PREDICCIÓN

Informe sobre el *dataset* 'Fondos'

Álvaro Ferro Pérez

Máster en Data Science para Finanzas

Introducción

El objetivo es conseguir proponer un modelo de regresión para una variable dependiente sobre un conjunto de datos que contiene variables relacionadas con la renta, la inversión y el retorno esperado. Este conjunto de datos es de sección cruzada, usando a diferentes individuos para su elaboración.

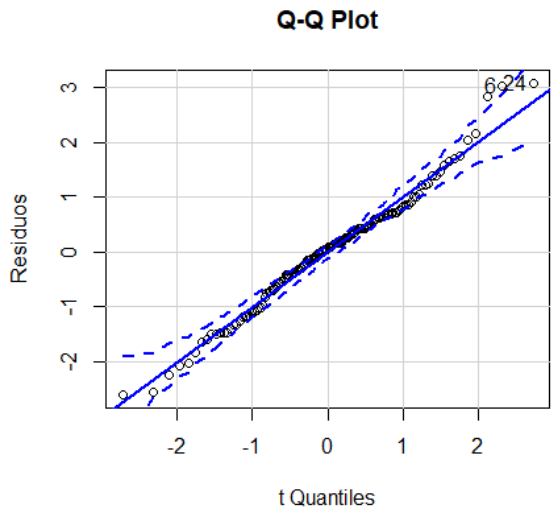
Para ello se deberán desechar de primera mano aquellas columnas de este conjunto que no nos aporten información (columnas cuyo número de NA sea alto) o columnas de tipo texto cuya información no es numérica.

En primer lugar, debemos identificar las variables explicativas y describir su relación con las demás, además de distinguir la variable dependiente (sobre la cual se podrá realizar una predicción en base a aquellas que son independientes).

En base a los métodos se selección se aplicarán técnicas (Best Subset y Stepwise) para tratar de determinar aquellas variables que aportan o no información sobre la regresión completa. Este tipo de algoritmos iteran de manera continua sobre la regresión, de atrás hacia adelante, al contrario, o usando ambas (Mixto). Este último probablemente sea el más fiable.

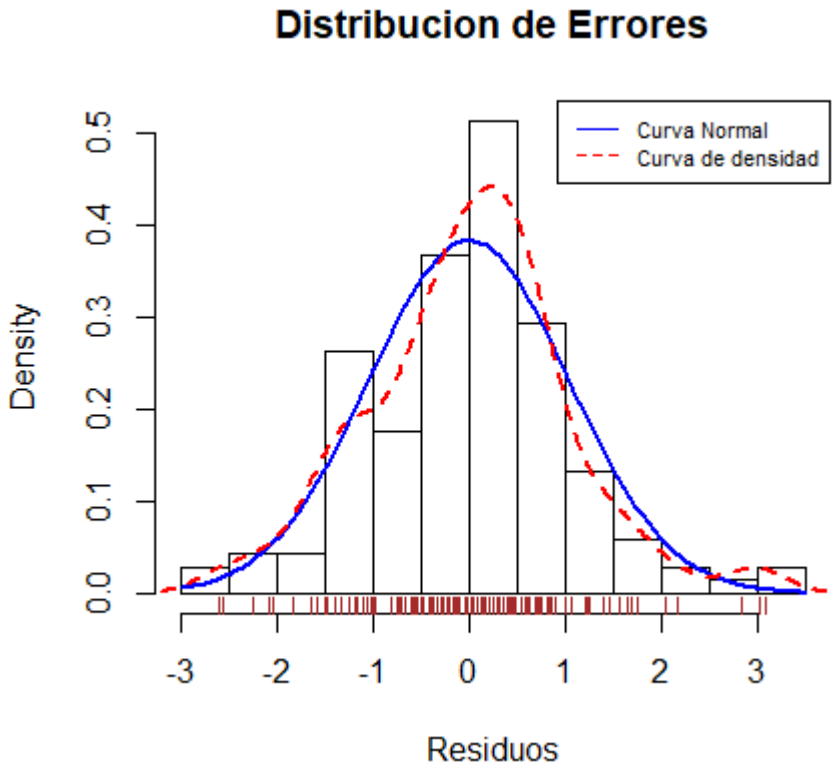
Desarrollo

Antes de proceder a realizar lo explicado anteriormente, debemos saber a que tipo de datos nos estamos enfrentando esto es, determinar el modelo de regresión construido.



