A3 - Modelización predictiva

Leonardo Segovia Vilchez

Diciembre 2021

Contents

Introducción En esta actividad se usará el fichero de datos (dat_Air) que contiene información sobre diferentes parámetros sobre la calidad del aire de una determinada ciudad europea en el año 2021. Estos datos han sido medidos en tiempo real en diferentes estaciones distribuidas en distintas zonas. Para nuestro estudio se ha seleccionado los datos recopilados de este año por una de las estaciones móviles. Se muestran las medidas de una serie de variables, tanto meteorológicas como de los principales contaminantes del aire (gases y partículas).

Todas ellas contribuyen para determinar el Índice de Calidad del Aire (ICA).

Las variables del fichero de datos son:

- Estacion: Estación móvil.
- Latitud: Latitud del lugar de medición.
- Longitud: Longitud del lugar de medición.
- Fecha: Fecha de medición.
- Periodo: Mediciones cada hora. Periodo de 1 a 24 horas (diarias).
- SO2: Concentración de SO2 (dióxido de azfre) en m g /m 3 .
- H2S: Concentración de H2s (ácido sulfhidrico) en m g /m 3 .
- NO: Concentración de NO (óxido nítrico) en m g /m 3 .
- NO2: Concentración de (dióxido de nitrógeno) en m g /m 3 .
- NOX: Concentración de NOX (óxidos de nitrógeno) en m g /m 3 .
- O3: Concentración de Ozono en m g /m 3 .
- PM10: Partículas en suspension <10 en m g /m 3 .
- PM25: Partíulas en Suspension PM 2,5 en m g /m 3 .
- BEN: Concentración de benceno en m g /m 3 .
- TOL: Tolueno en m g /m 3 .
- MXIL: MXileno en m g /m 3 .
- Dir_Aire: Dirección del viento en grados.
- Vel: Velocidad del viento en m/Sg.
- Tmp: Temperatura en grados centígrados.
- HR: Humedad relativa en % de hr.
- PRB: Presión Atmosférica en mb.
- RS: Radiación Solar W /m 2 .
- LL: Precipitación en l/m 2 .

```
# Librerias
# Paquetes y librerías.
#install.packages("car")
#install.packages("corrplot")
library(pROC)
library(tibble)
library(dplyr)
library(ggplot2)
library(scales)
library(knitr)
library(tidyr)
library(car)
library(stringr)
library(grid)
library(gridExtra)
library(nortest)
library(BSDA)
library(RColorBrewer)
library(caret)
library(corrplot)
library(ResourceSelection)
```

```r
# Cargamos el fichero de datos.
dat_Air <- read.csv('dat_Air.csv',stringsAsFactors = FALSE)

# Mostramos si el dataset se ha cargado correctamente.
head(dat_Air,2)
```

```
##   Estacion  latitud  longitud     Fecha Periodo SO2  H2S NO NO2 NOX O3 PM10
## 1       12 43.52096 -5.690707 11/7/2021      24   1 18.3  5  36  43  5   95
## 2       12 43.52096 -5.690707 11/7/2021      23   1  6.2  2  22  24 10   80
##   PM25 BEN TOL MXIL Dir_Aire  Vel  Tmp HR  PRB RS LL
## 1   30 1.6 2.3  3.2      260 1.49 11.8 93 1025 36  0
## 2   23 0.7 1.8  2.7      201 0.98 11.6 93 1025 36  0
```

Examinamos la interpretación que hace R de cada una de las variables.

```r
# Ejecutamos la función *class* a cada variable del dataset
sapply(dat_Air, class)
```

```
##    Estacion     latitud    longitud       Fecha     Periodo         SO2
##   "integer"   "numeric"   "numeric" "character"   "integer"   "integer"
##         H2S          NO         NO2         NOX          O3        PM10
##   "numeric"   "integer"   "integer"   "integer"   "integer"   "integer"
##        PM25         BEN         TOL        MXIL    Dir_Aire         Vel
##   "integer"   "numeric"   "numeric"   "numeric"   "integer"   "numeric"
##         Tmp          HR         PRB          RS          LL
##   "numeric"   "integer"   "integer"   "integer"   "numeric"
```

Lectura del fichero y preparación de los datos

Leed el fichero y guardad . A continuación, verificad que los datos se han cargado correctamente.

```r
# Valores resumenes.
str(dat_Air)
```

```
## 'data.frame':    7464 obs. of  23 variables:
##  $ Estacion: int  12 12 12 12 12 12 12 12 12 12 ...
##  $ latitud : num  43.5 43.5 43.5 43.5 43.5 ...
##  $ longitud: num  -5.69 -5.69 -5.69 -5.69 -5.69 ...
##  $ Fecha   : chr  "11/7/2021" "11/7/2021" "11/7/2021" "11/7/2021" ...
##  $ Periodo : int  24 23 22 21 20 19 18 17 16 15 ...
##  $ SO2     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ H2S     : num  18.3 6.2 4.3 2.7 2.1 ...
##  $ NO      : int  5 2 1 4 2 1 1 1 1 2 ...
##  $ NO2     : int  36 22 19 21 18 10 8 8 10 13 ...
##  $ NOX     : int  43 24 20 26 20 10 7 9 10 16 ...
##  $ O3      : int  5 10 14 11 16 24 36 55 53 52 ...
##  $ PM10    : int  95 80 38 36 30 21 15 24 14 15 ...
##  $ PM25    : int  30 23 15 12 12 17 12 10 6 10 ...
##  $ BEN     : num  1.6 0.7 0.6 0.9 0.2 ...
##  $ TOL     : num  2.3 1.8 1.1 1.3 0.6 ...
##  $ MXIL    : num  3.2 2.7 3.1 2.2 0.8 ...
##  $ Dir_Aire: int  260 201 296 185 297 295 309 23 241 24 ...
##  $ Vel     : num  1.49 0.98 0.98 0.99 0.99 ...
##  $ Tmp     : num  11.8 11.6 11.2 11.9 12.4 ...
##  $ HR      : int  93 93 92 88 85 83 76 71 67 61 ...
```

```
##  $ PRB     : int  1025 1025 1025 1026 1026 1026 1026 1026 1026 1026 ...
##  $ RS      : int  36 36 37 37 37 37 37 47 116 194 ...
##  $ LL      : num  0 0 0 0 0 0 0 0 0 0 ...
```

```r
# Valores resumenes.
summary(dat_Air)
```

```
##     Estacion     latitud         longitud         Fecha
##  Min.   :12   Min.   :43.52   Min.   :-5.691   Length:7464
##  1st Qu.:12   1st Qu.:43.52   1st Qu.:-5.691   Class :character
##  Median :12   Median :43.52   Median :-5.691   Mode  :character
##  Mean   :12   Mean   :43.52   Mean   :-5.691
##  3rd Qu.:12   3rd Qu.:43.52   3rd Qu.:-5.691
##  Max.   :12   Max.   :43.52   Max.   :-5.691
##
##     Periodo           SO2               H2S               NO
##  Min.   : 1.00   Min.   :  1.000   Min.   : 0.800   Min.   :  1.00
##  1st Qu.: 6.75   1st Qu.:  2.000   1st Qu.: 1.000   1st Qu.:  1.00
##  Median :12.50   Median :  3.000   Median : 2.000   Median :  1.00
##  Mean   :12.50   Mean   :  4.804   Mean   : 2.526   Mean   :  4.14
##  3rd Qu.:18.25   3rd Qu.:  6.000   3rd Qu.: 3.400   3rd Qu.:  3.00
##  Max.   :24.00   Max.   :105.000   Max.   :33.000   Max.   :111.00
##                  NA's   :54        NA's   :50       NA's   :63
##       NO2             NOX               O3             PM10
##  Min.   : 2.00   Min.   :  2.00   Min.   : 1.0   Min.   :  0.00
##  1st Qu.: 5.00   1st Qu.:  5.00   1st Qu.:20.0   1st Qu.: 15.00
##  Median :10.00   Median : 11.00   Median :47.0   Median : 24.00
##  Mean   :12.96   Mean   : 18.29   Mean   :42.2   Mean   : 42.83
##  3rd Qu.:18.00   3rd Qu.: 23.00   3rd Qu.:61.0   3rd Qu.: 45.00
##  Max.   :68.00   Max.   :221.00   Max.   :98.0   Max.   :724.00
##  NA's   :63      NA's   :52       NA's   :62     NA's   :43
##       PM25            BEN               TOL               MXIL
##  Min.   : 0.00   Min.   : 0.1000   Min.   :  0.100   Min.   : 0.100
##  1st Qu.: 7.00   1st Qu.: 0.1000   1st Qu.:  0.200   1st Qu.: 0.400
##  Median :11.00   Median : 0.2000   Median :  0.500   Median : 0.900
##  Mean   :12.88   Mean   : 0.6108   Mean   :  1.883   Mean   : 1.935
##  3rd Qu.:16.00   3rd Qu.: 0.5000   3rd Qu.:  1.400   3rd Qu.: 2.000
##  Max.   :96.00   Max.   :21.8000   Max.   :160.300   Max.   :53.600
##  NA's   :142     NA's   :88        NA's   :86        NA's   :86
##     Dir_Aire          Vel              Tmp              HR
##  Min.   :  1.0   Min.   :0.730   Min.   : 1.10   Min.   : 30.00
##  1st Qu.: 73.0   1st Qu.:1.000   1st Qu.:11.70   1st Qu.: 81.00
##  Median :215.0   Median :1.650   Median :15.20   Median : 92.00
##  Mean   :181.8   Mean   :2.409   Mean   :14.69   Mean   : 88.03
##  3rd Qu.:268.0   3rd Qu.:3.530   3rd Qu.:18.20   3rd Qu.: 99.00
##  Max.   :360.0   Max.   :9.760   Max.   :26.40   Max.   :100.00
##
##       PRB             RS              LL
##  Min.   : 991   Min.   : 24.0   Min.   : 0.00000
##  1st Qu.:1012   1st Qu.: 37.0   1st Qu.: 0.00000
##  Median :1016   Median : 46.0   Median : 0.00000
##  Mean   :1016   Mean   :133.5   Mean   : 0.08355
##  3rd Qu.:1020   3rd Qu.:184.0   3rd Qu.: 0.00000
##  Max.   :1031   Max.   :689.0   Max.   :13.00000
##
```

