Modelo de Regresión Lineal

PROGRAMA MATEMÁTICA DUOC UC

Estadística Descriptiva

Al trabajar con dos variables, naturalmente surge la pregunta acerca de si hay, y cuál es, la relación entre ellas. Por ejemplo, cuando aumentan las temperaturas máximas, las ventas de bebidas también tienden a aumentar. ¿Es posible cuantificar esta relación? ¿Definirla de manera más precisa?

Hay distintos modelos que podemos usar, pero uno de los más comunes es el modelo de regresión lineal.

Modelos de Regresión

Un modelo de regresión es una función Y que intenta predecir el valor estimado de una variable respuesta, en función de uno o más predictores de interés, es decir, nos permite predecir el valor de Y, dados valores de la variable X.

En tal caso, diremos que:

Y = Variable dependiente.

X = Variable independiente, explicativa o predictora.

Tipos de modelos

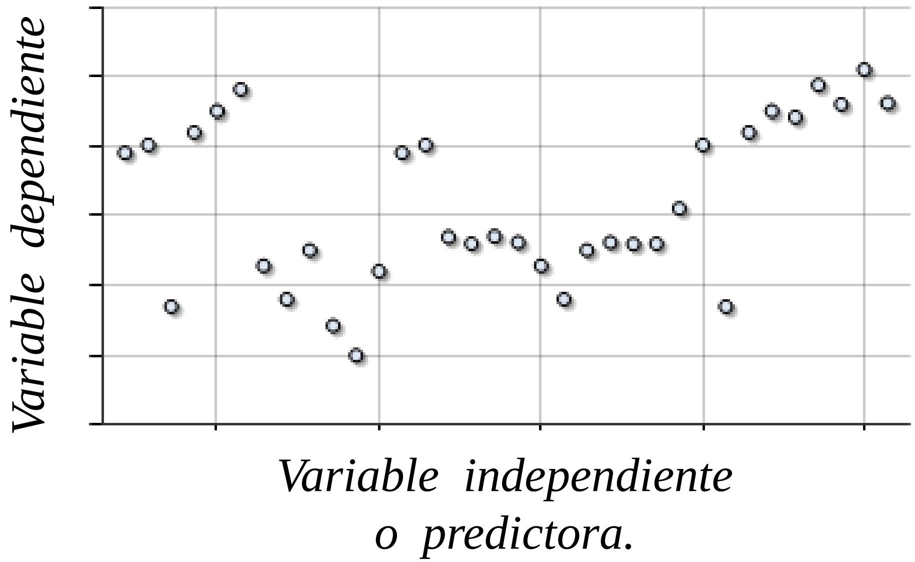
Existen diferentes tipos de modelos de regresión. Entre los modelos de regresión más conocidos están los siguientes:

- Modelo de regresión lineal.
- Modelo de regresión exponencial.
- Modelo de regresión logarítmica.

El modelo que convenga usar dependerá de la forma en que se relacionen las variables.

Diagrama de dispersión

Es un diagrama en que los puntos representan valores de dos variables, una representada en el eje X y la otra en el eje Y. Su propósito es ayudar a ver la existencia y tipo de relación que puede existir entre las dos variables representadas.



Es importante conocer y distinguir distintos modelos funcionales, como por ejemplo:

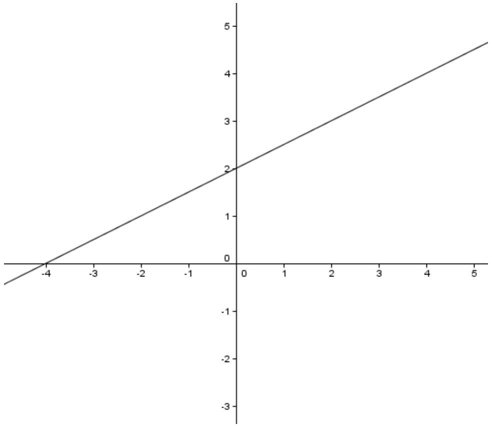


Gráfico de una función lineal

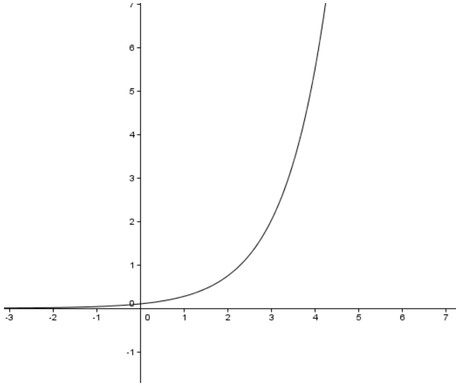


Gráfico de una función exponencial

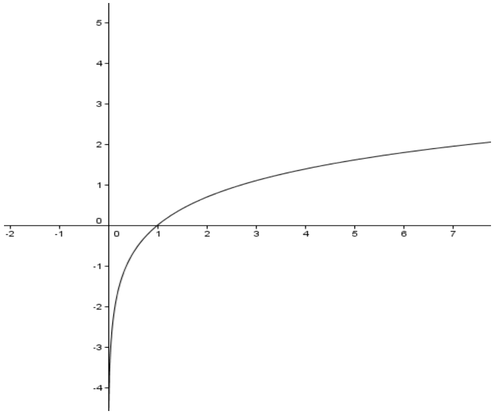


Gráfico de una función logarítmica

Medidas de Asociación entre dos variables

Al estudiar la relación entre dos variables debemos ser capaces de responder las siguientes preguntas:

- ¿Las variables están relacionadas o no?
- ¿Es lineal esta relación y cuál es su grado?

Las preguntas anteriores pueden ser respondidas utilizando las medidas de Covarianza y Coeficiente de Correlación Lineal de Pearson.

Covarianza

La **covarianza** (S_{xy}), nos indica si la posible relación entre dos variables es directa o inversa. Por lo que:

- Si $S_{xy} > 0$; La relación entre la variable X e Y es **directa** o **creciente**, es decir, si una de las variables aumenta la otra también aumenta y viceversa.

- Si $S_{xy} < 0$; La relación entre la variable X e Y es **inversa** o **decreciente**, es decir, si una de las variables aumenta la otra disminuye y viceversa.

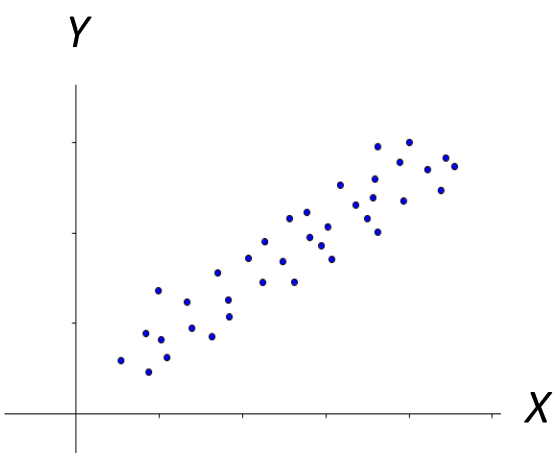
El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o decreciente, pero **no** nos da mayor información con respecto al grado de la relación.

Covarianza

Fórmula en Excel de la covarianza: **=COVAR(matriz1; matriz2)**

Ejemplos de representaciones gráficas de la covarianza.

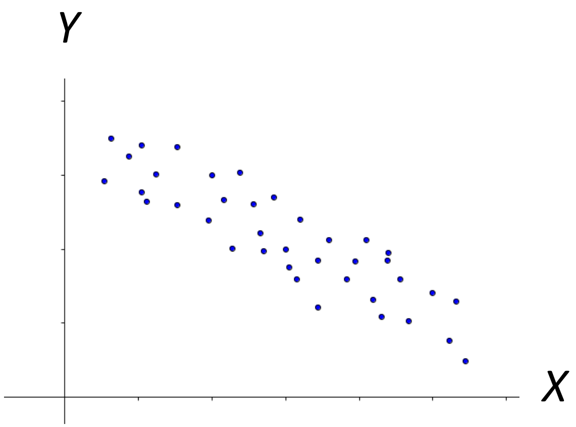
Ejemplo 1



En este caso la covarianza será positiva.

