Relación de ejercicios 2 EDIP

Carlos García, Bora Goker, Javier Gómez,
Ana Graciani, J.Alberto Hoces

2020/2021

Ejercicio 1.

Ejercicio 2.

Ejercicio 3. En una encuesta de familias sobre el número de individuos que la componen (X) y el número de personas activas en ellas (Y) se han obtenido los siguientes resultados:

X/Y	1	2	3	4
1	7	0	0	0
2	10	2	0	0
3	11	5	1	0
4	10	6	6	0
5	8	6	4	2
6	1	2	3	1
7	1	0	0	1
8	0	0	1	1

a) Calcular la recta de regresión de Y sobre X.

X/Y	1	2	3	4	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
1	7	0	0	0	7	7	7
2	10	2	0	0	12	24	48
3	11	5	1	0	17	51	153
4	10	6	6	0	22	88	352
5	8	6	4	2	20	100	500
6	1	2	3	1	7	43	252
7	1	0	0	1	2	14	98
8	0	0	1	1	2	16	128
$n_{.j}$	48	21	15	5	89		
$n_{.j}y_j$	48	42	45	20			
$n_{.j}y_j^2$	48	84	135	80			

La recta de regresión lineal de Y sobre X viene dada por la expresión:

$$y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x}) \Rightarrow y = \frac{\sigma_{xy}}{\sigma_x^2}x - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x} + \overline{y}$$

Por lo tanto, comencemos calculando las medias aritméticas y la varianza de x y la covarianza:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{8} x_i n_{i.} = 3{,}8427 \text{ individuos} \qquad \overline{y} = \frac{1}{n}\sum_{j=1}^{4} y_j n_{.j} = 1{,}7416 \text{ personas activas}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{8} n_{i.}x_i^2 - \overline{x}^2 = 2{,}5146 \text{ individuos}^2$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{8}\sum_{j=1}^{4} n_{ij}x_i y_j - \overline{x}\ \overline{y} = 0{,}7907$$

Por lo tanto, la recta de regresión de Y sobre X quedaría:

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x} + \overline{y} = 0{,}3144x + 0{,}5333$$

b) ¿Es adecuado suponer una relación lineal para explicar el comportamiento de Y a partir de X?

Para ver cómo de adecuado es suponer dicha relación calculamos el coeficiente de correlación lineal:

$$r^2 = \sqrt{\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}} \qquad r = \sqrt{r^2}$$

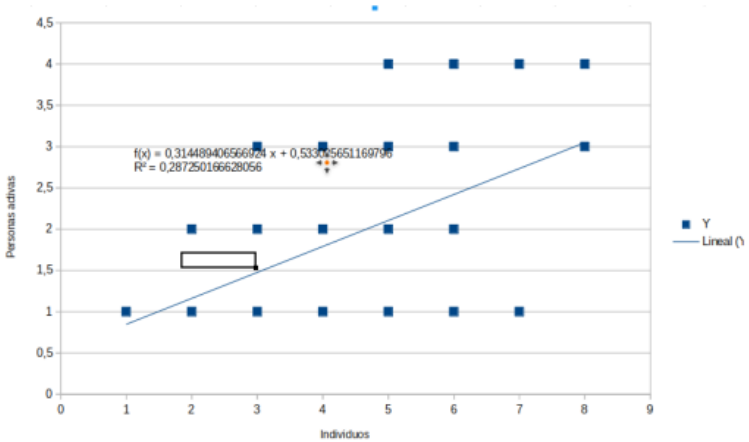
Ahora calculamos la varianza de Y:

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{4} n_{.j}y_j^2 - \overline{y}^2 = 0{,}8657 \text{ personas activas}^2$$

Por tanto:

$$r^2 = \frac{0{,}7907^2}{2{,}5146 \cdot 0{,}8657} = 0{,}2872 \qquad r = 0{,}536$$

Observando estos resultados podemos afirmar que no es adecuado suponer esta relación linear puesto que el coeficiente de correlación lineal está demasiado alejado de 1.



Ejercicio 4. Medidos los pesos, X (en Kg), y las alturas, Y (en cm), a un grupo de individuos, se han obtenido los siguientes resultados:

X/Y	160	162	164	166	168	170
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

a) Calcular el peso medio y la altura media y decir cuál es más representativo.

