

Introducción a la Estadística

Tema 4

Regresión lineal

4.01	Introducción	2
4.02	Rectas de regresión	3
4.03	Varianza residual	6
4.04	Descomposición de la varianza	7
4.05	Coefficiente de determinación	8
4.06	Coefficiente de correlación lineal	8
4.07	Posición relativa de las rectas de regresión	9

4.1 INTRODUCCIÓN

Siendo $(X;Y)$ una variable estadística o población bidimensional, nos plantemos estudiar la **relación de dependencia** entre "X" e "Y".

La relación entre "X" e "Y" puede obedecer a tres motivos:

- 1) Una de las variables influye en la otra. Por ejemplo, del nivel "X" de renta influye en el gasto "Y" en alimentación.
- 2) Una tercera variable influye en "X" e "Y". Por ejemplo, el número "X" de teléfonos móviles en un país y el número "Y" de coches que se venden en ese país pueden mostrar cierta relación debido a la influencia que en "X" e "Y" tiene la renta disponible "Z".
- 3) El azar. Por ejemplo, podemos encontrar cierta relación entre el número "X" de matrimonios en Moldavia en los últimos 124 meses y el número "Y" de vacas locas en Inglaterra en el mismo periodo de tiempo pero es evidente que no tiene sentido plantearse estudiar la relación entre "X" e "Y".

Dependencia funcional

Diremos que la variable "Y" depende funcionalmente de la variable "X" si podemos establecer una función $f: \mathcal{R} \mapsto \mathcal{R}$ que a cada valor de "X" le asocia un valor de "Y". Por ejemplo, siendo "X" la velocidad de un móvil e "Y" el espacio que recorre en un tiempo dado "t", sucede que $Y = f(X) = X \cdot t$.

Dependencia estadística

Es claro que hay variables (como peso y estatura, renta y ahorro, horas de estudio y calificaciones en los exámenes, etc.) que están relacionadas, pero la relación entre ellas no puede expresarse mediante una función "F". De este tipo de relación, que no puede expresarse de modo funcional, se dice que es una relación estadística.

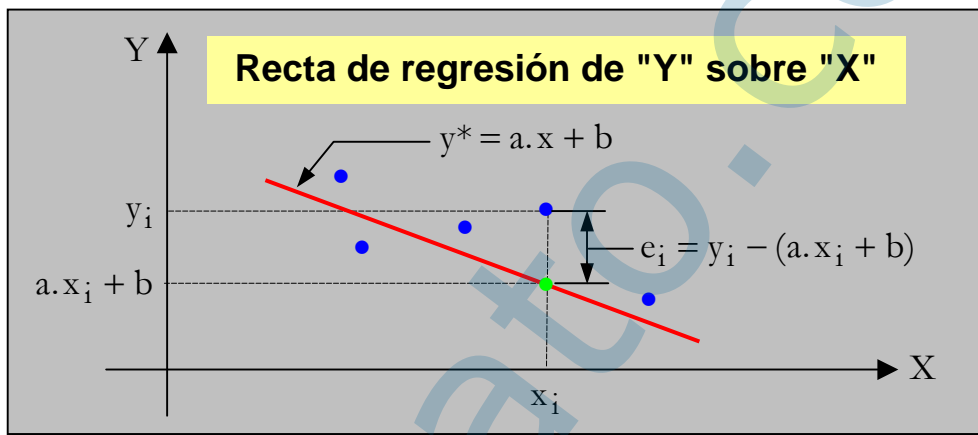
4.2 RECTAS DE REGRESIÓN

Sea $(X;Y)$ una variable estadística o población bidimensional representada por "n" pares $(x_i; y_i)$ de frecuencia absoluta unitaria.

Regresión de "X" sobre "Y": pretende **explicar** el comportamiento de "Y" (variable dependiente) en función del comportamiento de "X" (variable independiente) mediante una relación lineal $y^* = a \cdot x + b$.

