Tema 2: Análisis de datos bivariantes

En este tema:

- Tabla de contingencia, tabla de doble entrada, distribución conjunta.
- Frecuencias relativas, marginales, condicionadas.
- Diagrama de dispersión.
- Tipos de relación entre las variables (lineal, no lineal y no relación).
- Covarianza, correlación, bondades y propiedades.
- Recta de regresión, interpretación de los coeficientes, predicción.
- Residuos, desviación típica residual, varianza explicada y no explicada, R^2.

Datos bivariantes

- Datos bivariantes provienen de la observación simultánea de dos variables (x, y) en una muestra de n individuos. Los datos serán parejas de valores, numéricos o no numéricos, de la forma:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Se usarán para describir las dos variables conjuntamente o una variable en función de la otra
- En los estudios de relaciones entre variables, una de las dos variables juega un papel más importante que la otra, ésta será la variable dependiente que denotaremos por y, cuyo comportamiento se intentará describir en función de otra variable x que llamaremos variable independiente o explicativa

Tabulación de datos

- En la tabla de doble entrada los valores de las variables x e y se representan en los márgenes y la frecuencia de cada pareja de clases se representa en la casilla correspondiente
- Cuando la variable es cualitativa la tabla de doble entrada se denomina tabla de contingencia

Ejemplo

x = Color de ojos de la madre $\{Claros, Oscuros\}$

y = Color de ojos del hijo $\{Claros, Oscuros\}$

$$(x_1, y_1) = (C, C), (x_2, y_2) = (C, O), \ldots, (x_{64}, y_{64}) = (O, O)$$

		x		
		Claros	**Oscuros**	Total
y	**Claros**	23	12	35
	Oscuros	17	12	29
	Total	40	24	64

Distribución conjunta

- Cuando la variable es cuantitativa la tabla de doble entrada se denomina distribución conjunta de frecuencias. Si las variables son discretas y no toman demasiados valores la tabla es inmediata.

Ejemplo

x = Asistencia semanal al teatro

y = Asistencia semanal al cine

(frecuencias absolutas en negro, frecuencias relativas en rojo)

		x				
		0	**1**	**2**	**3**	**4**
	0	12 0,279	5 0,116	4 0,093	2 0,047	1 0,023
	1	4 0,093	3 0,070	2 0,047	1 0,023	0 0,000
y	**2**	3 0,070	3 0,070	2 0,047	0 0,000	0 0,000
	3	1 0,023	0 0,000	0 0,000	0 0,000	0 0,000

Frecuencias marginales

- Se obtienen de sumar frecuencias conjuntas (absolutas o relativas)
- Se corresponden con las frecuencias univariantes de cada una de ellas, cuando no se tiene en cuenta el valor de la otra variable
- Si denominamos mediante la expresión $fr(x_i, y_i)$ a la frecuencia relativa correspondiente a los valores $(x = x_i, y = y_j)$, tendremos que

$$\sum_i \sum_j fr(x_i, y_j) = 1$$

- las frecuencias marginales de x se obtienen como

$$fr(x_i) = \sum_j fr(x_i, y_j)$$

y las frecuencias marginales de y como

$$fr(y_j) = \sum_i fr(x_i, y_j)$$

Frecuencias marginales

Ejemplo

Tomamos una muestra de un cierto número de empresas y analizamos dos variables:

x = Número de trabajadores

y = Número de ventas

		x				
		1-24	**25-49**	**50-74**	**75-99**	Total
y	**1-100**	0,293	0,122	0,098	0,049	0,561
	101-200	0,098	0,073	0,049	0,024	0,244
	201-300	0,073	0,073	0,049	0,000	0,195
	Total	0,463	0,268	0,195	0,073	1,000

Frecuencias marginales

Trabajadores	**1-24**	**25-49**	**50-74**	**75-99**	
$fr(x)$	0,463	0,268	0,195	0,073	1

Ventas	**1-100**	**101-200**	**201-300**	
$fr(y)$	0,561	0,244	0,195	1

Frecuencias condicionadas

- Se construyen para una de las dos variables, cuando fijamos un valor concreto que ha sido observado en la otra
- Si fijamos el valor de $x = x_i$, podemos construir la distribución de frecuencias de la variable y, condicionada al valor x_i de x, frecuencias que representaremos por

$$fr(y_j|x_i) = \frac{fr(x_i, y_j)}{fr(x_i)}$$

- Se verifica que

$$\sum_j fr(y_j|x_i) = \frac{\sum_j fr(x_i, y_j)}{fr(x_i)} = 1$$

Frecuencias condicionadas

Ejemplo

x = Número de trabajadores

y = Número de ventas

Halla la distribución de las ventas para las empresas de entre 50 y 74 trabajadores.

		x				
		1-24	**25-49**	**50-74**	**75-99**	Total
y	**1-100**	0,293	0,122	0,098	0,049	0,561
	101-200	0,098	0,073	0,049	0,024	0,244
	201-300	0,073	0,073	0,049	0,000	0,195
	Total	0,463	0,268	0,195	0,073	1,000

La distribución condicional $fr(y|50 \leq x \leq 74)$

Ventas	**1-100**	**101-200**	**201-300**
$fr(y\|50 \leq x \leq 74)$	$0{,}500(=\frac{0{,}098}{0{,}195})$	$0{,}250(=\frac{0{,}049}{0{,}195})$	$0{,}250(=\frac{0{,}049}{0{,}195})$

Frecuencias condicionadas

Ejemplo

x = Trabajadores

y = Ventas

Análogamente, la distribución del número de trabajadores entre las empresas cuyas ventas están, entre 101 y 200 artículos.

		x				
		1-24	**25-49**	**50-74**	**75-99**	Total
y	**1-100**	0,293	0,122	0,098	0,049	0,561
	101-200	0,098	0,073	0,049	0,024	0,244
	201-300	0,073	0,073	0,049	0,000	0,195
	Total	0,463	0,268	0,195	0,073	1,000

La distribución condicional $fr(x|101 \leq y \leq 200)$

Trabajadores	**1-24**	**25-49**	**50-74**	**75-99**	Total
$fr(x\|101 \leq y \leq 200)$	0,402	0,299	0,201	0,098	1

Representaciones gráficas

- La representación gráfica más útil para dos variables continuas es a través del llamado diagrama de dispersión

m2	Precio
107	62657
114	65554
91	54506
100	62103
96	58271
107	66925
104	61917
100	61149
80	52263
81	51878
105	65678
111	66696
108	65387
97	61806
106	63824

