Situación problema: Contaminación del Aire en las grandes ciudades

Susana Leija (A00834061), Jorge Vincenzo Rizo (A00836399), José Rogelio Ruiz (A00835536)

2023-02-07

Análisis Estadístico Grupo 101

Porcentaje de participación:

Susana Leija: 33% Jorge Vincenzo: 33% *José Rogelio Ruiz*: 33%

Fecha de Entrega: 07 de febrero de 2023

Introducción

La contaminación del aire consiste en la presencia de particulas sólidas y gases en el aire que alteran las proporciones de lo que normalmente se encuentra en la atmósfera. Estos gases y particulas contaminantes causan problemas en los seres vivos expuestos a ellos, causando alrededor de 12.6 millones de muertes al año en el mundo, y cerca 9 mil 300 muertes al año asociadas con la contaminación del aire en México, según datos de 2019 de la OMS.

La ciudad de México en particular tiene graves problemas de contaminación, siendo la quinta ciudad más contaminada del mundo según Greenpace, por lo que se buscan soluciones. Para ello se realizan estudios anuales de la Calidad del Aire que se obtienen de 34 estaciones meteorológicas y se guardan en seis bases de datos según sus características.

En este reporte se analiza la información obtenida en la estación de "Miguel Hidalgo" en 2022, guardada en la base de datos MGH2022.csv, con el propósito de buscar una relación entre los niveles de diferentes variables de gases con una variable dependiente.

: ¿los niveles de 03 se ven influenciados ya sea de manera positiva o negativa por los niveles de CO, NO, NO2 y NOX? Teniendo como variable de respuesta o dependiente los niveles de Ozono en la atmósfera y de variables independientes o regresoras al Monóxido de carbono, Óxido nítrico, Dióxido de nitrógeno y Óxidos de Nitrógeno diferentes.

Se decidió investigar sobre el ozono como variable respuesta debido a su importancia en la protección de la atmosfera de rayos ultravioleta emitidos por el sol. Aunque la relevancia del mantenimiento de esta capa de gas en la atmósfera ya fue atendida en los 70s, se estima que el agujero producido en la Antártida por el uso de clorofluorocarbonos en aerosoles ya ha causado un daño que se tardará hasta después de la mitad de este siglo en recuperarse, y la fecha solo sigue siendo retrasado, lo que, según NASA , se debe al incremento de la temperatura del planeta.

Debido a esto, elegimos este gas como variable respuesta, para detectar alguna posible relación entre este y los demás gases y poder hacer algo al respecto, en caso de que sea perjudicial para el ambiente.

Para realizar este análisis, se preparará una muestra aleatoria de 800 datos, se harán análisis de correlación entre las variables, se compararán en gráficas, se harán análisis de regresión lineales y curvilíneos, así como un análisis cúbico y otro múltiple. Al final se sacarán conclusiones en base a los resultados.

Desarrollo

Preparación de los datos

Para poder utilizar los datos de la estación Miguel Hidalgo, primero tenemos que asegurarnos de que estén como los queremos. Teniendo el archivo MGH2022.csv con la base de datos lo cargamos a una variable "M" con el comando read.csv()

Leyendo los datos

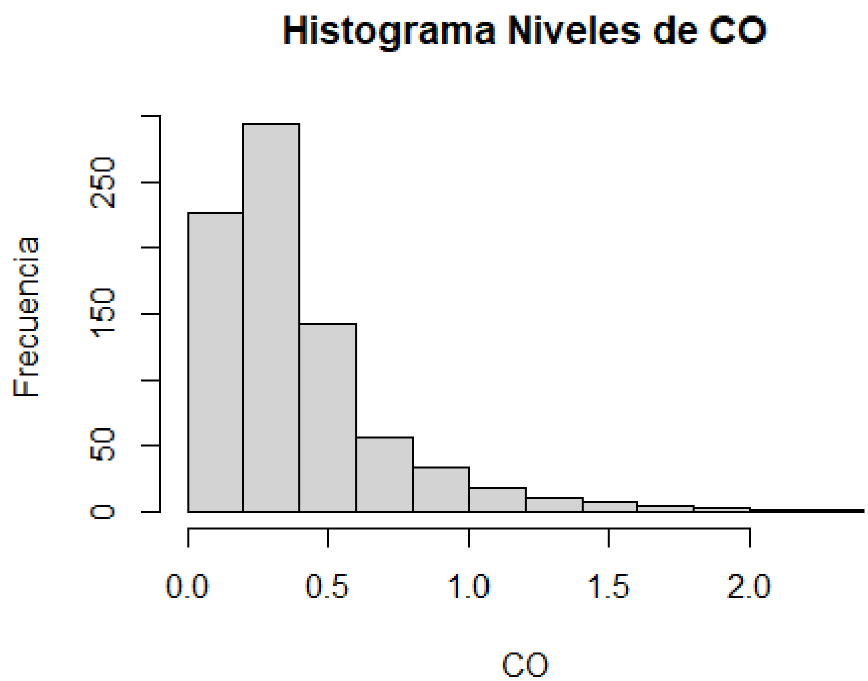
Podemos ver que se trata de una base de datos de 8016 observaciones y 7 variables, de las cuales 5 son numéricas y 2 son categóricas (fecha y hora). Para tener solamente las variables que nos interesan, seleccionamos entre corchetes y con la función "c()" solo de las columnas 3 a 7 que contienen las 5 variables numéricas con los niveles de cada gas.

Depués, analizando la base de datos, pudimos ver que los valores "-99" representaban valores no capturados, por lo que los reemplazamos con "NA", que representa valores faltantes en r, para poder utilizar la función na.omit() y evitar que afecten nuestros resultados.

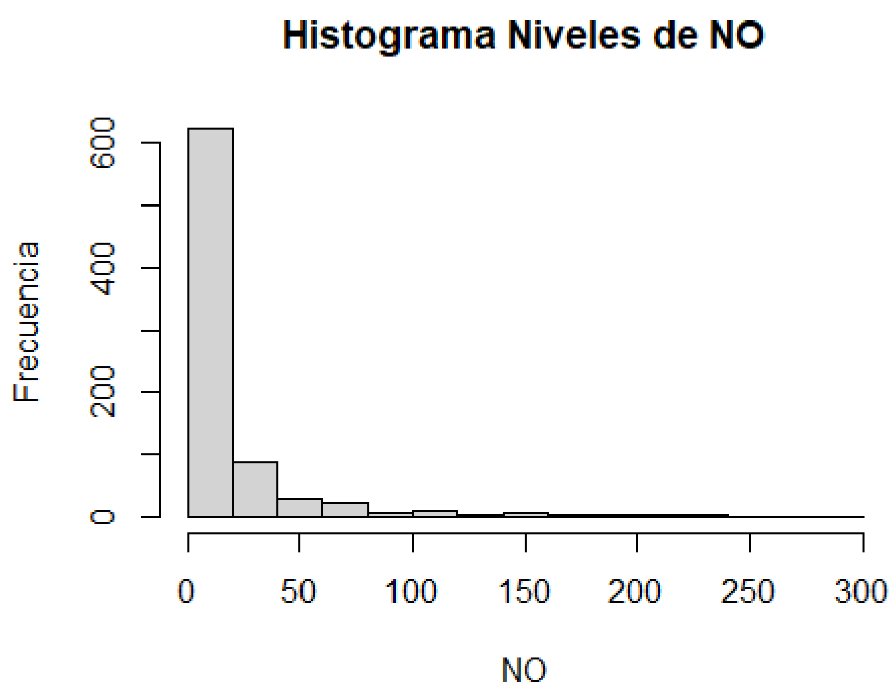
Hacemos una lista aleatoria de 800 filas utilizando la función sample() para sacar 800 valores aleatorios entre 1 y 6845, lo que representa los datos que quedaron después de eliminar los datos faltantes, después utilizamos esa lista cómo índice para sacar la muestra de la base de datos y la nombramos MGH2022_muestra-csv. Con esta muestra se trabajará en los análisis.

Exploración de los datos

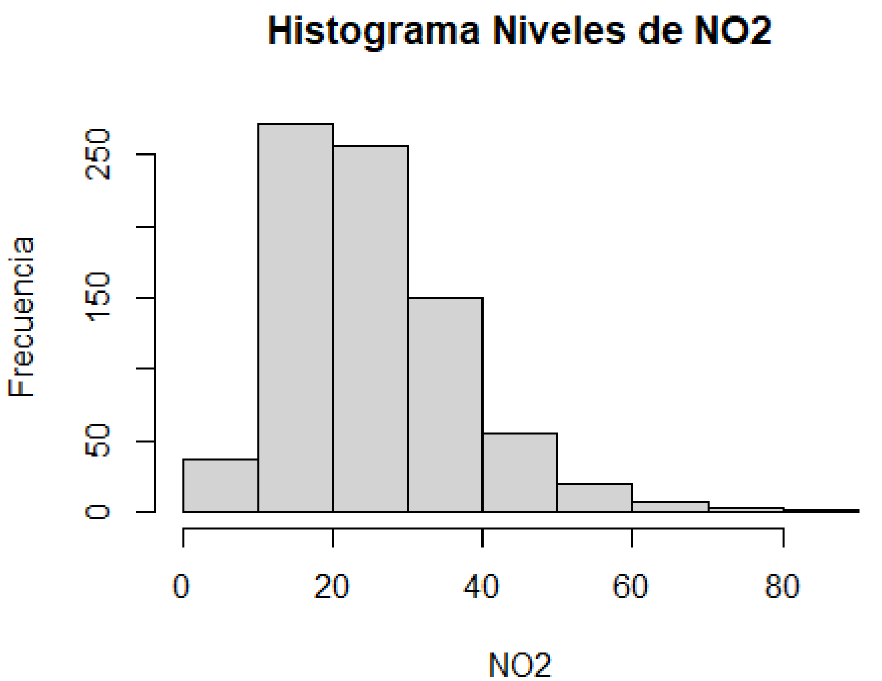
Para explorar los datos y darnos una idea de sus magnitudes y frecuencia, realizamos histogramas para cada una de las variables de los gases en la muestra aleatoria.

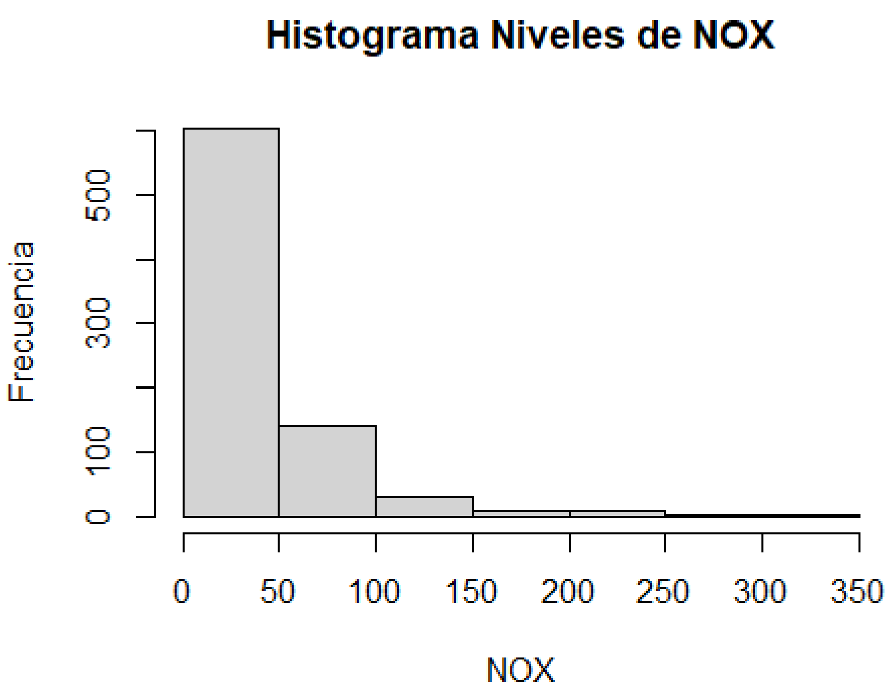


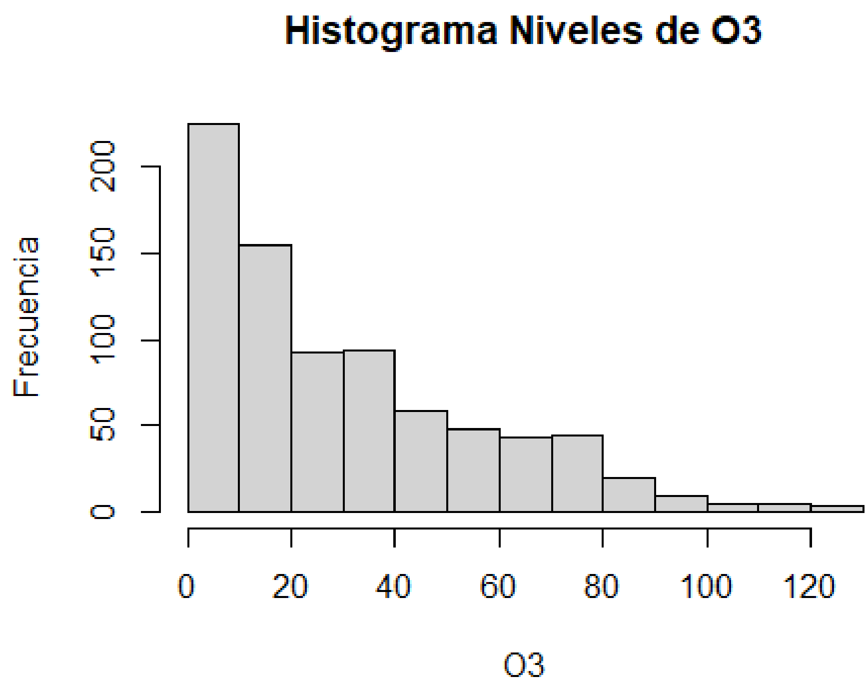
Histograma Niveles de CO
Frecuencia
0
50
150
250
0.0
0.5
1.0
1.5
2.0
CO



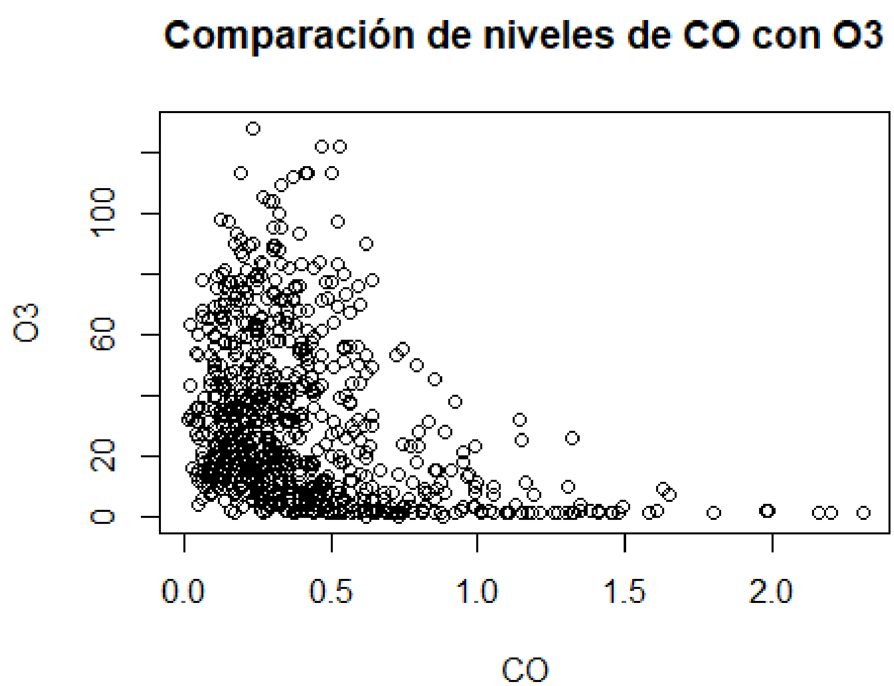
Histograma Niveles de NO
Frecuencia
0
200
400
600
0
50
100
150
200
250
300
NO

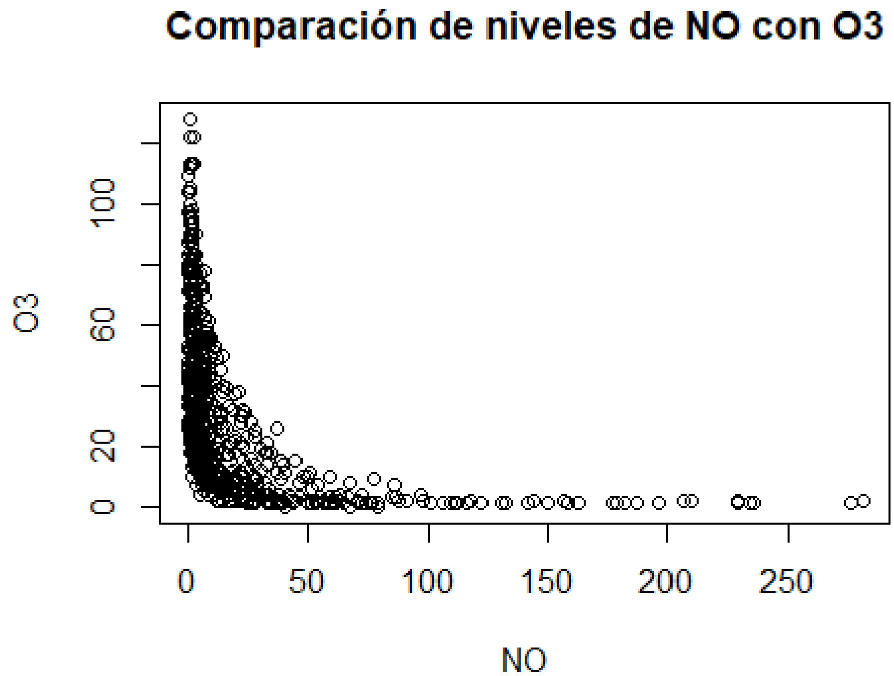

Histograma Niveles de NO2
Frecuencia
0
50
150
250
0
20
40
60
80
NO2

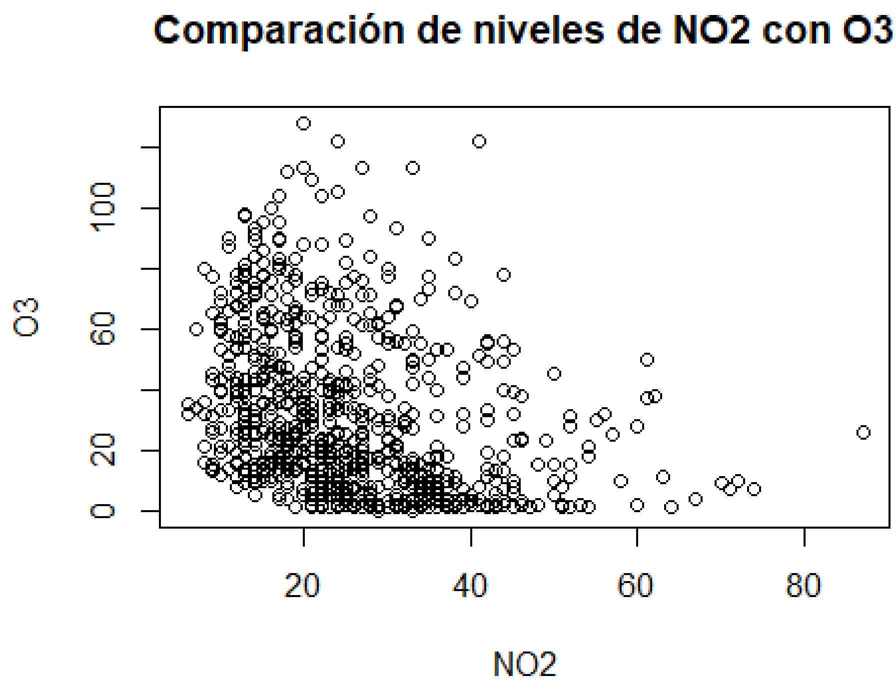

Histograma Niveles de NOX
Frecuencia
0
100
300
500
0
50
100
150
200
250
300
350
NOX

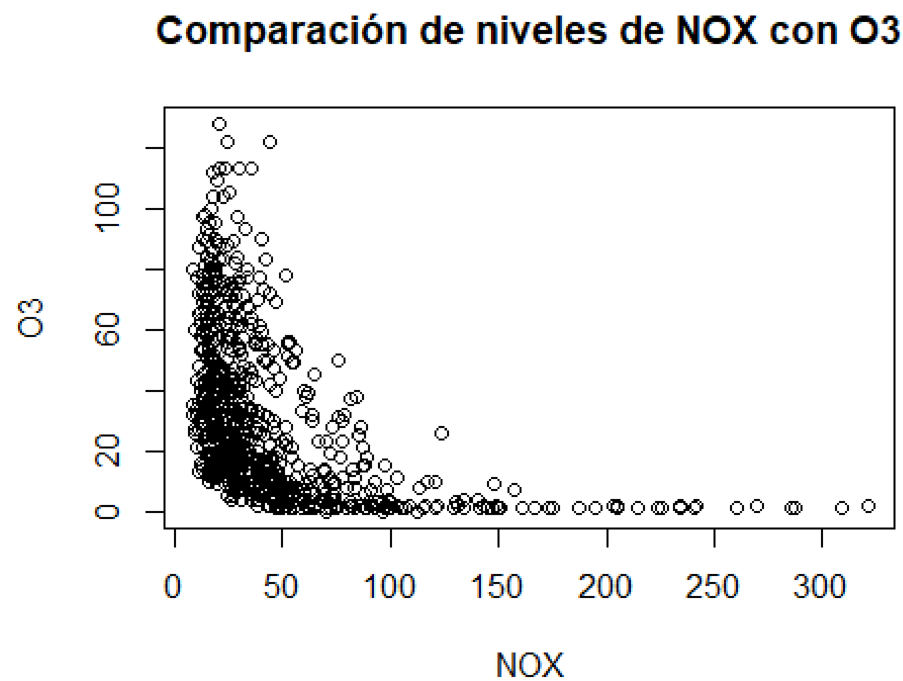


Se compararon también cada una de las variables independientes (CO, NO, NO2, NOX) con la variable dependiente elegida (O3), para observar si había alguna tendencia muy evidente, lo cuál no fue el caso.




Comparación de niveles de NO con O3
O3
0
20
60
100
0
50
100
150
200
250
NO


Comparación de niveles de NO2 con O3
O3
0
20
60
100
20
40
60
80
NO2



Calculamos la varianza de la media para cada una de las variables.

```
## [1] 0.1092641

## [1] 1252.627

## [1] 137.4833

## [1] 1773.139

## [1] 703.2597
```

Y calculamos las desviaciones estándar.

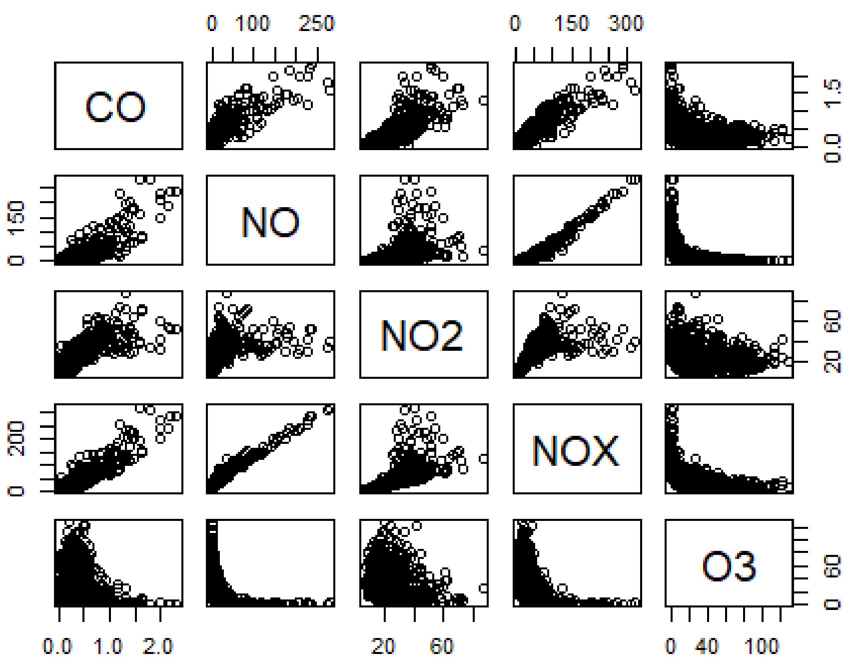
```
## [1] 0.3305513

## [1] 35.39248

## [1] 11.72533

## [1] 42.10866

## [1] 26.51904
```

Con este análisis de exploración de los datos pudimos ver que los datos que más se repiten son los más bajos o cercanos a 0, lo que tiene sentido ya que ver niveles anormalmente altos de estos gases en el ambiente no es tan común, o al menos eso se esperaría. También podemos ver que hay una varianza bastante alta en todas las variables menos monóxido de carbono, pero no podemos saber todavía si esto tiene relevancia a nuestra pregunta a analizar.

Análisis de correlación

Para determinar si las variables tienen correlación entre sí o no se realizó un análisis de correlación de pearson a través de r, utilizando la función cor() con la muestra aleatoria, y se graicaron todas las combinaciones entre las variables para comparar los resultados.

```
##             CO         NO        NO2        NOX         O3
## CO   1.0000000  0.8121584  0.7654448  0.8957388 -0.3158941
## NO   0.8121584  1.0000000  0.4616217  0.9688747 -0.4210994
## NO2  0.7654448  0.4616217  1.0000000  0.6665968 -0.3291815
## NOX  0.8957388  0.9688747  0.6665968  1.0000000 -0.4456237
## O3  -0.3158941 -0.4210994 -0.3291815 -0.4456237  1.0000000
```



Se realizó la prueba de correlación de pearson entre O3 y cada variable individualmente para corroborar que sean los mismos y ver los resultados uno por uno.

```
## [1] -0.3158941

## [1] -0.4210994

## [1] -0.3291815

## [1] -0.4456237
```

Se puede observar que si existen correlaciones significativamente diferentes a 0 entre cada variable y el O3, con la más baja siendo -0.31, y que en todos los casos la correlación es negativa, por lo que mientras los niveles de ozono aumentan las demás variables disminuyen un poco y viceversa.

Pruebas de Hipótesis de Correlación

Para determinar si las correlaciones entre las variables y el O3 son significativas se realiza una prueba de hipótesis para cada una apoyandonos de la librería Hmisc. Esta librería da como output dos matrices, una con los coeficientes de correlación entre cada variable y otra con los valores p para cadauna de las correlaciones.

$H_0: \rho = 0$ $H_1: \rho \neq 0$ $\alpha = 0.05$

Regla de decisión: Si valor p es menor a alfa entonces se rechaza Ho.

Realiza el análisis del resultado

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

##        CO    NO   NO2   NOX    O3
## CO   1.00  0.81  0.77  0.90 -0.32
## NO   0.81  1.00  0.46  0.97 -0.42
## NO2  0.77  0.46  1.00  0.67 -0.33
## NOX  0.90  0.97  0.67  1.00 -0.45
## O3  -0.32 -0.42 -0.33 -0.45  1.00
##
## n= 800
##
##
## P
##     CO NO NO2 NOX O3
## CO      0  0   0   0
## NO   0     0   0   0
## NO2  0  0      0   0
## NOX  0  0  0       0
## O3   0  0  0   0
```

Tomando en cuenta la prueba de hipótesis con nivel de significación de 0.05, podemos concluir que, cómo el valor p es < a alfa en todos los test de correlacion, entonces se rechaza Ho para todas las variables, lo que significa que si existe una correlacion significativa respecto a O3 con CO, NO, NO2 y NOX.

Análisis de Regresión

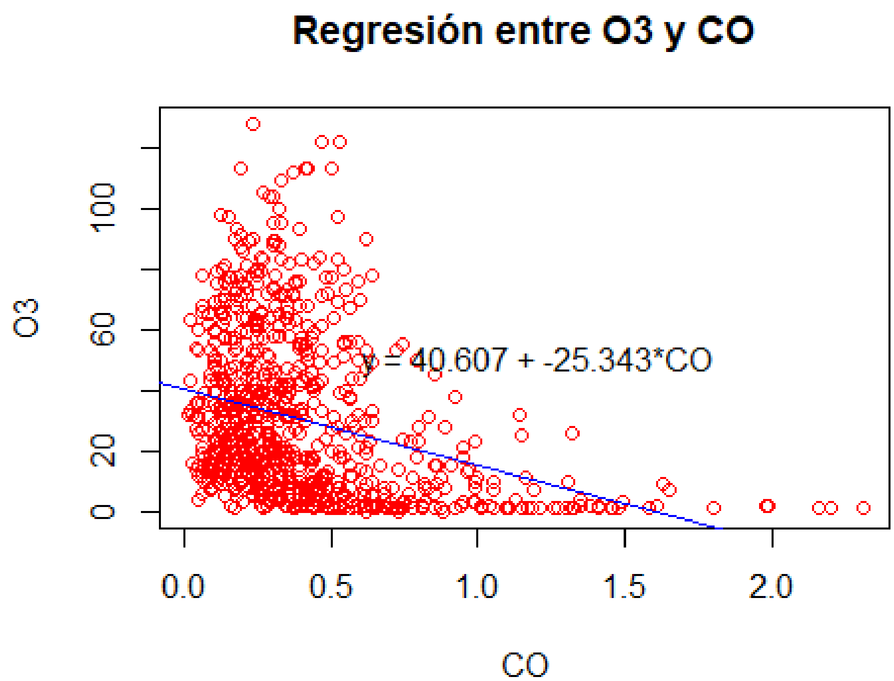
Ahora se realizarán varios análisis de regresión comparando cada una de las variables con la variable respuesta O3 utilizando técnicas de regresión lineal, múltiple, curvilíneo y cúbico. Para determinar la validez y precisión de los modelos debemos fijarnos en el coeficiente de determinación y las betas deben de ser significativas.

Regresion lineal simple

Se realizan modelos de regresión lineal simple entre O3 y cada una de las 4 variables, para utilizar el más acertado para la comparación con los siguientes modelos. Para cada uno de los casos se determina la variable dependiente e independiente, se utiliza la función lm() para sacar el valor del coeficiente y la constante del modelo lineal y se utiliza la función summary() para obtener el valor p y más información.

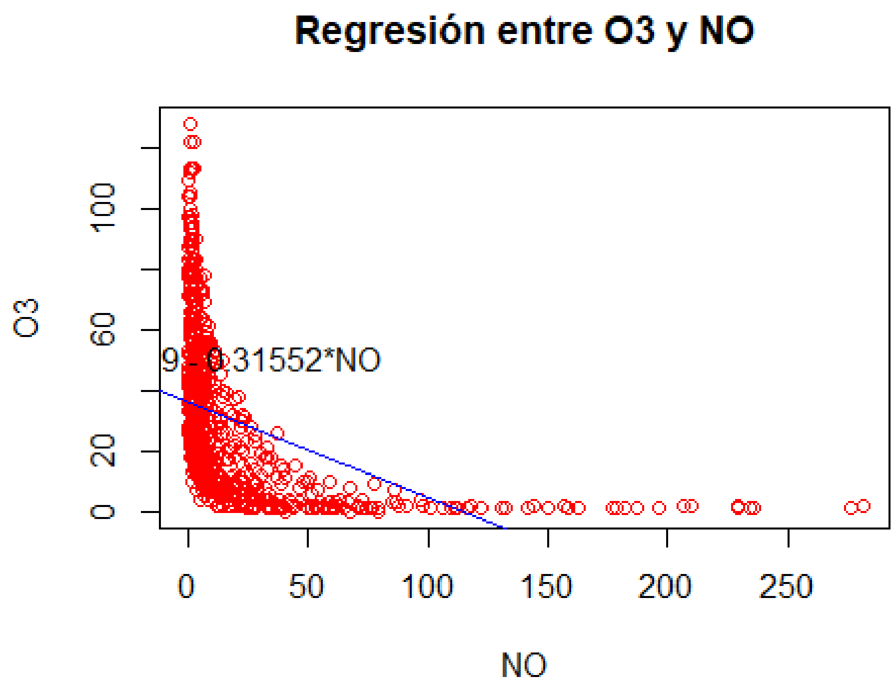
CO y O3

```
## 
## Call:
## lm(formula = y1 ~ x1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -35.340 -19.847  -7.299  13.568  94.824 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   40.607      1.401  28.987   <2e-16 ***
## x1           -25.343      2.695  -9.405   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 25.18 on 798 degrees of freedom
## Multiple R-squared:  0.09979,    Adjusted R-squared:  0.09866 
## F-statistic: 88.46 on 1 and 798 DF,  p-value: < 2.2e-16
```



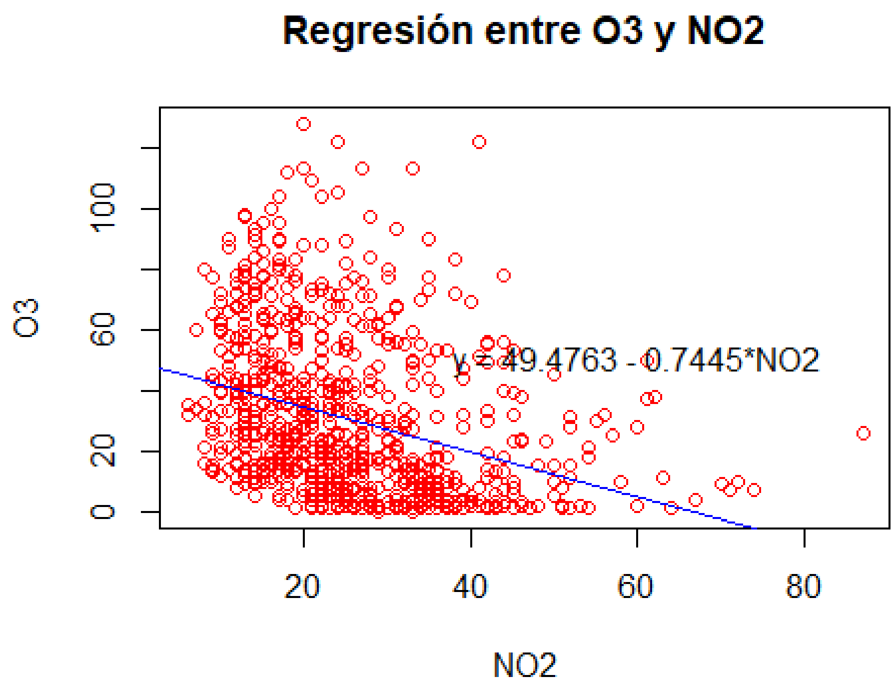
NO y 03

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.703 -19.414  -7.077  14.406  92.035
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.28079    0.96066   37.77   <2e-16 ***
## x2          -0.31552    0.02406  -13.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.07 on 798 degrees of freedom
## Multiple R-squared:  0.1773, Adjusted R-squared:  0.1763
## F-statistic:   172 on 1 and 798 DF,  p-value: < 2.2e-16
```



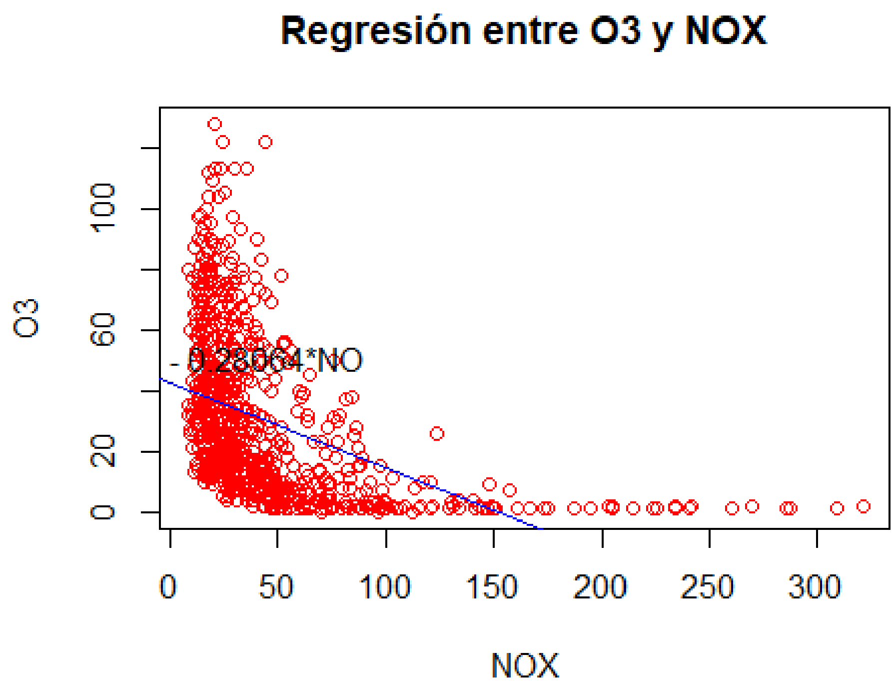
NO2 y 03

```
##
## Call:
## lm(formula = y3 ~ x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.331 -19.441  -8.097  14.386 103.048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.4763     2.1269  23.262   <2e-16 ***
## x3           -0.7445     0.0756  -9.848   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.06 on 798 degrees of freedom
## Multiple R-squared:  0.1084, Adjusted R-squared:  0.1072
## F-statistic: 96.98 on 1 and 798 DF,  p-value: < 2.2e-16
```



NOX y 03

```
##
## Call:
## lm(formula = y4 ~ x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.516 -19.047  -6.537  14.061  91.535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.81293    1.21671   35.19   <2e-16 ***
## x4          -0.28064    0.01996  -14.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.76 on 798 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1976
## F-statistic: 197.7 on 1 and 798 DF,  p-value: < 2.2e-16
```



Análisis de Modelo de Regresión Lineal Simple

Como solo queremos el mejor modelo, nos fijamos en el que tiene el mayor coeficiente de determinación de los 4 modelos sacados, en este caso es el de NOX y O3 con 0.1962, por lo que para este hacemos la validación de modelo a continuación:

Hipótesis:

$H_0: \beta = 0$ $H_1: \beta \neq 0$ $\alpha = 0.05$

Regla de decisión: Si valor p < alfa, se rechaza Ho, y por tanto, $$1 es significativa.

Se obtiene nuevamente el estadístico de prueba:

```
##
## Call:
## lm(formula = y4 ~ x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.516 -19.047  -6.537  14.061  91.535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.81293    1.21671   35.19   <2e-16 ***
## x4          -0.28064    0.01996  -14.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 23.76 on 798 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1976
## F-statistic: 197.7 on 1 and 798 DF,  p-value: < 2.2e-16
```

Se puede observar que el valor p: <2e-16 es considerablemente más pequeño que α por lo que podemos concluir que se rechaza la hipótesis nula Ho, y que la variable NOX tiene influencia significativa en O3.

Regresion curvilinea

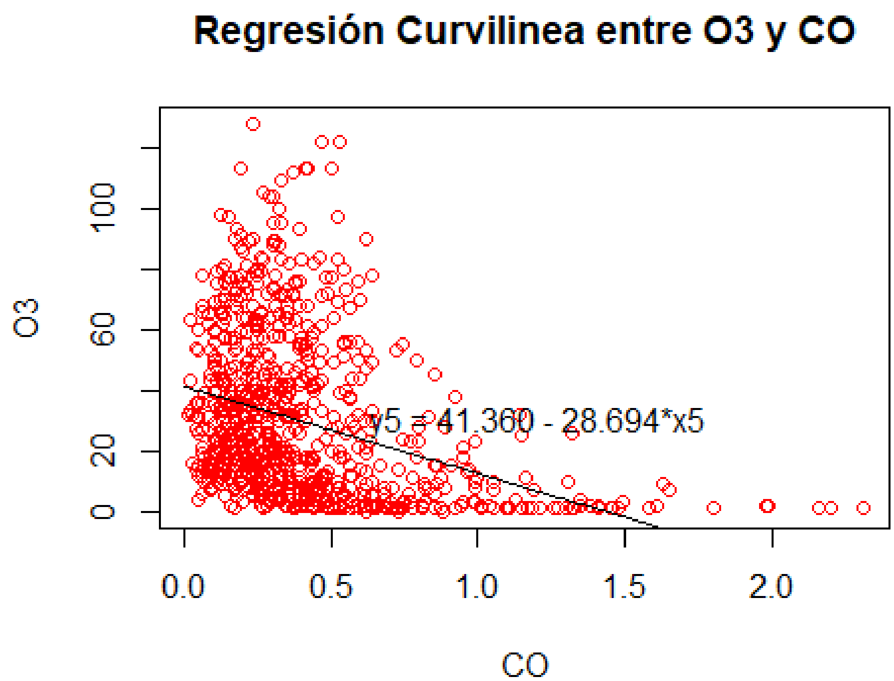
Ahora se realizan los modelos de regresión curvilineos entre O3 y cada una de las 4 variables, para utilizar el más acertado para la comparación con los demás modelos. Para esto se toman como variables independientes la variable correspondiente (CO, NO, NO2, NOX) y su mismo valor pero elevado al cuadrado. Utilizando esto se realiza el modelo con la función lm() y la misma variable dependiente en todos los casos. También se utiliza la función summary() para obtener los coeficientes de determinación y los valores p. Para todos los casos $\alpha = 0.05$

CO y 03

```
##
## Call:
## lm(formula = y5 ~ x5 + x6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.930 -19.545  -7.348  13.426  95.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.360      2.032  20.354  < 2e-16 ***
## x5           -28.694      7.086  -4.049 5.64e-05 ***
## x6             2.193      4.289   0.511    0.609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.19 on 797 degrees of freedom
## Multiple R-squared:  0.1001, Adjusted R-squared:  0.09783
## F-statistic: 44.32 on 2 and 797 DF,  p-value: < 2.2e-16
```

$$y5 = 41.360 - 28.694 * x5 - 2.193 * x5^2$$

Para este modelo curvilíneo entre O3 y CO, solo el valor de beta1 es significativo, con un valor p = 0.0023 < alfa = 0.05. Beta2 no es significativa ya que su valor p = 0.7735 > alfa = 0.05, por lo que esta se elimina del modelo.

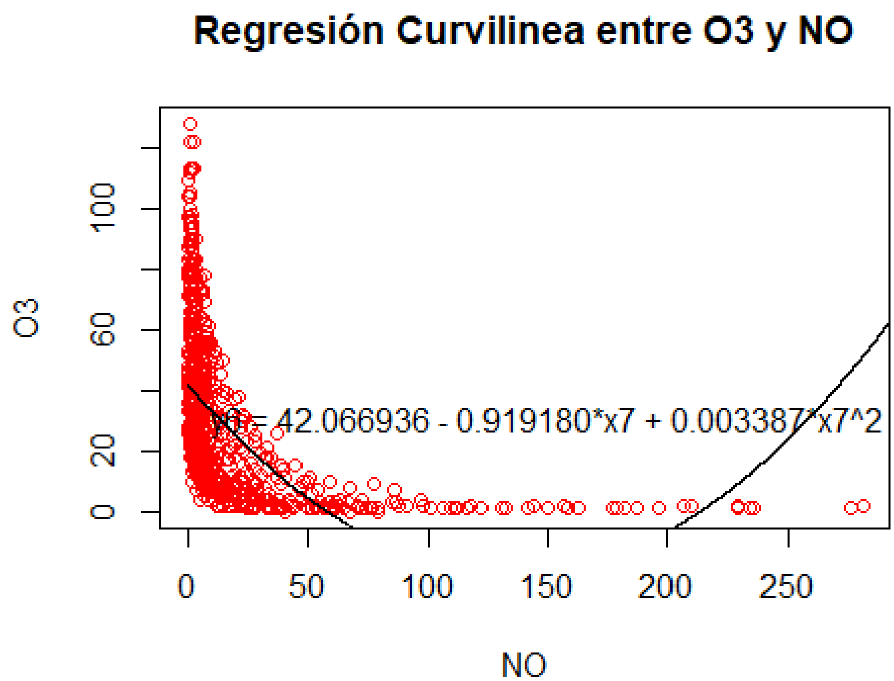


NO y 03

```
##
## Call:
## lm(formula = y6 ~ x7 + x8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.214 -18.267  -4.615  14.556  86.849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.066936   1.008412   41.72   <2e-16 ***
## x7          -0.919180   0.055087  -16.69   <2e-16 ***
## x8           0.003387   0.000283   11.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.17 on 797 degrees of freedom
## Multiple R-squared:  0.3027, Adjusted R-squared:  0.3009
## F-statistic:   173 on 2 and 797 DF,  p-value: < 2.2e-16
```

$$y6 = 42.066936 - 0.919180 * x7 + 0.003387 * x7^2$$

Para este modelo curvilíneo entre O3 y NO, los valores de $\beta 1$ y $\beta 2$ son significativos, ya que ambos tienen un valor p = 0.0000 < alfa = 0.05, por lo que se quedan en la ecuación.

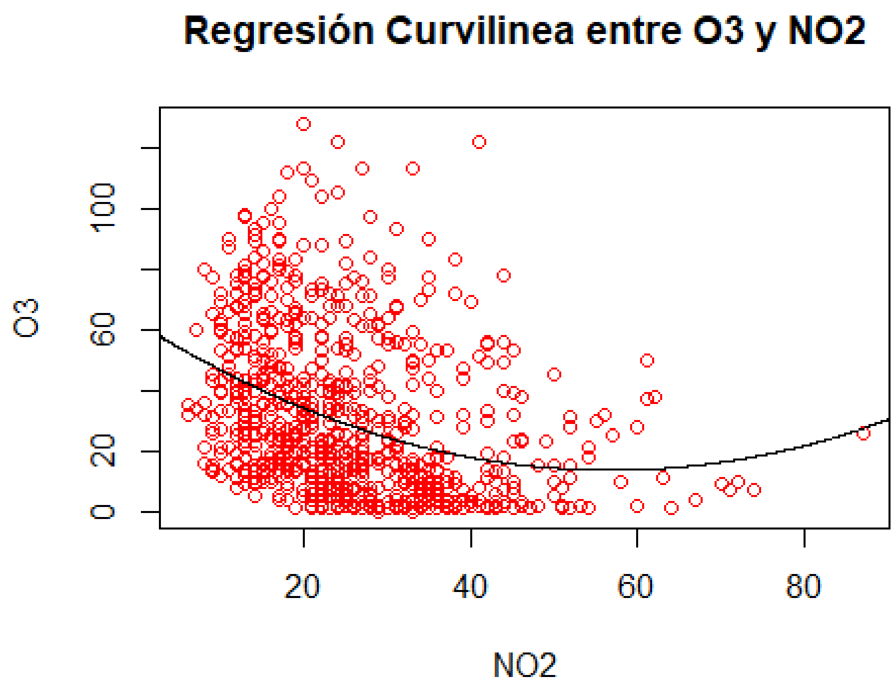


NO2 y 03

```
##
## Call:
## lm(formula = y7 ~ x9 + x10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.298 -18.269  -7.962  13.243 104.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.35659    3.97233  15.698  < 2e-16 ***
## x9          -1.71581    0.26469  -6.482 1.58e-10 ***
## x10          0.01511    0.00395   3.826  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.85 on 797 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1222
## F-statistic: 56.64 on 2 and 797 DF,  p-value: < 2.2e-16
```

$$y7 = 62.35659 - 1.71581 * x9 + 0.01511 * x9^2$$

Para este modelo curvilíneo entre O3 y NO2, los coeficientes $$1 y $$2 son significativos, ya que ambos tienen un valor p = 0.0000 < alfa = 0.05, por lo que se quedan en la ecuación.

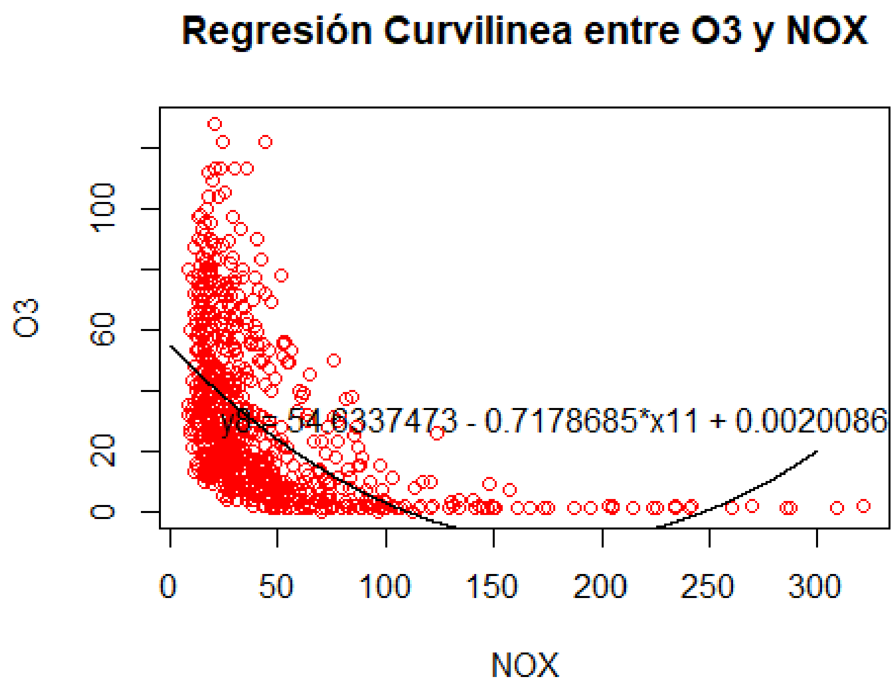


NOX y 03

```
##
## Call:
## lm(formula = y8 ~ x11 + x12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.980 -16.957  -4.336  11.734  95.064
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.6337473  1.7035758  32.070   <2e-16 ***
## x11         -0.7178685  0.0500514 -14.343   <2e-16 ***
## x12          0.0020086  0.0002128   9.437   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.54 on 797 degrees of freedom
## Multiple R-squared:  0.2791, Adjusted R-squared:  0.2773
## F-statistic: 154.3 on 2 and 797 DF,  p-value: < 2.2e-16
```

$$y8 = 54.6337473 - 0.7178685 * x11 + 0.0020086 * x11^2$$

Para este modelo curvilíneo entre O3 y NOX, ambos coeficientes $$1 y $$2 se quedan en la ecuación, ya que ambos tienen un valor p = 0.0000 < alfa = 0.05.



Análisis de Modelo de Regresión Curvilineo

De los 4 modelos de regresión curvilinea aneteriores, el que tiene un mayor coeficiente de determinación es el de NO y O3 con 0.2862, por lo que para este hacemos la validación de modelo a continuación:

Hipótesis:
$H_0: \beta i = 0$ $H_1: \beta i \neq 0$ $\alpha = 0.05$

Regla de decisión: Si valor p < alfa = 0.05, se rechaza Ho, y por tanto, βi es significativa.

Se saca nuevamente el estadístico de prueba:

```
##
## Call:
## lm(formula = y6 ~ x7 + x8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.214 -18.267  -4.615  14.556  86.849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.066936   1.008412   41.72   <2e-16 ***
## x7          -0.919180   0.055087  -16.69   <2e-16 ***
## x8           0.003387   0.000283   11.97   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.17 on 797 degrees of freedom
## Multiple R-squared:  0.3027, Adjusted R-squared:  0.3009
## F-statistic:   173 on 2 and 797 DF,  p-value: < 2.2e-16
```

Valor p: $$1 = <2e-16, $$2 = <2e-16

Conclusión: Como el valor p para ambas betas, uno y dos,, que en este caso son NO y NO^2 respectivamente, es 0.0000, mucho menor a alfa = 0.05, entonces se rechaza la hipótesis nula Ho para ambas betas, por lo tanto ambas son significativas.

Regresion cúbica

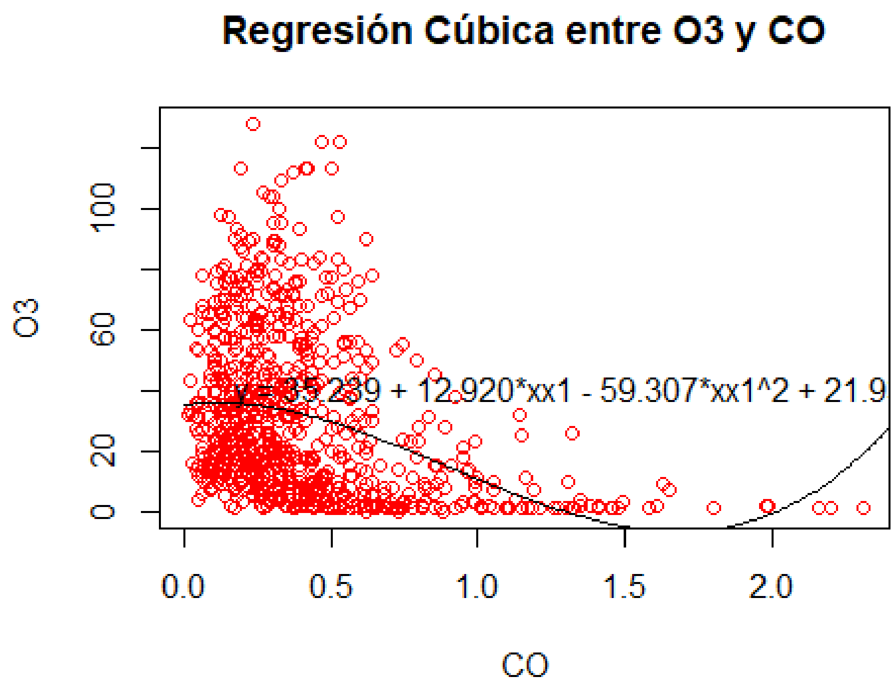
Ahora se realizan los modelos de regresión cúbicos entre O3 y cada una de las 4 variables, para comparar con los demás modelos el mejor. Las variables independientes son (CO, NO, NO2, NOX) y su mismo valor pero elevado al cuadrado y además su mismo valor pero elevado al cubo. Se utiliza nuevamente la función lm() y la misma variable dependiente en todos los casos. También se utiliza la función summary() para obtener los coeficientes de determinación y los valores p. Para todos los casos $\alpha = 0.05$

CO y 03

```
##
## Call:
## lm(formula = y ~ xx1 + xx2 + xx3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.829 -20.348  -5.859  13.246  93.307
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.239      2.782  12.666  < 2e-16 ***
## xx1           12.920     14.791   0.873  0.38266
## xx2          -59.307     19.688  -3.012  0.00267 **
## xx3           21.938      6.856   3.200  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.04 on 796 degrees of freedom
## Multiple R-squared:  0.1115, Adjusted R-squared:  0.1082
## F-statistic:  33.3 on 3 and 796 DF,  p-value: < 2.2e-16
```

$$y = 35.239 + 12.920 * xx1 - 59.307 * xx1^2 + 21.938 * xx1^3$$

Para este modelo cúbico entre O3 y CO, solo los valores de $$2 y $$3 son significativo, con un valor p < α. $$1 no es significativa ya que su valor p = 38266 > α = 0.05, por lo que esta se elimina del modelo.

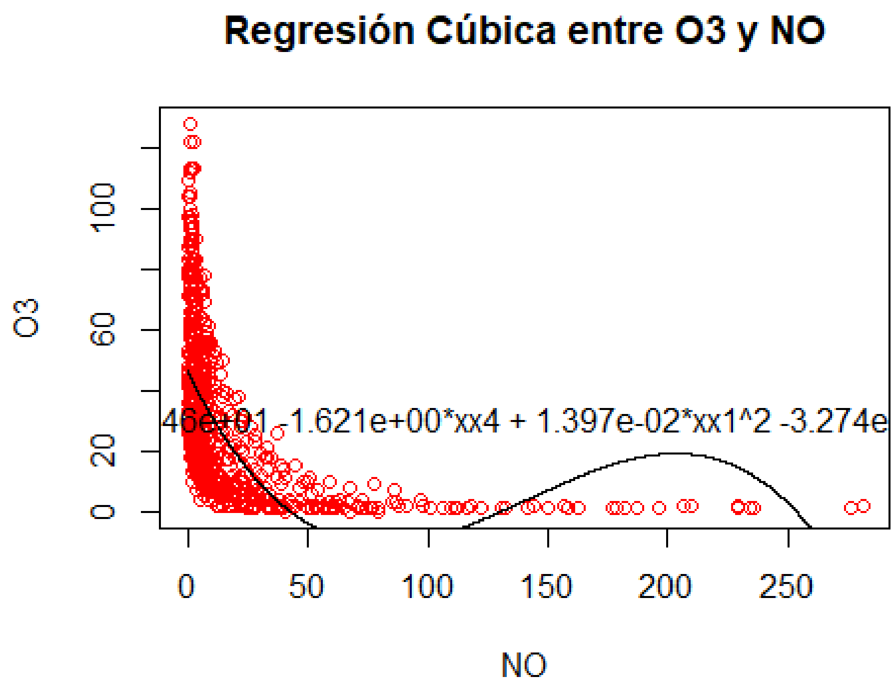


NO y 03

```
##
## Call:
## lm(formula = y ~ xx4 + xx5 + xx6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.700 -16.988  -3.285  12.607  83.149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.646e+01  1.076e+00  43.184   <2e-16 ***
## xx4         -1.621e+00  9.349e-02 -17.336   <2e-16 ***
## xx5          1.397e-02  1.198e-03  11.662   <2e-16 ***
## xx6         -3.274e-05  3.611e-06  -9.067   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.12 on 796 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.3656
## F-statistic: 154.5 on 3 and 796 DF,  p-value: < 2.2e-16
```

$$y = 4.646e + 01 - 1.621e + 00 * xx4 + 1.397e - 02 * xx1^2 - 3.274e - 05 * xx1^3$$

Para este modelo cúbico entre O3 y NO, todos los valores p de βi son menores a α por lo que son significativos y por lo que se quedan en la ecuación.

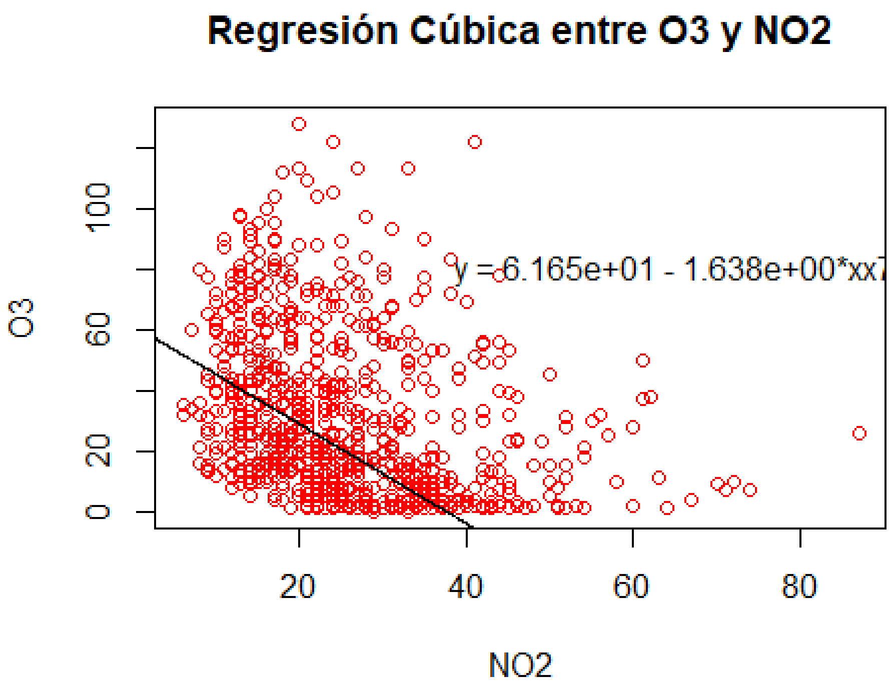


NO2 y 03

```
##
## Call:
## lm(formula = y ~ xx7 + xx8 + xx9)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.27 -18.29   -7.98   13.27  104.68
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.165e+01  7.246e+00   8.509   <2e-16 ***
## xx7         -1.638e+00  7.216e-01  -2.270   0.0235 *
## xx8          1.272e-02  2.105e-02   0.604   0.5457
## xx9          2.092e-05  1.807e-04   0.116   0.9078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.86 on 796 degrees of freedom
## Multiple R-squared:  0.1245, Adjusted R-squared:  0.1212
## F-statistic: 37.72 on 3 and 796 DF,  p-value: < 2.2e-16
```

$$y = 6.165e + 01 - 1.638e + 00 * xx7 + 1.272e - 02 * xx7^2 + 2.092e - 05 * xx7^3$$

Para este modelo cúbico entre O3 y NO2, solo el coeficiente $\beta1$ es significativo con un valor p = 0.02, $\beta2$ = 0.54 y $\beta3$ = 0.90 son menores a α = 0.05, por lo que se sacan de la ecuación.

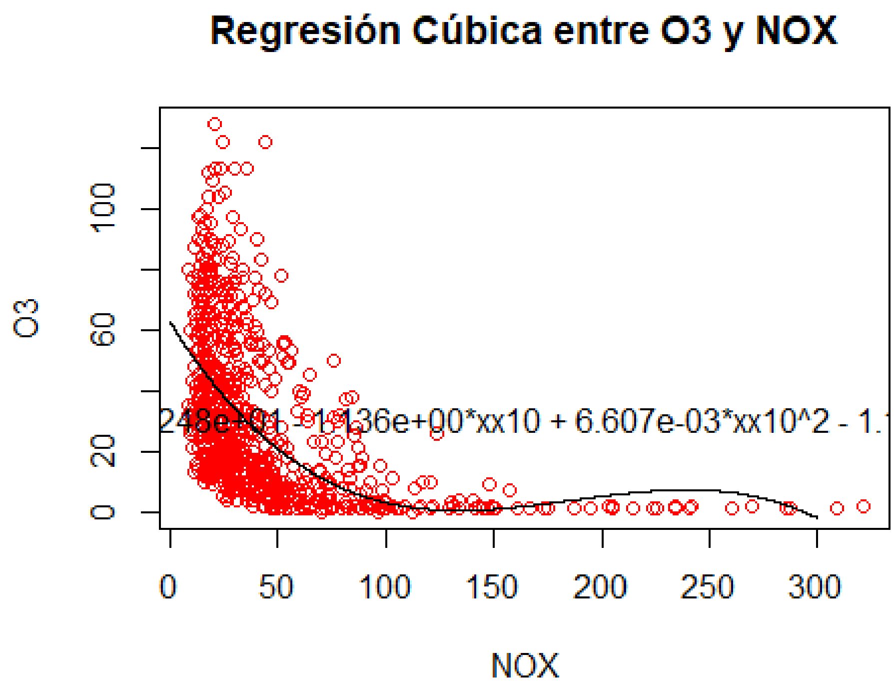


NOX y 03

```
##
## Call:
## lm(formula = y ~ xx10 + xx11 + xx12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.761 -15.850  -4.159  11.129  97.733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.248e+01  2.442e+00  25.589  < 2e-16 ***
## xx10        -1.136e+00  1.065e-01 -10.669  < 2e-16 ***
## xx11         6.607e-03  1.058e-03   6.247 6.83e-10 ***
## xx12        -1.179e-05  2.657e-06  -4.436 1.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.28 on 796 degrees of freedom
## Multiple R-squared:  0.2965, Adjusted R-squared:  0.2939
## F-statistic: 111.8 on 3 and 796 DF,  p-value: < 2.2e-16
```

$$y = 6.248e + 01 - 1.136e + 00 * xx10 + 6.607e - 03 * xx10^2 - 1.179e - 05 * xx10^3$$

Para este modelo cúbico entre O3 y NOX, todos los coeficientes βi tienen un valor p = 0.0000 < alfa = 0.05, por lo que se quedan en la ecuación.



Análisis de Modelo de Regresión Cúbico

De los 4 modelos de regresión cúbica aneteriores, el que tiene un mayor coeficiente de determinación es el de NO y O3 con 0.368, por lo que para este hacemos la validación de modelo a continuación:

Hipótesis:
$H_0: \beta i = 0$ $H_1: \beta i \neq 0$ $\alpha = 0.05$

Regla de decisión: Si valor p < alfa = 0.05, se rechaza Ho, y por tanto, Betai es significativa.

Se saca nuevamente el estadístico de prueba:

```
## 
## Call:
## lm(formula = y ~ xx4 + xx5 + xx6)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -34.700 -16.988  -3.285  12.607  83.149 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.646e+01  1.076e+00  43.184   <2e-16 ***
## xx4         -1.621e+00  9.349e-02 -17.336   <2e-16 ***
## xx5          1.397e-02  1.198e-03  11.662   <2e-16 ***
```

```
## xx6         -3.274e-05  3.611e-06  -9.067   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.12 on 796 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.3656
## F-statistic: 154.5 on 3 and 796 DF,  p-value: < 2.2e-16
```

Valor p: β 1 = <2e-16, β 2 = <2e-16, β 3 = <2e-16

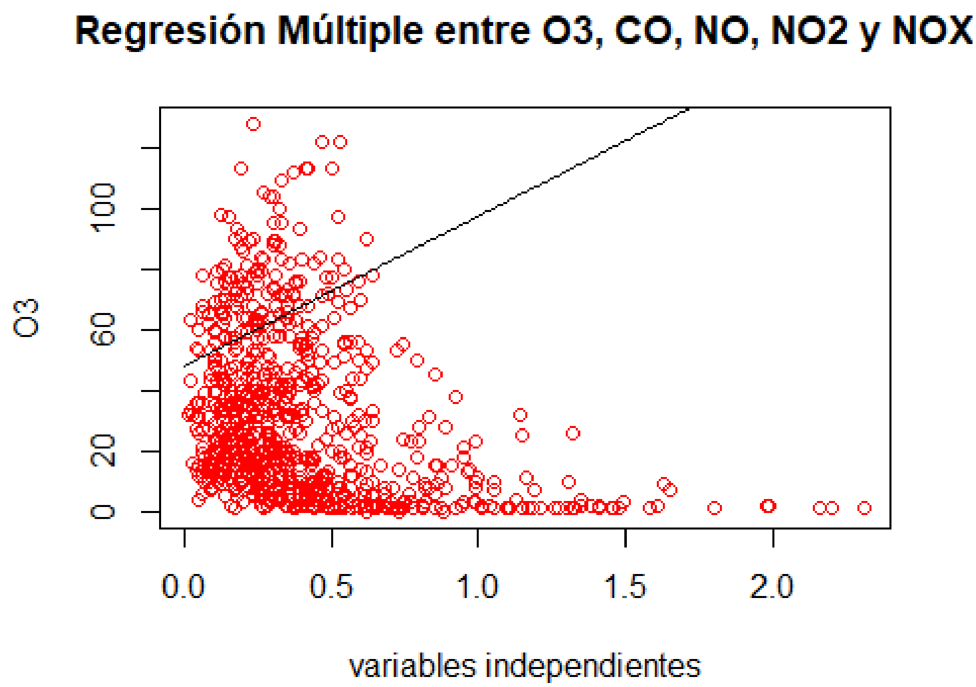
Conclusión: Como el valor p para todos los valores de β, es 0.0000, mucho menor a alfa = 0.05, entonces se rechaza la hipótesis nula H_0 y se consideran como significativas.

Regresion múltiple

Para la realización del modelo de regresión múltiple lineal se utilizan todas las variables independientes en la función lm() para determinar sus coeficientes, después se utiliza la función summary() para sacar el coeficiente de determinación y sus valores p.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.750 -17.329  -5.981  13.873  93.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.0054     2.0252  23.704  < 2e-16 ***
## x1           50.9026     6.3955   7.959 5.97e-15 ***
## x2           -0.2703     1.6143  -0.167    0.867
## x3           -0.8360     1.6203  -0.516    0.606
## x4           -0.2633     1.6143  -0.163    0.870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.88 on 795 degrees of freedom
## Multiple R-squared:  0.2594, Adjusted R-squared:  0.2557
## F-statistic: 69.63 on 4 and 795 DF,  p-value: < 2.2e-16
```

$$y = 48.0054 - 50.9026 * x1 - 0.2703 * x2 - 0.8360 * x3 - 0.2633 * x4$$



Validación del modelo

Hipótesis: $H_0: \beta i = 0$ $H_1: \beta i \neq 0$

Regla de decisión: Si valor p < alfa = 0.05, se rechaza Ho, y por tanto, Betai es significativa.

Estadístico de prueba:

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.750 -17.329  -5.981  13.873  93.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.0054     2.0252  23.704  < 2e-16 ***
## x1           50.9026     6.3955   7.959 5.97e-15 ***
## x2           -0.2703     1.6143  -0.167    0.867
## x3           -0.8360     1.6203  -0.516    0.606
## x4           -0.2633     1.6143  -0.163    0.870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

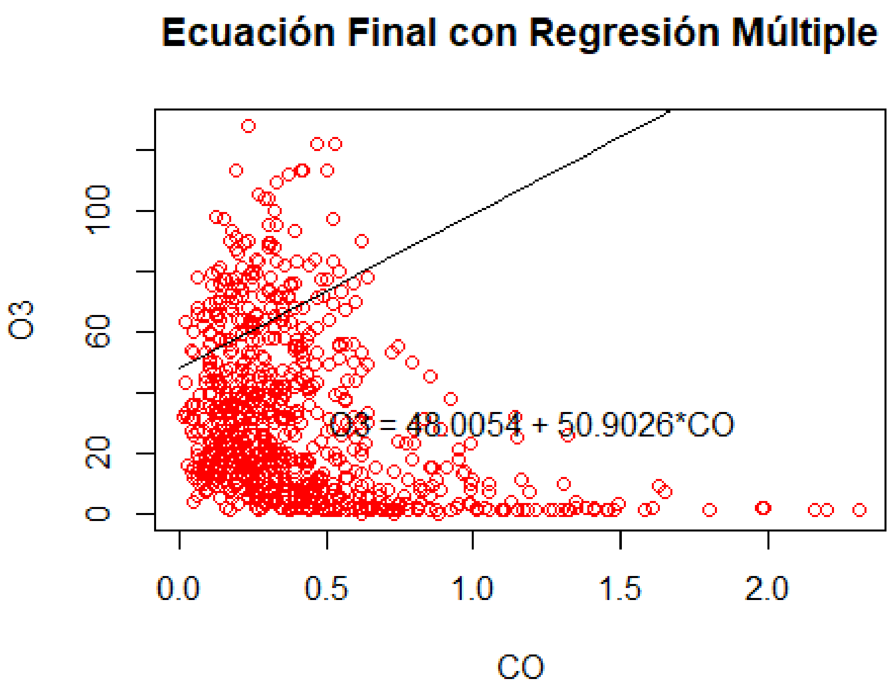
```
## Residual standard error: 22.88 on 795 degrees of freedom
## Multiple R-squared:  0.2594, Adjusted R-squared:  0.2557
## F-statistic: 69.63 on 4 and 795 DF,  p-value: < 2.2e-16
```

Análisis de Modelo de Regresión Múltiple

Valor p: Beta 1 = 5.97e-15, Beta 2 = 0.867, Beta 3 = 0.606, Beta 4 = 0.870

Resultado: El valor p para beta 1 es el único que es menor a alfa, por lo tanto esa beta es significativa, y los valores de la beta 2, 3, y 4 son superiores a alfa, por lo tanto estas no son significativas y podrían sacarse del modelo.

Ecuación Final: O3 = 48.0054 + 50.9026*CO



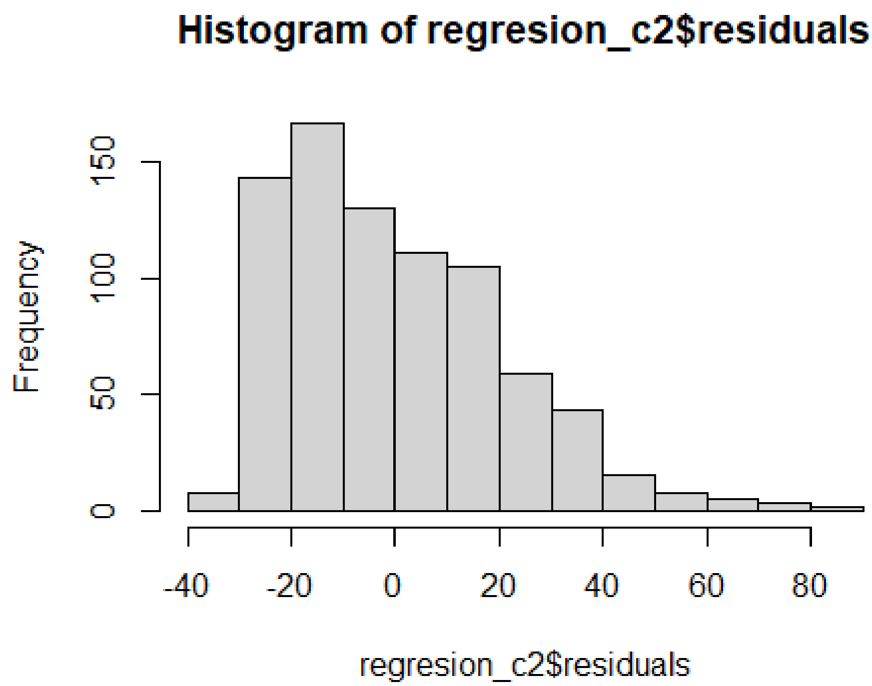
Modelo Elegido

De todos los modelos de regresion creados, comparamos los coeficientes de determinación de todos para encontrar cual es el que mejor representaría los datos.

Finalmente encontramos que el mejor modelo, por su coeficiente de deterinación seria el del modelo de regresión cúbica de NO y 03 (regresion_c2). De este modelo anteriormente se comprobo la significancia de sus betas, por lo tanto solo faltaría hacer el análisis residual, el cual se presenta a continuacion:

Análisis residual del modelo escogido

Graficas de residuos



Prueba de normalidad de residuos

H_0: Los datos si se distribuyen normalmente H_1: Los datos no se distribuyen normalmente

α = 0.05

Regla de desición: Si valor p < alfa se rechaza Ho

```
##
##  Shapiro-Wilk normality test
##
## data:  regresion_c2$residuals
## W = 0.94149, p-value < 2.2e-16
```

Como el valor p = 0.0000 < alfa = 0.05, se rechaza Ho, es decir, los datos no provienen de población normal.

Prueba de hipotesis para media de residuos 0

H_0: Media de los residuos = 0 H_1: Media de los residuos != 0 α = 0.05

Regla de decisión: si valor p < 0.05, se rechaza Ho.

```
## [1] 1.203914e-16
```

```
##
##  One Sample t-test
##
## data:  regresion_c2$residuals
## t = 1.6151e-16, df = 799, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.463159  1.463159
## sample estimates:
##    mean of x
## 1.203914e-16
```

Como el valor p = 1 > alfa = 0.05, no se rechaza Ho. Por tanto, se mantiene que la media de los residuos es 0.

Argumentación de homogeneidad de varianza e independencia

Al graficar las valores predichos con los residuales en el modelo elegido se observa que la variación a lo largo del eje horizontal no es la misma. Además, pareciera que tienden exponencialmente cuando el valor de la predicción se pasa un poco de 40, por lo que podemos concluir que no hay homocedasticidad.

Conclusiones

En conclusión se puede observar que el modelo con mayor precisión es el modelo de regresión cúbica entre O3 y N0 (regresion_c2), con un coeficiente de determinación = 0.368. Al igual que lo obtenido en el análisis por equipo, este nuevo modelo más preciso pasa las pruebas de residuos de media = 0 y beta1 significativa, pero no pasa las pruebas de homocedasticidad ni normalidad de los datos.

Investigando pude ver que se suele tomar como relativamente bueno un modelo si tiene un coeficiente de determinación entre 0.4 y 0.7, pero no fue el caso con ninguno de estos análisis. También se puede observar que hay otros modelos con coeficientes de determinación cercanos a 0.3, por lo que no se puede asumir que este toma en cuenta todos los factores. Sin embargo, viendo que si se pudo encontrar un modelo considerablemente mejor al obtenido con solo regresión curvilinea y lineal, se puede predecir que se encontraría un mejor modelo si se combinan y aplican más técnicas.

Respecto a la pregunta que se quería responder en este reporte, sí se puede decir que existe correlación entre las variables independientes elegidas y la dependiente (O3), ya que esto es comprobado en los test de correlación y sus análisis de hipótesis, donde todos dieron correlación negativa significativa. Pero para poder predecir con certeza la influencia de los niveles de estos gases en los niveles de ozono, de manera que se obtenga información útil sobre la que actuar en la problmática climática actual, se requieren más análisis y pruebas con diferentes técnicas, que den un modelo que tenga mayor precisión, si es que existe.

Referencias

Data Camp. (2023). Rdocumentation.org. https://www.rdocumentation.org/ Enders, F. Boyd (2022). coefficient of determination. Encyclopedia Britannica. https://www.britannica.com/science/coefficient-of-determination Gildardo, I. (2015, 17 julio). El Agujero de Ozono y El Calentamiento Global. https://www.mundohvacr.com.mx/2007/06/el-agujero-de-ozono-y-el-calentamiento-global/ IBM. Modelos estadísticos. (2021). https://www.ibm.com/docs/es/spss-modeler/saas?topic=nodes-statistical-models Investopedia. R-Squared Formula, Regression, and Interpretations. (2021, 12 septiembre). https://www.investopedia.com/terms/r/r-squared.asp MACROPROCESO: DOCENCIA PROCESO: LINEAMIENTOS CURRICULARES PROCEDIMIENTO: APROBACIÓN Y REVISIÓN DEL PLAN ACADÉMICO EDUCATIVO CONTENIDOS PROGRAMATICOS Código: D-LC-P02-F01 Versión: 03 Página 1 de 4 PRESENTACIÓN. (s/f). Edu.co. Recuperado el 5 de febrero de 2023, de http://www.uptc.edu.co/export/sites/default/facultades/f_educacion/pregrado/matematicas/inf_general/documentos/ELECTIVA_INTERDISCIPLINAR_I.pdf US EPA. Basic Ozone Layer Science. (2021, 7 octubre). https://www.epa.gov/ozone-layer-protection/basic-ozone-layer-science Wegener, R. (2013, 11 septiembre). Big Data: The Organizational Challenge. Bain. https://www.bain.com/insights/big_data_the_organizational_challenge

Autoevaluación

Interpretción de los resultados: Al realizar y ver los resultados de los análisis pude observar que la variable más significativa en el modelo múltiple lineal era la variable CO, sin embargo al tratar de predecir los niveles de O3, se acercó más el modelo curvilíneo entre O3 y NO. Además, los tests indican que solo el 36% de la variación es explicada por el modelo. Lo realizado en este reporte me da a entender que la significancia de una variable depende mucho de la técnica utilizada para sacar el modelo, por lo que es importante realizar varias pruebas y conocer diferentes técnicas, ya que, aunque existan técnicas que puedan ofrecer mejores resultados a través de Machine Learning (como redes neuronales), podemos usar modelos estadísticos básicos para juzgar el rendimiento de técnicas más avanzadas (IBM, 2021). Y como sé que esta clase de análisis son muy relevantes en el desempeño de empresas (Wegner, 2013), me queda perfectamente claro que debo entender como aplicarlas y que representan.

Respecto a lo que pienso de las actividades realizadas y mi desempeño, considero que si pude aprender bastante de lo que imagino era el propósito de la clase, el cómo realizar un análisis estadístico para llegar a conclusiones útiles en el mundo real. Para realizar este trabajo tuve que analizar y entender mayormente lo hicimos en las actividades de clase y tuve que poderlo aplicar en la realización de los análisis de este reporte. Sin embargo, si considero que me falta entendimiento de varios de los conceptos abstractos que vimos en este bloque, sin embargo sé que para mejorar eso solo es necesaria práctica y estudio, por lo que me queda claro que tengo que seguir haciendo eso. Aún así, definitivamente entendí bastante de lo necesario para realizar modelos de regresión básicos como los practicados en este reporte, y considero muy útil el haber aprendido de la importancia y relevancia de estos análisis en el mundo laboral, gracias a lo que nos comentaba el profesor en clase, además, el haber incrementado mi conocimiento de estadística me ha hecho percatarme de que existe toda un área de conocimiento útil y práctico que puedo aprender y utilizar en muchísimas áreas en mi futuro laboral, por lo que definitivamente la tendré en cuenta.

Anexo de Código utilizado

Se ponen los mismos subtítulos en los que el código fue relevante

Desarrollo

Preparación de los datos

M = read.csv(“MGH2022.csv”)

M1 = M[, c(3, 4, 5, 6, 7)]

M1[M1 == “-99”] <- NA #View(M1) M2 = na.omit(M1) View(M2) si se quiere sacar otra muestra aleatoria, quitar la comentarización de la linea de código correspondiente

indice = sample(1:nrow(M2), 800, replace = FALSE) ######### muestra_aleatoria = M2[indice,] ######### Guardando la nueva BD de la muestra: ######### write.csv(muestra_aleatoria, “MGH2022_muestra.csv”, row.names = FALSE) muestra_aleatoria = read.csv(“MGH2022_muestra.csv”) View(muestra_aleatoria)

Exploración de los datos

hist(x = muestra_aleatoria$CO, main = "Histograma Niveles de CO", xlab = "CO", ylab = "Frecuencia") hist(x = muestra_aleatoria$NO, main = “Histograma Niveles de NO”, xlab = “NO”, ylab = “Frecuencia”) hist(x = muestra_aleatoria$NO2, main = "Histograma Niveles de NO2", xlab = "NO2", ylab = "Frecuencia") hist(x = muestra_aleatoria$NOX, main = “Histograma Niveles de NOX”, xlab = “NOX”, ylab = “Frecuencia”) hist(x = muestra_aleatoria$O3, main = “Histograma Niveles de O3”, xlab = “O3”, ylab = “Frecuencia”)

plot(x = muestra_aleatoria$CO, y = muestra_aleatoria$O3, main = “Comparación de niveles de CO con O3”, xlab = “CO”, ylab = “O3”) plot(x = muestra_aleatoria$NO, y = muestra_aleatoria$O3, main = “Comparación de niveles de NO con O3”, xlab = “NO”, ylab = “O3”) plot(x = muestra_aleatoria$NO2, y = muestra_aleatoria$O3, main = “Comparación de niveles de NO2 con O3”, xlab = “NO2”, ylab = “O3”) plot(x = muestra_aleatoria$NOX, y = muestra_aleatoria$O3, main = “Comparación de niveles de NOX con O3”, xlab = “NOX”, ylab = “O3”)

var(x = muestra_aleatoria$CO, y = NULL, na.rm = FALSE) var(x = muestra_aleatoria$NO, y = NULL, na.rm = FALSE) var(x = muestra_aleatoria$NO2, y = NULL, na.rm = FALSE) var(x = muestra_aleatoria$NOX, y = NULL, na.rm = FALSE) var(x = muestra_aleatoria$O3, y = NULL, na.rm = FALSE)

sd(muestra_aleatoria$CO, na.rm = FALSE) sd(muestra_aleatoria$NO, na.rm = FALSE) sd(muestra_aleatoria$NO2, na.rm = FALSE) sd(muestra_aleatoria$NOX, na.rm = FALSE) sd(muestra_aleatoria$O3, na.rm = FALSE)

Análisis de correlación

cor(muestra_aleatoria) plot(muestra_aleatoria)

cor(x = muestra_aleatoria$CO, y = muestra_aleatoria$O3, method = c(“pearson”)) cor(x = muestra_aleatoria$NO, y = muestra_aleatoria$O3, method = c(“pearson”)) cor(x = muestra_aleatoria$NO2, y = muestra_aleatoria$O3, method = c(“pearson”)) cor(x = muestra_aleatoria$NOX, y = muestra_aleatoria$O3, method = c(“pearson”))

Pruebas de Hipótesis de Correlación

Realiza el análisis del resultado

library(Hmisc) rcorr(as.matrix(muestra_aleatoria))

Análisis de Regresión

Regresion lineal simple

CO y 03

x1 = muestra_aleatoria$COy1 = muestra_{a}leatoria$O3 regresion1= lm(y1 ~ x1) summary(regresion1)

plot(x1, y1, main = “Regresión entre O3 y CO”, xlab = “CO”, ylab = “O3”, col = “red”) abline(regresion1, col = “blue”) text(1.2, 50, “y = 38.089 + -23.464*CO”)

NO y 03

x2 = muestra_aleatoria$NOy2 = muestra_{a}leatoria$O3 regresion2= lm(y2 ~ x2) summary(regresion2)

plot(x2, y2, main = “Regresión entre O3 y NO”, xlab = “NO”, ylab = “O3”, col = “red”) abline(regresion2, col = “blue”) text(1.2, 50, “y = 35.23215 + -0.35842*NO”)

NO2 y 03

x3 = muestra_aleatoria$NO2y3 = muestra_{a}leatoria$O3 regresion3= lm(y3 ~ x3) summary(regresion3)

plot(x3, y3, main = “Regresión entre O3 y NO2”, xlab = “NO2”, ylab = “O3”, col = “red”) abline(regresion3, col = “blue”) text(1.2, 50, “y = 44.43073 + -0.61659*NO2”)

NOX y 03

x4 = muestra_aleatoria$NOXy4 = muestra_{a}leatoria$O3 regresion4= lm(y4 ~ x4) summary(regresion4)

plot(x4, y4, main = “Regresión entre O3 y NOX”, xlab = “NOX”, ylab = “O3”, col = “red”) abline(regresion4, col = “blue”) text(1.2, 50, “y = 35.23215 + -0.35842*NO”)

Análisis de Modelo de Regresión Lineal Simple

summary(regresion4)

Regresion curvilinea

CO y 03

y5 = muestra_aleatoria$O3x5 = muestra_{a}leatoria$CO x6 = (muestra_aleatoria$CO)^2 regresion5= lm(y5 ~ x5 + x6) summary(regresion5)

plot(x5, y5, main = “Regresión Curvilinea entre O3 y CO”, xlab = “CO”, ylab = “O3”, col = “red”) z1 = seq(0,2.5, 0.01) prediccion1 = 37.682 - 21.572*z1 #- 1.286*z1^2 lines(z1,prediccion1) text(1.2,30, “y5 = 37.682 - 21.572x5”)

NO y 03

y6 = muestra_aleatoria$O3x7 = muestra_{a}leatoria$NO x8 = (muestra_aleatoria$NO)^2 regresion6 = lm(y6 ~ x7 + x8) summary(regresion6)

plot(x7, y6, main = “Regresión Curvilinea entre O3 y NO”, xlab = “NO”, ylab = “O3”, col = “red”) z2 = seq(0,300, 0.01) prediccion2 = 39.5488 - 0.8034*z2 + 0.00284*z2^2 lines(z2,prediccion2) text(1.2,30, “y6 = 39.5488 - 0.8034*x7 - 0.00284*x7^2”)

NO2 y 03

y7 = muestra_aleatoria$O3x9 = muestra_{a}leatoria$NO2 x10 = (muestra_aleatoria$NO2)^2 regresion7= lm(y7 ~ x9 + x10) summary(regresion7)

plot(x9, y7, main = “Regresión Curvilinea entre O3 y NO2”, xlab = “NO2”, ylab = “O3”, col = “red”) z3 = seq(0,100, 0.01) prediccion3 = 56.36 - 1.4646*z3 + 0.012009*z3^2 lines(z3,prediccion3) text(1.2,30, “y7 = 56.36 - 1.4646*x9 + 0.012009*x9^2”)

NOX y 03

y8 = muestra_aleatoria$O3x11 = muestra_{a}leatoria$NOX x12 = (muestra_aleatoria$NOX)^2 regresion8= lm(y8 ~ x11 + x12) summary(regresion8)

plot(x11, y8, main = “Regresión Curvilinea entre O3 y NOX”, xlab = “NOX”, ylab = “O3”, col = “red”) z4 = seq(0,300, 0.01) prediccion4 = 50.6705 - 0.638589*z4 + 0.001737*z4^2 lines(z4,prediccion4) text(1.2,30, “y8 = 50.6705 - 0.638589*x11 + 0.001737*x11^2”)

Análisis de Modelo de Regresión Curvilineo

summary(regresion6)

Regresion cúbica

CO y 03

y = muestra_aleatoria$O3xx1 = muestra_{a}leatoria$CO xx2 = (muestra_aleatoria$CO)^{2}xx3 = (muestra_{a}leatoria$CO)^3 regresion_c1= lm(yy1 ~ xx1 + xx2 + xx3) summary(regresion_c1)

plot(xx1, y, main = “Regresión Cúbica entre O3 y CO”, xlab = “CO”, ylab = “O3”, col = “red”) z1 = seq(0,2.5, 0.01) prediccion_c1 = 35.239 + 12.920*z1 - 59.307*z1^2 + 21.938*z1^3*
*lines(z1,prediccion_c1) text(2,30, “yy1 = 35.239 + 12.920*xx1 - 59.307*xx1^2 + 21.938*xx1^3”)

NO y 03

y = muestra_aleatoria$O3xx4 = muestra_{a}leatoria$NO xx5 = (muestra_aleatoria$NO)^{2}xx6 = (muestra_{a}leatoria$NO)^3 regresion_c2= lm(y ~ xx4 + xx5 + xx6) summary(regresion_c2)

plot(xx4, y, main = “Regresión Cúbica entre O3 y NO”, xlab = “NO”, ylab = “O3”, col = “red”) z2 = seq(0,300, 0.01) prediccion_c2 = 4.646e+01 -1.621e+00*z2 + 1.379e-02*z2^2 -3.274e-05*z2^3*
*lines(z2,prediccion_c2) text(150,30, “y = 4.646e+01 -1.621e+00*xx4 + 1.379e-02*xx1^2 - 3.274e-05*xx1^3”)

NO2 y 03

y = muestra_aleatoria$O3xx7 = muestra_{a}leatoria$NO2 xx8 = (muestra_aleatoria$NO2)^{2}xx9 = (muestra_{a}leatoria$NO2)^3 regresion_c3= lm(y ~ xx7 + xx8 + xx9) summary(regresion_c3)

plot(xx7, y, main = “Regresión Cúbica entre O3 y NO2”, xlab = “NO2”, ylab = “O3”, col = “red”) z3 = seq(0,100, 0.01) prediccion_c3 = 6.165e+01 - 1.638e+00*z3*
*lines(z3,prediccion_c3) text(65,80, “y = 6.165e+01 - 1.638e+00*xx7”)

NOX y 03

y = muestra_aleatoria$O3xx10 = muestra_{a}leatoria$NOX xx11 = (muestra_aleatoria$NOX)^{2}xx12 = (muestra_{a}leatoria$NOX)^3 regresion_c4= lm(y ~ xx10 + xx11 + xx12) summary(regresion_c4)

plot(xx10, y, main = “Regresión Cúbica entre O3 y NOX”, xlab = “NOX”, ylab = “O3”, col = “red”) z4 = seq(0,300, 0.01) prediccion_c4 = 6.248e+01 - 1.136e+00*z4 + 6.607e-03*z4^2 - 1.179e-05*z4^3*
*lines(z4,prediccion_c4) text(1.2,30, “y = 6.248e+01 - 1.136e+00*xx10 + 6.607e-03*xx10^2 - 1.179e-05*xx10^3 “)

Análisis de Modelo de Regresión Cúbico

summary(regresion_c2)

Regresion múltiple

x1 = muestra_aleatoria$COx2 = muestra_{a}leatoria$NO x3 = muestra_aleatoria$NO2x4 = muestra_{a}leatoria$NOX y = muestra_aleatoria$O3 regresion_m = lm(y ~ x1 + x2 + x3 + x4) summary(regresion_m)

plot(x1, y, main = “Regresión Múltiple entre O3, CO, NO, NO2 y NOX”, xlab = “variables independientes”, ylab = “O3”, col = “red”) z5 = seq(0,300, 0.01) prediccion_m = 48.0054 + 50.9026*z5 - 0.2703*z5 - 0.8360*z5 - 0.2633*z5 lines(z5,prediccion_m) text(180,30, “y = 48.0054 - 50.9026*x1 - 0.2703*x2 - 0.8360*x3 - 0.2633*x4”)

summary(regresion_m)

Análisis de Modelo de Regresión Múltiple

plot(x1, y, main = “Ecuación Final con Regresión Múltiple”, xlab = “CO”, ylab = “O3”, col = “red”) z1 = seq(0,300, 0.1) prediccion5 = 48.0054 + 50.9026*z1 lines(z1,prediccion5) text(1.2,30, “O3 = 48.0054 + 50.9026*CO”)

Análisis residual del modelo escogido

Graficas de residuos

hist(regresion_c2$residuals)

plot(regresion_c2$fitted. values, regresion_c$2residuals, col = “pink”) abline(h = 0)

Prueba de normalidad de residuos

library(nortest) shapiro.test(regresion_c2$residuals)

Prueba de hipotesis para media de residuos 0

m = mean(regresion_c2$residuals)mt. test(regresion_c$2residuals)