Tema 2: Análisis de datos bivariantes

Los contenidos a desarrollar en este tema son los siguientes:

1. Tablas de doble entrada.
2. Diagramas de dispersión.
3. Covarianza y Correlación.
4. Regresión lineal.

Lecturas recomendadas:

- Capítulos 2 y 3 del libro de Newbold, Carlson y Thorne (2009).
- Capítulo 3 del libro de Peña (2001).
- Capítulos 7 a 9 del libro de Peña y Romo (1997).

Tema 2: Análisis de datos bivariantes

Despues de estudiar este tema conoceremos:

- Tabla de doble entrada y distribución conjunta
- Frecuencias relativas, marginales y condicionadas
- Diagrama de dispersión
- Tipos de relación entre las variables (lineal, no lineal y no relación)
- Covarianza y correlación. Propiedades
- Recta de regresión, interpretación de los coeficientes y 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 $\boldsymbol{y}$, cuyo comportamiento se intentará describir en función de otra variable $\boldsymbol{x}$ 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 $\{\mathbf{C}laros, \mathbf{O}scuros\}$

$y =$ Color de ojos del hijo $\{\mathbf{C}laros, \mathbf{O}scuros\}$

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

Distribución conjunta

Ejemplo

x = Asistencia semanal al teatro

y = Asistencia semanal al cine

		x				
		0	**1**	**2**	**3**	**4**
	0	12	5	4	2	1
	1	4	3	2	1	0
y	**2**	3	3	2	0	0
	3	1	0	0	0	0

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

Frecuencias marginales

- Se obtienen de sumar frecuencias conjuntas (absolutas o relativas) por filas y por columnas.

- Si denominamos $fr(x_i, y_i)$ 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

x = Trabajadores
y = 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

$x =$ Trabajadores
$y =$ 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** $\boldsymbol{fr(y|50 \leq x \leq 74)}$

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

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

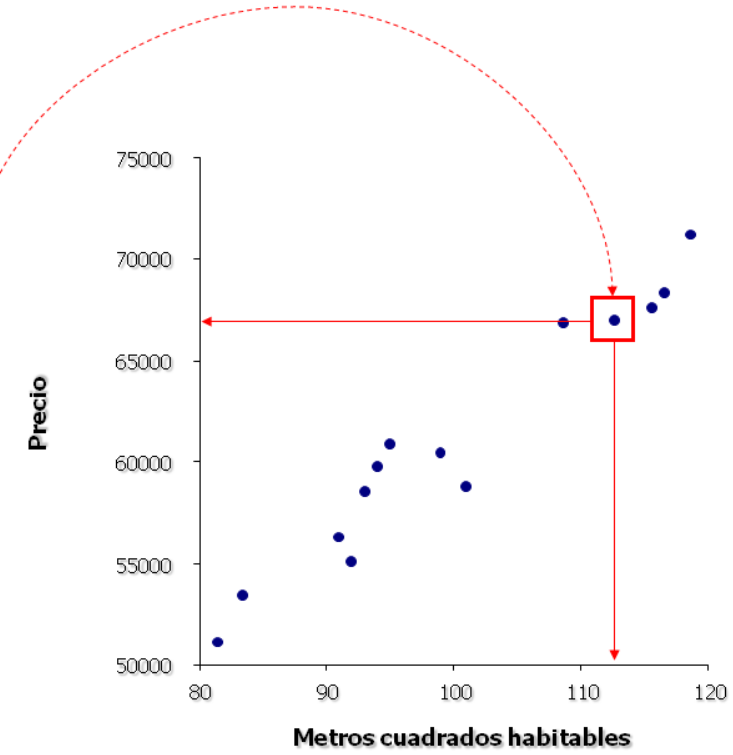
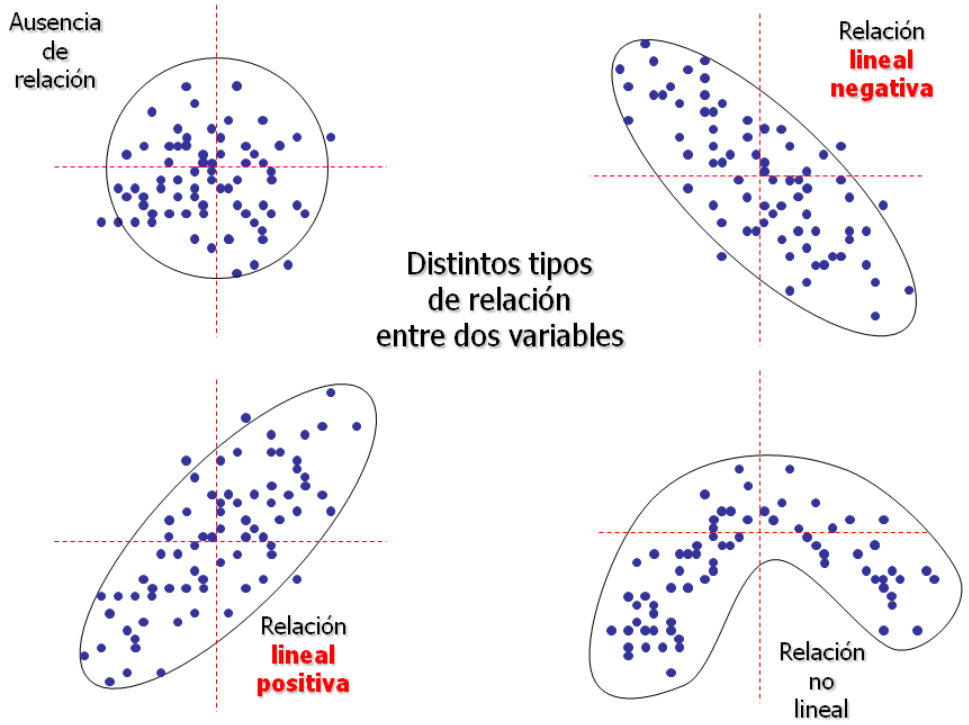


Diagrama de dispersión

Tipos de relación

- Existen distintas formas en que dos variables pueden estar relacionadas: ausencia de relación, relación **lineal** positiva, relación **lineal** negativa, relación no lineal



Medidas de dependencia lineal

- Buscamos una medida descriptiva que, mediante un único valor, nos indique si entre dos variables x e y existe una relación de tipo lineal o no.