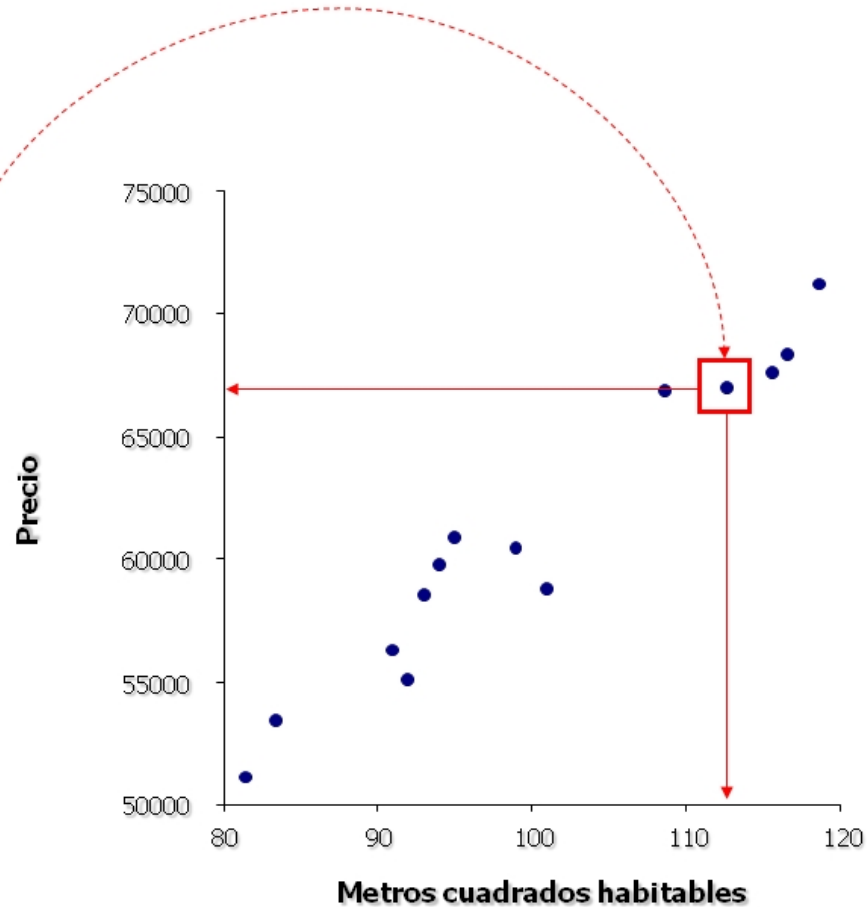
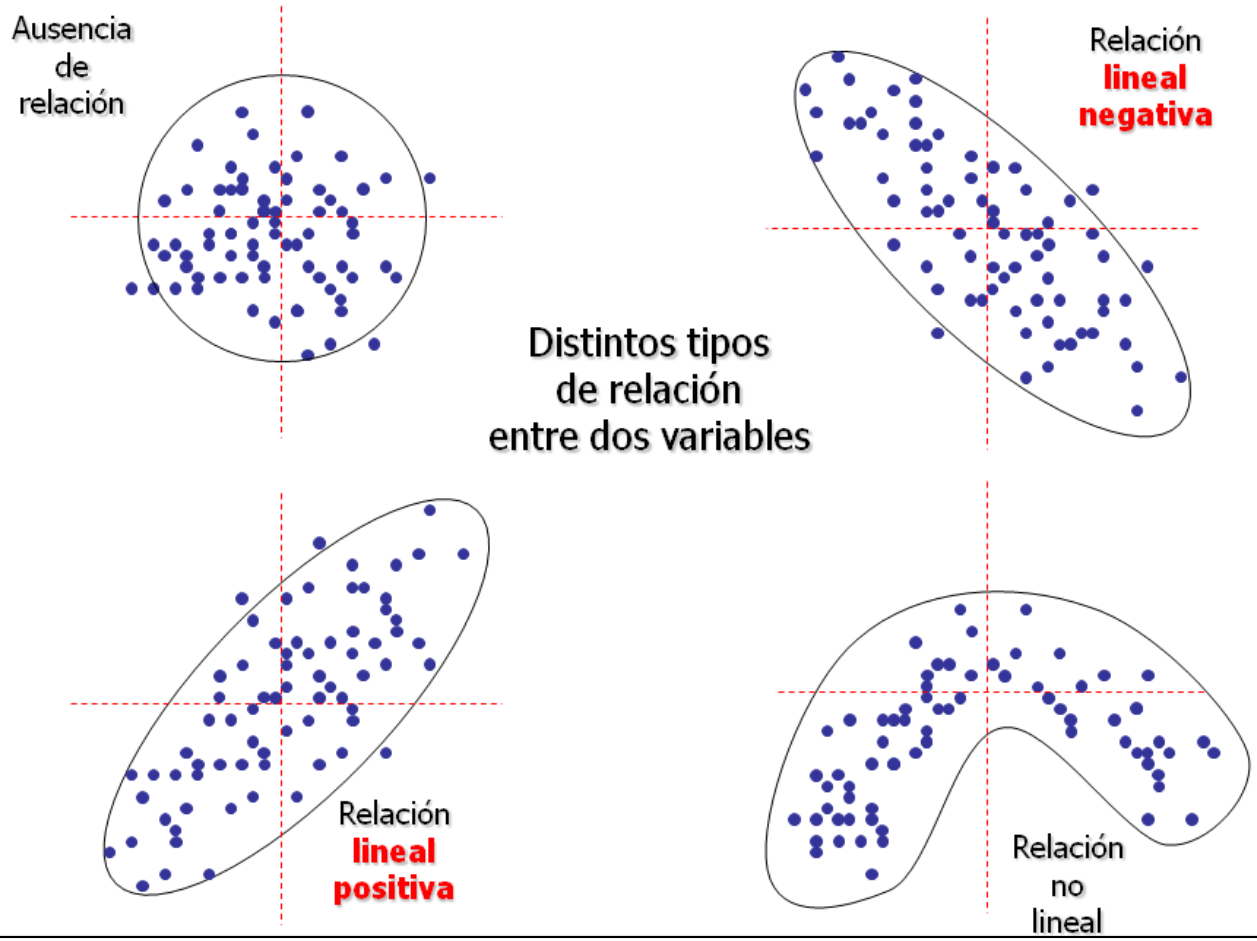


Diagrama de dispersión

Tipos de relación

- Existen distintas formas en la cual dos variables pueden estar relacionadas: ausencia de ninguna relación, relación lineal positiva, relación lineal negativa, relación no lineal



Medidas de dependencia lineal

- Buscamos una medida descriptiva que, mediante una única cifra, nos indique si entre dos variables x e y existe una relación de tipo lineal o no. Covarianza se obtiene como

$$Cov(x,y) = \frac{1}{n-1}\left(\overbrace{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}^{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}\right)$$

Propiedades de la covarianza

- Es una medida de la asociación lineal entre dos variables que resume la información existente en un gráfico de dispersión (¡la cifra se supedita al gráfico!)
- Si la covarianza es mayor que cero y 'grande' es porque existe una relación lineal positiva
- Si la covarianza es menor que cero y 'grande' es porque existe una relación lineal negativa
- Si la covarianza es 'pequeña' es porque bien no existe una relación lineal o bien porque existiendo relación, ésta es no lineal
- Inconvenientes
  - ¿Qué significa grande o pequeña?
  - ¿En qué unidades está medida?