Cuando aumenta X, también tiende a aumentar Y.

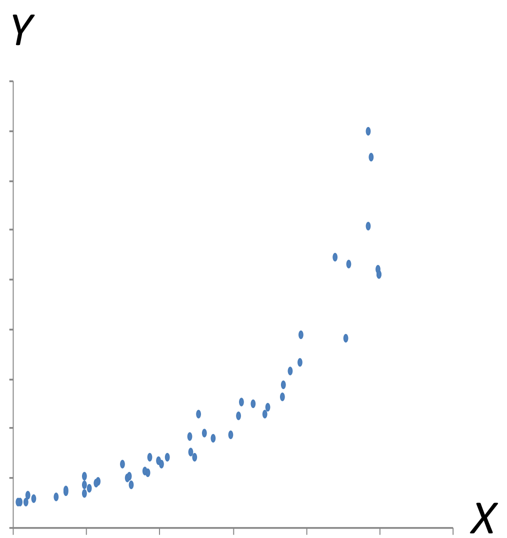
Ejemplo 2



En este caso la covarianza será negativa.

Cuando aumenta X, Y tiende a disminuir.

Ejemplo 3



En este caso la covarianza será positiva.

Cuando aumenta X, Y tiende a aumentar

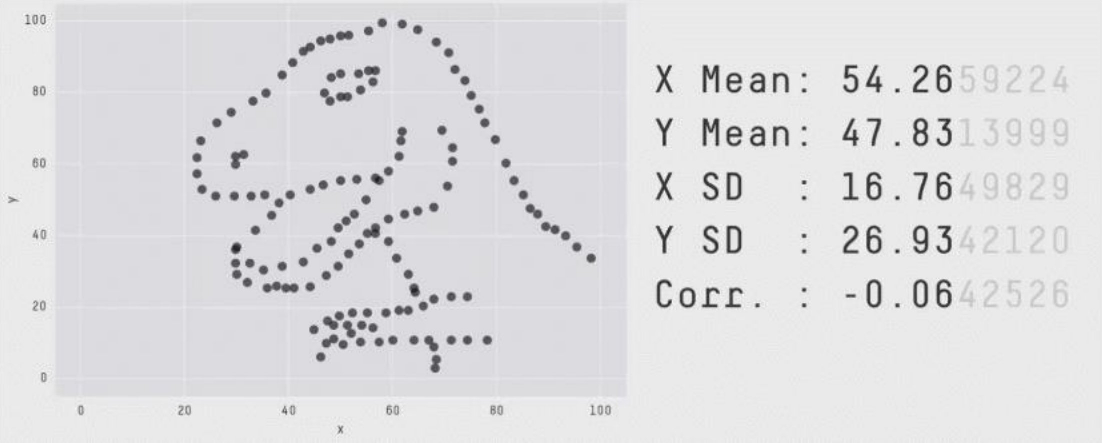
Coeficiente de correlación lineal de Pearson R_{xy}

El **Coeficiente de Correlación Lineal** entre dos variables (R_{xy}) muestra la linealidad de la relación entre las variables X e Y. En otras palabras, trata de responder la pregunta "¿Puedo representar a los datos mediante una línea recta?".

- Si $R_{xy} \approx 1$; La relación entre la variable X e Y es **lineal con pendiente positiva**.
- Si $R_{xy} \approx -1$; La relación entre la variable X e Y es **lineal con pendiente negativa**.
- Si $R_{xy} \approx 0$; No existe relación lineal entre la variable X e Y, aunque podría existir otro tipo de correlación (exponencial, logarítmica, etc.).

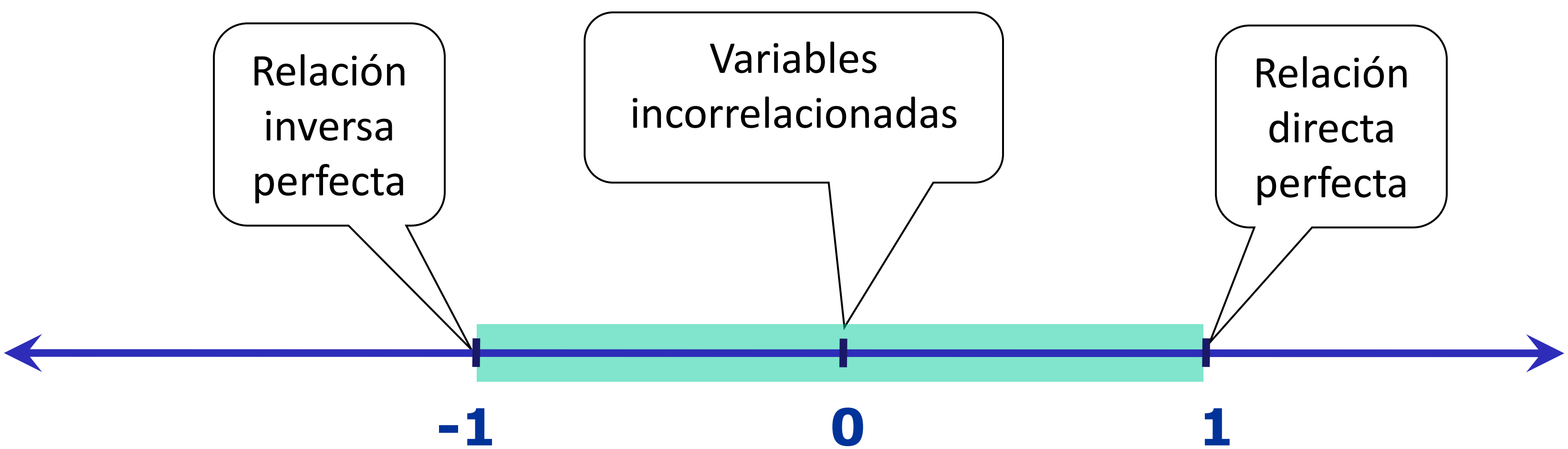
Coeficiente de correlación lineal de Pearson R_{xy}

Es importante usar este coeficiente junto con el diagrama de dispersión para tener una idea fidedigna de lo que ocurre con los datos. Por ejemplo, todos los conjuntos de datos tienen los mismos promedios, desviaciones estándar y correlación, pero claramente son muy distintos. Además, la baja correlación indica que no hay una relación lineal, pero visiblemente hay otras relaciones.



Coeficiente de correlación lineal de Pearson R_{xy}

- Es adimensional (no posee unidades de medida).
- Sólo toma valores en $[-1, 1]$.
- No hay correlación lineal si y sólo si R=0.
- Relación lineal perfecta entre dos variables ⇔ R = 1 o R = -1.
- Cuanto más cerca esté R de 1 o -1, mayor será el grado de relación lineal.



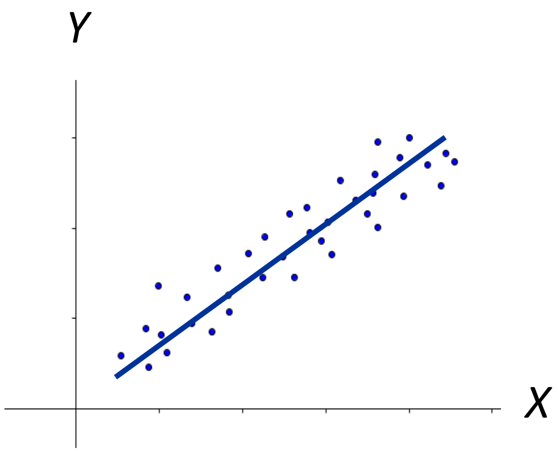
Coeficiente de correlación lineal de Pearson R_{xy}

Fórmula en Excel del Coeficiente de Correlación Lineal:

=COEF.DE.CORREL(matriz1; matriz2)

Ejemplos de representaciones gráficas del R_{xy}.