- La **covarianza** se obtiene como

$$\begin{aligned} Cov(x,y) &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i \;-\; n\bar{x}\bar{y}\right) \end{aligned}$$

Propiedades de la covarianza

- Es una medida de la asociación lineal entre dos variables que resume la información existente en un gráfico de dispersión
- 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 se expresa la covarianza?

Medidas de dependencia lineal

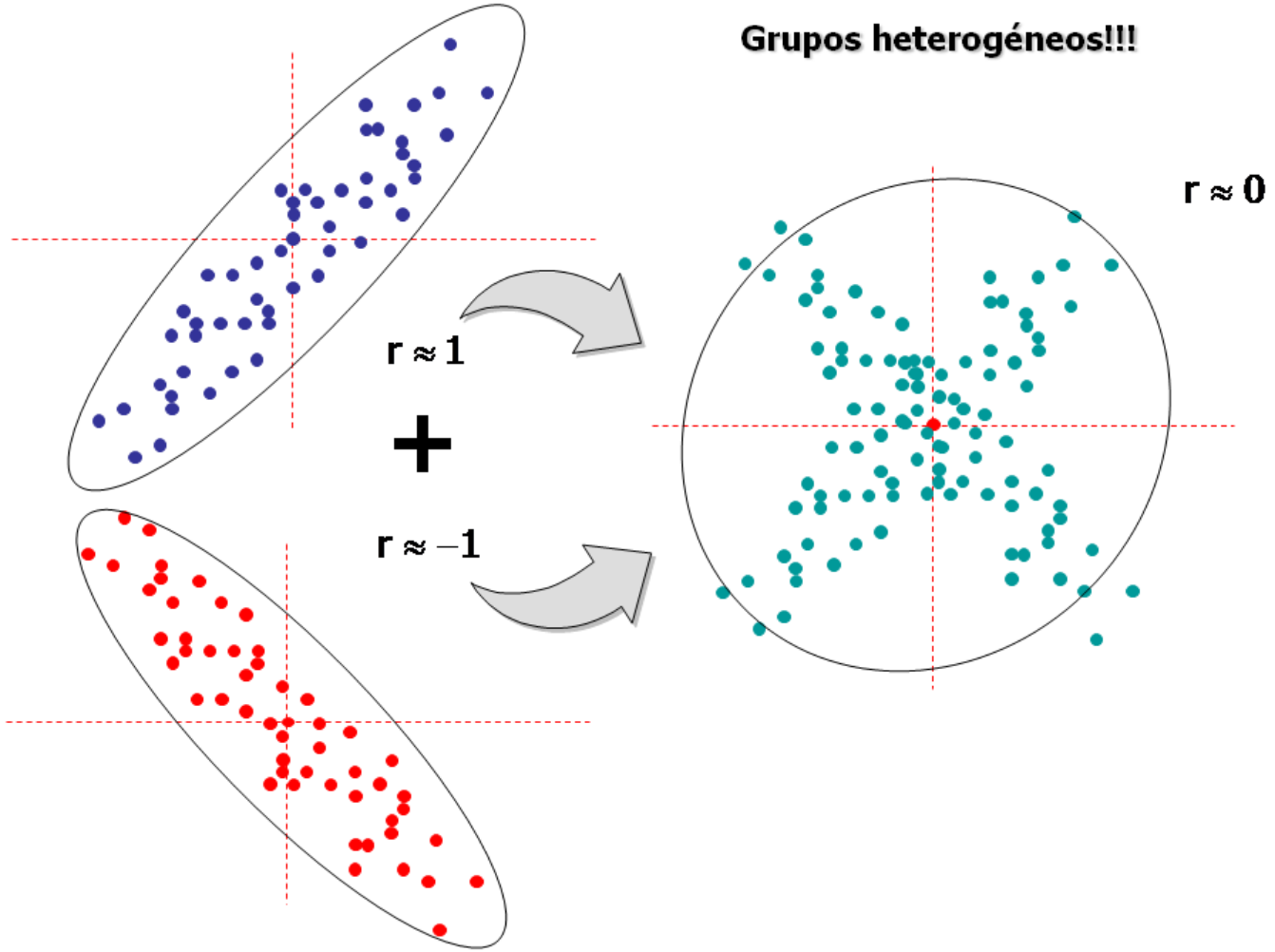
