# Predicción-Fondos.csv

## Analisis de regresión multiple

Se tiene una muestra de 500 observaciones de un total de 29 variables. El análisis consiste en estimar un modelo de regresión adecuado para la variable rent_1:

- Datos: Fondos.csv
- Variable dependiente: rent_1
- Variables explicativas: resto de variables

En general, un modelo de regresión lineal múltiple viene dado por la siguiente expresión(k el número de variables independientes y n la dimensión de los vectores)

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e$$

En donde:

- Y es la variable dependiente y *X1, X2,…,Xk* variables independientes.
- Los coeficientes a, *b1,…, bk* son parámetros desconocidos a estimar y se interpretan como la variación de la variable dependiente debida a la variable explicativa considerando las demás constantes.
- *e* es el error del modelo, diferencia entre los valores observados y los pronosticados.

Las hipótesis en las que se basa un modelo de regresión lineal múltiple son:

1. Normalidad de los residuos (sigue una distribución $N(0, \sigma^2)$)
2. Linealidad: la variable dependiente está generada por el modelo lineal expresado
3. Homocedasticidad ($Var(e) = \sigma^2$)

4. Independencia entre los errores (e)

La metodología para abordar este analisis de regresión multiple ha sido:

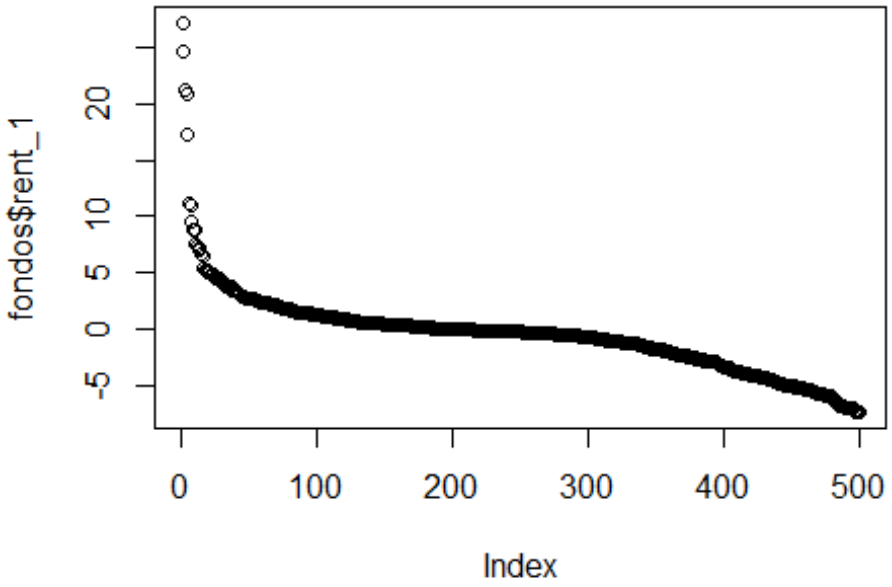
i) Selección de la muestra de entrenamiento y la de validación

ii) Selección de las variables para el modelo de regresión

iii) Analisis del modelo de regresión multiple

iv) Validación del modelo

```
# Instalo librerias


    #install.packages("fBasics")
    library(fBasics)

    #install.packages("akima")
    library(akima)

    #install.packages("car")
    library(car)

    #install.packages("gvlma")
    library(gvlma)
    #install.packages("rminer")
    library(rminer)

    #install.packages("ISLR")
    library(ISLR)

   #install.packages("lmtest")
   library(lmtest)


#cargo los datos

setwd("C:/Users/usuario/Desktop/prediccion/MDSF_Prediccion-
master/Clase01/Datos")
fondos<- read.csv("Fondos.csv", header=TRUE, sep = ";", dec=",")


#dibujo de la variable dependiente:
plot(fondos$rent_1)
```




