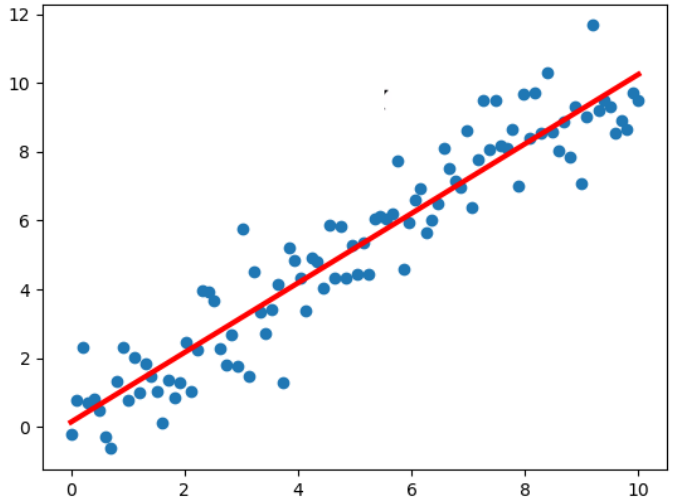
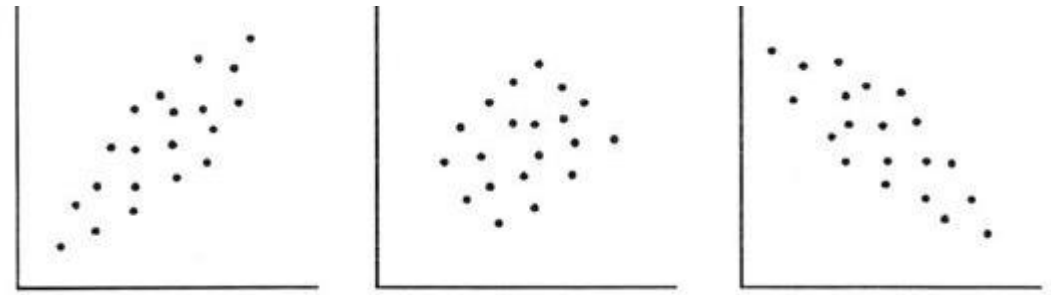
REGRESIÓN
Y
CORRELACIÓN





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

- **T**eoría de probabilidades
- **D**istribuciones de probabilidad
- **D**istribuciones 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(*)**

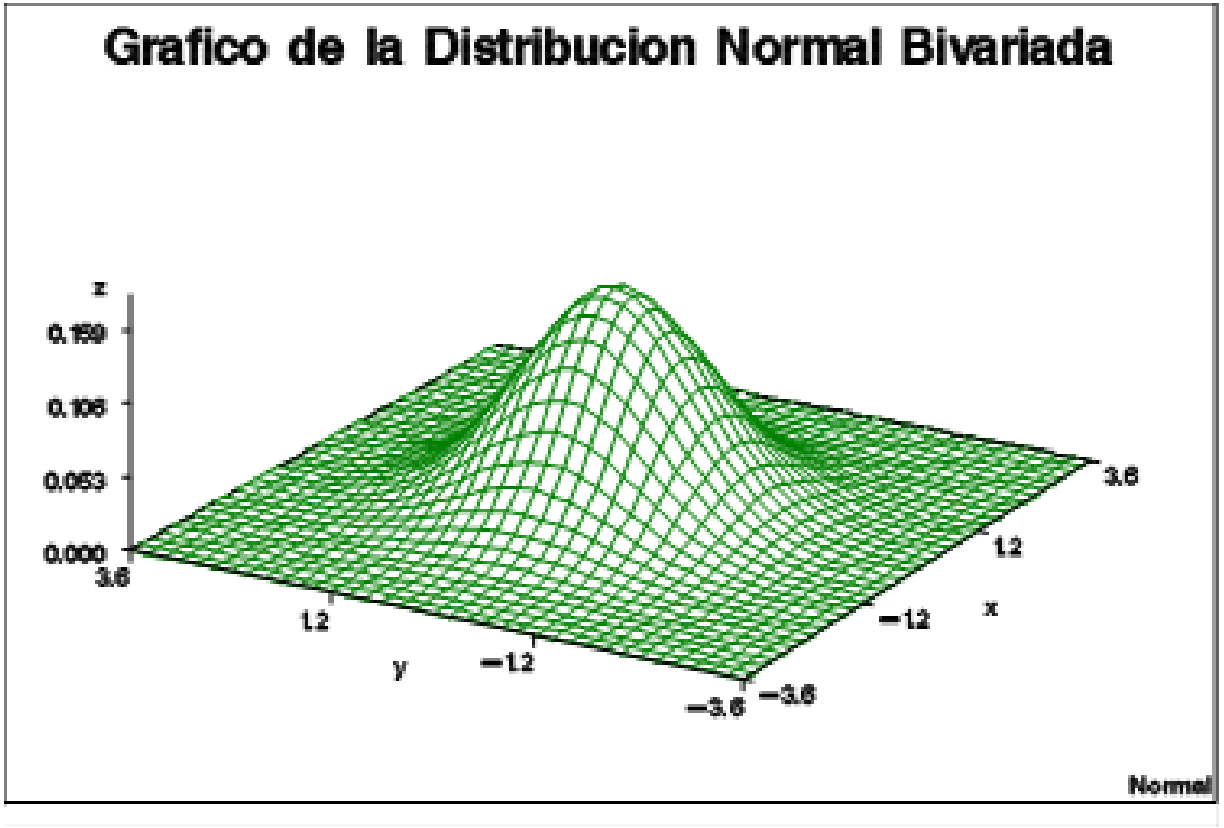
(*) También pueden ser no paramétricos

Ahora comenzamos con el estudio de 2 o más variables en simultáneo

Regresión Lineal Simple: 1 variable fijada por el investigador (medida sin error) y la otra es aleatoria. Se busca un modelo de función que explique la relación entre variables, me ayuda a inferir valores de Y.

Correlación: 2 variables aleatorias. Mide la magnitud, grado o fuerza de asociación entre variables.

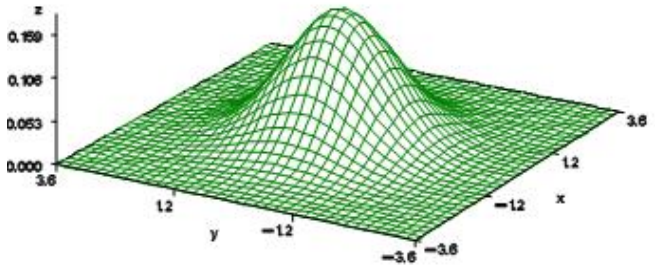
CORRELACIÓN



Supuestos de Correlación:

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

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



Coeficiente de correlación

Poblacional: ρ

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

Muestral: r

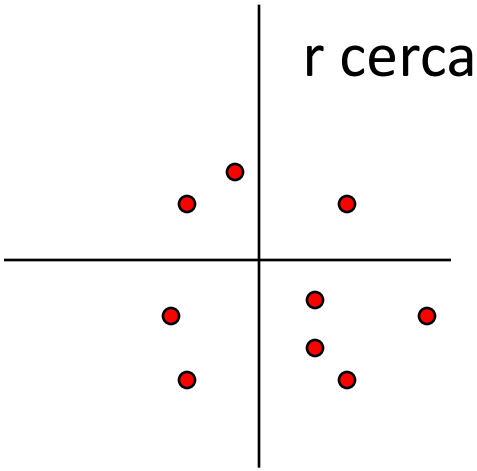
$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Fórmula recomendada de trabajo

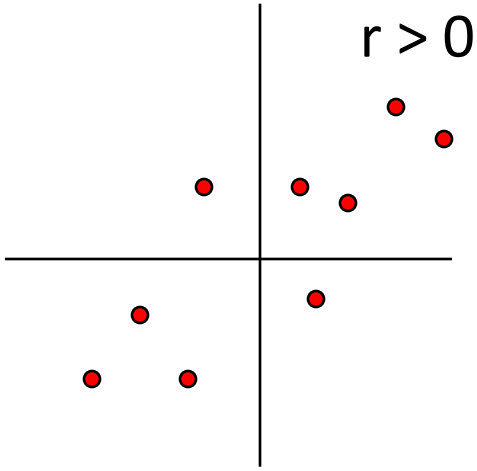
$$r = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

n: pares de valores

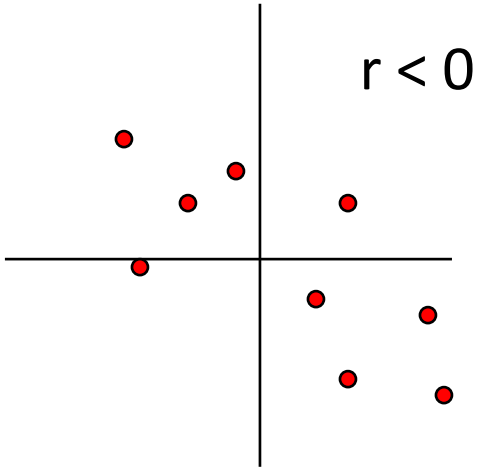
Ejemplos de correlación



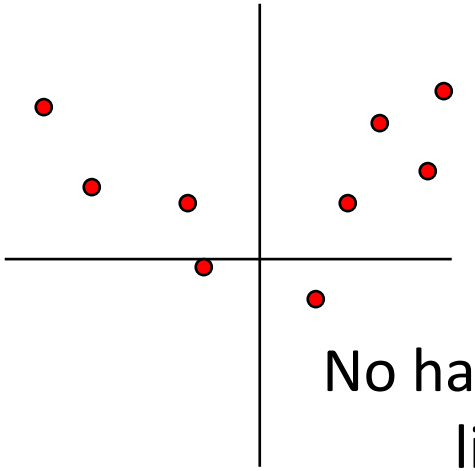
r cerca de 0



r > 0 (+)



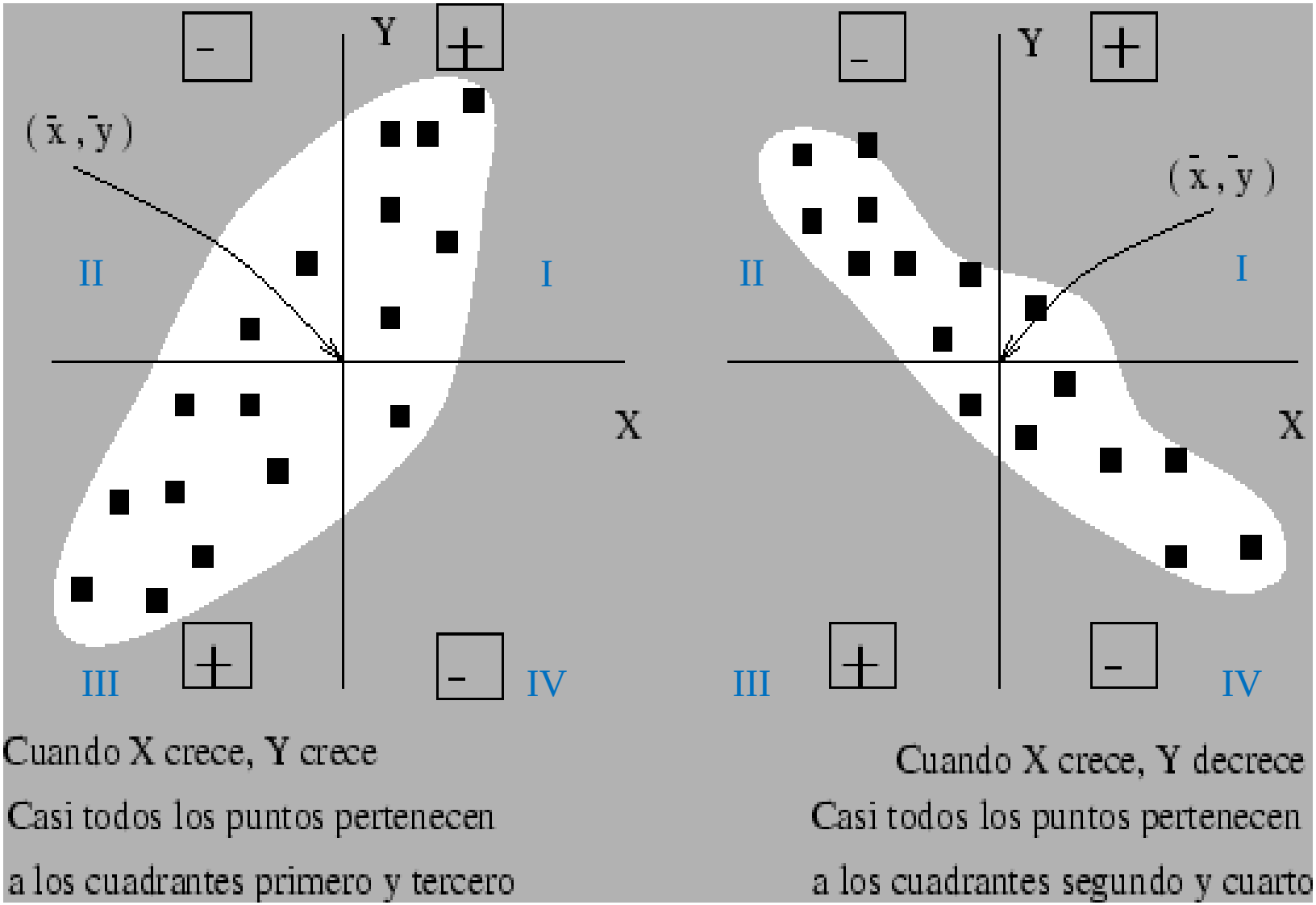
r < 0 (-)



No hay relación lineal

CORRELACIÓN

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



Coeficiente de correlación lineal de Pearson (ρ)

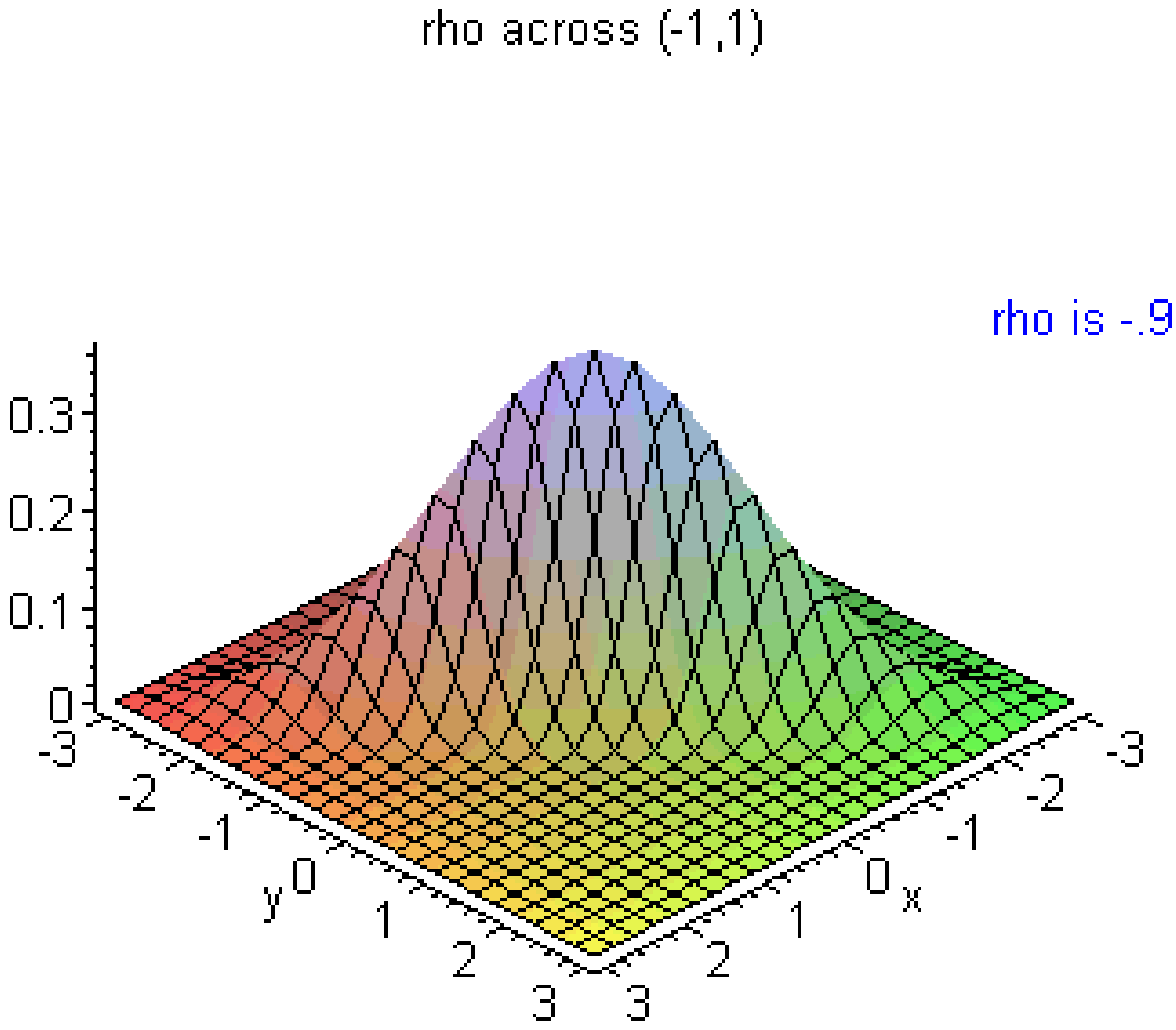
Su estimador es r → (coeficiente de correlación muestral)

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

Propiedades de ρ

- Es adimensional.
- Sólo toma valores entre [-1,1].
- Cuando 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.

CORRELACIÓN



EJEMPLO

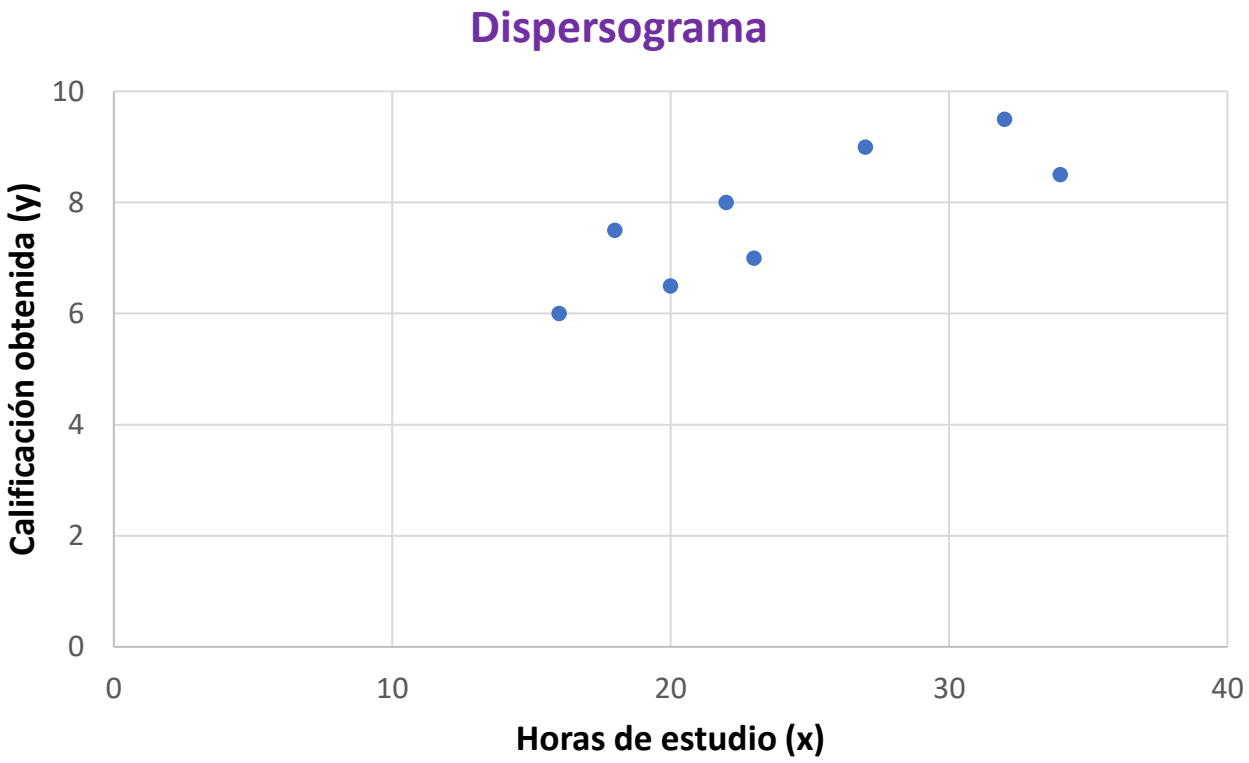
Se seleccionan al azar 8 estudiantes a los que se les consulta sobre: El número de horas dedicadas al estudio de una asignatura y la calificación obtenida en el examen correspondiente. Designamos con (X) a las Horas y con (Y) a la Calificación obtenida.
Los datos obtenidos se muestran en la siguiente 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? Es decir, ¿ambas variables son aleatorias?, si es así ¿Cuánto vale el coeficiente de correlación lineal?

CORRELACIÓN

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



CORRELACIÓN

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			

CORRELACIÓN

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	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

CORRELACIÓN

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

n=8	∑x=192	∑y=62	∑xy=1534	∑x^2= 4902	∑y^2=491

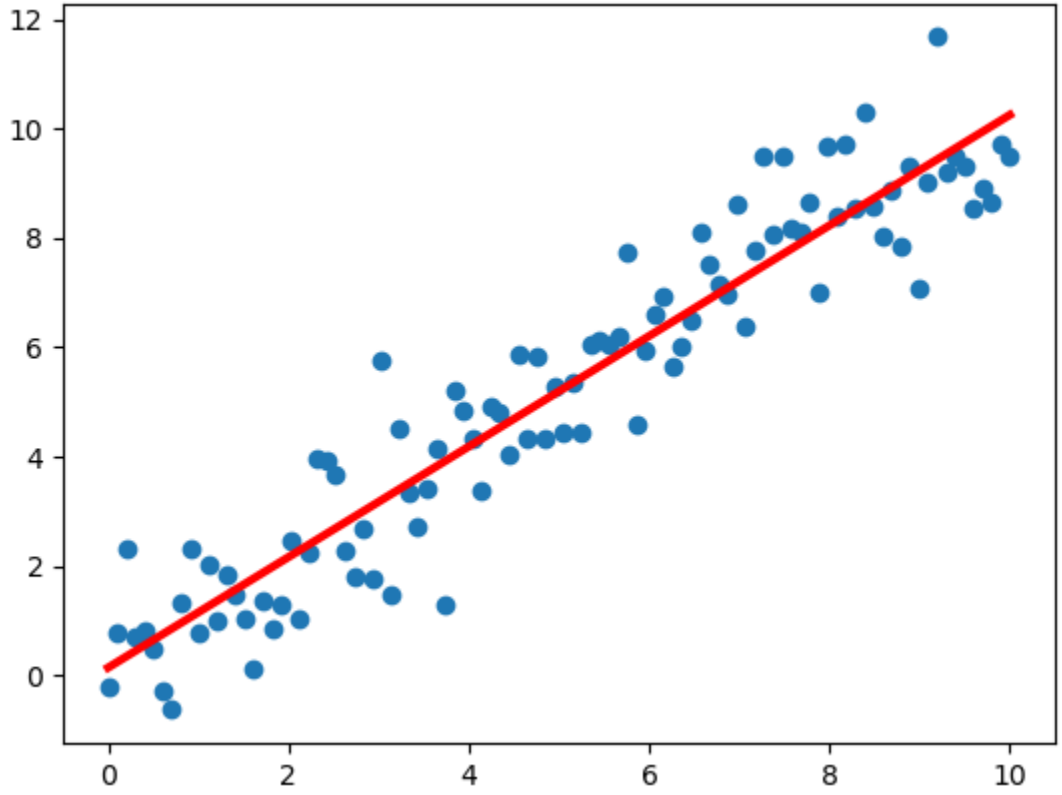
$$r = \frac{\left(\sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\sqrt{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

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

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

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



CONCEPTOS:

- **Regresión simple:** interviene una sola variable independiente (→ este caso veremos nosotros)
- **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

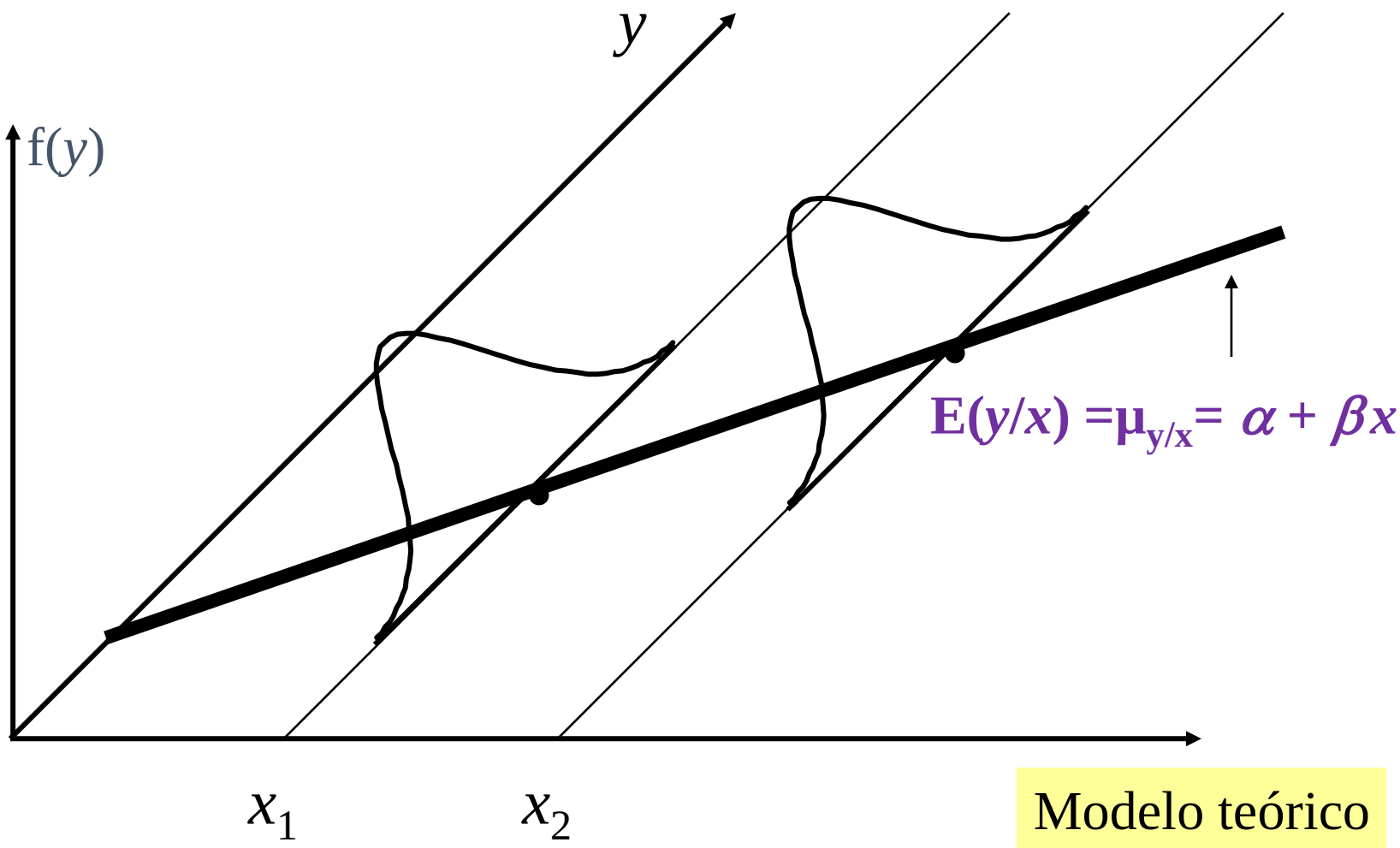
OBJETIVO:

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. ALEATORIA.

•La **variable X**: que es la independiente (predictora, explicativa, exógena). Es la variable que provee las **bases para estimar Y**. FIJADA POR EL INVESTIGADOR, MEDIDA SIN ERROR.

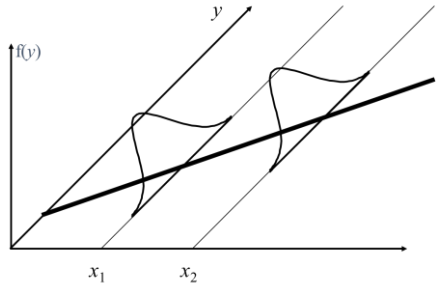
Regresión Lineal Simple



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

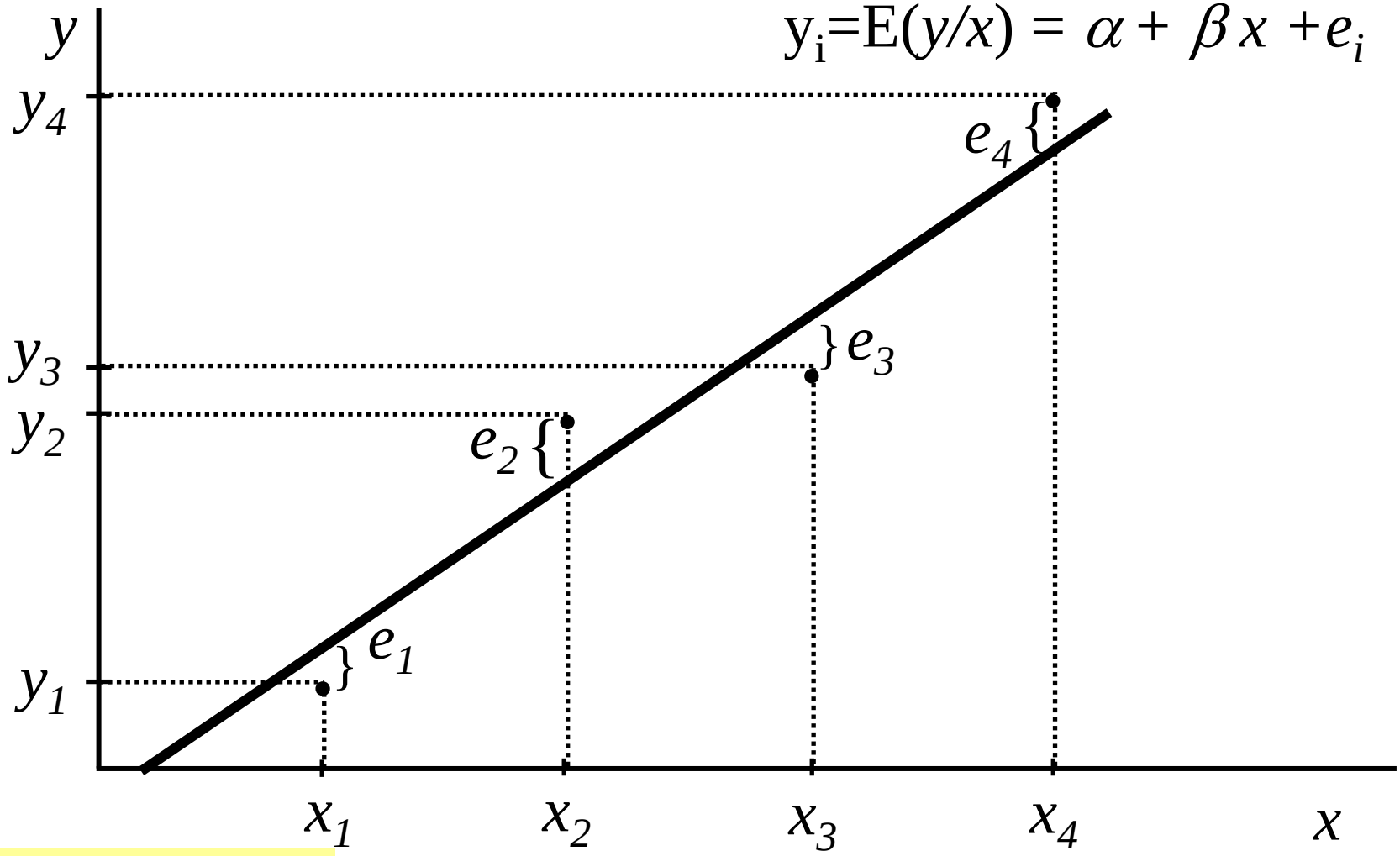
$$\mu_{y/x} = \alpha + \beta x$$

Ordenada al origen

Pendiente de la recta

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

Se eleva al cuadrado para que no se compensen las diferencias

Derivando $\frac{\partial \sum e^2}{\partial \alpha} = 0$ $\frac{\partial \sum e^2}{\partial \beta} = 0$

Hacemeos las derivadas parciales e igualamos a 0 para minimizar el error

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

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

Ordenada al origen (α)

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

Pendiente de la recta (β)

Modelo estimado:

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

Donde:

(a) es un estimador de α (ordenada al origen)

(b) es un estimador de β (pendiente de la recta)

Además

$$e \cong N(0, \sigma^2)$$

Interpretación de los coeficientes de regresión estimados

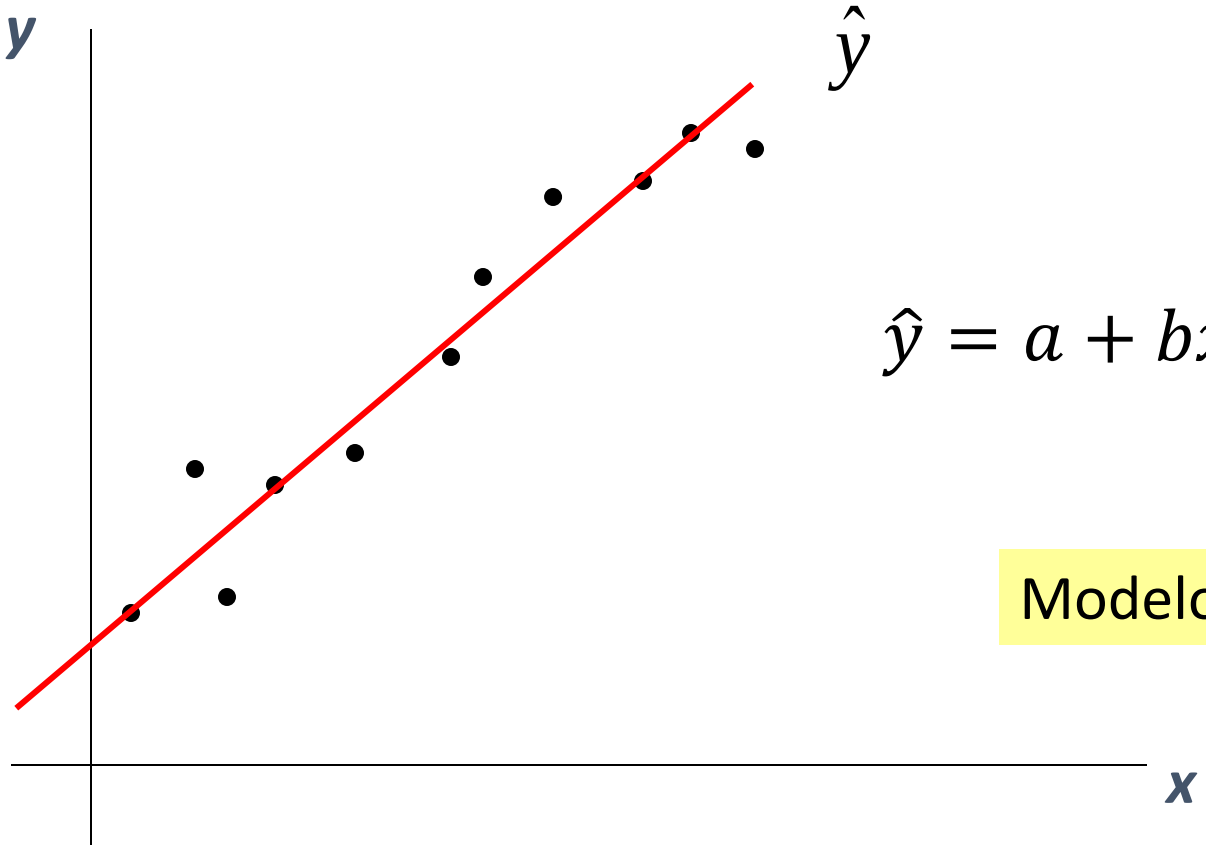
La pendiente "b" indica el cambio promedio estimado en la variable respuesta cuando la variable predictora aumenta en una unidad adicional.

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.

La recta de regresión de mejor ajuste es válida entre el rango de valores que nosotros tomamos, si extrapolamos hacia valores menores o mayores no sabemos si el comportamiento sigue siendo lineal o no.

REGRESION LINEAL SIMPLE

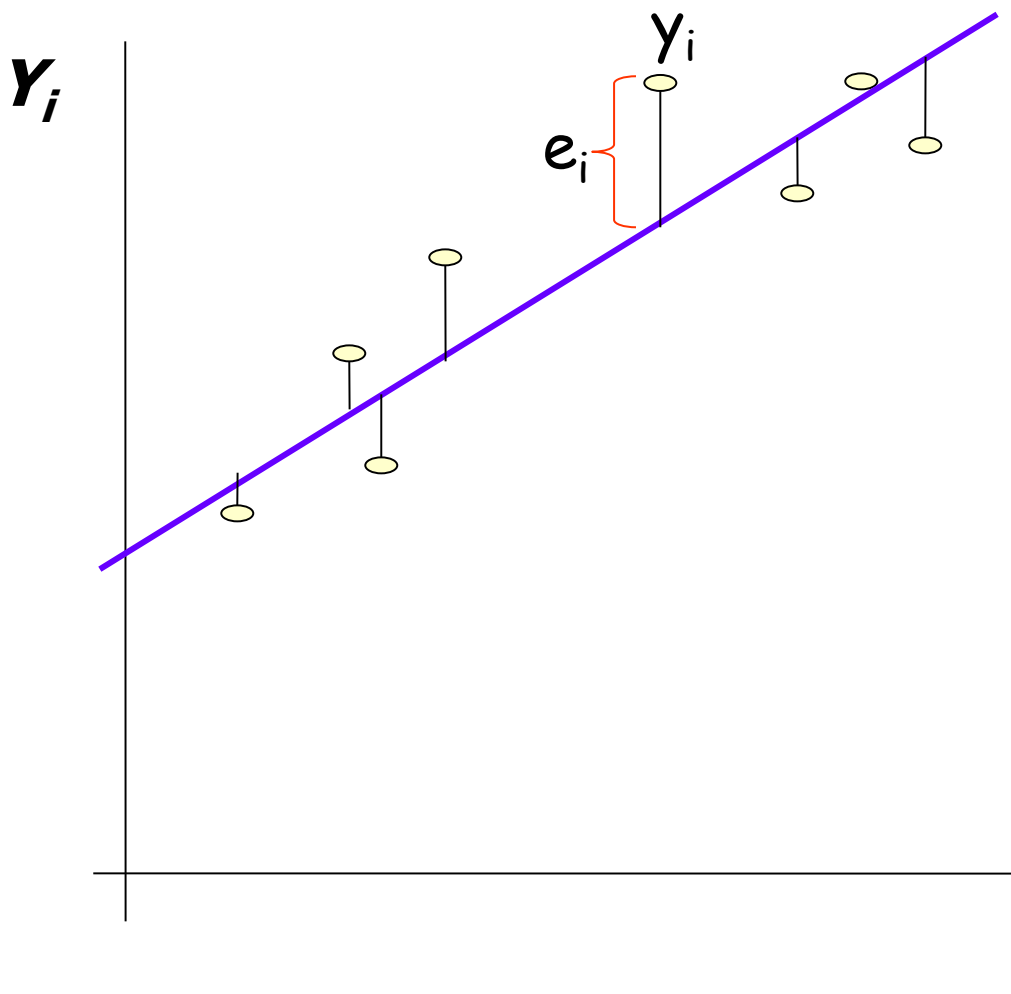
Estimar los valores de *y* (variable dependiente) a partir de los valores de *x* (variable independiente)



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

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 de regresión

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

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

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

Derivación

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

$$\frac{\partial \sum e_i^2}{\partial a} = 0$$

$$\frac{\partial \sum e_i^2}{\partial b} = 0$$

Derivada respecto de *a*

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

divido ambos miembros por -2= $\sum (y_i - a - bx_i) = 0$

distribuyo la sumatoria $\sum y_i - a\sum 1 - b\sum x_i = 0$

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

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

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

x/n y/n son las medias $\bar{y}_i - b\bar{x}_i = a$

Derivada respecto de b

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

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

divido ambos miembros por -2 y distribuimos la sumatoria

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

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

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

reacomodamos

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

reemplazamos medias por x/n o y/n
dejamos términos con b a la derecha

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

reacomodamos

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

sacamos factor común b

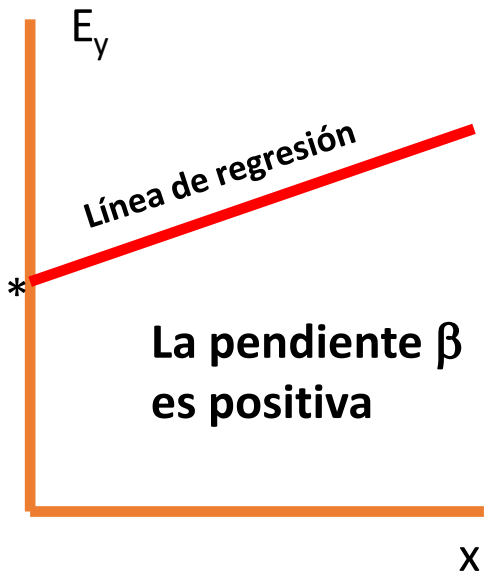
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\left(\sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n} \right)}{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n} \right)}$$

CoVar xy

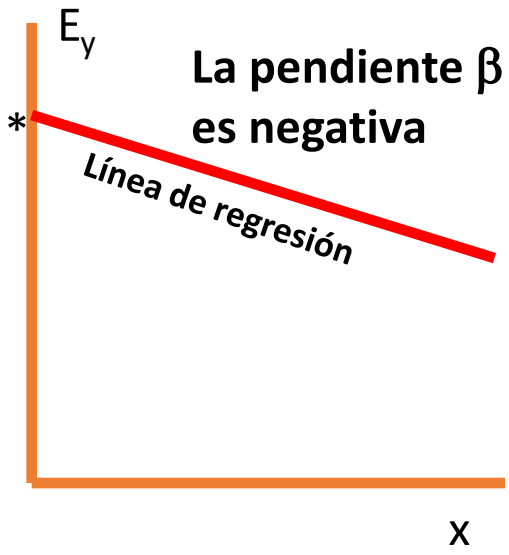
Var x

Casos posibles de Regresión Lineal Simple

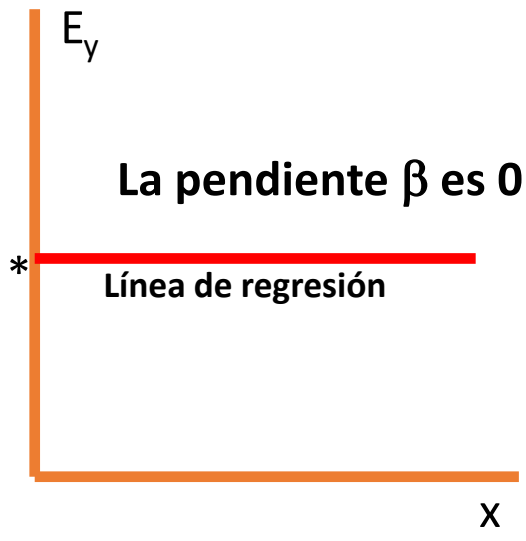
Relación lineal positiva



Relación lineal negativa



No hay relación



* **Ordenada al origen α**

Retomemos el ejemplo visto en Correlación…

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

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

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

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

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

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

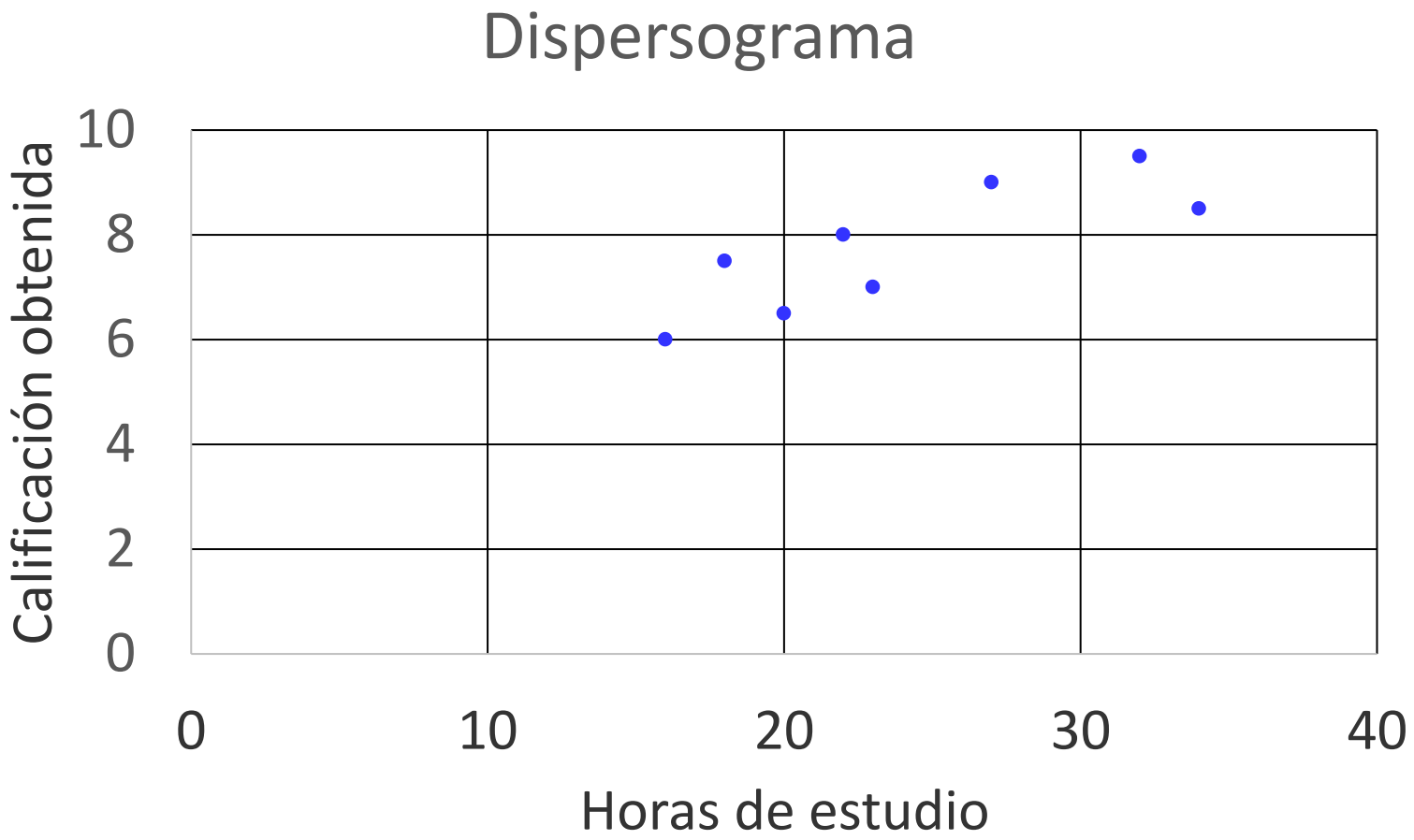
Calculamos a: $a = \bar{y}_i - b\bar{x}_i = \frac{62}{8} - 0{,}1565\,\frac{192}{8} = 3{,}99$

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

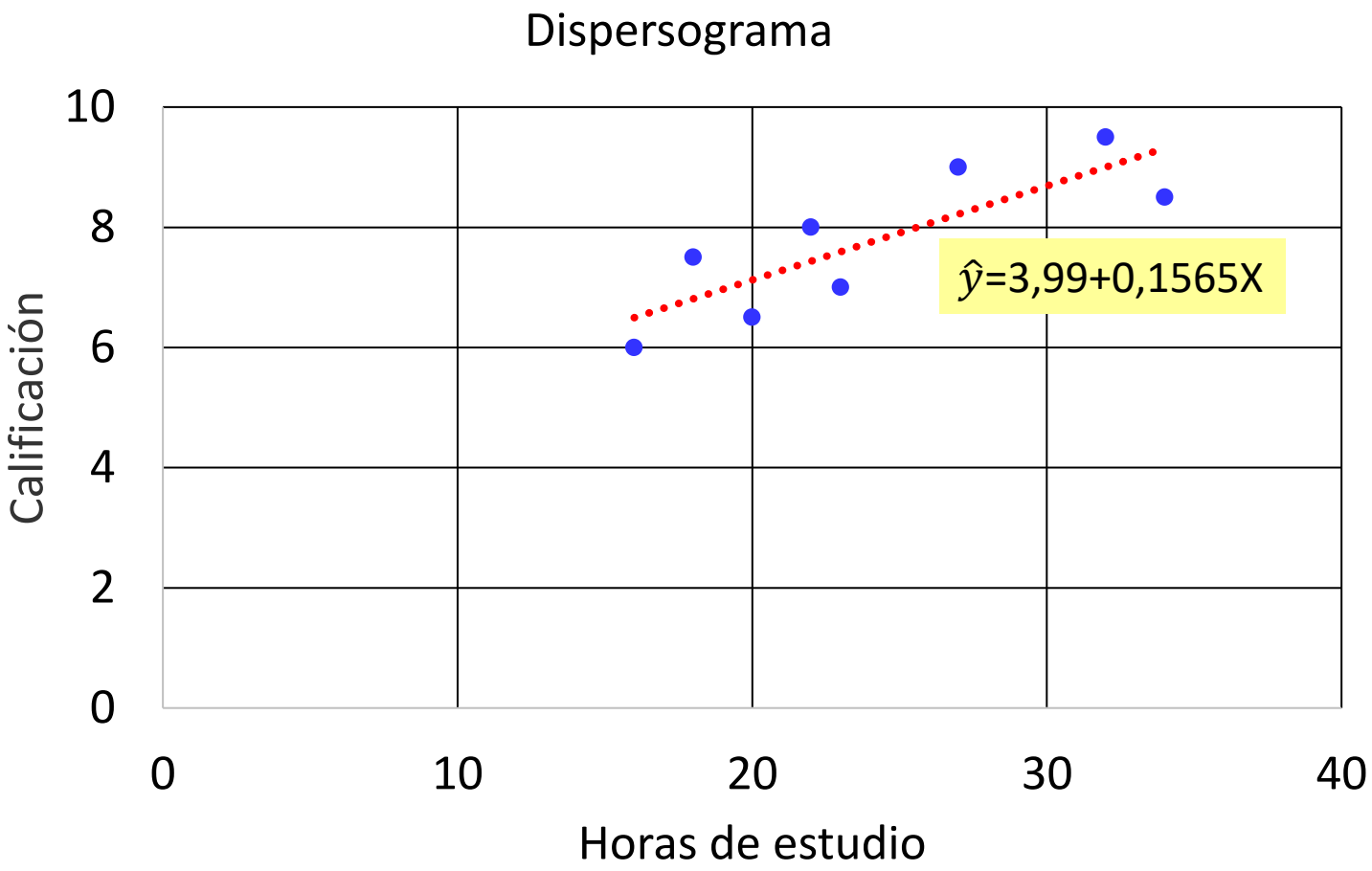
Reemplazamos x por 28 y obtenemos:

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

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


Dispersograma
Calificación obtenida
10
8
6
4
2
0
0
10
20
30
40
Horas de estudio

REGRESIÓN


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

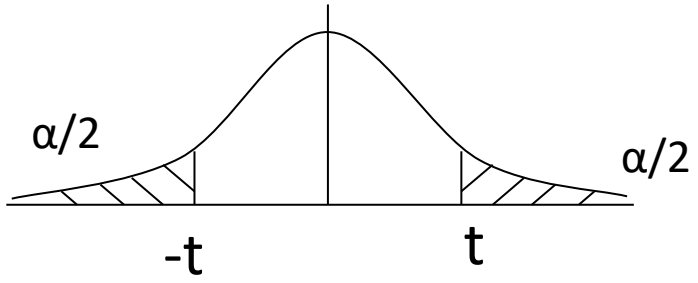
INFERENCIA EN CORRELACIÓN

Correlación:

hipótesis para el coeficiente de correlación

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

4-



2- α =.....

Regla de decisión…

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

En correlación sólo podemos contrastar la hipótesis de $\rho = 0$ contra $\rho \neq 0$; no se pueden usar valores puntuales. Para eso se usa otra variable pivotal (no la vemos en esta materia)

gl: n-2 (por los parámetros α y β)
n: cantidad de pares de valores

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			

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 ∑x=24 ∑y=60 ∑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%

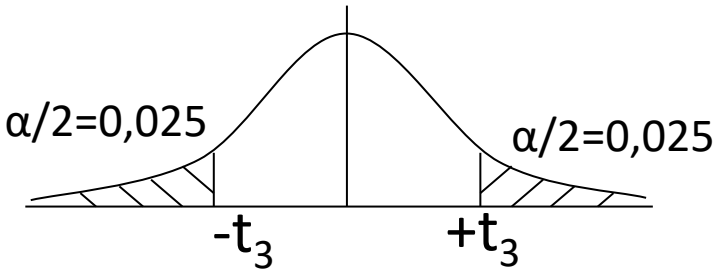
1- $\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$

2- α =0,05

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

gl: n-2
n: 5 (en este caso)
gl: 3

4-



t= +/- 3,1824

Regla de decisión

Rechazo H_0 si tcal ≥3,184 ó tcal ≤ -3,1824

No rechazo H_0 si -3,1824< tcal < 3,1824

5- 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 $t_{calculado}$ se encuentra -3,1824<t cal < 3,1824

No rechazo H_0

<u>Conclusión</u>: Con un nivel de significación del 5% no tengo evidencias suficientes 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 EN REGRESIÓN

Estimadores

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

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

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

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

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

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 (b)

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

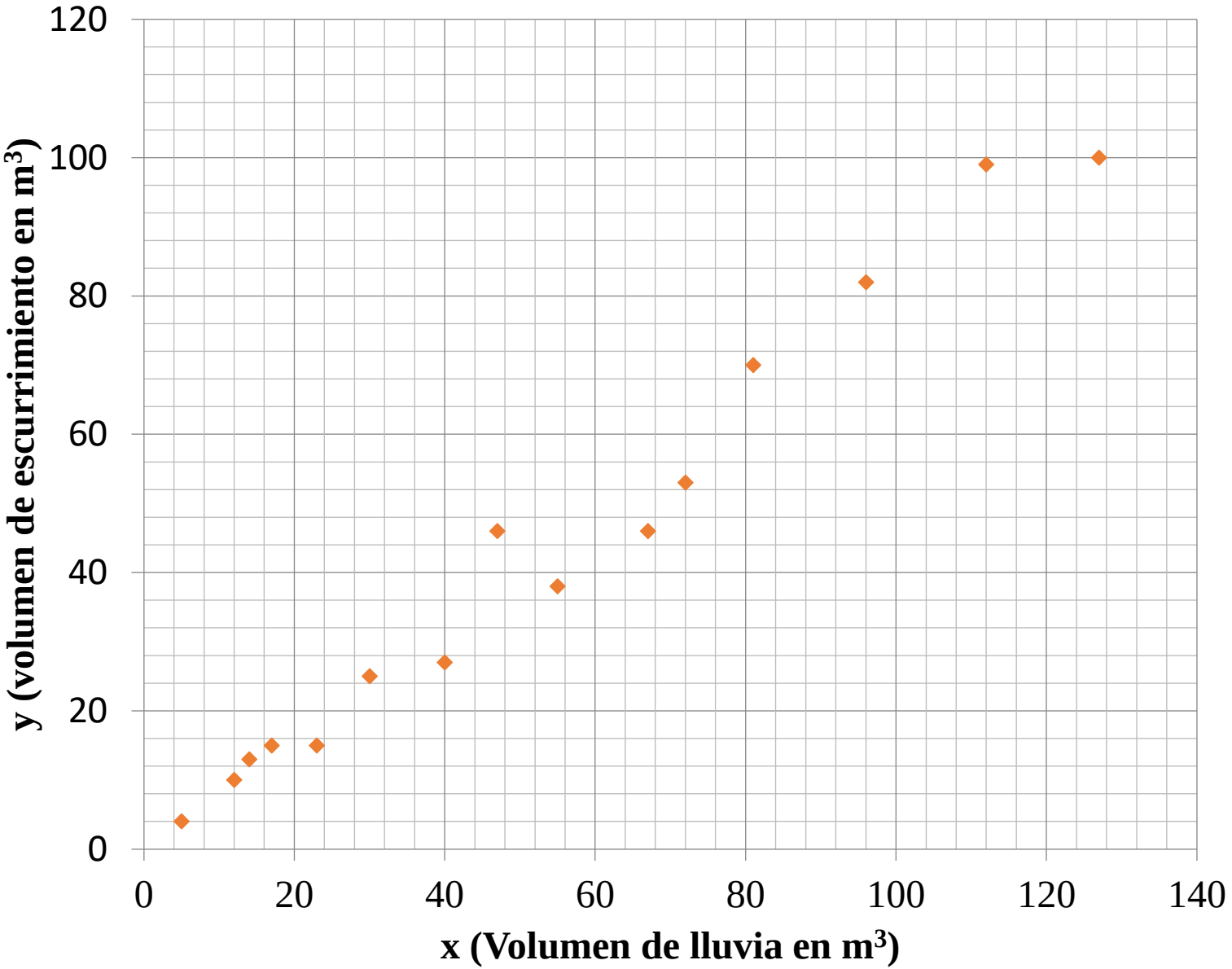
Sea X el volumen de precipitación pluvial -lluvia (m^3) e Y el volumen de escurrimiento (m^3) en determinado lugar. Se realizaron 15 observaciones que se muestran en la tabla siguiente:

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.


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

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**

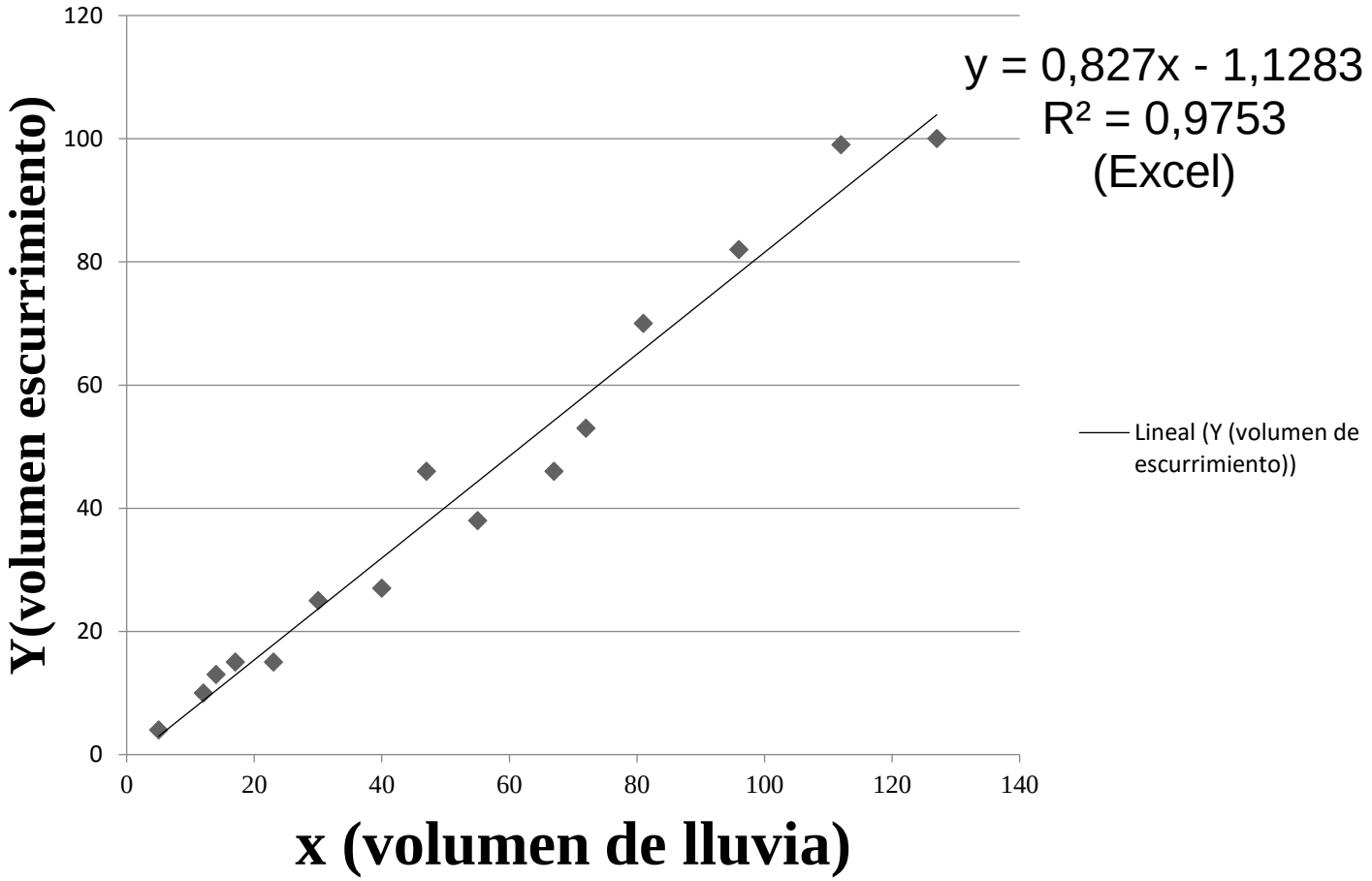
Coeficiente de correlación

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

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- α =0,05

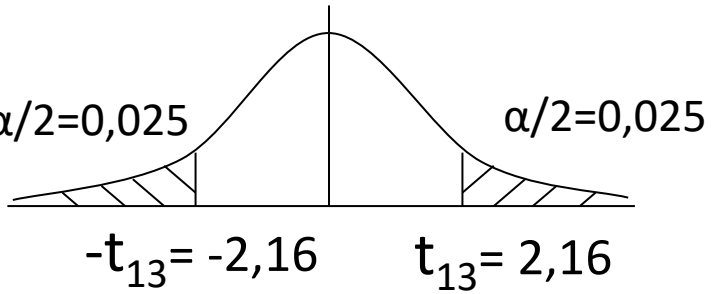
3- Variable pivotal

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

También se puede usar para hipótesis con valores puntuales

gl: n-2
n: 15 (en este caso)
gl: 13

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 ∑x=798 ∑y=643 ∑xy=51232 ∑x^2= 63040 ∑y^2= 41999

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

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

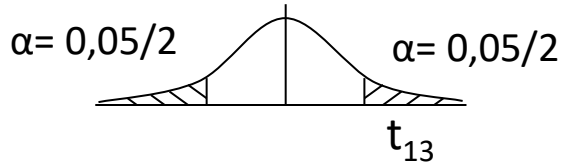
$$S_b = \frac{S_e}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sqrt{27{,}46}}{\sqrt{20586{,}4}} = 0{,}0365$$

Prueba de hipótesis para el coeficiente de regresión β

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

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

α= 0,05



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

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

Excel

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

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

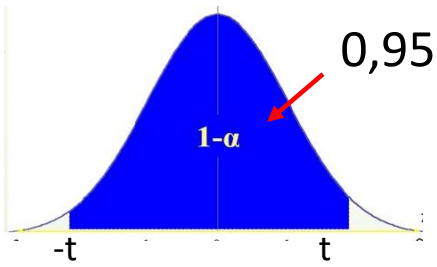
Conclusión: Con un nivel de significación del 5% tengo evidencias suficientes para suponer que existe una relación funcional poblacional donde, el volumen de escurrimiento está en función del volumen de precipitación pluvial, o que sea, por cada metro^3 que se incrementa la precipitación el volumen de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

ANÁLISIS DE LA VARIANZA (ANOVA) EN REGRESIÓN LINEAL SIMPLE

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

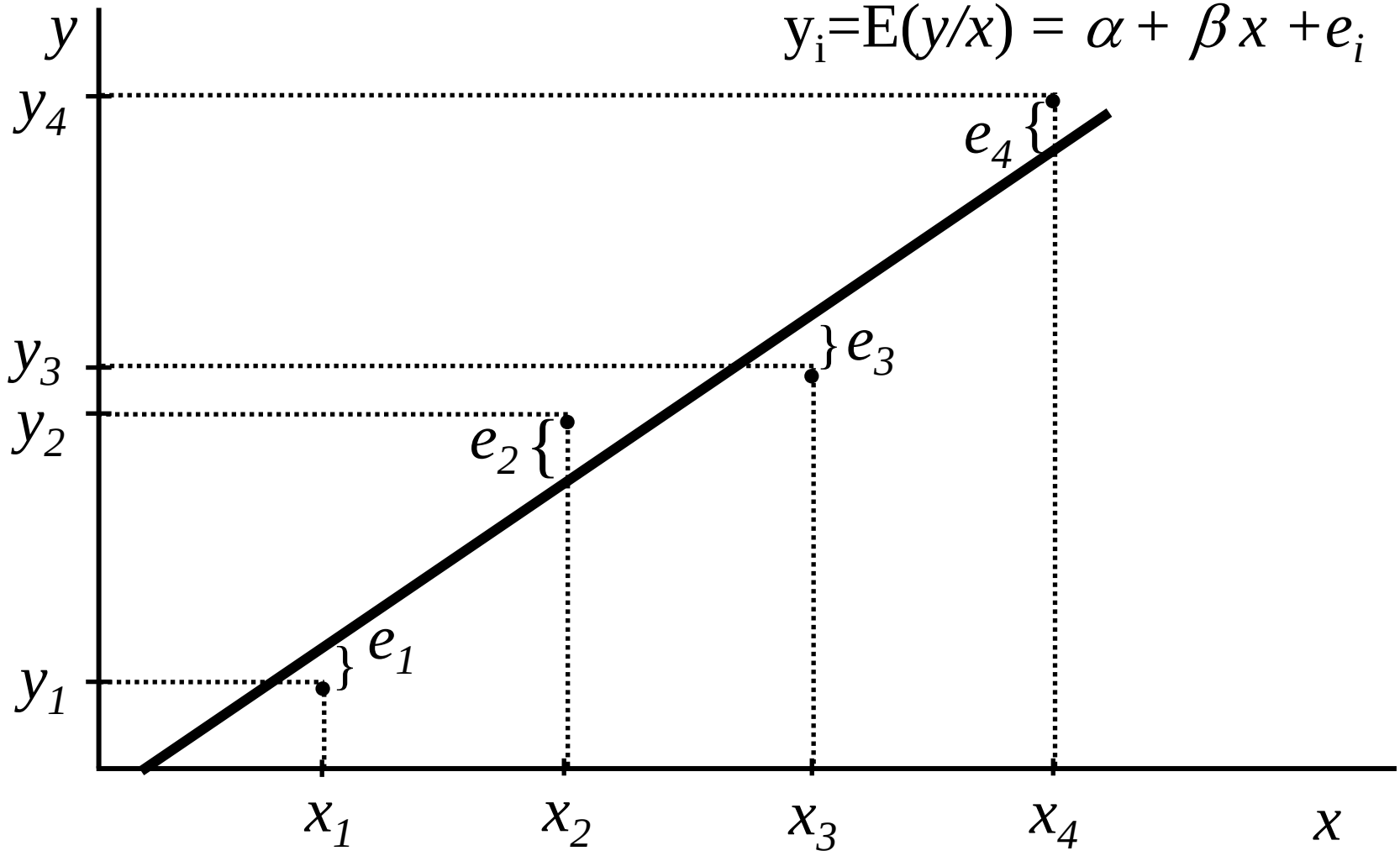
La variable X: que es la independiente (predictora, explicativa, exógena). Es la variable que provee las bases para estimar

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 estimado:

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

Donde:

(a) es un estimador de α

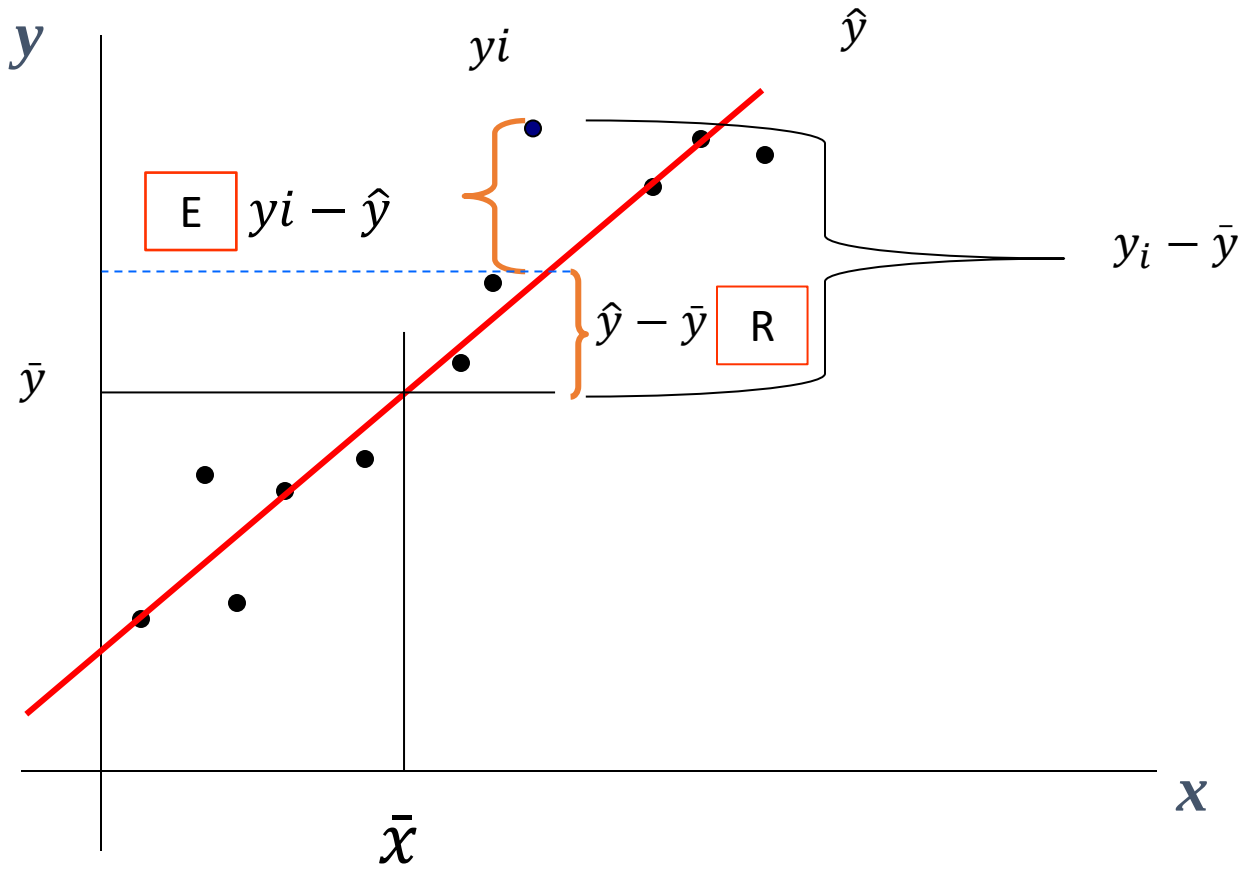
(b) es un estimador de β

Además

$$e \cong N(0, \sigma^2)$$

ANALISIS DE LA VARIANZA EN REGRESION LINEAL SIMPLE

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



Análisis de Varianza en el análisis de regresión

- El enfoque desde el análisis de varianza se basa en la partición de sumas de cuadrados y grados de libertad asociados con la variable respuesta Y.
- La variación de los Y_i se mide convencionalmente en términos de las desviaciones

$$(Y_i - \bar{Y}_i)$$

- La medida de la variación total SC_{tot}, es la suma de las desviaciones al cuadrado

$$\sum (Y_i - \bar{Y}_i)^2$$

Desarrollo formal de la partición

Consideremos la desviación

$$(Y_i - \bar{Y}_i)$$

Podemos descomponerla en

$$\underset{(T)}{(Y_i - \bar{Y})} = \underset{(R)}{(\hat{Y}_i - \bar{Y})} + \underset{(E)}{(Y_i - \hat{Y}_i)}$$

(T): desviación total

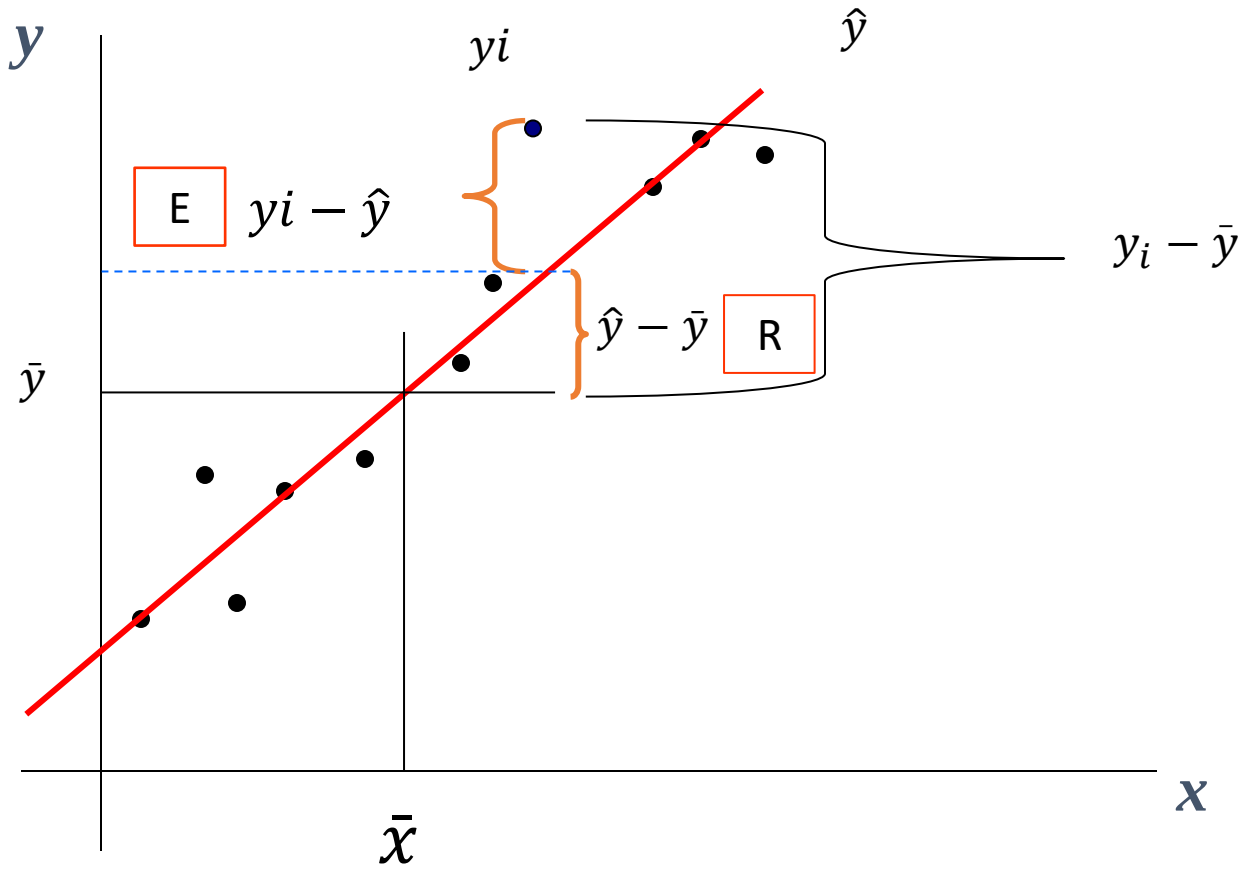
(R): es la desviación del valor ajustado por la regresión con respecto a la media general

(E): es la desviación de la observación con respecto a la línea de regresión

REGRESION LINEAL SIMPLE

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

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



Desarrollo formal de la partición

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Si consideramos todas las observaciones y elevamos al cuadrado para que los desvíos no se anulen:

$$\sum (Y_i - \bar{Y})^2 = \sum [\ (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

Desarrollando el cuadrado del binomio obtenemos:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum \left(\hat{Y}_i - \bar{y}\right)^2$$

Si consideramos en el segundo término: $e_i = y_i - \hat{y}_i$

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum e_i(\hat{y}_i - \bar{y}) = \sum e_i(a + bx_i) - \bar{y}\ \sum e_i =$$

$$= a\sum e_i + b\sum e_i x_i = 0$$ e_i es independiente de x_i

Aplicando propiedad distributiva el 2do término se anula quedando

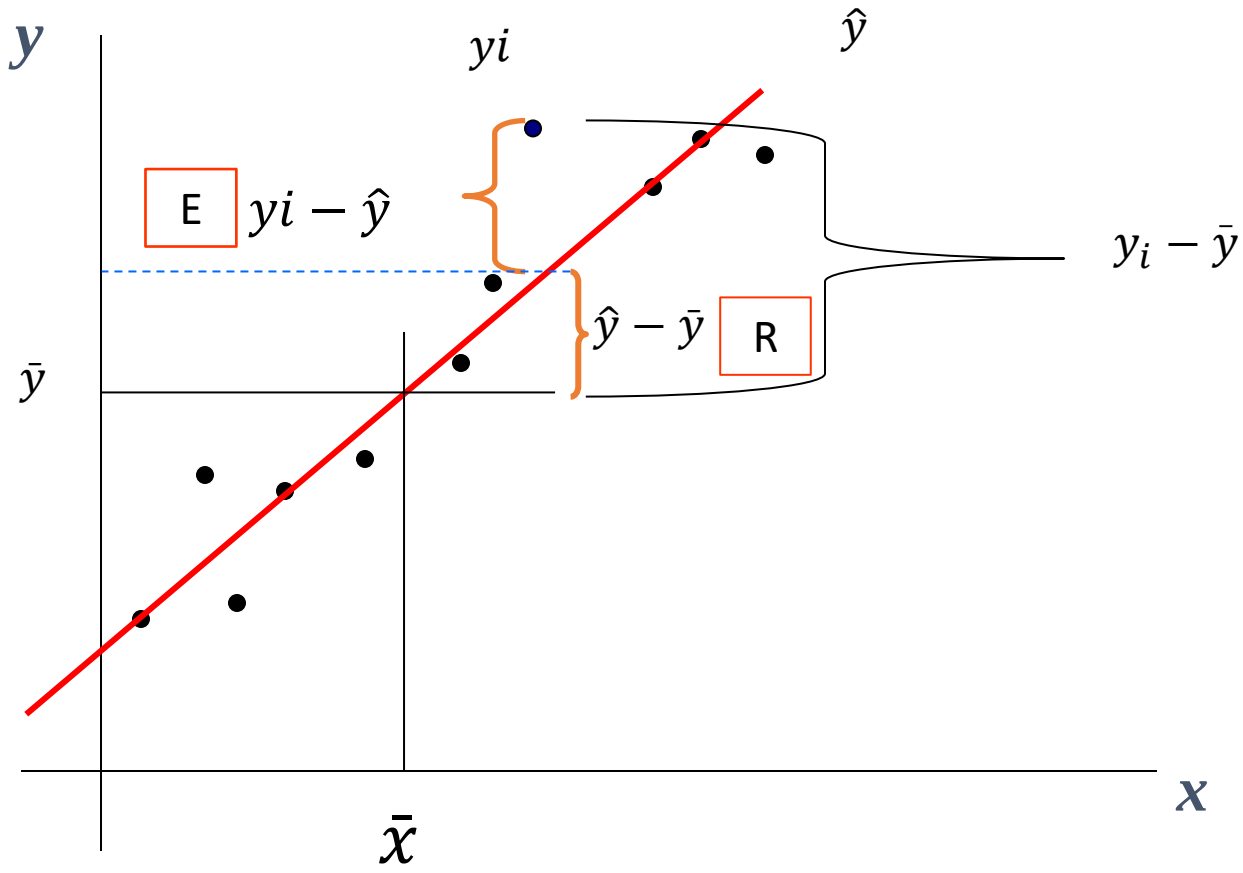
$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{y})^2 + \sum \left(\hat{Y}_i - \bar{y}\right)^2$$

E R

REGRESION LINEAL SIMPLE

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

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



Desarrollo formal de la partición

Si consideremos todas las observaciones y elevamos al cuadrado para que los desvíos no se anulen

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{SC_{tot}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{SC_{reg}} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SC_{er}}$$

(SC_{tot}): Suma de cuadrados total

(SC_{reg}): Suma de cuadrados de la regresión

(SC_{er}): Suma de cuadrados del error

Dividiendo por los grados de libertad, (n-1), (1) y (n-2), respectivamente cada suma de cuadrados, se obtienen los cuadrados medios del análisis de variancia.

Cada uno de estos cuadrados medios tiene una distribución χ^2

Suma de cuadrados debido a la regresión

$$\boldsymbol{SCregresión} = \sum (\hat{Y}_i - \bar{y})^2$$

Reemplazamos $\hat{y} = a + bx$ y también $\bar{y} = a + b\bar{x}$

Obtenemos $SCregresión = \sum (a + bx_i - (a + b\bar{x}\,))^2$

Cancelamos "a" $SCregresión = \sum (bx_i - b\bar{x})^2$

$$\boldsymbol{SCregresión} = \boldsymbol{b^2 \sum (x_i - \bar{x})^2}$$

Estimación de la varianza de los términos del error (σ^2)

$$SC_{error}=\sum(Y_i-\hat{y})^2=\sum(Y_i-a-bx_i)^2=\sum e_i{}^2$$

También se puede obtener por diferencia

$$SC_{error}=SC_{total}-SC_{regresión}$$

$$SC_{error}=\sum(Y_i-\hat{y})^2=\sum(Y_i-\bar{Y})^2-\sum\left(\hat{Y}_i-\bar{y}\right)^2$$

Reemplazando $$SC_{error}=\sum(Y_i-\bar{Y})^2-b^2\sum(x_i-\bar{x})^2$$

Estimación de la varianza de los términos del error (σ^2)

La suma de cuadrados del error, tiene n-2 grados de libertad asociados con ella, ya que se tuvieron que estimar dos parámetros.

Por lo tanto, las desviaciones al cuadrado dividido por los grados de libertad, se denomina cuadrados medios

$$CM_e = \frac{SC_e}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

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

Donde CM es el Cuadrado medio del error o cuadrado medio residual. Es un estimador insesgado de $\boldsymbol{\sigma^2}$

Tabla del análisis de varianza

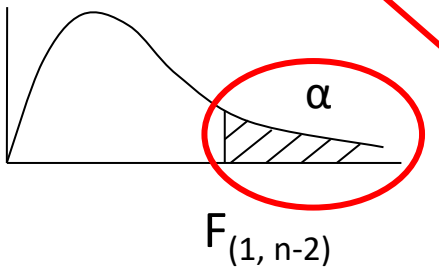
Fuentes de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	E(CM)	F
Debido a la regresión	1	SCRegresión	CMReg= SCReg/1	$\sigma^2{}_e + \beta^2 \sum (x - \bar{x})^2$	$\frac{CMReg}{CMe}$
Debido al Error	n-2	SCerror	CMe= SCe/(n-2)	$\sigma^2{}_e$	
Total	n-1	SCTotal	-----	-------	

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

Sólo se puede usar para estas hipótesis

$\alpha = \ldots$

$$F_{1;(n-2)} \cong \frac{CM_{Regresión}}{CM_{Error}}$$



Regla de decisión

Rechazo H_0 si Fcal ≥ Ftabla

No rechazo H_0 si Fcal <Ftabla

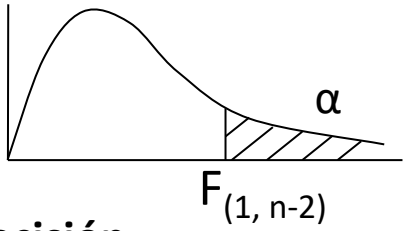
Tabla del análisis de varianza

Fuentes de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
Debido a la regresión	1	$b^2 \sum (x_i - \bar{x})^2$	CMReg=SCReg/1	$\frac{CMReg}{CMe}$
Debido al Error	n-2	SCerror	CMe=SCe/(n-2)	
Total	n-1	$\sum (Y_i - \bar{Y})^2$	-----	

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

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

$\alpha = \ldots$



$$F_{1;(n-2)} \cong \frac{CM_{Regresión}}{CM_{Error}}$$

Regla de decisión

Rechazo H_0 si Fcal ≥ Ftabla

No rechazo H_0 si Fcal <Ftabla

ANOVA EN REGRESIÓN

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:

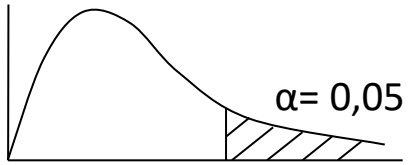
No. Observ.	X	Y
1	5	4
2	12	10
3	14	13
4	17	15
5	23	15
6	30	25
7	40	27
8	47	46
9	55	38
10	67	46
11	72	53
12	81	70
13	96	82
14	112	99
15	127	100

ANOVA en Regresión

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

α= 0,05

$$F_{1;(n-2)} \cong \frac{CM_{Regresión}}{CM_{Error}}$$



Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrados Medios	F	Valor crítico de F
Regresión	1	14078,72	14078,7216	512,65	4,667
Error	n-2=15-2=13	357,0116	27,4624372		
Total	n-1=15-1=14	14435,73			

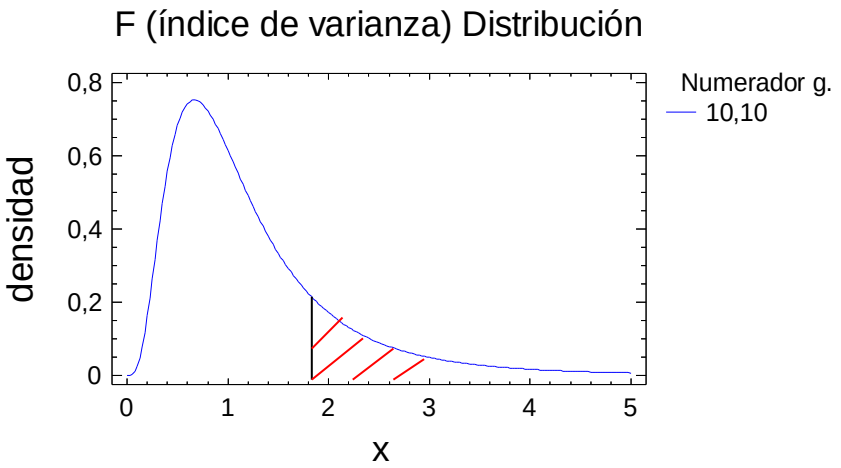
Ftabla: F1,13;0,95= 4,667. Como el valor de F calculado es mayor que 4,667 rechazo la hipótesis nula.

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, se modifica o cambia el valor medio poblacional del volumen de escurrimiento.

F de Snedecor

La distribución F es una distribución continua que relaciona dos variables aleatorias independientes con distribuciones de chi-cuadrado, cada una dividida entre sus grados de libertad. La distribución F es asimétrica hacia la derecha y es descripta por los grados de libertad de su numerador (ν_1) y denominador (ν_2).

$$P(F_{n_1,n_2} > F_{n_1,n_2,p}) = p$$



n_1=Grados de libertad del numerador

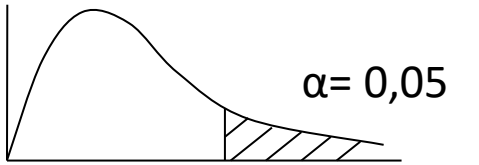
n_2=Grados de libertad del denominador

Está tabulada

$$F_{n_1,n_2,p}$$

En ANOVA en regresión
F=CMreg/CMerror

Búsqueda en la tabla F



$F_{(1, 13)}=4,667$

Valores críticos de la distribución F (cola superior)

alfa= 0,05

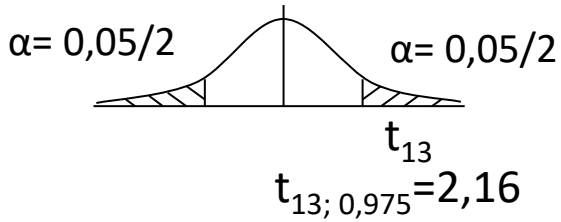
n_2 \ n_1 grados de libertad	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,91	245,36	246,46	248,01
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396	19,405	19,413	19,424	19,433	19,446
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,763	8,745	8,715	8,692	8,660
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	5,873	5,844	5,803
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	4,636	4,604	4,558
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	3,956	3,922	3,874
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	3,529	3,494	3,445
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,313	3,284	3,237	3,202	3,150
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	3,025	2,989	2,936
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	2,865	2,828	2,774
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	2,739	2,701	2,646
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	2,637	2,599	2,544
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604	2,554	2,515	2,459
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534	2,484	2,445	2,388
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	2,424	2,385	2,328

Prueba de hipótesis para el coeficiente de regresión β

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

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

α= 0,05



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

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

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 se modifica o cambia el valor medio poblacional del volumen de escurrimiento.

Coeficiente de Determinación (R^2)

Esta suma de variaciones tiene una propiedad que permite escribir:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{y})^2 + \sum \left(\hat{Y}_i - \bar{y}\right)^2$$

$SC_{total} = SC_{error} + SC_{regresión}$

Dividimos a ambos miembros por SC_{total}

$$\frac{SC_{total}}{SC_{total}} = \frac{SC_{error}}{SC_{total}} + \frac{SC_{regresión}}{SC_{total}}$$

$$1 = \frac{SC_{error}}{SC_{total}} + \frac{SC_{regresión}}{SC_{total}}$$

R^2

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{y}\right)^2}{\sum (Y_i - \bar{Y})^2} = \frac{b^2 \sum (x_i - \bar{x})^2}{\sum (Y_i - \bar{Y})^2}$$

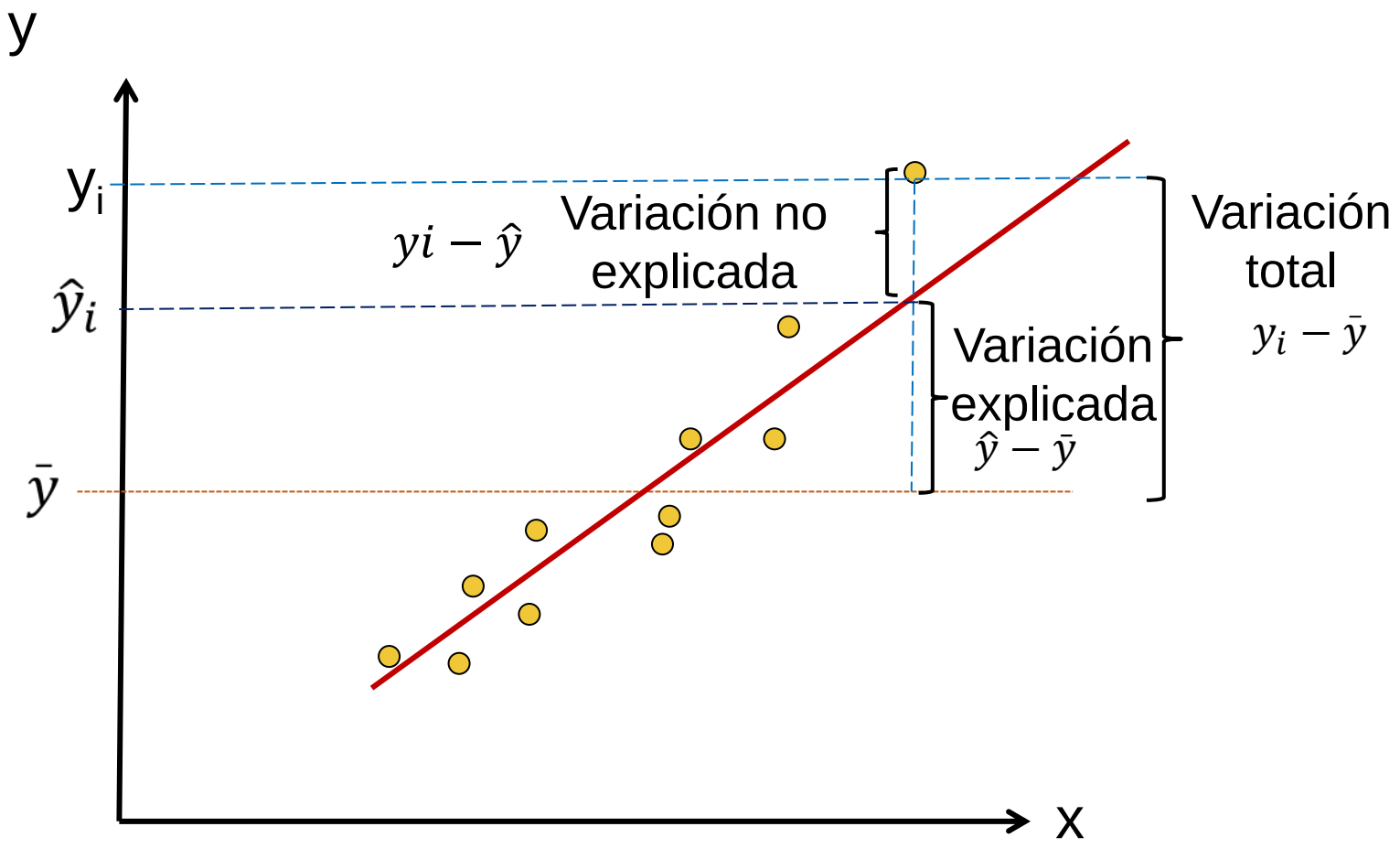
Coeficiente de Determinación (R^2)

La razón de la variación explicada a la variación total se llama Coeficiente de Determinación.

$$R^2 = \frac{\sum\left(\hat{Y}_i - \bar{y}\right)^2}{\sum(Y_i - \bar{Y})^2} = \frac{b^2 \sum(x_i - \bar{x})^2}{\sum(Y_i - \bar{Y})^2}$$

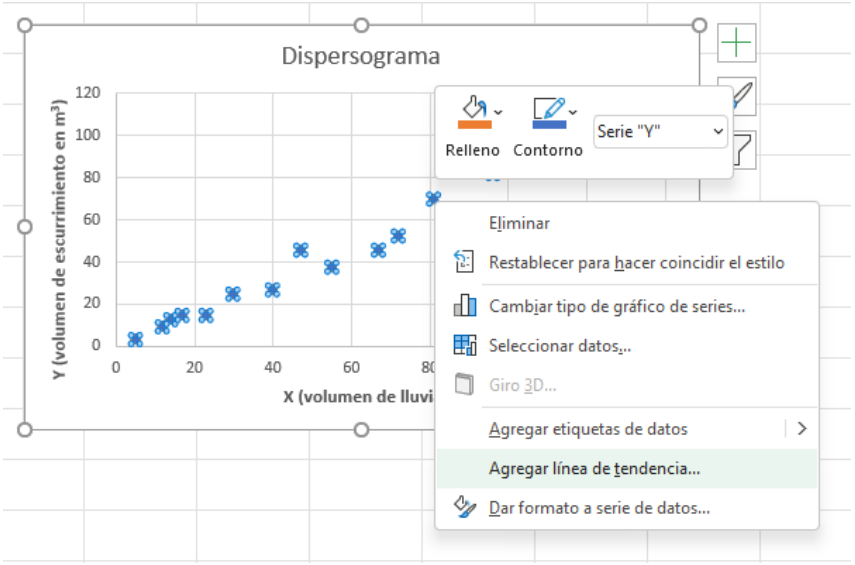
Si la variación explicada es cero, es decir, la variación total es toda no explicada, esta razón es cero.
Si la variación no explicada es cero, es decir, la variación total es toda explicada, la razón es uno.
En los demás casos **la razón se encuentra entre 0 y 1**. Puesto que la razón es siempre no negativa, se denota por R^2

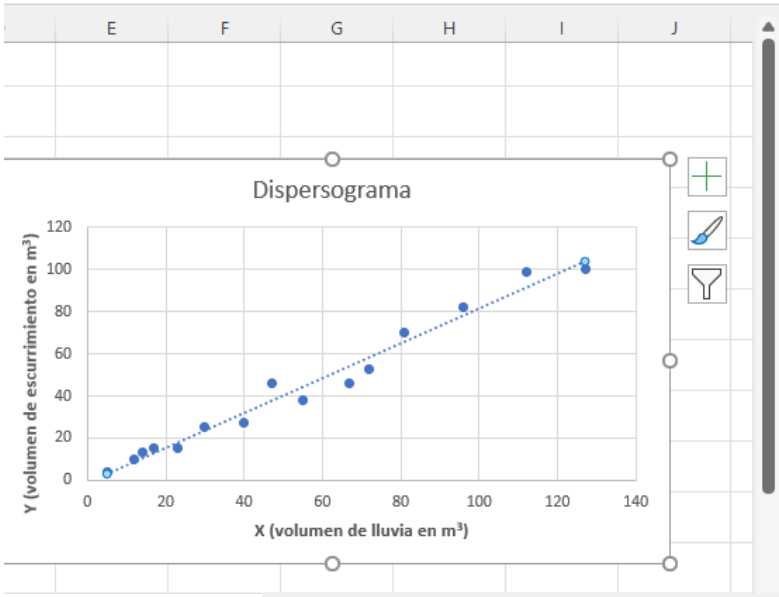
Coeficiente de Determinación (R^2)

