pacorabadan.com

7. Regresión y Correlación

ESTADÍSTICA DESCRIPTIVA

DR. FRANCISCO RABADÁN PÉREZ

Índice

Regresión

- Regresión Tipo I
- Regresión Tipo II
- Regresión Lineal
- Coeficientes de las rectas de regresión

Correlación:

- Correlación General
- Correlación Lineal
- Interpretación de r
- Correlación lineal e independencia estadística

9.1. Regresión

Objetivo de cualquier investigador: Encontrar relaciones entre variables en la forma

$$Y = f(x_1, x_2, \ldots x_n)$$

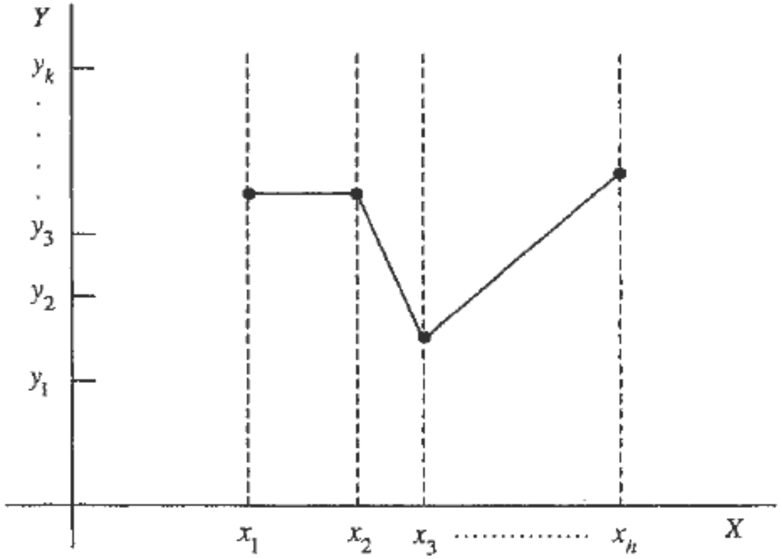
Por dos motivos:

1. Desconocimiento del investigador (análisis exploratorio)
2. Confirmar estructura (análisis confirmatorio)

Enfoque	Teoría
Grado de dependencia (medida)	Teoría de la Correlación
Estructura de dependencia (ajuste)	Teoría de la Regresión

9.1.1. Regresión Tipo I

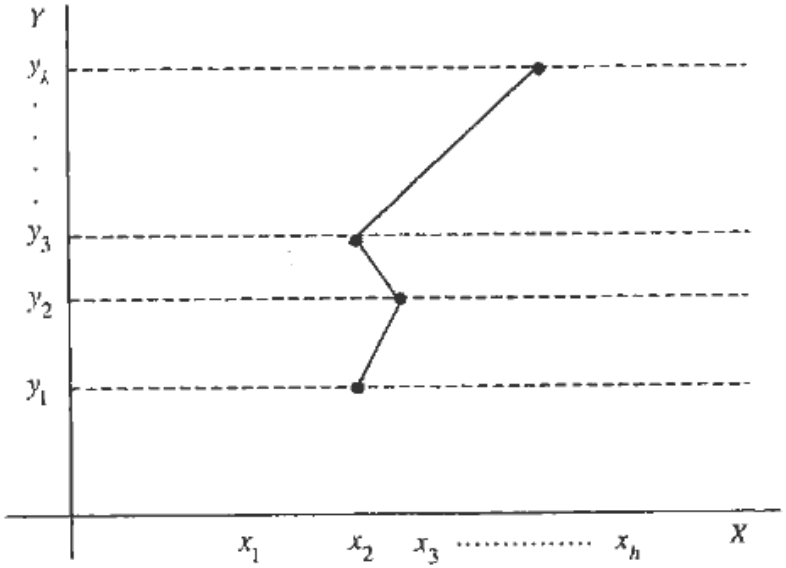
Basada en la media condicionada de una variable respecto a otra.