## Selección de las variables

- Sustituyo los NA por el método del vecino más cercano.
- Calculo la matriz de correlaciones para tener una idea de cuales variables están mas correlacionadas con la variable explicativa y la correlación entre ellas.
- Hago una regresión con las variables que a priori pienso que podrían influir en el modelo de las variables para tener una idea de la colinealidad existente entre variables y poder seleccionar un modelo en el que no exista colinealidad.

```
#elimino las variables que no sean numéricas para poder hacer la matriz
de correlaciones
fondos<- fondos[,-2] #columna donde se encuentra el nombre::2
fondos<- fondos[,-4] #columna donde se encuentra el ISN::4
fondos<- fondos[,-4] #columna donde se encuentra el Gestora en el nuevo
dataset fondos::4
#sustituyo los na por el metodo del vecino mas proximo
fondos<-imputation("hotdeck", fondos)
```

```
a<-cor(fondos)
View(a)
```

Se puede observar en la matriz de correlaciones que rent_en_el_anio es la variable más correlacionada con rent_1, además de las otras rentabilidades que también tienen una correlación significativa pero en menor proporción. Sin embargo, entre ellas están correlacionadas por lo que es consecuente que exista colinealidad entre ellas.

## Selección de la muestra de estimación y la de validación

```
set.seed(250)
numData=nrow(fondos)
train=sample(numData ,numData/2)

fondosTrain<-fondos[train,]
View(fondosTrain)
```

## Selección de las variables a estimar

Tomo las variables que he considerado que pueden influir en rent_1 y estimo un modelo de regresión para ver la existencia de correlación entre variables:

```
regres01=lm(rent_1~0+
rent_en_el_anio+rent_6_meses+Capitaliz_media_bursatil+Patrimonio+Volatili
dad_3+Sharpe_.3+Ratio_de_informacion+Media_3+Com_Gestion+Media_3+Estilo_i
nversion_.RF, data=fondosTrain)


summary(regres01)

##
## Call:
## lm(formula = rent_1 ~ 0 + rent_en_el_anio + rent_6_meses +
Capitaliz_media_bursatil +
##     Patrimonio + Volatilidad_3 + Sharpe_.3 + Ratio_de_informacion +
##     Media_3 + Com_Gestion + Media_3 + Estilo_inversion_.RF, data =
fondosTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6277 -0.3289 -0.0067  0.3325  3.9738
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## rent_en_el_anio             9.833e-01  2.299e-02  42.764  < 2e-16 ***
## rent_6_meses               -2.438e-01  2.928e-02  -8.327 6.34e-15 ***
## Capitaliz_media_bursatil -8.121e-06  3.862e-06  -2.102   0.0366 *
## Patrimonio                  7.645e-05  1.185e-04   0.645   0.5194
## Volatilidad_3              -2.167e-01  2.680e-02  -8.086 3.05e-14 ***
## Sharpe_.3                  -1.692e-01  1.054e-01  -1.606   0.1096
## Ratio_de_informacion        1.113e-01  6.266e-02   1.776   0.0770 .
## Media_3                     3.709e+00  3.713e-01   9.990  < 2e-16 ***
## Com_Gestion                -2.033e-01  1.544e-01  -1.317   0.1892
## Estilo_inversion_.RF        2.240e-02  2.642e-02   0.848   0.3974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9659 on 240 degrees of freedom
## Multiple R-squared:  0.9293, Adjusted R-squared:  0.9263
## F-statistic: 315.3 on 10 and 240 DF,  p-value: < 2.2e-16
```

Como se ha visto en la matriz de correlaciones, la rentabilidad en el año es significativa en el modelo, al igual que la rentabilidad en 6 meses. Además, se puede observar que la volatilidad y la Media_3 también son significativas en el modelo regres01.

Para elegir las variables explicativas hay que analizarla existencia de correlación entre ellas y su significatividad en el modelo. Es conveniente que no exista multicolinealidad entre ellas, en caso contrario, afectará a la predicción de la variable dependiente.

Veamos el diagnóstico de colinealidad:

```
fiv<-vif(regres01)

fiv
```

```
##          rent_en_el_anio             rent_6_meses
Capitaliz_media_bursatil
##                 2.239550                5.221833
2.764575
##               Patrimonio            Volatilidad_3
Sharpe_.3
##                 1.174577                9.994144
4.650398
##     Ratio_de_informacion                 Media_3
Com_Gestion
```

```
##                 2.498061                  4.433956
9.885473
##     Estilo_inversion_.RF
##                 9.463090

sqrt(fiv)>2

##          rent_en_el_anio             rent_6_meses
Capitaliz_media_bursatil
##                    FALSE                     TRUE
FALSE
##               Patrimonio           Volatilidad_3
Sharpe_.3
##                    FALSE                     TRUE
TRUE
##     Ratio_de_informacion                 Media_3
Com_Gestion
##                    FALSE                     TRUE
TRUE
##     Estilo_inversion_.RF
##                     TRUE
```

Cuanto mayor es el FIV de una variable, mayor es la varianza del correspondiente coeficiente de regresión.

Existen índices de condición con su raíz mayor que dos, esto indica presencia de colinealidad entre las variables. Ante este problema de colinealidad, existen diversas alternativas:

Eliminar variables colineales paso a paso y realizar el proceso hasta que no exista ninguna colinealidad y las variables sean significativas en el modelo. En este proceso se pueden utilizar diversos algoritmos como el de pasos hacia delante o hacia atrás.

A continuación planteo un modelo con las variables más significativas de la regresión regres01:

```
regres03=lm(rent_1~0+rent_en_el_anio+Volatilidad_3+Media_3,
data=fondosTrain)
summary(regres03)

##
## Call:
## lm(formula = rent_1 ~ 0 + rent_en_el_anio + Volatilidad_3 + Media_3,
##     data = fondosTrain)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2704 -0.4121 -0.0166  0.2393  5.2579
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## rent_en_el_anio  0.87580    0.01867  46.909  < 2e-16 ***
## Volatilidad_3   -0.33729    0.01730 -19.498  < 2e-16 ***
## Media_3          2.79800    0.35457   7.891 9.72e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.108 on 247 degrees of freedom
## Multiple R-squared:  0.9043, Adjusted R-squared:  0.9031
## F-statistic: 778.1 on 3 and 247 DF,  p-value: < 2.2e-16
```

```
gvmodel <- gvlma(regres03)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = rent_1 ~ 0 + rent_en_el_anio + Volatilidad_3 + Media_3,
##     data = fondosTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2704 -0.4121 -0.0166  0.2393  5.2579
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## rent_en_el_anio  0.87580    0.01867  46.909  < 2e-16 ***
## Volatilidad_3   -0.33729    0.01730 -19.498  < 2e-16 ***
## Media_3          2.79800    0.35457   7.891 9.72e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.108 on 247 degrees of freedom
## Multiple R-squared:  0.9043, Adjusted R-squared:  0.9035
## F-statistic: 778.1 on 3 and 247 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = regres03)
##
##
```

```
          Value p-value                    Decision
## Global Stat        489.52325 0.00000 Assumptions NOT satisfied!
## Skewness             3.45487 0.06307    Assumptions acceptable.
## Kurtosis           485.09267 0.00000 Assumptions NOT satisfied!
## Link Function        0.00326 0.95447    Assumptions acceptable.
## Heteroscedasticity   0.97245 0.32407    Assumptions acceptable.
```

```
fiv<-vif(regres03)
```

```
## Warning in vif.default(regres03): No intercept: vifs may not be
sensible.
```

```
fiv
```

```
## rent_en_el_anio   Volatilidad_3         Media_3
##        1.123025        3.167248        3.075025
```

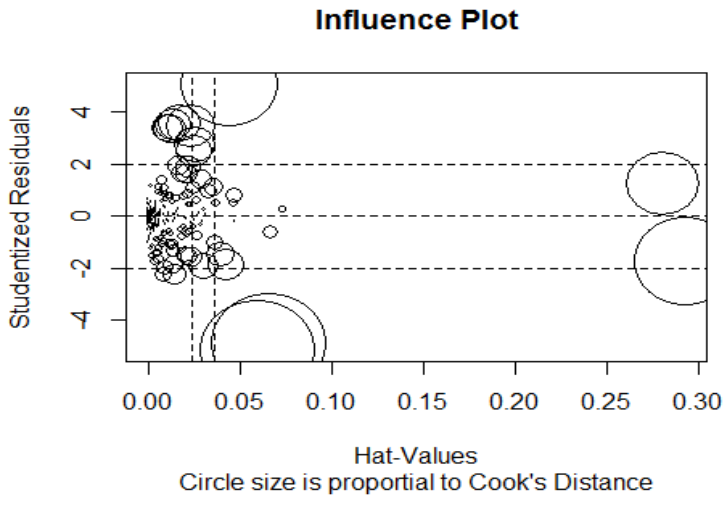
```
sqrt(fiv)>2 #son todos menores a 2
```

```
## rent_en_el_anio   Volatilidad_3         Media_3
##           FALSE           FALSE           FALSE
```

```
#outlier y valores influyentes
outlierTest(regres03)
```

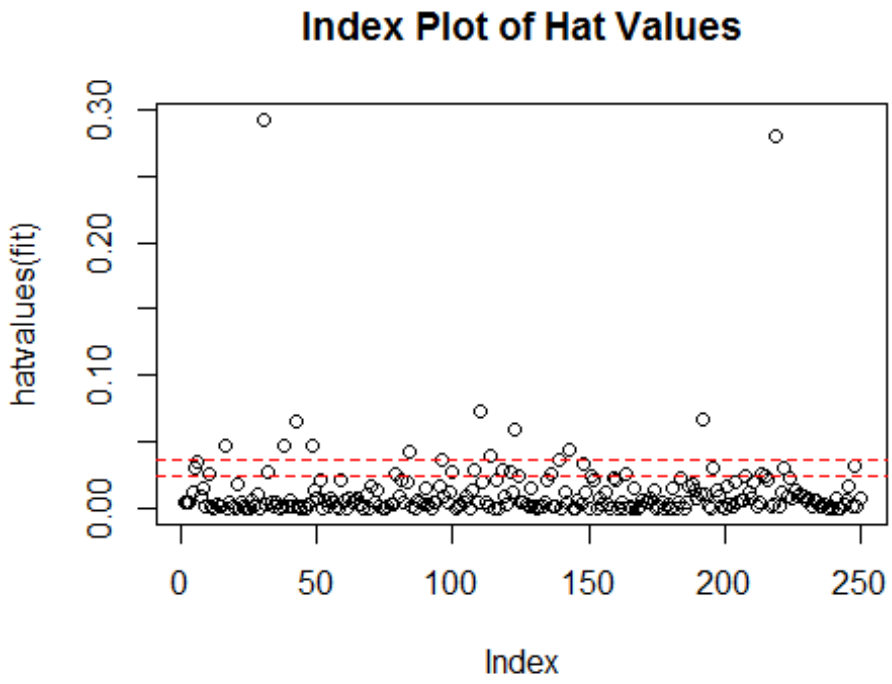
```
##      rstudent unadjusted p-value Bonferonni p
## 476 -5.154765         5.2184e-07   0.00013046
## 6    5.096132         6.9123e-07   0.00017281
## 48  -4.873102         1.9700e-06   0.00049251
```

```
# Influence Plot
influencePlot(regres03, id.method="identify", main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )
```




```
#  Identifying high leverage points
hat.plot <- function(fit) {
```

```
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main="Index Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red", lty=2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(regres03)
```




Coeficientes del modelo de regresion:

```
coefficients(regres03)

## rent_en_el_anio   Volatilidad_3         Media_3
##       0.8758003      -0.3372941       2.7979985

AIC(regres03)

## [1] 765.5069
```

La ecuación de regresión será:

***Y = 0.8758003*rent_en_el_anio-0.3372941*Volatilidad_3+2.7979985*Media_3***

Finalmente, el modelo regres03 consta de tres variables explicativas:

- rent_en_el_anio
- volatilidad_3
- media_3

Su coeficiente de determinación $R^2$ es de 0.9043 con una $R^2$ corregida de 0,9035 y no difieren demasiado, por lo que el modelo regres03 explica aproximadamente un 90% de la variable rent_1. El estadístico F indica que el modelo de regresión es estadísticamente significativo al rechazarse la hipótesis nula.

También se ha mostrado el coeficiente AIC y $C_p$ de Mallows. Cuando existen varios modelos planteados y se quiere elegir el mejor, se miran estos coeficientes.

El criterio de información AIC se basa en el error cuadrático medio residual con una penalización que crece con el aumento del número de coeficientes del modelo, resultando que el mejor modelo es el que minimiza el criterio de información. El Cp de Mallows es una medida del sesgo en el modelo, basado en la comparación entre el cuadrado medio del error total y la varianza del error verdadero. Se deben buscar modelos con valores de Cp cercanos a p (número de variables independientes del modelo).

