UNIVERSIDAD
NACIONAL DEL OESTE

Explotación de Datos

Actividad N^{o} 3

Regresión Lineal Múltiple

Profesores:

Dejean, Gustavo
Españadero, Juan
Mendoza, Dante

Integrantes Grupo B:

Benitez, Nicolas
Garcia Ravlic, Ignacio Agustin
Rechimon, Pablo Hernan
Rodriguez, Miguel Angel

Fecha de entrega:

26 de Septiembre de 2020

Resumen

A partir del Dataset de Informacion Meteorológica 2012, y el lenguaje de programación R, nos propusimos realizar un modelo de predicción de la temperatura utilizando tecnicas estadísticas como la Regresión Lineal Multiple, donde logramos obtener diferentes modelos y compararlos para seleccionar al que mejor se adecue.

Palabras Clave:
Análisis de Datos - Ciencia de Datos - Correlación - Explotación de datos - Lenguaje R - Meteorología - Modelo - Predicción - Regresión Lineal - Residuo

---

Índice

Graficos

1 Introducción

1.1 Datos a utilizar

Utilizaremos el dataset *Información meteorológica 2012*, provisto por el *Ministerio de Ambiente y Espacio Público. Agencia de Protección Ambiental (APRA)*, el cual cuenta con 11 variables y cerca de 30.000 muestras.

1.2 Objetivo

Nuestro objetivo consiste en aplicar un análisis de regresión multilineal para predecir el valor de nuestra variable dependiente seleccionada, que representa la temperatura en grados celsius, por lo que crearemos diversos modelos con el fin de alcanzar el mayor porcentaje de acierto en la predicción.

2 Desarrollo

2.1 Análisis de los datos

Nuestro dataset consta de 11 variables:

- FECHA - día, mes y año de la toma de la muestra
- HORA - Hora en la que fue tomada la medición
- ESTACION - estación de medición de donde se tomo la muestra
- VV_(M/S) - Velocidad del viento
- DV - Dirección del viento
- TEMP_C - Temperatura (en grados celsius)
- HR_PORC - Humedad relativa
- PRESS_MBAR - Presión en hpa
- PLUV_MM - Pluviometro (en milimetros)
- RAD_SOL_W/M^2 - Radiación solar (en vatios por m^2)
- UV_UVINDEX - Rayos ultravioletas

2.2 Preparación de los datos

Preparamos los datos en base a nuestro requerimientos:

- Eliminamos las variables no numéricas(Fecha y Estacion).
- Renombramos nuestras variables con colnames().
- Visualizamos un resumen estadístico de los datos con summary().

```
> summary(data) #Resumen estadistico de los datos
      Hora       Velocidad_Viento_MS  Direc_Viento   Temperatura_C     Humedad_Relativ
 Min.   : 0.00   Min.   :0.000       Min.   :  0.0   Min.   :   1.10   Min.   : 20.0
 1st Qu.: 5.00   1st Qu.:0.000       1st Qu.: 67.5   1st Qu.:  16.10   1st Qu.: 57.0
 Median :11.00   Median :0.400       Median :135.0   Median :  21.30   Median : 67.0
 Mean   :11.49   Mean   :0.653       Mean   :155.5   Mean   :  56.72   Mean   : 69.4
 3rd Qu.:17.00   3rd Qu.:0.900       3rd Qu.:247.5   3rd Qu.:  25.60   3rd Qu.: 78.0
 Max.   :23.00   Max.   :9.800       Max.   :337.5   Max.   :1802.60   Max.   :255.0
  Presion_hPA     Pluviometro_MM      Rad_Solar_WM2       UV_Rayos
 Min.   : 994.4   Min.   :  0.0000   Min.   :   0.0   Min.   : 0.000
 1st Qu.:1007.9   1st Qu.:  0.0000   1st Qu.:   0.0   1st Qu.: 0.000
 Median :1011.4   Median :  0.0000   Median :   3.0   Median : 0.000
 Mean   :1011.6   Mean   :  0.3961   Mean   : 138.8   Mean   : 1.744
 3rd Qu.:1014.8   3rd Qu.:  0.0000   3rd Qu.: 147.0   3rd Qu.: 2.200
 Max.   :1033.7   Max.   :237.0000   Max.   :1145.0   Max.   :16.000
```

Fig. 1: Primer Resumen Estadístico

- Creamos un nuevo subset de datos quitando los outliers de la Humedad Relativa y la Temperatura.

También realizamos un análisis de correlación y lo visualizamos:

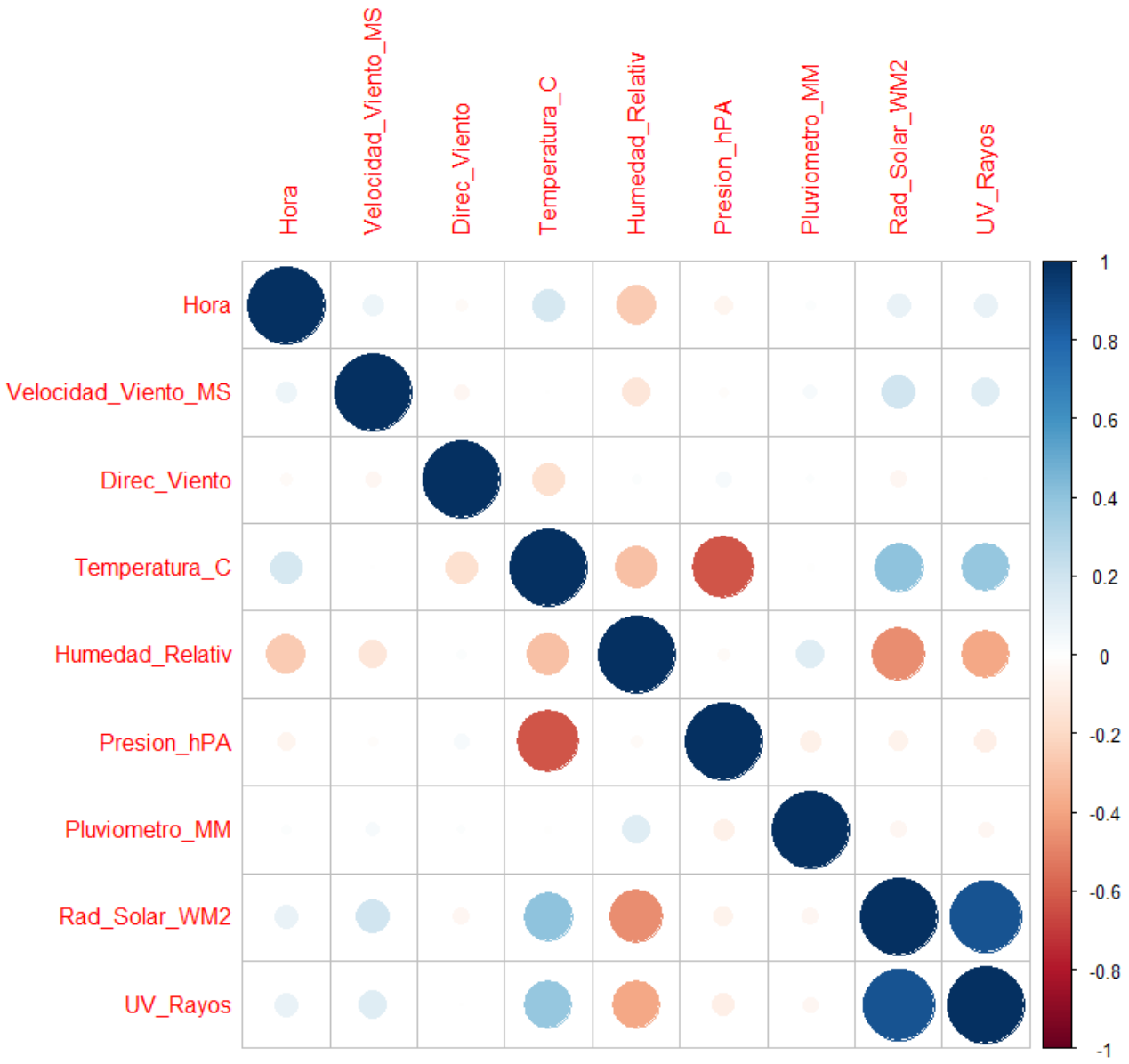

Fig. 2: Correlaciones

2.3 Análisis de los modelos

2.3.1 Modelo 1

Creamos nuestro primer modelo con la variable *Temperatura_C* como nuestra variable dependiente, dejando al resto como variables predictoras. Luego hacemos un summary para analizar el modelo:

```
> summary(lm.full)

Call:
lm(formula = f, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0227  -2.7177   0.7608   3.0957  12.3641

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)         8.201e+02  5.058e+00  162.149  < 2e-16 ***
Hora                6.560e-02  3.744e-03   17.520  < 2e-16 ***
Velocidad_Viento_MS -7.131e-01  2.857e-02  -24.961  < 2e-16 ***
Direc_Viento        -8.733e-03  2.441e-04  -35.771  < 2e-16 ***
Humedad_Relativ     -7.996e-02  2.021e-03  -39.558  < 2e-16 ***
Presion_hPA         -7.853e-01  4.986e-03 -157.495  < 2e-16 ***
Pluviometro_MM      -1.861e-02  5.110e-03   -3.642 0.000271 ***
Rad_Solar_WM2        5.060e-03  2.120e-04   23.868  < 2e-16 ***
UV_Rayos             2.222e-01  1.502e-02   14.791  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.223 on 28712 degrees of freedom
Multiple R-squared:  0.5862,    Adjusted R-squared:  0.5861
F-statistic:  5085 on 8 and 28712 DF,  p-value: < 2.2e-16
```

Fig. 3: summary() del Modelo 1

Donde podemos ver varias cuestiones importanes, los p-values de los coeficientes y su calidad, el error estándar (4.223), el valor de R^2 (0.5861), y el p-value del modelo, el cual se encuentra muy cercano al 0.

Realizamos las visualización que describen al modelo:

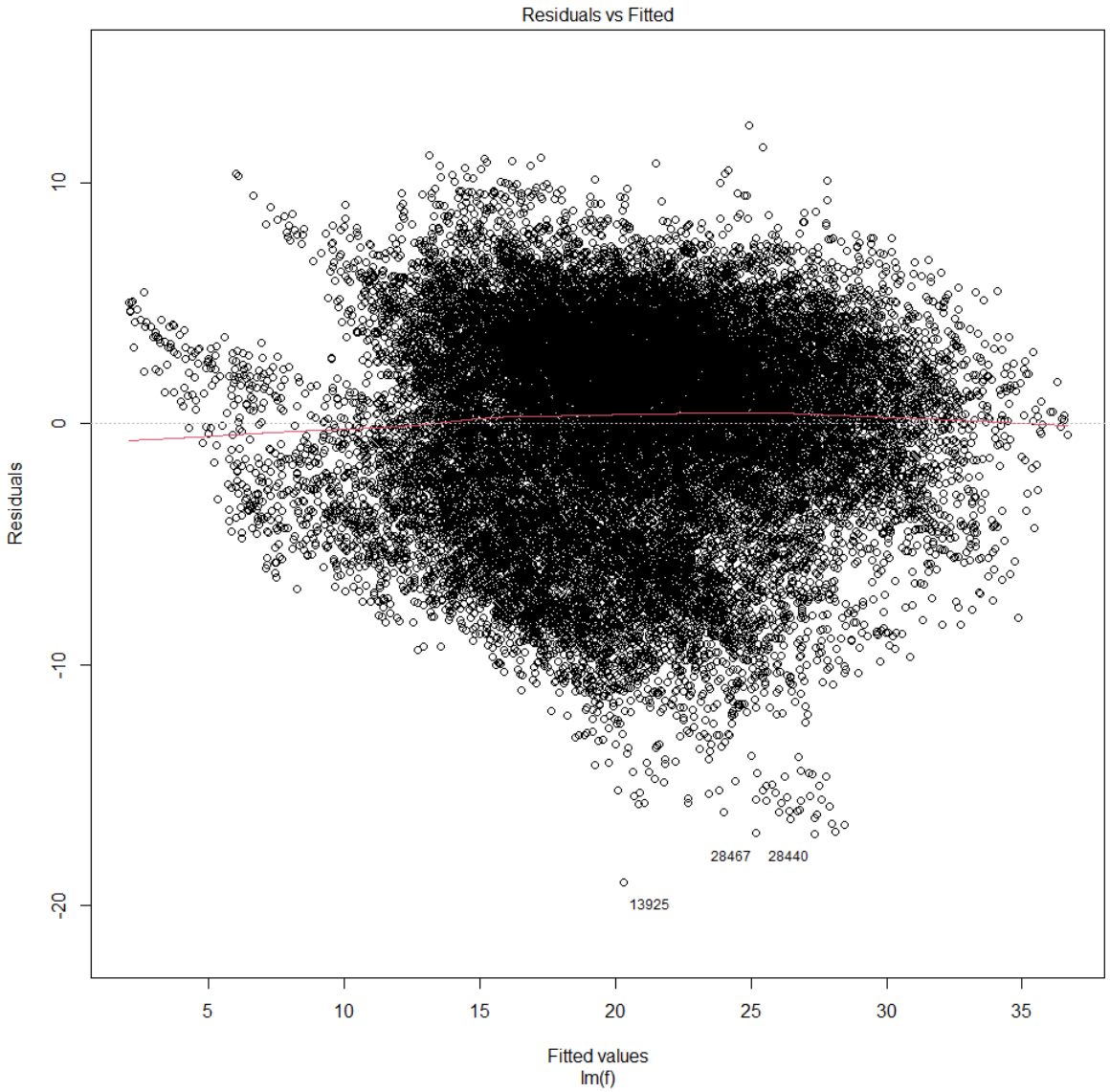

Fig. 4: Valores residuales contra valores ajustados del Modelo 1

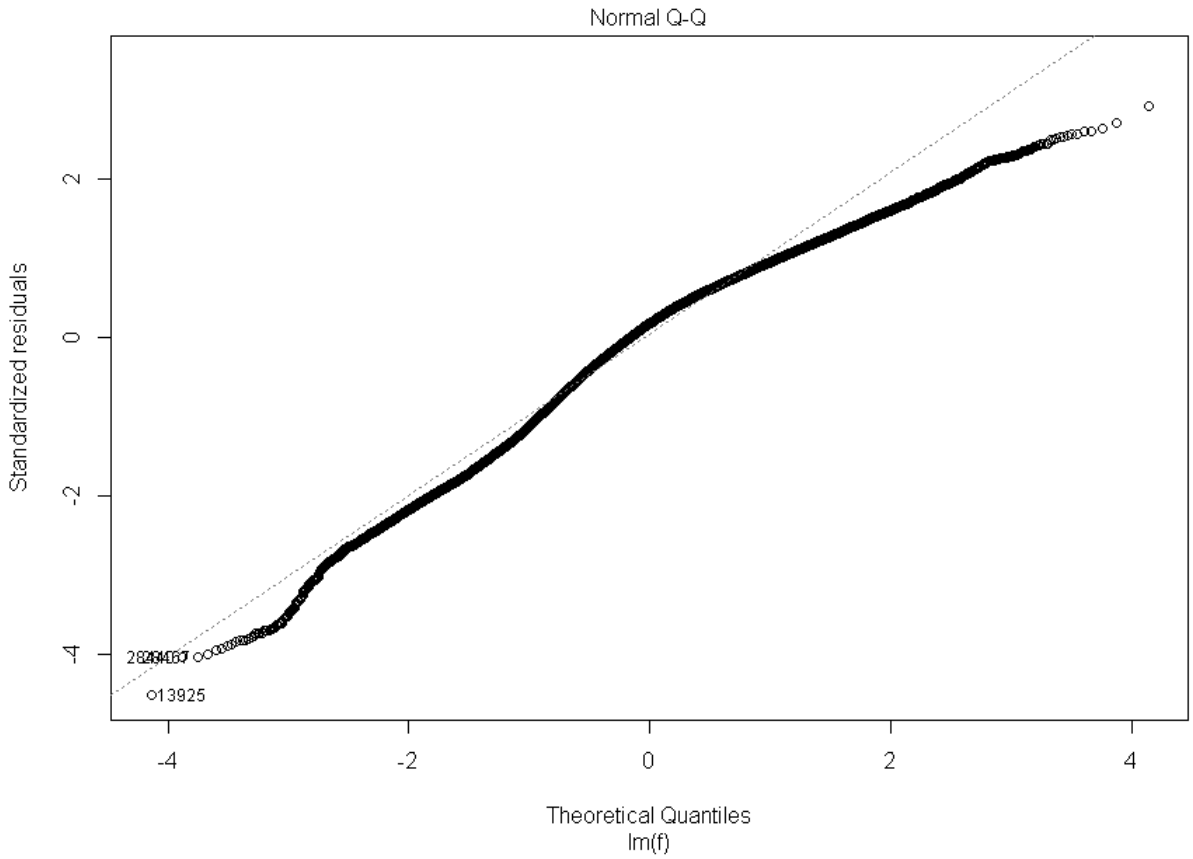

Fig. 5: Cuantil-cuantil del Modelo 1

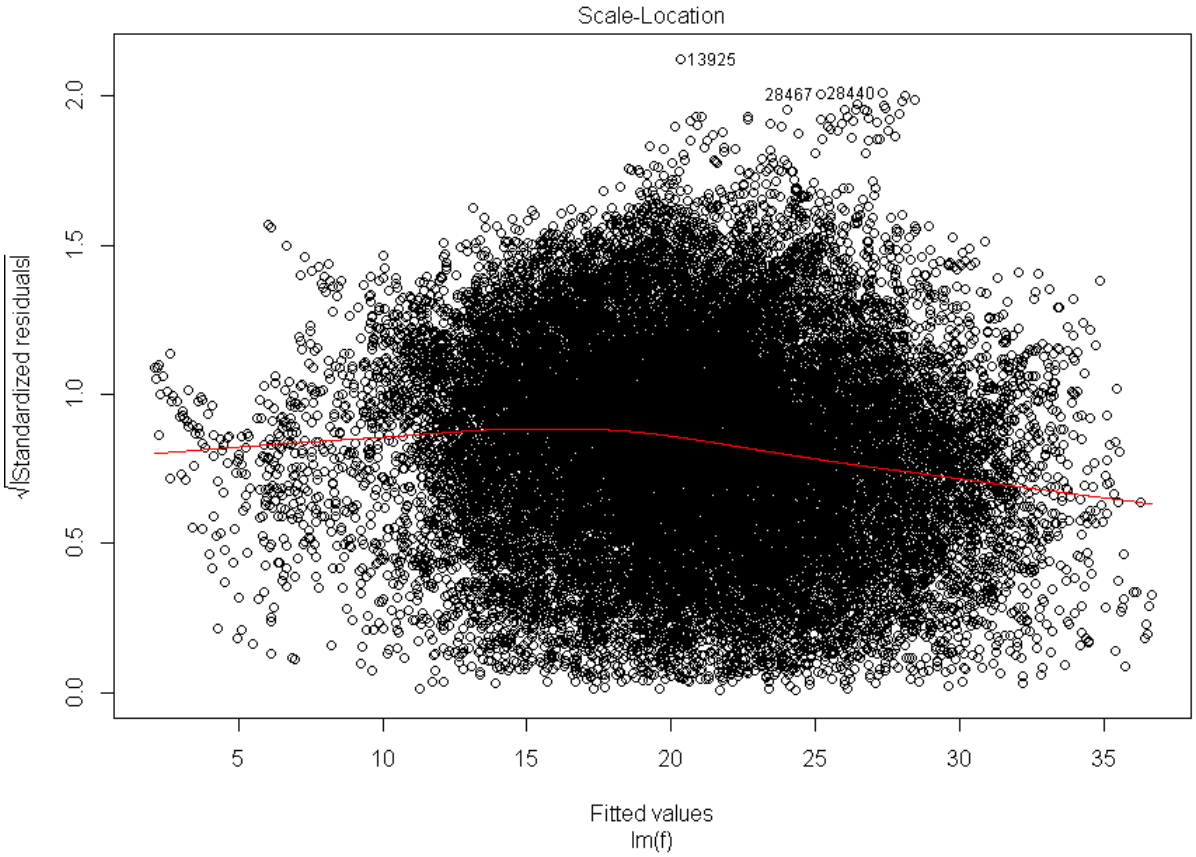

Fig. 6: Scale location

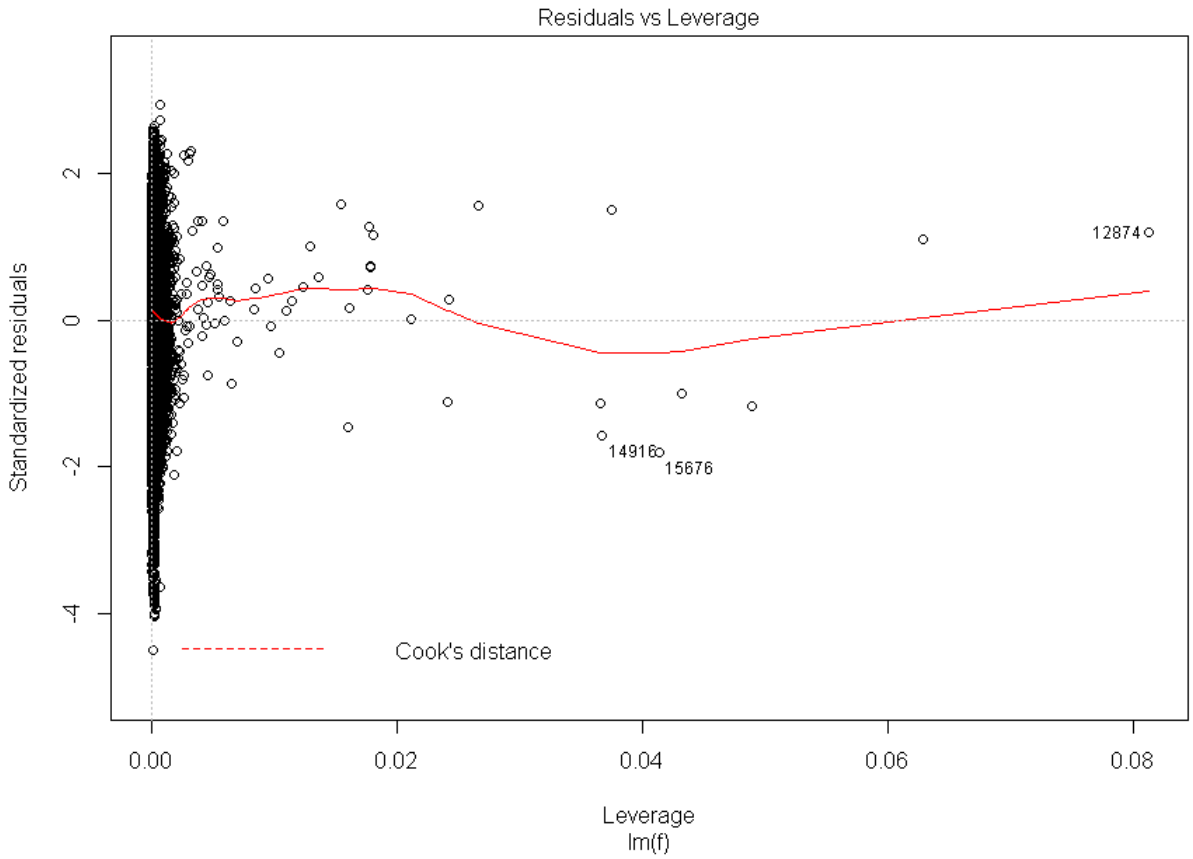

Fig. 7: Valores residuales contra valores de apalancamiento del Modelo 1

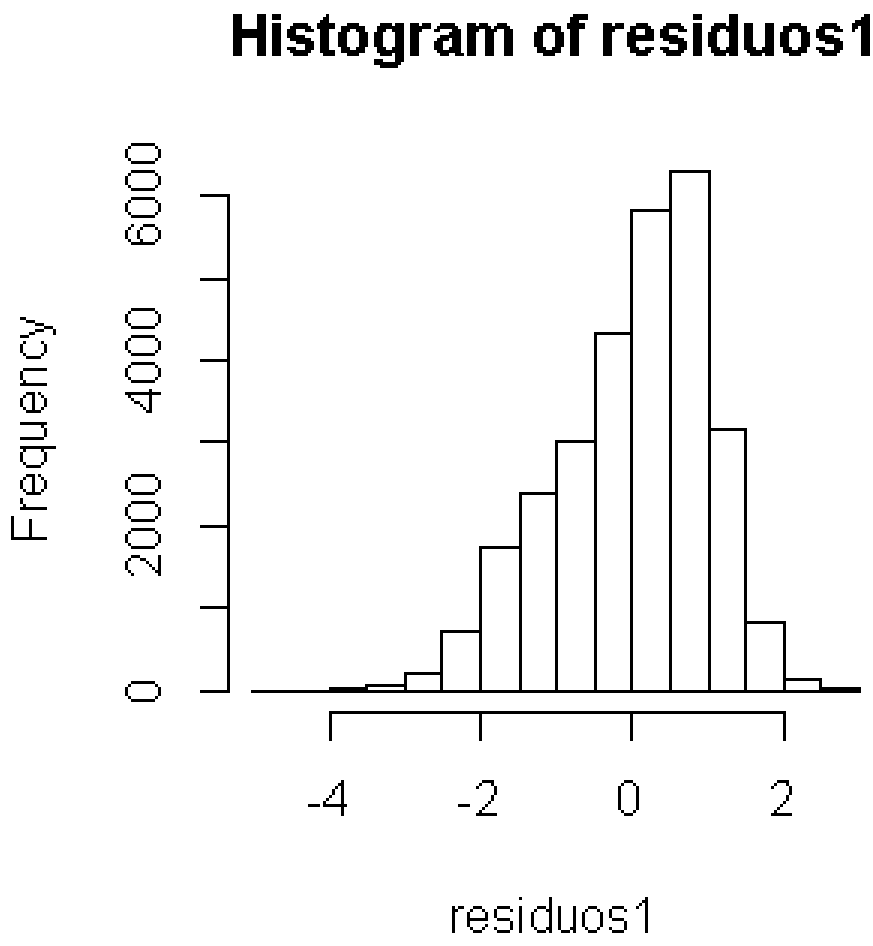

Fig. 8: Histograma de residuos del Modelo 1

2.3.2 Métodos de selección de variables

Luego de visualizar el primer modelo procedemos a realizar diferentes técnicas de selección de variables, para crear un modelo que cumpla con el Principio de Parsimonia.

Para esto hacemos uso de la función **step()**, en la cual utilizamos el método hacia adelante, hacia atras y ambos, y luego con la función **regsubsets()** buscamos cuales son los mejores modelos con diferentes cantidades de variables.

El resultado de esto se puede expresar en el siguiente gráfico, donde vemos como el porcentaje de R^2 va aumentando de a poco y se tiñe en negro la columna de la variable utilizada.

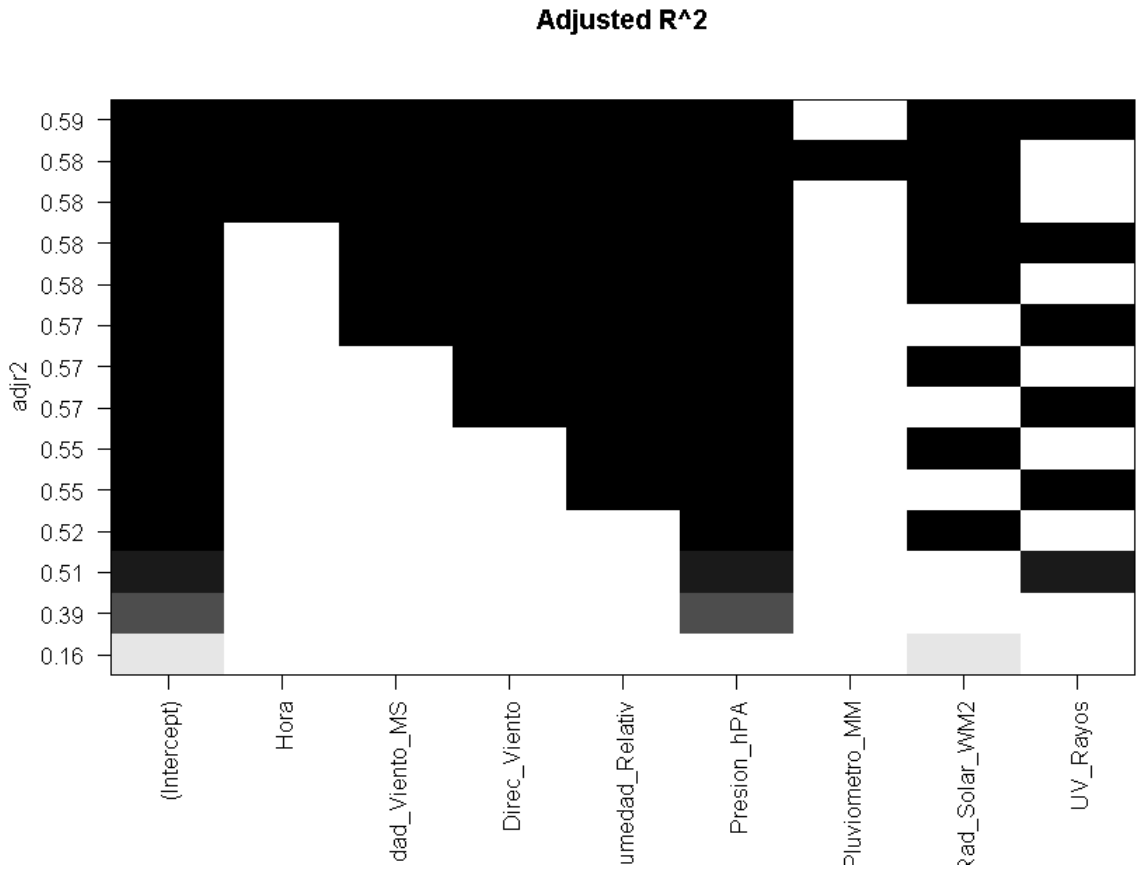

Fig. 9: R cuadrado

2.3.3 Modelo 2

A partir del análisis previamente hecho podemos crear un nuevo modelo, tomando menos variables, excluyendo aquellas que sean menos significativas.

Las variables a utilizar seran: Dirección del viento, Humedad Relativa, Presion en hPA y Radiación solar en WM2.

