Título

Titulación:

06MBID_10_B_2022_23_Estadistica_avanzada

Curso académico

2022 – 2023

Alumno/a: Garrido Diez, Verónica

D.N.I: 3469477L

Universidad Internacional de Valencia

ACTIVIDAD 1

De:
Planeta Formación y Universidades

Universidad Internacional de Valencia

Índice

1. Introducción

1.1. Contexto y motivación

Los datos que vamos a analizar provienen de 517 incendios del parque natural de Montesinho, en Portugal. En este estudio aparecen varios datos meteorológicos como lluvia, temperatura, humedad relativa del aire y viento, que son en los que estoy interesada.

Quizás podríamos buscar relaciones entre estas variables meteorológicas que son fáciles de medir en la naturaleza y que para su medición no necesitan aparatos que no estén conectados a ninguna fuente de energía, a ninguna red de datos.

Una vez analizada la relación entre las variables meteorológicas podríamos establecer también algún tipo de relación con el hecho de que haya ocurrido el incendio y de qué magnitud ha sido. Aunque en este estudio no lo vamos a abordar en esta práctica. Aquí sólo vamos a buscar relaciones entre las variables meteorológicas.

Con esto mi idea sería poner estaciones meteorológicas sencillas que constarías de 4 elementos: higrómetro (humedad del aire), termómetro (temperatura), anemómetro (velocidad viento) y pluviómetro (precipitación). Estas estaciones estarías repartidas por el parque de forma que los visitantes pudieran ver los datos y según unas tablas que podríamos elaborar con los datos sacados del estudio vieran el riesgo de incendio para ese día según las condiciones climáticas.

1.2. Objeto del análisis

Estudiar la relación entre variables meteorológicas con el fin de informar a los visitantes del parque la probabilidad de incendio que hay

2. Descripción de los datos a analizar

Los datos los hemos extraído del siguiente estudio:

https://www.kaggle.com/datasets/uttam94/forest-forest-dataset

En este estudio se entrena un modelo de regresión basado en las variables espaciales, temporales y meteorológicas para predecir la superficie quemada en función de la fecha actual y de las coordenadas en las que se ha avistado el incendio.

La información sobre los atributos del dataset es la siguiente:

X – coordenada espacial del eje x dentro del parque: 1 to 9
Y – coordenada espacial del eje y dentro del parque: 2 to 9
month – mes del año : "Jan" to "dec"
day – día de la semana: "mon" to "sun"
FFMC - FFMC índice FWI system: 18.7 to 96.20(**)
DMC - DMC índice FWI system: 1.1 to 291.3(**)
DC - DC índice FWI system: 7.9 to 860.6(**)
ISI - ISI índice FWI system: 0.0 to 56.10(**)
temp – temperatura en grados Celsius: 2.2 to 33.30
RH – humedad relativa del aire in %: 15.0 to 100

wind – velocidad del viento in km/h: 0.40 to 9.40
rain – lluvia caída en mm/m2 : 0.0 to 6.4
área – Superficie quemada del bosque (en ha): 0.00 a 1090.84

(**) Son índices sobre el sustrato y la propagación del fuego, no los vamos a tomar en cuenta para nuestro estudio.

3. Análisis

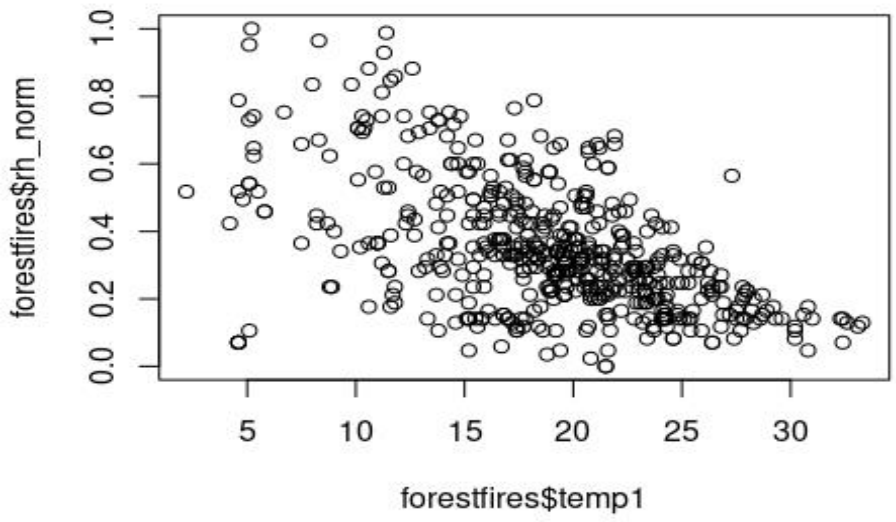
3.1. Regresión lineal/polinómica

Hacemos regresión lineal con las variables RH y temp. Intentamos explicar la variable RH en función de temp por lo tanto Rh será nuestra variable dependiente y temp será nuestra variable independiente.

Lo primero que hacemos es importar el dataset que vamos a utilizar y ver cómo se nos ha importado y describir los datos que traemos:

```
#RH va a ser la variable dependiente
#temp va a ser la variable independiente
#explicar la humedad en función de la temperatura
#Vemos primero cómo se han importado en R nuestros datos para tratar los
str(forestfires)

## tibble [517 × 15] (S3: tbl_df/tbl/data.frame)
##  $ X           : num [1:517] 7 7 7 8 8 8 8 8 8 7 ...
##  $ Y           : num [1:517] 5 4 4 6 6 6 6 6 6 5 ...
##  $ month       : chr [1:517] "mar" "oct" "oct" "mar" ...
##  $ n_month     : num [1:517] 3 10 10 3 3 8 8 8 9 9 ...
##  $ day         : chr [1:517] "fri" "tue" "sat" "fri" ...
##  $ FFMC        : chr [1:517] "86.2" "90.6" "90.6" "91.7" ...
##  $ DMC         : chr [1:517] "26.2" "35.4" "43.7" "33.3" ...
##  $ DC          : chr [1:517] "94.3" "669.1" "686.9" "77.5" ...
##  $ ISI         : chr [1:517] "5.1" "6.7" "6.7" "9" ...
##  $ temp        : chr [1:517] "8.2" "18" "14.6" "8.3" ...
##  $ RH          : num [1:517] 51 33 33 97 99 29 27 86 63 40 ...
##  $ wind        : chr [1:517] "6.7" "0.9" "1.3" "4" ...
##  $ rain        : chr [1:517] "0" "0" "0" "0.2" ...
##  $ bolean_rain: num [1:517] 0 0 0 1 0 0 0 0 0 0 ...
##  $ area        : chr [1:517] "0" "0" "0" "0" ...

##los datos que nos interesan los ha exportado como caracter, vamos a recodificar de la siguente forma

forestfires$temp1<-as.numeric(forestfires$temp)

##normalizacion de la variable humedad relativa para poder utilizarla
#define la función de normalización Min-Max
min_max_normRh <- function (x) {
  (x - minRh) / (maxRh - minRh)
}
min_max_normRhTr <- function (x) {
  (x - minRhTr) / (maxRhTr - minRhTr)
```

```
}
forestfires$rh1<-as.double(forestfires$RH)

minRh<-min(forestfires$rh1)
maxRh<-max(forestfires$rh1)
#aplicar la normalización Min-Max a la cOLUMNA rh ya transformada en
numerica
forestfires$rh_norm <- as.double(lapply (forestfires$rh1, min_max_no
rmRh))

###dibujamos la  nube de puntos de la humedad relativa frente a la t
emperatura
plot(forestfires$temp1,forestfires$rh_norm)
```



