
# PREDICCIÓN

## Máster en Data Science

**Profesor: Ricardo Queralt**

**Marta Divassón Carribero**

**Fecha: 25 octubre 2018**

# Índice

En la observación de los datos se representan diferentes fondos de entidades financieras con la rentabilidad total de cada uno de ellos en una secuencia temporal. Cuya proyección es desde 1 día hasta 10 años y datos como la volatilidad o la capitalización media bursátil que afectan a los datos en cuestión.

En el modelo de regresión la variable de fondos(rent_1) es la variable dependiente condicionada por las demás variables o variables predictivas. Bien, el modelo debe explicar las relaciones y la capacidad explicativa de las variables para obtener resultados precisos, mediante la obtención del parámetro que relaciona todos los elementos.

Para comenzar, tras definir la variable dependiente se crea una matriz con las variables predictivas que son las rentabilidades desde 1 mes hasta 10 años en diferentes escalas. Tras conocer la matriz de factores independientes, observamos como en 3,5 y 10 años hay una influencia de variables negativas o atípicas. La finalidad es conseguir una variable respuesta, es decir, un modelo de regresión, capaz de predecir rentabilidades futuras de los fondos en cuestión.

En la estimación de los coeficientes se persigue minimizar la suma de los cuadrados de los residuos. Es decir, cuando las variables que afectan sean representadas gráficamente, al ser unidas deberán estar lo mas cerca posible del eje lineal, minimizando esa distancia. Estos estimadores vienen acompañados de las varianzas, para así calcular la precisión de estimación en el estudio.

La construcción del modelo de regresión indica un modelo lineal múltiple en el que a partir de unas variables predictivas vamos a estimar el modelo, eliminando cualquier dato no disponible (NA). Es decir, la acción con NA implicará que se omitan todos los valores sin información que no serán útiles a la hora de predecir el modelo.

Ahora bien, al ejecutar los modelos de regresión, se estima la influencia que tiene cada variable sobre la variable dependiente. Es decir, la rentabilidad en 1 año por ejemplo y su relación (positiva o negativa) con la variable dependiente (Rent_1). Cuando llevamos a cabo los modelos, el valor de significación será determinante a la hora de excluir o utilizar las variables que explicarán la variable dependiente. Cuando el p valor es inferior a 0.05 indica que esas variables predictivas están relacionadas con la variable dependiente.

En el primer modelo, se contrasta con todas las variables predictivas y aquellas que tienen relación con la variable dependiente son la rentabilidad a 6 meses, a 1 año y menos relación, pero aún así un p valor inferior a 0.05 son la rentabilidad semanal, a 3 meses, el estilo de inversión y la volatilidad.

En el segundo modelo de regresión, se utilizan las variables predictivas más relacionadas con la variable independiente y dan un alto nivel de relación. Al observar el "t-value", los valores son distintos a 0, por tanto, se rechaza la hipótesis nula y de esta forma explican el valor dependiente. Las hipótesis, en el estudio de predicción, pierden importancia si el modelo predictivo es muy acertado. En este caso el "Multiple R Squared" obtendremos el porcentaje de la variabilidad observada que es capaz de explicar el modelo, en este caso es de 0.8635, es decir de 86.35%.