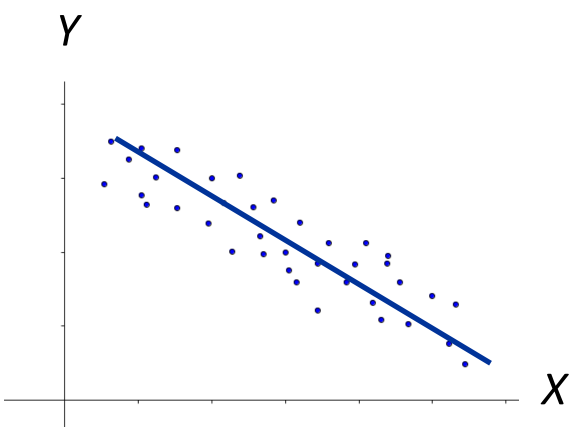
Ejemplo 1



En este caso el $R_{xy} \approx 1$.

La nube de puntos puede aproximarse por una recta creciente. Por ejemplo, la altura (X) y el peso (Y) de los alumnos, ya que los alumnos más altos suelen pesar más.

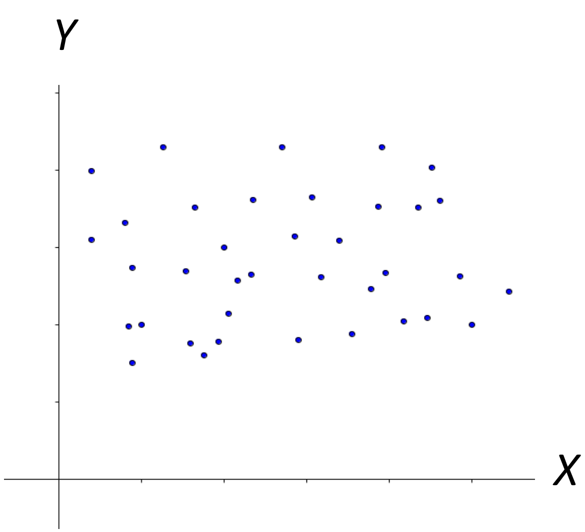
Ejemplo 2



En este caso $R_{xy} \approx -1$.

La nube de puntos puede aproximarse por una recta decreciente. Por ejemplo, el peso de una carga (X) y el rendimiento de combustible (Y) de un camión, ya que un camión con mayor carga tiene menor rendimiento de combustible.

Ejemplo 3



En este caso el $R_{xy} \approx 0$.

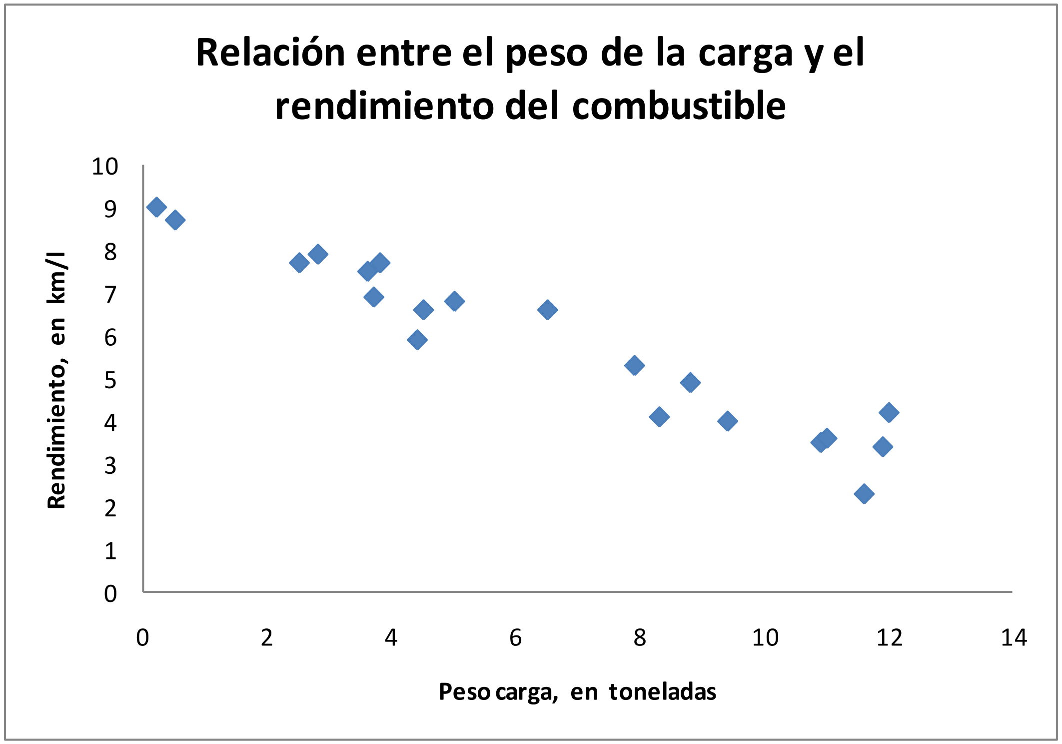
No hay correlación lineal.

