EJERCICIOS RESUELTOS DE VARIABLE ESTADÍSTICA BIDIMENSIONAL

1. Dada la variable estadística bidimensional (X, Y) con la tabla de frecuencias

X \ Y	1	2	4	6
1	2	0	1	1
3	3	1	0	1
5	0	1	0	5

Se pide:

a) $\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}$ b) f_{23}, f_{34}, f_{21} c) $\sum_{i=1}^{3} n_{i\bullet}$ y $\sum_{j=1}^{4} n_{\bullet j}$ d) $f(x_i / Y=2)$ y $f(y_j / X=3)$

e) a_{10} y a_{01} f) a_{11} g) s_{xy}

Solución:

a)

$$\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij} = \sum_{i=1}^{3}\left[n_{i1}+n_{i2}+n_{i3}+n_{i4}\right] = \left[n_{11}+n_{12}+n_{13}+n_{14}\right]+\left[n_{21}+n_{22}+n_{23}+n_{24}\right]+\left[n_{31}+n_{32}+n_{33}+n_{34}\right] =$$

$$=\left[2+0+1+1\right]+\left[3+1+0+1\right]+\left[0+1+0+5\right]=15$$

b) Cada n_{ij} representa la frecuencia absoluta del par (x_i, y_j), la frecuencia relativa se define $f_{ij}=\frac{n_{ij}}{N}$,

donde $N=\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}=15$

$$f_{23}=\frac{n_{23}}{N}=\frac{0}{15}=0 \qquad f_{34}=\frac{n_{34}}{N}=\frac{5}{15} \qquad f_{21}=\frac{n_{21}}{N}=\frac{3}{15}$$

c)

X \ Y	1	2	4	6	$n_{i\bullet}$
1	2	0	1	1	4
3	3	1	0	1	5
5	0	1	0	5	6
$n_{\bullet j}$	5	2	1	7	15

$$\sum_{i=1}^{3} n_{i\bullet}=\left[n_{1\bullet}+n_{2\bullet}+n_{3\bullet}\right]=\left[4+5+6\right]=15=\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}$$

$$\sum_{j=1}^{4} n_{\bullet j}=\left[n_{\bullet 1}+n_{\bullet 2}+n_{\bullet 3}+n_{\bullet 4}\right]=\left[5+2+1+7\right]=15=\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}$$

d)

X \ Y	1	2	4	6	$n_{i\bullet}$
1	2	0	1	1	4
3	3	1	0	1	5
5	0	1	0	5	$n_{3\bullet}=6$
$n_{\bullet j}$	5	$n_{\bullet 2}=2$	1	7	15

Las frecuencias relativas condicionadas $f(x_i/Y=2)$ y $f(y_j/X=3)$:

X	$n(x_i/Y=2)$	$f(x_i/Y=2)=\frac{n(x_i/Y=2)}{n_{\bullet 2}}$
1	0	0
2	1	1/2
3	1	1/2
	$n_{\bullet 2}=2$	1

Y	$n(y_j/X=3)$	$f(y_j/X=3)=\frac{n(y_j/X=3)}{n_{3\bullet}}$
1	0	0
2	1	1/6
4	0	0
6	5	5/6
	$n_{3\bullet}=6$	1

e)

$$a_{10}=\frac{\sum_{i=1}^{3}\sum_{j=1}^{4}x_i n_{ij}}{N}=\frac{\sum_{i=1}^{3}x_i\left[n_{i1}+n_{i2}+n_{i3}+n_{14}\right]}{N}=\frac{1}{N}\Big(\left[x_1 n_{11}+x_1 n_{12}+x_1 n_{13}+x_1 n_{14}\right]+$$

$$+\left[x_2 n_{21}+x_2 n_{22}+x_2 n_{23}+x_2 n_{24}\right]+\left[x_3 n_{31}+x_3 n_{32}+x_3 n_{33}+x_3 n_{34}\right]\Big)=$$

$$=\frac{\left[1.2+1.0+1.1+1.1\right]+\left[3.3+3.1+3.0+3.1\right]+\left[5.0+5.1+5.0+5.5\right]}{15}=\frac{49}{15}=3{,}26$$

o también, $a_{10}=\frac{\sum_{i=1}^{3}x_i n_{i\bullet}}{N}=\frac{1.4+3.5+5.6}{15}==\frac{49}{15}=3{,}26$

$$a_{01}=\frac{\sum_{j=1}^{4}y_j n_{\bullet j}}{N}=\frac{1.5+2.2+4.1+6.7}{15}=\frac{55}{15}=3{,}6$$

f)

$$a_{11}=\frac{\sum_{i=1}^{3}\sum_{j=1}^{4}x_i\,y_j\,n_{ij}}{N}=$$

$$= \frac{[1.1.2+1.2.0+1.4.1+1.6.1]+[3.1.3+3.2.1+3.4.0+3.6.1]+[5.1.0+5.2.1+5.4.0+5.6.5]}{15} = \frac{205}{15} = 13{,}66$$

g) $s_{xy} = a_{11} - a_{10}\, a_{01} = 13{,}66 - 3{,}26.3{,}6 = 1{,}924$

2. Las calificaciones obtenidas por un grupo de alumnos en Estadística (E) y Macroeconomía (M):

E	3	4	6	7	5	8	7	3	5	4	8	5	5	8	8	8	5
M	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

a) Hallar la tabla de frecuencias
b) Hallar las distribuciones marginales, media y varianza de las mismas
c) Covarianza

Solución:

a) La variable E (Estadística) toma seis valores diferentes. La variable M (Macroeconomía) toma siete valores distintos, por lo que para formar la tabla bastará hacer el recuento de las veces que se repite cada par.

E \ M	4	5	6	7	8	9	10	$n_{i\bullet}$
3	1	1						2
4	1	1						2
5		1		4				5
6					1			1
7				1			1	2
8		1				2	2	5
$n_{\bullet j}$	2	4	0	5	1	2	3	17

b)

E_i	$n_{i\bullet}$	$E_i\, n_{i\bullet}$	$E_i^2\, n_{i\bullet}$
3	2	6	18
4	2	8	32
5	5	25	125
6	1	6	36
7	2	14	98
8	5	40	320
	17	99	629

M_j	$n_{\bullet j}$	$M_j\, n_{\bullet j}$	$M_j^2\, n_{\bullet j}$
4	2	8	32
5	4	20	100
6	0	0	0
7	5	35	245
8	1	8	64
9	2	18	162
10	3	30	300
	17	119	903

- Distribución Marginal de Estadística:

$$\overline{E} = a_{10} = \frac{\sum_{i=1}^{6} E_i\, n_{i\bullet}}{N} = \frac{99}{17} = 5{,}82 \qquad a_{20} = \frac{\sum_{i=1}^{6} E_i^2\, n_{i\bullet}}{N} = \frac{629}{17} = 37 \qquad s_E^2 = a_{20} - a_{10}^2 = 37 - 5{,}82^2 = 3{,}13$$

- Distribución Marginal de Macroeconomía:

$$\overline{M} = a_{01} = \frac{\sum_{j=1}^{7} M_j n_{\bullet j}}{N} = \frac{119}{17} = 7 \qquad a_{02} = \frac{\sum_{j=1}^{7} M_j^2 n_{\bullet j}}{N} = \frac{903}{17} = 53{,}11 \qquad s_M^2 = a_{02} - a_{01}^2 = 53{,}11 - 7^2 = 4{,}11$$

c) Para hallar la covarianza: $s_{xy} = a_{11} - a_{10}\, a_{01}$

$$a_{11} = \frac{\sum_{i=1}^{6}\sum_{j=1}^{7} E_i M_j n_{ij}}{N} = \frac{3.4.1 + 3.5.1 + 4.4.1 + 4.5.1 + 5.5.1 + 5.7.4 + 6.8.1 + 7.7.1 + 7.10.1 + 8.5.1 + 8.9.2 + 8.10.2}{17}$$

$$a_{11} = \frac{739}{17} = 43{,}47 \qquad s_{xy} = a_{11} - a_{10}\, a_{01} = 43{,}47 - 5{,}82.7 = 2{,}73$$

3. Dada la tabla de correlaciones. Hallar n_{21} para que las dos variables sean estadísticamente independientes y calcular su covarianza en este caso.

