test, type="response")
prediction<-rep(1, length(prob))
prediction[prob<0.5]<-0
sum((prediction-(as.numeric(test$MORT)-1))==0)/length(prob)
```

```
## [1] 0.9333333
```

Responderemos a continuación a las preguntas del ejercicio 5. -¿Quienes tienen más probabilidad de morir, los que tienen ECG normal o ECG frontera? Como ya hems comentado al analizar los coeficientes, la proporción de muertos frente a vivos aumenta al pasar del grupo ECG normal al grupo ECG frontera por un factor de 40.93793. Por lo tanto, será más probable morir al tener una ECG frontera.

-Calcula la probabilidad de morir para un paciente de 40 años para los distintos valores de ECG.

```
# Para poder hacer el cálculo, imputaremos los datos que nos faltan empleando
# la media o la moda (según sea numérica o categórica respectivamente) de los
# hombre de 40 años que tengamos en la base de datos.

indices<-diabetes["EDAT"]==40
BMI_mean<-mean(diabetes["BMI"][indices])
TABAC_moda<-"ex-fumador"
SBP_mean<-mean(diabetes["SBP"][indices])
DBP_mean<-mean(diabetes["DBP"][indices])
CHD_moda<-"No"

# Si ECG Anormal
prob1 <- predict(fit2, newdata = data.frame("EDAT"=40, "ECG"="Anormal",
                                            "BMI"=BMI_mean, "TABAC"=TABAC_moda, "SBP"=SBP_mean
prob1
```

```
##         1
## 0.8257066
```

```
# Si ECG Frontera
prob2 <- predict(fit2, newdata = data.frame("EDAT"=40, "ECG"="Frontera",
                                            "BMI"=BMI_mean, "TABAC"=TABAC_moda, "SBP"=SBP_mean
prob2
```

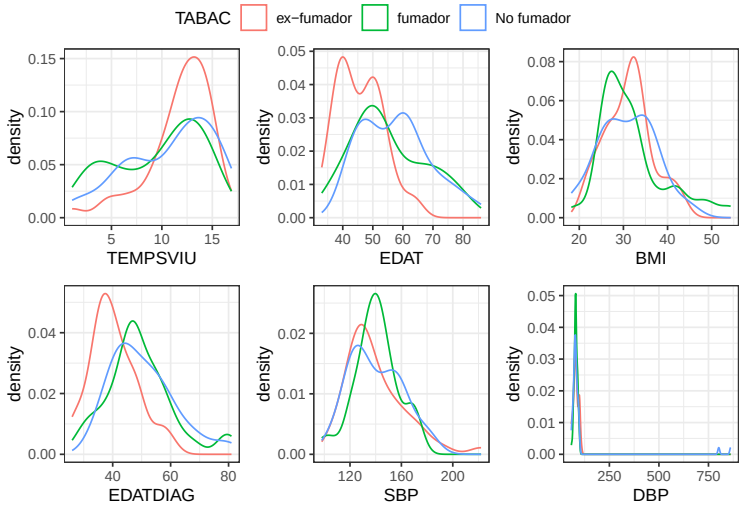
```
##         1
## 0.622822
```

```
# Si ECG Normal
prob3 <- predict(fit2, newdata = data.frame("EDAT"=40, "ECG"="Normal",
                                            "BMI"=BMI_mean, "TABAC"=TABAC_moda, "SBP"=SBP_mean
prob3
```

```
##          1
## 0.03877199
```

Ejercicio 2: Imagina que ahora estamos interesados en la variable TABAC. Realiza un análisis discriminante completo (considerando las variables adecuadas). Comprueba si se cumplen las condiciones de aplicabilidad.

```
n <- nrow(train)
m <- nrow(test)
# Veamos a priori que sucede gráficamente

library(ggplot2)
library(ggpubr)


plot1 <- ggplot(data = train, aes(x = TEMPSVIU)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot2 <- ggplot(data = train, aes(x = EDAT)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot3 <- ggplot(data = train, aes(x = BMI)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot4 <- ggplot(data = train, aes(x = EDATDIAG)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot5 <- ggplot(data = train, aes(x = SBP)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot6 <- ggplot(data = train, aes(x = DBP)) +
  geom_density(aes(colour = TABAC)) + theme_bw()