- La **correlación** se obtiene como

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

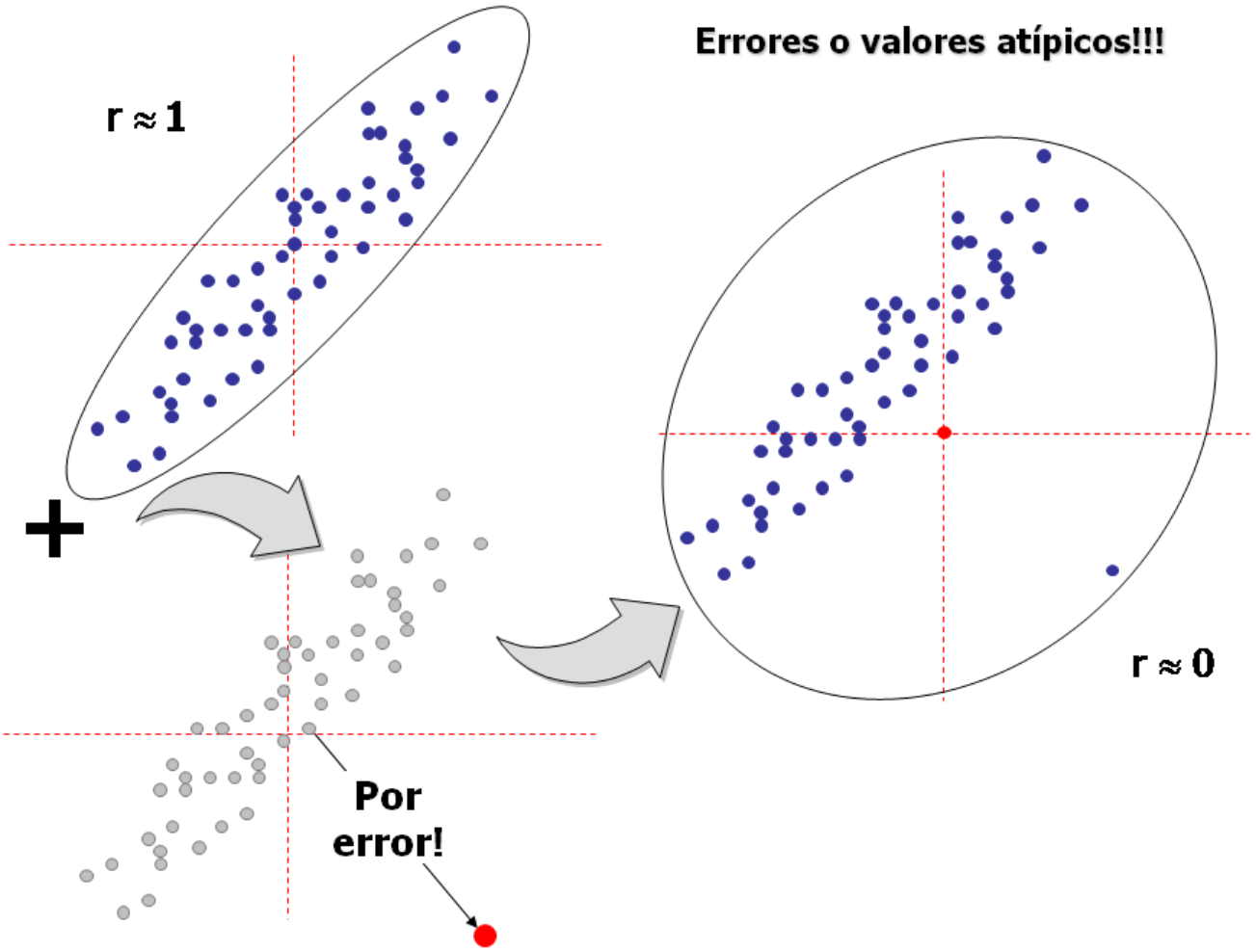
- ¿Inconvenientes? Ninguno.

- ¿Ventajas?
  - Está **acotada**: $\boxed{-1 \leq r_{(x,y)} \leq 1}$.
  - Ahora los términos grande y pequeño tienen sentido.
  - Es **adimensional**.

Correlación y heterogeneidad

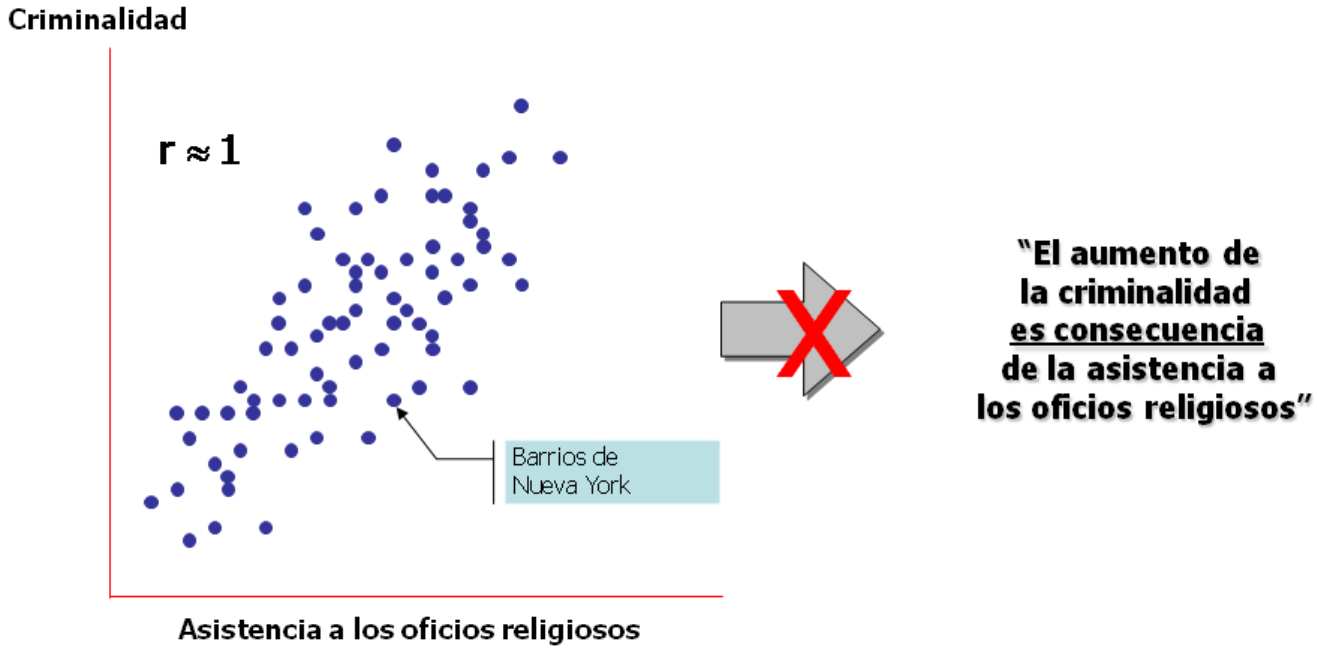


Correlación y datos atípicos



Correlación ≠ Causalidad

Correlación no es causalidad!!



