Práctica 1: Programación en R

Probabilidad y Estadística

Introducción

El objetivo de esta práctica es realizar un estudio muy superficial de modelos de regresión simple. Para ello, necesitaremos un conjunto de datos (*dataset*) sobre el que hacer el estudio. En la primera parte de la práctica, cargaremos dicho *dataset* y realizaremos una exploración estadística de ellos. En la segunda parte, realizaremos un estudio de regresión más específico.

Información del *dataset*

El dataset elegido es "*Boston Housing*". Consiste en un *dataset* público con información de viviendas en Boston. En concreto, el *dataset* contine las siguientes columnas:

- **CRIM**: Per capita crime rate by town
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft
- **INDUS**: Proportion of non-retail business acres per town
- **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- **NOX**: Nitric oxide concentration (parts per 10 million)
- **RM**: Average number of rooms per dwelling
- **AGE**: Proportion of owner-occupied units built prior to 1940
- **DIS**: Weighted distances to five Boston employment centers
- **RAD**: Index of accessibility to radial highways
- **TAX**: Full-value property tax rate per $10,000
- **PTRATIO**: Pupil-teacher ratio by town
- **B**: 1000(Bk - 0.63)?, where Bk is the proportion of [people of African American descent] by town
- **LSTAT**: Percentage of lower status of the population
- **MEDV**: Median value of owner-occupied homes in $1000s"'

El *dataset* lo cargaremos en R mediante la instalación de un paquete y su posterior importación:

```
install.packages("mlbench")
library(mlbench)
#Libreria para calculos (Asimetria, apuntamiento)
install.packages("moments")
library(moments)
#Libreria correlacion
install.packages("corrplot")
library(corrplot)
```

Ya cargado en R, procedemos a guardar el *dataset* en la variable `housing` para que realicéis el resto de la práctica sobre ella:

```
data("BostonHousing")
housing <- BostonHousing
```

Información de la práctica.

- La práctica se calificará sobre 10 puntos.

- La práctica se resuelve sobre este mismo código.
- Antes de entregarlo, habrá que cambiar el nombre del fichero, sustituyendo *nombre1* y *apellido1* por los propios del alumno.
- Puntuará más aquellas descripciones
- Puntuarán más aquellos gráficos que muestren más elaboración.
- Si el alumno considera relevante la instalación de alguna librería extra que pueda mostrar mejores resultados o gráficos, es libre para hacerlo. Puntuará más si se realiza un trabajo óptimo en esta parte.
- Sobre todo en la parte final (regresión), puntuará más aquellos comentarios del alumno que muestren haber investigado el significado de todos los análisis realizados.
- Si el alumno se encuentra con errores durante la ejecución del código, tiene que aprender a lidiar con ellos como futuro ingeniero informático.

1) Análisis exploratorio inicial:

Se pide utilizar los comandos `str`, `head`, `dim`, `summary` sobre `housing` para explorar distribución inicial de los datos en el *dataset*.

```
# CODIGO DEL ALUMNO
#Estructura interna del dataset
str(housing)
#Obtenemos primeros objetos del dataset
head(housing)
#Dimension del dataset (obj var)
dim(housing)
#Sumarios del dataset. Nos muestra sumarios genericos de las variables
summary(housing)
```

Comentarios del alumno (máximo 100 palabras):

```
# TEXTO DEL ALUMNO
El comando str() nos muestra que tenemos 506 objetos con 14 variables.
Ya sabemos que nuestro espacio muestral es n=506.
Tambien nos detalla las variables (Tipo de variable y algunos de sus valores)
Con el head() vemos los valores de las primeras 6 muestras.
dim(): dimensiones del dataset 506 objetos 14 var
summary(): podemos observar minimo, P25, mediana, media, P75 y
el maximo de cada variable en nuestro espacio muestral.
```

2) Análisis exploratorio de la variable objetivo:

En esta práctica, trataremos de predecir el valor del precio medio de la vivienda MEDV. Para ello, primero exploraremos la distribución de sus datos. Se pide dibujar un histograma, calcular asimetría y apuntamiento de MEDV. Se pide dibujar un boxplot, calcular los cuartiles y los percentiles 10-90 sobre MEDV. Se pide describir los elementos más importantes de ambas gráficas.

```
# CODIGO DEL ALUMNO
MEDV =  housing[,14]
#Histograma
hist(MEDV)
print("Asimetria")
skewness(MEDV)
print("Apuntamiento")
kurtosis(MEDV)
#Boxplot
boxplot(MEDV)
print("Cuartiles")
```

```
quantile(MEDV)
print("Percentiles")
quantile(MEDV, c(.10, .20, .30, .40, .50, .60, .70, .80, .90))
```

Comentarios del alumno (máximo 100 palabras):

```
# TEXTO DEL ALUMNO
Tanto en el histograma como en skewness() se observa una asimetria
positiva (por la derecha).
El calculo de la asimetria como el de apuntamiento (skewness y kurtosis)
utilizan los coeficientes de Fisher
El apuntamiento es >3 por lo que tenemos una distribucion leptocurtica
En cuanto al boxplot se observa la presencia de valores atipicos y
llama la atencion que los extremos del rango intercuartilico (Q1 Y Q3)
estan bastante alejados del maximo y el minimo.
```