Para ello hemos de trabajar con las distribuciones marginales del carácter X (peso en Kg) y del carácter Y (altura en cm). Tras efectuar los cálculos pertinentes, la tabla anterior queda de la siguiente forma:

X/Y	160	162	164	166	168	170	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
48	3	2	2	1	0	0	8	384	18432
51	2	3	4	2	2	1	14	714	36414
54	1	3	6	8	5	1	24	1296	69984
57	0	0	1	2	8	3	14	798	45486
60	0	0	0	2	4	4	10	600	36000
$n_{.j}$	6	8	13	15	19	9			
$n_{.j}y_j$	960	1296	2132	2490	3192	1530			
$n_{.j}y_j^2$	153600	209952	349648	413340	536256	260100			

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i n_{i.} = \frac{384 + 714 + 1296 + 798 + 600}{70} = 54{,}1714 \text{ kilogramos}$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{6} y_j n_{.j} = \frac{960 + 1296 + 2132 + 2490 + 3192 + 1530}{70} = 165{,}7143 \text{ centímetros}$$

Para poder determinar cuál de las dos medias es más representativa, haremos uso del coeficiente de variación de Pearson, el cual nos permitirá interpretar independientemente de la escala la variabilidad de los datos respecto de su media:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{5} n_{i.}x_i^2 - \overline{x}^2} = 3{,}582 \text{ kilogramos} \qquad C.V(X) = \frac{\sigma_x}{\bar{x}} = 0{,}0661$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{j=1}^{6} n_{.j}y_j^2 - \overline{y}^2} = 2{,}9519 \text{ centímetros} \qquad C.V(Y) = \frac{\sigma_y}{\bar{y}} = 0{,}0178$$

En vista de los resultados, se deduce que la distribución marginal del carácter Y (la altura en cm) es más homogénea, por lo que la altura media es la más representativa de las dos.

b) Calcular el porcentaje de individuos que pesan menos de 55 Kg y miden más de 165 cm.

Si miden más de 165 cm y pesan menos de 55 kg, hemos de tener en cuenta las frecuencias absolutas de todos los pares (x_i, y_j) con $i \in \{1,2,3\}$ e $j \in \{4,5,6\}$. Efectuamos el cálculo:

$$100 \cdot \frac{1}{n} \cdot \sum_{j=4}^{6} \sum_{i=1}^{3} n_{ij} = 100 \cdot \frac{1}{70} \cdot (1+2+2+1+8+5+1) = 28{,}571\,\% \text{ de los individuos}$$

c) Entre los que miden más de 165 cm, ¿cuál es el porcentaje de los que pesan más de 52 Kg?

Si miden más de 165 cm y pesan más de 52 kg, hemos de tener en cuenta las frecuencias absolutas de todos los pares (x_i, y_j) con $i \in \{3,4,5\}$ e $j \in \{4,5,6\}$. Efectuamos el cálculo:

$$100 \cdot \frac{1}{n} \cdot \sum_{j=4}^{6} \sum_{i=3}^{5} n_{ij} = 100 \cdot \frac{1}{70} \cdot (8+5+1+2+8+3+2+4+4) = 52{,}857\,\% \text{ de los individuos}$$

d) ¿Cuál es la altura más frecuente entre los individuos cuyo peso oscila entre 51 y 57 Kg?

Esto es equivalente a hallar la modalidad y_j a la que le corresponde el máximo de $n_{2j} + n_{3j} + n_{4j}$:

y_j	$n_{2j} + n_{3j} + n_{4j}$
160	3
162	6
164	11
166	12
168	15
170	5

De la tabla se obtiene que la altura más frecuente entre los individuos cuyo peso oscila entre 51 y 57 kg es $y_5 = 168$ cm.

e) ¿Qué peso medio es más representativo, el de los individuos que miden 164 cm o el de los que miden 168 cm?