Determinaremos "a" y "b" mediante el **método de los mínimos cuadrados**; o sea, "a" y "b" son los que minimizan la suma "S" de los cuadrados de los **residuos** $e_i = y_i - (a \cdot x_i + b)$; o sea, minimizan

$$S = \sum_{i=1}^n (y_i - (a \cdot x_i + b))^2$$



Como veremos en la página siguiente, resulta ser:

$$a = S_{XY} / S_X^2 \quad ; \quad b = \bar{y} - a \cdot \bar{x}$$

Así, la llamada **recta de regresión de "Y" sobre "X"** es:

$$y^* = a \cdot x + b \Rightarrow y^* = \frac{S_{XY}}{S_X^2} \cdot x + \left(\bar{y} - \bar{x} \cdot \frac{S_{XY}}{S_X^2} \right) \Rightarrow$$

$$a = S_{XY} / S_X^2 \quad ; \quad b = \bar{y} - a \cdot \bar{x} = \bar{y} - \bar{x} \cdot S_{XY} / S_X^2$$

$$\Rightarrow y^* - \bar{y} = \frac{S_{XY}}{S_X^2} \cdot (x - \bar{x})$$

Observa: la recta de regresión pasa por el punto $(\bar{x}; \bar{y})$.

De $a = S_{XY} / S_X^2$, que es la **pendiente** de la recta de regresión, se dice que es el **coeficiente de regresión** de "Y" sobre "X", e indica la variación que sufre "Y" cuando "X" varía una unidad, supuesto cierta la relación lineal $y^* = a \cdot x + b$.

Como $S_X^2 > 0$, el signo de "a" es el de la covarianza S_{XY} ; por tanto, si la covarianza es positiva (negativa), la recta de regresión tiene pendiente positiva (negativa); o sea, es ascendente (descendente).

$$S = \sum_{i=1}^n (y_i - (a \cdot x_i + b))^2 \text{ es m\u00ednimo si } a = \frac{S_{XY}}{S_X^2} \text{ y } b = \bar{y} - a \cdot \bar{x}$$

$$\nabla S = \bar{0} \Rightarrow \left\{ \begin{array}{l} \frac{\partial S}{\partial a} = 2 \cdot \sum_{i=1}^n (y_i - (a \cdot x_i + b)) \cdot (-x_i) = 0 \\ \frac{\partial S}{\partial b} = 2 \cdot \sum_{i=1}^n (y_i - (a \cdot x_i + b)) \cdot (-1) = 0 \end{array} \right\} \Rightarrow$$

Ecuaciones normales

$$\Rightarrow \left\{ \begin{array}{l} \sum_{i=1}^n x_i \cdot y_i - a \cdot \sum_{i=1}^n x_i^2 - b \cdot \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n y_i - a \cdot \sum_{i=1}^n x_i - n \cdot b = 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} a \cdot \sum_{i=1}^n x_i^2 + b \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n x_i \cdot y_i \\ a \cdot \sum_{i=1}^n x_i + n \cdot b = \sum_{i=1}^n y_i \end{array} \right.$$

De la 2^a resulta $b = \frac{1}{n} \cdot \left(\sum_{i=1}^n y_i - a \cdot \sum_{i=1}^n x_i \right) = \bar{y} - a \cdot \bar{x}$, que sustituido en la 1^a:

$$\begin{aligned} a \cdot \sum_{i=1}^n x_i^2 + (\bar{y} - a \cdot \bar{x}) \cdot \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i \cdot y_i \Rightarrow \\ \Rightarrow a \cdot \left(\sum_{i=1}^n x_i^2 - \bar{x} \cdot \sum_{i=1}^n x_i \right) &= \sum_{i=1}^n x_i \cdot y_i - \bar{y} \cdot \sum_{i=1}^n x_i \Rightarrow \\ \Rightarrow a = \frac{\sum_{i=1}^n x_i \cdot y_i - \bar{y} \cdot \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \cdot \sum_{i=1}^n x_i} &= \frac{\frac{1}{n} \cdot \sum_{i=1}^n x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - (\bar{x})^2} = \frac{S_{XY}}{S_X^2} \end{aligned}$$

dividiendo numerados y denominador por "n"

