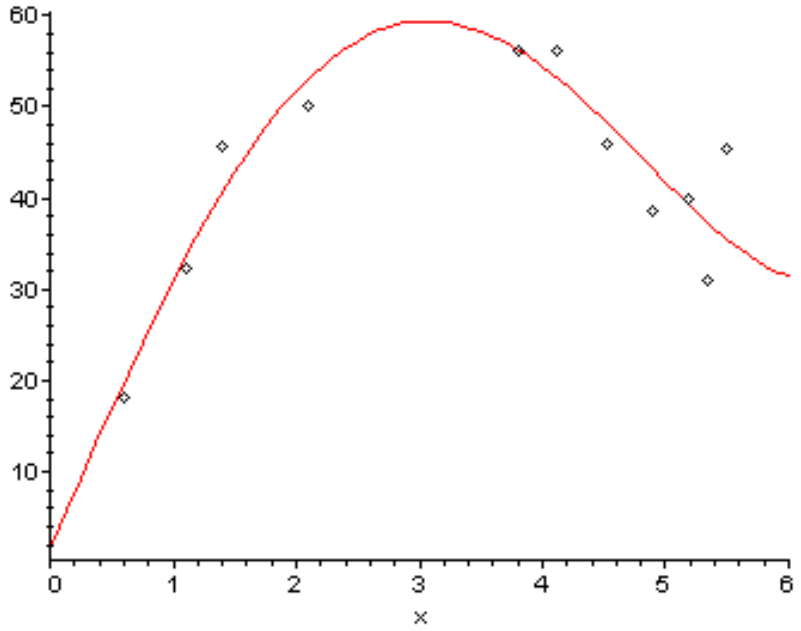
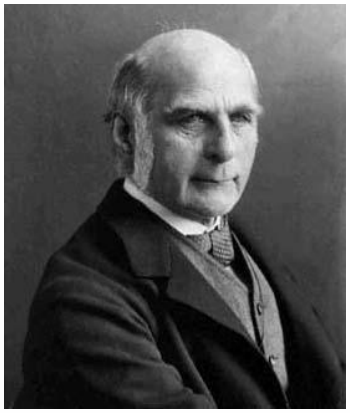
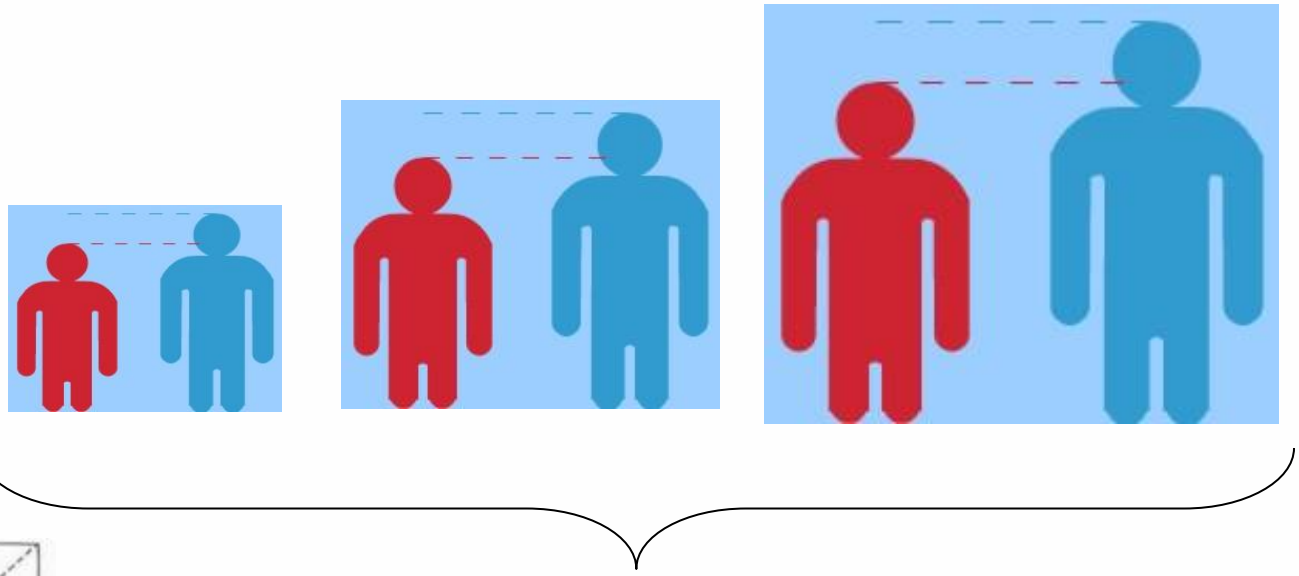