## Validación del modelo

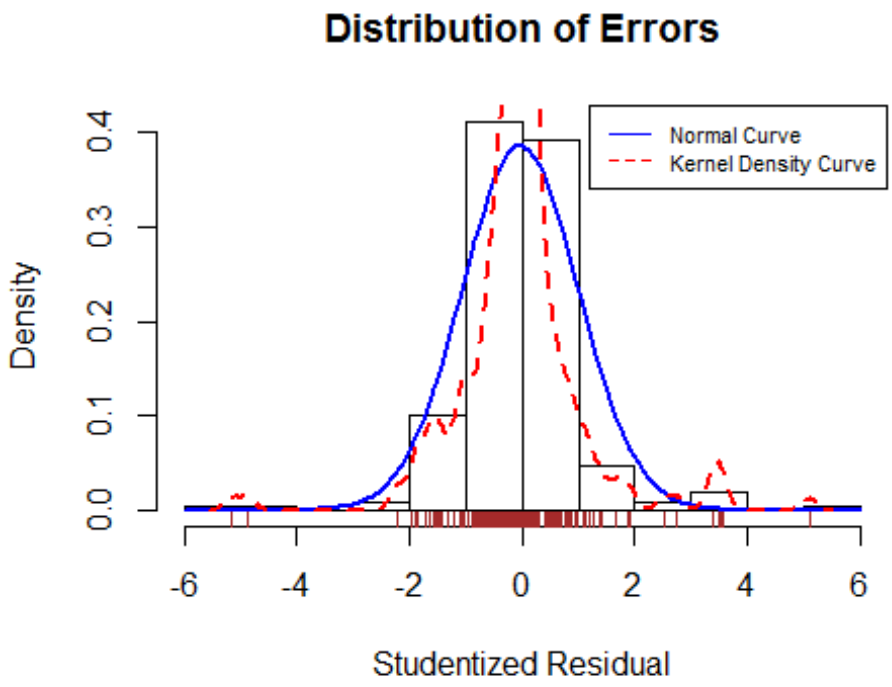
### Verificación de las hipótesis bajo las que se estima el modelo

```
vResid<-resid(regres03)
kurtosis(vResid)
```

```
## [1] 6.804909
## attr(,"method")
## [1] "excess"
```

```
residplot <- function(fit, nbreaks=10) {
  z <- rstudent(fit)
  hist(z, breaks=nbreaks, freq=FALSE,
       xlab="Studentized Residual",
       main="Distribution of Errors")
  rug(jitter(z), col="brown")
  curve(dnorm(x, mean=mean(z), sd=sd(z)),
        add=TRUE, col="blue", lwd=2)
```

```
  lines(density(z)$x, density(z)$y,
        col="red", lwd=2, lty=2)
  legend("topright",
         legend = c( "Normal Curve", "Kernel Density Curve"),
         lty=1:2, col=c("blue","red"), cex=.7)
}

residplot(regres03)
```

