Contenido

[1. Obtenga el mejor modelo de regresión lineal simple basado en las variables meteorológicas. 3](#_Toc173264036)

[Validación de los supuestos para el **“Modelo3”** 6](#_Toc173264037)

[**Explicación del Modelo3** 9](#_Toc173264038)

[2. Obtenga el mejor modelo de regresión lineal simple basado en los contaminantes atmosféricos. 11](#_Toc173264039)

[Validación de los supuestos para el **“Modelo2”** 14](#_Toc173264040)

[**Explicación del Modelo2** 16](#_Toc173264041)

[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. 18](#_Toc173264042)

[Resumen con comando summary para el M2 18](#_Toc173264043)

[Detalle del R2 Ajustado por paso a paso del step 19](#_Toc173264044)

[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. 20](#_Toc173264045)

[Resumen de los R2 para variables meteorológica 20](#_Toc173264046)

[Resumen de los R2 para variables contaminantes 20](#_Toc173264047)

[**Modelo con 3 predictores:** 21](#_Toc173264048)

[Validación de los supuestos para el ModeloEscogido PM2.5 ~ TMin + NO2 + CO 22](#_Toc173264049)

[Análisis de multicolinealidad 24](#_Toc173264050)

[Cálculo de Correlaciones 26](#_Toc173264051)

[Matriz de correlacion 26](#_Toc173264052)

[Coeficiente de determinación del modelo 26](#_Toc173264053)

[Pruebas de Hipótesis para Correlaciones (𝜌) 26](#_Toc173264054)

[Valores átipicos o influyentes 27](#_Toc173264055)

[Cook 28](#_Toc173264056)

[Leverage 28](#_Toc173264057)

**Alumno: Rodrigo Jeldres Carrasco**

Control N3

Variable respuesta PM2.5

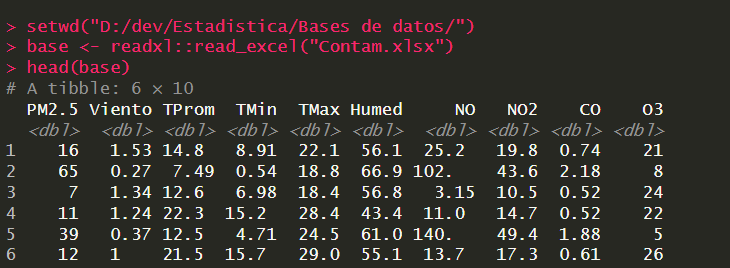
El objetivo es entender y explicar el comportamiento de los niveles de contaminación del aire en la Región Metropolitana. Para ello, desde el Sistema de Información Nacional de Calidad del Aire (sinca.mma.gob.cl), se seleccionó una muestra de la información histórica de la Estación Parque O’Higgins en 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 (%) |  |

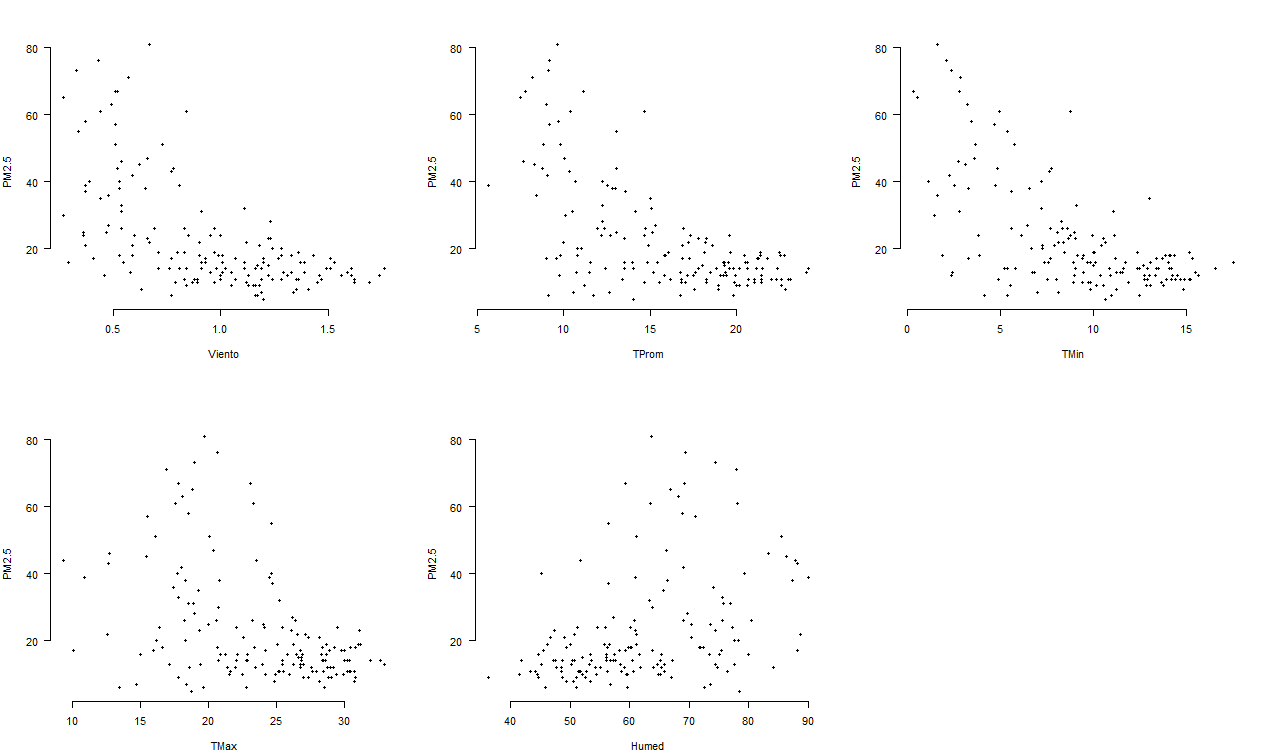
1. Obtenga el mejor modelo de regresión lineal simple basado en las variables meteorológicas.