Es un m\u00ednimo, pues a matriz hessiana de "S" es definida positiva en todo punto:

$$H(S) = \begin{bmatrix} \frac{\partial^2 S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial b \partial a} & \frac{\partial^2 S}{\partial b^2} \end{bmatrix} = \begin{bmatrix} 2 \cdot \sum_{i=1}^n x_i^2 & 2 \cdot \sum_{i=1}^n x_i \\ 2 \cdot \sum_{i=1}^n x_i & 2 \cdot n \end{bmatrix}$$

ya que

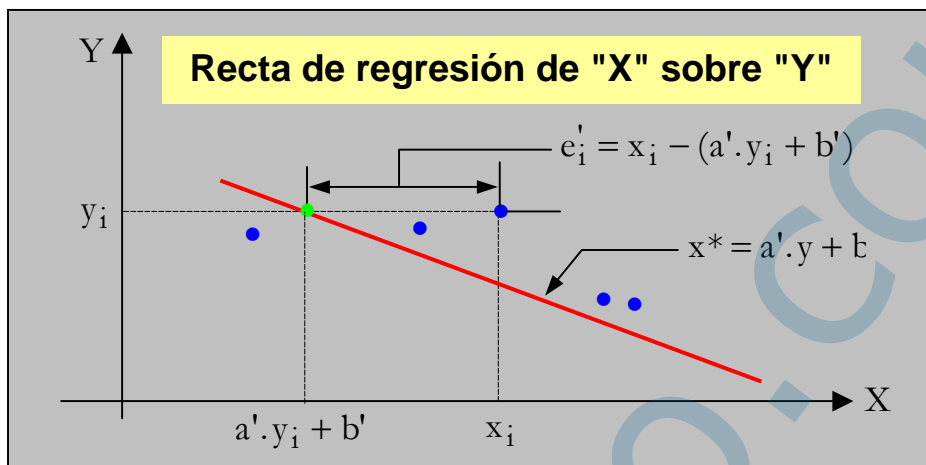
$$H_1 = 2 \cdot \sum_{i=1}^n x_i^2 > 0$$

$$H_2 = |H(S)| = 4 \cdot n \cdot \sum_{i=1}^n x_i^2 - 4 \cdot \left(\sum_{i=1}^n x_i \right)^2 = 4 \cdot n^2 \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - (\bar{x})^2 \right) = 4 \cdot n^2 \cdot S_X^2 > 0$$

Regresión de "X" sobre "Y": pretende **explicar** el comportamiento de "X" (variable dependiente) en función del comportamiento de "Y" (variable independiente) mediante una relación lineal $x^* = a' \cdot y + b'$.

Determinamos a' y b' mediante el **método de los mínimos cuadrados**; o sea, a' y b' son los que minimizan la suma S' de los cuadrados de los **residuos** $e'_i = x_i - (a' \cdot y_i + b')$; o sea, minimizan

$$S' = \sum_{i=1}^n (x_i - (a' \cdot y_i + b'))^2$$



Resulta ser:

$$a' = S_{XY} / S_Y^2 ; b' = \bar{x} - a' \cdot \bar{y}$$

Así, la llamada **recta de regresión de "Y" sobre "X"** es:

$$x^* = a' \cdot y + b' \Rightarrow x^* = \frac{S_{XY}}{S_Y^2} \cdot y + \left(\bar{x} - \bar{y} \cdot \frac{S_{XY}}{S_Y^2} \right) \Rightarrow$$

$$a' = S_{XY} / S_Y^2 ; b' = \bar{x} - a' \cdot \bar{y} = \bar{x} - \bar{y} \cdot S_{XY} / S_Y^2$$

$$\Rightarrow x^* - \bar{x} = \frac{S_{XY}}{S_Y^2} \cdot (y - \bar{y})$$

Observa: la recta de regresión pasa por el punto $(\bar{x}; \bar{y})$.

De $a' = S_{XY} / S_Y^2$, que, con los ejes dibujados, es la inversa de la **pendiente** de la recta de regresión de "X" sobre "Y", se dice que es el **coeficiente de regresión** de "X" sobre "Y", e indica la variación que sufre "X" cuando "Y" varía una unidad, supuesto cierta la relación lineal $x = a' \cdot y + b'$.

Como $S_Y^2 > 0$, el signo de a' es el de la covarianza S_{XY} ; por tanto, si la covarianza es positiva (negativa), la recta de regresión tiene pendiente positiva; o sea, es ascendente (descendente).

4.3 VARIANZA RESIDUAL

La recta de regresión de "Y" sobre "X", es la que, según el método de los mínimos cuadrados, mejor representa o explica la relación entre la variable dependiente "Y" y la independiente "X", pues para esa resta se minimiza la suma $S = \sum (y_i - (a \cdot x_i + b))^2$ de los cuadrados de los residuos $e_i = y_i - (a \cdot x_i + b)$.

Conocidos $a = S_{XY}/S_X^2$ y $b = \bar{y} - a \cdot \bar{x}$, queda determinada la **distribución de residuos** $e_i = y_i - (a \cdot x_i + b)$, que tiene media 0 y varianza S/n :

$$\begin{aligned} \bar{e} &= \frac{1}{n} \cdot \sum e_i = \frac{1}{n} \cdot \sum (y_i - (a \cdot x_i + b)) = \\ &= \left(\frac{1}{n} \cdot \sum y_i \right) - a \cdot \left(\frac{1}{n} \cdot \sum x_i \right) - b = \bar{y} - a \cdot \bar{x} - b = 0 \\ &\quad \boxed{b = \bar{y} - a \cdot \bar{x}} \end{aligned}$$

$$S_e^2 = \frac{1}{n} \cdot \sum (e_i - \bar{e})^2 = \frac{1}{n} \cdot \sum e_i^2 = \frac{1}{n} \cdot \sum (y_i - (a \cdot x_i + b))^2 = \frac{S}{n}$$

$$\boxed{\bar{e} = 0} \quad \boxed{e_i = y_i - (a \cdot x_i + b)}$$

De $S_e^2 = S/n$ se dice que es la **varianza residual**, y es una buena medida de la bondad o calidad del ajuste lineal $y^* = a \cdot x + b$ realizado.

Observa: es $S_e^2 = 0$ si todos los residuos $e_i = y_i - (a \cdot x_i + b)$ son nulos; o sea, si $y_i = a \cdot x_i + b$ para todo valor de "i", es decir, cada valor observado y_i de "Y" coincide con el valor teórico $a \cdot x_i + b$ que proporciona la recta de regresión (por tanto, la relación entre "X" e "Y" es de dependencia funcional). Cuanto mayor sea S_e^2 , menor es la representatividad de la recta de regresión.

¡Chollo! podemos determinar la varianza residual $S_e^2 = S/n$ sin calcular "S":

$$S_e^2 = \frac{1}{n} \cdot \sum e_i^2 = \frac{1}{n} \cdot \sum (y_i - (a \cdot x_i + b))^2 =$$

$$\boxed{\text{es } a = \frac{S_{XY}}{S_X^2} \text{ y } b = \bar{y} - a \cdot \bar{x} = \bar{y} - \frac{S_{XY}}{S_X^2} \cdot \bar{x}}$$

$$= \frac{1}{n} \cdot \sum \left(y_i - \frac{S_{XY}}{S_X^2} \cdot x_i - \bar{y} + \frac{S_{XY}}{S_X^2} \cdot \bar{x} \right)^2 = \frac{1}{n} \cdot \sum \left((y_i - \bar{y}) - \frac{S_{XY}}{S_X^2} \cdot (x_i - \bar{x}) \right)^2 =$$

$$= \frac{1}{n} \cdot \sum \left((y_i - \bar{y})^2 + \frac{S_{XY}^2}{S_X^4} \cdot (x_i - \bar{x})^2 - 2 \cdot \frac{S_{XY}}{S_X^2} \cdot (y_i - \bar{y}) \cdot (x_i - \bar{x}) \right) =$$

