UNIVERSIDAD NACIONAL EXPERIMENTAL DE GUAYANA
VICERRECTORADO ACADÉMICO
COORDINACIÓN GENERAL DE PREGRADO
ASIGNATURA: PROBABILIDAD Y ESTADÍSTICA

Correlación lineal

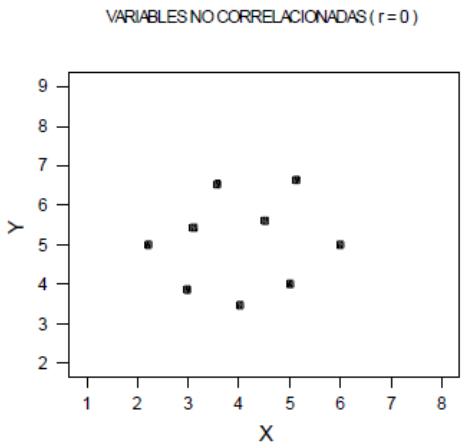
En ocasiones nos puede interesar estudiar si existe o no algún tipo de relación entre dos variables aleatorias. Así, por ejemplo, podemos preguntarnos si hay alguna relación entre las notas de la asignatura Estadística I y las de Matemáticas I. parámetro que nos da tal cuantificación es el **coeficiente de correlación lineal de Pearson r**, cuyo valor oscila entre –1 y +1:

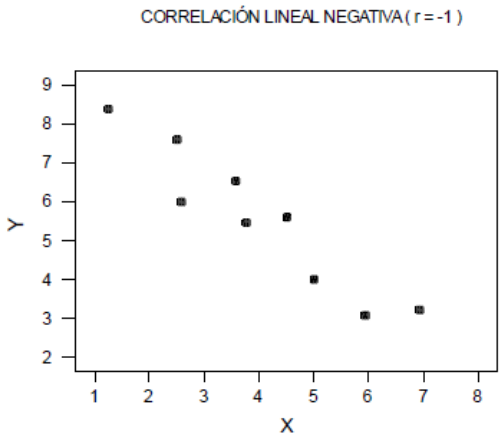
$$Cr = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

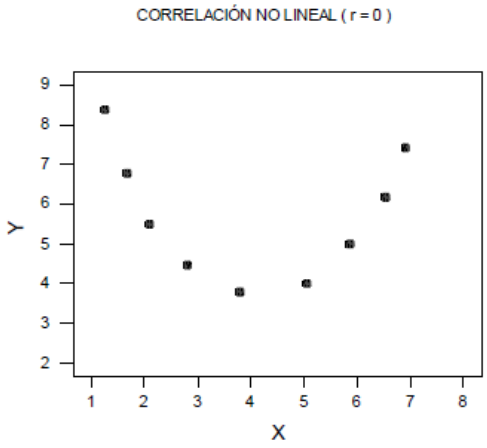
n: cantidad de pares de datos
y: variable dependiente
x: variable independiente

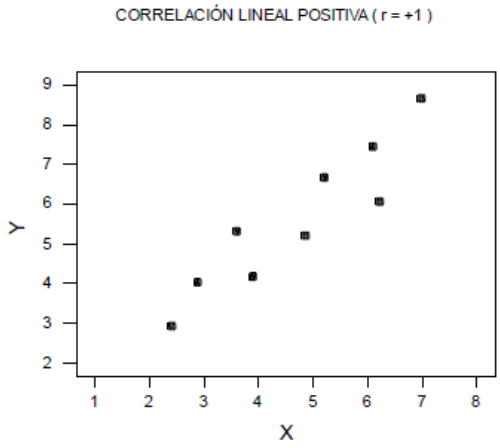
Diagrama de dispersión

Es la representación de una nube de puntos en el plano cartesiano, la cual podría indicarnos visualmente la existencia o no de algún tipo de relación (lineal, parabólica, exponencial, etc.) entre ambas variables.



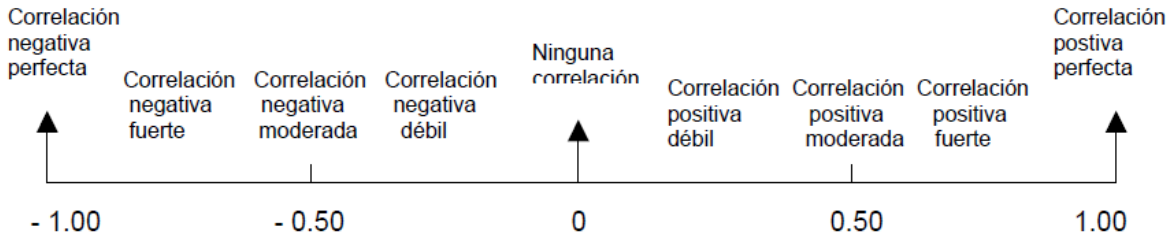






Como se observa en los diagramas anteriores, el valor de *r* se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de *X* significan mayores valores de *Y*), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa.
Es importante notar que la existencia de correlación entre variables no implica causalidad. ¡Atención!: si no hay correlación de ningún tipo entre dos variables aleatorias., entonces tampoco habrá correlación lineal, por lo que r = 0. Sin embargo, el que ocurra r = 0 sólo nos dice que no hay correlación lineal, pero puede que la haya de otro tipo.