Tema 2: Análisis de datos bivariantes

Los contenidos a desarrollar en este tema son los siguientes:

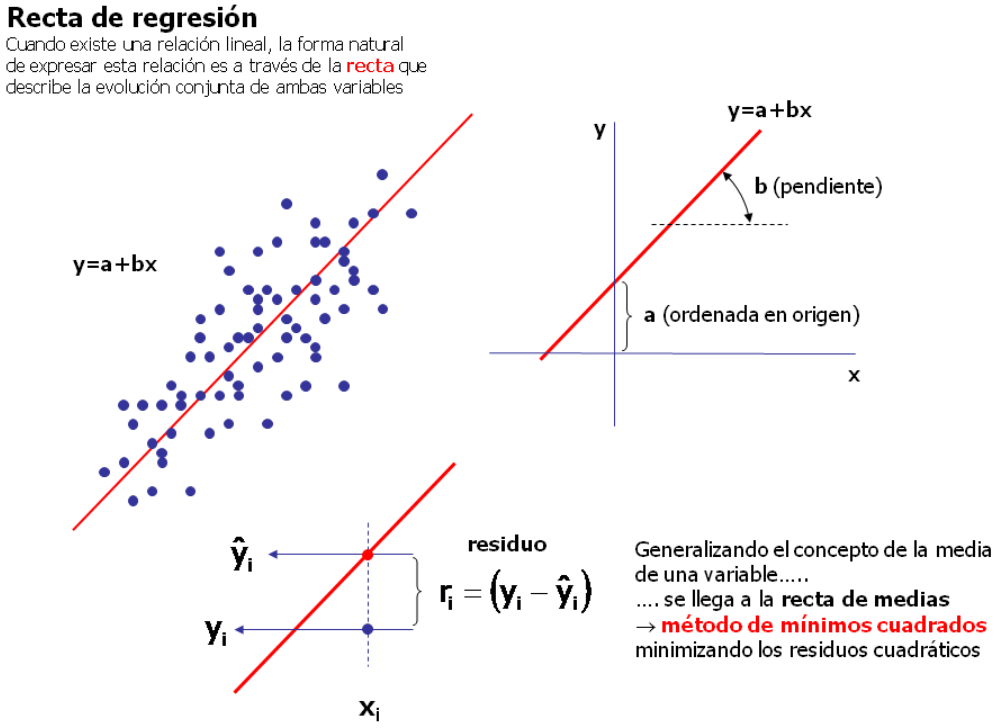
1. Tablas de doble entrada. ✓
2. Diagramas de dispersión. ✓
3. Covarianza y Correlación. ✓
4. Regresión lineal.

Lecturas recomendadas:

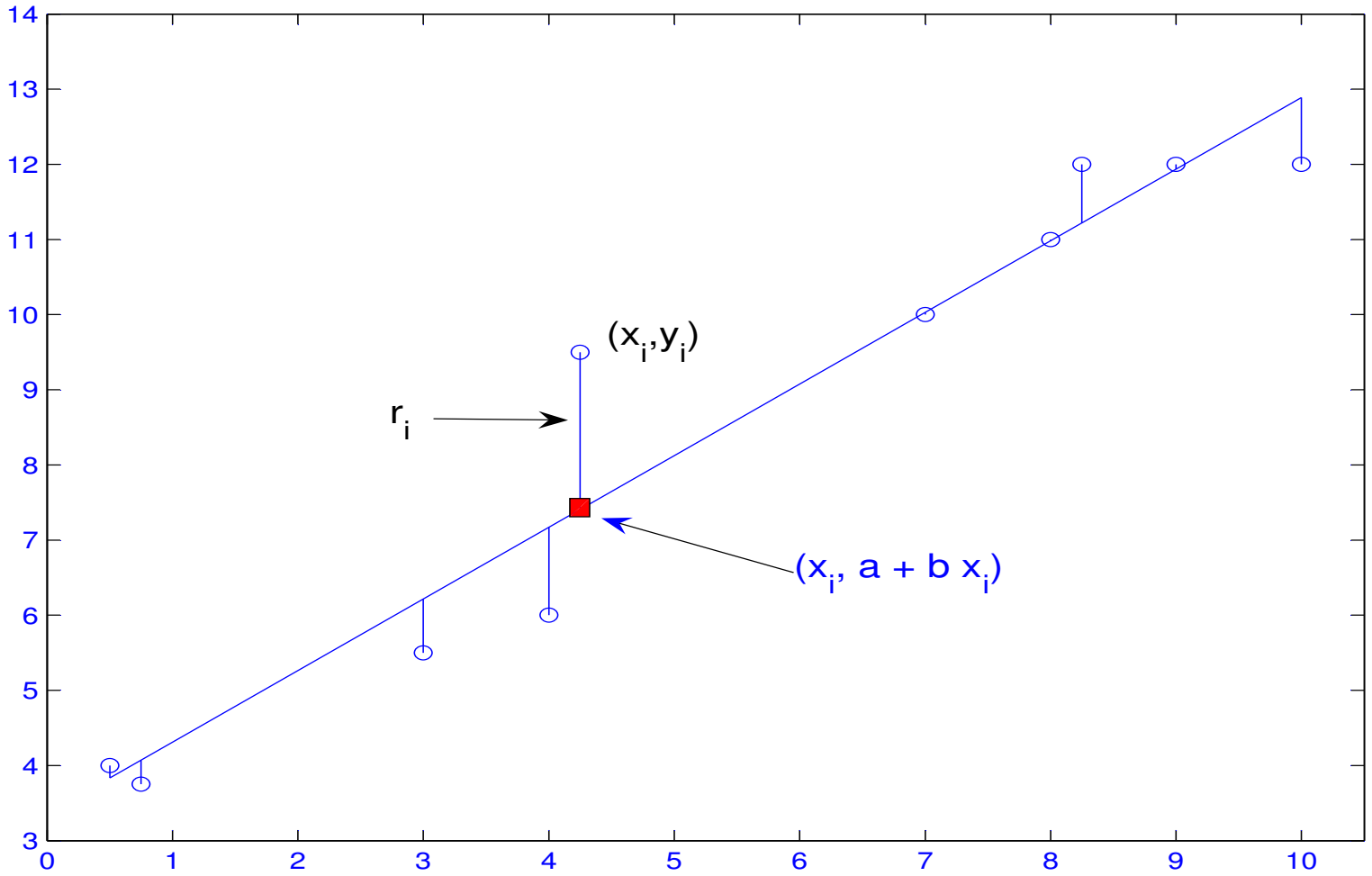
- Capítulo 3 del libro de Newbold, Carlson y Thorne (2009).
- Capítulo 3 del libro de Peña (2001).
- Capítulo 9 del libro de Peña y Romo (1997).

Recta de regresión

- El modelo poblacional es $y_i = \alpha + \beta x_i + \epsilon_i$.
- El modelo estimado es $\hat{y}_i = \mathbf{a} + \mathbf{b}x_i$.



Cálculo de la recta de regresión



$$r_i = y_i - (a + bx_i) = y_i - \widehat{y}_i.$$

Recta de regresión

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

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

$$= \min_{a,b} \sum_i (y - (a + bx_i))^2$$

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

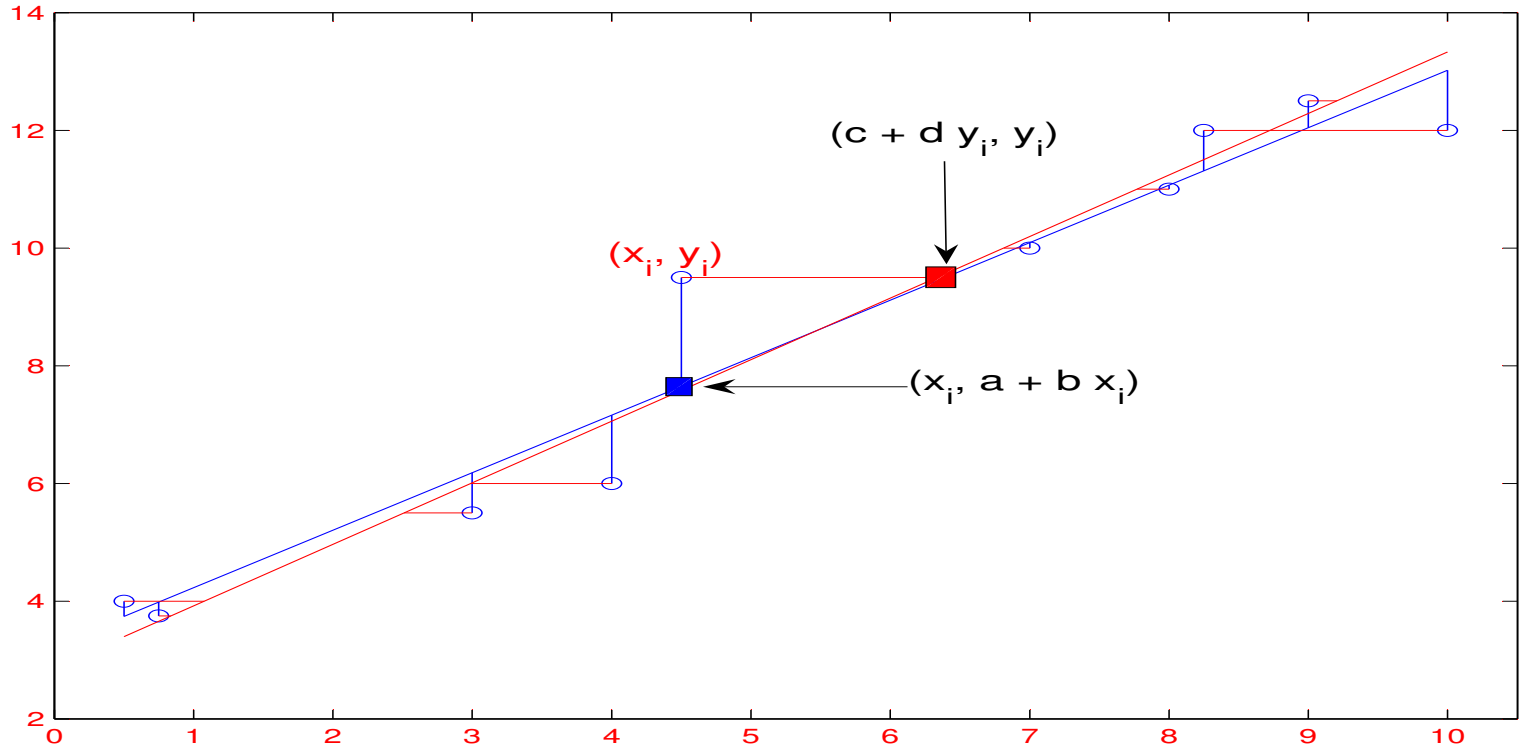
$$\mathbf{b} = \frac{\mathbf{Cov(x,y)}}{\mathbf{s_x^2}} = \mathbf{r_{(x,y)}} \frac{\mathbf{s_y}}{\mathbf{s_x}}$$