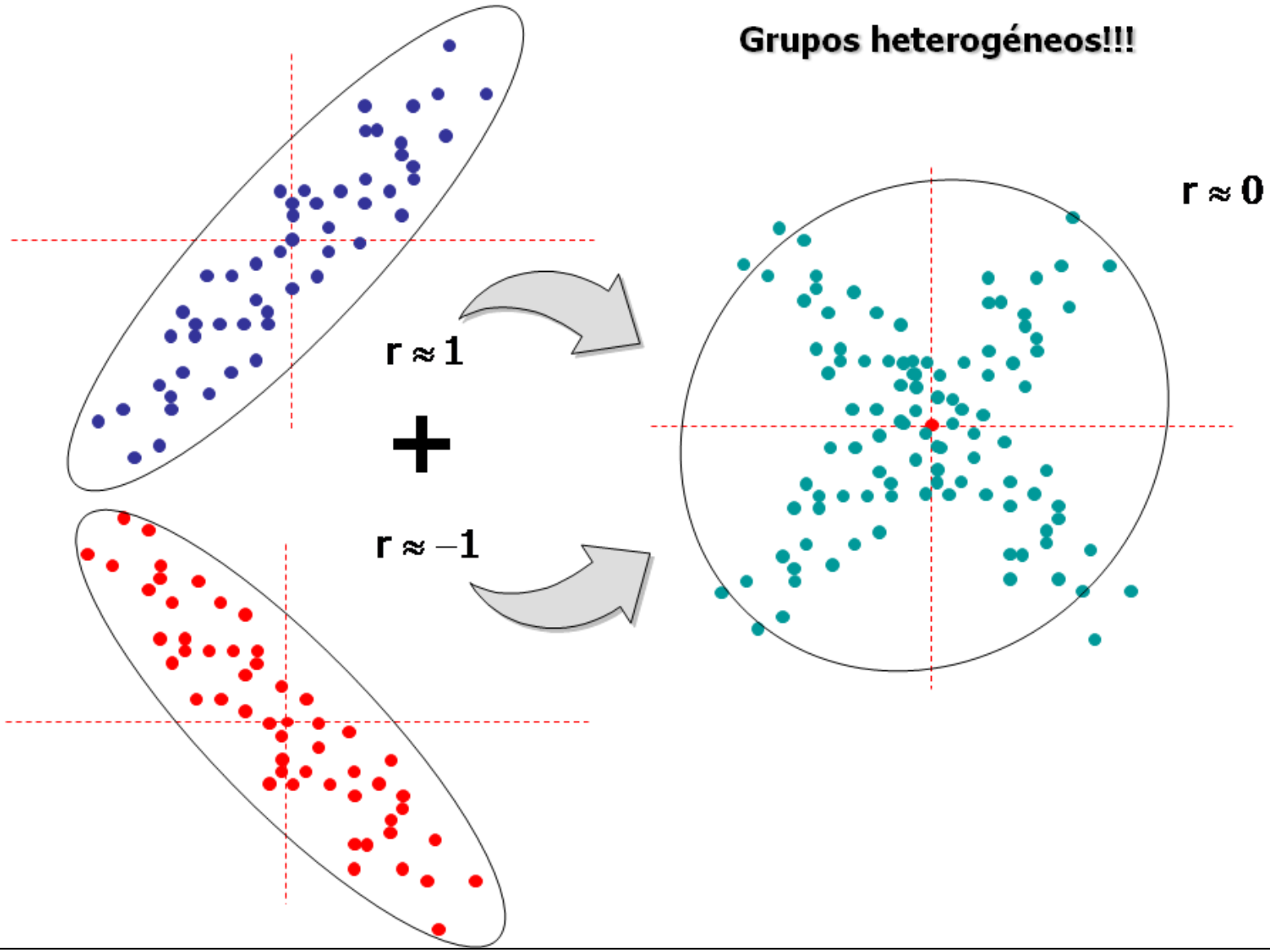
Medidas de dependencia lineal

- Correlación se obtiene como

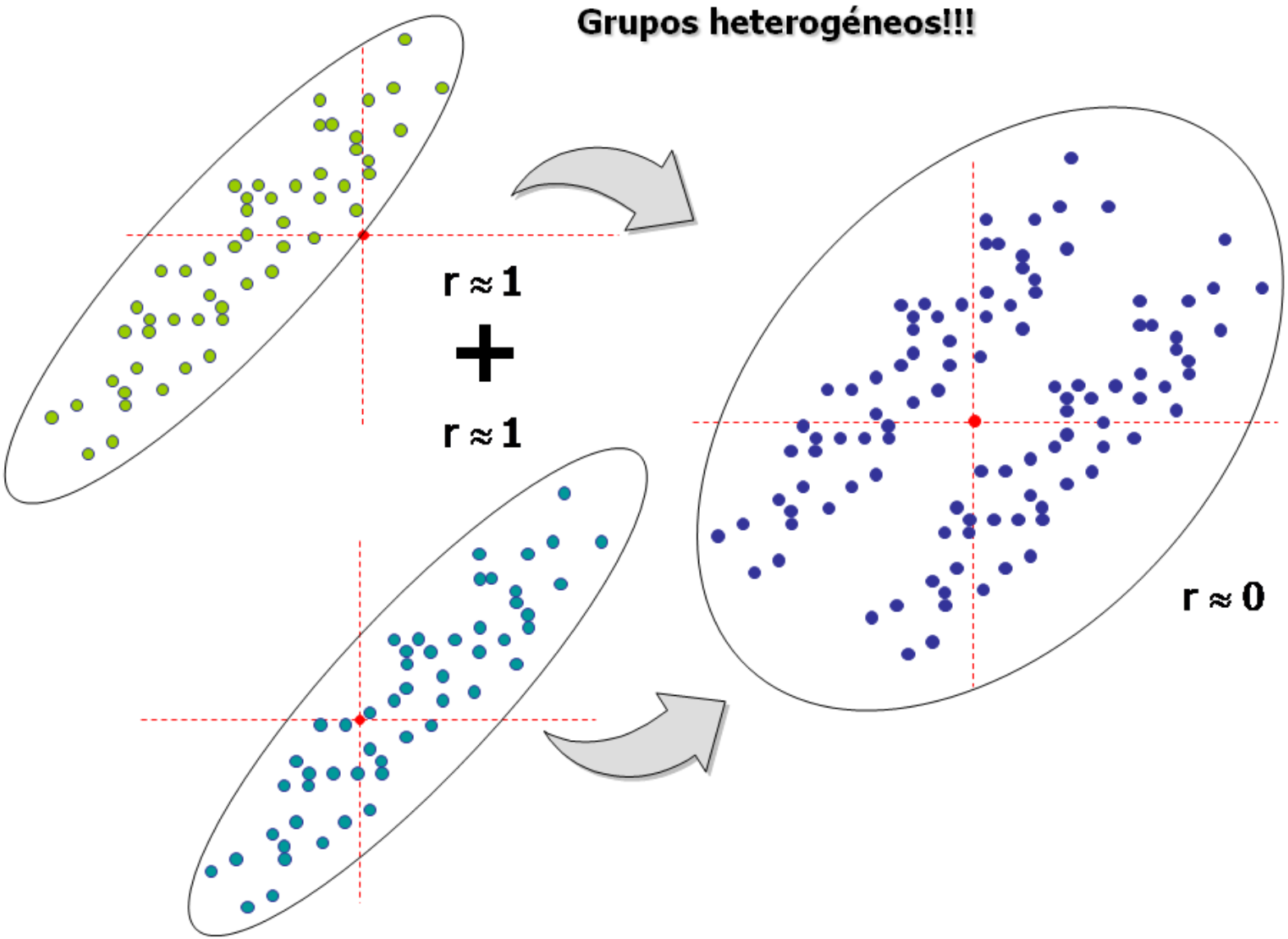
$$r_{(x,y)} = Cor(x,y) = \frac{Cov(x,y)}{s_x s_y}$$

- ¿Inconvenientes? Ninguno
- ¿Ventajas?
  - Está acotado $-1 \leq r_{(x,y)} \leq 1$, ahora los términos grande y pequeño tienen sentido
  - Es adimensional