Ejercicio:

El dueño de una flota de camiones desea estudiar si hay relación lineal entre el peso de la carga y el rendimiento del combustible, para ello consideró una muestra de 20 viajes, de los cuales se registró el peso de la carga (en toneladas) y el rendimiento del combustible (en km/litro). Utilizando la información, responda las siguientes preguntas:

Peso de la carga	Rendimiento combustible
7,9	5,3
0,2	9
6,5	6,6
2,5	7,7
8,8	4,9
4,5	6,6
12	4,2
5	6,8
11,9	3,4
0,5	8,7
2,8	7,9
3,7	6,9
8,3	4,1
4,4	5,9
3,8	7,7
9,4	4
3,6	7,5
11,6	2,3
10,9	3,5
11	3,6

1) Determine la variable dependiente e independiente, y construya un gráfico de dispersión que relacione las variables peso de la carga y el rendimiento del combustible.



Y = Rendimiento del combustible, en km/l (variable dependiente).

X = Peso de la carga, en tn (variable independiente).

2) Calcule la covarianza e interprete su resultado.

Fórmula	**Valor**	**Interpretación**
=COVAR(matriz1; matriz2)	**-7,00195**	La covarianza indica que hay una relación o influencia inversa entre las variables, por lo que a mayor peso de carga, menor será el rendimiento del combustible.

3) Calcule el coeficiente de correlación e interprete su resultado.

Fórmula	**Valor**	**Interpretación**
=COEF.DE.CORREL(matriz1; matriz2)	**-0,9633**	El grado de correlación lineal entre las variables es de -0,96 e indica que hay alta correlación lineal con pendiente negativa, entre las variables peso de la carga y rendimiento del combustible.

Coeficiente de determinación

El **coeficiente de determinación** (R^2) indica en qué proporción o porcentaje se puede explicar la variación de la variable Y a partir de la variación de la variable X.

Ejemplo

Se establece una correlación entre las variables cantidad de agua caída (mm) y número de eventos ambientales (alertas o preemergencias). Se obtiene un coeficiente de determinación de 75,6%. Interprete el coeficiente de determinación.

Interpretación

El 75,6% de la variabilidad del N° de eventos (alertas ambientales), es explicado por la variación de la cantidad de agua caída en un año. Además hay un 24,4% de variabilidad no explicada.

Fórmula en Excel del coeficiente de determinación lineal:

=COEFICIENTE.R2(conocido_y; conocido_x)

Modelo de regresión lineal simple

El modelo de regresión lineal simple consiste en obtener una aproximación a los valores de la variable dependiente (Y) a partir de la variable independiente (X). Esta aproximación se realiza encontrando una función lineal

$$\hat{Y} = b \cdot X + a$$

donde a y b son los parámetros del modelo lineal.

