ANÁLISIS ESTADÍSTICO

REGRESIÓN LINEAL SIMPLE

Jorge Fallas
jfallas56@gmail.com

2010

Temario

- Introducción: correlación y regresión
- Supuestos del análisis
- Variación total de Y y variación explicada por el modelo
- Modelos de regresión
- Estimación de la ecuación de una regresión lineal simple
  - Parámetros: intercepto, pendiente, error de estimación, IC
- Medidas de variación en Regresión
  - Suma de cuadrados totales, suma cuadrados regresión, error
- Evaluación supuestos del análisis de regresión
- Estimación de valores esperados (predicción)
- Uso de XLSTats

Regresión Lineal: supuestos

1. Normalidad

   Los valores de Y están normalmente distribuidos para cada valor de *X*

   La distribución de probabilidad del error es normal $E(e_i^2)=\sigma^2$

2. Homocedasticidad (varianza constante)
3. Errores independientes **$E(e_i e_j)=0$ (i<>j)**
4. Linealidad $Y_i = \alpha + \beta X_i$
5. Variables se miden sin error (No estocásticas)

Regresión Lineal: supuestos

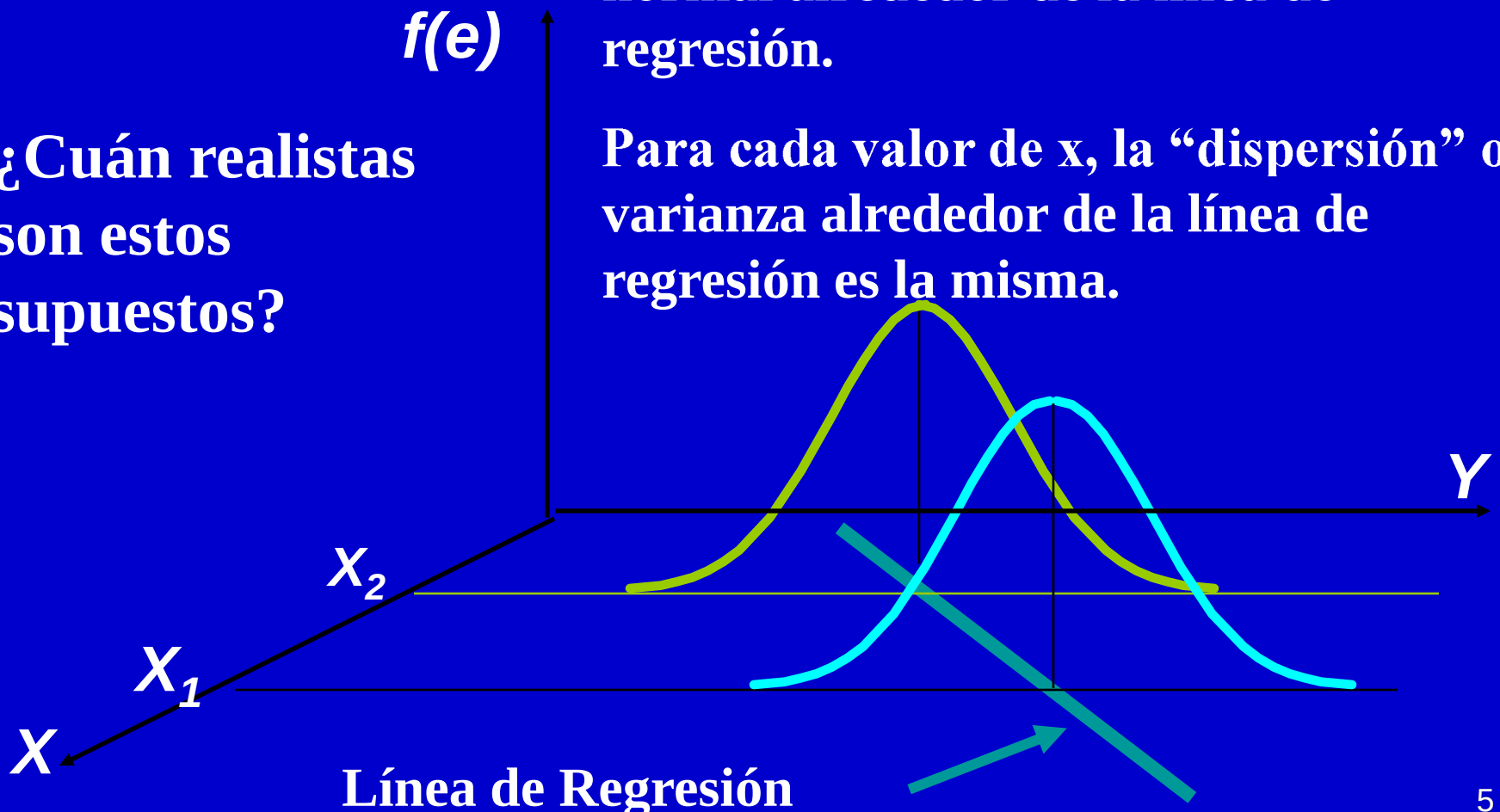
- Dado que los supuestos sean verdaderos....
- Las fórmulas utilizadas para estimar los coeficientes de regresión son **BLUE** (Best Linear Unbiased Estimators)
- Best = “El estimador más eficiente” = varianza más pequeña
- Linear: La media poblacional de Y es una función lineal de X
- Unbiased (insesgado)= Valor esperado del estimador = al parámetro poblacional

Variación de los errores alrededor de la línea de regresión

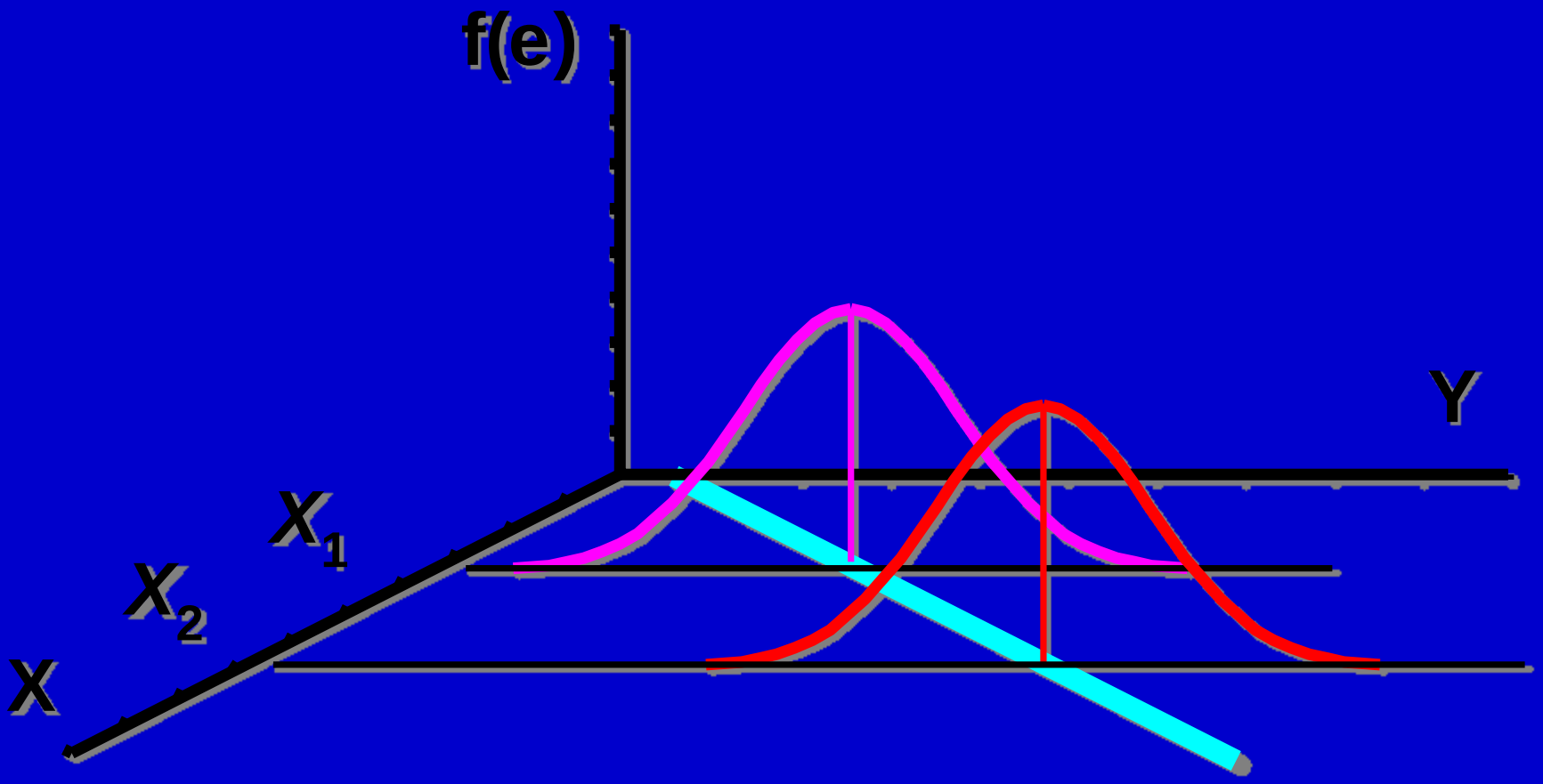
La distribución de valores de Y es normal alrededor de la línea de regresión.

Para cada valor de x, la "dispersión" o varianza alrededor de la línea de regresión es la misma.

¿Cuán realistas son estos supuestos?



Supuestos de Normalidad y Varianza Constante



Modelos de regresión

- Responden a la pregunta ¿Cuál modelo describe la relación entre las variables?
- Ecuaciones utilizadas
  - 1 variable cuantitativa predictora (X)
  - 1 variable dependiente (respuesta) (Y)

    Variable a predecir: Y
  - 1 o más variables predictoras numéricas o cualitativas (explicativas): X

    Variable a predecir: Y

¿Cuál modelo elegir?

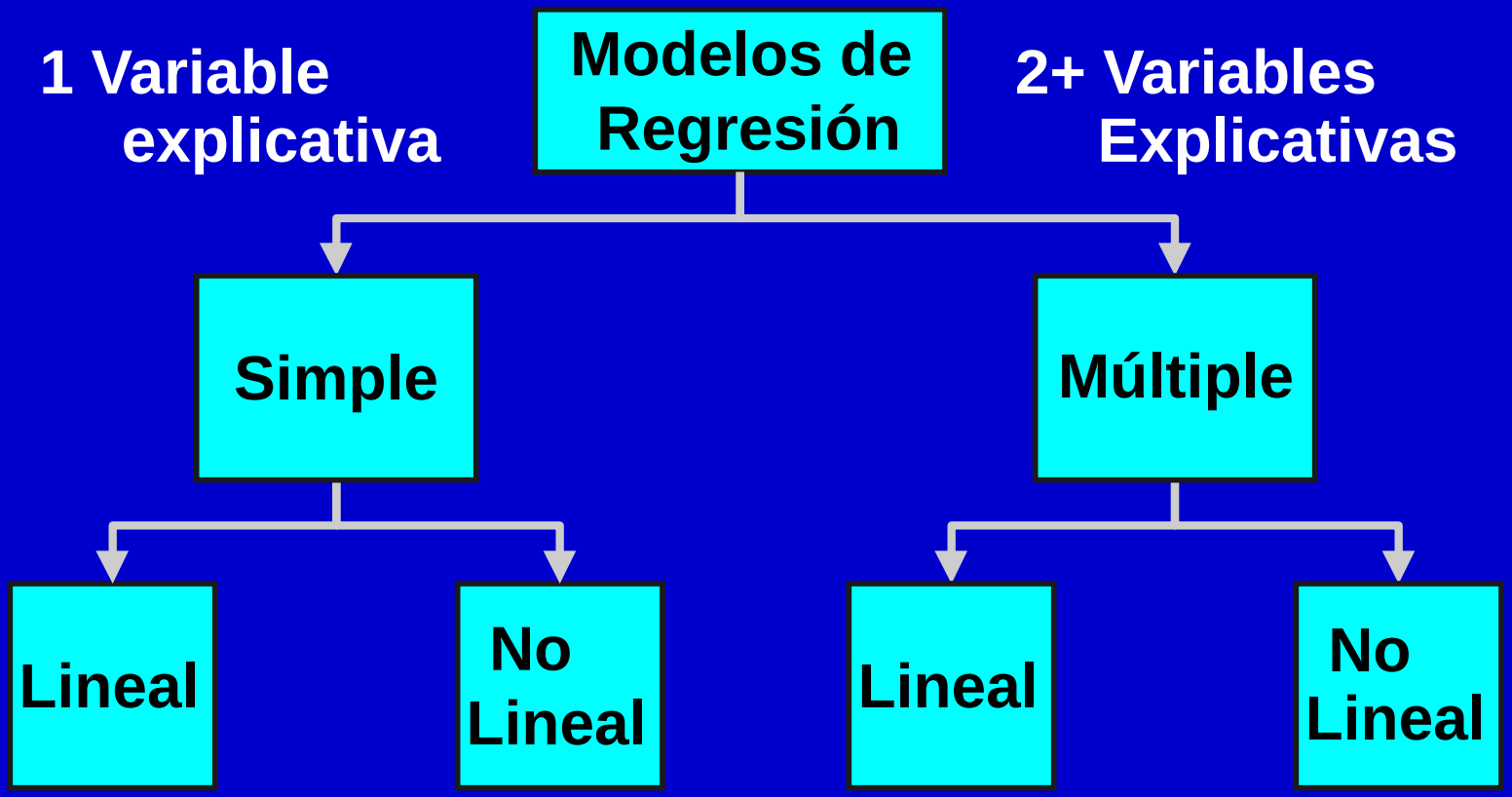
- Se basa en conocimiento teórico de los datos a analizar :
  - Económicos, psicológicos, biológicos
- Teoría matemática
- Investigación previa
- “Sentido común”