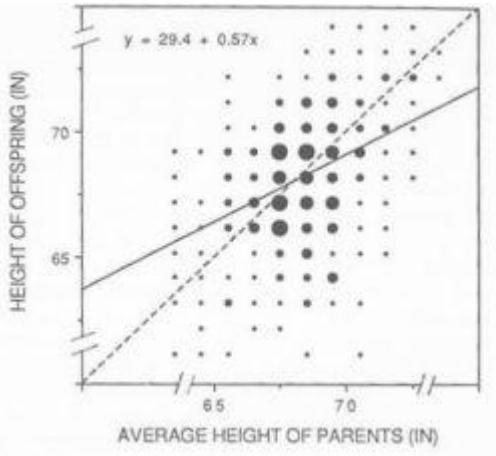
ANÁLISIS DE REGRESIÓN

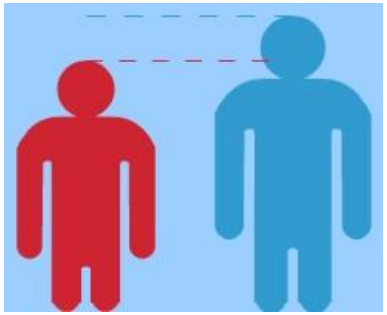
INTRODUCCIÓN



Francis Galtón

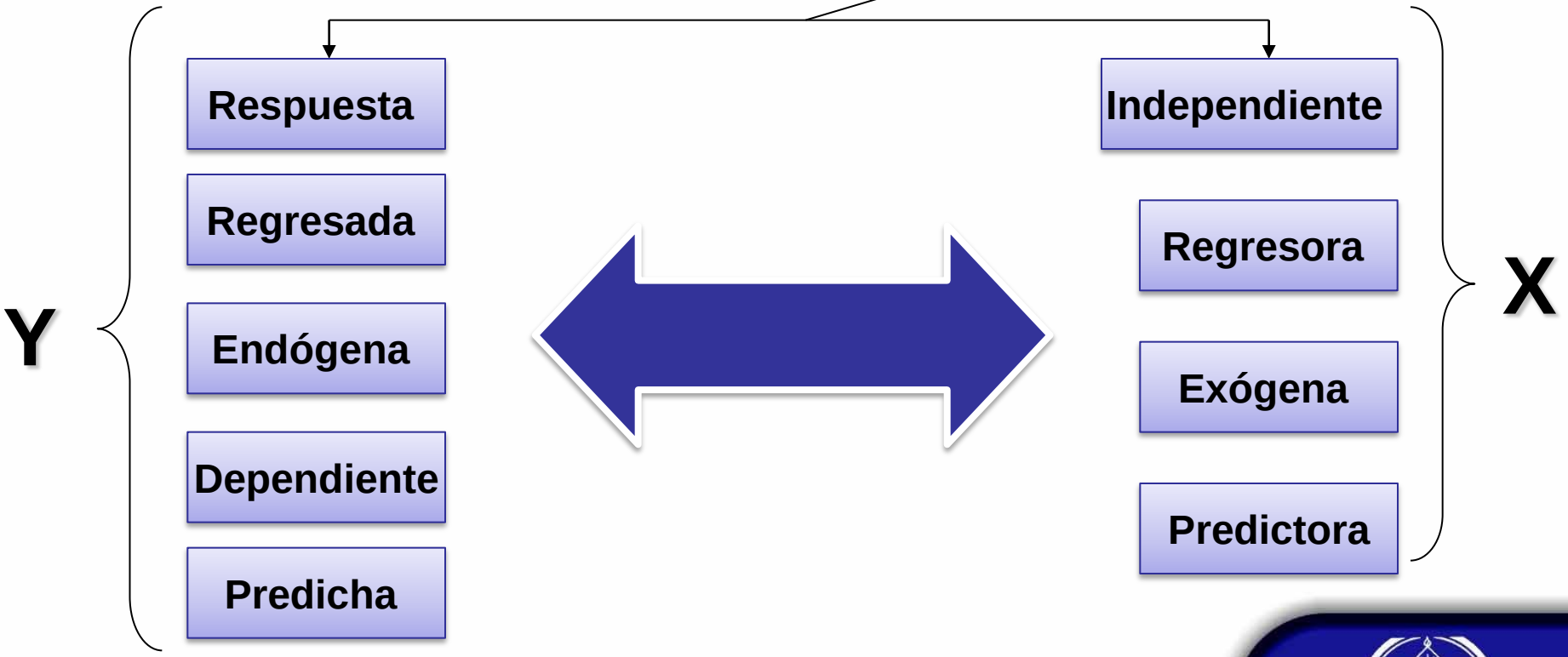






DEFINICIÓN

Es una técnica estadística que se usa para investigar y modelar la relación entre variables.