Dibujamos el gráfico de la línea de puntos para hacernos una idea de cómo están distribuidos y si estamos planteando algo que tiene sentido:

A primera vista vemos que puede existir una relación entre las dos variables. También vemos que la mayoría de valores se concentran en unos rangos. En este punto he pensado en eliminar los outliers, pero es un dataset pequeño, con sólo 517 observaciones, y en la descripción del dataset nos indican que en esas observaciones hubo incendio así que tenemos en cuenta todas las observaciones. Más adelante podemos estudiar el gráfico residuals vs leverage para ver los valores influyentes.

3.1.1. Estudio del resumen (summary)

```
##Aplicamos el modelo de regresión lineal a ambas variables
regresRh=lm(forestfires$rh_norm~forestfires$temp1,data=forestfires)
summary(regresRh)
## Call:
## lm(formula = forestfires$rh_norm ~ forestfires$temp1, data = fore
stfires)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52312 -0.09510 -0.01065  0.08442  0.51309
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.673915   0.024458   27.55   <2e-16 ***
## forestfires$temp1 -0.017436   0.001238  -14.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1633 on 515 degrees of freedom
## Multiple R-squared:  0.2781, Adjusted R-squared:  0.2767
## F-statistic: 198.4 on 1 and 515 DF,  p-value: < 2.2e-16
```

La recta de regresión de estas dos variables tiene la siguiente forma:

Rh_norm = 0.673915 -0.017436Temp

El valor de la pendiente es negativo y pequeño, lo cual quiere decir que si la variable rh_norm aumenta en una unidad la temperatura disminuye en -0.017436.

Ahora tememos que hacer dos validaciones, una global para ver si el modelo es bueno o no y las variables que estoy relacionando tienen sentido y otra validación individual para ver si la variable explicativa temp también es buena y sirve para explicar la variable RH.

Para hacer la validación individual y ver si los valores son estadísticamente significativos nos fijamos en los valores Pr(>|t) <2e-16 ***. Este valor es menor de 0.05, el valor que hemos decidido tomar como referencia, por lo tanto podemos decir que la variable independiente temp es buena

Para hace la validación global del modelos nos fijamos en el valor de p-value: < 2.2e-16. Este valor también nos dice que modelo es bueno si es menor a 0.05, y es este caso lo es por lo tanto la validacón golbal del modelo es buena.

Otro valor en el que nos fijamos es en el coeficiente de determinación Multiple R-squared: 0.2781. Para que el modelo sea bueno este coeficiente tiene que ser mayor de 0.5. En este caso es menor, esto nos quiere decir que las variables no están muy relacionadas entre sí. La variable humedad relativa sólo es explicada por la temperatura un 27,81%.

Para hacer un mejor análisis dividimos el dataset entre entrenamiento (80%) y test(20%). Fijamos una semilla para que los conjuntos de datos siempre sean los mismos y volvemos hacer regresión sobre cada parte a ver qué obtenemos.

Para la parte del entrenamiento:

```
regres_trainRh=lm(train$rh_norm~train$temp1,data=train)
summary(regres_trainRh)
```

```
##
## Call:
## lm(formula = train$rh_norm ~ train$temp1, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51890 -0.09296 -0.01199  0.08593  0.51639
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669076   0.027639   24.21   <2e-16 ***
## train$temp1 -0.017301   0.001392  -12.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1628 on 411 degrees of freedom
## Multiple R-squared:  0.2731, Adjusted R-squared:  0.2713
## F-statistic: 154.4 on 1 and 411 DF,  p-value: < 2.2e-16
```

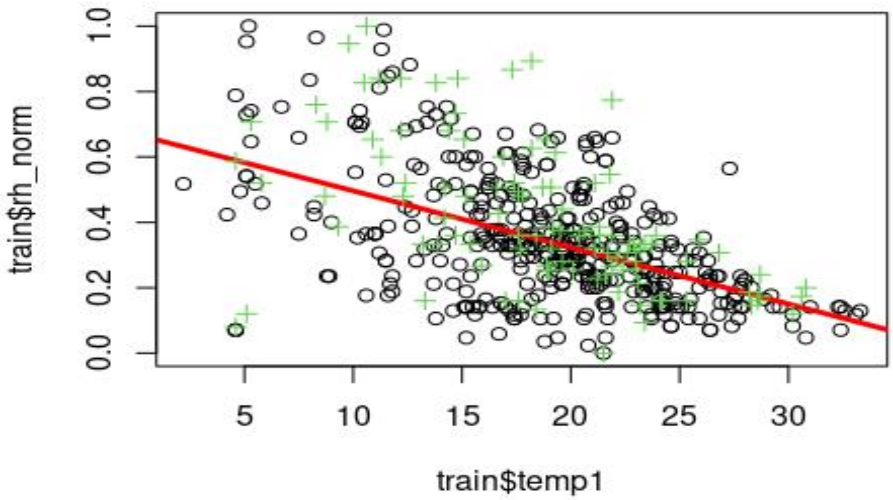
Para la parte de test:

```
regres_test= lm(test$rh_norm~test$temp1,data=test)
summary(regres_test)
```

```
##
## Call:
## lm(formula = test$rh_norm ~ test$temp1, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60880 -0.10748 -0.01078  0.08799  0.47887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.781591   0.060094  13.006  < 2e-16 ***
## test$temp1  -0.020172   0.003098  -6.512 2.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1884 on 102 degrees of freedom
## Multiple R-squared:  0.2936, Adjusted R-squared:  0.2867
## F-statistic:  42.4 on 1 and 102 DF,  p-value: 2.84e-09
```

Obtenemos valores muy parecidos para los datos de train y la muestra entera y un poco distintos para los de test, pero todos los datos significativos del resumen se interpretan de la misma forma.

En el siguiente gráfico podemos ver de negro los puntos del train y en verde los del test. En color rojo la línea de regresión. Podemos observar que los datos del test y train son parecidos y se distribuyen de la misma forma sobre la línea de regresión.



3.1.2. Estudio de los residuos

En una primera aproximación vemos cómo es la media tanto en la muestra entera como en la parte de train y test.

