SESIÓN 3: REGRESION

1. **Enumere cuales son las hipótesis que asumimos al hacer una regresión múltiple entre una variable de respuesta y y unas variables predictoras, $x_1, \ldots, x_p$.**

   La variable dependiente (resultado) debe ser escalar (numérica) o bien ordinal de más de 5 categorías, es decir, las categorías de la variable dependiente deben tener un orden interno o jerarquía.

   Las variables independientes (explicaciones) deben ser escalares (numérica), ordinales o dummy (variables de dos categorías donde una indica existencia u otra no-existencia).

   Las variables independientes no pueden estar altamente correlacionadas entre sí, las relaciones entre las variables independientes y la variable dependiente deben ser lineales, todos los residuales deben seguir la distribución normal y deben tener varianzas iguales.

2. **En un modelo de regresión, como se calcula y como se interpreta el coeficiente de determinación R^2.**

   ```
   R2 = anova(regval)$"Sum Sq"[1] / (var(x) * (length(x) - 1))
   ```

   El coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca del 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de 0, menos ajustado estará el modelo y, por lo tanto, menos fiable será.

3. **Lea el fichero "BCN_pisos.txt". Del fichero resultante seleccione 2/3 partes como muestra de training y la tercera parte restante como muestra test.**

   Asumiendo que queremos particionar las filas del dataframe de manera secuencial y no aleatoria, calculamos el 2/3 del número de filas total y seleccionamos las primeras resultantes como training, la resta, es decir el 1/3, como test.

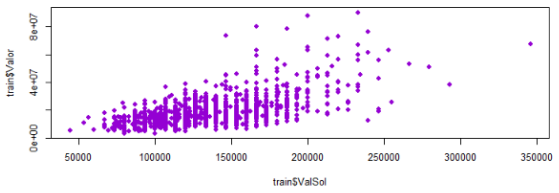
   ```
   dd <- read.table("bcn_pisos.txt", header=TRUE)
   ntr = nrow(dd)* (2/3) #1553
   train = dd[1:1553,]
   test = dd[1554:2329,]
   ```

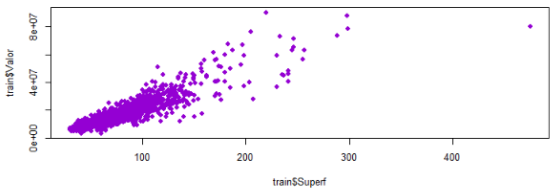
4. **Con la muestra de training, efectúe la representación gráfica de la variable "Valor" respecto del resto de variables del fichero. Calcule la correlación entre la variable "Valor" y el resto de variables numéricas.**

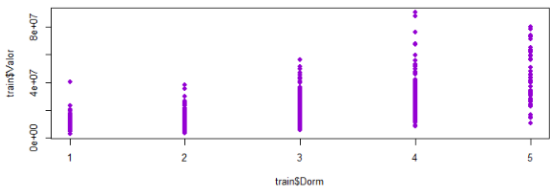
Utilizando la muestra training y escogiendo únicamente variables numéricas obtenemos la siguiente representación gráfica de "Valor" y el resto de variables.

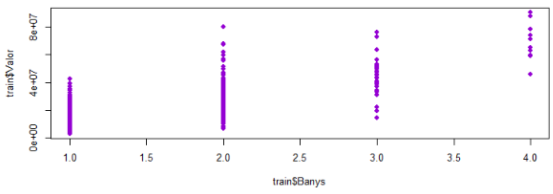
Calculamos la correlación de las variables numéricas y obtenemos que hay una relación lineal más fuerte con la variable Superficie ya que obtenemos un valor muy próximo a 1 (0.89). Seguidamente a esta variable podemos observar que hay una fuerte relación con las variables Baños y Valor del suelo (0.74 y 0.62 respectivamente).

