

A3 - Modelització predictiva

Xavier Vizcaino Gascon

7 de mayo, 2022

Contents

1. Regressió lineal	1
2. Regressió logística	18
3. Conclusions de l'anàlisi	28

1. Regressió lineal

L'exposició a les partícules (PM10), a l'ozó (O₃), al diòxid de nitrogen (NO₂) i el diòxid de sofre (SO₂), plantegen greus riscos per a la salut. Les directrius de l'OMS sobre la qualitat de l'aire estableixen els límits sobre aquests principals contaminants atmosfèrics.

- PM10: Límit de 45 micrograms de partícules per cada metre cúbic μ/m^3 .
- SO₂: Límit de 40 μ/m^3
- NO₂: Límit de 25 μ/m^3
- O₃: Límit de 60 μ/m^3

L'índex de qualitat de l'aire es calcula de manera individual tenint en compte cadascun de dites contaminants. Tots aquests valors estan referits a la mitjana diària. Amb referència a valors màxims diaris es prendran els valors de 100 μ/m^3 per a O₃ i de 120 μ/m^3 per a NO₂. Tant per a PM10 i SO₂, es prendran com a referència únicament els valors mitjans diaris per a comparar.

1.1. Estudi comparatiu entre estacions

a) Estudi dels valors mitjans i màxims diaris de cada contaminant. Per a cadascuna de les estacions de monitorització, es calcularan els valors màxims i mitjans diaris de cada contaminant. Posteriorment es farà una comparativa entre les cinc estacions sobre la base d'aquests valors. Interpreteu tenint en compte els límits esmentats anteriorment.

Es carreguen les dades i s'adapta el tipus de dades en algunes variables.

```
Air <- read.csv("dat_Air_Stations.csv")
Air$Estacion<-as.factor(Air$Estacion)
Air$Nombre<-as.factor(Air$Nombre)
Air$Fecha<-as.Date(Air$Fecha, format = "%d/%m/%Y")
```

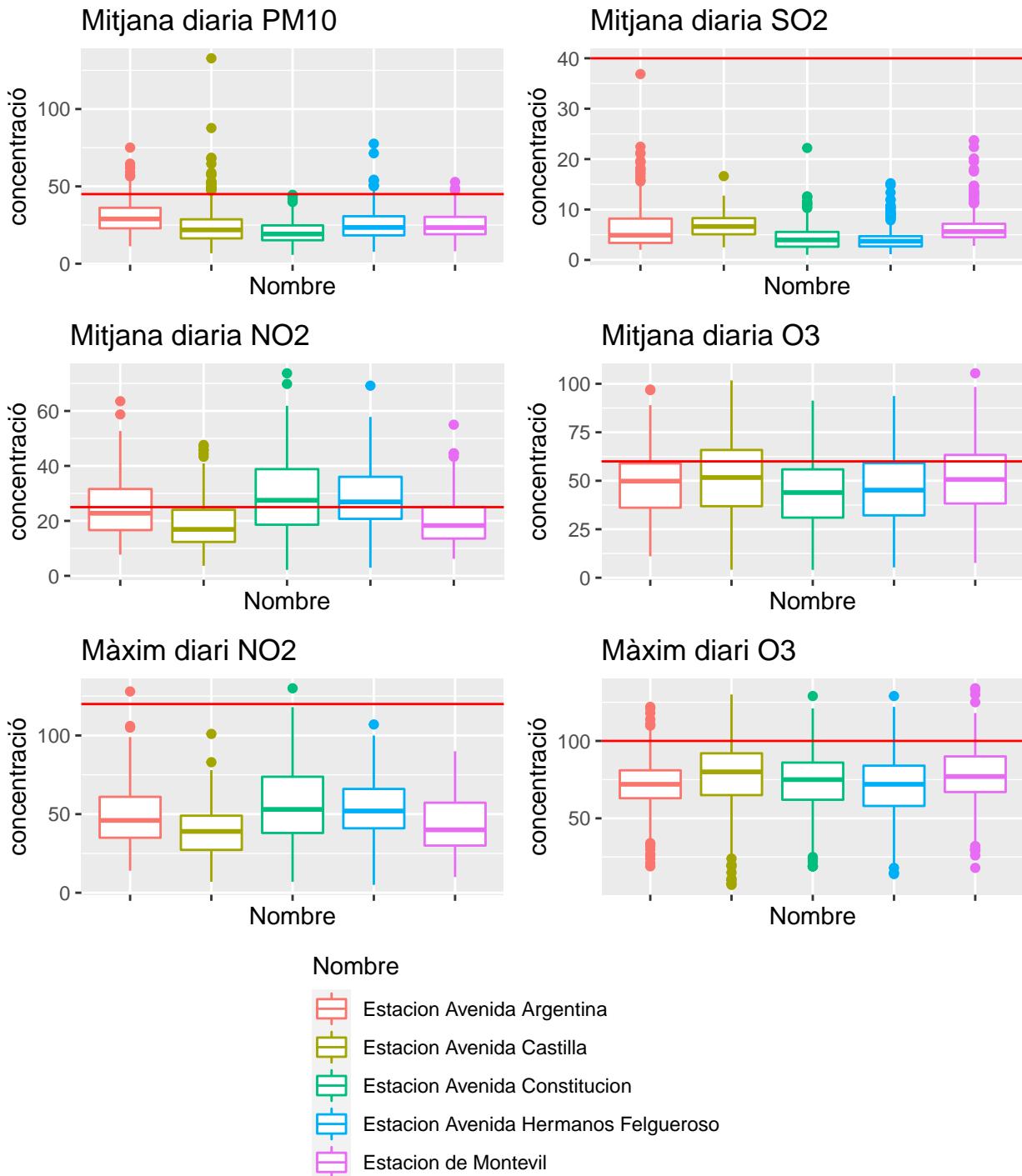
S'agrupen les dades per *Nombre* i *Fecha* per a poder agregar resultats i extreure el màxim i el promig diari per cada estació per a tots els contaminants del dataset.

```
daily_max<-Air %>%
  group_by(Nombre, Fecha) %>%
  summarise_at(vars(SO2:PM25), max)

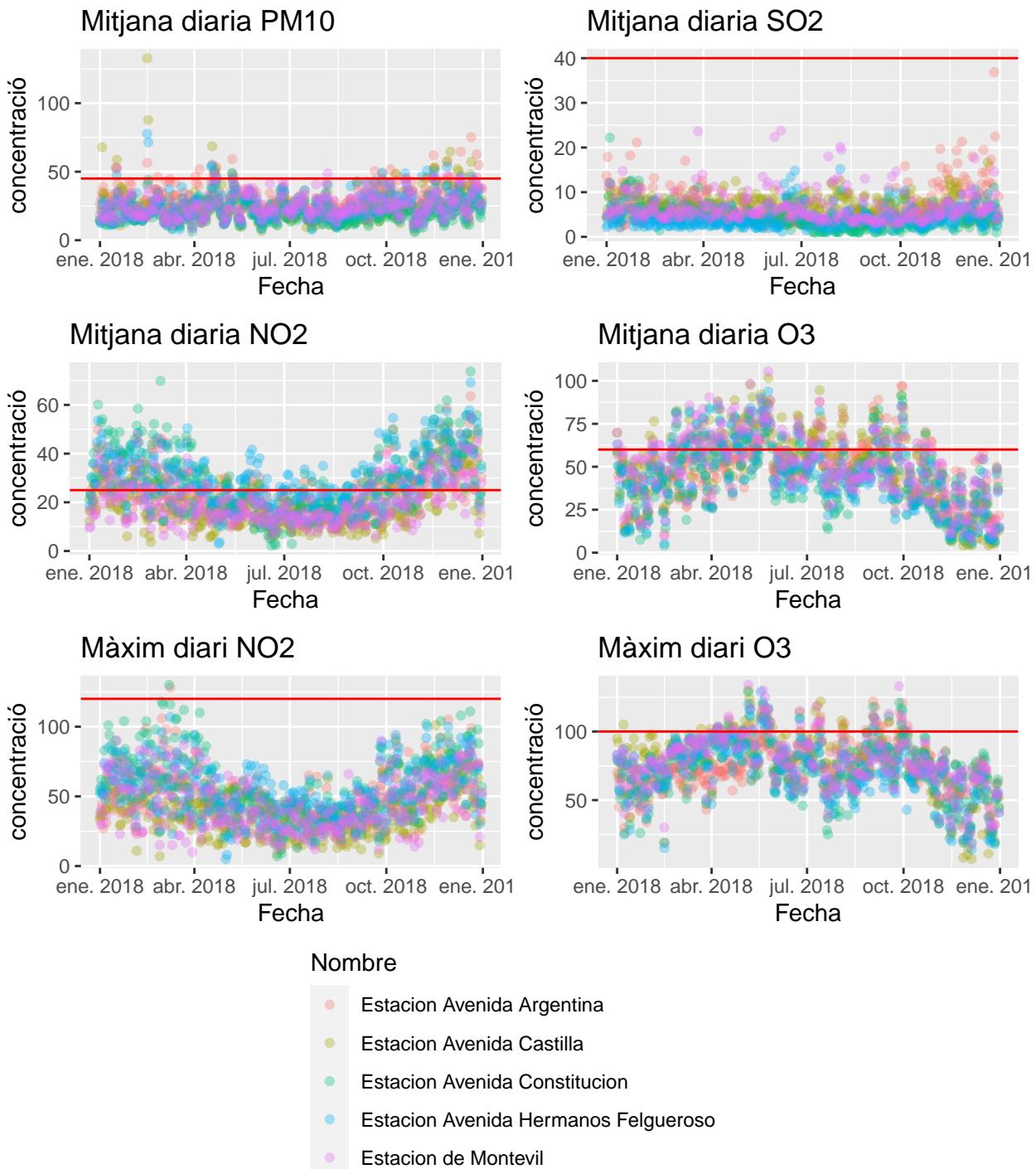
daily_mean<-Air %>%
```

```
group_by(Nombre, Fecha) %>%
summarise_at(vars(S02:PM25), mean)
```

Es grafiquen les dades de cada contaminant en format boxplot separant per colors les estacions de lectura. També s'afegeixen els límits proporcionats per als contaminants, tant per la lectura de promig com per la lectura de màxim diari.



Per obtenir informació de la estacionalitat es grafiquen els valors promig i màxim de cada contaminant per data.



