Contenido

Alumno: Rodrigo Jeldres Carrasco

Control N3

Variable respuesta PM2.5

El objetivo es entender y explicar el comportamiento de los niveles de contaminación del aire en la Región Metropolitana. Para ello, desde el Sistema de Información Nacional de Calidad del Aire (sinca.mma.gob.cl), se seleccionó una muestra de la información histórica de la Estación Parque O'Higgins en Santiago, la cual ha sido almacenada en la base ***Contam.xlsx.*** Se dispone de las siguientes variables:

- **PM2.5** – Materia particulada de 2.5 mg/m3. El PM2.5 son partículas muy pequeñas suspendidas en el aire que tienen un diámetro de menos de 2.5 micras. La materia particulada incluye sustancias químicas orgánicas, polvo, hollín y metales. Es nuestra variable respuesta.
- Potenciales variables explicativas:

Variables meteorológicas	**Contaminantes atmosféricos**
Viento – Velocidad del viento (m/s)	NO – Monóxido de nitrógeno (ppb)
TProm – Temperatura promedio (° Celsius)	NO2 – Dióxido de nitrógeno (ppb)
TMin – Temperatura mínima (° Celsius)	CO – Monóxido de carbono (ppm)
TMax – Temperatura máxima (° Celsius)	O3 – Ozono (ppb)
Humed – Humedad relativa del aire (%)	

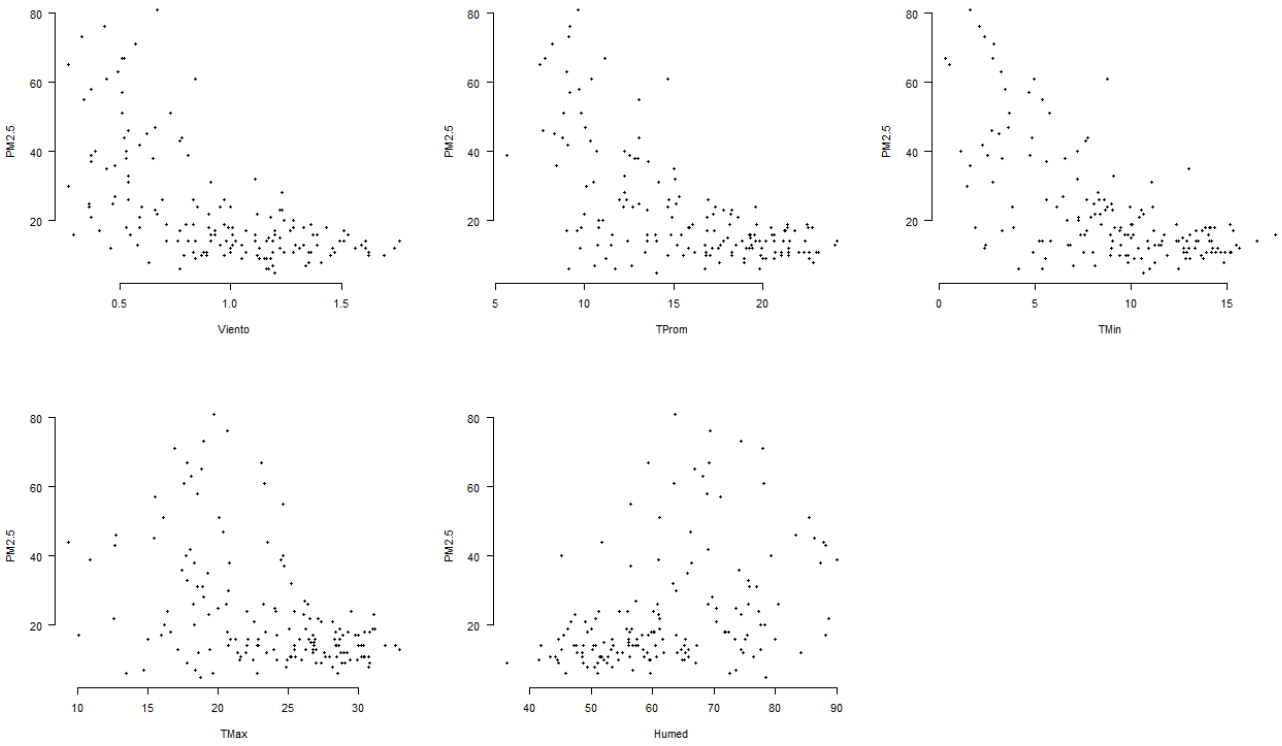
1. Obtenga el mejor modelo de regresión lineal simple basado en las variables meteorológicas.

Primero generar toda la selección de variable a explicar PM2.5 contra las **variables meteorológicas**

```
> setwd("D:/dev/Estadistica/Bases de datos/")
> base <- readxl::read_excel("Contam.xlsx")
> head(base)
# A tibble: 6 × 10
  PM2.5 Viento TProm  TMin  TMax Humed    NO   NO2    CO    O3
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    16   1.53 14.8   8.91  22.1  56.1  25.2  19.8  0.74    21
2    65   0.27  7.49  0.54  18.8  66.9 102.   43.6  2.18     8
3     7   1.34 12.6   6.98  18.4  56.8   3.15 10.5  0.52    24
4    11   1.24 22.3  15.2   28.4  43.4  11.0  14.7  0.52    22
5    39   0.37 12.5   4.71  24.5  61.0 140.   49.4  1.88     5
6    12   1    21.5  15.7   29.0  55.1  13.7  17.3  0.61    26
```

En este caso es Viento, TProm, TMin, TMax, Humed

Gráficamente:



Modelos para las distintas variables meteorológicas:

1. **Viento**

```
> Modelo1 <- lm(PM2.5 ~ Viento, data = base)
> summary(Modelo1)

Call:
lm(formula = PM2.5 ~ Viento, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-24.271  -8.885  -2.062   6.343  50.900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.033      2.865  16.767   <2e-16 ***
Viento       -26.767      2.822  -9.486   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.17 on 158 degrees of freedom
Multiple R-squared:  0.3628,    Adjusted R-squared:  0.3588
F-statistic: 89.98 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.3628

2. **TProm**

```
> Modelo2 <- lm(PM2.5 ~ TProm, data = base)
> summary(Modelo2)

Call:
lm(formula = PM2.5 ~ TProm, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-31.632  -6.993  -0.014   5.711  44.483

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.000      3.671    15.8   <2e-16 ***
TProm         -2.231      0.223   -10.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.91 on 158 degrees of freedom
Multiple R-squared:  0.3878,    Adjusted R-squared:  0.3839
F-statistic: 100.1 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.3878

3. **TMin**

```
> Modelo3 <- lm(PM2.5 ~ TMin, data = base)
> summary(Modelo3)

Call:
lm(formula = PM2.5 ~ TMin, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-29.089  -7.363  -0.288   5.466  39.578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.4178     2.4680    18.4   <2e-16 ***
TMin         -2.5131     0.2489   -10.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.87 on 158 degrees of freedom
Multiple R-squared:  0.3921,    Adjusted R-squared:  0.3883
F-statistic: 101.9 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.3921

4. **TMax**

```
> Modelo4 <- lm(PM2.5 ~ TMax, data = base)
> summary(Modelo4)

Call:
lm(formula = PM2.5 ~ TMax, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-31.212  -8.989  -2.195   4.068  52.602

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.2855     5.3891  10.444  < 2e-16 ***
TMax         -1.4170     0.2221  -6.379 1.88e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.72 on 158 degrees of freedom
Multiple R-squared:  0.2048,    Adjusted R-squared:  0.1998
F-statistic: 40.69 on 1 and 158 DF,  p-value: 1.878e-09
```

Multiple R-squared: 0.2048

5. **Humed**

```
> Modelo5 <- lm(PM2.5 ~ Humed, data = base)
> summary(Modelo5)

Call:
lm(formula = PM2.5 ~ Humed, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-26.538  -8.628  -3.668   4.124  57.430