Llegamos a la conclusión de que la relación que presentan los datos es lineal y por lo tanto podemos continuar con el análisis.

Al igual que la distribución de errores:



Fuente: Elaboración propia

Podemos ver que se aproxima a una campana normal, aunque con ciertos matices. Ahora que tenemos claro el tipo de modelo que tenemos podemos proceder al estudio de las variables.

Como se ha explicado al inicio de este informe, la variable dependiente será la renta (columna 1) y las independientes irán variando en función de los resultados obtenidos en cada prueba a través del criterio de Akaike y Schwarz. Se seleccionará aquel que presente menor valor.

Nuestro modelo de regresión incluirá las variables renta a 6 meses y renta a 1 año ya que son las que presentan el menor p valor.

Aquellos valores que se muestren con (***) indicarán la existencia de un p valor muy próximo a 0 y por tanto serán variables candidatas a ser escogidas.

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.341e+00  4.696e-01  -2.855 0.005069 **
Inv_minima_inicial       -6.184e-07  3.477e-07  -1.778 0.077855 .
X1_Day_Return             1.223e-02  1.100e-01   0.111 0.911653
X1_Week_Return           -2.930e-01  8.224e-02  -3.563 0.000526 ***
rent_1_mes               -2.504e-02  1.085e-01  -0.231 0.817855
rent_3_meses              1.335e-01  7.845e-02   1.702 0.091411 .
rent_6_meses             -1.976e-01  3.542e-02  -5.579  1.5e-07 ***
rent_en_el_anio           9.443e-01  2.990e-02  31.579  < 2e-16 ***
rent_5_anios              1.394e-01  5.709e-02   2.442 0.016037 *
rent_10_anios             7.017e-02  7.182e-02   0.977 0.330485
Capitaliz_media_bursatil  5.957e-06  5.080e-06   1.172 0.243308
Patrimonio               -2.862e-05  2.329e-04  -0.123 0.902409
Volatilidad_3            -1.630e-01  4.968e-02  -3.281 0.001354 **
Sharpe_.3                 1.036e+00  5.610e-01   1.846 0.067294 .
Ratio_de_informacion      4.727e-02  1.641e-01   0.288 0.773743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se ha descartado la renta a 10 años por presentar un número significativo de *NA.*

El problema surge al realizar el AIC y el BIC, cuyos valores nos indican que el modelo válido (el que tiene menor valor) corresponde al que contiene todas las variables. Jugando con aquellos valores con menor p valor, sacándolos e introduciéndolos en nuestro modelo no se consigue una combinación cuyo valor sea menor al del modelo completo.

Siguiendo las técnicas de selección de variables:

- Best Subset: Se estiman todas las regresiones posibles combinando los regresores usando el dataframe creado anteriormente con las variables de tipo no string.

```
         Inv_minima_inicial X1_Day_Return X1_Week_Return rent_1_mes rent_3_meses rent_6_meses rent_en_el_anio rent_5_anios
1  ( 1 ) " "                " "           " "           " "        " "          " "          "*"            " "
2  ( 1 ) " "                " "           " "           " "        " "          "*"          "*"            " "
3  ( 1 ) " "                " "           " "           " "        " "          "*"          "*"            " "
4  ( 1 ) " "                " "           " "           " "        " "          "*"          "*"            " "
5  ( 1 ) " "                " "           "*"           " "        " "          "*"          "*"            "*"
6  ( 1 ) " "                " "           "*"           " "        " "          "*"          "*"            "*"
7  ( 1 ) " "                " "           "*"           " "        "*"          "*"          "*"            "*"
8  ( 1 ) "*"                " "           "*"           " "        "*"          "*"          "*"            "*"
         rent_10_anios Capitaliz_media_bursatil Patrimonio Volatilidad_3 Sharpe_.3 Ratio_de_informacion
1  ( 1 ) " "           " "                      " "        " "          " "        " "
2  ( 1 ) " "           " "                      " "        " "          " "        " "
3  ( 1 ) " "           " "                      " "        " "          "*"        " "
4  ( 1 ) " "           " "                      " "        "*"          "*"        " "
5  ( 1 ) " "           " "                      " "        "*"          " "        " "
6  ( 1 ) " "           " "                      " "        "*"          "*"        " "
7  ( 1 ) " "           " "                      " "        "*"          "*"        " "
8  ( 1 ) " "           " "                      " "        "*"          "*"        " "
```

Aquella variable que mejor valor presenta sería por tanto la renta a un año y el BIC total estaría representado por el valor -436.86

```
> reg.summary$bic
[1] -328.2826 -370.5174 -430.9200 -432.6613 -436.4863 -436.8609 -435.4570 -434.6386
```

Dado que ninguno de los modelos que hemos propuesto es válido (por lo comentado anteriormente) debemos usar las técnicas de selección, las cuales irán introduciendo y descartando variables hasta llegar a una última iteración donde se presenta el menor valor del criterio de Akaike posible, así como las variables que se han seleccionado y/o descartado.

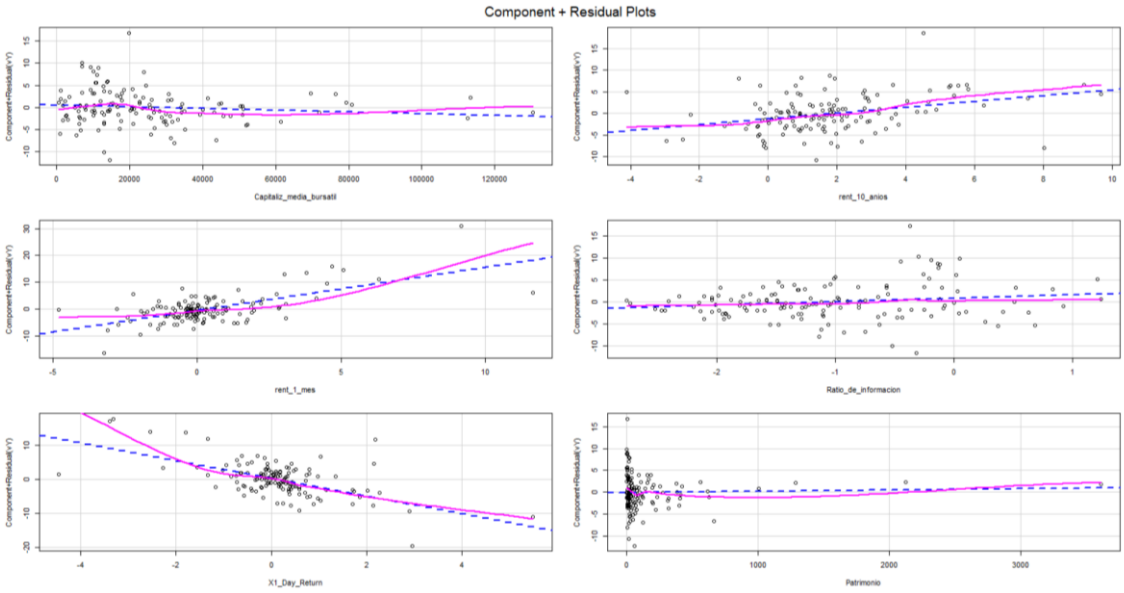
- Usando el Forward Stepwise la iteración comenzará con un modelo sin regresores al cual se irán añadiendo poco a poco.
- Con el Backward Stepwise el modelo incluye todos los regresores y se van eliminando poco a poco.

Ya que ninguno de los anteriores garantiza la selección del mejor modelo, normalmente se utilizan modelos mixtos que incluyen los anteriores.

El AIC obtenido en la primera iteración tiene un valor de 15.99, pasando en la última a 7.44

Por tanto, usando el modelo con las variables que nos devuelve este algoritmo el AIC baja de manera considerable pero aún así no hemos conseguido que sea el menor de ambos.

Para contrastar esto de manera gráfica usamos el gráfico de residuos, en el cuál se aprecia que no hay relación lineal con la variable principal.



Fuente: Elaboración propia

Conclusiones

A pesar de haber propuesto varios modelos cuyas variables parecían ser las indicadas, hemos comprobado que el modelo total (el que incluía todas las variables) es el que mejor se ajusta a la variable dependiente.

Esto nos hace pensar que la influencia de datos *NA* ha podido causar tales consecuencias y por tanto no se ha podido demostrar con total claridad cuales son las variables independientes relacionadas estrechamente con la renta.