Práctica: Regresión múltiple

1. Implemente una función que dado un vector **y** de valores de respuesta y una matriz **X** de valores observados, mediante las ecuaciones normales, el estimador de cuadrados mínimos $\hat{\beta}$.

2. Implemente una función que dado un vector **y** de valores de respuesta, un vector **x** de valores observados y un entero n, devuelva los estimadores de cuadrados mínimos basados en la siguiente regresión:

$$Y \cong \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n.$$

   (Sugerencia, haga uso de la función anterior).

3. Se sospecha que hay una relación entre el precio de un inmueble (Y) en miles de pesos y la cantidad de metros cuadrados (X). Se tienen registradas 120 ventas efectuadas con anterioridad, almacenadas en el archivo **inmuebles.txt**.

   (a) Realizar un *scatter plot* donde en el eje X figure los metros cuadrados y en el eje Y el precio (en miles de pesos).

   (b) Realizar una regresión lineal simple mediante cuadrados mínimos.

   (c) Graficar, sobre el *scatter plot* la recta de regresión.

   (d) Realizar una regresión cuadrática mediante cuadrados mínimos y grafíquela superpuesta en el *scatter plot*.

4. Se desea estudiar el vínculo entre la edad y la presión arterial sistólica. Para eso, se tiene el archivo **presion.txt** con esos datos, entre otras cosas.

   (a) Levante los datos de interés del archivo **presion.txt**. No todos los datos allí presentes son relevantes para el problema en cuestión, haga la "limpieza" necesaria.

   (b) Ensaye diversos modelos de regresión, graficando los resultados sobre un *scatter plot* y determine, en su opinión, la mejor forma de explicar ese vínculo. Haga los arreglos que considere pertinentes.

5. Volviendo al archivo **abalone.txt**, observe que el conjunto de datos tiene información del peso total de cada especimen junto con un desagregado por partes. Ajuste un modelo multilineal que explique el peso total en función del peso del caparazón, las vísceras y la carne.

6. Bajo el mismo archivo (**abalone.txt**), trate de establecer una relación entre el peso total y el diámetro del especimen. Empezar dibujando en un *scatter plot* ambos parámetros. Basándose en eso, considere los siguientes modelos:

   - Modelo lineal simple, $Peso = b + a * Diametro$.
   - Modelo cuadrático, $Peso = c + b * Diametro + a * Diametro^2$.
   - Modelo cúbico sin términos de orden inferior, $Peso = a * Diametro^3$.
   - Modelo exponencial, $\log(Peso) = b + a * Diametro$.

   (a) Efectúe en cada caso una regresión y grafique las curvas superpuestas sobre el *scatter plot*.

   (b) Identifique el "mejor" modelo mediante un esquema de **training** y **testing set** y quedándose con el que minimiza el error cuadrático medio en el conjunto de testing, calculando el error de la forma usual para los tres primeros modelos , y usando $Peso = e^{b+a*Diametro}$ para determinarlo en el cuarto.

7. (Regresión lineal con variables categóricas / factores) Abra el archivo **Credit.csv** (fuente: `http://www-bcf.usc.edu/~gareth/ISL/data.html`), que tiene información crediticia de diversos individuos. Inspeccione a ojo los datos para determinar su tipo.

   (a) Regresione multilinealmente **balance** en función de **income**, **limit**, **rating** y **student** (categórica). Sugerencia, codificar con 1 a un estudiante y con un 0 a un no estudiante.

(b) Haga lo mismo adicionando la variable categórica **ethnicity**. Sugerencia, agregar dos variables dicotómicas (0 o 1) para dos etnias en particular (la tercer etnia sale por "descarte" de las otras dos).

(c) Utilizando como base las variables **income**, **rating** y **student**, determine el subconjunto que mejor explica la situación haciendo un esquema de partición entre **training** y **testing set** y eligiendo aquel que minimiza el error cuadrático medio de predicción.