RELACIONES BIVARIADAS

Penna – Cobos – Vázquez Ferrero – Ulagnero

RELACIONES BIVARIADAS

El objetivo del presente tema es introducir el análisis simultáneo de dos variables cuantitativas y para ello nos preguntamos:

- Si se conoce el comportamiento de una de ellas, ¿se puede predecir el comportamiento de la otra?
- ¿Existe alguna relación entre las variables?

La estadística aplicada ofrece dos herramientas que permiten dar respuesta a dichas cuestiones: el *Análisis de Regresión lineal* y el *Análisis de Correlación lineal*.

Regresión lineal

Estudia la relación funcional que existe entre dos variables. Identifica el **modelo o función** lineal que une a las variables, estima sus parámetros y, eventualmente, prueba hipótesis acerca de ellos. Una vez estimado el modelo es posible predecir el valor de la **variable dependiente** en función de la **variable independiente**.

Correlación lineal

Estudia el **grado** y **sentido** de la asociación lineal que hay entre variables y, a diferencia del análisis de regresión, no se identifica ni se estima explícitamente un modelo funcional para las variables, pues este se supone lineal. El interés principal es **medir la asociación** entre dos variables aleatorias cualesquiera, sin necesidad de distinguir variables dependientes e independientes, sólo enfatiza la forma en que se comporta una variable en relación a la otra y se centra en medir la **intensidad** de esta asociación.

REGRESIÓN LINEAL SIMPLE

Al estudiar la relación entre dos variables surge la idea de encontrar una expresión matemática que la describa. Si se denota como *y* a la variable que se supone **dependiente** y como *x* a la variable que se postula como **independiente,** resulta familiar utilizar el concepto de función y decir "*y es función de x*", para indicar que de acuerdo a los valores asignados a *x* se pueden **predecir** los valores que tomará *y*. Dicho de otra manera, se puede conocer el comportamiento de *y* a través de un modelo que relaciona la variación en *y* con la variación de *x*.

El análisis de regresión tiene por objetivo **identificar** un modelo funcional que describa cómo varía, en promedio, la variable dependiente, *y*, frente a cambios en *x*. El modelo para *y* presenta constantes desconocidas que se llaman parámetros, por lo que otro objetivo del análisis es la **estimación** de los parámetros a partir de una muestra aleatoria de observaciones en *y* y en *x*. El análisis de regresión se ocupa también de la **validación** del modelo propuesto y de las **pruebas de hipótesis** sobre los parámetros del modelo; por último, la modelación por regresión también tiene como objetivo la **predicción**, es decir el uso del modelo para dar el valor esperado de *y* cuando *x* toma un valor particular.

Regresión lineal simple (cont.)

El modelo de regresión lineal más sencillo es el que se presenta en la siguiente definición:

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

y = variable dependiente o respuesta
x = variable independiente o regresora
α = ordenada al origen
β = pendiente
ε = variable aleatoria no observable (o no explicable por el modelo) ~ $N(0; \sigma)$

El modelo anterior incluye solamente una variable independiente y establece que la variable dependiente cambia con tasa constante, según crece o decrece el valor de la variable independiente. Siendo la recta estimada:

$$\hat{y} = a + bx$$

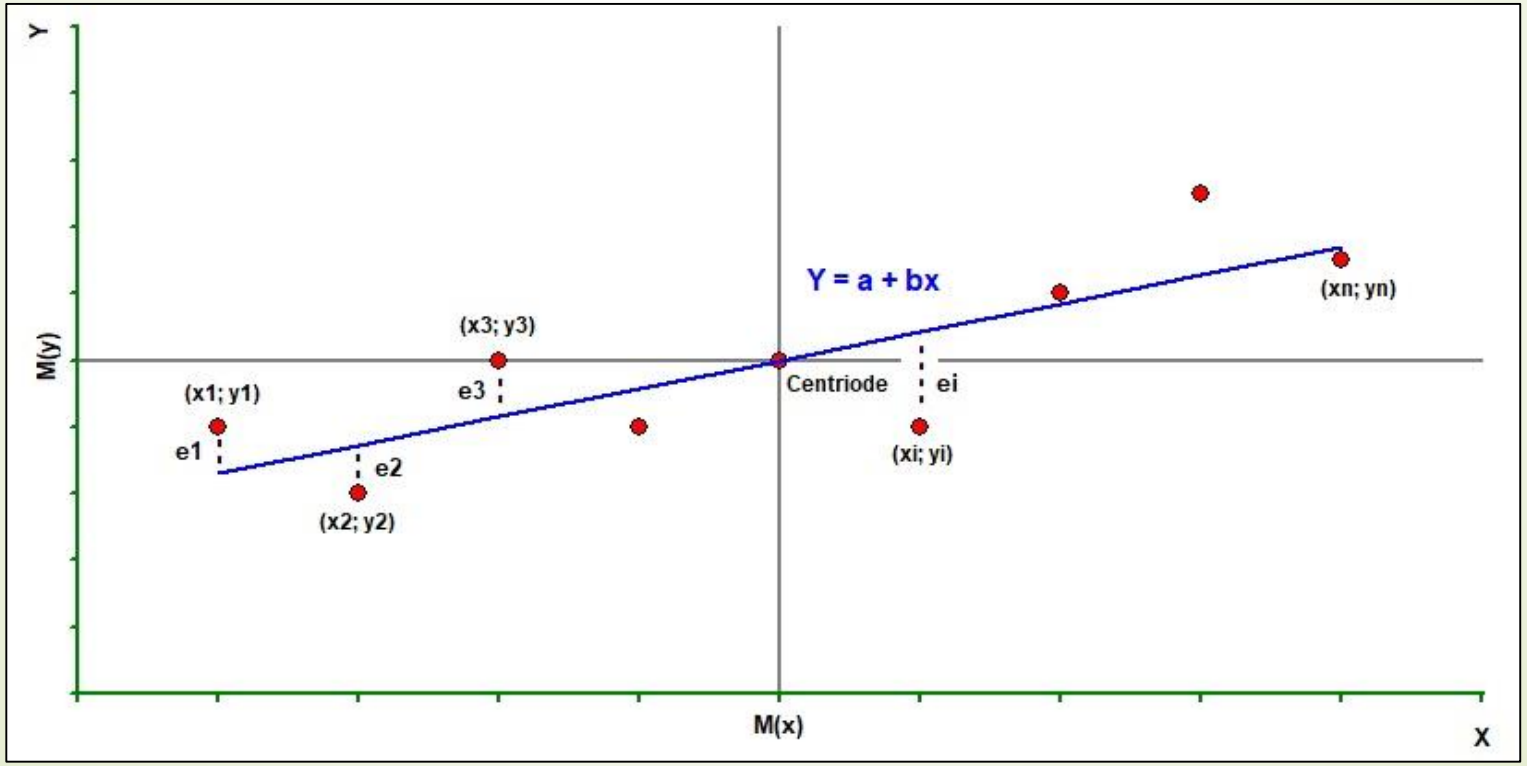
$\hat{y}$ = variable dependiente o respuesta
x = variable independiente o regresora
a = ordenada al origen
b = pendiente

REGRESIÓN LINEAL SIMPLE (CONT.)

Donde, por el método de mínimos cuadrados, obtenemos los coeficientes de regresión:

$$\mathbf{b} = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n} \qquad \mathbf{a} = \bar{y} - b\bar{x}$$

Gráficamente:



REGRESIÓN LINEAL SIMPLE (CONT.)

PRUEBA DE HIPÓTESIS PARA β

Hemos visto como estimar los parámetros de un modelo de regresión lineal simple y estos son: la ordenada al origen (α) y la pendiente (β). Abordaremos la problemática de la prueba de hipótesis sobre $\beta = \beta_0$

En primer lugar, vamos a estimar la varianza de la pendiente:

Sea: $S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-2} \Rightarrow \mathbf{V(b)} = \frac{S_e^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}$

Y ahora vamos a considerar la prueba de hipótesis para β (sólo para el caso $\beta_0 = 0$)

$H_0: \beta = 0$ vs. $H_1: \beta \neq 0$ Siendo la variable pivotal: $\mathbf{t_{H_0}} = \frac{b - \beta_0}{\sqrt{V(b)}} = \frac{b}{\sqrt{V(b)}} \sim t_{n-2}$

REGRESIÓN LINEAL SIMPLE (CONT.)

COEFICIENTE DE DETERMINACIÓN

Una medida muestral de la capacidad predictiva del modelo es el coeficiente de determinación, denotado por $\mathbf{R^2}$.

$$\mathbf{R^2} = \frac{\left[\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n\right]^2}{\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n\right] \times \left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2/n\right]} \qquad \mathbf{0 \leq R^2 \leq 1}$$

Este coeficiente se interpreta como la proporción de la variabilidad total en *y* explicable por la variación de *x* o como la proporción de la variabilidad total explicada por el modelo. Cuanto más próximo esté a 1, mayor valor predictivo tendrá el modelo en el sentido que los valores observables estarán muy próximos a la esperanza estimada por la regresión.

Es frecuente ver al $\mathbf{R^2}$ usado como una medida de adecuación del modelo, es decir que la relación funcional y los supuestos sobre los errores son correctos. Esto es **absolutamente incorrecto** ya que $\mathbf{R^2}$ puede ser alto y el modelo ser inapropiado. Luego, $\mathbf{R^2}$ es válido como medida de ajuste o valor predictivo si el modelo es correcto tanto en su parte determinística como en su parte aleatoria.

Regresión lineal simple (ejemplo)

Se llevó a cabo un estudio para determinar la relación entre la antigüedad laboral –en años– (X) y el sueldo mensual –en miles de pesos– (Y) entre docentes de una determinada escuela. Para ello se tomó una muestra aleatoria de 17 enseñantes, obteniendo los siguientes datos:

Antigüedad	13	16	30	2	8	6	31	19	20	1	4	10	27	25	7	15	13
Sueldo	26	33	36	16	26	19	36	34	37	17	29	25	36	37	21	31	31

a) Realice un dispersograma.

b) En el supuesto de una relación lineal, utilizando el método de mínimos cuadrados, calcule los coeficientes de regresión.

c) Interprete el significado de pendiente.

d) ¿La pendiente es significativa al 1%?

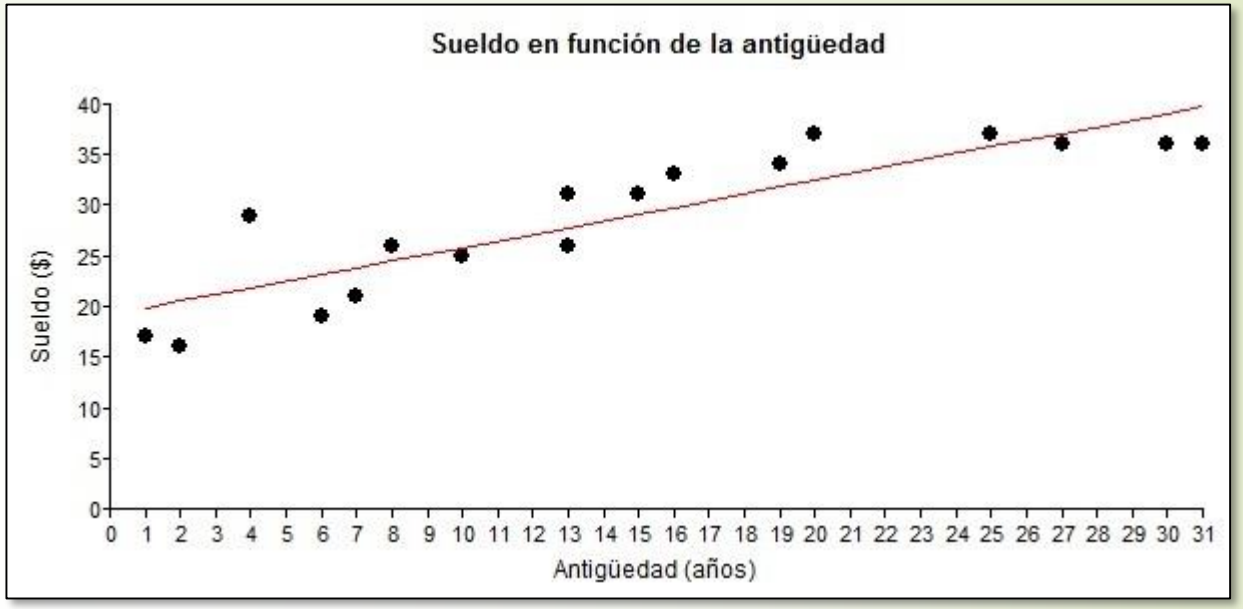
e) ¿Qué sueldo mensual (en promedio) predeciría usted que va a ganar un docente, con 18 años de antigüedad laboral?

f) Calcule el coeficiente de determinación, R^2. ¿Cómo interpretaría el resultado obtenido?

REGRESIÓN LINEAL SIMPLE (EJEMPLO)

Solución:

a) Dispersograma



b) Para calcular los coeficientes de regresión estimados (a y b), necesitamos los siguientes datos parciales:

$$\sum_{i=1}^{17} x_i = 247; \quad \sum_{i=1}^{17} y_i = 490; \quad \sum_{i=1}^{17} x_i\, y_i = 8097; \quad \sum_{i=1}^{17} x_i^2 = 5065; \quad \sum_{i=1}^{17} y^2 = 14958$$

$$\Rightarrow b = \frac{8097 - 247 \times 490/17}{5065 - (247)^2/17} = \frac{977{,}59}{1476{,}24} = \mathbf{0,66} \qquad a = \frac{490}{17} - 0{,}66 \times \frac{247}{17} = \mathbf{19,23}$$

$$\Rightarrow \mathbf{Sueldo} = \mathbf{19,23} + \mathbf{0,66} \times (\mathbf{Antigüedad})$$

REGRESIÓN LINEAL SIMPLE (EJEMPLO)

c) De acuerdo al modelo presentado en el punto anterior y siendo la pendiente positiva e igual a 0,66, podemos decir que –en promedio– por cada año de antigüedad el docente incrementará su salario mensual en $660.

d) $H_0: \beta = 0$ vs. $H_1: \beta \neq 0 \quad (\alpha = 0{,}01)$

$$S_e^2 = \frac{14958 - 19{,}23 \times 490 - 0{,}66 \times 8097}{15} = 12{,}75 \Rightarrow V(b) = \frac{17 \times 12{,}75}{17 \times 5065 - (247)^2} = \frac{216{,}75}{25096} = 0{,}0086$$

$$\Rightarrow t_{H_0} = \frac{0{,}66}{\sqrt{0{,}0086}} \cong 7{,}12$$

Buscando, con 15 gl, el 7,12 –o el número más próximo– en la tabla 2, obtenemos que, por ser bilateral, $p_v \cong 2 \times 0{,}0005 = 0{,}001 < \alpha = 0{,}01 \Rightarrow$ rechazamos H_0, la pendiente es significativa al 1%.

gl \ α	0,25	0,20	0,15	0,10	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015

REGRESIÓN LINEAL SIMPLE (EJEMPLO)

e) Cómo:

$$\text{Sueldo} = 19{,}23 + 0{,}66 \times (\text{Antigüedad}) \Rightarrow \text{Sueldo}_{(18)} = 19{,}23 + 0{,}66 \times 18 = \mathbf{31{,}11}$$

Esto quiere decir que el sueldo mensual promedio de un docente con 18 años de antigüedad laboral será de, aproximadamente, $31.110

f) $$R^2 = \frac{[8097 - 247 \times 490/17]^2}{[5065 - (247)^2/17] \times [14958 - (490)^2/17]} = \frac{(977{,}588)^2}{1476{,}235 \times 834{,}471} = \frac{955678{,}298}{1231875{,}297} = \mathbf{0{,}7758}$$

Esto nos dice que el 77,58% del comportamiento de los datos está explicado por la recta; el resto (22,42%), de dicho comportamiento, se debe al azar.

InfoStat

CORRELACIÓN LINEAL SIMPLE

En el análisis de regresión, la variable *x* es usualmente fija, mientras que la variable dependiente *y* es aleatoria. Si *x* e *y* son **ambas** variables aleatorias observables sobre una misma unidad o elemento de la población, podría ser de interés medir el grado en que estas variables covarían ya sea positiva o negativamente.

La simple observación de que dos variables parecen estar relacionadas, no revela gran cosa. Dos importantes preguntas se pueden formular al respecto:

a) ¿*Qué tan estrechamente relacionadas se encuentran las variables*? o ¿*cuál es el grado de asociación que existe entre ambas*? [Para responder esta pregunta se necesita una medida del grado de asociación entre las dos variables. Esta medida es el **coeficiente de correlación,** que se denota con la letra griega ρ (rho)].

b) ¿*Es real la asociación observada o podría haber ocurrido solo por azar*? [Para este caso, se precisa una prueba estadística de hipótesis para ρ].

CORRELACIÓN LINEAL SIMPLE (CONT.)

Siendo el estimador del coeficiente de correlación:

$$\mathbf{r_{xy}} = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n\right] \times \left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2/n\right]}}$$

Dicho coeficiente puede ser calculado bajo los siguientes supuestos:

a) Linealidad: la relación entre las dos variables tiene que ser lineal.

b) Homocedasticidad (homogeneidad de varianzas): las varianzas de los grupos tienen que ser homogéneas.

c) Normalidad: las muestras deben provenir de poblaciones distribuidas normalmente.

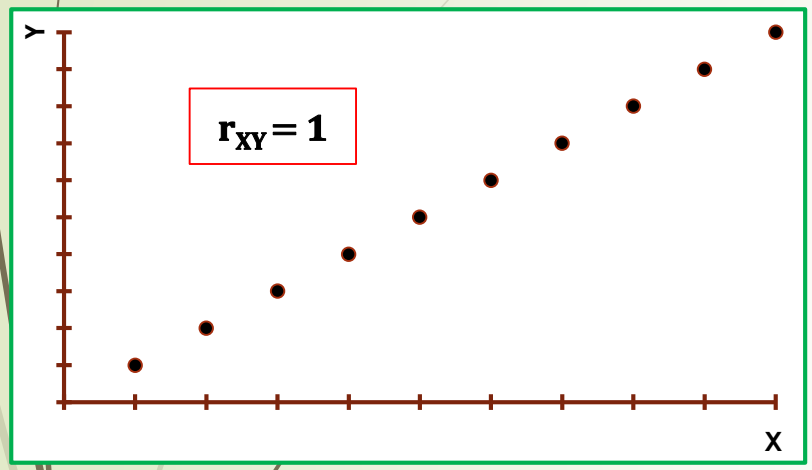
CORRELACIÓN LINEAL SIMPLE (CONT.)

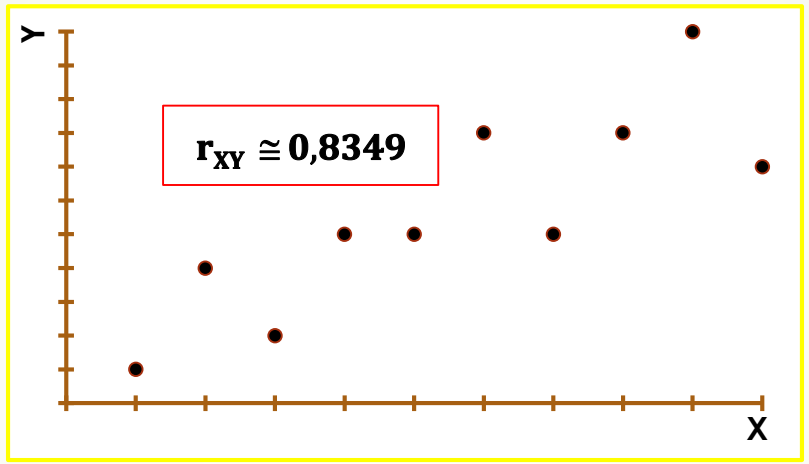
Siendo las características principales del coeficiente r_{xy}, las siguientes:

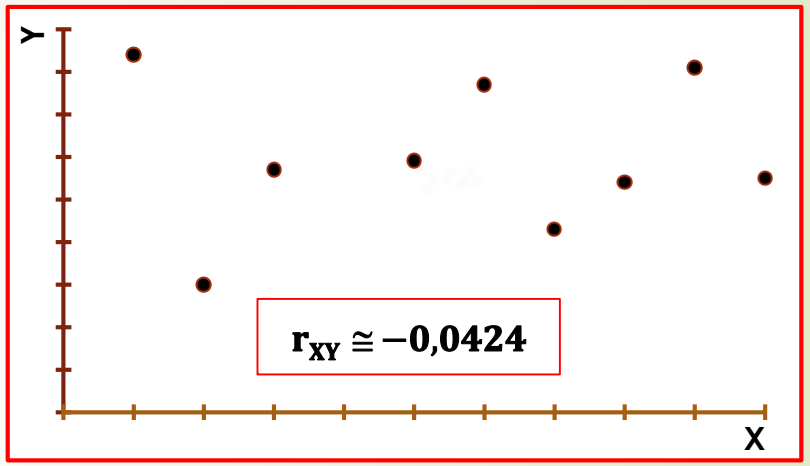
- Es un número adimensional (sin unidades).
- Su valor no puede superar +1 ni ser inferior a –1, es decir: $\mathbf{-1 \leq r_{xy} \leq +1}$.
- Si el signo es (+), las dos variables tienden a variar en el mismo sentido, o sea, si se incrementa el valor de una de ellas, aumenta el de la otra. Si el signo es (–), las variables varían en sentido contrario, o sea, si se incrementa el valor de una, disminuye la otra.
- La relación entre ambas variables es más estrecha, cuanto el valor del coeficiente r_{xy} se acerque a +1 ó –1; por el contrario, si $r_{xy} \rightarrow 0$ (o a un entorno próximo a cero), las mismas tienden a ser independientes o el comportamiento no es lineal.
- Si la relación es perfecta, r_{xy} será igual a +1 ó –1, según sea positiva o negativa la relación; si no hay relación, r_{xy} deberá ser cero.
- El valor de r_{xy} no está influido por el "tamaño" de las unidades de medida empleadas para medir las variables en estudio. Como consecuencia, si previo a los cálculos se simplifican o redondean las cifras, r_{xy} no variará significativamente.