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.7739     6.4138  -1.680    0.095 .
Humed         0.5398     0.1015   5.316 3.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.2 on 158 degrees of freedom
Multiple R-squared:  0.1517,    Adjusted R-squared:  0.1464
F-statistic: 28.26 on 1 and 158 DF,  p-value: 3.562e-07
```

Multiple R-squared: 0.1517

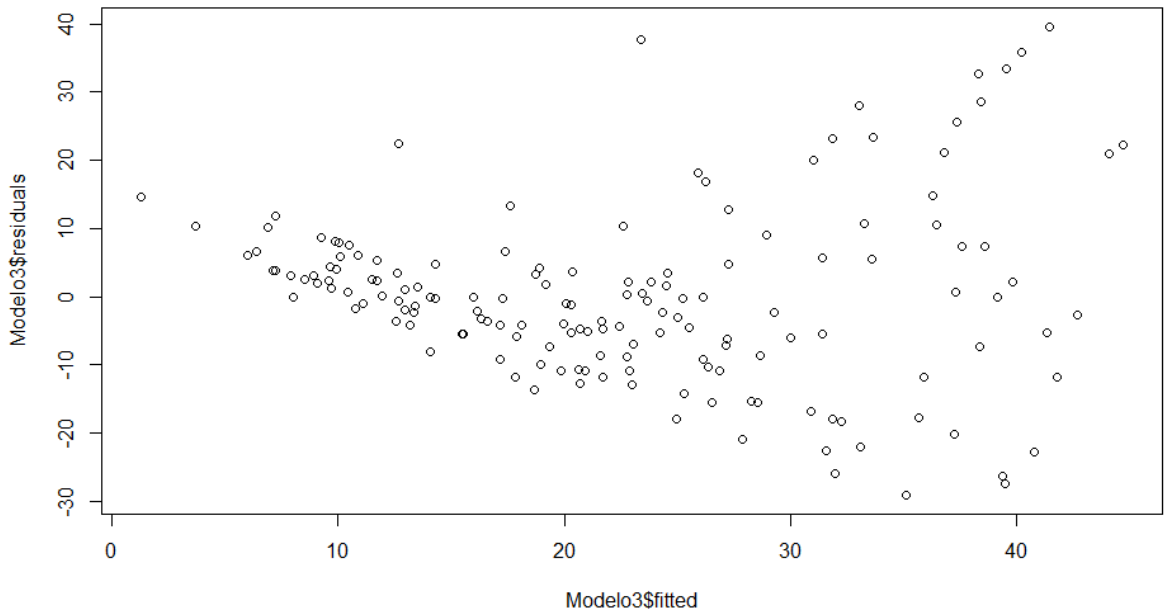
Resp: el mejor modelo simple sin transformación con variable meteorológica es presentado con Modelo3 usando "**Tmin**" dado que tiene la mayor variable explicativa que es 39,21% (en R da con decimales: 0.3921017)

Modelo3 <- lm(PM2.5 ~ TMin, data = base)

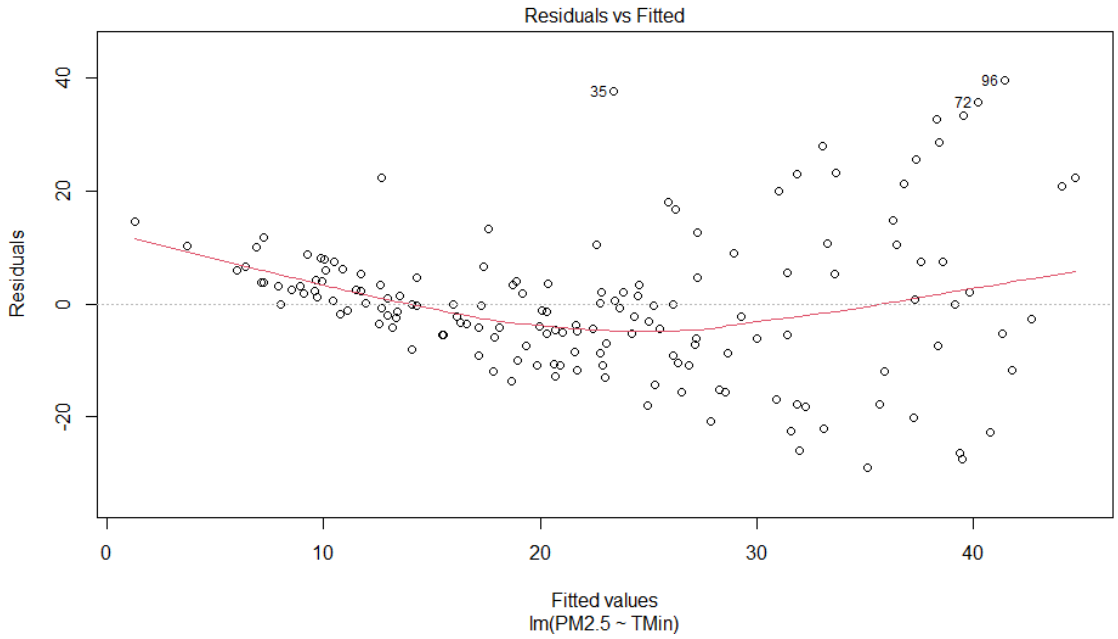
Validación de los supuestos para el "Modelo3"

1. **Linealidad**

plot(Modelo3$fitted, Modelo3$residuals)



plot(Modelo3, 1)



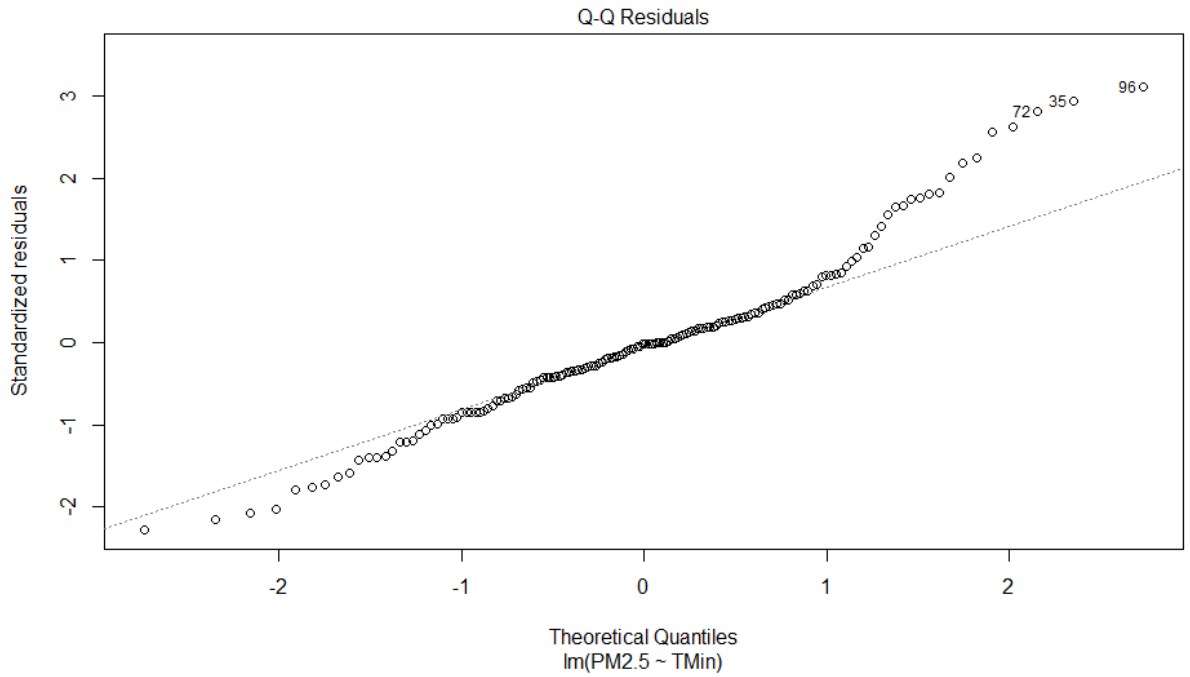
Resp: Cumple linealidad

2. Normalidad

H0: residuos distribuyen normales
H1: residuos no distribuyen normales

Gráficamente:

plot(Modelo3, 2)



> nortest::lillie.test(Modelo3$residuals)

data: Modelo3$residuals

D = 0.093753, p-value = 0.001558

Para un alfa < 0.05, se rechaza H0, es decir, no existe normalidad en los residuos según lillie.test

Test Ks

```
> ks.test(Modelo3$residuals, "pnorm", mean(Modelo3$residuals), sd(Modelo3$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Modelo3$residuals
D = 0.093753, p-value = 0.1201
alternative hypothesis: two-sided
```

D = 0.093753, p-value = 0.1201, SÍ existe normalidad según Test KS

Resp: no hay normalidad en los residuos

3. **Homocedasticidad**

H0: Sí existe Homocedasticidad
H1: No existe Homocedasticidad

```
> lmtest::bptest(Modelo3)

        studentized Breusch-Pagan test

data:  Modelo3
BP = 41.966, df = 1, p-value = 9.285e-11
```

BP = 41.966, df = 1, p-value = 9.285e-11 < alfa = 0.05

Resp: Se rechaza H0, no existe homocedasticidad

4. Independencia

H0: no hay autocorrelación vs H1: hay autocorrelación

```
> lmtest::dwtest(Modelo3)

        Durbin-Watson test

data:  Modelo3
DW = 2.039, p-value = 0.5968
alternative hypothesis: true autocorrelation is greater than 0
```

DW = 2.039, p-value = 0.5968 > alfa = 0.05

Resp: se acepta H0, existe independencia

Explicación del Modelo3

