ANÁLISIS NUMÉRICO

Mag. Carlos Alberto Ardila Albarracín

BLOQUE 2. AJUSTE DE CURVAS

2.1. CONCEPTO DE CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

DEFINICIÓN DE CORRELACIÓN

En ocasiones nos puede interesar estudiar si existe o no algún tipo de relación entre dos variables aleatorias:

Estudiar cómo influye la estatura del padre sobre la estatura del hijo.

Estudiar cómo influyen los gastos de promoción y publicidad en el volumen de facturación de una empresa.

Estimar el precio de una vivienda en función de su superficie.

DEFINICIÓN DE CORRELACIÓN

Un modelo de regresión 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

Objetivo: obtener estimaciones razonables de Y para distintos valores de X
a partir de una muestra de n pares de valores

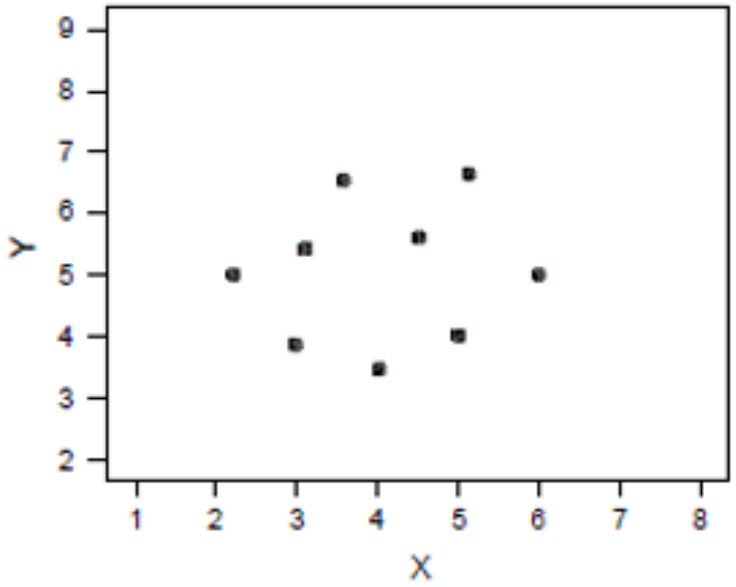
$$(x_1, y_1), \ldots, (x_n, y_n)$$

DEFINICIÓN DE CORRELACIÓN

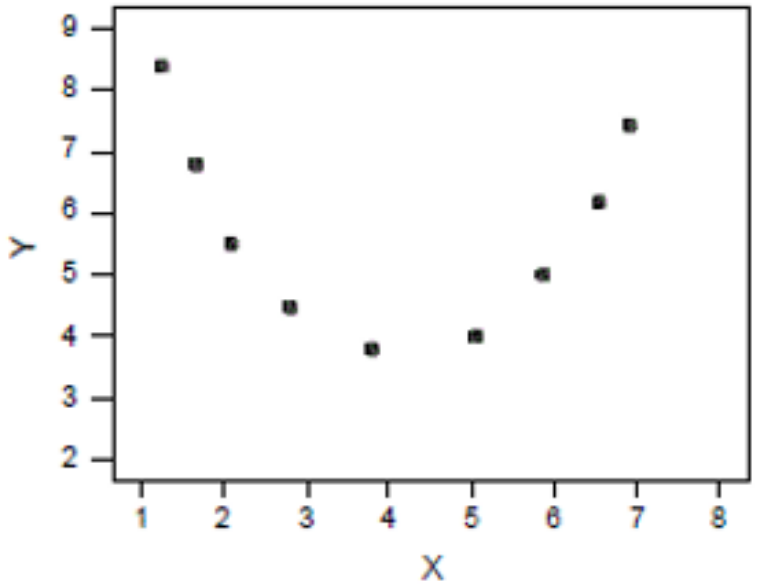
Interesa cuantificar la intensidad de la relación lineal entre dos variables.

El parámetro que nos da tal cuantificación es el coeficiente de correlación lineal de Pearson r, cuyo valor oscila entre -1 y +1.