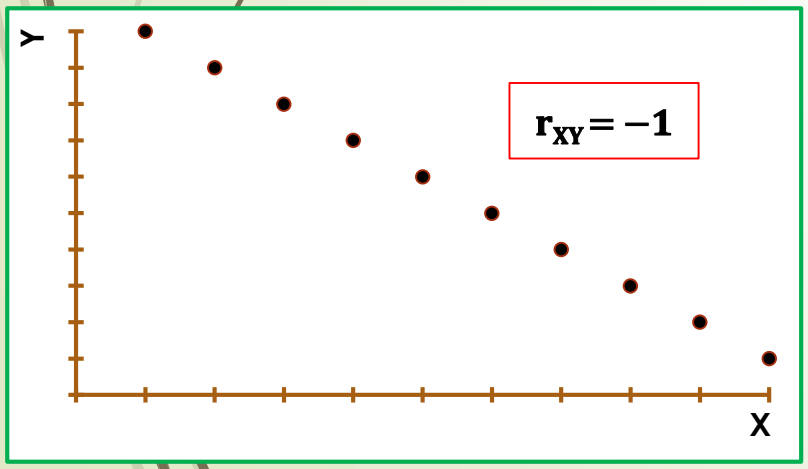
CORRELACIÓN LINEAL SIMPLE (CONT.)

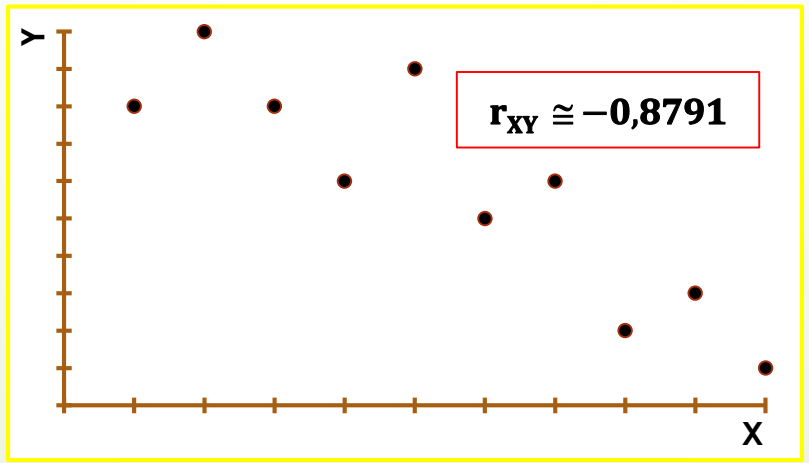
Gráficamente:

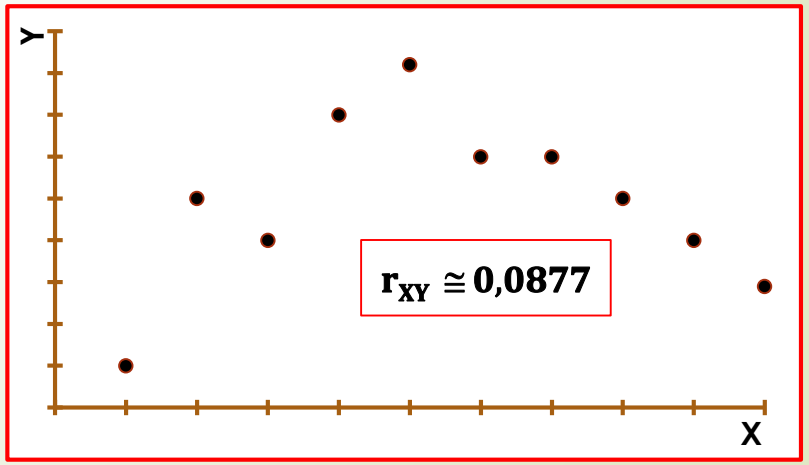












CORRELACIÓN LINEAL SIMPLE (CONT.)

PRUEBA DE HIPÓTESIS PARA ρ

Si se satisfacen las suposiciones de normalidad bivariada y se tiene una muestra aleatoria de n pares de valores (x_i, y_i), es posible utilizar el coeficiente de correlación muestral r_{xy}, para probar la independencia entre las variables x e y, siendo la hipótesis:

$$H_0: \rho = \rho_0 \text{ vs. } H_1: \rho \neq \rho_0$$

a) Para $\rho_0 = 0$

Siendo la variable pivotal o estadístico de prueba: $t_{H_0} = \dfrac{r_{xy}}{\sqrt{\dfrac{1 - r_{xy}^2}{n-2}}} = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} \sim t_{n-2}$

b) Para $\rho_0 \neq 0$

Siendo la variable pivotal o estadístico de prueba: $Z_{H_0} = \dfrac{Z - \mu_Z}{\sigma_Z} \sim N(0,1)$

Donde: $Z = \frac{1}{2}\ln\left(\frac{1 + r_{xy}}{1 - r_{xy}}\right)$; $\mu_Z = \frac{1}{2}\ln\left(\frac{1 + \rho_0}{1 - \rho_0}\right)$; $\sigma_Z = \sqrt{\frac{1}{n-3}}$

CORRELACIÓN LINEAL SIMPLE (EJEMPLO)

A doce alumnos de un centro de estudios se les preguntó a qué distancia (X) estaba su residencia del Instituto, con fin de estudiar si esta variable estaba relacionada con la nota media (Y) obtenida. Se obtuvieron los datos que figuran en la siguiente tabla:

Distancia (en km)	0,05	0,1	0,12	0,4	0,5	0,7	1	1,2	2,1	2,5	3	3
Nota media	8,4	4,0	5,7	9,1	6,3	6,7	4,3	5,4	7,8	4,5	7,2	8,1

Siendo algunos de los resultados parciales, los siguientes:

$$\sum_{i=1}^{12} x_i = 14{,}67; \quad \sum_{i=1}^{12} y_i = 77{,}5; \quad \sum_{i=1}^{12} x_i\, y_i = 97{,}29; \quad \sum_{i=1}^{12} x_i^2 = 32{,}03; \quad \sum_{i=1}^{12} y^2 = 532{,}63$$

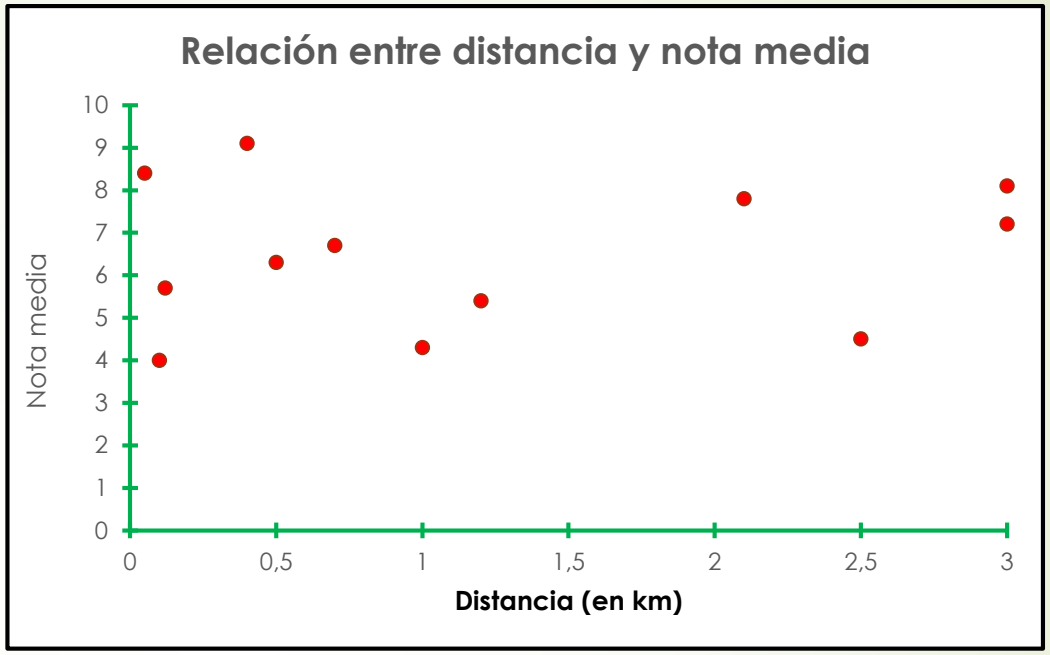
a) Realice un dispersograma de las variables X e Y.
b) Calcule el coeficiente de correlación lineal. ¿Cómo interpretaría el resultado obtenido, en función de las variables estudiadas?
c) ¿Es significativo, al 5%, dicho coeficiente? Interprete.