Observamos que las siguientes variable contienen valores nulos: SO2, H2S, NO, NO2, NOX, O3, PM10, PM25, BEN, TOL y MXIL

Imputamos los valores nulos con la mediana

```
sel<-which(is.na(dat_Air$SO2))
dat_Air[sel,"SO2"]<-median(dat_Air$SO2,na.rm = T)

sel<-which(is.na(dat_Air$H2S))
dat_Air[sel,"H2S"]<-median(dat_Air$H2S,na.rm = T)

sel<-which(is.na(dat_Air$NO))
dat_Air[sel,"NO"]<-median(dat_Air$NO,na.rm = T)

sel<-which(is.na(dat_Air$NO2))
dat_Air[sel,"NO2"]<-median(dat_Air$NO2,na.rm = T)

sel<-which(is.na(dat_Air$O3))
dat_Air[sel,"O3"]<-median(dat_Air$O3,na.rm = T)

sel<-which(is.na(dat_Air$PM10))
dat_Air[sel,"PM10"]<-median(dat_Air$PM10,na.rm = T)

sel<-which(is.na(dat_Air$TOL))
dat_Air[sel,"TOL"]<-median(dat_Air$TOL,na.rm = T)

sel<-which(is.na(dat_Air$PM25))
dat_Air[sel,"PM25"]<-median(dat_Air$PM25,na.rm = T)

sel<-which(is.na(dat_Air$BEN))
dat_Air[sel,"BEN"]<-median(dat_Air$BEN,na.rm = T)

sel<-which(is.na(dat_Air$MXIL))
dat_Air[sel,"MXIL"]<-median(dat_Air$MXIL,na.rm = T)

sel<-which(is.na(dat_Air$NOX))
dat_Air[sel,"NOX"]<-median(dat_Air$NOX,na.rm = T)
```

Regresión lineal

La calidad del aire ha sufrido cambios que afectan a nuestro modo de vida, por lo que resulta necesario estudiarlo. Para ello se toman medidas de la emisión de diferentes contaminantes y de factores metereológicos como por ejemplo el viento, la precipitación, radiación solar o la temperatura, con el fin de buscar relaciones entre dichas variables.

En este estudio se quiere demostrar la existencia de relación lineal entre los contaminantes atmosféricos y las variables metereológicas.

Estudio de correlación lineal

Se pide calcularla matriz de correlación entre las variables siguientes: Contaminantes: O3, NO2 y PM10, junto con las variables metereológicas: Tmp, HR, RS, Vel y Dir.

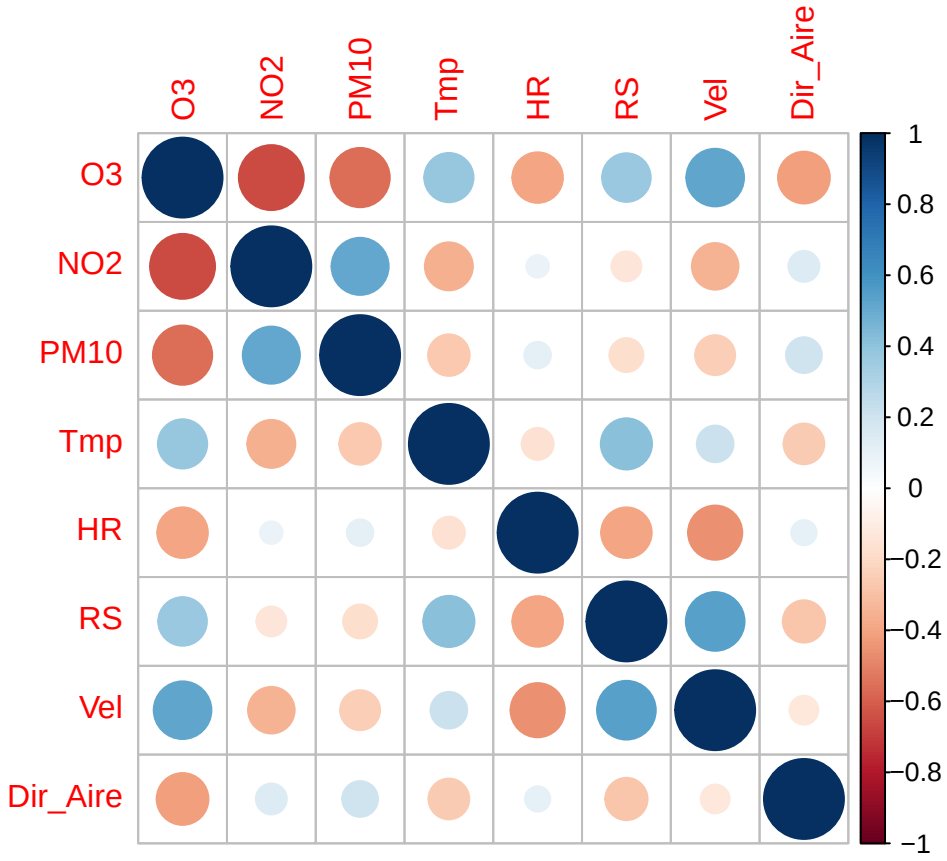
```
index <- c('O3','NO2','PM10','Tmp','HR','RS','Vel','Dir_Aire')
cor_dat_Air <- dat_Air[, index]
head(cor_dat_Air,3)
```

```
##   O3 NO2 PM10  Tmp HR RS  Vel Dir_Aire
## 1  5  36   95 11.8 93 36 1.49      260
## 2 10  22   80 11.6 93 36 0.98      201
## 3 14  19   38 11.2 92 37 0.98      296
```

```
cor_dat_Air <- cor(cor_dat_Air, method='pearson')
cor_dat_Air <- round(cor_dat_Air, digits=2)
cor_dat_Air
```

```
##             O3   NO2  PM10   Tmp    HR    RS   Vel Dir_Aire
## O3        1.00 -0.66 -0.55  0.38 -0.40  0.37  0.52    -0.42
## NO2      -0.66  1.00  0.51 -0.36  0.08 -0.14 -0.34     0.15
## PM10     -0.55  0.51  1.00 -0.27  0.11 -0.18 -0.25     0.20
## Tmp       0.38 -0.36 -0.27  1.00 -0.16  0.41  0.21    -0.26
## HR       -0.40  0.08  0.11 -0.16  1.00 -0.40 -0.46     0.10
## RS        0.37 -0.14 -0.18  0.41 -0.40  1.00  0.54    -0.28
## Vel       0.52 -0.34 -0.25  0.21 -0.46  0.54  1.00    -0.13
## Dir_Aire -0.42  0.15  0.20 -0.26  0.10 -0.28 -0.13     1.00
```

```
corrplot(cor_dat_Air)
```