- Asumimos que la relación va de X a Y

¿Qué le parece más lógico?

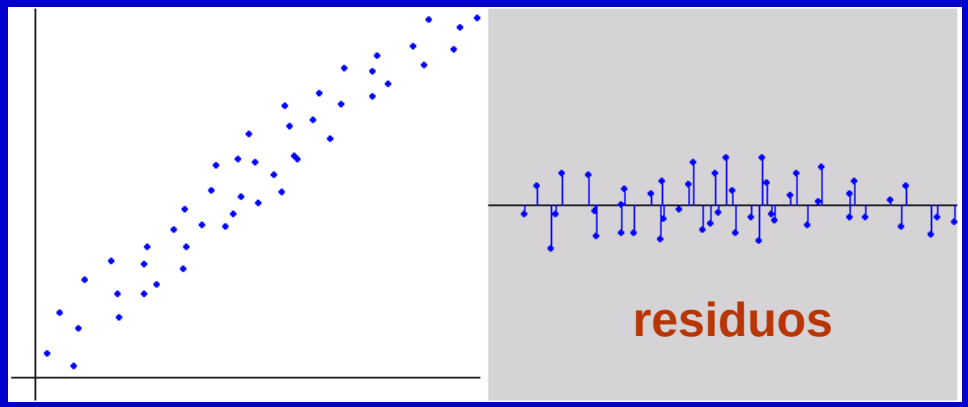
¿Cuál modelo elije usted? ¿Porqué?