Nota; suponga que las variables presentadas son homocedásticas y provienen de poblaciones normalmente distribuidas.

CORRELACIÓN LINEAL SIMPLE (CONT.)

Solución:

a) Dispersograma



b) Cálculo de r_{xy} e interpretación:

$$r_{xy} = \frac{97{,}29-(14{,}67)(77{,}5)/12}{\sqrt{[32{,}03-(14{,}67)^2/12]\times[532{,}63-(77{,}5)^2/12]}} = \frac{2{,}546}{\sqrt{14{,}096\times 32{,}109}} = \frac{2{,}546}{21{,}275} \cong \mathbf{0,1197}$$

Siendo $r_{xy} \cong 0{,}1197$ –aunque positivo– muy próximo a cero, no estamos en condiciones de asegurar que ambas variables (distancia y nota media) están correlacionadas. Por ello, tendremos que realizar una Prueba de Hipótesis para ρ.

CORRELACIÓN LINEAL SIMPLE (CONT.)

c) Prueba de hipótesis para ρ (con $\rho_0 = 0$)

$H_0: \rho = 0$ vs. $H_1: \rho \neq 0 \quad (\alpha = 0{,}05)$

$$t_{H_0} = \frac{0{,}1197\sqrt{10}}{\sqrt{1-0{,}0143}} = \frac{0{,}3785}{0{,}9928} \cong 0{,}3812$$

InfoStat

Buscando, con 10 gl, 0,3812 –o el número más próximo– en la tabla 2, obtenemos que, por ser una prueba bilateral, $p_v \cong 2 \times 0{,}25 = 0{,}5 > \alpha = 0{,}05 \Rightarrow$ no rechazamos H_0. Es independiente, al 5%, la nota media obtenida por los alumnos respecto a la distancia en la que viven.

gl \ α	0,25	0,20	0,15	0,10	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624		
15	0,691	0,866	1,074	1,341	1,753				
16	0,690	0,865	1,071	1,337					

Referencias Bibliográficas

- Bologna, E. (2011). *Estadística para Psicología y Educación*. Córdoba: Brujas.
- Gorgas García, J., Cardiel López, N. & Zamorano Calvo, J. (2009). *Estadística Básica para Estudiantes de Ciencias*. Madrid: Departamento de Astrofísica y Ciencias de la Atmósfera. Facultad de Ciencias Físicas. Universidad Complutense de Madrid.
- Hernández Sampieri R., Fernández–Collado C. & Baptista Lucio P. (2010). *Metodología de la investigación* (6ª ed.). México: McGraw-Hill Interamericana.
- Penna, F.O., Esteva, G.C., Cobos, O.H. & Ulagnero, C.A. (2018). *Fórmulas y Tablas III (para cursos de Estadística básica)* (2ª ed.). San Luis: Nueva Editorial Universitaria.
- Triola M. (2018). *Estadística* (12ª ed.). México: Pearson Educación.
- Wackerly, D.D., Mendenhall, W, & Scheaffer, R.L. (2010). *Estadística matemática con aplicaciones* (7ª ed.). México: Cengage Learning Editores S.A. de C.V.