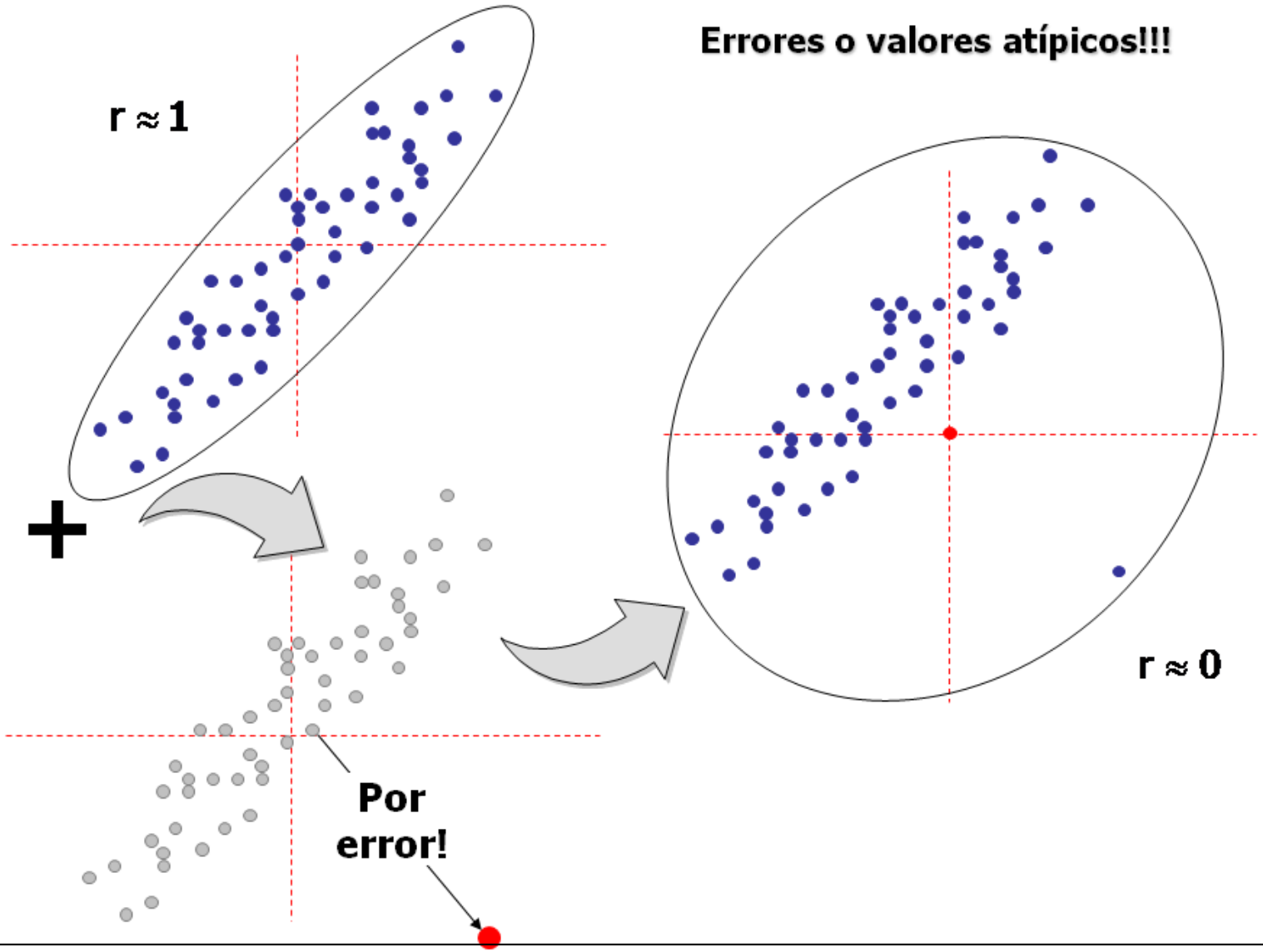
Correlación y heterogeneidad



Correlación y heterogeneidad

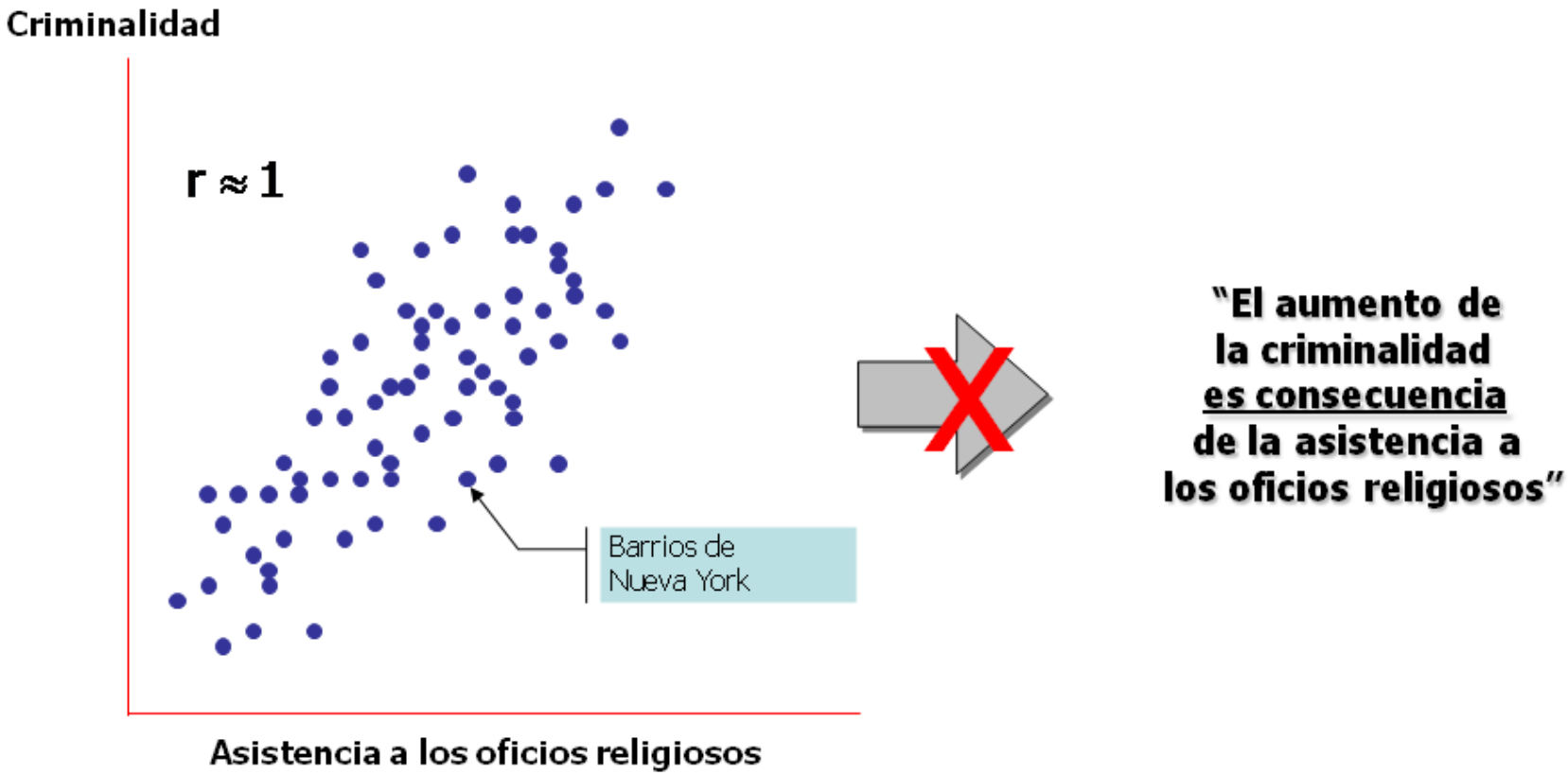


Correlación y datos atípicos



Correlación ≠ Causalidad

Correlación no es causalidad!!



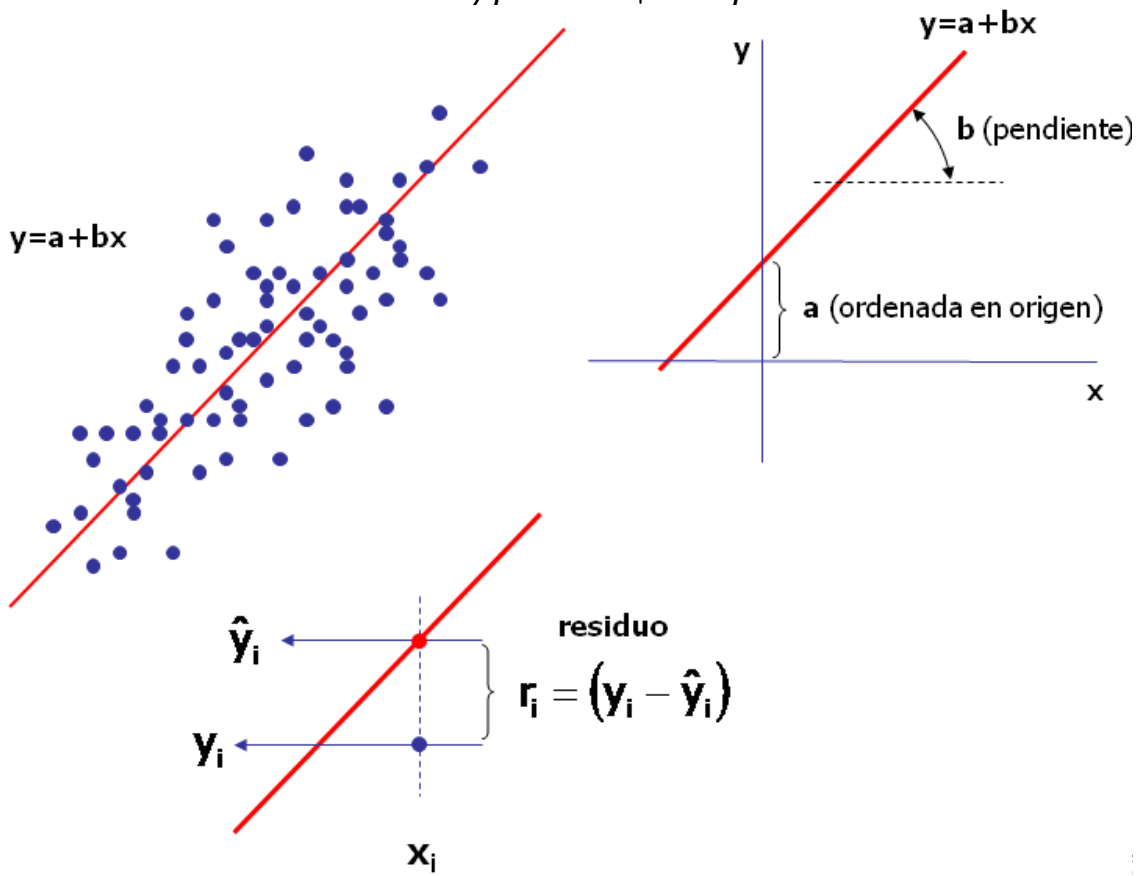
Recta de regresión

- El modelo poblacional es

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- El modelo estimado es

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



Recta de regresión

- Calculamos la recta imponiendo la condición de que los residuos (errores) cuadráticos sean mínimos (método de mínimo cuadrados)

$$\min_{a,b} \sum_i r_i^2 = \min_{a,b} \sum_i (\overbrace{y - \hat{y}_i}^{\text{residuo } i})^2$$
$$= \min_{a,b} \sum_i (y - (a + bx_i))^2$$

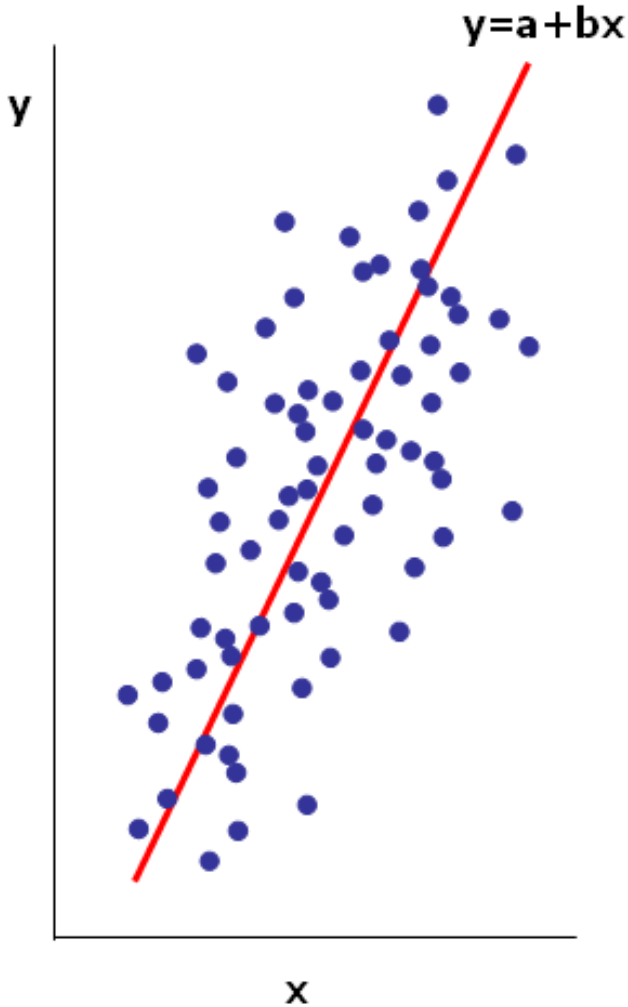
- Obtenemos los estimadores de la pendiente (b) y la ordenada en origen (a)

