Situacion Problema 1

Ivan Romero

7/3/2022

# Situación problema

Tan solo en México, alrededor de 9,300 personas mueren anualmente por causa de la contaminación en grandes ciudades. La contaminación es sin duda alguna, uno de los mayores problemas con los que contamos actualmente como sociedad, ya que este mismo, conlleva múltiples otras consecuencias, como alteraciones a la salud, problemas económicos, sociales, entre otros.

 El Dióxido de Nitrógeno es uno de los mayores contaminantes en la actualidad, ya que es producto del exceso de automóviles, efectos negativos de la industria y la calefacción en los hogares y edificios. Es sumamente dañino a la salud en niveles excesivos y puede llegar hasta causar la muerte.

En este proyecto, se analizará una muestra de mediciones de contaminación en la Ciudad de México, para encontrar el comportamiento y relación de ciertos contaminantes presentes en el aire.

## Lectura de datos

## X FECHA HORA CO CO2 NO NO2 NOX O3  
## 1 1 01/01/2021 1 0.62 3 4 36 39 7  
## 2 2 01/01/2021 2 0.70 3 11 38 49 4  
## 3 3 01/01/2021 3 0.95 2 20 40 60 1  
## 4 4 01/01/2021 4 0.98 2 28 38 66 2  
## 5 5 01/01/2021 5 0.91 2 11 35 46 4  
## 6 6 01/01/2021 6 0.71 1 9 31 40 3

## X FECHA HORA CO   
## Min. : 1 Length:7598 Min. : 1.00 Min. :0.0000   
## 1st Qu.:1971 Class :character 1st Qu.: 7.00 1st Qu.:0.2100   
## Median :3948 Mode :character Median :13.00 Median :0.3400   
## Mean :4340 Mean :12.65 Mean :0.4224   
## 3rd Qu.:6747 3rd Qu.:19.00 3rd Qu.:0.5400   
## Max. :8760 Max. :24.00 Max. :2.7600   
## CO2 NO NO2 NOX   
## Min. : 0.000 Min. : 0.00 Min. : 4.00 Min. : 5.00   
## 1st Qu.: 1.000 1st Qu.: 2.00 1st Qu.: 17.00 1st Qu.: 21.00   
## Median : 2.000 Median : 6.00 Median : 24.00 Median : 31.00   
## Mean : 3.269 Mean : 19.47 Mean : 26.65 Mean : 46.12   
## 3rd Qu.: 3.000 3rd Qu.: 19.00 3rd Qu.: 33.00 3rd Qu.: 55.00   
## Max. :99.000 Max. :370.00 Max. :114.00 Max. :414.00   
## O3   
## Min. : 0.00   
## 1st Qu.: 8.00   
## Median : 22.00   
## Mean : 30.62   
## 3rd Qu.: 46.00   
## Max. :148.00

La base de datos presentada, contiene 6 variables numéricas, con la información de los gases en cuestión. Así como una variable categórica (‘FECHA’), y dos variables continuas(‘X’ y ‘HORA’).

## Muestreo

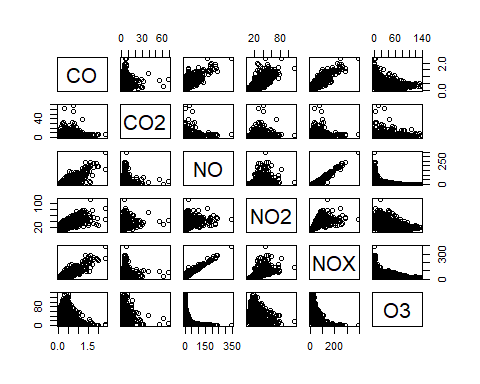
Hacemos una lectura de datos de nuestra muestra.

Para analizar y visualizar las correlaciones de los datos, se seleccionarán únicamente las variables numéricas discretas: #### seleción de las columnas haciendo uso de la librería dplyr hacemos uso de la librería de dplyr para seleccionar únicamente las columnas con valores numérico que nos sean de utilidad.