- RS: Radiación Solar W /m 2.
- NO2: Concentración de (dióxido de nitrógeno) en m g /m 3 .
- O3: Concentración de Ozono en m g /m 3 .
- PM10: Partículas en suspension <10 en m g /m 3.
- Dir_Aire: Dirección del viento en grados.
- Vel: Velocidad del viento en m/Sg.
- Tmp: Temperatura en grados centígrados.
- HR: Humedad relativa en % de hr.

a) ¿Cual de los contaminantes atmosféricos citados anteriormente, tienen una mayor relación lineal con la RS? Interpretar las relaciones de dicho contaminante con la RS y también con el resto de variables metereológicas.

Podemos observar que la relación lineal más fuerte con la variable 'Radiación Solar' coresponde a variable de 'Concentración de Ozono', teniendo está una correlación positiva. Los dos contaminantes restantes, concentración de dióxido de nitrógeno y partículas en suspension tienen un coorelación inferior, aproximadamente la mitad, y negativa. Notamos también que la relacion RS <-> O3 es muy parecida a Tmp <-> O3.

También podemos observar fenomenos más compresibles como la relación positiva de la radiación solar con la temperatura y la relación negativa respecto a la humedad. También es relevante la fuerte relación positiva entre el viento y la radiación solar / temperatura. Aunque cabe remarcar que la corelación no significa causalidad sino simplemente que las dos variables tienen una tendencia parecidad pudiendo ser está positiva o negativa

b) Se toma la media diaria de cada una de las variables del apartado a) y posteriormente se estudia de nuevo la relación pedida en dicho apartado. ¿Existe alguna diferencia en la relación entre las nuevas variables construídas con los valores medios diarios, con respecto a los resultados obtenidos anteriormente?

```
cor_dat_Air <- dat_Air %>% group_by(Fecha)
```

```r
cor_mean_dat_Air <- cor_dat_Air %>% summarise(
  NO2 = mean(NO2),
  O3 = mean(O3),
  PM10 = mean(PM10),
  Dir_Aire = mean(Dir_Aire),
  RS = mean(RS),
  Vel = mean(Vel),
  Tmp = mean(Tmp),
  HR = mean(HR)
)

head(cor_mean_dat_Air,3)
```

```
## # A tibble: 3 x 9
##   Fecha       NO2    O3  PM10 Dir_Aire    RS   Vel   Tmp    HR
##   <chr>     <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1/1/2021  10.2   51.2  28.7    260.   57.6  2.61  5.69  91.4
## 2 1/10/2021  8.04  56.3  27.8     88.2  60.0  4.03  7.69  74.7
## 3 1/11/2021 27.6   21.6  85.6    219.   75.6  1.38  6.73  81.2
```

```r
index <- c('O3','NO2','PM10','Tmp','HR','RS','Vel','Dir_Aire')
cor_mean_dat_Air <- cor_mean_dat_Air[, index]
cor_mean_dat_Air <- cor(cor_mean_dat_Air, method='pearson')
cor_mean_dat_Air <- round(cor_mean_dat_Air, digits=2)
cor_mean_dat_Air
```

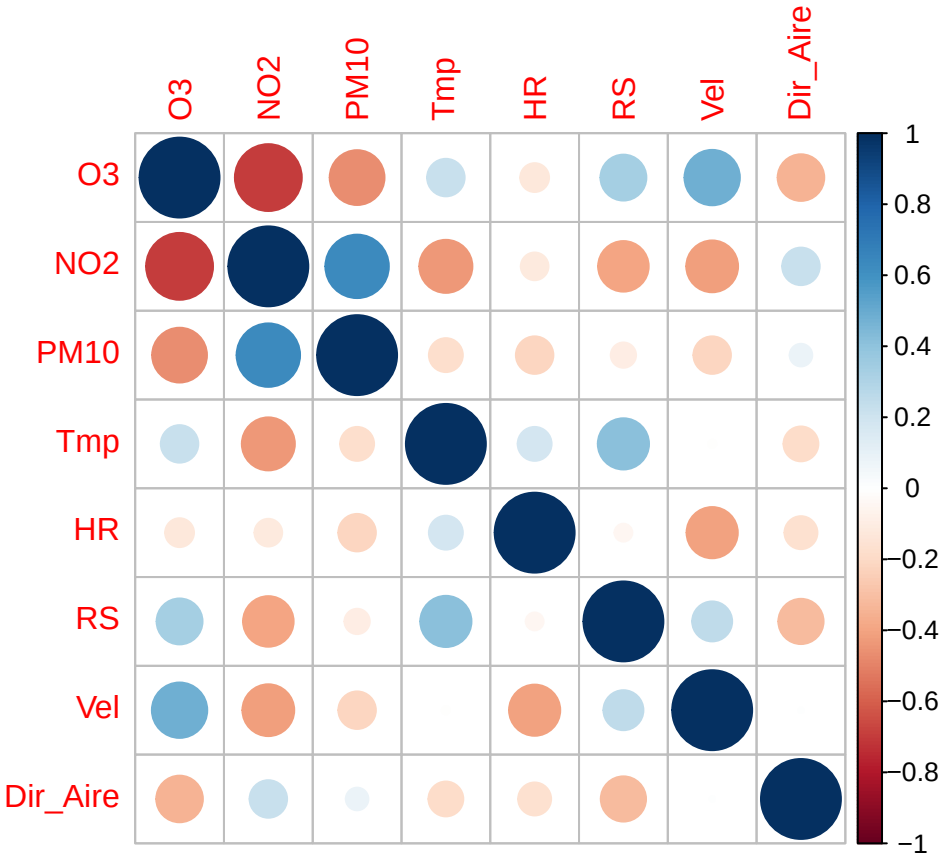
```
##             O3   NO2  PM10   Tmp    HR    RS   Vel Dir_Aire
## O3        1.00 -0.70 -0.47  0.22 -0.13  0.33  0.48    -0.34
## NO2      -0.70  1.00  0.63 -0.44 -0.12 -0.40 -0.42     0.22
## PM10     -0.47  0.63  1.00 -0.18 -0.22 -0.10 -0.22     0.08
## Tmp       0.22 -0.44 -0.18  1.00  0.18  0.41 -0.01    -0.19
## HR       -0.13 -0.12 -0.22  0.18  1.00 -0.05 -0.41    -0.17
## RS        0.33 -0.40 -0.10  0.41 -0.05  1.00  0.25    -0.32
## Vel       0.48 -0.42 -0.22 -0.01 -0.41  0.25  1.00     0.00
## Dir_Aire -0.34  0.22  0.08 -0.19 -0.17 -0.32  0.00     1.00
```