Para que el modelo sea bueno los residuos tiene que cumplir que siguen una distribución normal de media 0.

Para ver este supuesto lo podemos hacer de varias formas, podemos mirara la media de los residuos y ver si se aproximan a 0, podemos hacer una inspección gráfica y compararlo con una distribución normar. Para nuestra distribución hemos lo hemos estudiado de estas dos formas y hemos obtenido lo siguiente:

Media(muestra_entera)= -3.169886e-17

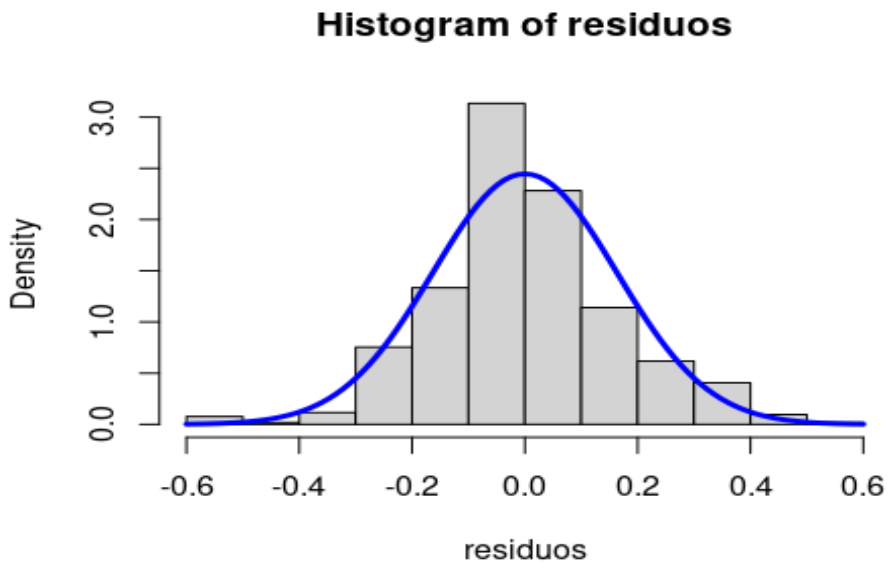
Media(train) =-2.162112e-18

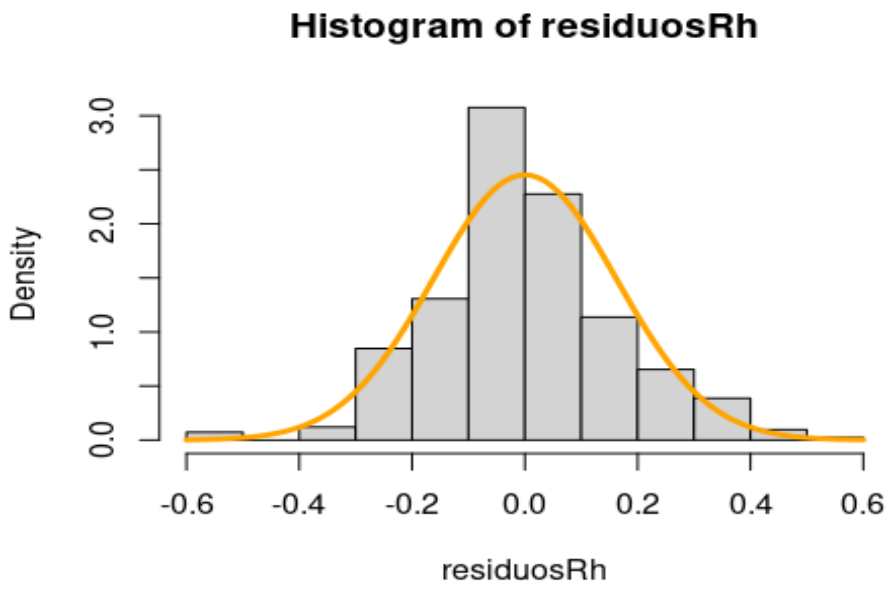
Media (test) = 3.802657e-18

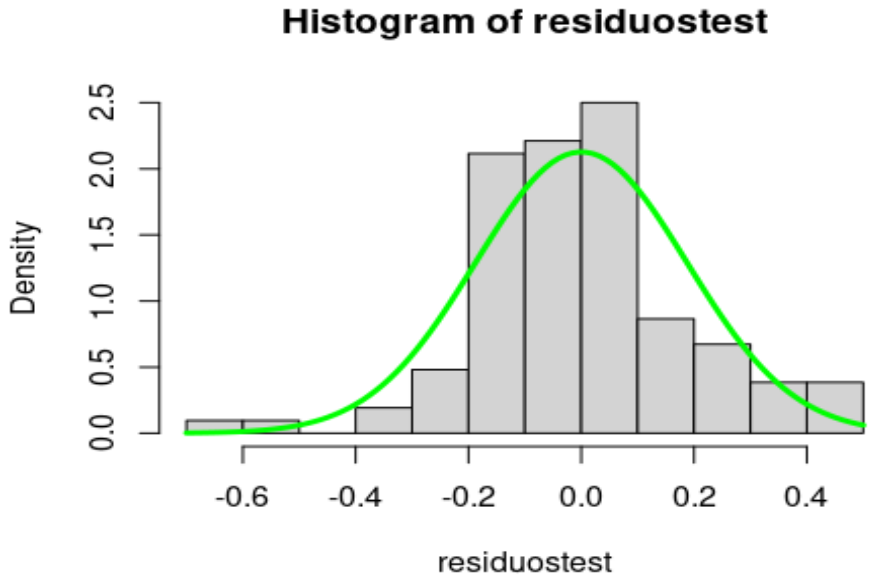
Las medias se aproximan a 0.

Si lo visualizamos en forma de gráfico:

```
# Histograma de los residuos con la curva de normalidad

(*resisuosRh son los residuos del train)
```