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



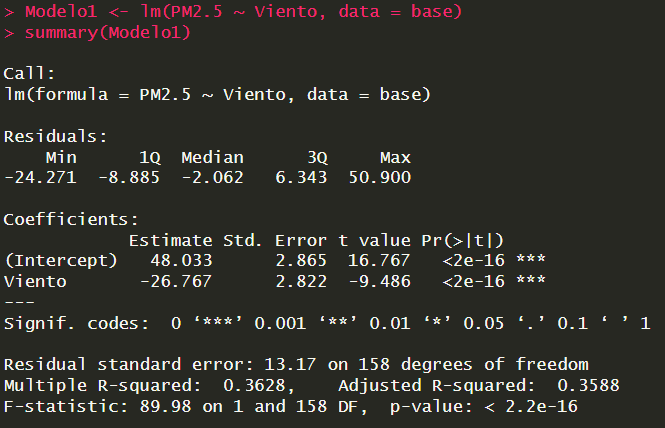
En este caso es Viento, TProm, TMin, TMax, Humed

Gráficamente:



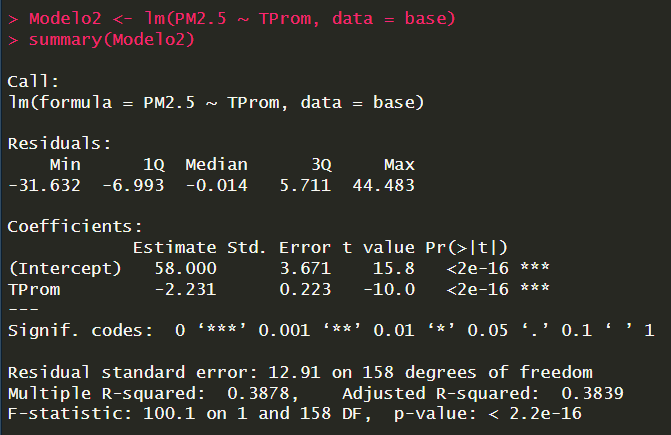
Para las distintas variables meteorológicas:

1. **Viento**



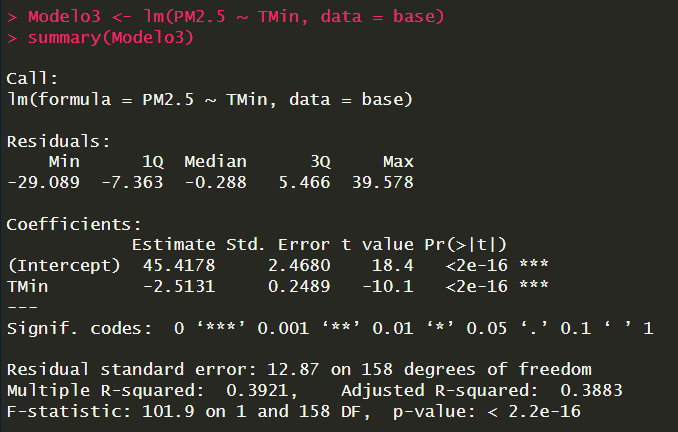
Multiple R-squared: 0.3628

1. **TProm**



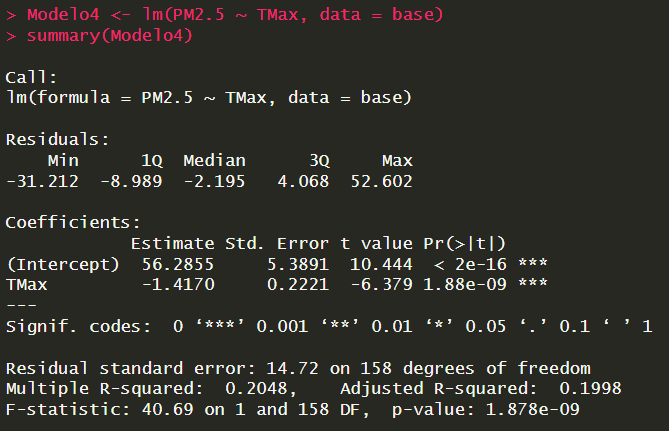
Multiple R-squared: 0.3878

1. **TMin**



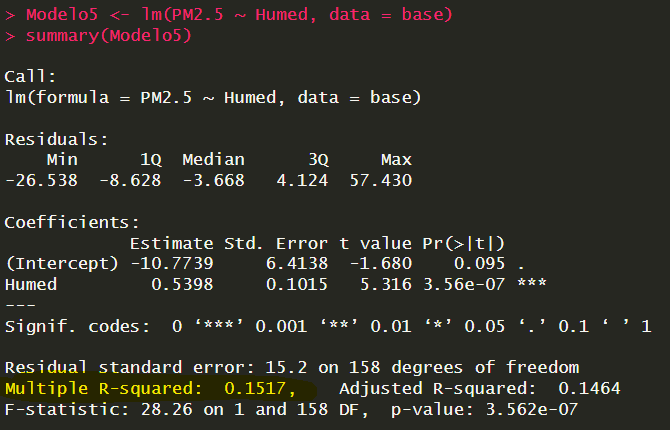
Multiple R-squared: 0.3921

1. **TMax**



Multiple R-squared: 0.2048

1. **Humed**



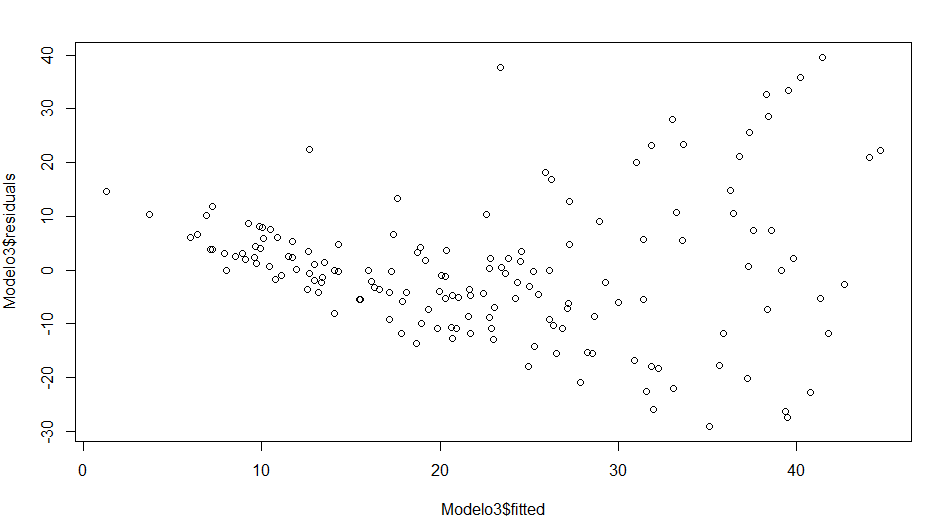
Multiple R-squared: 0.1517

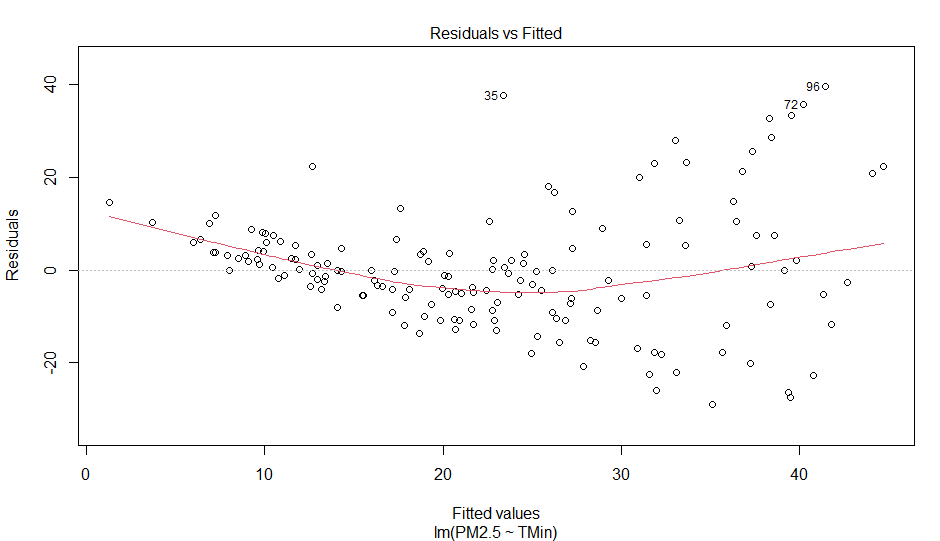
Resp: el mejor modelo simple con variable meteorológica es presentado con Modelo3 usando “Tmin” dado que tiene la mayor variable explicativa que es 39% (en R da con decimales: 0.3921017) Modelo3 <- lm(PM2.5 ~ TMin, data = base)

