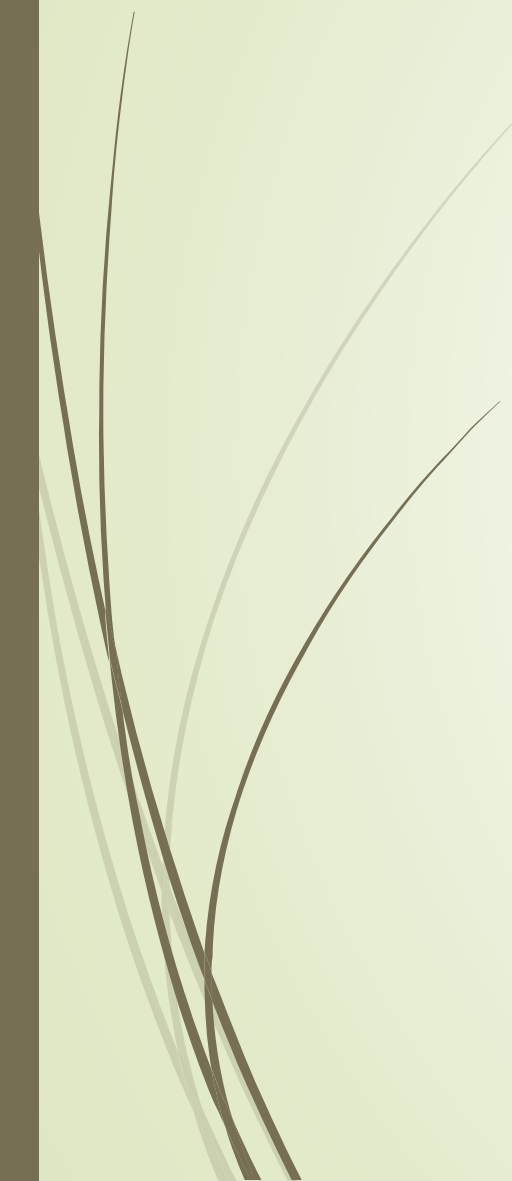


REGRESIÓN LINEAL SIMPLE





Regresión

- Técnica que permite identificar relaciones entre una variable dependiente y una o mas variables independientes y formula modelos matemáticos con el fin de estimar resultados.
 - y la correlación permite medir el grado de asociación entre dos o mas variables
- 



¿Qué es regresión lineal ?

Es un modelo que permite describir cómo influye una variable X sobre otra variable Y .

- X : Variable independiente o explicativa o exógena.
- Y : Variable dependiente o respuesta o endógena.
- El objetivo es obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores $(x_1, y_1), \dots, (x_n, y_n)$.



Ejemplo:

- Estudiar cómo influye la estatura del padre sobre la estatura del hijo.
- Estimar el precio de una vivienda en función de su superficie.
- Aproximar la calificación obtenida en una materia según el número de horas de estudio semanal.
- Prever el tiempo de computación de un programa en función de la velocidad del procesador.