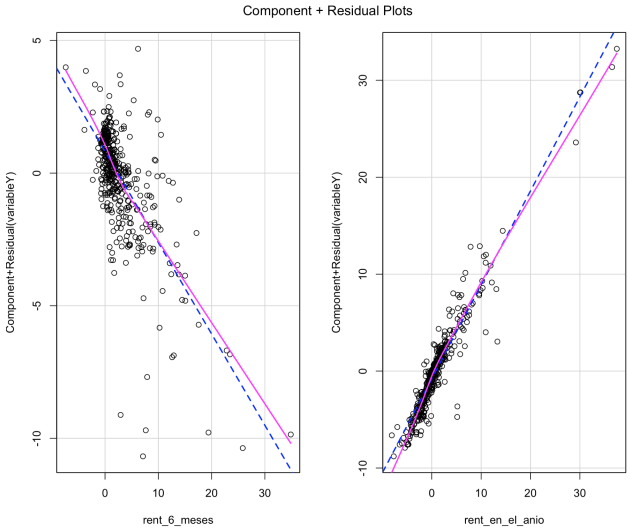
Es en el tercer modelo donde se obtiene un alto nivel de significación, pero el p valor no es tan significativo como en el segundo caso, por tanto, utilizaremos el segundo y el tercer modelo, como los más predictivos. Igualmente, como en el segundo modelo, el "t-value" es diferente a 0, por tanto se rechaza la hipótesis nula. El "Multiple R Squared" en este caso es de 0.2484, es decir, de 24.84%, menos significativo que en el segundo modelo.

Bien, los tests de anova sirven para medir la variabilidad, para comparar de igual forma que anteriormente los valores de significación. Al no utilizar las mismas variables en el modelo 2 y 3 no podemos anidar modelos, por tanto, no sería significativa.

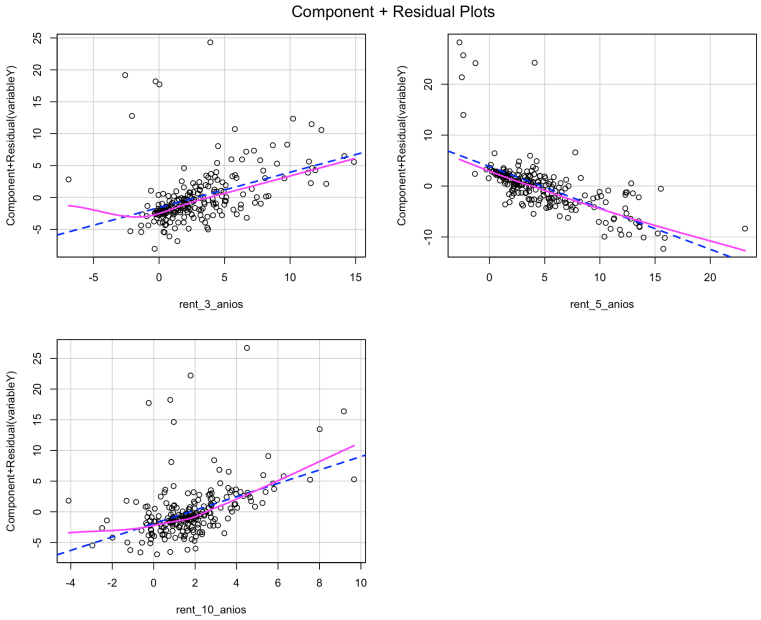
Aquellos cuyo p-valor esté por encima de 0.05 serán excluidos del modelo, aunque para la selección de variables se debe proceder al uso de modelos como **"Best Subset Selection"** o **"forward, backward e hybrid".** Es entonces uno de estos modelos de selección los que identifican las variables independientes que mejor predicen el modelo.

Posteriormente, se procede al estudio de la linealidad, ya que hay que descartar cualquier combinación de variables de indique colinealidad o multicolinealidad. Para ello, se precisa calcular el coeficiente de correlación. Si es muy alta y por tanto de alguna forma redundante, hay que ir excluyendo valores independientes para ver si afectan al modelo y empeoran tras eliminarlas y pierden relación.

En el siguiente gráfico, se estudia la correlación de la variable predictiva respecto a la independiente y se ve claramente una relación entre ambas.

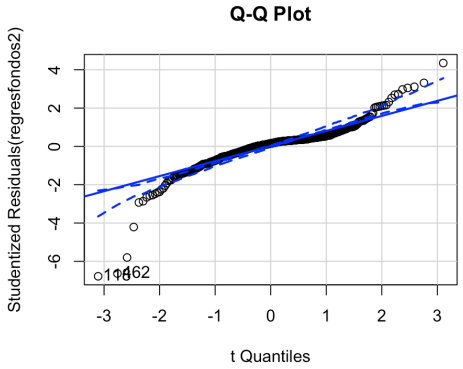



En el gráfico 2, se obtiene la correlación respecto al modelo 3 de regresión y claramente se observa linealidad.