##   
## Attaching package: 'dplyr'

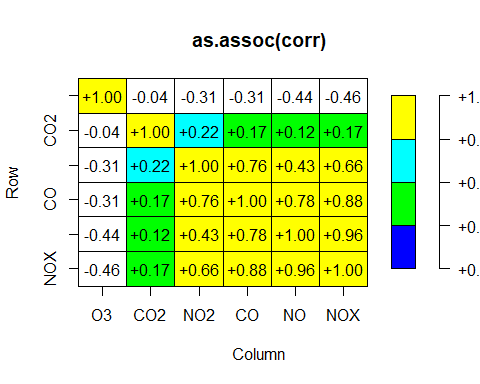
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Una vez obtenidos los datos, se analizarán las correlaciones entre nuestras variables representando las mediciones de los contaminantes en cuestión. #### Visualización gráfica de las correlaciones 

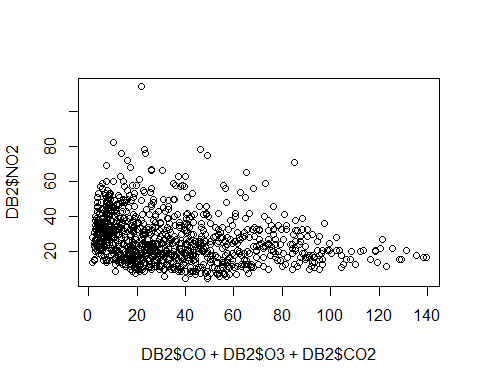
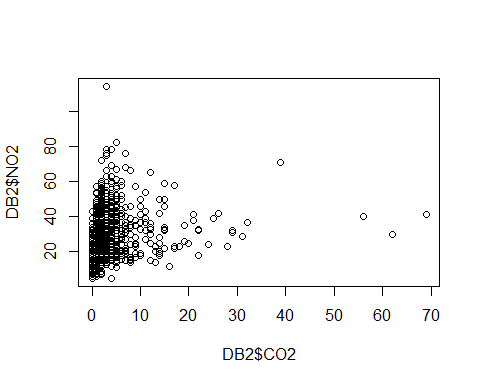
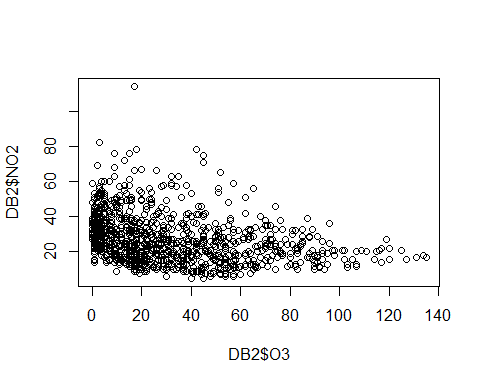
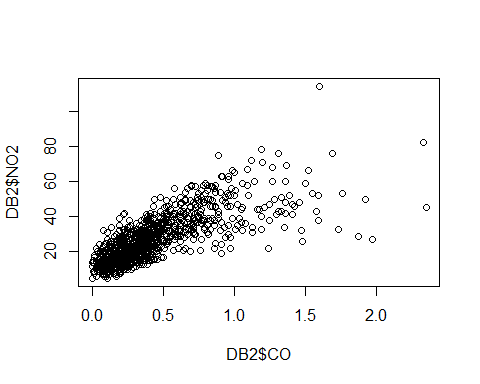
#### Tabla de correlaciones:

## CO CO2 NO NO2 NOX O3  
## CO 1.0000000 0.17085549 0.7760443 0.7600319 0.8770294 -0.31040186  
## CO2 0.1708555 1.00000000 0.1215341 0.2204943 0.1684025 -0.03658549  
## NO 0.7760443 0.12153411 1.0000000 0.4294466 0.9604718 -0.44196639  
## NO2 0.7600319 0.22049431 0.4294466 1.0000000 0.6636486 -0.31112246  
## NOX 0.8770294 0.16840245 0.9604718 0.6636486 1.0000000 -0.46201478  
## O3 -0.3104019 -0.03658549 -0.4419664 -0.3111225 -0.4620148 1.00000000

 Como se mencionó anteriormente, este análisis se centrará en el , ya que la CDMX, presenta un gran producción del mismo, producto del tráfico exesivo y la alta concentración de industria con efectos dañinos en el ambiente.