VARIABLES NO CORRELACIONADAS (r = 0)



CORRELACIÓN NO LINEAL (r = 0)

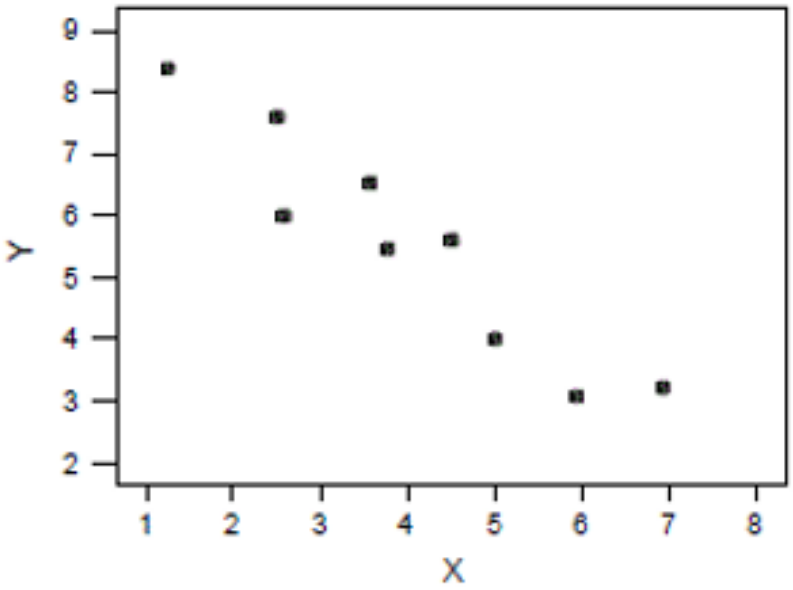


DEFINICIÓN DE CORRELACIÓN

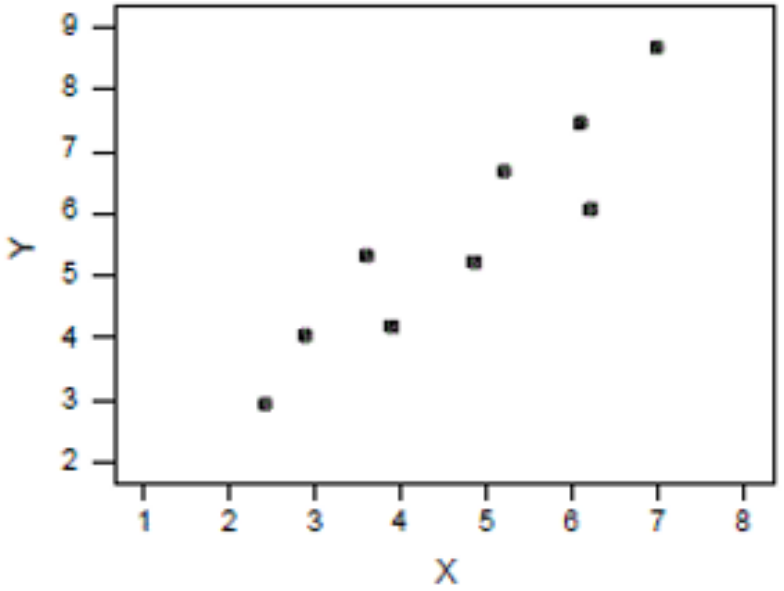
Interesa cuantificar la intensidad de la relación lineal entre dos variables.

El parámetro que nos da tal cuantificación es el coeficiente de correlación lineal de Pearson r, cuyo valor oscila entre -1 y +1.

