Actividad 3: Modelización predictiva

Solución

Semestre 2021.2

Índice

En esta actividad se usará el fichero de datos (dat_Air_Stations) que contiene información de diferentes parámetros sobre la calidad del aire de una determinada ciudad del Norte de España en el año 2018. Los datos nos muestran concentraciones por hora de varios contaminantes atmosféricos (gases y partículas) como SO2, NO2, O3 y PM10, entre otros, monitoreados en cinco estaciones. Por otro lado en dos de las cinco estaciones, se han recogido medidas de variables meteorológicas. El periodo que abarca este estudio está comprendido entre el 1 de Enero 2018 al 31 de Diciembre de 2018. Estos datos han sido medidos en tiempo real.

La contaminación del aire representa un importante riesgo medioambiental y para la salud, tanto en los países desarrollados como en los países en desarrollo, por lo que su estudio es muy necesario.

Las variables del fichero de datos son:

- Estación: Estación móvil.
- Nombre: Nombre de la 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 azufre) en μ_g/m^3.
- NO: Concentración de NO (óxido nítrico) en μ_g/m^3.
- NO2: Concentración de (dióxido de nitrógeno) en μ_g/m^3.
- CO: Concentración de CO en μ_g/m^3.
- O3: Concentración de Ozono en μ_g/m^3.
- PM10: Partículas en suspensión <10 en μ_g/m^3.
- PM25: Partículas en Suspensión PM 2,5 en μ_g/m^3.
- BEN: Concentración de benceno en μ_g/m^3.
- TOL: Tolueno en μ_g/m^3.
- MXIL: MiXileno en μ_g/m^3.
- dd: Dirección del viento en grados.
- vv: 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 en W/m^2.
- LL: Precipitación en l/m^2.

---

1 Regresión

La exposición a la materia particulada (PM10), al ozono (O3), al dióxido de nitrógeno (NO2) y el dióxido de azufre (SO2), plantean graves riesgos para la salud. Las directrices de la OMS sobre la calidad del aire establecen los límites sobre estos principales contaminantes atmosféricos.

PM10: Límite de 45 microgramos de partículas por cada metro cúbico μ_g/m^3. SO2: Límite de 40 μ_g/m^3. NO2: Límite de 25 μ_g/m^3. O3: Límite de 60 μ_g/m^3.

El índice de calidad del aire se calcula de forma individual teniendo en cuenta cada uno de dichos contaminantes. Todos estos valores están referidos a la **media diaria**. Con referencia a **valores máximos diarios** se tomarán los valores de 100 μ_g/m^3 para O3 y de 120 μ_g/m^3 para NO2. Tanto para PM10 y SO2, se tomarán como referencia únicamente los valores medios diarios para comparar.

1.1 Estudio comparativo entre estaciones.

a) Estudio de los valores medios y máximos diarios de cada contaminante. Para cada una de las estaciones de monitoreo, se calcularán los valores máximos y medios diarios de cada contaminante. Posteriormente se hará una comparativa entre las cinco estaciones en base a dichos valores. Interpretad teniendo en cuenta los límites mencionados anteriormente.

b) Representad gráficamente la evolución de cada uno de los contaminantes en cada estación. Se tomarán los valores máximos diarios.

Nota: El motivo de estos primeros apartados es tomar un primer contacto sobre las posibles diferencias entre estaciones, así como hacerse una idea de las relaciones existentes entre las variables, pero para construir los modelos de regresión se tomarán los datos por hora.

```
dat_max <- ddply(dat, c("Fecha", "Nombre"), summarise,N= length(O3),O3_max = max(O3),
                 NO2_max = max(NO2), SO2_max=max(SO2),PM10_max = max(PM10))
table.max<-by (dat_max, dat_max$Nombre, summary )
table.max
```

```
## dat_max$Nombre: Estacion Avenida Argentina
##     Fecha              Nombre                N          O3_max
##  Length:365         Length:365         Min.   :24   Min.   : 19.00
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 63.00
##  Mode  :character   Mode  :character   Median :24   Median : 72.00
##                                        Mean   :24   Mean   : 72.12
##                                        3rd Qu.:24   3rd Qu.: 81.00
##                                        Max.   :24   Max.   :122.00
##                                                     NA's   :32
##     NO2_max          SO2_max         PM10_max
##  Min.   : 14.00   Min.   : 2.00   Min.   : 21.20
##  1st Qu.: 35.00   1st Qu.: 5.00   1st Qu.: 41.40
##  Median : 46.00   Median : 8.00   Median : 54.20
##  Mean   : 49.05   Mean   :13.62   Mean   : 58.66
##  3rd Qu.: 61.00   3rd Qu.:19.00   3rd Qu.: 70.10
##  Max.   :128.00   Max.   :90.00   Max.   :201.00
##  NA's   :27       NA's   :31      NA's   :23
## ------------------------------------------------------------
## dat_max$Nombre: Estacion Avenida Castilla
##     Fecha              Nombre                N          O3_max
##  Length:365         Length:365         Min.   :24   Min.   :  7.00
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 65.00
##  Mode  :character   Mode  :character   Median :24   Median : 80.00
##                                        Mean   :24   Mean   : 77.05
##                                        3rd Qu.:24   3rd Qu.: 92.00
##                                        Max.   :24   Max.   :130.00
##                                                     NA's   :28
##     NO2_max          SO2_max         PM10_max
##  Min.   :  7.00   Min.   : 4.00   Min.   : 11.70
##  1st Qu.: 27.25   1st Qu.: 8.00   1st Qu.: 31.90
##  Median : 39.00   Median :10.00   Median : 43.50
##  Mean   : 39.17   Mean   :11.06   Mean   : 57.81
##  3rd Qu.: 49.00   3rd Qu.:12.50   3rd Qu.: 61.60
##  Max.   :101.00   Max.   :50.00   Max.   :395.00
##  NA's   :27       NA's   :30      NA's   :18
```

```
## ------------------------------------------------------------
## dat_max$Nombre: Estacion Avenida Constitucion
##     Fecha              Nombre                N             O3_max
##  Length:365         Length:365         Min.   :22.00   Min.   : 19.00
##  Class :character   Class :character   1st Qu.:24.00   1st Qu.: 62.00
##  Mode  :character   Mode  :character   Median :24.00   Median : 75.00
##                                        Mean   :23.99   Mean   : 73.34
##                                        3rd Qu.:24.00   3rd Qu.: 86.00
##                                        Max.   :24.00   Max.   :129.00
##                                                        NA's   :40
##     NO2_max          SO2_max          PM10_max
##  Min.   :  7.00   Min.   : 1.000   Min.   : 11.70
##  1st Qu.: 38.00   1st Qu.: 5.000   1st Qu.: 26.60
##  Median : 53.00   Median : 8.000   Median : 35.00
##  Mean   : 56.50   Mean   : 9.506   Mean   : 38.02
##  3rd Qu.: 73.75   3rd Qu.:12.000   3rd Qu.: 44.60
##  Max.   :130.00   Max.   :60.000   Max.   :116.00
##  NA's   :31       NA's   :37       NA's   :22
## ------------------------------------------------------------
## dat_max$Nombre: Estacion Avenida Hermanos Felgueroso
##     Fecha              Nombre                N          O3_max
##  Length:365         Length:365         Min.   :23   Min.   : 14.00
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 58.00
##  Mode  :character   Mode  :character   Median :24   Median : 72.00
##                                        Mean   :24   Mean   : 70.63
##                                        3rd Qu.:24   3rd Qu.: 84.00
##                                        Max.   :24   Max.   :129.00
##                                                     NA's   :33
##     NO2_max          SO2_max          PM10_max
##  Min.   :  5.00   Min.   : 2.000   Min.   : 14.90
##  1st Qu.: 41.00   1st Qu.: 6.000   1st Qu.: 34.00
##  Median : 52.00   Median : 9.000   Median : 44.60
##  Mean   : 53.33   Mean   : 9.994   Mean   : 52.36
##  3rd Qu.: 66.00   3rd Qu.:12.000   3rd Qu.: 58.40
##  Max.   :107.00   Max.   :58.000   Max.   :361.00
##  NA's   :27       NA's   :36       NA's   :19
## ------------------------------------------------------------
## dat_max$Nombre: Estacion de Montevil
##     Fecha              Nombre                N          O3_max
##  Length:365         Length:365         Min.   :24   Min.   : 18.00
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 67.00
##  Mode  :character   Mode  :character   Median :24   Median : 77.00
##                                        Mean   :24   Mean   : 77.67
##                                        3rd Qu.:24   3rd Qu.: 90.00
##                                        Max.   :24   Max.   :134.00
##                                                     NA's   :30
##     NO2_max          SO2_max          PM10_max
##  Min.   :10.00   Min.   : 3.00   Min.   : 14.90
##  1st Qu.:30.00   1st Qu.: 7.00   1st Qu.: 32.90
##  Median :40.00   Median :11.00   Median : 41.40
##  Mean   :44.54   Mean   :15.11   Mean   : 46.12
##  3rd Qu.:57.25   3rd Qu.:18.75   3rd Qu.: 54.20
##  Max.   :90.00   Max.   :71.00   Max.   :156.00
##  NA's   :29      NA's   :31      NA's   :12
```