S'estudia el nombre de dies en que el valor promig diari dels contaminants definits (PM10, SO2, NO2 i O3) està per sobre del límit establert, per cada estació d'enregistrament.

```
res_mean<-daily_mean %>%
  ungroup() %>%
  group_by(Nombre) %>%
  summarise(
    d_PM10_mean = sum(PM10>45, na.rm = TRUE),
    d_SO2_mean = sum(SO2>40, na.rm = TRUE),
    d_NO2_mean = sum(NO2>25, na.rm = TRUE),
    d_O3_mean = sum(O3>60, na.rm = TRUE)
  )

kable(res_mean, caption = "Dies amb promig diari superior al límit")
```

Table 1: Dies amb promig diari superior al límit

Nombre	d_PM10_mean	d_SO2_mean	d_NO2_mean	d_O3_mean
Estacion Avenida Argentina	27	0	147	78
Estacion Avenida Castilla	21	0	76	115
Estacion Avenida Constitucion	0	0	185	59
Estacion Avenida Hermanos Felgueroso	17	0	204	78
Estacion de Montevil	9	0	84	100

S'estudia el nombre de dies en que el valor màxim diari dels contaminants definits (O3 i NO2) està per sobre del límit, per cada estació d'enregistrament.

```
res_max<-daily_max %>%
  ungroup() %>%
  group_by(Nombre) %>%
  summarise(
    d_NO2_max = sum(NO2>120, na.rm = TRUE),
    d_O3_max = sum(O3>100, na.rm = TRUE)
  )

kable(res_max, caption = "Dies amb màxim diari superior al límit")
```

Table 2: Dies amb màxim diari superior al límit

Nombre	d_NO2_max	d_O3_max
Estacion Avenida Argentina	1	19
Estacion Avenida Castilla	0	43
Estacion Avenida Constitucion	1	20
Estacion Avenida Hermanos Felgueroso	0	18
Estacion de Montevil	0	36

Preneint la informació dels boxplots i de les taules resum s'observa clarament que existeixen problemes **greus** de compliment amb els límits per als contaminants:

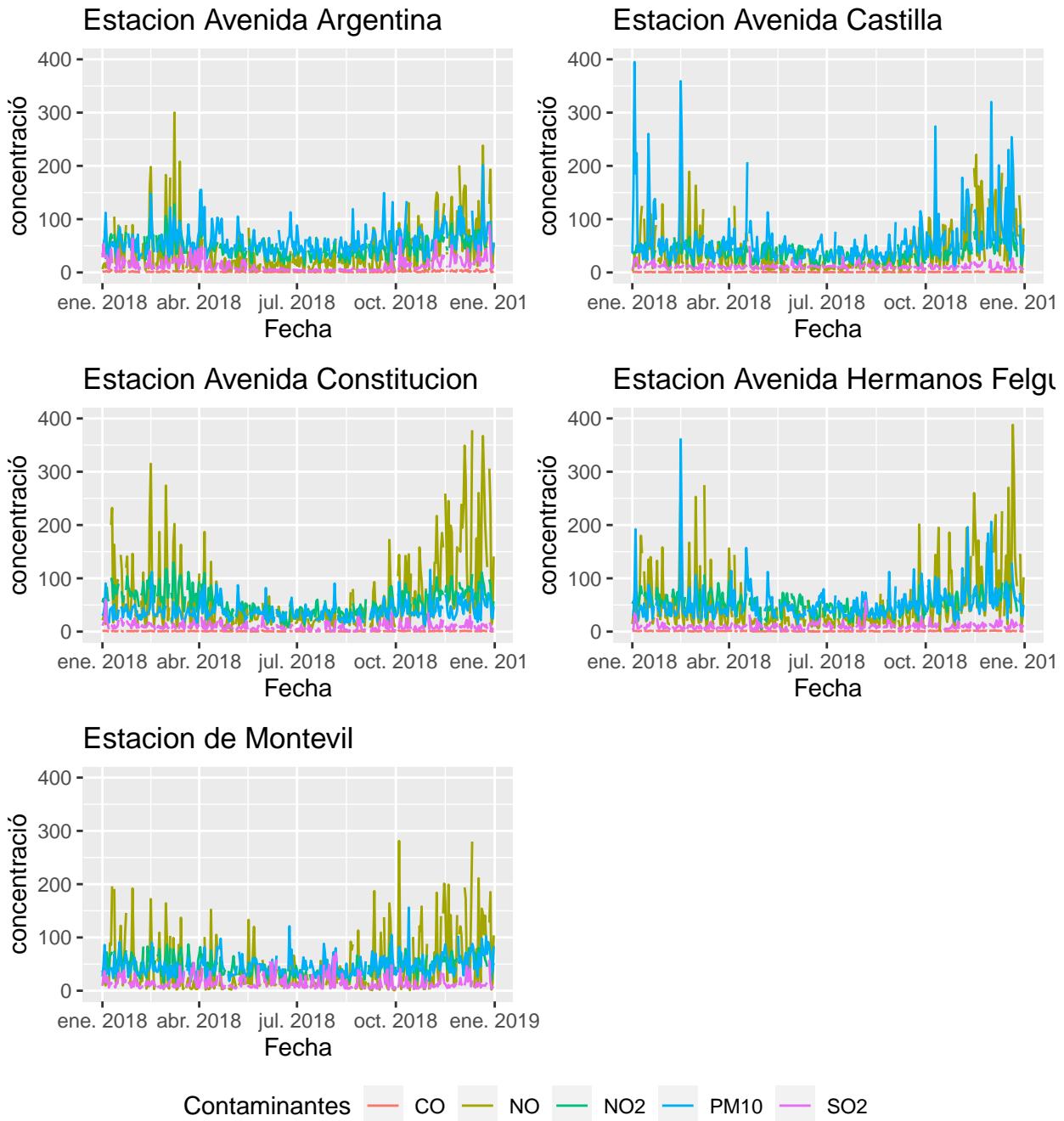
- NO2 en el valor de la mitja diària
- O3 en el valor de la mitja diària
- O3 en el valor màxim diari

i problemes **moderats** en el compliment per als contaminants:

- PM10 en el valor de la mitja diària

b) Representeu gràficament l'evolució de cadascun dels contaminants en cada estació. Es prendran els valors màxims diaris.

```
daily_max_Long<-daily_max %>%
  select(Nombre:PM10) %>%
  pivot_longer(cols = S02:PM10, names_to = "Contaminantes", values_to = "valor") %>%
  ungroup()
```



c) Estudi de correlació lineal. Per a això se seleccionen les dues estacions amb registres meteorològics: Estació de Montevil i Estació Avinguda Constitució. Per a cadascuna de les estacions, calcular la matriu de correlació entre els contaminants citats anteriorment i les variables meteorològiques: Temperatura (TMP), Humitat Relativa (HR), Radiació solar (RS), velocitat del vent (vv), precipitacions (LL) i Pressió baromètrica (PRB). Interpreteu.

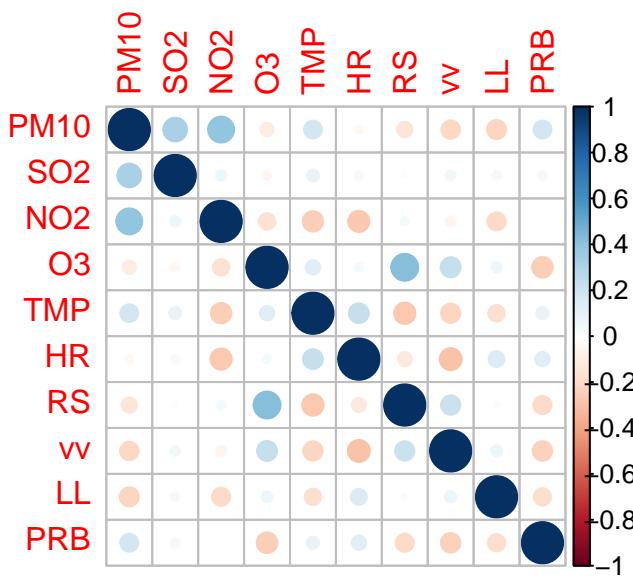
Nota: La matriu de correlació serà calculada sobre la base dels valors màxims de cada contaminant.

Nota2: El motiu d'aquests primers apartats és prendre un primer contacte sobre les possibles diferències entre estacions, així com fer-se una idea de les relacions existents entre les variables, però per a construir els models de regressió es prendran les dades per hora.

```
#desagrupem les dades
daily_max<-ungroup(daily_max)
#filtrem per E.Montevil i obtenim correlació
dm_Montevil<-daily_max %>%
  filter(Nombre=="Estacion de Montevil") %>%
  select(PM10, SO2, NO2, O3, TMP, HR, RS, vv, LL, PRB)
cor_Montevil<-cor(dm_Montevil, use = "pairwise.complete.obs")
kable(cor_Montevil, caption = "Matriu de correlació - Montevil", digits = 2)
```

Table 3: Matriu de correlació - Montevil

	PM10	SO2	NO2	O3	TMP	HR	RS	vv	LL	PRB
PM10	1.00	0.33	0.40	-0.10	0.19	-0.04	-0.14	-0.20	-0.21	0.19
SO2	0.33	1.00	0.06	-0.04	0.08	0.04	0.01	0.06	0.04	0.05
NO2	0.40	0.06	1.00	-0.16	-0.24	-0.26	0.05	-0.06	-0.20	0.00
O3	-0.10	-0.04	-0.16	1.00	0.12	0.04	0.42	0.23	0.07	-0.25
TMP	0.19	0.08	-0.24	0.12	1.00	0.23	-0.27	-0.22	-0.16	0.09
HR	-0.04	0.04	-0.26	0.04	0.23	1.00	-0.12	-0.28	0.15	0.12
RS	-0.14	0.01	0.05	0.42	-0.27	-0.12	1.00	0.22	-0.02	-0.19
vv	-0.20	0.06	-0.06	0.23	-0.22	-0.28	0.22	1.00	0.07	-0.22
LL	-0.21	0.04	-0.20	0.07	-0.16	0.15	-0.02	0.07	1.00	-0.18
PRB	0.19	0.05	0.00	-0.25	0.09	0.12	-0.19	-0.22	-0.18	1.00



Per l'estació de Montevil, s'observen certs nivells de correlació positiva entre les variables:

- PM10, SO2 i NO2
- O3 i RS

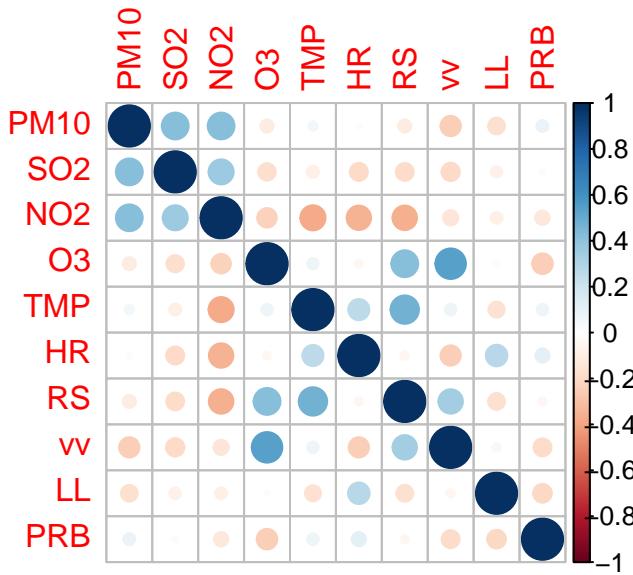
Així com certs nivells de correlació negativa entre les variables:

- PM10 i LL
- NO2, TMP i HR
- O3 i PRB

```
#filtrem per E. Constitució i obtenim correlació
dm_Constitucio<-daily_max %>%
  filter(Nombre=="Estacion Avenida Constitucion") %>%
  select(PM10, S02, N02, O3, TMP, HR, RS, vv, LL, PRB)
cor_Constitucio<-cor(dm_Constitucio, use="pairwise.complete.obs")
kable(cor_Constitucio, caption = "Matriu de correlació - Constitució", digits = 2)
```

Table 4: Matriu de correlació - Constitució

	PM10	SO2	N02	O3	TMP	HR	RS	vv	LL	PRB
PM10	1.00	0.42	0.42	-0.10	0.05	-0.02	-0.10	-0.24	-0.17	0.08
SO2	0.42	1.00	0.37	-0.18	-0.09	-0.19	-0.19	-0.20	-0.08	0.02
N02	0.42	0.37	1.00	-0.22	-0.37	-0.35	-0.35	-0.14	-0.08	-0.12
O3	-0.10	-0.18	-0.22	1.00	0.08	-0.04	0.42	0.55	0.02	-0.25
TMP	0.05	-0.09	-0.37	0.08	1.00	0.26	0.47	0.08	-0.15	0.08
HR	-0.02	-0.19	-0.35	-0.04	0.26	1.00	-0.04	-0.24	0.27	0.12
RS	-0.10	-0.19	-0.35	0.42	0.47	-0.04	1.00	0.35	-0.17	-0.04
vv	-0.24	-0.20	-0.14	0.55	0.08	-0.24	0.35	1.00	-0.04	-0.19
LL	-0.17	-0.08	-0.08	0.02	-0.15	0.27	-0.17	-0.04	1.00	-0.21
PRB	0.08	0.02	-0.12	-0.25	0.08	0.12	-0.04	-0.19	-0.21	1.00



Per l'estació d'Avinguda Constitució, s'observen certs nivells de correlació positiva entre les variables:

- PM10, SO2 i NO2
- O3, RS i vv

Així com certs nivells de correlació negativa entre les variables:

- PM10 i vv
- NO₂, TMP, HR i RS

1.2. Model de regressió lineal

Com he esmentat a dalt, per a construir els models de regressió, es prendran els valors de les variables triades per hora, tal com apareixen en la base de dades original.

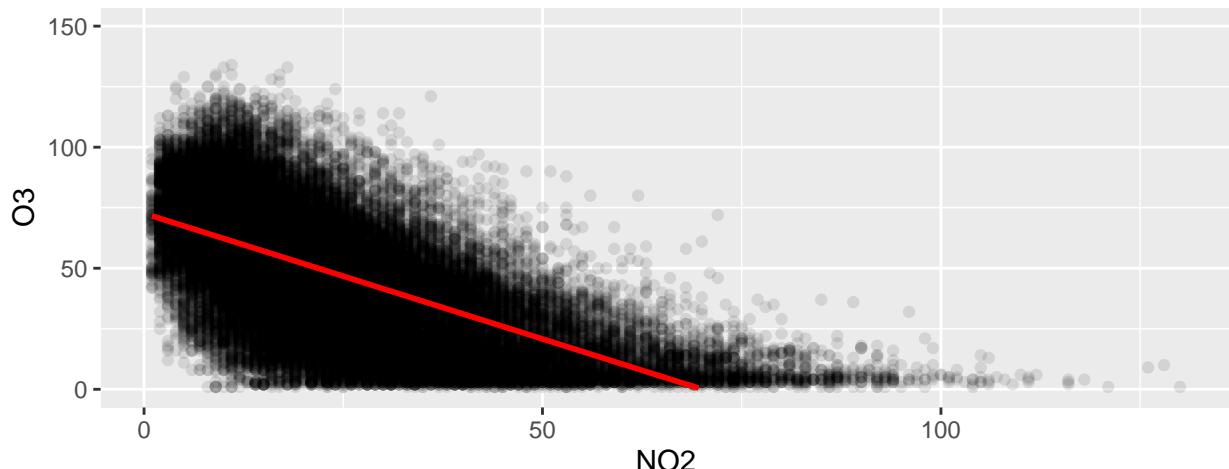
a) Es demana crear un model de regressió lineal, prenent com a variable dependent (O_3) i variable explicativa (NO_2). S'avaluarà la bondat de l'ajust, a partir del coeficient de determinació. Interpreteu.

```
model_03<-lm(O3~NO2, Air)
summary(model_03)
```

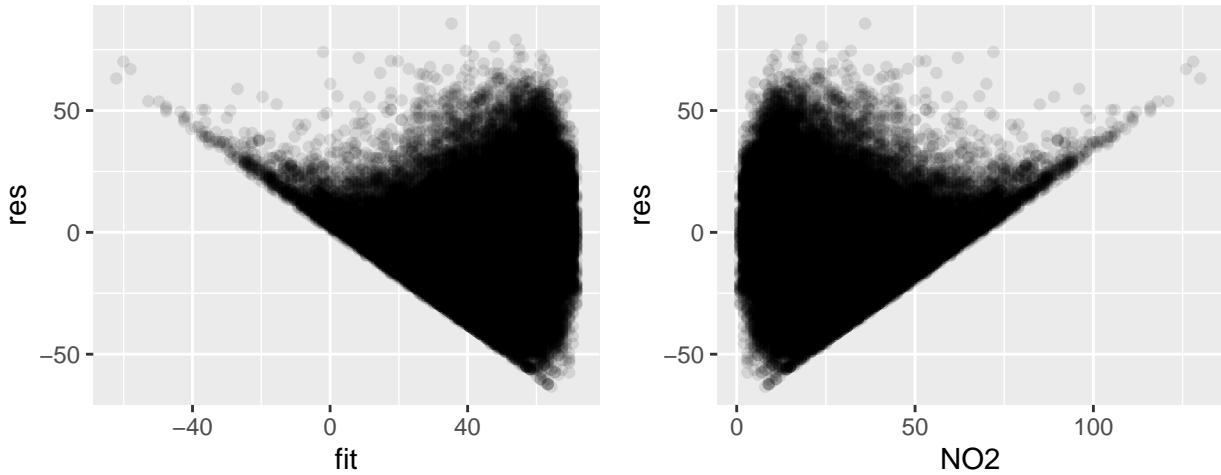
```
##
## Call:
## lm(formula = O3 ~ NO2, data = Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -63.353 -13.910   0.828  13.868  85.682 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 72.649230  0.174380  416.6   <2e-16 ***
## NO2        -1.036976  0.005861  -176.9   <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.53 on 43118 degrees of freedom
## (679 observations deleted due to missingness)
## Multiple R-squared:  0.4207, Adjusted R-squared:  0.4206 
## F-statistic: 3.131e+04 on 1 and 43118 DF,  p-value: < 2.2e-16
```

El coeficient de determinació ajustat és 0.4206, fet que implica que menys de la meitat de la variança de les dades és explicada pel model, per tant podem concloure que la bondat de l'ajust és certament força limitada.

Aquest fet es pot refermar amb l'observació de la representació gràfica de les dades i la recta de regressió següent.



Graficant els valors ajustats pel model enfront els residus s'obté el següent:



La forma triangular del gràfic de residus enfront els valors ajustats s'explica per la aplicació del model sobre el set de dades amb les següents característiques:

- Els valors de concentració de NO₂ en les dades originals pertanyen a l'interval [0, INF).
- Els valors de concentració de O₃ en les dades originals pertanyen a l'interval [0, INF).
- En les dades originals, podem tenir concentracions de O₃ ~ 0, per a concentracions de NO₂ en l'interval ~(10, 125)

b) S'afegeix al model anterior el nom de les estacions (Nom). Interpreteu.

```
model_03_u<-update(model_03, .~. + Nombre, Air)
summary(model_03_u)

##
## Call:
## lm(formula = O3 ~ NO2 + Nombre, data = Air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -60.684 -13.871   0.679  13.893  85.154 
##
## Coefficients:
## (Intercept)          NO2        NombreEstacion Avenida Castilla 
##                   75.305599    -1.065195    -5.100212    -1.112099 
##                   (Avenida Constitucion) (Avenida Hermanos Felgueroso) (NombreEstacion de Montevil) 
##                   -1.112099    0.231370   -3.788765 
##                   (Intercept)          NO2        NombreEstacion Avenida Castilla 
##                   < 2e-16 ***    < 2e-16 ***    < 2e-16 *** 
##                   NombreEstacion Avenida Constitucion (Avenida Hermanos Felgueroso) (NombreEstacion de Montevil) 
##                   0.000378    0.458300    < 2e-16 *** 
##                   --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

## 
## Residual standard error: 20.42 on 43114 degrees of freedom
##   (679 observations deleted due to missingness)
## Multiple R-squared:  0.4266, Adjusted R-squared:  0.4265
## F-statistic:  6415 on 5 and 43114 DF,  p-value: < 2.2e-16

```

El coeficient de determinació és ara 0.4265 quan anteriorment era 0.4206, fet que indica que la introducció de la variable *nom* millora el model, però no de manera notable.

```

#obtenim coeficients dels dos models
c1<-data.frame(model=model_03$coefficients)
c2<-data.frame(model_updated=model_03_u$coefficients)
#a juntem en un únic dataframe
model_03_summary<-merge(c1, c2, by="row.names", all = "TRUE")
kable(model_03_summary, caption = "Comparativa coeficients models 03")

```

Table 5: Comparativa coeficients models O3

Row.names	model	model_updated
(Intercept)	72.649230	75.3055990
NO2	-1.036976	-1.0651950
NombreEstacion Avenida Castilla	NA	-5.1002121
NombreEstacion Avenida Constitucion	NA	-1.1120988
NombreEstacion Avenida Hermanos Felgueroso	NA	0.2313699
NombreEstacion de Montevil	NA	-3.7887646

Si comparem els coeficients estimates dels dos models; l'inicial i l'actualitzat observem que la introducció de la variable categòrica *Nombre* gairebé no modifica els coeficients de les variables en el model inicial. Aquest fet suporta la afirmació anteriorment realitzada que la introducció de la nova variable no modifica/millora notablement el model.

Addicionalment observem que la introducció de la variable *nom*, donat que es tracta d'una variable tipus *factor* ha fet que el programa generi variables *dummy* per a representar cada un dels possibles nivells de la variable. Observem també que tot i tenir 5 possibles nivells en la variable *Nombre*, el model només introduceix 4 nivells, deixant *Estacion Avenida Argentina* fora. Evidentment, com que la variable *Nombre* només pot prendre 5 valors, per a aquelles observacions on les 4 variables relacionades amb el *Nombre* siguin 0, s'assumirà que es tracta del nivell restant o referència. Aquesta manera de procedir garanteix que no hi hagi problemes de multicolinealitat.

Finalment Observant el t-value (0.7417) i el p-valor (0.4583) de la variable *NombreEstacion Avenida Hermanos Felgueroso* podem afirmar que aquesta variable no és estadísticament rellevant i es podria treure del model.

1.3. Model de regressió lineal múltiple

Es vol construir un model de regressió múltiple amb el qual puguem predir la concentració d'ozó (*O3*) en les zones de Montevil i Avinguda de la Constitució.

a) Es demana dos models (un per a cada estació) prenent com a variable dependent el nivell d'ozó (*O3*) en funció de la concentració de diòxid de nitrogen (*NO2*) i diferents variables meteorològiques com *vv* (velocitat del vent), *RS* (radiació solar), *HR* (humitat relativa) i *LL* (precipitacions).

Estació de Montevil

```
#filtrem dades Montevil i generem model
Air_Montevil<-Air %>%
  filter(Nombre=="Estacion de Montevil") %>%
  select(PM10, S02, NO2, O3, TMP, HR, RS, vv, LL, PRB)
model_03_m_Montevil<-lm(O3~NO2+vv+RS+HR+LL, Air_Montevil)
summary(model_03_m_Montevil)

##
## Call:
## lm(formula = O3 ~ NO2 + vv + RS + HR + LL, data = Air_Montevil)
##
## 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 ***
## vv          4.743372   0.198630 23.880 <2e-16 ***
## RS          0.031481   0.001496 21.042 <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
```

Estació Avinguda de la constitució

```
#filtrem dades Constitució i generem model
Air_Constitucio<-Air %>%
  filter(Nombre=="Estacion Avenida Constitucion") %>%
  select(PM10, S02, NO2, O3, TMP, HR, RS, vv, LL, PRB)
model_03_m_Constitucio<-lm(O3~NO2+vv+RS+HR+LL, Air_Constitucio)
summary(model_03_m_Constitucio)

##
## Call:
## lm(formula = O3 ~ NO2 + vv + RS + HR + LL, data = Air_Constitucio)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -79.604 -12.997 -0.311 11.341 64.797 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 83.105742   1.503345 55.281 < 2e-16 ***
## NO2        -0.770074   0.011100 -69.374 < 2e-16 ***
## vv          15.193432   0.477656 31.808 < 2e-16 ***  
## RS          0.007788   0.001378  5.651 1.65e-08 ***
```

```

## HR          -0.352352  0.016727 -21.065 < 2e-16 ***
## LL          2.846492  0.352262  8.081 7.35e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.83 on 8276 degrees of freedom
##   (478 observations deleted due to missingness)
## Multiple R-squared:  0.5904, Adjusted R-squared:  0.5902
## F-statistic:  2386 on 5 and 8276 DF,  p-value: < 2.2e-16

```

b) S'afegeix als models anteriors la variable Temperatura (TMP). De ser necessari, es demana comprovar la presència o no de colinealitat entre les variables (vv) i (TMP). Podeu usar la llibreria (faraway) i estudiar el FIV (factor d'inflació de la variància). Discutiu si seria indicat o no afegir la variable (TMP) a cadascun dels models.

Estació de Montevil

```
model_03_m_Montevil_2<-update(model_03_m_Montevil, .~. + TMP, Air_Montevil)
summary(model_03_m_Montevil_2)
```

```

##
## Call:
## lm(formula = O3 ~ NO2 + vv + RS + HR + LL + TMP, data = Air_Montevil)
##
## 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 ***
## vv          4.900684  0.199737 24.536 < 2e-16 ***
## RS          0.030898  0.001496 20.660 < 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

```

Inicialment s'analitza la correlació entre les diferents variables del model actualitzat.

```
cor_Montevil<-cor(Air_Montevil[c("NO2","vv","RS","HR","LL","TMP")],  
                     use = "pairwise.complete.obs")  
kable(cor_Montevil, caption = "Matriu de correlació - Montevil", digits = 2)
```

Table 6: Matriu de correlació - Montevil

	NO2	vv	RS	HR	LL	TMP
NO2	1.00	-0.37	-0.17	0.08	-0.05	-0.26
vv	-0.37	1.00	0.29	-0.43	0.08	0.05
RS	-0.17	0.29	1.00	-0.42	-0.07	0.15
HR	0.08	-0.43	-0.42	1.00	0.16	-0.18
LL	-0.05	0.08	-0.07	0.16	1.00	-0.11
TMP	-0.26	0.05	0.15	-0.18	-0.11	1.00

A través de la taula de correlació s'observa que no existeix correlació forta entre cap parell de variables independents utilitzades en la definició del model. Tanmateix, a continuació s'analitzen els coeficients del model inicial i del model actualitzat buscant canvis substancials.

```
#obtenim coeficients dels dos models  
c1<-data.frame(model=model_03_m_Montevil$coefficients)  
c2<-data.frame(model_updated=model_03_m_Montevil_2$coefficients)  
#ajuntem en un únic dataframe  
model_03_montevil_summary<-merge(c1, c2, by="row.names", all = "TRUE")  
kable(model_03_montevil_summary, caption = "Comparativa coeficients models Montevil")
```

Table 7: Comparativa coeficients models Montevil

Row.names	model	model_updated
(Intercept)	87.0835157	81.4986470
HR	-0.2933396	-0.2790891
LL	2.5863805	2.7477518
NO2	-1.0827110	-1.0573474
RS	0.0314808	0.0308975
TMP	NA	0.2371410
vv	4.7433723	4.9006840

Es pot concloure que la introducció de la variable *TMP* no produceix canvis significatius en els coeficients de les altres variables, comparant el model de regressió inicial amb l'actualitzat.

Com a darrer pas, s'analitza el factor d'inflació de la variança.

```
vif(model_03_m_Montevil_2)
```

```
## NO2      vv      RS      HR      LL      TMP  
## 1.284549 1.504052 1.258373 1.499060 1.062499 1.136578
```

Amb l'anàlisi del factor d'inflació de la variança podem afirmar que no hi ha presencia de colinealitat, doncs tots els factors retornats per a les diferents variables independents son relativament propers a la unitat.

Finalment, es pot concloure que la introducció de la variable *TMP* en el model per a la Estació de Montevil:

- Millora lleugerament el model doncs el coeficient de determinació ajustat creix de 0.6107 a 0.6124.
- Redueix lleugerament l'error de predicció de 17.47 a 17.43
- Es considera estadísticament significativa doncs el P-valor per a la variable *TMP* és $2.5888754 \times 10^{-10}$
- No presenta un alt grau de correlació amb altres variables independents utilitzades en el model.
- No presenta colinealitat

Amb les conclusions anteriors la variable *TMP* es podria incloure en el model per a la Estació de Montevil.

Estació Avinguda de la Constitució

```
model_03_m_Constitucio_2<-update(model_03_m_Constitucio, .~. + TMP, Air_Constitucio)
summary(model_03_m_Constitucio_2)
```

```
##
## Call:
## lm(formula = O3 ~ NO2 + vv + RS + HR + LL + TMP, data = Air_Constitucio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -79.610  -13.003  -0.309  11.341  64.784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.091223   1.595550  52.077 < 2e-16 ***
## NO2        -0.769998   0.011445 -67.276 < 2e-16 ***
## vv         15.194078   0.478275  31.768 < 2e-16 ***
## RS         0.007775   0.001468   5.296 1.21e-07 ***
## HR        -0.352388   0.016780 -21.000 < 2e-16 ***
## LL         2.847457   0.354070   8.042 1.01e-15 ***
## TMP        0.001180   0.043431   0.027    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.83 on 8275 degrees of freedom
## (478 observations deleted due to missingness)
## Multiple R-squared:  0.5904, Adjusted R-squared:  0.5901
## F-statistic:  1988 on 6 and 8275 DF,  p-value: < 2.2e-16
```

Anàlogament a l'anàlisi realitzat per l'estació de Montevil, per a l'estació d'Avinguda de la Constitució s'analitza la correlació entre les diferents variables del model actualitzat

```
cor_Constitucio<-cor(Air_Constitucio[c("NO2", "vv", "RS", "HR", "LL", "TMP")]),
use = "pairwise.complete.obs")
kable(cor_Constitucio, caption = "Matriu de correlació - Constitució", digits = 2)
```

Table 8: Matriu de correlació - Constitució

	NO2	vv	RS	HR	LL	TMP
NO2	1.00	-0.35	-0.23	-0.10	-0.01	-0.34
vv	-0.35	1.00	0.52	-0.41	0.02	0.19
RS	-0.23	0.52	1.00	-0.31	-0.09	0.38
HR	-0.10	-0.41	-0.31	1.00	0.16	0.01
LL	-0.01	0.02	-0.09	0.16	1.00	-0.12
TMP	-0.34	0.19	0.38	0.01	-0.12	1.00

Novament, s'observa que no existeix correlació fort entre cap parell de variables independents utilitzades en la definició del model. Tanmateix, a continuació s'analitzen els coeficients del model inicial i del model actualitzat buscant canvis substancials.

```
#obtenim coeficients dels dos models
c1<-data.frame(model=model_03_m_Constitucio$coefficients)
c2<-data.frame(model_updated=model_03_m_Constitucio_2$coefficients)
#ajuntem en un únic dataframe
model_03_constitucio_summary<-merge(c1, c2, by="row.names", all = "TRUE")
kable(model_03_constitucio_summary, caption = "Comparativa coeficients models Constitucio")
```

Table 9: Comparativa coeficients models Constitucio

Row.names	model	model_updated
(Intercept)	83.1057419	83.0912229
HR	-0.3523518	-0.3523878
LL	2.8464917	2.8474572
NO2	-0.7700740	-0.7699983
RS	0.0077885	0.0077747
TMP	NA	0.0011802
vv	15.1934322	15.1940778

Es pot concloure que la introducció de la variable *TMP* no produeix canvis significatius en els coeficients de les altres variables, comparant el model de regressió inicial amb l'actualitzat.

Com a darrer pas, s'analitza el factor d'inflació de la variança.

```
vif(model_03_m_Constitucio_2)
```

```
##      NO2        vv        RS        HR        LL        TMP
## 1.332209 1.760187 1.606862 1.378497 1.058712 1.297911
```

Novament, a través de l'anàlisi del factor d'inflació de la variança també podem afirmar que en aquest cas tampoc hi ha presència de colinealitat, doncs tots els factors retornats per a les diferents variables independents son relativament propers a la unitat.

Finalment, es pot concloure que la introducció de la variable *TMP* en el model de regressió múltiple per a l'estació d'Avinguda de la Constitució:

- No millora el model doncs el coeficient de determinació ajustat decreix lleugerament de 0.5902 a 0.5901.
- L'error de predicció es manté constant a 17.83 (17.83 per al model actualitzat)
- No es considera estadísticament significativa doncs el P-valor és 0.9783212

Amb les conclusions anteriors la variable *TMP* no s'hauria d'incloure en el model per a la Estació d'Avinguda de la Constitució.

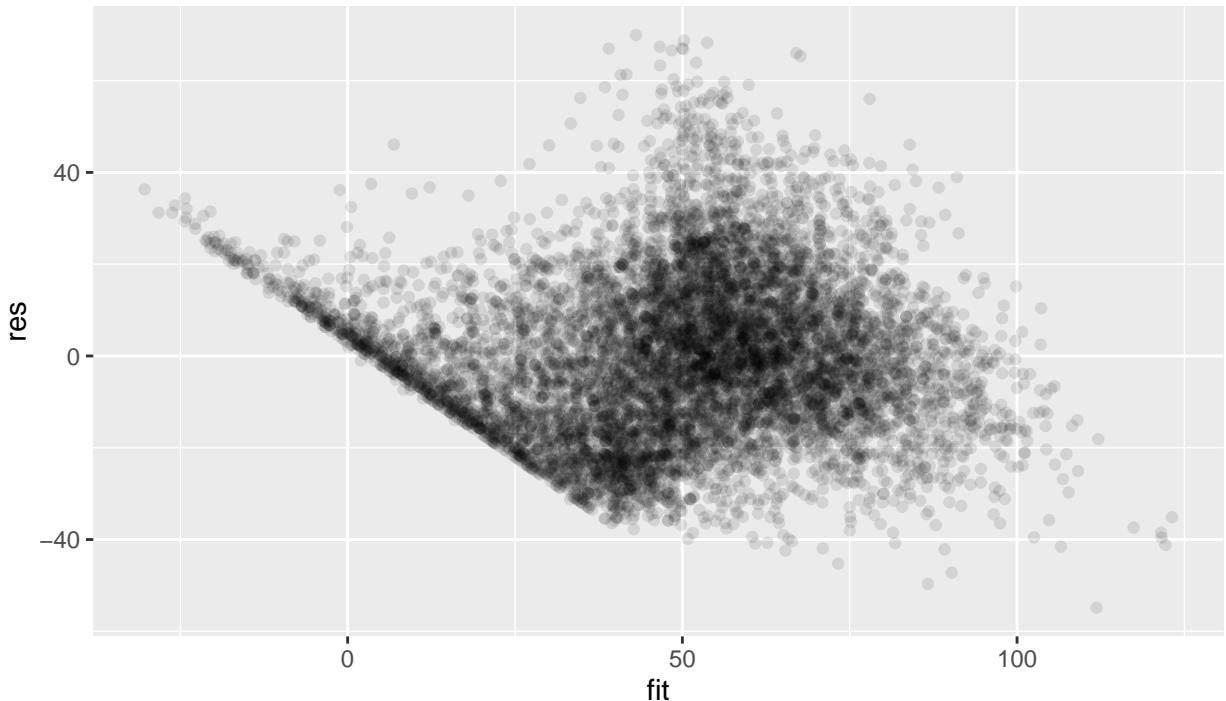
1.4. Diagnosi del model

Per a la diagnosi es tria l'últim model construït per a l'estació de Montevid i es demanen dos gràfics: un amb els valors ajustats enfront dels residus (que ens permetrà veure si la variància és constant) i el gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment (QQplot). Interpreteu els resultats.

```

model_fit<-data.frame(fit=model_03_m_Montevil_2$fitted.values,
                       res=model_03_m_Montevil_2$residuals,
                       NO2=model_03_m_Montevil_2$model$NO2,
                       vv=model_03_m_Montevil_2$model$vv,
                       RS=model_03_m_Montevil_2$model$RS,
                       HR=model_03_m_Montevil_2$model$HR,
                       LL=model_03_m_Montevil_2$model$LL,
                       TMP=model_03_m_Montevil_2$model$TMP)
ggplot(data=model_fit, aes(x=fit, y=res), color="black") +
  geom_point(alpha=1/10)

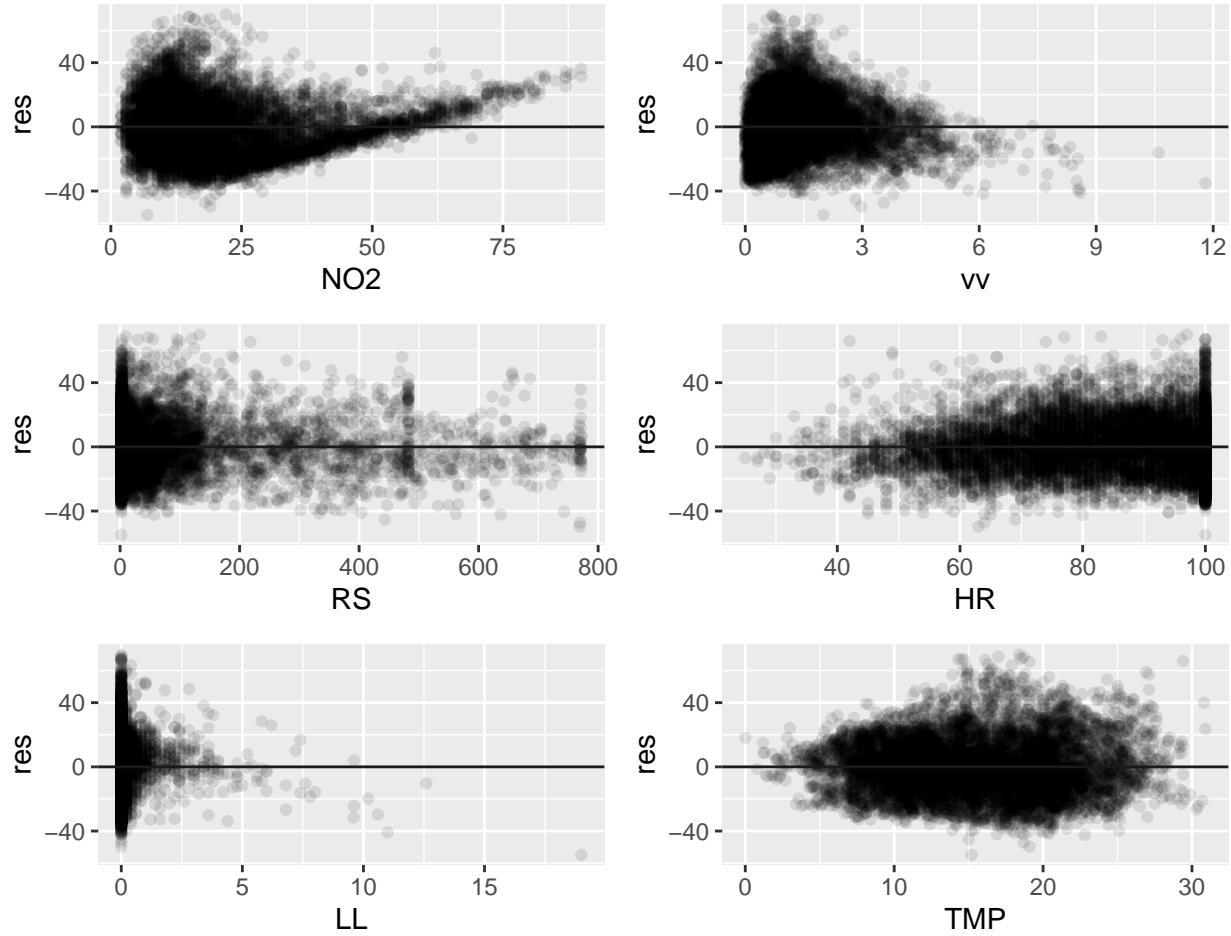
```



El gràfic de valors ajustats enfront els residus mostra dos comportaments diferenciats de les dades. Un comportament lineal d'un grup de punts entre les coordenades (-25, 30) i (40, -40) i un comportament aleatori general en la resta del gràfic.

Aquest comportament ja s'ha observat en la regressió lineal simple realitzada amb entre les variables O3~NO2 i com s'ha explicat anteriorment es deu principalment al fet que en les dades originals tenim valors de concentració de O3~0 per a tot l'interval de valors de concentració de NO2.

Per analitzar millor els residus, es procedeix a graficar-los enfront cada una de les variables independents del model i observar si existeixen relacions en la variança.

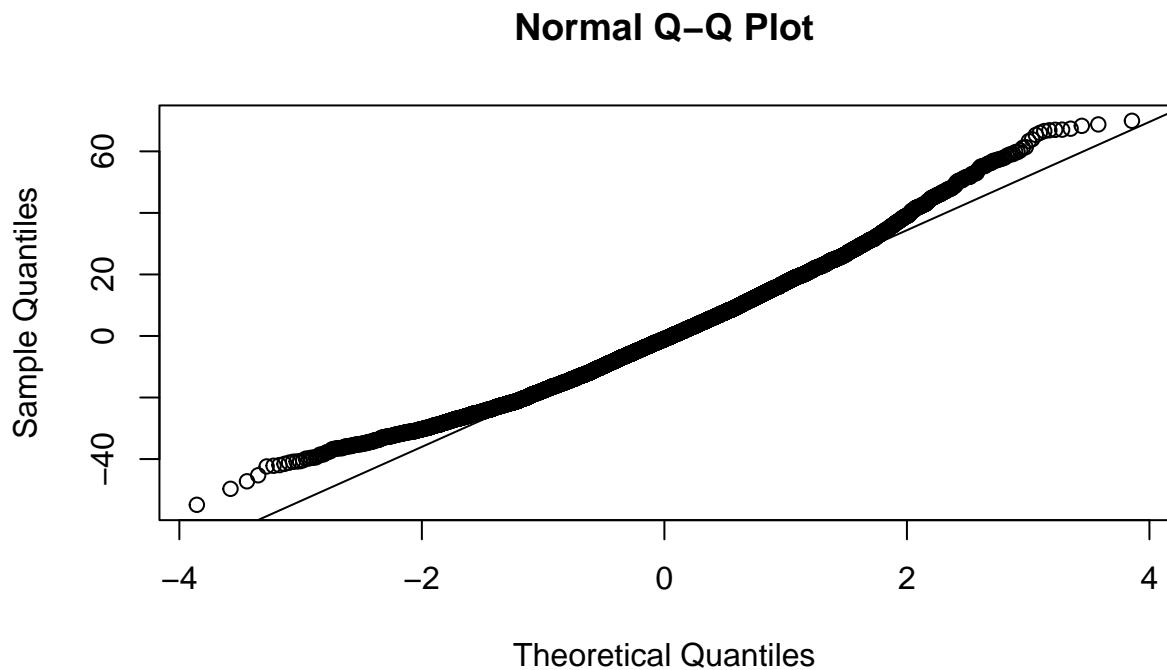


En base als gràfics anteriors s'observen distribucions aleatòries i aproximadament simètriques al voltant de l'eix d'abscisses (línia residus = 0) en els gràfics de residus enfront les variables *vv*, *RS*, *HR*, *LL* i *TMP*.