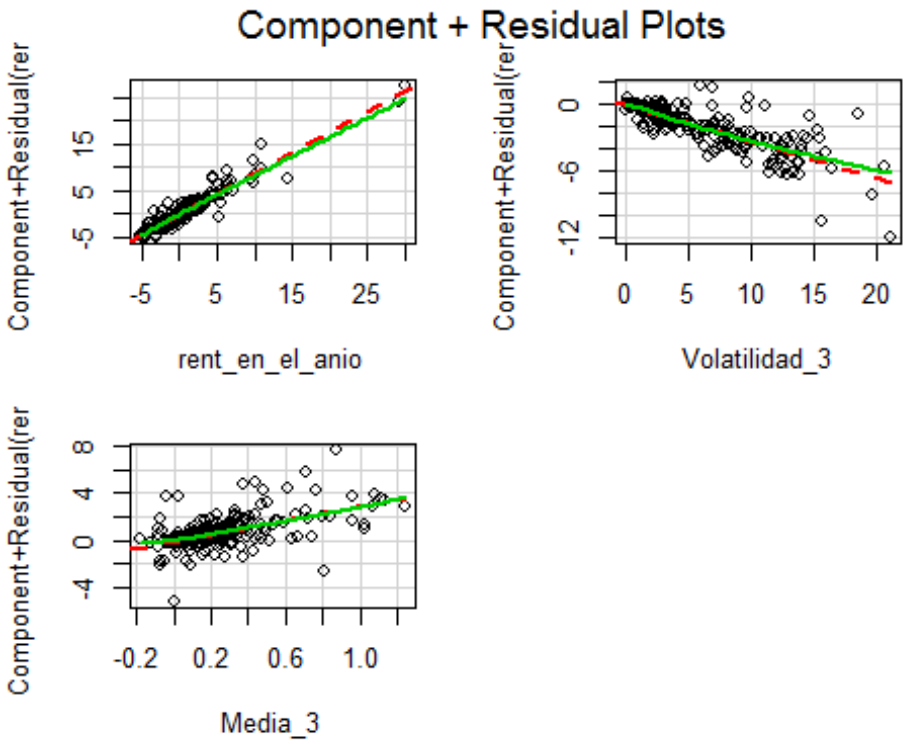



```
#Normalidad
jbTest(vResid)

## Title:
##  Jarque - Bera Normality Test
##
## Test Results:
##   PARAMETER:
##     Sample Size: 250
##   STATISTIC:
##     LM: 499.65
##     ALM: 533.871
##   P VALUE:
##     Asymptotic: < 2.2e-16
##
```

```
## Description:
##  Mon Oct 30 08:33:52 2017 by user: usuario
```

Se muestra la existencia de un exceso de kurtosis teniendo un valor de 6 y en el histograma se aprecia que la distribución de los residuos tiene una estructura más leptocúrtica, pudiendo seguir otras distribuciones como la t-student con éstas características.

```
#linealidad
crPlots(regres03)
```




```
#incorrelacion: no se puede rechazar la hipótesis nula


dwtest(regres03)

##
##  Durbin-Watson test
##
## data:  regres03
```

```
## DW = 1.9227, p-value = 0.2804
## alternative hypothesis: true autocorrelation is greater than 0
```

El test de Durbin Watson es cercano a 2, por lo que no se puede rechazar la hipótesis nula de incorrelación de los residuos.

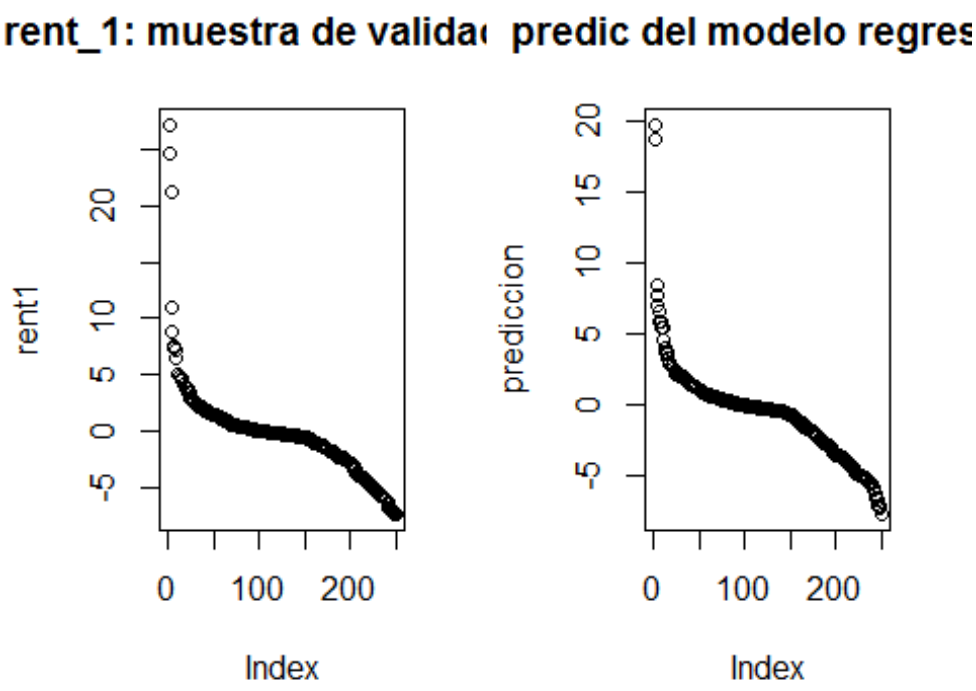
Como se ha visto en la validación global del modelo, no existe heterocedasticidad, no rechazando la hipótesis de homocedasticidad del modelo.