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

Dos rectas de regresión

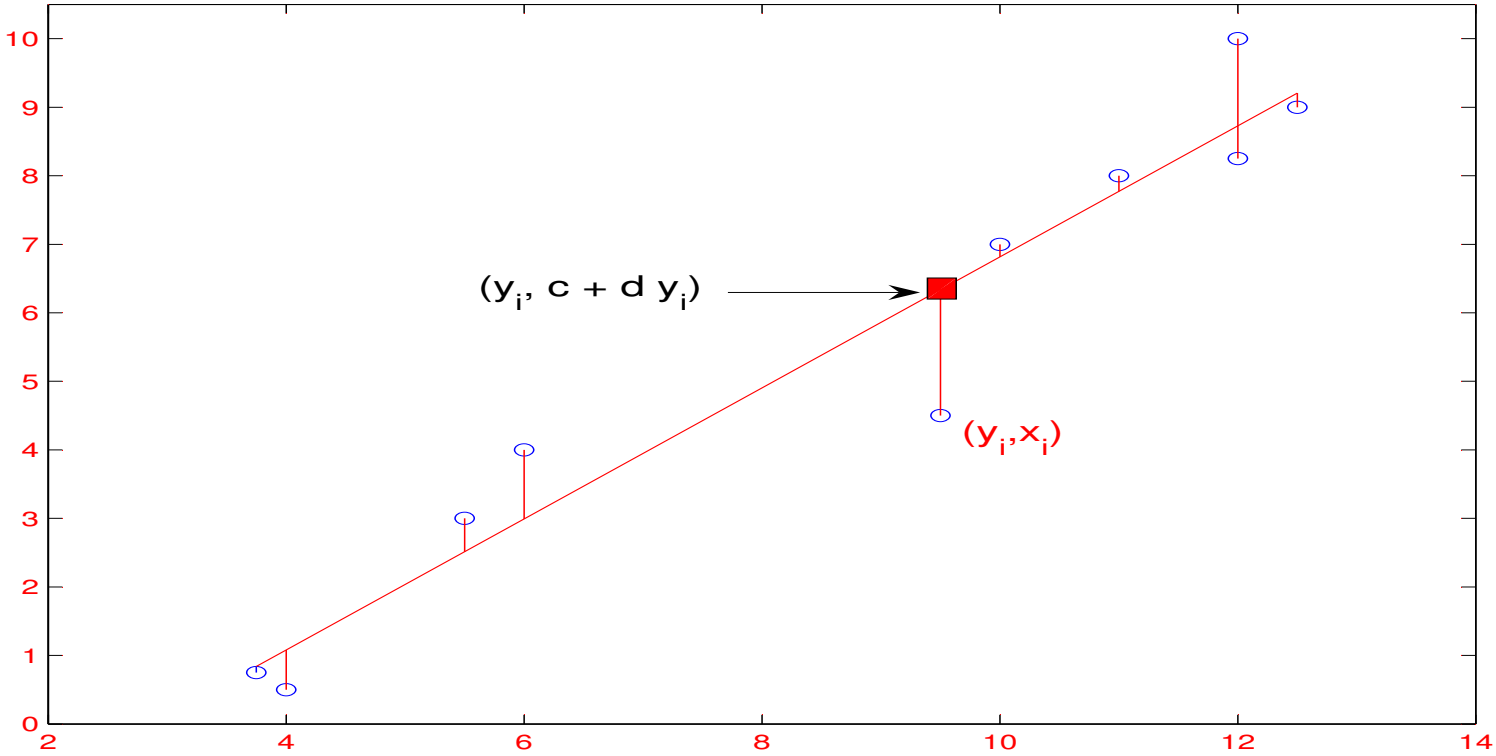
- Dependiendo de la dirección en que se miden los errores, r_i, existen **dos rectas**:

Recta de regresión de y sobre x y Recta de regresión de x sobre y



Dos rectas de regresión

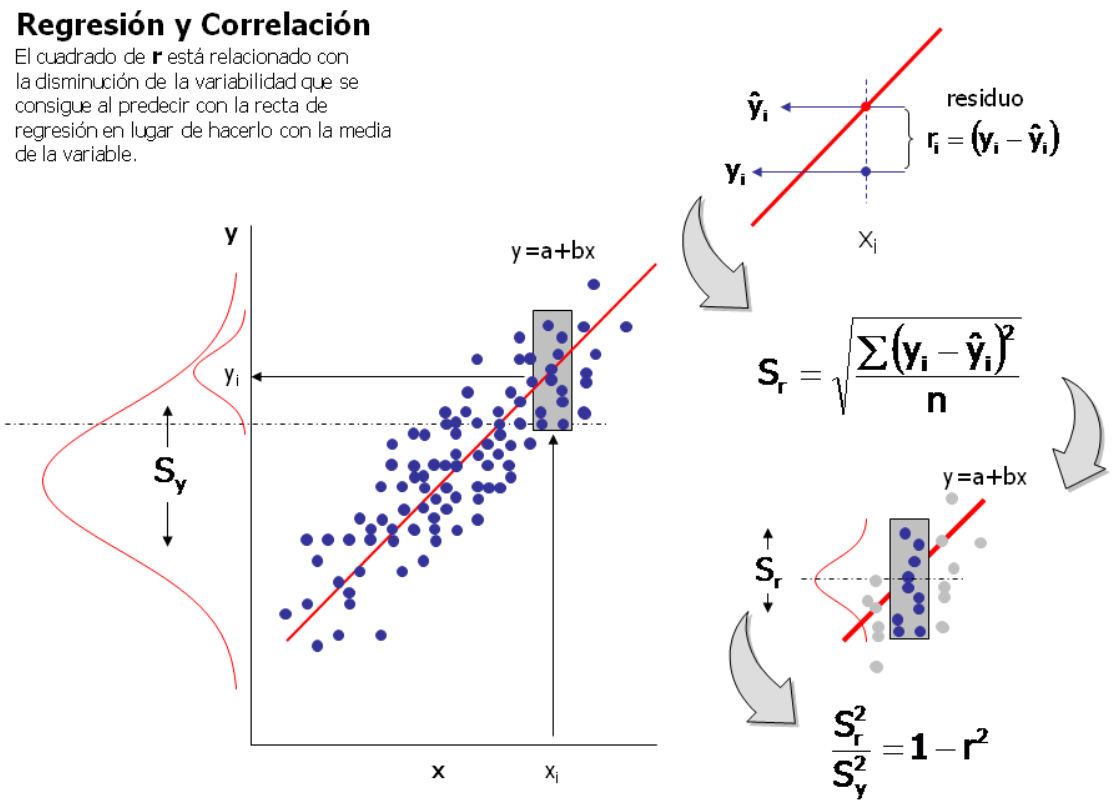
Recta de regresión de x sobre y



- (x, y) pasa a (y, x).
- Los errores se toman en horizontal, $r_i = x_i - c - dy_i$.

Recta de regresión y correlación

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



Análisis de los residuos

- Se puede utilizar los residuos para ver si el modelo de regresión es adecuado.
- Casi siempre es útil hacer gráficos de los residuos (frente x, y o $\hat{y}$) para ver si los supuestos del modelo lineal de regresión son adecuados o no.
- Si los puntos de **gráfico de los residuos** (residuos frente x, y o $\hat{y}$) parecen aleatorios, tenemos una buena indicación de que el modelo de regresíon se ajusta correctamente.