## Validación de los supuestos para el **“Modelo3”**

1. **Linealidad**

**plot(Modelo3$fitted, Modelo3$residuals)**



**plot(Modelo3, 1)**

Resp: Cumple linealidad

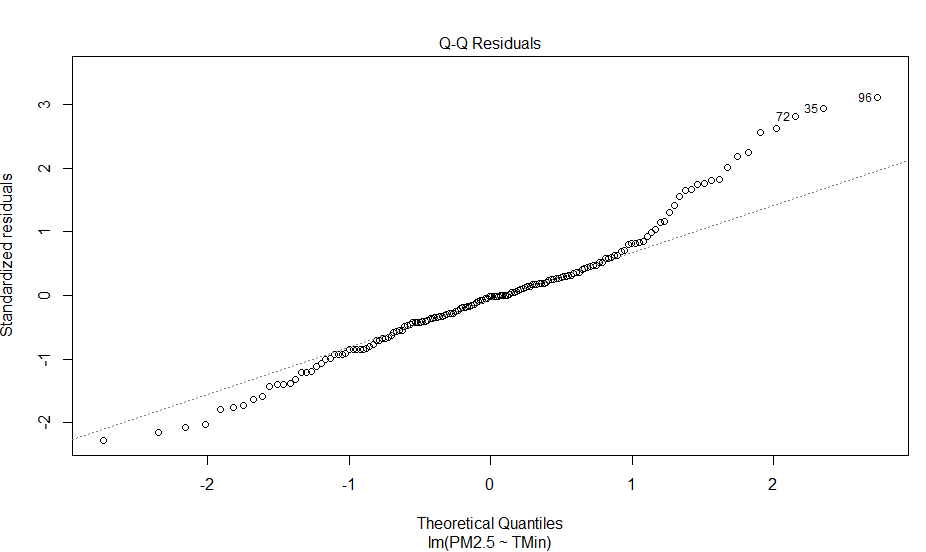
1. **Normalidad**

H0: residuos distribuyen normales

H1: residuos no distribuyen normales

Gráficamente:

**plot(Modelo3, 2)**



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

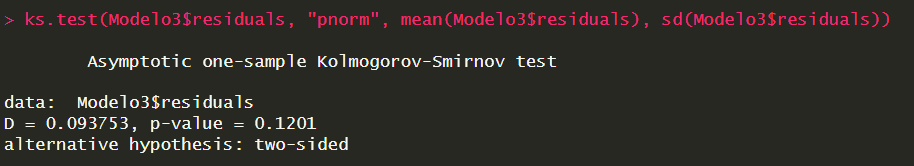
Lilliefors (Kolmogorov-Smirnov) normality test

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



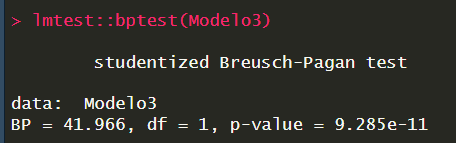
D = 0.093753, p-value = 0.1201, SÍ existe normalidad según Test KS

Resp: no hay normalidad en los residuos

1. **Homocedasticidad**

H0: Sí existe Homocedasticidad

H1: No existe Homocedasticidad

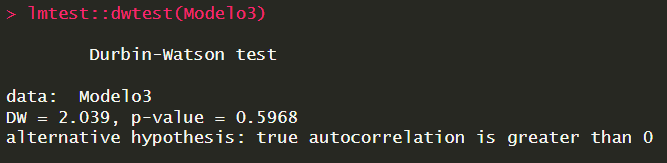


BP = 41.966, df = 1, p-value = 9.285e-11 < alfa = 0.05 ,

Resp: Se rechaza H0, no existe homocedasticidad

1. **Independencia**

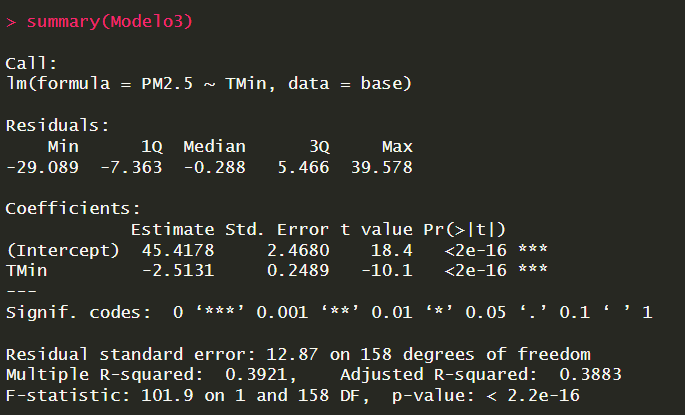
H0: no hay autocorrelación vs H1: hay autocorrelación



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

Resp: se acepta H0, existe independencia

## **Explicación del Modelo3**



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

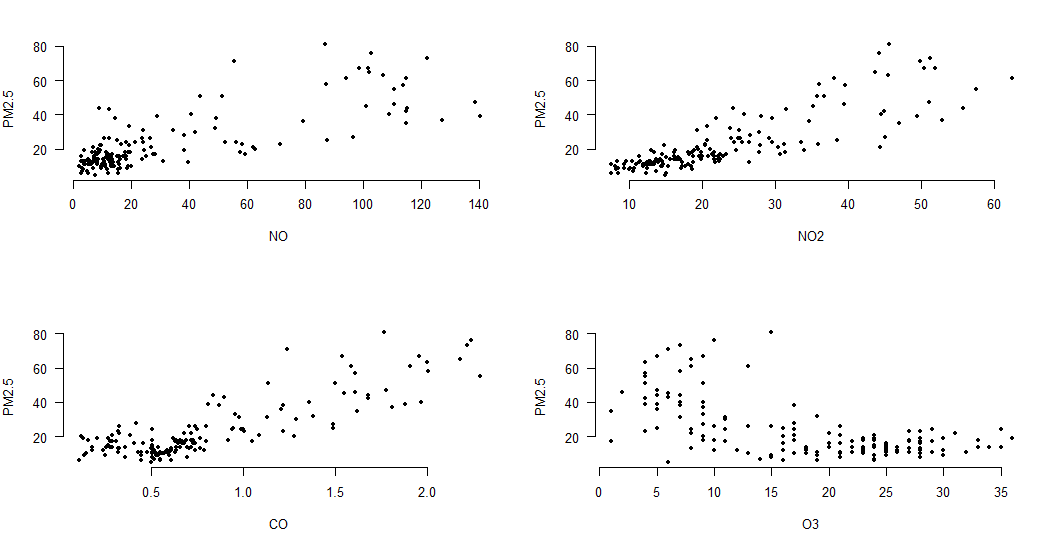
< alfa = 0.05, se rechaza H0, es decir, existe regresión

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

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

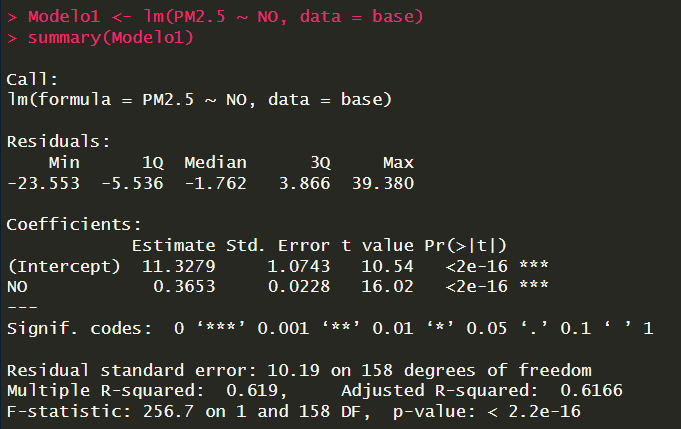
|  |
| --- |
| **Contaminantes atmosféricos** |
| NO – Monóxido de nitrógeno (ppb) |
| NO2 – Dióxido de nitrógeno (ppb) |
| CO – Monóxido de carbono (ppm) |
| O3 – Ozono (ppb) |

Graficamente:



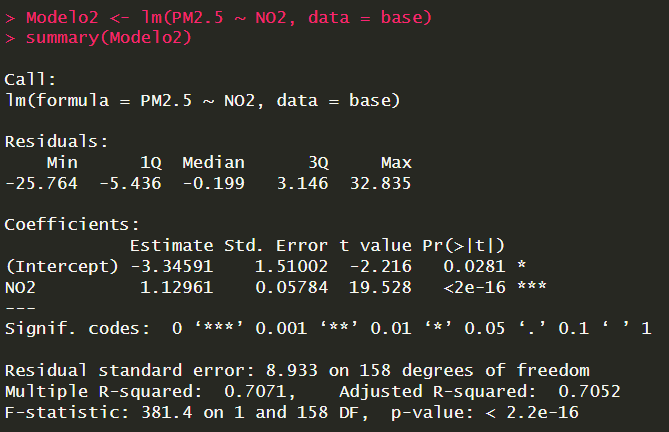
Para los distintos contaminantes atmosféricos:

1. **NO**



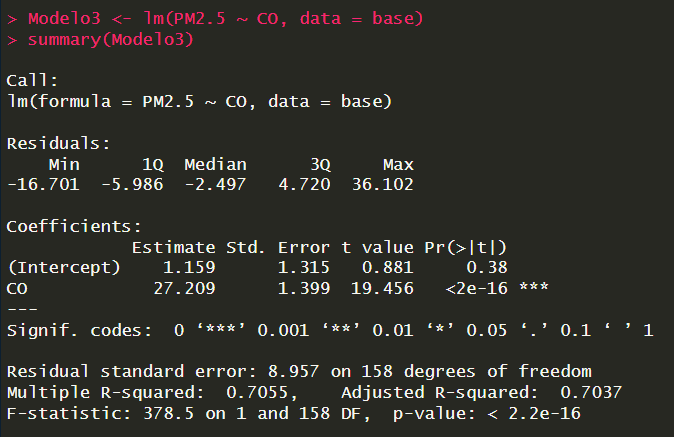
Multiple R-squared: 0.619

1. **NO2**



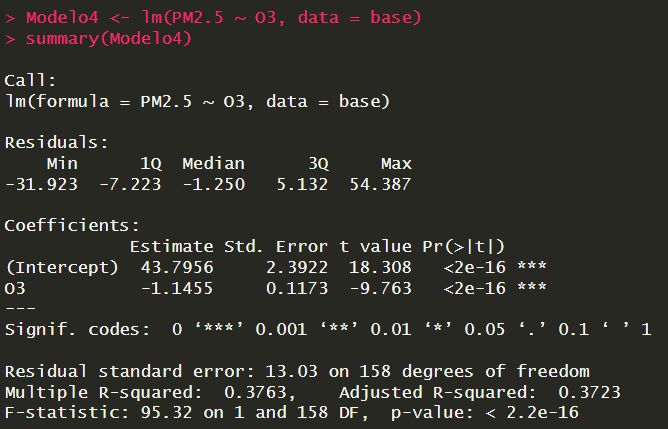
Multiple R-squared: 0.7071

1. **CO**



Multiple R-squared: 0.7055

1. **O3**



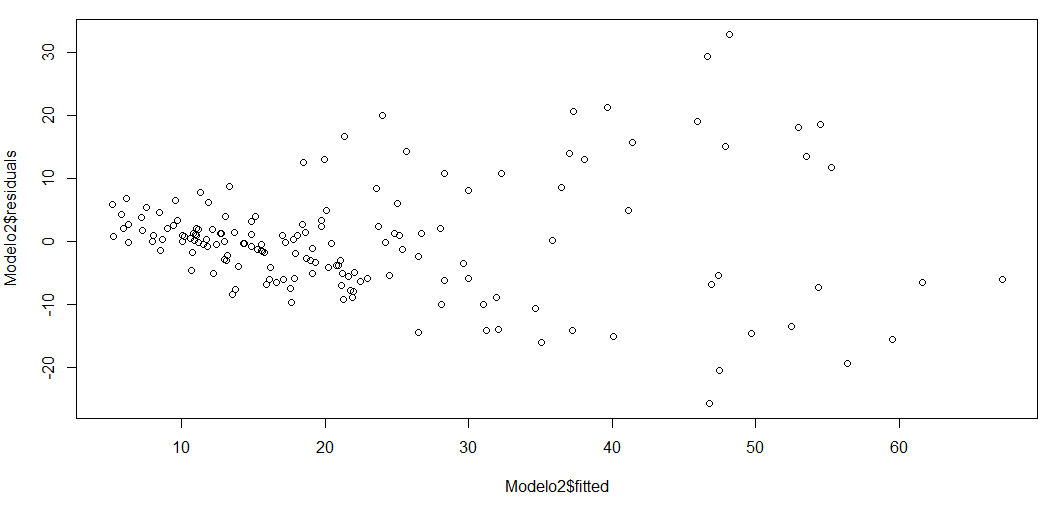
Multiple R-squared: 0.3763

Resp: el mejor modelo simple 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)