Tomando a como nuestro término dependiente Término independiente principal Segundo término independiente:

### Visualización de los datos

 Como se puede observar en las gráficas anteriores, existen ciertas variables que si sostienen una alta correlación con el como lo puede ser . Pero por otro lado, en suma, nuestros datos tienen una tendencia que puede llegar a ser muy difusa. De igual manera, se harán los modelos de regresión lineal para observar si un modelo se puede ajustar a nuestros datos. ## Modelo lineal

##   
## Call:  
## lm(formula = DB2$NO2 ~ DB2$CO + DB2$O3 + DB2$CO2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.929 -4.962 -1.021 4.594 52.816   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.28698 0.63189 24.192 < 2e-16 \*\*\*  
## DB2$CO 28.67231 0.86245 33.245 < 2e-16 \*\*\*  
## DB2$O3 -0.04103 0.01028 -3.992 7.02e-05 \*\*\*  
## DB2$CO2 0.23963 0.05189 4.618 4.38e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.601 on 996 degrees of freedom  
## Multiple R-squared: 0.5926, Adjusted R-squared: 0.5914   
## F-statistic: 483 on 3 and 996 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = DB2$NO2 ~ DB2$CO + DB2$O3 + DB2$CO2)  
##   
## Coefficients:  
## (Intercept) DB2$CO DB2$O3 DB2$CO2   
## 15.28698 28.67231 -0.04103 0.23963

Observamos que todas nuestras ’s son significativas y tenemos un coeficiente

Una vez obtenido nuestro modelo lineal, analizamos nuestros residuos para evaluar la eficiencia del mismo.

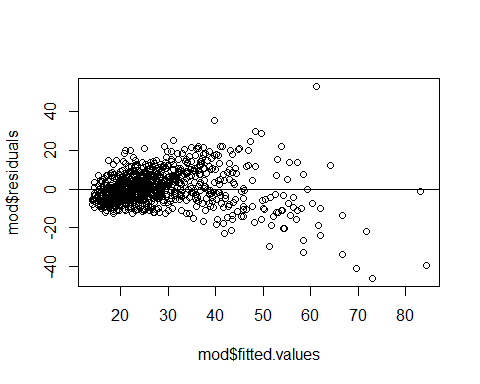
## Analisis de los residuos

##   
## One Sample t-test  
##   
## data: mod$residuals  
## t = -3.8427e-16, df = 999, p-value = 1  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.5329498 0.5329498  
## sample estimates:  
## mean of x   
## -1.043647e-16

Como el valor

, no se rechaza

### visualización de los residuos

 Notamos tendencia en los residuos, de forma parabólica, por lo tanto, más adelante se propondrá un modelo cuadrático.

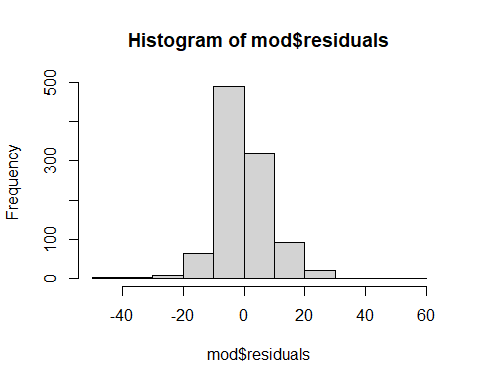
### Prueba de normalidad

Los residuos se distribuyen normalmente. Los residuos no se distribuyen normalmente.

##   
## Anderson-Darling normality test  
##   
## data: mod$residuals  
## A = 9.2047, p-value < 2.2e-16

Como el valor se rechaza .

### Distribución de los residuos

 Los residuos no se distribuyen de manera normal.

## Modelo cuadrático

##   
## Call:  
## lm(formula = y ~ z1 + z2 + z3 + z12 + z22 + z32)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.282 -4.648 -0.783 4.498 60.138   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.974e+00 8.654e-01 8.059 2.19e-15 \*\*\*  
## z1 5.329e+01 2.048e+00 26.024 < 2e-16 \*\*\*  
## z2 1.221e-01 2.838e-02 4.301 1.87e-05 \*\*\*  
## z3 4.662e-01 9.200e-02 5.067 4.81e-07 \*\*\*  
## z12 -1.613e+01 1.268e+00 -12.713 < 2e-16 \*\*\*  
## z22 -1.742e-03 2.762e-04 -6.306 4.30e-10 \*\*\*  
## z32 -7.008e-03 2.031e-03 -3.451 0.000581 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.85 on 993 degrees of freedom  
## Multiple R-squared: 0.6617, Adjusted R-squared: 0.6597   
## F-statistic: 323.8 on 6 and 993 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = y ~ z1 + z2 + z3 + z12 + z22 + z32)  
##   
## Coefficients:  
## (Intercept) z1 z2 z3 z12 z22   
## 6.974159 53.288608 0.122066 0.466185 -16.125618 -0.001742   
## z32   
## -0.007008