```
> summary(Modelo3)

Call:
lm(formula = PM2.5 ~ TMin, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-29.089  -7.363  -0.288   5.466  39.578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.4178     2.4680    18.4   <2e-16 ***
TMin         -2.5131     0.2489   -10.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.87 on 158 degrees of freedom
Multiple R-squared:  0.3921,    Adjusted R-squared:  0.3883
F-statistic: 101.9 on 1 and 158 DF,  p-value: < 2.2e-16
```

Para el Beta1: B1 = - 2.5131

Modelo3 quedaría con:

Y = B0 + B1 * X

PM2.5 = 45,42 – 2,51 * TMin

Test de significancia del B1

H0: **B1 = 0**
H1: **B1 distinto de 0**

Salida R:

t-value = -10.1 // p-value: <2e-16 < alfa = 0.05, se rechaza H0

Para un alfa tan pequeño, la variable **TMin** es significativa

Test de significancia del modelo

H0: no existe regresión vs H1: existe regresión

```
F-statistic: 101.9 on 1 and 158 DF,  p-value: < 2.2e-16
```

< alfa = 0.05, se rechaza H0, es decir, existe regresión

Multiple R-squared: 0.3921: el modelo explica en un 39,21% el nivel de contaminación PM2.5

Test para ro usando cor.test

H0: ro = 0
H1: ro distinto de 0

```
# Correlación entre PM2.5 y Viento
cor.test(base$PM2.5, base$Viento)
# t = -9.4857, p-value < 2.2e-16

# Correlación entre PM2.5 y TProm
cor.test(base$PM2.5, base$TProm)
# t = -10.004, p-value < 2.2e-16

# Correlación entre PM2.5 y TMin
cor.test(base$PM2.5, base$TMin)
# t = -10.095, p-value < 2.2e-16

# Correlación entre PM2.5 y TMax
cor.test(base$PM2.5, base$TMax)
# t = -6.3791, p-value = 1.878e-09

# Correlación entre PM2.5 y Humed
cor.test(base$PM2.5, base$Humed)
# t = 5.3161, p-value = 3.562e-07
```

Interpretación:

PM2.5 y Viento: Correlación negativa significativa (t = -9.4857, p-value < 2.2e-16).

PM2.5 y TProm: Correlación negativa significativa (t = -10.004, p-value < 2.2e-16).

PM2.5 y TMin: Correlación negativa significativa (t = -10.095, p-value < 2.2e-16).

PM2.5 y TMax: Correlación negativa significativa (t = -6.3791, p-value = 1.878e-09).

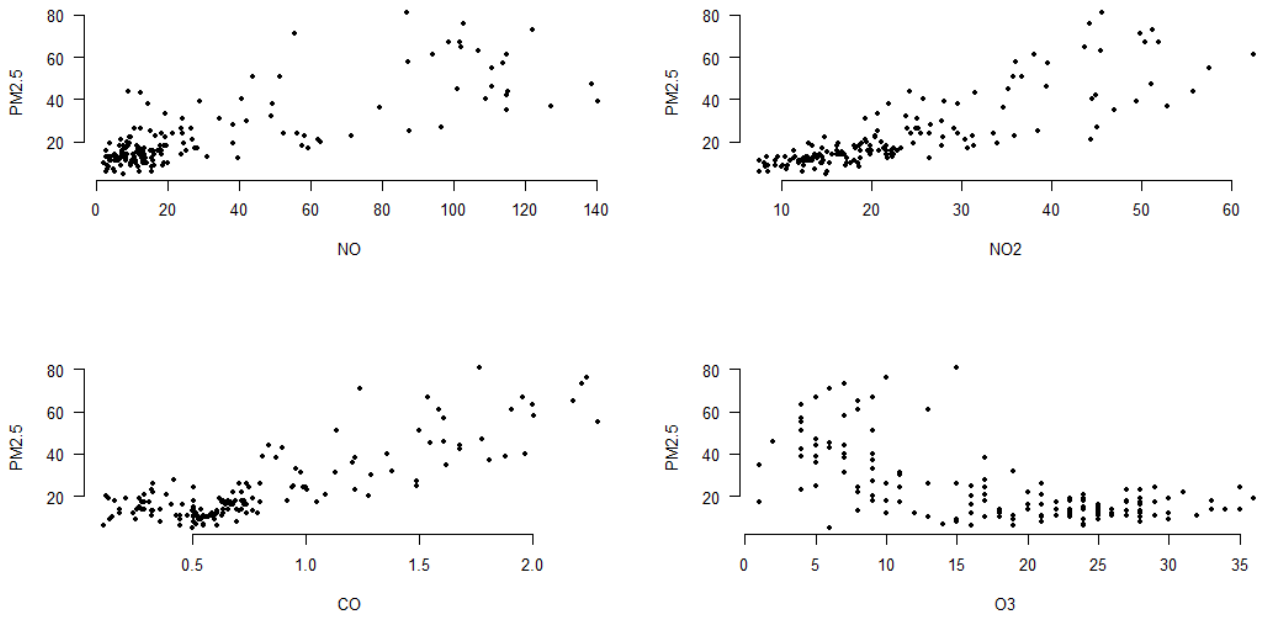
PM2.5 y Humed: Correlación positiva significativa (t = 5.3161, p-value = 3.562e-07).

Cada uno de estos tests sugiere que hay una correlación significativa entre las variables analizadas y **PM2.5**, dado que todos los p-values son mucho menores que el nivel de significancia típico de 0.05. En todos los casos, H0 es rechazado.

2. Obtenga el mejor modelo de regresión lineal simple basado en los contaminantes atmosféricos.

Contaminantes atmosféricos
NO – Monóxido de nitrógeno (ppb)
NO2 – Dióxido de nitrógeno (ppb)
CO – Monóxido de carbono (ppm)
O3 – Ozono (ppb)

Gráficamente:



Modelos para los distintos contaminantes atmosféricos:

1. NO

```
> Modelo1 <- lm(PM2.5 ~ NO, data = base)
> summary(Modelo1)

Call:
lm(formula = PM2.5 ~ NO, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-23.553  -5.536  -1.762   3.866  39.380

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.3279     1.0743   10.54   <2e-16 ***
NO            0.3653     0.0228   16.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.19 on 158 degrees of freedom
Multiple R-squared:  0.619,     Adjusted R-squared:  0.6166
F-statistic: 256.7 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.619

2. NO2

```
> Modelo2 <- lm(PM2.5 ~ NO2, data = base)
> summary(Modelo2)

Call:
lm(formula = PM2.5 ~ NO2, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-25.764  -5.436  -0.199   3.146  32.835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.34591    1.51002  -2.216   0.0281 *
NO2          1.12961    0.05784  19.528   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.933 on 158 degrees of freedom
Multiple R-squared:  0.7071,    Adjusted R-squared:  0.7052
F-statistic: 381.4 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.7071

3. CO

```
> Modelo3 <- lm(PM2.5 ~ CO, data = base)
> summary(Modelo3)

Call:
lm(formula = PM2.5 ~ CO, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-16.701  -5.986  -2.497   4.720  36.102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.159      1.315   0.881     0.38
CO            27.209      1.399  19.456   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.957 on 158 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.7037
F-statistic: 378.5 on 1 and 158 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.7055

4. **O3**

```
> Modelo4 <- lm(PM2.5 ~ O3, data = base)
> summary(Modelo4)

Call:
lm(formula = PM2.5 ~ O3, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-31.923  -7.223  -1.250   5.132  54.387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.7956     2.3922  18.308   <2e-16 ***
O3           -1.1455     0.1173  -9.763   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.03 on 158 degrees of freedom
Multiple R-squared:  0.3763,    Adjusted R-squared:  0.3723
F-statistic: 95.32 on 1 and 158 DF,  p-value: < 2.2e-16
```