Modelos de Regresión

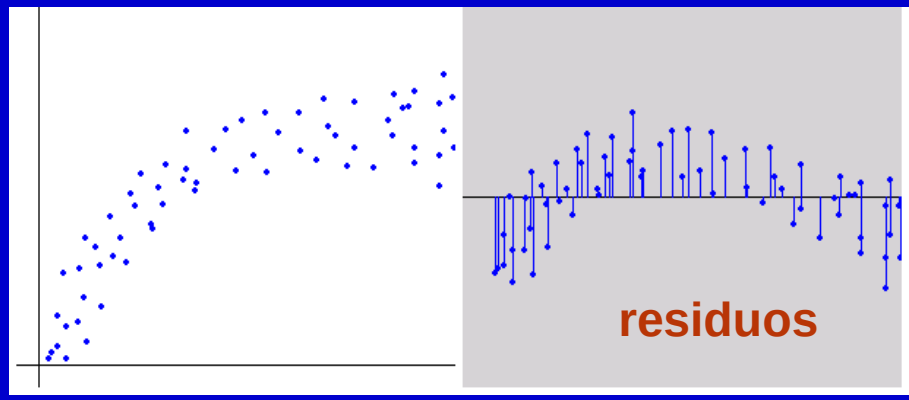


Ejemplos de Modelos de Regresión

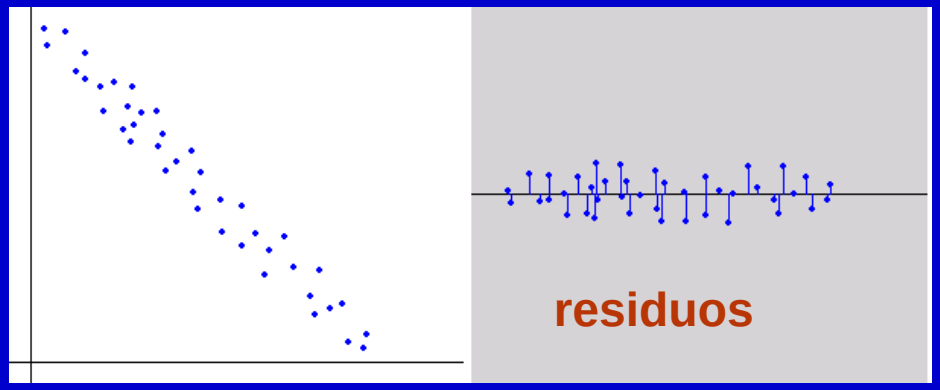
Relación lineal Positiva



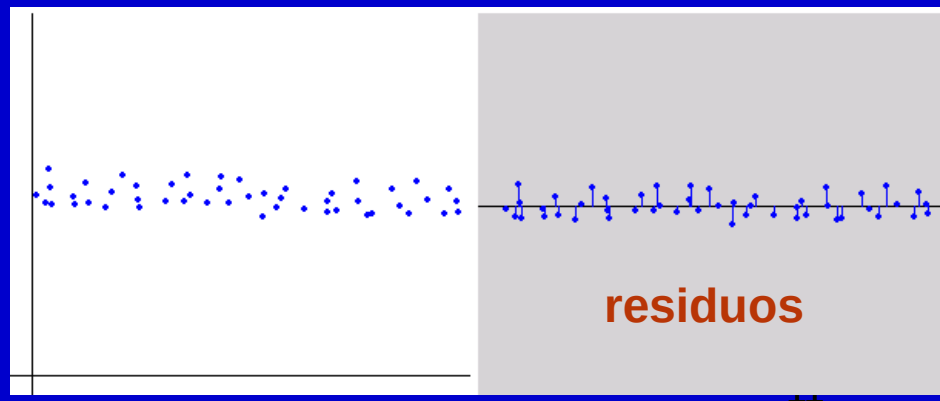
Relación NO lineal



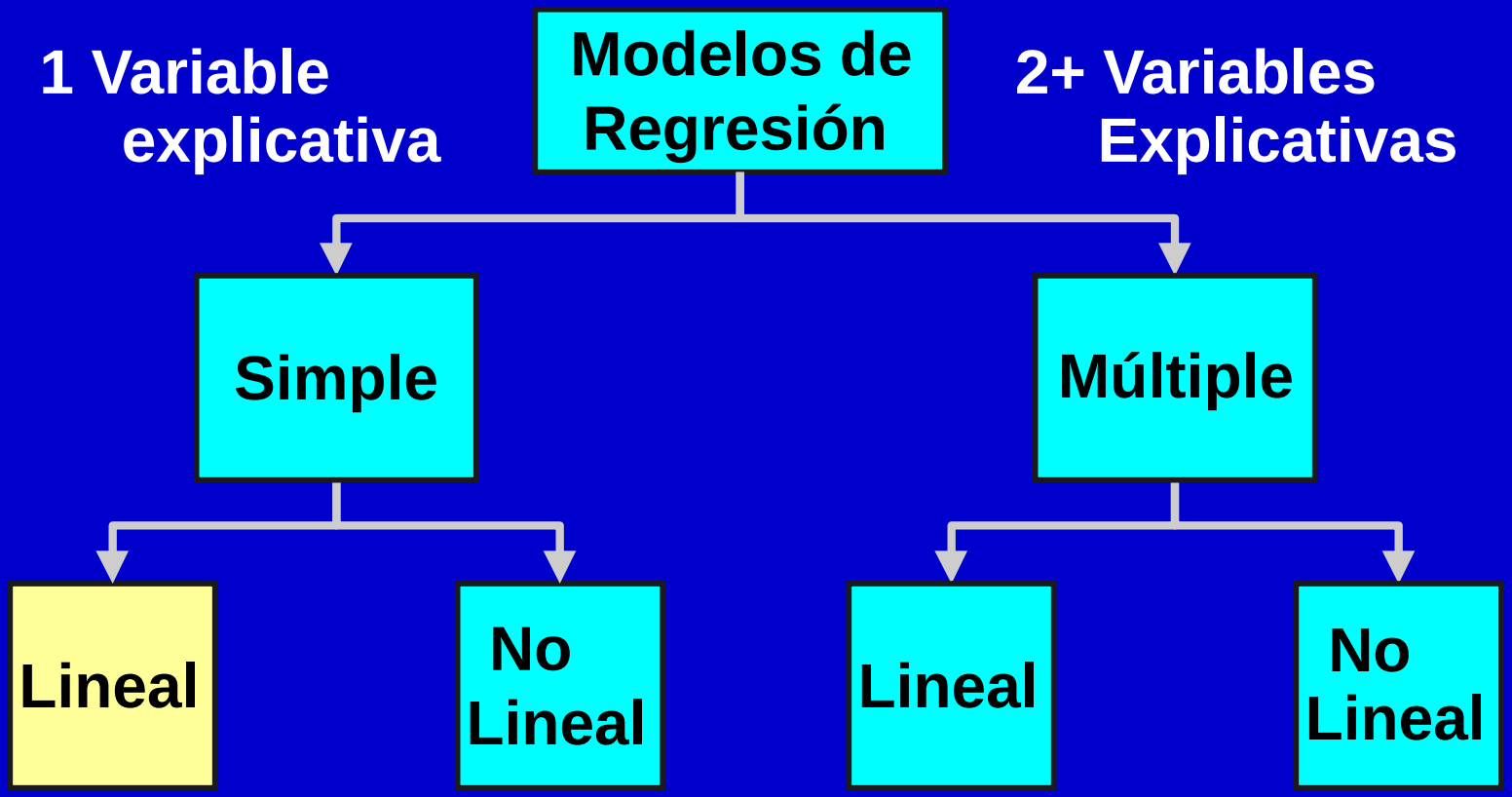
Relación lineal Negativa



Ausencia de Relación



Modelos de Regresión



El Modelo de Regresión Lineal simple

Ecuación Lineal

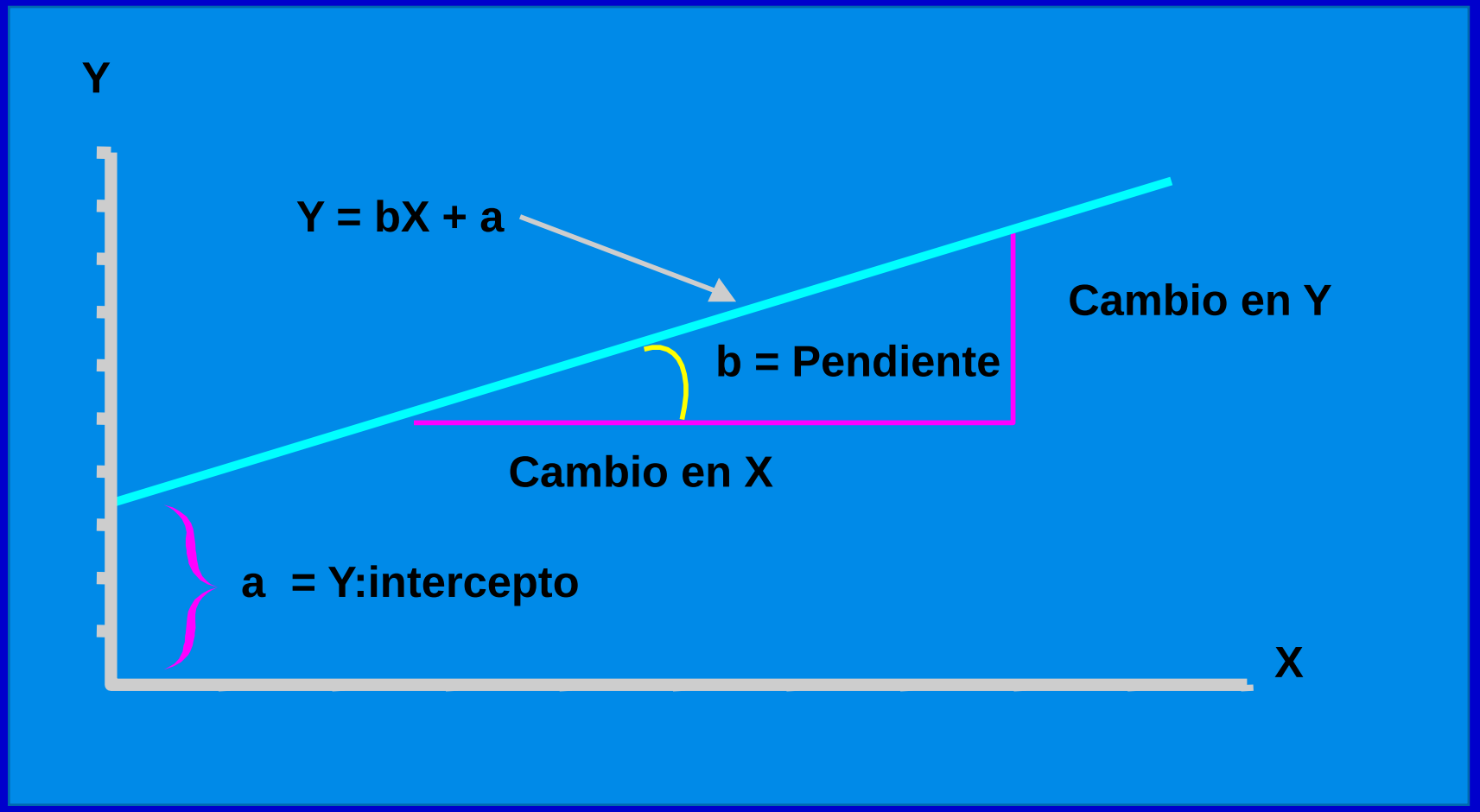
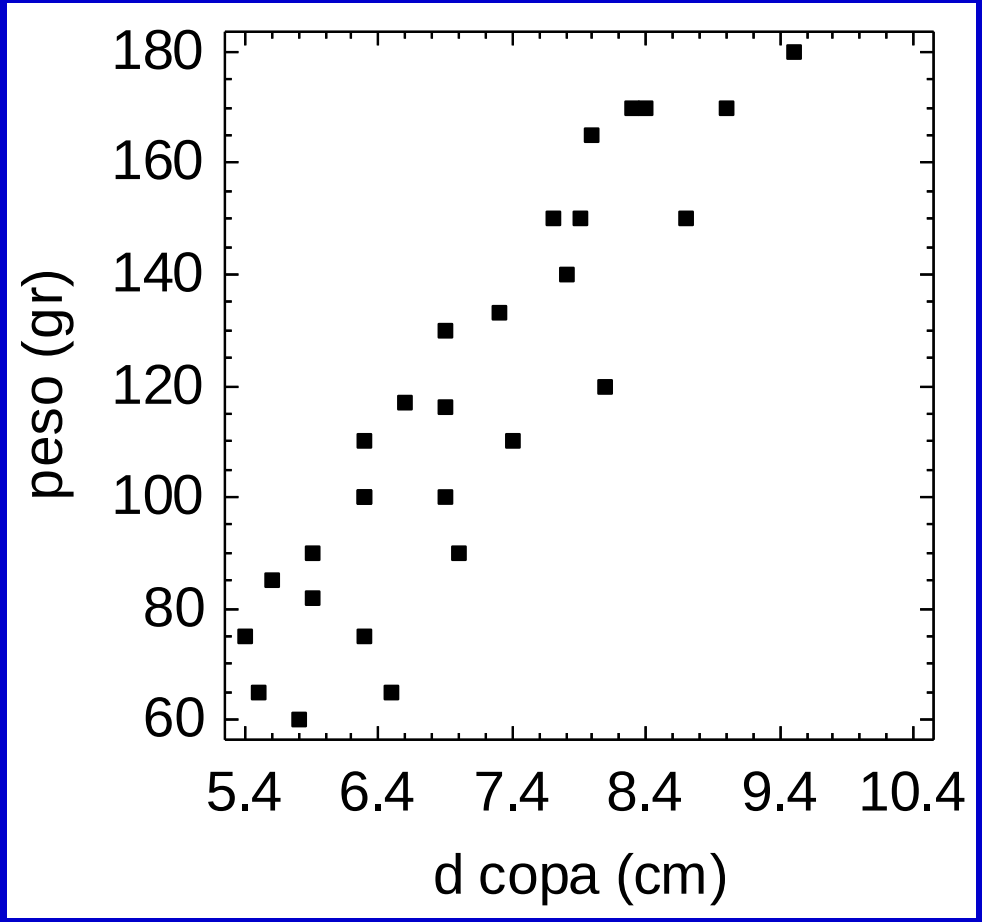


Diagrama de Dispersión

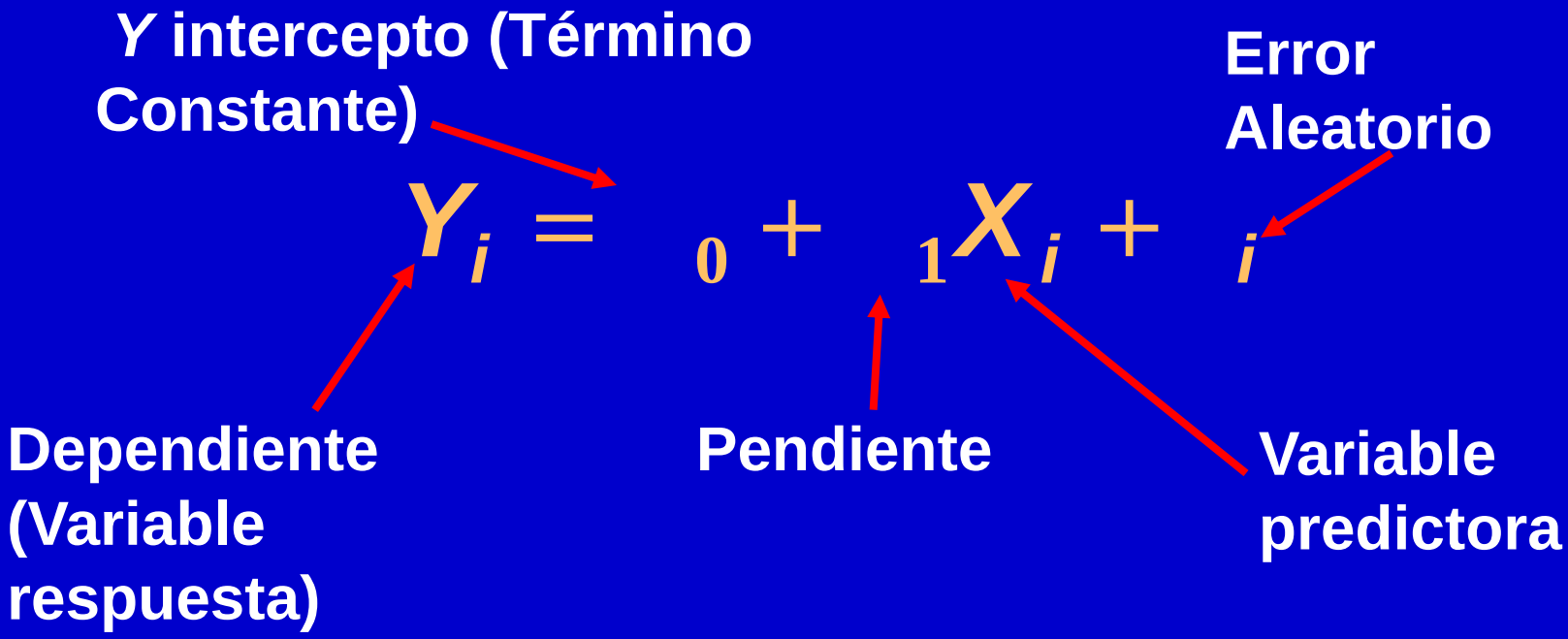
Graficar pares (X_i, Y_i)



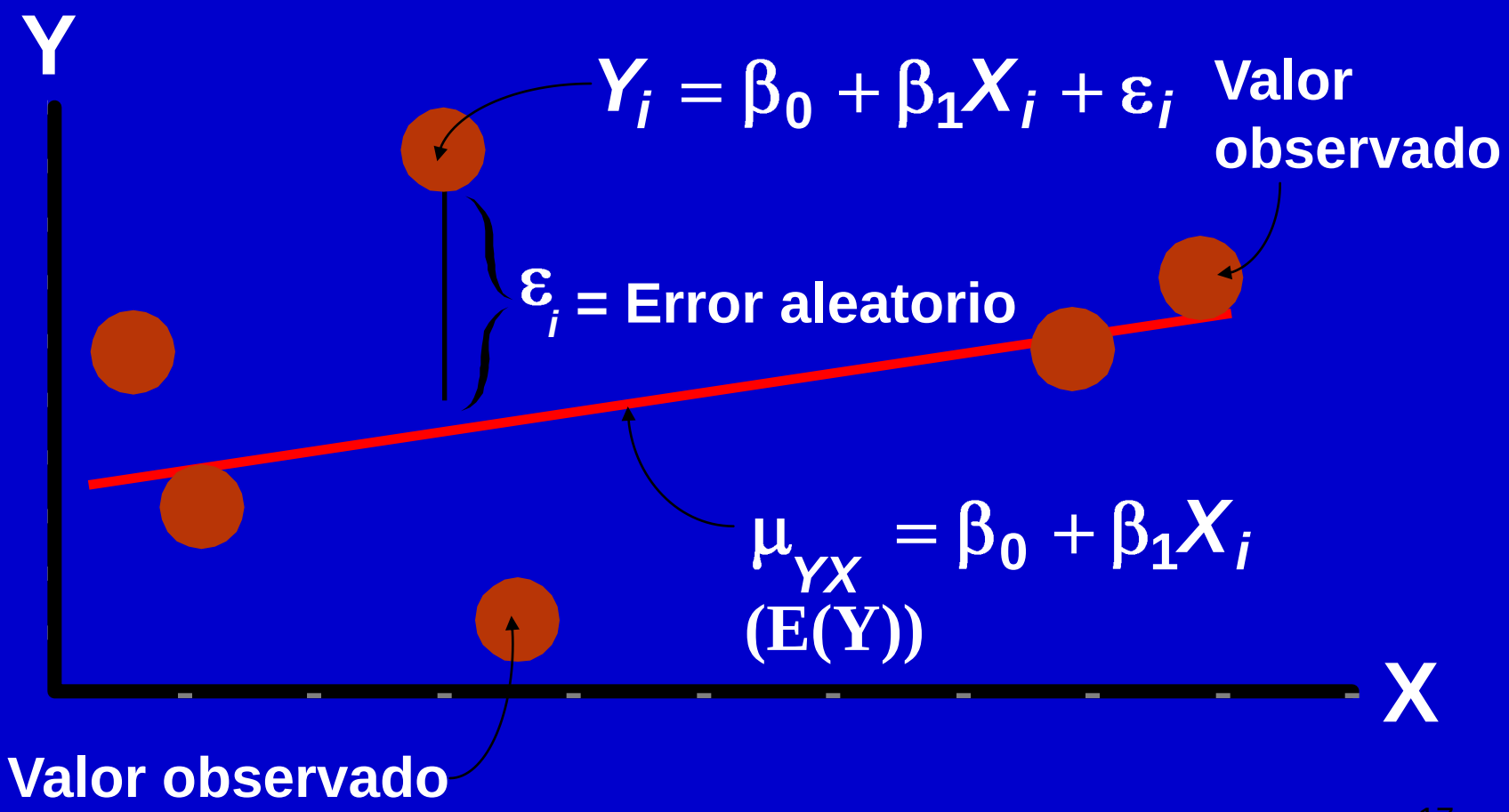
Modelo de regresión lineal simple

- **Relación entre variables es una función lineal**

- **La recta que mejor ajusta a los datos es:**



Modelo de Regresión Lineal de la Población



Modelo de Regresión lineal Simple: Estimación

$$\hat{Y}_i = b_0 + b_1 X_i$$

$\hat{Y}_i$ = Valor estimado de Y para observación i

X_i = Valor de X para observación i

b_0 = intercepto eje Y, estimador del parámetro β_0

b_1 = *Pendiente*: estimador del parámetro β_1

Estimación de parámetros: Método de Mínimos cuadrados

¿Cómo hacerlo?