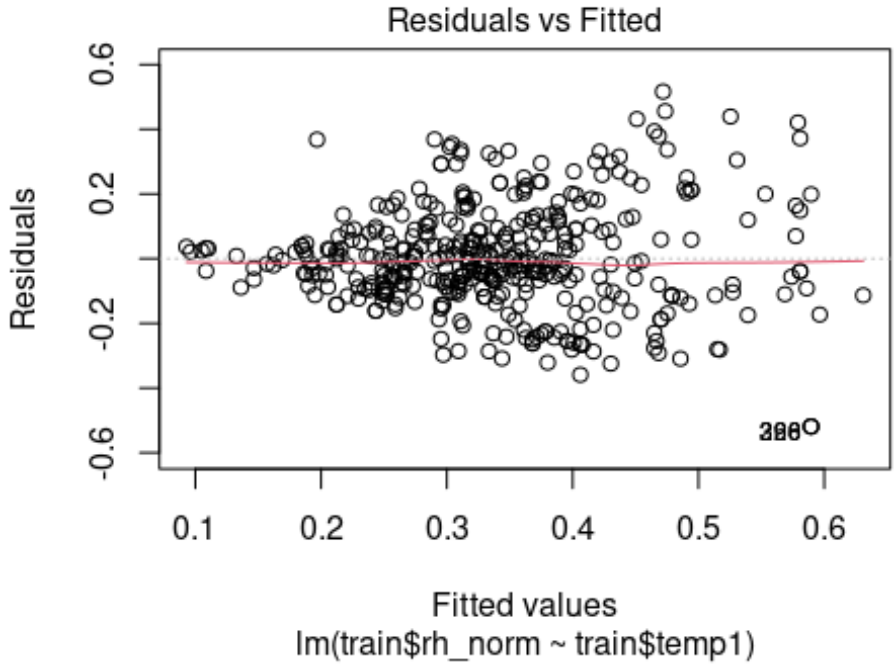


Parece que no se cumple muy bien el supuesto de normalidad. La media sí se aproxima a cero, pero viendo las gráficas no parece que sigan una distribución normal. Además observamos que la parte de los datos des test es diferente las otras dos.

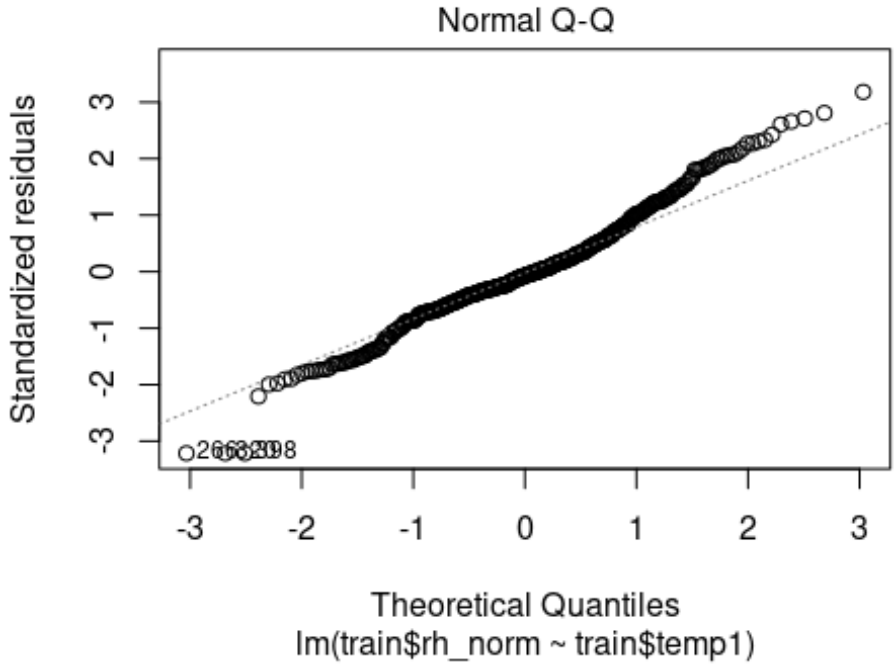
Analizamos la normalidad en los residuos mirando los gráficos Residuals vs fitted, Normal Q-Q, scale-location y residualsvs leverage:

Los mostramos sólo de la parte del train.

En estos gráficos vemos ya que nuestro modelo no es muy bueno. El primer gráfico muestra si los residuos tienen patrones no lineales, en nuestro caso los residuos no se distribuyen muy bien alrededor de la línea horizontal, esto nos hace sospechar que algo raro pasa, puede haber relaciones que no sean lineales.

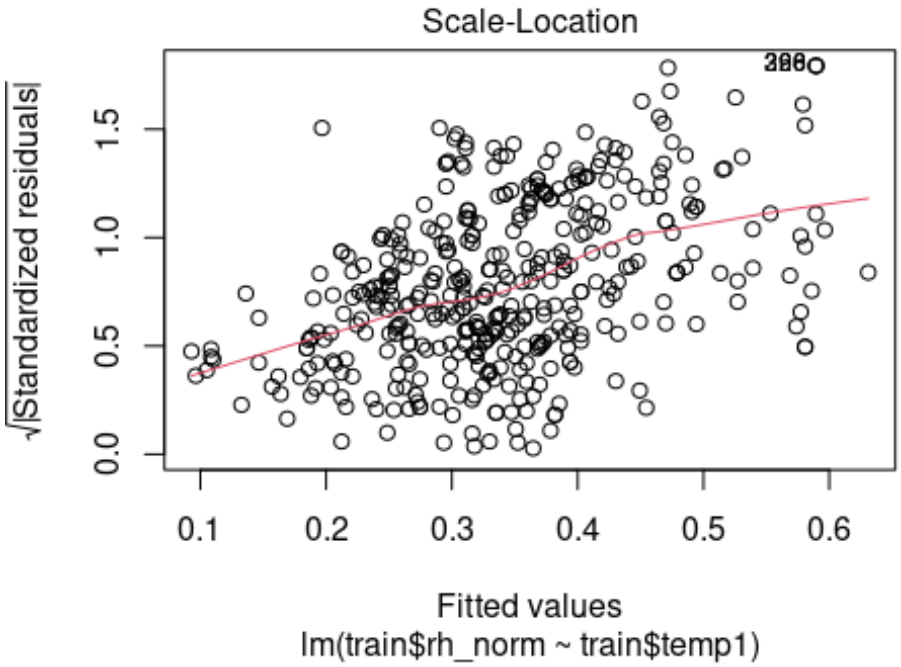


El Segundo gráfico nos muestra si los residuos se distribuyen normalmente, vemos que los valores centrales si se ajustan a la recta, pero los valores finales se van dispersando de la recta, esto quiere decir que a partir de un punto la distribución deja de ser normal.

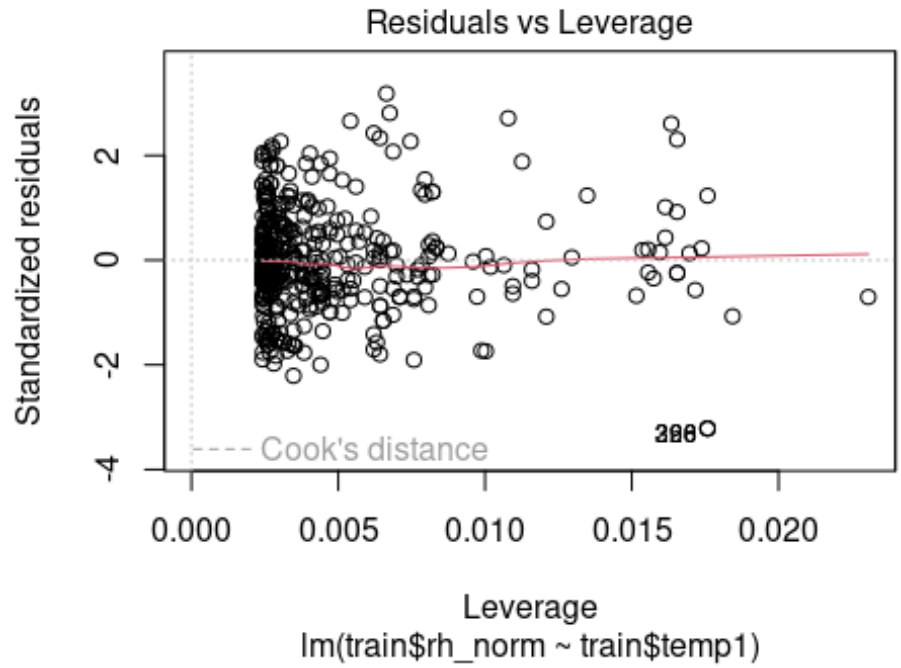


En el tercer gráfico podemos ver si los residuos se distribuyen por igual a lo largo de los rangos predictores. Aquí se verifica el supuesto homocedastididad (varianza constante). El resultado es correcto si se ve una línea horizontal con puntos disperses al azar. En nuestro gráfico podemos ver que la línea no es horizontal, tiene cierta

tendencia ascendente, aunque parece que los puntos si se distribuyen al azar, aunque se encuentran más concentrados en la zona central



El ultimo gráfico nos muestras si podemos tener observaciones influyentes, valores atípicos. Es el más difícil de interpretar. En nuestro caso se ve que hay valore que están bastante alejados de la recta de color rojo, habría que revisarlos o eliminarlos, pero esto no nos garantiza que el modelo mejore.



Otros indicadores para ver si nuestro modelo cumple la normalidad lo dan test de normalidad como el shapiro y Kolmogorov-Smirnov. Podemos aplicar estos test a nuestro modelo a ver qué dan. Los valores que obtuvimos fueron los siguientes:

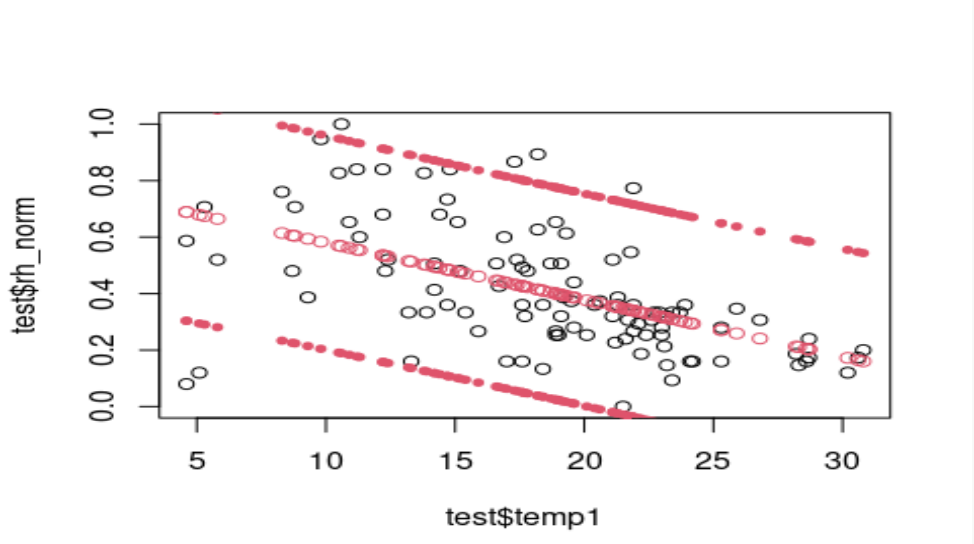
```
##aplicamos el shapiro menos 50 casos
##hipotesis nula es que la distribución es normal
shapiro.test(residuals(regres_trainRh))##
##  Shapiro-Wilk normality test
## data:  residuals(regres_trainRh)
## W = 0.98373, p-value = 0.0001357

##aplicamos el test de Kolmogorov-Smirnov para ver si los residuos
##cumplen la distribución normal.
#install.packages('nortest')
library(nortest)
lillie.test(residuals(regres_trainRh))
##  Lilliefors (Kolmogorov-Smirnov) normality test
## data:  residuals(regres_trainRh)
## D = 0.068144, p-value = 9.169e-05
```

En los dos text el p_value es menor de 0.05, por lo tanto no cumple el supuesto de normalidad.

3.1.3. Predicción del modelo e intervalos de confianza

Aunque nuestro modelo parece que no es bueno, vamos hacer una predicción sobre el modelo y a mostrar los intervalos de confianza. Para ello utilizamos la parte de los datos que hemos apartado para el test.

```
##intervalos de confianza
predConfiTest<-predict(regres_test, newdata = test, interval = 'pred
iction')
plot(test$temp1, test$rh_norm)
#en la primera columna (fit) tenemos el ajuste
points(test$temp, predConfiTest[,1], col=2)
#en la segunda y tercera los limites inferior y superior del interva
lo
points(test$temp1, predConfiTest[,2], col=2, pch=20)
points(test$temp1, predConfiTest[,3], col=2, pch=20)
```



El intervalo de confianza está entre la recta superior y la inferior. Las rectas que marcan los intervalos de confianza tienen el mismo comportamiento y sus pendientes son negativas.

3.2. Regresión multilineal

Vamos hacer regresión multilineal con las variables asociadas al clima, la humedad relativa del aire, temperatura, viento y precipitación.

Como variable dependiente vamos a tomar otra vez la humedad relativa del aire, RH, y como variables independientes tomamos la temperatura (temp), viento (wind) y precipitación (rain).

Para el estudio separamos el dataset en un 80% de entrenamiento y en un 20% de test como lo hemos hecho para la parte de regresión multilineal fijando de nuevo la semilla.

Para este caso vamos a estudiar primero los valores del train

3.2.1. Estudio del resumen (summary)

```
##Hacemos Regresion multiple
multiAll<-lm(rh_norm~., data = trainMulti)
summary(multiAll)

##
## Call:
## lm(formula = rh_norm ~ ., data = trainMulti)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49188 -0.09247 -0.01466  0.08969  0.49128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.721385   0.036907  19.546  < 2e-16 ***
## temp1       -0.018298   0.001415 -12.928  < 2e-16 ***
## wind1       -0.008794   0.004586  -1.917  0.05588 .
## rain1        0.085890   0.024239   3.543  0.00044 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1602 on 409 degrees of freedom
## Multiple R-squared:  0.2993, Adjusted R-squared:  0.2942
## F-statistic: 58.23 on 3 and 409 DF,  p-value: < 2.2e-16
```

La información que podemos extraer de este resumen es la siguiente:

Analizamos el valor p-value: < 2.2e-16, esto nos indica si el modelo es significativo de forma global, en este caso el p-value es menor del valor de referencia que hemos tomado (0.05) y por lo tanto sería un modelo significativo. En la regresión multinelial es importante hace este análisis de varianza, ya que estamos comparando más de 2 variables.

