Tema 2. análisis bidimensional. Regresión y correlación

1. Cuando sobre cada individuo se observan simultáneamente dos características ¿cómo se organizan y representan gráficamente esos datos bidimensionales?

2. ¿Cómo se puede saber si dos variables estadísticas están relacionadas de forma lineal, exponencial, potencial o parabólica?

3. ¿Se puede predecir el valor de una variable sabiendo el valor de otra variable que está relacionada con ella de forma lineal, exponencial, potencial o parabólica?

7.1. INTRODUCCIÓN

¿Cómo se relacionan unos fenómenos con otros?

Descubrimiento de relaciones

Realizar predicciones

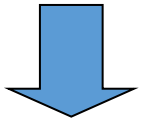
Identificar posibles causas de un fenómeno

La observación de relaciones entre VV ayuda a comprender los fenómenos y a encontrar explicaciones a los mismos.

1. Existe relación entre la ansiedad pre-competitiva y el rendimiento
2. La envergadura de un sujeto y el tiempo de nado
3. La edad del sujeto y la tensión asistólica que tiene
4. La educación de los padres y de sus hijos

Existe relación entre dos variables

Dos variables covarían



Si ciertos valores de una de las variables están ligados con más probabilidad a ciertos valores de la otra

Valores de X ESTATURA	Valores de Y PESO
Altos	Altos
Medios	Medios
Bajos	Bajos

análisis bidimensional. Regresión y correlación

- Análisis conjunto de dos variables
  - Dos caracteres cualitativos
  - Dos caracteres cuantitativos
    - Dos discretas
    - Dos continuas
    - Una discreta y la otra continua.
  - Uno cualitativo y otro cuantitativo

Variables cualitativas	Categórica / Categórica	Sexo y clase social
Variables cuantitativas	Discreta / Discreta	Número de hermanos y número de hijos.
	Continua / Continua	Peso y altura
	Discreta / Continua	Pulsaciones y temperatura cuerpo
Cualitativa y cuantitativa	Categórica / Discreta	Sexo y número de cigarrillos
	Categórica / Continua	Sexo e ingresos

Variables Estadísticas Bidimensionales

Tenemos una muestra bidimensional cuando sobre cada elemento de la muestra se realiza la observación simultánea de dos caracteres.

Ejemplo: color de ojos y color de pelo; peso y altura

La variable estadística es bidimensional

$(x,y)=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)..(x_k,y_k)\}$

La tabla de frecuencias correspondiente se denomina **Tabla de doble Entrada**

Estudio conjunto de dos variables

• A la derecha tenemos una posible manera de recoger los datos obtenido observando dos variables o características X e Y en varios individuos de una muestra.

• (X, Y) es una variable bidimesional
– En cada **fila** tenemos los datos de un individuo
– Cada **columna** representa los valores que toma una variable aleatoria sobre los mismos.
– Las individuos no se muestran en **ningún orden** particular.

• Dichas observaciones pueden ser representadas en un **diagrama de dispersión (*'scatterplot') o nube de puntos.*** En ellos, cada individuo es un punto cuyas coordenadas son los valores de las variables: *(x1,y1), (x2,y2),…. (xn,yn)*

• Nuestro objetivo será intentar **reconocer** a partir del mismo si hay **relación** entre las variables, de qué **tipo**, y si es posible **predecir** el valor de una de ellas en función de la otra.

Altura en cm.	**Peso en Kg.**
162	61
154	60
180	78
158	62
171	66
169	60
166	54
176	84
163	68
…	…

Variables Estadísticas Bidimensionales.- Tablas de correlación / contingencia

Tabla de Doble Entrada

$x \backslash y$	y_1	y_2	y_3	$\cdots$	y_j	$\cdots$	y_l	Suma
x_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{x_2}
x_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3j}	$\cdots$	n_{3l}	n_{x_3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	n_{i3}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{x_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{x_k}
Suma	n_{y_1}	n_{y_2}	n_{y_3}	$\cdots$	n_{y_j}	$\cdots$	n_{y_l}	N

$$f_{ij} = \frac{n_{ij}}{N}; \quad \sum_{i=1}^{K}\sum_{j=1}^{l} n_{ij} = N; \quad \sum_{i=1}^{K}\sum_{j=1}^{l} f_{ij} = 1$$

Por ejemplo, n_{11} muestra el número de veces que x_1 aparece conjuntamente con y_1; n_{12} es la frecuencia conjunta de x_1 y y_2, etc.
En el caso de datos cualitativos, la llamaremos tabla de contingencia.

Variables Estadísticas Bidimensionales

En el caso en que las variables X e Y sean continuas, éstas se suelen agrupar en clases. Por ejemplo, en la siguiente tabla se resumen las frecuencias absolutas de la variable bidimensional estatura peso, (E, P), en metros y en kilos de un conjunto de 100 individuos:

	$P \leq 50$	$50 < P \leq 70$	$70 < P \leq 90$	$90 < P$
$E \leq 1'5$	2	1	1	0
$1'5 < E \leq 1'65$	2	7	25	6
$1'65 < E \leq 1'8$	0	6	15	5
$1'8 < E \leq 1'95$	1	4	12	4
$E > 1'95$	0	1	2	6

Representaciones Gráficas

Diagrama de Barras e histograma tridimensional
Diagrama de Dispersión

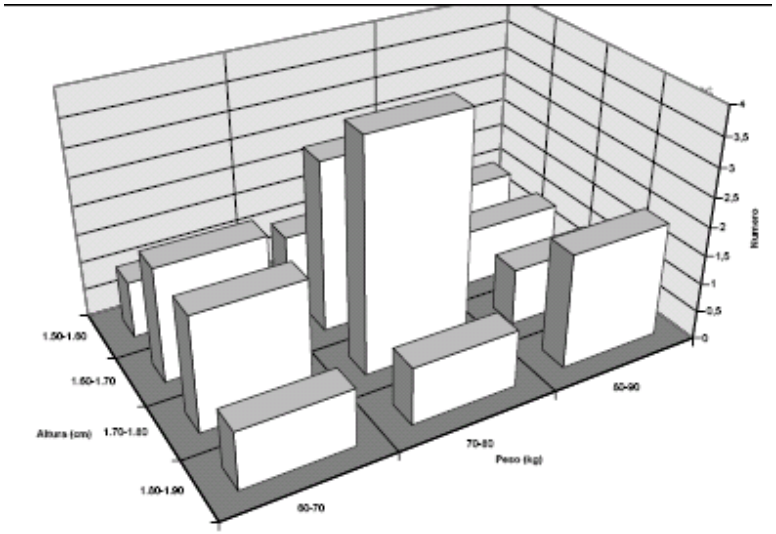


Figura 10: Diagrama tridimensional para la muestra de pesos y alturas de 20 personas.

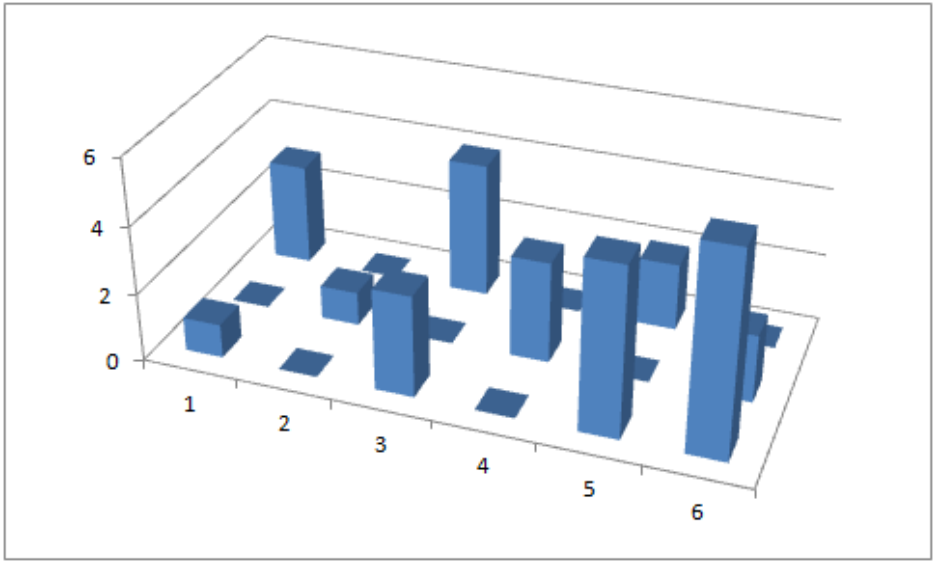
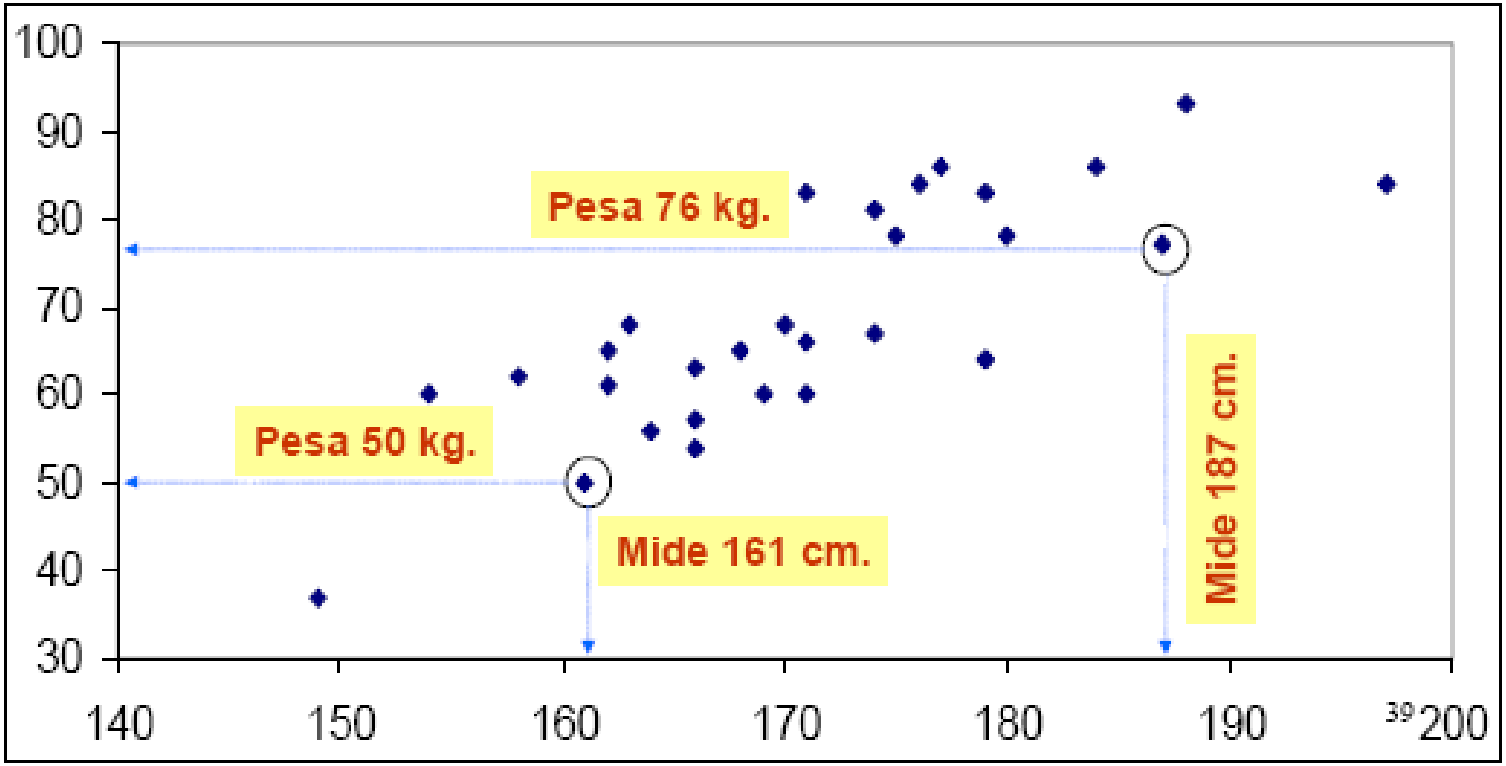
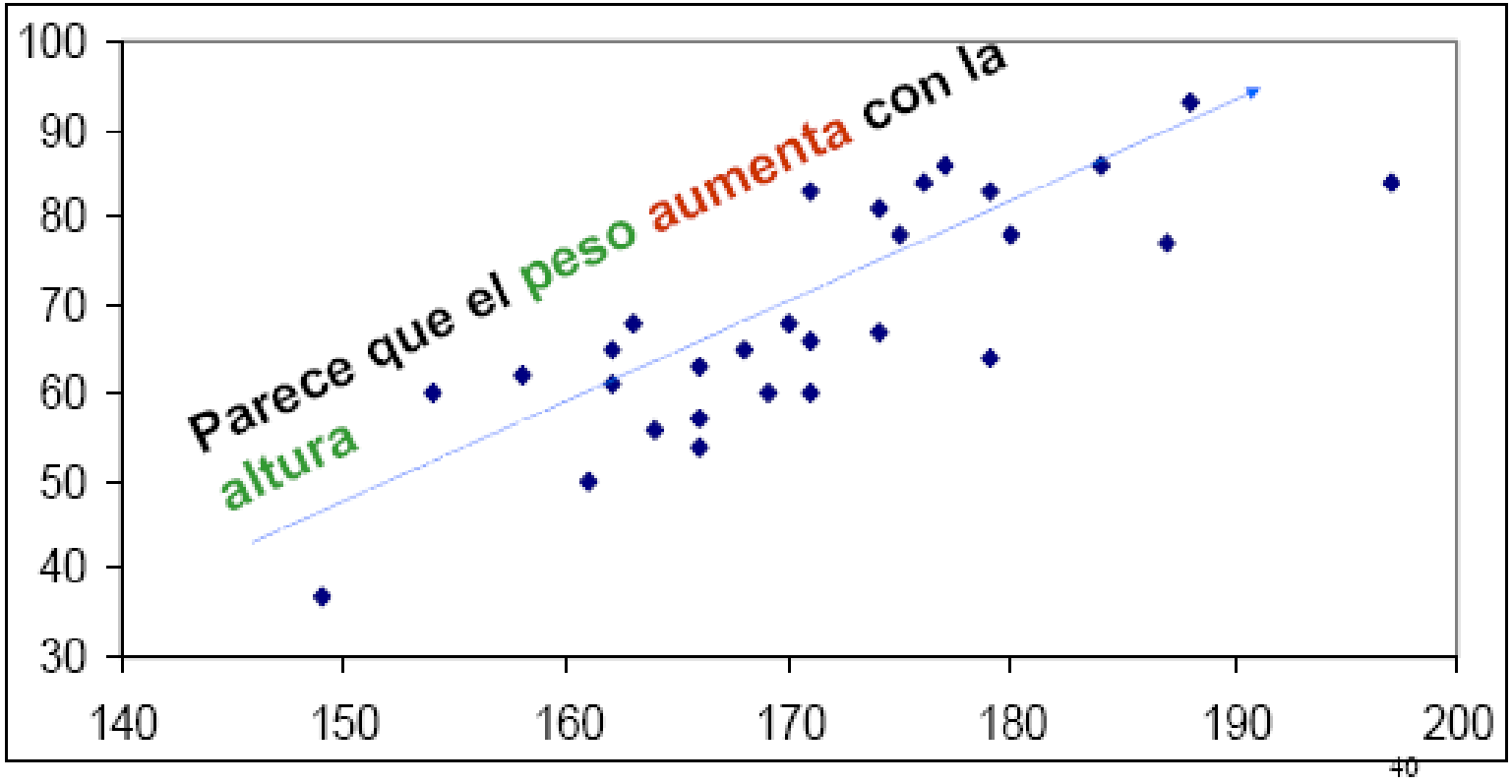


DIAGRAMA DE DISPERSIÓN O NUBE DE PUNTOS



RELACIÓN ENTRE VARIABLES



VARIABLES ESTADÍSTICAS BIDIMENSIONALES

- La presentación de una tabla bidimensional permite obtener tablas de las variables unidimensionales correspondientes sumando las filas o columnas según convenga.

- Estas variables se suelen denominar marginales, porque habitualmente se presentan en los márgenes de la tabla bidimensional.

DISTRIBUCIONES MARGINALES

A partir de la distribución bidimensional, podría interesarnos estudiar solo una variable. En este sentido, de la distribución bidimensional se obtienen 2 distribuciones unidimensionales (la de X y la de Y).

- Para el i-ésimo valor de X, la frecuencia marginal es:

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ij} + \cdots + n_{ik} = \sum_{j=1}^{k} n_{ij}$$

- La frecuencia marginal del j-ésimo valor de Y es:

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{ij} + \cdots + n_{hj} = \sum_{i=1}^{h} n_{ij}$$

DISTRIBUCIONES MARGINALES

Distribución de frecuencias de la variable X

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k	
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}	$\sum_{s=1}^{k} f_{1s}$
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}	$\sum_{s=1}^{k} f_{2s}$
	$\vdots$				$\vdots$	$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}	$\sum_{s=1}^{k} f_{is}$
	$\vdots$				$\vdots$	$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}	$\sum_{s=1}^{k} f_{rs}$

DISTRIBUCIONES MARGINALES

Distribución de frecuencias de la variable Y

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$		$\vdots$		$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$		$\vdots$		$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}
	$\sum_{s=1}^{r} f_{s1}$	$\cdots$	$\sum_{s=1}^{r} f_{sj}$	$\cdots$	$\sum_{s=1}^{r} f_{sk}$

Frecuencias marginales: son distribuciones de frecuencias en cada una de las variables sin tener en cuenta las otras. Son por tanto, distribuciones **Unidimensionales**

$$n_{xj} = \sum_{j=1}^{l} n_{ij} \quad ; \quad n_{yi} = \sum_{i=1}^{k} n_{ij}$$

Distribución marginal de X:

$$f_{xj} = \sum_{j=1}^{l} f_{ij} = \frac{n_{xj}}{N} \quad ;$$

Distribución marginal de y:

$$f_{yi} = \sum_{i=1}^{k} f_{ij} = \frac{n_{yi}}{N} \quad ;$$

Media y varianza:

$$\overline{X} = \frac{\sum_{i=1}^{k} X_i n_{xi}}{N}; \overline{Y} = \frac{\sum_{i=1}^{k} Y_i n_{yi}}{N}$$

$$s_x^2 = \frac{\sum_{i=1}^{k} (X_i - \overline{X}) n_{xi}}{N}; \; s_y^2 = \frac{\sum_{j=1}^{l} (Y_i - \overline{Y}) n_{yi}}{N}$$

DISTRIBUCIONES CONDICIONADAS

Si se observan los distintos valores de la variable X para un valor fijo de la variable Y , Y_j , **se obtiene la distribución de X condicionada a Y = Y_j** . Esta variable es unidimensional. En la columna marcada aparecen las frecuencias absolutas de la misma.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$				$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$				$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}

DISTRIBUCIONES CONDICIONADAS

Distribución de Y condicionada a X = Xi .

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$		$\vdots$		$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$		$\vdots$		$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}

Distribuciones condicionadas: estudio de una variable condicionado a que la otra variable tome un cierto valor.

- **Distribución de y condicionada a $X=x_i$:** $Y/X=x_i$
- **Distribución de x condicionada a $Y=y_i$:** $X/Y=y_j$

Tabla de frecuencias condicionada

$$n(x_i \,|\, y=y_j)=n_{ij} \quad ; \quad n(y_i \,|\, x=x_i)=n_{ij}$$

$$f(x_i \,|\, y=y_j)=\frac{1}{n_{yj}}n(x_i \,|\, y=y_i)=\frac{1}{n_{yj}}n_{ij}$$

$$\sum_{i=1}^{k} n(x_i \,|\, y=y_j)=n_{yj} \quad ; \quad \sum_{j=1}^{l} n(y_i \,|\, x=x_i)=n_{xi}$$

$$\sum_{i=1}^{k} f(x_i \,|\, y=y_j)=1$$

x	$n(x\|y=y_j)$	$f(x\|y=y_j)$
x_1	n_{1j}	f_{1j}
x_2	n_{2j}	f_{2j}
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_{ij}
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	f_{kj}
	n_{y_j}	1

edad	nº cirugías
18	0
25	2
28	1
32	3
43	3
49	2
55	1
15	1
59	3
31	3
43	3
46	0
56	2
36	2
16	0
18	2
52	0
29	0
44	0
42	1
26	2
43	0
21	2
49	2
42	2

	0	**1**	**2**	**3**	n_{xi}
(10-20]	2	1	1	0	4
(20-30]	1	1	3	0	5
(30-40]	0	0	1	2	3
(40-50]	3	1	3	2	9
(50-60]	1	1	1	1	4
n_{yj}	7	4	9	5	**25**

Media aritmética de cirugías = 1,48

Media aritmética de edad = 36,6

¿Cuál es el promedio de cirugías cuando la edad varía entre 10 y 40 años?

¿Cuál es el promedio de cirugías cuando la edad varía entre 10 y 40 años?

Y	0	1	2	3	total
n(y|x=10-40)	3	2	5	2	12
f(y|x=10-40)	0.25	0.17	0.42	0.17	1

Media aritmética = 1,5

Una variable (x,y) tiene una distribución bidimensional contenida en la siguiente tabla

x\y	10	20	30	n_i
15	2	4	6	12
25	3	6	9	18
35	1	2	3	6
n_j	6	12	18	36

- Determinar la distribución bidimensional y las distribuciones marginales de frecuencias relativas
- Calcular las distribuciones Y condicionadas para los distintos valores de X
- Calcular las distribuciones X condicionadas para los distintos valores de Y

RELACIONES DE DEPENDENCIA

- Dependencia funcional:

Y depende funcionalmente de X si existe una función que relaciona los elementos de X y los elementos de Y : Y = f (X)

- Dependencia estadística:

Y depende estadísticamente de X si las variables están relacionadas, pero la relación no puede expresarse mediante una función matemática.

La dependencia estadística puede medirse gradualmente, ya que puede haber relaciones más débiles o más fuertes. Llamamos a esta relación correlación entre variables cuantitativas y contingencia entre variables cualitativas.

- Independencia:

Dos variables X y Y son independientes cuando no existe ninguna relación entre ellas.

Independencia estadística

- Dos variables son ***estadísticamente independientes*** cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra.
- No existe relación entre las variables si:

$$f_{ij} = f_{i.} \cdot f_{.j} \quad \Rightarrow n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \qquad \forall i, j$$

Una variable (x,y) tiene una distribución bidimensional contenida en la siguiente tabla

x\y	10	20	30	n_i
15	2	4	6	12
25	3	6	9	18
35	1	2	3	6
n_j	6	12	18	36

$$f_{22} = f_{2.} \cdot f_{.2} \quad \Rightarrow n_{22} = \frac{n_{2.} \cdot n_{.2}}{n} \qquad \forall i, j$$

$$6/36 = 18/36 \cdot 12/36 \quad \Rightarrow 6 = \frac{18 \cdot 12}{36}$$

Determinar si ambas variables son independientes

X \ Y	1	2	3	n_{ix}
5	1	10	5	16
10	2	20	10	32
15	4	40	20	64
n_{yj}	7	70	35	112

- EJEMPLO

X\Y	19	23	25	31	$n_{i.}$
28	0	25	31	7	**63**
29	2	15	0	0	**17**
$n_{.j}$	**2**	**40**	**31**	**7**	**80**

DEPENDENCIA DE VARIABLES

Definición

Dos variables son dependientes cuando el conocimiento del valor de una de ellas en un individuo aporta información sobre el valor de la otra en ese individuo.

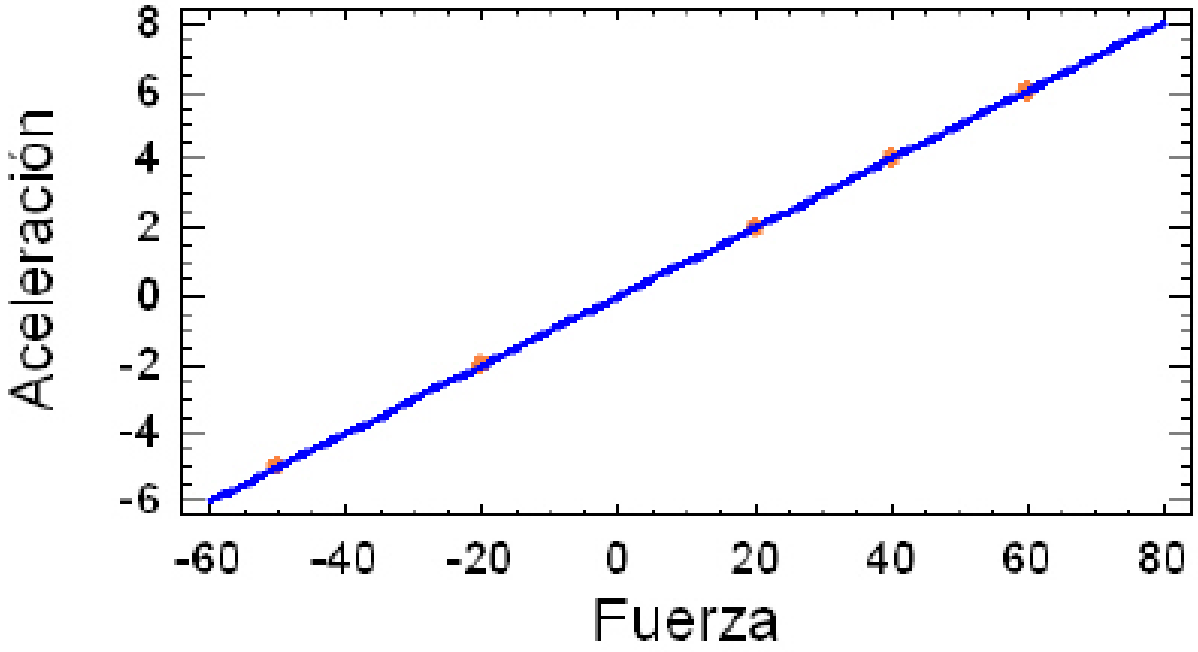
Dependencia determinista

Si a un cuerpo de masa **m** se le aplica una fuerza *F*, esta fuerza comunica una aceleración al cuerpo, cuyo módulo viene expresado por la ecuación:

$$a = \frac{F}{m}.$$

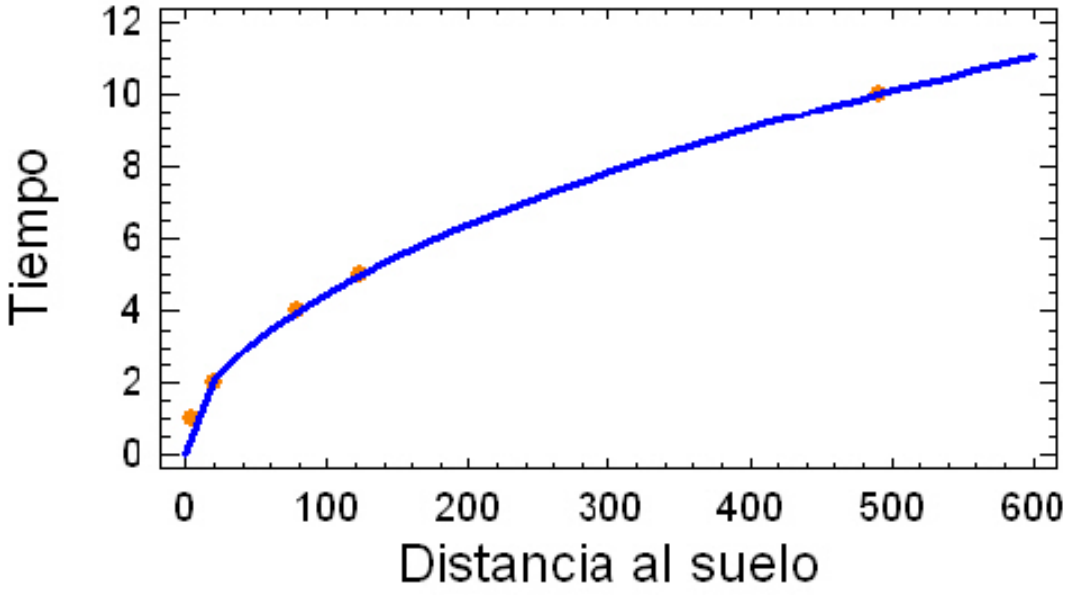
DEPENDENCIA DE VARIABLES

El siguiente gráfico muestra los distintos valores de las aceleraciones provocadas sobre un cuerpo de masa 10 Kg, por distintas fuerzas ejercidas sobre él.



DEPENDENCIA DE VARIABLES

El siguiente gráfico muestra los distintos valores del tiempo transcurrido hasta que un cuerpo en caída libre alcanza el suelo, en función de la distancia entre éste y el punto en el que inicia la caída.



$$t = \sqrt{\frac{2e}{g}}$$

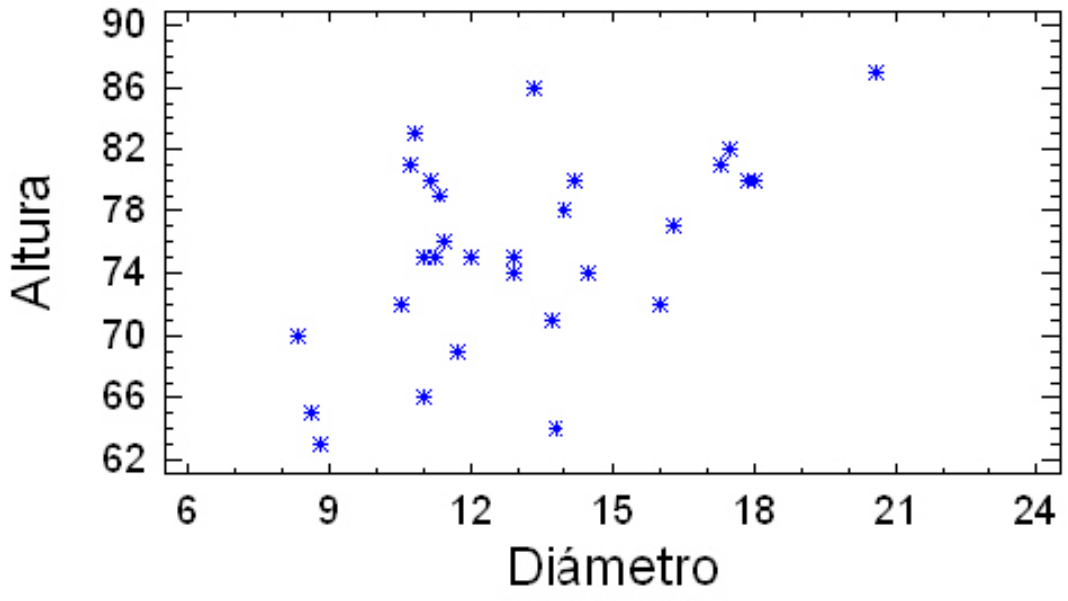
DEPENDENCIA DE VARIABLES

En ocasiones, cuando dos variables son **dependientes**, **NO** se puede calcular con exactitud el valor de una variable cuando el de la otra es conocido.

En estos casos se dice que la relación de dependencia entre las variables es **estadística** o **aleatoria**.

DEPENDENCIA DE VARIABLES

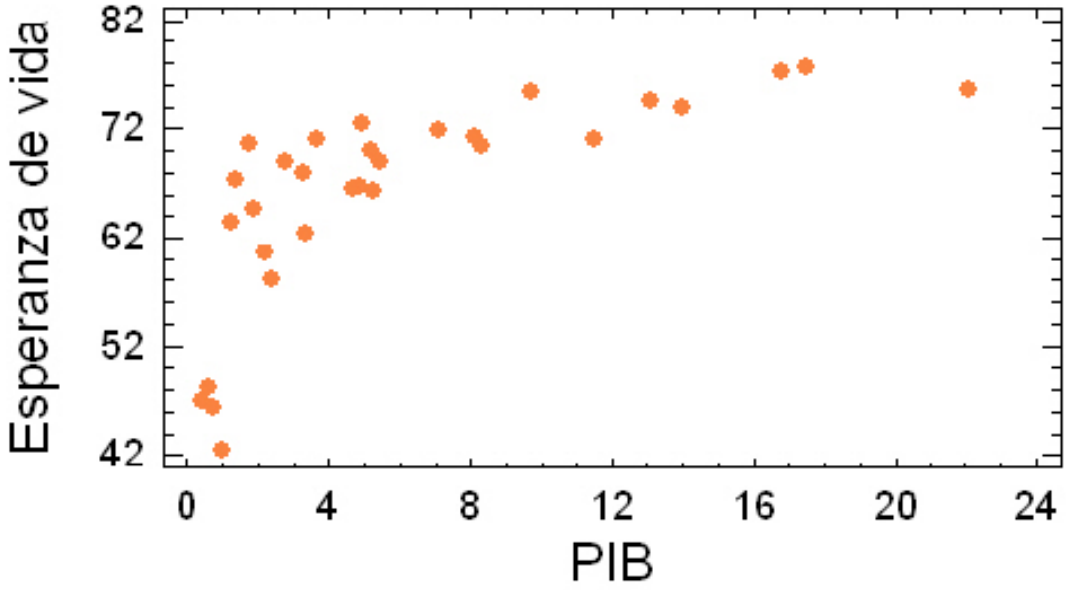
El siguiente gráfico representa los diámetros en la base del tronco, y las alturas, de un conjunto de cerezos.



¿Qué altura le corresponde a un cerezo que tenga un diámetro en la base de 14 unidades?

DEPENDENCIA DE VARIABLES

El siguiente gráfico representa la **esperanza de vida** en un conjunto de países en función de su **producto interior bruto**, (en el gráfico las unidades del PIB son miles de millones de dólares).



¿Qué esperanza de vida le corresponde a un país que tenga un PIB de 15 unidades?
¿Y a otro con un *PIB de 5 unidades?*

DEPENDENCIA DE VARIABLES

Problema

En los casos de dependencia estadística **no existe** un modelo **matemático** (ecuación) que permita calcular con exactitud el valor de una variable, cuando la otra es conocida.

Solución

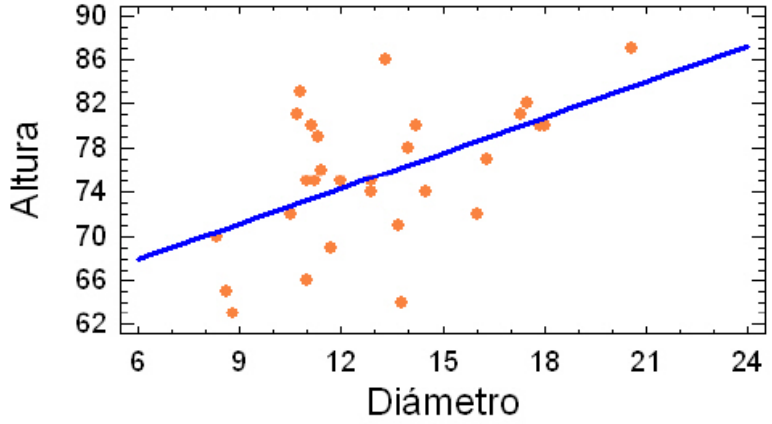
En ocasiones se puede establecer un modelo que permita calcular, de **manera aproximada**, el valor de una variable aleatoria, cuando el de la otra, también aleatoria, es conocida.

Regresión

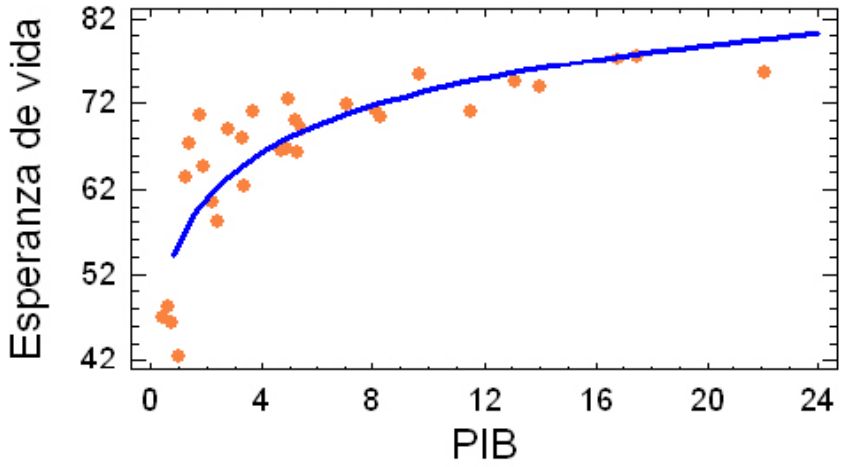
Búsqueda de una función que relacione ambas variables y sirva para predecir una variable a partir de la otra

$$y=f(x)$$

DEPENDENCIA DE VARIABLES



y=61'55+1'06X
*y=61'55+1'06*14=76'47*



*y=2'03+7'76*ln(x)*
*y=2'03+7'76*ln(5000)=68'12*

DEPENDENCIA DE VARIABLES

Resumen

- *Cuando dos variables son dependientes, el conocimiento del valor de una de ellas aporta información sobre el valor de la otra.*

- *En el caso de dependencia funcional, conocido el valor de una de las variables, la ecuación del modelo, y = f (x), permite el cálculo exacto del valor de la otra.*

- *En el caso de dependencia estadística, el conocimiento del valor de una variable aleatoria permite, sólo, el cálculo aproximado del valor de la otra.*

DEPENDENCIA DE VARIABLES

En el caso en que la nube de puntos sugiera una **relación lineal, con forma de recta**, entre las variables, existen dos coeficientes que complementan la información gráfica:

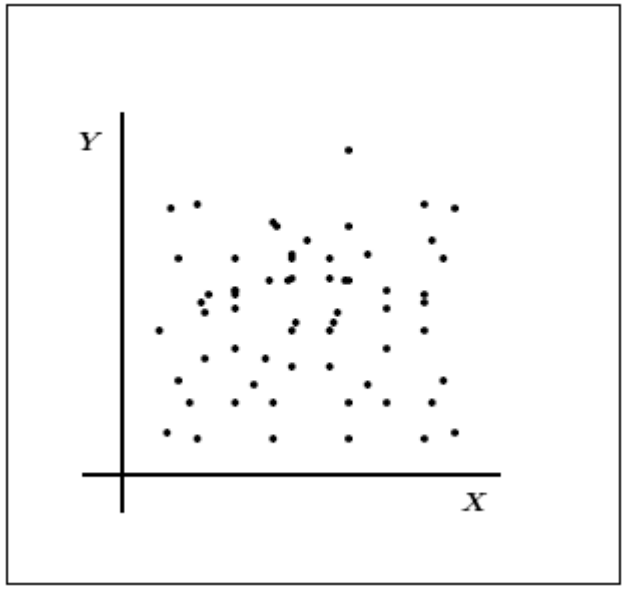
- Covarianza
- Coeficiente de correlación lineal.

COVARIANZA Y SUS PROPIEDADES

- El coeficiente de covarianza se construye para medir la intensidad de la dependencia lineal entre dos variables. Mide la manera en que dos variables varían juntas.
- Es una medida de variación conjunta de X e Y
- La covarianza es una medida del grado de variación conjunta entre dos variables estadísticas, respecto a sus medias. Se obtiene mediante la siguiente fórmula:

$$Cov(X,Y) = m_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})(y_j - \overline{y})n_{ij}}{N} = \frac{\sum_{i=1}^{n} x_i y_i}{N} - \overline{xy}$$

COVARIANZA Y SU INTERPRETACIÓN



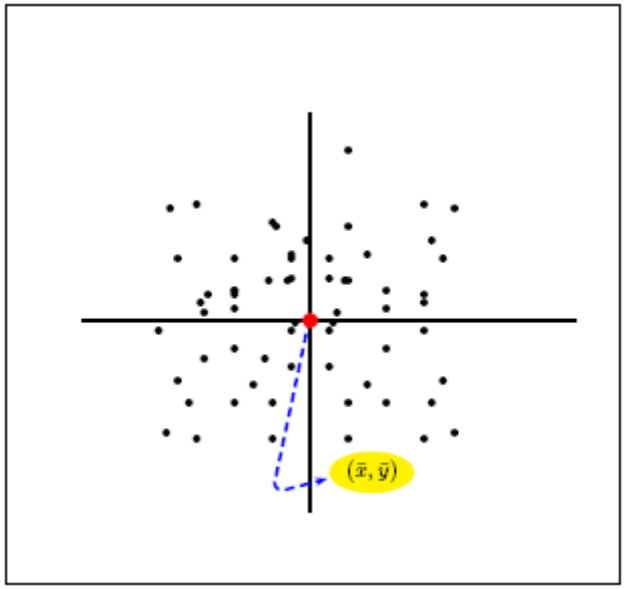
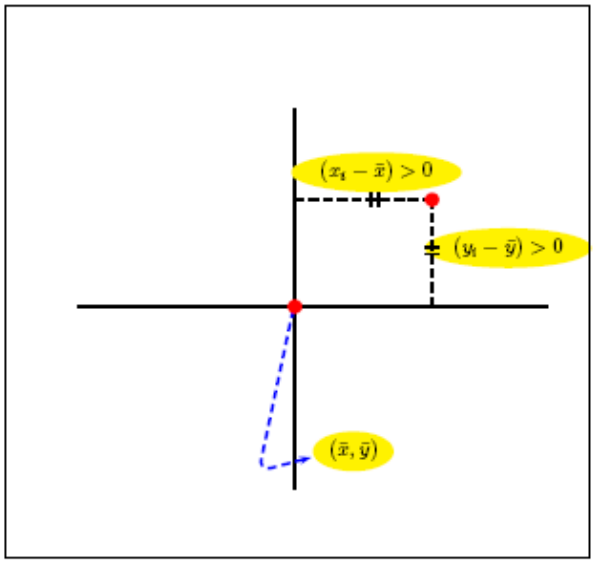
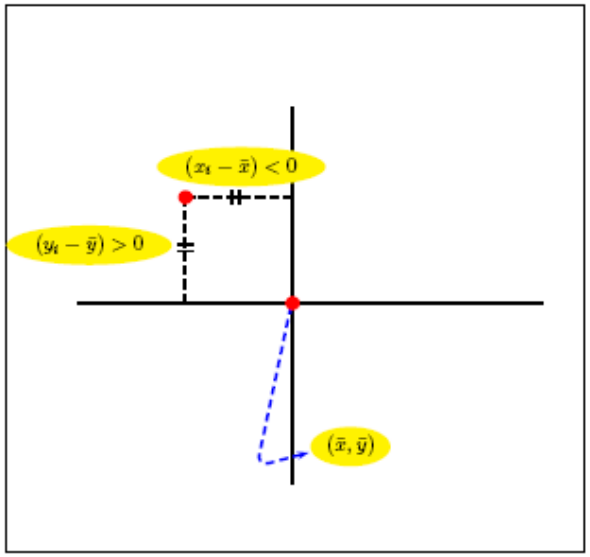
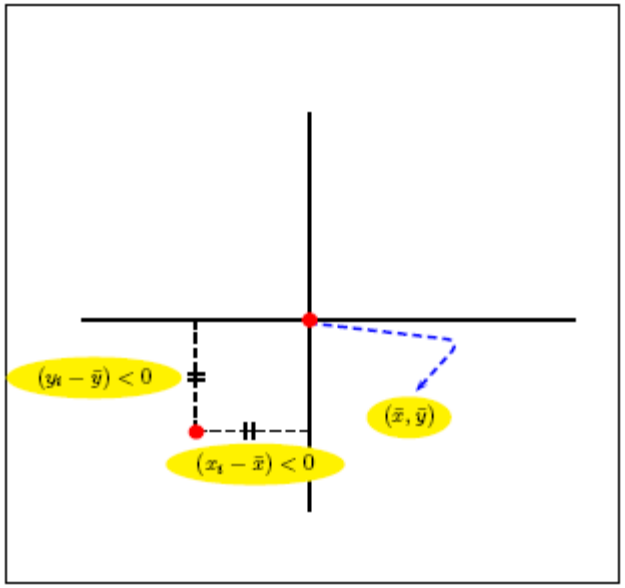


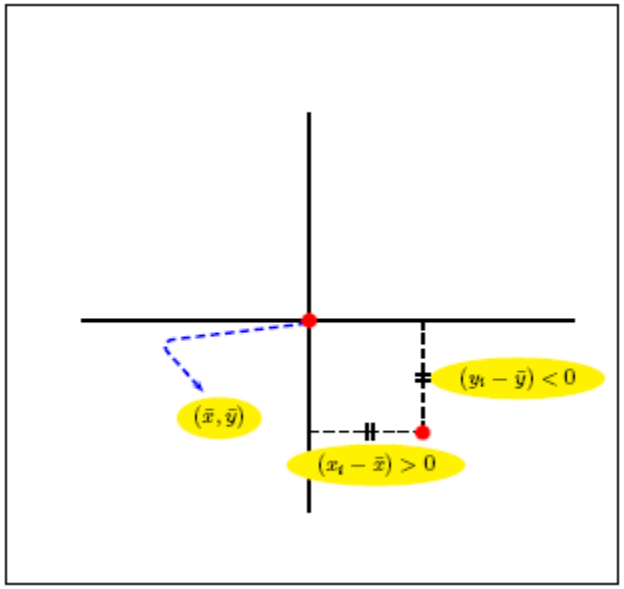
Diagrama de dispersión → Traslación a las medias

COVARIANZA Y SU INTERPRETACIÓN



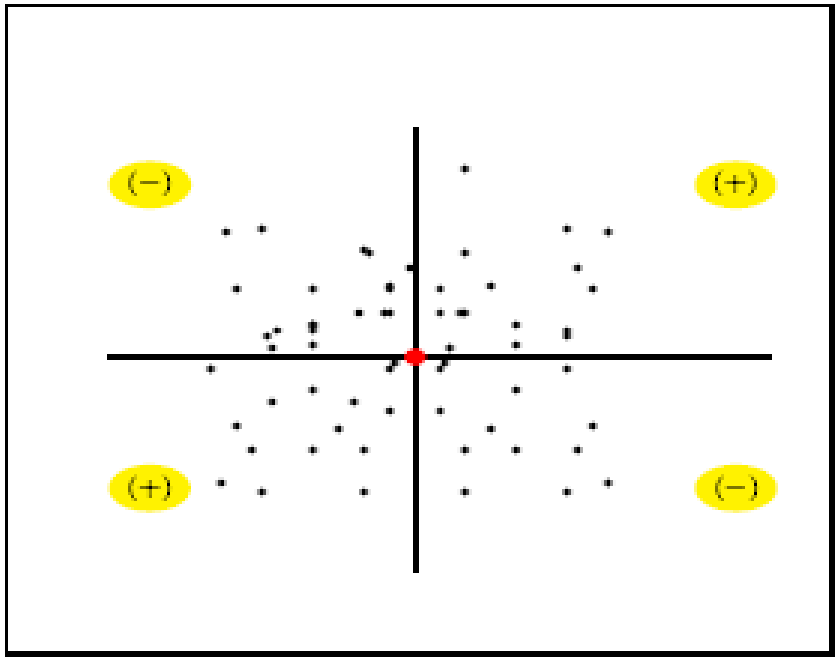






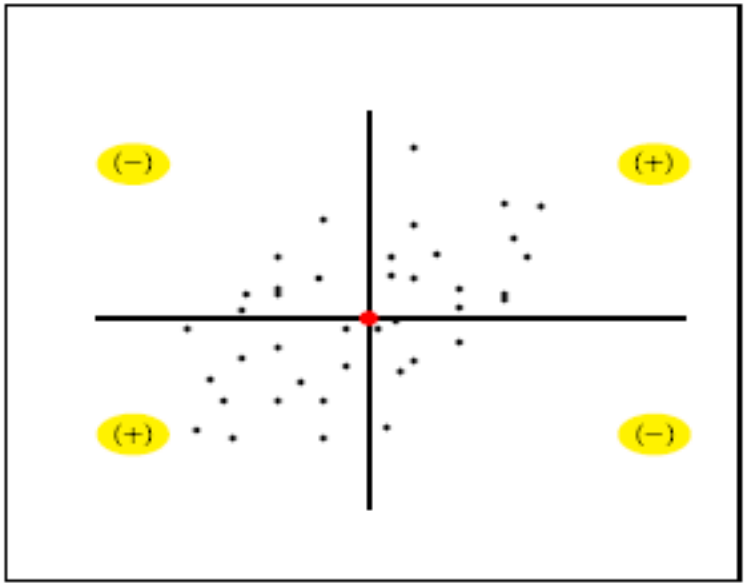
COVARIANZA Y SU INTERPRETACIÓN

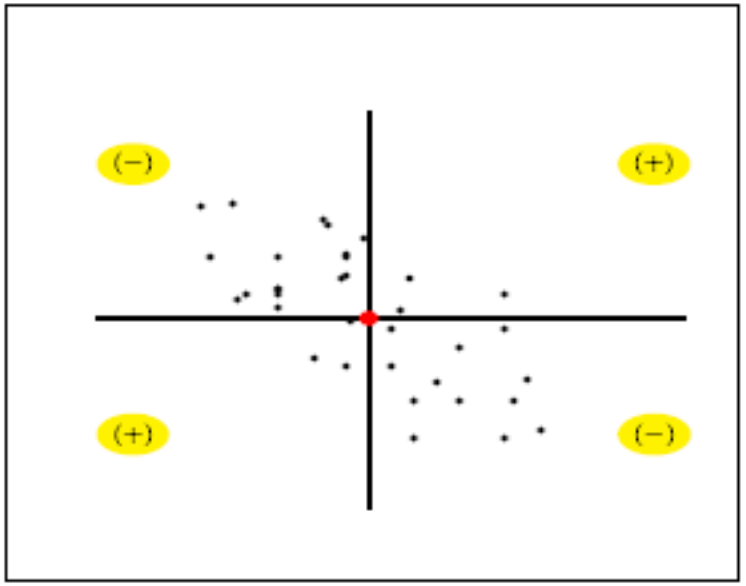
Por lo tanto, en distribuciones de puntos como las de las cabe esperar un coeficiente de covarianza **próximo a cero**.



COVARIANZA Y SU INTERPRETACIÓN

Sin embargo, en distribuciones de puntos como las de estas figuras cabe esperar un coeficiente de **covarianza alto** en valor absoluto.





COVARIANZA Y SU INTERPRETACIÓN

- El valor de la covarianza en caso de independencia estadística es $S_{xy} = 0$.
- Lo contrario es no necesariamente cierto, es decir, una covarianza nula no implica necesariamente independencia.
- Si las variables presentan una relación positiva (cuando una crece la otra también crece) la covarianza será positiva. Si la relación entre las variables es negativa, la covarianza también lo será.

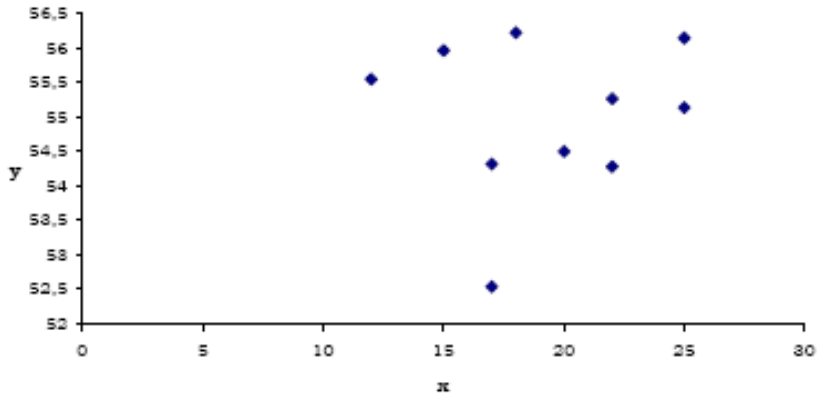
EJERCICIO DE REPASO

Una empresa ha registrado los siguientes datos sobre las variables X, gastos mensuales destinados a la promoción del producto, en miles de euros; e Y, cantidad de unidades vendidas al mes, en unidades de millar. Los datos se refieren a los últimos diez meses, siendo la última columna los del mes pasado.

X_i	17	20	22	25	18	15	12	17	25	22
Y_i	52.54	54.50	55.27	56.15	56.23	55.97	55.55	54.32	55.14	54.28

Se trata de determinar la correlación entre ambas variables y comentar el valor obtenido, indicando el valor que resultaría si se manejasen datos en euros y unidades de artículo.

EJERCICIO DE REPASO



$$Cov(X,Y)=\frac{\sum_{i=1}^{N} x_i y_i}{N}-\overline{xy}=1061.73\text{-}19.3\times54.995=0.3265$$

Se trata de un valor que no podemos precisar si es alto o bajo ya que se refiere a las unidades manejadas.

COVARIANZA Y SU INTERPRETACIÓN

La covarianza entre dos variables:

- Cov(x,y) > 0: X e Y tienden a moverse en la misma dirección
- Cov(x,y) < 0: X e Y tienden a moverse en direcciones opuestas.
- Cov(x,y) = 0: X e Y no están relacionadas linealmente

COVARIANZA Y SU INTERPRETACIÓN

La covarianza tiene unidades, las de la variable X multiplicadas por las de la variable Y .

La covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables.

La covarianza no tiene escala y se puede hacer, en valor absoluto, arbitrariamente grande o pequeña con el mismo conjunto de datos.

$$Cov(aX+b, cY+d) = acCov(X,Y)$$

$$Cov(X,X) = S_x^2$$

COEFICIENTE DE CORRELACIÓN LINEAL

Para evitar los problemas con la covarianza, Pearson define el coeficiente de correlación lineal entre dos variables como:

$$r = \frac{Cov(X,Y)}{S_x S_y}$$

Siendo S_x y S_y las desviaciones típicas de cada una de las variables

- Nos indica si los puntos tienen tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales)
- Tiene el mismo signo que la covarianza
- No sirve para otro tipo de relaciones

COEFICIENTE DE CORRELACIÓN LINEAL

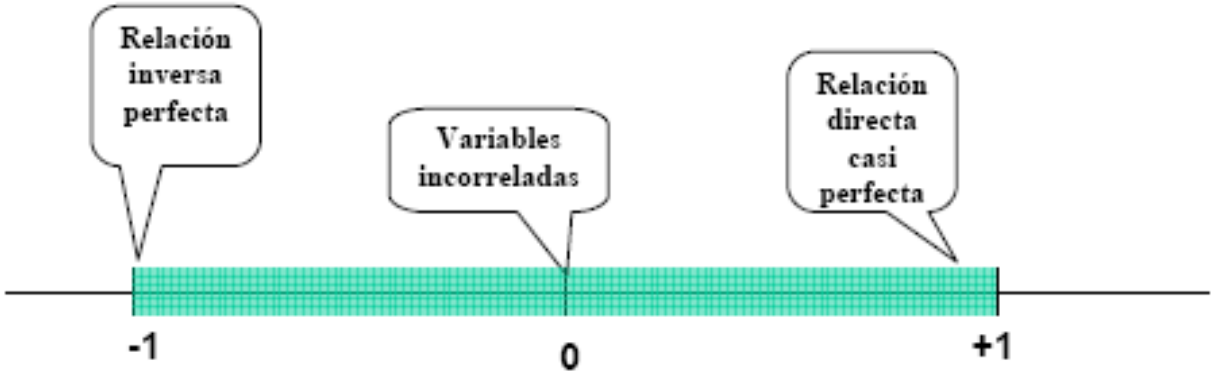
El coeficiente de correlación tiene las siguientes propiedades:

- Es un número adimensional.
- Su valor se encuentra siempre entre:

$$-1 \leq r \leq 1$$

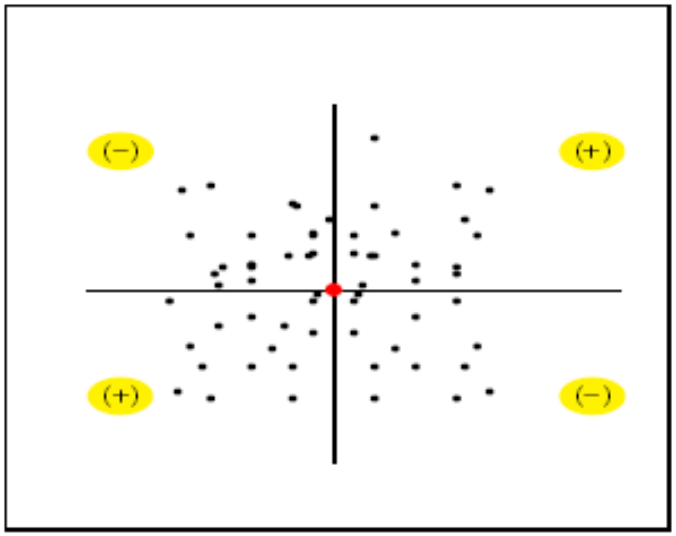
- r = +1 implica dependencia lineal positiva exacta entre X e Y .
- r= -1 implica dependencia lineal negativa exacta entre X e Y
- r= 0 implica falta de dependencia lineal entre X e Y .