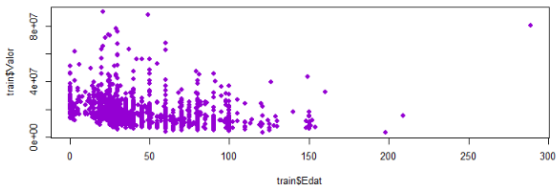
```
cor(train$Valor,train$Superf)
[1] 0.8906862
cor(train$Valor,train$ValSol)
[1] 0.6288794
cor(train$Valor,train$Edat)
[1] -0.2625441
cor(train$Valor,train$Dorm)
[1] 0.5825866
cor(train$Valor,train$Banys)
[1] 0.7404788
```











5. **Efectúe la regresión simple de la variable "Valor" respecto de la "Superficie". A continuación añada a la regresión la variable "Número de dormitorios". Es significativa esta variable una vez que el modelo ya contiene la variable "Superficie".**

Una vez efectuada la regresión simple de la variable "Valor" respecto de la "Superfice" en R obtenemos un Intercept y un Slope de y = 257308x -3795506, donde podemos estimar que por cada metro de superficie el valor aumentará 257308 pesetas. A continuación, añadimos la variable "Número de dormitorios" y obtenemos: y = $261676X_{sup}$ $-239362X_{Dor}$ -3476850. Junto con esta regresión obtenemos los p_values de los coeficientes que indican si el predictor es significativo y observamos que la variable "Número de dormitorios" tiene un p_value de 0.145, un valor alejado de p<0.05, por lo tanto, una variación en este predictor no tendrá un cambio significativo en nuestra respuesta, por lo tanto, podemos concluir que no es significativa esta variable una vez el modelo ya contiene la variable "Superficie".

```
> regdorm = lm(formula = Valor ~ Superf + Dorm, data = train )

> summary(regdorm)

Call:

lm(formula = Valor ~ Superf + Dorm, data = train)

Residuals:

     Min       1Q   Median       3Q      Max

-39573393  -2204467   -325492   1902878  37066258

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) -3476850     375129  -9.268   <2e-16 ***

Superf       261673       4480  58.403   <2e-16 ***

Dorm        -239362     164147  -1.458    0.145
```

6. **Efectúe la regresión múltiple del "Valor" respecto el resto de variables del fichero "BCN_pisos".¿Le parece que alguna variable predictora es no significativa?.**

Haciendo la regresión múltiple de "Valor" respecto de todas las variables del fichero y seguidamente analizando los p_value asociados a las variables utilizando *Anova* en R. Podemos comprobar variables predictoras no significativas como "Ascens" o "Dorm", ya que tienen un p_value mayor que nuestro nivel significativo o alpha 0.05.

```
names(train)

reg = lm(formula = Valor ~., data = train )

summary(reg)

anova(reg)

Analysis of Variance Table

Response: Valor

             Df     Sum Sq    Mean Sq    F value    Pr(>F)

Superf        1 1.2830e+17 1.2830e+17 21456.3861 < 2.2e-16 ***

Dorm          1 4.5791e+13 4.5791e+13     7.6581 0.0057203 **

Banys         1 4.2805e+15 4.2805e+15   715.8749 < 2.2e-16 ***

Edat          1 2.7376e+15 2.7376e+15   457.8308 < 2.2e-16 ***

Estat         4 2.2205e+15 5.5513e+14    92.8399 < 2.2e-16 ***

Planta        2 8.7699e+13 4.3850e+13     7.3335 0.0006767 ***

Dist          9 6.9748e+15 7.7497e+14   129.6076 < 2.2e-16 ***

ValSol        1 7.5178e+15 7.5178e+15  1257.2901 < 2.2e-16 ***

Tipus         1 1.3042e+14 1.3042e+14    21.8115 3.273e-06 ***

Ascens        1 6.7758e+11 6.7758e+11     0.1133 0.7364425

ExtInt        1 6.1683e+13 6.1683e+13    10.3159 0.0013465 **

Reforma       8 2.7195e+14 3.3994e+13     5.6852 3.839e-07 ***

Residuals  1521 9.0947e+15 5.9794e+12
```

7. **Encuentre la regresión óptima. ¿Cuál es el valor del R^2 alcanzado?. ¿Y cuál el valor del R^2 por validación cruzada "leave one out"?**

Utilizando la librería olsrr para poder calcular los mejores subconjuntos de la regresión conjuntamente con su summary, obtenemos que para 8 variables predictoras ("Superf", "Banys", "Edat", "Estat", "Dist", "Valsol", "Tipus" "Reforma") es el resultado más óptimo ya que es cuando la R2 de predicción deja de tener un cambio significativo. El valor R2 alcanzado es de 0.9395.

```
> names(train)
 [1] "Valor"   "Superf"  "Dorm"    "Banys"   "Edat"    "Estat"   "Planta"  "Dist"
 [9] "ValSol"  "Tipus"   "Ascens"  "ExtInt"  "Reforma"
```