Hacemos el análisis individual mirando para cada variable el valor del Pr(>|t|), de aquí podemos ver qué variables son significativas. Vemos que todas son menores de 0.05 (lo que significa que son estadísticamente significativas, excepto la variable wind.

Por lo tanto la podemos eliminar de nuestro modelo y así al eliminarla el modelo mejorará (parsimonia).

Por lo tanto al hacer regresión de nuevo eliminando la variable wind obtenemos lo siguiente:

```
##Hacemos regresion eliminando la variable wind
trainMulti2<-train[, c( 'rh_norm','temp1', 'rain1')]
multiAll2<-lm(rh_norm~., data = trainMulti2)
summary(multiAll2)

##
## Call:
## lm(formula = rh_norm ~ ., data = trainMulti2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52186 -0.09303 -0.01008  0.08932  0.51579
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.673627   0.027323   24.65  < 2e-16 ***
## temp1       -0.017648   0.001379  -12.80  < 2e-16 ***
## rain1        0.082463   0.024252    3.40 0.000739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1608 on 410 degrees of freedom
## Multiple R-squared:  0.293,  Adjusted R-squared:  0.2895
## F-statistic: 84.96 on 2 and 410 DF,  p-value: < 2.2e-16
```

El polinomio tiene la forma:

Rh_norm = 0.673627 -0.017648temp+0.082463rain

La validación global del modelo sigue siendo buena, p-value: < 2.2e-16.

La validación individual del modelo ahora está mas ajustada y todas las variables tienen un valor d Pr(>|t|)menor de 0.05 lo cual indica que son estadísticamente significativas.

También aquí en el resumen podemos miera el valor del Multiple R-squared: 0.293, que nos daría una idea de lo correlacionadas que están las variables, pero en el caso de regresión multilineal es mejor estudiar otro tipo de indicadores como el coeficiente de determinación ajustatado Adjusted R-squared: 0.2895. Este coeficiente tiene que estar por encima de 0.5 aunque en el caso de regresión multilineal hay que tener cuidado con la multicolineaidad.

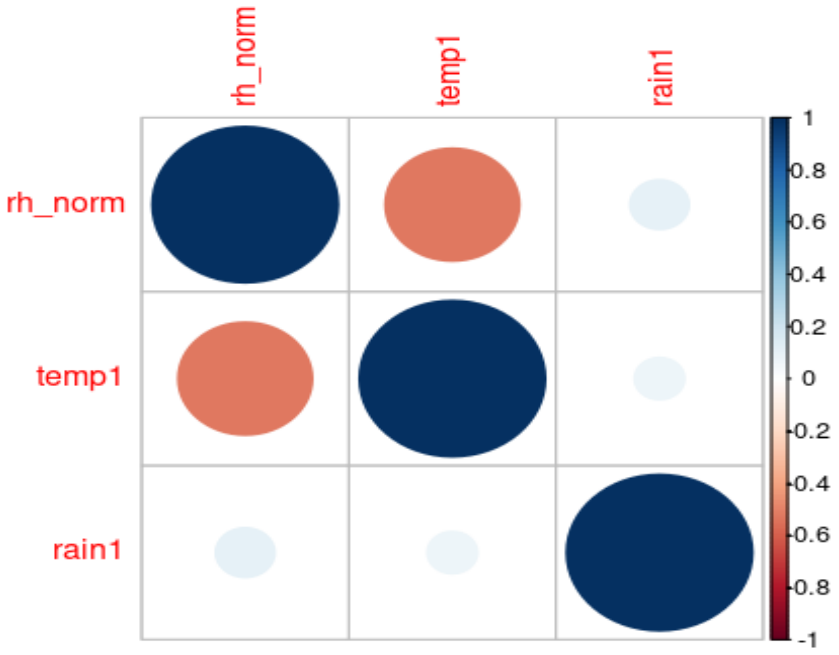
Para estudiar que el modelo cumple con el supuesto multicolinealidad estudiamos la matriz de correlaciones y el coeficiente VIF (variance inflation factor) en el siguiente punto.

3.2.2. Estudio matriz de correlaciones y VIF

La matriz de correlaciones nos sirve para mirar la correlación entre variables.

Los colores representa cómo de correlacionadas están las variables, el color azul muestra la máxima correlación, 1. Como es de esperar en la diagonal principal todos los colores son azules, ya que una la correlación entre una variable y ella misma es 1.

En nuestro modelo vemos que las variables rh_norm (humedad relativa del aire) y la temperatura tienen una correlacion de 0.52, lo que está muy bien, y no está en el rango del al multicolineidad.La correlación ente la variable temp1 y rain1es pequeña, de 0.1.



Además de la matriz de correlaciones el VIF nos da una idea de la correlación de las variables del modelo. Para nuestro modelo tenemos el siguiente valor:

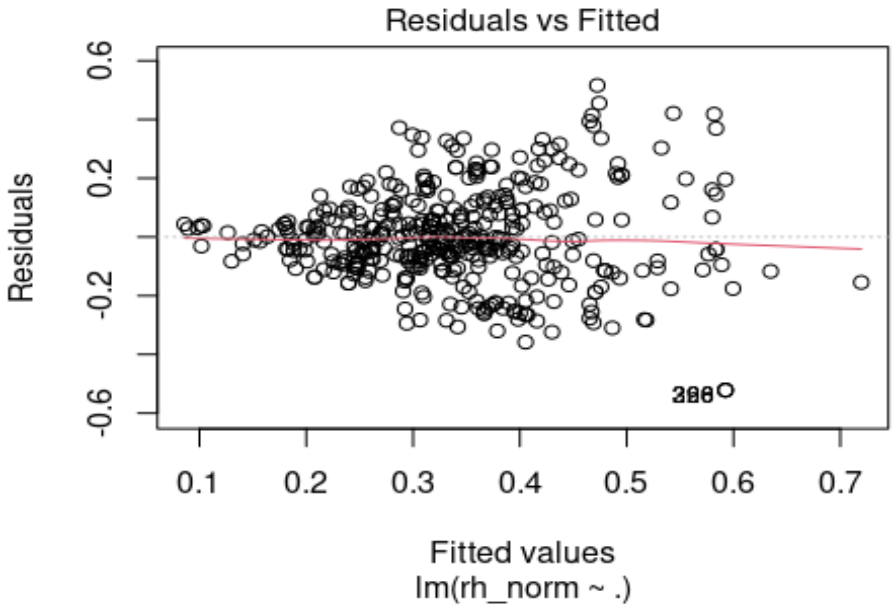
```
vif(multiAll2)

##    temp1    rain1
## 1.005503 1.005503
```

