

# REGRESIÓN

# 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

# Regresión Lineal Simple

## Conceptos:

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

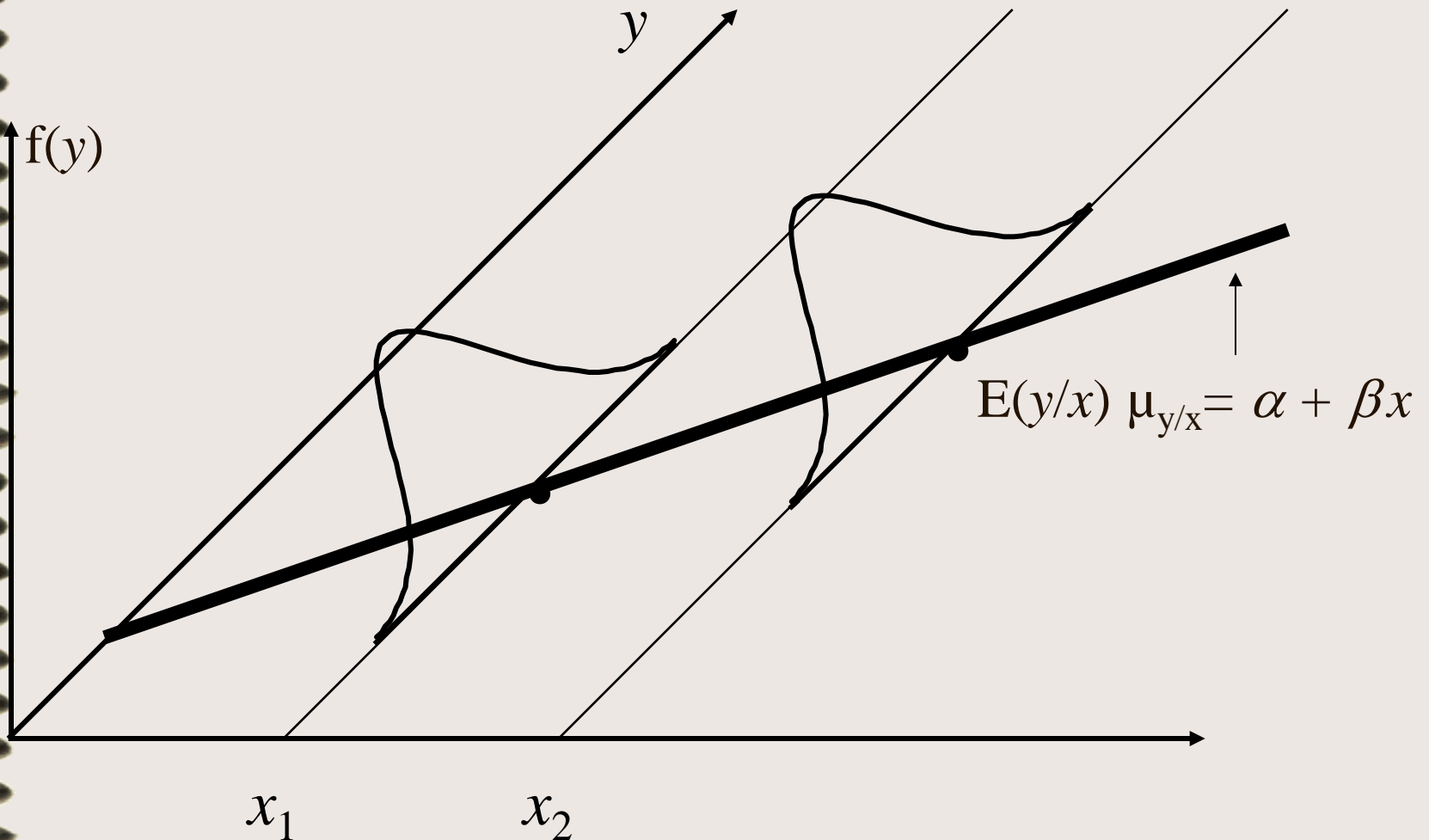
## Objetivo:

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

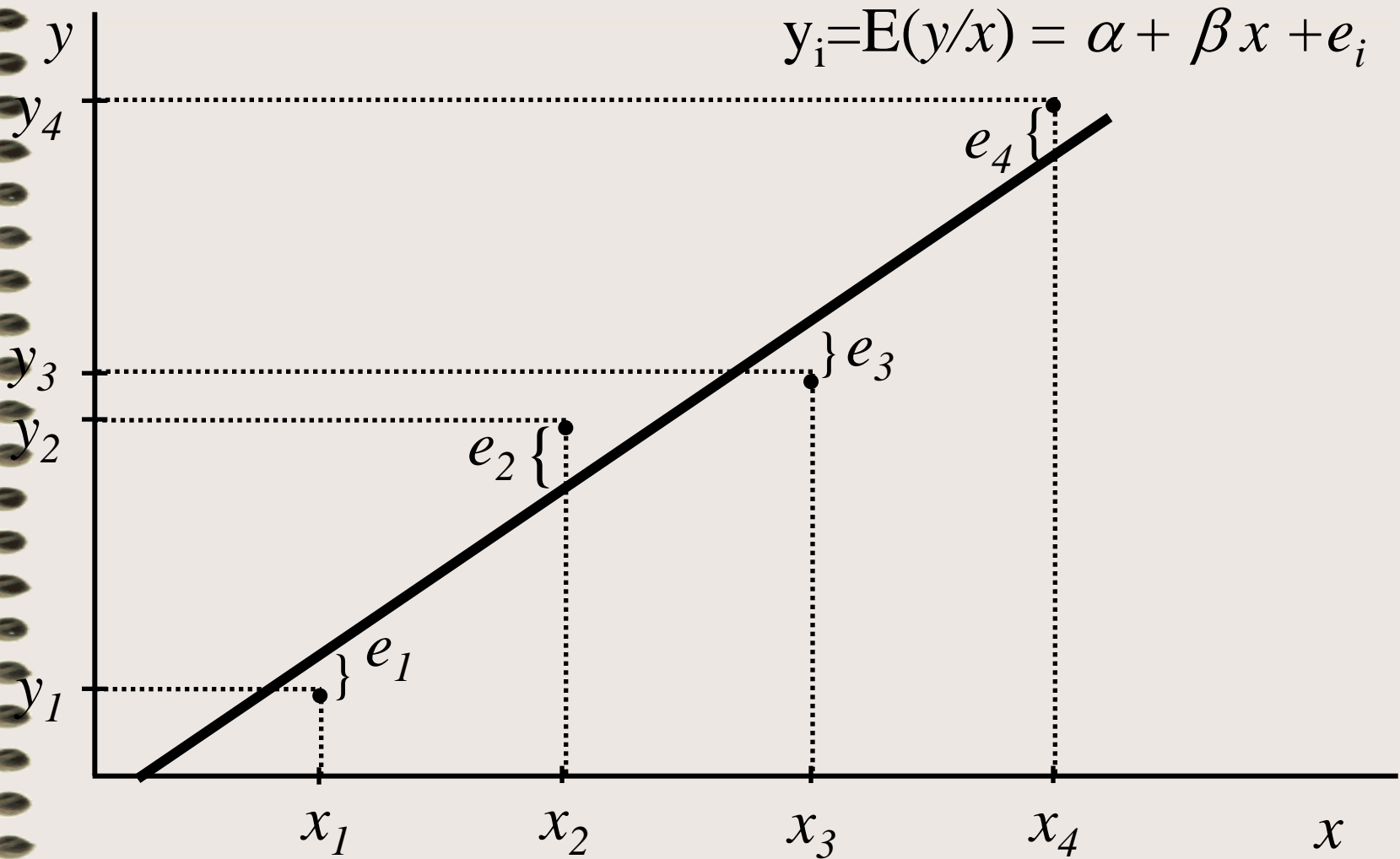
la variable X: que es la independiente (predictora, explicativa, exógena). Es la variable que provee las bases para estimar.

# Regresión Lineal Simple



Modelo teórico

# Regresión Lineal Simple



Modelo estadístico

# Regresión Lineal Simple

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

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

## Interpretación de los Coeficientes de Regresión:

$\alpha$ : 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.