CORRELACIÓN LINEAL NEGATIVA (r = -1)



CORRELACIÓN LINEAL POSITIVA (r = 1)



DEFINICIÓN DE CORRELACIÓN

Como se observa en los diagramas anteriores:

El valor de r se aproxima a +1
cuando la correlación tiende a ser lineal directa
(mayores valores de X significan mayores valores de Y)

El valor de r se aproxima a -1
cuando la correlación tiende a ser lineal inversa
(mayores valores de X significan MENORES valores de Y)

DEFINICIÓN DE CORRELACIÓN

¡Atención!
el que ocurra r = 0 sólo nos dice que no hay correlación lineal,
pero puede que la haya de otro tipo

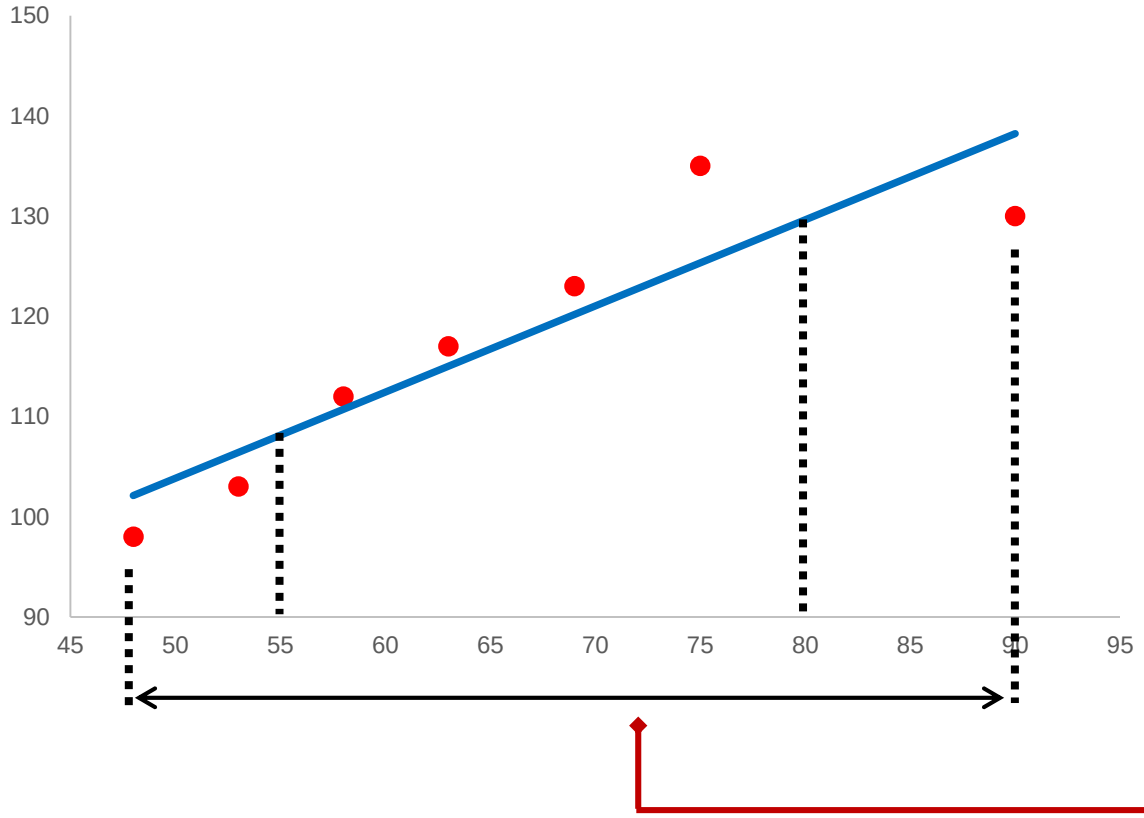
Es importante notar que:
La existencia de correlación entre variables
NO IMPLICA C A U S A L I D A D

Aquí lo máximo que podemos detectar
es si X influye en o explica Y (y cuantificar esa influencia),
pero NO se puede decir que X sea la CAUSA de Y

REGRESIÓN LINEAL SIMPLE

¿Qué buscamos?

La ecuación de la recta que "mejor se ajuste" (recta de mínimos cuadrados) a la nube de puntos (representada en el diagrama de dispersión):



Uno de los principales usos de dicha recta será

el de predecir o estimar los valores de **Y** que obtendríamos

para distintos valores de **X**

dentro del rango original

¡y diferentes a los representados en el diagrama!