Como podemos apreciar, nuevamente en nuestro modelo, se obtienen ’s significativas, y así mismo, se obtiene un mejor resultado para

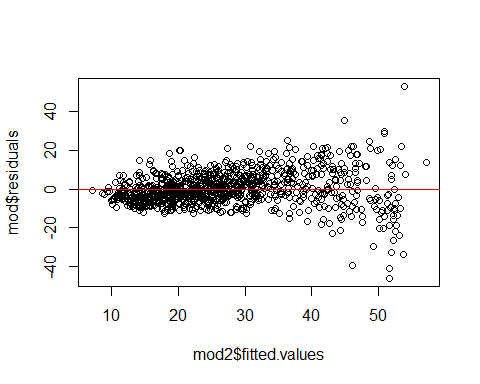
## Analisis de los residuos

##   
## One Sample t-test  
##   
## data: mod2$residuals  
## t = 4.4379e-16, df = 999, p-value = 1  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.4856435 0.4856435  
## sample estimates:  
## mean of x   
## 1.098292e-16

Como el valor

, no se rechaza

### visualización de los residuos

 Los residuos se distribuyen de manera más uniforme que en nuestro modelo anterior y no se observa una tendencia clara.

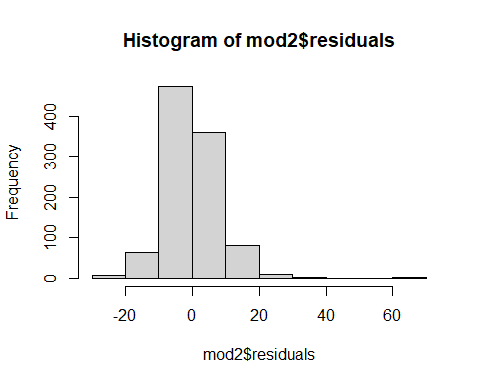
### Prueba de normalidad

Los residuos se distribuyen normalmente. Los residuos no se distribuyen normalmente.

##   
## Anderson-Darling normality test  
##   
## data: mod2$residuals  
## A = 6.2082, p-value = 2.948e-15

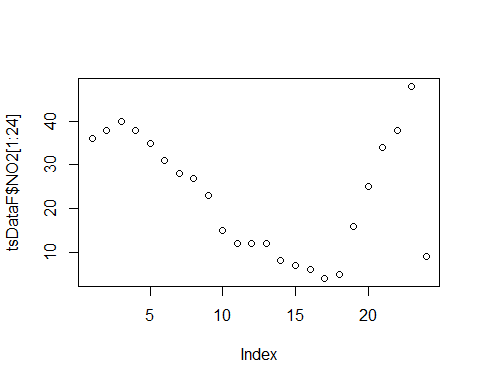
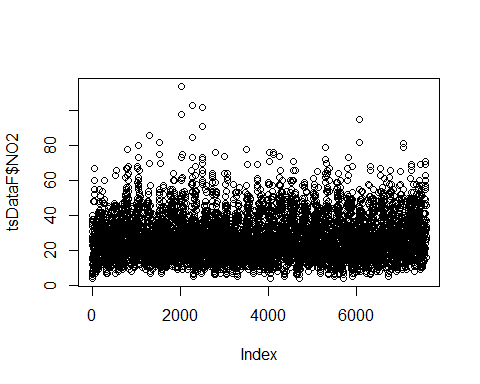
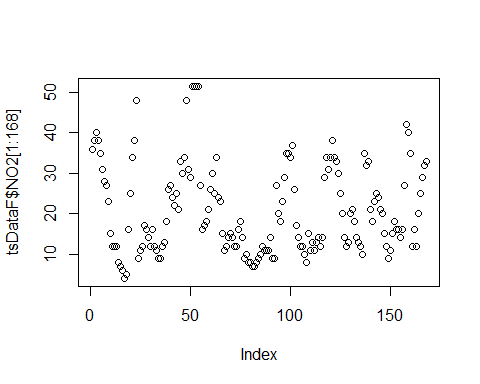
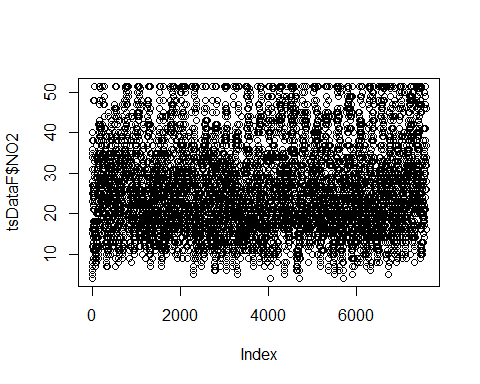
Como el valor se rechaza .

