RELACIÓN 2 ESTADÍSTICA DESCRIPTIVA E INTRODUCCIÓN A LA PROBABILIDAD

UNIVERSIDAD DE GRANADA

David Muñoz Sánchez
Higinio Paterna Ortiz
Tomás Rodríguez Hernáez
Mario Rubio Venzal
Hugo Teruel Muñoz

1. El primer ejercicio casi que mejor no lo hacemos, hágalo usted, bueno mejor dicho, hazlo tú.

2. Medidos los pesos, X (en kg), y las alturas, Y (en cm), a un grupo de individuos, se han obtenido los siguientes resultados (se incluyen en la tabla diversos cálculos para facilitar el cálculo de las medias y las desviaciones típicas):

X\Y	160	163	164	166	168	170	$n_{i.}$	$n_{i.} \cdot x_i$	$n_{i.}(x_i - \bar{x})^2$
48	3	2	2	1	0	0	8	384	304,689
51	2	3	4	2	2	1	14	714	140,809
54	1	3	6	8	5	1	24	1296	0,705
57	0	0	1	2	8	3	14	798	112,014
60	0	0	0	2	4	4	10	600	339,726
$n_{.j}$	6	8	13	15	19	9	70		
$n_{.j} \cdot y_j$	960	1296	2132	2490	3192	1530			
$n_{.j}(y_j - \bar{y})^2$	195,919	110,368	38,204	1,224	99,264	165,305			

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

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot n_{i.}}{n} = 54{,}1714 \; kg \qquad \bar{y} = \frac{\sum_{j=1}^{6} y_j \cdot n_{.j}}{n} = 165{,}7143 \; cm$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{5} n_{i.} \cdot x_i^2 - \bar{x}^2 = 12{,}828\, kg^2 \Rightarrow \sigma_x = +\sqrt{\sigma_x^2} = 3{,}582 \; kg$$

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{6} n_{.j} \cdot y_j^2 - \bar{y}^2 = 8{,}718\, cm^2 \Rightarrow \sigma_y = +\sqrt{\sigma_y^2} = 2{,}953 \; cm$$

Para ver cuál es más representativo, empleamos el coeficiente variación de Pearson y el que nos de un resultado menor, será el más representativo:

$$C.V.(X) = \frac{\sigma_x}{|\bar{x}|} = 0{,}066 \qquad C.V.(Y) = \frac{\sigma_y}{|\bar{y}|} = 0{,}01782$$

El más representativo es la altura media.

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

X\Y	166	168	170	$n_{i.}$
48	1	0	0	1
51	2	2	1	5
54	8	5	1	14
				20

$$\frac{20}{70} \cdot 100 = 28{,}5714\,\%$$

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

Ahora tengo X/Y>165cm:

X	n_i
48	1
51	5
54	14
87	13
60	10
	43

$$\frac{37}{43} \cdot 100 = 86{,}047\,\%$$

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

Tenemos Y/51<=x<=57:

Y	n_j
160	3
162	6
164	11
166	12
168	15
170	5

La moda (altura más frecuente) es 168 cm.

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

En este caso tenemos que calcular dos medias condicionadas a estos respectivos valores de Y y las varianzas, con objeto de calcular el C.V.

$$\bar{x}_{164} = \frac{48\cdot2+51\cdot4+54\cdot6+57}{13} = 52{,}3846 \ kg \qquad \bar{x}_{168} = \frac{51\cdot2+54\cdot5+57\cdot8+60\cdot4}{13} = 56{,}21 \ kg$$

$$\sigma^2_{x_{164}} = 6{,}3905 \ kg^2 \qquad \sigma^2_{x_{168}} = 7{,}4294 \ kg^2$$

$$C.V.(X_{164}) = \frac{\sigma_{x_{164}}}{|\bar{x}_{164}|} = 0{,}0483 \qquad C.V.(X_{168}) = \frac{\sigma_{x_{168}}}{|\bar{x}_{168}|} = 0{,}0485$$

Es mas representativo el de los individuos que miden 164 cm.

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 (introducimos columnas extras en la tabla para facilitar posteriores cálculos):

X\Y	1	2	3	4	$n_{i.}$	$n_{i.}\, x_i^2$
1	7	0	0	0	7	7
2	10	2	0	0	12	48
3	11	5	1	0	17	153
4	10	6	6	0	22	352
5	8	6	4	2	20	500
6	1	2	3	1	7	252
7	1	0	0	1	2	98
8	0	0	1	1	2	128
$n_{.j}$	48	21	15	5	89	
$n_{.j}\, y_j^2$	48	84	135	80		

a) Calcular la recta de regresión de Y sobre X.
b)¿Es adecuado suponer una relación lineal para explicar el comportamiento de Y a partir de X?

a) La recta de regresión lineal de Y sobre X viene dada por:

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

Comenzaremos calculando las medias aritméticas y la varianza de x y la covarianza:

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

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

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

La recta de regresión de Y sobre X quedaría:

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

b) Para ver cómo de adecuado calculamos el coeficiente de correlación lineal:

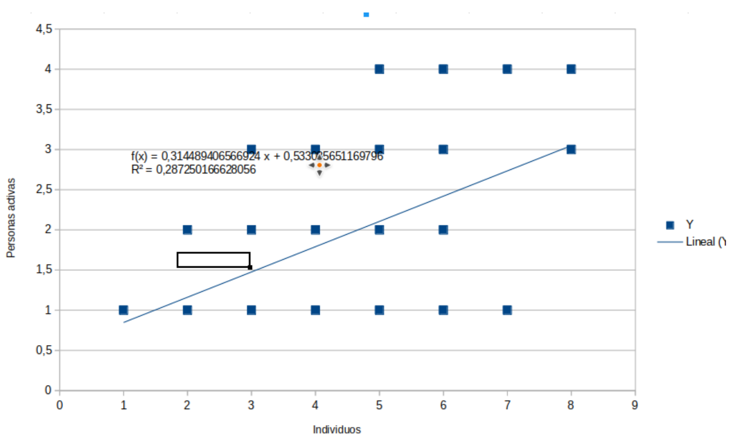
$$r = \sqrt{r^2} = \sqrt{\frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2}}$$

Calculamos la varianza de Y:

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

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}2872 \Rightarrow r = 0{,}536$$

No es adecuado suponer una relación lineal para explicar el comportamiento de Y a partir de X, el coeficiente está muy alejado de 1.



4. Se realiza un estudio sobre la tensión de vapor de agua (Y, en ml. De Hg.) a distintas temperaturas (X, en ºC). Efectuadas 21 medidas, explicar el comportamiento de la tensión de vapor en términos de la temperatura mediante una función lineal. ¿Es adecuado asumir este tipo de relación?

Habría que obtener la recta de regresión Y/X:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \bar{x})$$

Introducimos ya la tabla con las marcas de clase y columnas y filas auxiliares para facilitar el cálculo de la media, la covarianza y las varianzas.