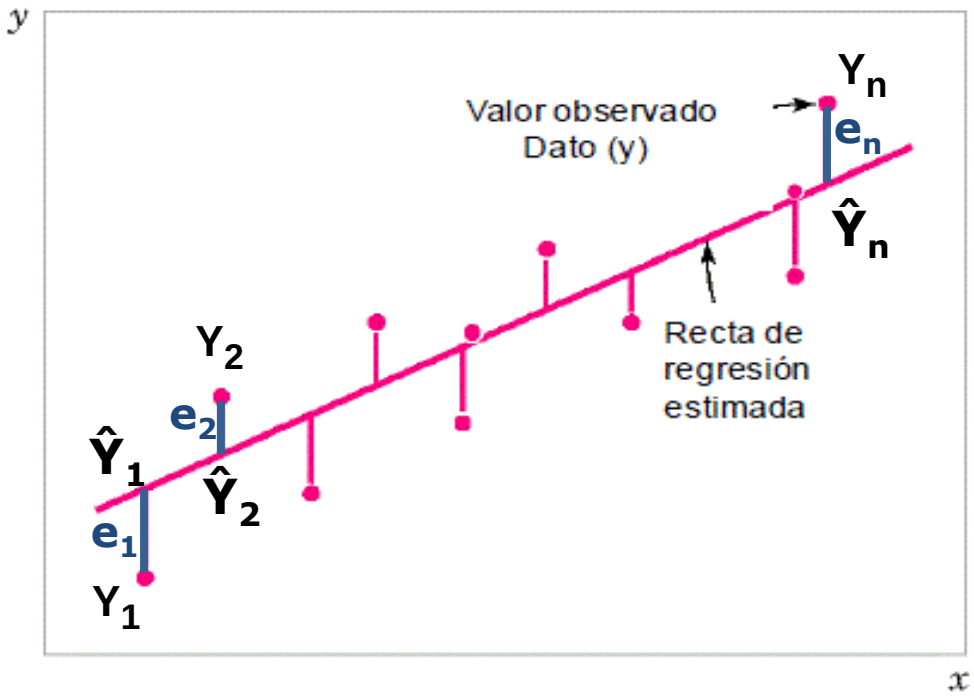
(ej: x=55, x=80)

REGRESIÓN LINEAL SIMPLE

La diferencia entre cada valor Y_i de la variable respuesta con su estimación $\hat{Y}_i$ se llama residuo:

$$e_i = y_i - \hat{y}_i$$

El modelo pretende dar el valor mínimo posible para cada e_i



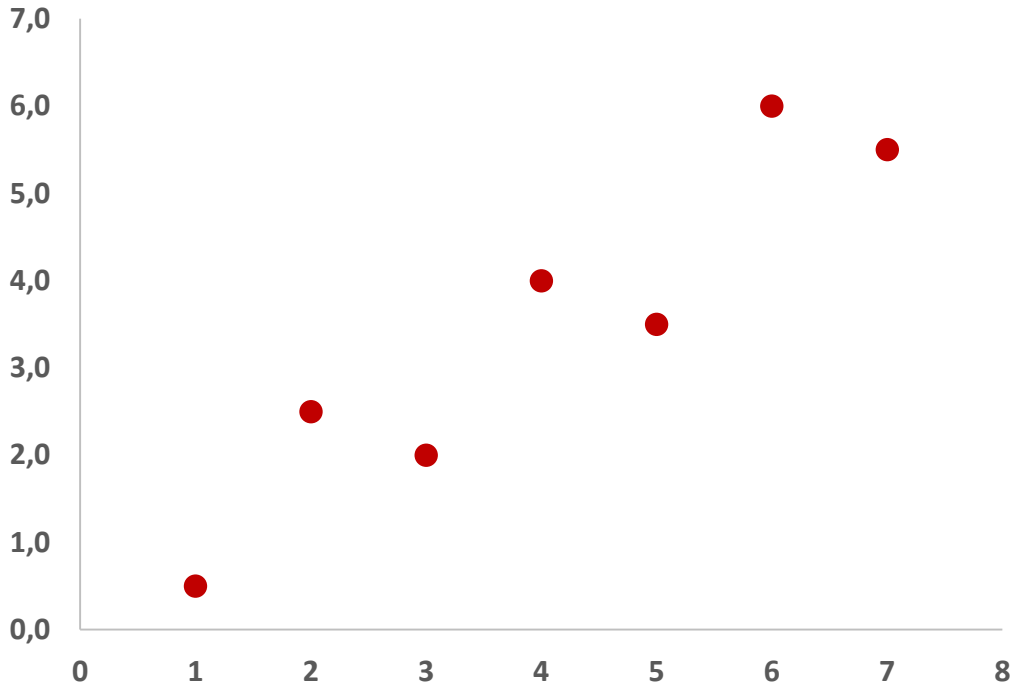
REGRESIÓN LINEAL SIMPLE

El procedimiento podemos verlo en el ejemplo siguiente:

Dado este conjunto de valores

X	Y
1	0,5
2	2,5
3	2,0
4	4,0
5	3,5
6	6,0
7	5,5

Generamos el diagrama de dispersión



REGRESIÓN LINEAL SIMPLE

El procedimiento podemos verlo en el ejemplo siguiente:

Dado este conjunto de valores

X	Y
1	0,5
2	2,5
3	2,0
4	4,0
5	3,5
6	6,0
7	5,5

Para estimar los coeficientes (DE LA RECTA DE REGRESIÓN)

$$Y = b1x + bo$$

por medio de mínimos cuadrados, se utilizan las siguientes fórmulas:

$$b_1 = \frac{\sum XY - \bar{y}\sum X}{\sum X^2 - \bar{x}\sum X}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

REGRESIÓN LINEAL SIMPLE

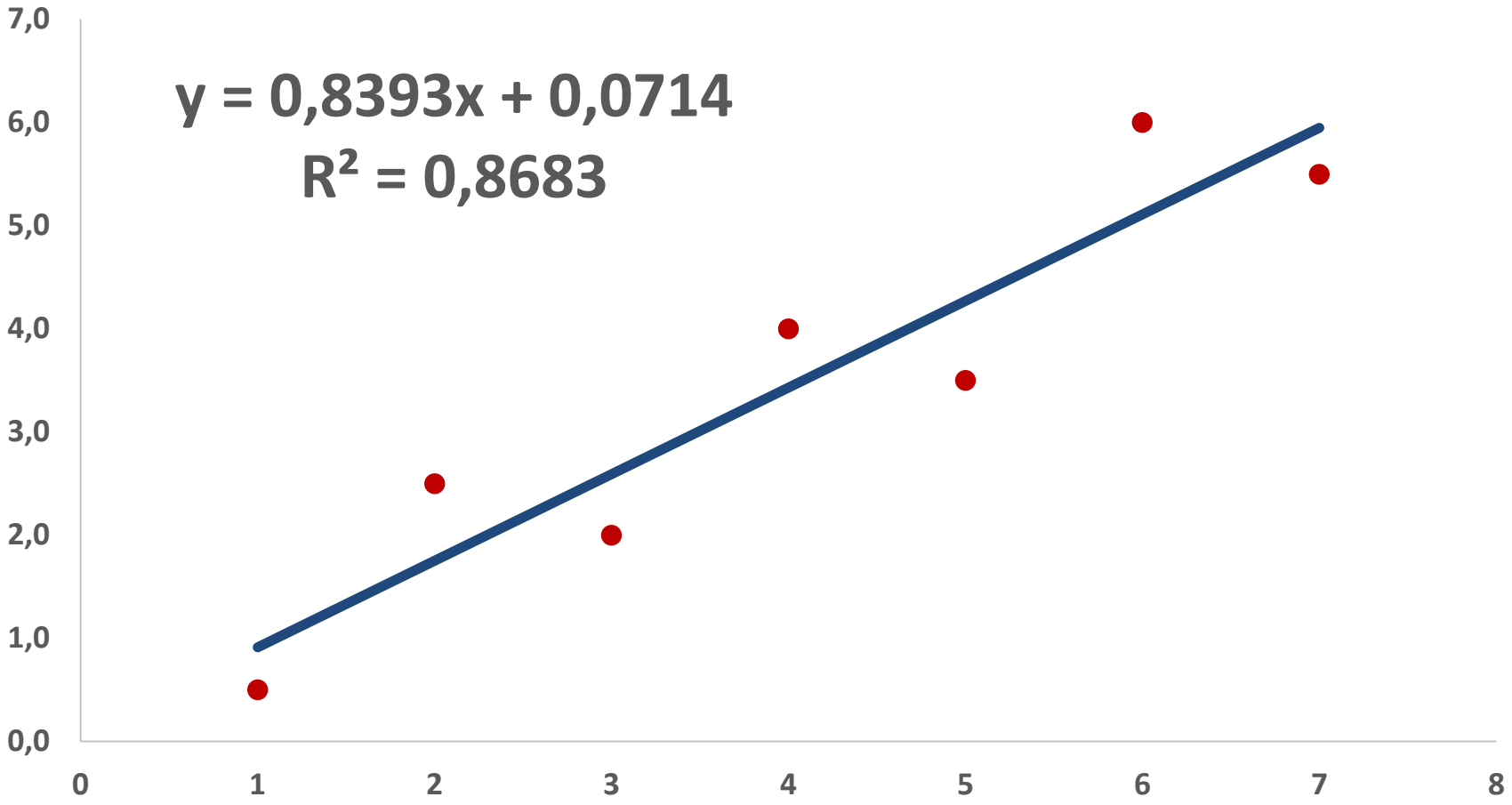
X	Y	X*Y	X2			
1	0,5	0,5	1,00		X trazo	4,0000
2	2,5	5,0	4,00		Y trazo	3,428571429
3	2,0	6,0	9,00		suma X*Y	119,5
4	4,0	16,0	16,00		suma x	28
5	3,5	17,5	25,00		suma x2	140
6	6,0	36,0	36,00			
7	5,5	38,5	49,00			