¿Cómo dibujaría usted una línea a través de la nube de puntos?

¿Cómo selecciona usted la línea que mejor ajusta a los datos?

Aplicaciones JAVA de Regresión

http://www.stat.sc.edu/~west/javahtml/Regression.html

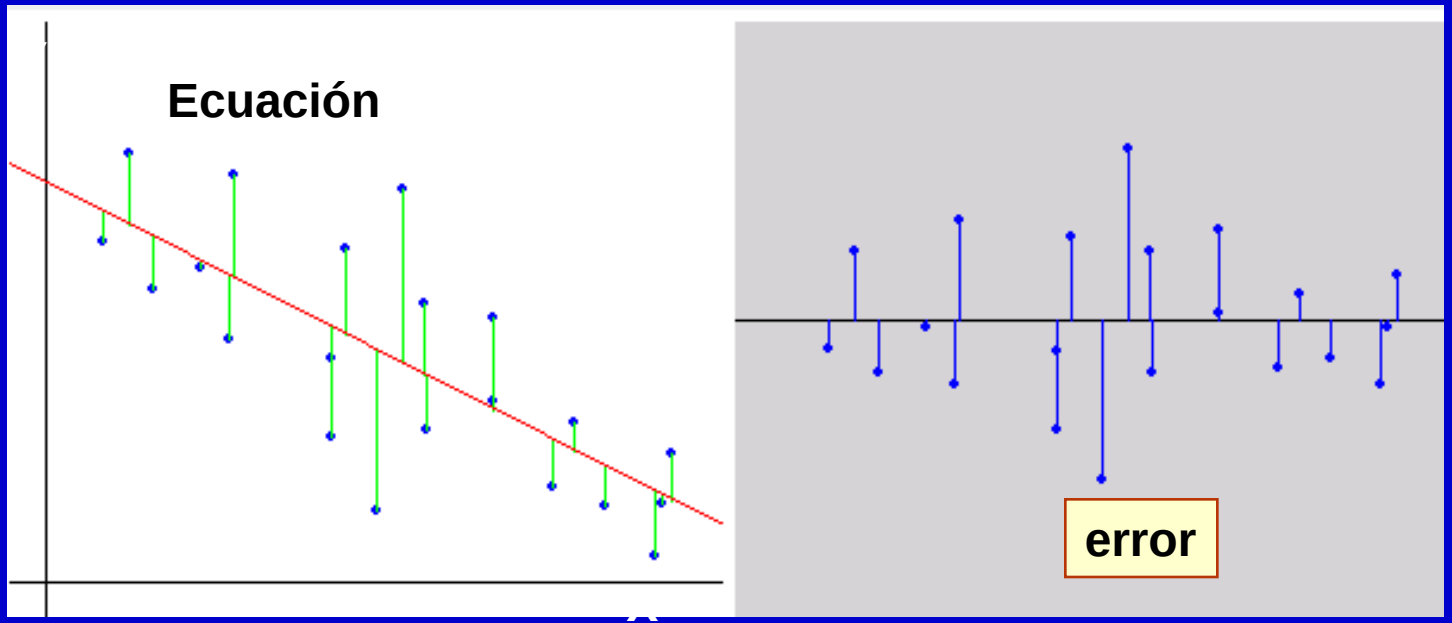
Ver video: **Observe el efecto de un punto en la recta**

http://www.math.csusb.edu/faculty/stanton/m262/regress/

Ver video 1 : **Observe el efecto de los datos en la recta**

Mínimos cuadrados

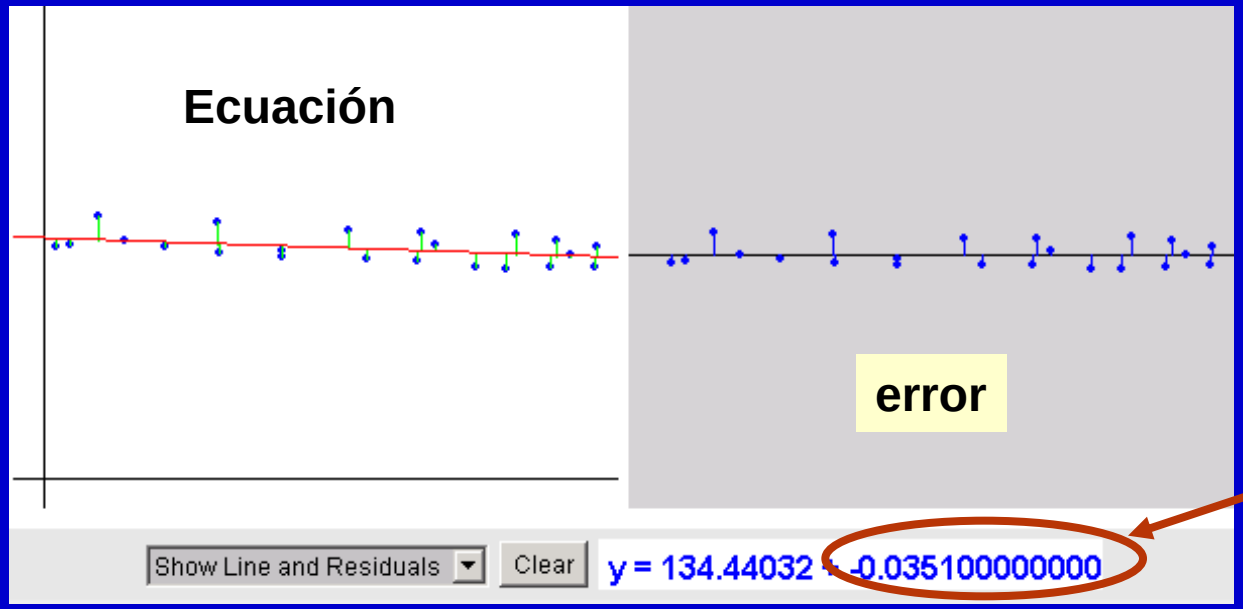
- ‘Mejor ajuste”: La diferencia entre el valor actual y el estimado (error) es un mínimo
  - *Sin embargo* diferencias positivas anulan diferencias negativas



Ausencia de correlación: ¿Qué debemos esperar?

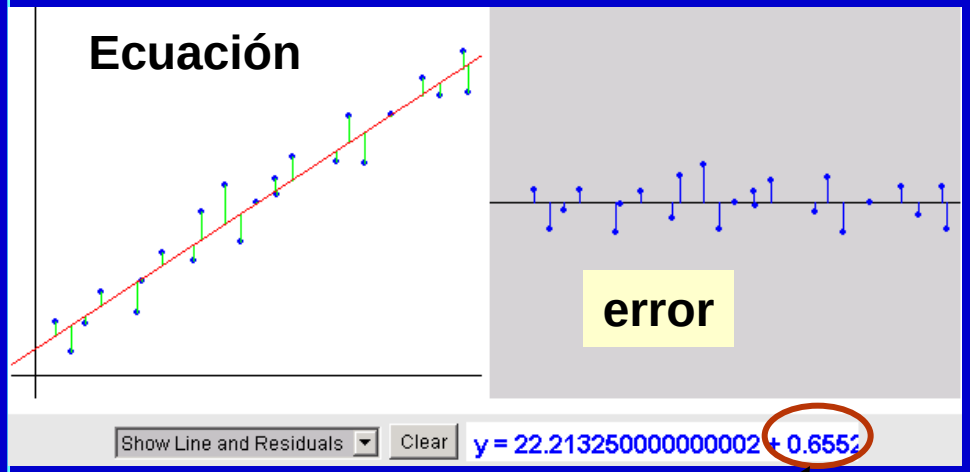
- Si Y y X no están relacionadas, entonces E(Y|X)= E(Y) – Deberíamos predecir el mismo valor de Y para todo valor de X.

Y= constante + 0*X= E (y)



Existe correlación: ¿Qué debemos esperar?

- Si Y y X están relacionadas, entonces E(Y|X)<>E (Y)
- Deberíamos predecir un valor diferente de Y para cada valor de X.
- La pendiente de la recta debe ser diferente de cero



Pendiente

¿Qué debemos esperar?

- Para el valor medio de X, predecimos la media de Y. Cuando X se desvía de la media, también el valor de Y se desviará de su media
- Esto nos indica que X explica la desviación de Y con respecto a su valor medio.

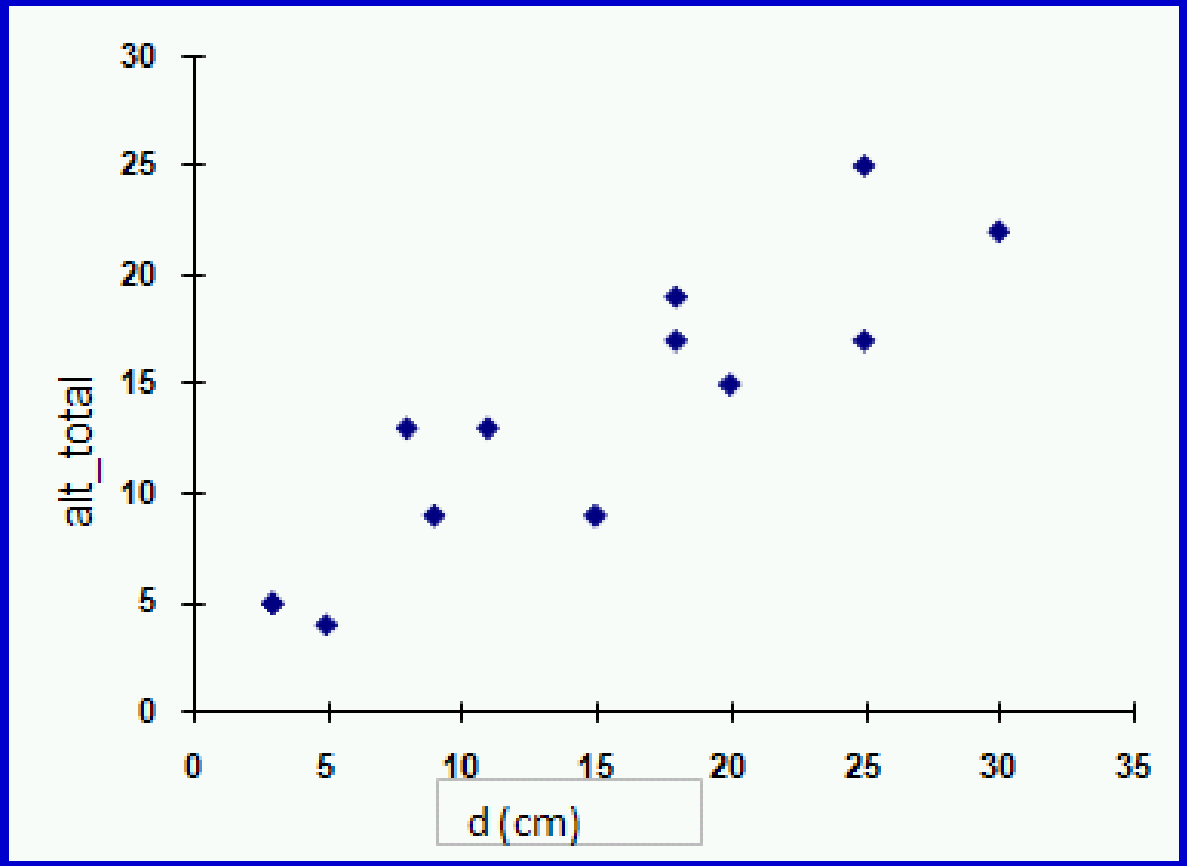
Modelo de Regresión lineal Simple : Ejemplo

Deseamos examinar la relación entre diámetro (cm) y altura total (m) para una muestra de 12 árboles.

