CORRELACIÓN

UNIDAD N° 1

RELACIONES ENTRE VARIABLES

Existen muchos casos en donde dos o más variables están relacionadas entre sí, y el conocimiento de una permite inferir el comportamiento de otra

Las técnicas de correlación y regresión sirven para identificar y modelar posibles relaciones entre variables

RELACIONES ENTRE VARIABLES

Análisis de Correlación

- Permite decidir si dos variables están relacionadas entre sí

Análisis de Regresión

- Permite determinar la relación funcional entre una variable dependiente (*efecto*) y una variable independiente (*causa*)

Tipos de Relaciones

Lineal

La relación se representa por una línea recta

Relación Positiva

Relación directa entre las variables

Relación Negativa

Relación inversa entre las variables

Curvilínea

Una línea curva se ajusta mejor a los datos

Perfecta

Todos los puntos caen sobre una recta

Imperfecta

No todos los puntos caen sobre la recta

OBTENCIÓN DE LA ECUACIÓN DE LA LÍNEA RECTA

Y = B.X + A

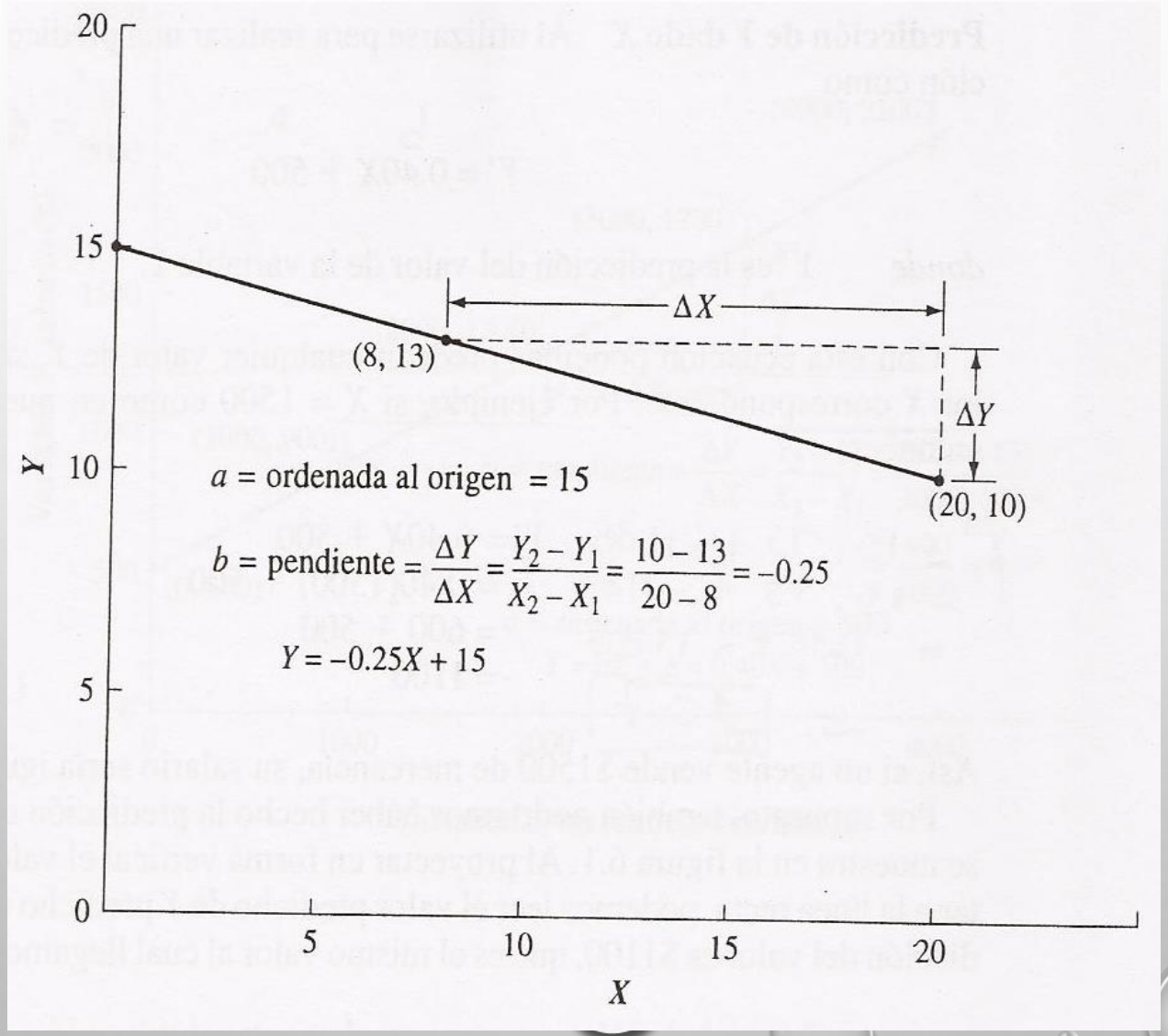
- A = ORDENADA AL ORIGEN (EL VALOR DE Y CUANDO X=0)
- B = PENDIENTE DE LA RECTA

- DETERMINACIÓN DE LA ORDENADA AL ORIGEN A: VALOR DE Y CUANDO LA RECTA CORTA AL EJE VERTICAL.

- DETERMINACIÓN DE LA PENDIENTE B: ES UNA MEDIDA DE SU RAZÓN DE CAMBIO. NOS DICE CUÁNTO CAMBIA UN DATO Y PARA CADA CAMBIO UNITARIO EN EL DATO X. EN FORMA DE ECUACIÓN:

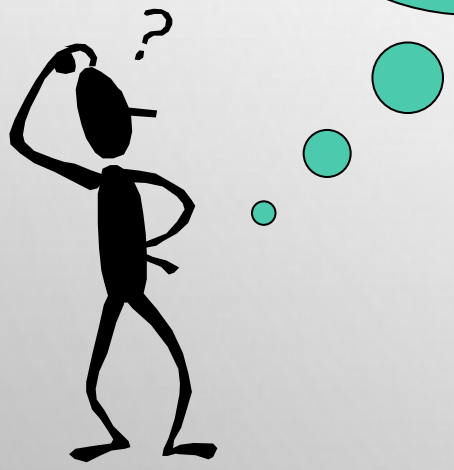
$$b = pendiente = \frac{\Delta Y}{\Delta X} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

OBTENCIÓN DE LA ECUACIÓN DE LA LÍNEA RECTA



CORRELACIÓN LINEAL

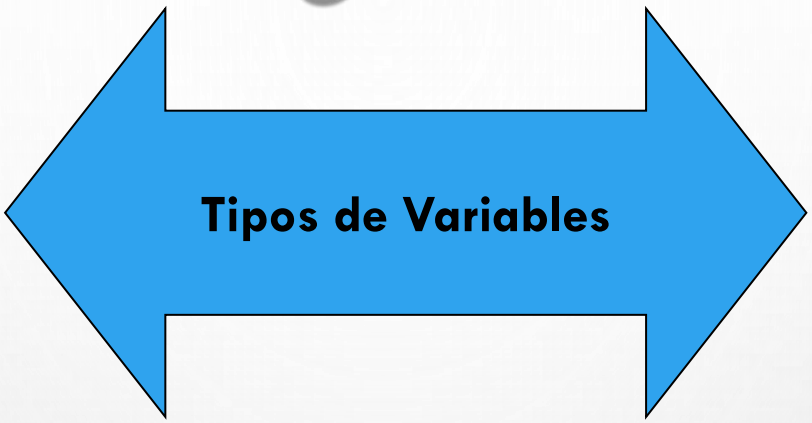
Qué es el análisis de correlación lineal ?



Es una herramienta estadística que podemos usar para describir el **grado de relación** lineal entre las variables.

CORRELACIÓN LINEAL

Variable Independiente (X)
(determinística, es decir **no aleatoria.**)



Variable Dependiente (Y)
aleatoria

Ejemplos

X: Número de llamadas telefónicas realizadas por un vendedor promocionando un producto.
Y: Unidades vendidas por el vendedor.

X: Tiempo que dedica un estudiante a una materia.
Y : Evaluación que obtiene el estudiante en la materia.

CORRELACIÓN

La correlación se ocupa de establecer la magnitud y dirección de las relaciones

Coeficiente de Correlación

- Expresa de manera cuantitativa la magnitud y dirección de una relación
- Varía entre +1 a -1
- El signo indica si es positiva o negativa
- La parte numérica describe la magnitud
- Si vale +1 o -1 la relación es perfecta
- Si vale 0 no existe relación
- Para valores intermedios la relación es imperfecta

COEFICIENTE DE CORRELACIÓN LINEAL R DE PEARSON

Medida en la que las parejas de datos ocupan posiciones iguales u opuestas dentro de sus propias distribuciones

Ecuación

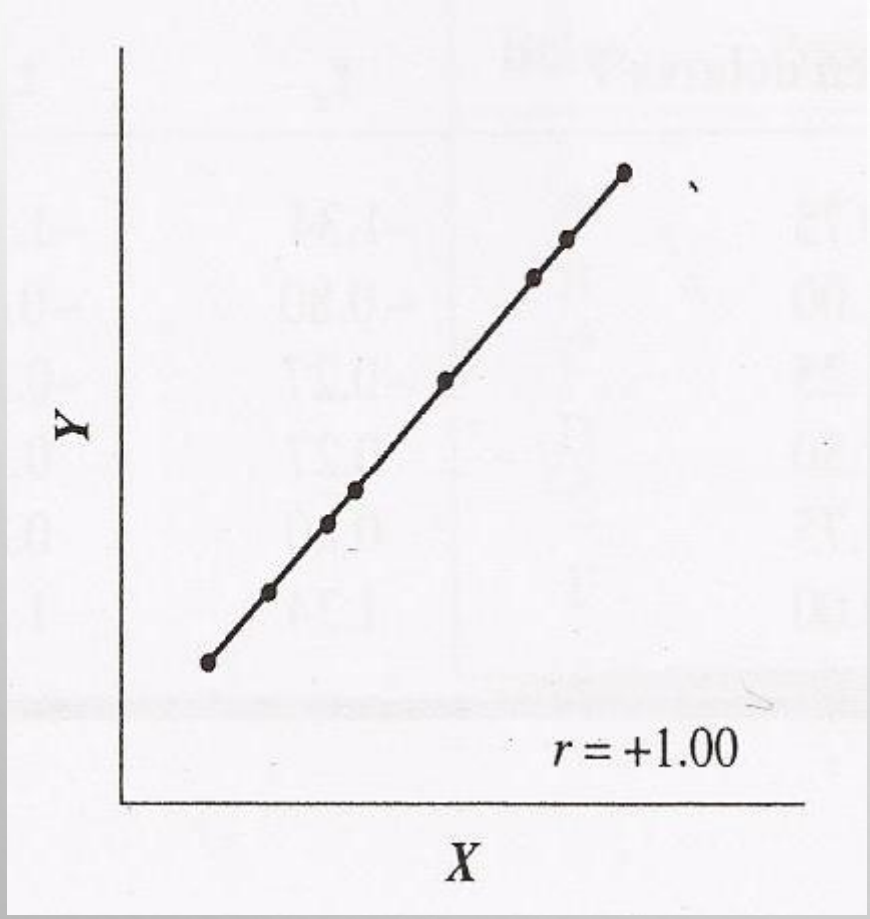
Para utilizar esta ecuación, primero hay que convertir cada dato en bruto en su valor z transformado:

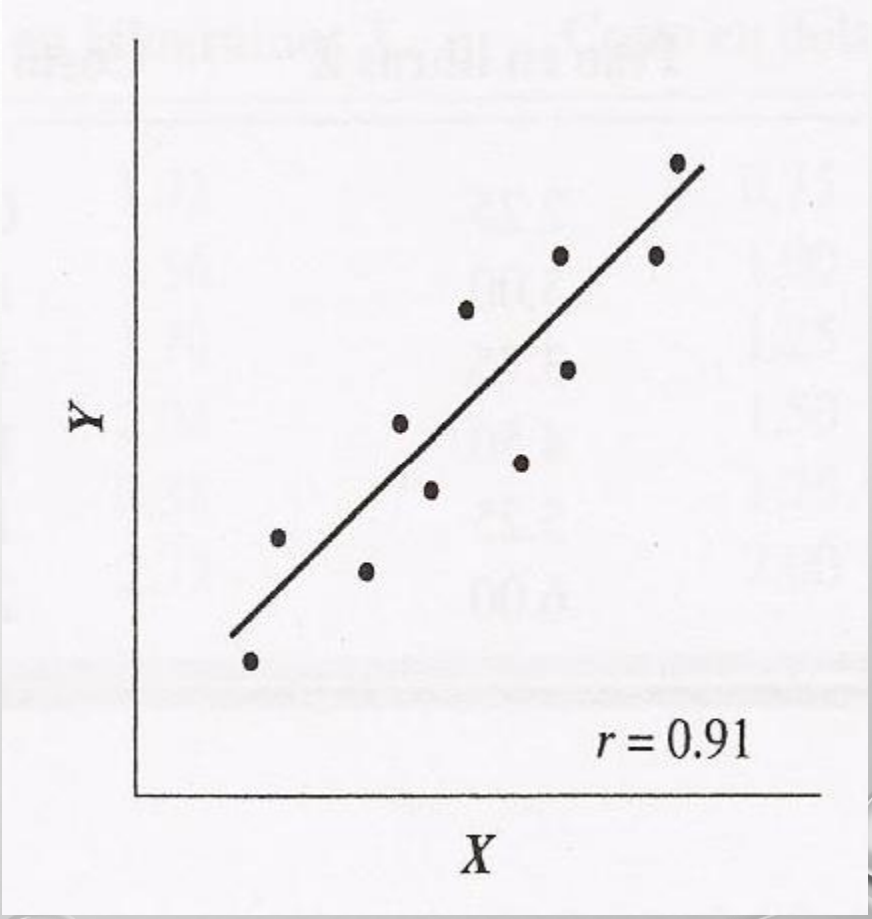
$$r = \frac{\sum z_x z_y}{N - 1}$$

Para datos en bruto:

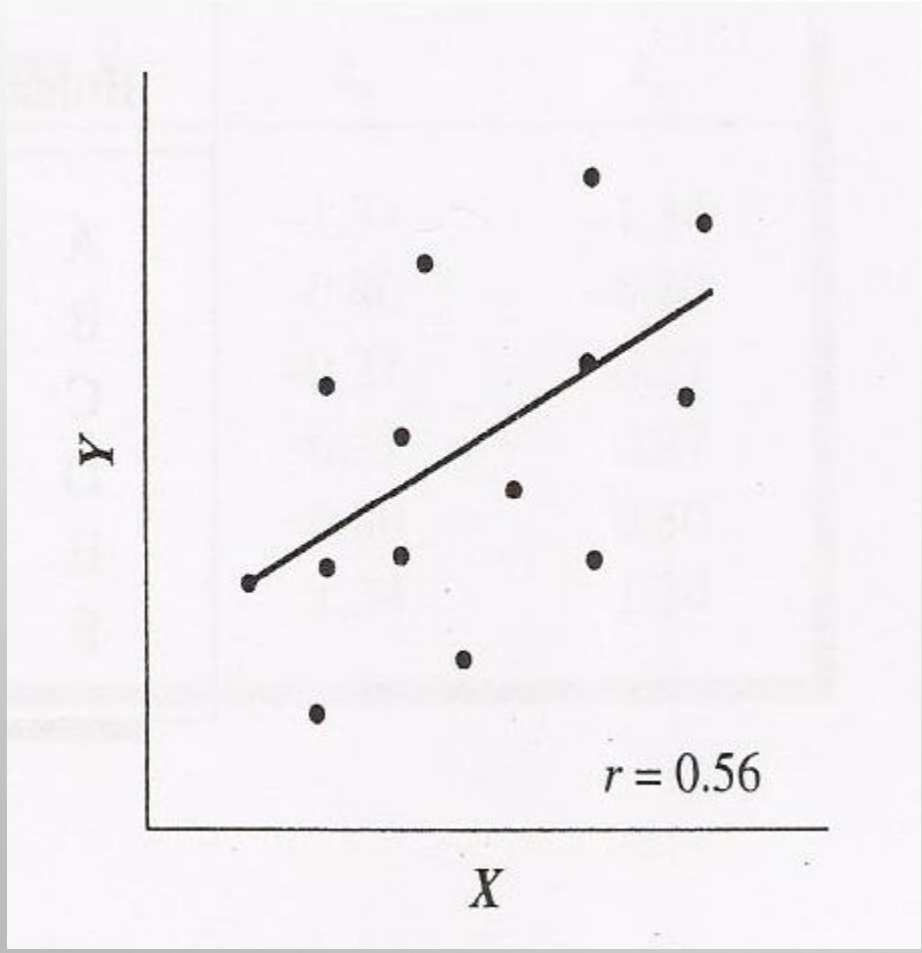
$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

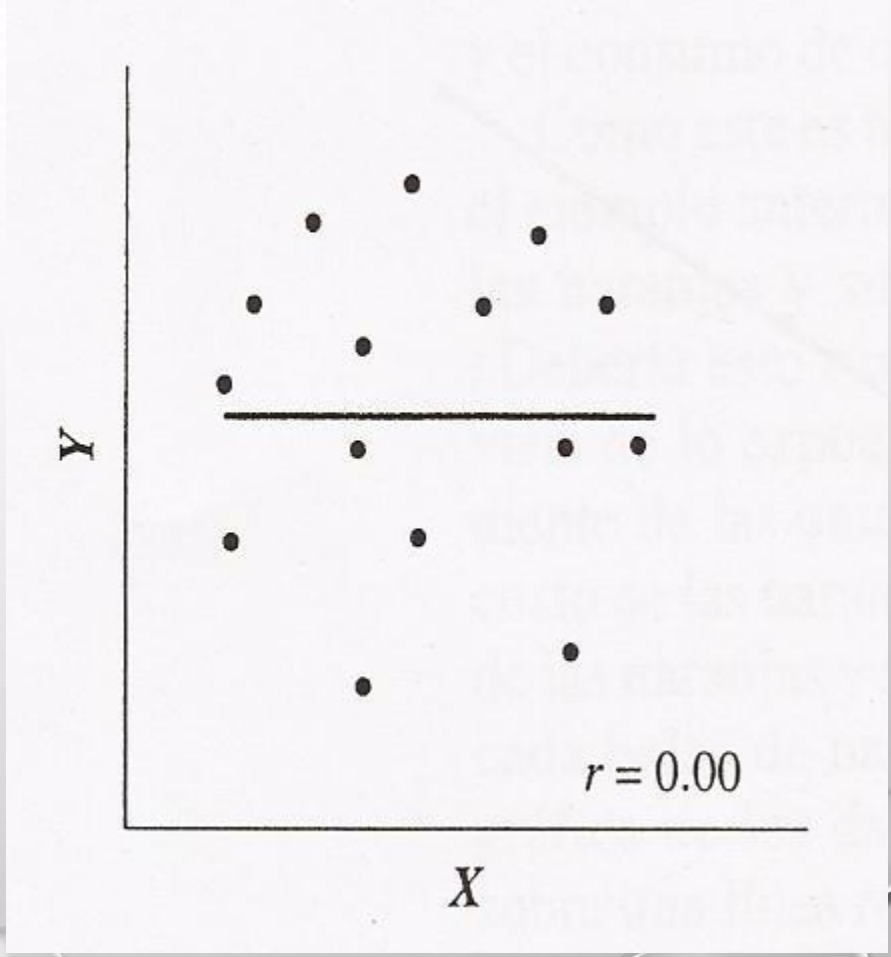
GRÁFICOS CON DIFERENTES COEFICIENTES DE CORRELACIÓN



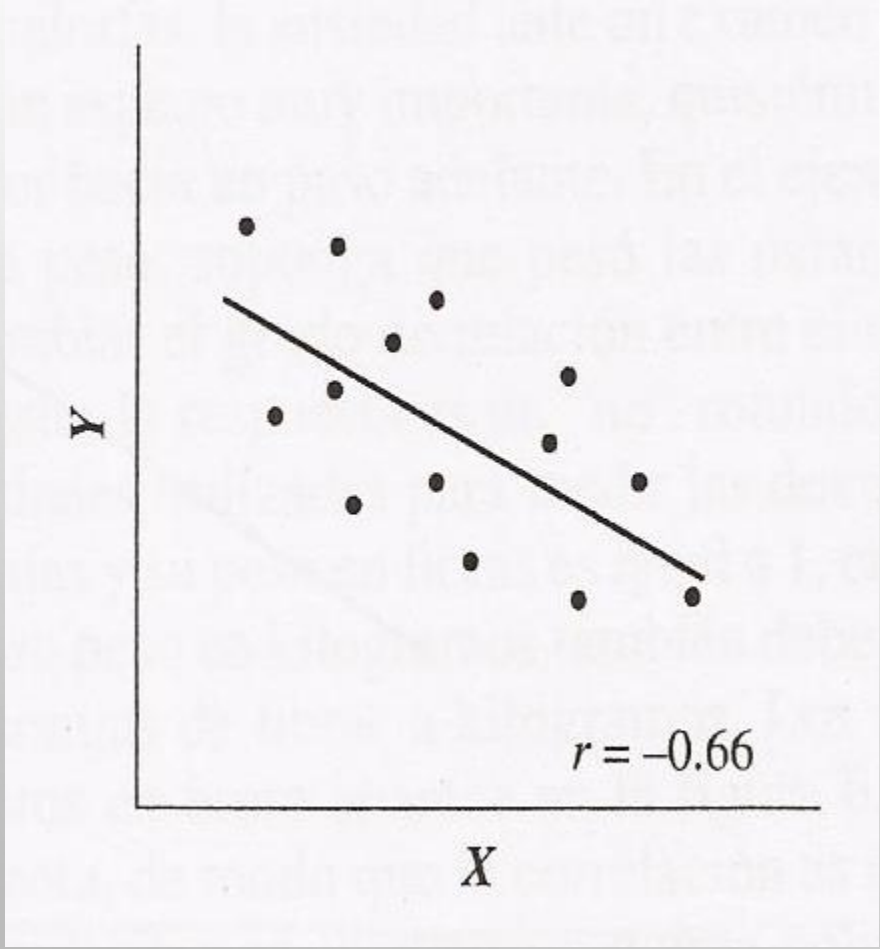


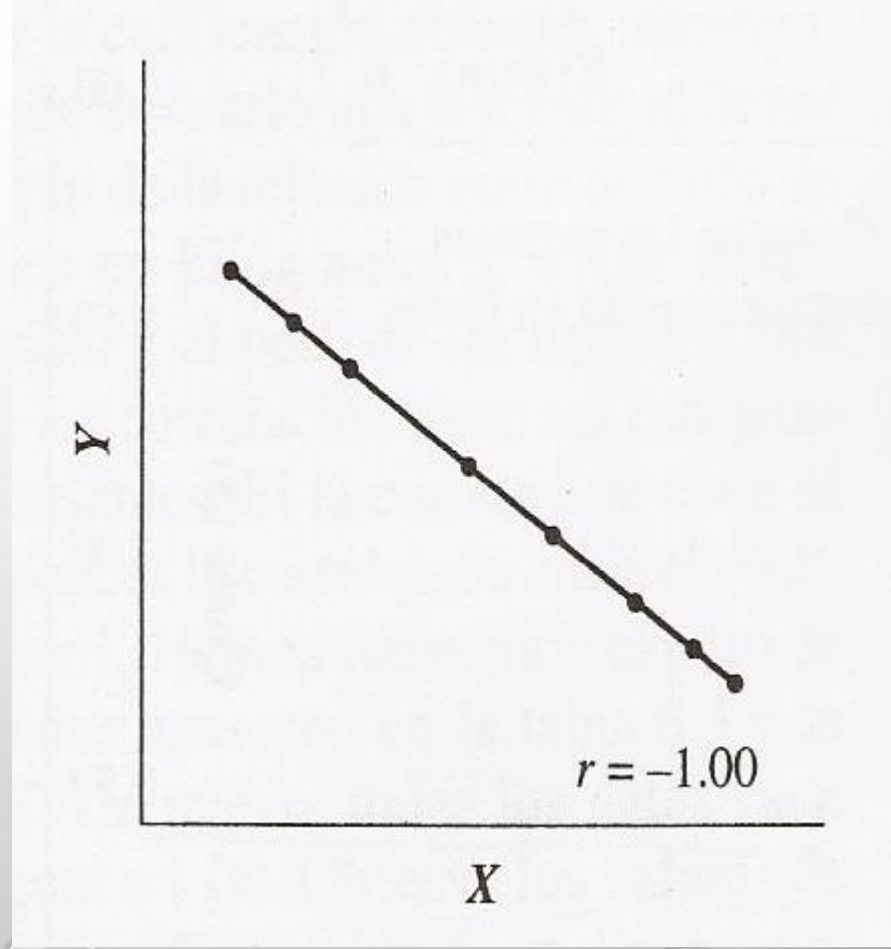
GRÁFICOS CON DIFERENTES COEFICIENTES DE CORRELACIÓN





GRÁFICOS CON DIFERENTES COEFICIENTES DE CORRELACIÓN





OTROS COEFICIENTES DE CORRELACIÓN

El coeficiente r de Pearson es el más utilizado en las investigaciones de las ciencias del comportamiento

Forma de relación: la elección depende de si la relación es lineal o curvilínea

Para relaciones curvilíneas se utiliza el coeficiente de correlación η (eta)

Escala de medición: la elección depende del tipo de escala de medición de datos

El coeficiente *r* de Pearson utiliza escala de intervalo o de razón

El coeficiente *rho* de Spearman (r_s) se utiliza cuando una o ambas variables tienen una escala ordinal

COVARIANZA

La **covarianza** de una variable bidimensional es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas

La **covarianza** se representa por s_{xy} o σ_{xy}

$$\sigma_{XY} = \frac{\sum f_i (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\sigma_{XY} = \frac{\sum f_i x_i y_i}{N} - \bar{x}\bar{y}$$

La **covarianza** indica el sentido de la correlación entre las variables

Si $\sigma_{xy} > 0$ la correlación es directa

Si $\sigma_{xy} < 0$ la correlación es inversa

La **covarianza** presenta como inconveniente, el hecho de que su valor depende de la escala elegida para los ejes

COEFICIENTE DE DETERMINACIÓN (BONDAD DE AJUSTE)

La bondad de la predicción depende de la relación entre las variables. Si dos variables no covarían, no podremos hacer predicciones válidas, y si la intensidad de la covariación es moderada, las predicciones no serán demasiado buenas. En consecuencia, hay que disponer de alguna medida de la capacidad de la ecuación de Regresión para obtener predicciones buenas (en el sentido de que sean lo menos erróneas posible).

Esta medida es el Coeficiente de Determinación, que es el cuadrado del coeficiente de correlación de Pearson, y da la proporción de variación de la variable Y que es explicada por la variable X (variable predictora o explicativa).

Cuanto mayor sea la proporción, mejor será la predicción. Si llegara a ser igual a 1 la variable predictora explicaría TODA la variación de Y, y las predicciones NO tendrían error.

El coeficiente de determinación es la proporción de la variable dependiente explicada por la variable independiente y por lo tanto está entre 0 y 1.

Es decir: $0 < R^2 < 1$

A medida que el R^2 se acerca a 1, la ecuación de regresión es más confiable, y entre más cercano esté el R^2 de cero, la ecuación es menos confiable

EJEMPLO: CORRELACIÓN

- EN UNA CIUDAD DE CANADÁ, LAS PERSONAS AL COMPRAR CASAS SE INTERESAN POR EL PRECIO DEL COSTO DE LA CALEFACCIÓN. SE HA DETERMINADO QUE UN GRUPO DE FACTORES PUEDEN ESTAR RELACIONADOS CON EL COSTO (EN DÓLARES):
  - TEMPERATURA EXTERIOR. (GRADOS FAHRENHEIT)
  - AISLANTE TÉRMICO EN EL DESVÁN. (EN PULGADAS)
  - ANTIGÜEDAD DEL CALEFACTOR.
  - ÁREA DE LA SALA PRINCIPAL DEL APARTAMENTO. (EN METROS CUADRADOS).
- UN CLIENTE LE HA PREGUNTADO A UN VENDEDOR:
- SI USTED ME BRINDA LA INFORMACIÓN DE LAS VARIABLES ANTERIORES DE UN APARTAMENTO, ¿CÓMO PUEDO SABER YO APROXIMADAMENTE CUANTO PAGARÉ EN CALEFACCIÓN?. ¿CUAN CONFIABLE SERÁ LA INFORMACIÓN QUE USTED ME BRINDE?

EJEMPLO: CORRELACIÓN

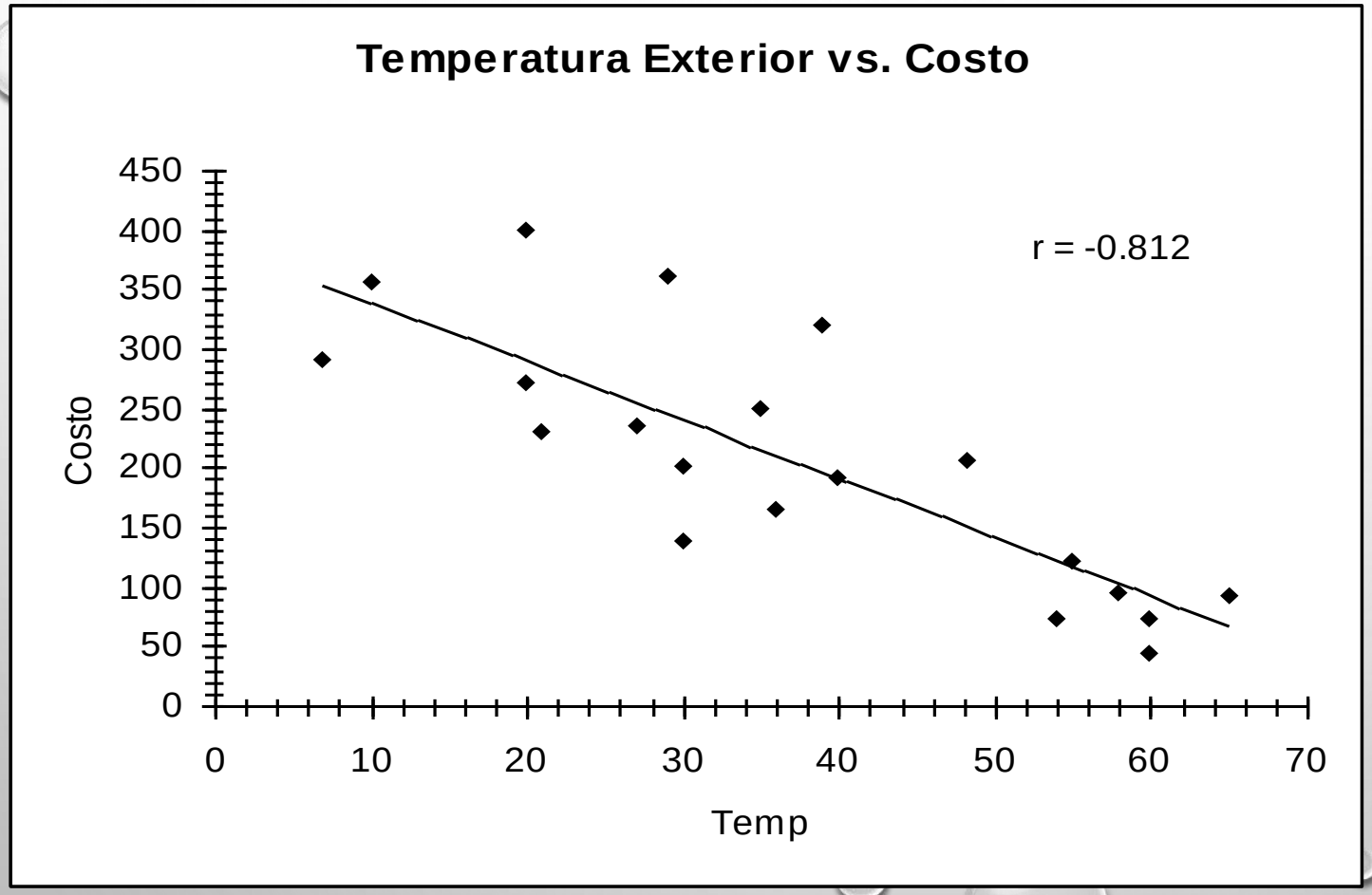
- **PASOS EN EL ANÁLISIS DE CORRELACIÓN:**

1. **DETERMINAR CUÁL ES LA VARIABLE DEPENDIENTE: Y:** COSTO.
2. **SELECCIONAR UNA MUESTRA DE TAMAÑO *N* DE AMBAS VARIABLES X E Y, CON LO QUE SE OBTIENEN N PARES DE OBSERVACIONES: $(X_1 , Y_1) , (X_2 , Y_2) \ldots (X_N , Y_N)$.**
3. **MOSTRAR LA RELACIÓN EN UN DIAGRAMA DE DISPERSIÓN: GRÁFICO DE X VS. Y. SE APRECIA DE MANERA DESCRIPTIVA EL SENTIDO Y LA INTENSIDAD DE RELACIÓN ENTRE LAS VARIABLES.** (SE REALIZARAN LOS 4 GRÁFICOS QUE CORRESPONDEN A CADA UNA DE LAS VARIABLES INDEPENDIENTES CONSIDERADAS)
4. **CALCULAR UN COEFICIENTE DE CORRELACIÓN LINEAL *R* A PARTIR DE LA MUESTRA**

- **EN NUESTRO EJEMPLO SE TOMO UNA MUESTRA DE 20 APARTAMENTOS. SE MIDIERON TODAS LAS VARIABLES INDEPENDIENTES PARA CADA UNO DE ELLOS.**

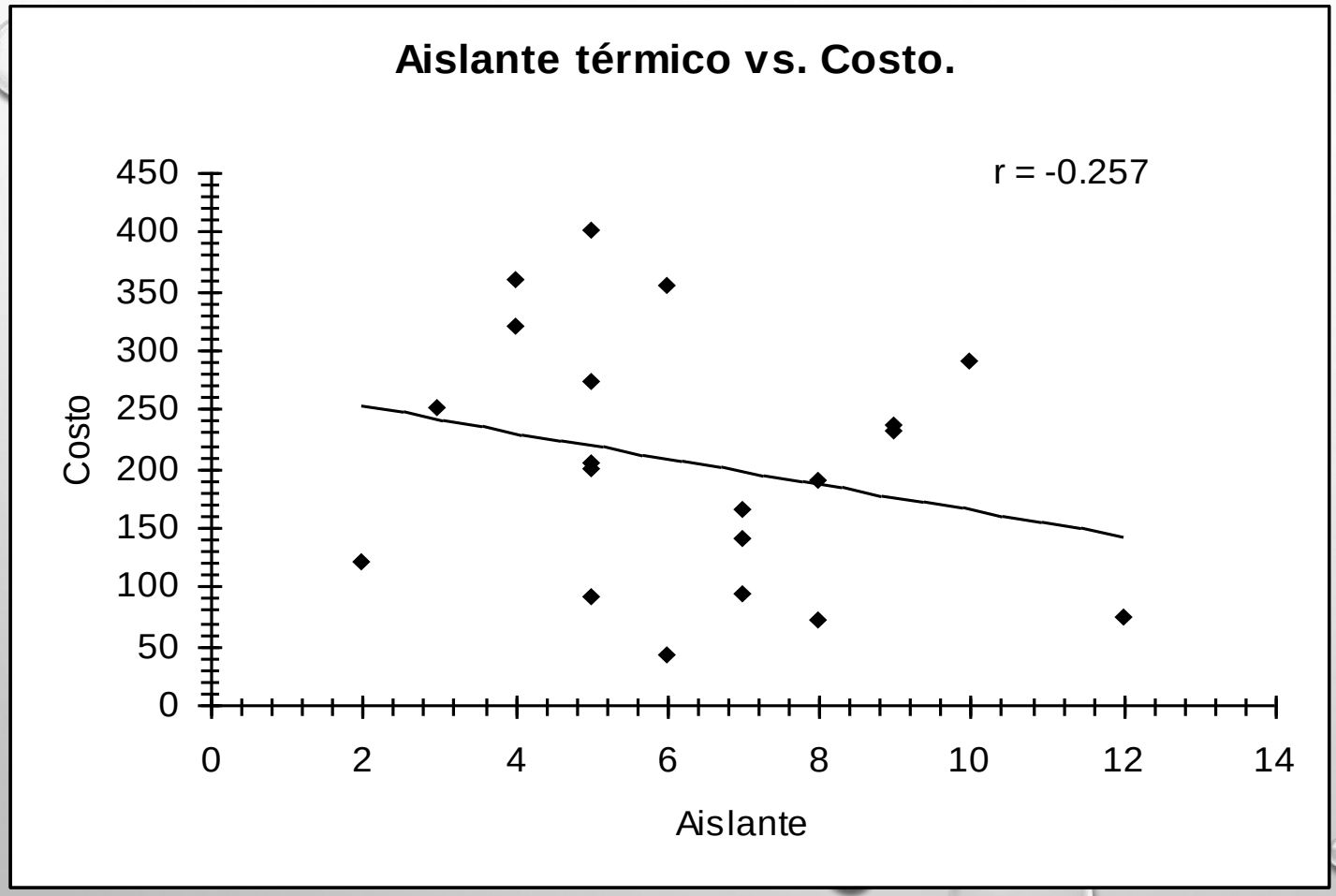
Costo	Temperatura	Aislante	Antigüedad	Tamaño
250	35	3	6	15,81139
360	29	4	10	18,97367
165	36	7	3	12,84523
43	60	6	9	6,557439
92	65	5	6	9,591663
200	30	5	5	14,14214
355	10	6	7	18,84144
290	7	10	10	17,02939
230	21	9	11	15,16575
120	55	2	5	10,95445
73	54	12	4	8,544004
205	48	5	1	14,31782
400	20	5	15	20
320	39	4	7	17,88854
72	60	8	6	8,485281
272	20	5	8	16,49242
94	58	7	3	9,69536
190	40	8	11	13,78405
235	27	9	8	15,32971
139	30	7	5	11,78983

Ejemplo: Correlación



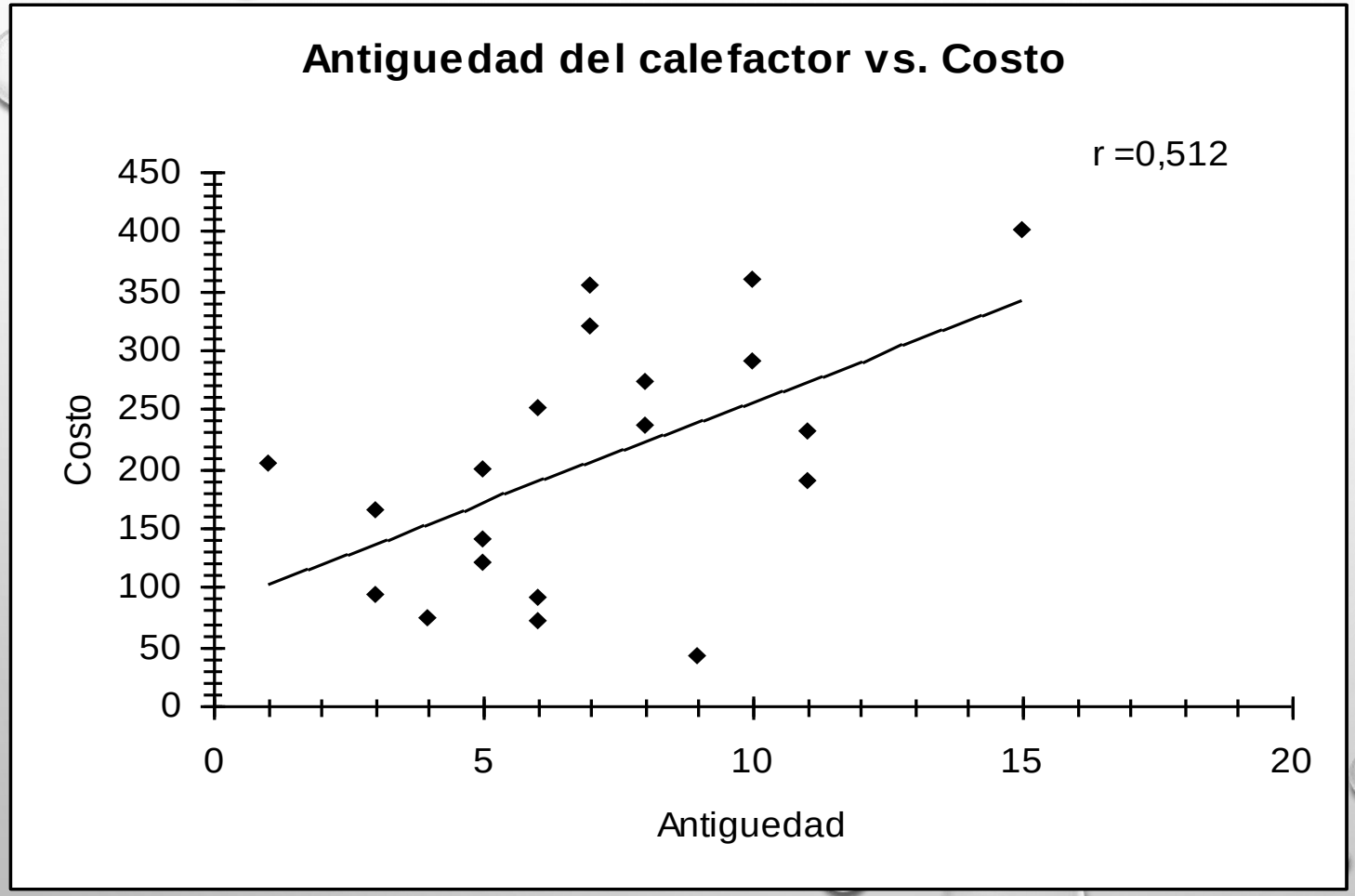
Relación Inversa fuerte

Ejemplo: Correlación



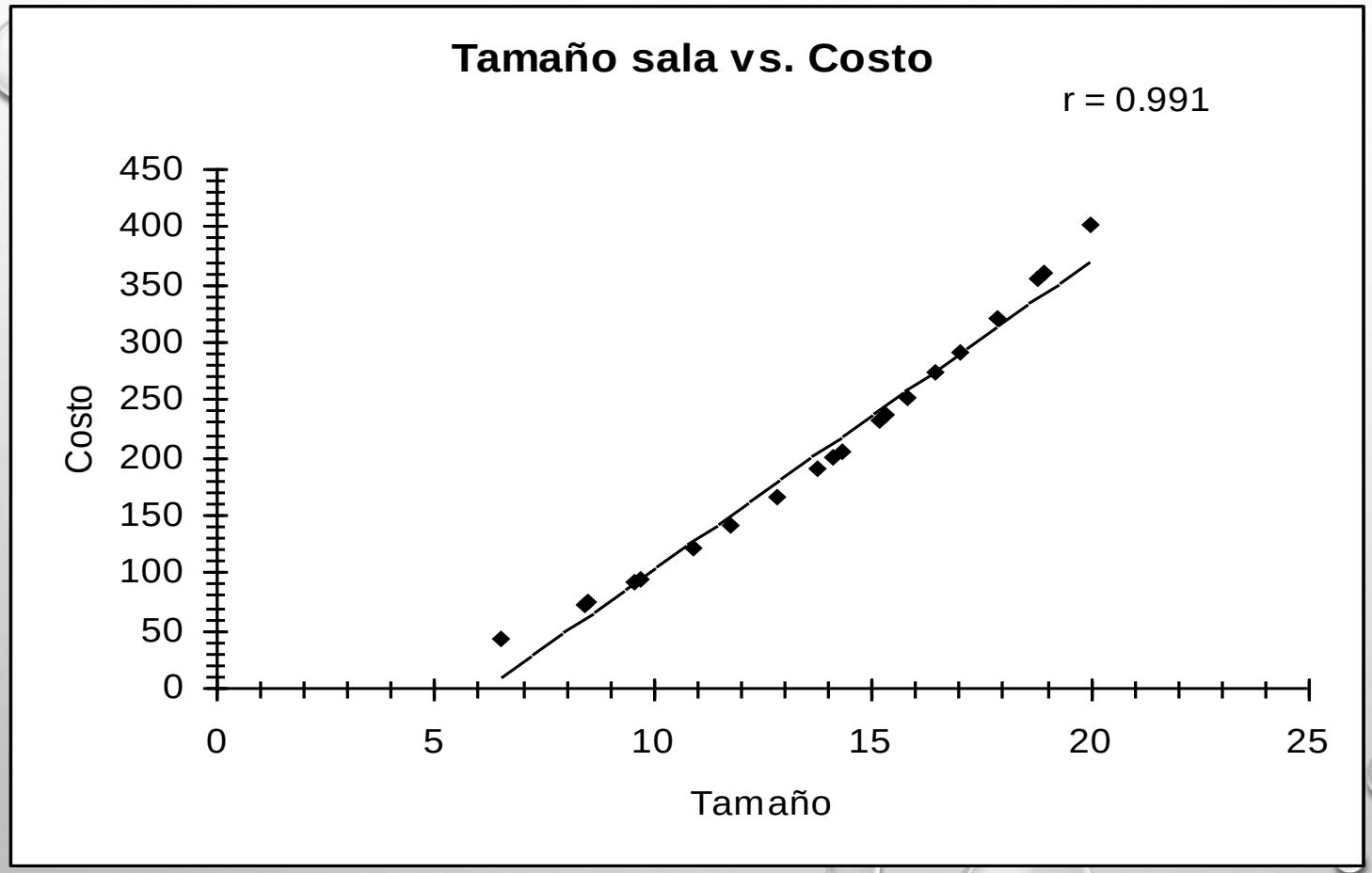
Relación Inversa débil

Ejemplo: Correlación



Relación Directa moderada

Ejemplo: Correlación



Relación directa fuerte

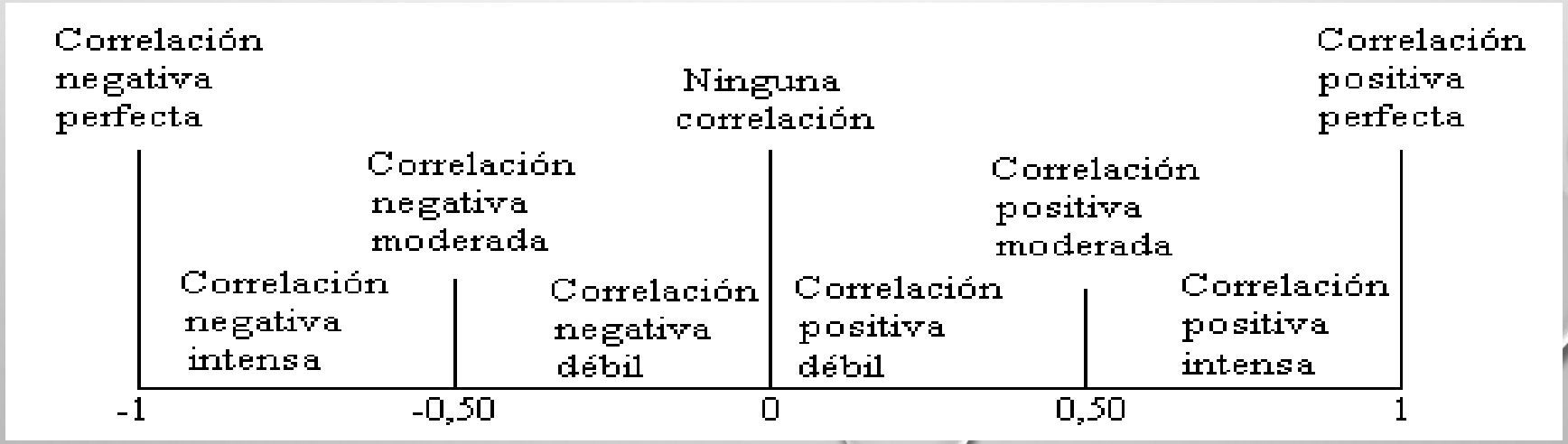
aunque se aprecia una tendencia no lineal

Ejemplo: Correlación

Interpretación:

1. **El valor absoluto de r indica la fuerza de la relación entre Y y X.**
2. **El signo la dirección de la relación (directa o inversamente proporcional) (tener cuidado con relaciones espúreas)**

r = 1 correlación positiva perfecta.
r = -1 correlación negativa perfecta.
r = 0 no hay relación lineal entre Y y X.



EJEMPLO: CORRELACIÓN

- **TEMPERATURA.**
  - UNA CORRELACIÓN DE **-**0,812 INDICA ALTA CORRELACIÓN, INVERSAMENTE PROPORCIONAL:
  - A MAYOR TEMPERATURA EXTERIOR, MENOR EL COSTO EN CALEFACCIÓN Y VICEVERSA.
- **AISLANTE.**
  - LA CORRELACIÓN DE 0,257 ES BAJA, ASÍ QUE NO EXISTE RELACIÓN LINEAL ENTRE LAS VARIABLES.

ANTIGÜEDAD.

- UNA CORRELACIÓN DE 0,512; ES MODERADA, DIRECTAMENTE PROPORCIONAL, A MAYOR ANTIGÜEDAD DEL CALEFACTOR, MAYOR COSTO Y VICEVERSA.

TAMAÑO DE LA SALA PRINCIPAL.

- UNA CORRELACIÓN DE 0,991; ES ALTA Y DIRECTAMENTE PROPORCIONAL: A MAYOR TAMAÑO DE LA SALA, MAYOR COSTO DE LA CALEFACCIÓN

CORRELACIÓN PARCIAL Y MÚLTIPLE

CORRELACIÓN PARCIAL

La correlación parcial se define como la correlación entre dos variables manteniendo las demás variables constantes, es decir, se eliminan los efectos de todas las demás variables

$r_{12.3}$ denota el coeficiente de correlación parcial entre X_1 y X_2 cuando X_3 permanece constante

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r^2{}_{13})(1-r^2{}_{23})}}$$

De manera similar, si $r_{12.34}$ denota el coeficiente de correlación parcial entre X_1 y X_2 cuando X_3 y X_4 permanecen constantes

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1-r^2{}_{13.4})(1-r^2{}_{23.4})}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1-r^2{}_{14.3})(1-r^2{}_{24.3})}}$$

EJEMPLO

	Calificación (Y)	Inteligencia (X_1)	Horas est. (X_2)
1	4,1	109	10
2	4,3	120	8
3	6,4	112	21
4	4,5	115	14
5	4,2	98	18
6	5,5	101	23
7	6,0	100	21
8	5,1	105	12
9	8,8	130	21
10	7,5	121	19