▶ La recta de regresión para los cuatro siguientes conjuntos de datos es la misma:

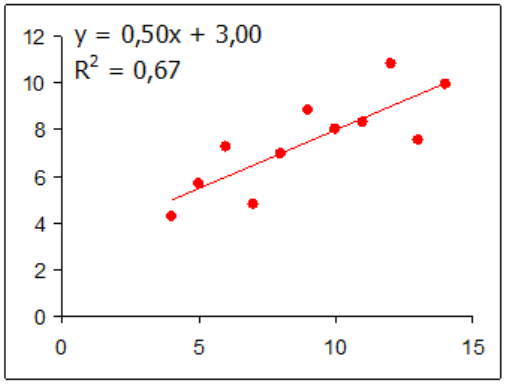
$$y = 0{,}5x + 3{,}00$$

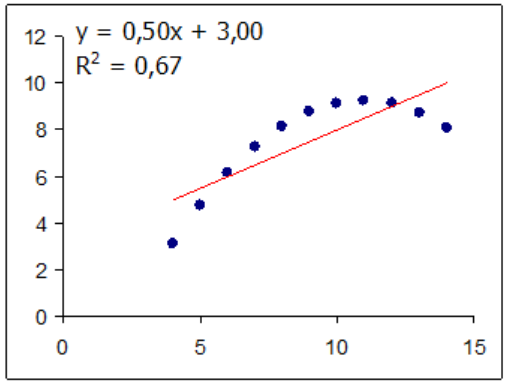
Ejemplo - Datos de Anscombe

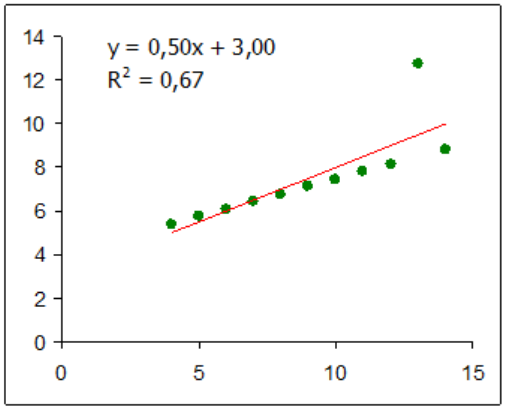
	Datos 1		Datos 2		Datos 3		Datos 4	
i	x	y	x	y	x	y	x	y
1	10	8,04	10	9,14	10	7,46	8	6,58
2	8	6,95	8	8,14	8	6,77	8	5,76
3	13	7,58	13	8,74	13	12,74	8	7,71
4	9	8,81	9	8,77	9	7,11	8	8,84
5	11	8,33	11	9,26	11	7,81	8	8,47
6	14	9,96	14	8,1	14	8,84	8	7,04
7	6	7,24	6	6,13	6	6,08	8	5,25
8	4	4,26	4	3,1	4	5,39	19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

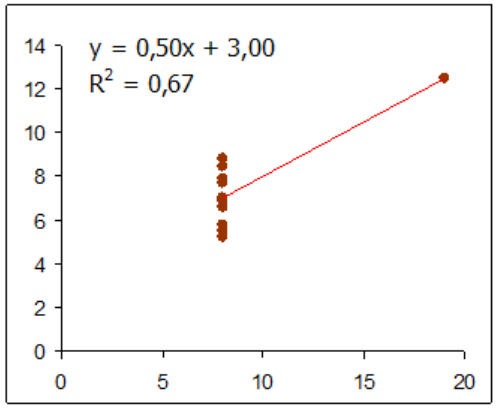
- Los datos de Anscombe son una demostración de la necesidad de interpretar, tanto la recta de regresión como el coeficiente de correlación, después de observar el diagrama de dispersión.

Ejemplo - Datos de Anscombe









Ejemplo - Datos de Anscombe

- El primer caso parece que la recta de regresión es adecuada.
- En el segundo caso, hay una relación no lineal.
- En el tercer gráfico, se ve la influencia de un dato atípico.
- En el último caso, se ve el efecto de un punto influyente.

- ¿Cómo son los gráficos de los residuos frente a las predicciones?

Interpretación y uso de la recta de regresión

- Para los **Datos 1** tenemos

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

- **Interpretación del pendiente**: un incremento de una unidad en x produce, **en promedio**, un incremento de **0,5** en y.

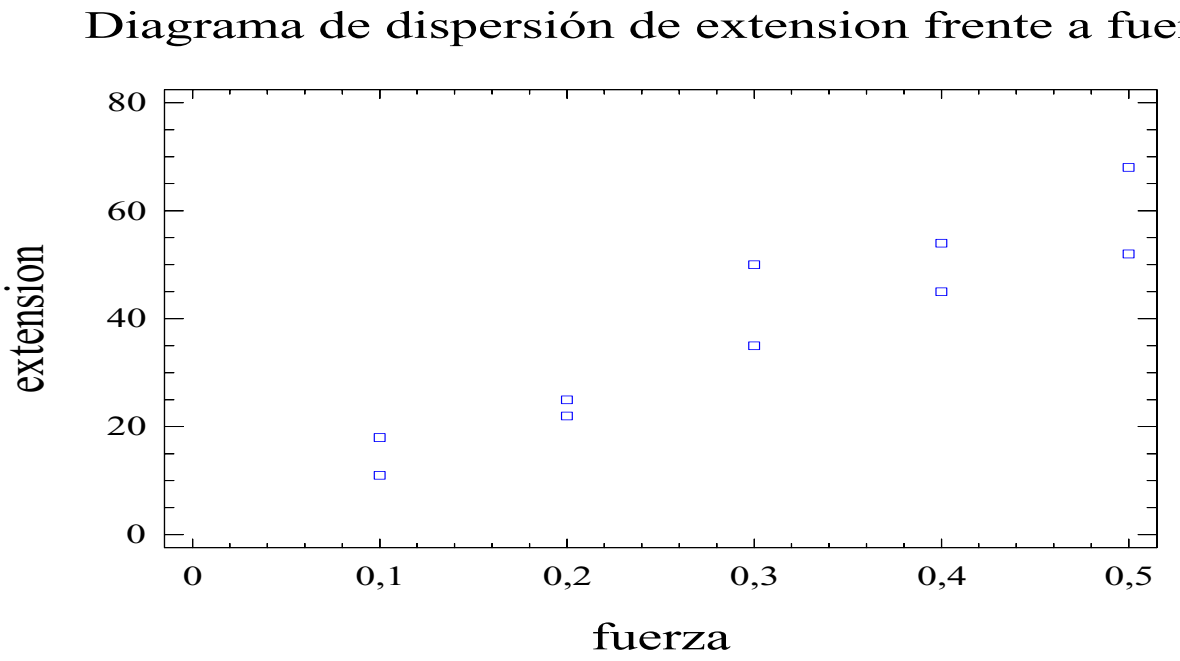
- **Predicción** para $\boldsymbol{x = 7,5}$ (dentro del rango de x):