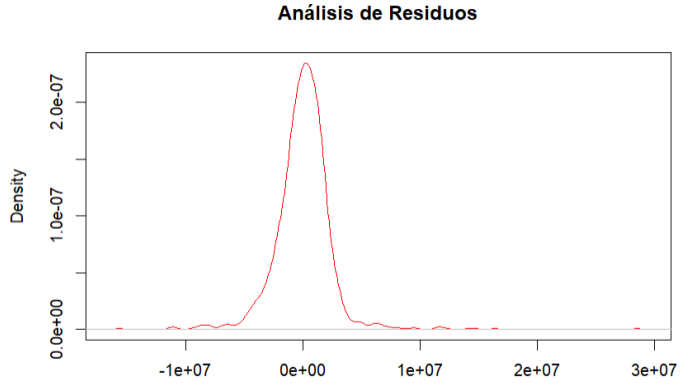
```
> train = dd[1:1553,]
> regval6 <-lm (Valor ~ ., data=train)
> bestval<-ols_step_best_subset(regval6)
> bestval
```

```
                                Best Subsets Regression
-------------------------------------------------------------------------------------------
Model Index    Predictors
-------------------------------------------------------------------------------------------
     1         Superf
     2         Superf ValSol
     3         Superf Estat ValSol
     4         Superf Banys Estat ValSol
     5         Superf Banys Estat Dist ValSol
     6         Superf Banys Edat Estat Dist ValSol
     7         Superf Banys Edat Estat Dist ValSol Reforma
     8         Superf Banys Edat Estat Dist ValSol Tipus Reforma
     9         Superf Banys Edat Estat Dist ValSol Tipus ExtInt Reforma
    10         Superf Banys Edat Estat Planta Dist ValSol Tipus ExtInt Reforma
    11         Superf Dorm Banys Edat Estat Planta Dist ValSol Tipus ExtInt Reforma
    12         Superf Dorm Banys Edat Estat Planta Dist ValSol Tipus Ascens ExtInt Reforma
-------------------------------------------------------------------------------------------
```

```
                                                                Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------
                      Adj.        Pred
Model    R-Square     R-Square    R-Square      C(p)          AIC          SBIC          SBC
-----------------------------------------------------------------------------------------------------------------
  1        0.7933       0.7932      0.7903    4040.8708    52090.4909    47679.3361    52106.5348
  2        0.9095       0.9094      0.9075     900.8803    50810.1265    46400.4427    50831.5183
  3        0.9308       0.9306      0.9287     325.5581    50400.3919    45985.7530    50443.1754
  4        0.9363       0.9360      0.9341     180.3917    50275.1350    45860.9037    50323.2665
  5        0.9389       0.9382      0.936      112.1462    50228.5085    45798.5511    50324.7715
  6        0.9406       0.9399      0.9378      67.7642    50186.3133    45756.6195    50287.9242
  7        0.9424       0.9415      0.939       20.4169    50153.8695    45710.6010    50298.2639
  8        0.9431       0.9421      0.9395       4.5822    50137.9827    45694.9164    50287.7252
  9        0.9434       0.9424      0.9398      -1.4542    50131.8550    45688.9117    50286.9454
 10        0.9437       0.9426      0.9399      -8.2550    50126.9052    45682.1184    50292.6914
 11        0.9438       0.9426      0.9398      -7.7955    50127.3333    45682.6110    50298.4675
 12        0.9438       0.9426      0.9397      -6.0000    50129.1245    45684.4476    50305.6067
```

8. Realice el análisis de los residuos. ¿Son normales los residuos?, ¿Existe alguna relación de dependencia con los valores ajustados?. ¿Existe heterocedasticidad?. ¿Existen observaciones influyentes?

Podemos observar haciendo una gráfica de densidad de los residuos de la regresión que siguen la tendencia de una distribución normal.

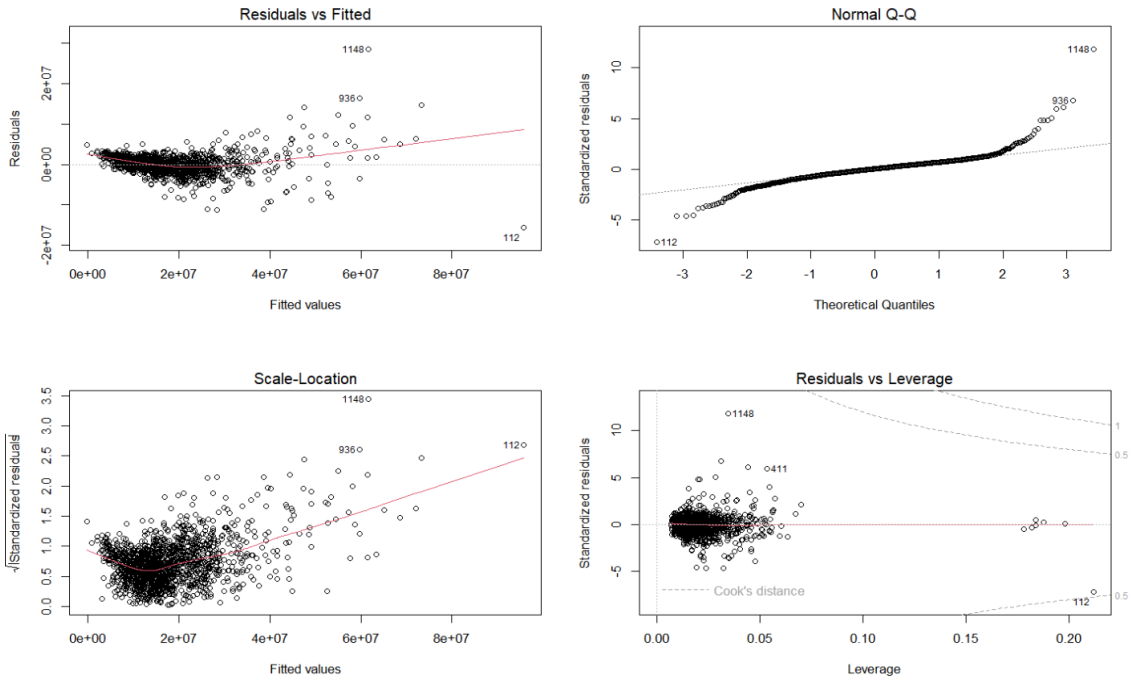


También aprovechando el siguiente plot de análisis de residuos del modelo, podemos ver el gráfico Q-Q Normal donde los puntos se aproximan razonablemente bien a la diagonal lo que confirma la hipótesis de normalidad.

Los residuos no deberían mostrar ninguna relación con otra variable. En la gráfica *Residuals vs Fitted* podemos observar que no hay un patrón discernible. Por lo tantol, concluimos que no hay dependencia entre los errores y los valores ajustados.

Observando la gráfica *Scale-Location* observamos que no hay heterocedasticidad ya que no existe un patrón definido en la nube de puntos.

Analizando la gráfica de *Residuals vs Leverage* observamos que hay aproximadamente 7 puntos que podrían ser outliers, ya que muestran un comportamiento anormal en la muestra estadística.



9. Obtenga el valor del R^2 de predicción en la muestra test.

R^2 es de 0.94, por lo tanto, podríamos decir tiene un ajuste lineal casi perfecto.

```
regval2<-lm(Valor ~ ., data=test)
PRESS <- sum((regval2$residuals/(1-ls.diag(regval2)$hat))^2)
R2loo <- 1 - PRESS/(var(test$Valor)*(nrow(test)-1)) R2loo
[1] 0.9414378
```

10. **Obtenga el fichero con las predicciones del valor de las viviendas con su intervalo de confianza del 95%, para los pisos de la muestra test.**

Se pueden ver el intervalo de confianza en nuestros valores predichos en la salida anterior.

	fit	lwr	upr
1554	15265238	11007189.10	19523287
1555	11462982	7229773.93	15696191
1556	17796479	13561245.36	22031712
1557	19295999	15038256.80	23553741
1558	26471119	22227205.48	30715033
1559	18259458	13961613.70	22557302
1560	22084041	17867043.92	26301037
1561	18360716	14152606.55	22568825
1562	10580055	6346362.99	14813747
1563	27194364	22993835.82	31394892

…