Per altra banda, s'observa una distribució no aleatòria en part del gràfic de residus enfront *NO2* que ja ha estat explicada anteriorment (veure 1.2. Model de regressió lineal simple)

Per finalitzar, el gràfic quantil-quantil de la següent pàgina mostra que els residus segueixen una distribució aproximadament normal, especialment en la zona central de dades, amb cues pesades o amb una freqüència més elevada que una distribució normal.

```
qqnorm(model_03_m_Montevil_2$residuals)
qqline(model_03_m_Montevil_2$residual)
```



1.5. Predicció del model

Segons el model de l'apartat anterior, calculeu la concentració d'O₃, si es tenen valors de NO₂ de 40, vv de 2, RS de 100, HR de 80, LL de 0.10 i TMP de 25.

```
predict(model_03_m_Montevil_2, newdata = data.frame(NO2=40, vv=2, RS=100,
                                                HR=80, LL=0.1, TMP=25))
```

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

La concentració d'O₃ en la estació de Montevil estimada pel model per als valors indicats a l'anunciat és 35.97 μm^3 .

2. Regressió logística

Per a construir les noves variables i els models de regressió logística, es prendran els valors de les variables triades per hora, tal com apareixen en la base de dades original. En aquest apartat es prendran com a contaminants la concentració de PM10 i d'O₃. Es procedirà a calcular els índexs de qualitat (icPM10 i icO₃) de la forma següent:

- PM10 re-codificada: (icPM10)
 - acceptable: valors de (0 a 45],
 - millorable: valors de (45 a 180]
- O₃ re-codificada: (icO₃)
 - acceptable: valors de (0 a 60],
 - millorable: valors de (60 a 170]
- RS re-codificada (RS_re):
 - normal_baixa:(0 a 100],
 - normal_alta: valors de (100 a 700]

Nota: Aquest índex de qualitat s'ha re-codificat conforme a les nostres dades.

```
#recodificacio de dades
Air$icPM10<-as.factor(ifelse(Air$PM10>0 & Air$PM10<=45,"acceptable",
                                ifelse(Air$PM10>45 & Air$PM10<=180,"millorable",NA)))
Air$icO3<-as.factor(ifelse(Air$O3>0 & Air$O3<=60,"acceptable",
                            ifelse(Air$O3>60 & Air$O3<=170,"millorable",NA)))
Air$RS_re<-as.factor(ifelse(Air$RS>0 & Air$RS<=100,"normal_baixa",
                              ifelse(Air$RS>100 & Air$RS<=700,"normal_alta",NA)))
```

2.1. Anàlisi cru de possibles factors de risc. Cálculo de OR

Es creerà una nova variable amb els mesos de l'any a partir de la variable Data, anomenada month.

```
Air$month<-as.factor(month(Air$Fecha))
Air_M<-Air[Air$Nombre=="Estacion de Montevil",]
```

a) Es calcularà les OR (Odds-Ràtio) entre cadascuna de les variables dependents i les variables explicatives en l'estació de Montevil. Important: Per al càlcul de les OR, es partirà de la taula de contingència i es calcularà a partir de la seva fórmula. Heu d'implementar aquesta fórmula en R. Es pot considerar que la radiació solar i el mes de l'any són factors de risc? Justifica la teva resposta i interpreta les OR.

```
odds_fun<-function(x,y){
  #x = variable explicativa // y = variable dependent

  #Taula de contingència
  tc<-table(y,x)
  tc<-cbind(tc, rowSums(tc))
  colnames(tc)[ncol(tc)]<-"sum_row"
  tc<-rbind(tc, colSums(tc))
  rownames(tc)[nrow(tc)]<-"sum_col"

  #Taula de percentatges
  tp<-tc
  for (i in 1:nrow(tc)-1){
    tp[i,]<-tc[i,]/tc[nrow(tc),]
    rownames(tp)[i]<-paste(rownames(tc)[i], "%", sep = " ")
  }
  tp<-tp[,ncol(tp)-1]
  tp<-tp[1:nrow(tp)-1,]

  #Taula odds - variable dependent
  #millorable sobre acceptable
  to<-tp
  for (i in 1:nrow(tp)-1){
    to[i,]<-tp[nrow(tp),]/tp[i,]
  }
  to<-to[1:nrow(tp)-1,]

  reference_odd<-to[1]
  #Array d'odds - variable explicativa
  #Months sobre gener // normal_baixa sobre normal_alta
  ao<-to/reference_odd
  ao<-ao[-1]
  return(ao)
}
```

```
#millorable sobre acceptable icPM10 i radiació baixa sobre alta
odds<-odds_fun(Air_M$RS_re, Air_M$icPM10)
odds
```

```
## normal_baixa
##      1.188835
```

NOTA: Per millorar la interpretació dels resultats, quan s'analitzi la variable *radiació solar* calcularem la inversa del resultat de la funció *odds_fun*. D'aquesta manera estarem analitzant la proporció de radiació solar alta, sobre radiació solar baixa, fet que millora la comprensió dels resultats.

```
#millorable sobre acceptable icPM10 i radiació alta sobre baixa
1/unname(odds)
```

```
## [1] 0.8411594
```

Si considerem com a factor la radiació solar alta, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icPM10, és de 0.84 vegades respecte al cas amb radiació solar baixa. Per tant, la radiació solar alta no és un factor de risc per a la concentració de icPM10 doncs el odd-ratio és inferior a 1.

```
#millorable sobre acceptable icPM10 i mesos sobre gener
```

```
odds<-odds_fun(Air_M$month, Air_M$icPM10)
odds
```

```
##      2      3      4      5      6      7      8      9
## 0.3516820 0.3453453 1.1941848 1.1257954 0.5227273 0.3898906 0.5182595 1.1211968
##      10     11     12
## 1.0601650 1.3130539 2.7053927
```

Quan estem al mes d'Abril, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icPM10, és de 1.19 vegades respecte al Gener, per tant la proporció és un 19.42% major al mes d'Abril que el mes de Gener.

Els mesos d'Abril, Maig, Setembre, Octubre, Novembre i Desembre es poden considerar factor de risc (comparant amb Gener) per a icPM10 doncs observem valors de *odds-ratio* superiors a la unitat.

```
#millorable sobre acceptable O3 i radiació alta sobre baixa
```

```
odds<-odds_fun(Air_M$RS_re, Air_M$icO3)
1/unname(odds)
```

```
## [1] 5.21607
```

Quan la radiació solar és alta, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icO3, és de 5.22 vegades respecte al cas amb radiació solar baixa. La radiació solar alta clarament és un factor de risc per a l'índex icO3.

```
#millorable sobre acceptable O3 i mesos sobre gener
```

```
odds<-odds_fun(Air_M$month, Air_M$icO3)
odds
```

```
##      2      3      4      5      6      7      8      9
## 2.4225943 5.7645615 5.8578161 7.2810016 2.4096680 1.8762714 1.2914707 2.7379369
##      10     11     12
## 1.7075986 0.4772135 0.3242287
```

Quan estem al mes de Maig, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icO3, és de 7.28 vegades respecte al Gener. Els mesos de febrer a Octubre son factors de risc per a icO3 (prenent com a referencia el mes de Gener).

b) Idem per a l'estació d'Avinguda Constitució.

```
Air_C<-Air[Air$Nombre=="Estacion Avenida Constitucion",]

#millorable sobre acceptable icPM10 i radiació alta sobre baixa
odds<-odds_fun(Air_C$RS_re, Air_C$icPM10)
1/unname(odds)

## [1] 0.7675727
```

Si considerem com a factor la radiació solar alta, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icPM10, és de 0.77 vegades respecte al cas amb radiació solar baixa. Per tant, la radiació solar alta no és un factor de risc en l'estació d'Avinguda Constitució per a la concentració de icPM10 doncs el *odd-ratio* és inferior a 1.

```
#millorable sobre acceptable icPM10 i mesos sobre gener
odds<-odds_fun(Air_C$month, Air_C$icPM10)
odds
```

```
##          2          3          4          5          6          7          8
## 0.91084626 0.18717099 2.18944099 0.82954339 0.31394045 0.22586110 0.04224608
##          9         10         11         12
## 0.51472022 0.64500717 1.85232827 2.79584951
```

Quan estem al mes de Desembre, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icPM10, és de 2.8 vegades respecte al Gener, per tant la proporció és un 179.58% major al mes de Desembre que el mes de Gener.

Els mesos d'Abril, Novembre i Desembre es poden considerar factor de risc, en comparació amb Gener per a icPM10 doncs observem valors de *odds-ratio* superiors a la unitat.

```
#millorable sobre acceptable O3 i radiació alta sobre baixa
odds<-odds_fun(Air_C$RS_re, Air_C$icO3)
1/unname(odds)
```

```
## [1] 2.487093
```

Quan la radiació solar és alta, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icO3, és de 2.49 vegades respecte al cas amb radiació solar baixa. La radiació solar alta és un factor de risc per a l'índex icO3.

```
#millorable sobre acceptable O3 i mesos sobre Gener
odds<-odds_fun(Air_C$month, Air_C$icO3)
odds
```

```
##          2          3          4          5          6          7          8
## 3.3519375 8.4539521 11.1618472 13.2265372 3.7453416 1.9248814 3.1414862
##          9         10         11         12
## 8.8143664 4.2270706 1.0075188 0.9670983
```

Quan estem al mes d'Abril, la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icO3, és de 11.16 vegades respecte al Gener. Els mesos de febrer a Octubre son factors de risc per a icO3 (prenen com a referència el mes de Gener) doncs tots mostren un *odd-ratio* superior a la unitat.

2.2. Model de regressió logística

Per a l'estació de Montevil de l'apartat anterior:

a) Es demana construir un model de regressió logística prenent com a variable dependent *icPM10* i variables explicatives (*RS_re*), (*vv*) i (*PRB*). Interpreteu i calculeu les OR.

Inicialment es comproven els nivells assignats a les variables categòriques:

```
#comprovació de nivells  
contrasts(Air_M$icPM10)
```

```
##           millorable  
## acceptable      0  
## millorable      1
```

```
contrasts(Air_M$ic03)
```

```
##           millorable  
## acceptable      0  
## millorable      1
```

```
contrasts(Air_M$RS_re)
```

```
##           normal_baixa  
## normal_alta      0  
## normal_baixa      1
```

```
contrasts(Air_M$month)
```

```
##   2 3 4 5 6 7 8 9 10 11 12  
## 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 1 0 0 0 0 0 0 0 0 0 0  
## 3 0 1 0 0 0 0 0 0 0 0 0  
## 4 0 0 1 0 0 0 0 0 0 0 0  
## 5 0 0 0 1 0 0 0 0 0 0 0  
## 6 0 0 0 0 1 0 0 0 0 0 0  
## 7 0 0 0 0 0 1 0 0 0 0 0  
## 8 0 0 0 0 0 0 1 0 0 0 0  
## 9 0 0 0 0 0 0 0 1 0 0 0  
## 10 0 0 0 0 0 0 0 0 1 0 0  
## 11 0 0 0 0 0 0 0 0 0 1 0  
## 12 0 0 0 0 0 0 0 0 0 0 1
```

La definició del model es realitza d'acord al següent:

```
model_logist_icPM10<-glm(formula = icPM10~RS_re+vv+PRB, data = Air_M,  
                           family = binomial(link = logit))  
summary(model_logist_icPM10)  
  
##  
## Call:  
## glm(formula = icPM10 ~ RS_re + vv + PRB, family = binomial(link = logit),  
##       data = Air_M)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -0.7573  -0.4331  -0.3618  -0.2676   4.4018  
##  
## Coefficients:
```

```

##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -36.605380   5.840926 -6.267 3.68e-10 ***
## RS_renornal_baixa  -0.319881   0.123158 -2.597  0.0094 **
## vv                  -0.576894   0.060108 -9.598 < 2e-16 ***
## PRB                 0.034458   0.005757  5.985 2.16e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4319.8 on 8590 degrees of freedom
## Residual deviance: 4150.6 on 8587 degrees of freedom
## (169 observations deleted due to missingness)
## AIC: 4158.6
##
## Number of Fisher Scoring iterations: 6

```

A partir dels signes dels coeficients podem dir que quan la *RS_re* és normal_baixa (normal_baixa=1) i/o quan incrementa la velocitat del vent *vv* la probabilitat de tenir un dia millorable (millorable=1) en icPM10 es redueix doncs el signe del coeficient és negatiu, per altra banda, quan la pressió atmosfèrica *PRB* incrementa, la probabilitat d'obtenir un dia millorable també ho fa.