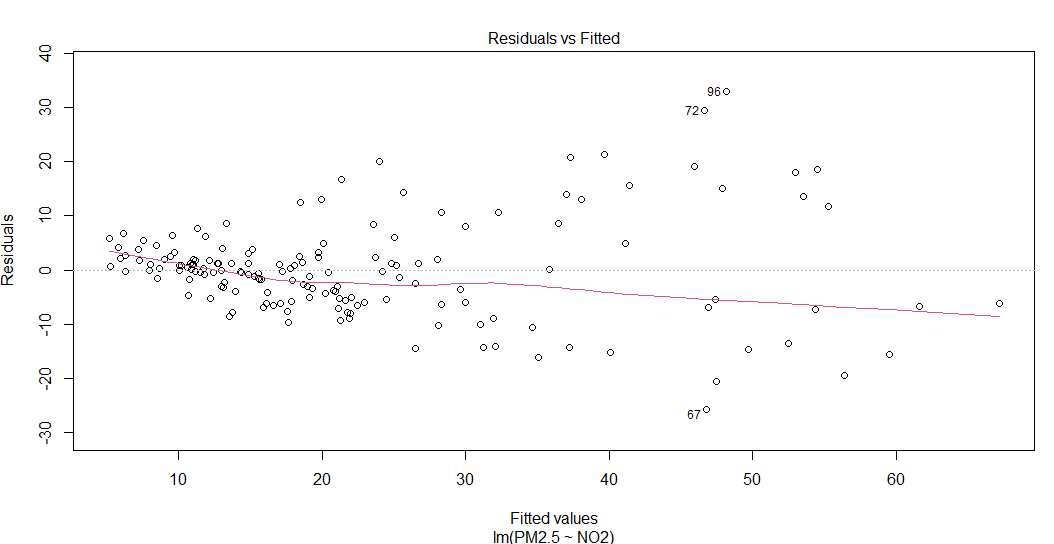
## Validación de los supuestos para el **“Modelo2”**

1. **Linealidad**

**plot(Modelo2$fitted, Modelo2$residuals)**



**plot(Modelo2, 1)**



Resp: Se acepta la linealidad

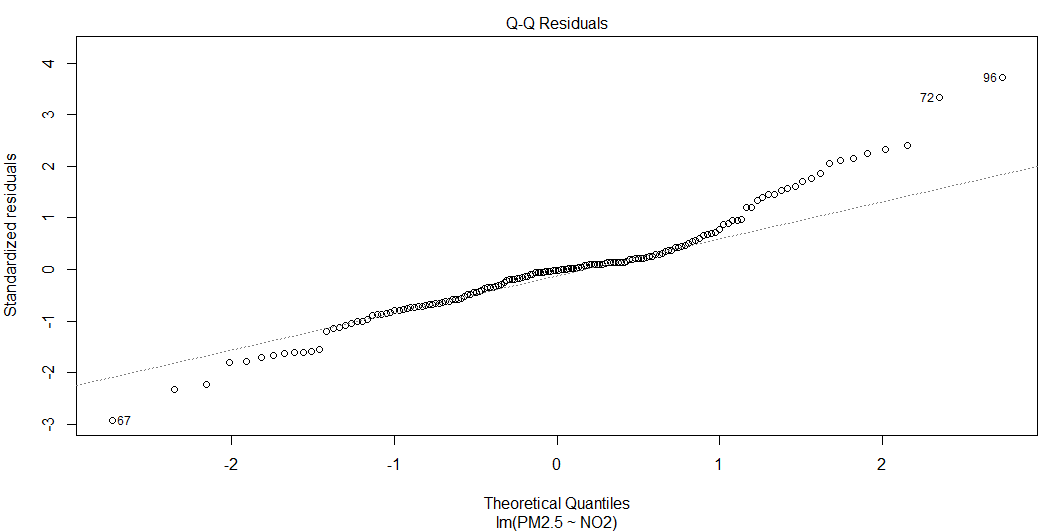
1. **Normalidad**

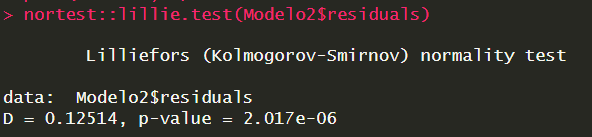
H0: residuos distribuyen normales

H1: residuos no distribuyen normales

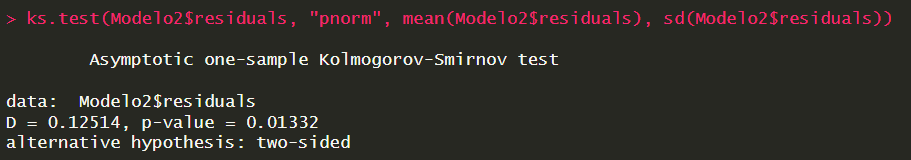
Gráficamente:

**plot(Modelo2, 2)**





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



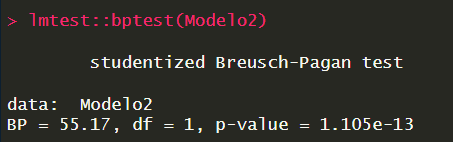
D = 0.12514, p-value = 0.01332 < alfa = 0.05 se rechaza H0

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

1. **Homocedasticidad**

H0: Sí existe Homocedasticidad

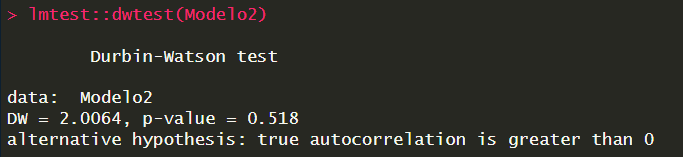
H1: No existe Homocedasticidad



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

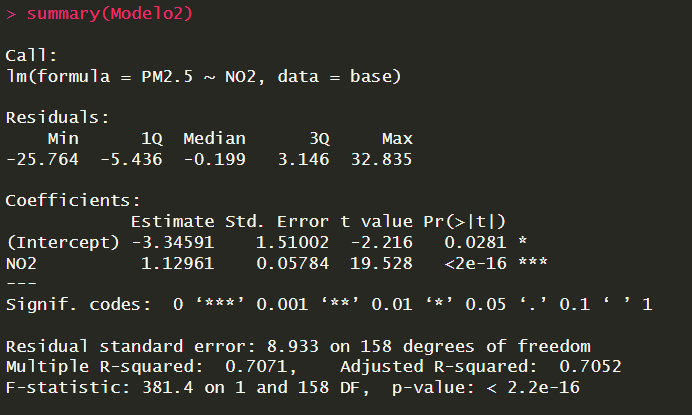
1. **Independencia**

H0: no hay autocorrelación vs H1: hay autocorrelación



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

## **Explicación del Modelo2**



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**

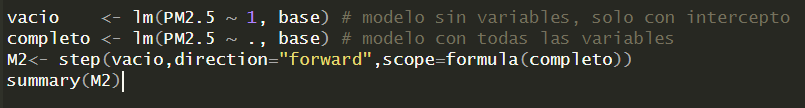
H0: no existe regresión vs H1: existe regresión