# la función grid.arrange del paquete grid.extra permite ordenar
# graficos de ggplot2

library(gridExtra)
ggarrange(plot1, plot2, plot3, plot4,plot5, plot6,  common.legend = TRUE)
```



Primero nos damos cuenta que en la gráfica DBP existen dos valores muy alejados, si nos detenemos en estos datos, encontramos que 2 personas tienen una presión arterial diastólica de 862 y 802, estos valores son claramente un error en la toma de los datos ya que un valor de 200 ya sería increíblemente inverosímil. Por este motivo procedemos a eliminar ambas entradas.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

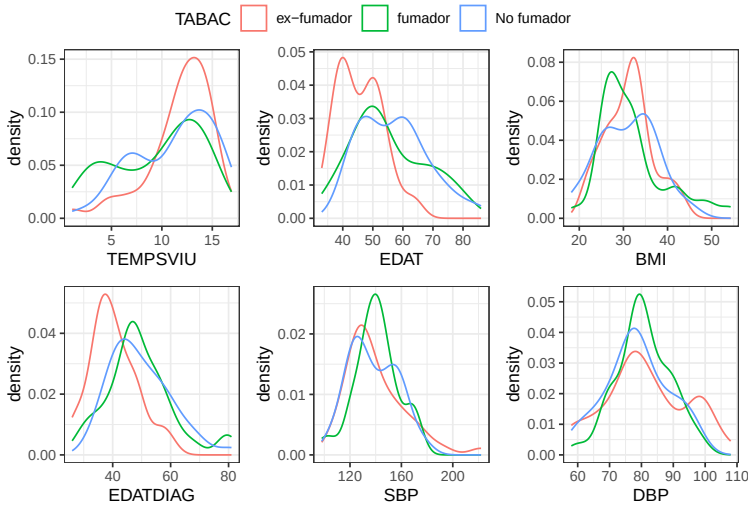
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
train <- train %>% filter(DBP!=862) %>% filter(DBP!=802)
```

Ahora volvamos a crear los gráficos de densidad otra vez.

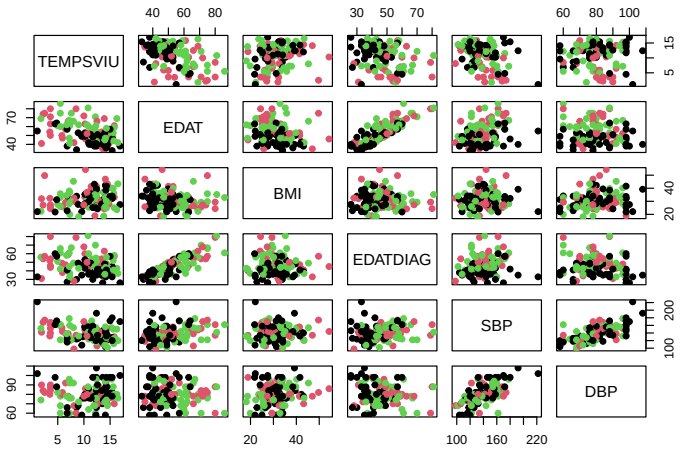
```
plot1 <- ggplot(data = train, aes(x = TEMPSVIU)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot2 <- ggplot(data = train, aes(x = EDAT)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
```

```
plot3 <- ggplot(data = train, aes(x = BMI)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot4 <- ggplot(data = train, aes(x = EDATDIAG)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot5 <- ggplot(data = train, aes(x = SBP)) +
  geom_density(aes(colour = TABAC)) + theme_bw()
plot6 <- ggplot(data = train, aes(x = DBP)) +
  geom_density(aes(colour = TABAC)) + theme_bw()

# la función grid.arrange del paquete grid.extra permite ordenar
# graficos de ggplot2

library(gridExtra)
ggarrange(plot1, plot2, plot3, plot4,plot5, plot6,  common.legend = TRUE)
```



Estas gráficas ya nos hacen dudar de la precisión que obtenedremos con un clasificador lda ya que no tenemos ninguna variable que esté formada por 3 normales con misma covarianza que separen claramente los 3 grupos. La variable más prometedora parece EDATDIAG, la cual separa el grupo ex-fumador de los otros dos. Veamos ahora las correlaciones lineales entre las diferentes variables.

```
pairs(x = train[, c("TEMPSVIU","EDAT","BMI","EDATDIAG","SBP","DBP")],
      col = c("1", "2","3")[train$TABAC], pch = 19)
```



En estas gráficas de correlación vemos que los datos no son muy separables linealmente. Además podemos observar algunas variables que están correlacionadas como EDAT y EDATDIAG o en menor medida SBP y DBP.