### Distribución de los residuos

 Los residuos se distribuyen de manera más uniforme que en nuestro modelo anterior, más no de una forma normal.

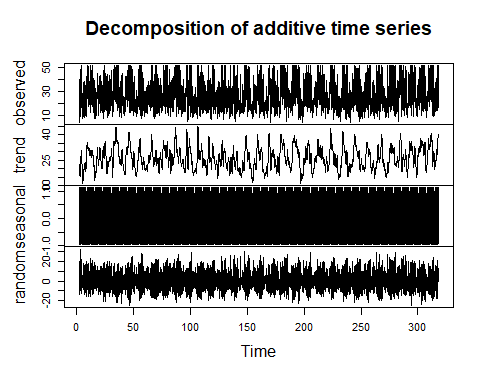
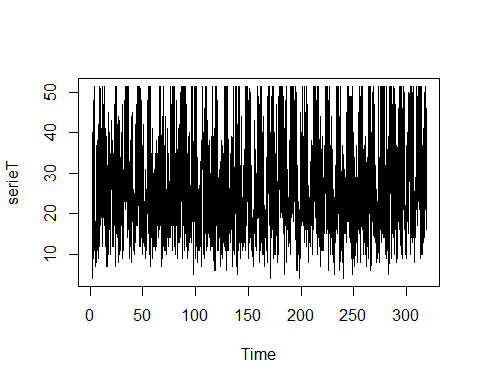
## intentando una serie de tiempo con respecto a

dado que para una serie de tiempo requerimos valores continuos, volveremos a usar la base de datos general.

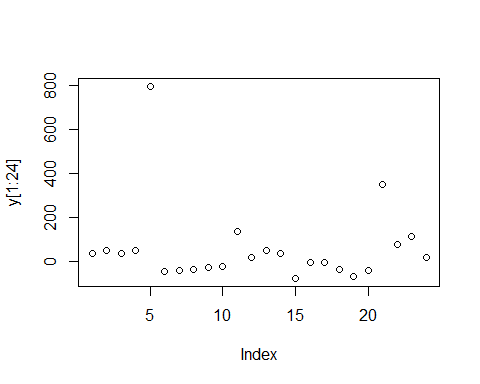
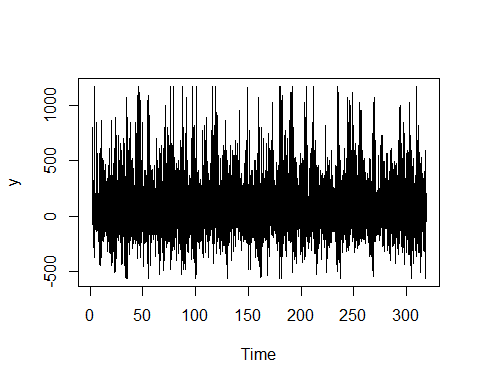
 Se observa una gran difusión para todos nuestros datos, pero observando de ‘cerca’, resulta más claro que es posible que existe una tendencia. ### limpieza de datos dado que se observa que existen una gran cantidad de datos que están muy alejados del resto de valores, se limpian los datos que quedan fuera de aproximadamente el 5%  Notamos que la serie de tiempo parece seguir una tendencia si lo observamos a lo largo de un día. ### Creacion de la serie de tiempo y su descomposición

