Universidad Nacional de La Plata

Facultad de Ciencias Agrarias y Forestales

CÁLCULO ESTADÍSTICO Y BIOMETRÍA

Técnicas de Análisis simultaneo de variables

ANALISIS DE REGRESION

Determinar como se relacionan funcionalmente dos o más variables

ANALISIS DE CORRELACION

Determinar si las variables están asociadas y como es esa asociación

Problemáticas

Muchas veces estamos interesados en describir cómo cambia una variable (que llamaremos **dependiente**) en función de una (o varias) llamada/s **independiente**/s.

- ✓ cuantificar cómo cambia la disponibilidad de agua a distintas profundidades del perfil analizado en un cultivo de soja.
- ✓ *evaluar la efectividad de un insecticida a distintas concentraciones sobre la sobrevida de un insecto.*
- ✓ *estudiar el comportamiento del aumento de peso de animales sobre distintas disponibilidades de una ración.*
- ✓ *estudiar el efecto de distintas disponibilidades de luz y/o la temperatura sobre el poder germinativo de una semilla forrajera.*

ANALISIS DE REGRESION

Objetivos

Identificar

- ... la mejor relación funcional entre una variable aleatoria (llamada variable dependiente y denotada generalmente con Y), y una o mas variables no aleatorias (independiente, indicada con X_i)

Estimar

- ... los parámetros del modelo

Probar hipótesis (evaluar la adecuación del modelo)

- ... sobre los parámetros del modelo

Predecir

- ... el nivel medio de la variable dependiente para valores determinados de las independientes.

Regresión lineal simple - MODELO

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- y_i: variable dependiente o respuesta,
- x_i: variable independiente,
- β_0 y β_1: parámetros desconocidos que representan la ordenada al origen y la pendiente respectivamente,
- ε_i: error aleatorio

Suposiciones del modelo

- Los errores son variables aleatorias no correlacionadas
- Los errores tienen distribución normal con media cero y varianza cte.