¿Cuál es la recta que mejor ajusta a los datos?

d (cm)	h tot(m)
3	5
5	4
8	13
9	9
11	13
15	9
18	17
18	19
20	15
25	17
25	25
30	22

Diagrama de dispersión



Video 1: Relación positiva

Video 2: relación negativa

Video 3: ausencia de relación

http://statweb.calpoly.edu/chance/applets/LRApplet.html

http://hadm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html

Ecuación de la recta que mejor ajusta a los datos

$$\hat{Y}_i = b_0 + b_1 X_i$$

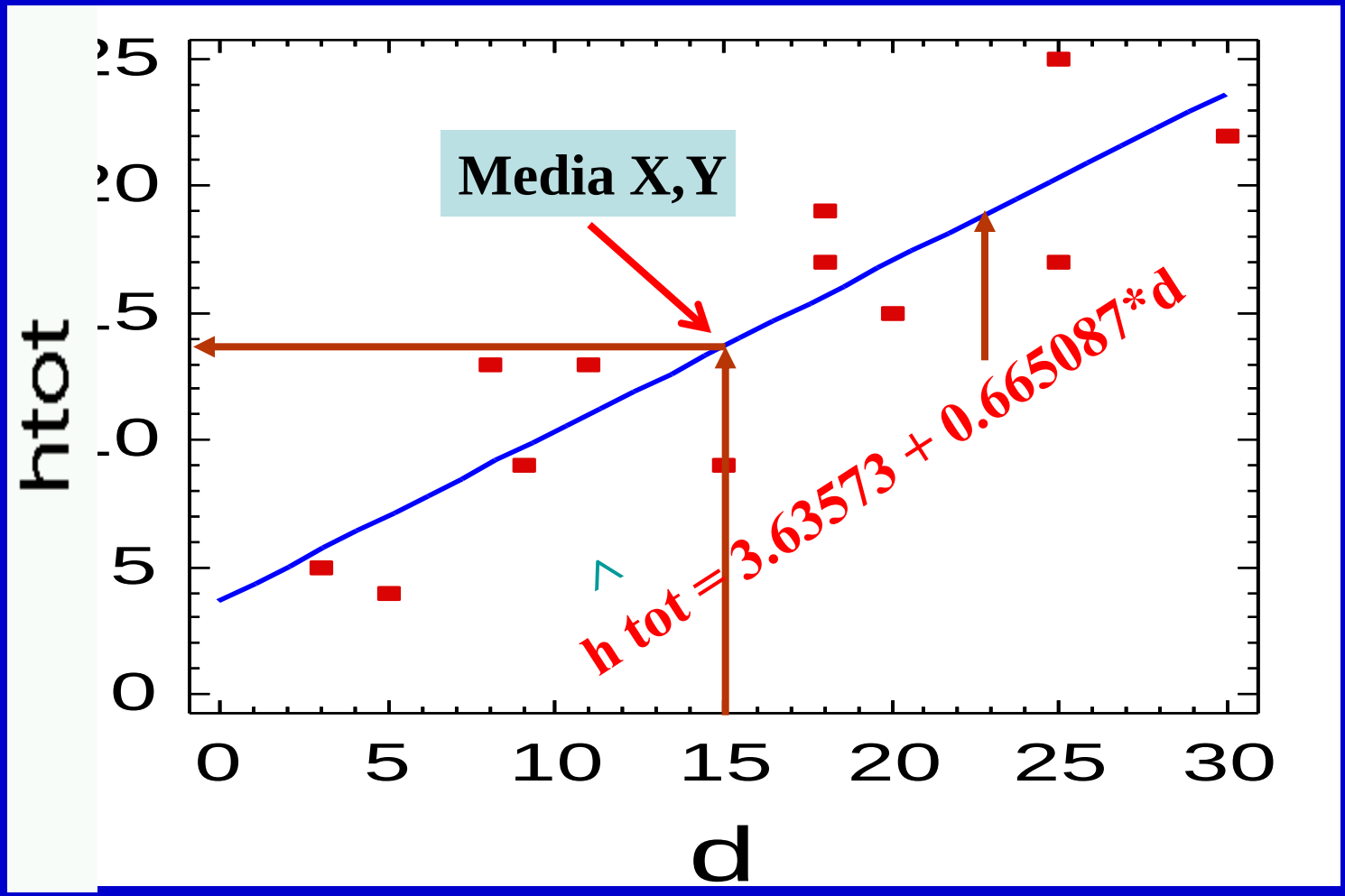
h tot = 3.63573 + 0.665087*d

Intercept	**3.63573 m**
Slope	**0.665087**

Si X=0, entonces $\hat{Y}$ = 3.63m

Un árbol con 0 metros tendrá una altura total de 3.6 m ¿Es esto realista?

Gráfico de la ecuación de regresión



Interpretación de resultados

Ecuación de regresión: pendiente

$$\hat{ht}_i = 3.63573 + 0.665087 * d_i$$

La pendiente 0.665 indica que por cada incremento de 1cm en diámetro, el valor de Ht incrementará en 0.665 metros. <u>La pendiente no tiene unidades</u>

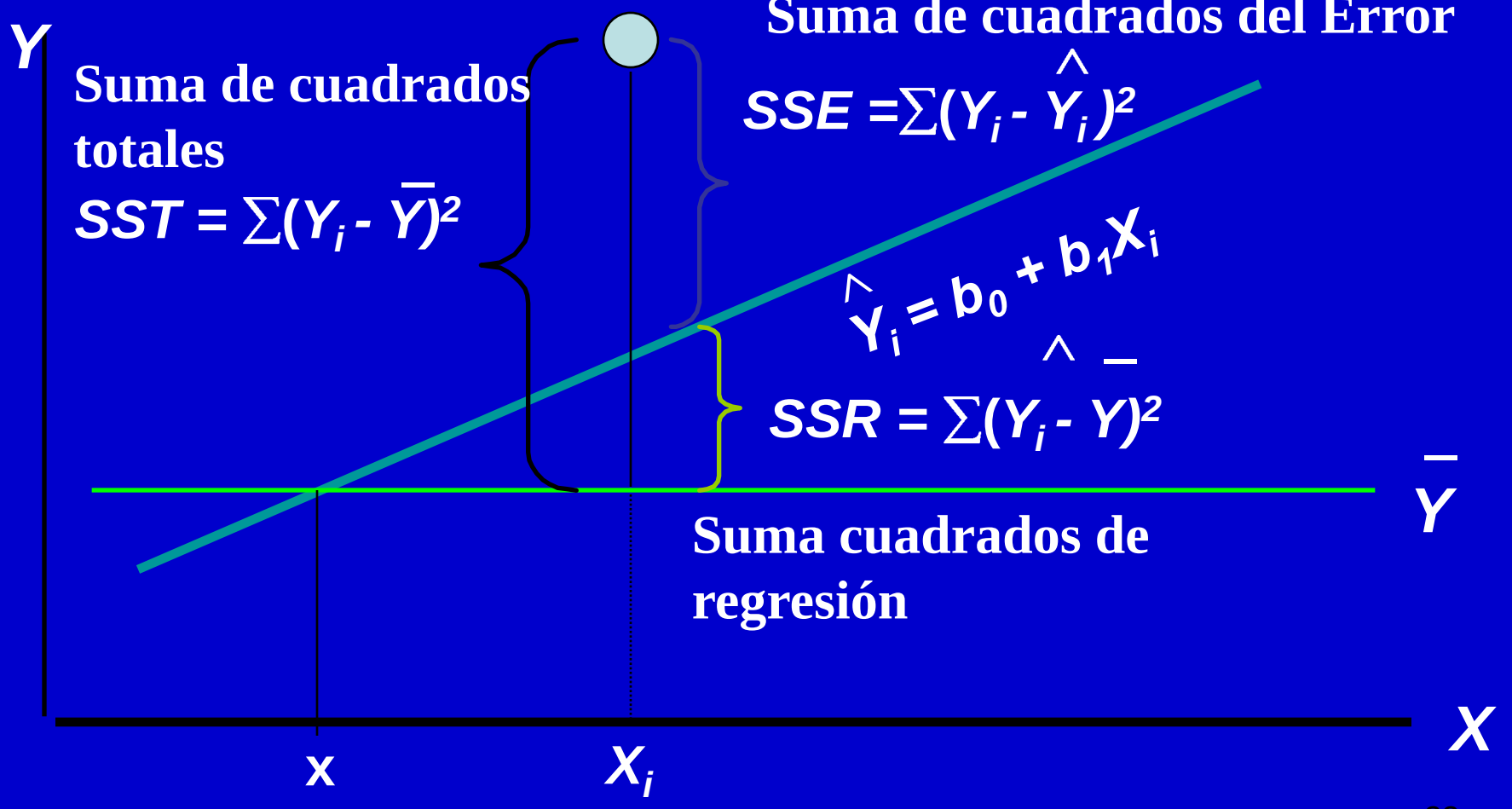
El modelo predice un incremento de 0.665 m en la altura del árbol por cada incremento de 1cm en su diámetro

Significancia del modelo: ¿Explica "X" una porción significativa de la variación en Y?

¿Qué explica la variación en Y?

- Si Y y X no están relacionadas, entonces E(Y|X)= E(Y) y deberíamos de predecir el mismo valor de Y para todo valor de X.
- De existir correlación entre X y Y, entonces debemos determinar si conociendo el valor de X podemos explicar porqué Y toma un valor diferente a su media

Medidas de Variación: Suma de cuadrados



Medidas de Variación: Suma de cuadrados

***SST* = Suma de cuadrados total**

- **Miden la variación de Y_i *alrededor de su media***

***SSR* = Suma de cuadrados de Regresión**

- **Variación explicada por la relación entre *X* y *Y***

***SSE* = Suma de cuadrados del Error**

- **Variación no explicada por la relación entre X y Y (se debe a otros factores no medidos)**

Medidas de Variación: Suma de cuadrados y ANOVA

***SST* = Suma de cuadrados total**

- **Medida de variación utilizada en el ANOVA**

***SSR* = Suma de cuadrados de Regresión**

- **Se denomina suma de cuadrados entre en el ANOVA**

***SSE* = Suma de cuadrados del Error**

- **Se denomina suma de cuadrados dentro en el ANOVA**

Medidas de Variación: Suma de cuadrados : Ejemplo

Datos generados por XLSTats

Tabla de Análisis de Varianza

Analysis of Variance

ANOVA Table

Source	DF	SS	MS	F	p-value
Regression	1	357.81683	357.81683	34.344976	0.0001593
Residual	10	104.18317	10.418317		
Total (corrected)	11	462			
Mean	1	2352			
Total (uncorrected)	12	2814			

SSR

SST

SSE

$R^2 = 357.817/462 = 0.77$

Interpretación del ANOVA: Significancia del modelo

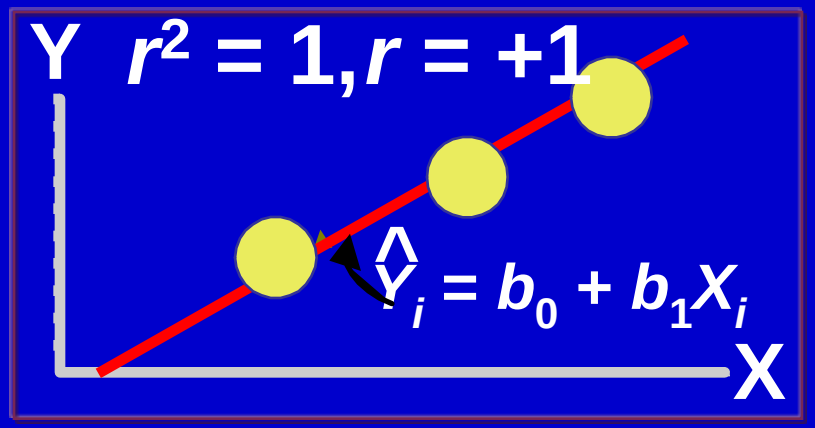
- El estadístico F somete a prueba la hipótesis nula de que el modelo de regresión no explica una proporción significativa de la variación en Y (No existe regresión de Y en X)
- Los grados de libertad de la prueba F para una regresión simple son 1 y n-2
- Para este ejemplo, F=34.34 con 1 y 10 grados de libertad.