Y trazo * suma x	96,0000000000	
suma X*Y - (Y trazo * suma x)	23,5000000000	N
X trazo * suma x	112,0000000000	
suma X2 - (X trazo * suma x)	28,0000000000	D
B1	0,8392857143	
B1* X trazo	3,3571428571	
B0	0,0714285714	

Por lo tanto, el ajuste con mínimos cuadrados (la ecuación de la recta) es :

$$y = 0.83929x + 0.07143$$

REGRESIÓN LINEAL SIMPLE



REGRESIÓN LINEAL SIMPLE

Para saber si el modelo es adecuado,
es necesario cuantificar el error en la regresión lineal

Lo primero que debemos calcular
es la Desviación Estándar Total

$$S_y = \sqrt{\frac{S_t}{n - 1}}$$

Donde:

n es la cantidad de puntos (los de la tabla)

S_t es la suma total de los cuadrados de las restas entre cada uno de los valores medidos y la media

$$S_t = \sum (y_i - \overline{y})^2$$

REGRESIÓN LINEAL SIMPLE

Para saber si el modelo es adecuado,
es necesario cuantificar el error en la regresión lineal

X	Y
1	0,5
2	2,5
3	2,0
4	4,0
5	3,5
6	6,0
7	5,5

Y trazo
3,428571429

	(Yi - Y trazo)2
	8,576530612
	0,862244898
	2,040816327
	0,326530612
	0,005102041
	6,612244898
	4,290816327
ST	22,714285714

$$S_y = \sqrt{\frac{22.714285714}{7 - 1}} = 1.94567$$

Tenemos el primer parámetro: Desviación Estándar Total

REGRESIÓN LINEAL SIMPLE

Para saber si el modelo es adecuado,
es necesario cuantificar el error en la regresión lineal

Lo segundo que debemos calcular es el Error Estándar de Aproximación

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

Donde:

n **es la cantidad de puntos (los de la tabla)**

Sr **es la discrepancia entre el valor verdadero de "y" con el valor aproximado que predice la ecuación lineal.**

$$S_r = (y_i - b_o - (b_1 * x_i))^2$$

REGRESIÓN LINEAL SIMPLE

Para saber si el modelo es adecuado,
es necesario cuantificar el error en la regresión lineal