Estimación de parámetros

Modelo Poblacional

$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i$$

Modelo Muestral

$$\hat{y}_i = a + b . x_i + \varepsilon_i$$

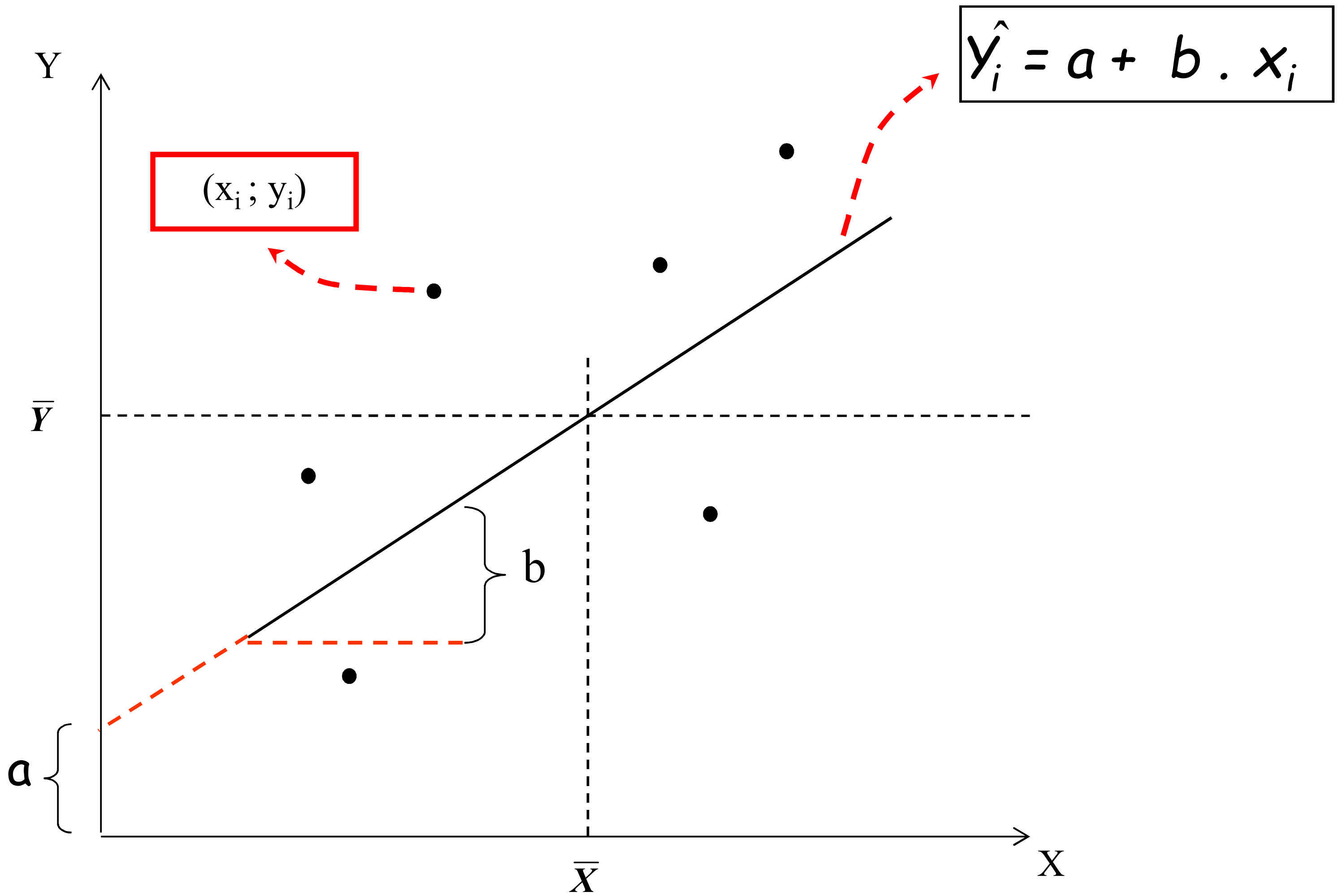
- ✓ Dados $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$
- ✓ Se estiman β_0 y β_1 minimizando la suma de los cuadrados de $(y_i - \hat{y}_i)$ donde $\hat{y}_i$ es el valor estimado de y_i.

$$a = \frac{\sum_{i=1}^{n} Y_i \quad b\sum_{i=1}^{n} X_i}{n} = \overline{Y} - b\,\overline{X}$$

$$b_{y/x} = \frac{\sum_{i=1}^{n} X\,Y - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}}$$

- ✓ La diferencia $(Y_i - \hat{Y}_i) = \varepsilon_i$ se conoce como el residuo de la i-ésima observación.

Gráfico de dispersión



Pruebas de hipótesis sobre los parámetros

Ordenada al origen

$$H_0: \quad \beta_0 = 0$$
$$H_1: \quad \beta_0 \neq 0$$

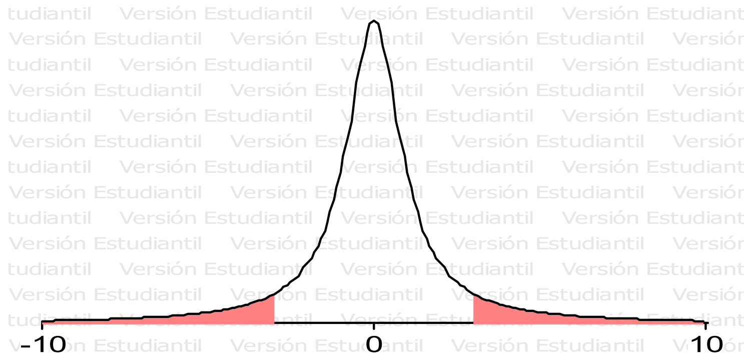
$$\boxed{t_a = \frac{a}{S_a}}$$

Pendiente

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

$$\boxed{t_b = \frac{b}{S_b}}$$

$$t_{\text{tabulado}} \begin{cases} gl = n - 2 \\ \alpha \end{cases}$$



Bondad de Ajuste al modelo

Coeficiente de determinación (R^2): porcentaje de la variabilidad total de y explicada por la regresión (modelo) o la variable dependiente x.