Como de costumbre, procedemos a realizar el resumen estadístico y analizamos:

```
> summary(m2)

Call:
lm(formula = Temperatura_C ~ Direc_Viento + Humedad_Relativ +
    Presion_hPA + Rad_Solar_WM2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.287  -2.769   0.815   3.151  13.392

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.281e+02  5.127e+00  161.52   <2e-16 ***
Direc_Viento    -8.350e-03  2.482e-04  -33.65   <2e-16 ***
Humedad_Relativ -8.557e-02  1.980e-03  -43.21   <2e-16 ***
Presion_hPA     -7.924e-01  5.058e-03 -156.67   <2e-16 ***
Rad_Solar_WM2    7.185e-03  1.176e-04   61.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.31 on 28716 degrees of freedom
Multiple R-squared:  0.5691,    Adjusted R-squared:  0.569
F-statistic:  9480 on 4 and 28716 DF,  p-value: < 2.2e-16
```

Fig. 10: summary() del Modelo 2

Comparando con el modelo anterior obtenemos un R^2 un tanto menor, un p-value igual, pero utilizando menos variables.

Visualizamos el modelo para comprobar sies acertado:

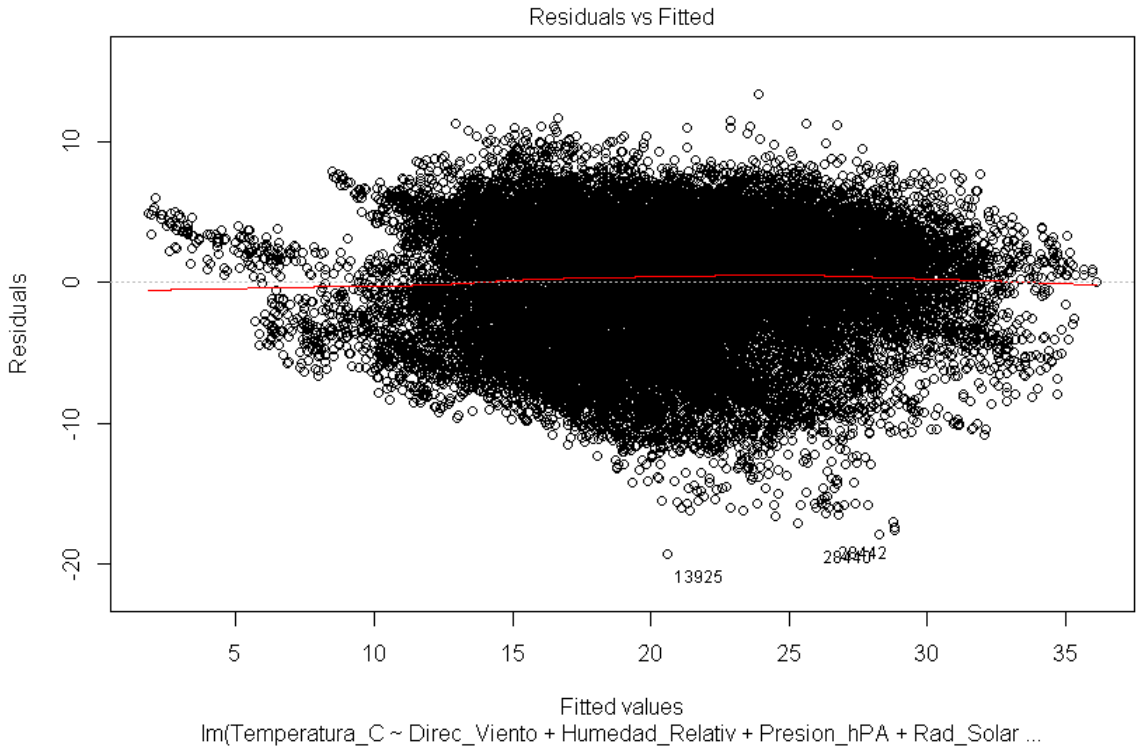

Fig. 11: Valores residuales contra valores ajustados del Modelo 2

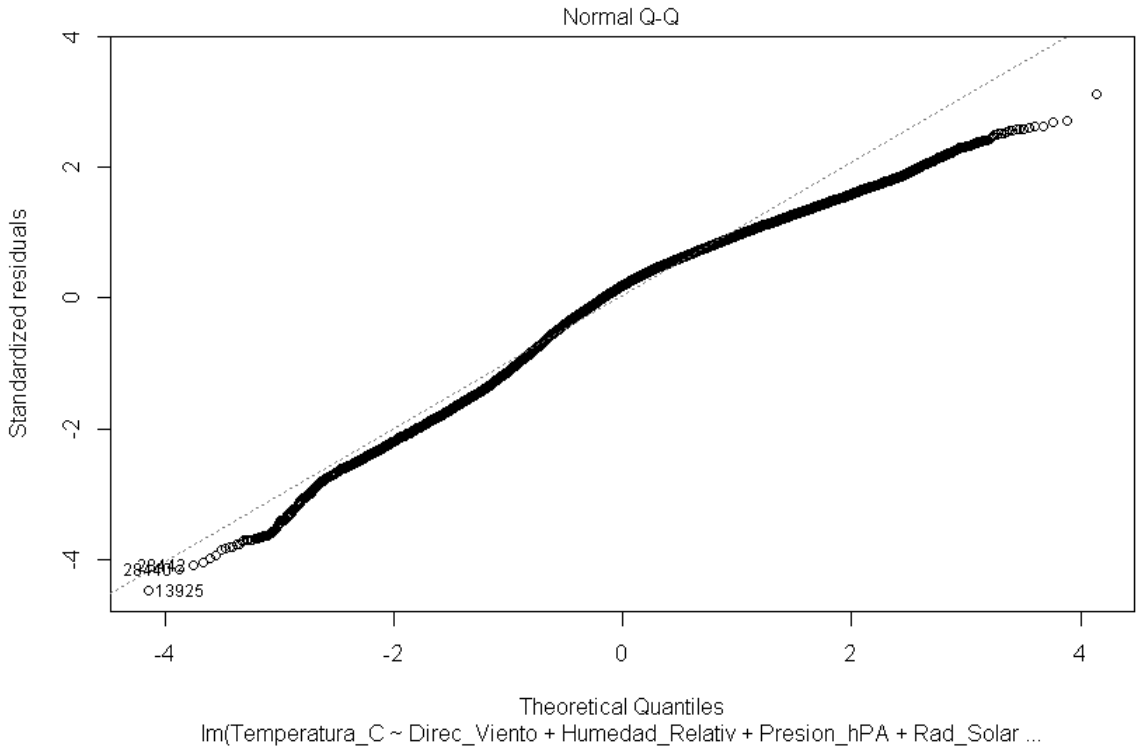

Fig. 12: Cuantil-cuantil del Modelo 2

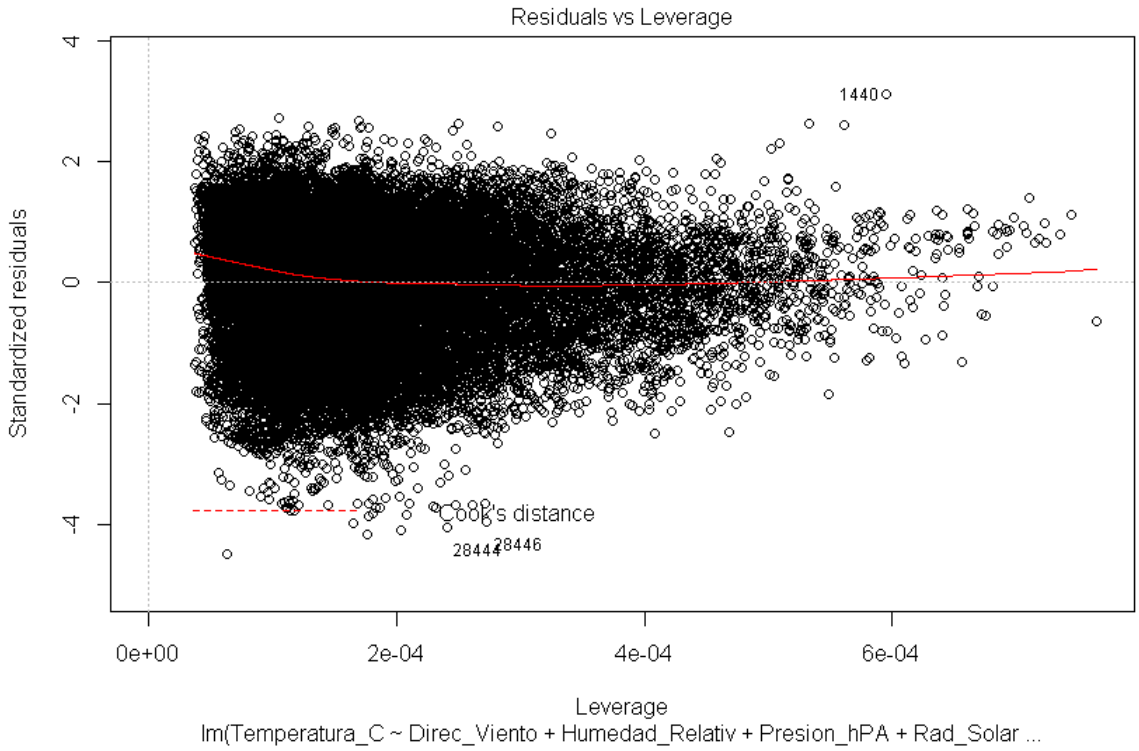

Fig. 13: Valores residuales contra valores de apalancamiento del Modelo 2

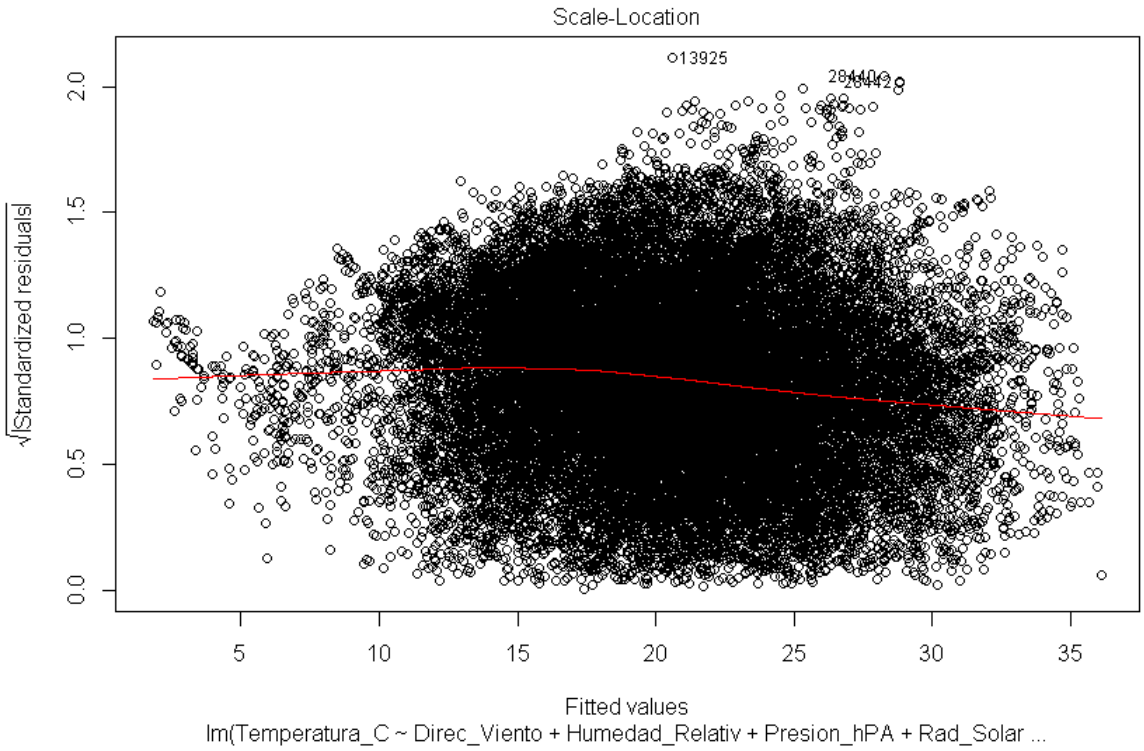

Fig. 14: Scale location del Modelo 2

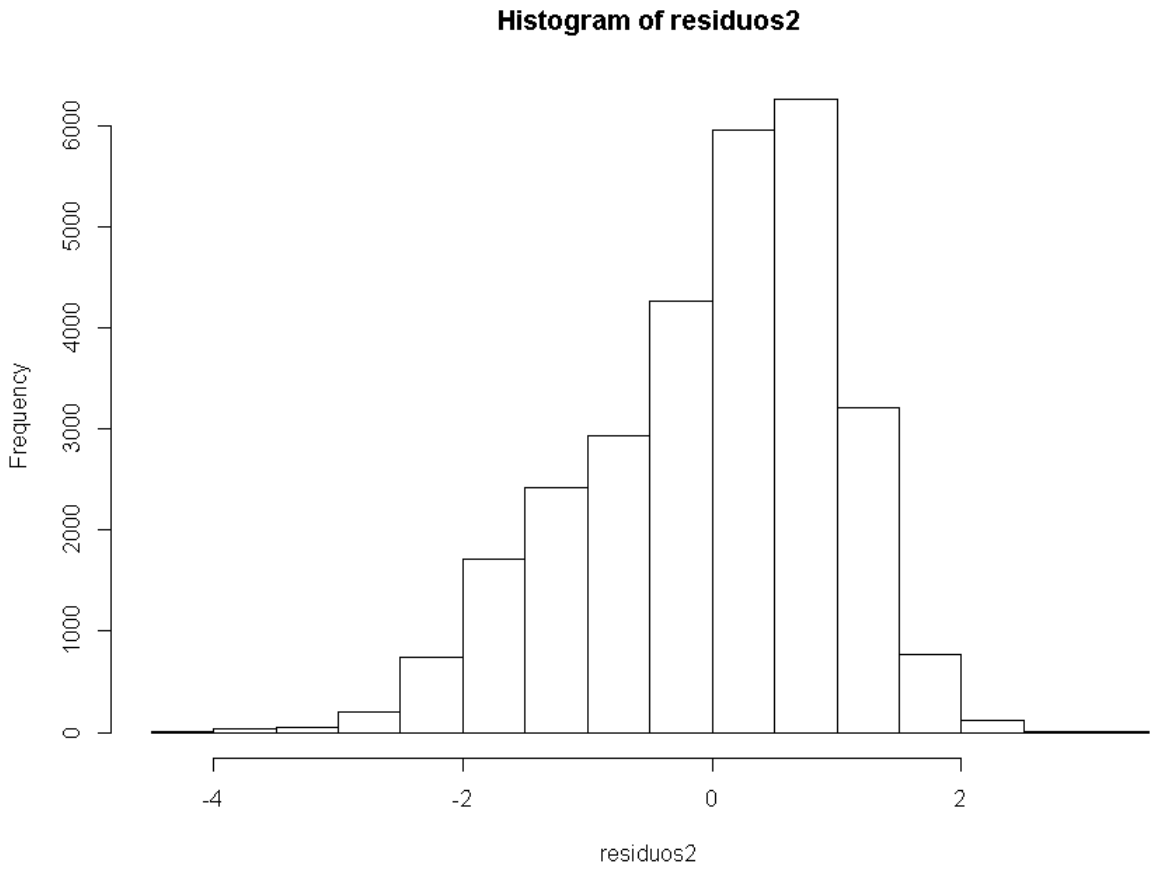

Fig. 15: Histograma de residuos del modelo 2

No se observan patrones que indiquen que el modelo este mal.
También visualizamos las distribuciones de los residuos:

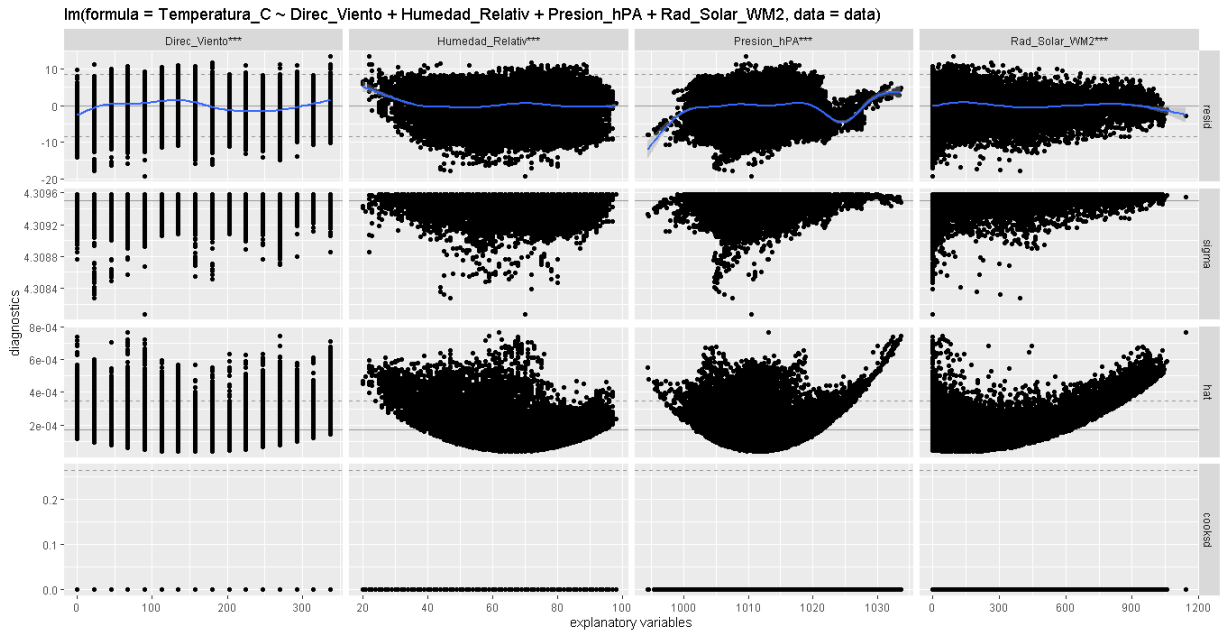

Fig. 16: Distribución de residuos de cada variable

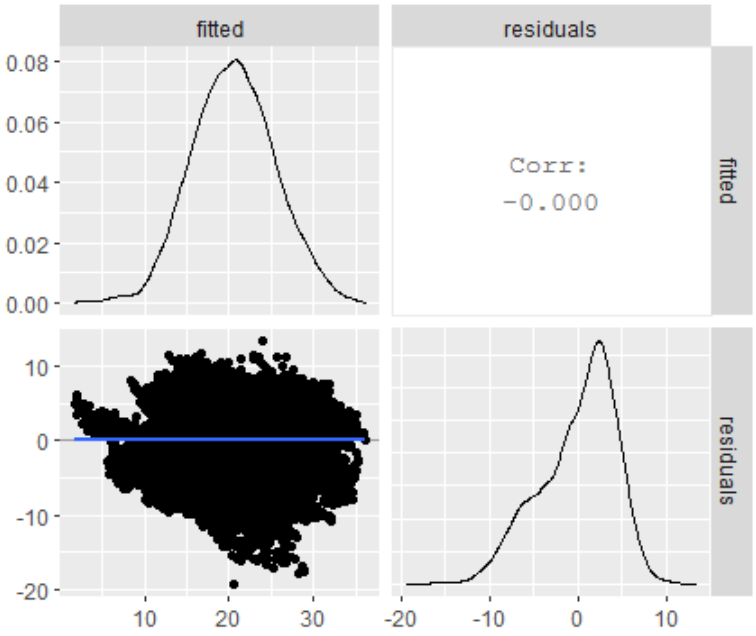

Fig. 17: Curva y distribución de residuos

Podemos observar que si bien no es una distribucion normal perfecta, se asemeja a como deberia ser.

Analizamos si hay colinealidad entre las variables predictoras

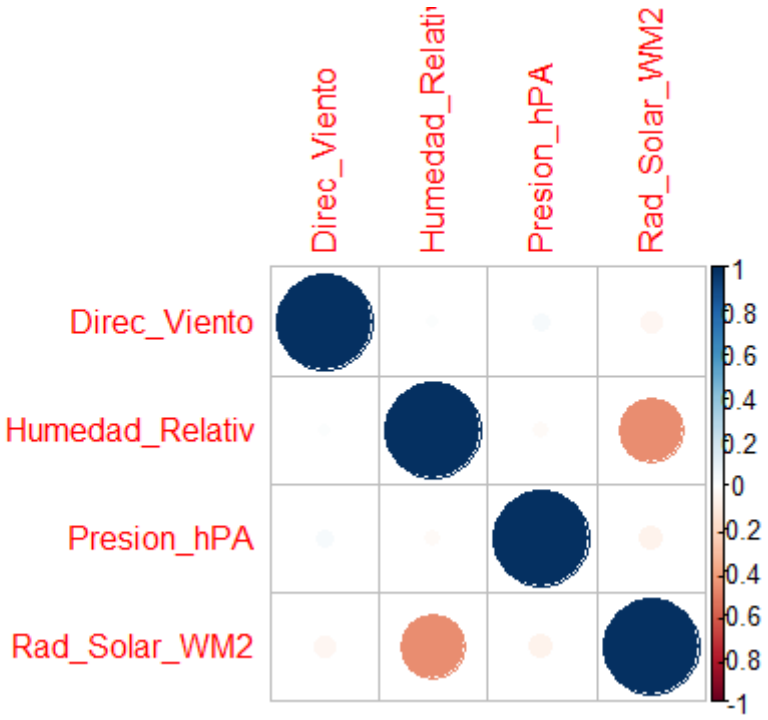

Fig. 18: Corrplot 2

2.3.4 Contrastación de $\hat{Y}$ con Y

Por último contrastamos nuestra variable dependiente (Y) y nuestra variable ya predecida ($\hat{Y}$) y podemos observar que la siguiente formula respalda y avala el modelo realizado.