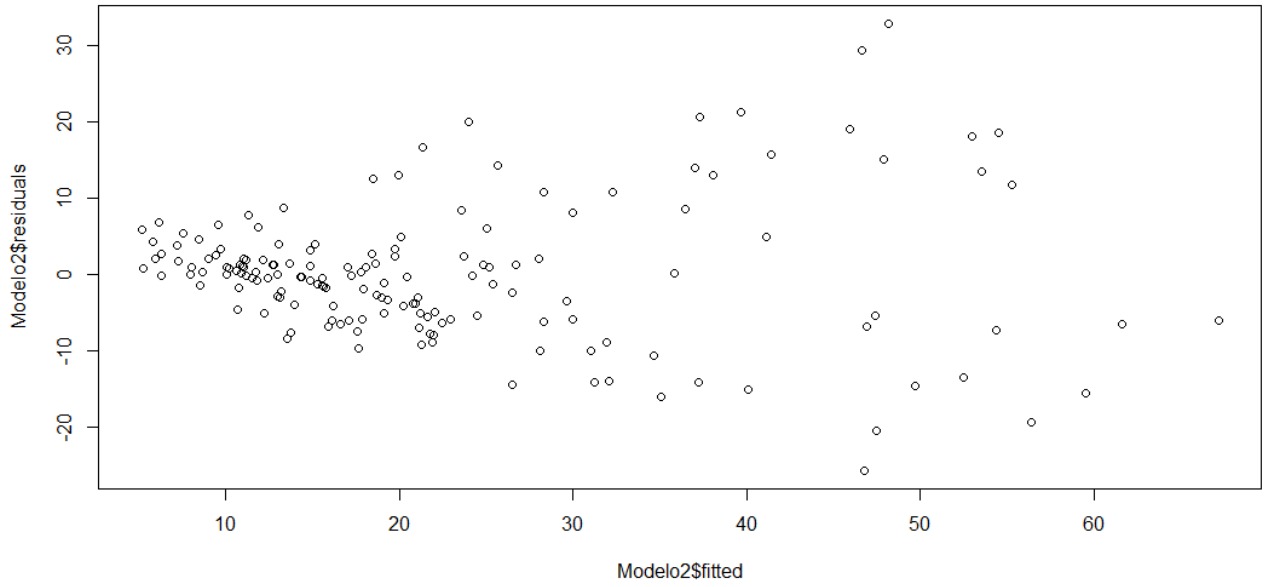
Multiple R-squared: 0.3763

Resp: el mejor modelo simple sin transformación con variable contaminantes atmosféricos es presentado con Modelo2 usando “NO2” dado que tiene la mayor R2 que es 70,7% (en R da con decimales: 0.7070606)

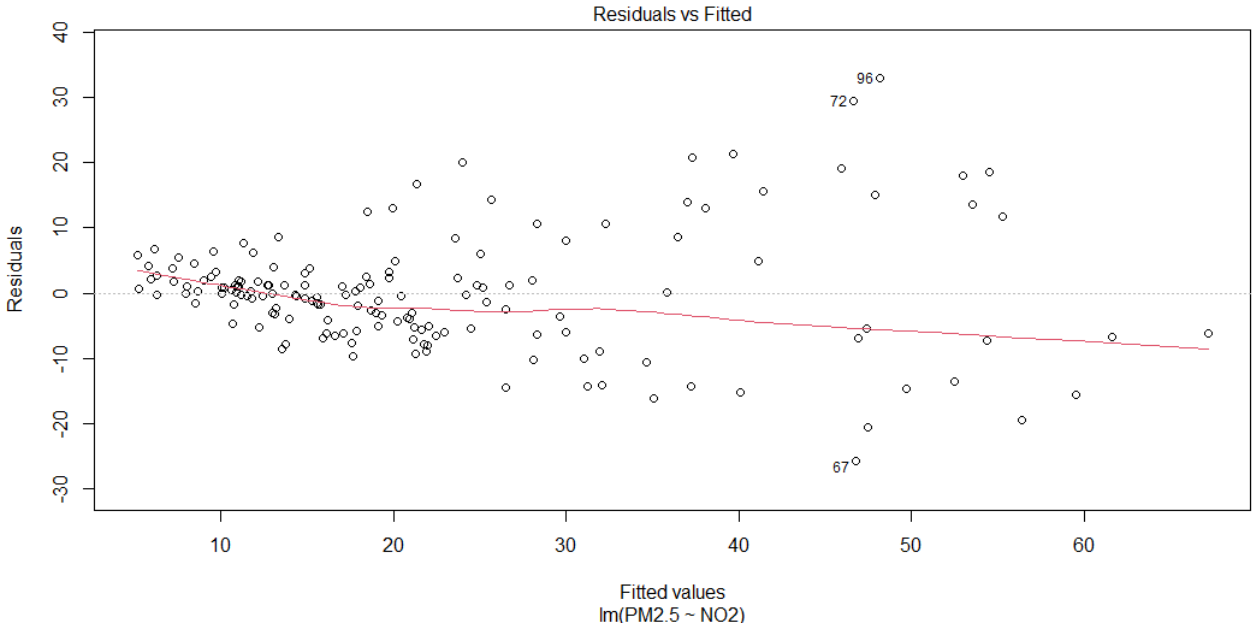
Validación de los supuestos para el "Modelo2"

1. **Linealidad**

plot(Modelo2$fitted, Modelo2$residuals)



plot(Modelo2, 1)

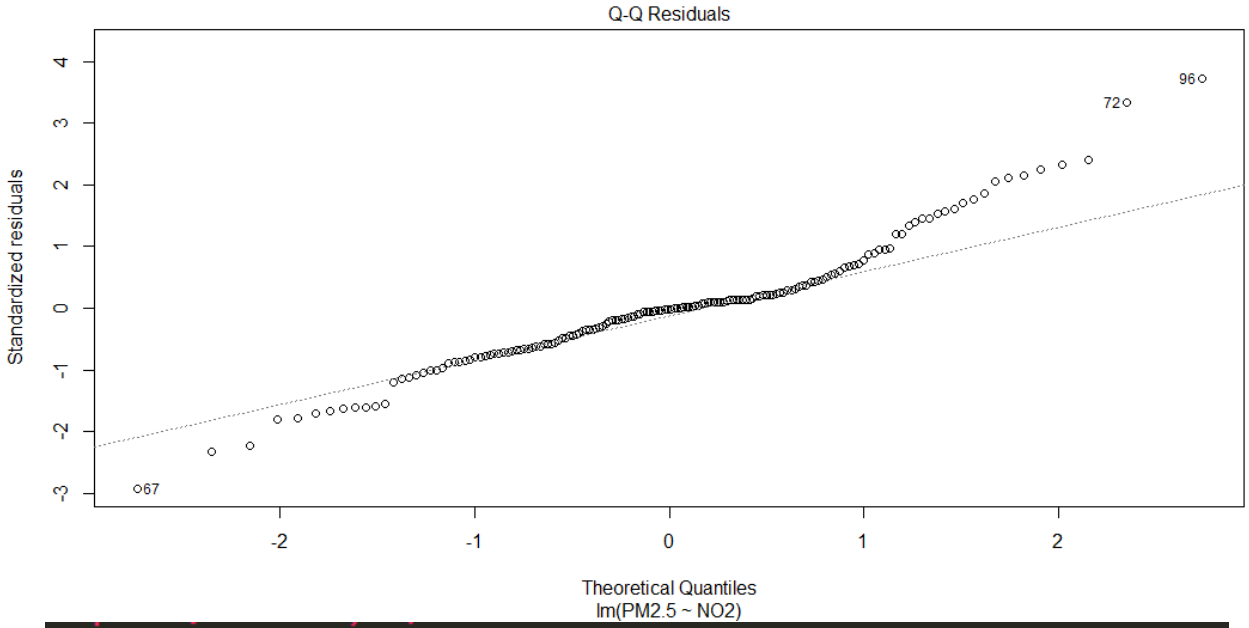


Resp: Se acepta la linealidad

2. **Normalidad**

H0: residuos distribuyen normales
H1: residuos no distribuyen normales
Gráficamente:

plot(Modelo2, 2)




```
> nortest::lillie.test(Modelo2$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Modelo2$residuals
D = 0.12514, p-value = 2.017e-06
```

D = 0.12514, p-value = 2.017e-06 < alfa = 0.05 Se rechaza H0

```
> ks.test(Modelo2$residuals, "pnorm", mean(Modelo2$residuals), sd(Modelo2$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Modelo2$residuals
D = 0.12514, p-value = 0.01332
alternative hypothesis: two-sided
```

D = 0.12514, p-value = 0.01332 < alfa = 0.05 se rechaza H0

Resp: En ambos test, se rechaza H0, es decir, residuos no distribuyen normales

3. **Homocedasticidad**

   H0: Sí existe Homocedasticidad

   H1: No existe Homocedasticidad

```
> lmtest::bptest(Modelo2)

        studentized Breusch-Pagan test

data:  Modelo2
BP = 55.17, df = 1, p-value = 1.105e-13
```

BP = 55.17, df = 1 ; p-value = 1.105e-13 < alfa = 0.05, se rechaza H0, es decir, No existe Homocedasticidad

4. **Independencia**

   H0: no hay autocorrelación vs H1: hay autocorrelación

```
> lmtest::dwtest(Modelo2)

        Durbin-Watson test

data:  Modelo2
DW = 2.0064, p-value = 0.518
alternative hypothesis: true autocorrelation is greater than 0
```

DW = 2.0064, p-value = 0.518 > alfa = 0.05 Se acepta H0, existe independencia

Explicación del Modelo2

```
> summary(Modelo2)

Call:
lm(formula = PM2.5 ~ NO2, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-25.764  -5.436  -0.199   3.146  32.835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.34591    1.51002  -2.216   0.0281 *
NO2          1.12961    0.05784  19.528   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.933 on 158 degrees of freedom
Multiple R-squared:  0.7071,    Adjusted R-squared:  0.7052
F-statistic: 381.4 on 1 and 158 DF,  p-value: < 2.2e-16
```

B1 = 1.12961

Modelo2 quedaría con:

Y = B0 + B1 * X

PM2.5 = -3.35 + 1.13 * NO2

Test de significancia del B1

H0: **B1 = 0**
H1: **B1 distinto de 0**

Salida R:

(NO2) t-value= 19,528; P-VALUE= < 2e-16 es menor a alfa = 0,05; Se rechaza H0

Para un alfa tan pequeño, la variable **NO2** es significativa

Test de significancia del modelo

H0: no existe regresión vs H1: existe regresión

F-statistic: 381,4; p-value: < 2.2e-16 ; es menor a alfa = 0,05, se rechaza H0 , es decir , existe regresión

Multiple R-squared: 0,7071: el modelo explica en un 70,71% el nivel de contaminación PM2.5

Test para ro usando cor.test

H0: ro = 0
H1: ro distinto de 0

```
# Correlación entre PM2.5 y NO
cor.test(base$PM2.5, base$NO)
# t = 16;021; p-value < 2.2e-16

# Correlación entre PM2.5 y NO2
cor.test(base$PM2.5, base$NO2)
# t = 19,528; p-value < 2.2e-16
```

```
# Correlación entre PM2.5 y CO
cor.test(base$PM2.5, base$CO)
# t = 19,456; p-value < 2.2e-16

# Correlación entre PM2.5 y O3
cor.test(base$PM2.5, base$O3)
# t = -9,7629; p-value < 2.2e-16
```

Interpretación: Dado un nivel de significancia $\alpha = 0.05$, todas las correlaciones entre las concentraciones de PM2.5 y las variables NO, NO2, CO y O3 son estadísticamente significativas. Esto se confirma por los p-values menores a 0.05 en cada caso, lo que indica que podemos rechazar la hipótesis nula de que no existe correlación entre PM2.5 y cada una de estas variables.

3. Con base a todas las variables (meteorológicas y contaminantes), mediante una técnica iterativa (forward o backward) seleccione el mejor modelo predictivo. Indique para cada paso qué variable entra/sale del modelo, indicando el aumento/disminución del R2-ajustado.

En este caso se usa "forward"

```
vacio    <- lm(PM2.5 ~ 1, base) # modelo sin variables, solo con intercepto
completo <- lm(PM2.5 ~ ., base) # modelo con todas las variables
M2<- step(vacio,direction="forward",scope=formula(completo))
summary(M2)
```

Step 1: PM2.5 ~ NO2: Se agrega NO2

Step 2: PM2.5 ~ NO2 + CO: Se agrega CO

Step 3: PM2.5 ~ NO2 + CO + Humed: Se agrega Humed

Step 4: PM2.5 ~ NO2 + CO + Humed + O3: Se agrega O3

Step 5: PM2.5 ~ NO2 + CO + Humed + O3 + TMin: Se agrega TMin

Step 6: PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO: Se agrega NO

Step 7: PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento: Se agrega Viento

Resumen con comando summary para el M2

```
> summary(M2)

Call:
lm(formula = PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento,
    data = base)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4803  -4.0512  -0.4822   4.4824  23.3433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.5243     6.9520  -4.822 3.41e-06 ***
NO2           0.8791     0.1234   7.124 3.92e-11 ***
CO           19.8313     2.7794   7.135 3.70e-11 ***
Humed         0.2983     0.0668   4.465 1.55e-05 ***
O3            0.4343     0.1218   3.565 0.000486 ***
TMin         -0.7535     0.1992  -3.782 0.000223 ***
NO           -0.1343     0.0529  -2.540 0.012090 *
Viento        5.0160     2.3258   2.157 0.032601 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.988 on 152 degrees of freedom
Multiple R-squared:  0.8276,    Adjusted R-squared:  0.8196
F-statistic: 104.2 on 7 and 152 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: 0,8196 = 81.96% de calidad del modelo

Calculo del aumento del R2-ajustado

ΔR2-Ajustado=R2(M2)−R2(Completo)

ΔR-Ajustado = 0.8196 − 0.8179=0.0017

```
summary(M2) #Adjusted R-squared:  0.8196
summary(completo) #Adjusted R-squared:  0.8179
print(0.8196 - 0.8179)
```

Resp: Para el caso “forward” hubo un aumento del R2 Ajustado de 0.0017 (0.17%). El modelo **M2**, que es más simple y se ha obtenido mediante una selección “forward”, proporciona un ajuste ligeramente mejor a los datos que el modelo **completo**.

Test con anova comparando modelo completo vs M2

H_0: El modelo completo no proporciona un ajuste significativamente mejor que el modelo reducido.
H_1: El modelo completo proporciona un ajuste significativamente mejor que el modelo reducido.

```
> anova(completo, M2)
Analysis of Variance Table

Model 1: PM2.5 ~ Viento + TProm + TMin + TMax + Humed + NO + NO2 + CO +
    O3
Model 2: PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    150 7393.5
2    152 7422.1 -2   -28.637 0.2905 0.7483
```

Conclusión: T = 0.29; p-valor = 0.7483: Esto es mayor que alfa = 0.05, por lo que no se rechaza la hipótesis nula. Esto significa que el modelo **completo** no mejora significativamente el ajuste en comparación con el modelo más sencillo (**M2**). Por lo tanto, el modelo **M2** es preferible en términos de simplicidad y ajuste.

Detalle del R2 Ajustado por paso a paso del step

```
modelos <- list(
  lm(PM2.5 ~ NO2, data = base),
  lm(PM2.5 ~ NO2 + CO, data = base),
  lm(PM2.5 ~ NO2 + CO + Humed, data = base),
  lm(PM2.5 ~ NO2 + CO + Humed + O3, data = base),
  lm(PM2.5 ~ NO2 + CO + Humed + O3 + TMin, data = base),
  lm(PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO, data = base),
  lm(PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento, data = base)
)
```

Step 1: 0.7052066

Step 2: 0.7627398

Step 3: 0.7877649

Step 4: 0.8011568

Step 5: 0.8115301

Step 6: 0.8153084

Step 7: 0.8196134

Step final con modelo **PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento**

R2 = 0.8196134

Modelo para este caso:

$$\hat{Y} = \beta_0 + \beta_1 NO2 + \beta_2 CO + \beta_3 Humed + \beta_4 O3 + \beta_5 TMin + \beta_6 NO + \beta_7 Viento$$

$$\hat{Y} = -33.5243 + 0.8791 \times NO2 + 19.8313 \times CO + 0.2983 \times Humed + 0.4343 \times O3 - 0.7535 \times TMin - 0.1343 \times NO + 5.0160 \times Viento$$

4. Basado en los resultados previos, proponga un modelo con tres predictores (debe incluir una variable meteorológica y dos contaminantes), revise supuestos y evalúe con especial énfasis el problema de multicolinealidad. Apóyese de tablas de correlación, gráficos y métricas respectivas.

Resumen de los R2 para variables meteorológica

Modelo	Multiple R-squared
Modelo1 <- lm(PM2.5 ~ **Viento**, data = base)	0.3628462
Modelo2 <- lm(PM2.5 ~ **TProm**, data = base)	0.3877903
Modelo3 <- lm(PM2.5 ~ **TMin**, data = base)	0.3921017
Modelo4 <- lm(PM2.5 ~ **TMax**, data = base)	0.2048031
Modelo5 <- lm(PM2.5 ~ **Humed**, data = base)	0.1517276

Resumen de los R2 para variables contaminantes

Modelo	Multiple R-squared
Modelo1 <- lm(PM2.5 ~ **NO**, data = base)	0.6189904
Modelo2 <- lm(PM2.5 ~ **NO2**, data = base)	0.7070606
Modelo3 <- lm(PM2.5 ~ **CO**, data = base)	0.7055177
Modelo4 <- lm(PM2.5 ~ **O3**, data = base)	0.3762709

Supuestos: como el enunciado pide seleccionar una variable meteorológica y dos contaminantes, se escoge los que tienen mejores R2, es decir, el más alto ya que con eso indica un mayor % de explicación del modelo.

Para el caso de variable meteorológica: **Tmin**

Para el caso de variable contaminante: **NO2** y **CO**

ModeloEscogido <- lm(PM2.5 ~ TMin + NO2 + CO, data = base)

Modelo con 3 predictores:

summary(ModeloEscogido)

```
> summary(ModeloEscogido)

Call:
lm(formula = PM2.5 ~ TMin + NO2 + CO, data = base)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4605  -4.5763  -0.1581   3.7906  28.6098

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.75361    2.94728   1.613  0.10879
TMin        -0.57425    0.19231  -2.986  0.00328 **
NO2          0.53292    0.09862   5.404 2.40e-07 ***
CO          13.69980    2.30419   5.946 1.74e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.819 on 156 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7741
F-statistic: 182.6 on 3 and 156 DF,  p-value: < 2.2e-16
```

Definición de $\hat{Y}$ con los 3 predictores

PM2.5 = 4.75 - 0.57 * TMin + 0.53 * NO2 + 13.70 * CO

Adjusted R-squared: 0.7741

Todas las variables son significativas, calidad del modelo ajustado es de 77,41%

Test de significancia del Bi

Para cada Bi:

H0: **Bi = 0**
H1: **Bi distinto de 0**

Coeficiente del Intercepto (constante)

- Resultado del test: t = 1.613, p-value = 0.10879

Conclusión: No se rechaza la hipótesis nula, ya que el p-valor es mayor que el nivel de significancia de 0.05. No hay suficiente evidencia para concluir que el intercepto es significativamente diferente de cero.

Coeficiente de TMin

- Resultado del test: t = -2.986, p-value = 0.00328

Conclusión: Se rechaza la hipótesis nula, ya que el p-valor es menor que el nivel de significancia de 0.05. Hay suficiente evidencia para concluir que el coeficiente de TMin es significativamente diferente de cero.

Coeficiente de NO2

- Resultado del test: t = 5.404, p-value = 2.40e-07

Conclusión: Se rechaza la hipótesis nula, ya que el p-valor es mucho menor que 0.05. Hay suficiente evidencia para concluir que el coeficiente de NO2 es significativamente diferente de cero.

Coeficiente de CO

- Resultado del test: t = 5.946, p-value = 1.74e-08

Conclusión: Se rechaza la hipótesis nula, ya que el p-valor es mucho menor que 0.05. Hay suficiente evidencia para concluir que el coeficiente de CO es significativamente diferente de cero.

Conclusión: Los coeficientes de las variables TMin, NO2 y CO son significativamente diferentes de cero, lo que indica que estas variables tienen un impacto significativo en la variable dependiente PM2.5

Test de significancia del modelo

H0: no existe regresión vs H1: existe regresión

F-statistic: 182,6 ; p-value: < 2.2e-16

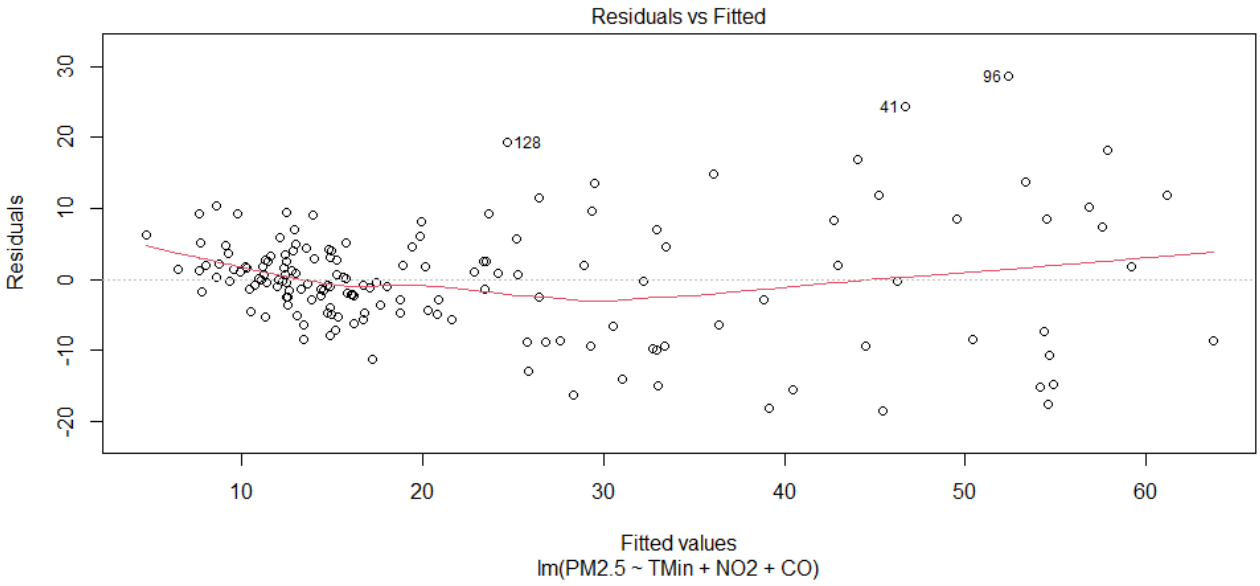
es menor a alfa = 0.05. Se rechaza la hipótesis nula. Esto implica que al menos una de las variables independientes en el modelo tiene un efecto significativo sobre la variable dependiente PM2.5

Adjusted R-squared: 0.7741. El modelo con los 3 predictores explica en un 77,41% el nivel de contaminación PM2.5

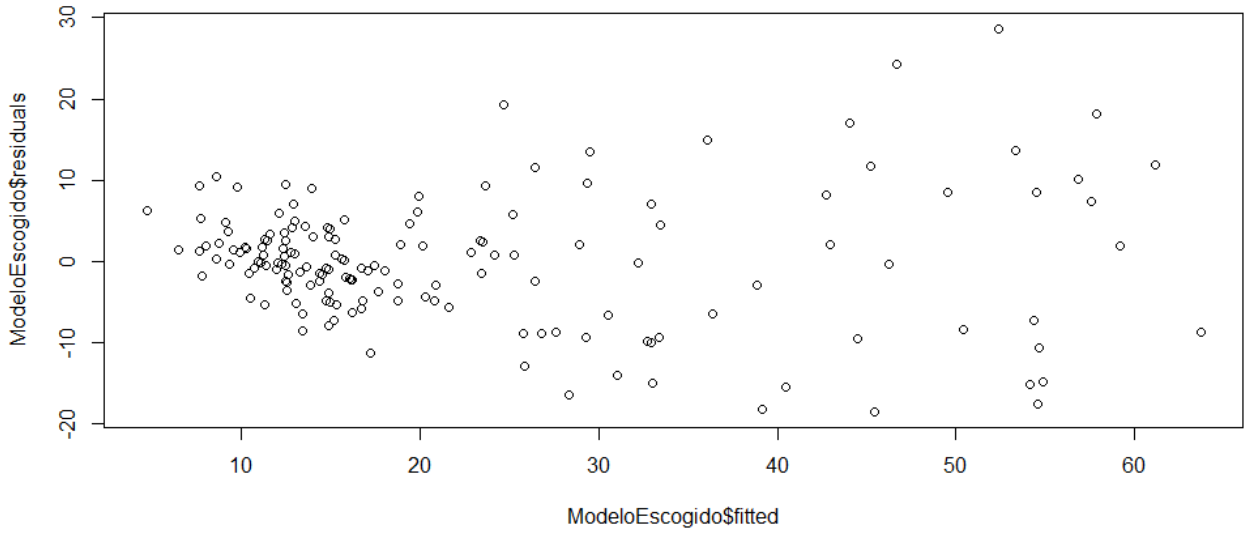
Validación de los supuestos para el ModeloEscogido PM2.5 ~ TMin + NO2 + CO

1. Linealidad

plot(ModeloEscogido, 1)



plot(ModeloEscogido$fitted, ModeloEscogido$residuals)



Resp: cumple linealidad

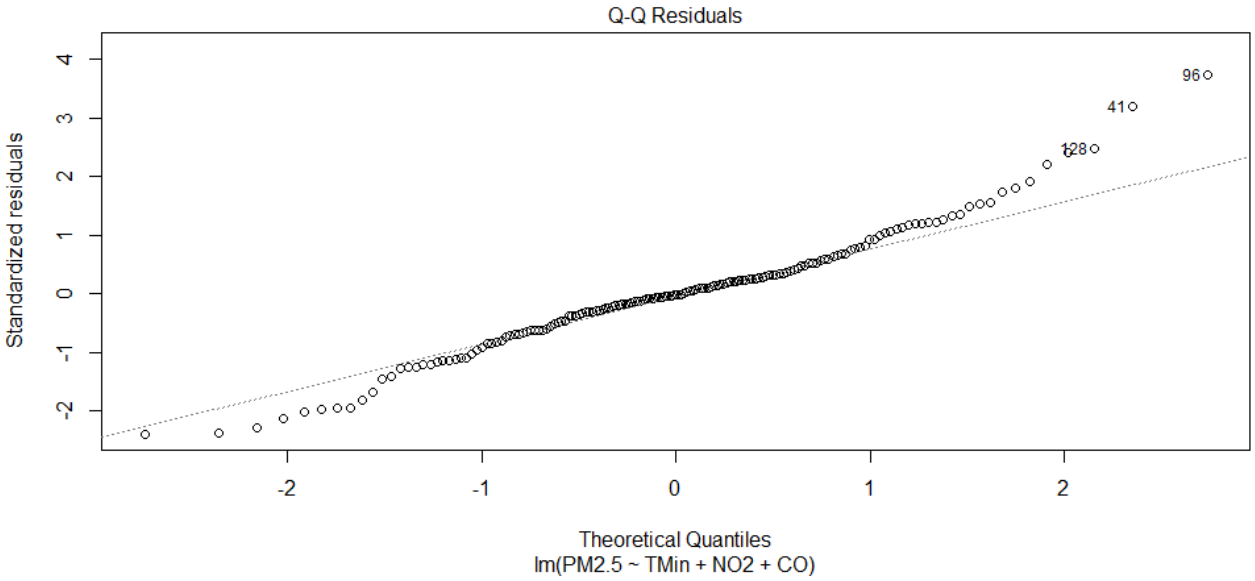
2. **Normalidad**

H0: residuos distribuyen normales

H1: residuos no distribuyen normales

Gráficamente:

plot(ModeloEscogido, 2)