Modelo de regresión lineal

$$Y = \beta_0 + \beta_1 x + e,$$

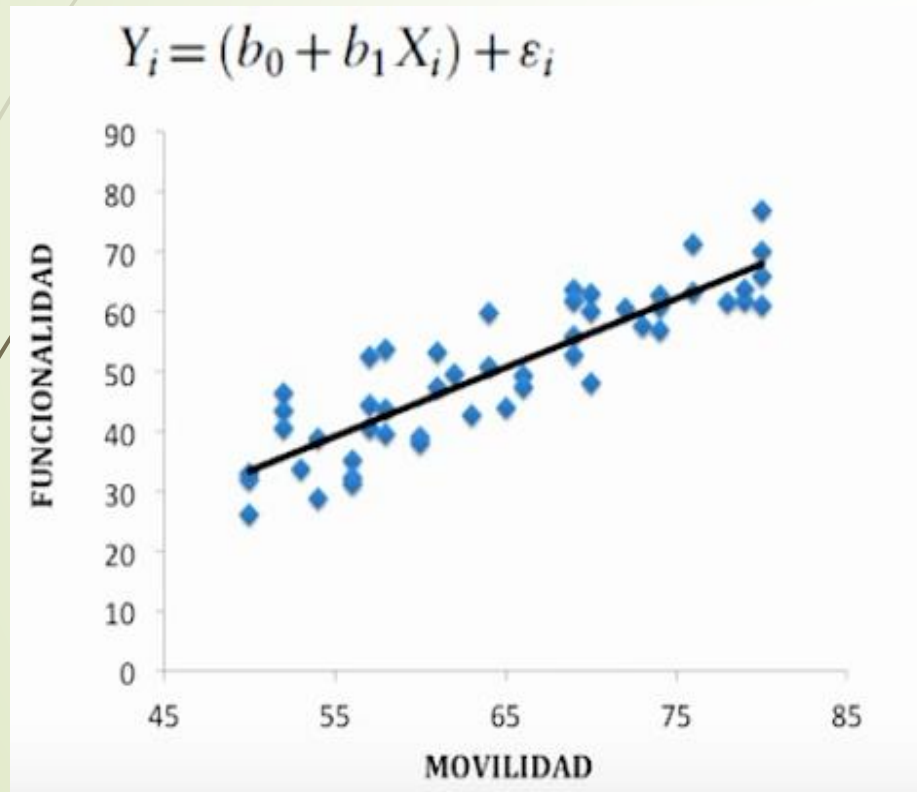
e es el error aleatorio

β_0 es la ordenada en el origen

β_1 es la pendiente

Permite predecir o estimar el comportamiento de la variable independiente.

Mínimos cuadrados



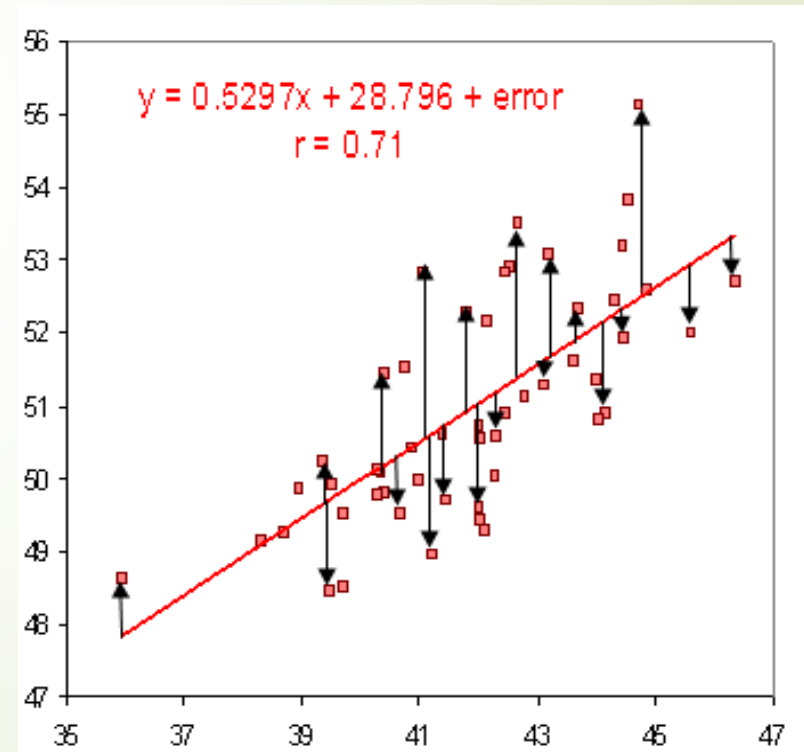
- Hay infinitas posibles rectas.
- Necesitamos encontrar aquella que minimice las distancias entre los valores observados y los predichos por la ecuación.
- Esta recta se obtiene mediante un proceso matemático conocido como mínimos cuadrados.

Se usa para buscar valores desconocidos usando como referencia otras muestras del mismo evento.

El método consiste en acercar una línea, según se escoja, lo más posible a los puntos determinados por la coordenadas $[x, f(x)]$, que normalmente corresponden a muestras de algún experimento

$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} :$$

$$b = \bar{y} - m\bar{x}$$



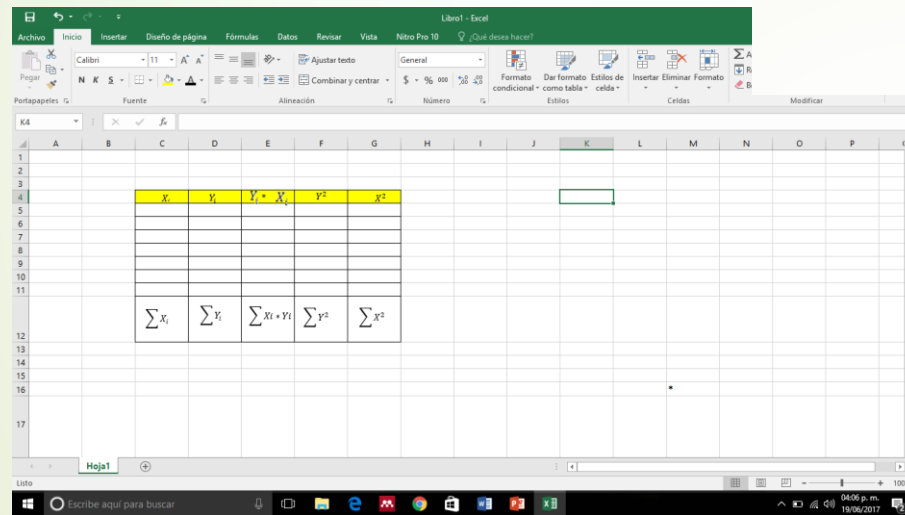
e es el error aleatorio

β_0 es el la ordenada en el origen

β_1 es la pendiente

$$Y = \beta_0 + \beta_1 x + e,$$

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$



The screenshot shows an Excel spreadsheet with a table for linear regression calculations. The table is located in the range B4 to G11. The columns are labeled X, Y, Y * X, Y^2, and X^2. The rows contain data points and summary statistics.