Ahora vamos a crear un modelo lda con todas las variables anteriores.

```
library(MASS)
#### en el train
m_lda_1 <- lda(TABAC ~ TEMPSVIU+EDAT+BMI+EDATDIAG+SBP+DBP, data=train)
predLDA<-predict(m_lda_1, newdata=train)
t<-table(train$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##              Clase predicha
## Clase real    ex-fumador fumador No fumador
##   ex-fumador          29       6          8
##   fumador             11      15          9
##   No fumador           9       8         22
```

```
100*sum(diag(t))/n
```

```
## [1] 55.46218
```

Obtenemos así una precisión en el grupo train del $55\,\%$ aproximadamente. Esta precisión es bastante baja, pero veamos que está ocurriendo en el grupo de test.

```
#### en el test
m_lda_1 <- lda(TABAC ~ TEMPSVIU+EDAT+BMI+EDATDIAG+SBP+DBP, data=train)
predLDA<-predict(m_lda_1, newdata=test)
t<-table(test$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##              Clase predicha
## Clase real    ex-fumador fumador No fumador
```

```
##   ex-fumador          4       3          1
##   fumador             2       4          0
##   No fumador          4       7          5
```

```
100*sum(diag(t))/m # Aquí simplemente vemos cuantos hemos "acertado". Es decir,
```

```
## [1] 43.33333
```

```
# que precisión tenemos
```

Obtenemos así una precisión del aproximadamente $43\,\%$. Estos resultados son bastante malos, obtenemos una precisión muy baja para ambos conjuntos (train y test). Esto ya nos dice que el modelo no es muy prometedor pero vamos a intentar eliminar alguna variable que peda mejorar el resultado.

Primero, de las gráficas de correlación deducimos que EDATDIAG y EDAT están muy correlacionadas, por eso, tomaremos solamente una de estas variables. Viendo que realmente ninguna de las gráficas de correlación sirve para separar de forma clara los 3 grupos vamos a quedarnos con EDATDIAG, ya que su densidad según la variable TABAC es más parecido a 3 normales con diferente media y misma covarianza.

```
m_lda_2 <- lda(TABAC ~ TEMPSVIU+EDATDIAG+BMI+SBP+DBP, data=train)
predLDA<-predict(m_lda_2, newdata=test)
t<-table(test$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##             Clase predicha
## Clase real   ex-fumador fumador No fumador
##   ex-fumador          3       2          3
##   fumador             1       1          4
##   No fumador          5       7          4
```

```
100*sum(diag(t))/m
```

```
## [1] 26.66667
```

Hemos empeorado considerablemente el modelo para el conjunto de test, con lo que deducimos que no deberíamos eliminar la variable EDAT. Vamos a probar a eliminar EDATDIAG en lugar de EDAT.

```
m_lda_3 <- lda(TABAC ~ TEMPSVIU+EDAT+BMI+SBP+DBP, data=train)
predLDA<-predict(m_lda_3, newdata=test)
t<-table(test$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##             Clase predicha
## Clase real   ex-fumador fumador No fumador
##   ex-fumador          6       1          1
##   fumador             3       1          2
##   No fumador          4       9          3
```

```r
100*sum(diag(t))/m
```

```
## [1] 33.33333
```

Como cabía esperar obtenemos un resultado parecido a eliminar EDAT por su alta correlación. Por tanto no podemos eliminar ninguna de estas variables.