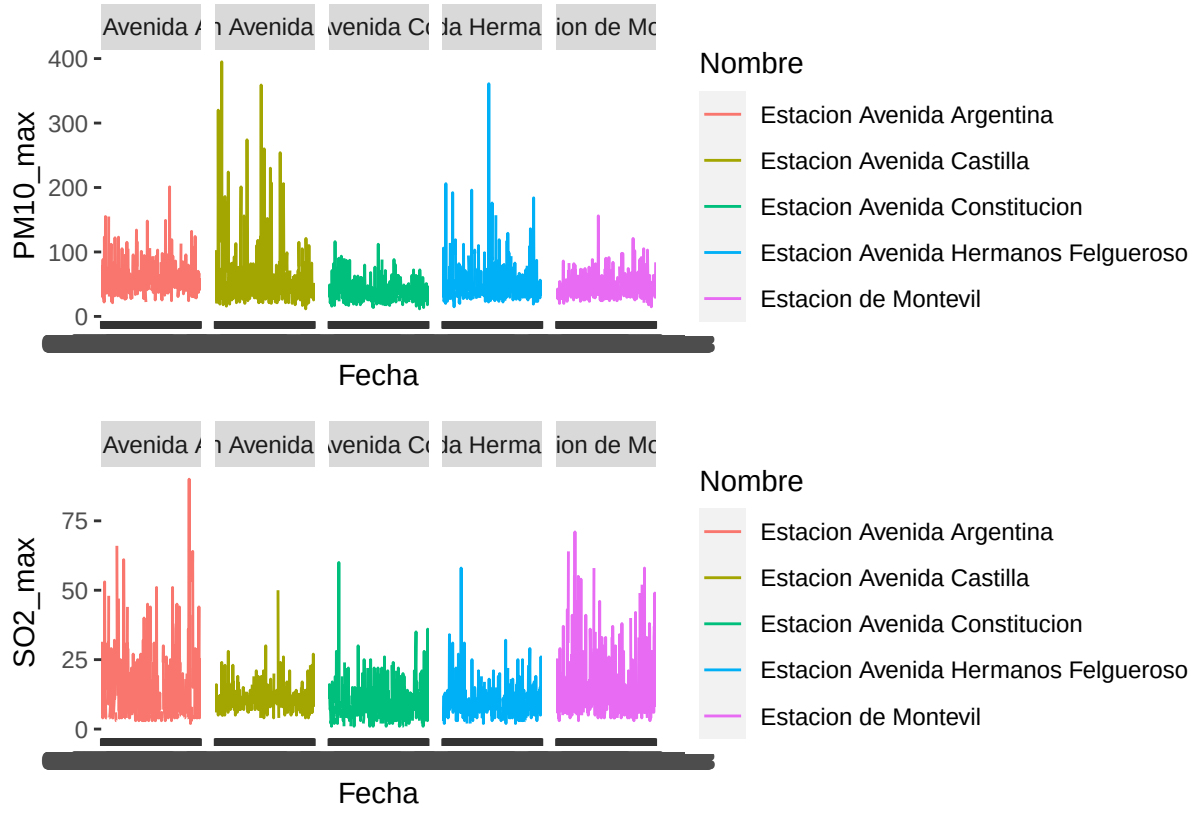
```r
dat_mean <- ddply(dat, c("Fecha", "Nombre"), summarise,N= length(O3),O3_m = mean(O3),
                  NO2_m = mean(NO2), SO2_m=mean(SO2),PM10_m = mean(PM10))
table.mean<-by (dat_mean, dat_mean$Nombre, summary )
table.mean
```

```
## dat_mean$Nombre: Estacion Avenida Argentina
##     Fecha              Nombre                N           O3_m
##  Length:365         Length:365         Min.   :24   Min.   :11.04
##  Class :character   Class :character   1st Qu.:24   1st Qu.:36.12
##  Mode  :character   Mode  :character   Median :24   Median :49.79
##                                        Mean   :24   Mean   :48.65
##                                        3rd Qu.:24   3rd Qu.:58.96
##                                        Max.   :24   Max.   :97.00
##                                                     NA's   :32
##      NO2_m            SO2_m            PM10_m
##  Min.   : 7.708   Min.   : 2.000   Min.   :11.27
##  1st Qu.:16.646   1st Qu.: 3.375   1st Qu.:22.89
##  Median :22.771   Median : 4.896   Median :28.97
##  Mean   :24.991   Mean   : 6.492   Mean   :30.60
##  3rd Qu.:31.594   3rd Qu.: 8.198   3rd Qu.:36.17
##  Max.   :63.542   Max.   :36.875   Max.   :75.07
##  NA's   :27       NA's   :31       NA's   :23
## ------------------------------------------------------------
## dat_mean$Nombre: Estacion Avenida Castilla
##     Fecha              Nombre                N           O3_m
##  Length:365         Length:365         Min.   :24   Min.   :  4.125
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 36.875
##  Mode  :character   Mode  :character   Median :24   Median : 51.667
##                                        Mean   :24   Mean   : 50.172
##                                        3rd Qu.:24   3rd Qu.: 65.875
##                                        Max.   :24   Max.   :101.750
##                                                     NA's   :28
##      NO2_m            SO2_m            PM10_m
##  Min.   : 3.667   Min.   : 2.500   Min.   :  6.725
##  1st Qu.:12.344   1st Qu.: 5.083   1st Qu.: 16.393
##  Median :16.917   Median : 6.625   Median : 21.867
##  Mean   :18.940   Mean   : 6.799   Mean   : 24.541
##  3rd Qu.:24.021   3rd Qu.: 8.292   3rd Qu.: 28.701
##  Max.   :47.625   Max.   :16.625   Max.   :132.812
##  NA's   :27       NA's   :30       NA's   :18
## ------------------------------------------------------------
## dat_mean$Nombre: Estacion Avenida Constitucion
##     Fecha              Nombre                N             O3_m
##  Length:365         Length:365         Min.   :22.00   Min.   : 4.042
##  Class :character   Class :character   1st Qu.:24.00   1st Qu.:31.000
##  Mode  :character   Mode  :character   Median :24.00   Median :43.917
##                                        Mean   :23.99   Mean   :43.405
##                                        3rd Qu.:24.00   3rd Qu.:55.875
##                                        Max.   :24.00   Max.   :91.333
##                                                        NA's   :40
##      NO2_m            SO2_m            PM10_m
##  Min.   : 2.167   Min.   : 1.000   Min.   : 5.709
##  1st Qu.:18.604   1st Qu.: 2.625   1st Qu.:15.199
```