X	Y	Y * X	Y^2	X^2
$\sum X_i$	$\sum Y_i$	$\sum X_i \cdot Y_i$	$\sum Y_i^2$	$\sum X_i^2$

$$b = \bar{y} - \beta_1 \cdot \bar{x}$$

Para encontrar el coeficiente de correlación se debe encontrar el error estándar y la varianza de la variable dependiente

$$Se = \sqrt{\frac{\sum y^2 - a \sum y - b \sum x \cdot y}{n - 2}}$$

$$S_y^2 = \frac{\sum y^2}{n} - \bar{y}^2$$

$$R^2 = 1 - \frac{Se^2}{S_y^2}$$

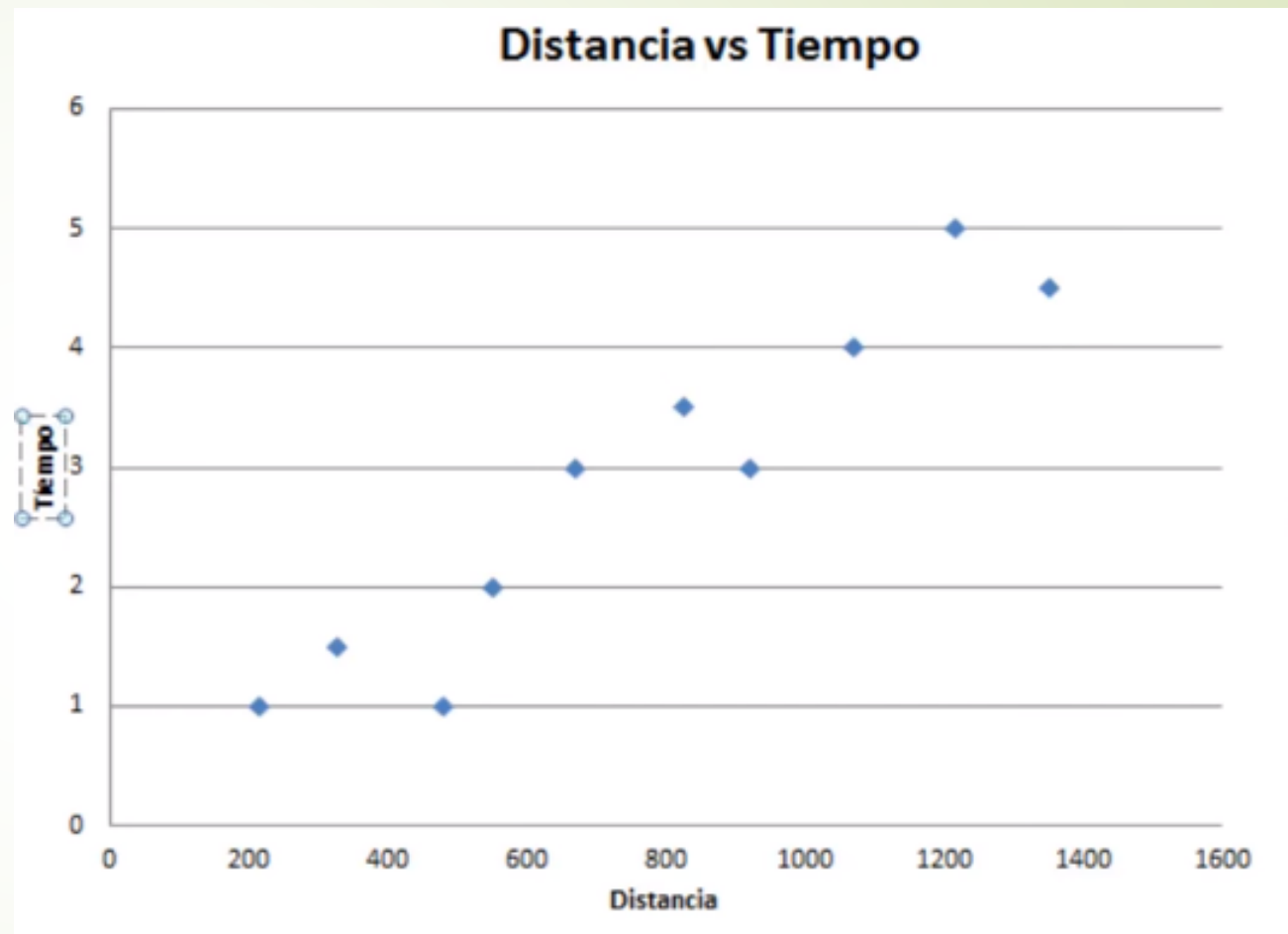
Ejemplo:

Una empresa de mensajería de entrega puerta a puerta, con el fin de mejorar la prestación del servicio desea establecer la relación que puede existir entre el tiempo empleado y la distancia recorrida para la entrega de un determinado producto.