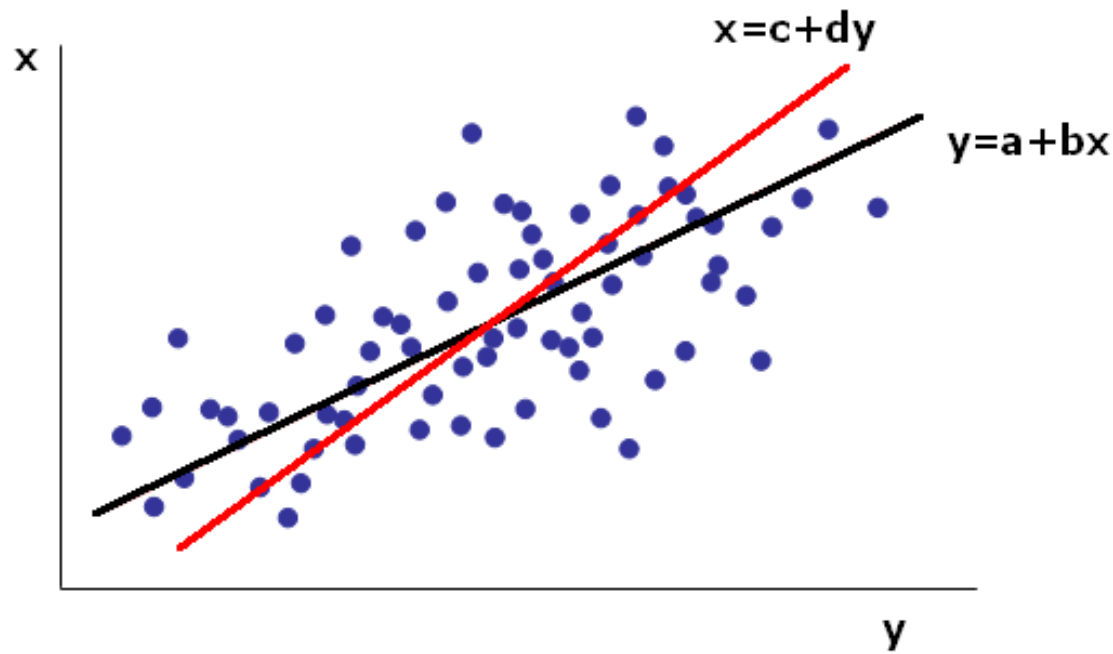
$$b = \frac{Cov(x,y)}{s_x^2} = r_{(x,y)} \frac{s_y}{s_x}$$
$$a = \bar{y} - b\bar{x}$$

- La recta de regresión pasa por el punto $(\bar{x}, \bar{y})$.

Recta de regresión

- Existen dos rectas

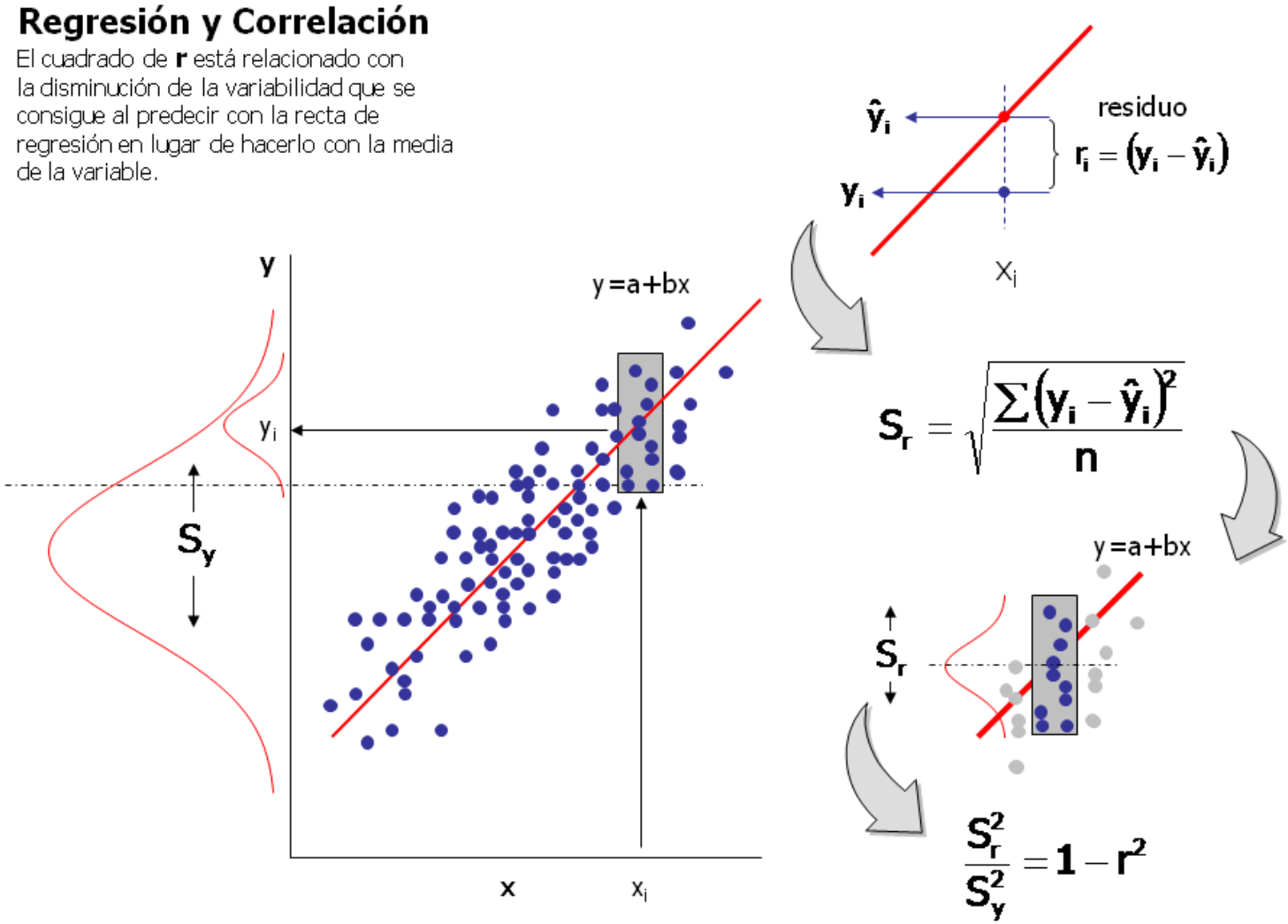




La recta de regresión no es simétrica

Recta de regresión y correlación

- Coeficiente de determinación $R^2 = r^2_{(x,y)}$



Análisis de los residuos en regresión

$$r_i = y_i - \hat{y}_i = y_i - (ax_i + b) \qquad i = 1, \ldots, n$$

La correlación entre dos variables puede ser alta a pesar de que la relación entre las dos sea fuertemente **no lineal**.