X \ Y	5	7
100	8	4
200	n_{21}	6

Solución:

X \ Y	5	7	$n_{i\bullet}$
100	8	4	12
200	n_{21}	6	$n_{21}+6$
$n_{\bullet j}$	$n_{21}+8$	10	$n_{21}+18$

Por ser independientes: $\frac{n_{ij}}{N} = \frac{n_{i\bullet}}{N}.\frac{n_{\bullet j}}{N} \quad \forall i, j$

$$\frac{4}{n_{21}+18} = \frac{12}{n_{21}+18}\ \frac{10}{n_{21}+18} \rightarrow 4 = \frac{120}{n_{21}+18} \rightarrow 4\left[n_{21}+18\right] = 120 \rightarrow \boxed{n_{21} = \frac{120-72}{4} = 12}$$

covarianza: $s_{xy} = a_{11} - a_{10}\, a_{01}$

X \ Y	5	7	$n_{i\bullet}$
100	8	4	12
200	12	6	18
$n_{\bullet j}$	20	10	30

$$a_{10} = \overline{x} = \frac{\sum_{i=1}^{2} x_i n_{i\bullet}}{N} = \frac{100.12 + 200.18}{30} = 160 \qquad a_{01} = \overline{y} = \frac{\sum_{j=1}^{2} y_j n_{\bullet j}}{N} = \frac{5.20 + 7.10}{30} = 5{,}67$$

$$a_{11} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} x_i y_j n_{ij}}{N} = \frac{100.5.8 + 100.7.4 + 200.5.12 + 200.7.6}{30} = \frac{27200}{30} = 906{,}67$$

$$s_{xy} = a_{11} - a_{10}\, a_{01} = 906{,}67 - 160\,.5{,}67 = -0{,}53$$

4. A partir de la siguiente distribución bidimensional $(X_i, Y_j; n_{ij})$, calcular: $\bar{x}, \bar{y}, s_x^2, s_y^2$ y s_{xy}. ¿Son independientes las variables X e Y?

X \ Y	1	2	3
-1	0	1	0
0	1	0	1
1	0	1	0

Solución:

X \ Y	1	2	3	$n_{i\bullet}$
-1	0	1	0	1
0	1	0	1	2
1	0	1	0	1
$n_{\bullet j}$	1	2	1	4

Las variables X e Y son independientes cuando se verifica $\frac{n_{ij}}{N} = \left(\frac{n_{i\bullet}}{N}\right)\left(\frac{n_{\bullet j}}{N}\right) \ \forall i, j$

No son independientes porque no se verifica la relación: $\frac{0}{4} \neq \frac{2}{4}.\frac{2}{4}$ $\left[\frac{n_{22}}{N} \neq \left(\frac{n_{2\bullet}}{N}\right)\left(\frac{n_{\bullet 2}}{N}\right)\right]$

$$a_{11} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} x_i y_j n_{ij}}{N} = \frac{1}{4}\left[-1.2.1 + 1.2.1\right] = 0$$

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{3} x_i n_{i\bullet}}{N} = \frac{1}{4}\left[-1.1 + 0.2 + 1.1\right] = 0 \qquad a_{20} = \frac{\sum_{i=1}^{3} x_i^2 n_{i\bullet}}{N} = \frac{1}{4}\left[(-1)^2.1 + 0.2 + 1^2.1\right] = \frac{2}{4} = 0{,}5$$

$$s_x^2 = a_{20} - a_{10}^2 = 0{,}5 - 0 = 0{,}5 \ \mapsto \ s_x = \sqrt{0{,}5} = 0{,}7$$

$$a_{01} = \bar{y} = \frac{\sum_{j=1}^{3} y_j n_{\bullet j}}{N} = \frac{1}{4}\left[1.1 + 2.2 + 3.1\right] = 2 \qquad a_{02} = \frac{\sum_{j=1}^{3} y_j^2 n_{\bullet j}}{N} = \frac{1}{4}\left[1^2.1 + 2^2.2 + 3^2.1\right] = \frac{18}{4} = 4{,}5$$

$$s_y^2 = a_{02} - a_{01}^2 = 4{,}5 - 2^2 = 0{,}5 \ \mapsto \ s_y = \sqrt{0{,}5} = 0{,}7$$

covarianza $s_{xy} = a_{11} - a_{10}.a_{01} = 0 - 0.2 = 0$

Adviértase que la covarianza es cero por la simetría de la distribución.

Si (X,Y) independientes	$\mapsto$	$s_{yx} = 0$
Si $s_{yx} = 0$	$\mapsto$	(X,Y) No independientes

5. Se han observado, durante un mes determinado, el gasto en el teléfono móvil y el ingreso total en seis familias. Los resultados obtenidos, expresados en unidades monetarias corrientes, han sido:

	Gasto teléfono móvil	Ingreso total (miles euros)
Familia 1	2	4
Familia 2	3	6
Familia 3	6	8
Familia 4	9	10
Familia 5	10	12
Familia 6	11	20

a) Calcular la covarianza entre el gasto y el ingreso. A la vista de este resultado, ¿puede afirmar que las variables sean dependientes e independientes?
b) Para estas 6 familias ¿Qué variable se distribuye de forma más homogénea, el gasto en móvil o en los ingresos totales?

Solución:

a)

Gasto teléfono móvil y_i	Ingreso total x_i	x_i^2	y_i^2	$x_i . y_i$
2	4	16	4	8
3	6	36	9	18
6	8	64	36	48
9	10	100	81	90
10	12	144	100	120
11	20	400	121	220
41	60	760	351	504

La primera columna (y_i), gasto del teléfono móvil, corresponde a la variable que se estudia, dependiendo de la variable ingreso total de las familias (x_i)

$$a_{01} = \bar{y} = \frac{\sum_{i=1}^{6} y_i}{N} = \frac{41}{6} = 6{,}83 \qquad a_{02} = \frac{\sum_{i=1}^{6} y_i^2}{N} = \frac{351}{6} = 58{,}5 \qquad s_y^2 = a_{02} - a_{01}^2 = 58{,}5 - 6{,}83^2 = 11{,}85$$

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{6} x_i}{N} = \frac{60}{6} = 10 \qquad a_{20} = \frac{\sum_{i=1}^{6} x_i^2}{N} = \frac{760}{6} = 126{,}67 \qquad s_x^2 = a_{20} - a_{10}^2 = 126{,}67 - 10^2 = 26{,}67$$

$$a_{11} = \frac{\sum_{i=1}^{6} x_i . y_i}{N} = \frac{504}{6} = 84 \qquad \boxed{s_{xy} = a_{11} - a_{10} . a_{01} = 84 - 10.6{,}83 = 15{,}7} \text{ covarianza}$$

b)

$$\bar{y} = 6{,}83 \qquad s_y = \sqrt{11{,}85} = 3{,}44 \qquad CV_y = \frac{s_y}{\bar{y}} = \frac{3{,}44}{6{,}83} = 0{,}5037 \text{ (50,37\% de dispersión)}$$

$$\bar{x} = 10 \qquad s_x = \sqrt{26{,}67} = 5{,}16 \qquad CV_x = \frac{s_x}{\bar{x}} = \frac{5{,}16}{10} = 0{,}516 \text{ (51,6\% de dispersión)}$$

Se distribuye de forma más homogénea el ingreso total de las familias.

6. Un psicólogo afirma, basándose en los datos obtenidos, que a medida que el niño crece menores son las respuestas inadecuadas que da en el transcurso de una situación experimental:

Edad	Número respuestas inadecuadas
2	11
3	12
4	10
4	13
5	11
5	9
6	10
7	7

Edad	Número respuestas inadecuadas
7	12
9	8
9	7
10	3
11	6
11	5
12	5

a) Determinar la validez de las conclusiones del psicólogo
b) María, de diez años y medio, participa en el experimento, ¿cuál es el número de respuestas inadecuadas que se puede predecir para ella?
c) Hallar la varianza residual

Solución:

a) La validez de la afirmación se obtendrá en función del coeficiente de correlación: $r = \dfrac{s_{xy}}{s_x\, s_y}$

Como no hay pares repetidos se entiende que son 15 pares de la forma (x_i, y_j) que representará x_i : edad e y_i : número respuestas inadecuadas de modo que la frecuencia de cada par es la unidad.

x_i	2	3	4	4	5	5	6	7	7	9	9	10	11	11	12
y_i	11	12	10	13	11	9	10	7	12	8	7	3	6	5	5

$$a_{11} = \frac{\sum_{i=1}^{15} x_i\, y_i}{N} = \frac{2.11 + 3.12 + 4.10 + \cdots + 11.5 + 12.5}{15} = \frac{789}{15} = 52{,}6$$

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{15} x_i}{N} = \frac{2 + 3 + 4 + 4 + 5 + \cdots + 11 + 11 + 12}{15} = \frac{105}{15} = 7$$

$$a_{01} = \bar{y} = \frac{\sum_{i=1}^{15} y_i}{N} = \frac{11 + 12 + 10 + 13 + \cdots + 6 + 5 + 5}{15} = \frac{129}{15} = 8{,}6$$

En consecuencia, $s_{xy} = a_{11} - a_{10}\, a_{01} = 52{,}6 - 7.8{,}6 = -7{,}6$

Para el cálculo de las desviaciones típicas (s_x, s_y):

$$a_{20} = \frac{\sum_{i=1}^{15} x_i^2}{N} = \frac{2^2 + 3^2 + 4^2 + 4^2 + 5^2 + \cdots + 11^2 + 11^2 + 12^2}{15} = \frac{877}{15} = 58{,}46$$

$$a_{02} = \frac{\sum_{i=1}^{15} y_i^2}{N} = \frac{11^2 + 12^2 + 10^2 + 13^2 + \cdots + 6^2 + 5^2 + 5^2}{15} = \frac{1237}{15} = 82{,}46$$

$$s_x^2 = a_{20} - a_{10}^2 = 58{,}46 - 7^2 = 9{,}46 \quad \mapsto \quad s_x = \sqrt{9{,}46} = 3{,}07$$

$$s_y^2 = a_{02} - a_{01}^2 = 82{,}46 - 8{,}6^2 = 8{,}5 \quad \mapsto \quad s_y = \sqrt{8{,}5} = 2{,}91$$

El coeficiente de correlación: $r = \frac{s_{xy}}{s_x\, s_y} = \frac{-7{,}6}{3{,}07.2{,}91} = -0{,}85$ correlación inversa del 85%

La validez solicitada es del 85% en correlación inversa, es decir, a medida que aumenta la edad del niño (X) disminuye las respuestas inadecuadas (Y).

b) Para poder predecir el número de respuestas para cada edad determinada (caso de María) será necesario hallar la ecuación de regresión de Y (nº respuestas inadecuadas) sobre X (edad del niño):

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad \text{pendiente de la recta} \equiv \text{coeficiente de regresión: } b_{yx} = \frac{s_{xy}}{s_x^2}$$

Adviértase que la pendiente de la recta o coeficiente de regresión b_{yx} viene determinado por el signo de la covarianza s_{xy}

$$b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{-7{,}6}{9{,}46} = -0{,}80 \quad \text{(recta de regresión decreciente)}$$

La ecuación de la recta de regresión será: $y - 8{,}6 = -0{,}80\,(x - 7) \quad \mapsto \quad y = 14{,}2 - 0{,}80x$

En consecuencia, para la edad de María $(x = 10{,}5)$ el número de respuestas inadecuadas que se puede predecir será:

$$y = 14{,}2 - 0{,}80.10{,}5 = 5{,}8 \cong 6 \text{ respuestas inadecuadas.}$$

c) La varianza residual $s_r^2 = s_y^2\,(1 - r^2)$

Coeficiente de Determinación: $r^2 = (-0{,}85)^2 = 0{,}7225$

$$s_r^2 = s_y^2\,(1 - r^2) = 8{,}50\,(1 - 0{,}7225) = 2{,}35875$$

$$\%\ \text{variaciones no explicado} = 100\,\frac{s_r^2}{s_y^2} = 100\,\frac{2{,}35875}{8{,}50} = 27{,}75\%$$

7. De una variable estadística bidimensional (X, Y) se conoce $s_x = 3$:

- Recta de regresión de Y sobre X: $y = 2 + \frac{1}{2}x$
- Recta de regresión de X sobre Y: $x = -4 + 2y$

a) Hallar el coeficiente de correlación
b) Si $\bar{x} = 2$, determinar $\bar{y}$, a_{20}, a_{02} y a_{11}

Solución:

a) La recta de regresión de Y sobre X: $y = 2 + \frac{1}{2}x$ puede escribirse:

$$y = 2 + \frac{1}{2}x \;\mapsto\; y - 0 = \frac{1}{2}(4 + x) \;\Rightarrow\; b_{yx} = \frac{1}{2}$$

Análogamente, la recta de regresión de X sobre Y: $x = -4 + 2y$

$$x = -4 + 2y \;\mapsto\; x - 0 = 2(-2 + y) \;\Rightarrow\; b_{xy} = 2$$

Sabemos que $\begin{cases} b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{1}{2} \;\rightarrow\; \frac{s_{xy}}{9} = \frac{1}{2} \;\rightarrow\; s_{xy} = 4{,}5 \\ b_{xy} = \frac{s_{xy}}{s_y^2} = 2 \;\mapsto\; \frac{4{,}5}{s_y^2} = 2 \;\mapsto\; s_y^2 = \frac{4{,}5}{2} = 2{,}25 \;\mapsto\; s_y = \sqrt{2{,}25} = 1{,}5 \end{cases}$

$r = \frac{s_{xy}}{s_x\, s_y} = \frac{4{,}5}{3\,.\,1{,}5} = 1$ con lo que existe una dependencia funcional, cosa que no es de extrañar por tratarse de única recta de regresión. Adviértase que las rectas: $\left\{ \begin{array}{l} y = 2 + \frac{1}{2}x \\ x = -4 + 2y \end{array} \right\}$ son la misma recta, basta con multiplicar la primera recta por 2 y despejar la x:

$$2y = 2\left[2 + \frac{1}{2}x\right] = 4 + x \;\mapsto\; x = -4 + 2y$$

b) $y = 2 + \frac{1}{2}x \;\mapsto\; \bar{y} = 2 + \frac{1}{2}\bar{x} \overset{\bar{x}=2}{\mapsto} \bar{y} = 2 + \frac{1}{2}2 = 3$

$$s_x^2 = a_{20} - a_{10}^2 \;\mapsto\; 3^2 = a_{20} - 2^2 \;\mapsto\; a_{20} = 3^2 + 2^2 = 13$$

$$s_y^2 = a_{02} - a_{01}^2 \;\mapsto\; 2{,}25 = a_{02} - 3^2 \;\mapsto\; a_{02} = 2{,}25 + 3^2 = 11{,}25$$

$$s_{xy} = a_{11} - a_{11}\, a_{01} \;\mapsto\; 4{,}5 = a_{11} - 2.3 \;\mapsto\; a_{11} = 4{,}5 + 6 = 10{,}5$$

8. En una experimentación sobre el sector turístico se han observado dos caracteres cuantitativos (X, Y), obteniéndose los siguientes resultados:

(0, 2), (1,6), (3, 14), (-1, -2), (2, 10)

a) Hallar las distribuciones marginales
b) Correlación entre ambos caracteres
c) ¿Cómo completaríamos los pares (-3, •), (•, 4)?. Utilizar para ello la recta de regresión ajustada a los datos observados.

Solución:

a) Como no hay repetición de los pares, la tabla de doble entrada de frecuencias absolutas vendrá dada de la forma:

X \ Y	2	6	14	-2	10	$n_{i\bullet}$
0	1					1
1		1				1
3			1			1
-1				1		1
2					1	1
$n_{\bullet j}$	1	1	1	1	1	5

Las distribuciones marginales de la X e Y, respectivamente, serán:

x_i	0	1	3	-1	2
$n_{i\bullet}$	1	1	1	1	1

y_j	2	6	14	-2	10
$n_{\bullet j}$	1	1	1	1	1

b) Para estudiar la correlación se forma la tabla adjunta, donde no figura la columna de las frecuencias absolutas por ser la unidad para todos los pares

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
0	2	0	0	4
1	6	6	1	36
3	14	42	9	196
-1	-2	2	1	4
2	10	20	4	100
5	30	70	15	340

$$a_{11} = \frac{\sum_{i=1}^{5} x_i y_i}{N} = \frac{70}{5} = 14$$

$$\overline{x} = a_{10} = \frac{\sum_{i=1}^{5} x_i}{N} = \frac{5}{5} = 1 \qquad a_{20} = \frac{\sum_{i=1}^{5} x_i^2}{N} = \frac{15}{5} = 3 \qquad s_x^2 = a_{20} - a_{10}^2 = 3 - 1^2 = 2 \qquad s_x = \sqrt{2} = 1{,}41$$

$$\bar{y} = a_{01} = \frac{\sum_{i=1}^{5} y_i}{N} = \frac{30}{5} = 6 \qquad a_{02} = \frac{\sum_{i=1}^{5} y_i^2}{N} = \frac{340}{5} = 68 \qquad s_y^2 = a_{02} - a_{01}^2 = 68 - 6^2 = 32 \qquad s_y = \sqrt{32} = 5{,}66$$

$$s_{xy} = a_{11} - a_{10}\, a_{01} = 14 - 1.6 = 8 \qquad r = \frac{s_{xy}}{\sqrt{s_x^2\, s_y^2}} = \frac{8}{\sqrt{2.32}} = 1$$

Como el coeficiente de correlación es igual a 1, indica que existe una dependencia funcional entre las variables (X, Y) estudiadas.

c) Para completar el par (-3, •) hay que hallar la ecuación de la recta de regresión de Y sobre X. Análogamente, para completar el par (•, 4) hay que hallar la ecuación de la recta de regresión de X sobre Y.

♦ Recta de regresión de Y sobre X:

$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$, donde el coeficiente de regresión $b_{yx} = \frac{s_{xy}}{s_x^2}$ (pendiente de la recta)

$\bar{x} = 1 \qquad \bar{y} = 6 \qquad b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{8}{2} = 4$

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \;\mapsto\; y - 6 = 4(x - 1) \;\mapsto\; y = 2 + 4x$$

El par (-3, •) se completa: $y = 2 + 4\,(-3) = -10 \;\rightarrow\; (-3, -10)$

♦ Recta de regresión de X sobre Y:

$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$, donde el coeficiente de regresión $b_{xy} = \frac{s_{xy}}{s_y^2}$ (pendiente de la recta)

$\bar{x} = 1 \qquad \bar{y} = 6 \qquad b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{8}{32} = \frac{1}{4}$

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \;\mapsto\; x - 1 = \frac{1}{4}(y - 6) \;\mapsto\; x = \frac{1}{4}(-2 + y)$$

El par (•, 4) se completa: $x = \frac{1}{4}\left[-2 + 4\right] = \frac{1}{2} \;\rightarrow\; \left[\frac{1}{2},\, 4\right]$

9. Se desea estudiar la relación que existe entre la variable X (porcentaje de la población urbana en las distintas provincias) e Y (renta media por hogar). La tabla adjunta contiene datos referentes a treinta provincias:

X \ Y	1 - 16	16 - 31	31 - 46	46 - 60
10 - 19	1	1	1	
19 - 28		8	3	
28 - 37		3	7	1
37 - 45		2	3	

a) Calcular las rectas de regresión

Solución:

a)

X \ Y	1 - 16	16 - 31	31 - 46	46 - 60	$n_{i\bullet}$
10 - 19	1	1	1		3
19 - 28		8	3		11
28 - 37		3	7	1	11
37 - 45		2	3		5
$n_{\bullet j}$	1	14	14	1	30

- Las distribuciones marginales de X e Y, respectivamente:

Intervalos	x_i	$n_{i\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$
10 - 19	14,5	3	43,5	630,75
19 - 28	23,5	11	258,5	6074,75
28 - 37	32,5	11	357,5	11618,75
37 - 45	41	5	205	8405
		30	864,5	26729,25

$$\bar{x} = a_{10} = \frac{\sum_{i=1}^{4} x_i n_{i\bullet}}{N} = \frac{864,5}{30} = 28,81 \qquad a_{20} = \frac{\sum_{i=1}^{4} x_i^2 n_{i\bullet}}{N} = \frac{26729,25}{30} = 890,975$$

$$s_x^2 = a_{20} - a_{10}^2 = 890,975 - 28,81^2 = 60,959 \qquad s_x = \sqrt{60,959} = 7,807$$

Intervalos	y_j	$n_{\bullet j}$	$y_j n_{\bullet j}$	$y_j^2 n_{\bullet j}$
1 - 16	8,5	1	8,5	72,25
16 - 31	23,5	14	329	7731,5
31 - 46	38,5	14	539	20751,5
46 - 60	53	1	53	2809
		30	929,5	31364,25

$$\bar{y} = a_{01} = \frac{\sum_{j=1}^{4} y_j n_{\bullet j}}{N} = \frac{929,5}{30} = 30,98 \qquad a_{02} = \frac{\sum_{j=1}^{4} y_j^2 n_{\bullet j}}{N} = \frac{31364,25}{30} = 1045,475$$

$$s_y^2 = a_{02} - a_{01}^2 = 1045{,}475 - 30{,}98^2 = 85{,}7146 \qquad s_y = \sqrt{85{,}7146} = 9{,}258$$

- La distribución conjunta

$x_i \setminus y_j$	8,5	23,5	38,5	53
14,5	1	1	1	
23,5		8	3	
32,5		3	7	1
41		2	3	

$$a_{11} = \frac{\sum_{i=1}^{4} x_i\, y_i\, n_{ii}}{N} = \frac{14{,}5.8{,}5.\,1 + 14{,}5.23{,}5.1 + 14{,}5.38{,}5.1 + 23{,}5.23{,}5.8 + \cdots + 41.38{,}5.3}{30} = \frac{27589{,}5}{30} = 919{,}65$$

$$s_{xy} = a_{11} - a_{10}\, a_{01} = 919{,}65 - 28{,}81.30{,}98 = 27{,}1162$$

✓ **Recta de regresión de Y sobre X:** $y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x}) \;\mapsto\; y - 30{,}98 = \frac{27{,}1162}{60{,}959}(x - 28{,}81)$

$$\boxed{y = 18{,}30 + 0{,}44\,x}$$

Coeficiente de regresión: $b_{yx} = \frac{m_{11}}{\sigma_x^2} = \frac{27{,}1162}{60{,}959} = 0{,}44 > 0$ (recta de regresión creciente)

✓ **Recta de regresión de X sobre Y:** $x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y}) \;\mapsto\; x - 28{,}81 = \frac{27{,}1162}{85{,}7146}(y - 30{,}98)$

$$\boxed{x = 19{,}20 + 0{,}31\,y}$$

Coeficiente de regresión: $b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{27{,}1162}{85{,}7146} = 0{,}31 > 0$ (recta de regresión creciente)

10. Justifique las razones por las cuales debe aceptarse o rechazarse que las dos rectas siguientes sean, respectivamente, las líneas de regresión mínimo-cuadráticas de Y sobre X y de X sobre Y de una serie de observaciones.

Y/X: $Y = 2X + 1$ X/Y: $X = -5Y + 10$

Solución:

$$\begin{cases} Y = 1 + 2X & \rightarrow \; b_{yx} = 2 > 0 \\ X = 10 - 5Y & \rightarrow \; b_{xy} = -5 < 0 \end{cases}$$

Los coeficientes de regresión deben tener el mismo signo, al depender ambos de la misma covarianza.
Con lo cual, no pueden ser rectas de regresión.

11. Justifique las razones por las cuales debe aceptarse o rechazarse que las dos rectas siguientes sean, respectivamente, las líneas de regresión mínimo-cuadráticas de Y sobre X y de X sobre Y de una serie de observaciones.

$$Y/X:\ Y = 2X + 1 \qquad X/Y:\ X = -5Y + 10$$

Solución:

$$\begin{cases} Y = 1 + 2X \ \rightarrow\ b_{yx} = 2 > 0 \\ X = 10 + 5Y \ \rightarrow\ b_{xy} = 5 > 0 \end{cases}$$

Los coeficientes de regresión tienen el mismo signo, lo que es lógico al depender ambos de la misma covarianza.

De otra parte, el coeficiente de correlación: $r = \sqrt{b_{yx}.b_{xy}} = \sqrt{2.5} = 3{,}16$, resultado absurdo cuando el coeficiente de correlación $-1 \le r \le 1$, concluyendo que no pueden ser rectas de regresión.

12. El coeficiente de correlación entre dos variables X e Y es 0,6. Sabiendo además que,
$\overline{x} = 10 \quad s_x = 1{,}5 \quad \overline{y} = 20 \quad s_y = 2$

a) Hallar las rectas de regresión de Y/X y de X/Y
b) Calcular la varianza residual para las dos regresiones anteriores

Solución:

- Recta de regresión de Y sobre X: $y - \overline{y} = \dfrac{s_{xy}}{s_x^2}(x - \overline{x}) \ \mapsto\ b_{yx} = \dfrac{s_{xy}}{s_x^2}$ (coeficiente regresión)
- Recta de regresión de X sobre Y: $x - \overline{x} = \dfrac{s_{xy}}{s_y^2}(y - \overline{y}) \ \mapsto\ b_{xy} = \dfrac{s_{xy}}{s_y^2}$ (coeficiente regresión)

El coeficiente de correlación: $r = \sqrt{b_{yx}.b_{xy}} = \dfrac{s_{xy}}{s_x.s_y} \ \mapsto\ 0{,}6 = \dfrac{s_{xy}}{1{,}5.2} \ \mapsto\ s_{xy} = 1{,}8$

En consecuencia, $b_{yx} = \dfrac{s_{xy}}{s_x^2} = \dfrac{1{,}8}{1{,}5^2} = 0{,}8 \qquad b_{xy} = \dfrac{s_{xy}}{s_y^2} = \dfrac{1{,}8}{2^2} = 0{,}45$

Las rectas de regresión serán: $\left| \begin{array}{l} Y/X:\ y - 20 = 0{,}8\,(x - 10) \ \rightarrow\ y = 12 + 0{,}8x \\ X/Y:\ x - 10 = 0{,}45\,(y - 20) \ \rightarrow\ x = 1 + 0{,}45y \end{array} \right.$

b) Varianza residual $\begin{cases} Y/X \quad s_r^2 = s_y^2\left[1 - r^2\right] \\ X/Y \quad s_r^2 = s_x^2\left[1 - r^2\right] \end{cases}$ Error típico estimación $\begin{cases} Y/X \quad s_r = s_y\sqrt{1 - r^2} \\ X/Y \quad s_r = s_x\sqrt{1 - r^2} \end{cases}$

por tanto, $\begin{cases} Y/X \quad s_r^2 = 2^2\left[1 - 0{,}6^2\right] \ \mapsto\ s_r^2 = 2{,}56 \ \rightarrow\ s_r = \sqrt{2{,}56} = 1{,}6 \\ X/Y \quad s_r^2 = 1{,}5^2\left[1 - 0{,}6^2\right] \ \mapsto\ s_r^2 = 1{,}44 \ \rightarrow\ s_r = \sqrt{1{,}44} = 1{,}2 \end{cases}$

13. En una distribución bidimensional se conoce:

$$R = 0{,}7 \qquad s_x = 1{,}2 \qquad \overline{y} = 4 \qquad X/Y:\ X = 0{,}6 + 0{,}44\,Y$$

Obtener:

a) Media de X
b) Recta de regresión de Y/X
c) Varianza de Y
d) Covarianza de ambas variables

Solución:

a) Recta de regresión de X sobre Y: $X = 0{,}6 + 0{,}44\,Y \ \mapsto\ \begin{cases} \overline{X} = 0{,}6 + 0{,}44\,\overline{Y} \\ \overline{X} = 0{,}6 + 0{,}44 \,.\, 4 = 2{,}36 \end{cases}$

b) La recta de regresión de Y/X:

siendo $X = 0{,}6 + 0{,}44\,Y \ \mapsto\ \begin{cases} a = 0{,}6 \\ b_{xy} = 0{,}44 \end{cases}$

$$r^2 = b_{yx} \,.\, b_{xy} \quad \mapsto \quad 0{,}7^2 = b_{yx} \,.\, 0{,}44 \quad \mapsto \quad b_{yx} = \frac{0{,}7^2}{0{,}44} = 1{,}114$$

con lo cual, la recta de regresión de Y sobre X: $y - \overline{y} = \frac{\overbrace{s_{xy}}^{b_{yx}}}{s_x^2}(x - \overline{x})$ será: $y - 4 = 1{,}114\ (x - 2{,}36)$

$$\boxed{y = 1{,}370 + 1{,}114\ x}$$

c) Varianza de la Y: Sabemos que, $s_x = 1{,}2 \qquad b_{xy} = 0{,}44 \qquad b_{yx} = 1{,}114$

$$b_{yx} = \frac{m_{11}}{\sigma_x^2} \quad \mapsto \quad 1{,}114 = \frac{s_{xy}}{1{,}2^2} \quad \mapsto \quad s_{xy} = 1{,}114 \,.\, 1{,}2^2 = 1{,}604$$

recurriendo a $b_{xy} = \frac{s_{xy}}{s_y^2} \ \mapsto\ 0{,}44 = \frac{1{,}604}{s_y^2} \ \mapsto\ \boxed{s_y^2 = \frac{1{,}604}{0{,}44} = 3{,}645}$

d) La covarianza de ambas ya se ha calculado: $\boxed{s_{xy} = 1{,}604}$

14. Sean las variables estadísticas bidimensionales (X, Y), donde X = "PIB per cápita (en miles de dólares) e Y = "Tasa natural de crecimiento demográfico de 162 países del mundo". Se conocen los datos siguientes:

$$\sum x = 978{,}9 \qquad \sum y = 2886{,}4 \qquad \sum xy = 8938{,}4$$
$$\sum x^2 = 17569{,}9 \qquad \sum y^2 = 172291{,}2$$

a) Obtener la recta de regresión que pretende explicar la tasa natural de crecimiento en función de la renta del país.
b) Interpretar los coeficientes de la recta estimada.
c) Obtener una medida de bondad del ajuste e interpretar si éste es bueno.

Solución:

a) Se trata de encontrar la recta de regresión de Y sobre X: $y - \overline{y} = \overbrace{\frac{s_{xy}}{s_x^2}}^{b_{yx}} (x - \overline{x})$

$$a_{10} = \overline{x} = \frac{\sum x}{N} = \frac{978{,}9}{162} = 6{,}04 \qquad a_{20} = \frac{\sum x^2}{N} = \frac{17569{,}9}{162} = 108{,}456$$

$$s_x^2 = a_{20} - a_{10}^2 = 108{,}456 - 6{,}04^2 = 71{,}97$$

$$a_{01} = \overline{y} = \frac{\sum y}{N} = \frac{2886{,}4}{162} = 17{,}82 \qquad a_{02} = \frac{\sum y^2}{N} = \frac{172291{,}2}{162} = 1063{,}526$$

$$s_y^2 = a_{02} - a_{01}^2 = 1063{,}526 - 17{,}82^2 = 745{,}97$$

$$a_{11} = \frac{\sum xy}{N} = \frac{8938{,}4}{162} = 55{,}175 \qquad s_{xy} = a_{11} - a_{10}\, a_{01} = 55{,}175 - 6{,}04 \,.\, 17{,}82 = -52{,}46$$

El coeficiente de regresión de Y sobre X (pendiente de la recta): $b_{yx} = \frac{s_{xy}}{s_x^2} = \frac{-52{,}46}{71{,}97} = -0{,}729$

Adviértase que la pendiente de la recta $(-0{,}729)$ en el signo depende de la covarianza (s_{xy}), al ser negativa la recta de regresión será decreciente, esto es, a medida que aumenta los valores de la variable X (PIB per cápita) disminuyen los valores de la variable Y (tasa natural de crecimiento demográfico).

La recta de regresión solicitada será: $y - 17{,}82 = -0{,}729\,(x - 6{,}04) \;\mapsto\; \boxed{y = 22{,}22 - 0{,}729\,x}$

c) El Coeficiente de determinación lineal: $r^2 = b_{yx} \,.\, b_{xy}$

$$b_{xy} = \frac{s_{xy}}{s_y^2} = \frac{-52{,}46}{745{,}97} = -0{,}07$$

con lo que, $r^2 = (-0{,}729).(-0{,}07) = 0{,}051$ (5,1% grado de fiabilidad)

El coeficiente de correlación lineal: $r = \sqrt{0{,}051} = 0{,}226$ (no existe apenas correlación lineal entre las variables, pudiendo existir otro tipo de correlación)

15. La siguiente distribución bidimensional se expresa en la siguiente tabla de correlaciones. La variable X representa los ingresos familiares mensuales en unidades de 10 euros. La variable Y representa, a su vez, los metros cuadrados de la vivienda familiar.

X/ Y	< 60	60 - 80	80 - 100	100 - 150	> 150
50 - 100	20	18	2	1	0
100 - 200	25	40	30	2	1
200 - 350	5	10	15	25	3
350 - 500	0	5	15	20	8
> 500	0	1	2	7	10

a) Calcular la distribución marginal de las dos variables. ¿Son independientes los ingresos familiares y el tamaño de la vivienda donde habitan?
b) Obtener la distribución de la superficie de la vivienda condicionada al intervalo modal de los ingresos familiares.
c) Calcular la distribución de los ingresos condicionada al intervalo mediano de la vivienda familiar.

Solución:

a)

X/ Y	< 60	60 - 80	80 - 100	100 - 150	> 150	$n_{i\bullet}$	$f_{i\bullet} = \frac{n_{i\bullet}}{N}$
50 - 100	20	18	2	1	0	41	0,155
100 - 200	25	40	30	2	1	98	0,370
200 - 350	5	10	15	25	3	58	0,219
350 - 500	0	5	15	20	8	48	0,181
> 500	0	1	2	7	10	20	0,075
$n_{\bullet j}$	50	74	64	55	22	N= 265	1
$f_{\bullet j} = \frac{n_{\bullet j}}{N}$	0,189	0,279	0,242	0,208	0,083	1	

Para que los ingresos familiares (X) y el tamaño de la vivienda familiar (Y) sean independientes debe verificarse $\frac{n_{ij}}{N} = \left(\frac{n_{i\bullet}}{N}\right)\left(\frac{n_{\bullet j}}{N}\right) \ \forall i, j$

No son independientes porque $\frac{n_{43}}{N} \neq \frac{n_{4\bullet}}{4}\frac{n_{\bullet 3}}{N} \ \mapsto \ \frac{15}{265} \neq \frac{48}{265}\frac{64}{265}$

DISTRIBUCIÓN MARGINAL DE LA VARIABLE X

Intervalos	x_i	$n_{i\bullet}$	c_i	$f_{i\bullet}=\frac{n_{i\bullet}}{N}$	N_i	$F_{i\bullet}=\frac{N_{i\bullet}}{N}$	$h_i=\frac{n_i}{c_i}$
50 - 100	75	41	50	0,155	41	0,155	0,82
100 - 200	150	98	100	0,370	139	0,525	0,98
200 - 350	275	58	150	0,219	197	0,744	0,39
350 - 500	425	48	150	0,181	245	0,925	0,32
> 500	-----	20	-----	0,075	265	1	-----
		265		1			

DISTRIBUCIÓN MARGINAL DE LA VARIABLE Y

Intervalos	y_j	$n_{\bullet j}$	c_j	$f_{\bullet j}=\frac{n_{\bullet j}}{N}$	N_j	$F_{\bullet j}=\frac{N_{\bullet j}}{N}$	$h_j=\frac{n_j}{c_j}$	
< 60	-----	50	-----	0,189	50	0,189	-----	
60 - 80	70	74	20	0,279	124	0,468	3,7	
80 - 100	90	64	20	0,242	188	0,71	3,2	mediano
100 - 150	125	55	50	0,208	243	0,918	1,1	
> 150	-----	22	-----	0,083	265	1	-----	
		265		1				

N/2=132,

b) X = "ingresos familiares" e Y = "metros cuadrados de la superficie"

y_j	$n_j/50-100$	$n_j/100-200$	$n_j/200-350$	$n_j/350-500$	$n_j/>500$
< 60	20	25	5	0	0
60 - 80	18	40	10	5	1
80 - 100	2	30	15	15	2
100 - 150	1	2	25	20	7
> 150	0	1	3	8	10
	41	98	58	48	20

Con los datos disponibles no se puede calcular el intervalo modal de la variable X, al no poder calcular todas las densidades de frecuencias marginales, es imposible hacerlo en el tramo (> 500) que tiene una amplitud ilimitada.

c) La distribución condicionada de la variable X al intervalo mediano de la Y (vivienda familiar):

X / Y	< 60	60 - 80	80 - 100	100 - 150	> 150
50 - 100	20	18	2	1	0
100 - 200	25	40	30	2	1
200 - 350	5	10	15	25	3
350 - 500	0	5	15	20	8
> 500	0	1	2	7	10

Intervalos	$n_{i3}\,(n_{i\bullet}/80-100)$
50 - 100	2
100 - 200	30
200 - 350	15
350 - 500	15
> 500	2

16. Se conocen las regresiones $\begin{cases} Y/X: \quad Y = 3 + 2X \\ X/Y: \quad X = 2 + 0{,}3Y \end{cases}$

Sabiendo además que $s_{xy} = 3{,}2$. Obtener la varianza residual de las dos rectas de regresión.

Solución:

$$\begin{cases} Y/X: \quad Y = 3 + 2X \\ X/Y: \quad X = 2 + 0{,}3Y \end{cases} \mapsto \begin{cases} b_{yx} = 2 \\ b_{xy} = 0{,}3 \end{cases} \mapsto \begin{cases} b_{yx} = s_{xy}/s_x^2 \xrightarrow{s_{xy}=3{,}2} s_x^2 = 3{,}2/2 = 1{,}6 \\ b_{xy} = s_{xy}/s_y^2 \xrightarrow{s_{xy}=3{,}2} s_y^2 = 3{,}2/0{,}3 = 10{,}67 \end{cases}$$

Por otra parte, el coeficiente de determinación: $R^2 = b_{yx}.b_{xy} = 2.0{,}3 = 0{,}6$

Varianza residual $\begin{cases} Y/X: \quad s_r^2 = s_y^2\left[1 - r^2\right] \rightarrow s_r^2 = 10{,}67\left[1 - 0{,}6\right] = 4{,}268 \\ X/Y: \quad s_r^2 = s_x^2\left[1 - r^2\right] \rightarrow s_r^2 = 1{,}6\left[1 - 0{,}6\right] = 0{,}64 \end{cases}$

Error típico estimación $\begin{cases} Y/X: \quad s_r = s_y\sqrt{1 - r^2} \rightarrow s_r = \sqrt{4{,}268} = 2{,}066 \\ X/Y: \quad s_r = s_x\sqrt{1 - r^2} \rightarrow s_r = \sqrt{0{,}64} = 0{,}8 \end{cases}$

17. Sean las siguientes ecuaciones las rectas de regresión de una variable bidimensional (Y, X; n_{ij})

$$\begin{cases} X - 2Y = 3 \\ X - 4Y = 2 \end{cases}$$

a) ¿Cuál de estas rectas corresponde a la regresión de Y/X y cuál a la regresión de X/Y?
b) Hallar las medias aritméticas de Y sobre X
c) ¿Cuánto vale el coeficiente de correlación lineal?

Solución:

a)

- Sea $\begin{cases} X - 2Y = 3 \xrightarrow{\text{recta regresión } X/Y} \\ X - 4Y = 2 \xrightarrow{\text{recta regresión } Y/X} \end{cases} \begin{cases} X = 3 + 2Y \mapsto \begin{cases} a = 3 \\ b_{xy} = 2 \end{cases} \\ Y = -\frac{1}{2} + \frac{1}{4}X \mapsto \begin{cases} a' = -1/2 \\ b_{yx} = 1/4 \end{cases} \end{cases} \mapsto$ signo $(b_{xy}) =$ signo (b_{yx})

Coeficiente de determinación $r^2 = b_{xy}.b_{yx} = 2.\frac{1}{4} = 0{,}5 < 1$

- Sea $\begin{cases} X-2Y=3 \\ X-4Y=2 \end{cases} \xrightarrow[\text{recta regresión X/Y}]{\text{recta regresión Y/X}} \begin{cases} Y=-\frac{3}{2}+\frac{1}{2}X \mapsto \begin{cases} a=-3/2 \\ b_{yx}=1/2 \end{cases} \\ X=2+4Y \mapsto \begin{cases} a'=2 \\ b_{xy}=4 \end{cases} \end{cases} \mapsto \text{signo}(b_{yx})=\text{signo}(b_{xy})$

Coeficiente de determinación $r^2=b_{yx}.b_{xy}=\frac{1}{2}.4=2>1$ cosa que no es posible $(0\le r^2\le 1)$

En consecuencia $\begin{cases} X/Y: \quad X=3+2Y \\ Y/X: \quad Y=-\frac{1}{2}+\frac{1}{4}X \end{cases}$

18. En una distribución bidimensional (X_i, Y_j , n_{ij}) se conoce $\overline{x}=10$ y $s_{xy}=10$. Ambas rectas de regresión pasan por el punto (0, 0). ¿Cuál es el grado de bondad del ajuste?.

Solución:

Las rectas de regresión de Y/X e X/Y se cortan en $(\overline{x},\overline{y})$, en este caso en el punto $(10,\overline{y})$.

Por otra parte, según el enunciado se cortan en (0, 0), por lo que se puede concluir que ambas rectas coinciden al tener dos puntos distintos en común.

En consecuencia, $R^2=1 \rightarrow R=1$ (100% grado de ajuste).

19. A partir de un conjunto de datos sobre las variables X e Y se ha calculado la regresión de Y sobre X, obteniéndose los siguientes resultados:

$$Y=10+0{,}45X \qquad r^2=0{,}9 \qquad \overline{x}=20$$

Calcular los parámetros de regresión de X sobre Y

Solución:

$$Y=10+0{,}45X \mapsto \begin{cases} a=10 \\ b_{yx}=0{,}45 \end{cases} \xrightarrow{r^2=b_{yx}.b_{xy}} r^2=0{,}9=0{,}45.b_{xy} \Rightarrow \underline{b_{xy}=\frac{0{,}9}{0{,}45}=2 \text{ (pendiente recta)}}$$

De otra parte, $y=a+\overset{b_{yx}}{\overset{\frown}{b}}.x \xrightarrow{\overline{y}=a+b.\overline{x}} \overline{y}=10+0{,}45\,.\,20=19$

Análogamente, $x=a'+\overset{b_{xy}}{\overset{\frown}{b'}}.y \xrightarrow{\overline{x}=a'+b'.\overline{y}} a'=\overline{x}-\overset{b_{xy}}{\overset{\frown}{b'}}.\overline{y} \Rightarrow \underline{a'=20-2.19=-18}$

La recta de regresión de X/Y: $\boxed{Y=-18+2X}$

20. ¿Cuáles de los siguientes pares de posibles rectas de regresión de Y/X y de X/Y realmente pueden serlo?. Razone la respuesta.

a) $Y = 3 + 4X$ siendo $X = 2 + Y$ b) $Y = 3 + 2X$ siendo $X = 2 - 0{,}3Y$ c) $Y = 3 + 2X$ siendo $X = 2 + 0{,}2Y$

Solución:

- $\begin{cases} Y/X: \quad Y = 3 + 4X \mapsto \begin{vmatrix} a = 3 \\ b_{yx} = 4 > 0 \end{vmatrix} \\ X/Y: \quad X = 2 + Y \mapsto \begin{vmatrix} a' = 2 \\ b_{xy} = 1 > 0 \end{vmatrix} \end{cases} \mapsto \begin{cases} \text{signo}(b_{yx}) = \text{signo}(b_{xy}) \\ r^2 = b_{yx}.b_{xy} = 4.1 = 4 > 1 \text{ contradicción} \end{cases}$

- $\begin{cases} Y/X: \quad Y = 3 + 2X \mapsto \begin{vmatrix} a = 3 \\ b_{yx} = 2 > 0 \end{vmatrix} \\ X/Y: \quad X = 2 - 0{,}3Y \mapsto \begin{vmatrix} a' = 2 \\ b_{xy} = -0{,}3 < 0 \end{vmatrix} \end{cases} \mapsto \text{signo}(b_{yx}) \neq \text{signo}(b_{xy}) \text{ contradicción}$

- $\begin{cases} Y/X: \quad Y = 3 + 2X \mapsto \begin{vmatrix} a = 3 \\ b_{yx} = 2 > 0 \end{vmatrix} \\ X/Y: \quad X = 2 + 0{,}2Y \mapsto \begin{vmatrix} a' = 2 \\ b_{xy} = 0{,}2 > 0 \end{vmatrix} \end{cases} \mapsto \left\{ \begin{array}{l} \text{signo}(b_{yx}) = \text{signo}(b_{xy}) \\ r^2 = b_{yx}.b_{xy} = 2.0{,}2 = 0{,}4 < 1 \end{array} \right\}$ coeficientes coherentes

21. Comprobar si son coherentes los resultados obtenidos al ajustar la recta de regresión:

a) $Y = A + bX \mapsto s_{xy} = 20 \quad s_x^2 = 10 \quad \bar{y} = 8 \quad \bar{x} = 4 \quad a = 3$

b) $Y = A + bX \mapsto s_y^2 = 4 \quad s_{xy} = 4 \quad s_{ry}^2 = 0{,}4 \quad s_x^2 = 5$

Solución:

a)

$$Y = A + bX \mapsto \begin{cases} b = b_{yx} = \dfrac{s_{xy}}{s_x^2} = \dfrac{20}{10} = 2 \\ \bar{y} = a + b\bar{x} \mapsto a = \bar{y} - b\bar{x} = 8 - 2.4 = 0 \neq 3 \end{cases} \longrightarrow \begin{array}{l} \text{Los datos no corresponden} \\ \text{a la recta de regresión} \end{array}$$

b) Los datos no corresponden a una recta de regresión como puede observarse.

$$Y=a+bX \;\mapsto\; \begin{cases} s_{ry}^2=s_y^2\,(1-r^2) \;\mapsto\; 0{,}4=4(1-r^2) \;\mapsto\; 0{,}1=(1-r^2) \;\mapsto\; r^2=0{,}9 \;\mapsto\; r=0{,}94 \\ b=b_{yx}=\dfrac{s_{xy}}{s_x^2}=\dfrac{4}{5}=0{,}8 \\ r^2=\dfrac{s_{xy}^2}{s_x^2.s_y^2}=\;1-\dfrac{s_{ry}^2}{s_y^2}\mapsto\; r^2=\dfrac{4^2}{5\,.\,4}=0{,}8\;\neq\;1-\dfrac{s_{ry}^2}{s_y^2}=1-\dfrac{0{,}4}{4}=0{,}9 \end{cases}$$

22. En una distribución bidimensional (X, Y) se ha ajustado una regresión lineal entre las dos variables. Se sabe que $r=0{,}8,\ s_x=4,\ \overline{y}=2$ y que la recta de regresión de X sobre Y ajustada es $Y=4X$. Se pide:

a) Calcular los valores de s_{xy}, s_y^2 y $\overline{x}$
b) Calcular la recta de regresión de Y sobre X
c) Calcular la varianza residual en la regresión de X sobre Y

Solución:

a)

Recta de regresión de X sobre Y $Y=4X$

$$\begin{cases} x-\overline{x}=\dfrac{s_{xy}}{s_y^2}(y-\overline{y}) \\ X=\dfrac{1}{4}Y \xrightarrow{x=a'+b'y} \begin{cases} a'=0 \\ b'=b_{xy}=1/4 \text{ (pendiente recta)} \end{cases} \end{cases}$$

covarianza (s_{xy})

$$\begin{cases} r^2=\overbrace{b_{yx}}^{b}.\overbrace{b_{xy}}^{b'} \;\mapsto\; 0{,}8^2=b_{yx}.\dfrac{1}{4} \;\mapsto\; b_{yx}=2{,}56 \\ \overbrace{b_{yx}}^{b}=\dfrac{s_{xy}}{s_x^2} \;\mapsto\; s_{xy}=b_{yx}.s_x^2 \;\mapsto\; s_{xy}=(2{,}56).4^2=40{,}96 \end{cases}$$

Varianza Y (s_y^2) $\quad \overbrace{b_{xy}}^{b'}=\dfrac{s_{xy}}{s_y^2} \;\mapsto\; s_y^2=\dfrac{s_{xy}}{b_{xy}} \;\mapsto\; s_y^2=\dfrac{40{,}96}{1/4}=163{,}84$

Media X $(\overline{x})$ $\quad x=a'+b'y \xrightarrow{E[x]=E[a'+b'y]\mapsto\ \overline{x}=a'+b'\overline{y}} \overline{x}=0+\dfrac{1}{4}.2=0{,}5$

b)

Recta de regresión de Y sobre X

$$\begin{cases} y-\overline{y}=\overbrace{\dfrac{s_{xy}}{s_x^2}}^{b=b_{yx}}(x-\overline{x}) \;\mapsto\; y=a+bx \\ y-2=\dfrac{40{,}96}{4^2}(x-0{,}5) \;\mapsto\; y=0{,}72+2{,}56x \end{cases}$$

c) Varianza residual de X: $\quad s_{rx}^2=s_x^2\,(1-r^2) \;\mapsto\; s_{rx}^2=16\,(1-0{,}64)=5{,}76$

23. Se desea estudiar la repercusión que tiene los días de lluvia en el número de visitas al zoo. Para ello, se observaron las siguientes variables, durante los últimos diez años, siendo Y="nº visitas anuales, en miles" y X="nº de días de lluvia al año":

Año	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
X	18	26	30	33	38	39	42	44	46	49
Y	107	105,5	105	104,4	104,3	104	103,7	103,4	103,1	103

a) Coeficiente de correlación lineal e interpretar el resultado.
b) Recta de regresión que explique el número de visitas anuales en función del número de lluvia.
c) ¿Qué previsión de visitas habrá para el año próximo si el Instituto Meteorológico informa que lloverá 40 días?. ¿Qué grado de fiabilidad tendrá esta predicción?.
d) Hallar la varianza residual del número de visitas anuales.
e) Obtener la recta de regresión X/Y.

Solución:

Año	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1994	18	107	1926	324	11449
1995	26	105,5	2743	676	11130,25
1996	30	105	3150	900	11025
1997	33	104,4	3445,2	1089	10899,36
1998	38	104,3	3963,4	1444	10878,49
1999	39	104	4056	1521	10816
2000	42	103,7	4355,4	1764	10753,69
2001	44	103,4	4549,6	1936	10691,56
2002	46	103,1	4742,6	2116	10629,61
2003	49	103	5047	2401	10609
10	365	1043,4	37978,2	14171	108881,96

Distribución marginal de X

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{10} x_i}{N} = \frac{365}{10} = 36,5 \qquad a_{20} = \frac{\sum_{i=1}^{10} x_i^2}{N} = \frac{14171}{10} = 1417,1$$

$$\begin{cases} s_x^2 = a_{20} - a_{10}^2 = 1417,1 - 36,5^2 = 84,85 \\ s_x = \sqrt{84,85} = 9,21 \end{cases}$$

Distribución marginal de Y

$$a_{01} = \bar{y} = \frac{\sum_{i=1}^{10} y_i}{N} = \frac{1043,4}{10} = 104,34 \qquad a_{02} = \frac{\sum_{i=1}^{10} y_i^2}{N} = \frac{108881,96}{10} = 10888,196$$

$$\begin{cases} s_y^2 = a_{02} - a_{01}^2 = 10888,196 - 104,34^2 = 1,36 \\ s_y = \sqrt{1,36} = 1,17 \end{cases}$$

Covarianza - Coeficientes regresión lineal - Coeficiente correlación lineal

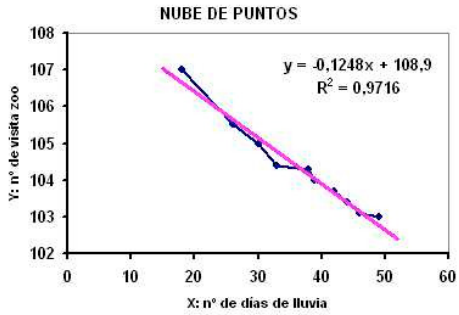
$$a_{11} = \frac{\sum_{i=1}^{10} x_i . y_i}{N} = \frac{37978,2}{10} = 3797,82$$

Covarianza: $s_{xy} = a_{11} - a_{10} . a_{01} = 3797,82 - 36,5 . 104,34 = -10,59$

Coeficientes regresión lineal:

$$\begin{cases} Y/X: & \overbrace{b_{yx}}^{b} = \dfrac{s_{xy}}{s_x^2} = \dfrac{-10,59}{84,85} = -0,125 \\ X/Y: & \overbrace{b_{xy}}^{b'} = \dfrac{s_{xy}}{s_y^2} = \dfrac{-10,59}{1,36} = -7,79 \end{cases}$$

Coeficiente de correlación lineal: $r = \sqrt{b_{yx} . b_{xy}} = \sqrt{(-0,125)(-7,79)} = 0,986$



Observando la gráfica de la nube de puntos a más días de lluvia menor número de visitas. El grado de ajuste entre la nube de puntos y la recta de regresión es del 98,6%.

b) Recta de regresión de Y sobre X:

$$y - \bar{y} = \overbrace{\frac{s_{yx}}{s_x^2}}^{b=b_{yx}} (x - \bar{x}) \;\mapsto\; y - 104,34 = -0,125\,(x - 36,5) \;\mapsto\; y = 108,90 - 0,125x$$

c) Si en 2007 se estiman 40 días de lluvia se estiman un número de visitas:

$$y = 108,90 - 0,125\,(40) \approx 104 \text{ días}$$

d) La varianza residual de la Y:

$$s_{ry}^2 = s_y^2(1 - r^2) \;\mapsto\; s_{ry}^2 = 1,36(1 - 0,986^2) = 0,0378 \;\; (3,78\% \text{ causas ajenas a la regresión})$$

e) Recta de regresión de X sobre Y:

$$x - \bar{x} = \overbrace{\frac{s_{yx}}{s_y^2}}^{b'=b_{xy}} (y - \bar{y}) \;\mapsto\; x - 36,5 = -7,79\,(y - 104,34) \;\mapsto\; x = 849,31 - 7,79\,y$$

$$X/Y:\; x = 849,31 - 7,79\,y \;\mapsto\; \hat{y} = \frac{849,31 - x}{7,79}$$

NOTA.- Para representar conjuntamente en EXCEL las dos rectas de regresión (Y/X, X/Y) se han de introducir dos series: Serie1 (X, Y), Serie2 (X, $\hat{Y}$)

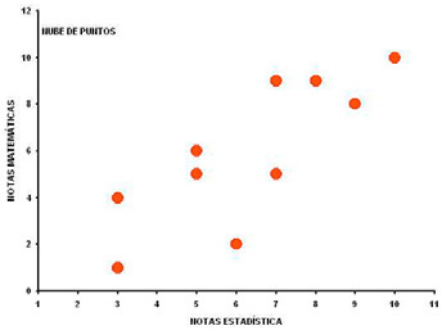
24. Las notas en Estadística (X) y en Matemáticas (Y) obtenidas por 10 alumnos elegidos al azar en un grupo de primer curso de la Facultad de Ciencias Económicas y Empresariales han sido las siguientes, según el orden de selección de la muestra:

Nº orden	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
X	9	7	3	6	7	5	10	8	3	5
Y	8	5	4	2	9	6	10	9	1	5

a) Representar la nube de puntos correspondiente a esta distribución. ¿Qué hipótesis pueden hacerse a la vista de la representación?.
b) Estimar los parámetros de la recta de regresión Y/X. Interpretar los coeficientes calculados.
c) Estimar los parámetros de la recta de regresión de X/Y y comparar ambas rectas.
d) Representar las dos rectas de regresión junto a la nube de puntos.
e) Calcular la varianza residual en la regresión Y/X. ¿Coincidirá con la varianza residual en la regresión X/Y?
f) Para un alumno que haya obtenido un 7 en Matemáticas, ¿qué nota se le pronosticaría en Estadística?
g) Para un alumno que haya obtenido un 4 en Estadística, ¿qué nota se le pronosticaría en Matemáticas?

Solución:

a)



Observando la nube de puntos (diagrama de dispersión) se puede establecer la hipótesis de que existe correlación lineal creciente entre las variables.

b) Estimar los parámetros de la recta de regresión Y/X

Nº orden	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	
x_i	9	7	3	6	7	5	10	8	3	5	63
y_i	8	5	4	2	9	6	10	9	1	5	59
$x_i . y_i$	72	35	12	12	63	30	100	72	3	25	424
x_i^2	81	49	9	36	49	25	100	64	9	25	447
y_i^2	64	25	16	4	81	36	100	81	1	25	433

Distribución marginal de X

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{10} x_i}{N} = \frac{63}{10} = 6{,}3 \qquad a_{20} = \frac{\sum_{i=1}^{10} x_i^2}{N} = \frac{447}{10} = 44{,}7$$

$$\begin{cases} s_x^2 = a_{20} - a_{10}^2 = 44{,}7 - 6{,}3^2 = 5{,}01 \\ s_x = \sqrt{5{,}01} = 2{,}24 \end{cases}$$

Distribución marginal de Y

$$a_{01} = \bar{y} = \frac{\sum_{i=1}^{10} y_i}{N} = \frac{59}{10} = 5{,}9 \qquad a_{02} = \frac{\sum_{i=1}^{10} y_i^2}{N} = \frac{433}{10} = 43{,}3 \qquad \begin{cases} s_y^2 = a_{02} - a_{01}^2 = 43{,}3 - 5{,}9^2 = 8{,}49 \\ s_y = \sqrt{8{,}49} = 2{,}91 \end{cases}$$

Covarianza - Coeficientes regresión lineal - Coeficiente correlación lineal

$$a_{11} = \frac{\sum_{i=1}^{10} x_i \cdot y_i}{N} = \frac{424}{10} = 42{,}4$$

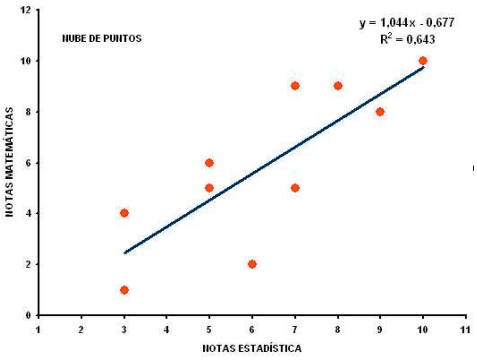
Covarianza: $s_{xy} = a_{11} - a_{10} . a_{01} = 42{,}4 - 6{,}3 \,.\, 5{,}9 = 5{,}23$

Parámetros regresión lineal Y/X

$Y = a + bX \;\mapsto\; Y = -0{,}677 + 1{,}044X$

$$\begin{cases} b = b_{yx} = \dfrac{s_{xy}}{s_x^2} = \dfrac{5{,}23}{5{,}01} = 1{,}044 > 0 \\ \bar{y} = a + b\bar{x} \;\mapsto\; a = \bar{y} - b\bar{x} = 5{,}9 - 1{,}044 \,.\, 6{,}3 = -0{,}677 \\ r^2 = \dfrac{s_{xy}}{s_x^2} . \dfrac{s_{xy}}{s_y^2} = \dfrac{5{,}23}{5{,}01} . \dfrac{5{,}23}{8{,}49} = 0{,}643 \;\mapsto\; r = \sqrt{0{,}643} = 0{,}80 \end{cases}$$

El coeficiente de regresión b es positivo, con lo que a mayor nota en estadística mayor nota en matemáticas. De otra parte, el coeficiente de correlación r es 0,80, con lo que la fiabilidad del modelo es del 80%.

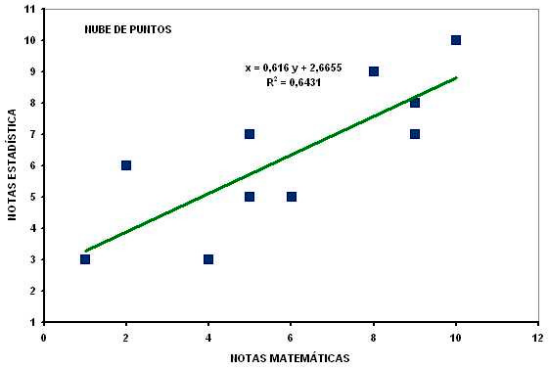


c)

Parámetros regresión lineal X/Y

$X = a' + b'\,Y \;\mapsto\; X = 2{,}665 + 0{,}616Y$

$$\begin{cases} b' = b_{xy} = \dfrac{s_{xy}}{s_y^2} = \dfrac{5{,}23}{8{,}49} = 0{,}616 > 0 \\ \bar{x} = a' + b'\,\bar{y} \;\mapsto\; a' = \bar{x} - b'\bar{y} = 6{,}3 - 0{,}616 \,.\, 5{,}9 = 2{,}665 \\ r^2 = \dfrac{s_{xy}}{s_x^2} . \dfrac{s_{xy}}{s_y^2} = \dfrac{5{,}23}{5{,}01} . \dfrac{5{,}23}{8{,}49} = 0{,}643 \;\mapsto\; r = \sqrt{0{,}643} = 0{,}80 \end{cases}$$



El coeficiente de regresión b' es positivo, con lo que a mayor nota en matemáticas mayor nota en estadística.

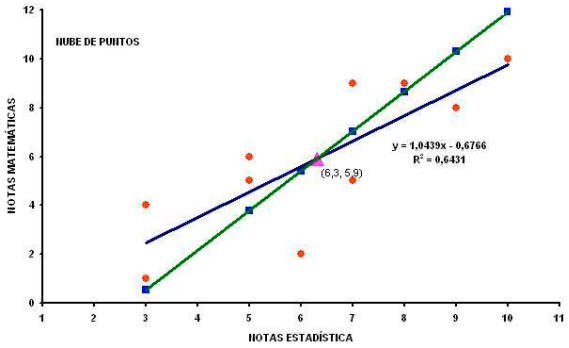
De otra parte, $X = 2{,}665 + 0{,}616Y \;\mapsto\; \hat{Y} = \dfrac{X - 2{,}665}{0{,}616}$ se utiliza para representar en Excel la serie $(X, \hat{Y})$, que junto a la serie (X, Y), permite la gráfica conjunta de la nube de puntos y las dos rectas de regresión.

d) Para representar en Excel las dos rectas de regresión junto a la nube de puntos.

X	9	7	3	6	7	5	10	8	3	5
Y	8	5	4	2	9	6	10	9	1	5
$\hat{Y}$	10,28	7,04	0,54	5,41	7,04	3,79	11,91	8,66	0,54	3,79

Diagrama dispersión: Series (X, Y), (X, $\hat{Y}$)

$\hat{Y} = (X - 2{,}665)/0{,}616$



e) Varianzas residuales

Varianza residual de Y/X: $r^2 = 0{,}643 \quad s_y^2 = 8{,}49 \qquad s_{ry}^2 = s_y^2\,(1 - r^2) \;\mapsto\; s_{ry}^2 = 8{,}49\,(1 - 0{,}643) = 3{,}03$

Varianza residual de X/Y: $r^2 = 0{,}643 \quad s_y^2 = 5{,}01 \qquad s_{rx}^2 = s_x^2\,(1 - r^2) \;\mapsto\; s_{rx}^2 = 5{,}01\,(1 - 0{,}643) = 1{,}79$

f) Un alumno con un 7 en Matemáticas $(\bullet, 7)$ para pronosticar la nota en Estadística habría que recurrir a la recta de regresión de X/Y: $X = 2{,}665 + 0{,}616\,Y$

$$X = 2{,}665 + 0{,}616.7 = 6{,}98 \text{ en estadística}$$

g) Un alumno con un 4 en Estadística $(4, \bullet)$ para pronosticar la nota en Matemáticas habría que recurrir a la recta de regresión de Y/X: $Y = -0{,}677 + 1{,}044\,X$

$$Y = -0{,}677 + 1{,}044\,.4 = 3{,}50 \text{ en matemáticas}$$