Regresión tipo I de y/x

La unión de puntos expresa la media de Y condicionada a que X tome el valor x_i.

$$\left(x_i; {}^{\bar{y}}/_{x_i}\right)$$



Regresión tipo I de x/y

La unión de puntos expresa la media de X condicionada a que Y tome el valor y_j

$$\left({}^{\bar{x}}/_{y_i}\,; y_j\right)$$

- Se considera que este método nos proporciona la auténtica regresión intrínseca.
- Si tuviéramos infinitos puntos estaríamos ante funciones continuas.
- En la práctica no es así.

9.1.2. Regresión Tipo II

Basada en el ajuste por Mínimos Cuadrados Ordinarios la función matemática que más se parezca a la nube de puntos.

Regresión II de Y sobre X: $$min\left[\sum_i\sum_j\left(y_j-\bar{y}\right)^2 n_{ij}\right]$$

Regresión II de X sobre Y: $$min\left[\sum_i\sum_j(x_i-\bar{x})^2 n_{ij}\right]$$

- El método de regresión tipo II:
  - se considera un método de aproximación a la regresión tipo I.
  - Nos proporciona una función continua.
  - Es el método más generalizado
  - El grado de ajuste será tanto mejor en la medida en que la curva describa la nube de puntos.
  - Mientras en la regresión tipo I no seleccionamos ningún tipo de curva, en la regresión tipo II es el primer paso.

(Martín-Pliego, 2011; pág. 239)

9.1.3. Regresión Lineal

Basada en el ajuste por Mínimos Cuadrados Ordinarios de una recta a la nube de puntos

$$min\left[\sum_i \sum_j \left(y_j - a - bx_i\right)^2 n_{ij}\right] \Longrightarrow \begin{cases} \sum_j y_j n_{.j} = aN + b\sum_i x_i n_{i.} \\ \sum_j x_i y_j n_{.j} = a\sum_i x_i n_{i.} + b\sum_i x_i^2 n_{i.} \end{cases}$$

Si dividimos por N, el sistema se expresa en función de los momentos; $por\ (-a_{10})$ y restando las dos ecuaciones

$$\left.\begin{array}{r} a_{01} = a + b \cdot a_{10} \\ a_{11} = a \cdot a_{10} + b \cdot a_{20} \end{array}\right\} \rightarrow a_{11} - a_{10} \cdot a_{01} = b(a_{20} - a_{10}^2)$$

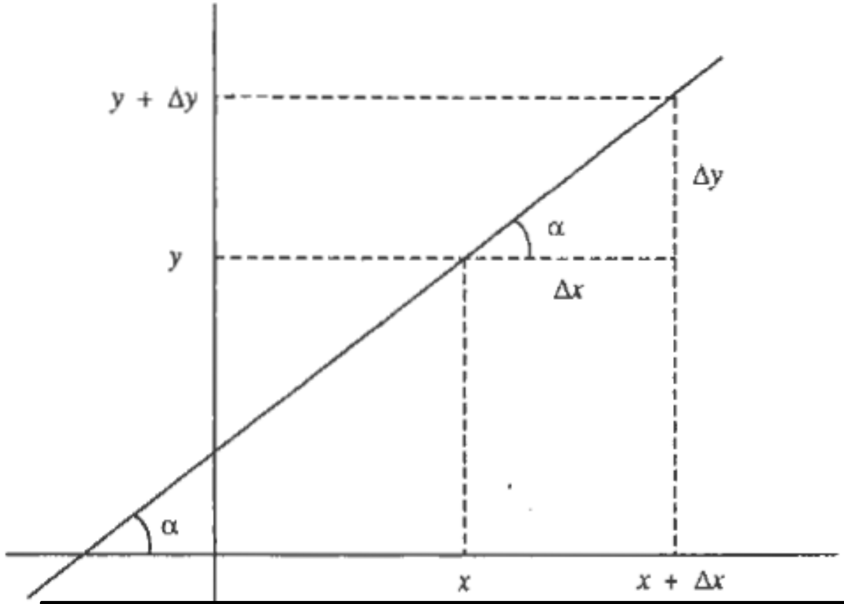
$Despejando$

$$b = \frac{a_{11} - a_{10}a_{01}}{a_{20} - a_{10}^2} = \frac{m_{11}}{m_{20}} = \frac{s_{xy}}{s_x^2}$$

$$a = a_{01} - \frac{s_{xy}}{s_x^2}a_{10} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$$