L'estimació dels OR es realitza calculant l'exponencial dels coeficients obtinguts del model de regressió logistica.

```
round(exp(coefficients(model_logist_icPM10)),4)
```

	(Intercept)	RS_renornal_baixa	vv	PRB
##	0.0000	0.7262	0.5616	1.0351

Així per la variable *RS_re* es té un OR de 0.7262 fet que indica que la probabilitat de trobar un dia millorable, sobre un dia acceptable, pel que fa a icPM10 és de 0.73 vegades respecte al cas amb radiació solar alta, afirmació alineada amb l'explicació anterior relacionada amb el signe dels coeficients de la regressió. Pel que fa a la velocitat del vent es pot concloure que per cada unitat que incrementa la velocitat del vent l'odds d'obtenir un dia millorable és 0.56 vegades menor. D'aquesta manera, si la variable vent augmenta en 3 unitats l'odds serà 0.18 vegades menor.

Com que icPM10 = millorable és pitjor (en termes de contaminació) podem resumir que la Radiació Solar baixa i l'increment del vent redueixen les probabilitats de que el dia sigui millorable, conseqüentment incrementen les probabilitats de que el dia sigui acceptable (menys contaminació).

b) S'afegeix al model de l'apartat anterior la variable (month). ¿Existeix una millora del model?. Justifiqueu i interpreteu.

```

model_logist_icPM10_b<-update(model_logist_icPM10, .~.+month, data = Air_M,
                                 family = binomial(link = logit))
summary(model_logist_icPM10_b)

##
## Call:
## glm(formula = icPM10 ~ RS_re + vv + PRB + month, family = binomial(link = logit),
##      data = Air_M)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.8233 -0.4295 -0.3181 -0.2326  4.3331
##
## Coefficients:

```

```

##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -23.697983   6.683215 -3.546 0.000391 ***
## RS_renormal_baixa  -0.541047   0.138300 -3.912 9.15e-05 ***
## vv                  -0.524236   0.061028 -8.590 < 2e-16 ***
## PRB                 0.021976   0.006542  3.359 0.000782 ***
## month2              -0.977476   0.281175 -3.476 0.000508 ***
## month3              -0.675468   0.291214 -2.319 0.020369 *
## month4              0.054912   0.215500  0.255 0.798868
## month5              -0.137286   0.208211 -0.659 0.509666
## month6              -0.759402   0.248728 -3.053 0.002265 **
## month7              -1.042109   0.264020 -3.947 7.91e-05 ***
## month8              -0.737021   0.240104 -3.070 0.002144 **
## month9              -0.018988   0.201249 -0.094 0.924829
## month10             0.020468   0.206402  0.099 0.921007
## month11             0.265336   0.209059  1.269 0.204373
## month12             0.754390   0.177076  4.260 2.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4319.8 on 8590 degrees of freedom
## Residual deviance: 4010.0 on 8576 degrees of freedom
## (169 observations deleted due to missingness)
## AIC: 4040
##
## Number of Fisher Scoring iterations: 6

```

L'AIC del model inicial és 4158.64 mentre que l'AIC del model actualitzat amb la variable *month* és 4040.02, per tant com que l'AIC disminueix podem afirmar que la introducció de la variable *month* millora el model.

Cal destacar que no totes les variables *dummy* generades a partir de la variable *month* són estadísticament significatives. Observant el valor de l'estadístic de Wald i el p-valor associat, podem afirmar que les variables *dummy month4, month5, month9, month10 i month11* no són estadísticament significatives.

Conseqüentment les dades registrades en els mesos citats no produeixen canvis significatius en el resultat del model per efecte de la estacionalitat..

c) *S'afegirà al model anterior com a variable explicativa la variable (TMP). Justifiqueu la presència o no d'una possible interacció amb (RS_re). Es podria estar davant una variable de confusió?. Raona la teva resposta.*

```

model_logist_icPM10_c<-update(model_logist_icPM10_b, .~.+TMP, data = Air_M,
                                 family = binomial(link = logit))
summary(model_logist_icPM10_c)

##
## Call:
## glm(formula = icPM10 ~ RS_re + vv + PRB + month + TMP, family = binomial(link = logit),
##      data = Air_M)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.1745  -0.4175  -0.3054  -0.2088   4.5513
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)

```

```

## (Intercept)      -31.081012   7.109828  -4.372 1.23e-05 ***
## RS_renormal_baixa -0.127501   0.143235  -0.890 0.373386
## vv              -0.679328   0.065607 -10.355 < 2e-16 ***
## PRB             0.027444   0.006937   3.956 7.61e-05 ***
## month2          -0.696578   0.285003  -2.444 0.014521 *
## month3          -0.536599   0.297425  -1.804 0.071208 .
## month4          -0.210656   0.219409  -0.960 0.337002
## month5          -0.622822   0.215863  -2.885 0.003911 **
## month6          -1.682838   0.264098  -6.372 1.87e-10 ***
## month7          -2.501829   0.297785  -8.401 < 2e-16 ***
## month8          -2.237429   0.280543  -7.975 1.52e-15 ***
## month9          -1.389952   0.240541  -5.778 7.54e-09 ***
## month10         -0.759848   0.223515  -3.400 0.000675 ***
## month11         -0.105645   0.216296  -0.488 0.625246
## month12         0.407189   0.182604   2.230 0.025754 *
## TMP              0.143061   0.013886  10.303 < 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: 4319.8 on 8590 degrees of freedom
## Residual deviance: 3903.0 on 8575 degrees of freedom
## (169 observations deleted due to missingness)
## AIC: 3935
##
## Number of Fisher Scoring iterations: 6

```

Per analitzar la possible presència d'interacció es crea un model adicional afegint la interacció entre les variables *TMP* i *RS_re*.

```

model_logist_icPM10_c_int<-update(model_logist_icPM10_c, .~.+RS_re:TMP, data = Air_M,
                                      family = binomial(link = logit))
summary(model_logist_icPM10_c_int)

```

```

##
## Call:
## glm(formula = icPM10 ~ RS_re + vv + PRB + month + TMP + RS_re:TMP,
##      family = binomial(link = logit), data = Air_M)
##
## Deviance Residuals:
##    Min      1Q      Median      3Q      Max
## -1.1629  -0.4173  -0.3053  -0.2085   4.5511
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -31.506046   7.123706 -4.423 9.75e-06 ***
## RS_renormal_baixa     0.326166   0.497542   0.656 0.512110
## vv                   -0.680998   0.065767 -10.355 < 2e-16 ***
## PRB                  0.027476   0.006936   3.961 7.45e-05 ***
## month2                -0.688972   0.285088  -2.417 0.015662 *
## month3                -0.523158   0.297674  -1.757 0.078835 .
## month4                -0.207235   0.219412  -0.945 0.344913
## month5                -0.628220   0.216006  -2.908 0.003633 **
## month6                -1.703365   0.264996  -6.428 1.29e-10 ***

```

```

## month7          -2.517265  0.298345 -8.437 < 2e-16 ***
## month8          -2.252664  0.281031 -8.016 1.10e-15 ***
## month9          -1.398486  0.240700 -5.810 6.24e-09 ***
## month10         -0.763308  0.223175 -3.420 0.000626 ***
## month11         -0.118217  0.216321 -0.546 0.584730
## month12          0.392126  0.182924  2.144 0.032061 *
## TMP              0.164786  0.026620  6.190 6.01e-10 ***
## RS_renormal_baixa:TMP -0.025027  0.026046 -0.961 0.336608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4319.8 on 8590 degrees of freedom
## Residual deviance: 3902.0 on 8574 degrees of freedom
## (169 observations deleted due to missingness)
## AIC: 3936
##
## Number of Fisher Scoring iterations: 6

```

S'observa que el terme d'interacció *RS_renormal_baixa:TMP* no és estadísticament significatiu doncs el p-valor associat al estadístic de wald és 0.3366.

Per analitzar si ens trobem davant una variable de confusió es comparen els paràmetres estimats dels dos models anteriors (amb i sense la variable *TMP*, sense el terme d'interacció).

```

#obtenim coeficients dels dos models
c1<-data.frame(model=model_logist_icPM10_b$coefficients)
c2<-data.frame(model_updated=model_logist_icPM10_c$coefficients)
#ajuntem en un únic dataframe
model_logist_icPM10_summary<-merge(c1, c2, by="row.names", all = "TRUE")
kable(model_logist_icPM10_summary,
      caption = "Comparativa paràmetres estimats regressió logística")

```

Table 10: Comparativa paràmetres estimats regressió logística

Row.names	model	model_updated
(Intercept)	-23.6979828	-31.0810120
month10	0.0204679	-0.7598480
month11	0.2653356	-0.1056454
month12	0.7543902	0.4071894
month2	-0.9774757	-0.6965780
month3	-0.6754677	-0.5365987
month4	0.0549120	-0.2106556
month5	-0.1372855	-0.6228220
month6	-0.7594022	-1.6828383
month7	-1.0421089	-2.5018292
month8	-0.7370215	-2.2374286
month9	-0.0189883	-1.3899515
PRB	0.0219762	0.0274442
RS_renormal_baixa	-0.5410471	-0.1275007
TMP	NA	0.1430612
vv	-0.5242362	-0.6793276

S'observa que la introducció de la variable TMP modifica el paràmetre estimat per $RS_renormal$ baixa doncs en el model inicial era -0.54 i en el model actualitzat amb la variable TMP és de -0.13. També s'observen canvis en els paràmetres estimats per algunes de les variables *dummy* de *month*.

Per tant es pot afirmar que la variable TMP és una variable de confusió.