- NO2: Concentración de (dióxido de nitrógeno) en m g /m 3 .
- O3: Concentración de Ozono en m g /m 3 .
- PM10: Partículas en suspension <10 en m g /m 3 .

En este caso, utilizando la media, el valor con más coorelación varía con respecto a los datos por hora. La relación lineal más fuerte con la variable 'Radiación Solar' coresponde a variable de 'Concentración de (dióxido de nitrógeno)'. Otros valores que varían significativamente son las relaciones con la variable Humedad relativa que pasa del -0.40 a -0.05 y la Velocidad del viento del 0.54 al 0.25

Se demuestra también de forma visual.

```r
corrplot(cor_mean_dat_Air)
```


O3
NO2
PM10
Tmp
HR
RS
Vel
Dir_Aire
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Modelo de regresión lineal

Se quiere explicar el nivel de ozono en función de la radiación solar.

a) Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable O3 en función de la radiación solar (RS). Se evaluará la bondad del ajuste, a partir del coeficiente de determinación.

```
Model_lineal_O3_RS <- lm(O3~RS, data=dat_Air )
summary(Model_lineal_O3_RS)
```

```
##
## Call:
## lm(formula = O3 ~ RS, data = dat_Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.421 -19.407   2.048  17.242  60.362
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.414838   0.341349  100.82   <2e-16 ***
## RS           0.058603   0.001705   34.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.98 on 7462 degrees of freedom
## Multiple R-squared:  0.1367, Adjusted R-squared:  0.1366
## F-statistic:  1182 on 1 and 7462 DF,  p-value: < 2.2e-16
```

El modelo de regresión es significativo (p-value: < 2.2e-16), con un R2 ajustado de 13.66%. RS explica el 13.66% de la variacion de O3. Es un modelo muy pobre

b) Para calcular el índice de calidad del aire, se establecen diferentes categorías, según sea la concentración de cada contaminante. En este apartado se tomará como contaminante la concentración de PM10 y se establecerán las siguientes categorías, para construir el PM10_cat (Indice de calidad del Aire, en función de PM10):

- Muy buena: valores de (0 a 40],
- Buena: valores de (40 a 60],
- Mejorable: valores de (60 a 120],
- Mala: valores de (120 a 160],
- Muy mala: valores de (160 a 724]

Se pide, construir un modelo de regresión lineal, tomando como variable dependiente (O3) y la variable explicativa PM10_cat. Interpretar los resultados.

Nota: Este apartado se podría interpretar también mediante el ANOVA. Dicho modelo se verá en la actividad A4.

```
dat_Air$PM10_cat[(0  <= dat_Air$PM10)   & (dat_Air$PM10 < 40 )] <- 'Muy buena'
dat_Air$PM10_cat[(40 <= dat_Air$PM10)   & (dat_Air$PM10 < 60 )] <- 'Buena'

dat_Air$PM10_cat[(60  <= dat_Air$PM10)  & (dat_Air$PM10 < 120)] <- 'Mejorable'
dat_Air$PM10_cat[(120  <= dat_Air$PM10) & (dat_Air$PM10 < 160)] <- 'Mala'
dat_Air$PM10_cat[(160  <= dat_Air$PM10) & (dat_Air$PM10 < 724)] <- 'Muy mala'
dat_Air$PM10_cat   <-as.factor(dat_Air$PM10_cat)

Model_lineal_O3_PM10_cat <- lm(O3~factor(PM10_cat), data=dat_Air )
```

```r
summary(Model_lineal_O3_PM10_cat)
```

```
##
## Call:
## lm(formula = O3 ~ factor(PM10_cat), data = dat_Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.663 -10.663   0.337  11.337  65.945
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 30.1898     0.6630   45.53   <2e-16 ***
## factor(PM10_cat)Mala       -22.3898     1.2966  -17.27   <2e-16 ***
## factor(PM10_cat)Mejorable  -14.9404     0.8961  -16.67   <2e-16 ***
## factor(PM10_cat)Muy buena   21.4729     0.7054   30.44   <2e-16 ***
## factor(PM10_cat)Muy mala   -23.1352     1.2243  -18.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.62 on 7458 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4455, Adjusted R-squared:  0.4452
## F-statistic:  1498 on 4 and 7458 DF,  p-value: < 2.2e-16
```

O3: Concentración de Ozono en m g /m 3 . PM10: Partículas en suspension <10 en m g /m 3.

Los residuos están bastante bien distribuidos sobre la mediana (cercana a cero) ya que 1Q y 3Q tienen valors parecidos.

Primero observamos que todas las estimaciones contiene valores significations (p-value: < 2.2e-16). Esto quiere decir que la probabilidad de que el valor obtenido se deba al azar es aceptable.

El fenomeno que se observa, es que el aumento en la mejora del Indice de calidad del Aire aumenta con una buena medida de partículas en suspension. Esto se ve claramente cuando las partículas en suspension es Muy buena. Sucede el efecto contrario para los escenarios de partículas en suspension inferiores, por el siguiente orden; Mejorable, Mala y Muy mala. Todos ellos con una relación negativa.

Modelo de regresión lineal múltiple

Se quiere explicar el nivel de ozono en función de la radiación solar (RS),concentración de dióxido de nitrógeno (NO2), temperatura (Tmp) y dirección del aire (Dir_Aire).

a) Primero, se añadirá al modelo del apartado a), la variable explicativa (Dir_Aire).¿El modelo ha mejorado?

```
Model_lineal_O3_RS_Dir_Aire <- lm(O3~RS+Dir_Aire, data=dat_Air )
summary(Model_lineal_O3_RS_Dir_Aire)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire, data = dat_Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.589 -17.017   1.021  15.458  57.875
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.962665   0.600493   84.87   <2e-16 ***
## RS           0.043566   0.001661   26.23   <2e-16 ***
## Dir_Aire    -0.079998   0.002458  -32.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.57 on 7461 degrees of freedom
## Multiple R-squared:  0.244,  Adjusted R-squared:  0.2438
## F-statistic:  1204 on 2 and 7461 DF,  p-value: < 2.2e-16
```

El modelo de regresión es significativo (p-value: < 2.2e-16) y ha mejorado pasando del R2 13.66% al 24.38%.