Para poder determinar lo que se nos pide, hemos de estudiar de estudiar dos distribuciones condicionadas: la del carácter X condicionada a la modalidad y_3 y la del carácter X condicionada a la modalidad y_5:

x_i	n_{i3}	$x_i n_{i3}$	$x_i^2 n_{i3}$
48	2	96	4608
51	4	204	10404
54	6	324	17496
57	1	57	3249
60	0	0	0
	13	681	35757
x_i	n_{i5}	$x_i n_{i5}$	$x_i^2 n_{i5}$
48	0	0	0
51	2	102	5202
54	5	270	14580
57	8	456	25992
60	4	240	14400
	19	1068	60174

$$\bar{x}_3 = \frac{1}{n}\sum_{i=1}^{5} x_i n_{i3} = \frac{681}{13} = 52{,}3846 \text{ kg} \qquad \bar{x}_5 = \frac{1}{n}\sum_{i=1}^{5} x_i n_{i5} = \frac{1068}{19} = 56{,}2105 \text{ kg}$$

$$\sigma_{x,3} = \sqrt{\frac{1}{n}\sum_{i=1}^{5} n_{i3}x_i^2 - \overline{x}_3^2} = 2{,}5283 \text{ kg} \qquad \sigma_{x,5} = \sqrt{\frac{1}{n}\sum_{i=1}^{5} n_{i5}x_i^2 - \overline{x}_5^2} = 2{,}7216 \text{ kg}$$

Al igual que se ha hecho en el apartado a), haremos uso del coeficiente de variación de Pearson para poder determinar qué peso medio es más representativo:

$$C.V_3(Y) = \frac{\sigma_{x,3}}{\bar{x}_3} = 0{,}0483 \qquad C.V_5(Y) = \frac{\sigma_{x,5}}{\bar{x}_5} = 0{,}0484$$

En vista de lo obtenido, podemos concluir que el peso medio de la distribución condicionada del carácter X a la modalidad y_3 es el más representativo de los dos, aunque en este caso los coeficientes de Pearson son prácticamente iguales.

Ejercicio 5. Estudiar la dependencia o independencia de las variables en cada una de las siguientes distribuciones. Dar, en cada caso las curvas de regresión y la covarianza de las dos variables.

Comencemos con la primera distribución:

X/Y	1	2	3	4	5
10	2	4	6	10	8
20	1	2	3	5	4
30	3	6	9	15	12
40	4	8	12	20	16

Debemos saber que el cáracter Y es independiente a nivel estadístico del carácter X si las distribuciones de Y condicionadas a cada valor de la variable X son iguales $\forall x_i \quad i = 1, 2, \ldots, k$:

$$f_j^i = f_{j/i}$$

Así, se tiene que cumplir lo siguiente:

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{ij}}{n_{i.}} = \cdots = \frac{n_{kj}}{n_{k.}} \qquad \forall j = 1, 2, \ldots, p \tag{1}$$

Ahora comprobemos si esto se cumple en la distribución representada por la tabla anterior. Es digno de mención que el recíproco también es cierto, es decir, si Y es independiente de X, X también lo es de Y.

$$j = 1 \qquad \frac{2}{30} = \frac{1}{15} = \frac{3}{45} = \frac{4}{60}$$

$$j = 2 \qquad \frac{4}{30} = \frac{2}{15} = \frac{6}{45} = \frac{8}{60}$$

$$j = 3 \qquad \frac{6}{30} = \frac{3}{15} = \frac{9}{45} = \frac{12}{60}$$

$$j = 4 \qquad \frac{10}{30} = \frac{5}{15} = \frac{15}{45} = \frac{20}{60}$$

$$j = 5 \qquad \frac{8}{30} = \frac{4}{15} = \frac{12}{45} = \frac{16}{60}$$

Podemos observar que, efectivamente, Y es independiente de X y por tanto carace de sentido estudiar la curva de regresión y afirmamos que:

$$\sigma_{xy} = 0$$

Ahora pasemos a estudiar la segunda distribución:

X/Y	1	2	3
-1	0	1	0
0	1	0	1
1	0	1	0

De entrada, podemos observar que estas dos variables no son independientes, puesto que si hubiese algún $n_{ij} = 0$, la igualdad (1) no se daría a no ser que $n_{ij} = 0 \quad \forall i, j$, lo cual volvería absurdo el estudio de esta tabla puesto que no representaría ninguna distribución de frecuencias.

Ahora observemos la posible dependencia funcional. Podemos afirmar que X no depende funcionalmente, pues $n_{12} = 1$ y $n_{22} = 0$, es decir, a la modalidad 2 de Y le corresponden dos posibles modalidades de X.