F-statistic: 381.4 on 1 and 158 DF, 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 70,7% el nivel de contaminación PM2.5

1. 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”**



**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

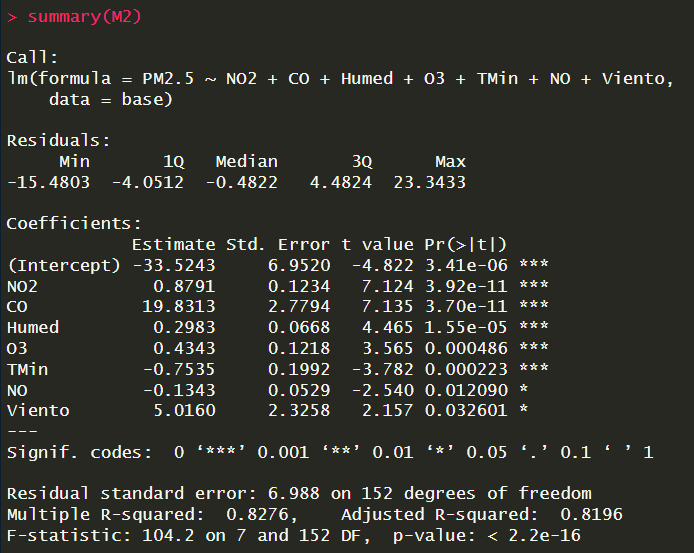
**Step 4: PM2.5 ~ NO2 + CO + Humed + O3**: Se agrega O3

**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

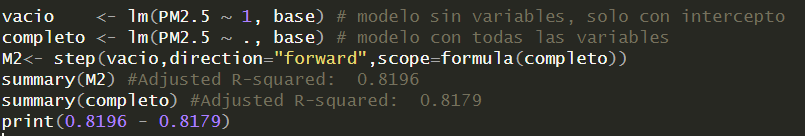
**Step 7: PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento**: Se agrega Viento

## Resumen con comando summary para el M2



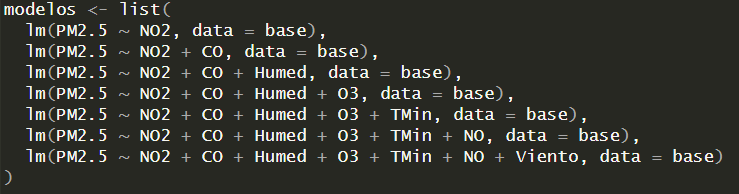
Adjusted R-squared: 0.8196 = 81.96% de calidad del modelo

Calculo del aumento del R2 ajustado



Resp: Para el caso forward hubo un aumento del R2 Ajustado de 0.0017 (0.17%)

## Detalle del R2 Ajustado por paso a paso del step



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

1. 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 | R2 |
| 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 | R2 |
| 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 % de explicación del modelo.

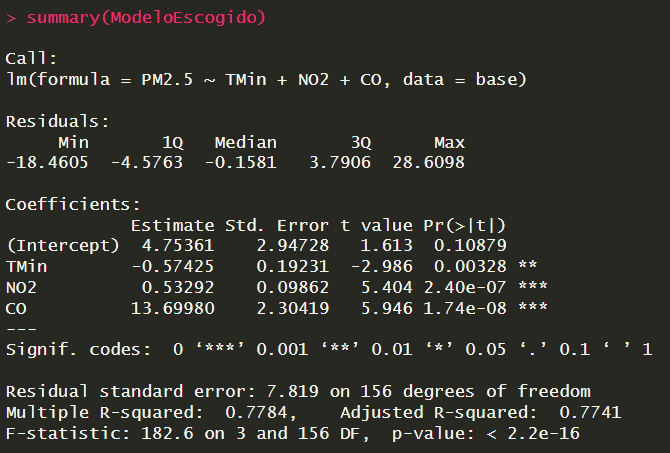
Para el caso de variable meteorológica: **Tmin**

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

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

## **Modelo con 3 predictores:**

**summary(ModeloEscogido)**



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

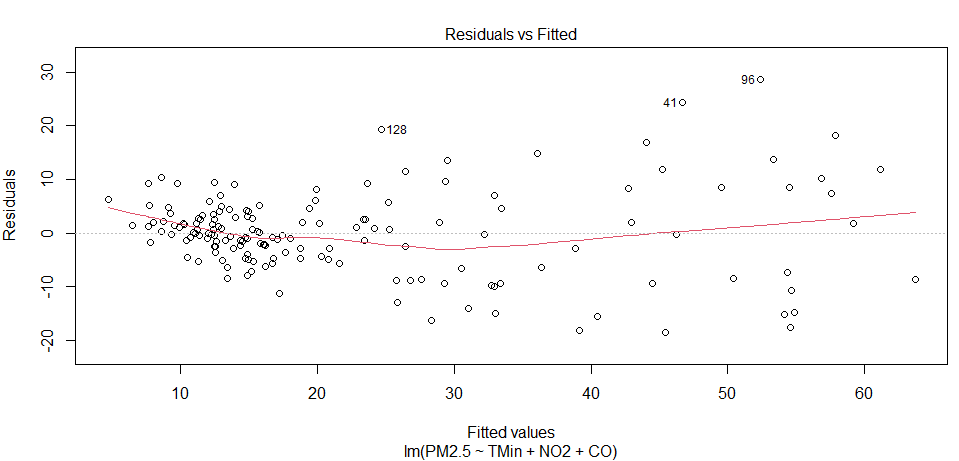
Adjusted R-squared: 0.7741

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

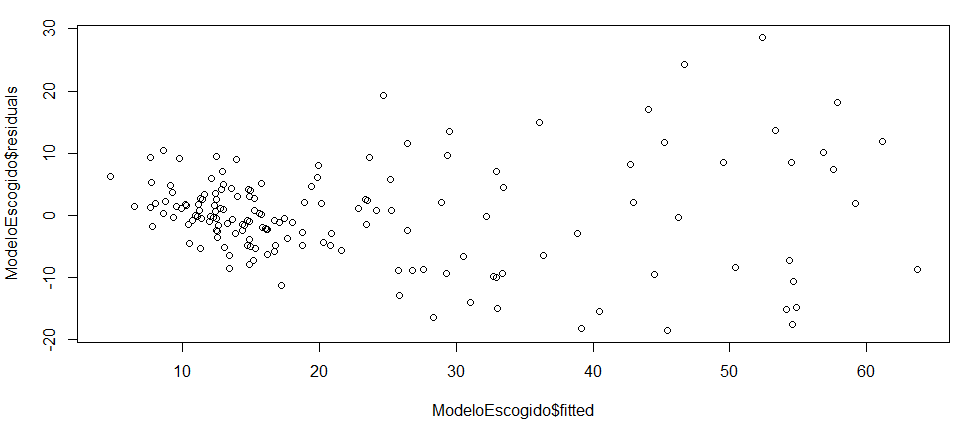
## Validación de los supuestos para el ModeloEscogido PM2.5 ~ TMin + NO2 + CO