$$\boxed{R^2 = \frac{\sum \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum \left(Y_i - \overline{Y}\right)^2} = \frac{\text{SCExp.}}{\text{SCT}}}$$

siendo $\mathbf{0 \leq R^2 \leq 1}$

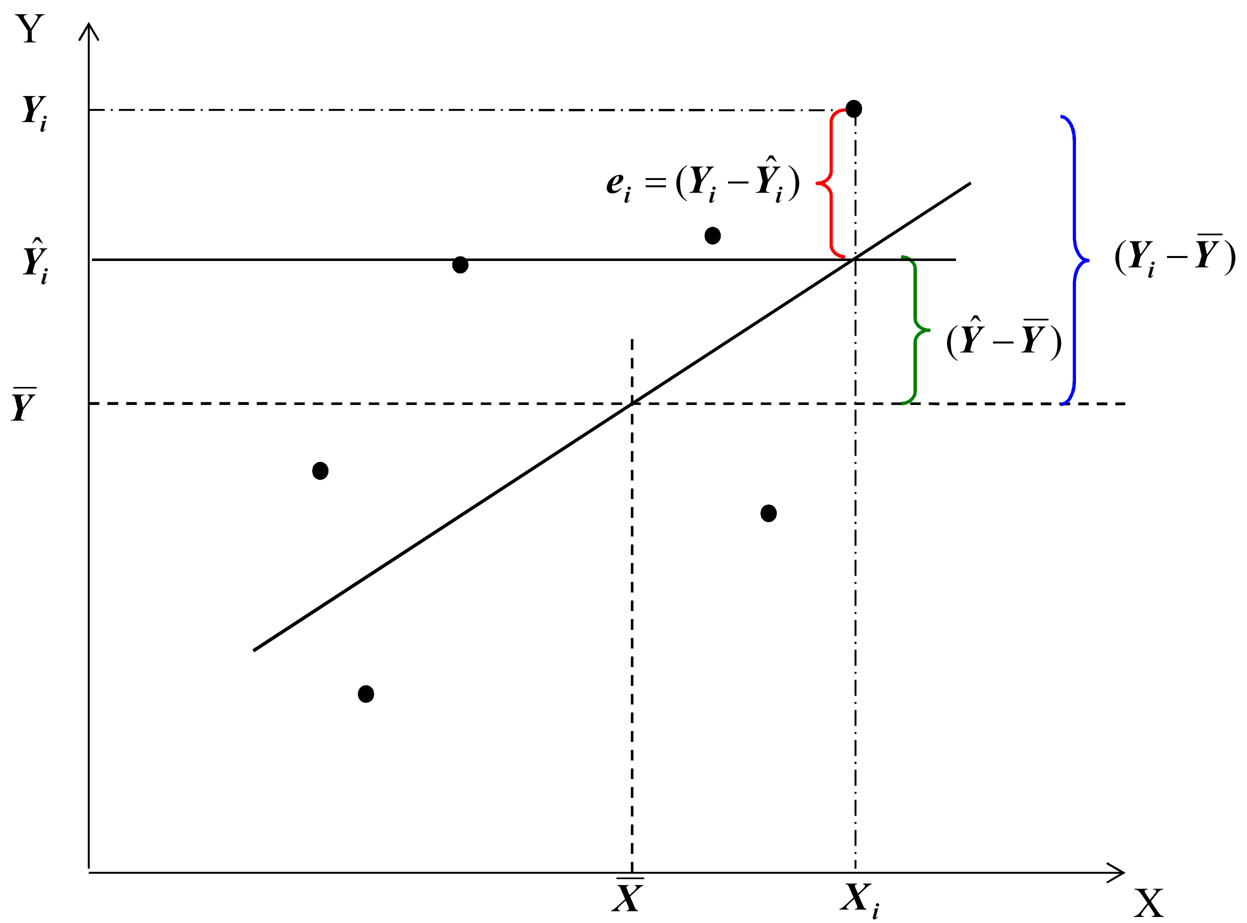
Objetivo: descomponer la variabilidad total de la variable Y en:

- la variabilidad explicada por el modelo
- la variabilidad no explicada o variabilidad residual

$$\text{SCT} = \sum \left(Y_i - \overline{Y}\right)^2 = \text{SCExp.} + \text{SCRes.}$$

$$\text{SCExp.} = \sum \left(\hat{Y}_i - \overline{Y}\right)^2 \qquad\qquad \text{SCRes.} = \sum \left(Y_i - \hat{Y}_i\right)^2$$

Gráfico de dispersión (Descomposición variabilidad)



ANOVA: Otra forma de probar la significancia de la regresión

En regresión lineal simple las hipótesis siguientes son EQUIVALENTES

$$H_0: R^2 = 0 \qquad\qquad \mathrm{H}_0: \quad \beta_1 = 0$$

$$H_1: R^2 \neq 0 \qquad\qquad \mathrm{H}_1: \quad \beta_1 \neq 0$$

Fuentes de Variación	**Suma de Cuadrados**	**Grados de Libertad**	**Cuadrado Medio**	**F Calculado**
Explicada por la regresión	*SCExp.*	$gl_{Exp.} = k$	$\mathrm{CM}_{\mathrm{Exp.}} = \dfrac{\mathrm{SCExp}}{\mathrm{gl}_{\mathrm{Exp}}}$	$\mathrm{F} = \dfrac{\mathrm{CM}_{\mathrm{Exp}}}{\mathrm{CM}_{\mathrm{Res}}}$
Error	*SCRes.*	$gl_{Res.} = n - k - 1$	$\mathrm{CM}_{\mathrm{Res.}} = \dfrac{\mathrm{SCRes.}}{\mathrm{gl}_{\mathrm{Res.}}}$	
Total	*SCT*	$glt = n - 1$		

$$\mathrm{F}_{\text{tabulado}} \begin{cases} \mathrm{gl}_{\text{(numerador)}} = \mathrm{k} = 1 \\ \mathrm{gl}_{\text{(denominador)}} = \mathrm{n} - 2 \\ \alpha \end{cases}$$

Ejemplo 1 - Regresión

Se determinó el crecimiento de una especie forestal en un suelo pobre en N al cual se le agregaron distintas cantidades de ese elemento.

Nitrógeno agregado (kg/ha)	0	50	100	150	200	250	300	350	400	450
Crecimiento (mg/día)	1.6	20.5	84.8	94.0	143.1	162.5	183.2	218.3	289.0	317.4

Operando con los datos del ejercicio se obtuvieron los siguientes valores:

a = -3,37 Sa = 8,91 b = 0,69 Sb = 0,03 R^2 = 0,98

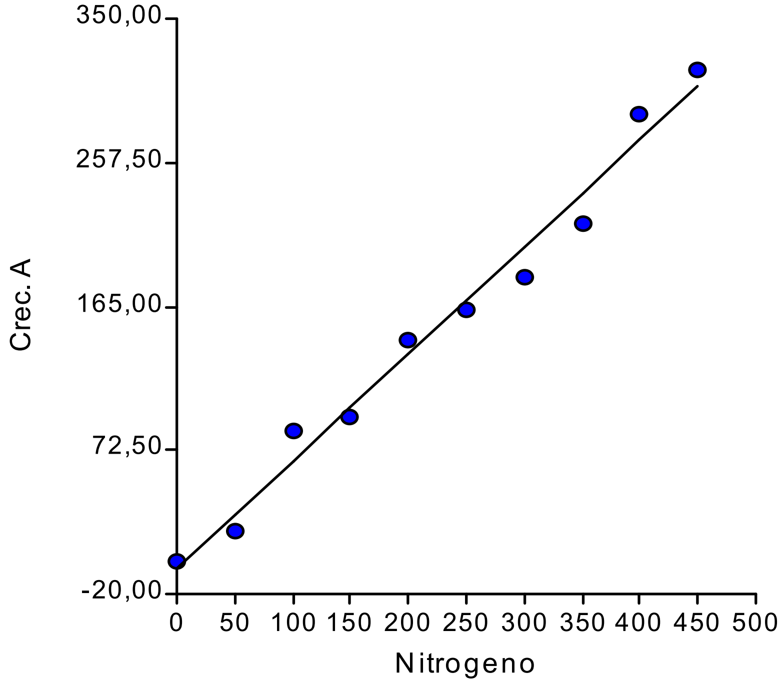
Responder:

a) Realizar el gráfico de dispersión. Explicitar el modelo de regresión e indicar el significado de cada termino.

b) Analizar la significancia de los parámetros del modelo, tomar un nivel de significancia del 5 %. Conclusiones

c) ¿Que es R^2 y que indica en este problema?

Ejemplo 1 - Gráfico



$$\hat{y}_i = -3{,}37 + 0{,}69\, x_i$$

Ejemplo 1 – Salida Infostat

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP
Crec. A	10	0,98	0,98	357,88

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	-3,37	8,91	-23,92	17,19	-0,38	0,7156	
Nitrogeno	0,69	0,03	0,61	0,77	23,00	<0,0001	378,49

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	97634,08	1	97634,08	424,55	<0,0001
Nitrogeno	97634,08	1	97634,08	424,55	<0,0001
Error	1839,78	8	229,97		
Total	99473,86	9			

ANALISIS DE CORRELACION

Generalidades

Finalidad

- medir el grado de asociación y el sentido del mismo entre variables aleatorias que se supone se relacionan linealmente.

Objetivos

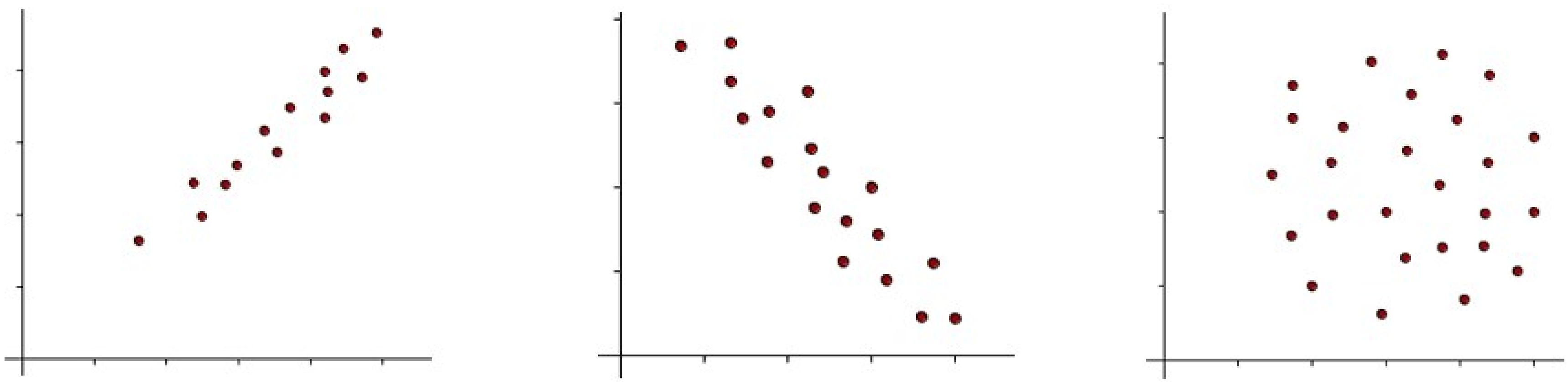
- **Estimar**...
  - ... el coeficiente de correlación ρ (rho)
- **Probar hipótesis** ...
  - ... sobre ρ (rho)

Estimación de ρ (rho)

$\rho \longrightarrow r$

$$\boxed{r = \frac{\sum \delta_x \, \delta_y}{\sqrt{\delta_x^{\,2} \, \delta_y^{\,2}}}}$$

