**Alumno: Rodrigo Jeldres Carrasco**

Control N2

Variable respuesta PM2.5

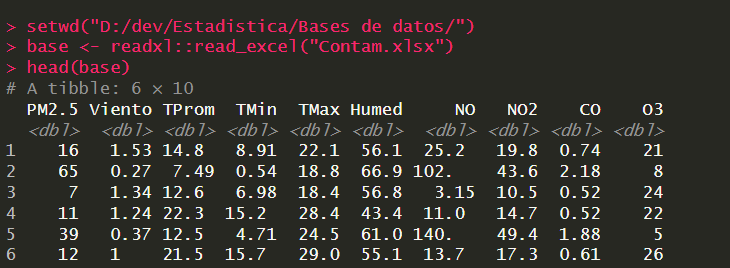
El objetivo es entender y explicar el comportamiento de los niveles de contaminación del aire en la Región Metropolitana. Para ello, desde el Sistema de Información Nacional de Calidad del Aire (sinca.mma.gob.cl), se seleccionó una muestra de la información histórica de la Estación Parque O’Higgins en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

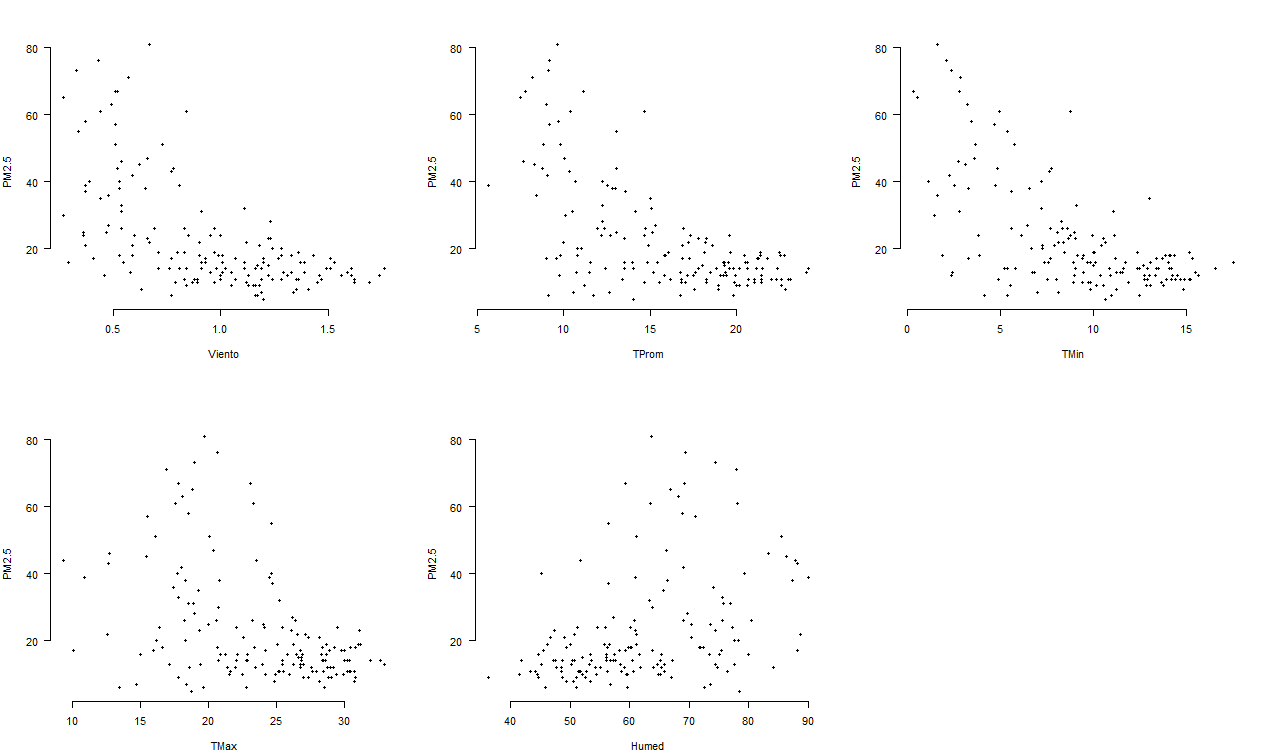
1. Obtenga el mejor modelo de regresión lineal simple basado en las variables meteorológicas.

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



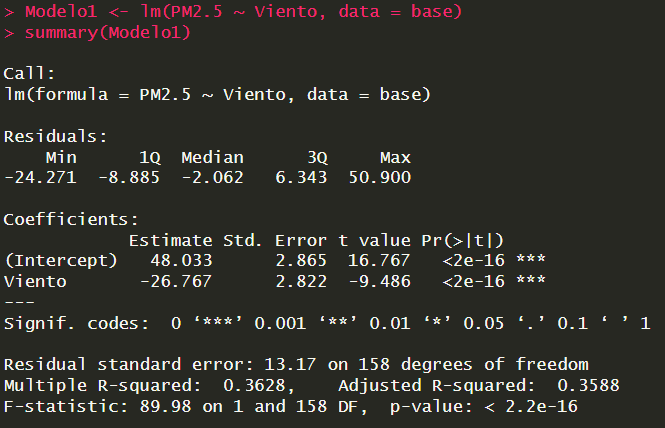
En este caso es Viento, TProm, TMin, TMax, Humed

Gráficamente:



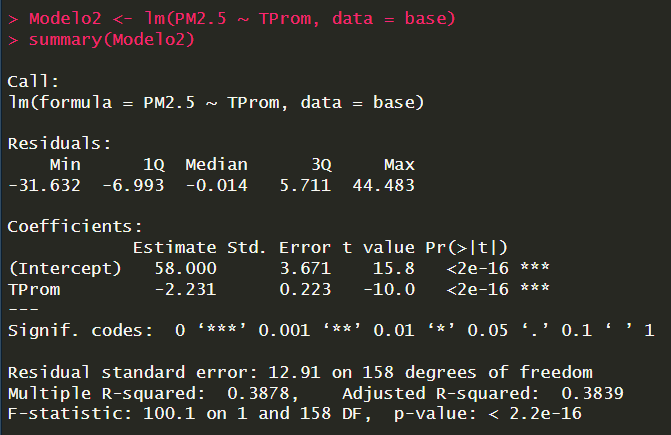
Para las distintas variables meteorológicas:

1. **Viento**



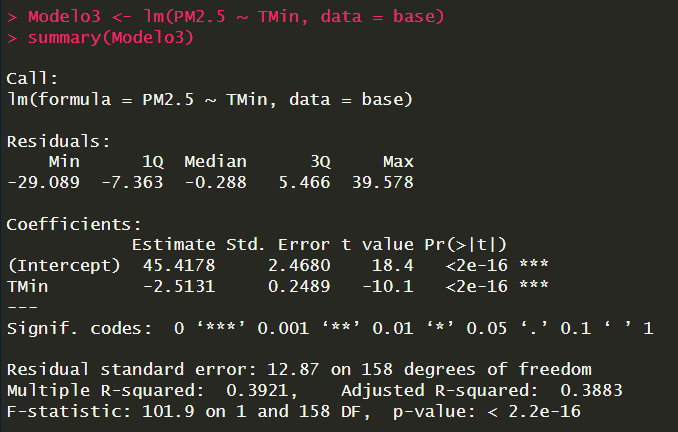
Multiple R-squared: 0.3628

1. **TProm**



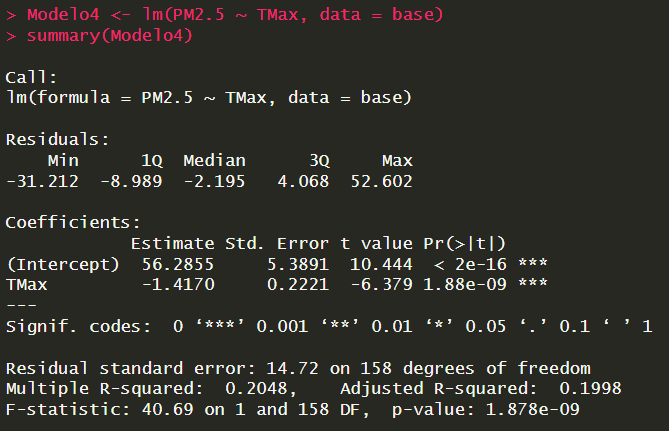
Multiple R-squared: 0.3878

1. **TMin**



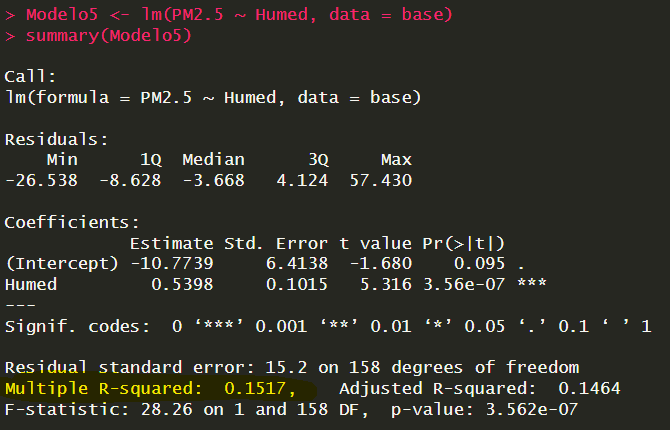
Multiple R-squared: 0.3921

1. **TMax**



Multiple R-squared: 0.2048

1. **Humed**



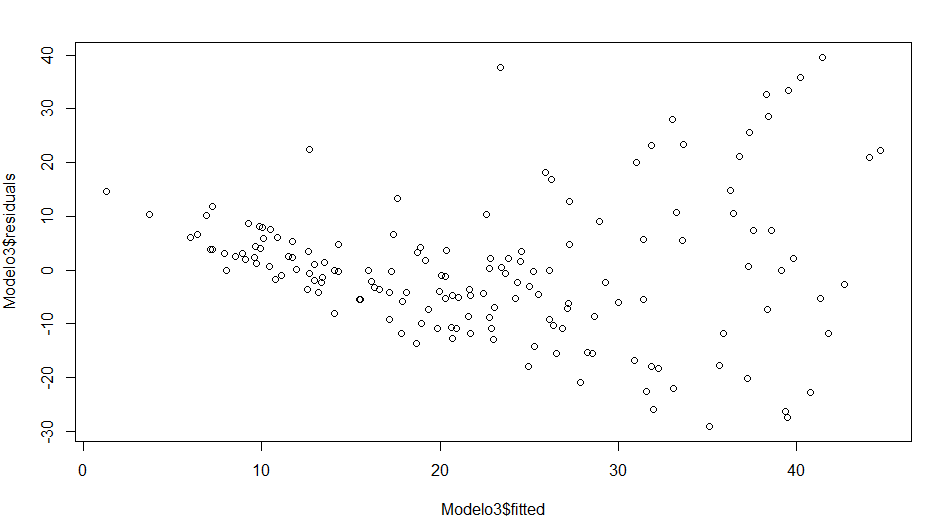
Multiple R-squared: 0.1517

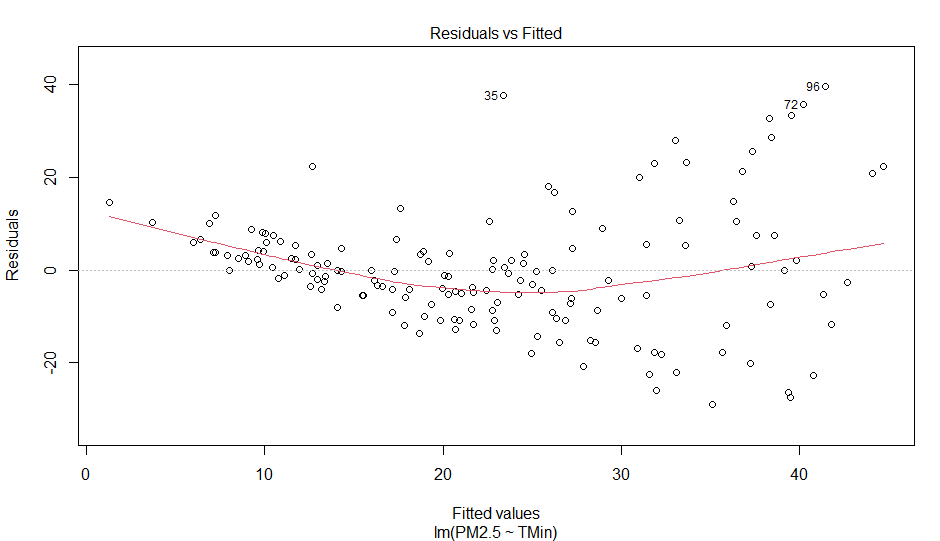
Resp: el mejor modelo simple con variable meteorológica es presentado con Modelo3 usando “Tmin” dado que tiene la mayor variable explicativa que es 39% (en R da con decimales: 0.3921017) Modelo3 <- lm(PM2.5 ~ TMin, data = base)

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

1. **Linealidad**

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



**plot(Modelo3, 1)**

Resp: Cumple linealidad

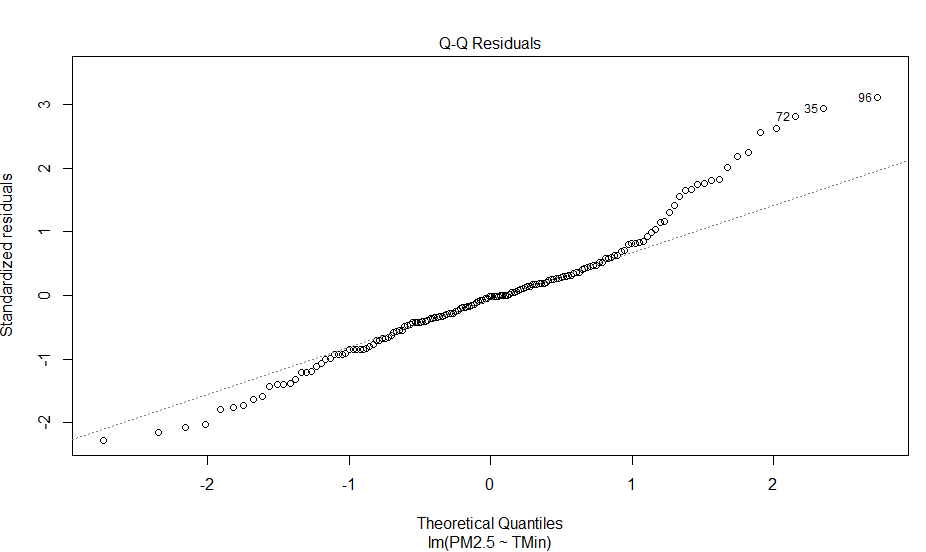
1. **Normalidad**

H0: residuos distribuyen normales

H1: residuos no distribuyen normales

Gráficamente:

**plot(Modelo3, 2)**



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

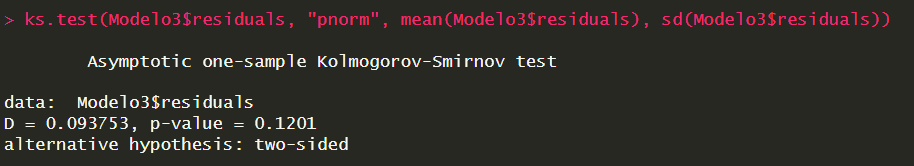
Lilliefors (Kolmogorov-Smirnov) normality test

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



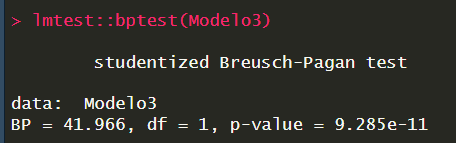
D = 0.093753, p-value = 0.1201, SÍ existe normalidad según Test KS

Resp: no hay normalidad en los residuos

1. **Homocedasticidad**

H0: Sí existe Homocedasticidad

H1: No existe Homocedasticidad

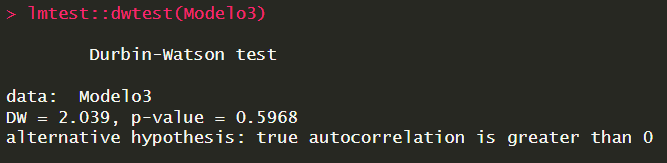


BP = 41.966, df = 1, p-value = 9.285e-11 < alfa = 0.05 ,

Resp: Se rechaza H0, no existe homocedasticidad

1. **Independencia**

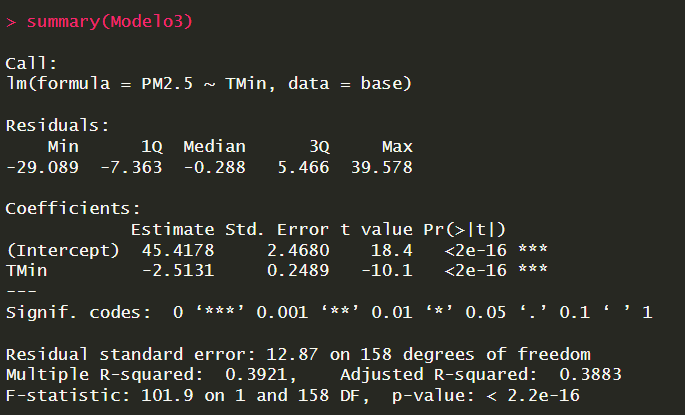
H0: no hay autocorrelación vs H1: hay autocorrelación