1. **Linealidad**

**plot(ModeloEscogido, 1)**



**plot(ModeloEscogido$fitted, ModeloEscogido$residuals)**



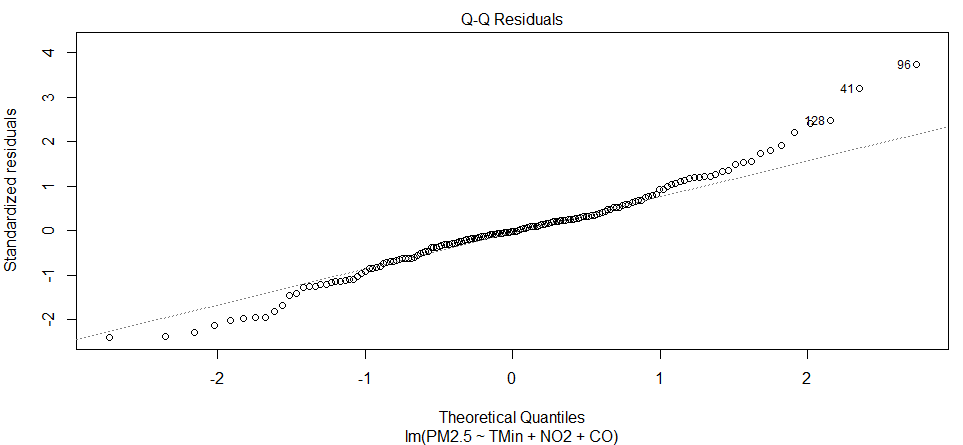
Resp: cumple linealidad

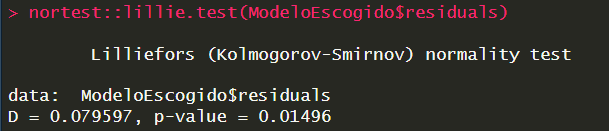
1. **Normalidad**

H0: residuos distribuyen normales

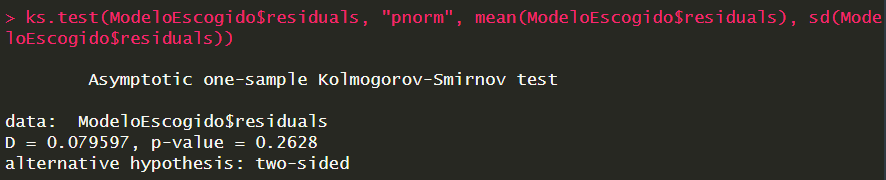
H1: residuos no distribuyen normales

Gráficamente:

**plot(ModeloEscogido, 2)**



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

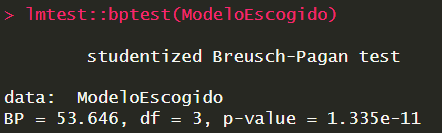


D = 0.079597, p-value = 0.2628 > alfa = 0.05 se acepta H0, hay normalidad según test ks

1. **Homocedasticidad**

H0: Sí existe Homocedasticidad

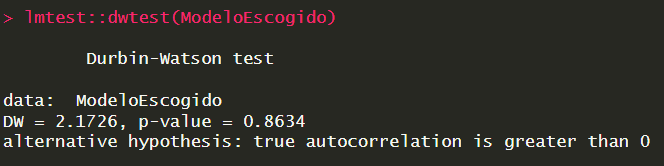
H1: No existe Homocedasticidad



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

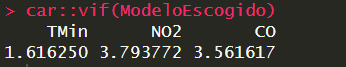
1. **Independencia**

H0: no hay autocorrelación vs H1: hay autocorrelación

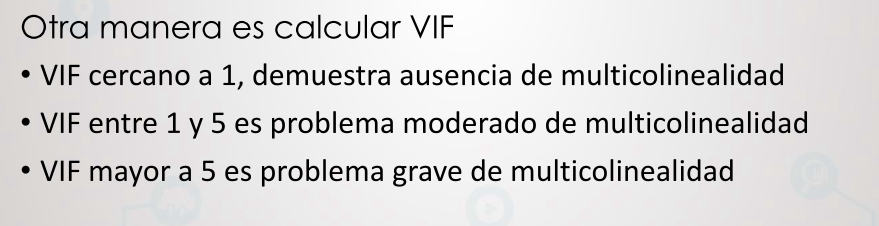


DW = 2.1726, p-value = 0.8634 > alfa = 0.05 , Se acepta H0, existe independencia

## Análisis de multicolinealidad



Según documentación



Resp: No hay multicolinealidad significativa en el modelo, ya que todos los VIF están por debajo de 5

Gráficos de relación entre todas las variables

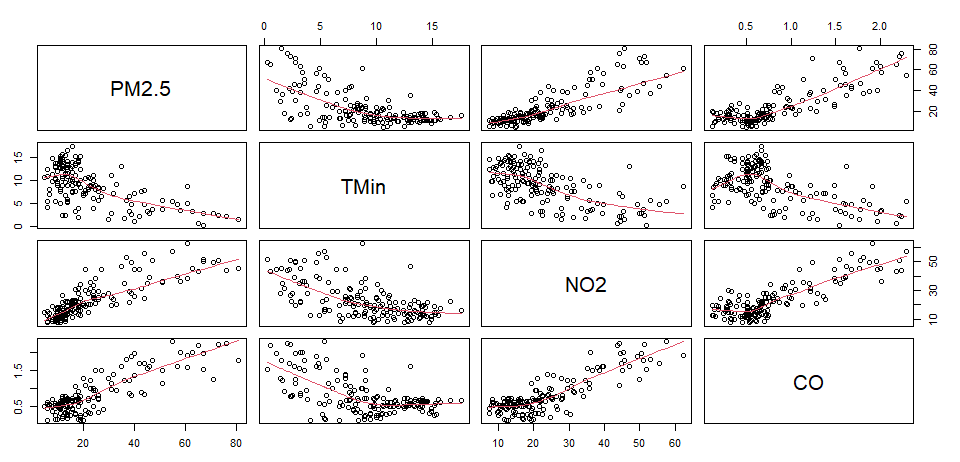
#Analisis de correlacion

base\_filtrada <- base[, c("PM2.5", "TMin", "NO2", "CO")]