## Contraste con la muestra de validación

```
#Validacion y prediccion


fondosValidacion<-fondos[-train,]
rent1<-fondosValidacion[,1]
prediccion<-predict(regres03)
prediccion<-sort(prediccion, decreasing = TRUE) #ordeno los datos de la
prediccion para que queden mejor en el dibujo
a<-length(prediccion);a
```

```
## [1] 250
```

```
par(mfrow = c(1,2))#dibujo conjunto
plot(rent1, main="rent_1: muestra de validacion")
plot(prediccion, main="predic del modelo regres03")
```




```
mean(rent1-prediccion)
```

```
## [1] 0.3015458
```

```
#mean((fondos$rent_1-predict(regres03 ,Auto))[-train ]^2)
```

La muestra de validación está bien predicha por la muestra de entrenamiento. La media de la diferencia de las observaciones es 0.30, valor que está considerablemente bien teniendo en cuenta que el intervalo de la variable rent_1 es [-8,27] aproximadamente. Además, el gráfico nos muestra que el valor predicho por el modelo frente al real son parecidos.

Por lo que se concluye que es un buen modelo para estimar la variable dependiente rent_1.

Éste modelo se ha estimado cambiando los NA por el método del vecino más cercano.

Si sustituimos el valor por la media y estimamos el mismo modelo:

```r
Media<-mean(na.omit(fondos$Media_3))
for (i in 1:length(fondos$Media_3)) {
  if (is.na(fondos$Media_3[i])==TRUE){
    fondos$Media_3[i]=Media

  }
}
mediavolatilidad<- mean(na.omit(fondos$Volatilidad_3))
for (i in 1:length(fondos$Volatilidad_3)) {
  if (is.na(fondos$Volatilidad_3[i])==TRUE){
    fondos$Volatilidad_3[i]=mediavolatilidad

  }
}
regres03=lm(rent_1~0+rent_en_el_anio+Volatilidad_3+Media_3,
data=fondosTrain)
summary(regres03)
```

```
##
## Call:
## lm(formula = rent_1 ~ 0 + rent_en_el_anio + Volatilidad_3 + Media_3,
##     data = fondosTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9746 -0.3517 -0.0197  0.2240  5.5314
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## rent_en_el_anio  0.86464    0.01897  45.582  < 2e-16 ***
## Volatilidad_3   -0.35256    0.01844 -19.116  < 2e-16 ***
## Media_3          2.94788    0.37124   7.941 1.45e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 199 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.916,  Adjusted R-squared:  0.9147
## F-statistic: 723.1 on 3 and 199 DF,  p-value: < 2.2e-16
```

```r
gvmodel <- gvlma(regres03)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = rent_1 ~ 0 + rent_en_el_anio + Volatilidad_3 + Media_3,
##     data = fondosTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.9746 -0.3517 -0.0197  0.2240  5.5314
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## rent_en_el_anio  0.86464    0.01897  45.582  < 2e-16 ***
## Volatilidad_3   -0.35256    0.01844 -19.116  < 2e-16 ***
## Media_3          2.94788    0.37124   7.941 1.45e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 199 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.9160,  Adjusted R-squared:  0.9149
## F-statistic: 723.1 on 3 and 199 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = regres03)
##
##                      Value p-value                   Decision
## Global Stat        604.443 0.00000 Assumptions NOT satisfied!
## Skewness             5.473 0.01932 Assumptions NOT satisfied!
## Kurtosis           596.374 0.00000 Assumptions NOT satisfied!
## Link Function        0.516 0.47256    Assumptions acceptable.
## Heteroscedasticity   2.081 0.14915    Assumptions acceptable.
```

```
fiv<-vif(regres03)
```

```
## Warning in vif.default(regres03): No intercept: vifs may not be
sensible.
```

```
fiv
```

```
## rent_en_el_anio   Volatilidad_3         Media_3
##        1.131197        3.436590        3.272957
```

```
sqrt(fiv)>2 #son todos menores a 2
```

```
## rent_en_el_anio   Volatilidad_3         Media_3
##           FALSE           FALSE           FALSE
```

Los resultados son muy parecidos. Este modelo explica un 91% frente al otro que explica un 90%. Además éste modelo tiene una media del error de los residuos estandarizados menor.