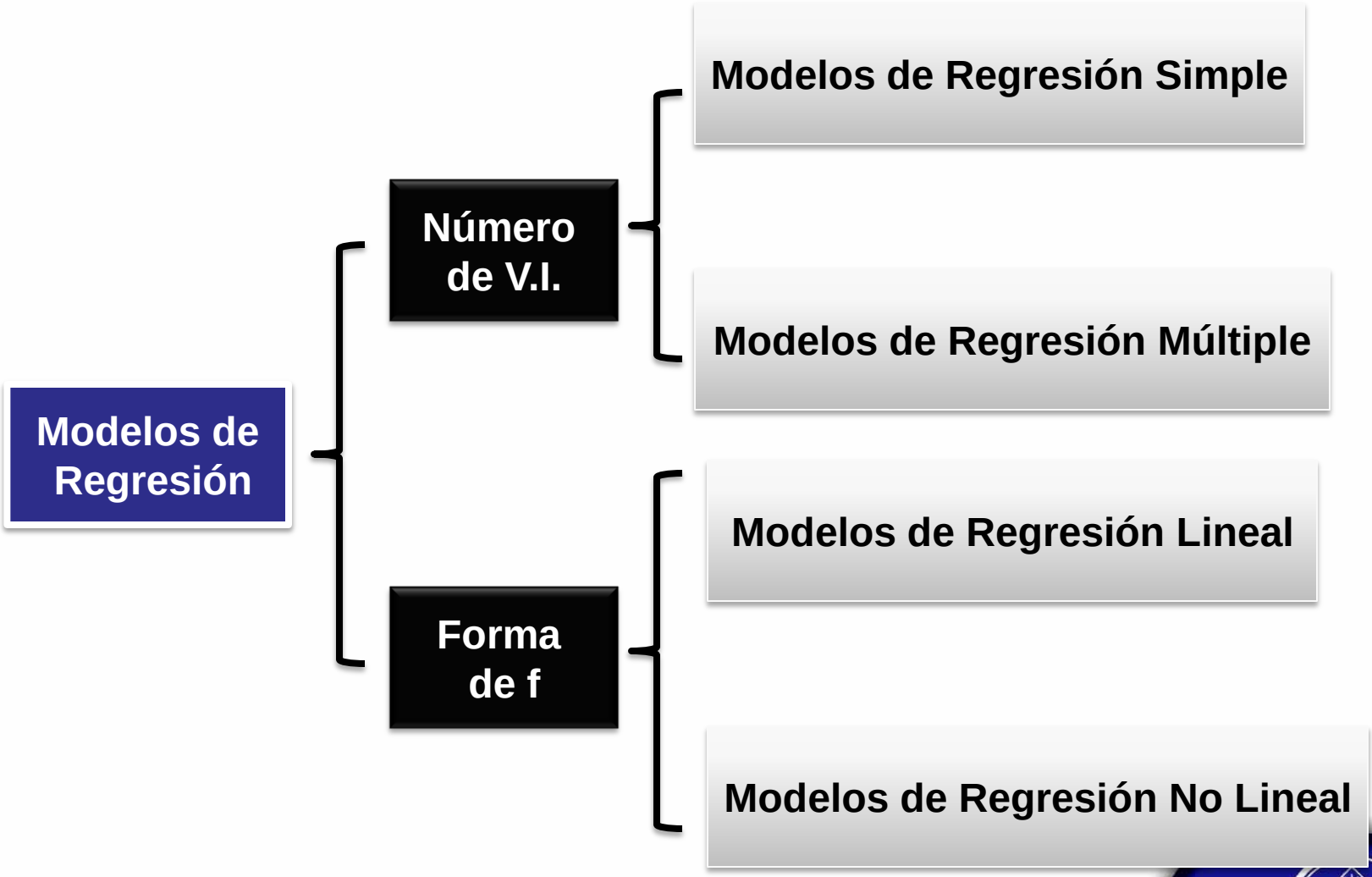
MODELO DE REGRESIÓN

Es un modelo matemático que explica la relación existente entre el conjunto de variables independientes y la variable dependiente.

$$Y = f(X)$$

$$Y = f(X_1, X_2, \ldots, X_n)$$

MODELO DE REGRESIÓN

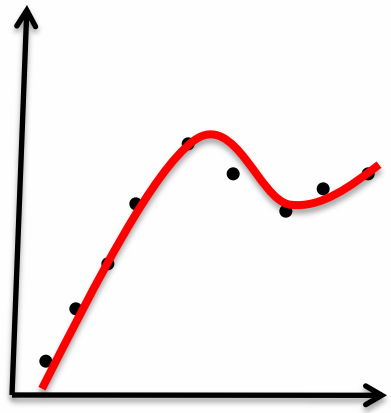


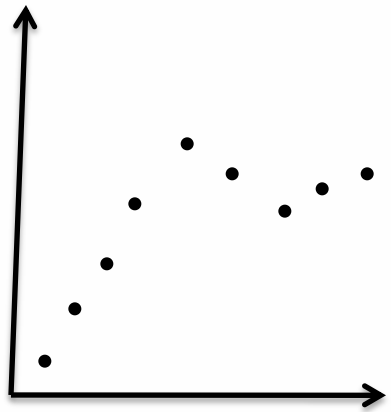
PROPÓSITOS

Describir Datos

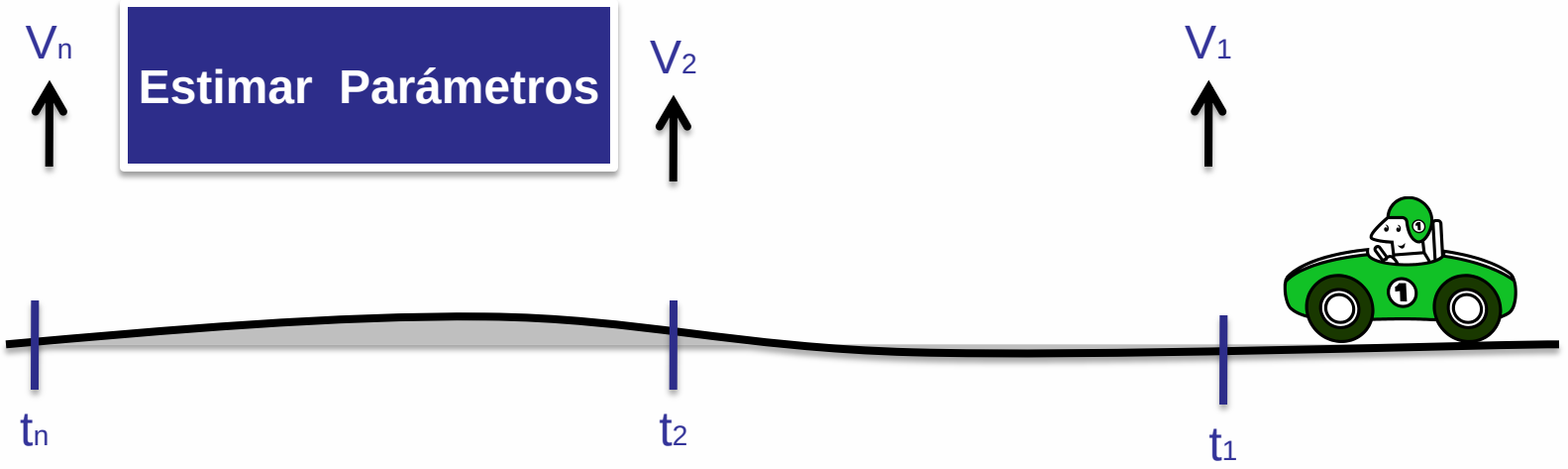
Y	X
y_1	x_1
y_2	x_2
y_3	x_3
y_4	x_4

$$Y = f(X1, X2, \ldots, Xn)$$

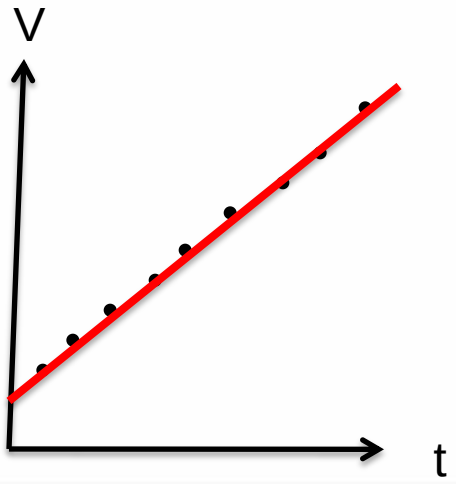




PROPÓSITOS



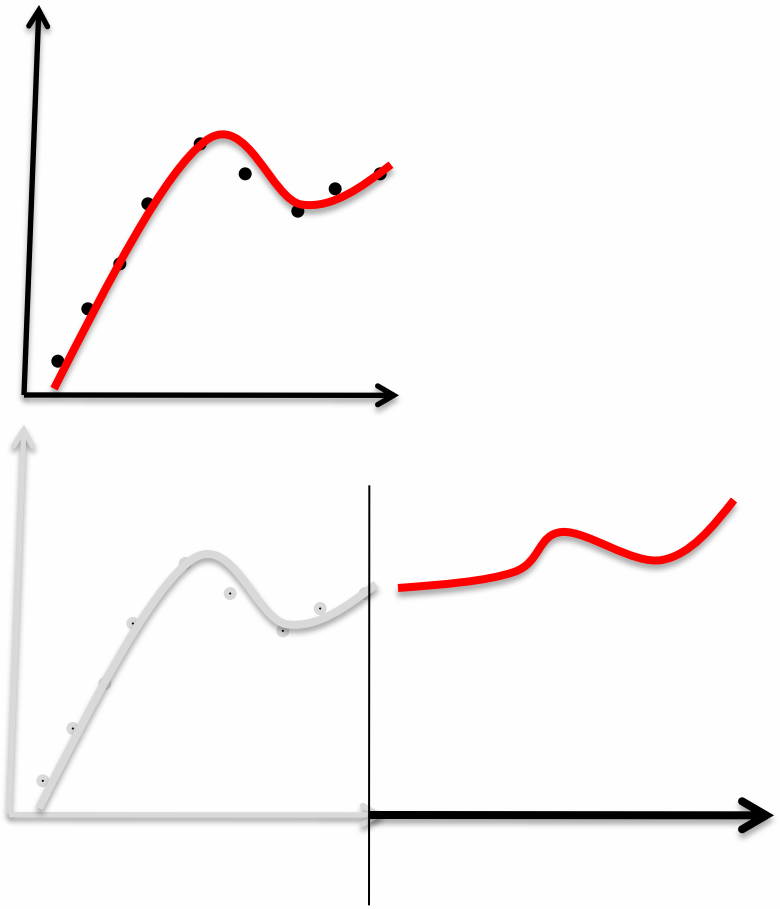
Tiempo	Velocidad
t_1	V_1
t_2	V_2
.	.
.	.
.	.
t_n	V_n

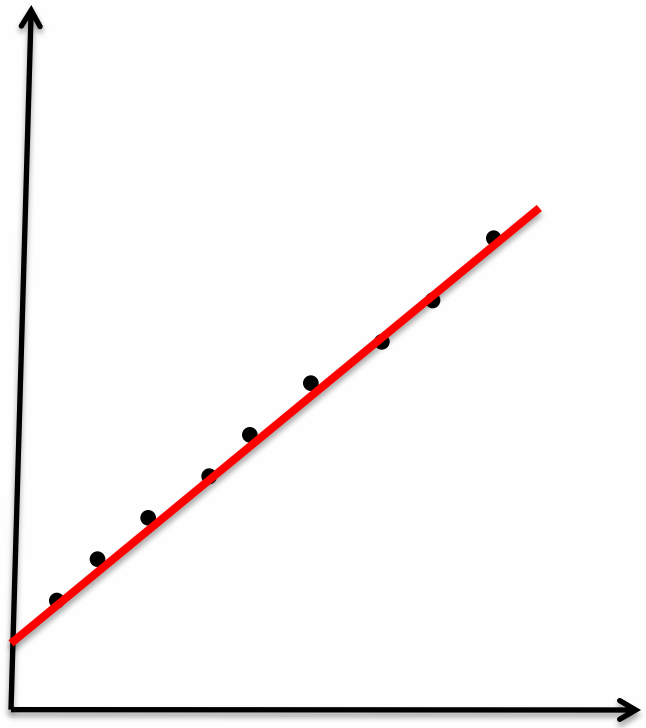


$$Vf = Vo + at$$

PROPÓSITOS

Realizar Pronóstico





REGRESIÓN LINEAL SIMPLE

MODELOS DE REGRESIÓN LINEAL

Variable Independiente ➡ X

Variable Dependiente ➡ Y

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 y β_1: son constantes no conocidas. Se conocen como coeficientes de regresión