X	Y
1	0,5
2	2,5
3	2,0
4	4,0
5	3,5
6	6,0
7	5,5

Y trazo
3,428571429

B1 Bo

$y = 0.83929x + 0.07143$

	(Yi - B0 - B1Xi)2
	0,168686224
	0,5625
	0,347257653
	0,326530612
	0,589604592
	0,797193878
	0,199298469
SR	2,991071429

$$S_{y/x} = \sqrt{\frac{2.991071429}{7 - 2}} = 0.77344$$

Tenemos el segundo parámetro:
Error Estándar de Aproximación

REGRESIÓN LINEAL SIMPLE

Para saber si el modelo es adecuado,
es necesario cuantificar el error en la regresión lineal

Los dos parámetros previamente calculados se utilizan para determinar si la aproximación se considera aceptable o no

El criterio es:

Si $(S_{y/x} < S_y)$ entonces la aproximación se considera aceptable

En nuestro ejemplo:

$S_y = 1.94567$

$S_{y/x} = 0.77344$

Como se cumple el criterio

$S_{y/x}$ S_y

$(0.77344 < 1.94567)$

entonces la aproximación se considera aceptable

REGRESIÓN LINEAL SIMPLE

Finalmente:

St → suma total de los cuadrados de las restas entre cada uno de los puntos y la media

Sr → suma de los cuadrados de las restas alrededor de la línea de regresión

La diferencia entre esas 2 cantidades (St – Sr) cuantifica la mejora en la reducción del error debido al modelo de la línea recta.

Esta diferencia se puede normalizar al error total y obtener:

$$r^2 = \frac{St - Sr}{St}$$

Donde:

r^2 es el coeficiente de determinación

REGRESIÓN LINEAL SIMPLE

Para un ajuste perfecto, Sr = 0 y $r^2 = 1$

indicando que la línea recta explica el 100% de la variabilidad.

Para el ejemplo que venimos tratando:

$$r^2 = \frac{22.714285714 - 2.991071429}{22.714285714} = 0.86831761$$

El resultado indica que el modelo lineal explica el

86.83%

de la incertidumbre original

REGRESIÓN LINEAL SIMPLE

¡A tener en cuenta!

Antes de elaborar un modelo de regresión lineal primero debemos revisar que se cumplen estas dos condiciones mínimas:

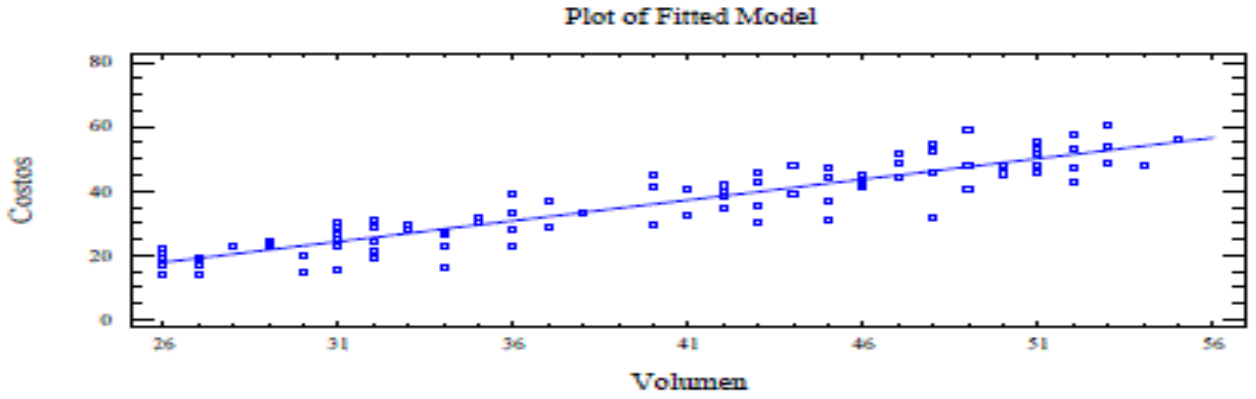
Linealidad: La relación entre X e Y es (o tiende a ser) lineal.