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

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

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

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

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



# Estimadores

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

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

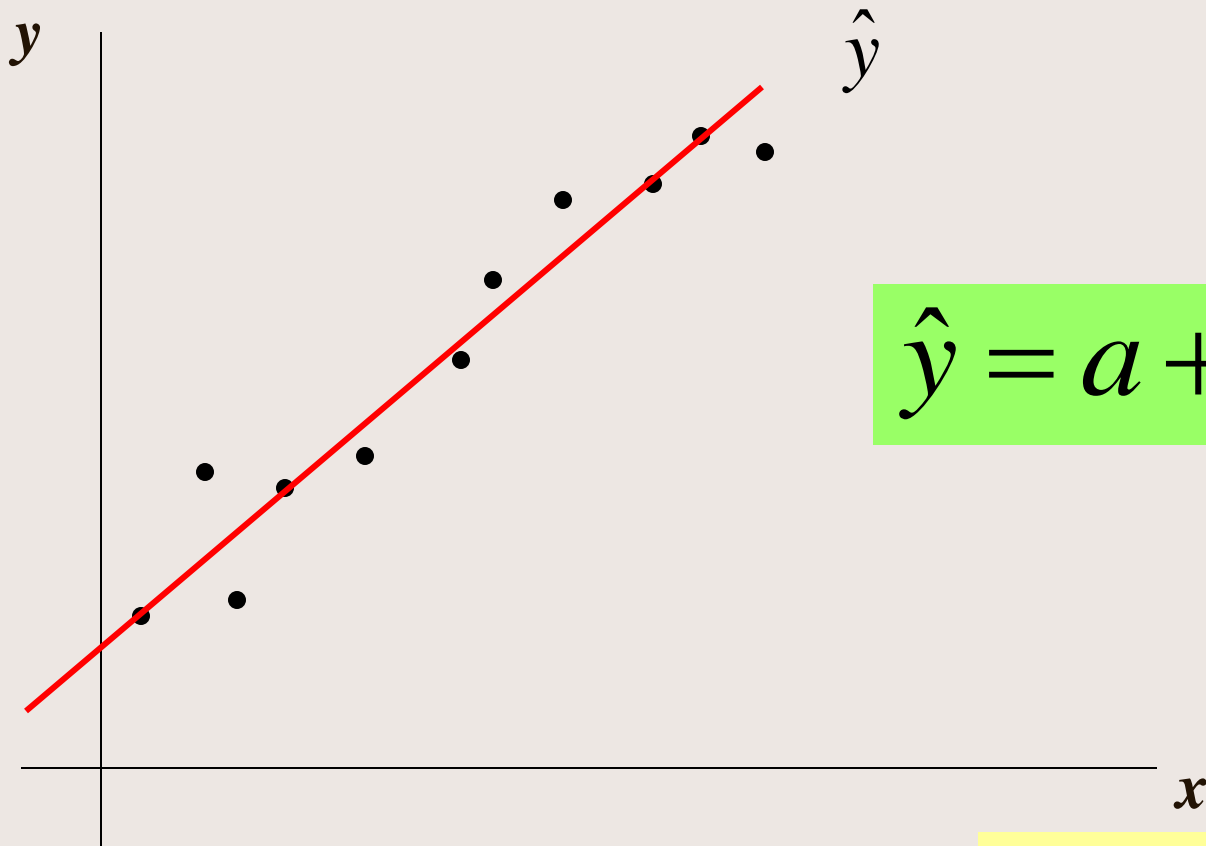
$$\hat{y} \cong N(E(\hat{y}) = \alpha + \beta x; V(\hat{y}) = S^2_e \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2} \right))$$

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

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

# 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

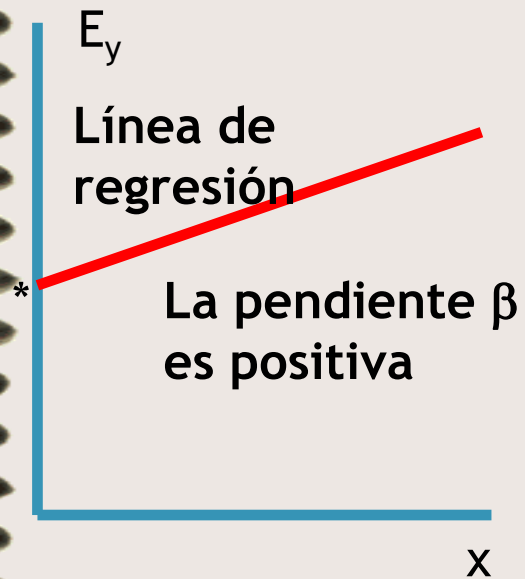
# Interpretación de los coeficientes de regresión estimados