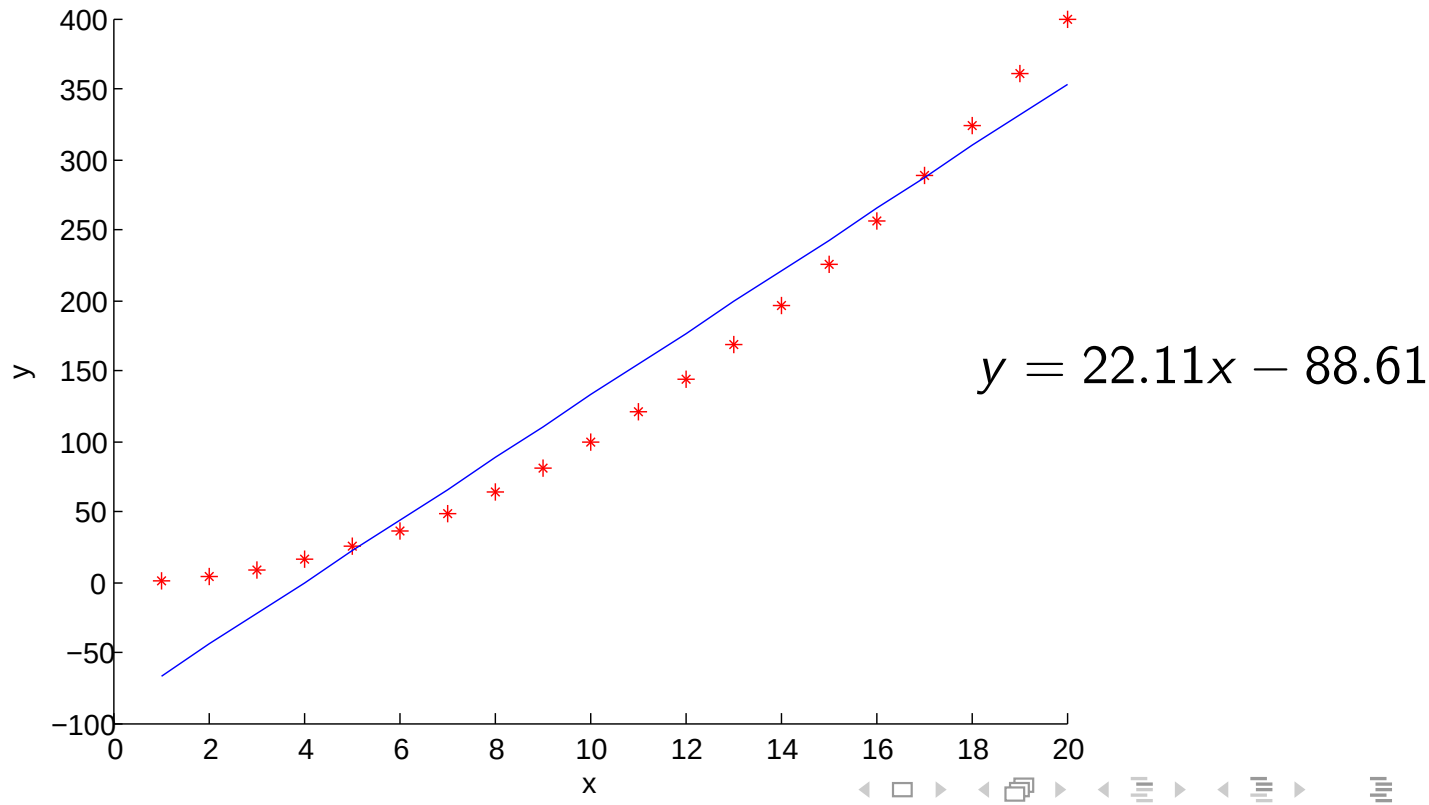
Se pueden utilizar los residuos para ver si el modelo de regresión lineal es adecuado.

Casi siempre es útil hacer gráficos de los residuos (frente a x, y o $\hat{y}$). Si los puntos del gráfico de los residuos parecen aleatorios, tenemos una buena indicación de que el modelo de regresión se ajusta correctamente.

Análisis de los residuos en regresión

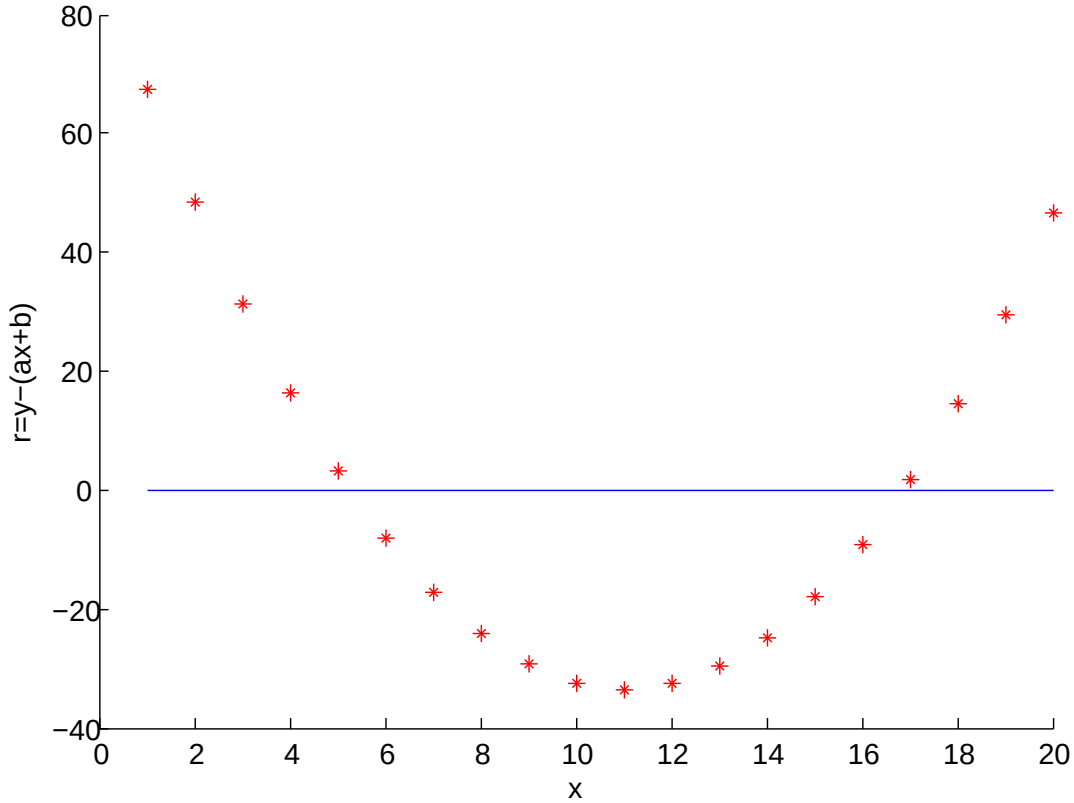
x	1	2	3	4	5	6	7	8	9	10	11	...	19	20
$y = x^2$	1	4	9	16	25	36	49	64	81	100	121	...	361	400

El coeficiente de correlación es $r_{xy} = 0.9713$.



Análisis de los residuos en regresión

Hacemos un gráfico de los residuos ($r_i = y_i - (a + bx_i)$) frente a la variable independiente (x).

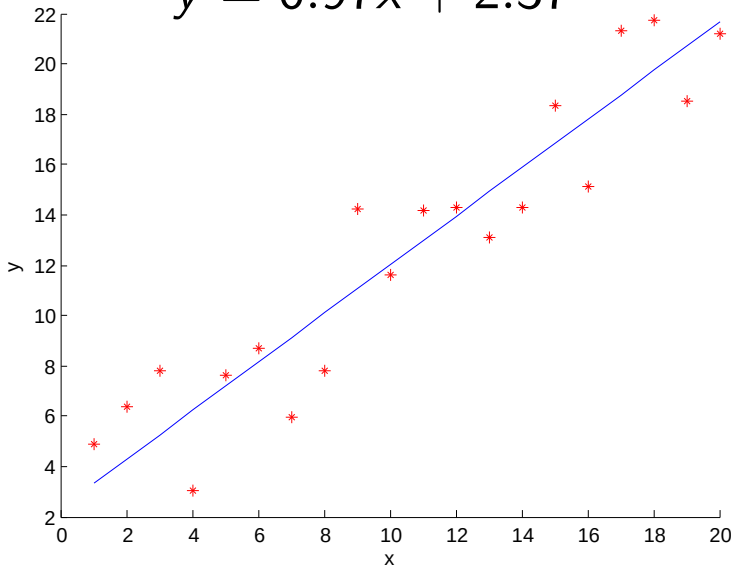


Los residuos no se reparten de forma aleatoria en torno a cero. Vemos que un modelo cuadrático sería más adecuado.