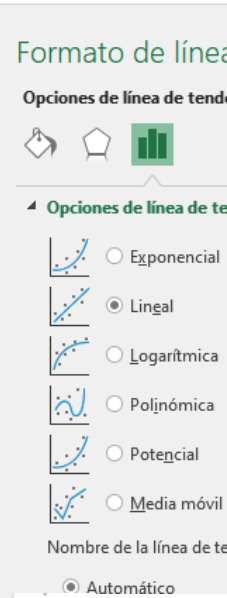


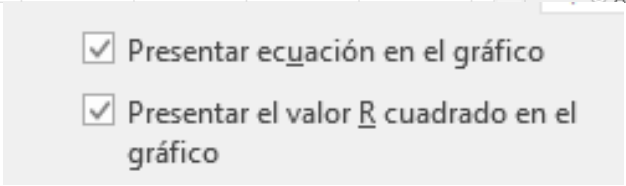
$$R^2 = \frac{\sum(\hat{Y}_i - \bar{y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{b^2 \sum(x_i - \bar{x})^2}{\sum(Y_i - \bar{Y})^2}$$

ANOVA EN REGRESIÓN

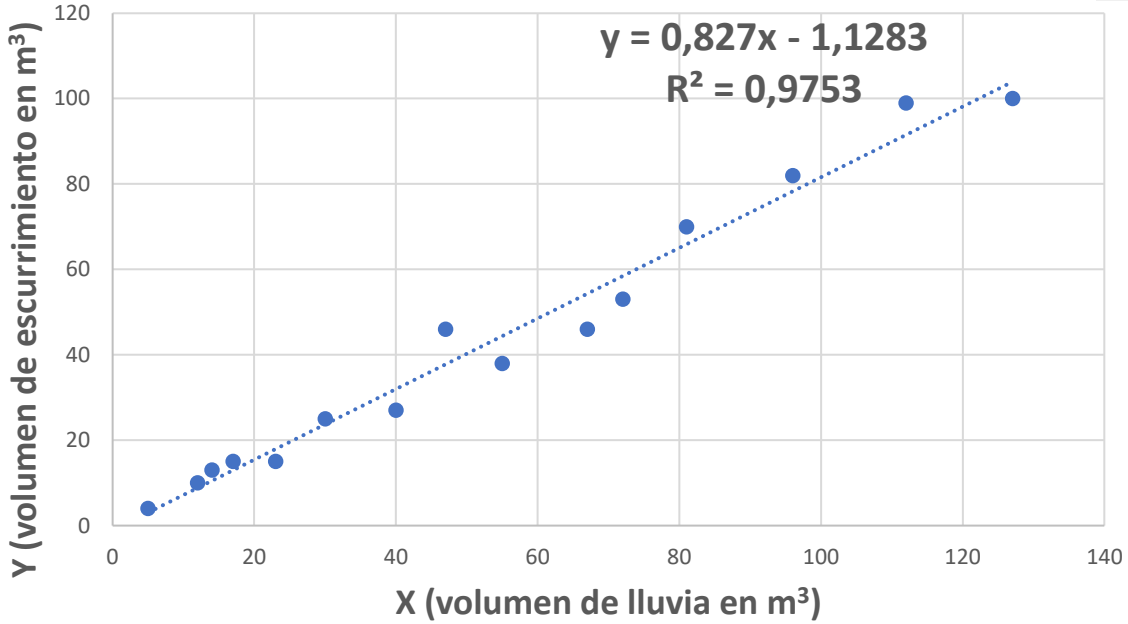








Dispersograma



Coeficiente de Determinación (R^2)

Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar.

El coeficiente de determinación, se define como la proporción de la varianza total de la variable explicada por la regresión.

Calcule el coeficiente de determinación del ejemplo que se viene desarrollando.

Coeficiente de Determinación (R^2)

$$R^2 = \frac{\sum\left(\hat{Y}_i - \bar{y}\right)^2}{\sum(Y_i - \bar{Y})^2} = \frac{b^2 \sum(x_i - \bar{x})^2}{\sum(Y_i - \bar{Y})^2}$$

Recordemos que: $\hat{y} = -1{,}1283 + 0{,}827x$

y que $\sqrt{\left(\sum x_i{}^2 - \frac{(\sum x_i)^2}{n}\right)} = \sqrt{20586{,}4}$ $\sqrt{\left(\sum y_i{}^2 - \frac{(\sum y_i)^2}{n}\right)} = \sqrt{14435{,}733}$

$$R^2 = \frac{b^2 \sum\left(\hat{Y}_i - \bar{y}\right)^2}{\sum(Y_i - \bar{Y})^2} = \frac{0{,}827^2 * 20586{,}4}{14435{,}733} = 0{,}9753$$

El 97,53% de la variabilidad está explicada por el modelo de regresión entre las variables volumen de lluvia y volumen de escurrimiento.
La variación total de la variable Y es explicada por el modelo de regresión.

Un ingeniero encargado del área de calidad de una empresa manufacturera, desea analizar la vida útil de una herramienta de corte (el tiempo que mantiene una calidad aceptable de funcionamiento) para presentar un plan de reemplazo. Ya que sin duda, las herramientas de corte pueden determinar el éxito o fracaso de un proceso de mecanizado.

Fresa

Las herramientas de corte más conocidas son: brocas, fresas, limas, sierras, herramientas de tornear, etc.

Brocas helicoidales

Teniendo en cuenta que la vida útil se ve afectada por varios aspectos como: el ambiente operacional, las condiciones de producción o de mantenimiento y el desgaste presentado por su uso, decide comenzar a investigar la relación funcional entre la velocidad de corte (metros por minuto) y el tiempo de vida (horas de uso) de la herramienta. Para ello tomó herramientas nuevas, del mismo tipo, y a cada una (al azar) las sometió a diferentes velocidades de corte registrando en cada caso la vida útil en horas. Los datos recogidos se muestran en la tabla:

ANOVA EN REGRESIÓN

Velocidad (Metros por minuto)	Vida (Horas)
20	8,7
20	9,5
25	8,5
25	7,7
25	8,4
30	7,3
30	6,1
30	7,3
35	6,8
35	5,7
35	6,1
40	4,3
40	4,2

$\Sigma x = 390$ $\Sigma x^2 = 12250$ $\Sigma y = 90,6$ $\Sigma y^2 = 663,9$

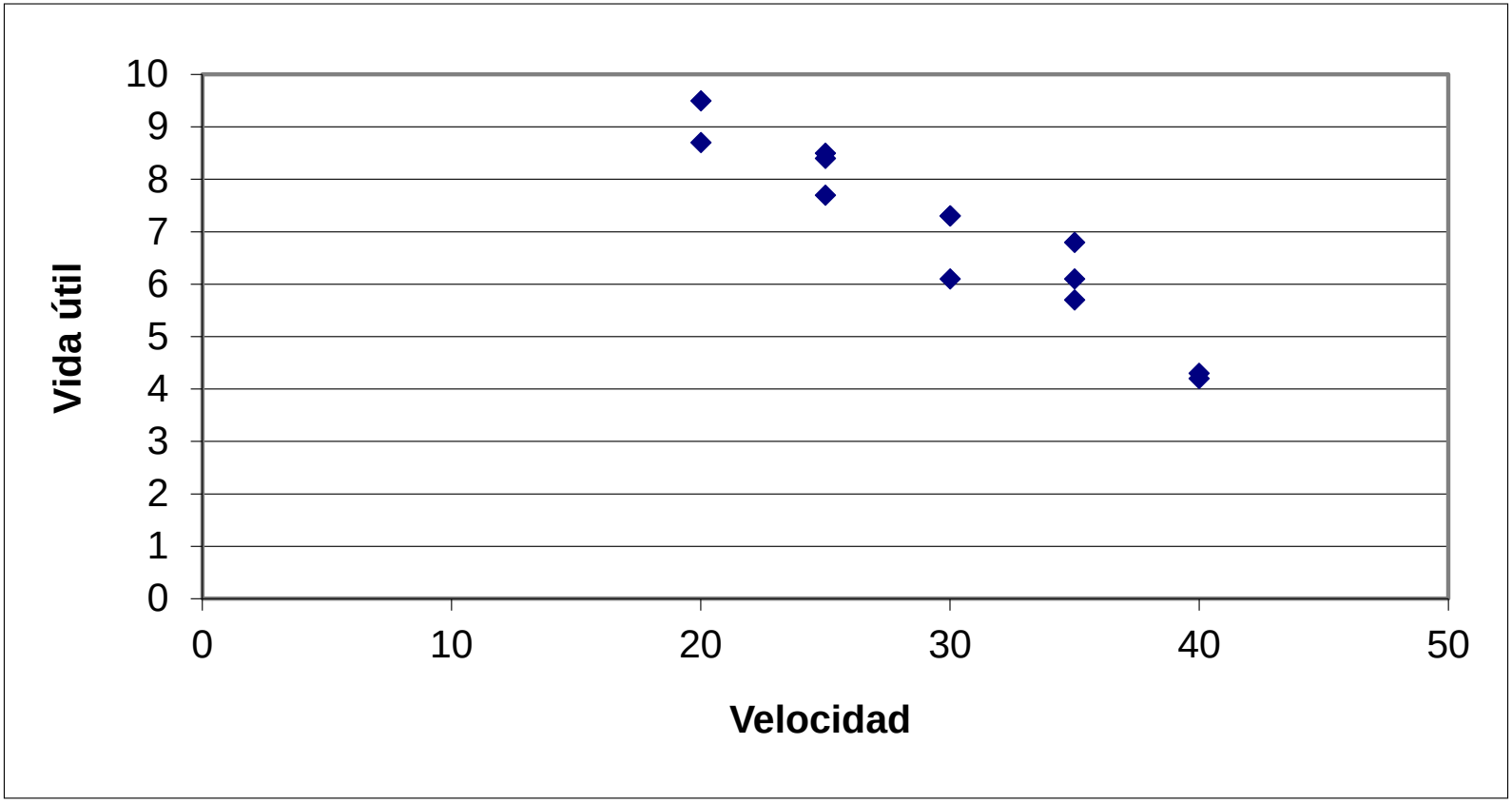
$n = 13$ $\Sigma xy = 2591$

Defina las variables. Realice el diagrama de dispersión. Analice la relación funcional entre estas dos variables.

Y: duración de la herramienta, es decir el tiempo de vida útil (en horas). Es la variable respuesta, dependiente, predicha o endógena. Es la variable que se desea predecir o estimar.

X: velocidad de corte (en m/min), que es la independiente (predictora, explicativa, exógena). Es la variable que provee las bases para estimar En este caso la variable independiente es controlable, es decir la fija el experimentador o el operario de la máquina.

Dibujar el diagrama de dispersión



Se observa que cuando aumenta la velocidad de corte (metros/minuto) disminuye el tiempo de vida (horas) de la herramienta. También se observa que la relación funcional podría ser estimada por una recta.

ANOVA EN REGRESIÓN

Suponiendo que la población sobre la que se ha tomado la muestra (de tiempos de vida tiene distribución normal para cada velocidad prefijada) realice las siguientes actividades:

a)Halle la recta de regresión por el método de los mínimos cuadrados luego dibújela en el diagrama de dispersión.

n=13 Σx = 390 Σx² =12250 Σy = 90,6 Σy² = 663,9 Σxy = 2591

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

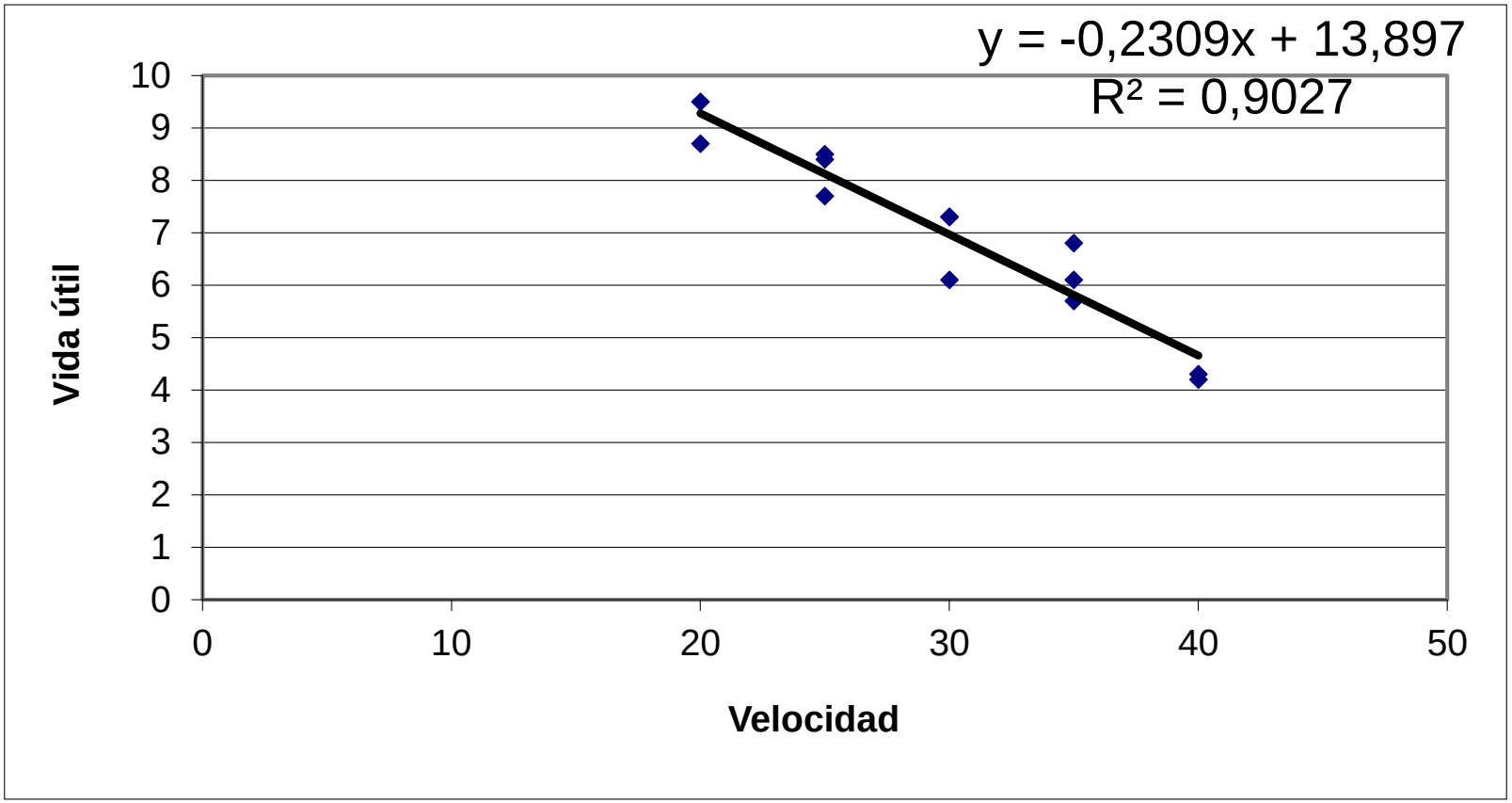
$$b = \frac{2591 - \frac{390 * 90,6}{13}}{12250 - \frac{390^2}{13}} = \frac{-127}{550} = -0,2309$$

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

$$a = \frac{90,6}{13} - 0,2309 * \frac{290}{13} = 13,897$$

$$\hat{y} = 13,897 - 0,2309x$$

ANOVA EN REGRESIÓN



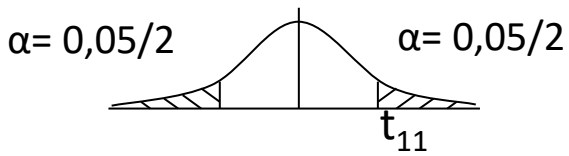
ANOVA EN REGRESIÓN

Pruebe la hipótesis más importante para decidir si continúa con el análisis del problema de regresión. Concluya e interprete. Use α= 0,05. La Prueba de hipótesis más importante es la del para el coeficiente de regresión β

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

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

α= 0,05



$t_{11;\ 0,975}$=2,201

Regla de decisión

Rechazo H_0 si tcal ≥ 2,201 ó tcal ≤ -2,201

No rechazo H_0 si
-2,201< t cal < 2,201

ANOVA EN REGRESIÓN

Cálculos

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

$$S^2{}_e = \frac{1}{11}\left[663{,}9 - \frac{90{,}6^2}{13} - (-0{,}2309) * (-127)\right] = 0{,}287$$

$$S^2_b = \frac{S_e^2}{\sum (x - \bar{x})^2} = \frac{S_e^2}{\sum x^2 - (\sum x)^2/n} = \frac{0{,}287}{550} = 0{,}000522$$

$S^2_e = 0{,}287$ $S_b = 0{,}02286$

$t_{cal} = (-0{,}2309 - 0)/0{,}02286$ $t_{cal} = -10{,}1006$

Como $t_{cal} < -2{,}201$ rechazo Ho

Conclusión: Con un nivel de significación del 5 % tengo evidencias suficientes para suponer que existe una relación funcional poblacional del tiempo de vida útil de la herramienta en función de la velocidad de corte, o que sea, por cada metro/minuto que se incrementa la velocidad de corte se modifica o cambia el valor medio poblacional del tiempo de vida útil de la herramienta.