X\Y	1	2	4	$n_{i.}$	$x_i \cdot n_{i.}$	$c_i^2 \cdot n_{i.}$	$c_i \sum_{i=1}^{5} n_{ij} \cdot c_j$
8	4	2	0	6	48	384	64
20	1	4	2	7	140	2800	340
27,5	0	3	5	8	220	6050	715
$n_{.j}$	5	9	7	21		9234	1119
$c_j^2 \cdot n_{.j}$	5	36	112	153			
$y_j \cdot n_{.j}$	5	18	28				

$$\bar{x} = \frac{\sum_{i=1}^{4} c_i \cdot n_{i.}}{n} = 19{,}429 \ ºC \qquad \bar{y} = \frac{\sum_{j=1}^{5} c_j \cdot n_{.j}}{n} = 2{,}429 \ ml.de\, Hg.$$

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{3} \sum_{j=1}^{3} n_{ij} \cdot c_i \cdot c_j - \bar{x}\bar{y} = 6{,}092$$

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{4} n_{i.} \cdot c_i^2 - \bar{x}^2 = 62{,}228 \Rightarrow \sigma_x = +\sqrt{\sigma_x^2} = 7{,}888 ºC^2$$

$$\sigma_y^2 = \frac{1}{n} \sum_{j=1}^{5} n_{.j} \cdot c_j^2 - \bar{y}^2 = 1{,}385 \Rightarrow \sigma_y = +\sqrt{\sigma_y^2} = 1{,}177 \ ml.de\, Hg.^2$$

La recta quedaría de la siguiente forma:

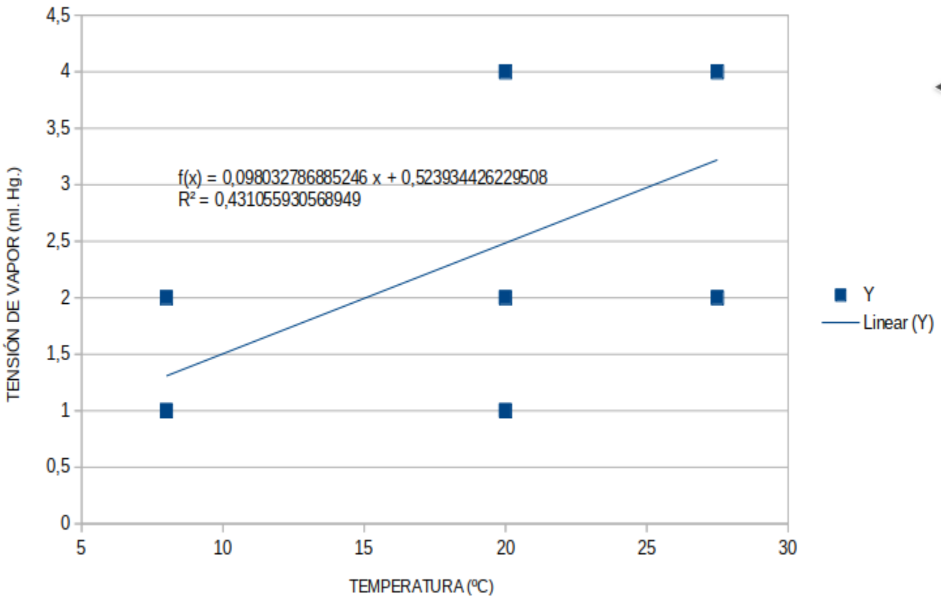
$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} \cdot x - \frac{\sigma_{xy}}{\sigma_x^2} \cdot \bar{x}$$

$$y = 0{,}979x + 0{,}527$$

Para determinar si es adecuado asumir una relación lineal, calculamos la razón de correlación lineal:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}431$$

Tenemos que la función explica menos del 50% de los casos, luego no será adecuado asumir este tipo de relación.



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.

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

Se dice que el carácter Y es independiente estadísticamente del carácter X, si las distribuciones de Y condicionadas a cada valor de x, son idénticas para todo $x_{i\ i=1,2,\ldots,k}$. Es decir:

$$f_j^i \equiv f_{j/i}$$

Por tanto, se debe cumplir lo siguiente:

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

En el caso de la tabla que tenemos arriba, vamos a comprobar la independencia de las dos variables. Cabe decir que si resulta que Y sea independiente al carácter X, el recíproco también es cierto.

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

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

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

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

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

Como se puede observar, Y es independiente del carácter X, así que no tiene sentido estudiar curvas de regresión. Además podemos afirmar:

$$\sigma_{xy} = 0$$

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

Es evidente que estas dos variables no son independientes, puesto que si algún $n_{ij} = 0$, la igualdad

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

no se va a dar a menos que $n_{ij} = 0 \quad \forall i, j$, lo cual no tendría sentido porque sería estudiar una variable que no ha presentado ninguna distribución de frecuencias.

Observando ahora la posible dependencia funcional, observamos que X no depende funcionalmente de Y, porque a la modalidad 2 de Y, le corresponden dos posibles modalidades de X (señalado en la tabla).

Procedemos a calcular la covarianza:

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

X\Y	1	2	3	$n_{i.} \cdot x_i$	$x_i \sum_{j=1}^{3} n_{ij} \cdot y_j$
-1	0	1	0	-1	-2
0	1	0	1	0	0
1	0	1	0	1	2
			$n = 4$	0	0

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

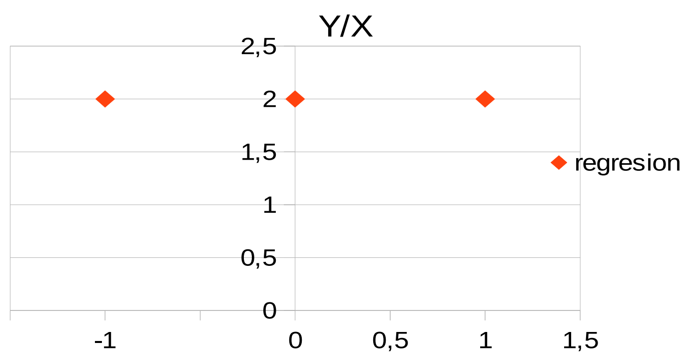
Como la media de x es 0, la covarianza será igual a m_{11}. Resulta que:

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{3} \sum_{j=1}^{3} x_i \cdot y_j \cdot n_{ij} = 0$$

Además, podemos observar que si las variables son independientes, la covarianza es 0, pero el recíproco no podemos asegurarlo. Aquí vemos un caso en el que las variables son dependientes y la covarianza resulta 0.

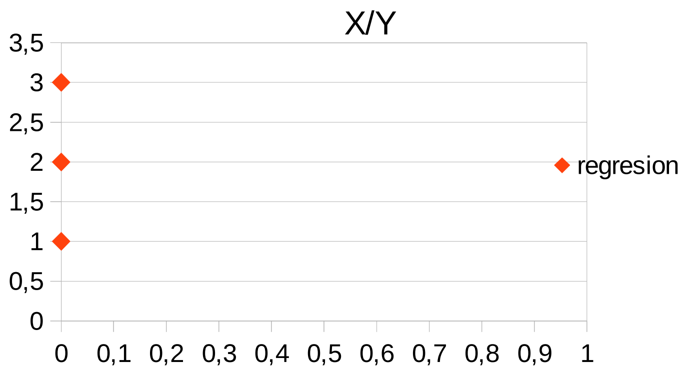
Ahora calculamos la curva de regresión de tipo 1 de Y/X, que sabemos que es la curva que pasa por los puntos $(x_i, \bar{y}_i)$; $i = 1, \ldots, k$.

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



La curva de regresión de tipo 1 de X/Y, es la que pasa por los puntos $(\bar{x}_j, y_j); \quad j = 1,\ldots,p$.

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



6. Dada la siguiente distribución bidimensional:

X\Y	1	2	3	4
10	1	3	0	0
12	0	1	4	3
14	2	0	0	2
16	4	0	0	0

a) ¿Son estadísticamente independientes X e Y?
b) Calcular y representar las curvas de regresión de X/Y e Y/X.
c) Cuantificar el grado en que cada variable es explicada por la otra mediante la correspondiente curva de regresión.
d) ¿Están X e Y correladas linealmente? Dar las expresiones de las rectas de regresión.