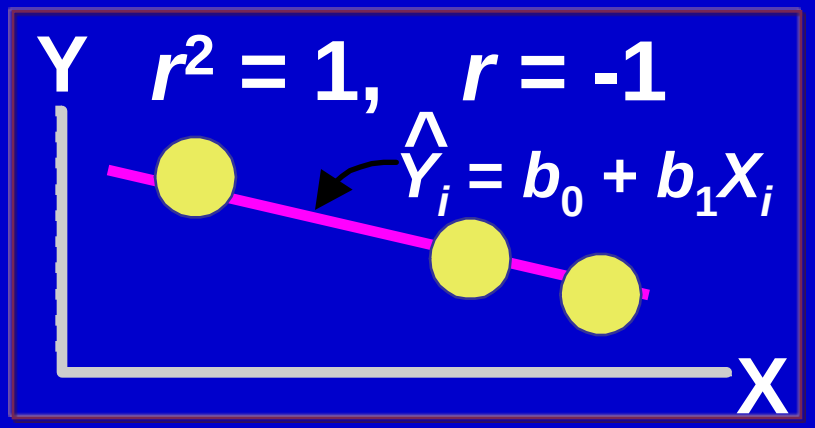
El coeficiente de determinación

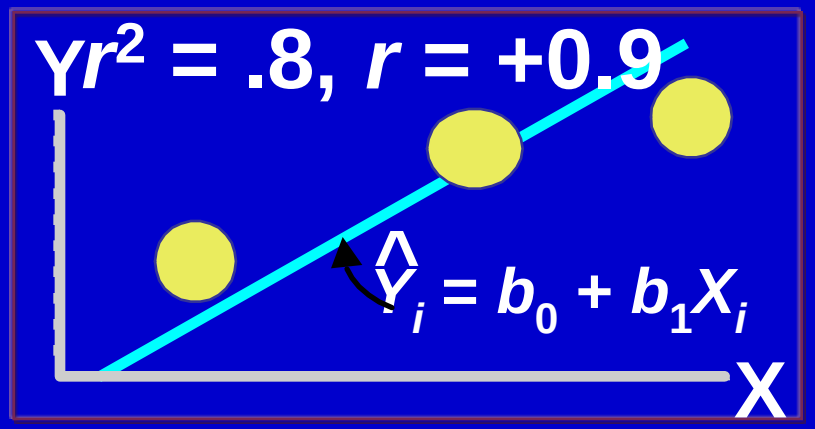
$$r^2 = \frac{SSR}{SST} = \frac{\textit{suma de cuadrados de regresión}}{\textit{suma de cuadrados total}}$$

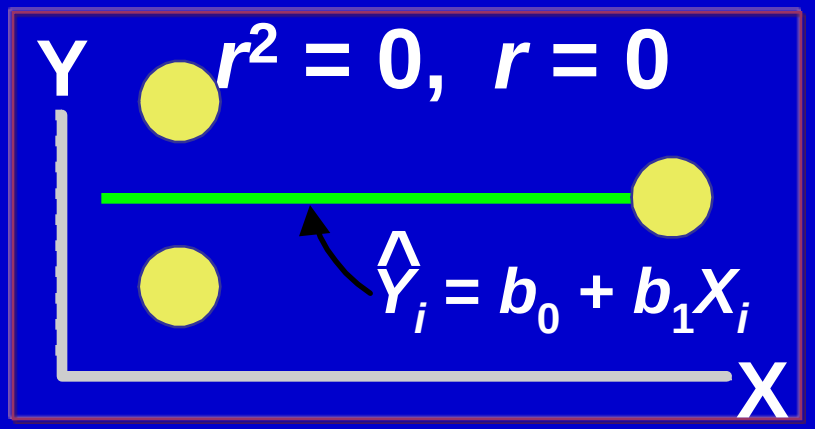
Este coeficiente mide la proporción de la variación de Y que es explicada por la variable predictora X en el modelo de regresión $Y_i = b_0 + b_1 X_i$

Coeficientes de Determinación (r^2) y de Correlación (r)







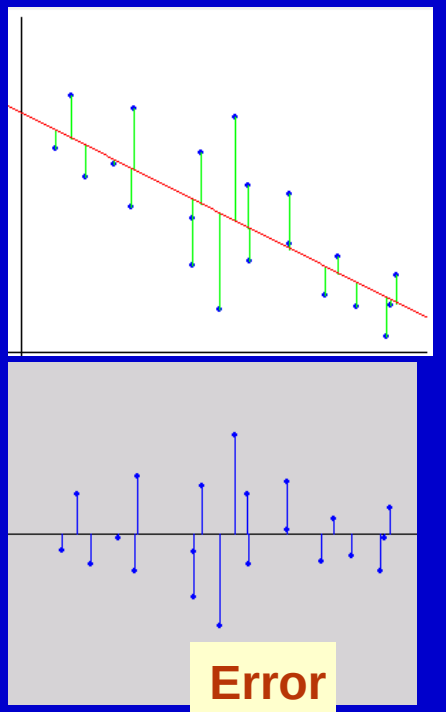


Error Estándar de Estimación

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Desviación estándar de la variación de las observaciones alrededor de la línea de regresión

Cuando la Suma Cuadrados Error (SSE) es grande el valor del Error Estándar de Estimación también será grande



Medidas de Variación : Ejemplo

Resultados de XLSTats

Error estándar de Estimación = 3.22774 m ← S_{yx}

Correlation Coeff	
Correlation	0.880054

Este valor es igual a la raíz cuadrada de 10.4183 (Cuadrado medio de regresión)

R^2 = 0.7745

75% de la variación en altura total es explicada por la variabilidad en el diámetro de los árboles

Inferencia sobre la pendiente: Prueba t

- **Prueba t para la Pendiente de la Población**

¿Existe una relación lineal entre *X* & *Y* ?

- **Hipótesis Nula y Alternativa**

H_0: $\beta_1 = 0$ (No existe relación lineal)
H_1: $\beta_1 \neq 0$ (Existe relación lineal)

- **Estadístico de prueba:** $t = \dfrac{b_1 - \beta_1}{S_{b_1}}$ Donde $S_{b_1} = \dfrac{S_{YX}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$

y df = n - 2

Ejemplo: Diámetro-altura total

Datos para 12 árboles:

d (cm)	htot (m)
3	5
5	4
8	13
9	9
11	13
15	9
18	17
18	19
20	15
25	17
25	25
30	22

Ecuación de Regresión:

$$\hat{Y}_i = 3.63573 + 0.665 X_i$$

La pendiente del modelo es 0.665087

¿Existe una relación lineal entre el diámetro (cm) y la altura total (m) de los árboles?

Inferencia sobre la pendiente: Ejemplo

- H_0: $\beta_1 = 0$
- H_1: $\beta_1 \neq 0$

  $\alpha = .05$
- df = **12** - 2 = 10
- Prueba hipótesis

Hypothesis Tests	
Slope	
H_0: Slope =	0
Alternative	≠ > <
H_1: Slope ≠	0
p-value =	0.000159

Rechazar H_0

De XLTats

Int. Conf.

	Estimate	SE	Lower	Upper
Slope	0.665087	0.11349	0.41222	0.91795

Decisión: Rechazar Ho

Conclusión:

La evidencia indica que existe una relación lineal entre el diámetro y altura total de los árboles

Relación entre F y t en regresión simple

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	3.63573	1.99895	1.81881	0.0990
Slope	0.665087	0.113487	5.86046	0.0002

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	357.817	1	357.817	34.34	0.0002
Residual	104.183	10	10.4183		
Total	462.0	11			

5.86*5.86= 34.34

La prueba t para B=0 es idéntica a la prueba F para r^2=0 en regresión simple. El valor del estadístico t es la raíz cuadrada del estadístico F (t=0.665087/0.113487= 5.86) $F_{1,n-2} = t^2_{n-2}$

Inferencia sobre la pendiente: IC

Intervalo de confianza para la pendiente

$$b_1 \pm t_{n-2}\, S_{b_1}$$

sociated with a model of the form Y = mX + c + error

Summary			Confidence Ints.			
			Level	0.95	R^2	0.7745
	Estimate	SE	Lower	Upper	s	3.22774
Slope	0.665087	0.11349	0.41222	0.91795		
Constant	3.635727	1.99895	-0.8182	8.08967		

XLSTats

A un nivel de confianza de 95% el intervalo de confianza para la pendiente es (0.412-0.917). No incluye cero.

Conclusión: Existe una relación lineal significativa entre el diámetro y la altura de los árboles (alfa 0.05).

Estimación de valores de Y

Intervalo de confianza para

μ_{XY} **Media de Y dado un valor medio de X_i**

El tamaño del intervalo depende de la distancia a $\bar{X}$

Error estándar de estimación

$$Y_i \pm t_{n-2} \bullet S_{yx} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

valor de t con n-2 grados de libertad

Estimación de predicciones de Y

Intervalo de confianza para valores individuales de Y_i a un valor particular de X_i

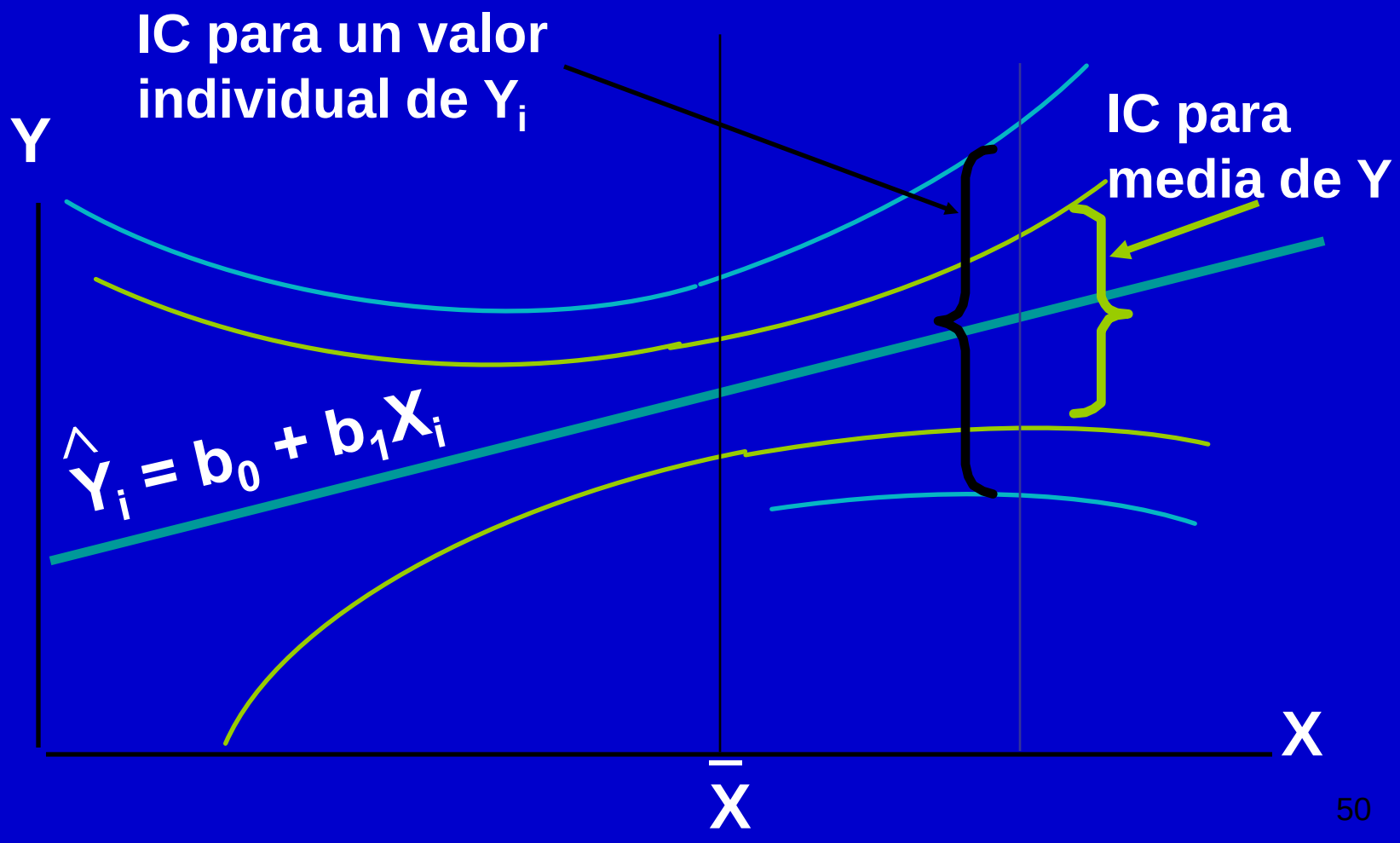
La adición de este 1 incrementa el ancho del intervalo con respecto al IC para la media

