Variables estadísticas bidimensionales: problemas resueltos

BENITO J. GONZÁLEZ RODRÍGUEZ (bjglez@ull.es)
DOMINGO HERNÁNDEZ ABREU (dhabreu@ull.es)
MATEO M. JIMÉNEZ PAIZ (mjimenez@ull.es)
M. ISABEL MARRERO RODRÍGUEZ (imarrero@ull.es)
ALEJANDRO SANABRIA GARCÍA (asgarcia@ull.es)

Departamento de Análisis Matemático
Universidad de La Laguna

Índice

6. Problemas resueltos

Ejercicio 6.1. *Se han tomado cinco muestras de glucógeno, de una cantidad fija cada una. Se les ha aplicado una cantidad X de glucogenasa (en milimoles por litro) anotando en cada caso la velocidad de reacción Y medida en micromoles por minuto, obteniéndose así la siguiente tabla:*

X	1	2	3	0.2	0.5
Y	18	35	60	8	10

Se pide:

a) *¿Se deduce de estos datos que la velocidad de reacción aumenta con la concentración de glucogenasa?*

b) *Si a una de las muestras le hubiésemos aplicado una concentración de glucogenasa de 5 milimoles por litro, ¿cuál hubiera sido la velocidad de reacción? ¿Con qué grado de predicción?*

c) *Dibujar la nube de puntos y las rectas de regresión.*

RESOLUCIÓN. Las variables $X :=$'concentración de glucogenasa (mmol/L)' e $Y :=$'velocidad de reacción (μmol/min)' son ambas cuantitativas. Aunque en los cálculos que siguen es suficiente la tabla de frecuencias para las variables marginales X e Y dada en el enunciado del ejercicio (véase la Observación 2.5 del desarrollo teórico), construiremos la tabla de frecuencias de la variable bidimensional (X,Y) (Cuadro 6.1).

X \ Y	8	10	18	35	60	n_{x_i}	f_{x_i}
0.2	1					1	0.2
0.5		1				1	0.2
1			1			1	0.2
2				1		1	0.2
3					1	1	0.2
n_{y_j}	1	1	1	1	1	5	
f_{y_j}	0.2	0.2	0.2	0.2	0.2		1

Cuadro 6.1. Tabla de frecuencias para la variable bidimensional del Ejercicio 6.1.

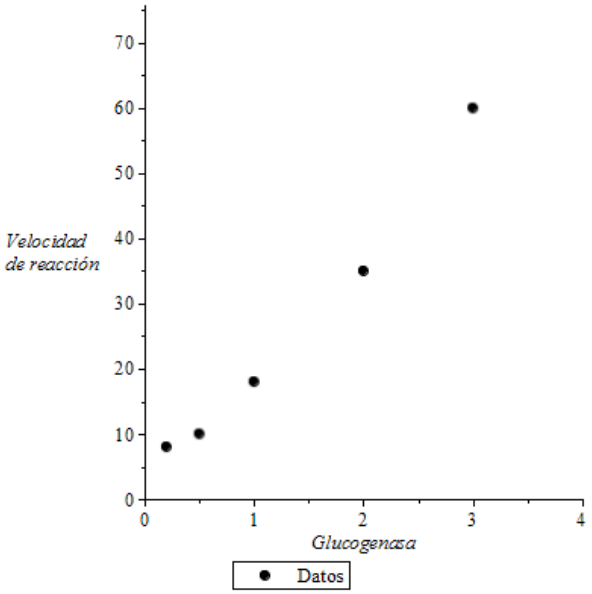

Figura 6.1. Diagrama de dispersión para la variable bidimensional del Ejercicio 6.1.

En primer lugar, trazamos un diagrama de dispersión para conjeturar una posible relación entre las variables (Figura 6.1).

El diagrama de la Figura 6.1 parece indicarnos que existe una relación lineal directa entre las variables, esto es, la velocidad de reacción Y aumenta a medida que aumenta la cantidad de glucogenasa y esta dependencia responde a un modelo lineal. Veamos si esta intuición que sugiere el diagrama puede ser confirmada.

a) Al objeto de analizar cuál es el grado de correlación entre las variables X e Y estudiamos el coeficiente de correlación lineal (o coeficiente de correlación de Pearson)

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y},$$

donde σ_{xy}, σ_x y σ_y denotan la covarianza y las desviaciones típicas marginales de las variables X e Y, respectivamente.