Con objeto de facilitar posteriores cálculos, se añadirán más columnas y filas a la tabla:

X\Y	1	2	3	4	$n_{i.}$	$n_{i.} \cdot x_i$	$n_{i.} \cdot x_i^2$	$x_i \sum_j n_{ij} \cdot y_j$
10	1	3	0	0	4	40	400	70
12	0	1	4	3	8	96	1152	312
14	2	0	0	2	4	56	784	140
16	4	0	0	0	4	64	1024	64
$n_{.j}$	7	4	4	5	20	256	3360	586
$y_j \cdot n.j$	7	8	12	20	47			
$y_j^2 \cdot n.j$	7	16	36	80	139			

Antes de comenzar, vamos a calcular la media de x, de y, las varianzas y la covarianza:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i \cdot n_{i.}}{n} = 12{,}8 \qquad \bar{y} = \frac{\sum_{j=1}^{4} y_j \cdot n_{.j}}{n} = 2{,}35$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{4} n_{i.} \cdot x_i^2 - \bar{x}^2 = 4{,}16$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{4} n_{.j} \cdot y_j^2 - \bar{y}^2 = 1{,}4275$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij} \cdot x_i \cdot y_j - \bar{x}\,\bar{y} = -0{,}78$$

a) Es evidente que estas dos variables no son independientes, puesto que si algún $n_{ij} = 0$, la igualdad

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

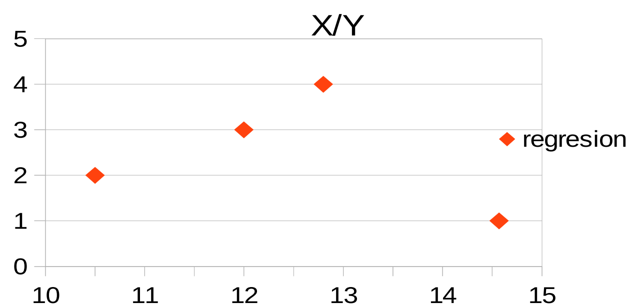
no se va a dar a menos que $n_{ij} = 0 \quad \forall i,j$, lo cual no tendría sentido porque sería estudiar una variable que no ha presentado ninguna distribución de frecuencias. Tampoco son funcionalmente dependientes, puesto que para cada modalidad de X no encontramos una única modalidad de Y con frecuencia no nula, e igual ocurre con Y.

b) La curva de regresión de tipo 1 de X/Y, es la que pasa por los puntos $(\bar{x}_j, y_j); \quad j = 1,\ldots,p$.

$$\bar{x}_1 = \frac{10+2\cdot 14+4\cdot 16}{7} = 14{,}57 \qquad \bar{x}_2 = \frac{3\cdot 10+12}{4} = 10{,}5 \qquad \bar{x}_3 = \frac{12\cdot 4}{4} = 12$$

$$\bar{x}_4 = \frac{3\cdot 12+2\cdot 14}{5} = 12{,}8$$

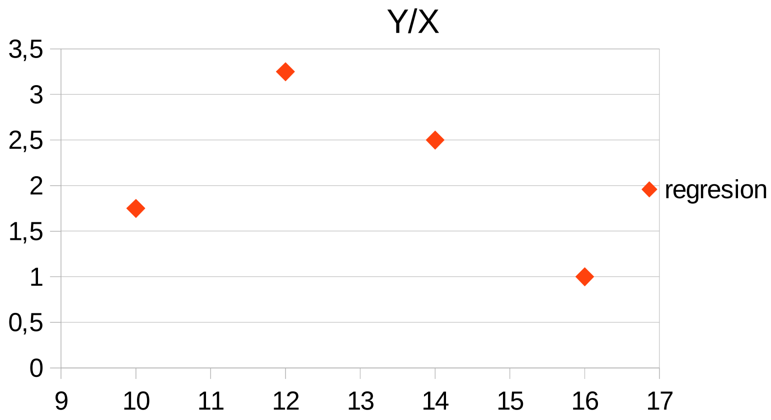
Punto 1: $(14{,}57\,,\ 1)$ Punto 2: $(10{,}5\,;\ 2)$ Punto 3: $(12\,,\ 3)$ Punto 4: $(12{,}8\,;\ 4)$



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

$$\bar{y}_{10} = \frac{1+2\cdot 3}{4} = 1{,}75 \quad \bar{y}_{12} = \frac{2+4\cdot 3+3\cdot 4}{8} = 3{,}25 \quad \bar{y}_{14} = \frac{2+2\cdot 4}{4} = 2{,}5 \quad \bar{y}_{16} = \frac{4}{4} = 1$$

Punto 1: $(10,\ 1{,}75)$ Punto 2: $(12,\ 3{,}25)$ Punto 3: $(14,\ 2{,}5)$ Punto 4: $(16;\ 1)$



c) <u>Para este ejercicio, las desviaciones típicas han sido calculadas en el apartado d).</u>

Calculamos los valores de la razón de correlación lineal de Y/X y X/Y:

$$\eta^2_{Y/X} = \frac{\sigma^2_{ey}}{\sigma^2_y}$$

$$\sigma_{ey} = \frac{1}{n}\sum_{1=1}^{k}\sum_{j=1}^{p} n_{ij}(\bar{y}_i - y_j)^2 = 0{,}765$$

$$\eta^2_{Y/X} = 0{,}5359$$

$$\eta^2_{X/Y} = \frac{\sigma^2_{ex}}{\sigma^2_x}$$

$$\sigma_{ex} = \frac{1}{n}\sum_{1=1}^{k}\sum_{j=1}^{p} n_{ij}(\bar{x}_j - x_i)^2 = 2{,}2843$$

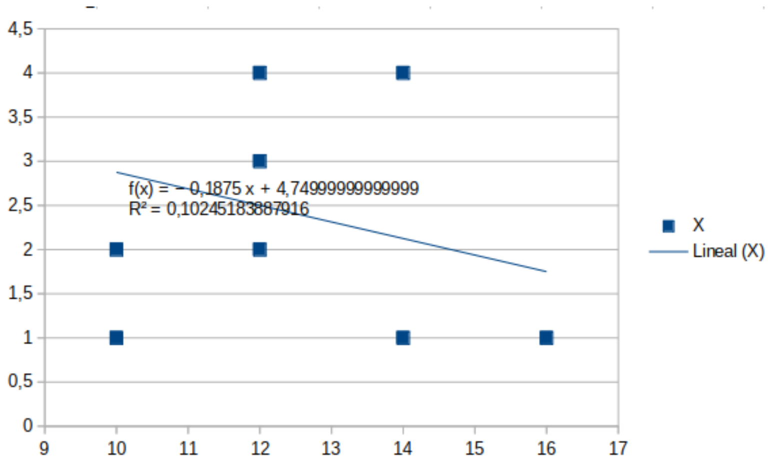
$$\eta^2_{X/Y} = 0{,}5491$$

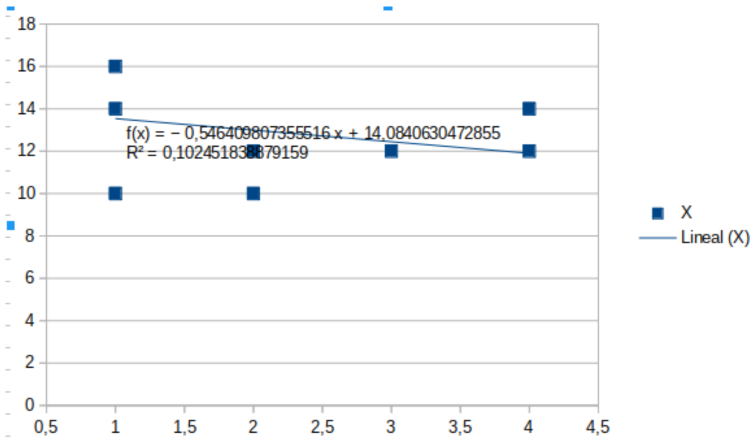
Cada variable explica en torno a un 54% de la otra.

d) Vamos a calcular la recta de regresión de Y sobre X y la recta de regresión de X sobre Y que vienen dadas respectivamente por:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x-\bar{x}) \qquad x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y-\bar{y})$$

$$y = -0{,}1875\,x + 4{,}75 \qquad x = -0{,}5464\,y + 14{,}084$$





Para ver si X e Y están correladas linealmente, calculamos el coeficiente de correlación lineal de Pearson. Hacemos primero la razón de correlación lineal y tomamos la raíz cuadrada negativa (puesto que la covarianza es negativa).

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}1024 \Rightarrow r = -\sqrt{r^2} = -0{,}320$$

Por tanto, no están correladas linealmente (valor bastante alejado de -1). Como se puede observar, la razón de correlación de nuestras rectas no supera a la de la recta de regresión de tipo 1. lo cual siempre se cumple.

7. Para cada una de las distribuciones:

Distribución A:

X\Y	10	15	20
1	0	2	0
2	1	0	0
3	0	0	3
4	0	1	0

Distribución B:

X\Y	10	15	20
1	0	2	0
2	1	0	0
3	0	0	3

Distribución C:

X\Y	10	15	20	25
1	0	3	0	1
2	0	0	1	0
3	2	0	0	0

a) ¿Dependen funcionalmente X de Y o Y de X?
b) Calcular las curvas de regresión y comentar los resultados.

a) En la distribución A, Y depende funcionalmente de X ya que para cada valor de X existe un único valor de Y.