b) Posteriormente se añade al modelo anterior la variable (NO2). ¿Existe una mejora del modelo?

```
Model_lineal_O3_RS_Dir_Aire <- lm(O3~RS+Dir_Aire+NO2, data=dat_Air )
summary(Model_lineal_O3_RS_Dir_Aire)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire + NO2, data = dat_Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.442 -11.319   0.043  10.760  56.618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.795445   0.498588  135.97   <2e-16 ***
## RS           0.033714   0.001246   27.06   <2e-16 ***
## Dir_Aire    -0.063247   0.001847  -34.24   <2e-16 ***
## NO2         -1.435581   0.018634  -77.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.35 on 7460 degrees of freedom
## Multiple R-squared:  0.579,  Adjusted R-squared:  0.5788
```

```
## F-statistic:  3420 on 3 and 7460 DF,  p-value: < 2.2e-16
```

El modelo de regresión es significativo (p-value: < 2.2e-16) y ha mejorado significativamente pasando del R2 24.38% al 57.88%.

c) Se toma la variable (Tmp) y se añade al modelo anterior. Se pide comprobar la presencia o no de colinealidad entre las variables (RS) y (Tmp). Podéis usar la librería (faraway) y estudiar el FIV (factor de inflación de la varianza). Según la conclusión obtenida, discutir si sería indicado o no añadir la variable (Tmp) al modelo. De ser afirmativa la respuesta, construye el modelo e interpreta el resultado.

```
Model_lineal_O3_RS_Dir_Aire_tmp<- lm(O3~RS+Dir_Aire+NO2+Tmp, data=dat_Air )
summary(Model_lineal_O3_RS_Dir_Aire_tmp)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire + NO2 + Tmp, data = dat_Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.409 -11.323   0.125  10.774  56.146
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.434021   0.902408   73.62   <2e-16 ***
## RS           0.032877   0.001329   24.74   <2e-16 ***
## Dir_Aire    -0.062803   0.001863  -33.70   <2e-16 ***
## NO2         -1.424119   0.019679  -72.37   <2e-16 ***
## Tmp          0.084726   0.046813    1.81   0.0704 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.35 on 7459 degrees of freedom
## Multiple R-squared:  0.5792, Adjusted R-squared:  0.5789
## F-statistic:  2566 on 4 and 7459 DF,  p-value: < 2.2e-16
```

El modelo de regresión es significativo (p-value: < 2.2e-16) y no ha mejorado con respecto al modelo anterior.

```
vif(Model_lineal_O3_RS_Dir_Aire_tmp)
```

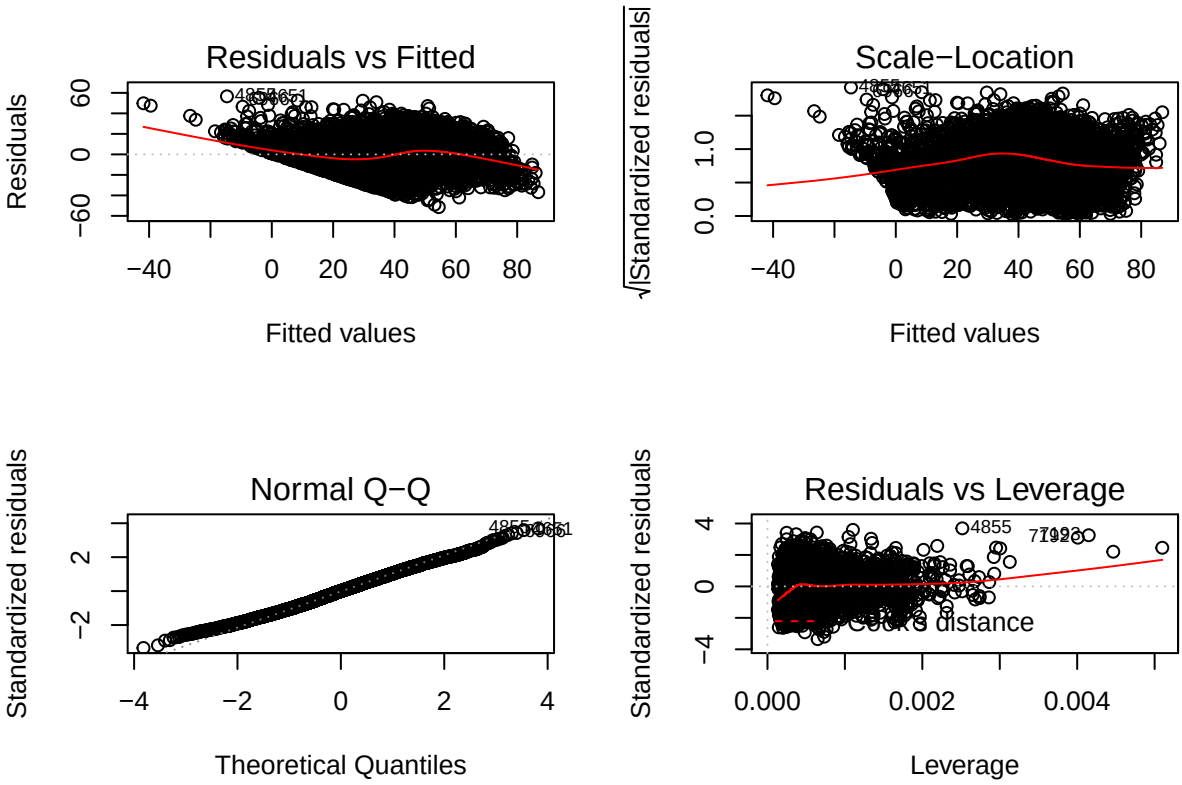
```
##       RS Dir_Aire      NO2      Tmp
## 1.246482 1.118493 1.153627 1.371999
```

Ningun valor supera el umbral de 5 que sería preocupante por la colinealidad. De todas maneras observamos que la variable Tmp contiene el valor más alto y no aporta nueva información al modelo (R2) por lo que podiamos excluirla del modelo final.

Diagnosis del modelo

Para la diagnosis se escoge el modelo construído en el apartado b) y se piden dos gráficos: uno con los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y el gráfico normalmente(QQ plot).Interpretar los resultados.

```
layout(matrix(c(1,2,3,4),2,2))
plot(Model_lineal_O3_RS_Dir_Aire)
```



El estudio de los residuos nos permite extraer información acerca del cumplimiento de las suposiciones de modelización. De la gráfica Normal Q-Q obtenemos la siguien información; la mayoría de datos centrales siguen la distribución normal, detectando algunas observaciones extremas que se desbían. Por otro lado, el gráfico de residuos frente valores ajustados, muestra un patrón "aleatorio" de los residuos, excepto en los extremos. Por lo que aceptaremos el cumplimiento de las suposiciones.

Predicción del modelo

Según el modelo del apartado c), calcular la concentración de O3, si se tienen valores de RS de 180, NO2 de 15, Dir_Aire de 250 grados y Tmp de 20 grados centígrados.

```
predict_data <-data.frame(RS =180, NO2=15 ,Dir_Aire=250 ,Tmp=20)
pre_Model_lineal_O3_RS_Dir_Aire_tmp<- predict(Model_lineal_O3_RS_Dir_Aire_tmp, predict_data ,interval =

pre_Model_lineal_O3_RS_Dir_Aire_tmp
```

```
##        fit      lwr      upr
## 1 36.98376 36.30034 37.66718
```

El 95% del intervalo de predición de la concentraciónn de O3 es de 36.98376 entre el intervalo de 37.66 como max y minimo de 36.30

Regresión logística

Se quiere estudiar la concentración de O3 del aire de una determinada ciudad. Primero se creará una nueva variable dicotómica llamada icO3 (índice de calidad del aire basado en O3). Se codificará de la siguiente manera:

buena: valores de (0 a 80], mejorable: valores de (80 a 100]

Posteriormente se recodificará como valor 0 la categoría " buena". En caso contrario se codificará con el valor 1.

Nota: Dicho índice de calidad se ha recodificado conforme a nuestros datos.

```
dat_Air$icO3[(0  <= dat_Air$O3)  & (dat_Air$O3 < 80 )]    <- 0
dat_Air$icO3[(80 <= dat_Air$O3) & (dat_Air$O3 <= 100 )]   <- 1
dat_Air$icO3   <-as.factor(dat_Air$icO3)
levels(dat_Air$icO3)
```

```
## [1] "0" "1"
```

```
levels(dat_Air$icO3) <- c('buena', 'mejorable')
levels(dat_Air$icO3)
```

```
## [1] "buena"     "mejorable"
```

Estudio de relaciones entre variables. Análisis crudo de posibles factores de riesgo

a) Se visualiza la relación entre ic O3 y las variables independientes: RS, Vel y HR. Para ello se recodificaran las variables RS y Vel, dejando la variable cuantitativa HR, tal como está en la base de datos.

Para comprobar si existe asociación entre las variable dependiente y cada una de las variables explicativas, se aplicará el test Chi-cuadrado de Pearson. Un resultado significativo nos dirá que existe asociación. Se procederán a categorizar las variables explicativas de la siguiente forma:

Radiación solar (RS_cat2): * normal_baja:(0 a 100], * normal_alta: valores de (100 a 700]

Velocidad del viento (Vel_cat2): * flojo: valores de (0 a 3], * moderado: valores de (3 a 10]

```
dat_Air$RS_cat2[(0  <= dat_Air$RS)  & (dat_Air$RS < 100 )]    <- 0
dat_Air$RS_cat2[(100 <= dat_Air$RS) & (dat_Air$RS < 700 )]    <- 1
dat_Air$RS_cat2   <-as.factor(dat_Air$RS_cat2)
levels(dat_Air$RS_cat2)
```

```
## [1] "0" "1"
```

```
levels(dat_Air$RS_cat2) <- c('normal_baja', 'normal_alta')
levels(dat_Air$RS_cat2)
```

```
## [1] "normal_baja" "normal_alta"
```

```
dat_Air$Vel_cat2[(0  <= dat_Air$Vel)  & (dat_Air$Vel < 3 )]    <- 0
dat_Air$Vel_cat2[(3 <= dat_Air$Vel) & (dat_Air$Vel < 10 )]    <- 1
dat_Air$Vel_cat2 <-as.factor(dat_Air$Vel_cat2)
levels(dat_Air$Vel_cat2)
```

```
## [1] "0" "1"
```

```
levels(dat_Air$Vel_cat2) <- c('flojo', 'moderado')
levels(dat_Air$Vel_cat2)
```

```
## [1] "flojo"    "moderado"
```

```
model_icO3_RS_cat2 = glm(formula=icO3~RS_cat2, data=dat_Air, family=binomial (link=logit))
model_icO3_Vel_cat2 = glm(formula=icO3~Vel_cat2, data=dat_Air, family=binomial (link=logit))
model_icO3_HR = glm(formula=icO3~HR, data=dat_Air, family=binomial (link=logit))

1-pchisq(sum(residuals(model_icO3_RS_cat2,type="pearson")^2),1)
```

```
## [1] 0
```

```
1-pchisq(sum(residuals(model_icO3_Vel_cat2,type="pearson")^2),1)
```

```
## [1] 0
```

```
1-pchisq(sum(residuals(model_icO3_HR,type="pearson")^2),1)
```

```
## [1] 0
```

Para todas se obtiene un resultado significativo. Por lo que se demuestra que existe asociación.

b) Posteriormente, para conocer el grado de dicha asociación, se calculará las OR (Odds-Ratio). Importante: Para el cálculo de las OR, se partirá de la tabla de contingencia y se calculará a partir de su fórmula. Debéis implementar dicha fórmula en R. Interpretar las OR calculadas.

- O3: Concentración de Ozono en m g /m 3 .
- RS: Radiación Solar W /m 2 .

Lo realizamos solo a las variables cuantitativas.

```
table(dat_Air$icO3, dat_Air$RS_cat2)
```

```
##
##             normal_baja normal_alta
##   buena            4651        2615
##   mejorable          74         124
```

```
prop.table(table(dat_Air$icO3, dat_Air$RS_cat2))
```

```
##
##             normal_baja normal_alta
##   buena     0.623124330 0.350348339
##   mejorable 0.009914255 0.016613076
```

```
odds_normal_baja <-(0.623124330/0.009914255)
odds_normal_alta <- (0.350348339/0.016613076)

odds_normal_baja / odds_normal_alta
```

```
## [1] 2.980332
```

Medida de O3 Buena respecto O3 Mala.

La probabilidad de tener una medida O3 buena cuando RS_cat2 sea normal_baja, es de 2.98 veces mayor respecto al caso de RS_cat2 sea normal_alta.

- O3: Concentración de Ozono en m g /m 3 .
- Vel: Velocidad del viento en m/Sg.

```
prop.table(table(dat_Air$icO3, dat_Air$Vel_cat2))
```

```
##
##                 flojo   moderado
##   buena     0.67256163 0.30091104
##   mejorable 0.01312969 0.01339764
```

```
odds_moderado <- (0.30091104/0.01339764)
odds_flojo <- (0.67256163/0.01312969)

odds_moderado/odds_flojo
```

```
## [1] 0.4384622
```

Medida de O3 Buena respecto O3 Mala.

La probabilidad de tener una medida O3 buena cuando Vel_cat2 sea moderado, es de 0.43 veces respecto al caso de Vel_cat2 sea Flojo.

Modelo de regresión logística

a) Estimad el modelo de regresión logística tomando como variable dependiente icO3 y variable explicativa RS_cat2. Calculad la OR a partir de los resultados del modelo y su intervalo de confianza. ¿Se puede considerar que la radiación solar es un factor de riesgo? Justifica tu respuesta.

```
model_icO3_RS_cat2 <- glm(formula=icO3~RS_cat2, data=dat_Air, family=binomial (link=logit))
summary(model_icO3_RS_cat2)
```

```
##
## Call:
## glm(formula = icO3 ~ RS_cat2, family = binomial(link = logit),
##     data = dat_Air)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3044  -0.3044  -0.1777  -0.1777   2.8832
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.1408     0.1172 -35.340  < 2e-16 ***
## RS_cat2normal_alta   1.0920     0.1489   7.333 2.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1828.0  on 7463  degrees of freedom
## Residual deviance: 1771.9  on 7462  degrees of freedom
## AIC: 1775.9
##
## Number of Fisher Scoring iterations: 7
```

El signo positivo del coeficiente hace indicar que una radiación normal alta hace disminuir la calidad del aire, por lo que aumentaria la etiqueta O3 Mejoreable.

```
exp(cbind(coef(model_icO3_RS_cat2),confint(model_icO3_RS_cat2)))
```

```
##                                     2.5 %     97.5 %
## (Intercept)         0.01591056 0.01253437 0.01985449
## RS_cat2normal_alta 2.98033177 2.23242138 4.00604114
```

La occurrencia de que la calidad del aire empeore (o sea que la etiqueta normal_alta aumente) es de 2.98 veces mayor en relación cuando RS_cat es normal_baja

Por lo que considero RS de factor de riesgo.

b) Se crea un nuevo modelo con la misma variable dependiente y se añade al apartado a) la variable TMP. Interpretar si nos encontramos o no ante una posible variable de confusión.

```
model_ic03_RS_cat2_Tmp <- glm(formula=ic03~RS_cat2+Tmp, data=dat_Air, family=binomial (link=logit))
summary(model_ic03_RS_cat2_Tmp)
```

```
##
## Call:
## glm(formula = ic03 ~ RS_cat2 + Tmp, family = binomial(link = logit),
##     data = dat_Air)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4773  -0.2639  -0.2021  -0.1567   3.0783
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.64021    0.33091 -17.045  < 2e-16 ***
## RS_cat2normal_alta  0.76032    0.16076   4.729 2.25e-06 ***
## Tmp                 0.10474    0.02044   5.124 2.99e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1828.0  on 7463  degrees of freedom
## Residual deviance: 1743.2  on 7461  degrees of freedom
## AIC: 1749.2
##
## Number of Fisher Scoring iterations: 7
```

EL valor del coeficiente del RS_cat2normal_alta cambia significativamente al introducir el la nueva variable Tmp. Ademá, podemos ver que Tmp esta relacionada con la variable dependiente y la independiente RS_cat2. Veasé abajo los coeficientes cernanos a cero.

Por lo que estaíamos ante una posible variable de confusión.

```
model_ic03_Tmp <- glm(formula=ic03~Tmp, data=dat_Air, family=binomial (link=logit))
model_ic03_Tmp
```

```
##
## Call:  glm(formula = ic03 ~ Tmp, family = binomial(link = logit), data = dat_Air)
##
## Coefficients:
## (Intercept)          Tmp
##     -5.8676       0.1425
##
## Degrees of Freedom: 7463 Total (i.e. Null);  7462 Residual
## Null Deviance:       1828
## Residual Deviance: 1766  AIC: 1770
```

```
model_RS_cat2_Tmp <- glm(formula=RS_cat2~Tmp, data=dat_Air, family=binomial (link=logit))
model_RS_cat2_Tmp
```

```
##
## Call:  glm(formula = RS_cat2 ~ Tmp, family = binomial(link = logit),
##     data = dat_Air)
```

```
##
## Coefficients:
## (Intercept)          Tmp
##      -3.767        0.211
##
## Degrees of Freedom: 7463 Total (i.e. Null);  7462 Residual
## Null Deviance:       9812
## Residual Deviance: 8669  AIC: 8673
```

c) Se añade al modelo del apartado a) la variable HR. Estudiar la existencia o no de interacción entre las variables explicativas RS_cat2 y HR. Interpretar.

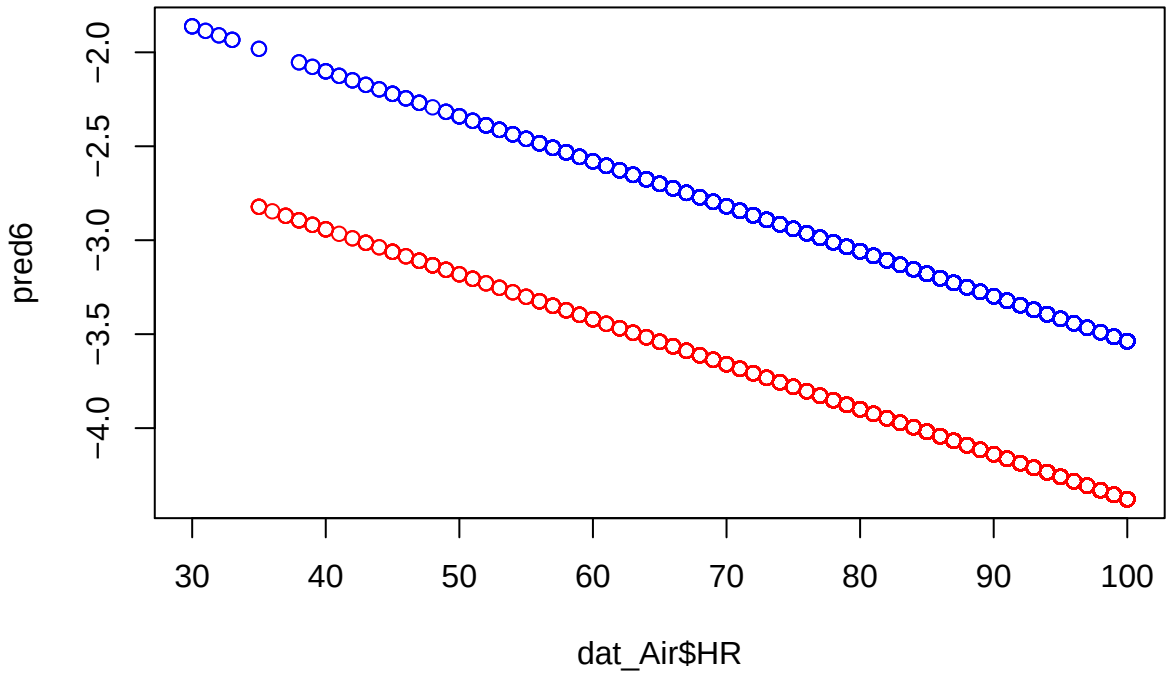
```
model_ico3_RS_cat2_HR <- glm(formula=ico3~RS_cat2+HR+RS_cat2:HR, data=dat_Air, family=binomial)
summary(model_ico3_RS_cat2_HR)
```

```
##
## Call:
## glm(formula = ico3 ~ RS_cat2 + HR + RS_cat2:HR, family = binomial,
##     data = dat_Air)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6790  -0.3016  -0.1775  -0.1288   3.1142
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.530099   0.517199   1.025    0.305
## RS_cat2normal_alta     -3.977190   0.796653  -4.992 5.96e-07 ***
## HR                     -0.053713   0.006259  -8.582  < 2e-16 ***
## RS_cat2normal_alta:HR   0.058579   0.009604   6.100 1.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1828.0  on 7463  degrees of freedom
## Residual deviance: 1714.2  on 7460  degrees of freedom
## AIC: 1722.2
##
## Number of Fisher Scoring iterations: 7
```

RS_cat2normal_alta:HR es estadísticamente significativo.También, observamos que la esimación cambia muy significativamente. En este caso no podemos saber si existe interacción o no, porque este cambio es algo normal para interaciones de variables continuas. Por lo que representamos el logit del modelo.

```
model_ico3_RS_cat2_HR <- glm(formula=ico3~RS_cat2+HR, data=dat_Air, family=binomial)
pred6=predict(model_ico3_RS_cat2_HR,type="link")
plot(dat_Air$HR,pred6,type="n",main="NO interaccion")

points(dat_Air$HR[dat_Air$RS_cat2=='normal_baja'],pred6[dat_Air$RS_cat2=='normal_baja'],col=2)
points(dat_Air$HR[dat_Air$RS_cat2=='normal_alta'],pred6[dat_Air$RS_cat2=='normal_alta'],col=4)
```



Se tiene que las rectas son paralelas para cada nivel del factor; normal_baja y normal_alta, por lo que la asociación entre O3 y RS_cat2 no varía por la HR, es decir, dicha variable no modifica el efecto del factor RS_cat2, por lo tanto, no existe interacción.

d) Se crea un nuevo modelo con las variables explicativas RS_cat2 y Dir_Aire. ¿Existe una mejora del modelo?

```
model_icO3_RS_cat2_Dir_Aire <- glm(formula=icO3~RS_cat2+Dir_Aire, data=dat_Air, family=binomial (link=l
summary(model_icO3_RS_cat2_Dir_Aire)
```

```
##
## Call:
## glm(formula = icO3 ~ RS_cat2 + Dir_Aire, family = binomial(link = logit),
##     data = dat_Air)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4911  -0.2732  -0.1507  -0.1138   3.4200
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.7331596  0.1639107 -16.675  < 2e-16 ***
## RS_cat2normal_alta  0.6873616  0.1532159   4.486 7.25e-06 ***
## Dir_Aire           -0.0087910  0.0009457  -9.296  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1828.0  on 7463  degrees of freedom
## Residual deviance: 1660.9  on 7461  degrees of freedom
```

```
## AIC: 1666.9
##
## Number of Fisher Scoring iterations: 7
```

Según el criterio de información de Akaike, esté modelo ajusta mejor los datos. Pasa de un valor AIC: 1775.9 a AIC: 1666.9.