$$\hat{Y}_i \pm t_{n-2} \bullet S_{yx} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

Intervalos de confianza

- El error asociado a una estimación tiene dos componentes:
  - Error en la media (Error estándar de estimación)
  - Error en la estimación de B
- Por tanto, el IC para las estimaciones será mayor cuando más nos alejemos de la media de X

IC para diferentes valores de X



Predicciones de Y: Ejemplo

IC para un valor Y dado un valor medio de X

Calcular el IC al 95% para la altura media de los árboles dado un valor medio de diámetro

Altura estimada $\hat{ht}_i$ **= 3.63573 + 0.665087*d**$_i$ ***= 13.99 m***

$\bar{X}$ **= 15.58 cm** S_{YX} **= 3.227 m** $t_{n-2} = t_{10} = 2.228$

$$\hat{Y}_i \pm t_{n-2} \bullet S_{yx} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

= 13.99 m ± 2.07 m

IC para la media de Y

Predicciones de Y: Ejemplo

IC para un valor de Y a un valor particular de X

Calcular el IC al 95% para la altura total de un árbol dado un valor particular de diámetro

Altura estimada $\hat{ht}_j$ **= 3.63573 + 0.665087*d** $_i$ **=** ***12.9 m***

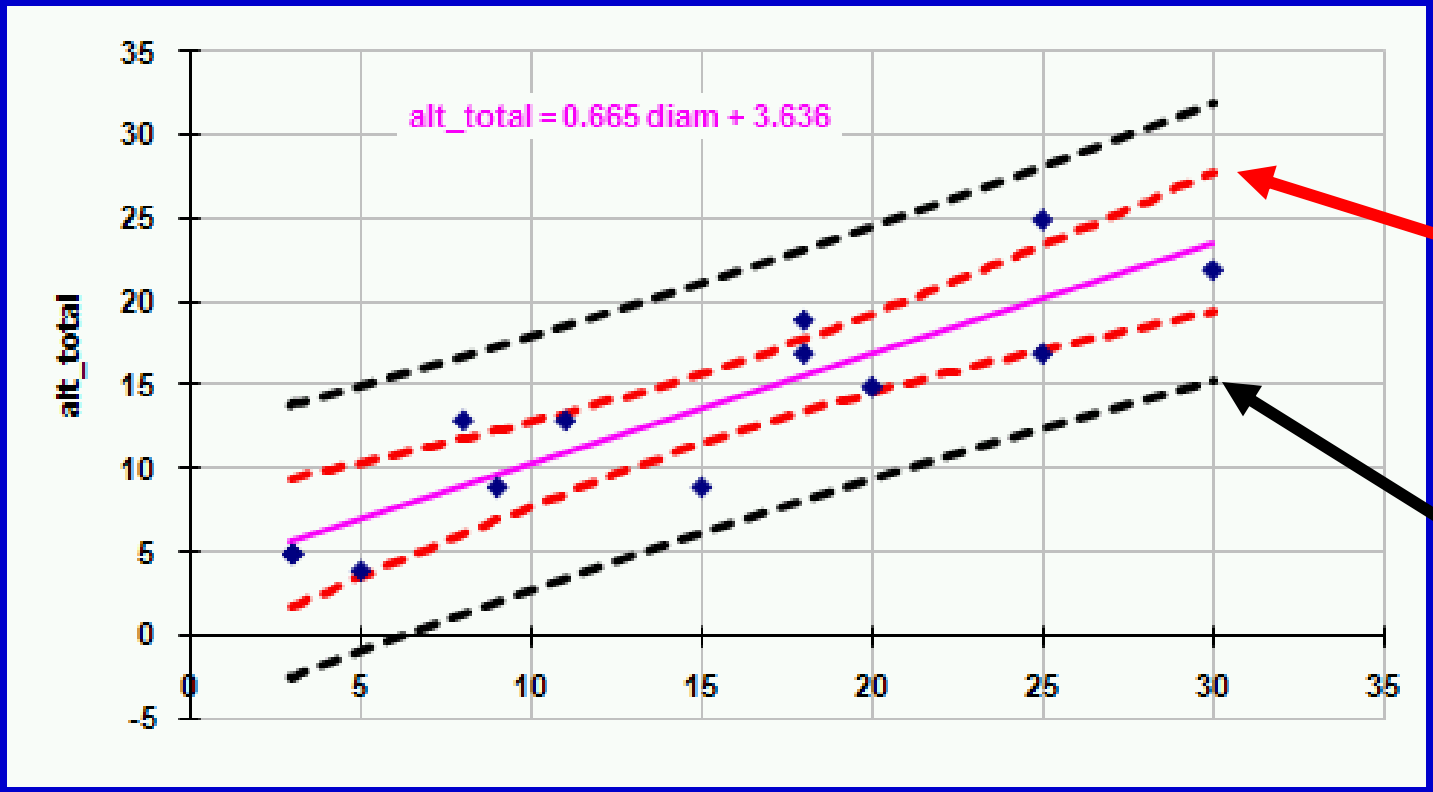
$\bar{X}$ **= 15.58 cm** S_{YX} **= 3.227 m** $t_{n-2} = t_{10} = 2.228$

$$\hat{Y}_i \pm t_{n-2} \bullet S_{yx} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