En la distribución B, X depende funcionalmente de Y, e Y depende funcionalmente de X, ya que para cada X existe un único Y y para cada Y existe un único X.

En la distribución C, X depende funcionalmente de Y ya que para cada Y existe un único X.

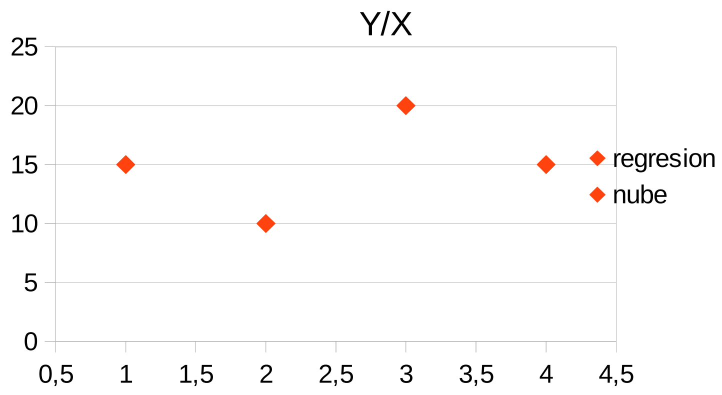
b) En adelante, para todas las distribuciones, esta es la descripción de los puntos por donde pasa la curva:

1. Regresión de tipo 1 de Y/X, que sabemos que es la curva que pasa por los puntos $(x_i, \bar{y}_i); \quad i = 1,\dots,k$.
2. Regresión de tipo 1 de X/Y, es la que pasa por los puntos $(\bar{x}_j, y_j); \quad j = 1,\dots,p$.

Distribución A
Y/X

$$\bar{y}_1 = \frac{2\cdot 15}{2} = 15 \qquad \bar{y}_2 = 10 \qquad \bar{y}_3 = \frac{3\cdot 20}{3} = 20 \qquad \bar{y}_4 = 15$$

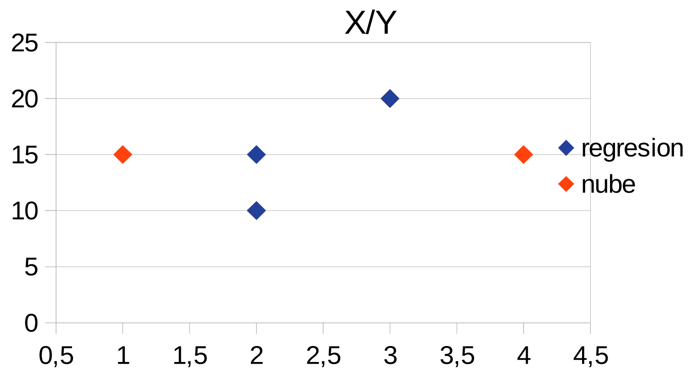
Pasará por los puntos: Punto 1: $(1,\ 15)$ Punto 2: $(2,\ 10)$ Punto 3: $(3,\ 20)$ Punto 4: $(4,\ 15)$.



X/Y

$$\bar{x}_{10} = 2 \qquad \bar{x}_{15} = \frac{2+4}{3} = 2 \qquad \bar{x}_{20} = 3$$

Pasará por los puntos: Punto 1: $(2,\ 10)$ Punto 2: $(2,\ 15)$ Punto 3: $(3,\ 20)$

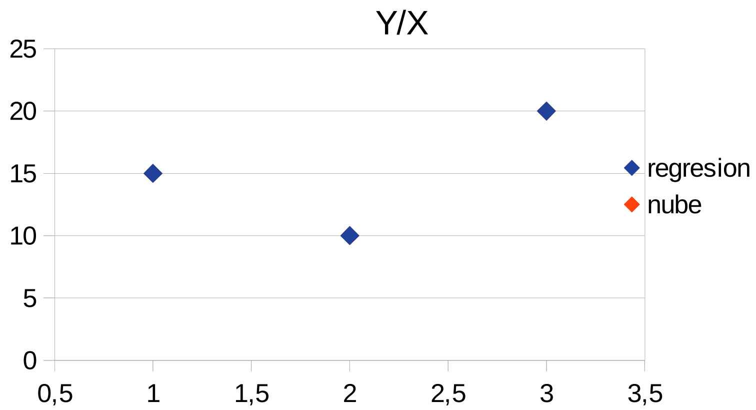


Distribución B

Y/X

$$\bar{y}_1 = 15 \qquad \bar{y}_2 = 10 \qquad \bar{y}_3 = 20$$

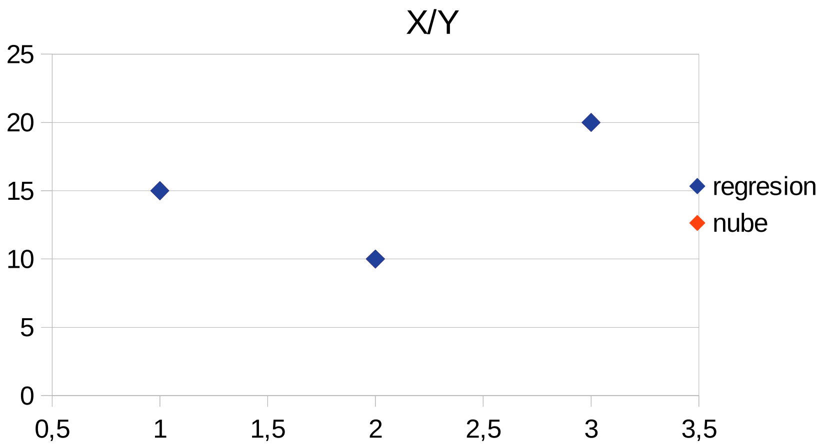
Pasará por los puntos: Punto 1: $(1,\ 15)$ Punto 2: $(2,\ 10)$ Punto 3: $(3,\ 20)$.



X/Y

$$\bar{x}_{10} = 2 \quad \bar{x}_{15} = 1 \quad \bar{x}_{20} = 3$$

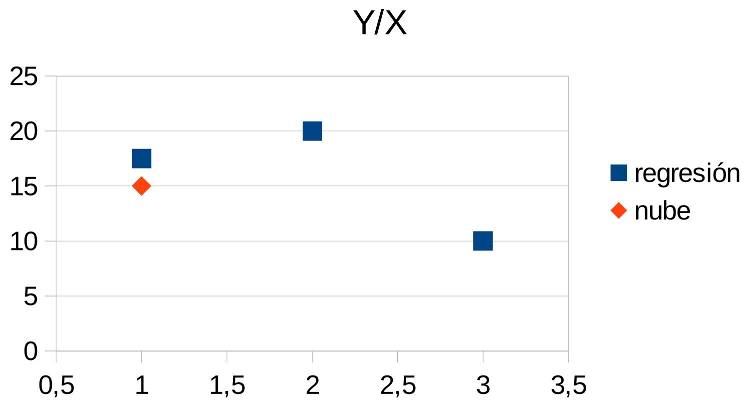
Pasará por los puntos: Punto 1: $(2,\ 10)$ Punto 2: $(1,\ 15)$ Punto 3: $(3,\ 20)$.



Distribución C
Y/X

$$\bar{y}_1 = \frac{3 \cdot 15 + 25}{4} = 17{,}5 \quad \bar{y}_2 = 20 \quad \bar{y}_3 = 10$$

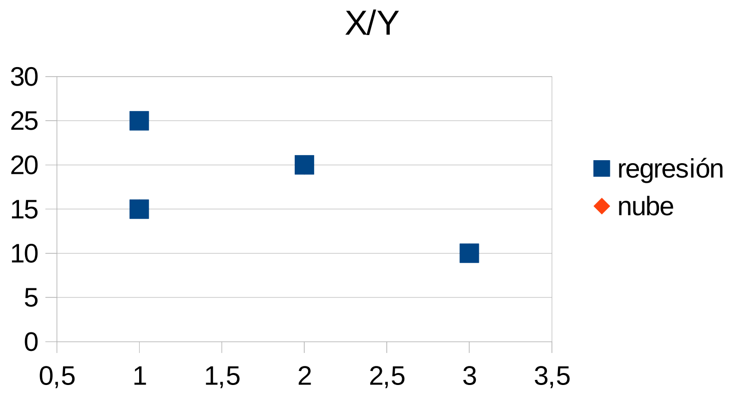
Pasará por los puntos: Punto 1: $(1,\ 17{,}5)$ Punto 2: $(2,\ 20)$ Punto 3: $(3,\ 10)$.



X/Y

$$\bar{x}_{10} = 3 \quad \bar{x}_{15} = 1 \quad \bar{x}_{20} = 2 \quad \bar{x}_{25} = 1$$

Pasará por los puntos: Punto 1: $(3,\ 10)$ Punto 2: $(1,\ 15)$ Punto 3: $(2,\ 20)$ Punto 4: $(1,\ 25)$.



Como se puede observar, cuando en una distribución hay dependencia funcional (caso b), ambas curvas de regresión coinciden con los puntos de la nube. En el caso a, donde Y depende funcionalmente de X, la curva de regresión Y/X coincide con la nube de puntos y, por último, en el caso c, donde X depende funcionalmente de Y, la curva de regresión X/Y coincide con la nube de puntos.

8) 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 (que tiene filas y columnas extras para facilitar el posterior cálculo de las varianzas, las medias y la covarianza):

X\Y	1	2	3	4	$n_{i.}$	$n_{i.}x_i^2$
1	1	2	0	0	3	3
2	1	2	3	1	7	28
3	0	1	2	6	9	81
4	0	0	2	3	5	80
$n_{.j}$	2	5	7	10	24	
$n_{.j}y_j^2$	2	20	63	160		

a) Determinar las rectas de regresión.

Vamos a calcular la recta de regresión de Y sobre X y la recta de regresión de X sobre Y que vienen dadas respectivamente por:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x-\bar{x}) \qquad x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y-\bar{y})$$

Comenzamos calculando las medias aritméticas, las varianzas y la covarianza de X y de Y:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i \cdot n_{i.}}{n} = 2{,}667\, balanzas \qquad \bar{y} = \frac{\sum_{j=1}^{4} y_j \cdot n_{.j}}{n} = 3{,}041\, dependientes$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{4} n_{i.} \cdot x_i^2 - \bar{x}^2 = 0{,}888\, balanzas^2$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{4} n_{.j} \cdot y_j^2 - \bar{y}^2 = 0{,}9563\, dependientes^2$$

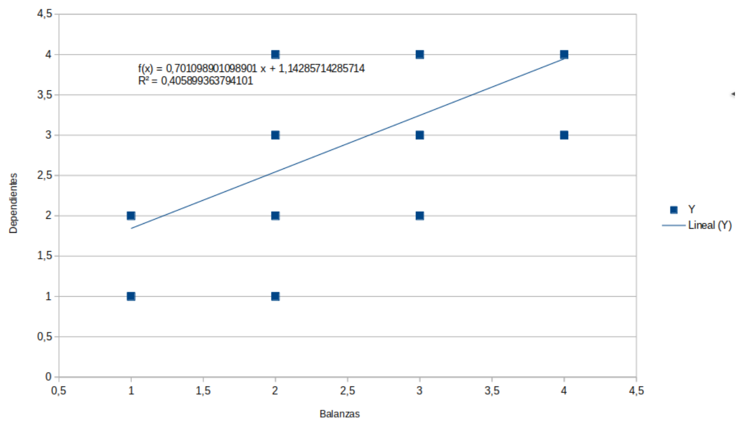
$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij} \cdot x_i \cdot y_j - \bar{x}\,\bar{y} = 0{,}5970$$

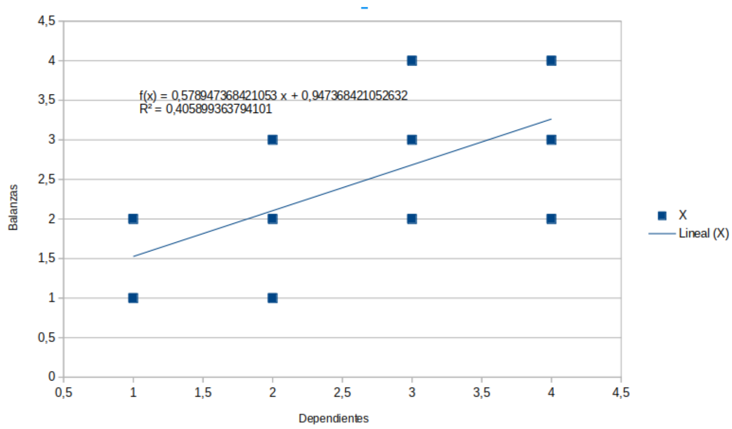
Recta de regresión de Y sobre X:

$$y = 0{,}6717\, x + 1{,}2505$$

Recta de regresión de X sobre Y:

$$x = 0{,}6243\, x + 0{,}7678$$





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

Para ello calculamos el coeficiente de correlación lineal:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}41934 \Rightarrow r = +\sqrt{r^2} = 0{,}6475$$

No, puesto que el grado de dependencia no es demasiado elevado.

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

No podemos salirnos de los límites de nuestra recta de regresión, que llega hasta 4. El cálculo de una predicción fuera de estos límites no sería de ninguna manera fiable.

9. Se eligen a 50 matrimonios al azar y se les pregunta la edad de ambos al contraer matrimonio. Los resultados se recogen en la siguiente tabla, en la que la X denota la edad del hombre e Y la de la mujer:

Introducimos la tabla ya con las marcas de clase calculadas:

X\Y	15	22,5	27,5	32,5	37,5	$n_{i.}$	$c_i \cdot n_{i.}$
16,5	3	2	3	0	0	8	132
19,5	0	4	2	2	0	8	156
22,5	0	7	10	6	1	24	540
25,5	0	0	2	5	3	10	255
$n_{.j}$	3	13	17	13	4	50	
$c_j \cdot n_{.j}$	45	292,5	467,5	422,5	150		

$$\bar{x} = \frac{\sum_{i=1}^{4} c_i \cdot n_{i.}}{n} = 21{,}66 \ \text{años hombre} \qquad \bar{y} = \frac{\sum_{j=1}^{5} c_j \cdot n_{.j}}{n} = 27{,}55 \ \text{años mujer}$$

Empleamos de nuevo la tabla para añadir datos que nos serán útiles para hallar la covarianza y las varianzas:

X\Y	15	22,5	27,5	32,5	37,5	$c_i^2 \cdot n_{i.}$	$c_i \sum_{i=1}^{5} n_{ij} \cdot c_j$
16,5	3	2	3	0	0	2178	2846,25
19,5	0	4	2	2	0	3042	4095
22,5	0	7	10	6	1	12150	14962,5
25,5	0	0	2	5	3	6502,5	8415
$c_j^2 \cdot n_{.j}$	675	6581,25	12836,3	13731,3	5625		30318,75

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{4} n_{i.}\cdot c_i^2 - \bar{x}^2 = 8{,}2944 \Rightarrow \sigma_x = +\sqrt{\sigma_x^2} = 2{,}88 \ \ años \ \ hombre^2$$

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{5} n_{.j}\cdot c_j^2 - \bar{y}^2 = 30{,}3745 \Rightarrow \sigma_y = +\sqrt{\sigma_y^2} = 5{,}5113 \ \ años \ \ mujer^2$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij}\cdot c_i\cdot c_j - \bar{x}\,\bar{y} = 9{,}642$$

Para ver el grado de dependencia lineal calculamos el coeficiente de correlación lineal:

$$r = \frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y} = 0{,}6075$$

La correlación es positiva y el grado de dependencia lineal es elevado.

10. Calcular el coeficiente de correlación lineal de dos variables cuyas rectas de regresión son:

$$x + 4y = 1$$

$$x + 5y = 2$$

Supongamos que la primera es la recta para X/Y y la segunda para Y/X.

Además, sabemos que ambas rectas tienen el mismo signo en la pendiente (que lo proporciona la covarianza).

$$(i) \quad x = -4y + 1 \qquad (ii) \quad y = -\frac{1}{5}x + \frac{2}{5}$$

$$a_1 = \frac{\sigma_{xy}}{\sigma_y^2} \qquad a_2 = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$a_1 \cdot a_2 = r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = \frac{4}{5}$$

Nos piden el coeficiente de correlación lineal, r, que sabemos que tiene que ser negativo por que la covarianza es negativa (en las dos rectas la pendiente es negativa).

$$r = -\sqrt{r^2} = -\frac{\sqrt{4}}{\sqrt{5}} = -\frac{2}{\sqrt{5}}$$

Dicho resultado tiene sentido por que $-1 \leq r \leq 1$.

Supongamos ahora que la primera es la recta para Y/X y la segunda para X/Y.

$$(i) \quad y = -\frac{1}{4}x + \frac{1}{4} \qquad (ii) \quad x = -5y + 2$$

$$a_1 = \frac{\sigma_{xy}}{\sigma_y^2} \qquad a_2 = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$a_1 \cdot a_2 = r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 1{,}25 > 1$$

Esta suposición no tiene sentido, puesto que se sale de los límites entre los que se tiene que situar el valor de la razón de correlación lineal ($0 \leq r \leq 1$ **).**

11. Consideremos una distribución bidimensional en la que la recta de regresión de Y sobre X es $y = 5x - 20$ y $\sum y_j^2 n_j = 3240$. Supongamos, además, que la distribución marginal de X es:

x_i	3	5	8	9
$n_{i.}$	5	1	2	1

Determinar la recta de regresión de X sobre Y, y la bondad de los ajustes lineales.

x_i	$n_{i.}$	$x_i \cdot n_{i.}$	$x_i^2 \cdot n_{i.}$	$n_{i.}(x_i - \bar{x})^2$
3	5	15	45	20
5	1	5	25	0
8	2	16	128	18
9	1	9	81	16
	9	45	279	54

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i \cdot n_{i.}}{n} = \frac{45}{9} = 5 \qquad \sigma_x^2 = \frac{\sum_{i=1}^{4} n_{i.}(xi - \bar{x})^2}{n} = 6$$

Tenemos como dato $y = 5x - 20$ y $\sum y_j^2 n_j = 3240$, siendo la primera la recta de Y sobre X. Debemos hallar $x = a'y + b'$. Consideremos nuestra primera recta $y = ax + b$:

$$a' = \frac{\sigma_{xy}}{\sigma_y^2}$$

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \Rightarrow \sigma_{xy} = a \cdot \sigma_x^2 = 6a = 5 \cdot 6 = 30$$

$$\bar{y} = 5\bar{x} - 20 = 5 \cdot 5 - 20 = 5$$

$$\sigma_{xy} = m_{11} - m_{10} \cdot m_{01}$$

$$m_{11} = \sigma_{xy} + m_{10} \cdot m_{01} = 30 + 5 \cdot 5 = 55$$

$$m_{02} = \frac{1}{n} \sum n_{.j} \cdot y_j^2 = \frac{3290}{9} = 360$$

$$m_{11} = a' m_{22} + b' m_{01}$$

$$m_{10} = a' m_{01} + b'$$

Si resolvemos el sistema (estas dos últimas ecuaciones), resulta:

$$a' = 0.0896$$

$$b' = 4{,}552$$

Por último, para calcular la bondad de los ajuste lineales, hará falta la razón de correlación lineal, pero antes hemos de hallar la desviación típica de y:

$$\sigma_y^2 = \frac{1}{n} \sum n_{.j} \cdot y_j^2 - \bar{y}^2 = 335$$

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = \frac{30^2}{6 \cdot 335} = 0{,}448$$

Nuestro ajuste explica menos del 50% de los casos, no es muy bueno.

12. De la 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:

$$\sum y_i = 219{,}719 \qquad \sum y_i^2 = 2396{,}504 \qquad \sum x_i \cdot y_i = 349{,}486$$

$$\sum x_i = 31{,}470 \qquad \sum x_i^2 = 51{,}075 \qquad \sum x_i^2 \cdot y_i = 633{,}993$$

$$\sum x_i^4 = 182{,}977 \qquad \sum x_i^3 = 93{,}6$$

La variable Y expresa el consumo total de combustible, en miles de libras, correspondiente a un vuelo de duración X (el tiempo se expresa en horas, y se utilizan como unidades de orden inferior fracciones decimales de la hora).

a) Ajustar un modelo del tipo $Y = ax + b$ ¿Qué consumo total se estimaría para un programa de vuelos compuestos de 100 vuelos de media hora, 200 de un hora y 100 de dos horas?¿Es fiable esta estimación?

Para calcular las medias, la covarianza y las desviaciones típicas, haré uso de los momentos, fácilmente calculables a partir de los datos que me dan en el enunciado.

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

$$m_{10} = \frac{31{,}470}{24} = 1{,}3113 \qquad m_{01} = \frac{219{,}719}{24} = 9{,}155$$

$$m_{11} = \frac{349{,}486}{24} = 14{,}562 \qquad \sigma_{xy} = m_{11} - m_{10} \cdot m_{01} = 2{,}557$$

$$m_{20} = \frac{51{,}075}{24} = 2{,}1281 \qquad m_{02} = \frac{2396{,}504}{24} = 99{,}8543$$

$$\sigma_x^2 = m_{20} - m_{10}^2 = 0{,}4086 \qquad \sigma_y^2 = m_{02} - m_{01}^2 = 16{,}0403$$

$$y = \frac{2{,}557}{0{,}4086}x - \frac{2{,}557}{0{,}4086} \cdot 1{,}3113 + 9{,}155 = 6{,}258\,x + 0{,}9489$$

Realicemos ahora el consumo estimado:

- Para 100 vuelos de media hora: $y = 6{,}258 \cdot 0{,}5 + 0{,}9489 = 4{,}0779 \cdot 100 = 407{,}79$.
- Para 200 vuelos de una hora: $y = 6{,}258 \cdot 1 + 0{,}9489 = 7{,}2069 \cdot 200 = 1441{,}38$.
- Para 100 vuelos de dos hora: $y = 6{,}258 \cdot 2 + 0{,}9489 = 13{,}4649 \cdot 100 = 1346{,}49$.
- El consumo total se estima en 3195,66 millones de libras.

Para ver como de bueno es el ajuste, calculamos la razón de correlación lineal:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}9975$$

Como se puede observar, el valor es muy cercano a 1, por lo tanto el ajuste es muy bueno y las estimaciones son fiables.

b) Ajustar un modelo del tipo $Y = a + bX + cX^2$ ¿Qué consumo total estimaría para el mismo programa de vuelos del apartado a)?

Para realizar este ejercicio, tenemos el siguiente sistema de ecuaciones:

$$\begin{aligned} m_{01} &= a + b\,m_{10} + c\,m_{20} \\ m_{11} &= a\,m_{10} + b\,m_{20} + c\,m_{30} \\ m_{21} &= a\,m_{20} + b\,m_{30} + c\,m_{40} \end{aligned}$$

Los diferentes momentos me los dan de partida en el enunciado. Obtenemos como soluciones:

$$a = 0{,}7923$$
$$b = 6{,}5549$$
$$c = -0{,}1004$$

El ajuste queda:

$$y = 0{,}7923 + 6{,}5549\,x - 0{,}1094\,x^2$$

Realicemos ahora el consumo estimado:

- Para 100 vuelos de media hora: $y = 0{,}7923 + 6{,}5549\cdot 0{,}5 - 0{,}1094\cdot 0{,}5^2 = 4{,}0424\cdot 100$.
- Para 200 vuelos de una hora: $y = 0{,}7923 + 6{,}5549\cdot 1 - 0{,}1094\cdot 1^2 = 7{,}2378\cdot 200$
- Para 100 vuelos de dos horas: $y = 0{,}7923 + 6{,}5549\cdot 2 - 0{,}1094\cdot 2^2 = 13{,}4645\cdot 100$.
- El consumo total se estima en 3198,25 millones de libras.

c) ¿Cuál de los dos modelos se ajusta mejor? Razonar la respuesta.

Si prestamos atención al segundo modelo, la ecuación tiene la misma forma que la recta pero con un coeficiente más, es decir, añadimos más información de la que teníamos con la recta y además no perdemos esta. Por tanto, la razón de correlación de la parábola siempre será mayor o igual que la razón de correlación lineal.

13. La curva de Engel, que expresa el gasto en un determinado bien en función de la renta, adopta en ocasiones la forma de una hipérbola equilátera. Ajustar dicha curva a los siguientes datos, en los que x denota la renta en miles de euros e Y el gasto en euros. Cuantificar la bondad del ajuste:

X	**10**	**12,5**	**20**	**25**
Y	**50**	**90**	**160**	**180**

Invertimos los valores de X de la forma $X' = \dfrac{1}{X}$:

X	$\frac{1}{10}$	$\frac{1}{12{,}5}$	$\frac{1}{20}$	$\frac{1}{25}$
Y	50	90	160	180

Calculamos la recta de regresión lineal de Y sobre X', que viene dada por:

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

Calculamos la medias aritméticas, las varianza de X y la covarianza:

$$\bar{x}' = \frac{\sum_{i=1}^{4} x'_i \cdot n_{i.}}{n} = 0{,}0675 \ \ miles \ \ de \ \ euros \qquad \bar{y} = \frac{\sum_{j=1}^{4} y_j \cdot n_{.j}}{n} = 120\, euros$$

$$\sigma_{x'}^2 = \frac{1}{n} \sum_{i=1}^{4} n_{i.} \cdot x'^2_i - \bar{x}'^2 = 0{,}0005687 \ \ miles \ \ de \ \ euros^2$$

$$\sigma_{x'y} = \frac{1}{n} \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} \cdot x'_i \cdot y_j - \bar{x}' \bar{y} = -1{,}25$$

La recta de regresión de Y sobre X quedaría:

$$y = \frac{\sigma_{x'y}}{\sigma_{x'}^2} x' - \frac{\sigma_{x'y}}{\sigma_{x'}^2} \bar{x}' + \bar{y} = -2197\, x' + 268{,}36$$

Deshacemos el cambio:

$$y = -\frac{2197{,}99}{x} + 268{,}36$$

Para cuantificar la bondad del ajuste calculamos la razón de correlación:

$$\eta_{Y/X}^2 = \frac{\sigma_{ey}^2}{\sigma_y^2}$$

Siendo σ_{ey}^2 la varianza explicada por la regresión.

$$\sigma_{ey}^2 = \frac{1}{n} \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} \left(f(x_i) - \bar{y}\right)^2$$

$$f(10) = 48{,}561$$
$$f(12{,}5) = 92{,}5208$$
$$f(20) = 158{,}4605$$
$$f(25) = 180{,}4404$$

$$\sigma_{ey}^2 = \frac{1}{n} \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} \left(f(x_i) - \bar{y}\right)^2 = 2747{,}72$$

Calculamos ahora la varianza de Y:

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{4} n_{.j} \cdot y_i^2 - \bar{y}^2 = 2750 \ \ euros^2$$

$$\eta^2_{Y/X} = \frac{\sigma^2_{ey}}{\sigma^2_y} = 0{,}99917$$

El ajuste es muy bueno, la razón de correlación es muy próxima a 1.

14. Se dispone de la siguiente información referente al gasto en espectáculos (Y, en euros) y la renta disponible mensual (X, en cientos de euros) de 6 familias.

Y	**30**	**50**	**70**	**80**	**120**	**140**
X	**9**	**10**	**12**	**15**	**22**	**32**

Explicar el comportanmiento de Y por X mediante:

a) **Relación lineal.**
b) **Hipérbola equilátera.**
c) **Curva potencial.**
d) **Curva exponencial.**

¿Qué ajuste es más adecuado? (Asumimos en todos los gráficos que el eje X representa la renta disponible mensual en cientos de euros y el eje Y el gasto en espectáculos en euros).

a) **Relación lineal:**

Tendremos que calcular la recta de regresión:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma^2_x} \cdot x - \frac{\sigma_{xy}}{\sigma^2_x} \cdot \bar{x}$$

Para hacer más sencillos los cálculos, empleará una tabla, con objeto de calcular las medias:

X\Y	30	50	70	80	120	140	$n_{i.}$	$n_{i.} \cdot x_i$
9	1	0	0	0	0	0	1	9
10	0	1	0	0	0	0	1	10
12	0	0	1	0	0	0	1	12
15	0	0	0	1	0	0	1	15
22	0	0	0	0	1	0	1	22
32	0	0	0	0	0	1	1	32
$n_{.j}$	1	1	1	1	1	1	6	
$n_{.j} \cdot y_j$	30	50	70	80	120	140		

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot n_{i.}}{n} = 16{,}667 \ \ cientos \ \ de \ \ euros \qquad \bar{y} = \frac{\sum_{i=1}^{6} y_j \cdot n_{.j}}{n} = 81{,}667 \ \ euros$$

Volvemos a copiar la tabla, esta vez con el objetivo de calcular las varianzas y la covarianza:

X\Y	30	50	70	80	120	140	$n_{i.} \cdot x_i^2$	$x_i \sum_{i=1}^{6} n_{ij} \cdot y_j$
9	1	0	0	0	0	0	81	270
10	0	1	0	0	0	0	100	500
12	0	0	1	0	0	0	144	840
15	0	0	0	1	0	0	225	1200
22	0	0	0	0	1	0	484	2640
32	0	0	0	0	0	1	1024	4480
$n_{.j} \cdot y_j^2$	900	2500	4900	6400	14400	19600		9930

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{6} n_{i.} \cdot x_i^2 - \bar{x}^2 = 65{,}221 \ \ cientos \ \ de \ \ euros^2$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{6} n_{.j} \cdot y_i^2 - \bar{y}^2 = 1447{,}217 \ \ euros^2$$

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} n_{ij} \cdot x_i \cdot y_j - \bar{x}\,\bar{y} = 293{,}884$$

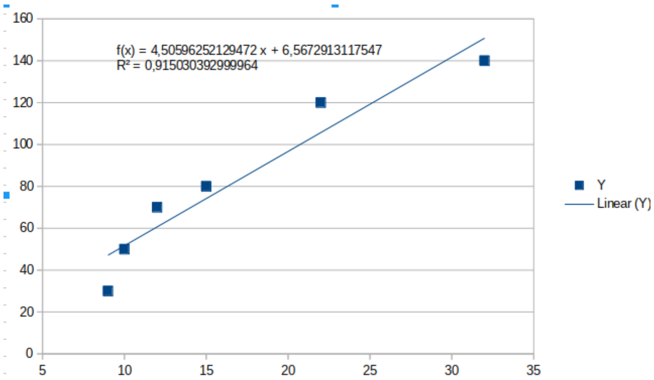
Sabiendo que:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} \cdot x - \frac{\sigma_{xy}}{\sigma_x^2} \cdot \bar{x}$$

$$y = 4{,}506\,x + 6{,}567$$

Ahora calculamos el coeficinete de determinación lineal para ver cómo de bueno es el ajuste:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = 0{,}915$$



b) Hipérbola equilátera:

Tenemos que obtener una expresión $y = \frac{a}{x} + b$, por lo que llamaremos $z = \frac{1}{x}$ y procederemos a obtener $y = az + b$.