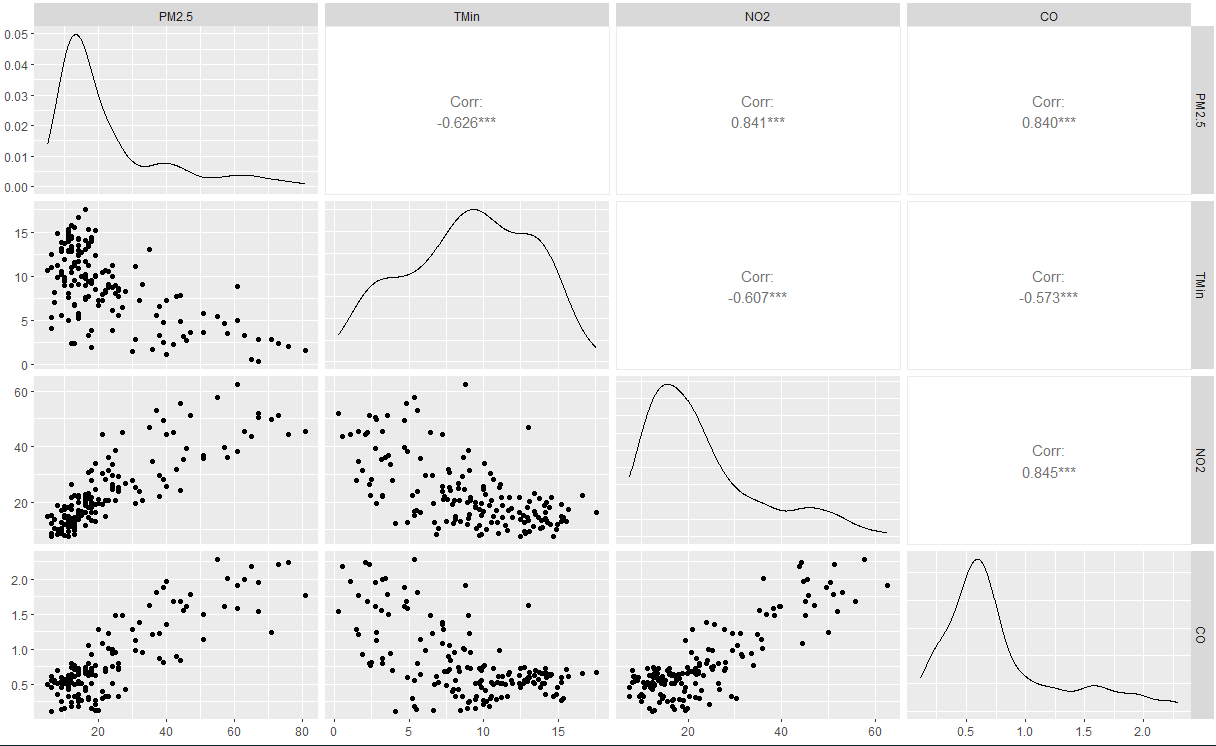
# Matriz de dispersión

library(GGally)

**pairs(base\_filtrada,upper.panel= panel.smooth, lower.panel = panel.smooth)**



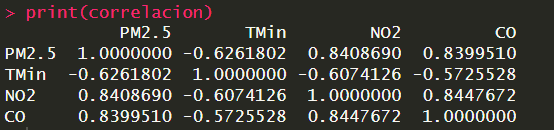
**ggpairs(base\_filtrada)**



## Cálculo de Correlaciones

correlacion <- cor(base\_filtrada)

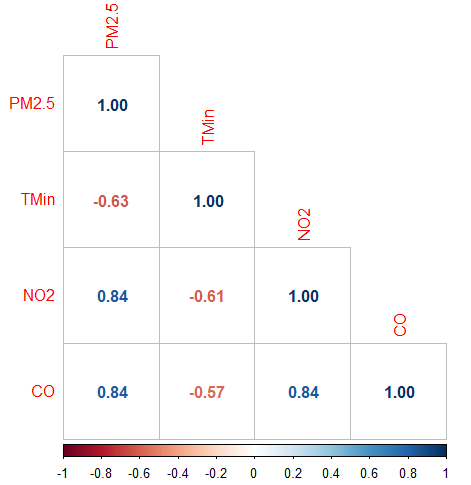
print(correlacion)



## Matriz de correlacion

library(corrplot)

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



## Coeficiente de determinación del modelo

Multiple R-squared: 0.7784, 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**

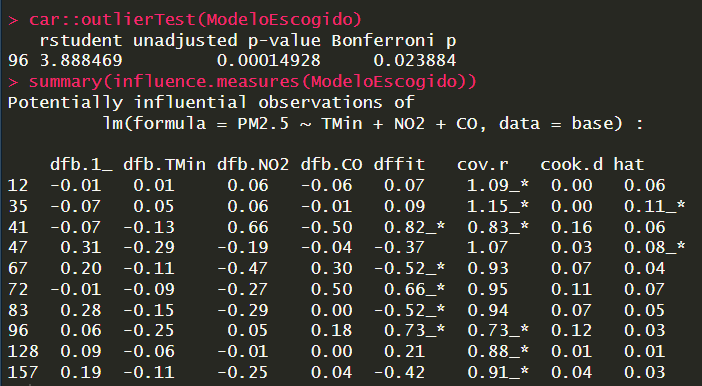
*> cor.test(base$NO2, base$CO)*

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.

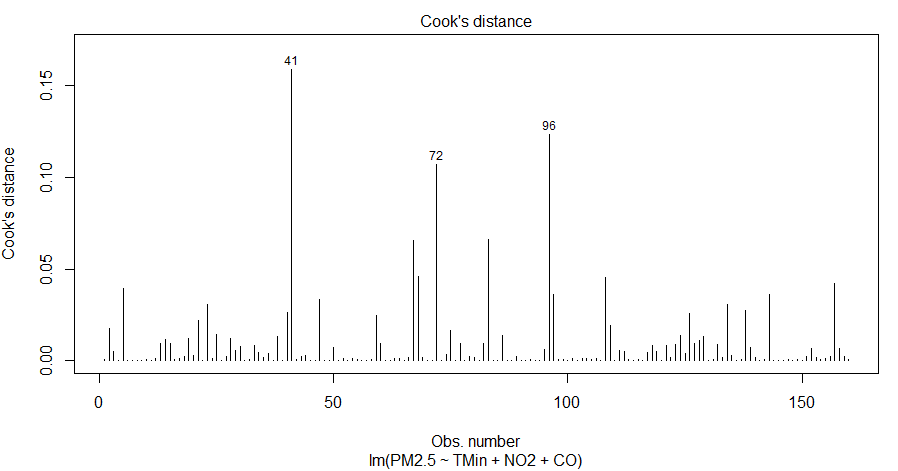
**Comentario general:** en todas las pruebas, los valores p son extremadamente bajos (mucho menores a 0.05), lo que indica que todas las correlaciones encontradas son estadísticamente significativas.

## Valores átipicos o influyentes



## Cook

**plot(ModeloEscogido,4)**



## Leverage

**plot(ModeloEscogido,5)**