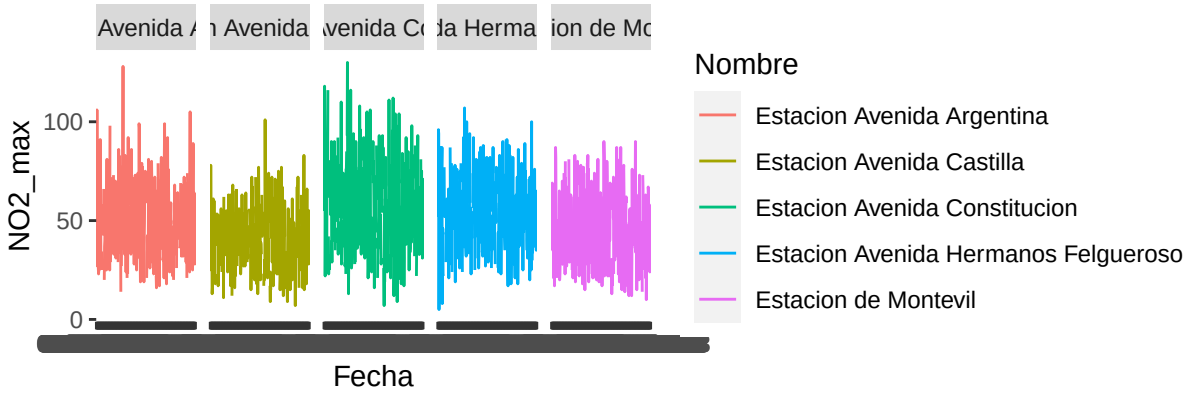
```
##  Median :27.500   Median : 3.958   Median :19.206
##  Mean   :28.684   Mean   : 4.293   Mean   :20.305
##  3rd Qu.:38.833   3rd Qu.: 5.552   3rd Qu.:24.785
##  Max.   :73.708   Max.   :22.208   Max.   :44.554
##  NA's   :31       NA's   :37       NA's   :22
## ------------------------------------------------------------
## dat_mean$Nombre: Estacion Avenida Hermanos Felgueroso
##     Fecha              Nombre                N           O3_m
##  Length:365         Length:365         Min.   :23   Min.   : 5.333
##  Class :character   Class :character   1st Qu.:24   1st Qu.:32.156
##  Mode  :character   Mode  :character   Median :24   Median :45.167
##                                        Mean   :24   Mean   :44.816
##                                        3rd Qu.:24   3rd Qu.:59.042
##                                        Max.   :24   Max.   :93.708
##                                                     NA's   :33
##      NO2_m            SO2_m            PM10_m
##  Min.   : 2.958   Min.   : 1.125   Min.   : 7.782
##  1st Qu.:20.740   1st Qu.: 2.667   1st Qu.:18.322
##  Median :26.917   Median : 3.708   Median :23.435
##  Mean   :28.864   Mean   : 4.171   Mean   :25.281
##  3rd Qu.:36.031   3rd Qu.: 4.708   3rd Qu.:30.701
##  Max.   :69.208   Max.   :15.208   Max.   :77.621
##  NA's   :27       NA's   :36       NA's   :19
## ------------------------------------------------------------
## dat_mean$Nombre: Estacion de Montevil
##     Fecha              Nombre                N           O3_m
##  Length:365         Length:365         Min.   :24   Min.   :  7.625
##  Class :character   Class :character   1st Qu.:24   1st Qu.: 38.271
##  Mode  :character   Mode  :character   Median :24   Median : 50.667
##                                        Mean   :24   Mean   : 50.100
##                                        3rd Qu.:24   3rd Qu.: 63.333
##                                        Max.   :24   Max.   :105.417
##                                                     NA's   :30
##      NO2_m            SO2_m            PM10_m
##  Min.   : 6.208   Min.   : 2.792   Min.   : 8.053
##  1st Qu.:13.573   1st Qu.: 4.500   1st Qu.:19.072
##  Median :18.312   Median : 5.625   Median :23.367
##  Mean   :20.072   Mean   : 6.514   Mean   :25.274
##  3rd Qu.:25.062   3rd Qu.: 7.167   3rd Qu.:30.275
##  Max.   :55.000   Max.   :23.750   Max.   :52.846
##  NA's   :29       NA's   :31       NA's   :12
```

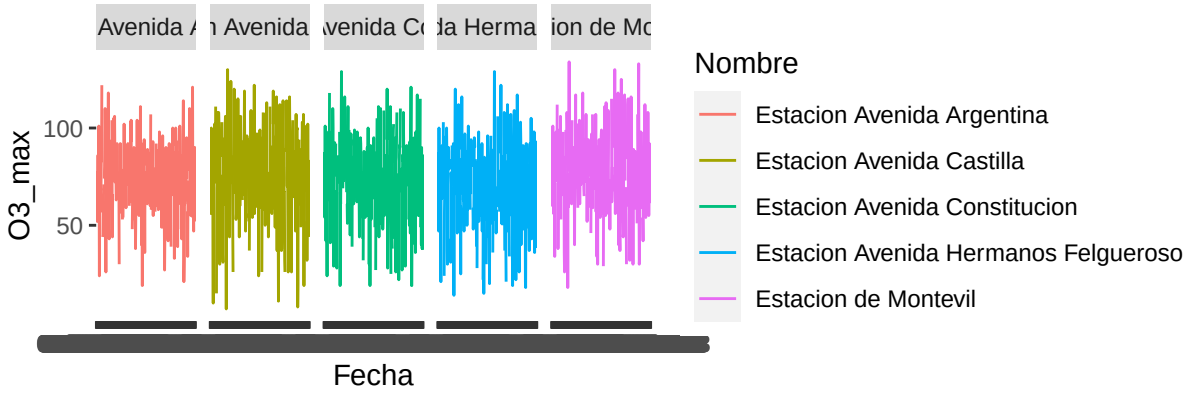
```
#b) Representad gráficamente la evolución
#Para PM10
g1<-ggplot(dat_max, aes(x = Fecha, y = PM10_max, group = Nombre, colour = Nombre)) +
  geom_line() +
  facet_grid(.~Nombre)
# Para SO2
g2<-ggplot(dat_max, aes(x = Fecha, y = SO2_max, group = Nombre, colour = Nombre)) +
  geom_line() +
  facet_grid(.~Nombre)
# Para NO2
g3<-ggplot(dat_max, aes(x = Fecha, y = NO2_max, group = Nombre, colour = Nombre)) +
  geom_line() +
```

```
  facet_grid(.~Nombre)
# Para O3
g4<-ggplot(dat_max, aes(x = Fecha, y = O3_max, group = Nombre, colour = Nombre)) +
  geom_line() +
  facet_grid(.~Nombre)
g1/g2
```




```
g3/g4
```





Las estaciones de Avenida Argentina y Avenida Constitución, superan el valor máximo 120 g/m3 de NO2. Con referencia a los valores medios, el valor límite de 25, ha sido superado también en todas las estaciones. En las que menos días se ha excedido este valor, han sido en Avenida Castilla y Montevil con el 25% de los días (Q3 próximo a 25).

Los valores máximos de O3, han sido superados en todas las estaciones, siendo el valor más alto registrado en la de Montevil. En las estación de avenida de la Argentina se observa una mayor concentración de dióxido de azufre. Esto puede ser debido a que está cerca de una zona más industrial y próxima al puerto. Con referencia a las partículas PM10 (menores de diez micras), el valor medio límite diario de 45, es superado en todas las estaciones, excepto en Avenida de Constitución, donde el valor más grande de la media sería 44,55.

c) Estudio de correlación lineal.Para ello se seleccionan las dos estaciones con registros meteorológicos: Estación de Montevil y Estación Avenida Constitución. Para cada una de las estaciones, calcular la matriz de correlación entre los contaminantes citados anteriormente y las variables meteorológicas: Temperatura (TMP), Humedad Relativa (HR), Radiación solar (RS), velocidad del viento (vv), precipitaciones (LL) y Presión barométrica (PRB).Interpretad.

