3. CORRELACIÓN Y REGRESIÓN

Objetivo

Medir y ajustar una relación lineal entre dos variables cuantitativas.

Bibliografia recomendada

Peña y Romo (1997), Capítulos 8 y 9.

Índice

Covarianza

Se ve en el Ejemplo 63 que existe una relación creciente y más o menos lineal entre el peso pérdido y el peso original de las pacientes. La covarianza es una medida de la fuerza de la relación lineal entre dos variables cuantitativas.

Definición 18 *Para una muestra de n datos bivariantes*

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

la covarianza *entre las dos variables es*

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

donde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ son las medias de ambas variables.

Es ineficiente calcular la covarianza directamente a través de esta definición.

Ejemplo 64 *Volvemos al Ejemplo 63. En primer lugar hallamos las medias de ambas variables.*

$$\begin{aligned} \bar{x} &= \frac{1}{16}(225 + 235 + \ldots + 149) \\ &= 181{,}375 \\ \bar{y} &= \frac{1}{16}(15 + 44 + \ldots + 10) \\ &= 18{,}125 \end{aligned}$$

Luego calculamos la covarianza.

$$\begin{aligned} s_{xy} &= \frac{1}{16}\{(225 - 181{,}375)(15 - 18{,}125) + \\ &\quad (235 - 181{,}375)(44 - 18{,}125) + \ldots + \\ &\quad (149 - 181{,}375)(10 - 18{,}125)\} \\ &\approx 361{,}64 \end{aligned}$$

La covarianza es positiva, que implica una relación creciente entre x e y.

Otra manera de calcular la covarianza

En la práctica, se cálcula la covarianza mediante la siguiente fórmula.

Teorema 5

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

El cálculo a través de este resultado es mucho más rápido, ya que no se tiene que restar las medias de todos los datos.

Demostración

$$
\begin{aligned}
s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}[x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}]\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\bar{y} - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\bar{x}\bar{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - n\bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i - n\bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right)
\end{aligned}
$$

$\diamond$

Ejemplo 65 *Retomando el Ejemplo 63, tenemos*

$$\begin{aligned}\sum_{i=1}^{16} x_i y_i &= 225 \times 15 + 235 \times 44 + \ldots + 149 \times 10 \\ &= 58385 \\ s_{xy} &= \frac{1}{16}(58385 - 16 \times 181{,}375 \times 18{,}125) \\ &= 361{,}64\end{aligned}$$

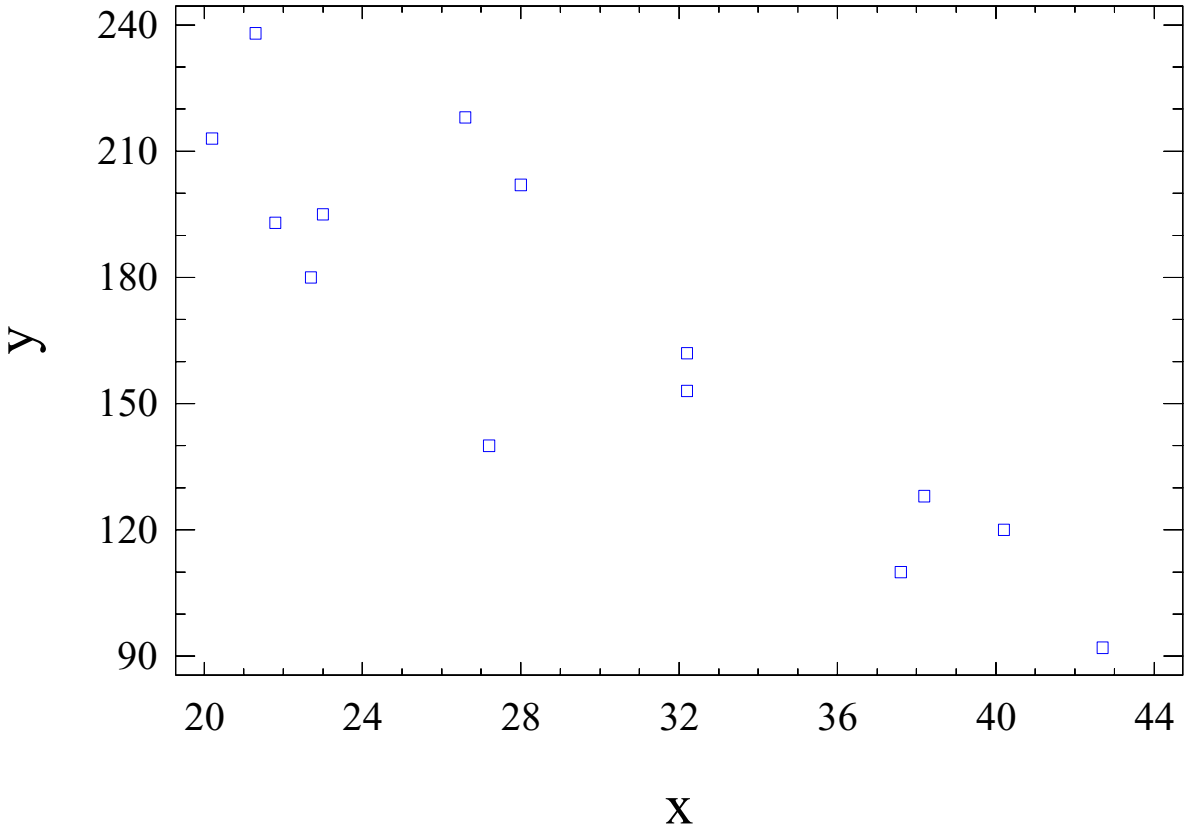
es decir el mismo resultado.

Ejemplo 66 *Se quería determinar la concentración de ácido úrico en la leche de una especie de vaca y se tomo una muestra de 14 vacas. Los datos son producción de leche (x kg/día) y concentración de ácido (y μmol/litro).*

Tiemeyer, Stohrer, W. y Giesecke, D. (1984). Metabolites of nucleic acids in bovine milk. J. Dairy Sci., 67, 723728.

x	42,7	40,2	38,2	37,6	32,2	32,2	28,0
y	92	120	128	110	153	162	202
x	27,2	26,6	23,0	22,7	21,8	21,3	20,2
y	140	218	195	180	193	238	213

Diagrama de dispersión



Vemos que existe una relación negativa entre las dos variables.

Calculamos ahora la covarianza.

Tenemos:

$$\begin{aligned}
\bar{x} &= \frac{1}{14}(42{,}7 + \ldots + 20{,}2) \\
&\approx 29{,}56 \\
\bar{y} &= \frac{1}{14}(92 + \ldots + 213) \\
&\approx 167{,}43 \\
\sum_{i=1}^{14} x_i y_i &= 42{,}7 \times 92 + \ldots + 20{,}2 \times 213 \\
&= 65334{,}2 \\
s_{xy} &= \frac{1}{14}(65334{,}2 - 14 \times 29{,}56 \times 167{,}43) \\
&\approx -283{,}2
\end{aligned}$$

La covarianza es positiva si existe una relación (lineal) creciente y negativa si existe una relación decreciente.

La cuasi covarianza

Igual que con la varianza, en muchos casos, se prefiere definir la covarianza con un denominador de $n-1$, es decir

$$s_{xy}^c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

En este caso, se suele llamar el resultado la **cuasi covarianza**.

Es importante observar que en Statgraphics se emplea esta definición.

Cálculo de la covarianza para datos agrupados

Dada la tabla de doble entrada,

$$
\begin{array}{cc|cccc|c}
 & & & & Y & & \\
 & & y_1 & y_2 & \cdots & y_J & \\
\hline
 & x_1 & f_{11} & f_{12} & \cdots & f_{1J} & f_{1\cdot} \\
 & x_2 & f_{21} & f_{22} & \cdots & f_{2J} & f_{2\cdot} \\
X & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & x_I & f_{I1} & f_{I2} & \cdots & f_{IJ} & f_{I\cdot} \\
\hline
 & & f_{\cdot 1} & f_{\cdot 2} & \cdots & f_{\cdot J} & 1
\end{array}
$$

la media de X es $\bar{x} = \sum_{i=1}^{I} f_{i\cdot} x_i$ con varianza

$$s_x^2 = \sum_{i=1}^{I} f_{i\cdot} x_i^2 - \bar{x}^2.$$

Igualmente se calculan la media y varianza de Y.

Ahora covarianza es

$$s_{xy} = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} x_i y_j - \bar{x}\bar{y}.$$

Ejemplo 67 *En el Ejemplo 57 tuvimos la siguiente tabla de frecuencias relativas.*

		Y				
		5	6	7	8	
	0	,3	,1	,06	,04	,5
	1	,08	,16	,04	,02	,3
X	2	0	,04	,02	,06	,12
	3	0	0	0	,08	,08
		,38	,3	,12	,2	1

y en el Ejemplo 58 demostramos que $\bar{x} = ,78$ *e* $\bar{y} = 6{,}14$*. Ahora, la covarianza es*

$$\begin{aligned} s_{xy} &= \sum_i \sum_j f_{ij} x_i y_j - \bar{x}\bar{y} \\ \sum_i \sum_j f_{ij} x_i y_j &= 0 \times 5 \times ,3 + 0 \times 6 \times ,1 + \ldots + \\ &\quad 3 \times 7 \times 0 + 3 \times 8 \times ,08 \\ &= 5{,}44 \\ s_{xy} &= 5{,}44 - ,78 \times 6{,}14 \\ &= 0{,}6508 \end{aligned}$$

Correlación

Si, por ejemplo las unidades de la variable X son centimetros y las unidades de la variable Y son gramos, entonces las unidades de la covarianza son $cm \times g$ y si cambiamos la escala de las variables, cambia la covarianza. Esto hace que el valor de la covarianza sea difícil de interpretar.

Una medida normalizada es la correlación.

Definición 19 *Para una muestra bivariante*

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

la **correlación** *entre las dos variables es*

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

donde s_x y s_y son las desviaciones típicas y s_x^2 e s_y^2 son las varianzas.

La correlación es independiente de las unidades de las variables.

Propiedades

- $-1 \leq r_{xy} \leq 1$.

- $r_{xy} = 1$ si y sólo si existen constantes α y $\beta > 0$ donde $y_i = \alpha + \beta x_i$ para $i = 1, \ldots n$. Es decir que existe una relación lineal positiva exacta entre las dos variables.

- $r_{xy} = -1$ si y sólo si existen constantes α y $\beta < 0$ donde $y_i = \alpha + \beta x_i$ para $i = 1, \ldots n$. Es decir que existe una relación lineal negativa exacta entre las dos variables.

- Si no existe ninguna relación entre las dos variables, la correlación se aproxima a 0.

Si la correlación está cerca de 1 o -1, entonces hay una relación aproximadamente lineal.

Ejemplo 68 *Retomamos el Ejemplo 66 sobre las vacas.*

Calculamos las medias y la covarianza anteriormente. Ya calculamos las varianzas, desviaciones típicas y la correlación.

$$\begin{aligned} s_x^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n \times \bar{x}^2\right) \\ &= \frac{1}{14}\left(42{,}7^2 + \ldots + 20{,}2^2 - 14 \times 29{,}56^2\right) \\ &\approx 54{,}43 \quad \textit{y de manera parecida,} \\ s_y^2 &\approx 1868{,}82. \end{aligned}$$

Entonces la correlación es

$$r_{xy} = \frac{-283{,}2}{\sqrt{54{,}43 \times 1868{,}82}} \approx -0{,}89$$

Existe una relación negativa aproximadamente lineal entre las dos variables.

Ejemplo 69 *Volvemos al Ejemplo 63 sobre los diabéticos. Calculamos la covarianza como* $s_{xy} =$ 361,64 *en el Ejemplo 64. Ahora, hallamos las varianzas y la correlación.*

Calculamos que $s_x^2 \approx$ 1261,98 *y* $s_y^2 \approx$ 211,23 *y luego* $s_x \approx$ 35,52 *y* $s_y \approx$ 14,53.

Entonces $r_{xy} = \frac{361,64}{35,52 \times 14,53} \approx 0,70$.

Hay una relación lineal positiva bastante fuerte entre las dos variables.

Ejemplo 70 *En el Ejemplo 67, calculamos la covarianza entre el número de suspensos en Introducción a la Estadística y el número de años en la licenciatura.*

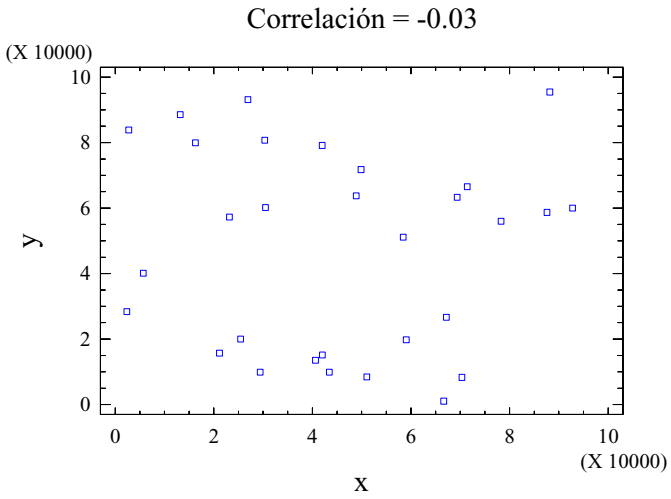
Recordando que las desviaciones típicas son $s_x = 0,9442$ *y* $s_y = 1,1315$, *la correlación es*

$$r_{xy} = \frac{0,6508}{0,9442 \times 1,1315} \approx 0,61.$$

Hay una correlación positiva entre las dos variables.

Si no hay relación entre las variables, la correlación es aproximadamente cero

Ejemplo 71 *Los datos son 30 parejas de números aleatorios.*

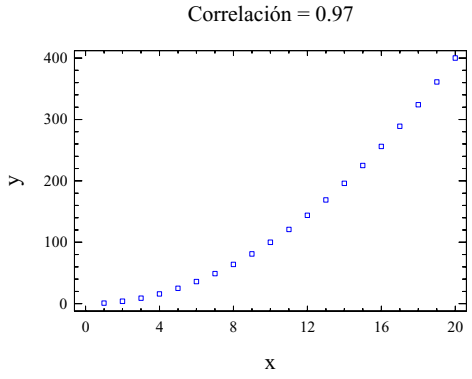


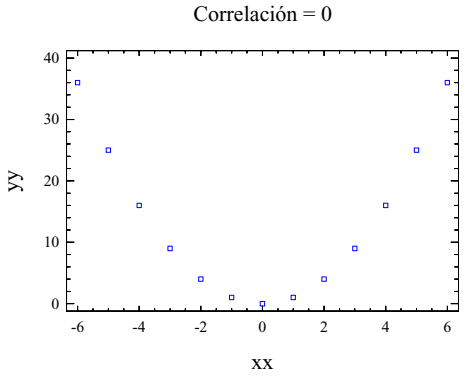
La correlación es casi cero.

Al revés no es verdad.

¡Ojo! Cero correlación no implica ninguna relación

Se ha visto que si hay una relación más o menos lineal, la correlación entre las dos variables es bastante alta pero ¿Qué pasa si hay una relación no lineal?





En ambas gráficas se ha utilizado la fórmula $y = x^2$ para generar los datos. Una fuerte relación no lineal.