$$= \underbrace{\left(\frac{1}{n} \cdot \sum (y_i - \bar{y})^2 \right)}_{S_Y^2} + \frac{S_{XY}^2}{S_X^4} \cdot \underbrace{\left(\frac{1}{n} \cdot \sum (x_i - \bar{x})^2 \right)}_{S_X^2} - 2 \cdot \frac{S_{XY}}{S_X^2} \cdot \underbrace{\left(\frac{1}{n} \cdot \sum (y_i - \bar{y}) \cdot (x_i - \bar{x}) \right)}_{S_{XY}} =$$

$$= S_Y^2 + \frac{S_{XY}^2}{S_X^4} \cdot S_X^2 - 2 \cdot \frac{S_{XY}}{S_X^2} \cdot S_{XY} = S_Y^2 - \frac{S_{XY}^2}{S_X^2}$$

4.4 DESCOMPOSICIÓN DE LA VARIANZA

Refiriéndonos a la regresión de "Y" sobre "X", la varianza S_Y^2 de la distribución de valores observados y_i de "Y", la varianza $S_{Y^*}^2$ de la distribución de valores teóricos $y_i^* = a \cdot x_i + b$ de "Y" y la varianza S_e^2 de la distribución de residuos $e_i = y_i - y_i^* = y_i - (a \cdot x_i + b)$ son tales que:

$$S_Y^2 = S_{Y^*}^2 + S_e^2$$

Es:
$$S_{Y^*}^2 = \frac{1}{n} \cdot \sum (y_i^* - \bar{y}^*)^2 = \frac{1}{n} \cdot \sum (y_i^* - \bar{y})^2$$

$$y_i^* = a \cdot x_i + b \Rightarrow \bar{y}^* = a \cdot \bar{x} + b = a \cdot \bar{x} + (\bar{y} - a \cdot \bar{x}) = \bar{y}$$

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

Es:
$$S_Y^2 = \frac{1}{n} \cdot \sum (y_i - \bar{y})^2 = \frac{1}{n} \cdot \sum (y_i^* - \bar{y} + e_i)^2 = \frac{1}{n} \cdot \sum ((y_i^* - \bar{y}) + e_i)^2 =$$

$$y_i = y_i^* + e_i \Rightarrow y_i - \bar{y} = y_i^* - \bar{y} + e_i$$

$$= \left(\frac{1}{n} \cdot \sum (y_i^* - \bar{y})^2 \right) + \left(\frac{1}{n} \cdot \sum e_i^2 \right) + 2 \cdot \left(\frac{1}{n} \cdot \sum (y_i^* - \bar{y}) \cdot e_i \right) =$$

$$\sum (y_i^* - \bar{y}) \cdot e_i = \sum y_i^* \cdot e_i - \bar{y} \cdot \sum e_i = \sum y_i^* \cdot e_i = \sum (a \cdot x_i + b) \cdot e_i =$$

$$\sum e_i = 0$$

$$= a \cdot \sum x_i \cdot e_i + b \cdot \sum e_i = a \cdot \sum x_i \cdot e_i = 0$$

según la 1ª de las ecuaciones normales

$$= \underbrace{\left(\frac{1}{n} \cdot \sum (y_i^* - \bar{y})^2 \right)}_{S_{Y^*}^2} + \underbrace{\left(\frac{1}{n} \cdot \sum e_i^2 \right)}_{S_e^2} = S_{Y^*}^2 + S_e^2$$

De $S_{Y^*}^2$ se dice que es la **varianza explicada por la regresión**, y expresa la variabilidad de "Y" debida a la supuesta relación lineal entre "Y" y "X" y la **varianza residual** S_e^2 expresa la variabilidad de "Y" que se debe a causas ajenas a la relación lineal entre "Y" y "X".

Es $S_Y^2 = S_{Y^*}^2$ si $S_e^2 = 0$; o sea, si todos los residuos $e_i = y_i - (a \cdot x_i + b)$ son nulos, es decir, para todo valor de "i" sucede que el valor observado y_i de "Y" coincide con el valor teórico $a \cdot x_i + b$ que proporciona la recta de regresión.