(Martín-Pliego, 2011; pág. 240,241)

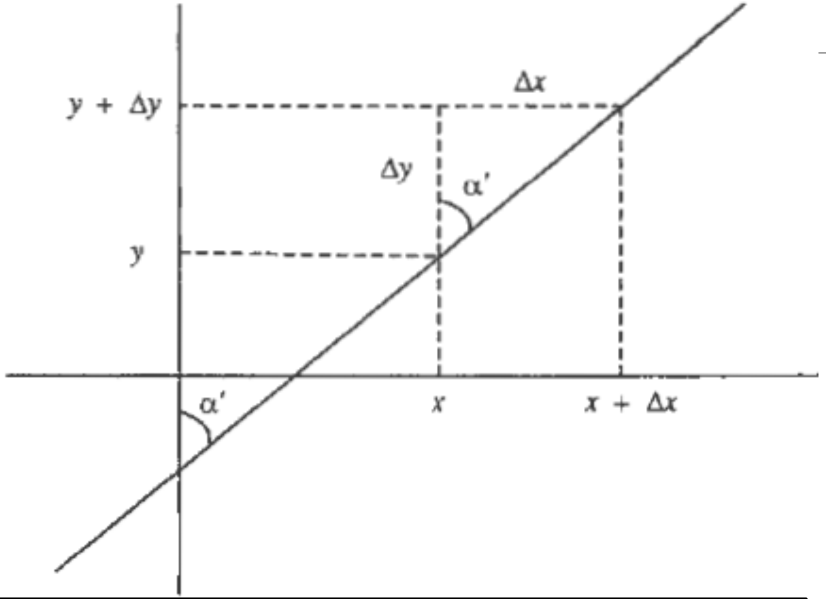
Recta regresión y/x



$$\hat{y}_j = \frac{s_{xy}}{s_x^2} x_i + \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right)$$

$$\hat{y}_j - \bar{y} = \frac{s_{xy}}{s_x^2} (x_i - \bar{x})$$

Recta regresión x/y



$$\hat{x}_i = \frac{s_{xy}}{s_i^2} y_j + \left(\bar{x} - \frac{s_{xy}}{s_y^2} \bar{y} \right)$$

$$\hat{x}_i - \bar{x} = \frac{s_{xy}}{s_y^2} (y_j - \bar{y})$$

9.1.3. Regresión lineal

Hay dos rectas de regresión porque

- La recta de regresión de Y sobre X (Y/X) : hace mínimos los errores cuadráticos al estimar Y si contamos con información de X

$$min(\hat{y}-\bar{y})^2 \text{ ; } X \ conocida$$

- La recta de regresión de X sobre Y: hace mínimos los errores al estimar X si contamos con información de Y

$$min(\hat{x}-\bar{x})^2 \text{ ; } Y \ conocida$$

Es importante detectar cual de las dos tiene sentido económico, o si las dos lo tienen.

Ambas rectas interseccionan en $(\bar{x},\bar{y})$: basta con formar con ellas un sistema de dos ecuaciones con dos incógnitas y despejar.

El producto de las pendientes de ambas rectas coincide con el coeficiente de determinación lineal (r^2) (lo veremos en el 9.2.2)