Nota: La matriz de correlación será calculada en base a los valores máximos de cada contaminante.

```
#c) Estudio de correlación lineal
#Estación de Montevil
dat_MO<-dat[dat$Nombre=="Estacion de Montevil",]
dat_max_MO <- ddply(dat_MO, c("Fecha"), summarise,N= length(O3),O3_max = max(O3),
                    NO2_max = max(NO2), SO2_max=max(SO2),PM10_max = max(PM10),
                    TMP_max= max(TMP), vv_max= max(vv), LL_max= max(LL),
                    HR_max= max(HR), RS_max=max(RS), PRB_max = max(PRB))
```

```
var.cor_max<- select(dat_max_MO,O3_max,NO2_max,SO2_max,PM10_max, TMP_max, vv_max,
                     LL_max,HR_max,RS_max,PRB_max)
cor(var.cor_max, method = "pearson", use="pairwise.complete.obs")
```

```
##               O3_max       NO2_max     SO2_max    PM10_max     TMP_max
## O3_max    1.00000000 -0.1594628193 -0.04046466 -0.10297347  0.12467359
## NO2_max  -0.15946282  1.0000000000  0.06492725  0.39566839 -0.24309536
## SO2_max  -0.04046466  0.0649272480  1.00000000  0.32952123  0.08324537
## PM10_max -0.10297347  0.3956683929  0.32952123  1.00000000  0.18695007
## TMP_max   0.12467359 -0.2430953575  0.08324537  0.18695007  1.00000000
## vv_max    0.23459075 -0.0590695216  0.05830618 -0.20071217 -0.21760757
## LL_max    0.06684187 -0.1954803228  0.04015035 -0.21305923 -0.16079166
## HR_max    0.04064338 -0.2625694698  0.03920936 -0.03783367  0.23020029
## RS_max    0.42391357  0.0464474735  0.01217166 -0.13848838 -0.26946880
## PRB_max  -0.24775837 -0.0006537712  0.04824672  0.18919651  0.08875524
##               vv_max      LL_max      HR_max      RS_max       PRB_max
## O3_max    0.23459075  0.06684187  0.04064338  0.42391357 -0.2477583700
## NO2_max  -0.05906952 -0.19548032 -0.26256947  0.04644747 -0.0006537712
## SO2_max   0.05830618  0.04015035  0.03920936  0.01217166  0.0482467187
## PM10_max -0.20071217 -0.21305923 -0.03783367 -0.13848838  0.1891965113
## TMP_max  -0.21760757 -0.16079166  0.23020029 -0.26946880  0.0887552369
## vv_max    1.00000000  0.07426965 -0.28229174  0.21932735 -0.2241969784
## LL_max    0.07426965  1.00000000  0.14908099 -0.02377485 -0.1757608827
## HR_max   -0.28229174  0.14908099  1.00000000 -0.11705944  0.1223903292
## RS_max    0.21932735 -0.02377485 -0.11705944  1.00000000 -0.1928144956
## PRB_max  -0.22419698 -0.17576088  0.12239033 -0.19281450  1.0000000000
```

```
# Estación de AC
dat_AC<-dat[dat$Nombre=="Estacion Avenida Constitucion",]
dat_max_AC <- ddply(dat_AC, c("Fecha"), summarise,N= length(O3),O3_max = max(O3),
                    NO2_max = max(NO2), SO2_max=max(SO2),PM10_max = max(PM10),
                    TMP_max= max(TMP), vv_max= max(vv), LL_max= max(LL),
                    HR_max= max(HR), RS_max=max(RS), PRB_max = max(PRB))
var.cor_max<- select(dat_max_AC,O3_max,NO2_max,SO2_max,PM10_max, TMP_max, vv_max,
                     LL_max,HR_max,RS_max,PRB_max)
cor(var.cor_max, method = "pearson", use="pairwise.complete.obs")
```

```
##               O3_max     NO2_max     SO2_max    PM10_max     TMP_max
## O3_max    1.00000000 -0.22381545 -0.17646292 -0.10275523  0.07620192
## NO2_max  -0.22381545  1.00000000  0.36814418  0.42024656 -0.37020587
## SO2_max  -0.17646292  0.36814418  1.00000000  0.42249953 -0.08530356
## PM10_max -0.10275523  0.42024656  0.42249953  1.00000000  0.05207180
## TMP_max   0.07620192 -0.37020587 -0.08530356  0.05207180  1.00000000
## vv_max    0.54534903 -0.13698660 -0.19763783 -0.24183391  0.07760621
## LL_max    0.01944144 -0.08250997 -0.07773028 -0.16607291 -0.15349870
## HR_max   -0.03918082 -0.34950296 -0.19038255 -0.01778824  0.26148533
## RS_max    0.42272015 -0.35233802 -0.18617885 -0.10138309  0.47274326
## PRB_max  -0.24884147 -0.12250536  0.02041717  0.08340804  0.07877364
##               vv_max      LL_max      HR_max      RS_max     PRB_max
## O3_max    0.54534903  0.01944144 -0.03918082  0.42272015 -0.24884147
## NO2_max  -0.13698660 -0.08250997 -0.34950296 -0.35233802 -0.12250536
## SO2_max  -0.19763783 -0.07773028 -0.19038255 -0.18617885  0.02041717
```

```
## PM10_max -0.24183391 -0.16607291 -0.01778824 -0.10138309  0.08340804
## TMP_max   0.07760621 -0.15349870  0.26148533  0.47274326  0.07877364
## vv_max    1.00000000 -0.04076384 -0.24092288  0.34906997 -0.18847643
## LL_max   -0.04076384  1.00000000  0.27328832 -0.16698414 -0.20519659
## HR_max   -0.24092288  0.27328832  1.00000000 -0.04374517  0.11851418
## RS_max    0.34906997 -0.16698414 -0.04374517  1.00000000 -0.04115076
## PRB_max  -0.18847643 -0.20519659  0.11851418 -0.04115076  1.00000000
```

Según el coeficiente de correlación, se puede apreciar en ambas estaciones que la concentración de O3 tiene relación lineal positiva con la RS y vv, así como relación inversa con PRB. En cambio si nos fijamos en el contaminante PM10, se observa relación lineal negativa con RS y vv, y positiva con PRB.

Por otro lado se tiene una correlación alta entre los contaminantes PM10 con NO2 y SO2, con un coeficiente de correlación aproximadamente del 40% para cada uno de ellos, en ambas estaciones. Esto es debido a que parte de los componentes del material particulado (PM10) se originan por la oxidación en la atmósfera de SO2 y NO2. Con relación al O3, se observa una relación lineal negativa con NO2, por lo que la disminución de NO2 en la atmósfera, favorece el aumento de concentrasción de O3.

1.2 Modelo de regresión lineal

Como he mencionado arriba, para construir los modelos de regresión, se tomarán los valores de las variables escogidas por hora, tal como aparecen en la base de datos original.

a) Se pide crear un modelo de regresión lineal, tomando como variable dependiente (O3) y variable explicativa (NO2). Se evaluará la bondad del ajuste, a partir del coeficiente de determinación. Interpretad.

b) Se añade al modelo anterior el nombre de las estaciones (Nombre). Interpretad.

```
# a) Estimacion del modelo (todas las estaciones)
Model_1<- lm(O3~NO2, data=dat)
summary(Model_1)
```

```
##
## Call:
## lm(formula = O3 ~ NO2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.354 -13.911   0.825  13.868  85.680
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.648967   0.174379   416.6   <2e-16 ***
## NO2         -1.036920   0.005861  -176.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 43116 degrees of freedom
##   (679 observations deleted due to missingness)
## Multiple R-squared:  0.4206, Adjusted R-squared:  0.4206
## F-statistic: 3.13e+04 on 1 and 43116 DF,  p-value: < 2.2e-16
```

```r
# b) Se añade al modelo anterior el nombre de las estaciones (Nombre).
Model_1<- lm(O3~NO2+Nombre, data=dat)
summary(Model_1)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + Nombre, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.683 -13.874   0.685  13.888  85.148
##
## Coefficients:
##                                           Estimate Std. Error  t value
## (Intercept)                              75.304590   0.266605  282.458
## NO2                                      -1.065155   0.006029 -176.686
## NombreEstacion Avenida Castilla          -5.099968   0.312237  -16.334
## NombreEstacion Avenida Constitucion      -1.106588   0.312837   -3.537
## NombreEstacion Avenida Hermanos Felgueroso  0.231214   0.311961    0.741
## NombreEstacion de Montevil               -3.788570   0.311478  -12.163
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## NO2                                       < 2e-16 ***
## NombreEstacion Avenida Castilla           < 2e-16 ***
## NombreEstacion Avenida Constitucion      0.000405 ***
## NombreEstacion Avenida Hermanos Felgueroso 0.458598
## NombreEstacion de Montevil                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.42 on 43112 degrees of freedom
##   (679 observations deleted due to missingness)
## Multiple R-squared:  0.4266, Adjusted R-squared:  0.4265
## F-statistic:  6414 on 5 and 43112 DF,  p-value: < 2.2e-16
```

a) Se observa que la variable NO2 es significativa con un p_valor de 2.2e-16 existiendo una relación líneal negativa entre ambas variables, con un coeficiente de determinación ajustado de 0.4206.

b) Si se toma como referencia la estación de Avenida Argentina, se observa que todas las estaciones son significativas, excepto en Hermanos Felgueroso. Esto indica que el hecho de estar en una u otra zona geográfica es relevante para la concentración de O3. Nota: En este modelo los efectos de dicha variable (Nombre) están ajustados por la variable explicativa NO2.