A la hora de lo cotidiano, si conocemos "a" y S_X^2 , calcularemos $S_{Y^*}^2$ teniendo en cuenta que como $y_i^* = a \cdot x_i + b$, es $S_{Y^*}^2 = a^2 \cdot S_X^2$.

4.5 COEFICIENTE DE DETERMINACIÓN

El **coeficiente de determinación** R^2 es el cociente entre la varianza $S_{Y^*}^2$ explicada por la regresión y la varianza S_Y^2 de "Y", y puede emplearse como medida de la bondad o calidad del ajuste $y^* = a \cdot x + b$ realizado.

$$R^2 = \frac{S_{Y^*}^2}{S_Y^2} = \frac{S_Y^2 - S_e^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2}$$

$$\boxed{S_Y^2 = S_{Y^*}^2 + S_e^2 \Rightarrow S_{Y^*}^2 = S_Y^2 - S_e^2}$$

El valor de R^2 está en $[0;1]$, pues R^2 es un cociente de varianzas (números no negativos) y el numerador no es superior al denominador. Si $R^2 = 1$ ($\Leftrightarrow S_e^2 = 0$), la relación entre "X" e "Y" es de dependencia funcional, como quedó dicho. Si $R^2 \leq 0.75$ consideraremos que la relación lineal $y^* = a \cdot x + b$ no es buena para explicar los valores que toma "Y" a través de los valores que toma "X".

4.6 COEFICIENTE DE CORRELACIÓN LINEAL

De $r = \frac{S_{XY}}{S_X \cdot S_Y}$, que tiene igual signo que la covarianza S_{XY} , se dice que es el **coeficiente de correlación lineal**.

Como $S_{XY}^2 = r^2 \cdot S_X^2 \cdot S_Y^2$, es $S_e^2 = S_Y^2 - \frac{S_{XY}^2}{S_X^2} = S_Y^2 - \frac{r^2 \cdot S_X^2 \cdot S_Y^2}{S_X^2} = (1 - r^2) \cdot S_Y^2$.

Es:
$$R^2 = \frac{S_{Y^*}^2}{S_Y^2} = \frac{S_Y^2 - S_e^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2} = 1 - \frac{S_Y^2 - (S_{XY}^2/S_X^2)}{S_Y^2} = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = r^2$$

$$\boxed{S_e^2 = S_Y^2 - (S_{XY}^2/S_X^2)}$$

Así, siendo $R^2 \in [0;1]$, es $r \in [-1;1]$. Como $r^2 = R^2 = 1 - \frac{S_e^2}{S_Y^2}$, se tiene que:

- Si $r = \left\{ \begin{matrix} 1 \\ -1 \end{matrix} \right\}$ ($\Rightarrow S_e^2 = 0$), la relación de dependencia funcional entre "X" e "Y" es $\left\{ \begin{matrix} \text{directa} \\ \text{inversa} \end{matrix} \right\}$, pues la pendiente $a = S_{XY}/S_X^2$ de la recta de regresión tiene igual signo que la covarianza S_{XY} , que es $\left\{ \begin{matrix} \text{positiva} \\ \text{negativa} \end{matrix} \right\}$ si $r = \left\{ \begin{matrix} 1 \\ -1 \end{matrix} \right\}$.
- Si $r = 0$ ($\Rightarrow S_e^2 = S_Y^2$), no hay relación lineal entre "X" e "Y".