Predicción

Según el modelo del apartado d), calculad la probabilidad de que la concentración de O3 sea o no superior a 80, con unos valores de RS_cat2="Normal_alta"y Dir_Aire=40.

```
pre_data_d<-data.frame(RS_cat2='normal_alta', Dir_Aire=c(40))
predict(model_icO3_RS_cat2_Dir_Aire,pre_data_d ,type="response")
```

```
##          1
## 0.08336825
```

Tendría una probabilidad de 0.083 de O3 ser Mejorable y 1-0.083 de ser Bueno.

Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste, tomando el modelo del apartado d). En la librería ResourceSelection hay una función que ajusta el test de Hosmer-Lemeshow.

```
hoslem.test(dat_Air$icO3,fitted(model_icO3_RS_cat2_Dir_Aire))
```

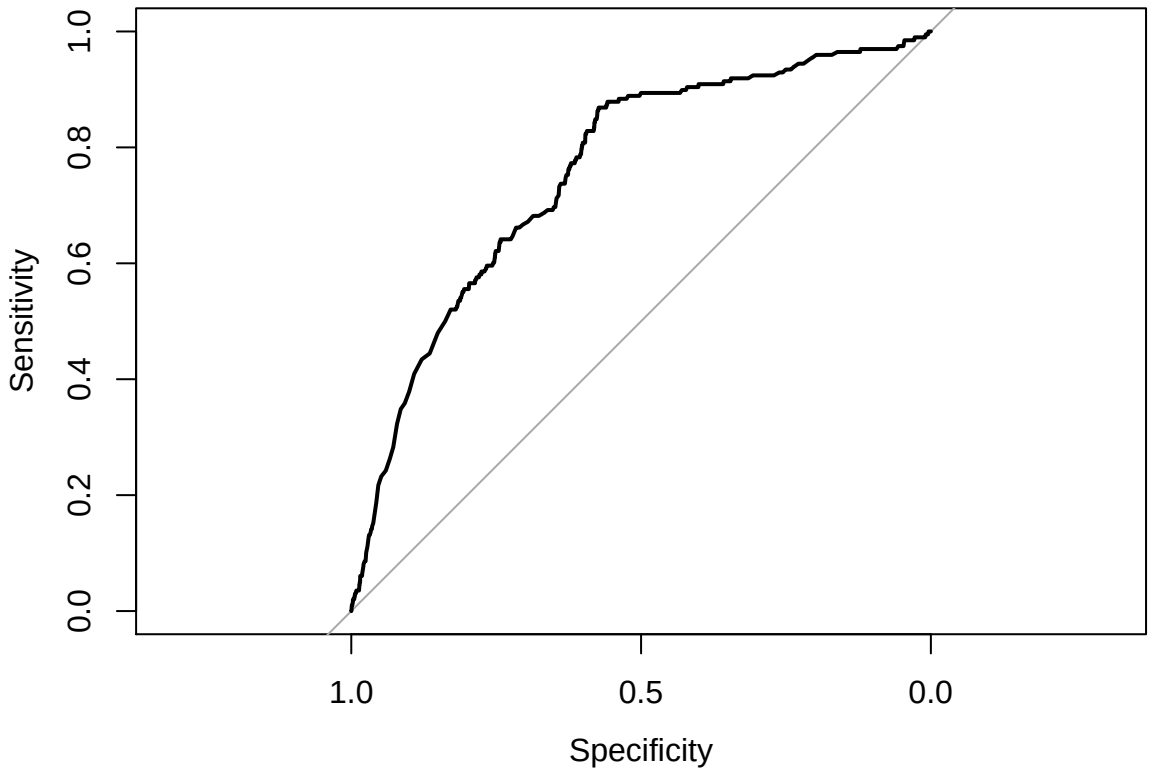
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  dat_Air$icO3, fitted(model_icO3_RS_cat2_Dir_Aire)
## X-squared = 7464, df = 8, p-value < 2.2e-16
```

Rechazamos la hipótesis nula (H0). El modelo no ajusta bien los datos

Curva ROC

Dibujar la curva ROC, y calcular el área debajo de la curva con el modelo del apartado d). Discutir el resultado.

```
prob=predict(model_icO3_RS_cat2_Dir_Aire, dat_Air, type="response")
r=roc(dat_Air$icO3, prob, data=dat_Air)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
plot (r)
```




```
auc(r)
```

```
## Area under the curve: 0.7572
```

Podemos decir que el modelo discrimina de manera adecuada ya que auc se encuentra entre 0.6 y 0.8.

Conclusiones del análisis

En este apartado se deberán exponer las conclusiones en base a los resultados obtenidos en todo el estudio.

Regresión lineal simple *El contaminantes atmosférico 'Concentración de Ozono' tiene una relación lineal fuerte con la variable 'Radiación Solar' siendo esta relación positiva. Pero utilizando la media diaria de las medidas, el valor con más coorelación varía con respecto a los datos por hora. La relación lineal más fuerte con la variable 'Radiación Solar' coresponde a variable de 'Concentración de (dióxido de nitrógeno)'.*

La explicación del nivel de ozono en función de la radiación solar genera un modelo lineal pobre en el que solo 13.66% de la variacion queda explicada por O3. Además, el estudio de la calidad del aire demuestra que, la mejora del Indice de calidad del Aire aumenta con una buena medida de partículas en suspension. Esto se ve claramente cuando las partículas en suspension tiene la eqiqueta de Muy buena. Sucede el efecto contrario para los escenarios de partículas en suspension inferiores, por el siguiente orden; Mejorable, Mala y Muy mala.

Regresión lineal multiple: *Observamos una mejora significativa del modelo anterior al añadir nuevas variables como son :Dir_Aire y NO2. Pasando del R2 13.66% al 57.88%. Se descarta la variable Tmp ya que no hace mejorar el modelo original.*

Regresion logistica: *2.1 La ocurrencia de que la calidad del aire empeora es de 2.98 veces mayor en relación cuando RS_cat es normal_baja, por lo que consideramos RS de factor de riesgo.*

2.2 Observamos la Tmp como una variable de confusion ya esta relacionada con la variable dependiente y la independiente.

2.3 No observamos una interacción entre RS_cat2 y HR.

2.4 Vemos una merora del modelo añdiendo las variables explicativas RS_cat2 y Dir_Aire. ya que según el criterio de información de Akaike, esté modelo ajusta mejor los datos. Pasa de un valor AIC: 1775.9 a AIC: 1666.9.