Para hallar los estadísticos anteriores, necesarios en la determinación del coeficiente de correlación lineal, calculamos en primer lugar la media de X:

$$\overline{x} = \frac{0.2+0.5+1+2+3}{5} = 1.340.$$

En segundo lugar, la varianza de X viene dada por:

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 n_{x_i} - \overline{x}^2 = \frac{(0.2^2 + 0.5^2 + 1^2 + 2^2 + 3^2)}{5} - 1.340^2 = \frac{14.29}{5} - 1.7956 = 1.0624,$$

de manera que su desviación típica es

$$\sigma_x = \sqrt{1.0624} \simeq 1.031.$$

Similarmente, la media de Y viene dada por:

$$\overline{y} = \frac{8 + 10 + 18 + 35 + 60}{5} = 26.200,$$

mientras que su varianza es

$$\sigma_y^2 = \frac{8^2 + 10^2 + 18^2 + 35^2 + 60^2}{5} - 26.200^2 = \frac{5313}{5} - 686.440 = 376.160$$

y su desviación típica,

$$\sigma_y = \sqrt{376.160} \simeq 19.395.$$

Finalmente, calculamos la covarianza entre las variables X e Y:

$$\begin{aligned}\sigma_{xy} &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{r} x_i y_j n_{ij} - \overline{x}\,\overline{y} \\ &= \frac{(0.2\cdot 8) + (0.5\cdot 10) + (1\cdot 18) + (2\cdot 35) + (3\cdot 60)}{5} - (1.340\cdot 26.200) \\ &= \frac{274.6}{5} - 35.108 = 19.812.\end{aligned}$$

En consecuencia, el coeficiente de correlación lineal toma el valor

$$\rho = \frac{19.812}{1.031\cdot 19.395} \simeq 0.99.$$

Al estar el coeficiente de Pearson muy cercano a 1 podemos garantizar que existe una muy buena correlación lineal entre las variables. Esta relación es, además, directa, como indica el hecho de que el coeficiente de correlación de Pearson tiene signo positivo, al igual que la covarianza. Luego, cabe afirmar que la velocidad de reacción aumenta con la concentración de glucogenasa y que este aumento es de tipo lineal.

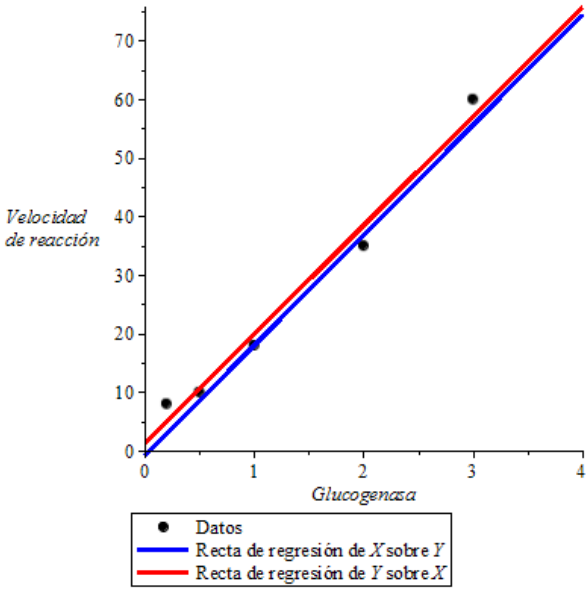

Figura 6.2. Diagrama de dispersión y rectas de regresión para la variable bidimensional del Ejercicio 6.1.

b) Calculamos primeramente la recta de regresión de Y sobre X:

$$r_{y_x}:\ y-\overline{y}=\beta_{y_x}(x-\overline{x}),$$

donde

$$\beta_{y_x}=\frac{\sigma_{xy}}{\sigma_x^2}=\frac{19.812}{1.0624}\simeq 18.648.$$

Por tanto, la recta de regresión de Y sobre X viene dada por:

$$r_{y_x}:\ y=26.200+18.648\,(x-1.340)=18.648\,x+1.212.$$

Si a una de las muestras le hubiésemos aplicado una concentración de glucogenasa de 5 mmol/L, la velocidad de reacción hubiese sido de

$$y=(18.648\cdot 5)+1.212\simeq 94.45\ \mu\text{mol/min},$$

siendo esta predicción muy buena pues, como hemos mencionado anteriormente, el coeficiente de correlación de Pearson es prácticamente 1 ($\rho=0.99$).