1.3 Modelo de regresión lineal múltiple

Se quiere construir un modelo de regresión múltiple con el que podamos predecir la concentración de ozono (O3) en las zonas de Montevil y Avenida de la Constitución.

a) Se pide dos modelos (uno para cada estación) tomando como variable dependiente el nivel de ozono (O3) en función de la concentración de dióxido de nitrógeno (NO2) y diferentes variables meteorológicas como vv (velocidad del viento), RS (radiación solar), HR (humedad relativa) y LL (precipitaciones).

b) Se añade a los modelos anteriores la variable Temperatura (TMP). De ser necesario, se pide comprobar la presencia o no de colinealidad entre las variables (vv) y (TMP). Podéis usar la librería (faraway) y estudiar el FIV (factor de inflación de la varianza). Discutid si sería indicado o no añadir la variable (TMP) a cada uno de los modelos.

```
#Estimacion del modelo Estación MO
Model_1MO<- lm(O3~NO2+RS+vv+HR+LL, data=dat_MO)
summary(Model_1MO)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + RS + vv + HR + LL, data = dat_MO)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.86 -12.66  -1.02  10.93  70.68
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.083516   1.446493  60.203   <2e-16 ***
## NO2         -1.082711   0.013968 -77.512   <2e-16 ***
## RS           0.031481   0.001496  21.042   <2e-16 ***
## vv           4.743372   0.198630  23.880   <2e-16 ***
## HR          -0.293340   0.014337 -20.461   <2e-16 ***
## LL           2.586381   0.305459   8.467   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.47 on 8661 degrees of freedom
##   (93 observations deleted due to missingness)
## Multiple R-squared:  0.6109, Adjusted R-squared:  0.6107
## F-statistic:  2720 on 5 and 8661 DF,  p-value: < 2.2e-16
```

```
#Estimacion del modelo Estación AC
Model_1AC<- lm(O3~NO2+RS+vv+HR+LL, data=dat_AC)
summary(Model_1AC)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + RS + vv + HR + LL, data = dat_AC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.551 -13.010  -0.295  11.324  64.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.400738   1.505634  55.392  < 2e-16 ***
## NO2         -0.770572   0.011096 -69.444  < 2e-16 ***
## RS           0.007818   0.001378   5.675 1.43e-08 ***
## vv          15.134840   0.477808  31.676  < 2e-16 ***
## HR          -0.355753   0.016755 -21.233  < 2e-16 ***
## LL           2.857869   0.352117   8.116 5.50e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 8274 degrees of freedom
##   (478 observations deleted due to missingness)
## Multiple R-squared:  0.5907, Adjusted R-squared:  0.5905
## F-statistic:  2389 on 5 and 8274 DF,  p-value: < 2.2e-16
```

```
# b) Se añade a los modelos anteriores la variable Temperatura (TMP)
Model_2MO<- lm(O3~NO2+RS+vv+HR+LL+TMP, data=dat_MO)
summary(Model_2MO)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + RS + vv + HR + LL + TMP, data = dat_MO)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.863 -12.705  -1.027  11.040  69.931
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.498647   1.691628  48.178  < 2e-16 ***
## NO2         -1.057347   0.014502 -72.912  < 2e-16 ***
## RS           0.030898   0.001496  20.660  < 2e-16 ***
## vv           4.900684   0.199737  24.536  < 2e-16 ***
## HR          -0.279089   0.014481 -19.273  < 2e-16 ***
## LL           2.747752   0.305837   8.984  < 2e-16 ***
## TMP          0.237141   0.037469   6.329 2.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.43 on 8660 degrees of freedom
##   (93 observations deleted due to missingness)
## Multiple R-squared:  0.6127, Adjusted R-squared:  0.6124
## F-statistic:  2283 on 6 and 8660 DF,  p-value: < 2.2e-16
```

```
Model_2AC<- lm(O3~NO2+RS+vv+HR+LL+TMP, data=dat_AC)
summary(Model_2AC)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + RS + vv + HR + LL + TMP, data = dat_AC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.543 -13.004  -0.297  11.321  64.721
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.423853   1.598385  52.193  < 2e-16 ***
## NO2         -0.770692   0.011442 -67.355  < 2e-16 ***
## RS           0.007840   0.001467   5.343 9.40e-08 ***
```

```
## vv          15.133798   0.478448  31.631  < 2e-16 ***
## HR          -0.355697   0.016806 -21.165  < 2e-16 ***
## LL           2.856342   0.353917   8.071 7.97e-16 ***
## TMP         -0.001871   0.043422  -0.043    0.966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 8273 degrees of freedom
##   (478 observations deleted due to missingness)
## Multiple R-squared:  0.5907, Adjusted R-squared:  0.5904
## F-statistic:  1990 on 6 and 8273 DF,  p-value: < 2.2e-16
```

a) Se observa que en ambas estaciones todas las varibles incluídas en el modelo son significativas con un coeficiente de determinación ajustado de 0.6107, en la estación de Montevil y de 0.5905 en Avenida Constitución. Además todas las variables explicativas tienen una relación lineal positiva con O3, excepto NO2 y HR.

b) Se añade la variable TMP, y se observa que existen diferencias según sea la estación de monitoreo escogida. Se tiene que la variable TMP es significativa en el primer modelo, cuando se toman los datos de Montevil y no en el segundo, por lo que no sería adecuado añadir esta variable explicativa al segundo modelo.

Se comprobará la presencia o no de colinealidad sólo para la estción de Montevil.

```
# Estación de Montevil:
cor(x = dat_MO$vv, y = dat_MO$TMP, method = "pearson", use="pairwise.complete.obs")
```

```
## [1] 0.05367551
```

```
#Veamos cómo difieren las estimaciones del modelo global con 'vv' y 'TMP', de los
#modelos de regresión lineal simple que podemos construir con cada una de
#las variables explicativas:

Model.g<-lm(O3~vv+TMP, data=dat_MO)
model.vv <- lm(O3~vv, data=dat_MO )
model.TMP <- lm(O3~TMP, data=dat_MO )
summary(Model.g)
```

```
##
## Call:
## lm(formula = O3 ~ vv + TMP, data = dat_MO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.764  -18.852   -1.296   15.850   80.885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.53659    0.80363   18.09   <2e-16 ***
## vv          12.59810    0.21488   58.63   <2e-16 ***
## TMP          1.14298    0.04637   24.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 22.96 on 8664 degrees of freedom
##   (93 observations deleted due to missingness)
## Multiple R-squared:  0.3277, Adjusted R-squared:  0.3276
## F-statistic:  2112 on 2 and 8664 DF,  p-value: < 2.2e-16
```

```
summary(model.vv)
```

```
##
## Call:
## lm(formula = O3 ~ vv, data = dat_MO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.621  -18.855   -0.872   16.254   84.998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.8763     0.4019   79.31   <2e-16 ***
## vv           12.9004     0.2219   58.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.75 on 8665 degrees of freedom
##   (93 observations deleted due to missingness)
## Multiple R-squared:  0.2806, Adjusted R-squared:  0.2805
## F-statistic:  3380 on 1 and 8665 DF,  p-value: < 2.2e-16
```

```
summary(model.TMP)
```

```
##
## Call:
## lm(formula = O3 ~ TMP, data = dat_MO)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.891 -23.081   1.024  21.496  78.446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.75952    0.89877   33.11   <2e-16 ***
## TMP          1.29815    0.05471   23.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.13 on 8665 degrees of freedom
##   (93 observations deleted due to missingness)
## Multiple R-squared:  0.06101,    Adjusted R-squared:  0.06091
## F-statistic:   563 on 1 and 8665 DF,  p-value: < 2.2e-16
```

```
# Cáculo de FIV
vif(Model.g)
```

```
##       vv      TMP
## 1.003268 1.003268
```

```
# Se compara con 1/(1-R2)
1/(1-summary(Model.g)$r.squared)
```

```
## [1] 1.487505
```

Por un lado el coeficiente de correlación entre ambas variables es de 0.053, por lo que no existe prácticamente relación lineal entre ambas variables, esto indicaría la no presencia de colinealidad. Comprobación: Primero se compara el modelo globlal, con cada uno de los modelos simples. Los coeficientes estimados para Tmp y vv no difieren mucho de los estimados en el modelo de la regresión múltiple. De forma numérica: Se procederá a detectar posibles efectos de multicolinealidad. Puesto que uno de los efectos principales de la multicolinealidad es la inflación de la varianza y covarianza de las estimaciones, se calculará el FIV(factor de inflación de la varianza).

El FIV = 1,003 resulta menor que su equivalente en el modelo global, 1/(1−R2) = 1,487. Ademá el valor de FIV, es muy bajo. A la vista de los últimos resultados, no se encuentra indicios de multicolinealidad entre los regresores 'vv' y 'Tmp', respecto a los los criterios de diagnóstico propuestos.