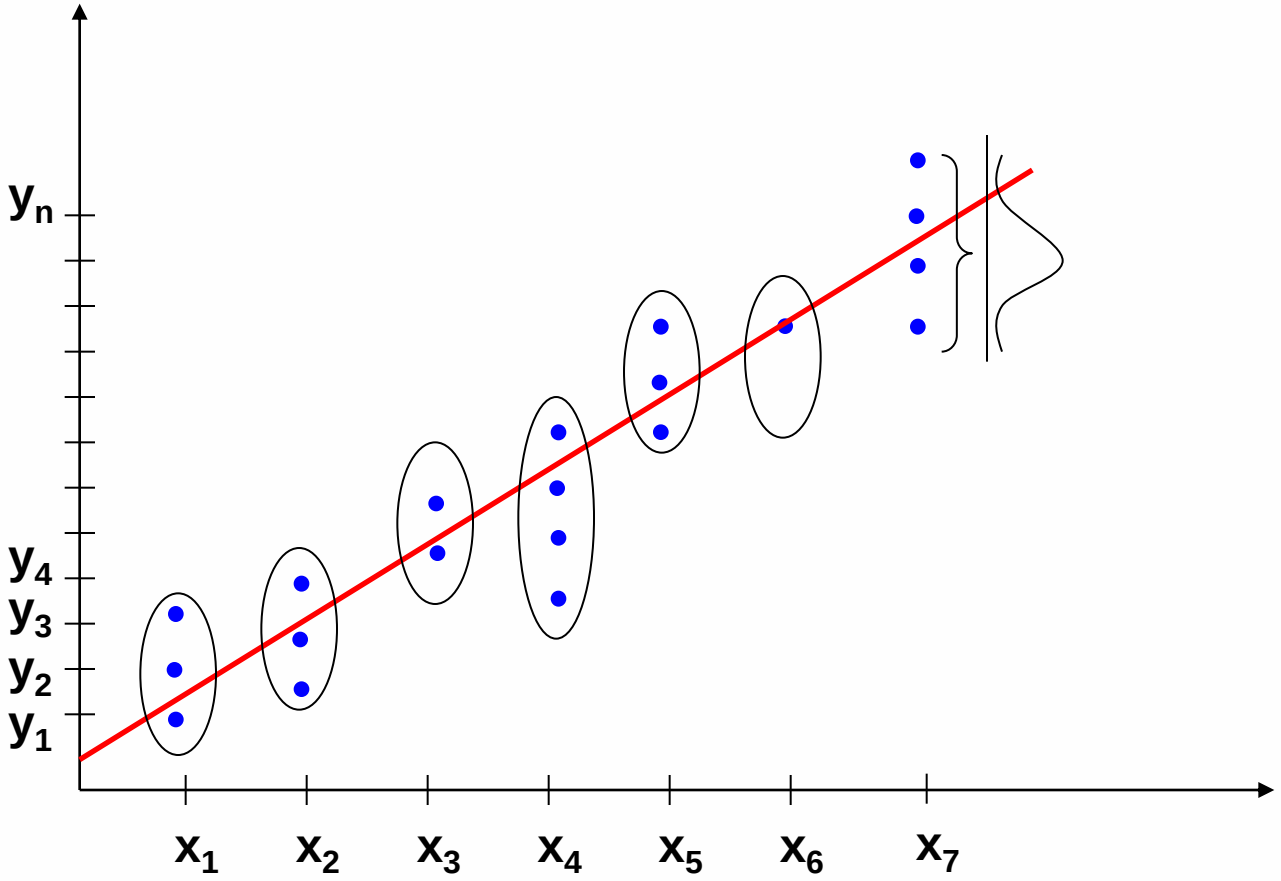
ε Es el componente de error aleatorio.

- Media Cero
- Varianza desconocida σ^2
- Descorrelacionados

X Fijada por el investigador

Y Puede variar para el mismo valor de X

MODELO DE REGRESIÓN LINEAL SIMPLE



MODELO DE REGRESIÓN LINEAL SIMPLE

Media

$$E(Y/x) = \beta_0 + \beta_1 x \longrightarrow$$ **Recta de regresión**

Varianza

$$V(Y/x) = \sigma^2$$

MODELO DE REGRESIÓN LINEAL SIMPLE

β_1

es el cambio en la media de la distribución de **Y** producido por un cambio en una unidad en **X**

β_0

El la media de la distribución de la respuesta **Y** cuando **X=0.** Si el rango de **X** no incluye al cero, entonces no tiene interpretación práctica.

ESTIMADORES MÍNIMOS CUADRADOS

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

MODELO DE REGRESIÓN LINEAL SIMPLE

Recta de regresión muestral

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

ESTIMACIÓN DE σ^2

El estimador de σ^2 se obtiene a partir de la suma de cuadrados del error

$$SS_E = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

Que puede escribirse como

$$SS_E = S_{yy} - \beta_1 S_{xy}$$

Cuyo valor esperado es $E(SS_E) = (n-2)\sigma^2$

Por lo tanto un estimador para σ^2 es

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

PROPIEDADES DE LOS ESTIMADORES

1 Son combinaciones lineales de las observaciones

$$\hat{\beta}_1 = \sum_{i=0}^{n} c_i y_i \qquad \text{Donde} \qquad c_i = \frac{x_i - \bar{x}}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2 Son estimadores insesgados.

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$$

PROPIEDADES DE LOS ESTIMADORES

3 Las varianzas son

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

4 La suma de los residuales es cero $\sum_{i=0}^{n} e_i = \sum_{i=0}^{n} (y_i - \hat{y}_i) = 0$

5 La suma de los valores observados es igual a la suma de los valores ajustados

$$\sum_{i=0}^{n} y_i = \sum_{i=0}^{n} \hat{y}_i$$

PROPIEDADES DE LOS ESTIMADORES

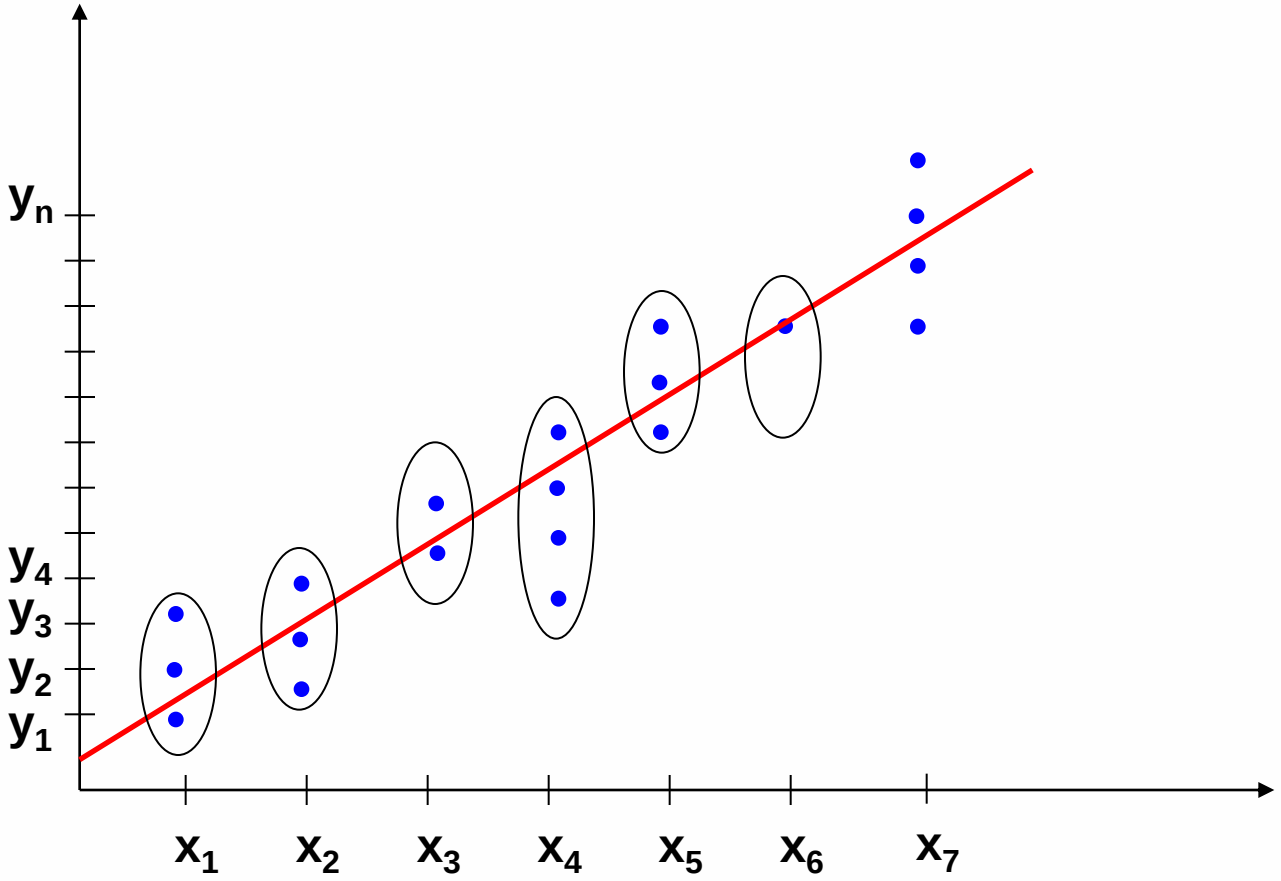
6 La suma de los residuos ponderados por los correspondientes valores de la variable regresora es siempre cero; esto es,

$$\sum_{i=0}^{n} x_i e_i = 0$$

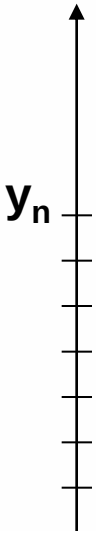
7 La suma de los residuos ponderados por los correspondientes valores ajustados es siempre cero; esto es,

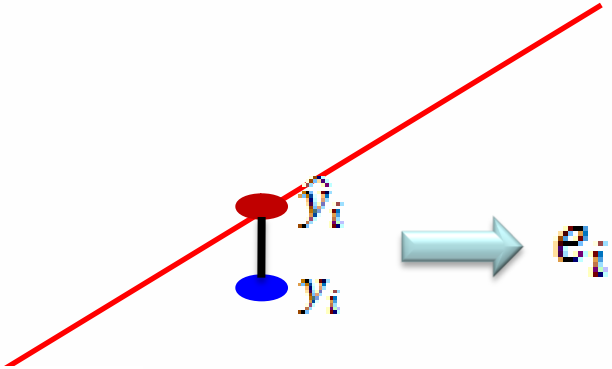
$$\sum_{i=0}^{n} \hat{y}_i e_i = 0$$

PRECISIÓN DE LA REGRESIÓN ESTIMADA

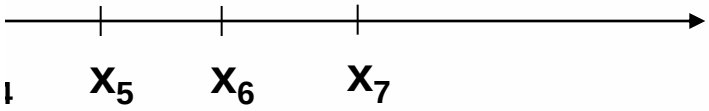


PRECISIÓN DE LA REGRESIÓN ESTIMADA





$$e_i \rightarrow \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$



PRECISIÓN DE LA REGRESIÓN ESTIMADA

$$\sum_{i=0}^{n}(y_i-\hat{y}_i)^2=\sum_{i=0}^{n}(y_i-\bar{y})^2-\sum_{i=0}^{n}(\hat{y}_i-\bar{y})^2$$

$$\sum_{i=0}^{n}(y_i-\bar{y})^2=\sum_{i=0}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=0}^{n}(y_i-\hat{y}_i)^2$$