```
> nortest::lillie.test(ModeloEscogido$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  ModeloEscogido$residuals
D = 0.079597, p-value = 0.01496
```

D = 0.079597, p-value = 0.01496 < alfa = 0.05, se rechaza H0, no hay normalidad usando lilli test

```
> ks.test(ModeloEscogido$residuals, "pnorm", mean(ModeloEscogido$residuals), sd(Mode
loEscogido$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  ModeloEscogido$residuals
D = 0.079597, p-value = 0.2628
alternative hypothesis: two-sided
```

D = 0.079597, p-value = 0.2628 > alfa = 0.05 se acepta H0, hay normalidad según test ks

Resp: Los dos tests ofrecen resultados contradictorios respecto a la normalidad de los residuos. El test de Lilliefors indica una falta de normalidad, mientras que el test de KS no encuentra evidencia suficiente para rechazar la normalidad. Se tomará como válido el test Lillie, Se rechaza H0, No existe normalidad de los residuos

3. **Homocedasticidad**

H0: Sí existe Homocedasticidad
H1: No existe Homocedasticidad

```
> lmtest::bptest(ModeloEscogido)

        studentized Breusch-Pagan test

data:  ModeloEscogido
BP = 53.646, df = 3, p-value = 1.335e-11
```

BP = 53.646, df = 3, p-value = 1.335e-11 < alfa = 0.05, se rechaza H0, no existe homocedasticidad

4. **Independencia**

H0: no hay autocorrelación vs H1: hay autocorrelación

```
> lmtest::dwtest(ModeloEscogido)

        Durbin-Watson test

data:  ModeloEscogido
DW = 2.1726, p-value = 0.8634
alternative hypothesis: true autocorrelation is greater than 0
```

DW = 2.1726, p-value = 0.8634 > alfa = 0.05, Se acepta H0, existe independencia

Análisis de multicolinealidad

```
> car::vif(ModeloEscogido)
    TMin      NO2       CO
1.616250 3.793772 3.561617
```

Según documentación

Otra manera es calcular VIF

- VIF cercano a 1, demuestra ausencia de multicolinealidad
- VIF entre 1 y 5 es problema moderado de multicolinealidad
- VIF mayor a 5 es problema grave de multicolinealidad

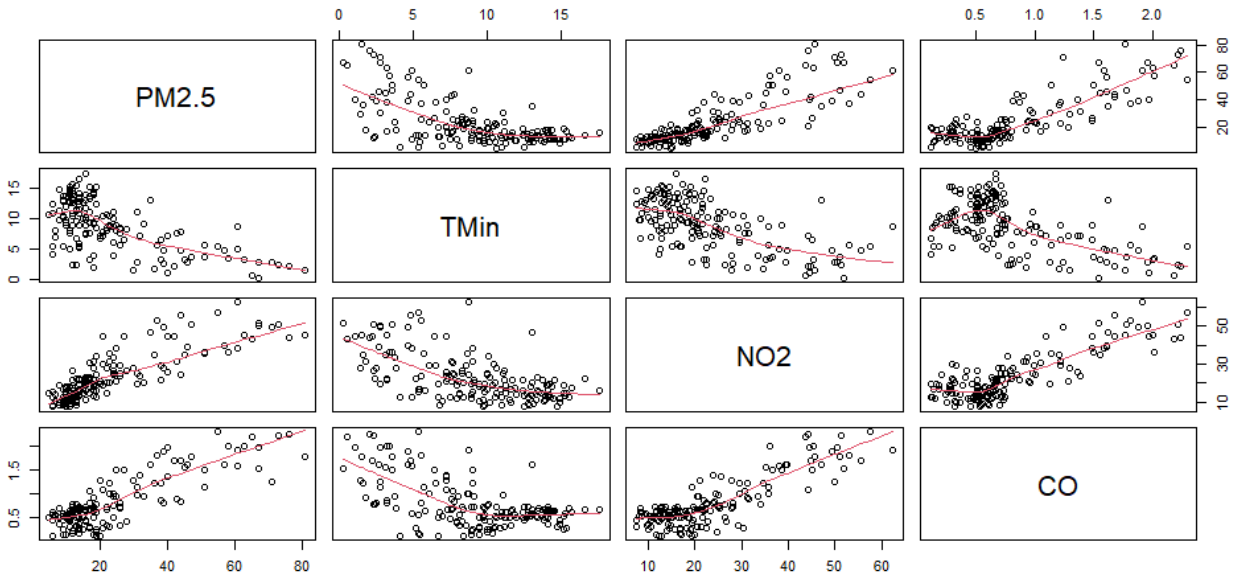
Resp: No hay multicolinealidad grave en el modelo, ya que todos los VIF están por debajo de 5. Dado el rango anterior, cae en la categoría de “moderado”

Gráficos de relación entre todas las variables

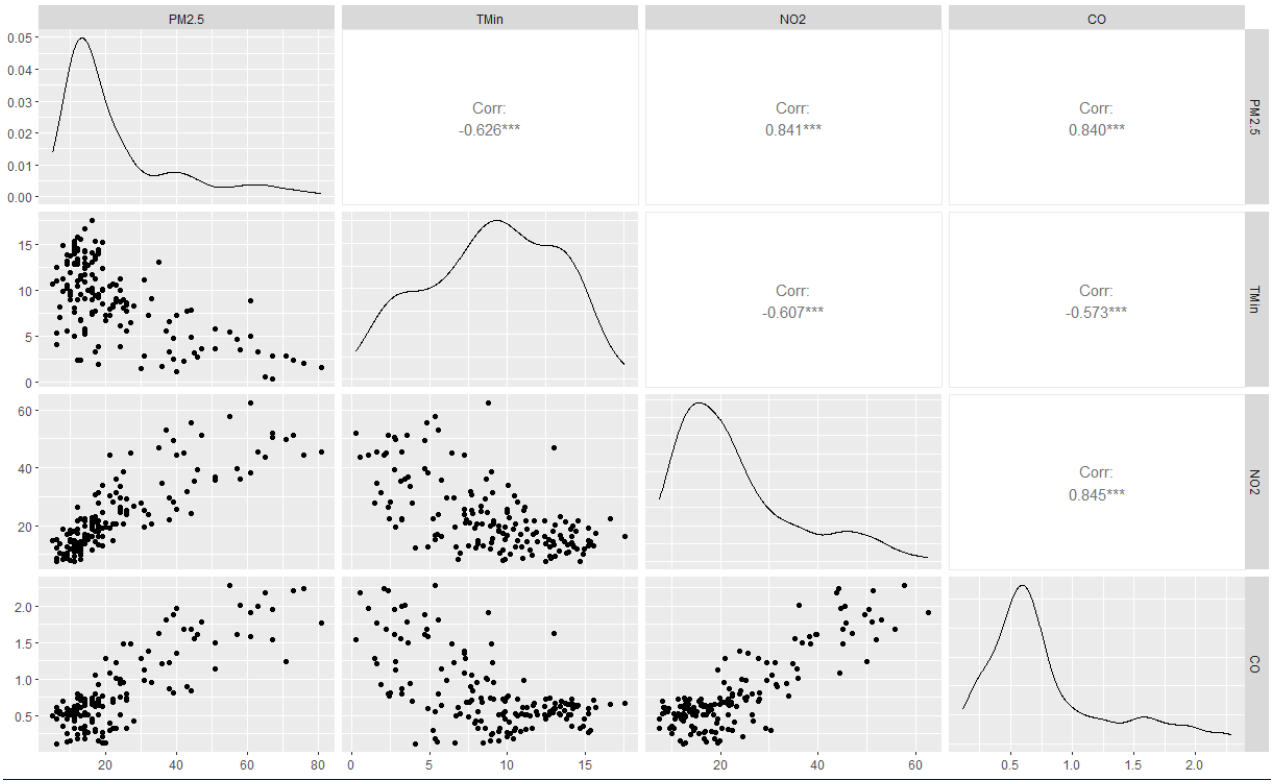
Análisis de correlación
base_filtrada <- base[, c("PM2.5", "TMin", "NO2", "CO")]

Matriz de dispersión

pairs(base_filtrada,upper.panel= panel.smooth, lower.panel = panel.smooth)



ggpairs(base_filtrada)



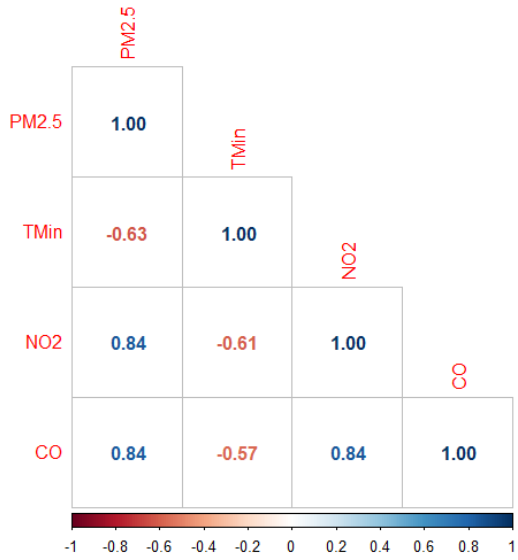
Cálculo de Correlaciones

correlacion <- cor(base_filtrada)
print(correlacion)