ANOVA EN REGRESIÓN

Obtenga un intervalo de confianza del 95% para el coeficiente de regresión.

Intervalo de confianza para el coeficiente de Regresión

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

(-0,281214< β < -0,180585)

Con una confianza de 95 %, podría decir que el intervalo (-0,2812 ; -0,1805) horas/(metros/minuto) encerraría al verdadero valor de la pendiente de la recta de regresión. Esto es, con una confianza de 95 %, podría decir que el intervalo (-0,2812 ; -0,1805) horas/(metros/minuto) encerraría al verdadero cambio del promedio poblacional del tiempo de vida de la herramienta, para un aumento unitario en la velocidad de corte.

a) Seleccione un punto del diagrama de dispersión e indique la desviación sin explicar, la explicada y la total.

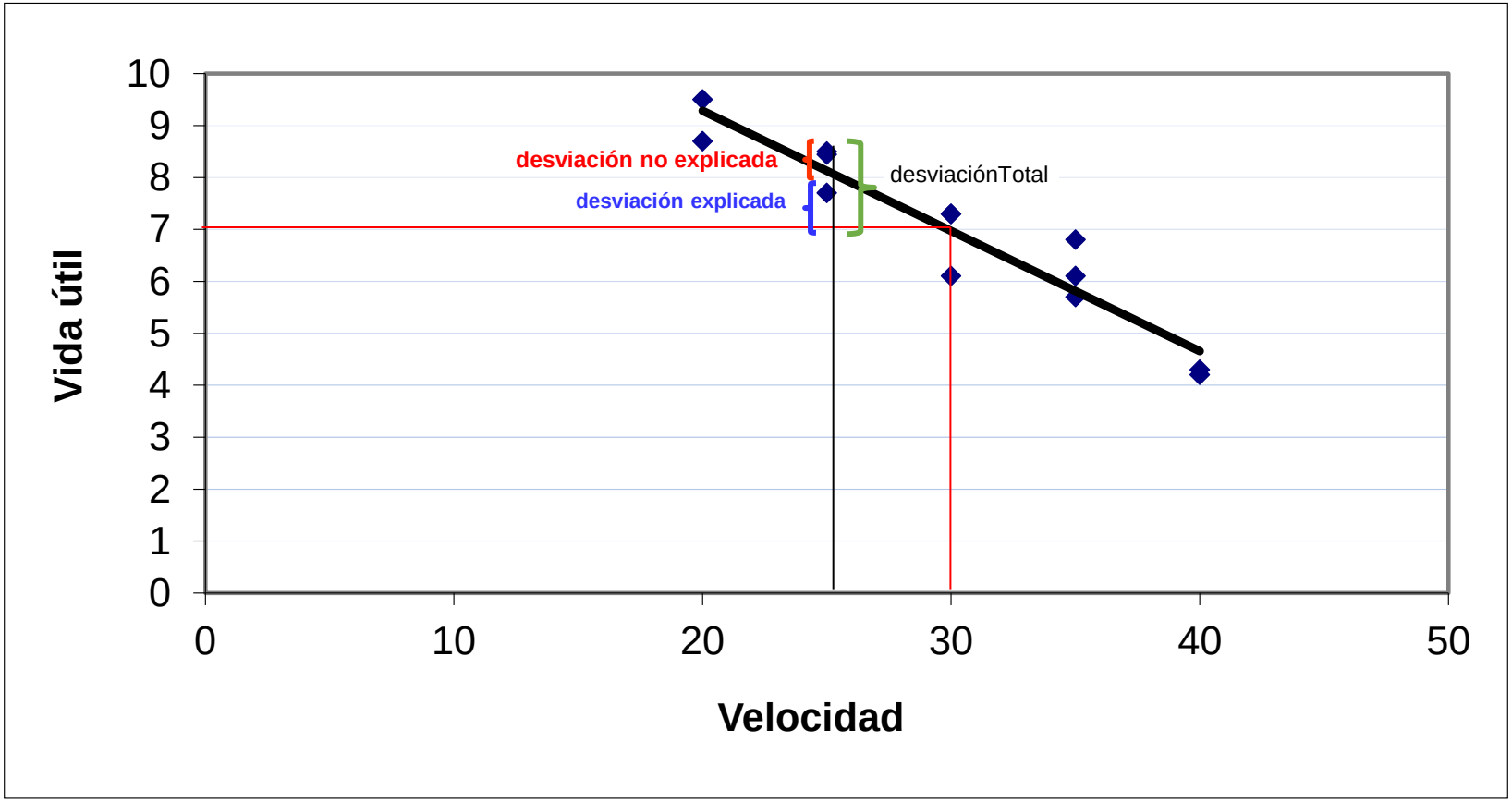
b) Plantee las hipótesis correspondientes.

c) Plantee la regla de decisión con α= 0,05

d) Concluya en términos del problema empleando los valores que se observan en la tabla.

e) Compare los resultados con el item b).

ANOVA EN REGRESIÓN



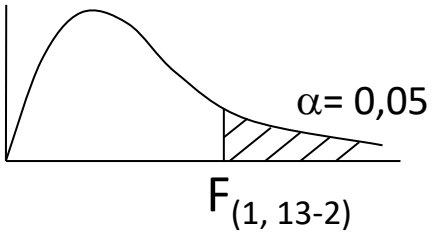
ANOVA en Regresión

$H_o: \beta = 0$
$H_1: \beta \neq 0$

$\alpha = 0{,}05$

$$F_{1;(n-2)} \cong \frac{CM_{Regresión}}{CM_{Error}}$$

ANÁLISIS DE VARIANZA					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	29,3254545	29,3254545	102,01004	4,844
Error	11	3,16223776	0,28747616		
Total	12	32,4876923			



Regla de decisión

F (1 ;11) y una probabilidad a la derecha de 0,05.

Rechazo Ho si $F_{cal} \geq 4{,}844$

No rechazo Ho si $F_{cal} < 4{,}844$

Conclusión: Con un nivel de significación del 5 % tengo evidencias suficientes para suponer que existe una relación funcional poblacional del tiempo de vida útil de la herramienta en función de la velocidad de corte, o que sea, por cada metro/minuto que se incrementa la velocidad de corte se modifica o cambia el valor medio poblacional del tiempo de vida útil de la herramienta.

Se obtiene la misma respuesta con ambos métodos.
Esto se debe, y teóricamente se demuestra que una (distribución t con n-2 grados de libertad)^2 es equivalente a una distribución F con 1 y n-2 grados de libertad.

Si toma el valor crítico de la t= 2,2010 y lo eleva al cuadrado obtiene el valor crítico de la F= 4,844401. Si ambas pruebas se realizan con un software estadístico veremos que los **p valores** (probabilidad a la derecha del punto calculado) son iguales.

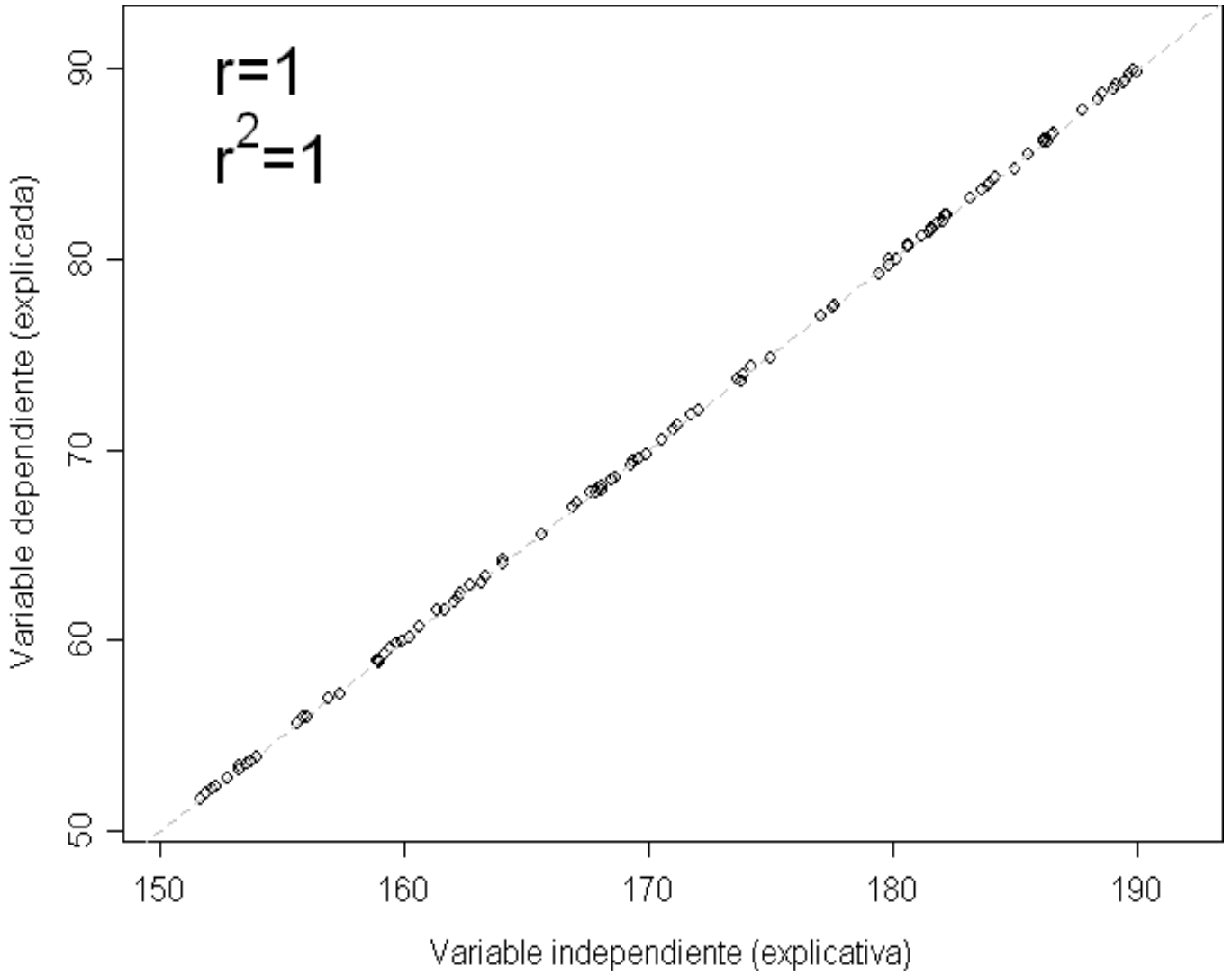
Modelo de Regresión Lineal Simple

- La ecuación de regresión lineal poblacional es:

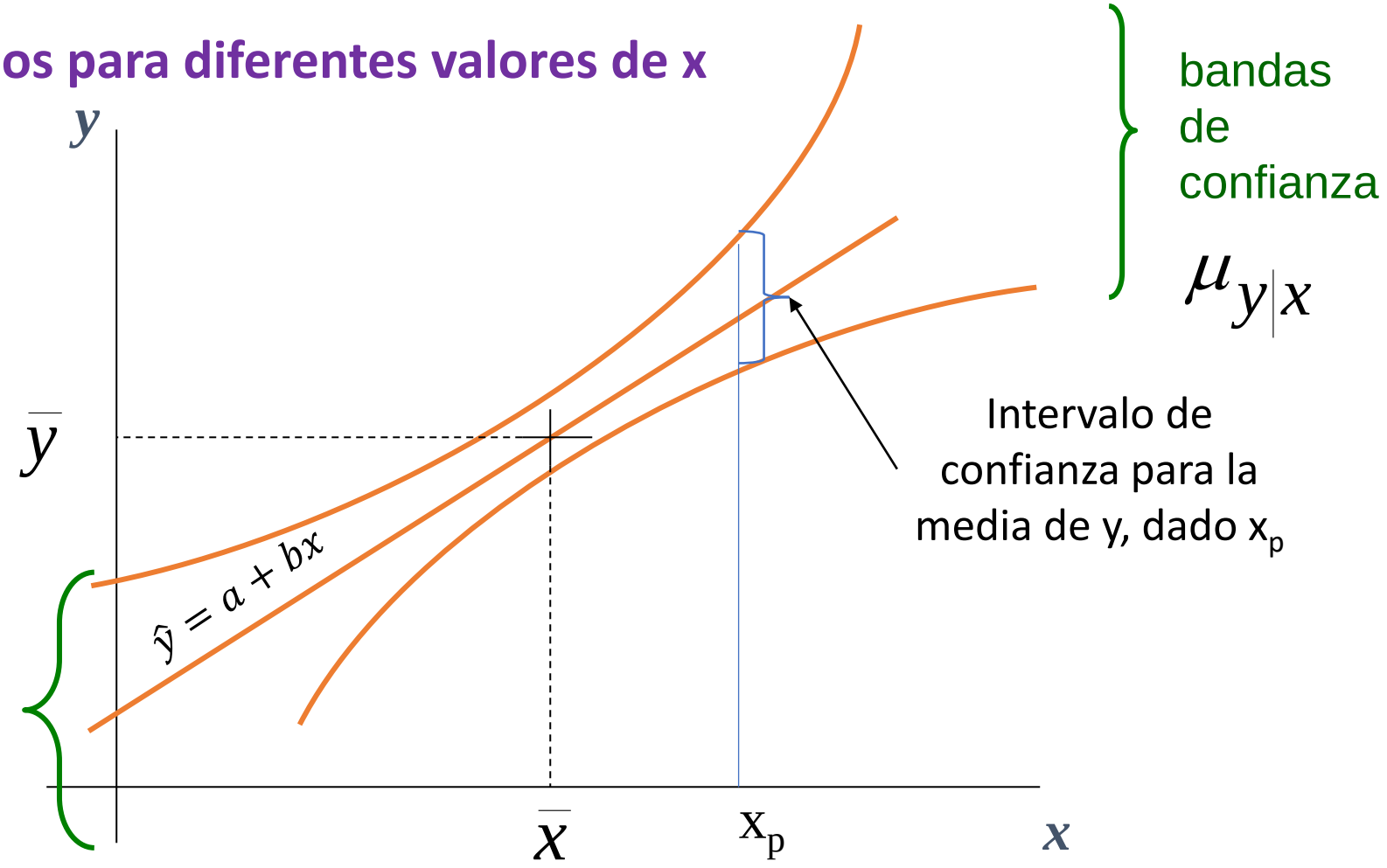
$$Y_i = \alpha + \beta x + \varepsilon_i$$

- Variable respuesta o dependiente: Y_i
- Coeficiente de intersección poblacional: α
- Coeficiente de regresión poblacional: β
- Variable predictora, regresora o independiente: X_i
- Error aleatorio no observable: ε_i

Animación:
Evolución de r y R^2, y el respectivo diagrama de dispersión



Intervalos para diferentes valores de x



A medida que los valores se alejan del centroide ($\bar{x}$, $\bar{y}$) las estimaciones de ***y*** son más imprecisas

¡Muchas Gracias por su atención!