c) La recta de regresión de X sobre Y tiene por ecuación

$$r_{x_y}:\ x-\bar{x}=\beta_{x_y}(y-\bar{y}),$$

donde

$$\beta_{x_y}=\frac{\sigma_{xy}}{\sigma_y^2}=\frac{19.812}{376.160}\simeq 0.053.$$

Es decir,

$$r_{x_y}:\ x=1.340+0.053\,(y-26.200)=0.053\,y-0.049.$$

Si representamos las rectas de regresión anteriores en el diagrama de dispersión de la Figura 6.1, obtenemos la Figura 6.2. □

Ejercicio 6.2. *Se ha medido, en miligramos por litro, el contenido de oxígeno Y del lago Worther, en Austria, a una profundidad de X metros, obteniéndose los siguientes datos:*

X	15	20	30	40	50	60	70
Y	6.5	5.6	5.4	6.0	4.6	1.4	0.1

Se pide:

a) *Ajustar una recta a los datos obtenidos.*

b) *Estudiar la correlación entre ambas variables.*

c) *Para una profundidad comprendida entre* 75 *y* 80 *metros, ¿qué contenido en oxígeno se podría predecir?*

d) *Dibujar la nube de puntos y las rectas de regresión.*

RESOLUCIÓN. Las variables X :='profundidad (m)' e Y :='cantidad de oxígeno (mg/L)' son ambas cuantitativas. De manera similar al problema anterior, la tabla de frecuencias para la variable bidimensional (X,Y) queda recogida en el Cuadro 6.2.

El diagrama de dispersión para la variable bidimensional (X,Y) de la Figura 6.3 nos indica de manera intuitiva que si existiese alguna relación entre las variables X e Y, ésta debiera ser inversa, esto es, el aumento de una de las variables implicaría la disminución de la otra y viceversa.

X \ Y	0.1	1.4	4.6	5.4	5.6	6.0	6.5	n_{x_i}	f_{x_i}
15							1	1	0.142
20					1			1	0.142
30				1				1	0.142
40						1		1	0.142
50			1					1	0.142
60		1						1	0.142
70	1							1	0.142
n_{y_j}	1	1	1	1	1	1	1	7	
f_{y_j}	0.142	0.142	0.142	0.142	0.142	0.142	0.142		1

Cuadro 6.2. Tabla de frecuencias para la variable bidimensional del Ejercicio 6.2.

Veamos si podemos garantizar lo afirmado en el párrafo anterior.

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

$$r_{y_x}:\ y-\overline{y}=\beta_{y_x}(x-\overline{x}),$$

donde

$$\beta_{y_x}=\frac{\sigma_{xy}}{\sigma_x^2}.$$

Calculemos en primer lugar la media y la varianza de la variable marginal X, y a continuación la media de Y y la covarianza entre ambas variables en estudio.

La media de X es:

$$\overline{x}=\frac{15+20+30+40+50+60+70}{7}\simeq 40.714.$$

La varianza de X viene dada por:

$$\begin{aligned}\sigma_x^2 &= \frac{15^2+20^2+30^2+40^2+50^2+60^2+70^2}{7}-40.714^2\\ &\simeq \frac{14125}{7}-1657.630\simeq 360.227.\end{aligned}$$

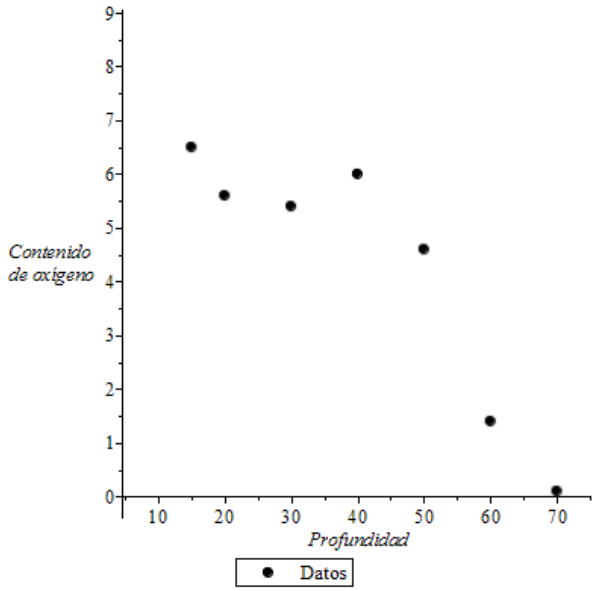

Figura 6.3. Diagrama de dispersión para la variable bidimensional del Ejercicio 6.2.

Por tanto, su desviación típica es:

$$\sigma_x = \sqrt{360.227} \simeq 18.980.$$

La media de Y viene dada por:

$$\bar{y} = \frac{0.1+1.4+4.6+5.4+5.6+6.0+6.5}{7} \simeq 4.228.$$

Hallamos ahora la covarianza entre las variables X e Y:

$$\sigma_{xy} = \frac{(15\cdot 6.5)+(20\cdot 5.6)+(30\cdot 5.4)+(40\cdot 6.0)+(50\cdot 4.6)+(60\cdot 1.4)+(70\cdot 0.1)}{7} - (40.714\cdot 4.228)$$
$$\simeq \frac{932.5}{7} - 172.139 = -38.925.$$

Estamos ya en disposición de calcular

$$\beta_{y_x} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-38.925}{360.227} \simeq -0.108.$$

Así pues, la recta de regresión de Y sobre X será

$$r_{y_x}:\ y = 4.228 - 0.108\,(x - 40.714) = -0.108\,x + 8.625.$$

b) Nótese que la covarianza σ_{xy} es negativa: esto nos indica que existe una relación inversa entre las variables. Para decidir si tal relación es de tipo lineal, estudiamos el coeficiente de Pearson

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}.$$

La varianza de Y es

$$\begin{aligned}\sigma_y^2 &= \frac{0.1^2 + 1.4^2 + 4.6^2 + 5.4^2 + 5.6^2 + 6.0^2 + 6.5^2}{7} - 4.228^2 \\ &\simeq \frac{161.9}{7} - 17.876 \simeq 5.252\end{aligned}$$

y su desviación típica,

$$\sigma_y = \sqrt{5.252} \simeq 2.292.$$

Por tanto, el coeficiente de correlación lineal toma el valor

$$\rho = \frac{-38.925}{18.980 \cdot 2.292} \simeq -0.895,$$

y podemos concluir que las variables presentan una muy buena correlación lineal inversa (o correlación lineal negativa).

c) La recta de regresión de Y sobre X, de ecuación

$$r_{y_x}: \; y = -0.108\,x + 8.625,$$

es un buen modelo de predicción, tal y como hemos visto en el apartado anterior.

Para una profundidad de $x = 75$ m el modelo predice una cantidad de oxígeno de

$$y = -0.108 \cdot 75 + 8.625 = 0.525 \simeq 0.5 \text{ mg/L},$$

mientras que para una profundidad de $x = 80$ m, predecimos

$$y = -0.108 \cdot 80 + 8.625 = -0.015 \simeq 0 \text{ mg/L}.$$

Podemos concluir entonces que para una profundidad comprendida entre 75 y 80 metros los niveles de oxígeno varían entre 0 y 0.5 mg/L.

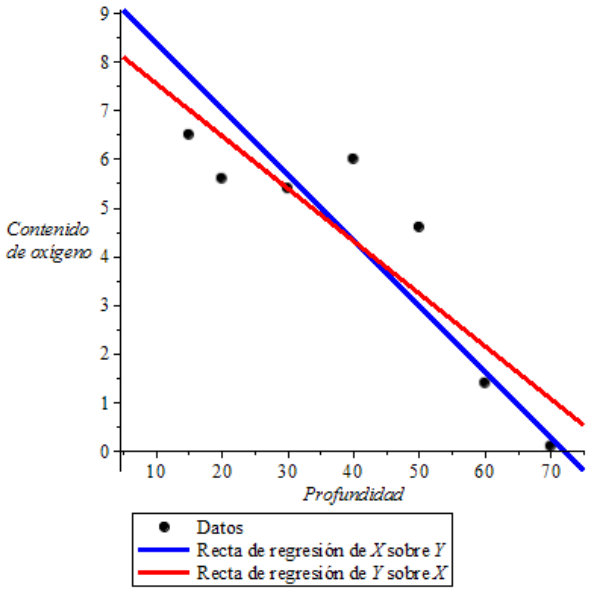

Figura 6.4. Diagrama de dispersión y rectas de regresión para la variable bidimensional del Ejercicio 6.2.

d) La recta de regresión de X sobre Y tiene por ecuación

$$r_{x_y}:\ x-\bar{x}=\beta_{x_y}(y-\bar{y}),$$

donde

$$\beta_{x_y}=\frac{\sigma_{xy}}{\sigma_y^2}=\frac{-38.925}{5.252}\simeq -7.411.$$