El siguiente diagrama resume el análisis del coeficiente de correlación entre dos variables:



Coeficiente de determinación

Denominamos **coeficiente de determinación R2** como el coeficiente que nos indica el porcentaje del ajuste que se ha conseguido con el modelo lineal, es decir el porcentaje de la variación de Y que se explica a través del modelo lineal que se ha estimado, es decir a través del comportamiento de X. A mayor porcentaje mejor es nuestro modelo para predecir el comportamiento de la variable Y.
También se puede entender este coeficiente de determinación como el porcentaje de varianza explicada por la recta de regresión y su valor siempre estará entre 0 y 1 y siempre es igual al cuadrado del coeficiente de correlación (Cr).
$R^2 = (Cr)^2 \times 100$

Regresion lineal

El **análisis de regresión** encuentra la ecuación de la recta que mejor describe la relación entre las dos variables. Un uso de esta ecuación es hacer predicciones. Hacemos uso de estas predicciones regularmente, por ejemplo, predecimos el éxito que un estudiante tendrá en la universidad con base en sus resultados de preparatoria y predecir la distancia necesaria para detener un auto con base en su velocidad. En general, el valor exacto de *y* no se puede pronosticar y nos satisface saber que las predicciones son razonablemente cercanas.
La relación entre dos variables será una expresión algebraica que describe la relación matemática entre *x* y *y*.

$$y = \alpha + \beta x$$

$$\beta = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Alfa" "α se obtiene de la siguiente manera: el promedio de "y" menos el producto de beta "β"por el promedio de "x".

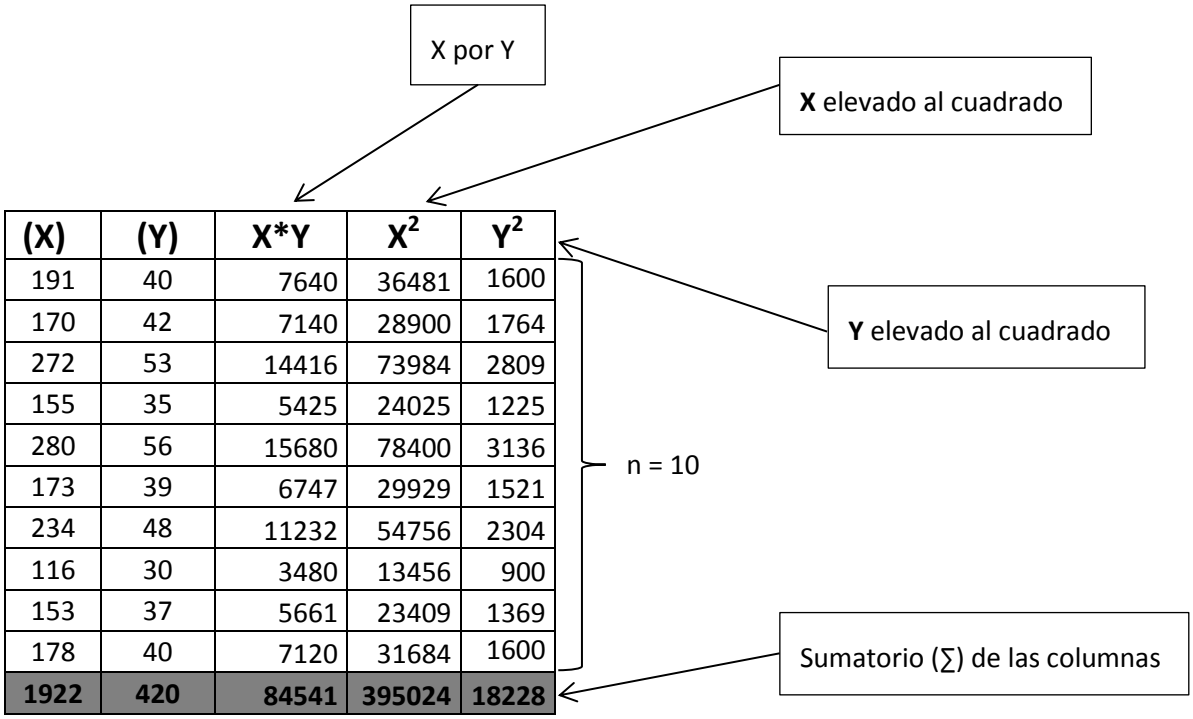
$$\alpha = \overline{y} - \beta\overline{x}$$