```
> print(correlacion)
             PM2.5       TMin        NO2         CO
PM2.5   1.0000000 -0.6261802  0.8408690  0.8399510
TMin   -0.6261802  1.0000000 -0.6074126 -0.5725528
NO2     0.8408690 -0.6074126  1.0000000  0.8447672
CO      0.8399510 -0.5725528  0.8447672  1.0000000
```

Matriz de correlación

corrplot(correlacion, type="lower", method = "number")



Coeficiente de determinación del modelo

Adjusted R-squared: 0.7741

Pruebas de Hipótesis para Correlaciones (ρ)

Hipótesis ro
H0: ro = 0
H1: ro distinto de 0

1. Correlación entre PM2.5 y TMin
> cor.test(base$PM2.5, base$TMin)
t = -10,095, df = 158, p-value < 2.2e-16
Conclusión: Existe una correlación negativa moderada significativa entre PM2.5 y TMin, con un valor t de -10,095 y un p-value < 2.2e-16.

2. Correlación entre PM2.5 y NO2
> cor.test(base$PM2.5, base$NO2)
t = 19,528, df = 158, p-value < 2.2e-16
Conclusión: Existe una correlación positiva fuerte significativa entre PM2.5 y NO2, con un valor t de 19,528 y un p-value < 2.2e-16.

3. Correlación entre PM2.5 y CO
> cor.test(base$PM2.5, base$CO)
t = 19,456, df = 158, p-value < 2.2e-16
Conclusión: Existe una correlación positiva fuerte significativa entre PM2.5 y CO, con un valor t de 19,456 y un p-value < 2.2e-16.

4. Correlación entre TMin y NO2
> cor.test(base$TMin, base$NO2)
t = -9,6113, df = 158, p-value < 2.2e-16
Conclusión: Existe una correlación negativa moderada significativa entre TMin y NO2, con un valor t de -9,6113 y un p-value < 2.2e-16.

5. Correlación entre TMin y CO
> cor.test(base$TMin, base$CO)
t = -8,7781, df = 158, p-value = 2.572e-15
Conclusión: Existe una correlación negativa moderada significativa entre TMin y CO, con un valor t de -8,7781 y un p-value de 2.572e-15.

6. Correlación entre NO2 y CO
> cor.test(base$NO2, base$CO)
t = 19,843, df = 158, p-value < 2.2e-16
Conclusión: Existe una correlación positiva fuerte significativa entre NO2 y CO, con un valor t de 19,843 y un p-value < 2.2e-16.

Comentario general: en todas las pruebas, los valores p son extremadamente bajos (mucho menores a 0,05), lo que indica que todas las correlaciones encontradas son estadísticamente significativas.

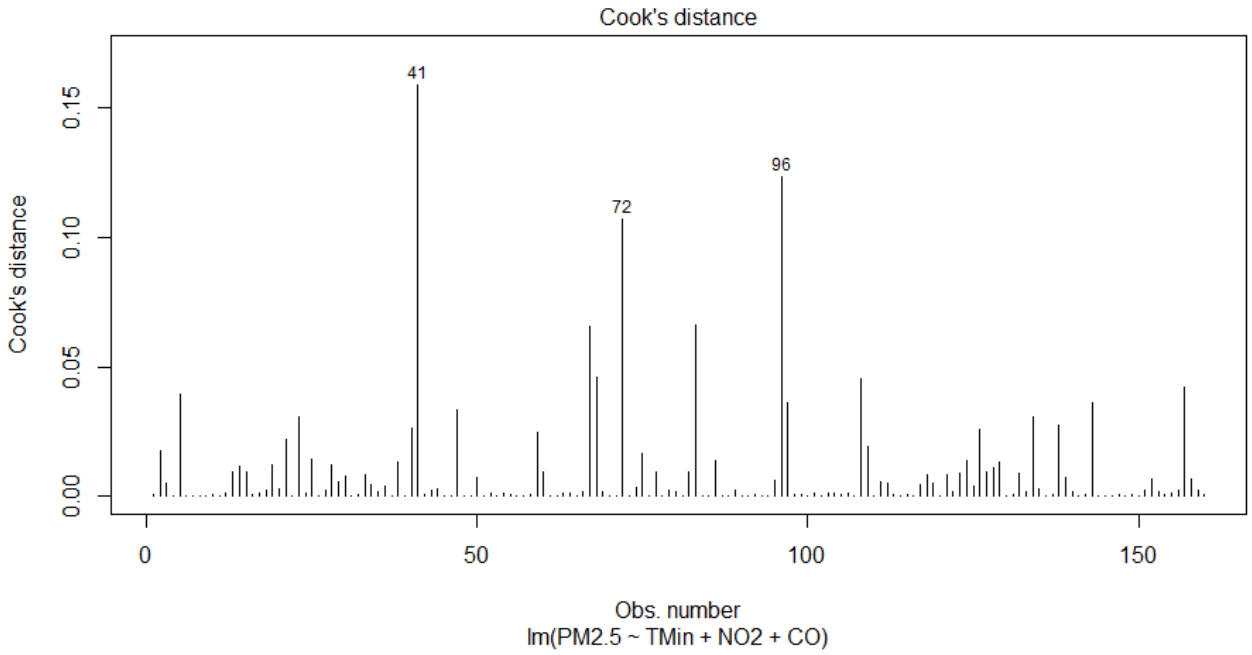
Valores atípicos o influyentes

```
> car::outlierTest(ModeloEscogido)
   rstudent unadjusted p-value Bonferroni p
96 3.888469         0.00014928     0.023884
> summary(influence.measures(ModeloEscogido))
Potentially influential observations of
         lm(formula = PM2.5 ~ TMin + NO2 + CO, data = base) :

    dfb.1_ dfb.TMin dfb.NO2 dfb.CO dffit   cov.r   cook.d hat
12  -0.01   0.01     0.06   -0.06   0.07    1.09_*  0.00   0.06
35  -0.07   0.05     0.06   -0.01   0.09    1.15_*  0.00   0.11_*
41  -0.07  -0.13     0.66   -0.50   0.82_*  0.83_*  0.16   0.06
47   0.31  -0.29    -0.19   -0.04  -0.37    1.07    0.03   0.08_*
67   0.20  -0.11    -0.47    0.30  -0.52_*  0.93    0.07   0.04
72  -0.01  -0.09    -0.27    0.50   0.66_*  0.95    0.11   0.07
83   0.28  -0.15    -0.29    0.00  -0.52_*  0.94    0.07   0.05
96   0.06  -0.25     0.05    0.18   0.73_*  0.73_*  0.12   0.03
128  0.09  -0.06    -0.01    0.00   0.21    0.88_*  0.01   0.01
157  0.19  -0.11    -0.25    0.04  -0.42    0.91_*  0.04   0.03
```

Cook

plot(ModeloEscogido,4)



Leverage

plot(ModeloEscogido,5)