Veamos el coeficiente AIC

```r
coefficients(regres03)
```

```
## rent_en_el_anio   Volatilidad_3         Media_3
##       0.8646434      -0.3525600       2.9478830
```

```r
AIC(regres03)
```

```
## [1] 607.3553
```

El coeficiente AIC es menor que el del modelo sustituyendo los NA por el método del vecino más cercano (valor de `765.5069`).

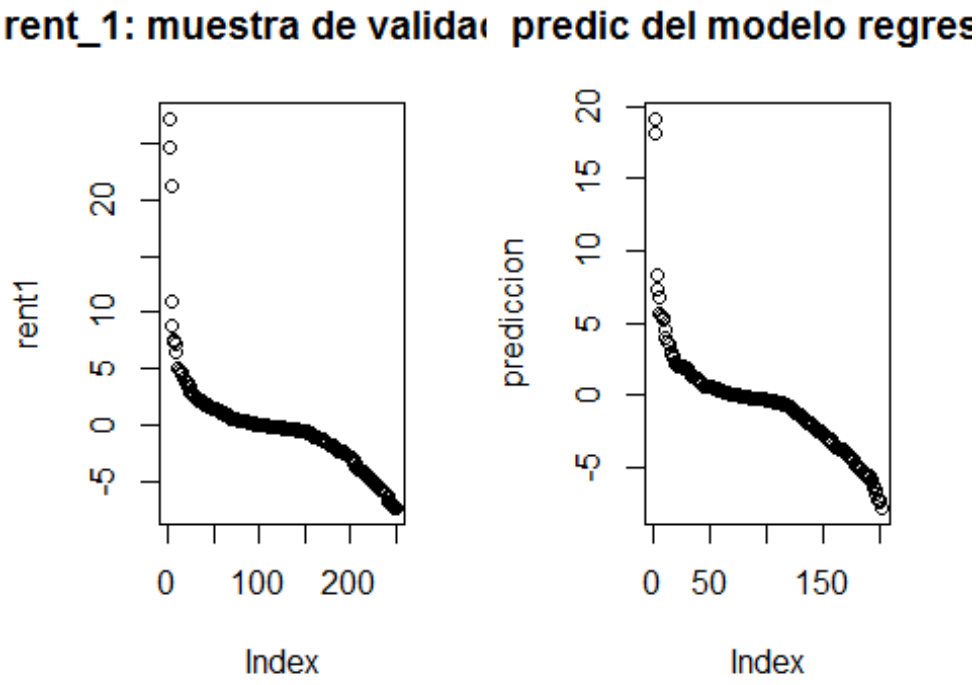
En general la verificación de las hipótesis bajo las que se estima el modelo tienen las mismas características que las del otro modelo.

Veamos si predice bien la muestra de validación:

```r
rent1<-fondosValidacion[,1]
prediccion<-predict(regres03)
prediccion<-sort(prediccion, decreasing = TRUE) #ordeno los datos de la prediccion para que queden mejor en el dibujo
a<-length(prediccion);a
```

```
## [1] 202
```

```r
par(mfrow = c(1,2))#dibujo conjunto
plot(rent1, main="rent_1: muestra de validacion")
plot(prediccion, main="predic del modelo regres03")
```




```
mean(rent1-prediccion)
```

```
## [1] -0.3936082
```

```
#mean((fondos$rent_1-predict(regres03 ,Auto))[-train ]^2)
```

Se observa que el valor predicho y el error es muy parecido también al del otro modelo.

Por tanto, se concluye que la estimación del modelo sustituyendo los NA por el método del vecino más cercano o sustituyéndolo por la media de los valores, en este caso, se obtienen resultados muy parecidos.

Sin embargo, si nos fijamos en el coeficiente AIC del modelo, se elegiría la sustitución de los NA por su valor medio, ya que es menor.