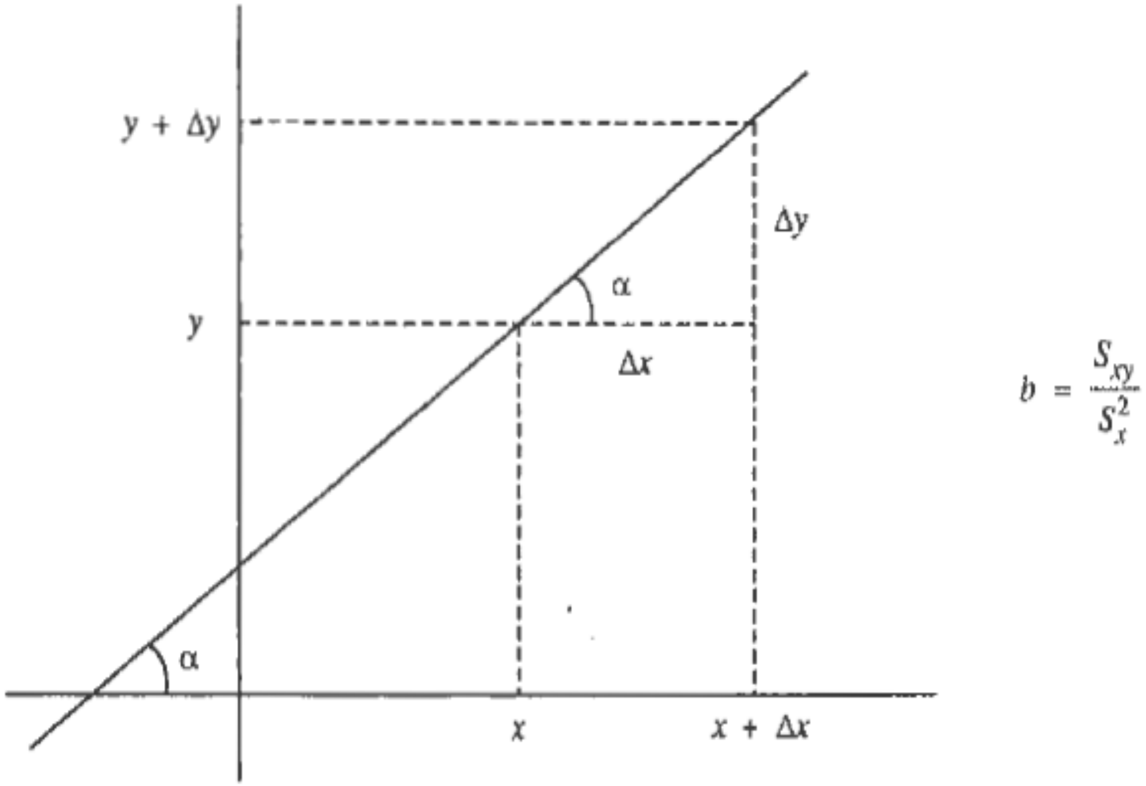
(Martín-Pliego, 2011; pág. 239)

9.1.4. Coeficientes de las rectas de regresión

El coeficiente de la recta de regresión de y/x se interpreta como:

- cuanto varía la variable Y ante una unidad de X
- Como $b = \tan\alpha = \frac{\Delta(y)}{\Delta(x)}$ (ver gráfico 9.1.3), también nos mide la tasa de crecimiento de y para variaciones de X.

La recta de regresión de x/y se interpreta de forma similar.



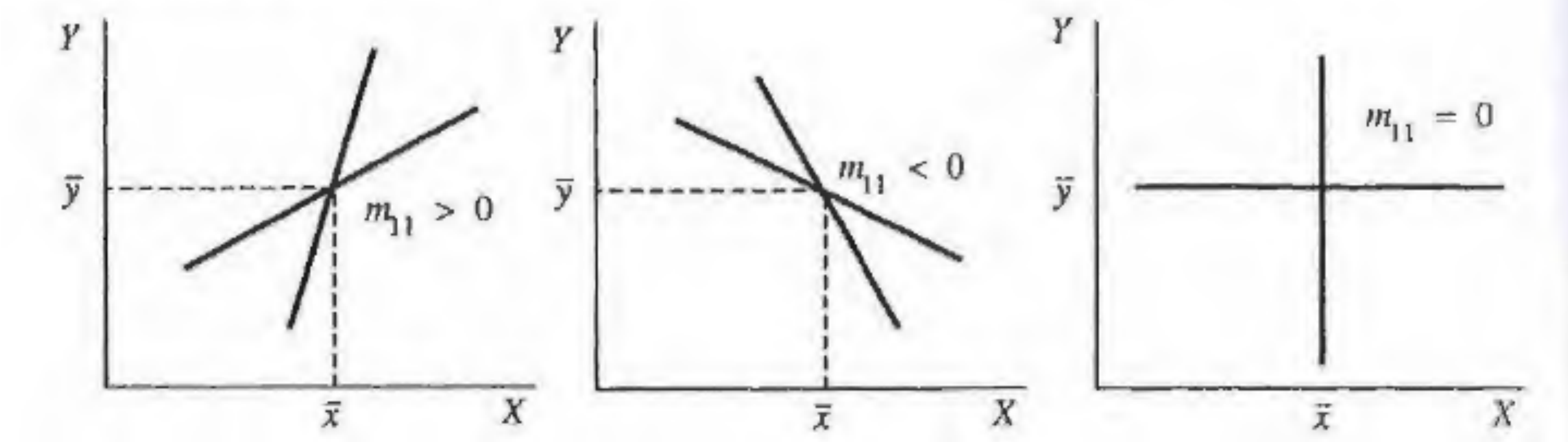
(Martín-Pliego, 2011; pág. 243)

9.1.5. Interpretación de la covarianza

- El signo de las pendientes depende de la covarianza
- **Dimensión: (ud. de X * ud. De Y)**

$$s_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) n_{ij}}{N} = m_{11}$$

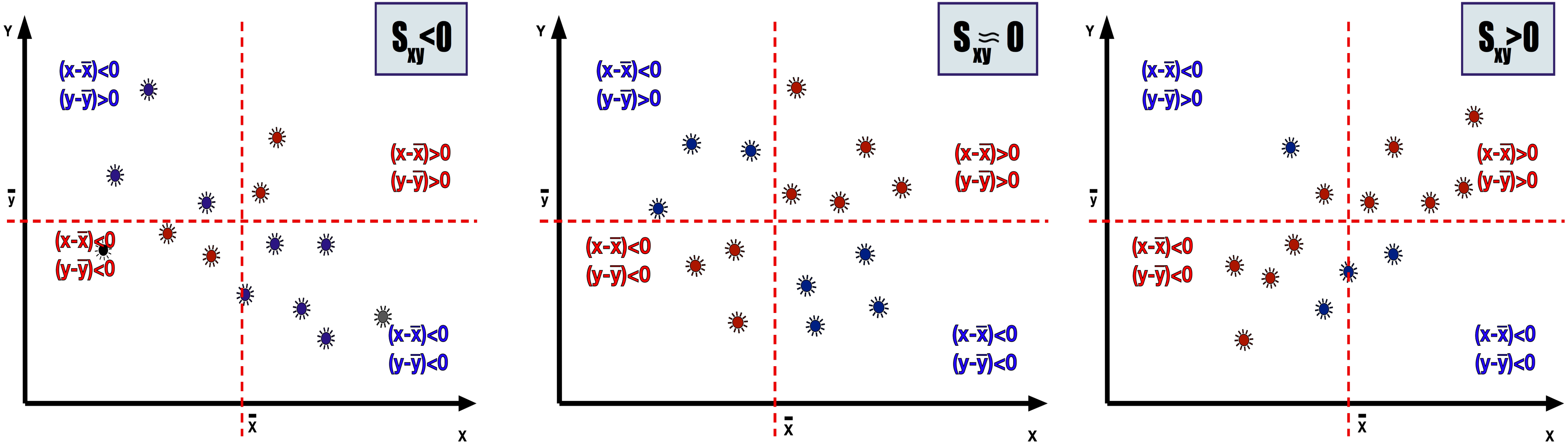
S_{xy}=0 implica independencia lineal, pero <u>no tiene porque haber independencia estadística</u>



(Martín-Pliego, 2011; pág. 244)

9.1.5. Interpretación de la covarianza

$$s_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) n_{ij}}{N} = m_{11}$$



(Martín-Pliego, 2011; pág. 244)

9.2. Correlación

Correlación: Grado de dependencia entre las variables

Varianza Residual: media de todos los residuos al cuadrado (dependerá de la curva de ajuste)

$$S_{ry}^2 = \sum_i \sum_j \left(y_j - \hat{y}_j\right)^2 \frac{n_{ij}}{N} = \sum_j \left(y_j - \hat{y}_j\right)^2 \frac{n_{.j}}{N}$$

