Tema: Regresión

Objetivo: El objetivo del **análisis de regresión** es describir la relación (Una función) entre un conjunto de variables (aleatorias) $\mathbf{Y} = (Y_1, X_2, ..., Y_m)$, llamadas variables dependientes, y otro conjunto de variables , llamadas variables independientes $\mathbf{X} = (X_1, X_2, ..., X_n)$, es decir, entender cómo el valor de la variables dependientes cambia cuando cualquiera de las variables independientes cambia.

$$\mathbf{Y} = (Y_1, X_2, ..., Y_m) = f(\mathbf{X}) = f(X_1, X_2, ..., X_n).$$

El resultado final de este análisis, es establecer una función f, llamada **función de Regresión.**

El tipo más básico de Regresión **(Lineal).**

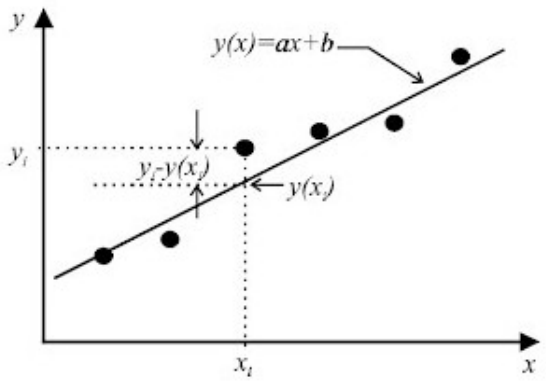
* Vector aleatoria $Z = (X, Y)$.

* $\mathbf{X} = (x_1, x_2, ...x_n)$ y $\mathbf{Y} = (y_1, y_2, ...y_n)$.

* Se asumen una relación lineal entre las y_i y x_i:

$$\widetilde{y_i} = f(x_i) = ax_i + b.$$

donde $y_i = \widetilde{y_i} + \varepsilon_i$, **con un error** ε_i.



* **Método**: *Uno de los métodos más populares para realizar regresión lineal es el de mínimos cuadrados ordinarios (**OLS**, por sus siglas en inglés). En este método,* $\mathbf{a}, \mathbf{b}$ *se* ***eligen*** *para* ***minimizar*** *(Usaremos Cálculo diferencial para esto) la suma de los cuadrados de los errores* $\sum_{i=1} \varepsilon_i^2$, ***de la distancia entre los valores estimados*** $\widetilde{y_i}$ ***y los valores reales*** y_i **.**

Definamos $h(a,b) := \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \Big(y_i - (ax_i + b) \Big)^2$

Tenemos que minimizar la función h, *por lo que tenemos que resolver el sistema de ecuaciones:*

A) $$\frac{\partial h}{\partial b} = \sum_{i=1}^{n} -2\Big(y_i - (ax_i + b) \Big) = 0$$

B) $$\frac{\partial h}{\partial b} = \sum_{i=1}^{n} -2\Big(y_i - (ax_i + b) \Big) = 0$$

El cual es equivalente al sistema:

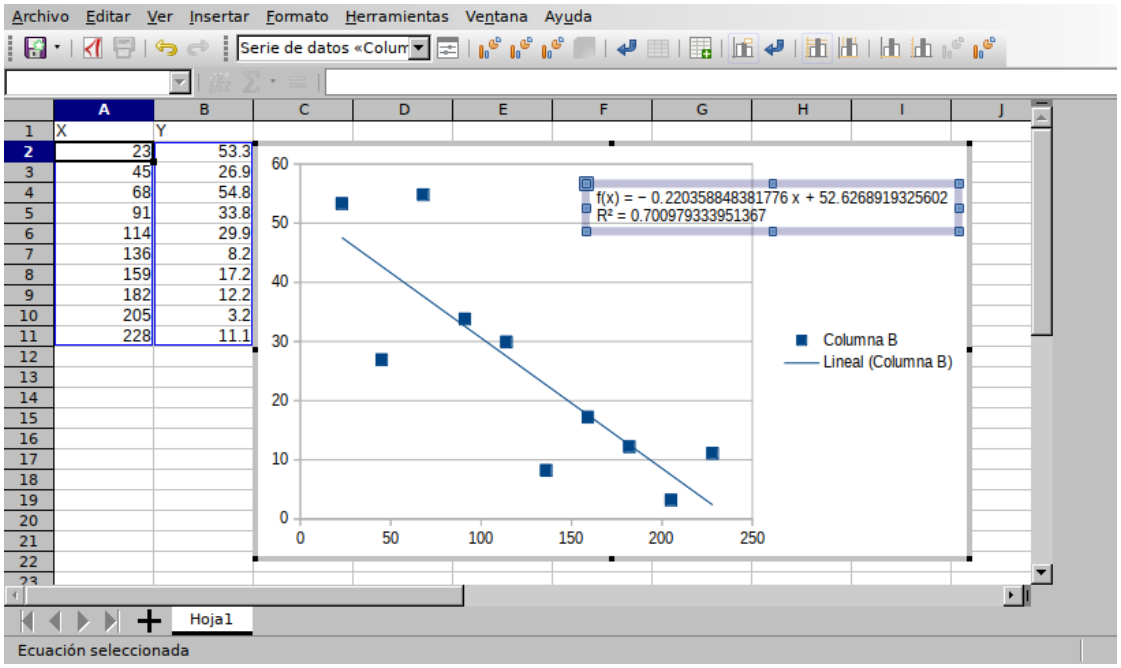
A) $$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i$$

B) $$\sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b$$

La solución de dicho sistema es

$$a = \frac{n\Big(\sum x_i y_i \Big) - \Big(\sum x_i \Big)\Big(\sum y_i \Big)}{n\Big(\sum x_i^2 \Big) - \Big(\sum x_i \Big)^2}.$$

$$b = \bar{y} - a\bar{x}.$$



	A	B
1	X	Y
2	23	53.3
3	45	26.9
4	68	54.8
5	91	33.8
6	114	29.9
7	136	8.2
8	159	17.2
9	182	12.2
10	205	3.2
11	228	11.1

Correlación de Variables aleatorias

Recordando algunas conceptos básicos de estadística

* **Media aritmética**: (Probabilidad $p_i := P(x_i)$, en nuestro caso asuminos $p_i = \frac{1}{n}$).

$$\mu = E(X) := \sum_{i=1}^{n} p_i x_i = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

* **Varianza**: La varianza intenta describir la dispersión de los datos.

$$\sigma_X^2 = \sigma^2 = Var(X) = E\Big[\Big(X - \mu\Big)^2\Big] = \sum_{i=1}^{n}(x_i - \mu)^2 p_i = E(X^2) - \mu^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.$$

* **Desviación estándar:**

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

Covarianza de variables aleatorias

$$\sigma_{XY}^2 := Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

Correlación: La correlación trata de establecer la relación o **dependencia** que existe entre las dos **variables aleatorias X, Y** (al menos una de ellas, $Z = (X, Y)$ será llamado un vector aleatoria). Es decir, determinar si los cambios en una de las variables influyen en los cambios de la otra. En caso de que suceda, diremos que las variables están correlacionadas o que hay correlación entre ellas.

La correlación es **positiva** (directa) cuando los valores de las variables aumenta juntos; y es **negativa** (inversa) cuando un valor de una variable se reduce cuando el valor de la otra variable aumenta.

La covarianza hay que entenderla en términos de probabilidad, una covarianza positiva sería el que a valores altos de una de la variables le corresponden **"con mayor probabilidad"** valores altos de la otra y a valores bajos valores bajos.(Una covariación negativa).

La Correlación de Variables aleatorias.

Se define, entonces ***el coeficiente de correlación*** *como:*

$$r = r_{XY} = \frac{Cov(X,Y)}{Var(x)Var(Y)} = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$$

Se tiene además que r_{XY} *esta acotada por* $|r_{XY}| < 1.$

Se tiene la siguiente relación entre a y r:

$$a = r\frac{S_Y}{S_X}.$$