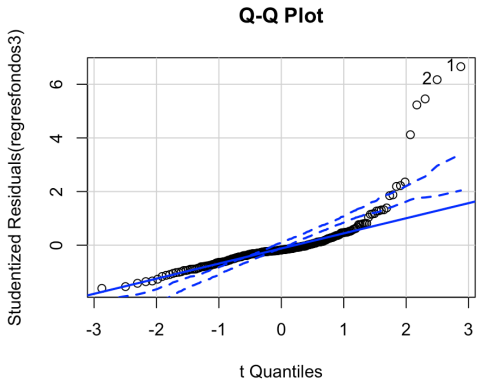



Para el estudio de la multicolinealidad, se utiliza el factor de inflación de varianza VIF. Esto se hace a partir de la librería "car" y con vif() indicando el modelo de regresión que se ha escogido. Esto nos puede dar una pista ya que si el valor está por encima de 10 es dispensable ya que es un valor problemático. En el modelo 2, los valores obtenidos son cercanos a 1 y en el modelo 2 no hay valores por encima o iguales a 10. Por tanto, se concluye que pueden ser utilizadas como variables predictivas.

En cuanto a la representación gráfica de los QQplots, se comparan las distribuciones para ver si coinciden. En ambos casos los cuantiles se distribuyen igual y los puntos forman aproximadamente una línea recta o una grafica normal bimodal.




Modelo 2.




Modelo 3.

Ahora bien, como he mencionado anteriormente, en la selección de valores predictivos se utiliza el método paso a paso o "**stepwise**", utilizado para conocer en qué orden se introducen dichas variables. El primero, es el método ***forward,*** esté consiste en ir introduciendo variable a variable de entre las seleccionadas para ver si el modelo va mejorando, hasta el punto en el que ninguna de las variables restantes pueda mejorar el modelo más. En la demostración, en el primer modelo, se introduce primero la rentabilidad a 6 meses y tras este la rentabilidad a 1 año. En el segundo modelo, se incliye primero la de 3 años, la de 5 años y finalmente la de 10 años.

En el caso del método ***backward***, se introducen todas las variables y se van excluyendo unas a unas hasta que el modelo se optimice. Se van eliminando aquellas variables que no aporten nada según lo observado en las "t values" y en la "F statistics".

En el caso del ***modelo híbrido***, se combinan los dos procedimientos, introduciendo variables con el menor "F statistics". Aquellas que ya están en el modelo pueden ser eliminadas. Se finaliza cuando ya no haya variables que se eliminen o que se introduzcan.

En ambos, obtenemos un dato de AIC. En el modelo 2, el AIC es de 344.73, mientras que en el modelo 3 el AIC es de 623.18. Por tanto, nos decantamos por el modelo 2, ya que tiene un AIC más pequeño.

## Conclusión

Tras el estudio en la selección de variables para determinar la relación con la rentabilidad de fondos, se obtiene que la rentabilidad a 6 meses y la rentabilidad a 1 año son las variables seleccionadas para escoger un modelo de predicción. Aunque la hipótesis nula no siempre es determinante a la hora de predecir, si el modelo final tiene una estructura sólida y las hipótesis fallan, entonces nos fiaríamos por tanto en el modelo de predicción.