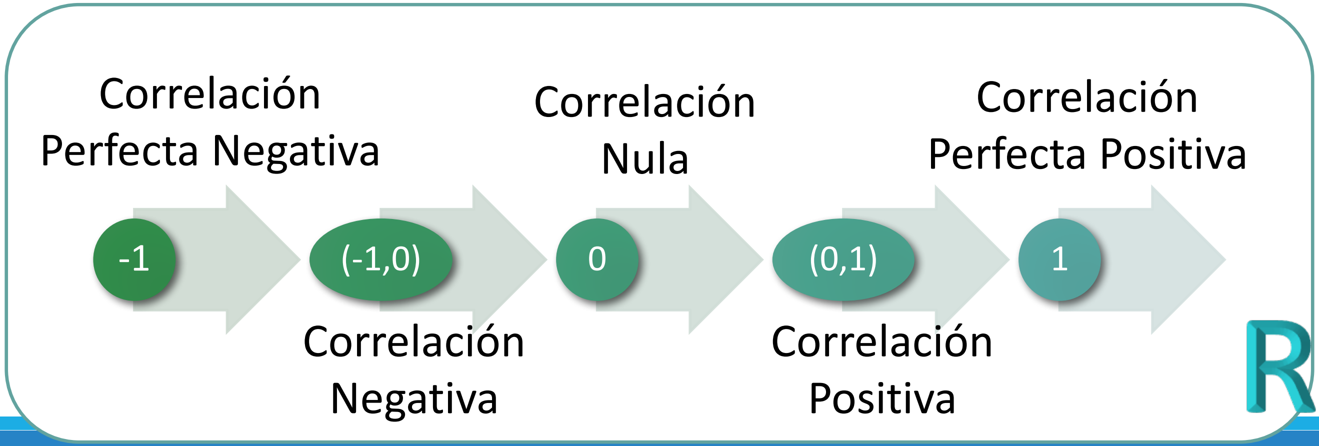
- Cuanto mayor es S_{ry}^2 peor es el ajuste a la curva
- Problema: unidades de medida

- **Coeficiente de Correlación General de Karl Pearson:**

$$R = \sqrt{1 - \frac{S_{ry}^2}{S_y^2}}$$

Grado de dependencia entre las variables; acotado (-1,1) con el signo de la covarianza

De R nos interesa el signo y los extremos

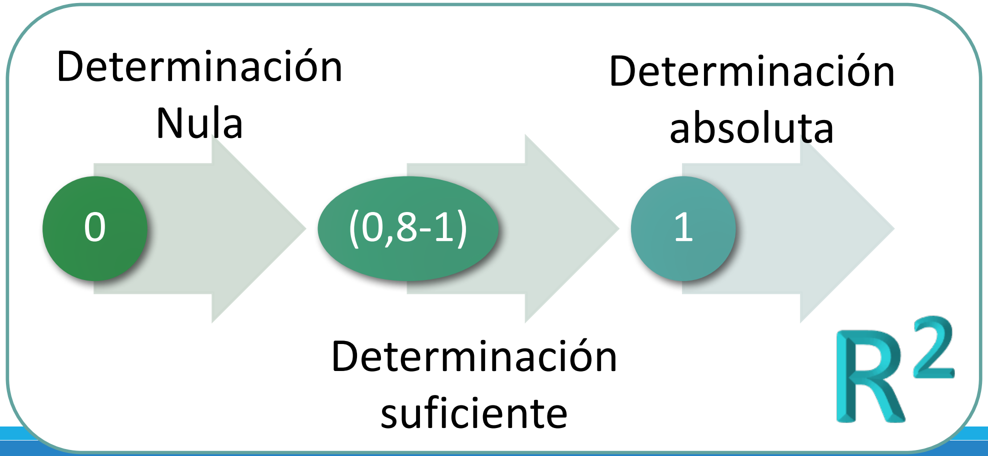


- **Coeficiente de Determinación General de Karl Pearson:**

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2}$$

<u>*Causas comunes entre variables*</u>; acotado (0,1) siempre >0

Auténtica Medida de la Bondad del Ajuste



Coeficiente de Correlación lineal de Karl Pearson:

$$r = \frac{S_{xy}}{S_x S_y}$$

Grado de dependencia entre las variables; acotado (-1,1); el signo depende de la covarianza