= 13.99 m $\pm$ **7.49 m**

IC para la valor individual de y

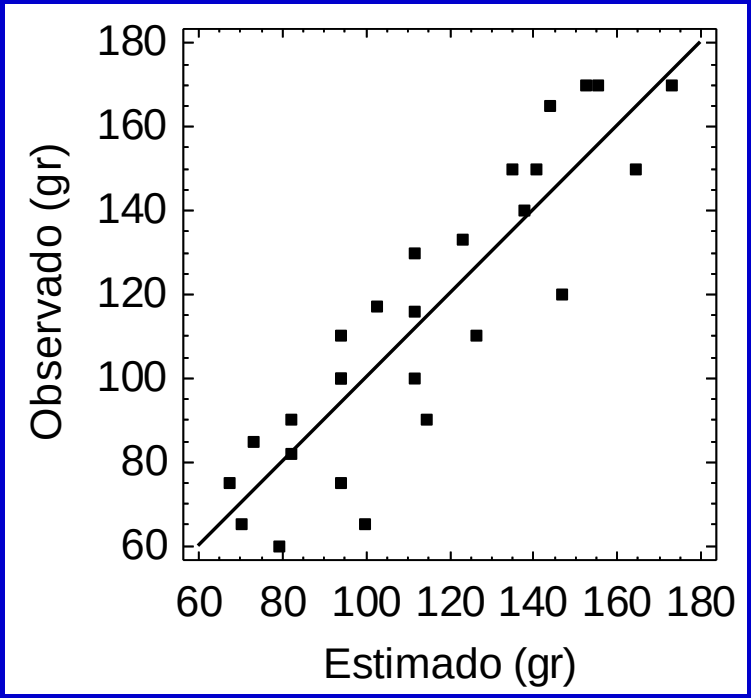
Intervalos de Confianza

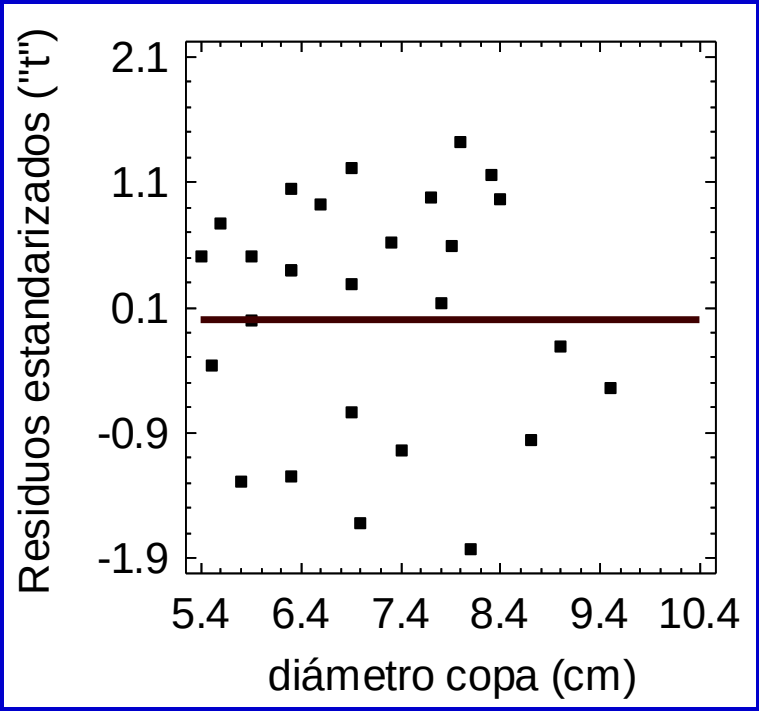


Valores medios

Valores individuales

Análisis de residuos

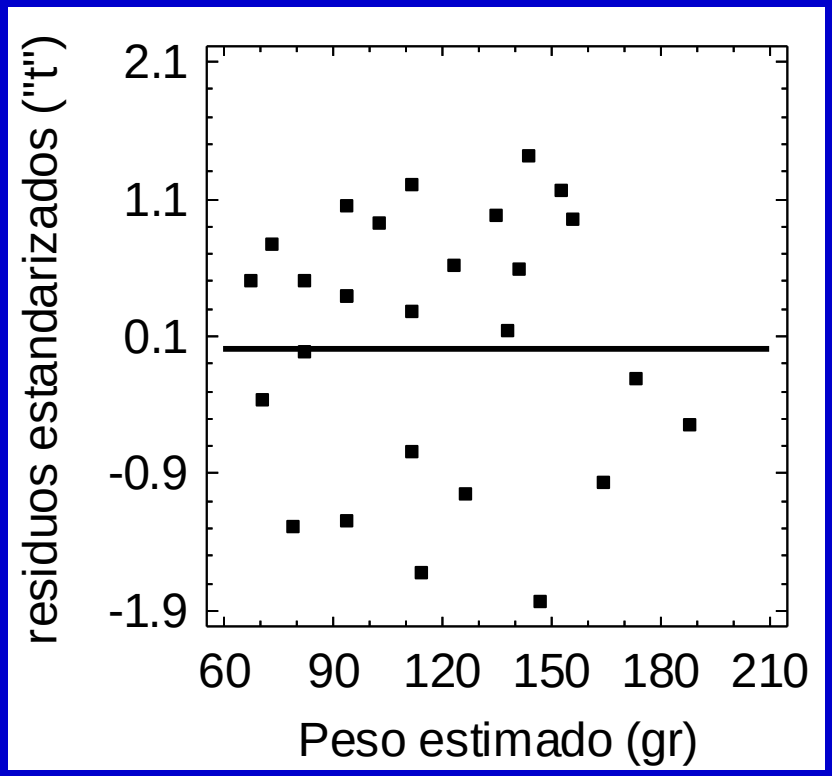




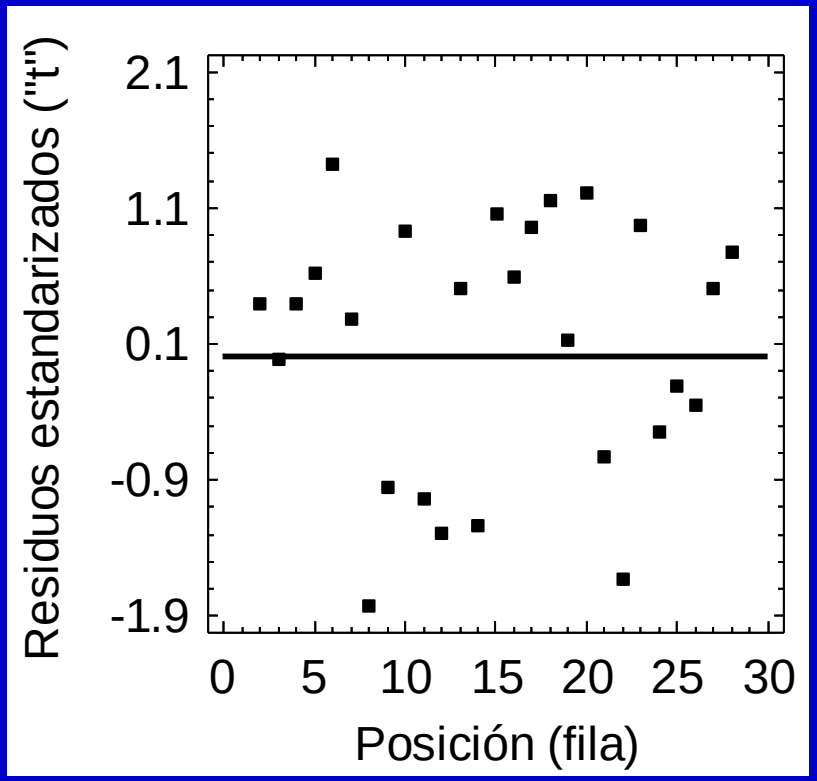
Si el modelo no exhibe ninguna anomalía los datos deben tender a una línea con una ángulo de 45 º (relación 1:1).

Si el modelo no exhibe ninguna anomalía los residuos deben formar una banda alrededor de cero.

Análisis de residuos



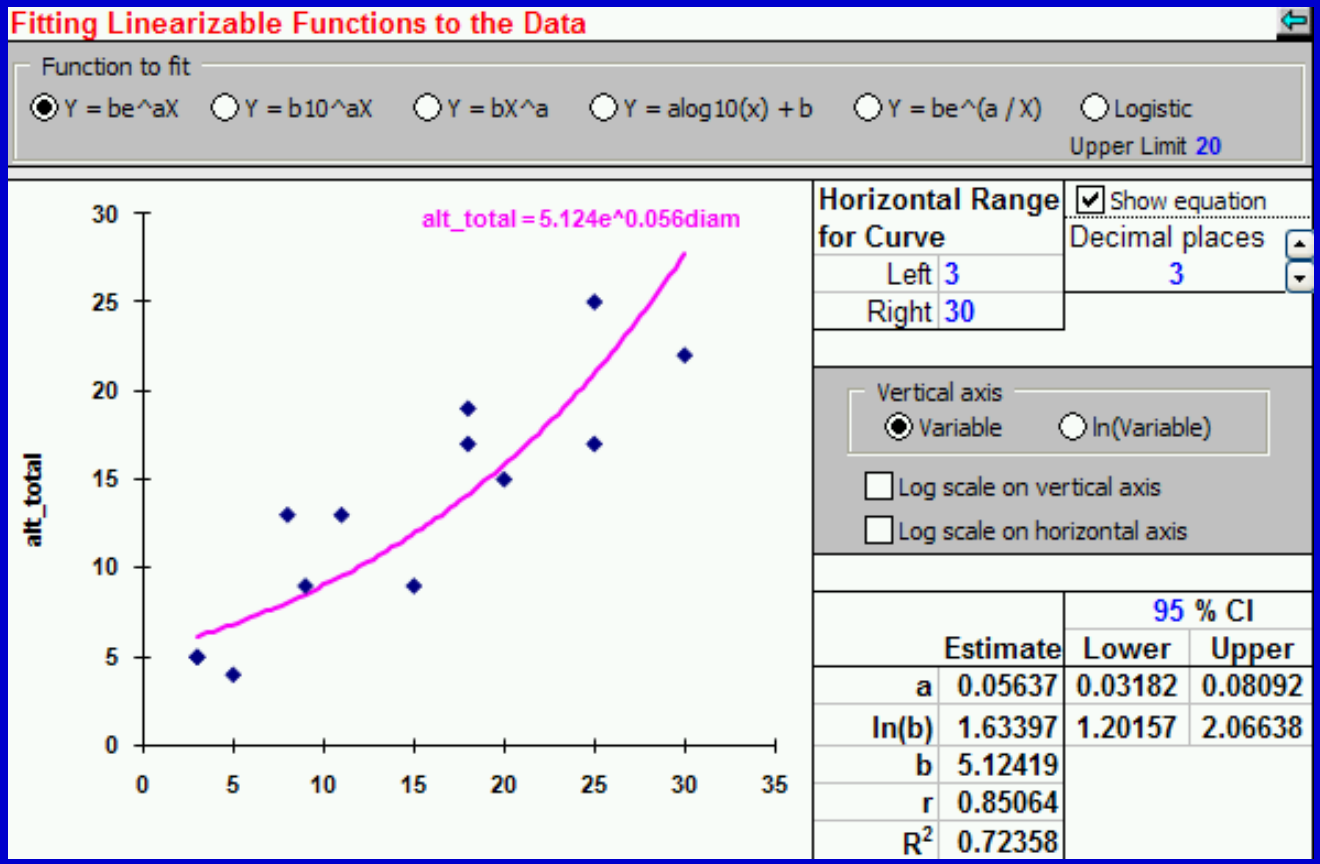
Si el modelo no exhibe ninguna anomalía los residuos deben formar una banda alrededor de cero.



Si el modelo no exhibe ninguna anomalía los residuos deben formar una banda alrededor de cero.

¿Es el modelo lineal el mejor?

- Una vez ajustado un modelo lineal, XLSTats le permite ajustar 5 modelos adicionales a sus datos



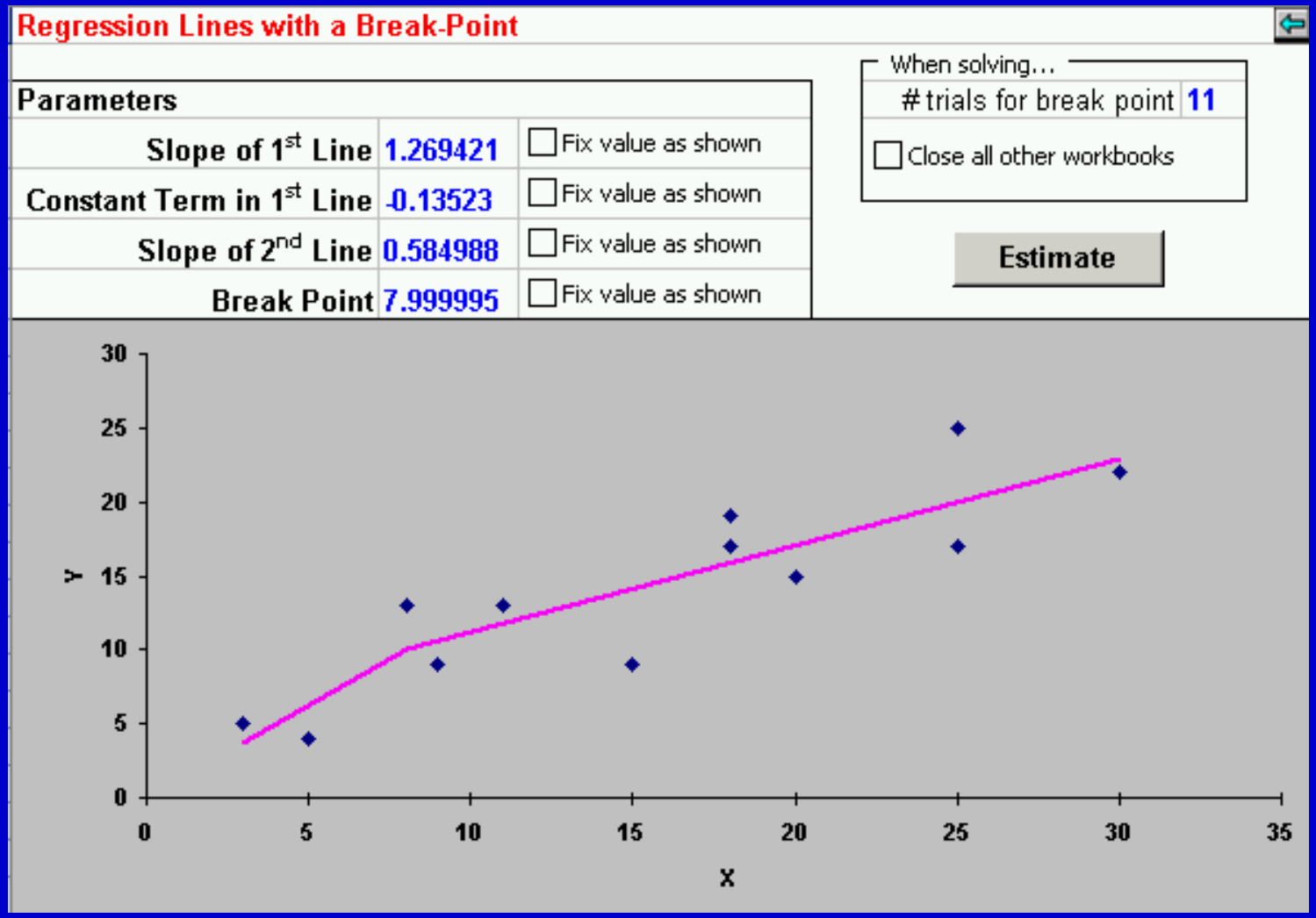
¿Es el modelo lineal el mejor?

Comparación de modelos alternativos

Model o	Correlación	R-Cuadrado
Multiplicative	0.8926	79.67%
Square root-X	0.8839	78.12%
Linear	0.8801	77.45%
Square root-Y	0.8735	76.30%
Logarithmic-X	0.8663	75.05%
Exponential	0.8506	72.36%
Double reciprocal	0.8424	70.97%
Reciprocal-Y	-0.7703	59.33%
Reciprocal-X	-0.7657	58.63%

El R^2 del modelo multiplicativo es 79.7% y explica un 2.2% adicional de la variabilidad en la altura de los árboles comparado con el modelo lineal

¿Es el modelo lineal el mejor?



XLSTats le permite ajustar una ecuación de regresión en segmentos

RESUMEN

- **Realizar análisis de correlación**
- **Graficar datos, calcular R**
- **Seleccionar modelo inicial que mejor se ajuste a los datos (Ej. Lineal, no lineal)**
- **Evaluar parámetros del modelo**
- **Evaluar residuos. Probar supuestos del modelo. ¿Existen residuos inusuales y/o puntos influyentes?** Transformar datos.
- **Comparar con modelos alternativos**
- **Predicción de valores de Y a partir de valores de X**