11	7,8	132	16
12	9,3	140	18
13	5,2	111	9
14	6,5	109	25
15	5,2	95	16

- SUPONGAMOS EN ESTE SENTIDO QUE TENEMOS UNA MUESTRA DE 15 SUJETOS Y DESEAMOS ESTUDIAR EL EFECTO QUE TIENE SOBRE LA CALIFICACIÓN DE UNA DETERMINADA ASIGNATURA (Y) LAS SIGUIENTES VARIABLES:
  INTELIGENCIA ($X1$)
  HORAS DE ESTUDIO ($X2$)

EJEMPLO

- SI DESEAMOS CONOCER LA CORRELACIÓN PARCIAL DE LA INTELIGENCIA CON LAS CALIFICACIONES ELIMINANDO EL EFECTO DE LAS HORAS DE ESTUDIO:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{(1 - r^2{}_{Y2})(1 - r^2{}_{12})}} = 0{,}8879$$

- SI DESEAMOS CONOCER LA CORRELACIÓN PARCIAL DE LAS HORAS DE ESTUDIO CON LAS CALIFICACIONES ELIMINANDO EL EFECTO DE LA INTELIGENCIA:

$$r_{Y2.1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{(1 - r^2{}_{Y1})(1 - r^2{}_{12})}} = 0{,}8059$$

COEFICIENTE DE CORRELACIÓN MÚLTIPLE

En el caso de tener dos variables independientes, el ***coeficiente de correlación múltiple*** está dado por:

$$R_{1.23} = \sqrt{1 - \frac{s^2{}_{1.23}}{s^2{}_1}}$$

La cantidad $R^2_{1.23}$ se conoce como ***coeficiente de determinación múltiple***

Cuando se emplea una ecuación de regresión lineal, al coeficiente de correlación múltiple se le llama ***coeficiente de correlación lineal múltiple***.

La ecuación puede expresarse en términos de r_{12}, r_{13}, r_{23} como:

$$R_{1.23} = \sqrt{\frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}{}^2}}$$

El valor de un coeficiente de correlación múltiple, como $R_{1.23}$, está entre 0 y 1, inclusive. Cuanto más cerca está de 1, mejor es la relación lineal entre las variables. Cuanto más cerca está de 0, peor será la relación lineal entre las variables. Si un coeficiente de correlación múltiple es 1, esa correlación es perfecta. Aunque un coeficiente de correlación sea 0, esto indica que no hay relación lineal entre las variables, pero puede que exista una relación no lineal.

EJEMPLO

- CONSIDERAMOS EL EJEMPLO ANTERIOR, SÓLO QUE AHORA VAMOS A SUPONER QUE LAS CALIFICACIONES (Y) DEPENDEN SIMULTÁNEAMENTE DE LA INTELIGENCIA (X1) Y LAS HORAS DE ESTUDIO (X2), POR LO QUE EL COEFICIENTE DE CORRELACIÓN MÚLTIPLE SERÍA:

$$R_{Y.12} = \sqrt{\frac{{r_{Y1}}^2 + {r_{Y2}}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - {r_{12}}^2}} = 0{,}9175$$