Homocedasticidad: La varianza de los errores es constante.

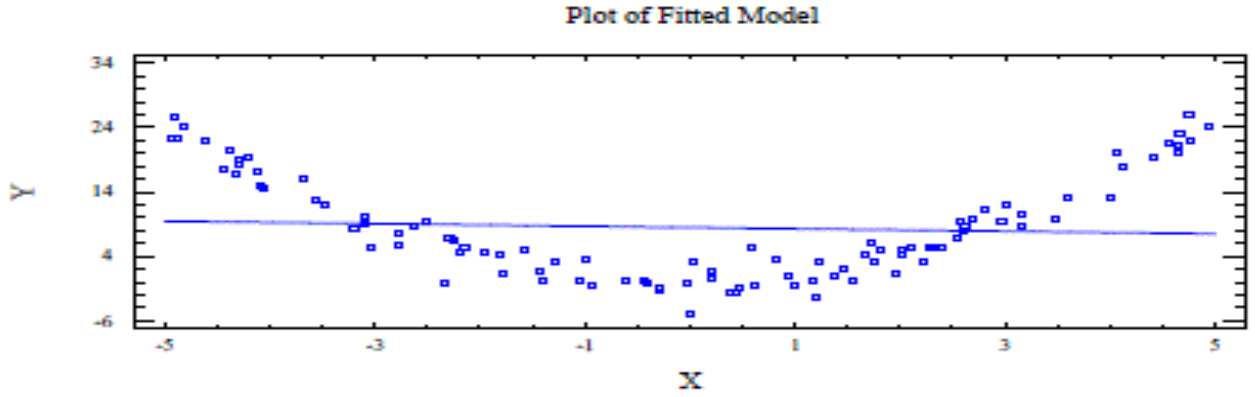
REGRESIÓN LINEAL SIMPLE

Linealidad

Los datos deben ser razonablemante rectos.



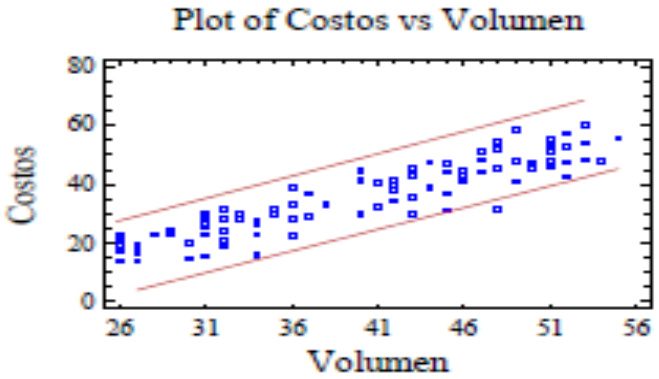
Si no, la recta de regresión no representa la estructura de los datos.



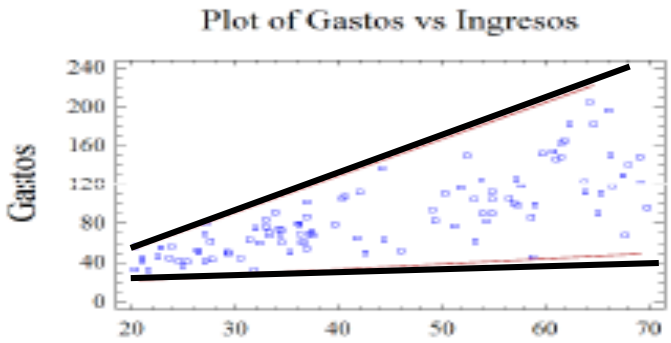
REGRESIÓN LINEAL SIMPLE

Homocedasticidad

La dispersión de los datos debe ser constante para que los datos sean homocedásticos.



Si no se cumple, los datos son heterocedásticos.



---------- *FIN DEL DOCUMENTO*