Probemos a eliminar una de las variables SBP o DBP del modelo original ya que existe cierta correlación lineal entre ellas. Vamos a probar a quedarnos con SBP por los mismo motivos de antes (las densidades de DBP tienen aproximadamente el mismo centro y covarianzas muy diferentes).

```r
m_lda_4 <- lda(TABAC ~ TEMPSVIU+EDAT+EDATDIAG+SBP, data=train)
predLDA<-predict(m_lda_4, newdata=test)
t<-table(test$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##             Clase predicha
## Clase real   ex-fumador fumador No fumador
##   ex-fumador          3       4          1
##   fumador             3       3          0
##   No fumador          8       2          6
```

```r
100*sum(diag(t))/m
```

```
## [1] 40
```

Volvemos a perder precisión (aunque poca). Podemos probar a eliminar SBP en lugar de DBP, y obtenemos:

```r
m_lda_5 <- lda(TABAC ~ TEMPSVIU+EDAT+BMI+DBP+EDATDIAG, data=train)
predLDA<-predict(m_lda_5, newdata=test)
t<-table(test$TABAC, predLDA$class, dnn = c("Clase real", "Clase predicha"))
t
```

```
##             Clase predicha
## Clase real   ex-fumador fumador No fumador
##   ex-fumador          5       2          1
##   fumador             2       4          0
##   No fumador          7       4          5
```

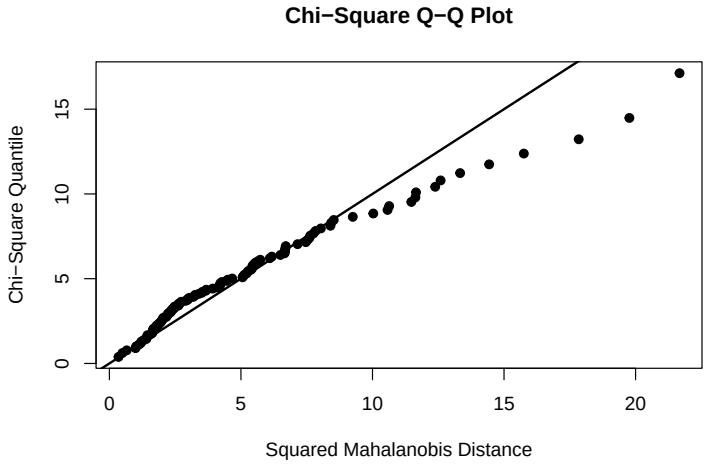
```r
100*sum(diag(t))/m
```

```
## [1] 46.66667
```

Hemos aumentado algo la precisión del grupo test. Por tanto vamos a quedarnos con el modelo TABAC $\sim$ TEMPSVIU + EDAT + BMI + DBP + EDATDIAG. Parece que nuestro modelo no es capaz de distinguir cuando se trata de un no fumador.

Dicho esto veamos si se cumplen las hipótesis de normalidad.

```
library(MVN)
royston_test <- mvn(data = train[,c('TEMPSVIU','EDAT', 'BMI', 'DBP', 'EDATDIAG')],
                    mvnTest = "royston", multivariatePlot = "qq")
```




```
royston_test$multivariateNormality
```

```
##      Test       H      p value MVN
## 1 Royston 49.9803 7.716004e-10  NO
```

```
royston_test$univariateNormality
```

```
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling TEMPSVIU     3.2935  <0.001      NO
## 2 Anderson-Darling   EDAT       1.6207   3e-04      NO
## 3 Anderson-Darling    BMI       0.8315  0.0311      NO
## 4 Anderson-Darling    DBP       0.9365  0.0171      NO
## 5 Anderson-Darling EDATDIAG     1.0557  0.0086      NO
```

Obtenemos que no se cumple la condición de normalidad para ninguna variable de nuestro modelo. Además también vemos que tampoco se puede asumir normalidad en la variable conjunto.

```
## Igualdad de matrices de covarianza

library(biotools)
```

```
## ---
## biotools version 4.2
```

```
boxM(data =train[,c('TEMPSVIU','EDAT', 'BMI', 'DBP', 'EDATDIAG')],
     grouping = train$TABAC) #se rechaza
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
```

```
##
## data:  train[, c("TEMPSVIU", "EDAT", "BMI", "DBP", "EDATDIAG")]
## Chi-Sq (approx.) = 57.292, df = 30, p-value = 0.001934
```

Obtenemos un p-valor significativo y por tanto rechazamos la hipótesis de homogeneidad de matrices de covarianza.

Es decir, acabamos de ver que no se cumple ninguna de las hipótesis necesarias para aplicar el modelo lda, esto explica en cierta forma los malos resultados obtenidos durante la definición del modelo. Dicho esto, no tiene sentido intentar ajustar un modelo cuadrático ya que las hipótesis de normalidad son las mismas para este modelo.