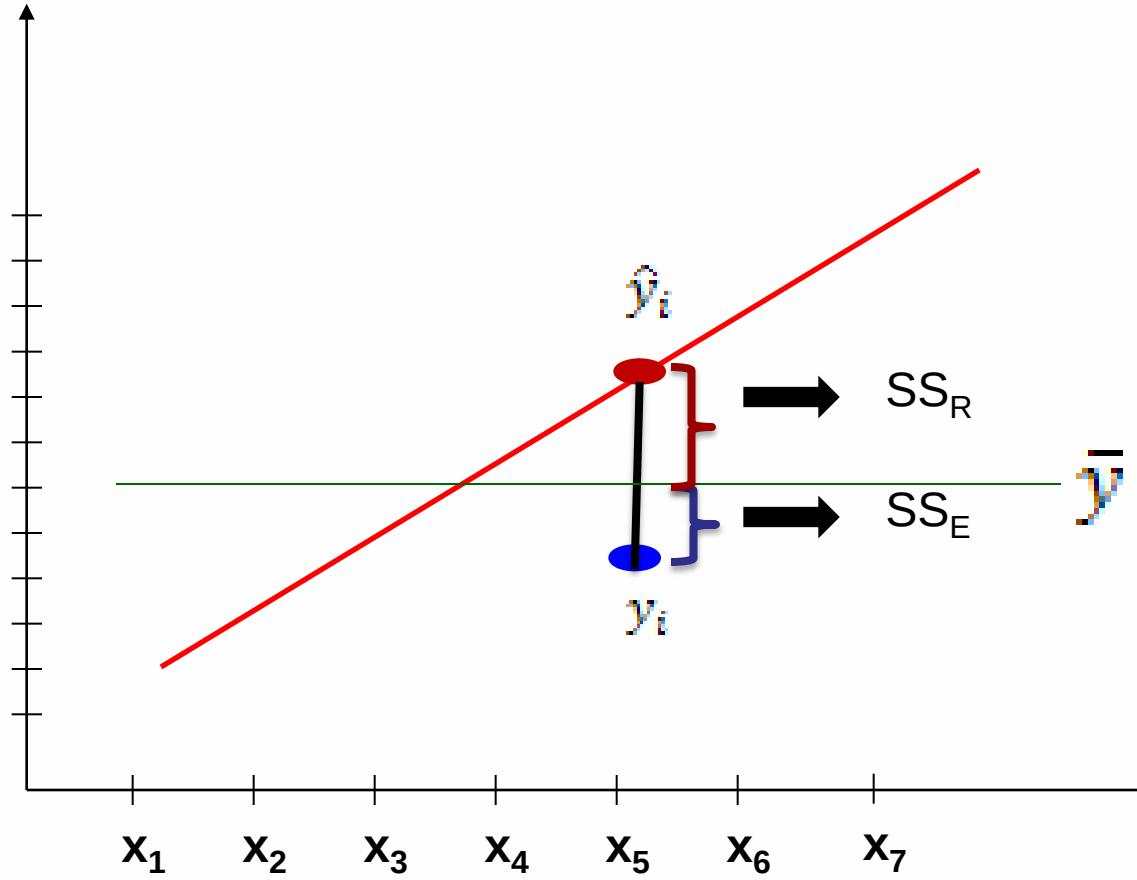
Suma de cuadrados sobre la media

Suma de cuadrados de regresión

Suma de cuadrados del error

$$S_{yy}=SS_R+SS_E$$

PRECISIÓN DE LA REGRESIÓN ESTIMADA



$$SS_R > SS_E$$

$$SS_R \rightarrow S_{yy}$$

$$\frac{SS_R}{S_{yy}} \rightarrow 1$$

$$R^2$$

PRECISIÓN DE LA REGRESIÓN ESTIMADA

El coeficiente de determinación calcula la proporción de la variación explicada por la variable x

- $0 \leq R^2 \leq 1$
- Valores cercanos a 1 implican que gran parte de la variabilidad de Y es explicada por el modelo de regresión.
- La magnitud de R^2 también depende del rango de variabilidad de la variable independiente.

$$E(R^2) = \frac{\beta_1^2 S_{xx}}{\beta_1^2 S_{xx} + \sigma^2}$$

→ Hahn

- En general, R^2 no mide la magnitud de la pendiente de la recta de regresión.

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

- $E(\varepsilon_i)=0$ y $Var(\varepsilon_i)=\sigma^2$

$\Rightarrow$ $E(y_i)=\beta_0 + \beta_1 x_i$

- $Cov(\varepsilon_i , \varepsilon_j)=0$

- $\varepsilon_i \sim N(0, \sigma^2)$

$\Rightarrow$ ε_i y ε_j no sólo están descorrelacionados sino que además son independientes

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_1.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \Rightarrow ee(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}} = \frac{\sigma}{\sqrt{S_{xx}}} \Rightarrow \widehat{ee(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Como

$$\varepsilon_i \sim N(0, \sigma^2) \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_1. | **Prueba de Hipótesis**

$H_0: \beta_1 = \beta_{10}$

$H_1: \beta_1 \neq \beta_{10}$

Como

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \Rightarrow Z = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$ Si H_0 *es cierta*

Debido a que por lo general σ^2 es desconocido usamos su estimador MS_E

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_E}{S_{xx}}}} \sim t_{n-2}$$ **Rechazar si** $|t| > t_{\alpha/2, n-2}$

>>?