Luego,

$$r_{x_y}:\ x=40.714-7.411\,(y-4.228)=-7.411\,y+72.048.$$

El diagrama de dispersión junto con las rectas de regresión aparece representado en el gráfico de la Figura 6.4. □

Ejercicio 6.3. *En un hospital se ha aplicado un medicamento A a* 100 *enfermos, y en otro hospital se ha aplicado un segundo medicamento B a otros* 100 *enfermos. El número diario de curados durante los* 10 *primeros días es el siguiente:*

medicamento A	8	7	6	5	4	3	3	2	1	1
medicamento B	4	4	6	7	2	5	1	2	2	2

Se pide:

a) *Rectas de regresión de Y sobre X y de X sobre Y.*

b) *Dibujar la nube de puntos y las rectas de regresión.*

c) *Hallar el coeficiente de correlación e interpretarlo.*

RESOLUCIÓN. Estamos interesados en determinar la relación que pueda existir entre el número de enfermos curados en un mismo día, en dos hospitales diferentes y mediante dos medicamentos distintos, para, por ejemplo, determinar la eficacia de ambos medicamentos. El estudio se hace durante 10 días (por lo que la población en este caso consta de $N = 10$ observaciones).

Para ello denotamos por $X :=$'número de pacientes curados por el medicamento A en un determinado día' y por $Y :=$'número de pacientes curados por el medicamento B en el mismo día'. Obsérvese que ambas variables marginales son cuantitativas discretas. La tabla de frecuencias para la variable bidimensional (X,Y) viene dada en el Cuadro 6.3.

X \ Y	1	2	4	5	6	7	n_{x_i}	f_{x_i}
1		2					2	0.2
2		1					1	0.1
3	1			1			2	0.2
4		1					1	0.1
5						1	1	0.1
6					1		1	0.1
7			1				1	0.1
8			1				1	0.1
n_{y_j}	1	4	2	1	1	1	10	
f_{y_j}	0.1	0.4	0.2	0.1	0.1	0.1		1

Cuadro 6.3. Tabla de frecuencias para la variable bidimensional del Ejercicio 6.3.

Al igual que hemos hecho en los ejercicios anteriores, trazamos un diagrama de dispersión para la variable bidimensional (Figura 6.5) que nos permita conjeturar una posible relación entre las variables marginales. Nótese que, aunque podemos intuir *algún* tipo de relación directa entre las variables (esto es, el aumento en la eficiencia de uno de los medicamentos en un determinado día implica también la del del otro), no está del todo

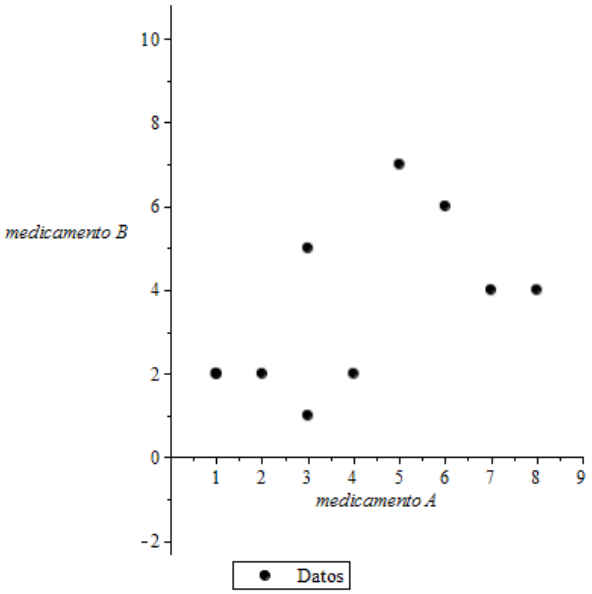

Figura 6.5. Diagrama de dispersión para la variable bidimensional del Ejercicio 6.3.

claro que esta relación sea de tipo lineal.

Pasamos a discutir ahora los modelos de regresión.

a) Para determinar las rectas de regresión, calculamos en primer lugar los estadísticos necesarios.

- Estadísticos de la variable marginal X.

  - Media de X:

$$\bar{x} = \frac{(2\cdot 1)+(1\cdot 2)+(2\cdot 3)+(1\cdot 4)+(1\cdot 5)+(1\cdot 6)+(1\cdot 7)+(1\cdot 8)}{10} = 4.000.$$

  - Varianza de X:

$$\sigma_x^2 = \frac{(2\cdot 1^2)+(1\cdot 2^2)+(2\cdot 3^2)+(1\cdot 4^2)+(1\cdot 5^2)+(1\cdot 6^2)+(1\cdot 7^2)+(1\cdot 8^2)}{10} - 4.000^2$$
$$= \frac{214}{10} - 16 = 5.400.$$

  - Desviación típica de X:

$$\sigma_x = \sqrt{5.400} \simeq 2.324.$$

- Estadísticos de la variable marginal Y.

- Media de Y:

$$\bar{y} = \frac{(1\cdot 1)+(4\cdot 2)+(2\cdot 4)+(1\cdot 5)+(1\cdot 6)+(1\cdot 7)}{10} = 3.500.$$

- Varianza de Y:

$$\begin{aligned}\sigma_y^2 =& \frac{(1\cdot 1^2)+(4\cdot 2^2)+(2\cdot 4^2)+(1\cdot 5^2)+(1\cdot 6^2)+(1\cdot 7^2)}{10} - 3.500^2\\ =& \frac{159}{10} - 12.25 = 3.650.\end{aligned}$$

- Desviación típica de Y:

$$\sigma_y = \sqrt{3.650} \simeq 1.910.$$

- Estadísticos de la variable bidimensional (X,Y).

- Covarianza entre las variables X e Y:

$$\begin{aligned}\sigma_{xy} =& \frac{(1\cdot 2\cdot 2)+(2\cdot 2)+3\cdot(1+5)+(4\cdot 2)+(5\cdot 7)+(6\cdot 6)+(7\cdot 4)+(8\cdot 4)}{10} - (4\cdot 3.500)\\ =& \frac{165}{10} - 14 = 2.500.\end{aligned}$$

Atendiendo a los datos calculados:

- Recta de regresión de Y sobre X:

$$r_{y_x}:\ y-\bar{y} = \beta_{y_x}(x-\bar{x}),$$

donde

$$\beta_{y_x} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{2.500}{5.400} \simeq 0.463.$$

Por tanto,

$$r_{y_x}:\ y = 3.500 + 0.463\,(x-4.000) = 0.463x + 1.648.$$

- Recta de regresión de X sobre Y:

$$r_{x_y}:\ x-\bar{x} = \beta_{x_y}(y-\bar{y}),$$

donde

$$\beta_{x_y} = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{2.500}{3.650} \simeq 0.685.$$

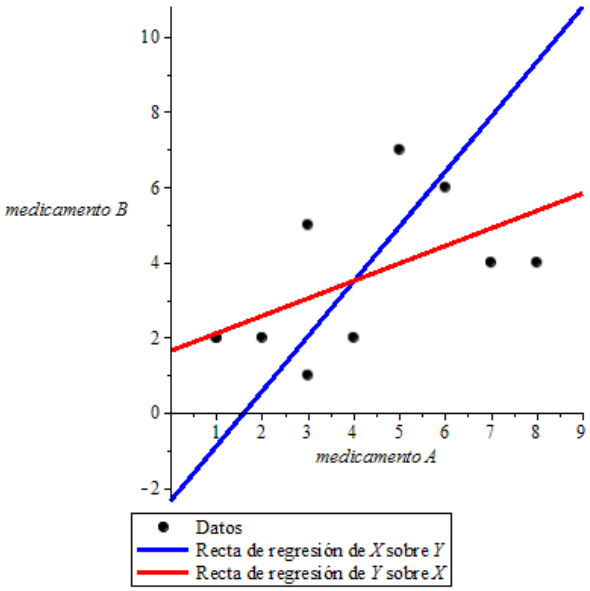

Figura 6.6. Diagrama de dispersión y rectas de regresión para la variable bidimensional del Ejercicio 6.3.

Por tanto,

$$r_{x_y}:\ x = 4.000 + 0.685\,(y - 3.500) = 0.685\,y + 1.602.$$

b) Las rectas de regresión están representadas en el gráfico de la Figura 6.6, junto con el diagrama de dispersión.

c) Finalmente, el coeficiente de correlación lineal ρ (o coeficiente de correlación de Pearson) toma el valor

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{2.500}{2.324 \cdot 1.910} \simeq 0.56.$$

Como el coeficiente de correlación de Pearson está comprendido entre 0.5 y 0.8, podemos afirmar que la relación entre las variables descritas en los modelos lineales anteriores es buena. □