Si la el valor del VIF está entre 0 y 1 podemos decir que la correlación es moderada, en nuestro caso el valor está muy próximo a 1, lo que quiere decir hay poca correlación entre los valores

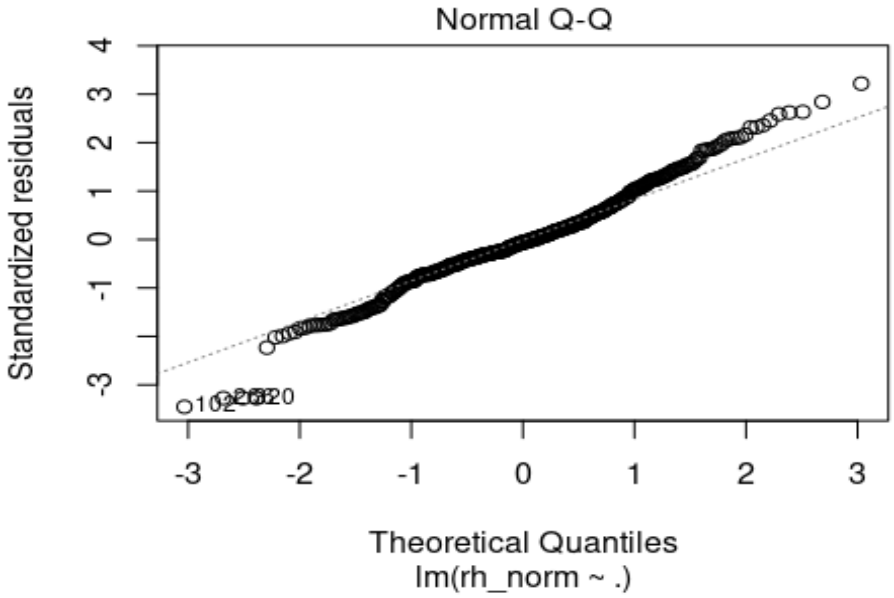
3.2.3. Residuos

Analizamos los residuos del modelo evaluando las 4 gráficas de los residuos.

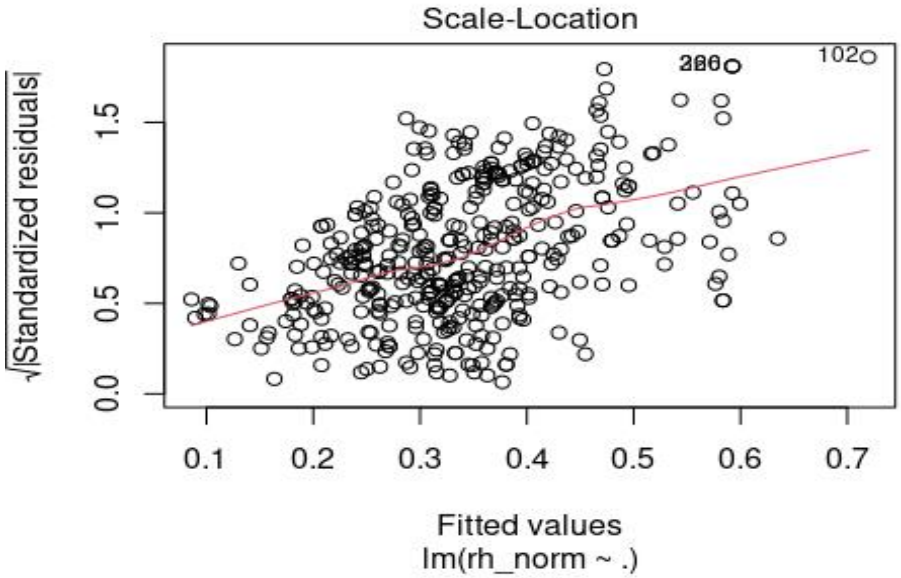
El primer gráfico en el que nos muestra la relación de multinealidad entre las variables, no es muy bueno. Los residuos no están muy bien distribuidos alrededor de la línea horizontal.



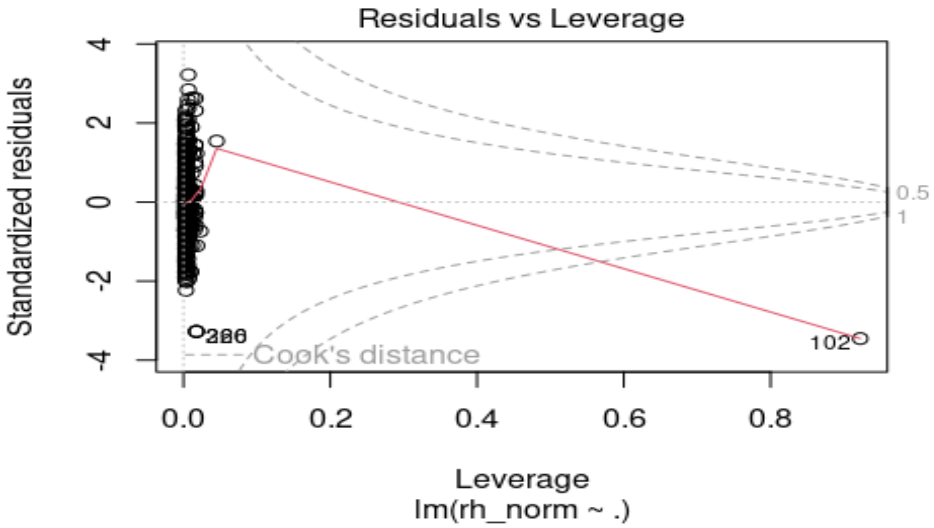
En el gráfico Q-Q, observamos que los residuos se distribuyen de forma normal hasta el valor 1.5 mas o menos.



En el tercer gráfico , que es de la varianza vemos que ésta tampoco es constante para este modelo.



En el ultimo gráfico vemos que hay varios puntos que difieren bastante del resto.



3.3. Regresión logística

Vamos a estudiar si hay alguna relación entre que la lluvia y la humedad relativa del aire. La variable rain la hemos transformado en una variable boleana si un día hay precipitación bolean_rain = 1, si no hay precipitación será 0.