Ejemplo:

Unidades(y)	40	42	53	35	56	39	48	30	37	40
Gastos(x)	191	170	272	155	280	173	234	116	153	178

Calcule:

a) La correlación lineal
b) El coeficiente de determinación
c) La recta del mejor ajuste
d) Diagrama de dispersión



(X)	(Y)	X*Y	X^2	Y^2
191	40	7640	36481	1600
170	42	7140	28900	1764
272	53	14416	73984	2809
155	35	5425	24025	1225
280	56	15680	78400	3136
173	39	6747	29929	1521
234	48	11232	54756	2304
116	30	3480	13456	900
153	37	5661	23409	1369
178	40	7120	31684	1600
1922	**420**	**84541**	**395024**	**18228**

Se sustituyen los valores en la formula.

Correlación lineal

$$Cr = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

$$Cr = \frac{10(84541) - (1922)(\sum 420)}{\sqrt{10(395024) - (1922)^2}\sqrt{10(18228) - 420^2}} = 0{,}98$$

Coeficiente de determinación
$R^2 = (Cr)^2$ x 100
$R^2 = (0{,}98)^2$ x 100 = 0,96%

Recta del mejor ajuste:
Se calcula α y β

$$\beta = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$\beta = \frac{10(84541) - (1922)(420)}{10(395024) - (1922)^2} = 0{,}15$$

Promedio de “Y”

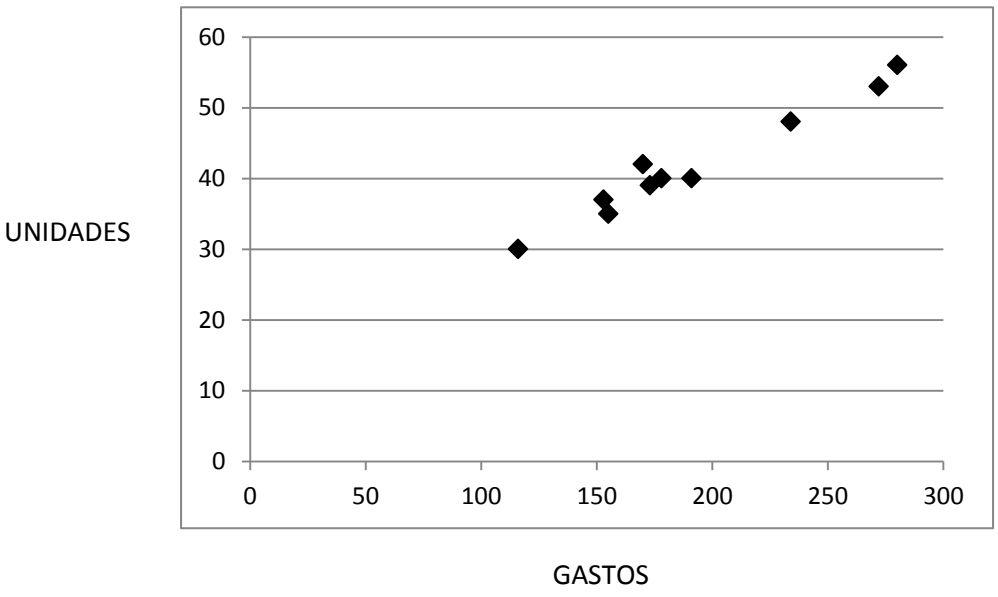
$$\alpha = \overline{y} - \beta\overline{x}$$ ← Promedio de “x”

α = (420/10) – 0,15(1922/10) = 13,17

Recta del mejor ajuste = $y = \alpha + \beta x$

Y = 13,17 + 0,15x

DIAGRAMA DE DISPERSIÓN



Resolver (20 ptos.):

Se realizó un estudio estadístico a unos estudiantes sobre el tiempo que le dedican a los estudios y su puntuación. Los resultados se muestran en la tabla:

HORAS	12	9	17	10	14	18	15	13	15	11
CALIFICACIONES	50	38	68	65	46	88	79	35	48	41

Calcule:

a) La correlación lineal (7 ptos.)
b) El coeficiente de determinación (2 ptos.)
c) La recta del mejor ajuste (8 ptos.)
d) Diagrama de dispersión (3 ptos.)