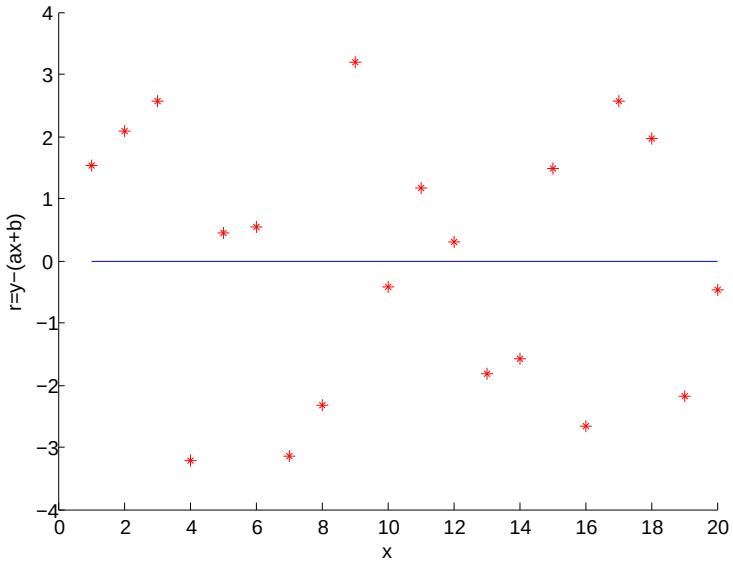
Análisis de los residuos en regresión

Diagrama de dispersión

$y = 0.97x + 2.37$



Residuos frente a x



Los residuos se reparten de forma aleatoria en torno a cero. El modelo lineal parece adecuado.

Recta de regresión

Ejemplo Los datos Anscombe:

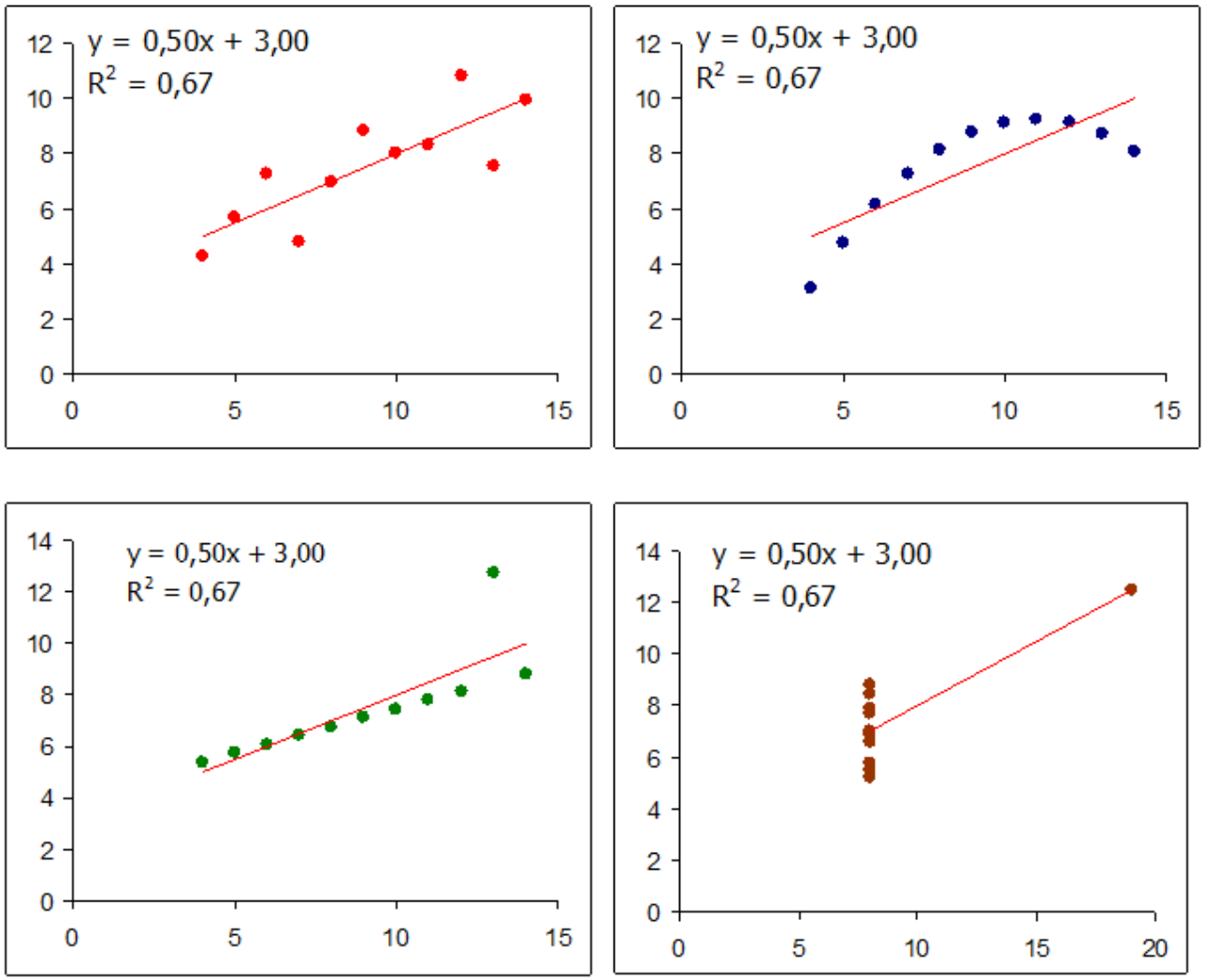
Estos datos están disponibles en RCommander:

`Datos` → `Conjunto de datos en paquetes` → `datasets` → `anscombe`

	Datos 1		Datos 2		Datos 3		Datos 4	
i	x	y	x	y	x	y	x	y
1	10	8,04	10	9,14	10	7,46	8	6,58
2	8	6,95	8	8,14	8	6,77	8	5,76
3	13	7,58	13	8,74	13	12,74	8	7,71
4	9	8,81	9	8,77	9	7,11	8	8,84
5	11	8,33	11	9,26	11	7,81	8	8,47
6	14	9,96	14	8,1	14	8,84	8	7,04
7	6	7,24	6	6,13	6	6,08	8	5,25
8	4	4,26	4	3,1	4	5,39	19	12,5
9	12	10,84	12	9,13	12	8,15	8	5,56
10	7	4,82	7	7,26	7	6,42	8	7,91
11	5	5,68	5	4,74	5	5,73	8	6,89

- Los datos Anscombe es un conjunto de cuatro pares de datos que tienen las mismas propiedades estadísticas, pero que evidentemente son distintos
- Son una demostración de la necesidad de interpretar, tanto la recta de regresión como el coeficiente de correlación, después de observar el diagrama de dispersión

Los datos Anscombe



Los datos Anscombe

- Para los Datos 1 tenemos

$$\hat{y} = 3 + 0,5x$$

- Interpretación de la pendiente: un incremento de una unidad en x produce, en promedio, un incremento de 0,5 en y
- Predicción para $x = 7,5$ (dentro del rango de x)

$$\hat{y} = 3 + 0,5(7,5) = 6,75$$