Estudiamos si la probabilidad de que un día llueva en función de la humedada relativa del aire sigue un modelo logarítmico representado por la ecuación:

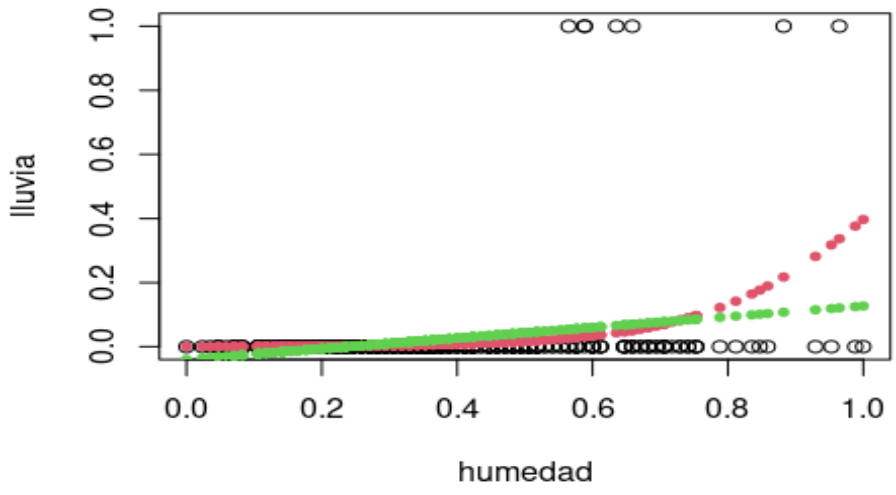
$$P(X) = \frac{e^{\beta 0+\beta 1x}}{1+e^{\beta 0+\beta 1x}}$$

3.3.1. Estudio gráfico, resumen y test *Likelihood* :

Hacemos regresión logarítmica con las varibles bolean_rain y la rh. Primero dibujamos gráficamente como es nuestro modelo y lo comparamos con el líneal y luego interpretamos el resumen para calcular la ecuación y sacar algunas conclusiones.

Hacemos regresión sobre la parte de datos de entrenamiento

```
##Si es lluvioso o no en funcion de la humedad relativa del aire
lluvia<-train$bolean_rain
humedad<-train$rh_norm
logist3 <- glm(lluvia~humedad, data = train, family = binomial)
plot(humedad, lluvia)
points(humedad, logist3$fitted.values, col=2, pch=20)
lin<-lm(lluvia~humedad, data=train)
points(humedad, lin$fitted.values, col=3, pch=20)
```



En esta grafica están representados con puntos rojos cómo sería una regression logistica delas valiables lluvia /humedad y con puntos verdes cómo sería una regression lineal e ambas variables.

Observamos que la regression logistica sí parece tener la forma de esa regression, pero nos faltarían más datos para estimar major el modelo.

```
#Vemos el resumen modelo
summary(logist3)

##
## Call:
## glm(formula = lluvia ~ humedad, family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.00558  -0.13823  -0.08992  -0.06102   2.69718
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.752      1.291  -6.004 1.92e-09 ***
## humedad        7.333      1.838   3.991 6.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.966  on 412  degrees of freedom
## Residual deviance: 52.080  on 411  degrees of freedom
## AIC: 56.08
##
## Number of Fisher Scoring iterations: 8
```

Los valores de β0 = -7.752 y β1= 7.333. Como era de esperar, la probabilidad de que llueva si la humedad es cero es muy baja, si sustituimos en la fórmula de arriba obtenemos p(lluvia) = 0.00043→0.043%

El valor de p-value de humedad es significativo 6.59e-05

El test *Likelihood ratio* calcula la significancia de la diferencia de residuos entre el modelo de interés y el modelo nulo. El estadístico sigue una distribución *chi-cuadrado* con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos.

```
anova(logist3, test = "Chisq")

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      412     70.966
humedad  1   18.886       411     52.080 1.388e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Podemos decir que el modelo es significativo a nivel global, ya que el valor de Pr(>Chi) es menor de 0.05.

3.3.2. Estudio matiz de confusión:

La matriz de confusión que obtenemos para este modelo es la siguiente:

```
predicciones <- ifelse(test = logist3$fitted.values > 0, yes = 1, no = 0)
```

```
matriz_confusion <- table(lluvia, predicciones,
                          dnn = c("observaciones", "predicciones"))
matriz_confusion
```

```
##              predicciones
## observaciones   1
##             0 406
##             1   7
```

Los valores de la diagonal principal se corresponden con los valores estimados por el modelo de forma correcta y la otra diagonal representa los casos en los que el modelo se ha equivocado

Podemos calcular la exactitud (porcentaje de predicciones correctas)de nuestro modelo a través de esta matriz.

$$exactitud = 100 * \frac{TP + TN}{N} = 100 * \frac{0 + 406}{0 + 406 + 1 + 7} = 98{,}07\%$$

Esto significa que el modelo es capad de clasificar correctamente el 98,07% de las observaciones. Estos datos los hemos obtenido con la parte del train, habría que hacer el mismo estudio con la parte del test para contrastar

4. Conclusiones

4.1. Conclusiones Regresión lineal

El modelo lineal que hemos encontrado para explicar la humedad relativa del aire sigue la siguiente ecuación:

Rh_norm = 0.673915 -0.017436Temp

Las validaciones globales e individuales del modelo nos indican que es estadísticamente significativo y podría ser bueno, los p-valores en todos los casos son menores de 0.05. Pero el valor del coeficiente de determinación es menor de 0.5 y ya empezamos a sospechar que algo no está bien en nuestro modelo, este valor tiene que estar para un modelo de regresión lineal por encima de 0.5.

Analizando los residuos vemos que no cumplen el modelo de normalidad, no siguen una distribución normal de media 0 y varianza constante.

Por lo tanto nuestro modelo no es bueno.

4.2. Conclusiones Regresión multilineal

La ecuación que hemos encontrado para el modelo que explicara la humedad relativa del aire en función de la temperatura y la lluvia es la siguiente:

Rh_norm = 0.673627 -0.017648temp+0.082463rain

Las validaciones globales e individuales son significativas, tanto el p-valor de la variable global como Pr(>|z|) para las variables individuales es menor de 0.05(Aunque hemos tenido que eliminarla variable el viento porque no era significativa para nuestro modelo).

Tampoco hemos encontrarlo multicolinealidad entre las variables, tanto los valores de la matriz de correlación como el coeficiente de determinación ajustado y el VIF son muy bajos (a lo mejor excesivamente bajos).

Pero al estudiar los residuos nos hemos encontrado con que no cumplen demasiado bien los supuestos de normalidad. Estudiando las 4 gráficas más significativas para los residuos vemos que no cumplen claramente una distribución norma.

Podemos decir que el modelo no es demasiado bueno.

4.3. Conclusiones Regresión logística

El modelo creado para predecir la probabilidad de que llueva a partir de la humedad relativa del aire parece estadísticamente significativo si miramos el test likelihood,su p-value = `1.388e-05. La validación individual también es significativa, el p-valor de la humedad es sinificativo, 6.59e-05.`

La ecuación del modelo es la siguiente:

$$\boldsymbol{P(lluvia) = \frac{e^{-7{,}752+7{,}333*humedad}}{1+e^{-7{,}752+7{,}333*humedad}}}$$

5. Limitaciones y trabajo futuro

Habría que analizar los datos un poco más. Hay algunos valores que parece que están fuera de rango y podrían haber influido en nuestros modelos, en especial en el multilineal. A lo mejor nos convendría estudiarlos de nuevo y aplicarlos agún tratamiento o relacionarlos con otras variables del dataset como la época del año.