$$\hat{y} = Y + \epsilon$$

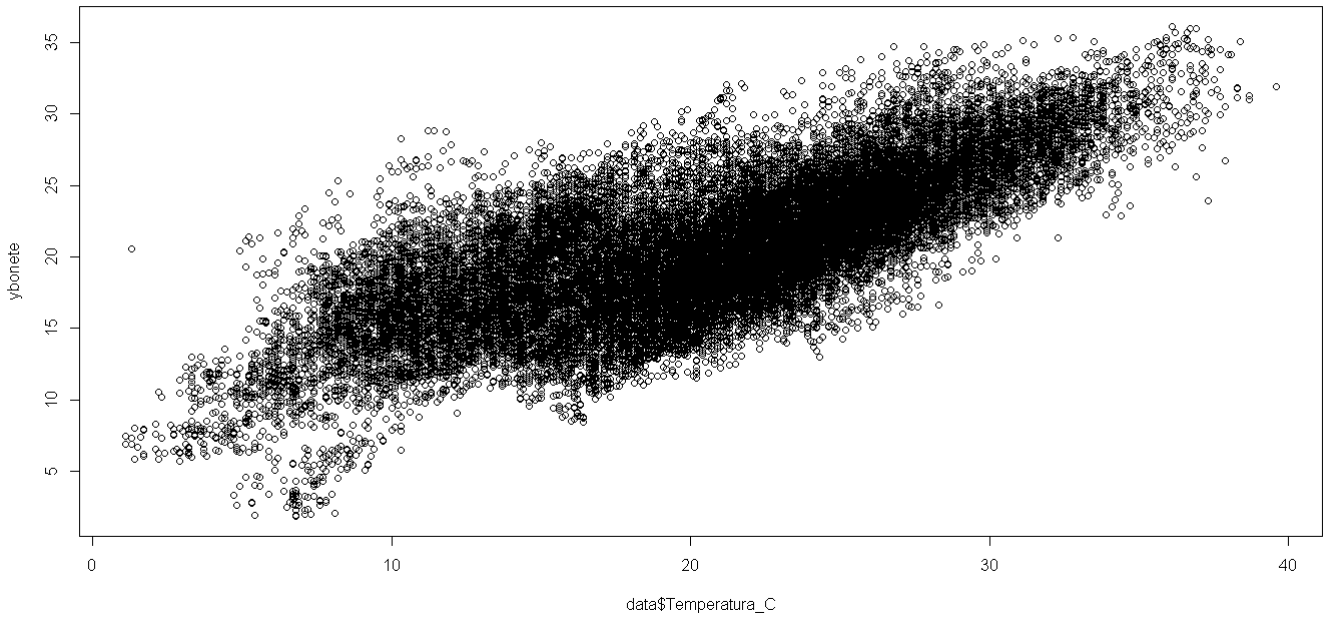

Fig. 19: Comparación entre Y contra $\hat{Y}$

3 Conclusión

Luego de aplicar el análisis de regresión multilineal obtuvimos 2 modelos, el primero con un acierto del 58,6% y el segundo con un acierto de 56,9%, aplicando el principio de parsimonia al tomar menos variables, lo que conlleva a una cantidad menor de errores residuales. Por otro lado, observamos que mediante nuestros β podremos predecir el valor de Y con una aproximación que será el valor de nuestro residuo $\pm\epsilon$.

4 Anexo

4.1 Código en R

```
##
# Modelo de Regresion lineal Multiple, Actividad 3
# Creado por Grupo B: Benitez, Garcia, Rodriguez, Rechimon
#
# Creado: 24/09/2020 v. 25/09/2020
# Ultima mod: resumen de distribucion de residuos
##

#### Importamos bibliotecas
library(readr)
library(dplyr)
library(corrplot)
library(data.table)
library(leaps)
library(ggplot2)
library(GGally)
library(lmtest)
library(nortest)
library(fmsb)
library(olsrr)


#### LEEMOS EL DATASET
temperaturas <- read_csv("informacion-meteorologica-2012.csv",
                         col_types = cols(HORA = col_number()))

# Corroboramos que se haya leido bien
head(temperaturas, 5)


##### PREPARAMOS LOS DATOS
data <- temperaturas[, c(2,4:11)] #Variables innecesarias
colnames(data) <- c("Hora", "Velocidad_Viento_MS", "Direc_Viento","
    Temperatura_C", "Humedad_Relativ", "Presion_hPA", "Pluviometro_
    MM", "Rad_Solar_WM2", "UV_Rayos" )

summary(data) #Resumen estadistico de los datos

data <- subset(data, data$Temperatura_C!=1802.6 & data$Humedad_
    Relativ!=255.0) #Outliers


#### Visualizamos las Correlaciones
cor_data <- cor(data[,])
corrplot(cor_data)


#### Creamos el MODELO 1
## Creamos la formula
```

```r
variables <- c("Hora","Velocidad_Viento_MS", "Direc_Viento", "Humedad_Relativ", "Presion_hPA", "Pluviometro_MM", "Rad_Solar_WM2", "UV_Rayos" )
y <- "Temperatura_C"  #  variable a predecir

f <- as.formula(paste(y,paste(variables, collapse = " + "),sep = " ~ "))
print(f) #  aqui vemos a todas las variables


## Creamos el Modelo FULL Variables
lm.full <- lm(f, data = data)
summary(lm.full)
print(lm.full)

#par(mfrow=c(2,2),4)
plot(lm.full)
residuos1 <- rstandard(lm.full)
mean(residuos1)
hist(residuos1)


#### Creamos el Modelo NULL Variables
lm.null <- lm(Temperatura_C ~ 1, data = data ) # Intercept-only
summary(lm.null)

anova(lm.full, lm.null ) #Contrastamos modelos, si p-value es chico podemos partir del null hacia full


#### Aplicamos Metodos de Seleccion de Variables
lm.step.bw   <- step(lm.full, direction = "backward")

step.fw      <- step(lm.null,
                     scope = ~ Hora + Velocidad_Viento_MS + Direc_Viento + Humedad_Relativ + Presion_hPA + Pluviometro_MM + Rad_Solar_WM2 + UV_Rayos,
                     direction = "forward")

lm.step.both <- step(lm.null,
                     scope = ~ Hora + Velocidad_Viento_MS + Direc_Viento + Humedad_Relativ + Presion_hPA + Pluviometro_MM + Rad_Solar_WM2 + UV_Rayos,
                     direction = "both")


#### Evaluamos el mejor modelo
regsubsets.out <- regsubsets(f
                             , data = data
                             , nbest = 2    #  cuantos  mejores modelos quiero por cada cantidad de variables
                             , nvmax = 7  # max tama?o del modelo
                             , force.in = NULL, force.out = NULL
                             , method = "exhaustive")
summary(regsubsets.out)

```

```
plot(regsubsets.out, scale = "adjr2", main = "Adjusted R^2")


#### Creamos Modelo 2 a partir de los analisis
m2      = lm(data = data, Temperatura_C ~ Direc_Viento + Humedad_
    Relativ + Presion_hPA + Rad_Solar_WM2)


#### Analisis de la Bondad del Modelo
#Resumen
summary(m2)
plot(m2)

#Residuos
residuos2 <- rstandard(m2)
mean(residuos2)
hist(residuos2)

#Mas Resumenes sobre residuos
p_ <- GGally::print_if_interactive
pm <- ggnostic(m2)

p_(pm)

pm <- ggpairs(
  m2, c(".fitted", ".resid"),
  columnLabels = c("fitted", "residuals"),
  lower = list(continuous = ggally_nostic_resid)
)
p_(pm)

## Agrupamiento de las direcciones del viento por patron raro
# library(sqldf)
# sqldf("SELECT Direc_Viento FROM data GROUP BY Direc_Viento")

#Test de Kolmogorov-Smirnov
lillie.test(residuos2) # si  p-value < 0.05 entonces NO hay
    distribucion normal
#   REVISAR

# Analisis de Colinealidad: Durbin-Watson Test
dwtest(m2) # si p-value < 0.05 entonces hay autocorrelacion
#   REVISAR

VIF(m2) # si p-value > 10 hay colinealidad
ols_vif_tol(m2) # si VIF> 4 hay que investigar si hay colinealidad

cor_model <- cor(data[, c(3,5,6,8)])
corrplot(cor_model)

# Breusch-Pagan Test:
bptest(m2)  # si  p-value < 0.05 entonces la varianza de los
    residuos es homocedastica
#   REVISAR

#### Comparacion de Y con Ybonete
```

```
147 ybonete <- predict(m2, data)
148 plot(x=data$Temperatura_C, y= ybonete)
```