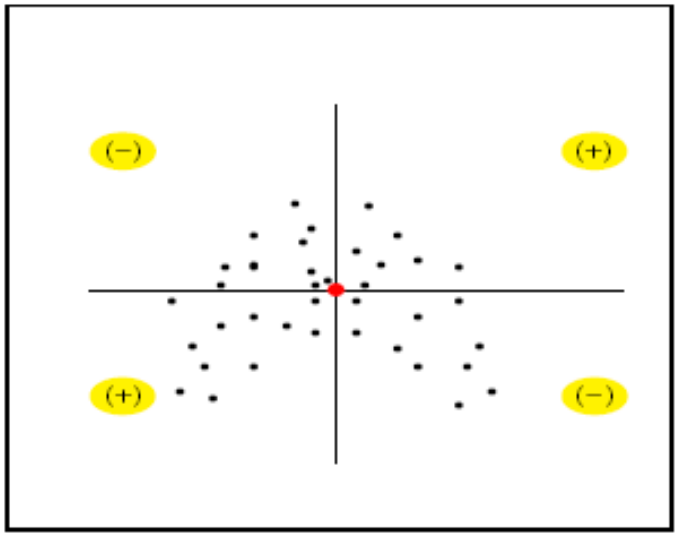
r=0,967: una relación fuerte y positiva



COEFICIENTE DE CORRELACIÓN LINEAL

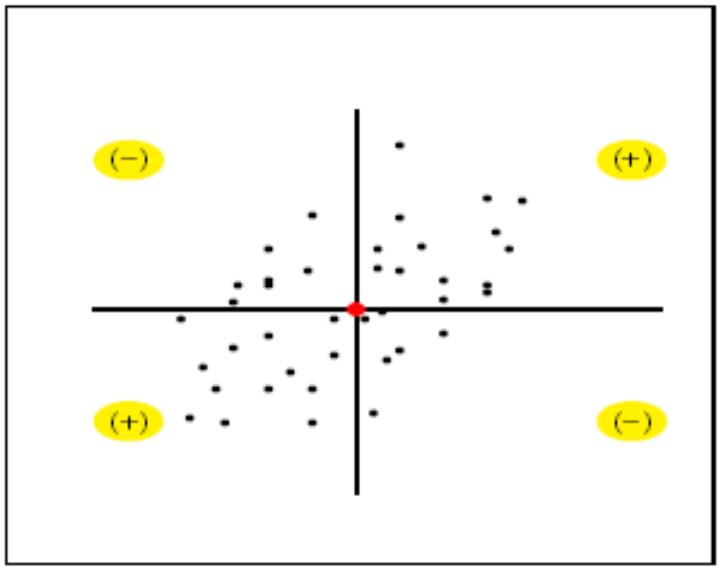
r estará próximo a cero

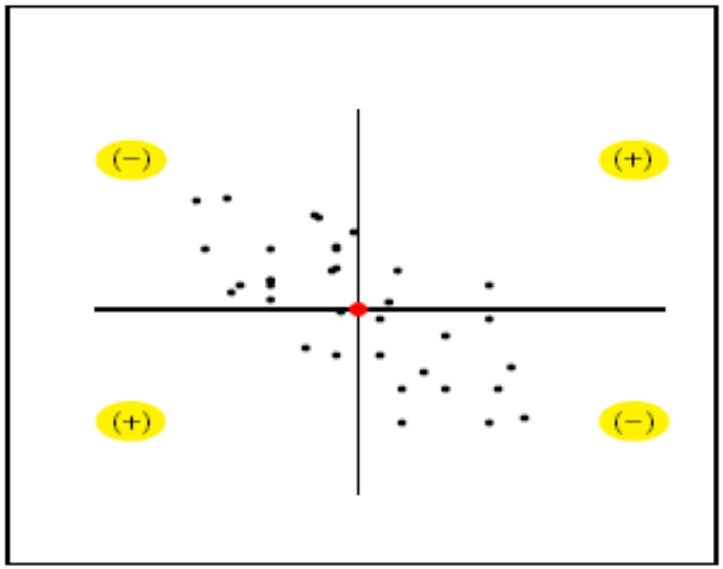




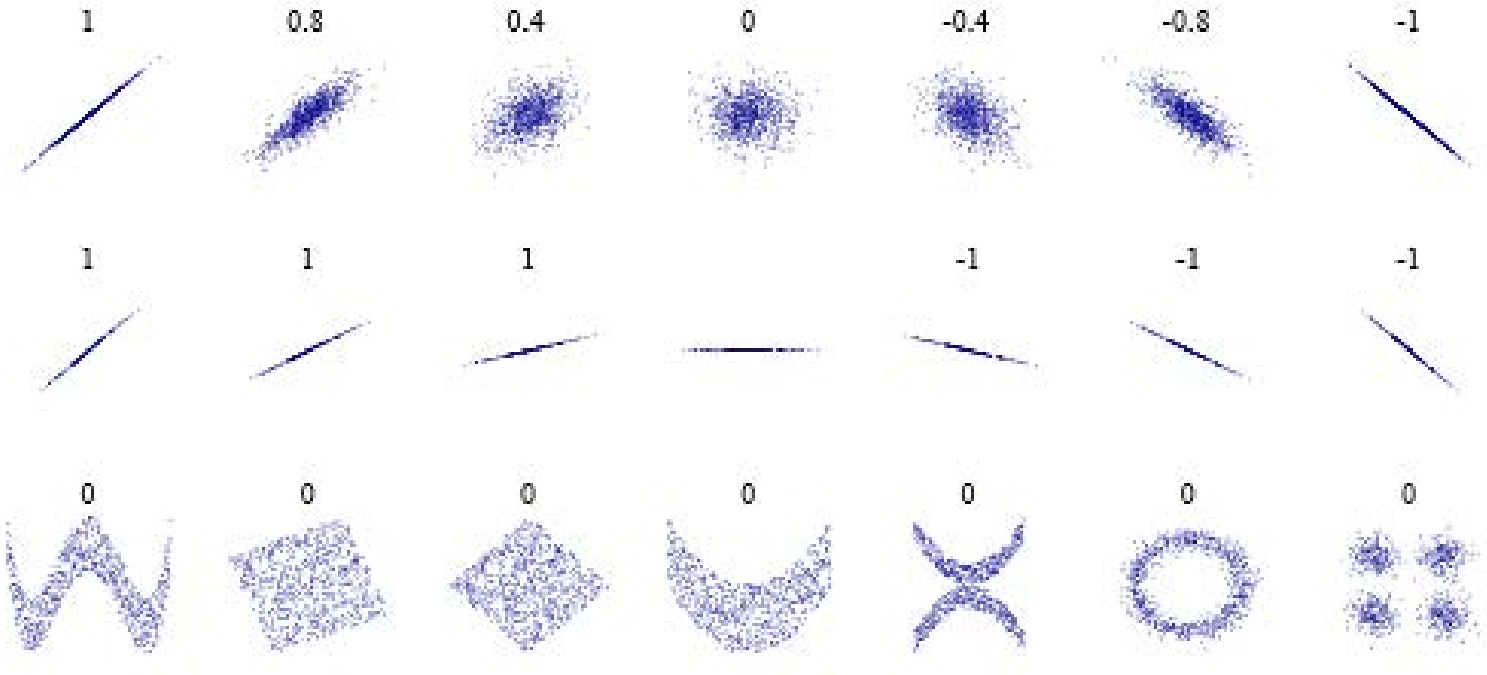
COEFICIENTE DE CORRELACIÓN LINEAL

r estará próximo a uno y menos uno





Coeficiente de correlación



Correlación y causalidad

Homer: *No hay siquiera un oso a la vista. ¡La "patrulla anti-osos" funciona de maravilla!*

Lisa: *Eso es un razonamiento falaz, Papá.*

Homer [sin comprender]: *Gracias, hija.*

Lisa: *Usando tu lógica, yo puedo afirmar que esta roca aleja a los tigres.*

Homer: *Hmmm, ¿y cómo funciona?*

Lisa: *No funciona. (pausa) ¡Es sólo una roca estúpida!*

Homer: *Ajá.*

Lisa: *Pero no veo ningún tigre alrededor, ¿y tú?*

Homer: *(. . . pausa . . .) Lisa, quiero comprar tu roca.*

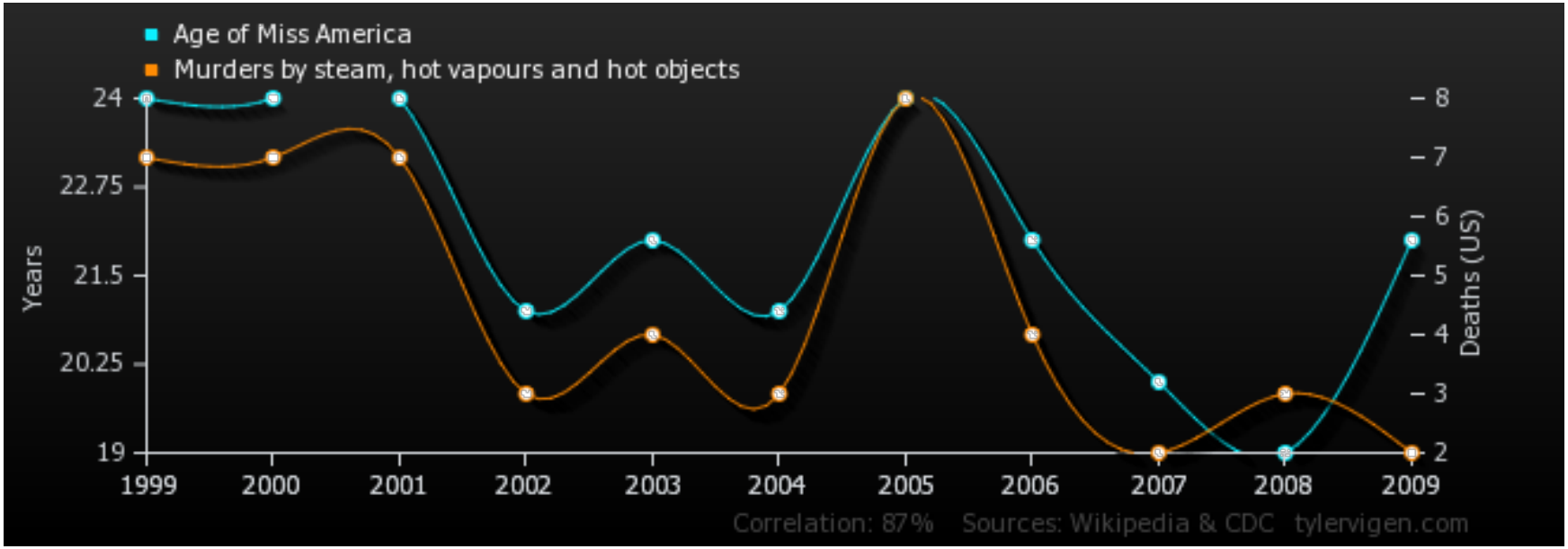
Correlación y causalidad

- ***correlación no implica causalidad*** (en latín, Cum hoc ergo procter hoc) es cierta lo que se quiere decir es que el hecho de **que haya correlación entre dos variables no significa que una provoque a la otra, pero eso no significa que si encontramos correlación entre dos variables automáticamente podamos descartar** que una sea causa de la otra.

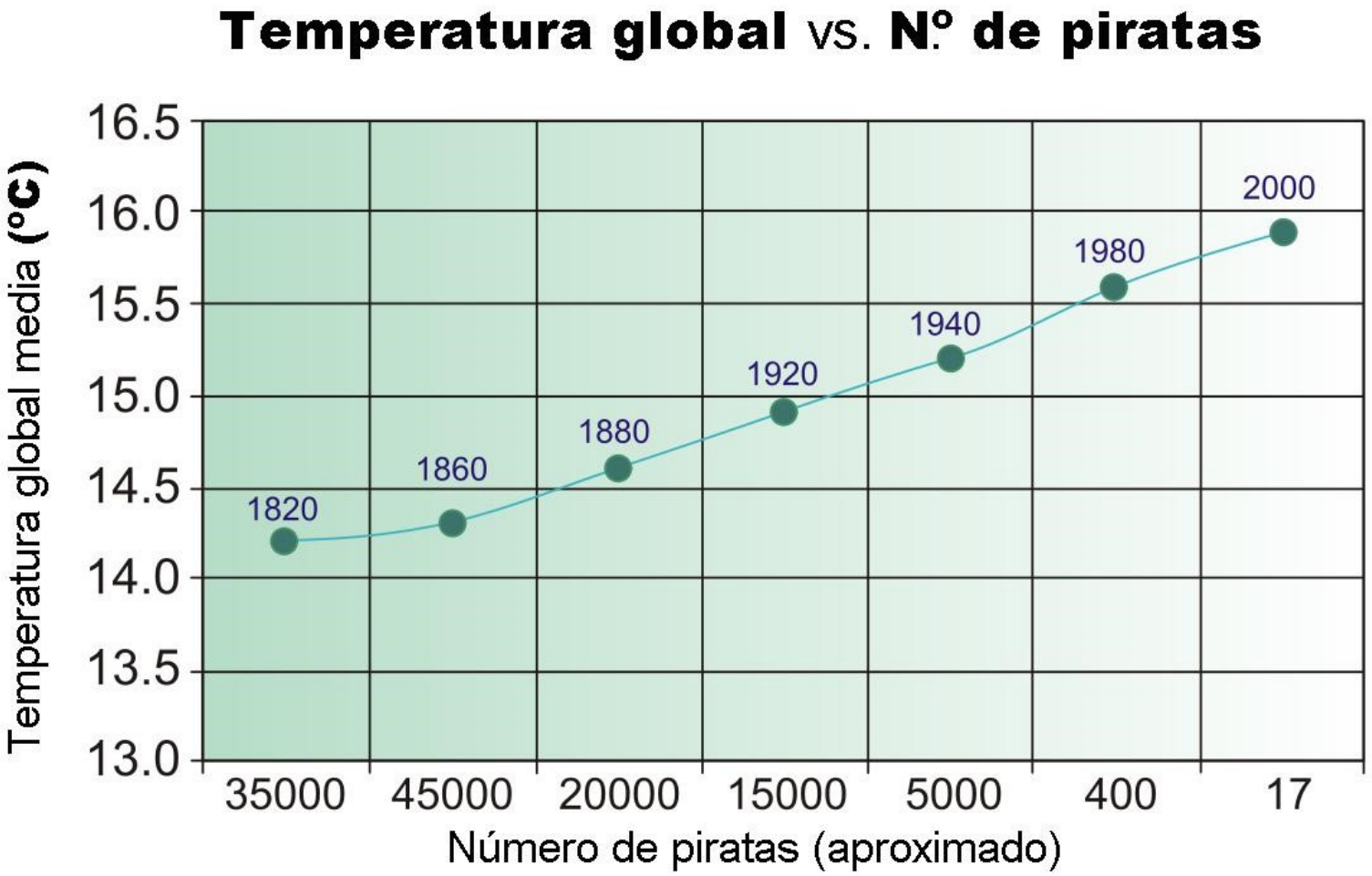
Motivos:

- Que A cause B
- Que B cause A (falacia de dirección)
- Que haya un tercer fenómeno, C, que provocara tanto A como B (relación espuria)
- Puro y duro azar.

Age of Miss America Years	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects Deaths (US) (CDC)	7	7	7	3	4	3	8	4	2	3	2
Correlation: 0.870127											



Correlación y causalidad



http://www.tylervigen.com/spurious-correlations

EJERCICIO DE REPASO

Determine el coeficiente de correlación de la siguiente tabla de datos donde figura los años de escolarización completados y el número de pulsaciones.

Años de escolarización	**12**	**16**	**13**	**18**	**19**	**12**	**18**	**19**	**12**	**14**
Número de pulsaciones	73	67	74	63	73	84	60	62	76	71

Cov = -15,29 R= -0,7638

La correlación mide la asociación de las variables **NO la causalidad**

Regresión

Definición.- Regresión

La regresión pretende encontrar la estructura de dependencia que mejor explique el comportamiento de una variable Y a la que denominaremos (variable dependiente, explicada o endógena) a partir de un conjunto de variables $X_1, X_2, \cdots, X_p$ (variables independientes, explicativas o exógenas) relacionadas con Y.

Definición.- Regresión lineal simple

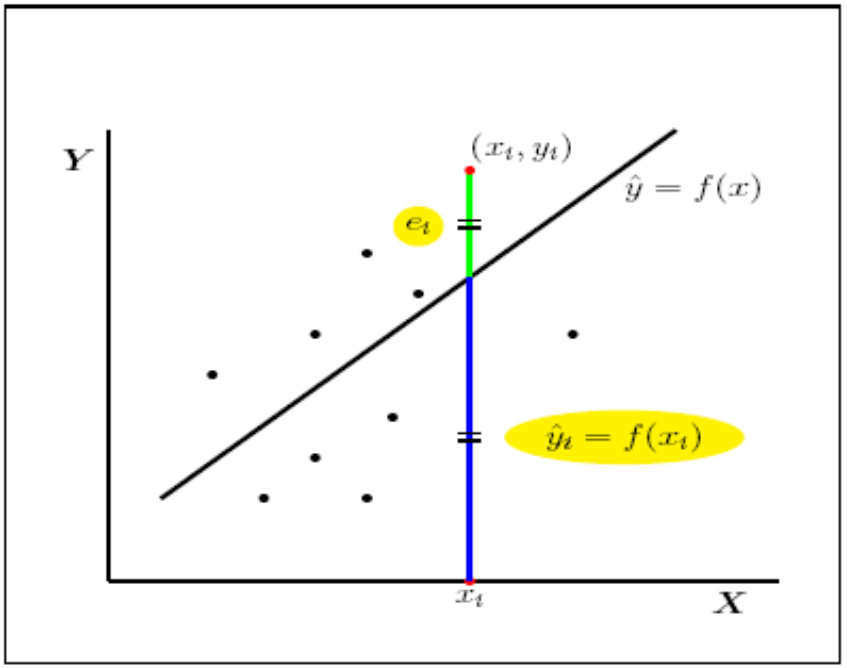
La regresión lineal simple pretende encontrar la recta que mejor explica el comportamiento de la variable dependiente Y a partir del comportamiento de una única variable X.

Romero-Ramos,

MODELO DE REGRESIÓN LINEAL Y CORRELACIÓN

Sea (x_i , y_i) un punto correspondiente a un dato cualquiera del conjunto:

y_i se puede descomponer como se describe en el gráfico:

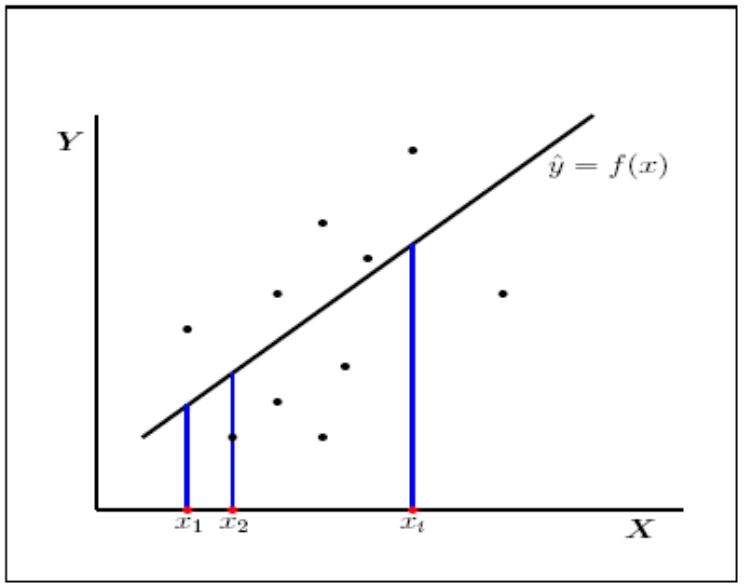


La parte inferior, representa el valor que el modelo prevé para la variable Y , en un individuo cuyo valor en X es x_i

La parte superior, e_i , es la diferencia entre el valor observado de Y en el individuo y_i , y el previsto por el modelo

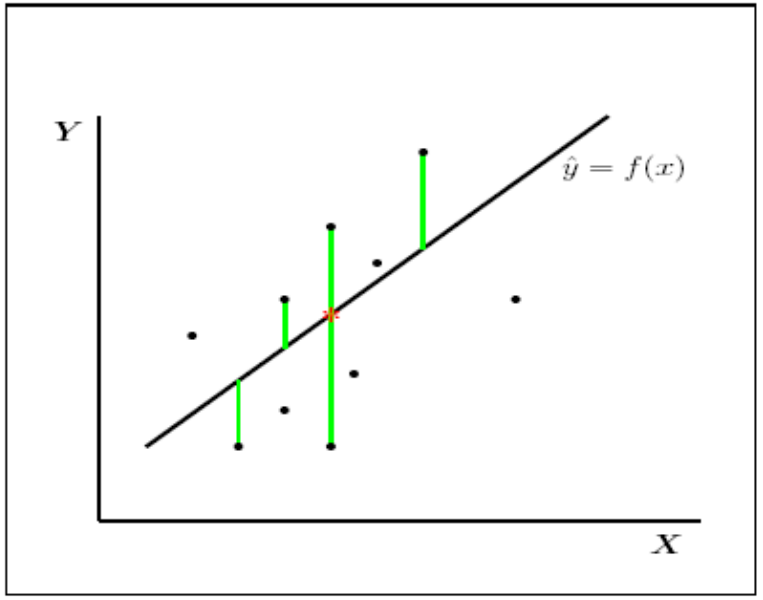
MODELO DE REGRESIÓN

Calculado el modelo, el valor de $\hat{y}_i$ queda determinado para cada x_i, $\hat{y}_i = f(x_i)$

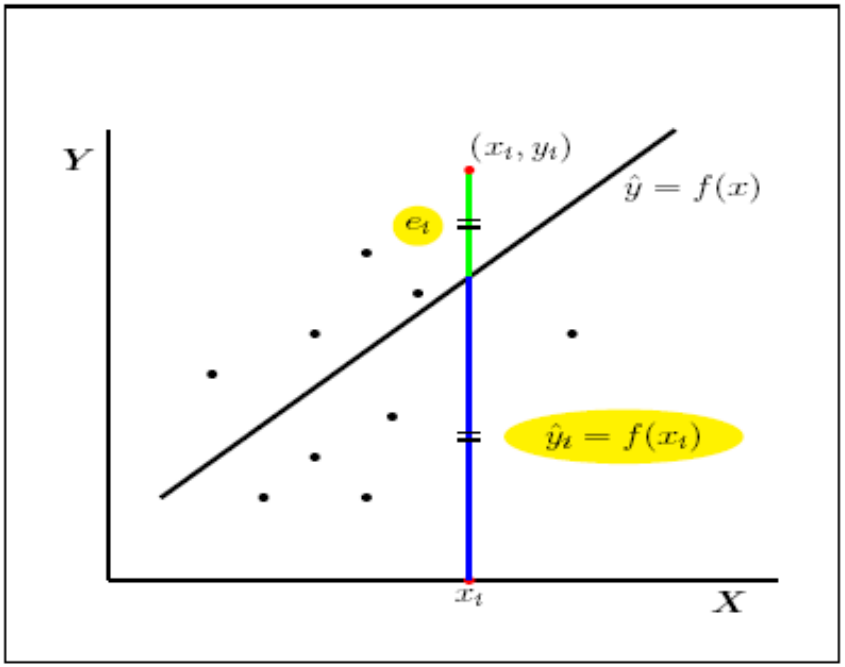


$\hat{y}_i = f(x_i)$ es la parte **determinista, o funcional** del modelo.

$e_i = y_i - \hat{y}_i$ es la parte **aleatoria** del modelo. (Error aleatorio.)



MODELO DE REGRESIÓN



El modelo de regresión simple vendrá determinado por

$$\underbrace{y_i}_{\text{Valor observado}} = \underbrace{f(x_i)}_{\text{Parte determinista, } \hat{y}_i} + \underbrace{e_i}_{\text{Error aleatorio}}$$

RECTAS DE REGRESIÓN

¿Cómo hacemos mínimos los errores del modelo?

Criterio de mínimos cuadrados:

- Sea $e = (e_1, e_2, \ldots, e_n)$ el vector de errores asociado al modelo.

- El módulo de este vector viene dado por la expresión:

$$|e| = \sqrt{e_1^2 + e_2^2 + \ldots + e_n^2}$$

- El criterio de mínimos cuadrados selecciona los valores de los parámetros del modelo que ***minimizan*** el módulo del vector error

RECTAS DE REGRESIÓN

Dado el conjunto de datos bidimensionales nos planteamos ajustarlo a un recta

Tenemos que minimizar:

$$\sum (y_i - (ax_i + b))^2 \equiv SCD(a,b)$$

Elegimos la recta que minimiza la suma de los cuadrados de la diferencia

$$\frac{\partial SCD(a,b)}{\partial a} = 0 \quad y \quad \frac{\partial SCD(a,b)}{\partial b} = 0$$

$$a_0 = \frac{S_{xy}}{S_{xx}} = \frac{Cov(X,Y)}{S_x^2} \quad y \quad b_0 = \overline{Y} - a_0 \overline{X}$$

RECTAS DE REGRESIÓN

Recta de regresión mínima cuadrática de Y sobre X

$y=a_0x+b_0$

$$y = \frac{S_{xy}}{S_{xx}}x + \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x}$$

$$y - \overline{y} = \frac{S_{xy}}{S_{xx}}(x - \overline{x}) = \frac{Cov(X,Y)}{S_x^2}(x - \overline{x})$$

RECTAS DE REGRESIÓN

Propiedades:

- La recta de regresión pasa necesariamente por $(\overline{x}, \overline{y})$
- El signo de su pendiente lo marca la covarianza
- La media de los errores es cero
- Se llama coeficiente de regresión Y/X:

$$b_{yx} = \frac{S_{xy}}{S_{xx}} = \frac{Cov(X,Y)}{S_x^2}$$

Que es la **pendiente** de la recta Y sobre X

RECTAS DE REGRESIÓN

Recta de regresión mínima cuadrática de X sobre Y

$X=c_0y+d_0$

$$c_0 = \frac{S_{xy}}{S_{yy}} = \frac{Cov(X,Y)}{S_y^2} \quad y \quad d_0 = \bar{X} - c_0\bar{Y}$$

$$x - \bar{x} = \frac{S_{xy}}{S_{yy}}(y - \bar{y}) = \frac{Cov(X,Y)}{S_y^2}(y - \bar{y})$$

Coeficiente de regresión de X sobre Y:

$$b_{xy} = \frac{S_{xy}}{S_{yy}} = \frac{Cov(X,Y)}{S_y^2}$$

Que es la pendiente de la recta X sobre Y

RECTAS DE REGRESIÓN

b_{YX}

- *>0 Recta de regresión de Y sobre X creciente*
- *=0 Recta de regresión de Y sobre X horizontal*
- *<0 Recta de regresión de Y sobre X decreciente*

b_{XY}

- *>0 Recta de regresión de X sobre Y creciente*
- *=0 Recta de regresión de X sobre Y vertical*
- *<0 Recta de regresión de X sobre Y decreciente*

En general ambas rectas no coinciden y se cruzan en el punto de las medias. Será necesario estudiar ambas rectas.

EJEMPLO

Xi	Yi
1	1,5
2	2
3	4
5	4,6
6	4,7
8	8,5
9	8,8
10	9,9

Calcular la recta de regresión de y vs x

CORRELACIÓN

Cuando las variables son estadísticamente independientes, su covarianza es cero. Por tanto, si las variables son independientes, también están incorreladas linealmente, es decir, $r = 0$.

Sin embargo, dos variables pueden estar incorreladas linealmente y ser (incluso fuertemente) dependientes, ya que cuando $r = 0$ lo único que podemos decir es que la dependencia estadística lineal es nula, pero las variables pueden estar relacionadas mediante otro tipo de función (exponencial, hiperbólica, etc.)

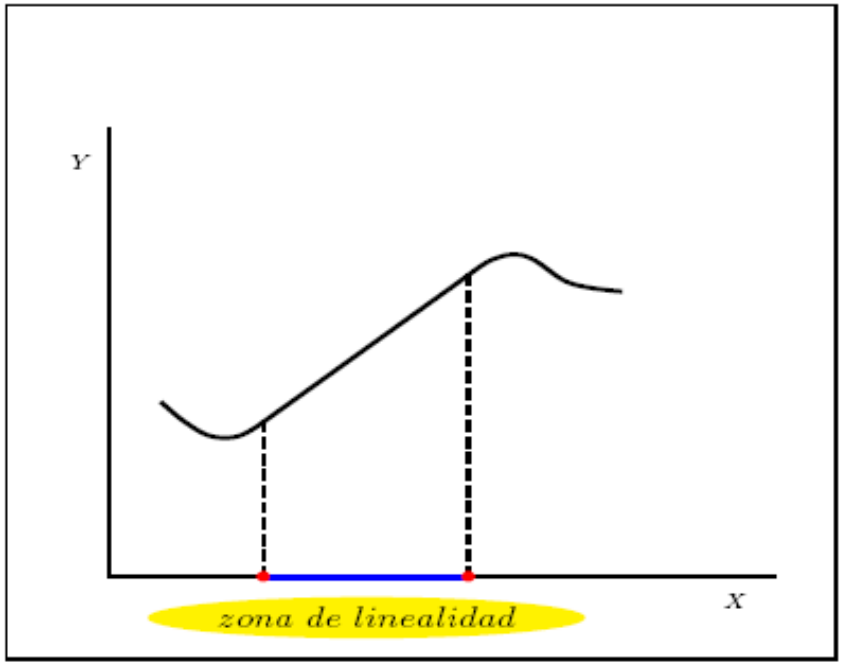
Se puede demostrar la invarianza de r ante transformaciones lineales.

PREDICCIÓN

Predicción de una nueva observación de Y dado un valor de X

$$\hat{y}_i = a_0 x + b_0$$

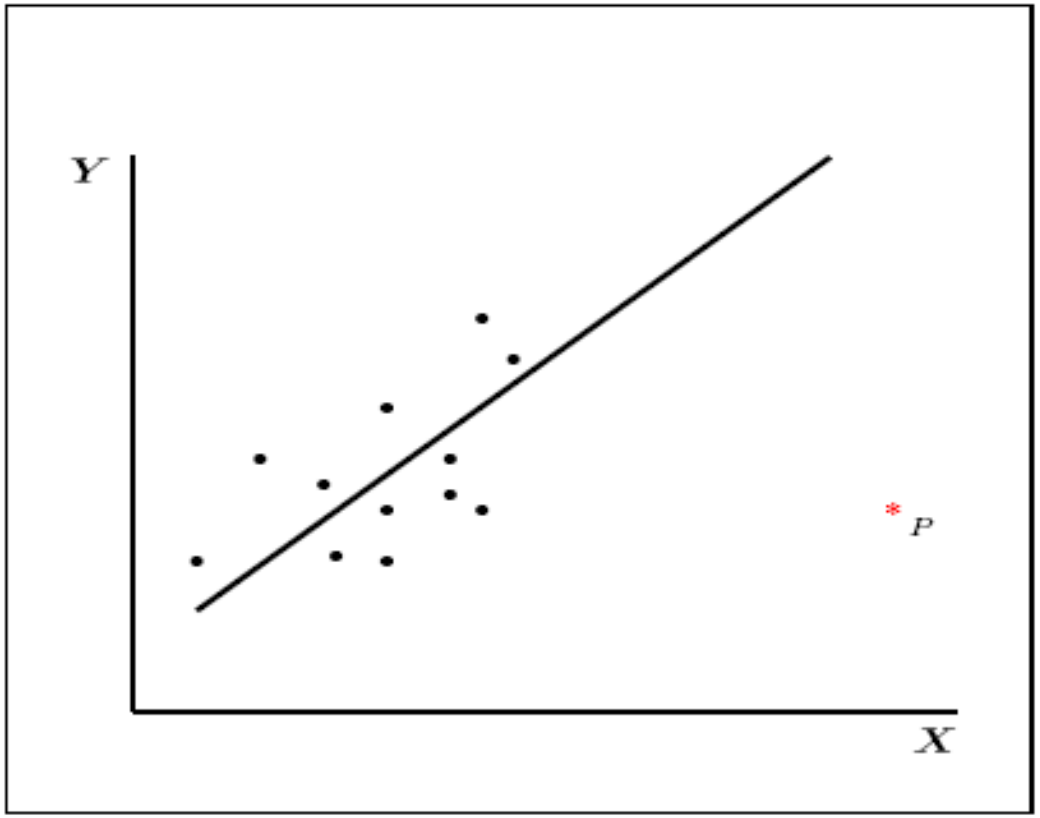
- X: variable independiente, explicativa o regresora
- Y : variable dependiente o respuesta
- La predicción sólo tiene sentido en el rango de valores X e Y



PREDICCIÓN

Valores atípicos en la regresión

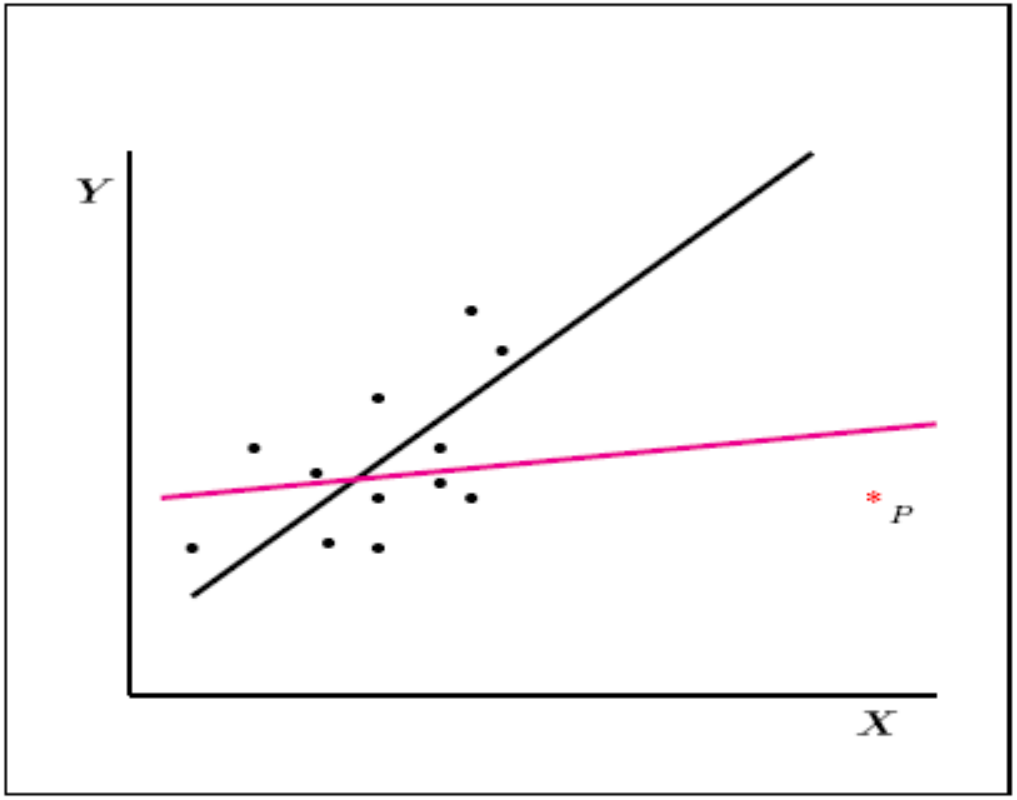
Recta de regresión calculada sin el valor atípico P



PREDICCIÓN

Valores atípicos en la regresión

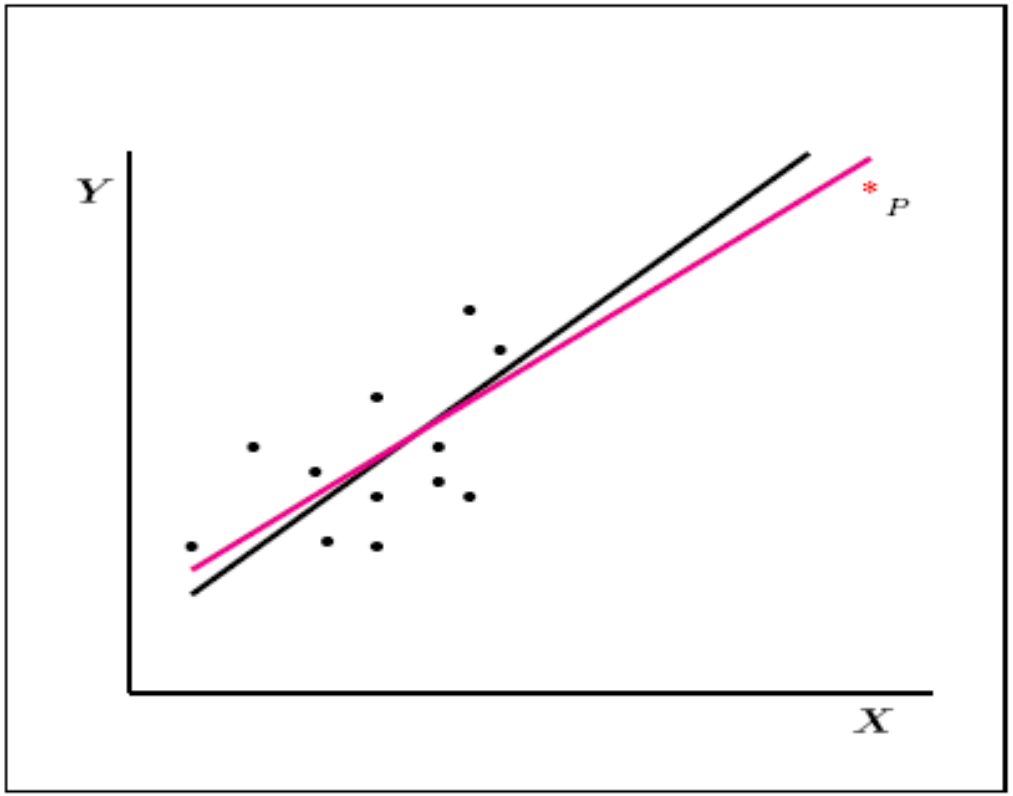
P es un valor atípico influyente



PREDICCIÓN

Valores atípicos en la regresión

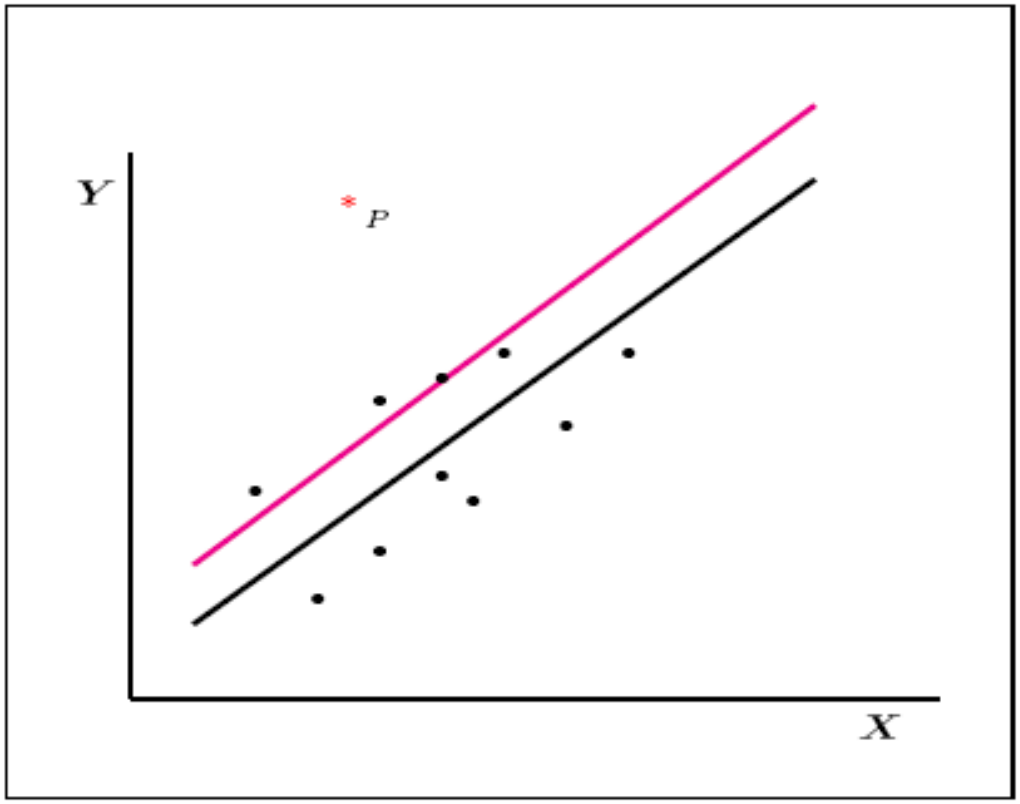
El punto P NO es influyente, puesto que su inclusión NO modifica sustancialmente la recta de regresión.



PREDICCIÓN

Valores atípicos en la regresión

La inclusión del punto P
no supone variación significativa en la pendiente de la recta de
regresión estimada.



Actuación ante valores atípicos

Si en un análisis se observan valores atípicos, una estrategia recomendable es la siguiente:

1. Descartar que se trata de un error.
2. Analizar si el punto es influyente.
3. Si el punto es influyente, calcular las rectas de regresión incluyéndole y excluyéndole, eligiendo la que mejor se adapte al conocimiento del problema y a las observaciones futuras.

Observación: En caso de duda, se debe utilizar el modelo con precaución. No se debe descartar, en ningún caso, recabar más información.

COEFICIENTE DE DETERMINACIÓN

- Una forma de medir la bondad del ajuste y por tanto la fiabilidad de las estimaciones
- El coeficiente de determinación se interpreta como el porcentaje de variación de la variable dependiente explicado por el modelo.
- En modelos de regresión lineal simple, se calcula simplemente como el cuadrado de coeficiente de correlación lineal:

Minimizamos la expresión anterior y tenemos el coeficiente de determinación

$$D = \sum_{i=1}^{n} [y_i - f(x_i, a, b, \ldots)]^2$$

Para una curva cualquiera

$$R^2 = 1 - \frac{\sum_{i=1}^{n} [y_i - f(x_i)]^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

DESCOMPOSICIÓN DE LA SUMA DE CUADRADOS

Variación total de los datos de respuesta:

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Si y_i =a (es una constante) entonces SCT=0
- A mayor dispersión de los datos mayor SCT
- $\frac{SCT}{N} = S_y^2$

DESCOMPOSICIÓN DE LA SUMA DE CUADRADOS

Variación entre el valor de la **predicción y la media** de los valores ajustados. **Suma de Cuadrados debida a la regresión.**

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Si los valores de la predicción coinciden con los ajustados para todo n entonces SCT=SCR
- Si los valores coinciden con la media entonces SCR=0 ➔Cov(X,Y)=0 ➔ ausencia total de relación lineal.
- $$SCR = \frac{S_{xy}^2}{S_{XX}}$$

DESCOMPOSICIÓN DE LA SUMA DE CUADRADOS

Variación en la respuesta alrededor de la recta de regresión.
Suma de Cuadrados debida al error (residual)

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Si SCE=0 ➔ los valores coinciden ➔relación lineal perfecta
- Si los valores coinciden con la media entonces SCE=SCT ➔ ausencia total de relación lineal.

$$e_i = y_i - \hat{y}_i \Rightarrow SCE = \sum_{i=1}^{N} e_i^{\,2}$$

DESCOMPOSICIÓN DE LA SUMA DE CUADRADOS

SCT=SCR+SCE o lo que es lo mismo *VT=VE+VNE*

Coeficiente de determinación entre Y y X

$$r^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

Mide el efecto que tiene la variable explicativa a la hora de reducir las variaciones en la variable respuesta➔ proporción en que se reduce la variabilidad de Y al introducir la relación lineal con la X

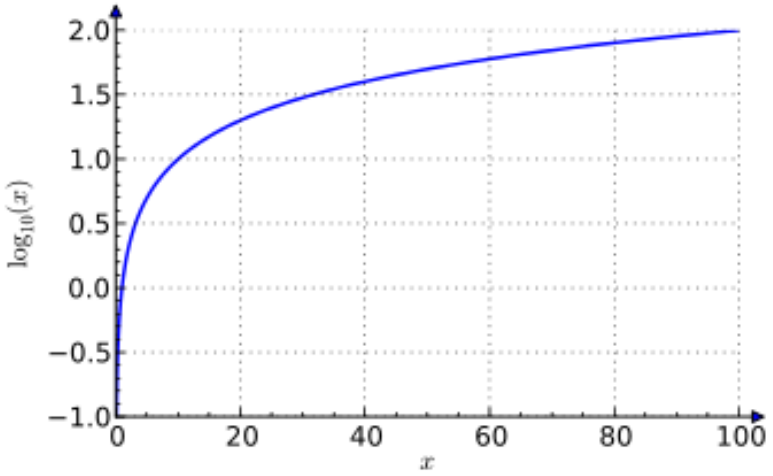
DESCOMPOSICIÓN DE LA SUMA DE CUADRADOS

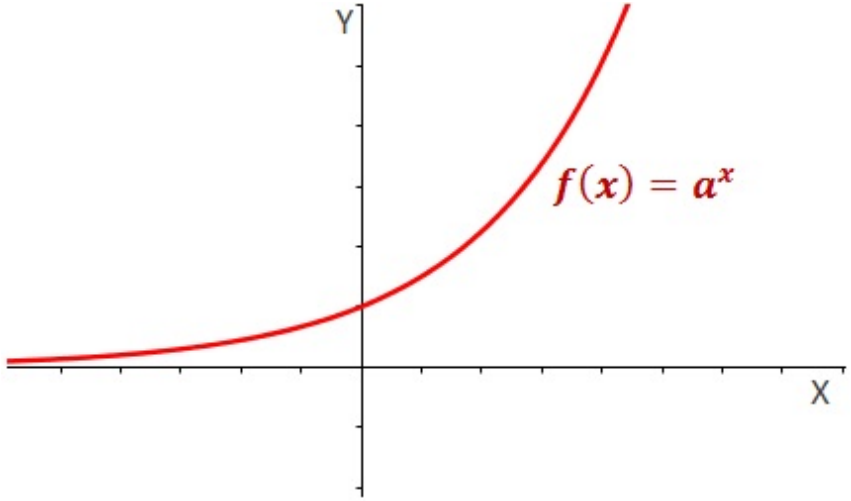
Propiedades

- Se verifica que su valor está entre 0 y 1
- $r^2=1 \Leftrightarrow$ SCE=0 $\Leftrightarrow$ SCT=SCR Toda la variación en los datos se debe a la relación lineal
- $r^2=0 \Leftrightarrow$ SCR=0 $\Leftrightarrow$ SCE=SCT $\Leftrightarrow$ No hay relación lineal. Toda la variación se debe al error.
- se verifica que $r^2 = b_{yx}b_{xy} = \frac{Cov^2(X,Y)}{S_x^2 S_y^2}$
- Definimos el coeficiente de correlación lineal de Pearson como

$$r = \pm\sqrt{r^2} = \frac{Cov(X,Y)}{\sqrt{S_x^2 S_y^2}}$$

Regresión no lineal





Regresión no lineal

$$\hat{y} = c \cdot e^{bx}$$

$$\ln \hat{y} = \ln c + bx \begin{cases} u = \ln \hat{y} \\ k = \ln c \end{cases}$$

$$\hat{u} = k + bx$$

Mediante regresión lineal se calculan b y k. C se halla a partir de k ($c=e^k$)

$$\hat{y} = c \cdot x^b$$

$$\ln \hat{y} = \ln c + b \ln x \begin{cases} u = \ln \hat{y} \\ v = \ln x \\ k = \ln c \end{cases}$$

$$\hat{u} = k + bv$$

$$\hat{y} = a + b \ln x$$

$$u = \ln x$$

$$\hat{y} = a + bu$$

EJERCICIO 1

Se ha realizado una encuesta a 474 empleados de una compañía multinacional. Entre los datos recogidos consta el salario anual (en miles) y los años de educación. Al realizar el diagrama de dispersión asumiendo que el salario depende de los años de educación se observa la siguiente nube de puntos:

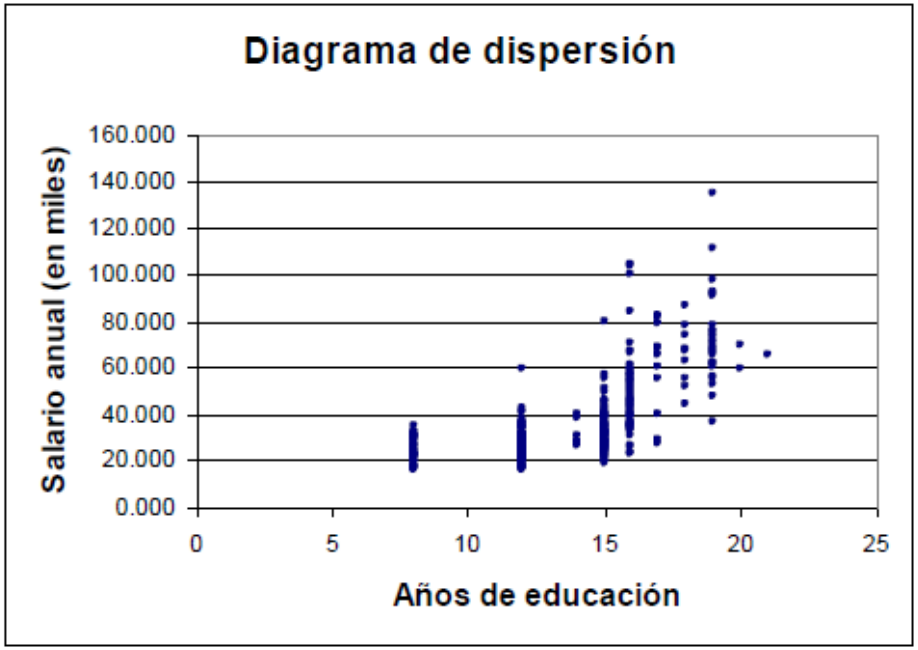
Señala cual de las siguientes opciones es la correcta:

a) La covarianza debe ser positiva y la correlación negativa.

b) La covarianza debe ser positiva y la correlación positiva.

c) La covarianza debe ser negativa y la correlación negativa.

d) La covarianza debe ser negativa y la correlación positiva.



EJERCICIO 2

Se ha realizado una encuesta a 474 empleados de una compañía multinacional. Entre los datos recogidos consta el salario anual (en miles) y los años de educación. Suponiendo Y=Salario, X=Años de educación

Varianza X = 8,305 Varianza Y = 290,963 Covarianza = 32,471

Señala cual es el valor correcto de la correlación:

a)-0,53
b) 0,066
c)-0,662
d) 0,662

EJERCICIO 3

En una oficina se desea conocer el grado de satisfacción de los empleados. Para ello se realiza un cuestionario de satisfacción a 10 de ellos y se les pide que valoren, en una escala continúa de 0 a 10, el ambiente en su puesto de trabajo. El valor 0 identifica un pésimo ambiente de trabajo y el 10 identifica un inmejorable ambiente de trabajo. Además se recoge la edad de los empleados.
Asumiendo que la valoración depende de la edad se ha estimado la recta de regresión obteniéndose: $\hat{y}_i = 6.13 - 0.087\, x_i$

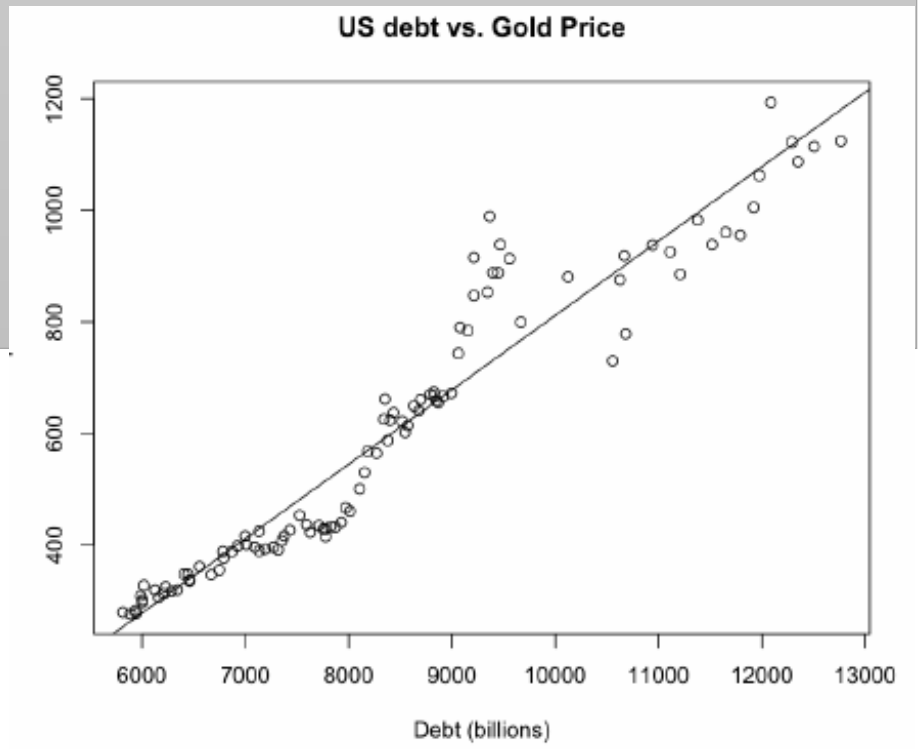
Ahora se desearía conocer cual es la valoración media para un nuevo trabajador cuya edad es 43 años. Di cual de las siguientes opciones es la correcta:

a)2.19 puntos
b)2.39 puntos
c)4.69 puntos
d)-2.05 puntos

Ejercicio 4

El diagrama muestra el nivel de la deuda Americana como función del precio de oro. La fórmula para la recta de regresión es:
PRECIO DE ORO (nominal) = -522,86 + (0,1334 * deuda en $ billones)
Si la deuda Americana es de $19000 billones, calcular la predicción para el precio de oro.

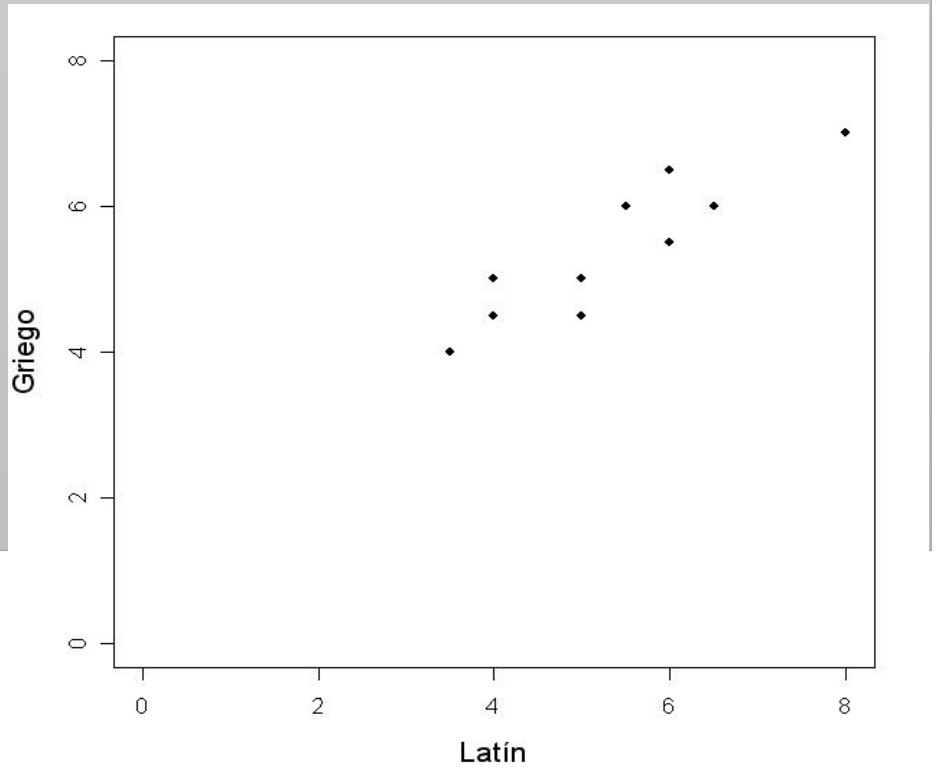
a)2011,74
b)2933,14
c)3057,46
d)-520,32



Ejercicio 5

El gráfico siguiente muestra los niveles de conocimiento de Griego y de Latín para **10 jueces**. Llamamos Y al nivel de conocimiento de Griego y X al nivel de conocimiento de Latín. Si utilizamos la nota de Latín para determinar la nota en Griego mediante una recta de regresión, observando el diagrama de dispersión, ¿cuál de las opciones mostradas abajo podría ser la recta correcta?

a)Y=1.97+0.64X
b)Y=1.97-0.64X
c)Y=-1.97+0.64X
d)Y=-1.97-0.64X



EJEMPLO

Xi	Yi
1	1,5
2	2
3	4
5	4,6
6	4,7
8	8,5
9	8,8
10	9,9

Calcular la recta de regresión de y vs x

EJERCICIO

Se realiza un estudio con la influencia de la dosis de un determinado fármaco en el nivel de glucosa en sangre al cabo de unos minutos desde la aplicación del fármaco. Los datos se recogen en la siguiente tabla:

Dosis	30	40	50	60	70
Glucemia	180	170	140	130	130

Medimos el nivel de glucemia de un paciente y es 165, teniendo en cuenta la tabla de datos anterior, ¿podríamos predecir cuál sería la dosis que debemos suministrar al paciente con ese nivel de glucemia? Justificar la respuesta. En caso afirmativo, cuál sería la dosis recomendada?

Algunas referencias

https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia-JS/interactivos2/coef_correlacion-JS/index.html

https://mlu-explain.github.io/linear-regression/