- a) Realice un diagrama de dispersión a partir de los datos obtenidos.
- b) Determine la mejor ecuación que se ajuste a los datos.
- c) Pronostique el tiempo de entrega para un pedido que esta a una distancia de 1000 km

Distancia(Km)	825	215	1070	550	480	920	1350	325	670	1215
Tiempo (Dias)	3,5	1,0	4,0	2,0	1,0	3,0	4,5	1,5	3,0	5,0

x	y	x*y	x ²	y ²
825	3,5	2887,5	680625	12,25
215	1	215	46225	1
1070	4	4280	1144900	16
550	2	1100	302500	4
480	1	480	230400	1
920	3	2760	846400	9
1350	4,5	6075	1822500	20,25
325	1,5	487,5	105625	2,25
670	3	2010	448900	9
1215	5	6075	1476225	25
7620	28,5	26370,0	7104300	99,75



$$b = \frac{n \sum x \cdot y - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{10 \cdot (26370) - (7620) \cdot (28.6)}{10 \cdot (7104300) - (7620)^2}$$

$$b = \frac{263700 - 217170}{71043000 - 58064400}$$

$$b = \frac{46530}{12978600}$$

$$b = 3.585 \times 10^{-3}$$

$$b = 0.003585$$

$$a = \frac{\sum y - b \sum x}{n}$$

$$a = \frac{28.5 - (0.003585)(7620)}{10}$$

$$a = \frac{28.5 - 27.3177}{10}$$

$$a = \frac{1.1823}{10}$$

$$a = 0.11823$$

$$Y = a + bX$$

Reemplazando se tiene

$$Y = 0.0036 x + 0.11823$$

Error estándar del estimado

$$Se = \sqrt{\frac{\sum y^2 - a \sum y - b \sum x \cdot y}{n - 2}}$$

$$Se = \sqrt{\frac{99.75 - 0.11823 \cdot (28.5) - 0.0036 \cdot (26370)}{10 - 2}}$$

$$Se = \sqrt{\frac{99.75 - 3.369555 - 94.932}{8}}$$

$$Se = \sqrt{\frac{1.448445}{8}}$$

$$Se = \sqrt{0.18105525}$$

$$Se = 0.4255$$

$$S_y^2 = \frac{99.75}{10} - (2.85)^2 = 9.975 - 8.1225$$

$$S_y^2 = 1.8525$$

Ahora si el de correlación

$$R^2 = 1 - \frac{Se^2}{S_y^2}$$

$$Se = 0.4255$$

$$R^2 = 1 - \frac{(0.4255)^2}{1.8525} = 1 - \frac{0.1810}{1.8525} = 1 - 0.0977$$

$$R^2 = 0.9023$$

Ejercicio

- La empresa Epsa, analiza la relación entre el consumo de energía (en miles de kilowatts - hora ,kwh) y el número de habitaciones en una residencia privada unifamiliar. Una muestra aleatoria de 10 casas produjo lo siguiente:

# habitac. (X)	Consumo (Y)
12	9
9	7
14	10
6	5
10	8
8	6
10	8
10	10
5	4
7	7

A-) Determina la ecuación de regresión

B-) Evalué el consumo, en miles de kilowatts - hora para una casa de 6 habitaciones

Solución

X_i	Y_i	$Y_i * X_i$	Y^2	X^2
12	9	144	81	108
9	7	81	49	63
14	10	196	100	140
6	5	36	25	30
10	8	10	64	80
8	6	64	36	48
10	8	100	64	80
10	10	100	100	100
5	4	25	16	20
7	7	49	49	49
91	74	895	584	718

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{10(718) - 91(74)}{10(895) - (91)^2} = \frac{7180 - 6734}{8950 - 8281} = \frac{446}{669} = 0.667$$

$$a = \frac{\sum Y}{n} - b \frac{\sum X}{n}$$

$$a = \frac{74}{10} - (0.667) \frac{91}{10} = 7.4 - (0.667)(9.1) = 7.4 - 6.0697 = 1.330$$

$$Y' = a + bx$$

$$Y' = 1.330 + 0.667 x$$