$$\hat{y} = 3 + 0,5 \times 7,5 \; = \; 6,75.$$

Ejemplo

Se quiere probar la elasticidad de un muelle. Con este objetivo, se sometió el muelle a varios niveles de fuerza (x Newtons) y se midió la extensión total del muelle (y mm) en cada caso.

fuerza	0,1	0,1	0,2	0,2	0,3	0,3	0,4	0,4	0,5	0,5
extensión	18	11	25	22	35	50	54	45	52	68



- El diagrama de dispersión sugiere que existe una relación casi lineal entre fuerza y extensión. Para predecir la extensión del muelle en torno de la fuerza aplicada, aplicamos el model de regresión

$$y = \alpha + \beta x + \varepsilon$$

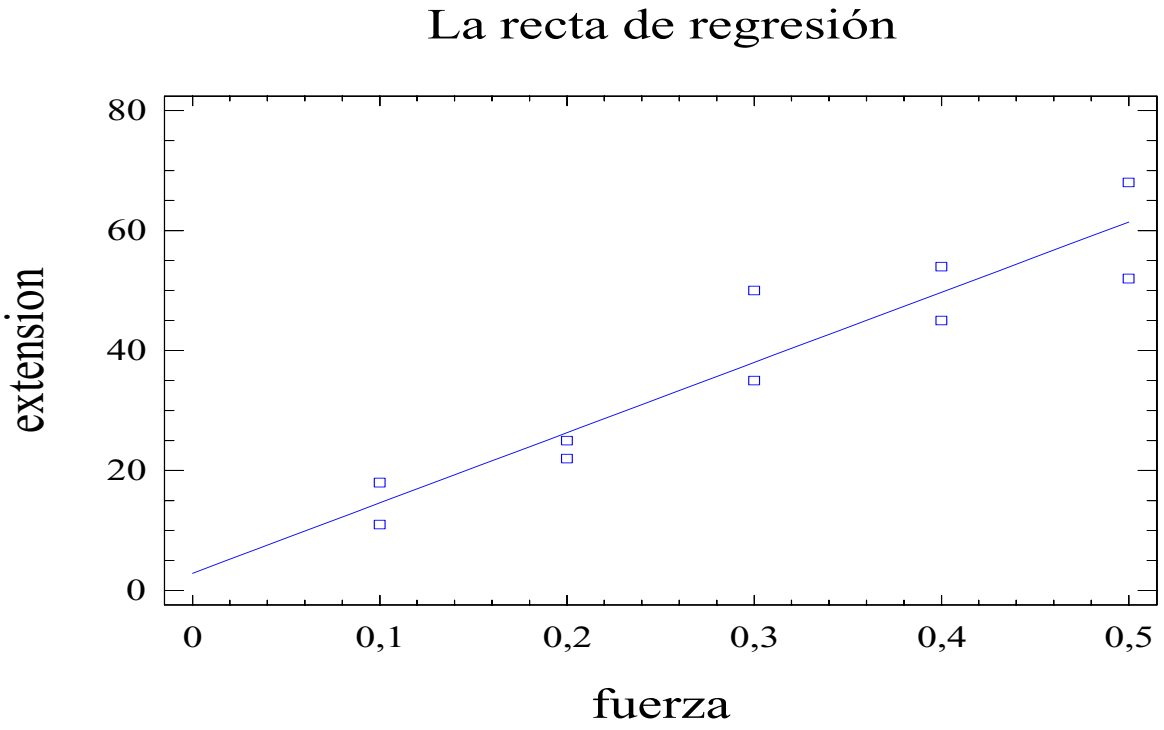
- Dados los datos de la muestra, hallamos la recta estimada por mínimos cuadrados. Tenemos:

$$\begin{aligned} \bar{x} &= 0{,}3 \\ s_x^2 &= 0{,}02 \\ \bar{y} &= 38 \\ s_y^2 &= 310{,}8 \\ s_{xy} &= 2{,}34 \end{aligned}$$

- Calculamos la recta de mínimos cuadrados.

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} = \frac{2{,}34}{0{,}02} = 117 \\ a &= \bar{y} - b\bar{x} = 38 - 117 \times 0{,}3 = 2{,}9 \end{aligned}$$

- La recta ajustada es $y = 2{,}9 + 117x$.



En este ejemplo, podemos predecir que la extensión del muelle si se aplica una fuerza de $0{,}4$ Newtons es:

$$\hat{y} = 2{,}9 + 117 \times 0{,}4 = 49{,}7\text{mm}.$$

¿Qué pasaría si ponemos una fuerza de 0?

- La extensión prevista por la recta de regresión en este caso es de $2{,}9$ mm.
- No obstante el resultado no tiene sentido. Con fuerza 0, la extensión del muelle debe ser cero.
- No es conveniente (arriesgado) hacer predicciones usando valores de x fuera del rango de los datos observados.