3) Correlación entre variables:

Se pide utilizar el comando `corrplot` de la libreria `corrplot` (que posiblemente haya que instalar) para mostrar la matriz de correlaciones entre todas las variables. Se pide describir los elementos de la gráfica que aportan mayor información.

```
# CODIGO DEL ALUMNO
#Correlacion entre todas las variables menos **CHAS** por que es booleana.
cor = cor(housing[,-4])
corrplot(cor, method = 'circle')
```

Comentarios del alumno (máximo 100 palabras):

```
# TEXTO DEL ALUMNO
La funcion cor() nos calcula el coeficiente de correlacion de
Pearson (-1 < r < 1) entre las variables
cooplot() nos dibuja una matrix grafica para observar
estas correlaciones mas facilmente
r > 0  => dependencia directa (Positiva); r  = 0 => independientes;
r < 0 => dependencia inversa (negativa)
La diagonal azul la obviamos pues nos indica interdependencia total de cada
variable con ella misma
Podemos observar interdependencia positiva casi total entre TAX y RAD.
Tambien dependencias positivas entre: medv-rm, tax-indus, age-nox, tax-nox,
rad-crim y crim-tax
Dependencias negativas: dis-indus, age-zn, ...
Dependencia negativa total: medv-istat, age-dis, dis-nox, ...
Dependencia casi nula: b-rm, rm-dis, b-zn, ...
```

4) Regresiones lineales simples:

Se pide escoger cuatro variables independientes (a vuestro juicio, las mejores) y realizar tres regresiones simples con cada una de ellas sobre la variable dependiente MEDV. Justificar qué criterio(s) habéis tomado para elegir dichas cuatro variables. Se pide dibujar los *scatterplots* de cada variable independiente con MEDV y la recta de regresión resultante sobre cada *scatterplot*. Describir brevemente el resultado de este análisis.

```
# CODIGO DEL ALUMNO
# Variables
ISTAT = housing[,13]
RM = housing[,6]
```

```
PTRATIO = housing[,11]
INDUS = housing[,3]
#Calculos
regresion1 <- lm(MEDV ~ ISTAT, data = housing)
plot(ISTAT, MEDV, main="ISTAT-MEDV")
abline(regresion1)
regresion2 <- lm(MEDV ~ RM, data = housing)
plot(RM, MEDV, main="RM-MEDV")
abline(regresion2)
regresion3 <- lm(MEDV ~ PTRATIO, data = housing)
plot(PTRATIO, MEDV, main="PTRATIO-MEDV")
abline(regresion3)
regresion4 <- lm(MEDV ~ INDUS, data = housing)
plot(INDUS, MEDV, main="INDUS-MEDV")
abline(regresion4)
```

Comentarios del alumno (máximo 300 palabras):

```
# TEXTO DEL ALUMNO
He escogido como variables independientes las que tienen mayor correlacion
con MEDV para obtener formulas
utilizables de la regresion lineal
En los dos primeros graficos se puede observar como la linea de regresion
corta por la mitad la nube de puntos.
Es asi por que las dos primeras variables tienen gran correlacion con MEDV.
Sin embargo, las dos ultimas tienen menor correlacion y en sus graficos se
observa como la linea de regresion ya no se aproxima tanto a los puntos
MEDV-ISTAT tienen correlacion negativa por lo tanto la pendiente de la
regresion lineal es negativa. Lo mismo ocurre con la coorelacion
positiva de MEDV-RM
```

5) Análisis de los residuos:

Se pide mostrar en un *scatterplot* los residuos e_i (eje-x) y la predicción que hace cada recta de regresión del apartado anterior con cada punto $\hat{y}_i$ (eje-y), también llamada variable ajustada. Se pide realizar un histograma de los residuos exclusivamente. Se pide investigar el significado y la importancia de este gráfico y comentarlo brevemente.

```
# CODIGO DEL ALUMNO
residuos = residuals(regresion1)
vars = fitted(regresion1)
plot(residuos,vars, main="RESIDUOS-VAR AJUSTADOS")
hist(residuos)
```

Comentarios del alumno (máximo 200 palabras):

```
# TEXTO DEL ALUMNO
La gran mayoria de los valores ajustados tienen un residuo entre (-10,0)
Por lo tanto la gran mayoria de los residuos estan ubicados entre (-10,0)
```

6) Regresión lineal múltiple:

Se pide realizar una regresión lineal múltiple con las cuatro variables independientes a la vez sobre la variable MEDV. Se pide, además, realizar un análisis de los residuos, similar al realizado en el apartado anterior.

```r
# CODIGO DEL ALUMNO
modelo <- lm(MEDV ~ ISTAT + RM + PTRATIO + INDUS , data = housing )
residuos = residuals(modelo)
vars = fitted(modelo)
plot(residuos,vars, main="RESIDUOS-VAR AJUSTADOS")
hist(residuos)
```

Comentarios del alumno (máximo 300 palabras):

```r
# TEXTO DEL ALUMNO
En esta regresion se pueden observar valores mas altos de residuos y un mayor
intervalo de valores para ellos (-10,10)
```