Si $U = (X - A)/B$ y $V = (X - A')/B'$, es:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{B \cdot B' \cdot S_{UV}}{(|B| \cdot S_U) \cdot (|B'| \cdot S_V)} = \frac{B \cdot B'}{|B| \cdot |B'|} \frac{S_{UV}}{S_U \cdot S_V} = \frac{B \cdot B'}{|B| \cdot |B'|} \cdot r_{UV}$$

$$\boxed{S_{XY} = B \cdot B' \cdot S_{UV} ; S_X = |B| \cdot S_U ; S_Y = |B'| \cdot S_V}$$

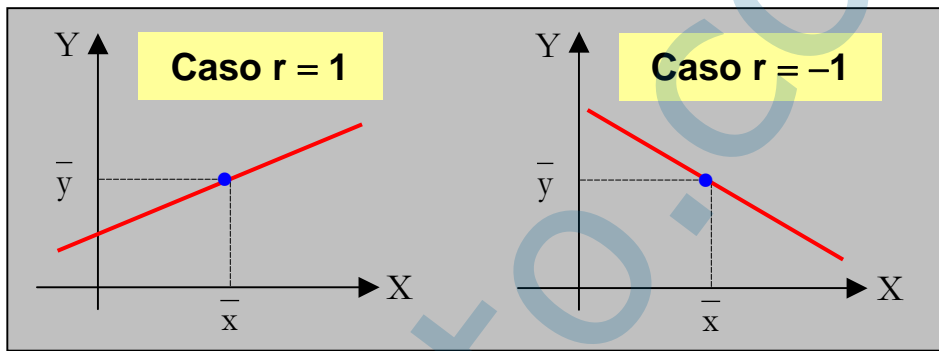
Por tanto, si B y B' tienen igual signo, es $r_{XY} = r_{UV}$, siendo $r_{XY} = -r_{UV}$ si B y B' tienen signo contrario.

4.7 POSICIÓN RELATIVA DE LAS RECTAS DE REGRESIÓN

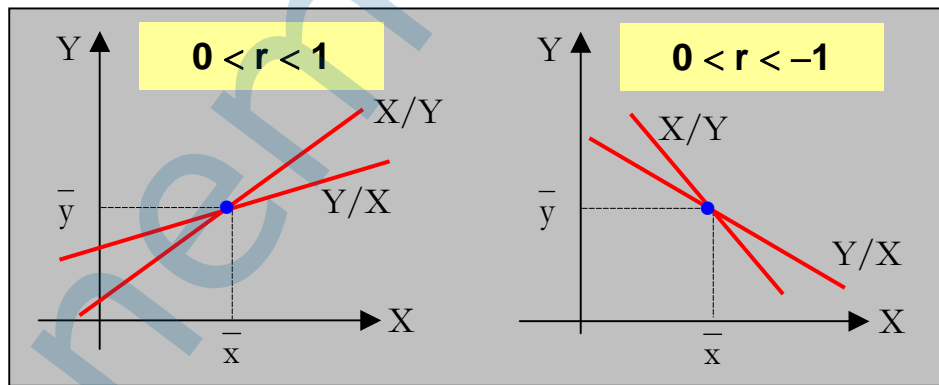
La pendiente $a = S_{XY}/S_X^2$ de la recta de regresión de "Y" sobre "X" tiene igual signo que la inversa $a' = S_{XY}/S_Y^2$ de la pendiente de la recta de regresión de "X" sobre "Y", siendo:

$$a \cdot a' = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = r^2$$

- Si $r = \pm 1 \Rightarrow a \cdot a' = r^2 = 1 \Rightarrow a = 1/a' \Rightarrow$ las dos rectas de regresión tienen igual pendiente y como ambas pasan por el punto $(\bar{x}; \bar{y})$, son la misma recta.



- Si $r \in (-1; 0) \cup (0; 1) \Rightarrow |a \cdot a'| = r^2 < 1 \Rightarrow |a| < \frac{1}{|a'|}$; o sea, el valor absoluto de la pendiente de la recta de regresión de "Y" sobre "X" es inferior al valor absoluto de la pendiente de la recta de regresión de "X" sobre "Y"



- Si $r = 0 \Rightarrow a = 0 \Rightarrow$ la recta de regresión de "Y" sobre "X" tiene pendiente 0 y pasa por el punto $(\bar{x}; \bar{y})$ y la recta de regresión de "X" sobre "Y" es la perpendicular a la anterior que pasa por el punto $(\bar{x}; \bar{y})$.