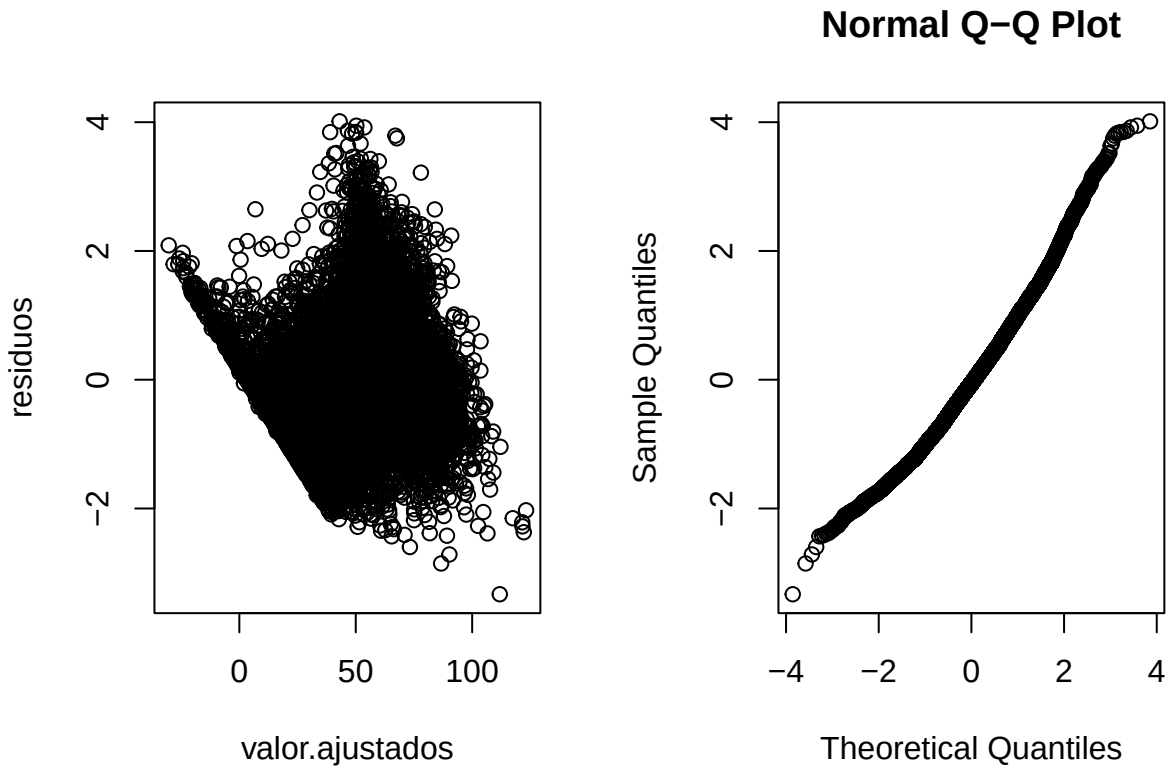
Si, sería adecuado incluir la TMP, como variable explicativa.

NOTA: Generalmente, valores de un FIV superiores a 10 dan indicios de un problema de multicolinealidad, si bien su magnitud depende del modelo ajustado.(Otros autores consideran valores por encima de 4).Es mejor compararlo con su equivalente en el modelo ajustado, esto es, 1/(1 − R2), donde R2 es el coeficiente de determinación del modelo. Los valores FIV mayores que esta cantidad implican que la relación entre las variables explicativas es mayor que la que existe entre la respuesta y los predictores, y por tanto dan indicios de multicolinealidad.

1.4 Diagnosis del modelo.

Para la diagnosis se escoge el último modelo construído para la estación de Montevil 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 cuantil-cuantil que compara los residuos del modelo con los valores de una variable que se distribuye normalmente(QQ plot). Interpretad los resultados.

```
residuos <- rstandard(Model_2MO)
valor.ajustados <- fitted(Model_2MO)
par(mfrow=c(1,2))
plot(valor.ajustados, residuos)
qqnorm(residuos)
```



A la vista del gráfico se observa un patrón de dispersión irregular. Es decir no es un patrón aleatorio de los residuos. Esto indica que no se cumple el supuesto de varianza constante en los errores del modelo. Por otro lado el QQ plot, muestra que los datos se ajustan bien a una normal.

1.5 Predicción del modelo

Según el modelo del apartado anterior, calculad la concentración de O3, si se tienen valores de NO2 de 40, vv de 2, RS de 100, HR de 80, LL de 0.10 y TMP de 25.

```
newdata = data.frame(NO2= 40, RS = 100, vv= 2 , HR= 80, LL= 0.10, TMP= 25)
predict(Model_2MO, newdata)
```

```
##        1
## 35.97204
```

Se obtiene un valor de 35.97

2 Regresión logística.

Para construir las nuevas variables y los modelos de regresión logística, se tomarán los valores de las variables escogidas por hora, tal como aparecen en la base de datos original.

En este apartado se tomarán como contaminantes la concentración de PM10 y de O3. Se procederá a calcular los índices de calidad (icPM10 e icO3) de la forma siguiente:

PM10 recodificada: (**icPM10**)

aceptable: valores de (0 a 45],
mejorable: valores de (45 a 180]

O3 recodificada: (**ic03**)

aceptable: valores de (0 a 60],
mejorable: valores de (60 a 170]

La variable RS también será recodificada:

RS recodificada (**RS_re**):

normal_baja:(0 a 100],
normal_alta: valores de (100 a 700]

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

2.1 Análisis crudo. Cáculo de OR

Se creará una nueva variable con los meses del año a partir de la variable Fecha, llamada **month**.

a) Se calculará las OR (Odds-Ratio) entre cada una de las variables dependientes **icPM10** y **icO3** y las variables explicativas radiación solar recodificada (RS_re) y (month) en la estación de Montevil. 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. ¿Se puede considerar que la radiación solar y el mes del año son factores de riesgo? Justifica tu respuesta e interpreta las OR.

b) Idem para la estación de Avenida Constitución.

```
dat_MO2<-dat_MO[,4:19]
noC<-c("CO","NO")
dat_MO2<-dat_MO2[ , !(names(dat_MO2) %in% noC)]
dat_MO2<-na.omit(dat_MO2)

# se crean las nuevas variables
dat_MO2[,"icO3"] <- cut(dat_MO2$O3, breaks = c(0,60,170),
                        labels = c("aceptable", "mejorable"))
dat_MO2[,"icPM10"] <- cut(dat_MO2$PM10,breaks = c(0,45,180),
                          labels = c("aceptable", "mejorable"))
dat_MO2[,"RS_re"] <- cut(dat_MO2$RS, breaks = c(0,100,700),
                         labels = c("normal_baja", "normal_alta"))
dat_MO2[,"icO3"]<-(dat_MO2$icO3=="mejorable")
dat_MO2[,"icO3"] <-ifelse(dat_MO2$icO3==TRUE, 1, 0)
dat_MO2$icO3 <- as.factor(dat_MO2$icO3)
dat_MO2[,"icPM10"]<-(dat_MO2$icPM10=="mejorable")
dat_MO2[,"icPM10"] <-ifelse(dat_MO2$icPM10==TRUE, 1, 0)
dat_MO2$icPM10 <- as.factor(dat_MO2$icPM10)

# Se toma la variable icO3 y se calcula la OR para RS_re
tab1 = table(dat_MO2$icO3,dat_MO2$RS_re)
chi.test<-chisq.test(tab1)
print(chi.test)
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab1
## X-squared = 828.18, df = 1, p-value < 2.2e-16
```

```
OR.RS <-(tab1[1]*tab1[4])/(tab1[2]*tab1[3])
OR.RS
```

```
## [1] 5.255848
```

```
# Se toma la variable icPM10 y se calcula la OR para RS_re
tab2= table(dat_MO2$icPM10, dat_MO2$RS_re)
chi.test<-chisq.test(tab2)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab2
## X-squared = 2.1259, df = 1, p-value = 0.1448
```

```
OR.Vel <-(tab2[1]*tab2[4])/(tab2[2]*tab2[3])
OR.Vel
```

```
## [1] 0.8389159
```

a) Con referencia al cálculo de las OR se tiene:

Estación de Montevil: Rs_re

Con referencia al icO3, la OR es de 5,25, por lo que nos indica que un lugar con RS más alta , tiene una probabilidad 5 veces mayor de que su concentración de O3 sea superior a 60. Por lo tanto, se puede considerar un factor de riesgo para el aumento de la concentración de O3. Con referencia al icPM10, la OR es de 0,84, por lo que nos indica que un lugar con RS más alta, sería un factor de protección para concentraciones altas de PM10.

Para calcular las OR de la variable mes (month) por el procedimiento anterior se deberían ir escogiendo subtablas de tamaño 2x2. Por ejemplo si se toma como referencia el mes de Julio, se calcularían las OR asociadas a:

Por ejemplo, para icPM10 y month:

1. icPM10 y Month: Enero y Julio
2. icPM10 y month: Febrero y Julio.

y así sucesivamente hasta completar todos los meses.