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_1. | **Prueba de Hipótesis**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Significancia de regresión.

1. **Usando el estadístico t.**
2. **Usando la tabla de análisis de Varianza**

$SS_R > SS_E$

$$\frac{SS_R}{SS_E} \gg ?$$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Tabla de Análisis de Varianza

Fuente de Variación	SS	G.L.	C.M.	F
Regresión	SS_R	1	CM_R	$\frac{CM_R}{CM_E}$
Error	SS_E	n-2	CM_E	
Total	S_{yy}	n-1		

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

$$F = \frac{SC_R/_1}{SC_E/_{n-2}} \Rightarrow F = \frac{CM_R}{CM_E} \sim F_{1,n-2}$$

$$F > F_{\alpha;1,n-2}$$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_1. | **Intervalo de confianza**

Bajo el supuesto de normalidad

$$\hat{\beta}_1 \pm \left| t_{\frac{\alpha}{2},n-2} \right| \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_0. | **Prueba de Hipótesis**

$H_0: \beta_0 = \beta_{00}$

$H_1: \beta_0 \neq \beta_{00}$

Como

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right) \Rightarrow Z = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$ **Si *H_0 es cierta***

Debido a que por lo general σ^2 es desconocido usamos su estimador MS_E

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$ **Rechazar si** $|t| > t_{\alpha/2, n-2}$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre β_1. | **Intervalo de confianza**

Bajo el supuesto de normalidad

$$\hat{\beta}_0 \pm \left| t_{\frac{\alpha}{2},n-2} \right| \sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

EXAMINANDO LA ECUACIÓN DE REGRESIÓN

Inferencias sobre σ^2. | **Intervalo de confianza**

Bajo el supuesto de normalidad

$$\frac{(n-2)MS_E}{\sigma^2} \sim \chi^2_{n-2}$$

En consecuencia un intervalo de confianza de 100(1 -α)% *para* σ^2 *es*

$$\frac{(n-2)MS_E}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_E}{\chi^2_{1-\alpha/2,n-2}}$$

Predicción

Predicción Media

Sea x_0 un valor de la variable independiente para el cual se desea estimar la respuesta media, es decir **$E(y/x_0)$.**

$$E(\widehat{y/x_0}) = \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_0 x_0$$

Estimador puntual

Como

$$\hat{y}_0 \sim N\left[E(y/x_0), \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right]$$

Predicción

$$Z = \frac{\hat{y}_0 - E(y/x_0)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim N(0,1)$$

$$t = \frac{\hat{y}_0 - E(y/x_0)}{\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

$$\hat{y}_0 \pm \left| t_{\frac{\alpha}{2},n-2} \right| \sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Predicción

Predicción Individual

Sea x_0 un valor de la variable independiente para el cual se desea estimar la respuesta **y.**

$$\hat{y}_0 \pm \left| t_{\frac{\alpha}{2},n-2} \right| \sqrt{MS_E \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Correlación entre X e Y

X
Y

→ Variables aleatorias → Distribución de probabilidad desconocida

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

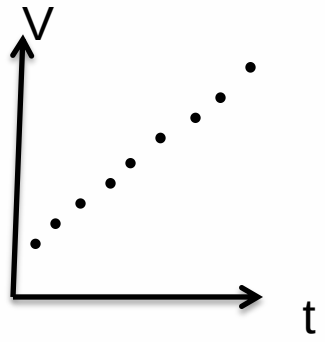
↓

Mide el grado de asociación lineal entre las variables

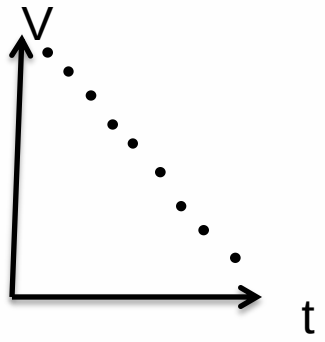
$$-1 \leq \rho_{XY} \leq 1$$

Correlación entre X e Y

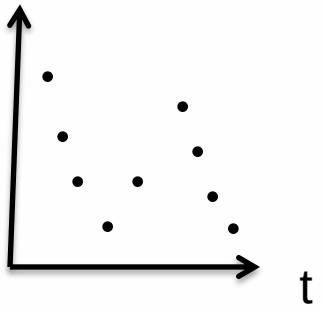
ρ_{XY} → 1 →



ρ_{XY} → -1 →



ρ_{XY} → 0 →



Correlación entre X e Y

Estimación

$$r = \frac{\sum_{i=0}^{n} y_i (x_i - \bar{x})}{\sqrt{\sum_{i=0}^{n} (x_i - \bar{x})^2 \sum_{i=0}^{n} (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Relaciones

$$\hat{\beta}_1 = \sqrt{\frac{S_{yy}}{S_{xx}}}\, r$$

$$r^2 = R^2$$

Correlación entre X e Y

Prueba de Hipótesis

H_0: $\rho = 0$

H_1: $\rho \neq 0$

El estadístico apropiado para esta prueba es $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Rechazar si $$|t| > t_{\alpha/2, n-2}$$