De r nos interesa el signo y los extremos

Correlación lineal Perfecta Negativa

-1

(-1,0)

Correlación lineal Negativa

Correlación lineal Nula

0

(0,1)

Correlación lineal Positiva

Correlación lineal Perfecta Positiva

1

r

Coeficiente de Determinación lineal de Karl Pearson:

$$r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

Causas comunes entre variables; acotado (0,1) siempre >0

Auténtica Medida de la Bondad del Ajuste

Determinación lineal Nula

0

(0,8-1)

Determinación lineal suficiente

1

Determinación lineal absoluta

r²

■ Coeficiente de Correlación lineal de Karl Pearson:

$$r = \frac{S_{xy}}{S_x S_y}$$

r=1

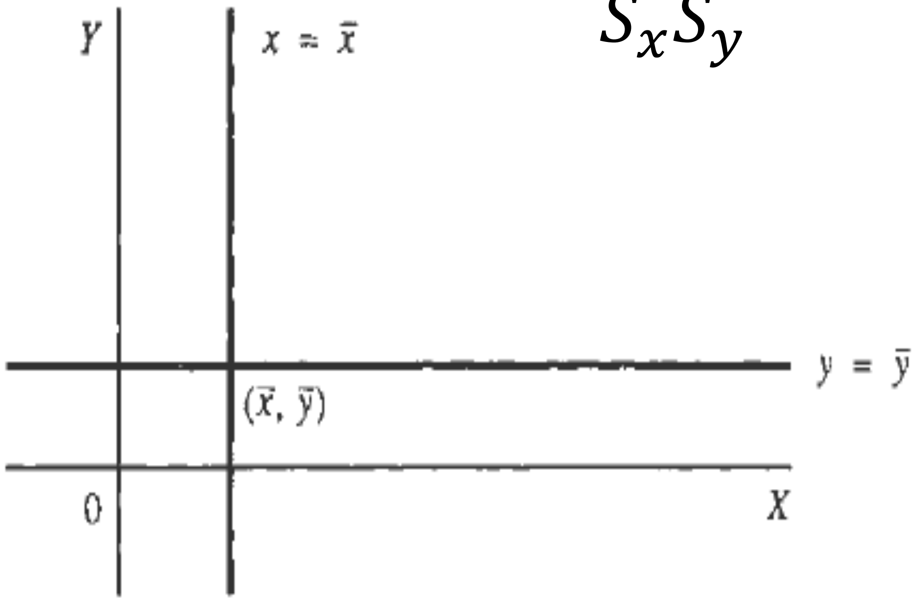
- la varianza residual es cero
- el ajustes es perfecto
- las rectas x/y e y/x coinciden
- A medida que crece X crece Y y viceversa

r=0

- La varianza residual es máxima
- Las rectas X/Y y e Y/X son tangenciales
- Variables ortogonales.

r=-1

- La varianza residual es cero
- El ajuste es perfecto
- las rectas x/y e y/x coinciden
- A medida que crece X decrece Y y viceversa



(Martín-Pliego, 2011; pág. 250)

9.2.4. Correlación Lineal e independencia estadística

- **La independencia lineal (r=0) no implica independencia estadística**
- **El modelo será capaz de explicar r^2 tanto por uno de causas comunes, o de variabilidad total del modelo.**
- **La varianza relativa no explicada por el modelo será $(1-r^2)$ tanto por uno de causas comunes**
- **(9.2.5) r es invariante ante transformaciones lineales (Martín-Pliego, 2011; pág. 252)**

IMPORTANTE

pacorabadan.com

Textos recomendados

- Martín-Pliego, *Introducción a la Estadística Económica y Empresarial*, Editorial AC, 2011, 3ª Edición

Prácticas recomendadas

Lectura 9.3 y 9.4 (Martín-Pliego,2011)

Ejercicios resueltos en clase

Prácticas y recursos web (aula virtual)