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

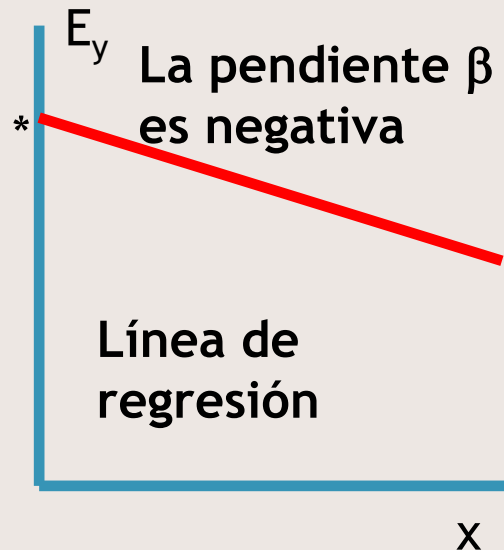
**La ordenada al origen “a”** indica el valor promedio estimado de la variable respuesta cuando la variable predictora vale 0. Sin embargo carece de interpretación práctica si es irrazonable considerar que el rango de valores de  $x$  incluye a cero.

# Líneas posibles de regresión en la regresión lineal simple

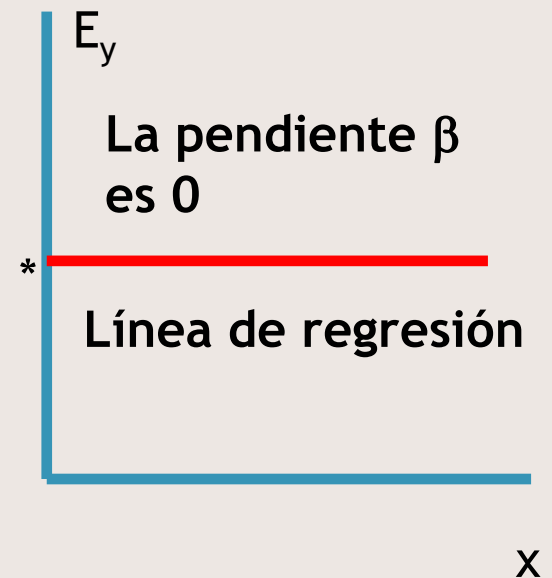
Relación lineal positiva



Relación lineal negativa



No hay relación

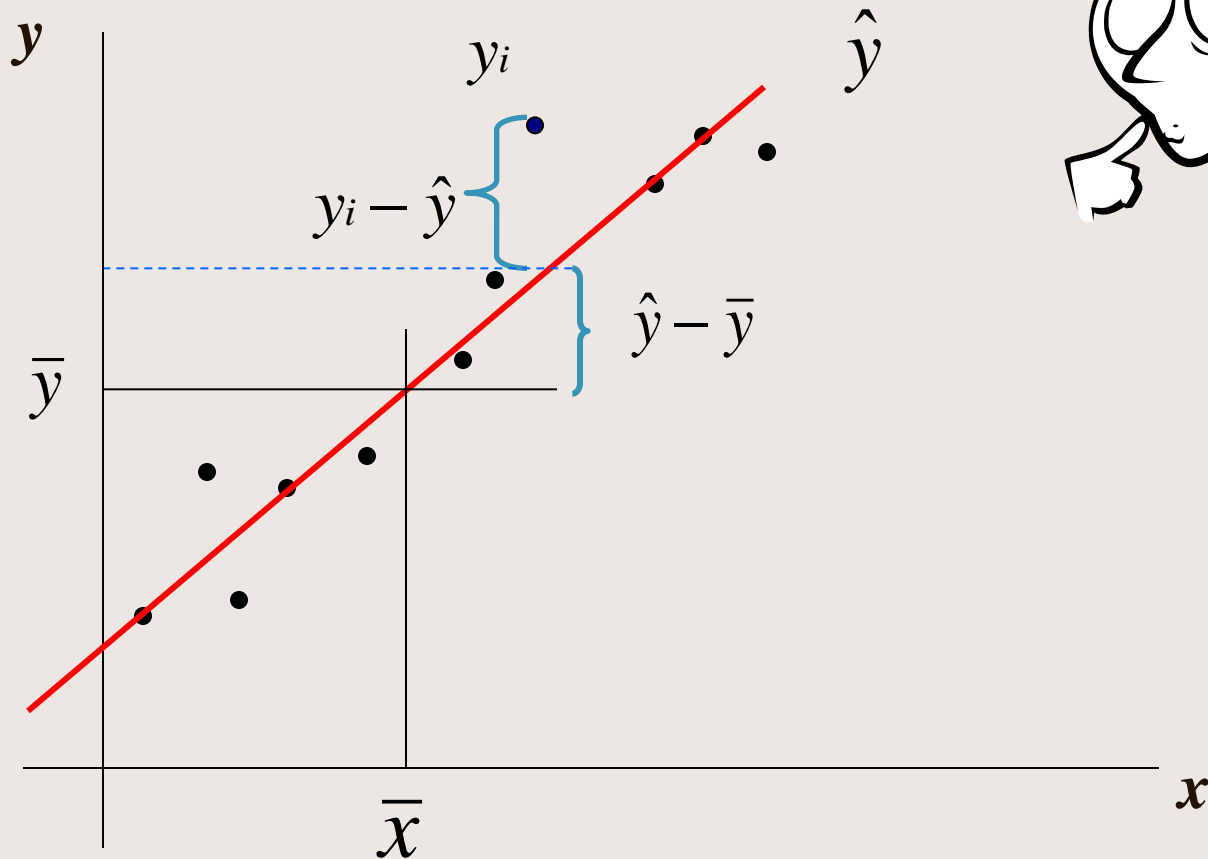


\* Ordenada al origen  $\alpha$

# REGRESION LINEAL SIMPLE

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

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



# Análisis de Variancia en el análisis de regresión

- ✘ El enfoque desde el análisis de variancia 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  $SC_{\text{tot}}$ , es la suma de las desviaciones al cuadrado

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

# Desarrollo formal de la partición

Consideremos la desviación

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

Podemos descomponerla en

$$\underbrace{(Y_i - \bar{Y})}_T = \underbrace{(\hat{Y}_i - \bar{Y})}_R + \underbrace{(Y_i - \hat{Y}_i)}_E$$

(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

# Desarrollo formal de la partición

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

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

( $SC_{\text{tot}}$ ): Suma de cuadrados total

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

( $SC_{\text{er}}$ ): 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 un de estos cuadrados medios tiene una distribución Ji Cuadrado.



## Estimación de la variancia de los términos del error ( $\sigma^2$ )

Dado que los  $Y_i$  provienen de diferentes distribuciones de probabilidades con medias diferentes que dependen del nivel de  $X$ , la desviación de una observación  $Y_i$  debe ser calculada con respecto a su propia media estimada  $\hat{Y}_i$ .

Por tanto, las desviaciones son los residuales

$$Y_i - \hat{Y}_i = e_i$$

Y la suma de cuadrados es:

$$SC_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2 = \sum_{i=1}^n e_i^2$$

## Estimación de la variancia de los términos del error ( $\sigma^2$ )

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

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

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

## Tabla del análisis de varianza

Fuentes de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
Debido a la regresión	1	SCR	CMR=SCR/1	CMR/CMEE
Debido al Error	n-2	SCEE	CMEE=SCEE/(n-2)	
Total	n-1	SCTot		

La hipótesis nula  $H_0: \beta = 0$  se rechaza si el “p-valor” de la prueba de F es menor que el nivel de significación.

## Error estándar de la estimación

Se o  $S_{y/x}$

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

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

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

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 torneear, etc.