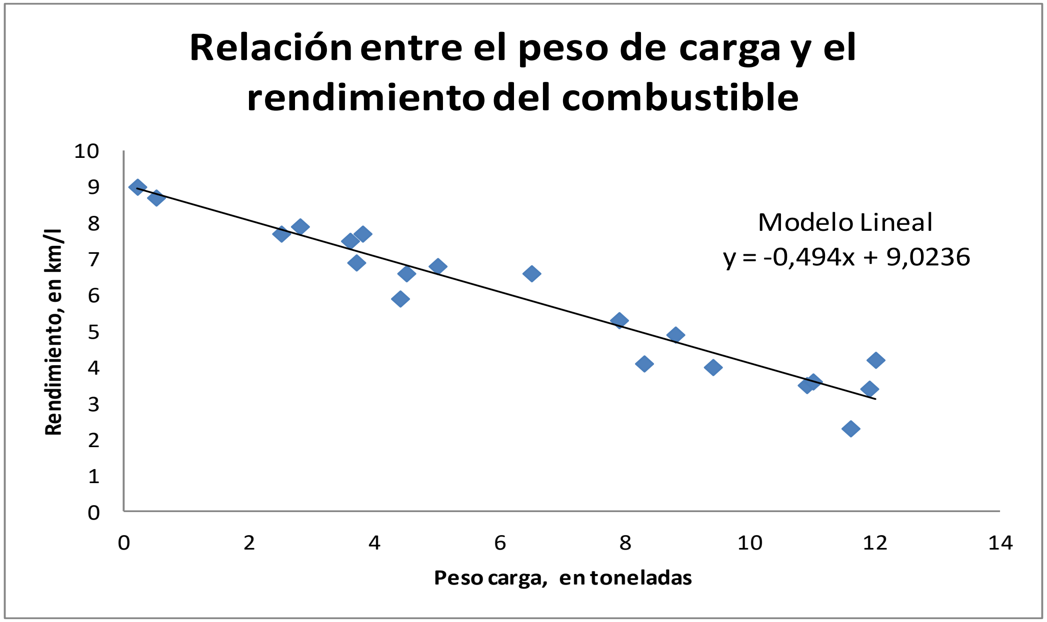
a = Coeficiente de posición.
b = Pendiente.

mientras que $\hat{Y}$ y X son
$\hat{Y}$ = Variable dependiente.
X = Variable independiente, explicativa o predictora.

Ejemplo

Considerando el ejercicio anterior cuyas variables eran el peso de la carga (en toneladas) y el rendimiento del combustible (en km/litro), de una muestra de 20 viajes.

1) Construya el modelo de regresión lineal.



Modelo de Regresión Lineal	$y = -0{,}494x + 9{,}0236$

2) Interprete la pendiente del modelo de regresión lineal.

	Fórmula	Valor	Interpretación
Pendiente	**=PENDIENTE(conocido_y; conocido_x)**	**-0,494**	Por cada tonelada que aumenta el peso de la carga, el rendimiento del combustible disminuye en 0,494 km/l.

3) Considere el modelo de regresión lineal y responda las siguientes preguntas:

a) Si en un viaje el peso de la carga es de 8 toneladas, estime el rendimiento del combustible.

x =	8
y =	**5,1**

? **=TENDENCIA(conocido_y;conocido_x;valor x)**

O bien se puede utilizar el modelo de regresión lineal, para estimar el valor de la variable dependiente.

Y = -0,494*8+9,0236 ≈ 5,1

Respuesta: El rendimiento del combustible se estima en 5,1 km/l.

b) Estime el peso de la carga de un viaje cuyo rendimiento de combustible es de 7,5 km/l.

x =	**3,1**	?
y =	7,5	

Usando el modelo de regresión lineal, se despeja la variable independiente.

X = (7,5-9,0236)/-0,494 ≈ 3,1

Respuesta: Se estima que el peso de la carga es de 3,1 toneladas.

4) Interprete el coeficiente de determinación lineal, cuyo valor es 92,8%.

Respuesta: El 92,8% de la variación del rendimiento del combustible, se debe a la variabilidad del peso de la carga. Además hay un 7,2% de variabilidad no explicada.