2.3. Predicció

Segons el model de l'apartat b), calculeu la probabilitat que la concentració de PM10 sigui o no superior a 45, amb uns valors de $vv= 0.6$, $RS_re="Normal_alta"$, $PRB= 1013$, el mes d'agost.

```
predict(model_logist_icPM10_b , newdata = data.frame(vv=0.6, RS_re="normal_alta",
                                                       PRB=1013,month="8"),
        type="response")
```

```
##           1
## 0.07673163
```

La probabilitat que l'índex icPM10 sigui “millorable” (concentració $PM10 \geq 45$) amb les dades proporcionades a l'enunciat és del 7.67%.

2.4. Bondat de l'ajust

Fes servir el test de Hosman-Lemeshow per veure la bondat d'ajust, prenent el model de l'apartat b). El paquet ResourceSelection té la funció que aplica el test de Hosmer-Lemeshow.

```
hoslem.test(model_logist_icPM10_b$y, model_logist_icPM10_b$fitted.values)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_logist_icPM10_b$y, model_logist_icPM10_b$fitted.values
## X-squared = 33.447, df = 8, p-value = 5.116e-05
```

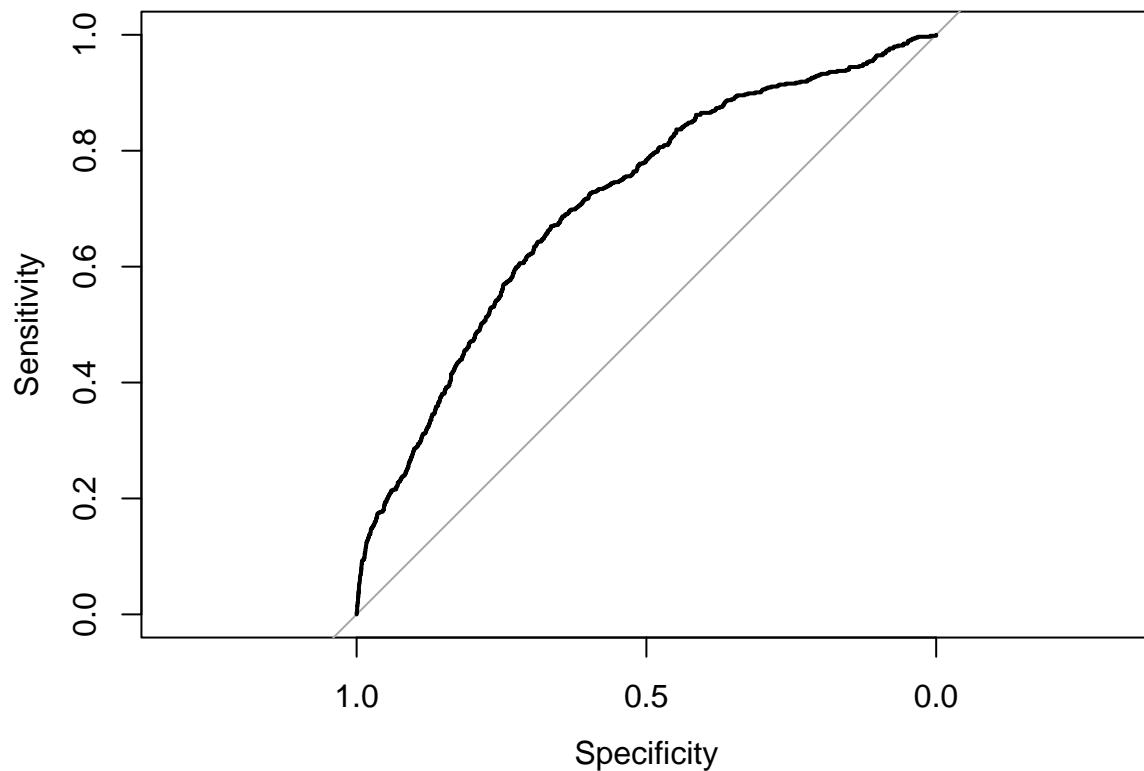
El test d'Hosmer-Lemeshow realitza un contrast d'hipòtesis on H_0 indica que no hi ha diferències entre els valors observats i els valors pronosticats. Observant els resultats del test s'obté un p-valor de 5.1158156×10^{-5} , molt petit. Conseqüentment es rebutja l'hipòtesis nul · la i s'affirma que si existeixen diferències entre els valors observats i els valors pronosticats pel model, per tant el model no està correctament ajustat.

2.5. Corba ROC

Dibuixeu la corba ROC, i calculeu l'àrea sota la corba amb el model de l'apartat b). Discutir el resultat.

```
prob<-predict(model_logist_icPM10_b, Air_M, type = "response")
r<-roc(Air_M$icPM10, prob)
```

```
## Setting levels: control = acceptable, case = millorable
## Setting direction: controls < cases
plot(r)
```



```
auc(r)
```

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

Amb els resultats d'àrea sota la corba és pot afirmar que el model té unes prestacions correctes, sense arribar a ser excel·lents.

3. Conclusions de l'anàlisi

En aquest apartat s'han d'exposar les conclusions sobre la base dels resultats obtinguts a tot l'estudi. Regressió lineal i logística

ANÀLISI DE DADES

L'anàlisi inicial de les dades mostra que existeixen problemes **greus** de compliment amb els límits per als contaminants:

- NO₂ en el valor de la mitja diària
- O₃ en el valor de la mitja diària
- O₃ en el valor màxim diari

i problemes **moderats** en el compliment per als contaminants:

- PM10 en el valor de la mitja diària

També mostra que per a l'estació de Montevil existeixen certs nivells de correlació positiva entre les variables:

- PM10, SO₂ i NO₂
- O₃ i RS

Així com certs nivells de correlació negativa entre les variables:

- PM10 i LL
- NO₂, TMP i HR
- O₃ i PRB

Mentre per l'estació d'Avinguda de la Constitució, existeixen certs nivells de correlació positiva entre les variables:

- PM10, SO₂ i NO₂
- O₃, RS i vv

Així com certs nivells de correlació negativa entre les variables:

- PM10 i vv
- NO₂, TMP, HR i RS

REGRESSIÓ LINEAL

Realitzant regressió lineal simple entre les variables O₃~NO₂ s'obté un coeficient de determinació de 0.4206 i quan s'afegeix la variable *nom* el coeficient de determinació és de 0.4265 fet que indica que la introducció de la variable *nom* millora el model, però no de manera notable.

S'observa que la distribució dels residus en front dels valors ajustats no segueix un comportament aleatori, causat pel fet que no existeix una forta correlació entre les variables O₃ i NO₂ en les variables originals. És especialment destacable el fet de tenir concentracions properes a 0 en O₃, per a tot l'interval de concentracions de NO₂.

REGESSIÓ LINEAL MULTIPLE

Comparant els dos models de regressió lineal múltiple per a l'estació de Montevil es pot conoure que la introducció de la variable *TMP* en el model:

- Millora lleugerament el model doncs el coeficient de determinació ajustat creix de 0.6107 a 0.6124.
- Redueix lleugerament l'error de predicció de 17.47 a 17.43
- Es considera estadísticament significativa doncs el P-valor per a la variable *TMP* és $2.5888754 \times 10^{-10}$
- No presenta un alt grau de correlació amb altres variables independents utilitzades en el model.
- No presenta colinealitat

Amb les conclusions anteriors la variable *TMP* es podria incloure en el model per a la Estació de Montevil.

Realitzant la mateixa comparació per a l'estació d'Avinguda de la Constitució s'obté que la introducció de la variable *TMP* en el model:

- No millora el model doncs el coeficient de determinació ajustat decreix lleugerament de 0.5902 a 0.5901.
- L'error de predicció es manté constant a 17.83 (17.83 per al model actualitzat)
- No es considera estadísticament significativa doncs el P-valor és 0.9783212

Amb les conclusions anteriors la variable *TMP* no s'hauria d'incloure en el model per a la Estació d'Avinguda de la Constitució.

Analitzant en profunditat les distribucions dels gràfics de residus per al model de regressió lineal múltiple (actualitzat amb la variable *TMP*) per a l'estació de Montevil s'observen:

- Distribucions aleatòries i aproximadament simètriques al voltant de l'eix d'abscisses (línia residus = 0) en els gràfics de residus enfront les variables vv, RS, HR, LL i TMP.
- Una distribució no aleatòria en el gràfic de residus enfront la variable NO₂, prèviament explicat (REGRESSIÓ LINEAL).

Finalment, el gràfic quantil-quantil mostra que els residus segueixen una distribució aproximadament normal, especialment en la zona central de dades, una freqüència més elevada en ambdues cues.

REGRESIÓ LOGISTICA

L'anàlisi cru indica:

- Estació de Montevil
 - La radiació solar alta no és un factor de risc per a icPM10 (odds = 0.84).
 - Els mesos d'Abril, Maig, Setembre, Octubre, Novembre i Desembre es poden considerar factor de risc per a icPM10 prenent com a referència Gener.
 - La radiació solar alta és un factor de risc per a icO3 (odds = 5.22).
 - Els mesos de febrer a Octubre son factors de risc per a icO3 prenent com a referència Gener.
- Estació d'Avinguda de la Constitució
 - La radiació solar alta no és un factor de risc per a icPM10 (odds = 0.77).
 - Els mesos d'Abril, Novembre i Desembre es poden considerar factor de risc, en comparació amb Gener per a icPM10.
 - La radiació solar alta és un factor de risc per a l'índex icO3 (odds = 2.49).
 - Els mesos de febrer a Octubre son factors de risc per a icO3 (prenent com a referència el mes de Gener)

L'anàlisi dels odds-ratio a partir del primer model de regressió logística (icPM10~RS_re + vv + PRB) per a l'estació de Montevil ens indica:

- Quan la radiació solar es baixa, la probabilitat de trobar un dia millorable sobre un dia acceptable, pel que fa a icPM10 és de 0.73 vegades respecte al cas amb radiació solar alta.
- Per cada unitat que incrementa la velocitat del vent la probabilitat de tenir un dia millorable és 0.56 vegades menor.

Es pot considerar que els odds obtinguts del model entren en contradicció amb l'exposat a partir de l'anàlisi cru, però l'introducció en el model d'altres variables explicatives (*vv* i *PRB*) fa que no es puguin comparar directament els odds per *RS_re* obtinguts en el model i en l'anàlisi cru.

Quan s'afegeix la variable *month* en el model per a considerar l'estacionalitat s'observa que el model millora ja que l'AIC del model inicial era 4158.64 i l'AIC del model actualitzat amb la variable *month* és 4040.02.

Quan s'afegeix la variable *TMP* s'observa un canvi en els paràmetres estimats del model, però quan es considera la interacció *RS_re* *baixa*:*TMP* s'obté que aquest terme no és estadísticament significatiu, per tant es pot afirmar que la variable *TMP* és una variable de confusió.

El test de Hosmer-Lemeshow per a la bondat de l'ajust dona un p-valor molt petit, fet que obliga a rebutjar la hipòtesi nula i afirmar que existeixen diferències entre els valors observats i els valors pronosticats pel model, conseqüentment el model no està correctament ajustat.

Finalment, l'àrea sota la corba ROC és de 0.71 que ens indica que el model té unes prestacions correctes sense arribar a ser excel·lents.