El procedimiento manual sería análogo al anterior. En este caso, para calcular las OR, sería más eficiente construir un modelo de regresión logística.

b) Idem Estación Avenida Constitución

```r
# se crean las nuevas variables
#(No sería necesario transformar la base de datos (dat_AC), ya que no se usará
#en los siguientes apartados.)
dat_AC[,"icO3_2"] <- cut(dat_AC$O3, breaks = c(0,60,170),
                         labels = c("aceptable", "mejorable"))
dat_AC[,"icPM10_2"] <- cut(dat_AC$PM10,breaks = c(0,45,180),
                           labels = c("aceptable", "mejorable"))
dat_AC[,"RS_re2"] <- cut(dat_AC$RS, breaks = c(0,100,700),
                         labels = c("normal_baja", "normal_alta"))

# Se toma la variable icO3_2 y se calcula la OR para RS_re2
tab1 = table(dat_AC$icO3_2,dat_AC$RS_re2)
chi.test<-chisq.test(tab1)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab1
## X-squared = 337.42, df = 1, p-value < 2.2e-16
```

```r
OR.RS <-(tab1[1]*tab1[4])/(tab1[2]*tab1[3])
OR.RS
```

```
## [1] 2.490348
```

```r
# Se toma la variable icPM10_2 y se calcula la OR para RS_re2
tab2= table(dat_AC$icPM10_2, dat_AC$RS_re2)
chi.test<-chisq.test(tab2)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab2
## X-squared = 3.092, df = 1, p-value = 0.07868
```

```r
OR.Vel <-(tab2[1]*tab2[4])/(tab2[2]*tab2[3])
OR.Vel
```

```
## [1] 0.7681564
```

Estación de Avenica Constitución: Rs_re2

Respecto al icO3_2, la OR es de 2,49 por lo que al igual que en la estación anterior,la RS más alta sería un factor de riesgo. La probabilidad de que la concentración de O3 sea superior a 60 sería 2,5 veces mayor. Con referencia al icPM10_2, la OR es de 0,76, por lo que también nos indica que un lugar con RS más alta, sería un factor de protección para concentraciones altas de PM10.

Respecto a la variable (month), idem que el apartado anterior.

2.2 Modelo de regresión logística

Para la estación de Montevil del apartado anterior:

a) Se pide construir un modelo de regresión logística tomando como variable dependiente **icPM10** y variables explicativas (RS_re), (vv) y (PRB). Interpretad y calculad las OR.

b) Se añade al modelo del apartado anterior la variable (month).¿Existe una mejora del modelo?. Justificad e interpretad.

```
#a)
logit_1a <- glm(formula=icPM10~vv+RS_re+PRB, data=dat_MO2, family=binomial)
summary(logit_1a)
```

```
##
## Call:
## glm(formula = icPM10 ~ vv + RS_re + PRB, family = binomial, data = dat_MO2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7515  -0.4310  -0.3606  -0.2677   4.3878
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -37.008786   5.874087  -6.300 2.97e-10 ***
## vv                -0.570249   0.060311  -9.455  < 2e-16 ***
## RS_renormal_alta   0.315206   0.124298   2.536   0.0112 *
## PRB                0.034526   0.005786   5.967 2.41e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4271.3  on 8542  degrees of freedom
## Residual deviance: 4106.2  on 8539  degrees of freedom
##   (65 observations deleted due to missingness)
## AIC: 4114.2
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coefficients(logit_1a))
```

```
##      (Intercept)               vv RS_renormal_alta              PRB
##     8.458407e-17     5.653845e-01     1.370542e+00     1.035129e+00
```

Comprobación de la existencia de variable de confusión. No se pide expresamente, por lo que no sería necesario.Si, se deberían dar cuenta que la OR de RS_re ajustada por la vv, difiere de la OR de RS_re sin ajustar e interpretar.

```
logit_1a <- glm(formula=icPM10~RS_re, data=dat_MO2, family=binomial)
exp(coefficients(logit_1a))
```

```
##      (Intercept) RS_renormal_alta
##       0.07591421       0.83891588
```

```
logit_1a <- glm(formula=icPM10~RS_re+PRB, data=dat_MO2, family=binomial)
exp(coefficients(logit_1a))
```

```
##      (Intercept) RS_renormal_alta              PRB
##     1.628530e-18     8.842446e-01     1.038576e+00
```

```
logit_1a <- glm(formula=icPM10~vv+RS_re+vv, data=dat_MO2, family=binomial)
exp(coefficients(logit_1a))
```

```
##      (Intercept)               vv RS_renormal_alta
##        0.1377214        0.5576251        1.3205845
```

a) Se obseva en el modelo que todas las variables escogidas como factores explicativos son significativas con un AIC de 4114,2. Por otro lado cuando se calculan las OR ajustadas, se tiene:

OR para vv de 0,565, por lo que el viento es un factor de protección para el aumento de concentración de PM10. OR para RS_re de 1,37. Al tomar un valor bastante diferente al obtenido en el apartado anterior, se observa que la asociación de la radiación solar (RS_re) con la variable dependiente icPM10 difiere significativamente según se considere, o no la velocidad del viento. Por lo que se podría estar ante una variable de confusión. Esto es posible ya que la vv, tanto está relacionada con icPM10, como con RS_re.

b) Se añade la variable month

```
Fecha1 <- as.Date(dat_MO2$Fecha,format = "%d/%m/%Y")
month<-as.factor(month (Fecha1))
month_Rel=relevel(month, ref = '7')
logit_1b <- glm(formula=icPM10~vv+RS_re+PRB+month_Rel, data=dat_MO2, family=binomial)
summary(logit_1b)
```

```
##
## Call:
## glm(formula = icPM10 ~ vv + RS_re + PRB + month_Rel, family = binomial,
##     data = dat_MO2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8166  -0.4292  -0.3180  -0.2307   4.3194
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -25.734476   6.660837  -3.864 0.000112 ***
## vv                -0.518576   0.061195  -8.474  < 2e-16 ***
## RS_renormal_alta   0.524926   0.139291   3.769 0.000164 ***
## PRB                0.022328   0.006564   3.401 0.000671 ***
## month_Rel1         1.138548   0.272529   4.178 2.94e-05 ***
## month_Rel2         0.165010   0.331476   0.498 0.618623
## month_Rel3         0.465525   0.332847   1.399 0.161928
## month_Rel4         1.200837   0.271257   4.427 9.56e-06 ***
```

```
## month_Rel5         1.009484   0.270058   3.738 0.000185 ***
## month_Rel6         0.349916   0.304965   1.147 0.251217
## month_Rel8         0.405959   0.296431   1.369 0.170847
## month_Rel9         1.100522   0.266990   4.122 3.76e-05 ***
## month_Rel10        1.162074   0.266504   4.360 1.30e-05 ***
## month_Rel11        1.388798   0.263857   5.263 1.41e-07 ***
## month_Rel12        1.868823   0.251861   7.420 1.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4271.3  on 8542  degrees of freedom
## Residual deviance: 3965.9  on 8528  degrees of freedom
##   (65 observations deleted due to missingness)
## AIC: 3995.9
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coefficients(logit_1b))
```

```
##      (Intercept)               vv RS_renormal_alta              PRB
##     6.662833e-12     5.953680e-01     1.690333e+00     1.022579e+00
##       month_Rel1       month_Rel2       month_Rel3       month_Rel4
##     3.122230e+00     1.179405e+00     1.592850e+00     3.322897e+00
##       month_Rel5       month_Rel6       month_Rel8       month_Rel9
##     2.744185e+00     1.418948e+00     1.500741e+00     3.005733e+00
##      month_Rel10      month_Rel11      month_Rel12
##     3.196556e+00     4.010028e+00     6.480667e+00
```

Se obseva que según el mes que estemos, la concentración de PM10 puede aumentar. Tomando como referencia el mes de Julio, los meses de septiembre a Enero, así como abril y mayo son significativos en el modelo, Si se calculan las OR se tiene que en los meses de Septiembre a Enero, con referencia a Julio la probabilidad de que la concentración de PM10 sea superior a 45, es como mínimo casi tres veces mayor, llegando a 6 veces en el mes de Diciembre. Existe una mejora del modelo con un AIC de 3995,9.

c) Se añadirá al modelo anterior como variable explicativa la variable (TMP). Justificad la presencia o no de una posible interacción con (RS_re). ¿Se podría estar ante una variable de confusión?. Razona tu respuesta.

```
logit_1c <- glm(formula=icPM10~vv+RS_re+PRB+month_Rel+TMP, data=dat_MO2, family=binomial)
summary(logit_1c)
```

```
##
## Call:
## glm(formula = icPM10 ~ vv + RS_re + PRB + month_Rel + TMP, family = binomial,
##     data = dat_MO2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1628  -0.4169  -0.3042  -0.2079   4.5357
##
```

