Aprendizaje Estadístico - FIUBA
Guía 1 - Regresión Lineal

---

**1.** La idea de este ejercicio es estudiar el comportamiento de las distintas graficas ante diversas situaciones simuladas.

1. Genere, utilizando el comando `rnorm(1000)`, 1000 elementos provenientes de una normal estandar, guarde los valores en una variable `x`. Obtenga un histograma, un boxplot y un qqnorm de `x`.

2. Repita el análisis anterior ante las siguientes situaciones:

   a) Binomial: `rbinom(1000,10,0.4)`.
   b) Chi-cuadrado: `rchisq(1000,50)`.
   c) F de Snedecor: `rf(1000,90,40)`.
   d) Gamma: `rgamma(1000,0.7)`.

---

**2.** El archivo `graduados.txt` contiene los promedios obtenidos en su carrera de grado de 30 inscriptos en el programa de postgrado del Departamento de Ingeniería Industrial de la Universidad de Berkeley.

1. Calcular la media, la mediana muestral y la media 10 %-podada.

2. Calcular el desvo estandar, la distancia intercuartil y la MAD.

3. Realizar un boxplot sobre este conjunto de datos. ¿Cuáles son las características mas sobresalientes? ¿Cómo relaciona lo observado en el boxplot con lo obtenido en los incisos anteriores?

4. ¿Es razonable suponer normalidad de los datos? ¿Con qué gráfico o herramienta lo podría verificar?

---

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

---

**4.** Se tiene en el archivo `girasol.txt` el rinde de diversas parcelas de girasol (en toneladas) según la cantidad de dinero invertida en fertilizantes (en miles de pesos).

1. Levante los datos del archivo y grafique en un diagrama de dispersión inversión vs rinde.

2. Bajo un modelo de regresión lineal simple obtenga el estimador de mínimos cuadrados.

3. Grafique la recta de regresión obtenida, ¿detecta algo sospechoso?
   Efectúe una "limpieza"de los datos y repita el procedimiento.

---

**5.** Considere el archivo `abalone.txt` que contiene información sobre distintas muestras de abalones. Los atributos están separados por coma, con los siguientes campos:
Sexo (categorica): M (masculino), F (femenino) o I (infante).
Longitud (continua), en milimetros.
Diametro (continua), en milimetros.
Altura (continua), en milimetros.
Peso completo del abalone (continua), en gramos.
Peso de la carne (continua), en gramos.
Peso de las visceras (continua), en gramos.
Peso del caparazon (continua), en gramos.
Anillos (entera).

Efectúe una regresión lineal simple por cuadrados mínimos para obtener el diámetro en función de la longitud usando todos los datos

---

**6.** Volviendo al archivo `abalone.txt`, observe que el conjunto de datos tiene información del peso total de cada espécimen 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.

---

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

a) Modelo lineal simple, $Peso = b + aDiametro$.

b) Modelo cuadrático, $Peso = c + bDiametro + aDiametro^2$.

c) Modelo cúbico sin términos de orden inferior, $Peso = aDiametro^3$.

d) Modelo exponencial, $log(Peso) = b + aDiametro$.

1. Efectue en cada caso una regresión y grafique las curvas superpuestas sobre el scatter plot.
2. Identifique el "mejor"modelo mediante un esquema de training y testing set, quedándose con el que minimiza el error cuadrático medio en el conjunto de testing.

---

**8.** Consideremos

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde $\bar{y}$ es el promedio de $\{y_i\}_{1\leq i\leq n}$, e $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la aproximación por cuadrados mínimos. Probar que

$$TSS - RSS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

---

**9.** Consideremos el modelo

$$Y = \beta_0^* + \beta_1^* X + \varepsilon, \qquad X \perp \varepsilon,\ E[\varepsilon] = 0,\ Var[\varepsilon] = \sigma^2.$$

Recordando lo visto en clase que

$$\mathcal{R}^2 = \frac{Var(E[Y|X])}{V(Y)},$$

probar que $\mathcal{R}^2 = \rho(X,Y)^2$.

---

**10.**

1. Calcular la matriz Hessiana de

$$J(\beta_1, \ldots, \beta_p) = \tfrac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2.$$

2. Probar que si $\mathbf{X^T X}$ es no singular, entonces la primer iteración del método de Newton da

$$\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X}^T\mathbf{Y}.$$