X/Y	1	2	3	$n_{i.}x_i$
-1	0	1	0	-1
0	1	0	1	0
1	0	1	0	1

Calculemos la covarianza:

$$\sigma_{xy} = m_{11} - \overline{x}\ \overline{y}$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{3} n_{i.}x_i = 0$$

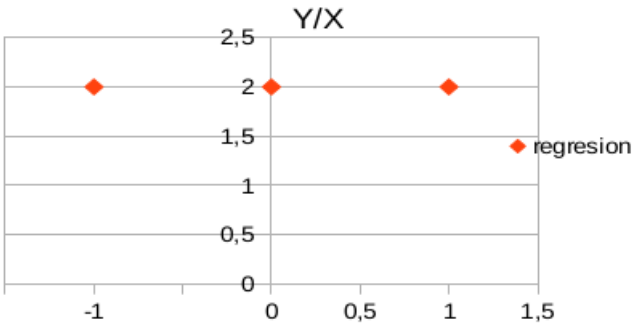
Al ser $\overline{x} = 0$, la covarianza será m_{11}:

$$\sigma_{xy} = \sum_{i=1}^{3}\sum_{j=1}^{3} f_{ij}x_i y_j = 0$$

Este resultado nos demuestra que si las variables son independientes, su covarianza es 0, pero el recíproco no es cierto. Aquí las variables no son independientes y su covarianza es 0.

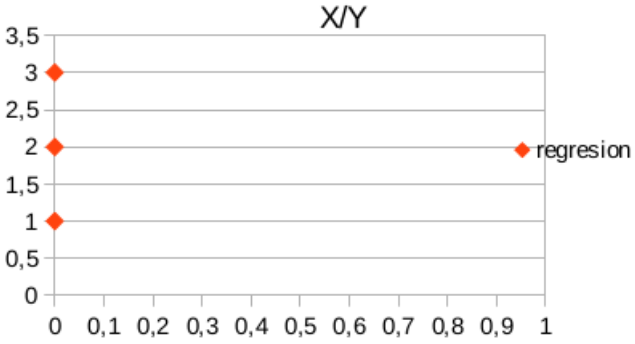
Ahora calcularemos la curva de regresión de tipo 1 de Y/X, que es la curva que pasa por los puntos $(x_i, \overline{y}_i) \quad i = 1, \ldots, k$.

Punto 1: $(-1, 2)$ Punto 2: $(0, 2)$ Punto 3: $(1, 2)$



Ahora calcularemos la curva de regresión tipo 1 de X/Y, la cual pasa por los puntos $\overline{x}_j, y_j \quad j = 1, \ldots, p$.

Punto 1: $(0, 1)$ Punto 2: $(0, 2)$ Punto 3: $(0, 3)$



Ejercicio 6.

Ejercicio 7.

Ejercicio 8.

Ejercicio 9.

Ejercicio 10. De una muestra de 24 puestos de venta en un mercado de abastos se ha recogido información sobre el número de balanzas (X) y el número de dependientes (Y). Los resultados aparecen en la siguiente tabla:

X/Y	1	2	3	4
1	1	2	0	0
2	1	2	3	1
3	0	1	2	6
4	0	0	2	3

a) Determinar las rectas de regresión.

Las rectas que se nos pide determinar son la recta de regresión lineal de Y sobre X y la recta de regresión lineal de X sobre Y, las cuales vienen dadas por las siguientes expresiones respectivamente:

$$y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x}) \Rightarrow y = \frac{\sigma_{xy}}{\sigma_x^2}x - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x} + \overline{y}$$

$$x - \overline{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \overline{y}) \Rightarrow x = \frac{\sigma_{xy}}{\sigma_y^2}y - \frac{\sigma_{xy}}{\sigma_y^2}\overline{y} + \overline{x}$$

Por lo tanto, será necesario calcular las medias y varianzas marginales de cada carácter y la covarianza. Antes de desarrollar los cálculos de la tabla, vamos a especificar cómo calcularemos la covarianza. Como se vio en clase, $\sigma_{xy} = \mu_{11} = m_{11} - m_{10}m_{01}$, siendo μ_{rs} el momento conjunto central de órdenes r y s y m_{rs} el momento conjunto respecto al origen de órdenes r y s. Recordemos que $m_{10} = \bar{x}$, $m_{01} = \bar{y}$ y $m_{11} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j$. Procedamos a efectuar los cálculos pertinentes en la tabla:

X/Y	1	2	3	4	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
1	1	2	0	0	3	3	3
2	1	2	3	1	7	14	28
3	0	1	2	6	9	27	81
4	0	0	2	3	5	20	80
$n_{.j}$	2	5	7	10	24	64	192
$n_{.j}y_j$	2	10	21	40	73		
$n_{.j}y_j^2$	2	20	63	160	245		
$\sum_{i=1}^{k} n_{ij}x_i$	3	9	20	32			

$$m_{10} = \bar{x} = \frac{1}{n}\sum_{i=1}^{4} x_i n_{i.} = \frac{64}{24} = 2{,}667 \text{ balanzas}$$

$$m_{01} = \bar{y} = \frac{1}{n}\sum_{j=1}^{4} n_{.j}y_j = \frac{73}{24} = 3{,}0417 \text{ dependientes}$$

$$m_{11} = \sum_{i=1}^{4}\sum_{j=1}^{4} f_{ij}x_iy_j = \frac{1}{n}\sum_{j=1}^{4}\sum_{i=1}^{4} n_{ij}x_iy_j = \frac{1}{n}\sum_{j=1}^{4} y_j \sum_{i=1}^{4} n_{ij}x_i =$$

$$= \frac{1\cdot 2 + 2\cdot 9 + 3\cdot 20 + 4\cdot 32}{24} = 8{,}708$$

$$\sigma_{xy} = m_{11} - m_{10}m_{01} = 0{,}5958$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{4} n_{i.}x_i^2 - \overline{x}^2 = 0{,}8871 \text{ balanzas} \qquad \sigma_y^2 = \frac{1}{n}\sum_{j=1}^{4} n_{.j}y_j^2 - \overline{y}^2 = 0{,}9564 \text{ dependientes}$$

Ahora que tenemos todos los datos que necesitábamos, sustituimos sus valores en las expresiones de las rectas expuestas al principio del ejercicio:

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x} + \overline{y} \Rightarrow y = 0{,}6716x + 1{,}2505$$

$$x = \frac{\sigma_{xy}}{\sigma_y^2}y - \frac{\sigma_{xy}}{\sigma_y^2}\overline{y} + \overline{x} \Rightarrow x = 0{,}623y + 0{,}7721$$

b) ¿Es apropiado suponer que existe una relación lineal entre las variables?

Para ello, calcularemos el coeficiente de correlación de Pearson y a través de su interpretación podremos contestar esta pregunta:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0{,}5958}{\sqrt{0{,}8871} \cdot \sqrt{0{,}9564}} = 0{,}6468$$

En nuestro caso, $0 < r < 1 \Rightarrow$ Cuanto más próximo se encuentre r de 1, mejor dependencia lineal existirá entre el carácter X y el carácter Y en estudio. Sin embargo, 0.6468 es un valor que dista mucho de 1 como para poder considerar la relación lineal como la que mejor describe la relación del número de balanzas y de dependientes (se empezaría a considerar el coeficiente alto a partir de 0.85). Así, concluimos que no es apropiado suponer una relación lineal entre las variables.

c) Predecir, a partir de los resultados, el número de balanzas que puede esperarse en un puesto con seis dependientes. ¿Es fiable esta predicción?

Para realizar la predicción haremos uso de la recta de regresión lineal de X sobre Y, ya que se nos está proporcionando el número de dependientes:

$$x = 0{,}623 \cdot 6 + 0{,}7721 = 4{,}5101 \text{ balanzas (entre 4 y 5)}$$

Por el apartado anterior, podemos afirmar que esta predicción no es fiable al estar basada en una relación lineal, pues esta no es la que mejor explica la relación entre las variables.

Ejercicio 11.

Ejercicio 12.

Ejercicio 13.

Ejercicio 14. De las estadísticas de "Tiempos de vuelo y consumos de combustible"de una compañía aérea, se han obtenido datos relativos a 24 trayectos distintos realizados por el avión DC-9. A partir de estos datos se han obtenido las siguientes medidas: