REGRESIÓN

Y

CORRELACIÓN

ESTADÍSTICA

DESCRIPTIVA

trabaja con

Poblaciones

Muestras

e incluye las siguientes etapas

- **R**ecopilación u obtención
- **O**rganización y reducción
- **P**resentación
- **A**nálisis e interpretaciones

a partir de las muestras se calculan

Estimadores

y con aporte de las

- **Teoría de probabilidades**
- **Distribuciones de probabilidad**
- **Distribuciones de muestreo**

INFERENCIAL

a partir de lo informado por la muestra,

llevará a hacer dos tipos de generalización

- **E**stimación
- **P**rueba de hipótesis

sobre los **Parámetros(*)**

(*) Tambien pueden ser no paramétricos

Estudio de 2 o más variables con diferentes objetivos.

Tipo de Variables

Análisis de Regresión Simple:……………..

Correlación:……………………

Objetivo

Análisis de Regresión Simple:……………..

Correlación:……………………

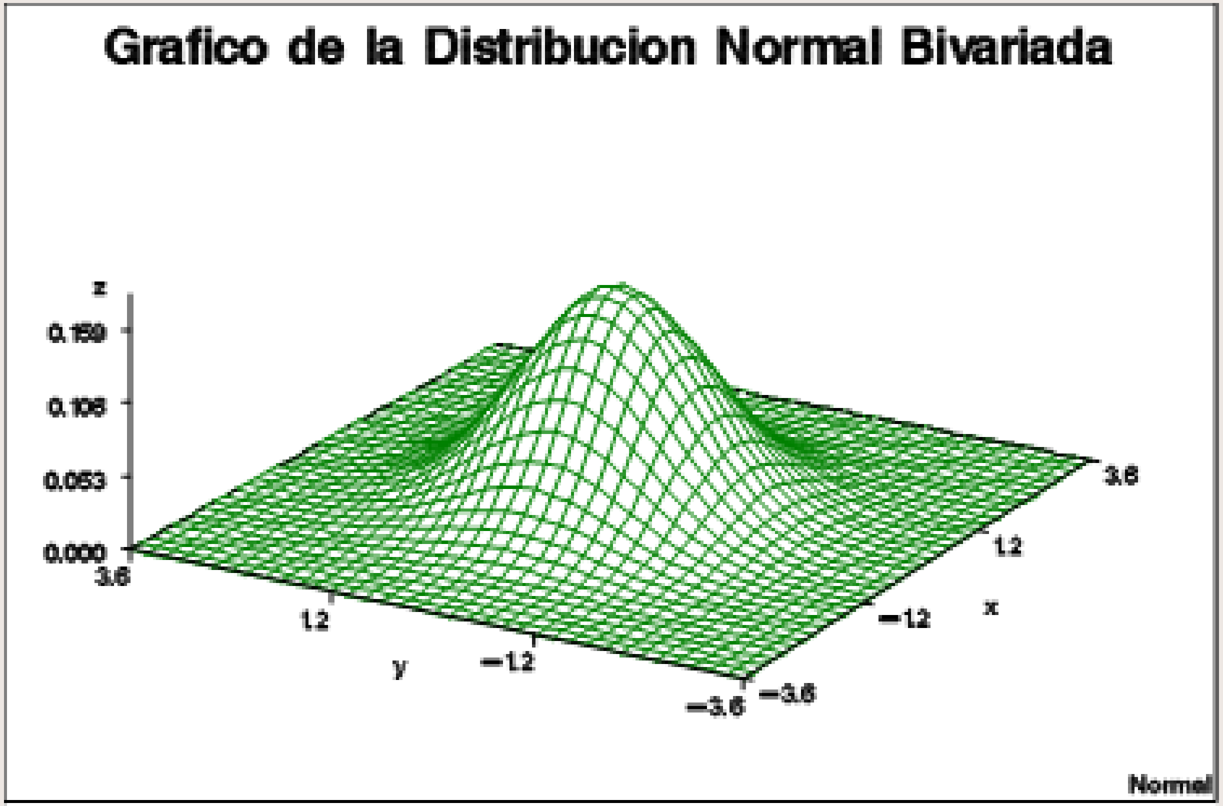
Regresión - Correlación

Análisis que requieren la consideración de 2 o más variables cuantitativas en forma simultánea.

Análisis de Regresión: estudia la relación funcional de una o más variables respecto de otra.

Análisis de Correlación: estudia la magnitud o grado de asociación entre las variables.

CORRELA CIÓN



SUPUESTOS CORRELACIÓN

X y Y tienen una distribución de probabilidad conjunta normal bivariada, esto es:

a. para cada valor de X existe una subpoblación de valores de Y que siguen una distribución normal.

b. Para cada valor de Y existe una subpoblación de valores de X que siguen una distribución normal.

Coeficiente de correlación

Poblacional: ρ

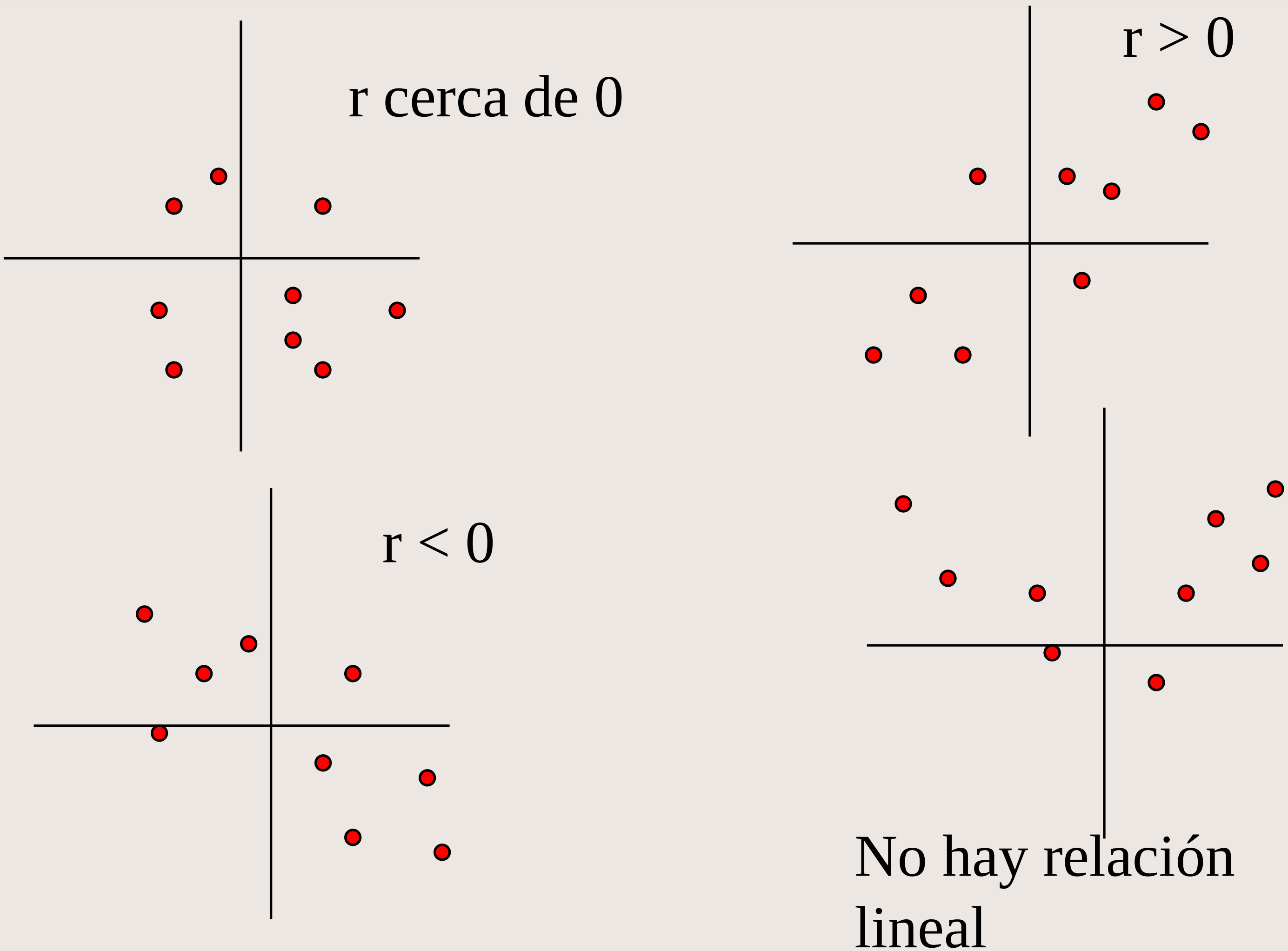
Muestral: r

$$\rho = \frac{Cov(xy)}{\sqrt{Var(x)Var(y)}}$$

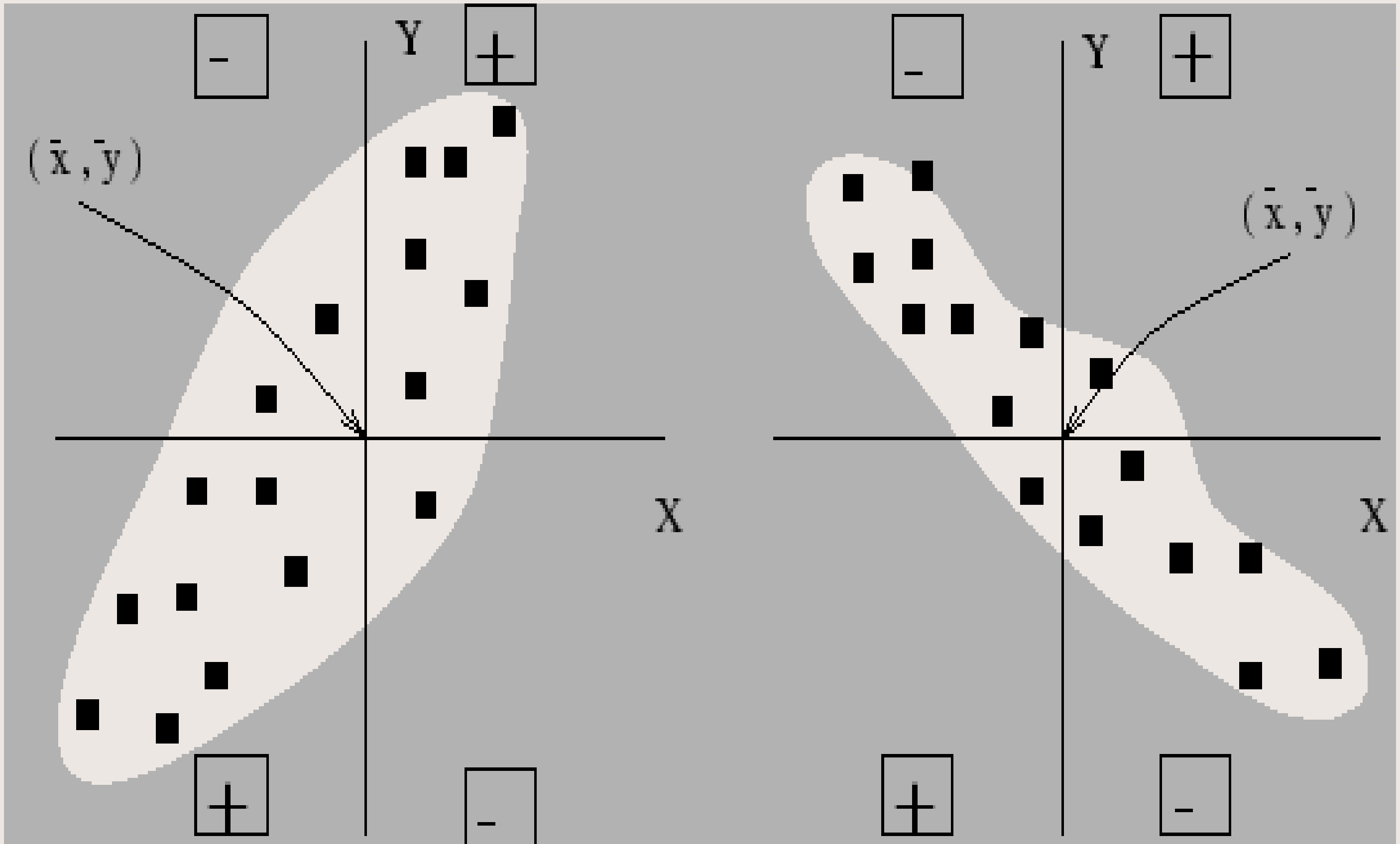
$$r = \frac{\mathrm{S_{xy}}}{\sqrt{\mathrm{S_{xx}}}\sqrt{\mathrm{S_{yy}}}} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

$$\boxed{r = \frac{\left(\sum \mathrm{x_i y_i} - \frac{\sum \mathrm{x_i} \sum \mathrm{y_i}}{\mathrm{n}}\right)}{\sqrt{\left(\sum \mathrm{x_i}^2 - \frac{(\sum \mathrm{x_i})^2}{\mathrm{n}}\right)}\sqrt{\left(\sum \mathrm{y_i}^2 - \frac{(\sum \mathrm{y_i})^2}{\mathrm{n}}\right)}}}$$

Ejemplos de correlación



$$r = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



Cuando X crece, Y crece

Casi todos los puntos pertenecen

a los cuadrantes primero y tercero

Cuando X crece, Y decrece

Casi todos los puntos pertenecen

a los cuadrantes segundo y cuarto

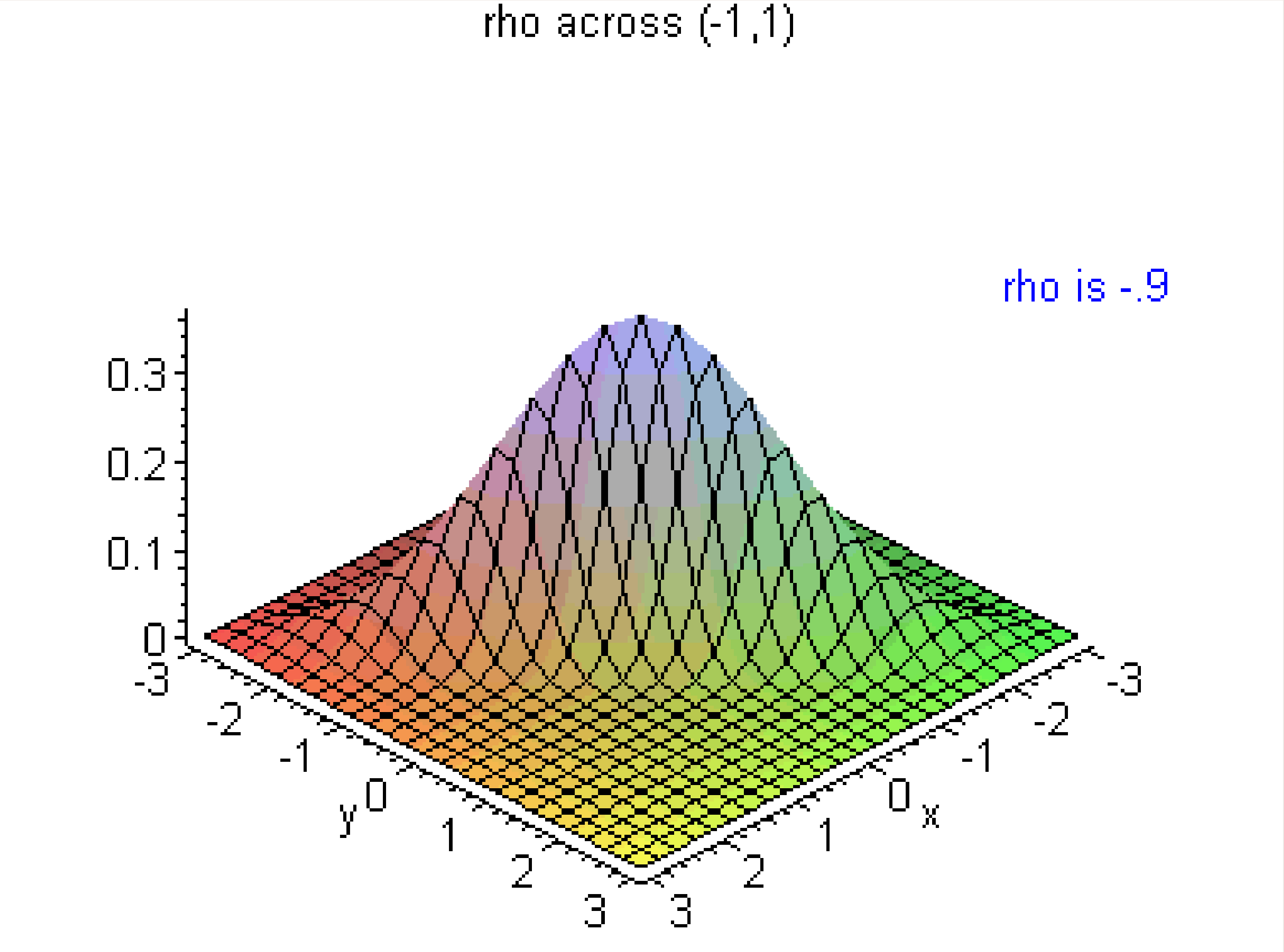
Coeficiente de correlación lineal de Pearson

Coeficiente de correlación lineal estimado

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\,\sqrt{\sum(y_i - \bar{y})^2}}$$

Propiedades de ρ

- Es adimensional.
- Sólo toma valores entre [-1,1].
- No hay correlación lineal entre las variables (incorrelacionadas)➔ ρ = 0.
- Relación lineal perfecta entre 2 variables ⇔ ρ= +1 o ρ = -1
  - Excluimos los casos de puntos alineados horizontal o verticalmente.
- Cuanto más cerca esté ρ de +1 o -1 mejor será el grado de relación lineal.
  - Siempre que se cumplan los supuestos teóricos.


rho across (-1,1)
rho is -.9
0.3
0.2
0.1
0
y
x
-3
-2
-1
0
1
2
3

EJEMPLO

Se seleccionan al azar 8 estudiantes a los que se les consulta sobre: El número de horas dedicadas al estudio de una asignatura y la calificación obtenida en el examen correspondiente. Designamos con (X) a las Horas y con (Y)a la Calificación obtenida.

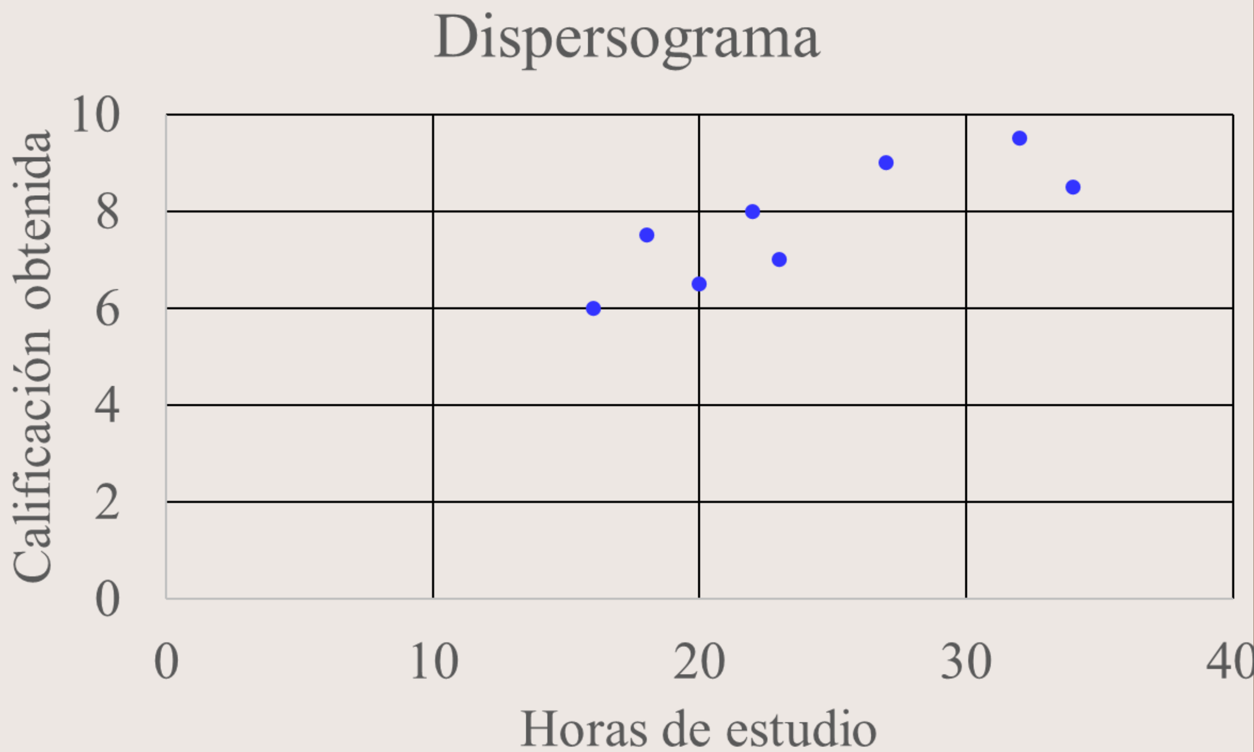
Los datos obtenidos se muestran en la siguiente tabla:

X_i	20	16	34	23	27	32	18	22
Y_i	6,5	6	8,5	7	9	9,5	7,5	8

a) Dibuje el dispersograma

b) ¿Se puede calcular el coeficiente de correlación?, si es así ¿Cuánto vale el coeficiente de correlación lineal?

Estudiante	X_i	Y_i
1	20	6,5
2	16	6
3	34	8,5
4	23	7
5	27	9
6	32	9,5
7	18	7,5
8	22	8
Total	192	62



Cálculos

$$r = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Estudiante	X_i	Y_i	x_i^2	Y_i^2	$x_i * y_i$
1	20	6,5			
2	16	6			
3	34	8,5			
4	23	7			
5	27	9			
6	32	9,5			
7	18	7,5			
8	22	8			
Total	192	62	4902	491	1534

Cálculos

Estudiante	X_i	Y_i	x_i^2	Y_i^2	$x_i *y_i$
1	20	6,5	400	42,25	130
2	16	6	256	36	96
3	34	8,5	1156	72,25	289
4	23	7	529	49	161
5	27	9	729	81	243
6	32	9,5	1024	90,25	304
7	18	7,5	324	56,25	135
8	22	8	484	64	176
Total	192	62	4902	491	1534

Al ser las dos variables aleatorias se puede calcular el coeficiente de correlación r

n=8	$\sum x=192$	$\sum y=62$	$\sum xy=1534$	$\sum x^2= 4902$	$\sum y^2=491$

$$r = \frac{\left(\sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

$$= \frac{\left(1534 - \frac{192 * 62}{8}\right)}{\sqrt{\left(4902 - \frac{(192)^2}{8}\right)}\sqrt{\left(491 - \frac{(62)^2}{8}\right)}} = \frac{46}{\sqrt{(294)}\sqrt{(10{,}5)}}$$

$$= \frac{46}{55{,}56} = 0{,}828$$

Como r=0,828 se puede interpretar que hay una fuerte asociación entre la variable (X) Horas de estudio y la variable (Y) Calificación obtenida.

REGRESIÓN

Regresión Lineal Simple

<u>Conceptos:</u>

Regresión simple: interviene una sola variable independiente.

Regresión múltiple: intervienen dos o más variables independientes.

Regresión no lineal: la función que relaciona los parámetros no es una combinación lineal en los parámetros.

Regresión Lineal Simple

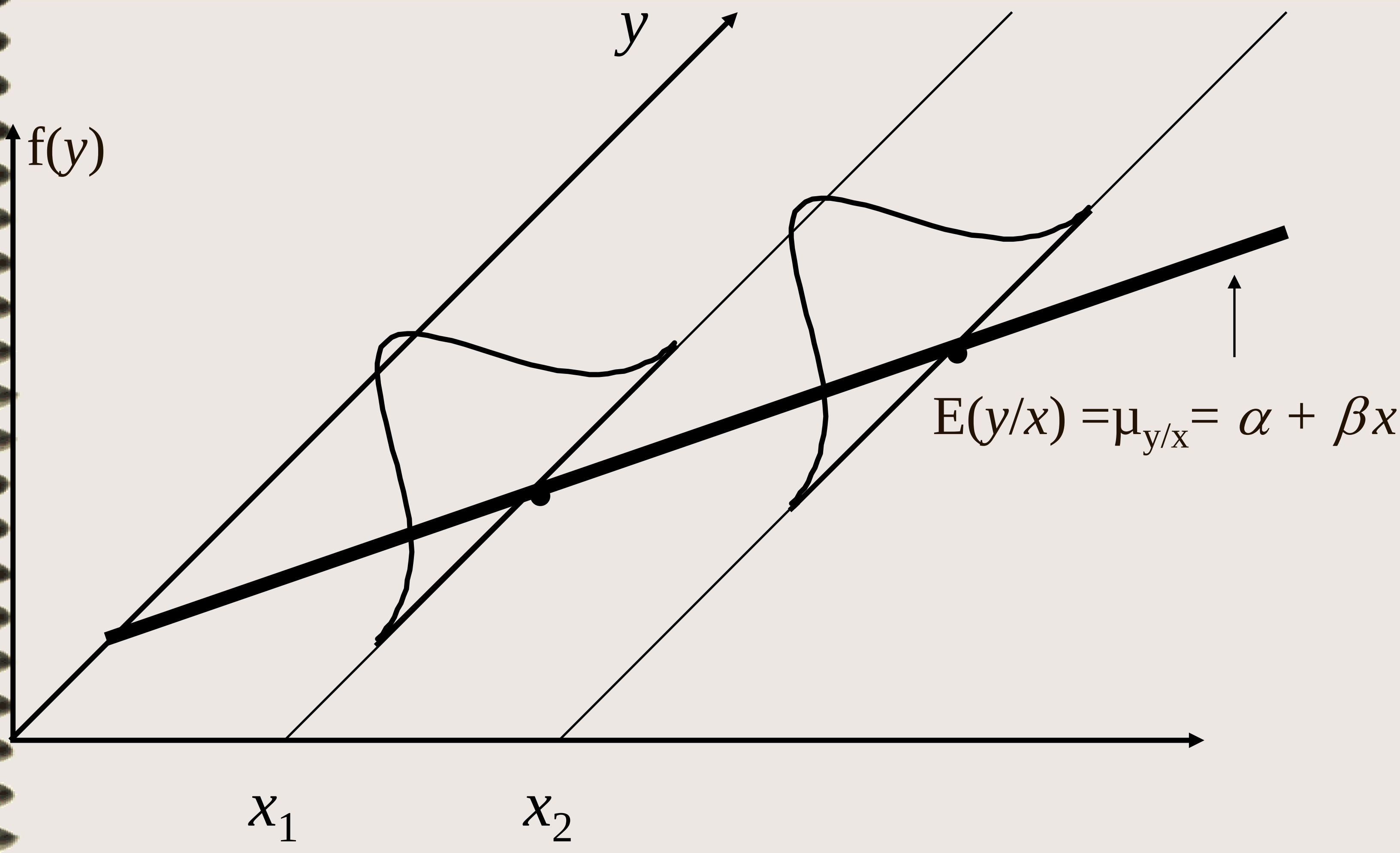
Objetivo:

Hallar una función o un modelo matemático para predecir y estimar el valor de una variable a partir de valores de otra, ambas cuantitativas.

La variable Y: que es la dependiente (respuesta, predicha, endógena). Es la variable que se desea predecir o estimar y

la variable X: que es la independiente (predictora, explicativa, exógena). Es la variable que provee las bases para estimar.

Regresión Lineal Simple



Modelo teórico

SUPUESTOS

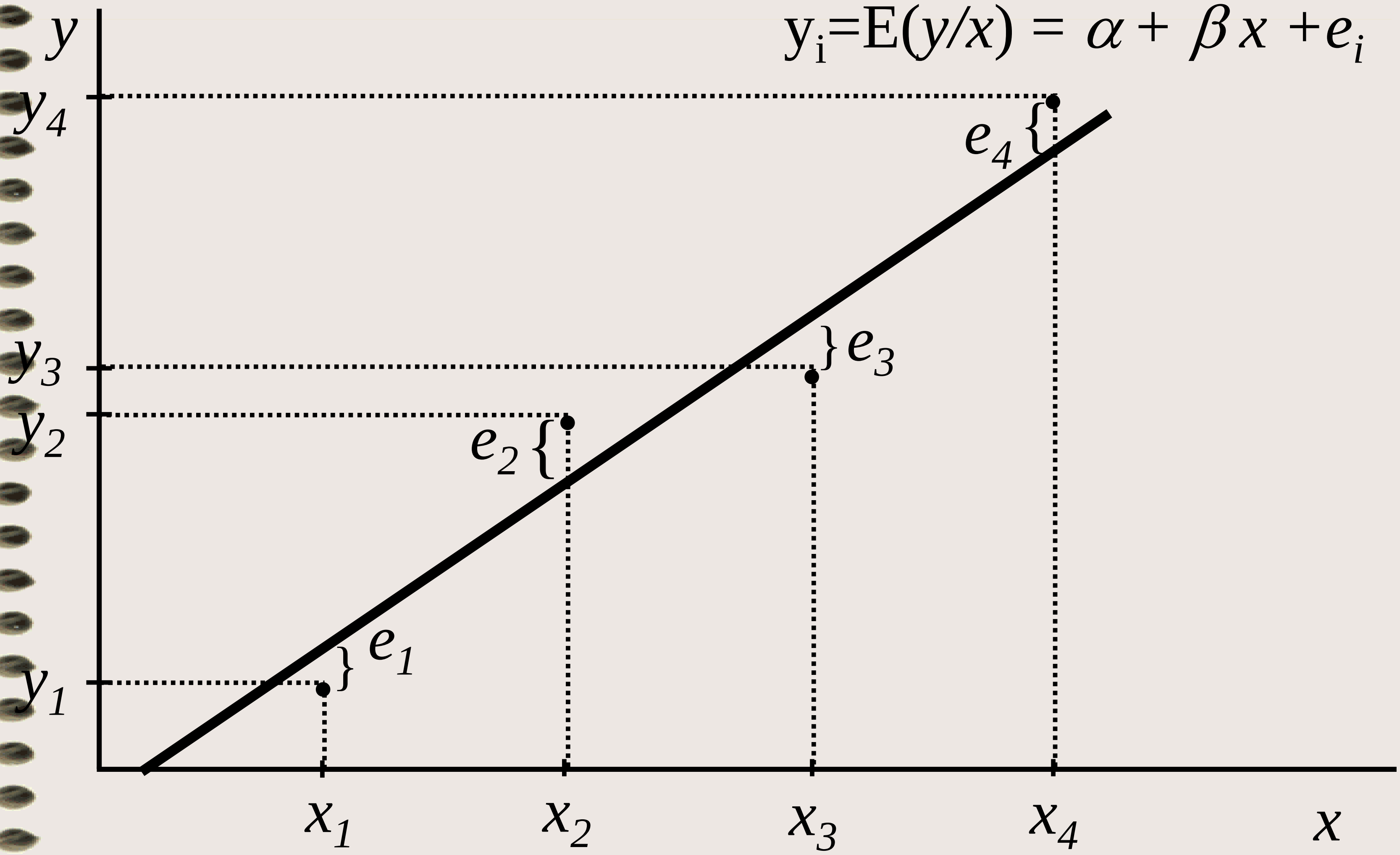
- Los valores de la variable independiente “x” son fijos, esto significa que son preseleccionados por el investigador, de modo que en la recolección de los datos no se permite que varíen. La variable “x” se mide sin error (se desprecia)
- Para cada valor de “x” existe una subpoblación de valores de “y”. Estas subpoblaciones deben estar normalmente distribuidas.
- Las varianzas de las subpoblaciones de “y” son todas iguales e iguales a la varianza del error.
- las medias de las subpoblaciones deben situarse sobre la línea recta del modelo teórico.
- Los valores de y son independientes de los valores de y de la otra población.

Modelo teórico:

$$\mu_{y/x} = \alpha + \beta x$$

- Este modelo implica que todas las medidas de las subpoblaciones de "y" están sobre la misma recta.
- α y β son los coeficientes de regresión de la población y geométricamente representan la ordenada al origen y la pendiente de la recta, respectivamente.

Regresión Lineal Simple



Modelo estadístico

Modelo estadístico:

$$\mu_{y/x} = \alpha + \beta x + e$$

- los valores de "y" son estadísticamente independientes.
- En esta ecuación se tiene en cuenta el término del error "e".
- Los errores para cada subpoblación están normalmente distribuidos con una varianza igual a la varianza común de las subpoblaciones de valores "y".

Regresión Lineal Simple

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

$$\mu_{y/x} = E_{(Y/X)} = \alpha + \beta x$$

Interpretación de los Coeficientes de Regresión:

α: es la ordenada al origen

Indica el valor medio poblacional de la variable respuesta Y cuando X es cero. Si se tiene certeza de que la variable predictora X no puede asumir el valor 0, entonces la interpretación no tiene sentido.

β: es la pendiente de la línea de regresión

Indica el cambio o modificación del valor medio poblacional de la variable respuesta Y cuando X se incrementa en una unidad.

e: es un error aleatorio

$$e = y - (\alpha + \beta x)$$

Estimación de la línea de regresión usando Mínimos Cuadrados

Se debe Minimizar $\sum_{i=1}^{n} e_i^2 \quad = \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

Derivando $\quad \dfrac{\partial \sum e^2}{\partial \alpha} = 0 \qquad \dfrac{\partial \sum e^2}{\partial \beta} = 0$

se obtiene un par de ecuaciones normales para el modelo, cuya solución produce

$$b = \frac{\left(\sum x_i\, y_i - \dfrac{\sum x_i \sum y_i}{n}\right)}{\left(\sum {x_i}^2 - \dfrac{\left(\sum x_i\right)^2}{n}\right)} \qquad a = \bar{y} - b\bar{x}$$

Modelo estimado:

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

Donde:

(a) es un estimador de α

(b) es un estimador de β

Además $e \cong N(0, \sigma^2)$

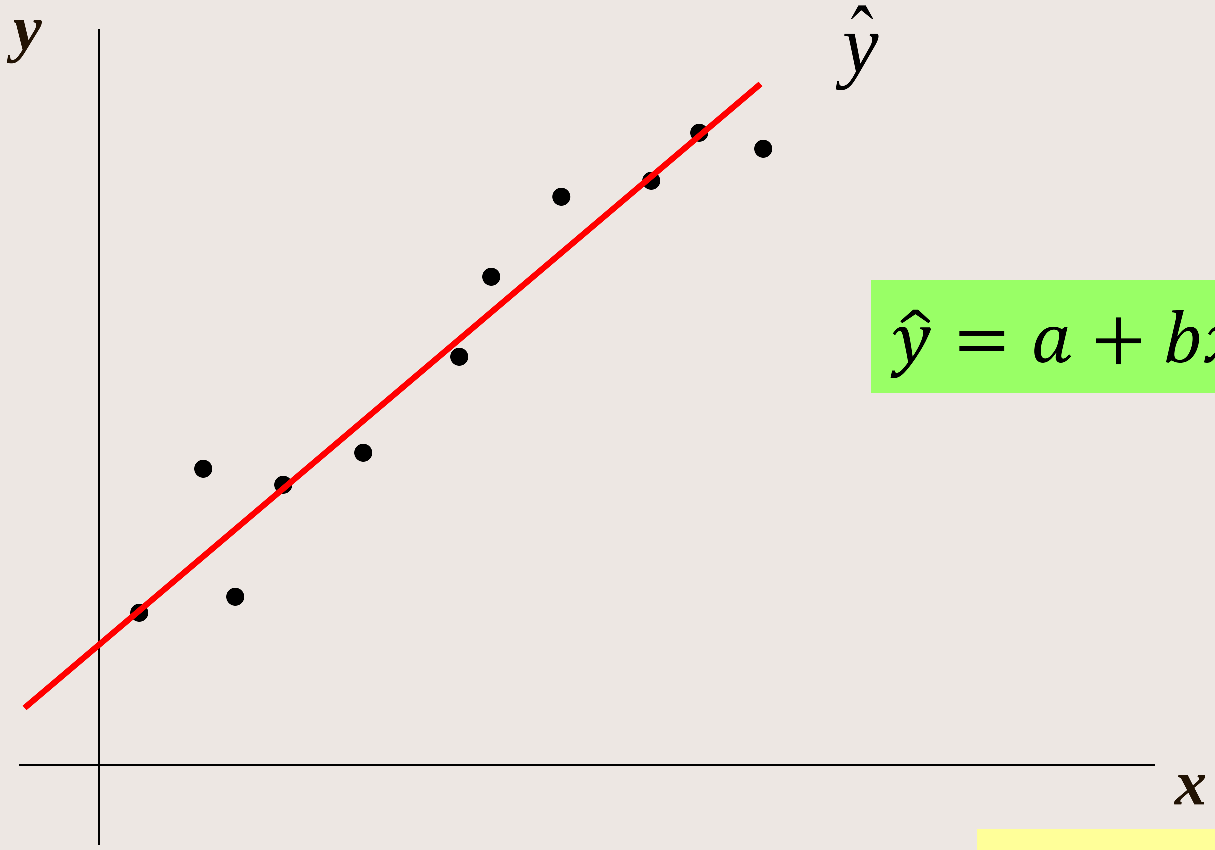
Interpretación de los coeficientes de regresión estimados

La pendiente "b" indica el cambio promedio estimado en la variable respuesta cuando la variable predictora aumenta en una unidad adicional.

La ordenada al origen "a" indica el valor promedio estimado de la variable respuesta cuando la variable predictora vale 0. Sin embargo carece de interpretación práctica si es irrazonable considerar que el rango de valores de x incluye a cero.

REGRESION LINEAL SIMPLE

Estimar los valores de *y* (variable dependiente) a partir de los valores de *x* (variable independiente)

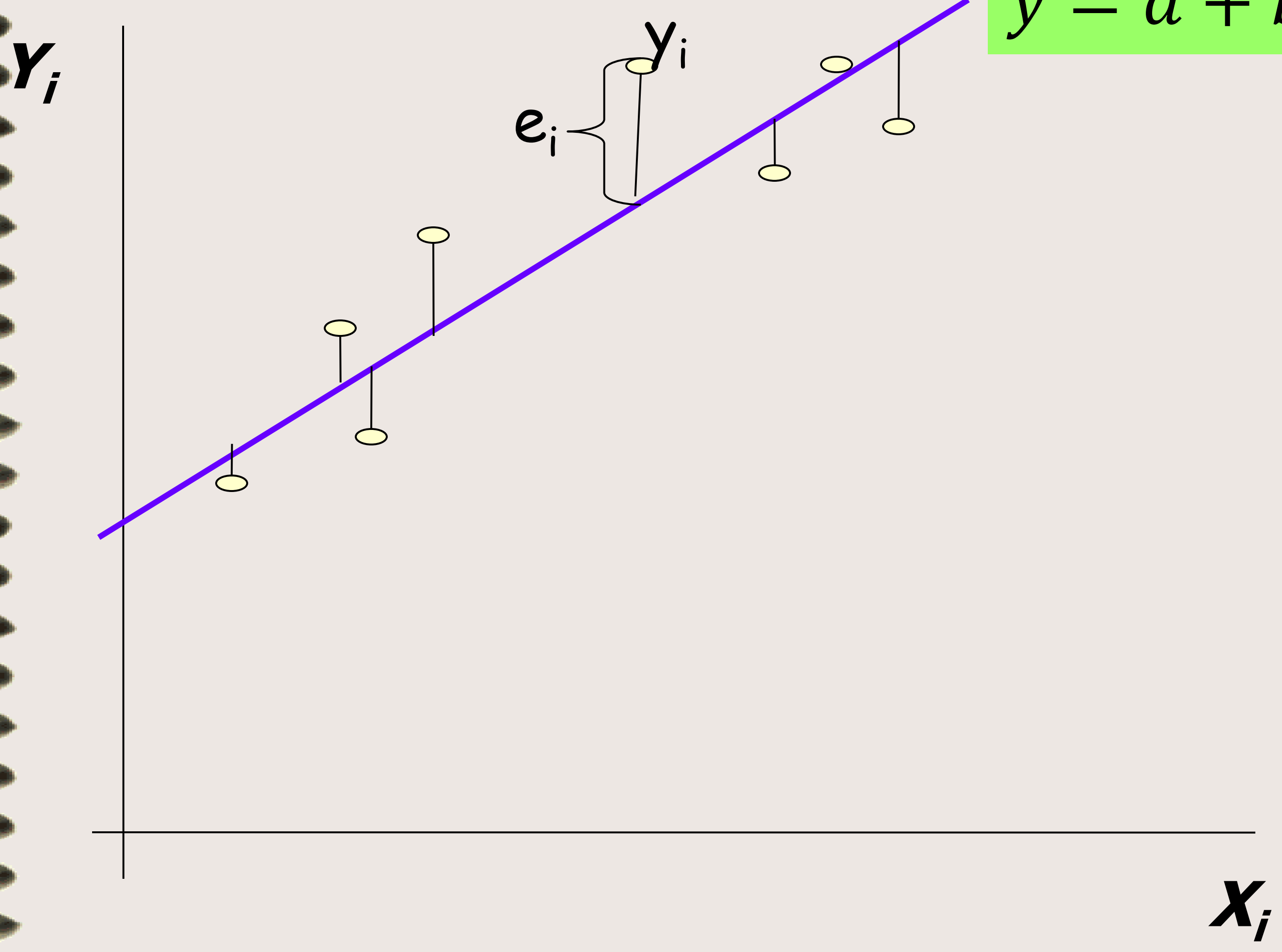


$$\hat{y} = a + bx$$

Modelo estimado

MÉTODO DE MÍNIMOS CUADRADOS

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



MÉTODO DE MÍNIMOS CUADRADOS

- Consiste en minimizar la distancia vertical entre los puntos y la recta.

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$\sum e_i^2 = \sum (y_i - a - bx_i)^2$$

$$\sum e_i^2 = f(a,b)$$

Derivación

$$Min = \sum e_i^2 = \sum (y_i - a - bx_i)^2$$

$$\left.\begin{array}{l} \dfrac{\partial \sum e_i^2}{\partial a} = 0 \\[2em] \dfrac{\partial \sum e_i^2}{\partial b} = 0 \end{array}\right\}$$

Derivada respecto de a

(1) $$\frac{\partial \sum e_i^2}{\partial a} = 2\sum (Y_i - a - bx_i)\,(-1) = 0$$

$$\sum (y_i - a - bx_i) = 0$$

$$\sum y_i - a\sum 1 - b\sum x_i = 0$$

$$\sum y_i - an - b\sum x_i = 0$$

$$\sum y_i - b\sum x_i = na$$

$$\frac{\sum y_i}{n} - \frac{b\sum x_i}{n} = a$$

$$\boxed{\bar{y}_i - b\bar{x}_i = a}$$

Derivada respecto de b

$$\frac{\partial \sum e_i^2}{\partial b} = \sum (y_i - a - bx_i)^2 = 0$$

$$\frac{\partial \sum e_i^2}{\partial b} = 2\sum (Y_i - a - bx_i)\,(x_i)(-1) = 0$$

(2) $$\sum y_i x_i - a\sum x_i - b\sum x_i^2 = 0$$

Sustituyendo $a = \bar{y} - b\bar{x}$

$$\sum y_i x_i - (\bar{y} - b\bar{x})\sum x_i - b\sum x_i^2 = 0$$

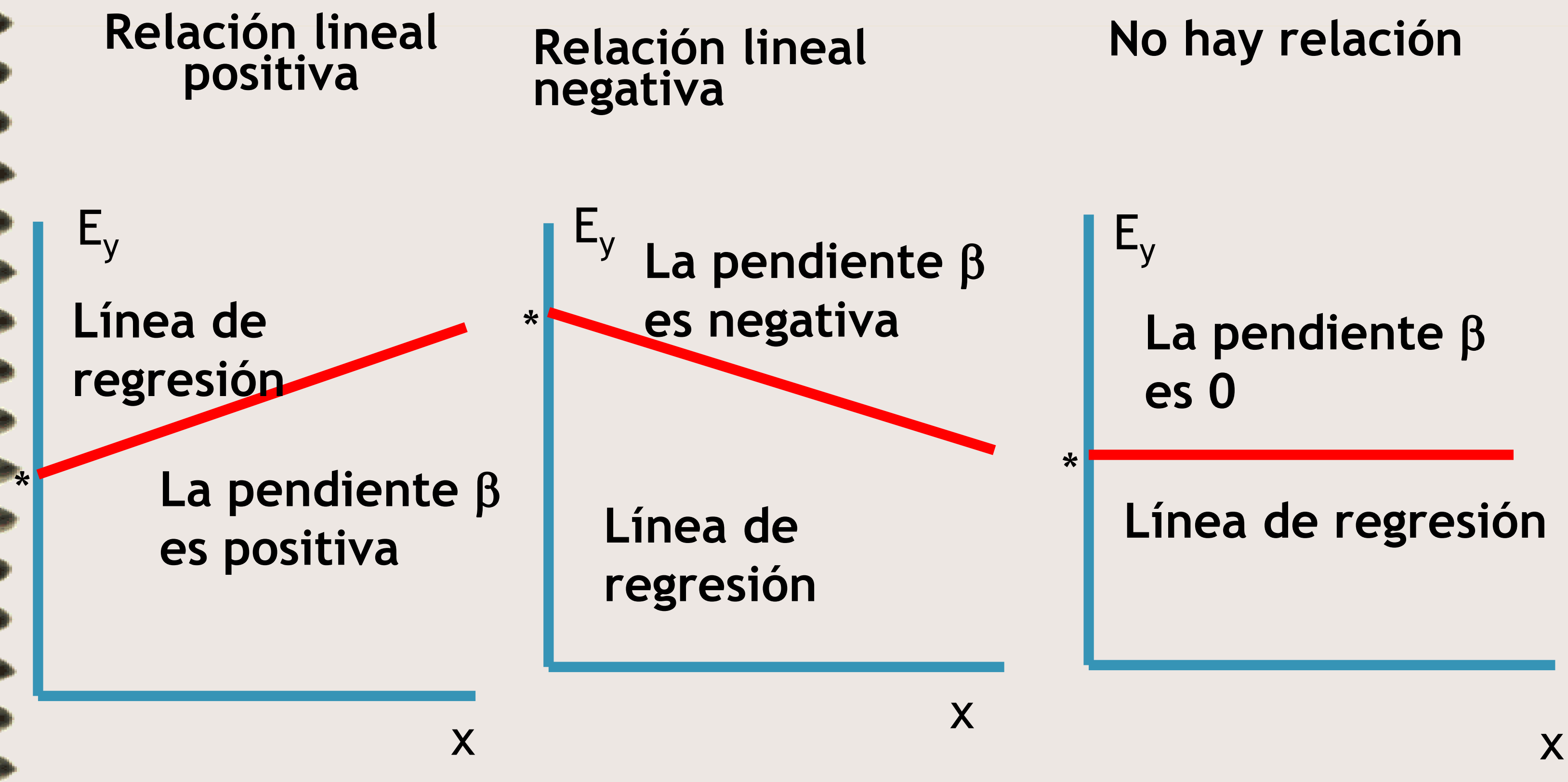
$$\sum y_i x_i - \bar{y}\sum x_i + b\bar{x}\sum x_i - b\sum x_i^2 = 0$$

$$\sum y_i x_i - \frac{\sum y_i}{n}\sum x_i = b\sum x_i^2 - b\frac{\sum x_i}{n}\sum x_i$$

$$\sum y_i x_i - \frac{\sum y_i \sum x_i}{n} = b\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$$

$$\boxed{b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\left(\sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right)}}$$

Líneas posibles de regresión en la regresión lineal simple



* Ordenada al origen α

Retomamos en EJEMPLO

Se seleccionan al azar 8 estudiantes a los que se les consulta sobre: El número de horas dedicadas al estudio de una asignatura y la calificación obtenida en el examen correspondiente. Designamos con (X) a las Horas y con (Y)a la Calificación obtenida.

Los datos obtenidos se muestran en la siguiente tabla:

X_i	20	16	34	23	27	32	18	22
Y_i	6,5	6	8,5	7	9	9,5	7,5	8

c) Encontrar la Recta de regresión de Y sobre X, es decir que permita estimar la calificación obtenida para una determinada cantidad de horas de estudio.

d) Estime la calificación obtenida para una persona que hubiese estudiado 28 horas.

Cálculos

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

$$b = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)}$$

$$a = \bar{y} - b\bar{x}$$

Estudiante	X_i	Y_i	x_i^2	Y_i^2	$x_i * y_i$
1	20	6,5			
2	16	6			
3	34	8,5			
4	23	7			
5	27	9			
6	32	9,5			
7	18	7,5			
8	22	8			
Total	192	62	4902	491	1534

Cálculos

Estudiante	X_i	Y_i	x_i^2	Y_i^2	$x_i * y_i$
1	20	6,5	400	42,25	130
2	16	6	256	36	96
3	34	8,5	1156	72,25	289
4	23	7	529	49	161
5	27	9	729	81	243
6	32	9,5	1024	90,25	304
7	18	7,5	324	56,25	135
8	22	8	484	64	176
Total	192	62	4902	491	1534

Calculamos b:

n=8 $\sum x=192$ $\sum y=62$ $\sum xy=1534$ $\sum x^2= 4902$ $\sum y^2=491$

$$b=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{\left(\sum x_i\,y_i-\frac{\sum x_i\sum y_i}{n}\right)}{\left(\sum x_i{}^2-\frac{(\sum x_i)^2}{n}\right)}$$

$$b=\frac{\left(1543-\frac{192*62}{8}\right)}{\left(4902-\frac{(192)^2}{8}\right)}=\frac{46}{294}=0{,}1565$$

Calculamos a:

$$a=\bar{y}_i-b\bar{x}_i=\frac{62}{8}-0{,}1565\,\frac{192}{8}=3{,}99$$

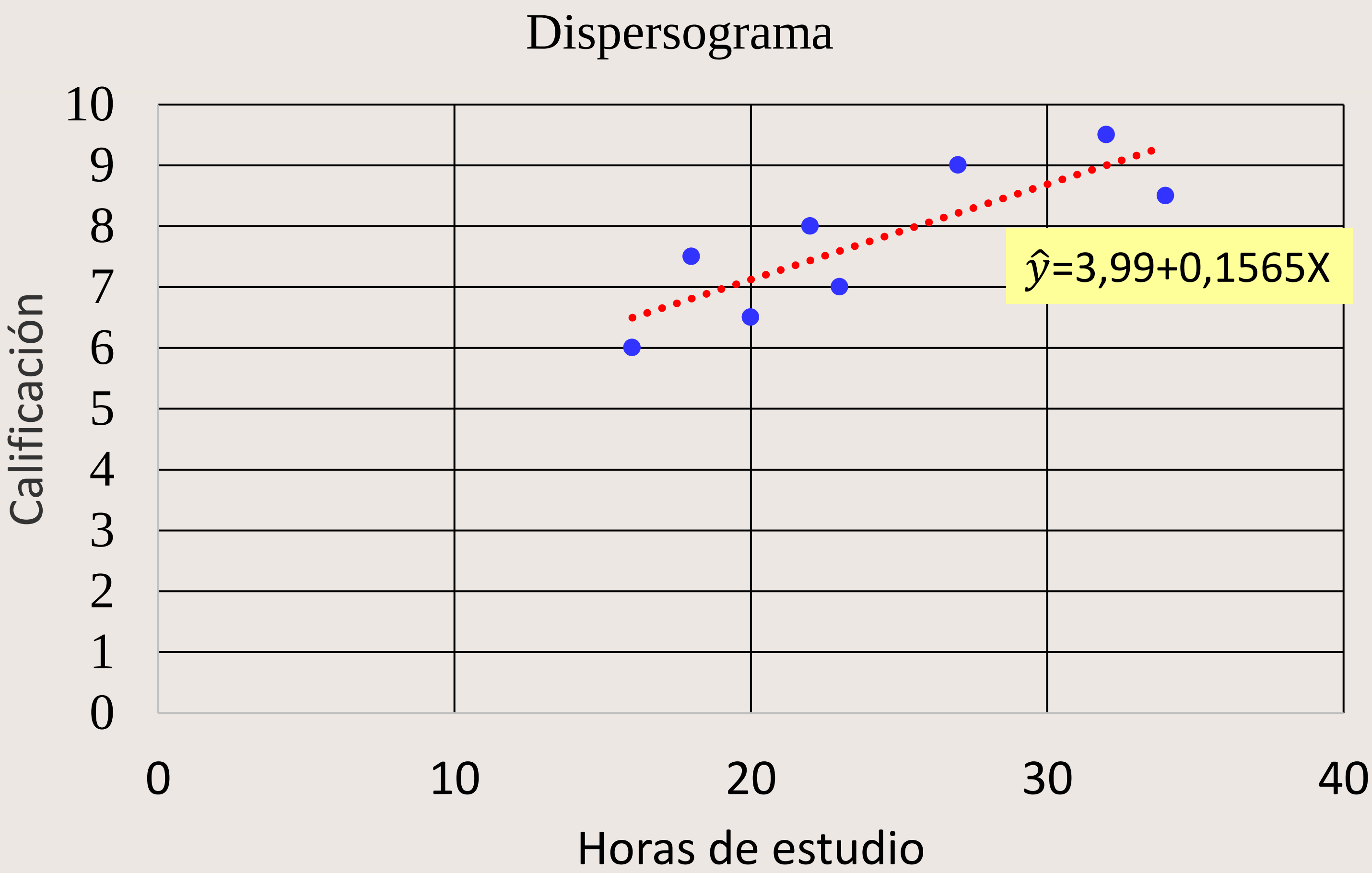
$$\hat{y}=a+bx=3{,}99+0{,}1565X$$

Reemplazamos x por 28 y obtenemos:

$$\hat{y}=3{,}99+0{,}1565*28=8{,}37$$

La calificación obtenida por un estudiante que dedicó 28 horas de estudio será de 8,37.

Dispersograma


Dispersograma
10
9
8
7
6
5
4
3
2
1
0
0
10
20
30
40
Calificación
Horas de estudio
$\hat{y}$=3,99+0,1565X

Ejemplo

Supongamos que conocemos los datos de producción y horas trabajadas de 10 trabajadores de una fábrica en un momento de tiempo.

Definimos:

Y=producto

X=horas de trabajo

Obs.	X	Y	X^2	Y^2	XY
1	10	11	100	121	110
2	7	10	49	100	70
3	10	12	100	144	120
4	5	6	25	36	30
5	8	10	64	100	80
6	8	7	64	49	56
7	6	9	36	81	54
8	7	10	49	100	70
9	9	11	81	121	99
10	10	10	100	100	100
Σ	80	96	668	952	789

Ejemplo

$\bar{X} = 8$ $\bar{Y} = 9{,}6$

$$b = \frac{789 - \dfrac{80 \cdot 96}{10}}{668 - \dfrac{80^2}{10}} = 0{,}75$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 9{,}6 - b \cdot 8$$

$$\boxed{\hat{Y}_i = a + bx_i = 3{,}6 + 0{,}75x}$$

INFERENCIA

CORRELACIÓN

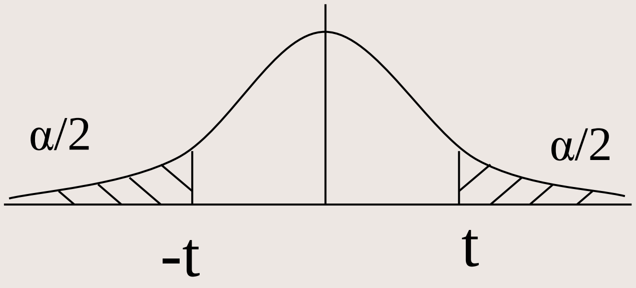
Correlación:

hipótesis para el coeficiente de correlación

1- $\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$

2- α =…….

3- $t_{n-2} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$



Ejemplo

Una compañía de seguros considera que el número de vehículos (Y) que circulan por una autopista, puede ponerse en función del número de accidentes (X) que ocurren en ella. Durante cinco días se obtuvo los siguientes resultados

x	5	7	2	1	9
y	15	7	10	8	20

a) Calcule el coeficiente de correlación.

b) Pruebe la hipótesis de que no existe correlación entre esas dos variables utilizando un nivel de significación del 5%

Cálculos

$$r = \frac{\left(\sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum x_i{}^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum y_i{}^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Días	**X_i**	**Y_i**	**x_i^2**	**Y_i^2**	**$x_i * y_i$**
1	5	15			
2	7	7			
3	2	10			
4	1	8			
5	9	20			
	24	60	160	838	332

Cálculos

Días	X_i	Y_i	x_i^2	Y_i^2	$x_i * y_i$
1	5	15	25	225	75
2	7	7	49	49	49
3	2	10	4	100	20
4	1	8	1	64	8
5	9	20	81	400	180
	24	60	160	838	332

a) Al ser las dos variables aleatorias se puede calcular el coeficiente de correlación r

n=5 $\sum x=24$ $\sum y=60$ $\sum xy=332$ $\sum x^2= 160$ $\sum y^2= 838$

$$r=\frac{\left(\sum x_i\,y_i-\frac{\sum x_i\sum y_i}{n}\right)}{\sqrt{\left(\sum x_i{}^2-\frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum y_i{}^2-\frac{(\sum y_i)^2}{n}\right)}}$$

$$=\frac{\left(332-\frac{24*60}{5}\right)}{\sqrt{\left(160-\frac{(24)^2}{5}\right)}\sqrt{\left(838-\frac{60^2}{5}\right)}}=\frac{44}{\sqrt{(44{,}8)}\sqrt{(118)}}=\frac{44}{72{,}71}$$

$$=0{,}605$$

Como r=0,6051 se puede interpretar que hay una débil asociación el número de vehículos (Y) que circulan por una autopista, y el número de accidentes ocurridos (X) .

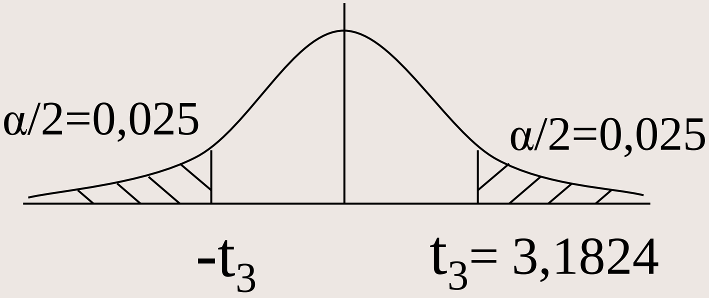
b) Pruebe la hipótesis de que no existe correlación entre esas dos variables utilizando un nivel de significación del 5%

1- $\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$

2- $\alpha = 0{,}05$

3- $t_{n-2} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

4-



Regla de decisión

Rechazo H_0 si tcal $\geq 3{,}184$ ó tcal $\leq -3{,}1824$

No rechazo H_0 si $-3{,}1824 <$ tcal $< 3{,}1824$

Cálculos

r=0,6051

$$t_{Calculado} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,6051\sqrt{5-2}}{\sqrt{1-0,6051^2}} = 1,316$$

6- Como $t_{\text{calculado}}$ se encuentra -3,1824<t cal < 3,1824

No rechazo H_0

Conclusión: Con un nivel de significación del 5% no tengo evidencias para rechazar la hipótesis nula por lo que se supone que no existe asociación entre las variables número de vehículos (Y) que circulan por una autopista, y el número de accidentes ocurridos (X)

INFERENCIA

REGRESIÓN

Estimadores

$$\boxed{a \cong N(E(a) = \alpha; V(a) = S^2{}_e(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2})}$$

$$\boxed{b \cong N(E(b) = \beta; V(b) = S^2{}_e(\frac{1}{\sum(x-\bar{x})^2})}$$

$$\hat{y} \cong N(E(\hat{y}) = \alpha + \beta x; V(\hat{y}) = S^2{}_e\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x-\bar{x})^2}\right)$$

$$S^2{}_e = \frac{1}{n-2}\left(\sum(y_i-\bar{y})^2 - b\sum(x_i-\bar{x})(y_i-\bar{y})\right)$$

$$S^2{}_e = \frac{1}{n-2}\left(\sum(y_i-\bar{y})^2 - b^2\sum(x_i-\bar{x})^2\right)$$

Error estándar de la estimación

Se o Sy/x

Mide la dispersión o alejamiento promedio de los puntos con respecto a la recta estimada.

$$S^2{}_e = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$$

$$S^2{}_e = \frac{1}{n-2}\left(\sum(y_i - \bar{y})^2 - b\sum(x_i - \bar{x})(y_i - \bar{y})\right)$$

$$S^2{}_e = \frac{1}{n-2}\left(\sum(y_i - \bar{y})^2 - b^2 \sum(x_i - \bar{x})^2\right)$$

Desviación típica del coeficiente de regresión

$$s_b = \frac{s_e}{\sqrt{\sum(x-\bar{x})^2}}$$

Intervalo de confianza para el coeficiente de regresión (β)

$$\beta \in \{\, b \pm t s_b \} \; ; \;\; t \;\; con \;\; (n-2) gl$$

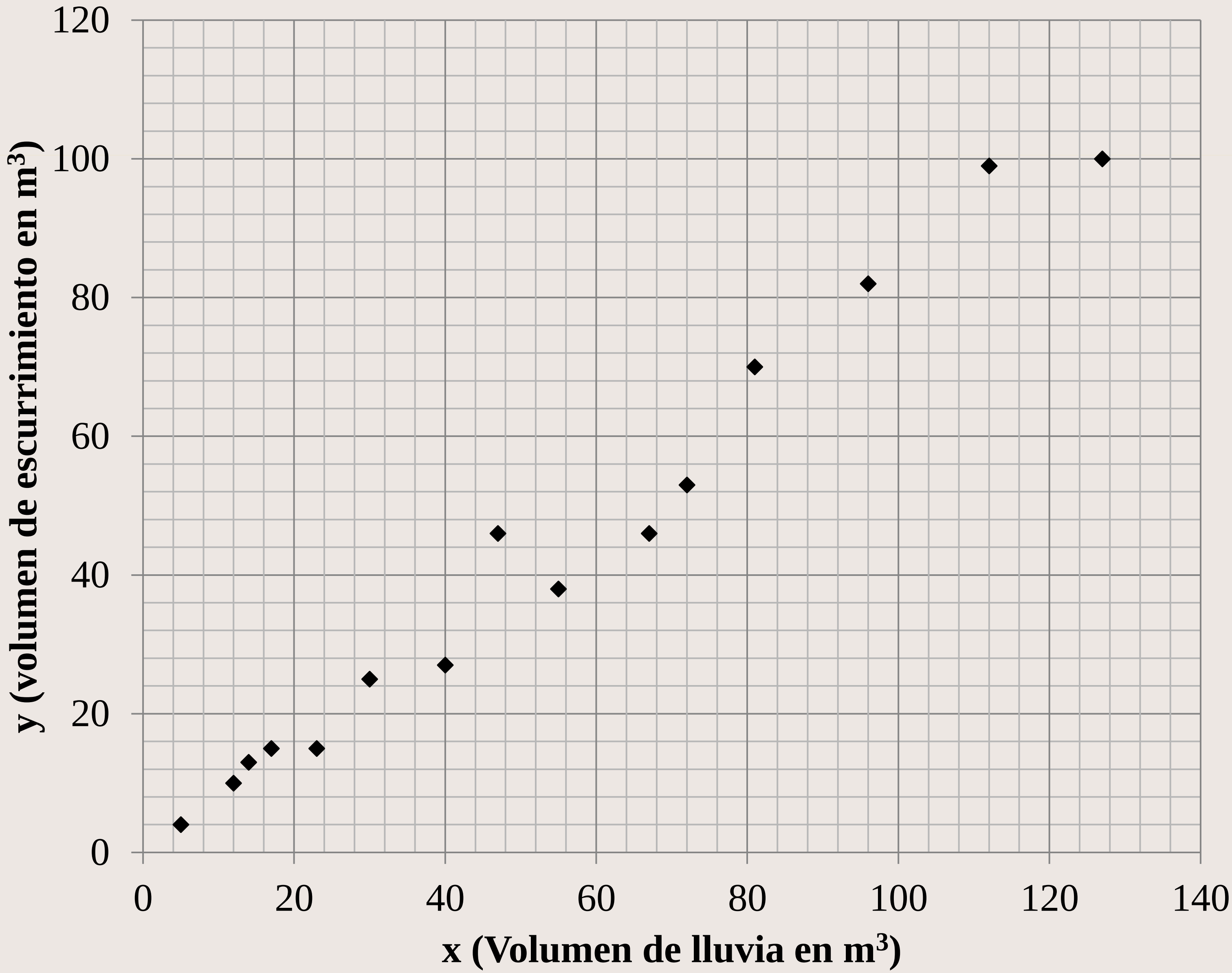
Sea X el volumen de precipitación pluvial -lluvia (m^3) e Y el volumen de escurrimiento (m^3) en determinado lugar. Se realizaron 15 observaciones que se muestran en la tabla siguiente:

No. Observ.	X	Y
1	5	4
2	12	10
3	14	13
4	17	15
5	23	15
6	30	25
7	40	27
8	47	46
9	55	38
10	67	46
11	72	53
12	81	70
13	96	82
14	112	99
15	127	100

a.-Dibuje el dispersograma.

b.-Calcule el coeficiente de correlación e interprete en términos del problema.

c.-Encuentre la recta de mínimos cuadrados. Realice una estimación puntual en x =50 e interprete en términos del problema.


y (volumen de escurrimiento en m³)
x (Volumen de lluvia en m³)
0
20
40
60
80
100
120
140

No. Observ.	X	Y	X^2	Y^2	XY
1	5	4	25	16	20
2	12	10	144	100	120
3	14	13	196	169	182
4	17	15	289	225	255
5	23	15	529	225	345
6	30	25	900	625	750
7	40	27	1600	729	1080
8	47	46	2209	2116	2162
9	55	38	3025	1444	2090
10	67	46	4489	2116	3082
11	72	53	5184	2809	3816
12	81	70	6561	4900	5670
13	96	82	9216	6724	7872
14	112	99	12544	9801	11088
15	127	100	16129	10000	12700
TOTAL	**798**	**643**	**63040**	**41999**	**51232**

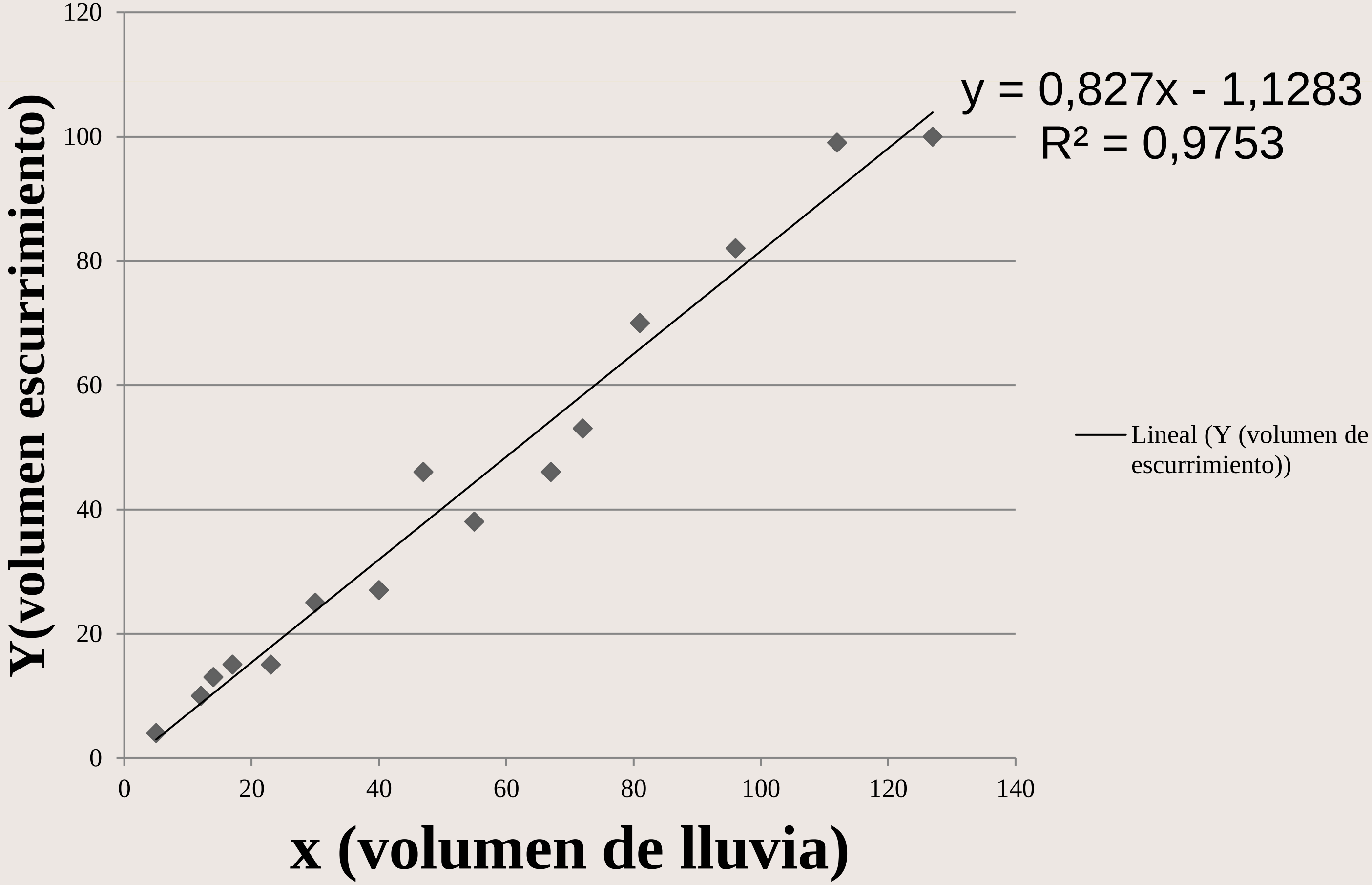
Coeficiente de correlación

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

$n=15$ $\sum x=798$ $\sum y=643$ $\sum xy=51232$ $\sum x^2= 63040$ $\sum y^2= 41999$

$$r = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}\right)}} = \frac{17024{,}4}{\sqrt{(20586{,}4)}\sqrt{(14435{,}7335)}}$$

$$r = \frac{17024{,}4}{143{,}479 * 120{,}1488} = 0{,}98755$$



Para x=50, La estimación puntual para la respuesta media de y es: 40,2548063 d- Realice la prueba de hipótesis más importante en Regresión utilizando un nivel de significación de 0,05

Prueba de hipótesis para el coeficiente de regresión β

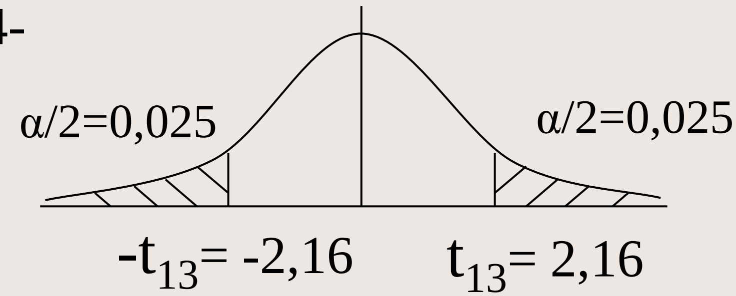
1- $\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$

2- $\alpha = 0{,}05$

3- Variable pivotal

$$t = \frac{b - \beta}{S_b} \approx t_{(n-2)}$$

4-



Regla de decisión

Rechazo H_0 si tcal ≥ 2,16 ó tcal ≤ -2,16

No rechazo H_0 si -2,16< t cal < 2,16

5-Cálculos

$n=15$ $\sum x=798$ $\sum y=643$ $\sum xy=51232$ $\sum x^2= 63040$ $\sum y^2= 41999$

$$S^2{}_e = \frac{1}{n-2}\left[\underbrace{\sum y_i{}^2 - \frac{(\sum y_i)^2}{n}} - b\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)\right]$$

$$S^2{}_e = \frac{1}{13}[\overbrace{14435{,}73} - 0{,}82697(17024{,}4)] = \frac{357{,}406}{13} = 27{,}46$$

$$S^2{}_b = \frac{S^2{}_e}{\sum(x_i - \bar{x})^2} = \frac{27{,}46}{20586{,}4} = 0{,}00133389$$

Prueba de hipótesis para el coeficiente de regresión β

$$\begin{cases} H_o: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$$

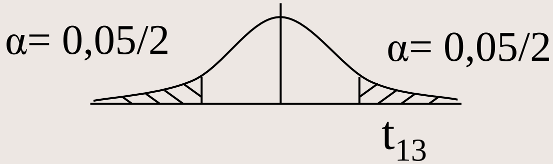
$$t = \frac{b - \beta}{S_b} \approx t_{(n-2)}$$

Regla de decisión

Rechazo H_0 si
tcal ≥ 2,16 ó
tcal ≤ -2,16

No rechazo H_0 si
-2,16< t cal < 2,16

α= 0,05



$t_{13;\,0{,}975}$=2,16

$$t_{calculado} = \frac{0{,}8269 - 0}{0{,}0365} = 22{,}64$$

6- Decisión: Como tcal es >2,16 Rechazo Ho.

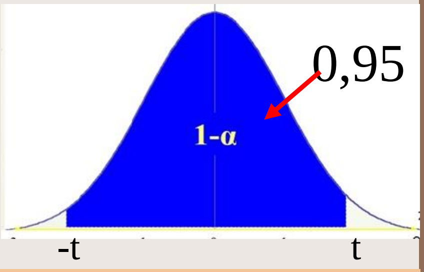
Conclusión: Con un nivel de significación del 5% tengo evidencias suficientes para suponer que existe una relación funcional poblacional donde, el volumen de escurrimiento está en función del volumen de precipitación pluvial, o que sea, por cada metro^3 que se incrementa la precipitación el volumen de escurrimiento se modifica o cambia el valor medio poblacional.

e-Encuentre un intervalo de confianza para la pendiente poblacional.
Utilice una confianza de 0,95 e interprete en términos del problema.

Intervalo de confianza para el coeficiente de Regresión

$$P\left(b - t_{(n-2;\alpha/2)} S_b < \beta < b + t_{(n-2;1-\alpha/2)} S_b\right) = 1 - \alpha$$

$$0{,}82697 - 2{,}16 * 0{,}0365 < \beta < 0{,}82697 + 2{,}16 * 0{,}0365$$



$$(0{,}74 < \beta < 0{,}9)$$

Con una confianza de 95 %, podría decir que el intervalo (0,74; 0,90) m^3/m^3 encerraría al verdadero valor de la pendiente de la recta de regresión. Esto es, con una confianza de 95 %, podría decir que el intervalo (0,74; 0,9) m^3/m^3 encerraría al verdadero cambio del promedio poblacional del volumen de escurrimiento, para un aumento unitario en el volumen de lluvia.

Error estándar de la estimación

Se o Sy/x

Mide la dispersión o alejamiento promedio de los puntos con respecto a la recta estimada.

$$S^2{}_e = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$$

$$S^2{}_e = \frac{1}{n-2}\left(\sum(y_i - \bar{y})^2 - b\sum(x_i - \bar{x})(y_i - \bar{y})\right)$$

$$\boxed{S^2{}_e = \frac{1}{n-2}\left(\sum(y_i - \bar{y})^2 - b^2\sum(x_i - \bar{x})^2\right)}$$