## Apéndice

- **Trabajo en Rstudio:**

```
#####----PREDICCIÓN----#######
##############################
###La regresion: tecnica de prediccion de una variable dependiente con una o más predictivas
##Se utiliza para:
#a.Identificar las variables explicativas.
#b.Describir la relación existente entre las variables.
#c.Predecir la variable dependiente a partir de las variables independientes.
library(tidyverse)
library(stats) ##estadisticos
library(dplyr) ##manipulacion de datos
library(car) ##modelo de regresion aplicada
library(leaps) ##regresion selecc
library(rminer)
library(MASS)
library(fBasics)

##a partir de las ecuaciones del modelo(y=Xβ+u) se puede representar la matriz.
Fondos=read.csv("~/Desktop/DATASCIENCE/PRACTICASR/Fondos.csv", sep=';',dec=',')
Fondos
##se leen los datos desde el archivo csv
str(Fondos)
View(Fondos)
##hacemos un head and tail para comparar rentabilidadesa simple observación
head(Fondos)
tail(Fondos)

##Antes de crear la matriz de variables independientes, filtro los datos para eliminar variables categóricas
Fondos2 <- Fondos[,-c(1,2,5,6,18)] ##de esta forma eliminamos variab.categoricas
View(Fondos2)

variableY=Fondos$rent_1 ##variable dependiente

matrizX=cbind(1,Fondos[,10:16]) ##la matriz de variables ind: desde rent_1 mes hasta rent_10 años

head(variableY)##vector Y
head(matrizX) #matriz X
```

```
#####MODELOS DE REGRESIÓN######
##Primer modelo de prueba
regresfondos=lm(variableY~., data =Fondos2)
summary(regresfondos)

##probamos con otro modelo
regresfondos2=lm(variableY~rent_6_meses+rent_en_el_anio, data =
Fondos2)
summary(regresfondos2)

##probamos el modelo con las variables más dispares
regresfondos3=lm(variableY~rent_3_anios+rent_5_anios+rent_10_anios,
data=Fondos2, na.action = 'na.exclude')
summary(regresfondos3)

##Se testea la linealidad del modelo 2 y 3
crPlots(regresfondos2)
crPlots(regresfondos3)

##estudio de VIF
vif(regresfondos2)
vif(regresfondos3)

##Seleccion de variables
##FORWARD MODELO 1
library(MASS)
regfit.fwd=regsubsets(variableY~Fondos2$rent_6_meses+Fondos2$rent_e
n_el_anio,Fondos2,method ="forward")
summary (regfit.fwd)
##FORWARD MODELO 2
regfit.fwd=regsubsets(variableY~Fondos$rent_3_anios+Fondos2$rent_5_a
nios+Fondos2$rent_10_anios,Fondos2,method ="forward")
summary (regfit.fwd )

##BARCKWARD
library(MASS)
##BACKWARD MODELO 1
stepAIC(regresfondos2, direction="backward")
regfit.bwd=regsubsets(variableY~Fondos2$rent_6_meses+Fondos2$rent_
en_el_anio,Fondos2,method ="backward")
summary (regfit.bwd )
##BACKWARD MODELO 1
regfit.bwd=regsubsets(variableY~Fondos$rent_3_anios+Fondos2$rent_5_
anios+Fondos2$rent_10_anios,Fondos2,method ="backward")
summary (regfit.bwd )
```

```
##HYBRID
stepAIC(regresfondos2, direction="both")
stepAIC(regresfondos3, direction="both")

##se comparan 2 distribuciones para ver si coinciden; gráfico de puntos
de distribucion normal
qqPlot(regresfondos2, labels=row.names(Fondos2), id.method="identify",
      simulate=TRUE, main="Q-Q Plot")
qqPlot(regresfondos3,labels=row.names(Fondos2), id.method="identify",
      simulate=TRUE, main="Q-Q Plot")


##se lleva a cabo un test AIC y BIC
AIC(regresfondos2,regresfondos3) ##AIC Del modelo 3 es inferior
BIC(regresfondos2,regresfondos3) ##BIC del modelo 3 es inferior


##summary de los componentes del modelo
reg.summary$rss ##queremos coger el menor rss
reg.summary$cp ##se selecciona el modelo con el menor cp
reg.summary$aic ##estadísticos que nos dicen cuanta información
tienen los errores
reg.summary$bic ##se selecciona el modelo con el menor BIC
```