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -34.019536   7.091923  -4.797 1.61e-06 ***
## vv                -0.671940   0.065704 -10.227  < 2e-16 ***
## RS_renormal_alta   0.118686   0.144161   0.823  0.41034
## PRB                0.027686   0.006954   3.981 6.85e-05 ***
## month_Rel1         2.582185   0.305551   8.451  < 2e-16 ***
## month_Rel2         1.886742   0.368723   5.117 3.11e-07 ***
## month_Rel3         2.044609   0.367446   5.564 2.63e-08 ***
## month_Rel4         2.380916   0.294334   8.089 6.01e-16 ***
## month_Rel5         1.971338   0.285613   6.902 5.12e-12 ***
## month_Rel6         0.878245   0.309046   2.842  0.00449 **
## month_Rel8         0.366688   0.296856   1.235  0.21674
## month_Rel9         1.190248   0.267571   4.448 8.65e-06 ***
## month_Rel10        1.834613   0.273304   6.713 1.91e-11 ***
## month_Rel11        2.465648   0.283867   8.686  < 2e-16 ***
## month_Rel12        2.972714   0.273212  10.881  < 2e-16 ***
## TMP                0.141496   0.013930  10.158  < 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: 4271.3  on 8542  degrees of freedom
## Residual deviance: 3862.0  on 8527  degrees of freedom
##   (65 observations deleted due to missingness)
## AIC: 3894
##
## Number of Fisher Scoring iterations: 6
```

```
#Se analiza el modelo con interacción
logit_1c <- glm(formula=icPM10~vv+RS_re+PRB+month_Rel+TMP+RS_re:TMP, data=dat_MO2, family=binomial)
summary(logit_1c)
```

```
##
## Call:
## glm(formula = icPM10 ~ vv + RS_re + PRB + month_Rel + TMP + RS_re:TMP,
##     family = binomial, data = dat_MO2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1523  -0.4168  -0.3041  -0.2078   4.5356
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -34.008522   7.091281  -4.796 1.62e-06 ***
## vv                   -0.673422   0.065847 -10.227  < 2e-16 ***
## RS_renormal_alta     -0.289956   0.498785  -0.581  0.56102
## PRB                   0.027716   0.006953   3.986 6.72e-05 ***
## month_Rel1            2.595507   0.306053   8.481  < 2e-16 ***
## month_Rel2            1.906787   0.369300   5.163 2.43e-07 ***
## month_Rel3            2.069835   0.368525   5.617 1.95e-08 ***
## month_Rel4            2.397088   0.294889   8.129 4.34e-16 ***
## month_Rel5            1.979407   0.285724   6.928 4.28e-12 ***
```

```
## month_Rel6             0.873160   0.309128   2.825  0.00473 **
## month_Rel8             0.366749   0.296933   1.235  0.21678
## month_Rel9             1.196028   0.267796   4.466 7.96e-06 ***
## month_Rel10            1.845083   0.273918   6.736 1.63e-11 ***
## month_Rel11            2.467927   0.284192   8.684  < 2e-16 ***
## month_Rel12            2.972661   0.273529  10.868  < 2e-16 ***
## TMP                    0.138503   0.014355   9.648  < 2e-16 ***
## RS_renormal_alta:TMP   0.022643   0.026237   0.863  0.38813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4271.3  on 8542  degrees of freedom
## Residual deviance: 3861.2  on 8526  degrees of freedom
##   (65 observations deleted due to missingness)
## AIC: 3895.2
##
## Number of Fisher Scoring iterations: 6
```

En este caso, no existe interacción entre dichas variables, ya que el término de interacción RS_re:TMP, no es estadísticamente significativo, pero si se observa que al añadir la variable TMP al modelo, RS_re deja de ser significativa, por lo que se podría estar ante una variable de confusión.

2.3 Predicción

Según el modelo del apartado b), calculad la probabilidad de que la concentración de PM10 sea o no superior a 45, con unos valores de vv= 0.6, RS_re="Normal_alta",PRB= 1013, en el mes de Agosto.

```
pred <-predict (logit_1b, data.frame (RS_re = "normal_alta",vv= 0.6, PRB=1013, month_Rel="8"),
                type = "response")
pred
```

```
##          1
## 0.07607242
```

Este modelo nos predice una probablidad del 7,6 % de que icPM10 sea superior a 45, suponiendo los valores que se han tomado para las variables explicativas.

2.4 Bondad del ajuste

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

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
logit_1b_dat<- model.frame(logit_1b)
hoslem.test(logit_1b_dat$icPM10,fitted(logit_1b))
```

```
## Warning in Ops.factor(1, y): '-' not meaningful for factors
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logit_1b_dat$icPM10, fitted(logit_1b)
## X-squared = 8543, df = 8, p-value < 2.2e-16
```

A la vista de los resultados el ajuste no es muy bueno con un p-value próximo a 0.

2.5 Curva ROC

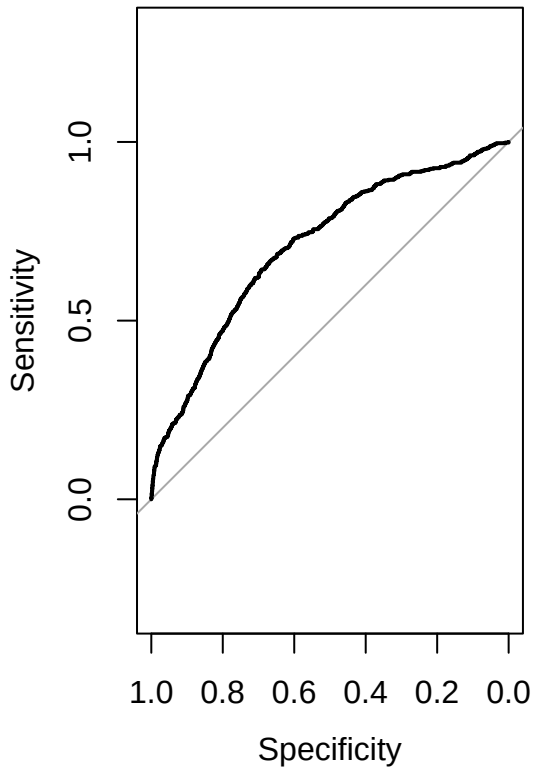
Dibujar la curva ROC y calcular el área debajo de la curva con el modelo del apartado b). Discutir el resultado.

```
library(pROC)
```

```
prob_low=predict(logit_1b, dat_MO2, type="response")
r=roc(dat_MO2$icPM10,prob_low, data=dat_MO2)
```

```
par(mfrow=c(1,2))
plot(r)
auc(r)
```

```
## Area under the curve: 0.7101
```



En el modelo, el área por debajo de la curva toma el valor de 0.71, por lo que la habilidad del modelo para predecir es buena.

2.6 Conclusiones

- Se puede concluir que la calidad de aire en las zonas evaluadas en este estudio, no es la adecuada. Para determinar la calidad del aite se han tomado datos de varios contaminantes, centrándose este análisis en NO2, PM10, O3 y SO2. Los valore medios permitidos de NO2 y PM10 han sido superados en todas las estaciones de monitoreo escogidas, y dos de ellas superan el valor máximo 120 g/m3 de NO2. Los valores máximos de O3, también han sido superados en todas las estaciones, siendo el valor más alto registrado en la de Montevil. Por otro lado en la zona más industrial se ha observado una mayor concentración de dióxido de azufre.

- Con referencia a los modelos de regresión lineal, si existe relación entre las variables meteorológicas y contaminantes atmosféricos. Con respecto al ozono, se observa que las variables más relacionadas en las estaciones escogidas son la vv (velocidad del viento), RS (radiación solar), HR (humedad relativa) y LL (precipitaciones). Una vez ajustado el modelo con dichas variables explicativas se ha obtenido un coeficiente de determinación ajustado de 0.6107, en la estación de Montevil y de 0.5905 en Avenida Constitución Se comprueba que al añadir la variable TMP es significativa en la estación de Montevil y no se observan problemas de colinealidad, por lo que sería adecuado añadirla al modelo anterior, para dicha estación de monitoreo.

- En vista a los resultados obtenidos con el cálculo de las OR, tomando como variables dependientes icO3 e icPM10, la variable explicativa RS_re puede considerarse factor de riesgo para el aumento de la concentración de O3 y factor de protección para el aumento de PM10, con OR de 5,25 y 0,83, en la estación de Montevil y de 2,49 y 0,76 para la estación de Avenida Constitución.

  Por otro lado al construir los modelos de regresión logística para la estación de Montevil, ajustados por varias variables explicativas, se observa que tanto RS_re, vv, PRB y month, son significativas y se obtiene un indicador AIC = 3995,9. Se ha comprobado que al incorporar la variable Tmp al modelo de regresión logística con las variables citadas, la asociación de la radiación solar con la variable dependiente difiere hasta el punto de dejar de ser significativa. Por lo que se podría estar ante un problema de confusión.

  Del estudio de la curva ROC, se puede deducir que la habilidad de dicho modelo para predecir es aceptable, aunque el test de Hosman-Lemeshow no haya dado significativo.