Y	30	50	70	80	120	140
Z	0,111	0,1	0,083	0,067	0,046	0,031

$$\bar{z} = \frac{\sum_{i=1}^{6} z_i \cdot n_{i.}}{n} = 0{,}073$$

$$\sigma_z^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.} \cdot z_i^2 - \bar{z}^2 = 0{,}0008$$

$$\sigma_{zy} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij} \cdot z_i \cdot y_j - \bar{z}\,\bar{y} = -1{,}069$$

$$y = \bar{y} + \frac{\sigma_{zy}}{\sigma_z^2} \cdot z - \frac{\sigma_{zy}}{\sigma_z^2} \cdot \bar{z}$$

$$y = -1333{,}712\,z + 179{,}25$$

Sustituyendo, resulta:

$$y = -\frac{1333{,}712}{x} + 179{,}2501$$

Debemos calcular ahora la razón de correlación para ver cómo de bueno es el ajuste:

$\eta_{Y/X}^2 = \frac{\sigma_{ey}^2}{\sigma_y^2}$ siendo σ_{ey}^2 la varianza explicada por la regresión.

$$\sigma_{ey}^2 = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}\left(f(x_i)-\bar{y}\right)^2$$

$$f(9) = 31{,}0598$$
$$f(10) = 45{,}8789$$
$$f(12) = 68{,}1074$$
$$f(15) = 90{,}3359$$
$$f(22) = 118{,}6268$$
$$f(32) = 137{,}5716$$

$$\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}\left(f(x_i)-\bar{y}\right)^2 = 2561{,}05833+1280{,}7666+183{,}8546+75{,}153+1366{,}049+3125{,}337 = 8592{,}2413$$

$$\sigma_{ey}^2 = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}\left(f(x_i)-\bar{y}\right)^2 = 1432{,}0402$$

$$\eta_{Y/X}^2 = \frac{\sigma_{ey}^2}{\sigma_y^2} = 0{,}9895$$

c) Curva potencial

Tenemos una función de la forma $y = bx^a$. Haremos la transformación:

$$\log y = \log bx^a \Leftrightarrow \log y = \log b + a\log x$$

$$y' = az + b'$$

$$y' = \log y$$

$$b' = \log b$$

$$z' = \log x$$

Y’	3,4012	3,912	4,2485	4,382	4,7875	4,9416
Z	2,1972	2,3026	2,4849	2,7081	3,091	3,4657

$$\bar{z} = \frac{\sum_{i=1}^{6} z_i \cdot n_{i.}}{n} = 2{,}7083$$

$$\bar{y}' = \frac{\sum_{i=1}^{6} y'_j \cdot n_{.j}}{n} = 4{,}2733$$

$$\sigma_z^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.} \cdot z_i^2 - \bar{z}^2 = 0{,}199$$

$$\sigma_{z y'} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij} \cdot z_i \cdot y'_j - \bar{z}\,\bar{y}' = 0{,}2166$$

$$y = \bar{y} + \frac{\sigma_{z y'}}{\sigma_z^2} \cdot z - \frac{\sigma_{z y'}}{\sigma_z^2} \cdot \bar{z}$$

$$y = 1{,}0884\, z + 1{,}331$$

$$b' = \log b \Leftrightarrow b = e^{b'} = 3{,}785$$

$$y = 3{,}785\, x^{1{,}0884}$$

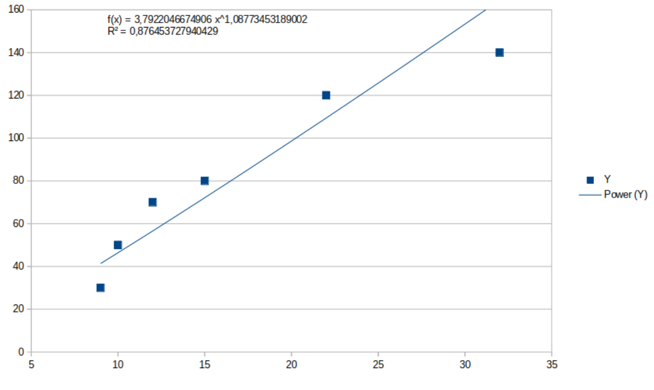
Debemos calcular ahora la razón de correlación para ver cómo de bueno es el ajuste:

$$\eta_{Y/X}^2 = 1 - \frac{\sigma_{r\,y}^2}{\sigma_y^2}$$

$$\begin{aligned} f(9) &= 41{,}3679 \\ f(10) &= 46{,}3944 \\ f(12) &= 56{,}5779 \\ f(15) &= 72{,}1313 \\ f(22) &= 109{,}4356 \\ f(32) &= 164{,}3399 \end{aligned}$$

$$\sigma_{r\,y}^2 = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}(y_j - f(x_i))^2 = \frac{1}{6}(129{,}2292+13{,}004+130{,}1528+61{,}9164+111{,}6065+602{,}2067) = 183{,}0187$$

$$\eta_{Y/X}^2 = 1 - \frac{\sigma_{r\,y}^2}{\sigma_y^2} = 0{,}8735$$



d) Curva exponencial

$$y = ba^x \Leftrightarrow \log y = \log b + x \log a$$

$$y' = b' + xa'$$

$$y' = \log y$$

$$b' = \log b$$

$$a' = \log a$$

Y’	3,4012	3,912	4,2485	4,382	4,7875	4,9416
X	9	10	12	15	22	32

Para proseguir, tomaré los datos referentes a la media de x y a la varianza del apartado a:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot n_{i.}}{n} = 16,667 \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.} \cdot x_i^2 - \bar{x}^2 = 65,221$$

$$\sigma_{xy'} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij} \cdot x_i \cdot y'_j - \bar{x}\,\bar{y}' = 3,6697$$

$$y' = 0,0563\,x + 3,341$$

$$a' = \log a \Leftrightarrow a = e^{a'} = 1,0579$$

$$b' = \log b \Leftrightarrow b = e^{b'} = 28,24736$$

$$y = 28,24736 \cdot 1,0579^x$$

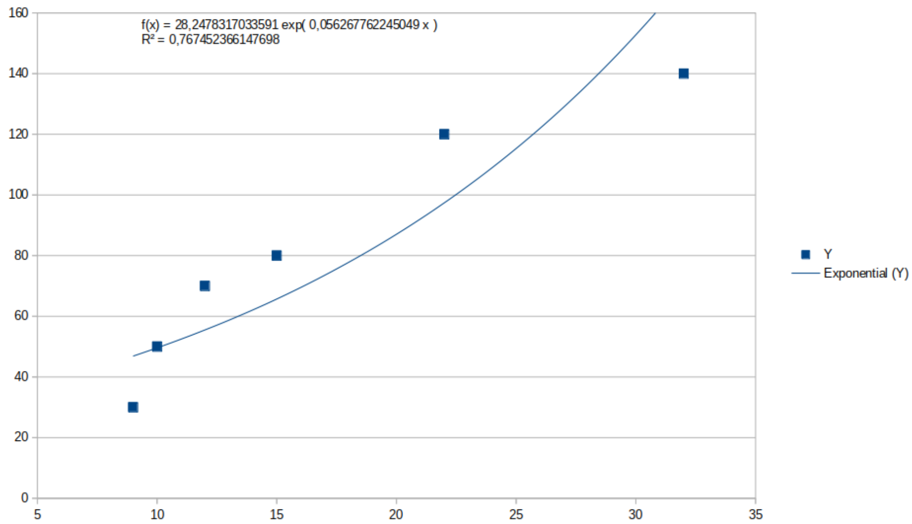
Calculamos ahora la razón de correlación para ver cómo de bueno es el ajuste:

$$\eta_{Y/X}^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$$

$$f(9) = 46,8791$$
$$f(10) = 49,5934$$
$$f(12) = 55,5026$$
$$f(15) = 65,7124$$
$$f(22) = 97,445$$
$$f(32) = 171,0825$$

$$\sigma_{ry}^2 = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}\left(y_j - f(x_i)\right)^2 = \frac{1}{6}\left(284{,}904+0{,}1653+210{,}1746+204{,}1355+508{,}728+966{,}1218\right) = 362{,}3715$$

$$\eta_{Y/X}^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2} = 0{,}7496$$



El ajuste más adecuado es el de la hipérbola equilátera, pues su razón de correlación es más próxima a 1 que las demás y por lo tanto es la que mejor se ajusta a la nube de puntos. No hubiera sido necesario calcular todas las razones de correlación, puesto que son todos los ajustes sobre la misma distribución, es decir, las varianzas no van a cambiar, así que lo único que tendríamos que tener en cuenta sería la que tuviera mayor varianza explicada o menor varianza residual.