## Warning: package 'tseries' was built under R version 4.1.3

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo



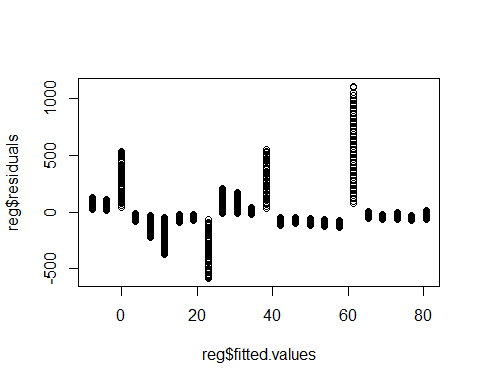
### Datos sin estacionalidad

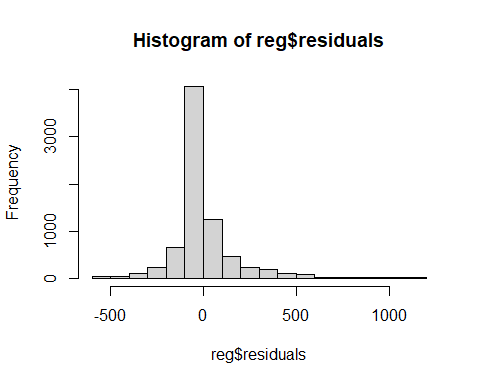
 la tendencia en los datos generales parece igual de difusa que antes, pero es posible observar la diferencia al acercarnos a 24h

#### Encontrando el modelo lineal para nuestros datos sin estacionalidad

##   
## Call:  
## lm(formula = z1 ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -584.83 -76.44 -42.52 46.10 1110.26   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 84.5068 4.2921 19.69 <2e-16 \*\*\*  
## x -3.8444 0.3005 -12.79 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 181.3 on 7596 degrees of freedom  
## Multiple R-squared: 0.02109, Adjusted R-squared: 0.02096   
## F-statistic: 163.6 on 1 and 7596 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = z1 ~ x)  
##   
## Coefficients:  
## (Intercept) x   
## 84.507 -3.844

Dado que nuestro valor de es tan bajo, es posible ver como nuestro modelo no se ajusta a nuestros datos ### Analisis de los residuos 

No se observa una tendencia clara 

##   
## Anderson-Darling normality test  
##   
## data: reg$residuals  
## A = 493.37, p-value < 2.2e-16

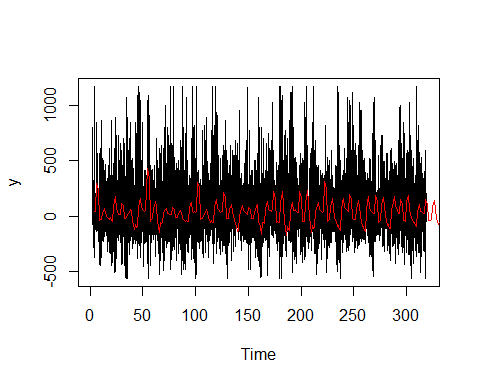
como se rechaza

Es fácil observar la gran variabilidad que nuestros datos presentan por lo cuál, hacer un modelo que se ajuste a estos datos será sumamente complicado.

## [1] 31347.91

Obtenemos un error sumamente alto, gracias a la dispersión de nuestros datos, ya aún sin estacionalidad.

### Promedios moviles

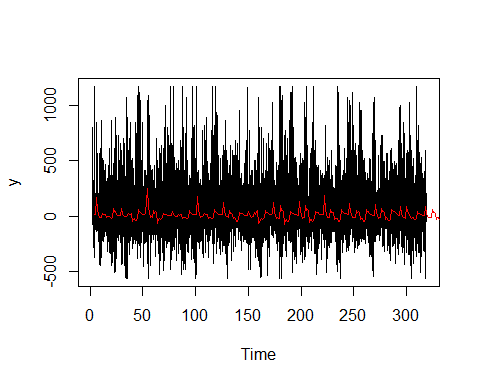


## [1] 45294.6

Nuestro error resulta sumamente alto y mayor a nuestro modelo anterior.

### Promedios moviles ponderados

Ponderación:



## [1] 35814.69

nuestro error sigue siendo alto, más no tan alto como nuestro modelo anterior, pero menor que nuestro primer modelo.

### Suavizamiento exponencial

##   
## Attaching package: 'collections'

## The following object is masked from 'package:utils':  
##   
## stack

## [1] 0.01 56154.06

Vemos que nuestro modelo de suavizamiento exponencial, nos deja con un coeficiente muyu bajo, de 0.01 y con un error más alto del visto en nuestros modelos anteriores, por lo cual, se considera que el modelo lineal propuesto es el mejor hasta el momento.

## referencias

* manatimx (2020, December 13). Registran más de 2 mil muertes por contaminación ambiental en Puebla. Manatí. <https://manati.mx/2020/12/13/muertes-por-contaminacion-del-aire-en-puebla/>
* CEMDA (s.f.). Impactos de la calidad del aire en la salud y en la economía de México. <https://www.cemda.org.mx/wp-content/uploads/2014/07/respira-mexico-baja-resolucion.pdf>