Donde: $-1 \leq \mathbf{r} \leq 1$



$r \longrightarrow 1$ $\qquad$ $r \longrightarrow -1$ $\qquad$ $r \longrightarrow 0$

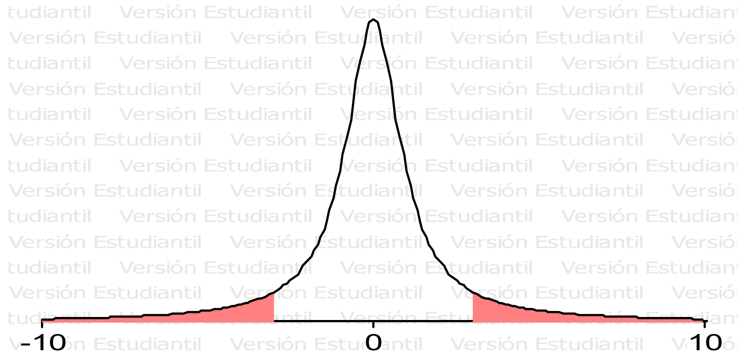
Prueba de Hipótesis

$$H_0 : \quad \rho = 0$$
$$H_1 : \quad \rho \neq 0$$

$$\boxed{t_r = \frac{r}{S_r}}$$

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$$t_{\text{tabulado}} \begin{cases} gl = n - 2 \\ \alpha \end{cases}$$



Ejemplo 2 - Correlación

Se registro el crecimiento de una especie forestal y la profundidad alcanzada por sus raíces en 10 ejemplares de dicha especie.

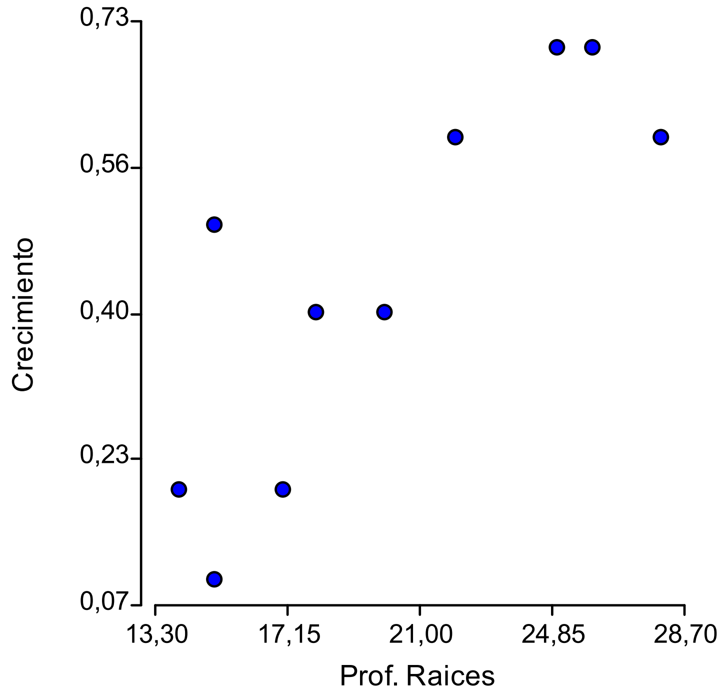
Planta	Crecimiento	Prof. Raíces
1	0,10	15,00
2	0,40	20,00
3	0,50	15,00
4	0,60	28,00
5	0,70	25,00
6	0,20	14,00
7	0,40	18,00
8	0,60	22,00
9	0,70	26,00
10	0,20	17,00

Operando con los datos del ejercicio se obtuvo un **r = 0,82**

Responder:

a) ¿ Que indica el coeficiente de correlación?. Gráficar
b) Analizar la significancia del coeficiente de correlación.

Ejemplo 2 - Gráfico



r → **1**

Ejemplo 2 – Salida InfoStat

Coeficientes de correlación

Correlacion de Pearson: coeficientes\probabilidades

	Crecimiento	Prof. Raices
Crecimiento	1,00	3,4E-03
Prof. Raices	0,82	1,00