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

Resp: se acepta H0, existe independencia

**Resumen 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

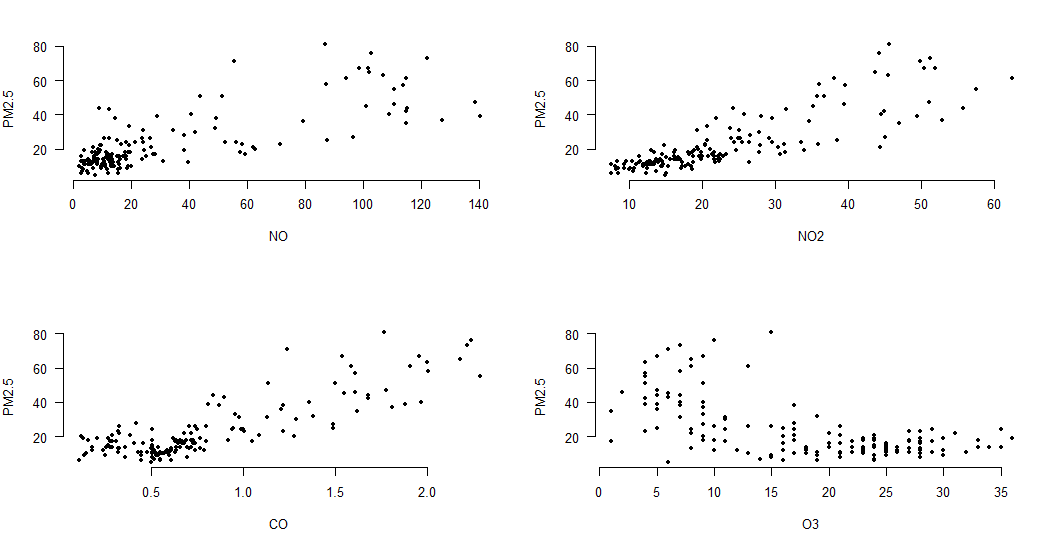
< alfa = 0.05, se rechaza H0, es decir, existe regresión

**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.

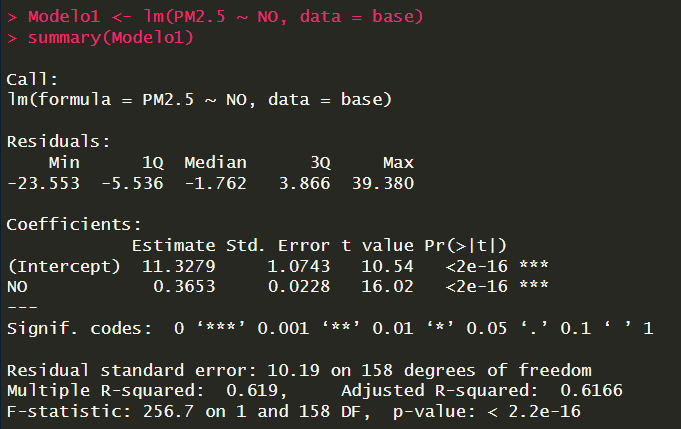
|  |
| --- |
| **Contaminantes atmosféricos** |
| NO – Monóxido de nitrógeno (ppb) |
| NO2 – Dióxido de nitrógeno (ppb) |
| CO – Monóxido de carbono (ppm) |
| O3 – Ozono (ppb) |

Graficamente:



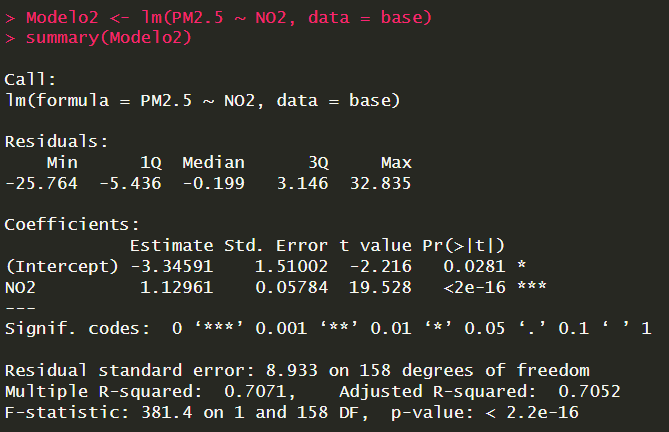
Para los distintos contaminantes atmosféricos:

1. **NO**



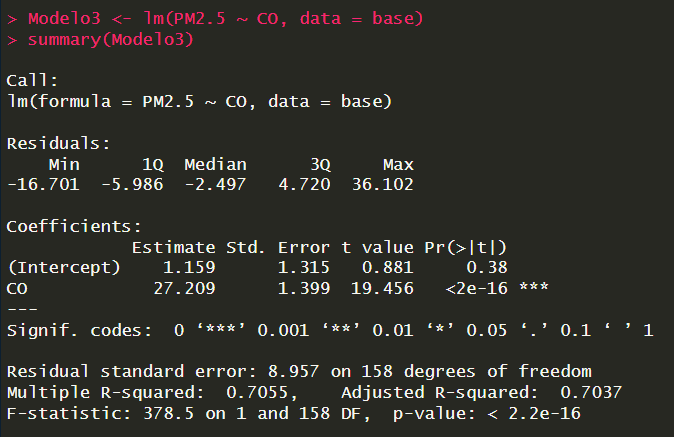
Multiple R-squared: 0.619

1. **NO2**



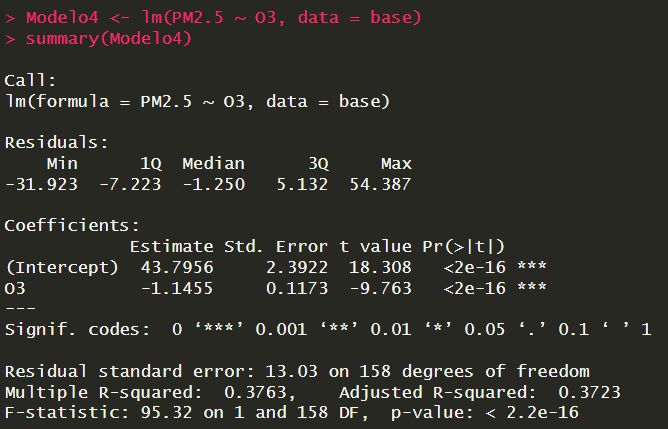
Multiple R-squared: 0.7071

1. **CO**



Multiple R-squared: 0.7055

1. **O3**



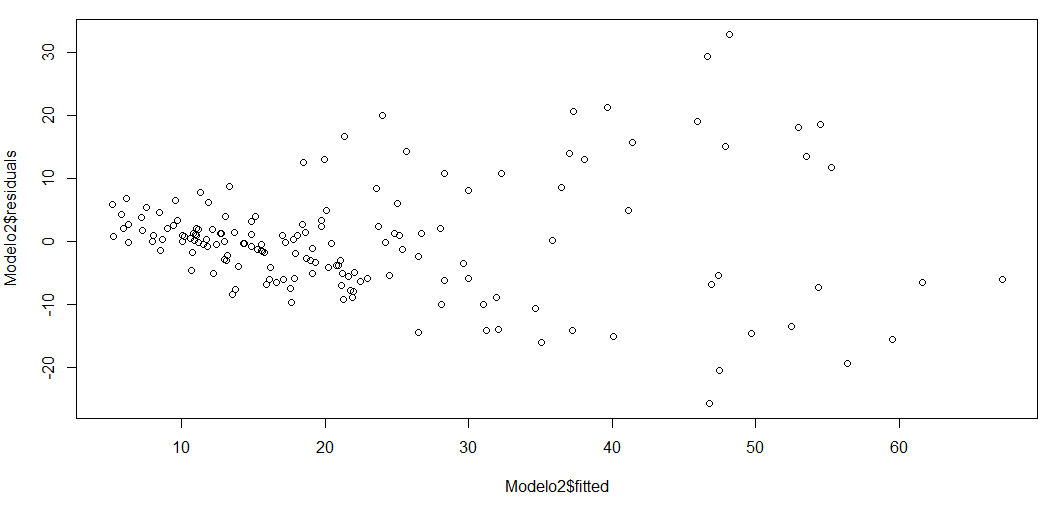
Multiple R-squared: 0.3763

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)

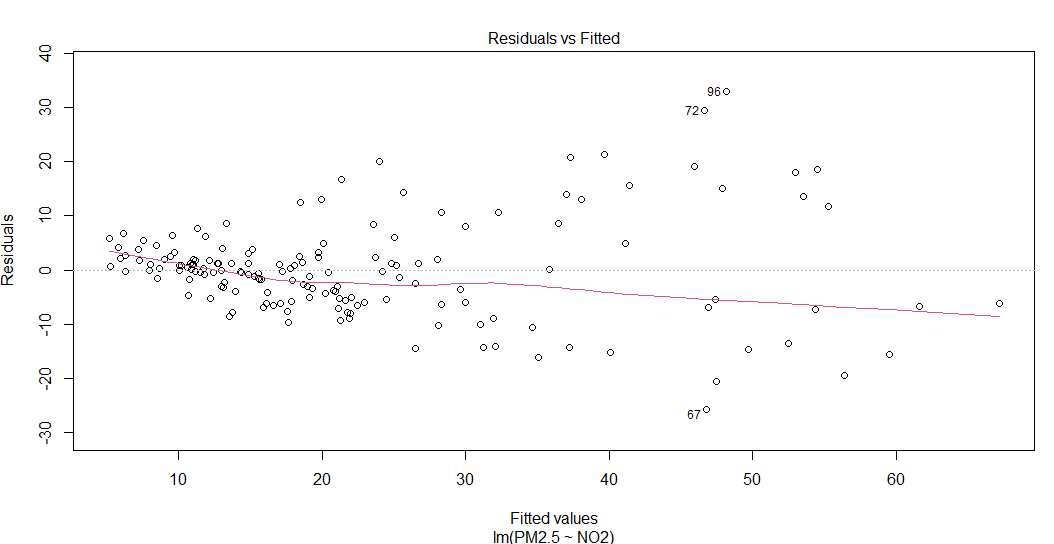
Validación de los supuestos para el **“Modelo2”**

1. **Linealidad**

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



**plot(Modelo2, 1)**



Resp: Se acepta la linealidad

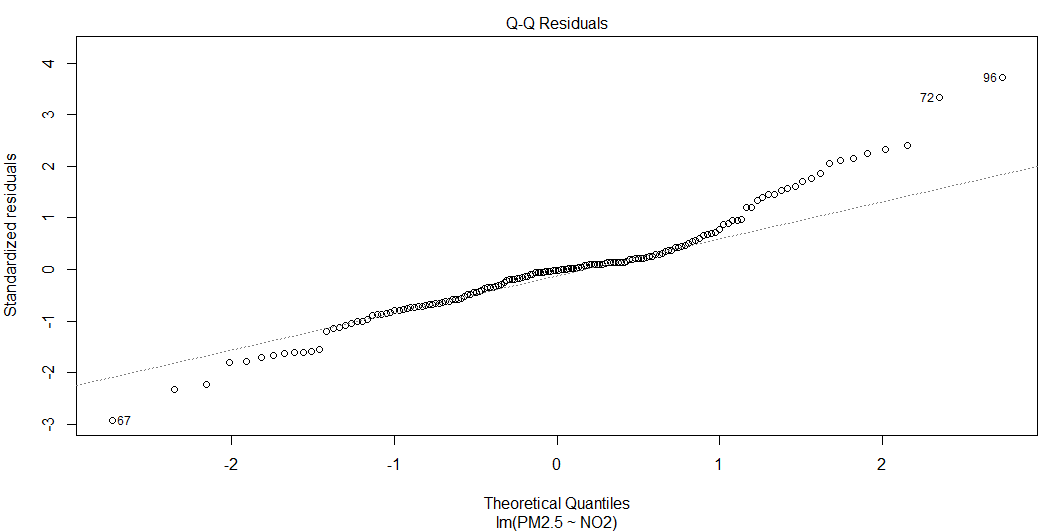
1. **Normalidad**

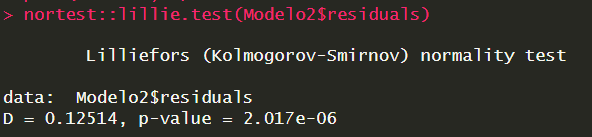
H0: residuos distribuyen normales

H1: residuos no distribuyen normales

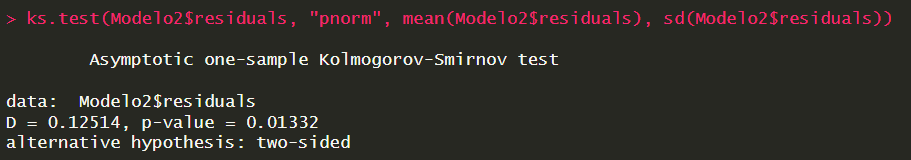
Gráficamente:

**plot(Modelo2, 2)**





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



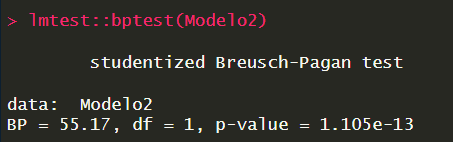
D = 0.12514, p-value = 0.01332 < alfa = 0.05 se rechaza H0

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

1. **Homocedasticidad**

H0: Sí existe Homocedasticidad

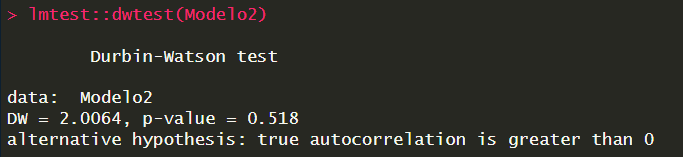
H1: No existe Homocedasticidad



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

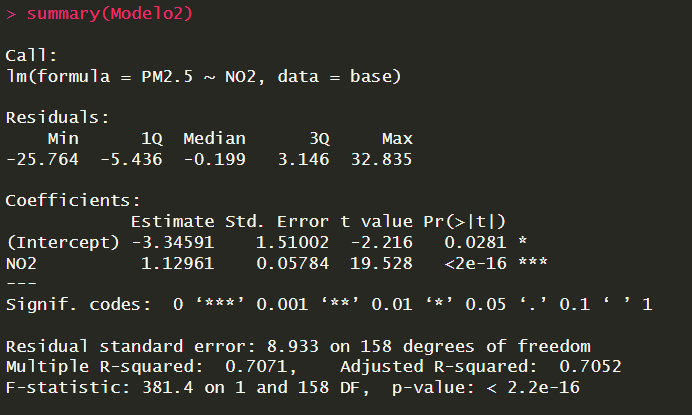
1. **Independencia**

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 a alfa = 0.05; Se rechaza H0

Para un alfa tan pequeño, la variable NO2 es significativa

**Test de significancia del modelo**

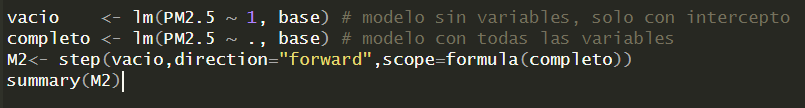
H0: no existe regresión vs H1: existe regresión

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 un70,7% 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

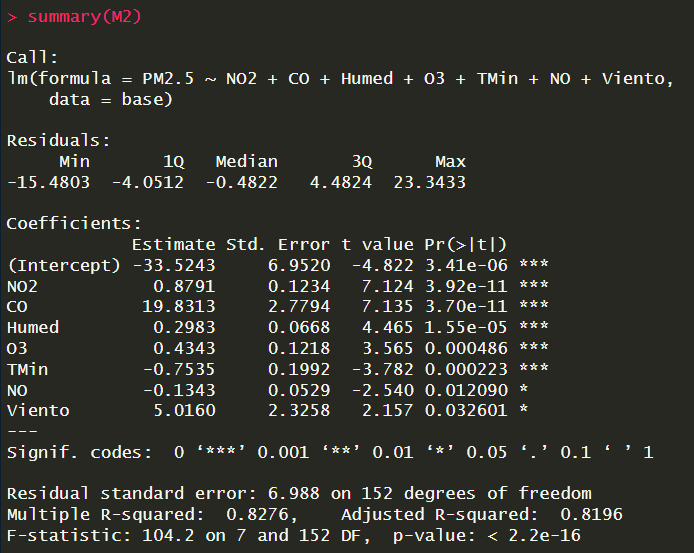
**Step 4: PM2.5 ~ NO2 + CO + Humed + O3**: Se agrega O3

**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

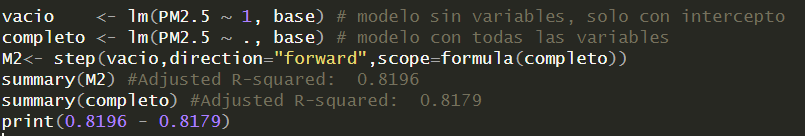
**Step 7: PM2.5 ~ NO2 + CO + Humed + O3 + TMin + NO + Viento**: Se agrega Viento

Resumen con comando summary



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%)