Brocas  
helicoidales

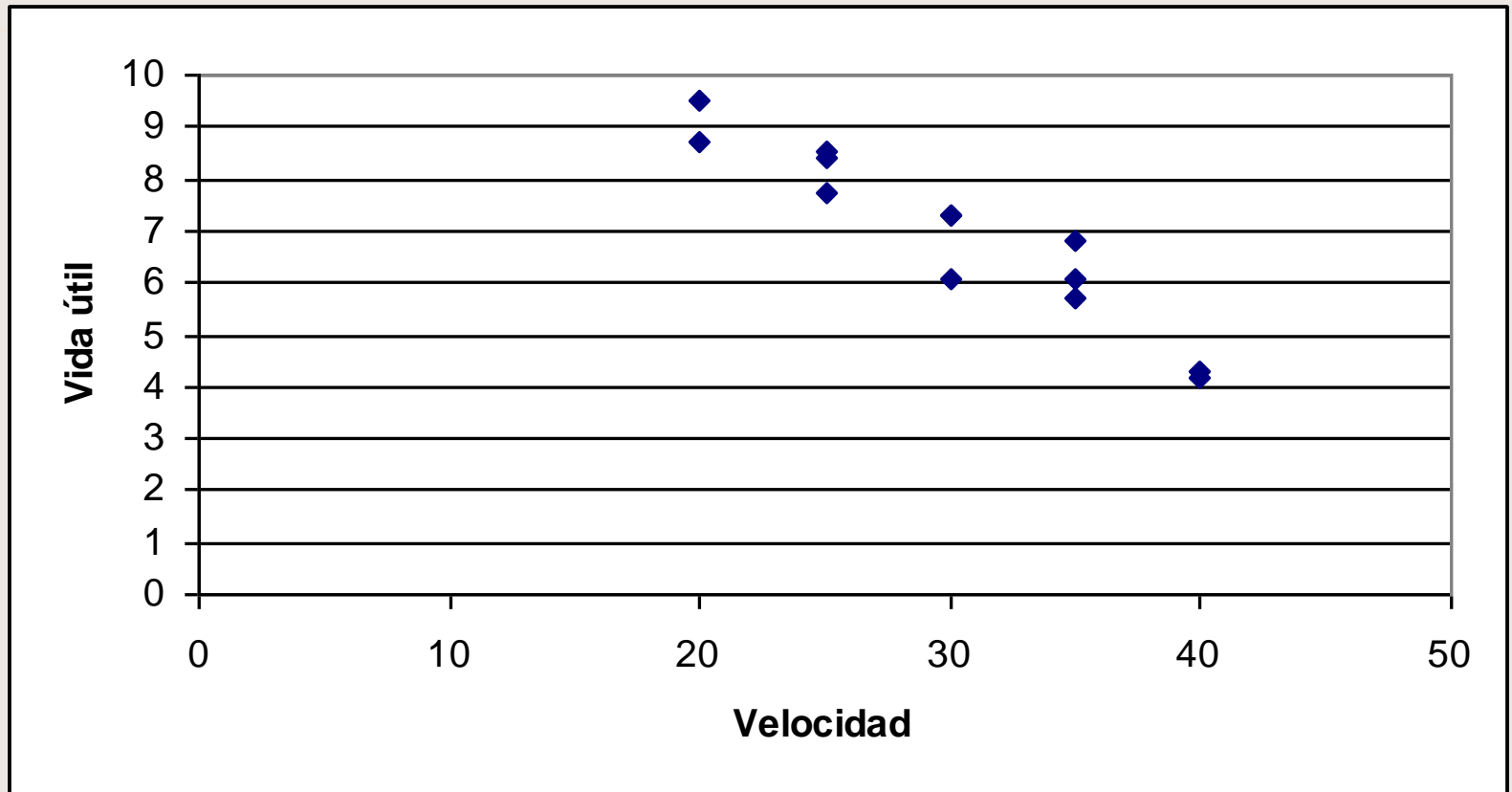


Teniendo en cuenta que la vida útil se ve afectada por varios aspectos como: el ambiente operacional, las condiciones de producción o de mantenimiento y el desgaste presentado por su uso, decide comenzar a investigar la relación funcional entre la velocidad de corte (metros por minuto) y el tiempo de 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:

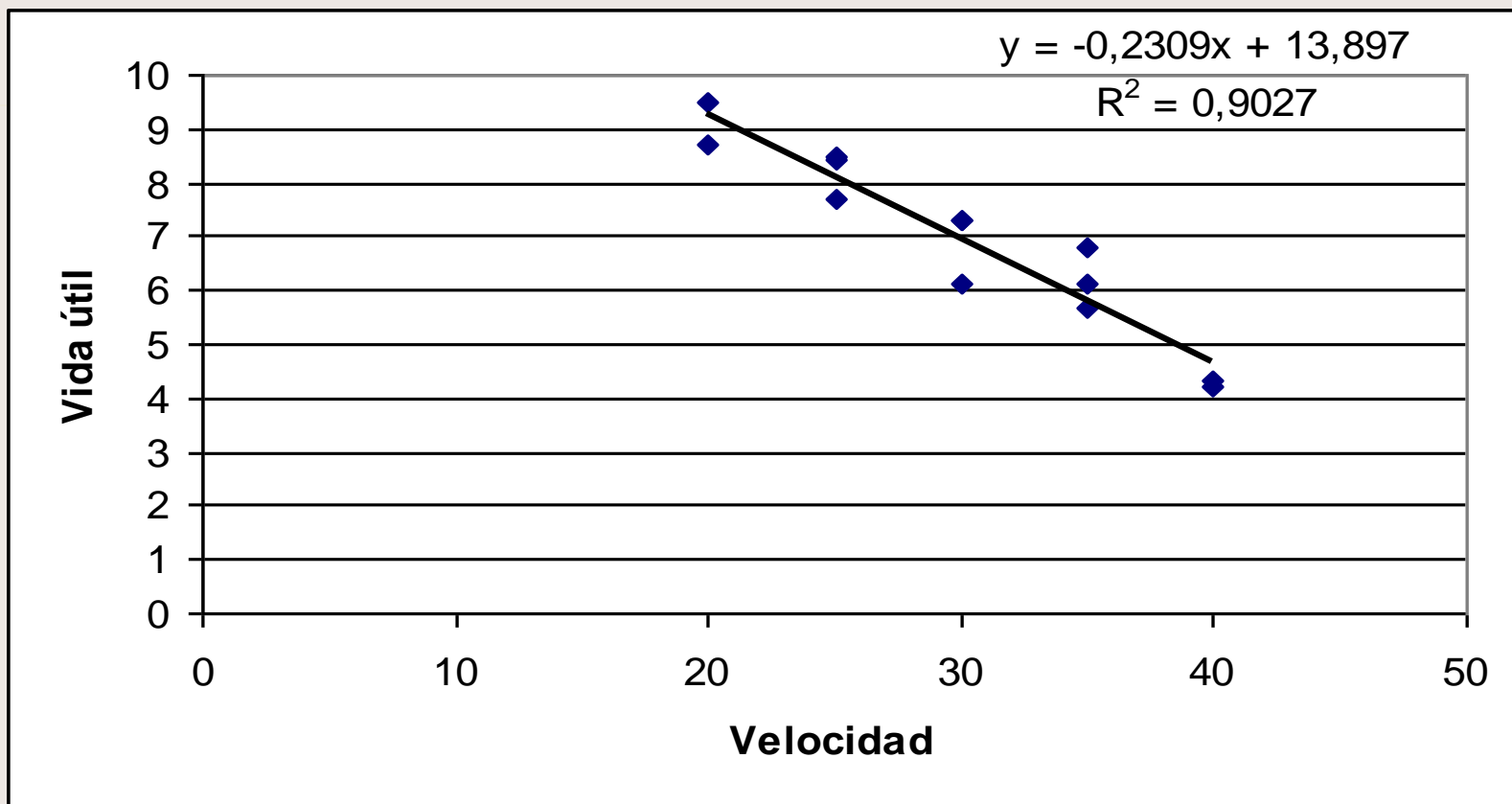
<b>Velocidad (Metros por minuto)</b>	<b>Vida (Horas)</b>
20	8,7
20	9,5
25	8,5
25	7,7
25	8,4
30	7,3
30	6,1
30	7,3
35	6,8
35	5,7
35	6,1
40	4,3
40	4,2

$$\begin{aligned}
 \Sigma x &= 390 & \Sigma x^2 &= 12250 & \Sigma y &= 90,6 & \Sigma y^2 &= 663,9 \\
 & & \Sigma xy &= 2591 & & & &
 \end{aligned}$$

a) Dibujar el diagrama de dispersión.







## Prueba de hipótesis para el coeficiente de regresión $\beta$

$$H_0: \beta = 0 \quad \text{vs} \quad H_1: \beta \neq 0$$

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

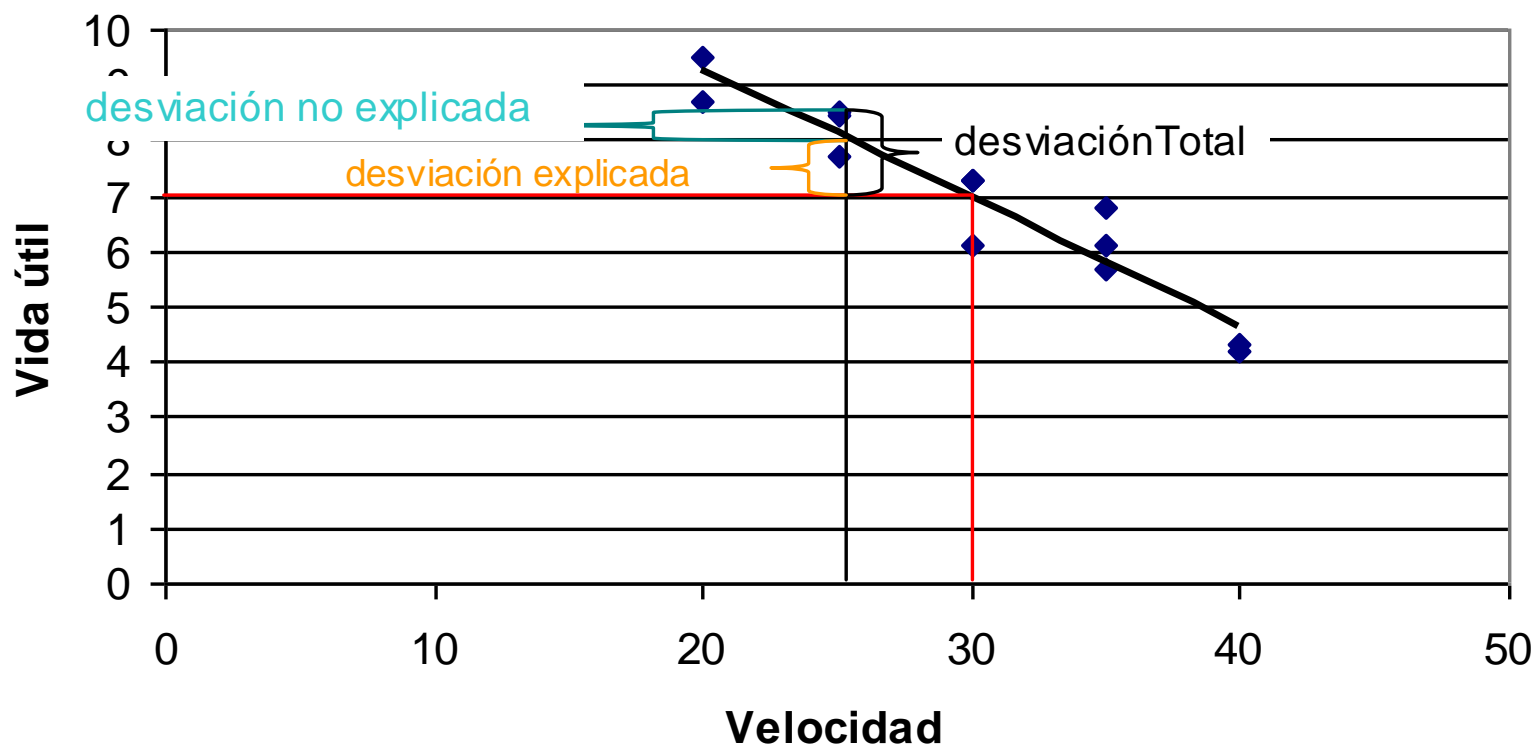
**Conclusión:** Con un nivel de significación del 5 % tengo evidencias suficientes para suponer que existe una relación funcional poblacional del tiempo de vida útil de la herramienta en función de la velocidad de corte, o que sea, por cada metro/minuto que se incrementa la velocidad de corte se modifica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

## Intervalo de confianza para el coeficiente de Regresión

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

$$(-0,281214 < \beta < -0,180585)$$

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.



# ANOVA en Regresión

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

## ANÁLISIS DE VARIANZA

	<i>Grados de libertad</i>	<i>Suma de cuadrados</i>	<i>Promedio de los cuadrados</i>	<i>F</i>	<i>Valor crítico de F</i>
Regresión	1	29,3254545	29,3254545	102,01004	6,6927E-07
Residuos	11	3,16223776	0,28747616		
Total	12	32,4876923			