Estadística Descriptiva e Introducción a la Probabilidad.
Relaciones de prácticas

ELÍAS MONGE SÁNCHEZ
DANIEL MORÁN SÁNCHEZ
JESÚS MUÑOZ VELASCO

Marzo 2023

Índice

1. Relación 2

1.1. Ejercicio 1:

Se han lanzado dos dados varias veces, obteniendo los resultados que se presentan en la siguiente tabla, donde X designa el resultado del primer dado e Y el resultado del segundo:

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

a) Construir la tabla de frecuencias

X/Y	1	2	3	4	5	6	$f_{i.}$
1	0	0.04167	0	0	0	0.125	0.16667
2	0.4167	0	0.4167	0	0.04167	0	0.125
3	0.4167	0	0	0.0833	0.04167	0.04167	0.20833
4	0.08333	0.125	0	0	0	0.04167	0.25
5	0.08333	0.4167	0.4167	0	0	0	0.1667
6	0	0.04167	0	0	0.04167	0	0.08333
$f_{.j}$	0.25	0.25	0.08333	0.08333	0.125	0.20833	1

b) Calcular las puntuaciones medias obtenidas con cada dado y ver cuales son más homogéneas.

Empezaremos calculando las medias:
Media de los resultados del dado x:

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6}(x_i * n_{i.}) = 3,375$$

Media de los resultados del dado y:

$$\bar{y} = \frac{1}{6}\sum_{j=1}^{6}(y_j * n_{.j}) = 3,208333333$$

desv(x):

$$\sigma^2 = \frac{1}{6}\sum_{i=1}^{6}(x_{i.} - \bar{x})^2 = 2,317708333$$

desv(y):

$$\sigma^2 = \frac{1}{6}\sum_{i=1}^{6}(y_{i.} - \bar{y})^2 = 3,692708333$$

coeficiente de variación:

$$C.V.(x) = \frac{\sigma_x}{\bar{x}} = 0,4510821211$$

$$C.V.(y) = \frac{\sigma_y}{\bar{y}} = 0,5989533796$$

El dado x tiene menor coeficiente de variación, por tanto sus resultados son mas homogéneos

c) ¿Qué resultado del segundo dado es más frecuente cuando en el primero se obtiene un 3?

El resultado mas frecuente de la distribución marginal cuando X=3 es Y=4 con una frecuencia de 0,0833333.

d) Calcular la puntuación máxima del 50 % de las puntuaciones más bajas obtenidas con el primer dado si con el segundo se ha obtenido un 2 o un 5.

X/Y	2	5	2+5	F_i
1	1	0	1	1
2	0	1	1	2
3	0	1	1	3
4	3	0	3	**6**
5	1	0	1	7
6	1	1	2	9
$n_{.j}$	6	3	9	

La mitad de 9 (4.5) se obtiene para $x_i = 4$ por lo que este será el valor de la mediana.

1.2. Ejercicio 2:

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.

Antes de nada vamos a añadir las frecuencias marginales a la tabla.

X/Y	160	162	164	166	168	170	$n_{i.}$
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
$n_{.j}$	6	8	13	15	19	9	70

El peso viene dado por la variable X con lo cual podemos calcular el peso medio tomando la media marginal de x:

$\bar{x} = \sum_{i=1}^{5} f_{i.} x_i = 54.1714\ kg$

La altura media la da la media marginal de la variable Y:

$\bar{y} = \sum_{j=1}^{6} f_{.j} y_j = 165.7142\ cm$

Para decir cuál es más representativa se pueden tener en cuenta diversos aspectos. Por ejemplo, será más representativa aquella media de la que se alejen menos los datos de forma relativa. Esta información la da el coeficiente de variación de Pearson:

$$\sigma_x = \sqrt{\sum_{i=1}^{5} f_{i.}(x_i - \bar{x})^2} = 3,6074\ kg \quad C.V.(X) = \frac{\sigma_x}{\bar{x}} = 0.0666$$

$$\sigma_y = \sqrt{\sum_{j=1}^{6} f_{.j}(y_j - \bar{y})^2} = 2,9740 cm \quad C.V.(Y) = \frac{\sigma_y}{\bar{y}} = 0.0179$$

Con esto podemos afirmar que la altura media es más representativa que el peso medio.

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

Vamos a hacer una tabla con la submatriz que recoge a 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
$n_{.j}$	11	7	2	20

Como vemos hay 20 individuos que cumplen los requisitos con lo que el porcentaje será:

$\frac{20}{70} \times 100 = 28.57\,\%$ de la población

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

Hay un total de $15 + 19 + 9 = 43$ individuos que miden más de 165 cm.

X/Y	166	168	170	$n_{i.}$
54	8	5	1	14
57	2	8	3	13
60	2	4	4	10
$n_{.j}$	12	17	8	37

Y como vemos en la tabla superior hay un total de 37 individuos que miden más de 165 cm y además pesan más de 52 kg. Por lo tanto el porcentaje es:

$\frac{37}{43} \times 100 = 86.05\,\%$

Entonces un $86.05\,\%$ de los individuos que miden más de 165cm pesan además más de 52 kg.

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

Primero reduzcamos la población respecto del carácter X:

X/Y	160	162	164	166	168	170	$n_{i.}$
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
$n_{.j}$	3	6	11	12	15	5	52

De la tabla de arriba, tan solo nos tenemos que fijar en los $n_{.j}$ y ver cuál es el máximo. Cómo vemos el máximo se alcanza para $j = 5$ con $y_5 = 168\ cm$, siendo esta la altura más frecuente entre los individuos de entre 51 y 57 kg.

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

Calculamos $\bar{x_j}$ para $j = 3$ e $j = 5$:

$$\bar{x_3} = \frac{1}{n_{.3}} \sum_{i=1}^{5} n_{i3} x_i = \frac{2 \times 48 + 4 \times 51 + 6 \times 54 + 1 \times 57 + 0 \times 60}{13} = 52.38\ kg$$

$$\bar{x_5} = \frac{1}{n_{.5}} \sum_{i=1}^{5} n_{i5} x_i = \frac{0 \times 48 + 2 \times 51 + 5 \times 54 + 8 \times 57 + 4 \times 60}{19} = 56.21\ kg$$

El peso medio condicionado más representativo será aquel que se aproxime más al peso medio, es decir, $\bar{x_3}$.

1.3. 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	$n_{i.}$
1	7	0	0	0	7
2	10	2	0	0	12
3	11	5	1	0	17
4	10	6	6	0	22
5	8	6	4	2	20
6	1	2	3	1	7
7	1	0	0	1	2
8	0	0	1	1	2
$n_{.j}$	48	21	15	5	89

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

Para ello calcularé los puntos $(x_i, \bar{y}_i)$:

$$\bar{y}_1 = \frac{1 \times 7 + 2 \times 0 + 3 \times 0 + 4 \times 0}{7} = 1 \Rightarrow (1; 1)$$

$$\bar{y}_2 = \frac{1 \times 10 + 2 \times 2 + 3 \times 0 + 4 \times 0}{12} = 1.1667 \Rightarrow (2; 1.1667)$$

$$\bar{y}_3 = \frac{1 \times 11 + 2 \times 5 + 3 \times 1 + 4 \times 0}{17} = 1.412 \Rightarrow (3; 1.412)$$

$$\bar{y}_4 = \frac{1 \times 10 + 2 \times 6 + 3 \times 6 + 4 \times 0}{22} = 1.818 \Rightarrow (4; 1.818)$$

$$\bar{y}_5 = \frac{1 \times 8 + 2 \times 6 + 3 \times 4 + 4 \times 2}{20} = 2 \Rightarrow (5; 2)$$

$$\bar{y}_6 = \frac{1 \times 1 + 2 \times 2 + 3 \times 3 + 4 \times 1}{7} = 2.571 \Rightarrow (6; 2.571)$$

$$\bar{y}_7 = \frac{1 \times 1 + 2 \times 0 + 3 \times 0 + 4 \times 1}{2} = 2.5 \Rightarrow (7; 2.5)$$

$$\bar{y}_8 = \frac{1 \times 0 + 2 \times 0 + 3 \times 1 + 1 \times 0}{2} = 3.5 \Rightarrow (8; 3.5)$$

La recta de regresión será aquella que pase por los puntos calculados.

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

Si existiese dependencia funcional lineal, la recta de regresión X/Y y la recta Y/X deberían coincidir con la recta de dependencia.

Esto además implicaría que $r = 1$ por lo que pasaré a calcular el coeficiente de correlación lineal. En este caso:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0.791}{1.586 \times 0.93} = 0.536 \neq 1$$

Por lo que no sería correcta tal suposición.

1.4. Ejercicio 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, los resultados son:

X/Y	(0.5,1.5]	(1.5,2.5]	(2.5,5.5]
(1,15]	4	2	0
(15,25]	1	4	2
(25,30]	0	3	5

Explicar el comportamiento de la tensión de vapor en términos de la temperatura mediante una función lineal. ¿Es adecuado asumir este tipo de relación?
Con las marcas de clase la distribución en la tabla de frecuencias quedaría:

X/Y	1	2	4	$n_{i.}$	$c_i \times n_{i.}$	$c_i \times n_{i.}^2$
8	4	2	0	6	48	384
20	1	4	2	7.	140	2800
27.5	0	3	5	8	220	6050
$n_{.j}$	5	9	7	21	408	9234
$d_j \times n_{.j}$	5	18	28	51	-	-
$d_j \times n_{.j}^2$	5	36	112	153	-	-

Media de x:

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4}(x_i * n_{i.}) =$$

Media de y:

$$\bar{y} = \frac{1}{4}\sum_{j=1}^{4}(y_i * n_{.j}) =$$

desv(x):

$$\sigma^2 = \frac{1}{4}\sum_{i=1}^{4}(x_i - \bar{x})^2 =$$

desv(y):

$$\sigma^2 = \frac{1}{4}\sum_{i=1}^{4}(y_i - \bar{y})^2 =$$

1.5. 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.

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

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

Completamos ambas tablas:

X/Y	1	2	3	4	5	$n_{i.}$
10	2	4	6	10	8	30
20	1	2	3	5	4	15
30	3	6	9	15	12	45
40	4	8	12	20	16	60
$n_{.j}$	10	20	30	50	40	150

Podemos ver que en la primera tabla se da independencia entre las variables por el **Teorema de caracterización de la independencia**, que dice lo siguiente:

$$X \text{ e } Y \text{ independientes} \iff n_{ij} = \frac{n_{i.}n_{.j}}{n} \ \forall i = 1, ..., k \ \forall j = 1, ..., p$$

Podemos comprobar que dicha caracterización se verifica $\forall i, j = 1, 2, 3, 4$ por lo que podemos afirmar que en esta primera distribución las variables son independientes.

X/Y	1	2	3	$n_{i.}$
-1	0	1	0	1
0	1	0	1	2
1	0	1	0	1
$n_{.j}$	1	2	1	4

Para esta segunda tabla, aplicando el mismo teorema, comprobamos que no se da dicha caracterización:

$$\frac{n_{1.}n_{.j}}{n} = \frac{1 \times 1}{4} = \frac{1}{4} \neq 0 = n_{11}$$

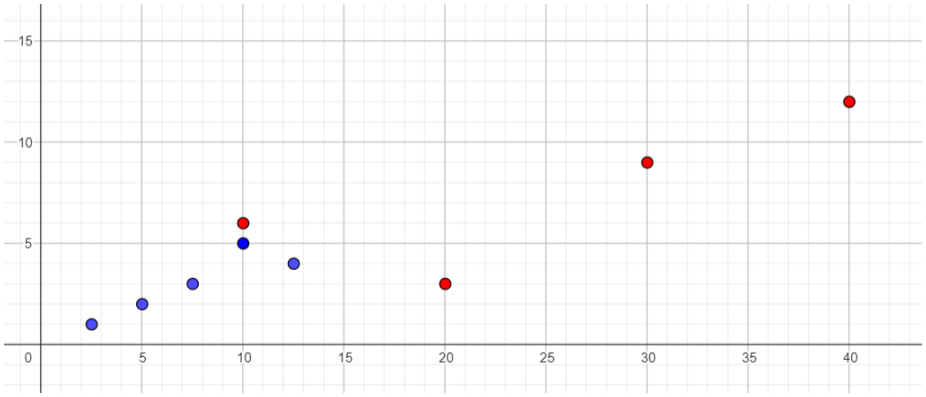
A continuación, procedemos a trazar las curvas de regresión de tipo I de ambas distribuciones. Se define la curva de regresión de tipo I de Y sobre X como aquella que pasa por todos los puntos $(x_i, \bar{y_i})$ y la curva de regresión de tipo I de X sobre Y como aquella que pasa por todos los puntos $(\bar{x_j}, y_j)$. Para el caso de variables discretas, las curvas de regresión no serán más que puntos aislados.

Curvas de regresión de tipo I:

Tabla 1:

x_i	$\bar{y_i}$
10	6
20	3
30	9
40	12

$\bar{x_j}$	y_j
2.5	1
5	2
7.5	3
12.5	4
10	5

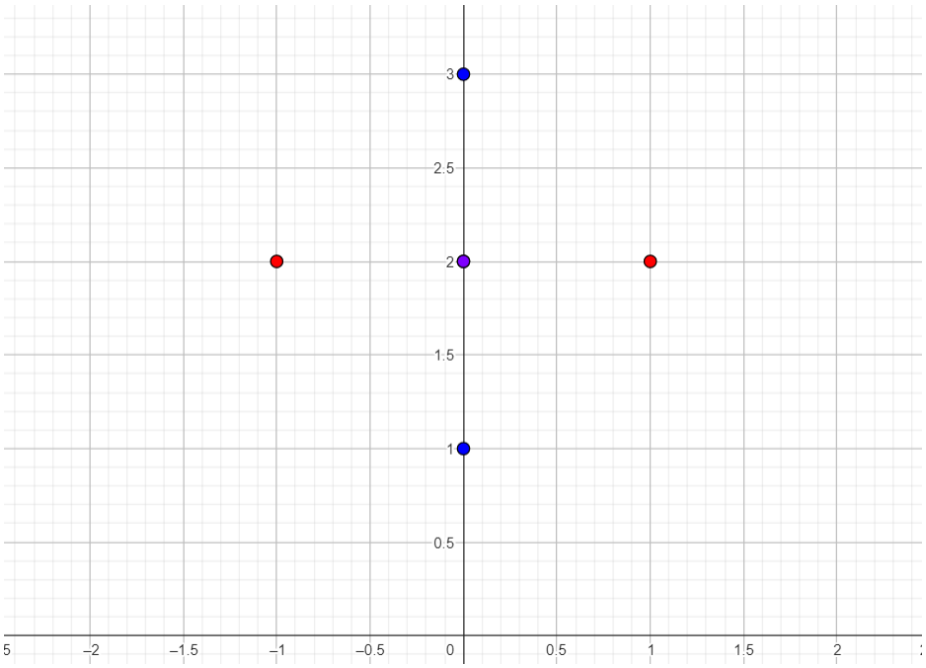


Curva de regresión tabla 1 (Y/X en rojo y X/Y en azul)

Tabla 2:

x_i	$\bar{y_i}$
-1	2
0	2
1	2

$\bar{x_j}$	y_j
0	1
0	2
0	3



Curva de regresión tabla 2 (Y/X en rojo y X/Y en azul)

En el anterior gráfico el punto (0,2) es doble, por eso está en morado.

Para terminar el ejercicio, calculamos las covarianzas de ambas distribuciones:

$\sigma_{xy}^{(1)} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = 0$

$\sigma_{xy}^{(2)} = \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = 0$

Se ha dado el caso de que ambas covarianzas son nulas, lo cual indica, por ejemplo, que el coeficiente de correlación es nulo, que las rectas de regresión son perpendiculares a los ejes y de la forma $y = \bar{y}$ y $x = \bar{x}$

Cabe mencionar que el valor de la primera covarianza ya lo sabíamos, puesto que las variables X e Y eran independientes y esto implica que la covarianza tiene que ser 0. Como vemos, no se da la implicación contraria, puesto que la segunda distribución también tiene covarianza nula pero las variables no son independientes.

1.6. Ejercicio 6:

Dada la siguiente distribución bidimensional:

X/Y	1	2	3	4	
10	1	3	0	0	4
12	0	1	4	3	8
14	2	0	0	2	4
16	4	0	0	0	4
	7	4	4	5	20

a) ¿Son estadísticamente independientes X e Y?

La independencia de 2 caracteres estadísticos viene dada por la caracterización:

$$\text{X e Y son independientes} \iff n_{ij} = \frac{n_{i.}n_{.j}}{n} \quad \forall i,j = 1,2,3,4$$

En el caso i=1 , j=1 tenemos:

$$n_{1,1} = 1 \neq \frac{n_{1.}n_{.1}}{n} = \frac{4 \times 7}{20} = \frac{28}{20} = 1.4$$

Por lo que ambos caracteres estadísticos no son independientes.

b) Calcular y representar las curvas de regresión de X/Y e Y/X.

Primero calcularé la curva de regresión de X/Y indicando por qué puntos deberá pasar:
Dichos puntos serán $(\bar{x}_j, y_j)$; $j = 1,2,3,4$;

$$\bar{x}_1 = \frac{10 \times 1 + 12 \times 0 + 14 \times 2 + 16 \times 4}{7} = 14.57 \Rightarrow (14.57; 1)$$

$$\bar{x}_2 = \frac{10 \times 3 + 12 \times 1 + 14 \times 0 + 16 \times 0}{4} = 10.5 \Rightarrow (10.5; 2)$$

$$\bar{x}_3 = \frac{10 \times 0 + 12 \times 3 + 14 \times 2 + 16 \times 0}{5} = 12 \Rightarrow (12; 3)$$

$$\bar{x}_4 = \frac{10 \times 1 + 12 \times 0 + 14 \times 2 + 16 \times 4}{7} = 12.8 \Rightarrow (12.8; 4)$$

Del mismo modo calculando la curva de Y/X calcularé $(x_i, \bar{y}_i)$:

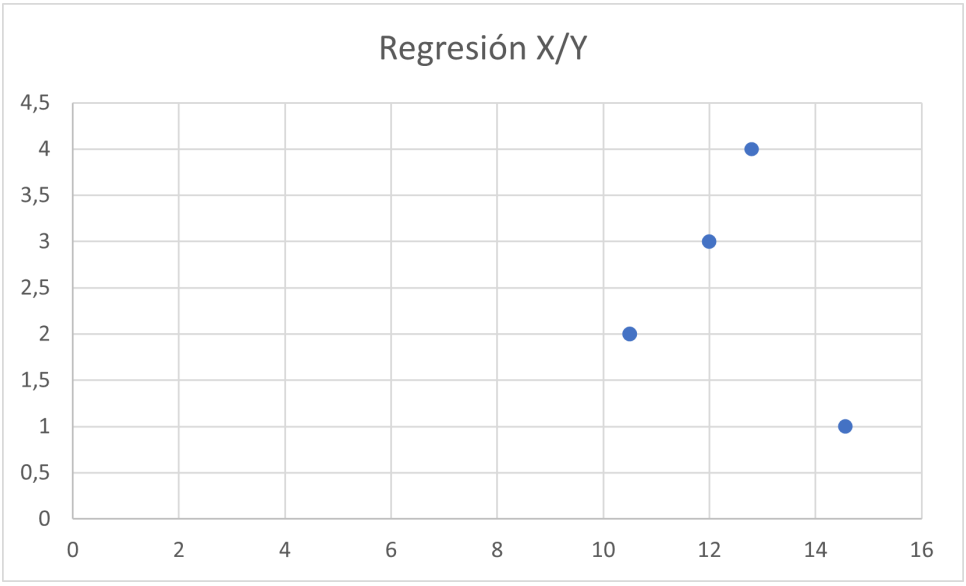
$$\bar{y}_1 = \frac{1 \times 1 + 2 \times 3 + 3 \times 0 + 4 \times 0}{4} = 1.75 \Rightarrow (1; 1.75)$$

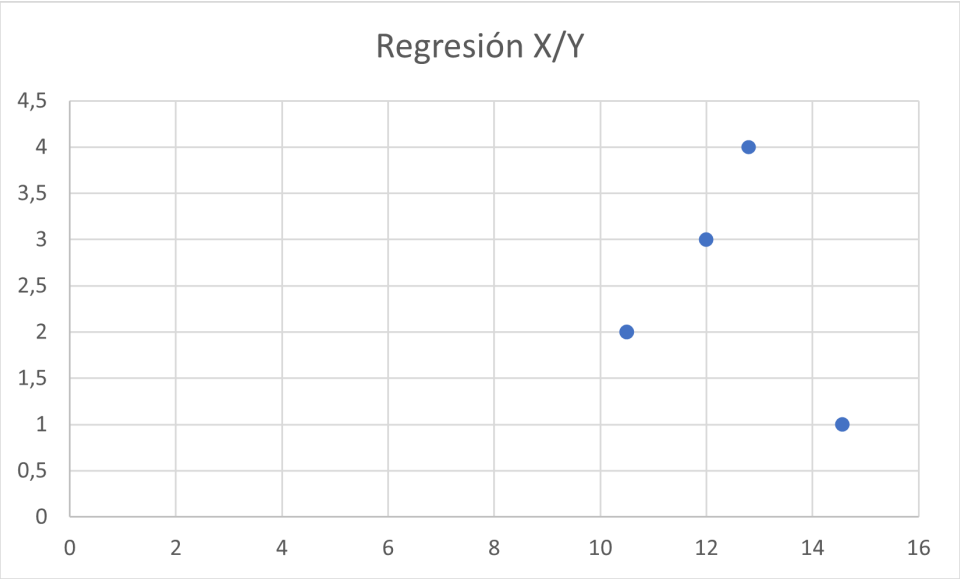
$$\bar{y}_2 = \frac{1 \times 0 + 2 \times 1 + 3 \times 4 + 4 \times 3}{8} = 3.25 \Rightarrow (2; 3.25)$$

$$\bar{y}_3 = \frac{1 \times 2 + 2 \times 0 + 3 \times 0 + 4 \times 2}{4} = 2.5 \Rightarrow (3; 2.5)$$

$$\bar{y}_4 = \frac{1 \times 4 + 2 \times 0 + 3 \times 0 + 4 \times 0}{4} = 1 \Rightarrow (4; 1)$$

Las representaciones gráficas quedarían:





c) Cuantificar el grado en que cada variable es explicada por la otra mediante la correspondiente curva de regresión.

Para ello calcularé la varianza explicada por cada una de las regresiones:

$$\sigma^2_{ey} = \sum_{i=1}^{4}\sum_{j=1}^{4} f_{ij}(\hat{y}_j - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}(\hat{y}_j - \bar{y})^2$$

$$\bar{y} = \frac{1}{n}\sum_{j=0}^{4} y_j n_{.j} = \frac{1\times 7 + 2\times 4 + 3\times 4 + 4\times 5}{20} = 2.35$$

$$\sigma^2_{ey} = \frac{(1+2+4)(1.75-2.35)^2 + (3+1)(3.25-2.35)^2 + 4(2.5-2.35)^2 + (3+2)(1-2.35)^2}{20} =$$

d) ¿Están X e Y correlacionadas linealmente? Dar las expresiones de las rectas de regresión.

Calcularé

1.7. Ejercicio 7:

Para cada una de las distribuciones:

a) ¿Dependen funcionalmente X de Y o Y de X?

b) Calcular curvas de regresión y comentar los resultados

Distribución A:

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

a) Es fácil ver que Y depende funcionalmente de X ya que a cada valor de X le corresponde un único valor no nulo de Y, pero no al revés.
b) Calcularé las curvas de regresión:
Y/X: puntos de la forma $(x_k, \bar{y_k})$

$$(x_1, \bar{y_1}) = (1, \frac{15 \times 2}{2}) = (1, 15)$$

$$(x_2, \bar{y_2}) = (2, \frac{10 \times 1}{1}) = (2, 10)$$

$$(x_3, \bar{y_3}) = (3, \frac{20 \times 3}{3}) = (3, 20)$$

$$(x_4, \bar{y_4}) = (4, \frac{5 \times 1}{1}) = (4, 15)$$

Y/X: puntos de la forma $(\bar{x_k}, y_k)$

$$(\bar{x_1}, y_1) = (\frac{2 \times 1}{1}, 10) = (2, 10)$$

$$(\bar{x_2}, y_2) = (\frac{1 \times 2 + 4 \times 1}{3}, 15) = (2, 15)$$

$$(\bar{x_3}, y_3) = (\frac{3 \times 3}{3}, 20) = (3, 20)$$

Distribución B:

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

a) En este caso Y depende funcionalmente de X por lo comentado anteriormente y además ahora también depende X de Y por el mismo motivo.
b) Calcularé las curvas de regresión:
Y/X: puntos de la forma $(x_k, \bar{y_k})$

$$(x_1, \bar{y_1}) = (1, \frac{15 \times 2}{2}) = (1, 15)$$

$$(x_2, \bar{y_2}) = (2, \frac{10 \times 1}{1}) = (2, 10)$$

$$(x_3, \bar{y_3}) = (3, \frac{20 \times 3}{3}) = (3, 20)$$

Y/X: puntos de la forma $(\bar{x_k}, y_k)$

$$(\bar{x_1}, y_1) = (\frac{2 \times 1}{1}, 10) = (2, 10)$$

$$(\bar{x_2}, y_2) = (\frac{1 \times 2}{3}, 15) = (1, 15)$$

$$(\bar{x_3}, y_3) = (\frac{3 \times 3}{3}, 20) = (3, 20)$$

Como se puede apreciar ambas curvas son iguales (o por lo menos en los puntos calculados).

Distribución C:

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

a) Es fácil ver que X depende funcionalmente de Y ya que a cada valor de Y le corresponde un único valor no nulo de X, pero no al revés.
b) Calcularé las curvas de regresión:

Y/X: puntos de la forma $(x_k, \bar{y_k})$

$$(x_1, \bar{y_1}) = (1, \frac{15 \times 3 + 25 \times 1}{4}) = (1; 17.5)$$

$$(x_2, \bar{y_2}) = (2, \frac{20 \times 1}{1}) = (2, 20)$$

$$(x_3, \bar{y_3}) = (3, \frac{10 \times 2}{2}) = (3, 10)$$

Y/X: puntos de la forma $(\bar{x_k}, y_k)$

$$(\bar{x_1}, y_1) = (\frac{3 \times 2}{2}, 10) = (3, 10)$$

$$(\bar{x_2}, y_2) = (\frac{1 \times 3}{3}, 15) = (1, 15)$$

$$(\bar{x_3}, y_3) = (\frac{2 \times 1}{1}, 20) = (2, 20)$$

$$(\bar{x_4}, y_4) = (\frac{1 \times 1}{1}, 25) = (1, 25)$$

1.8. Ejercicio 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:

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.

b) ¿Es apropiado suponer que existe 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?

a) Para las rectas de regresión, utilizaremos las fórmulas que se obtienen al realizar el método de mínimos cuadrados:

Recta de regresión Y/X

$$y = a + bx$$

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Recta de regresión X/Y

$$x = a' + b'y$$

$$a' = \bar{x} - b'\bar{y}$$

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

X/Y	1	2	3	4	$n_{i.}$
1	1	2	0	0	3
2	1	2	3	1	7
3	0	1	2	6	9
4	0	0	2	3	5
$n_{.j}$	2	5	7	10	24

Cálculo de medias:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} n_{i.}x_i = 2.667$

$\bar{y} = \frac{1}{n}\sum_{j=1}^{4} n_{.j}y_j = 3.042$

Cálculo de varianzas y covarianza:

$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{4} n_{i.}(x_i - \bar{x})^2 = 0.889$

$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{4} n_{.j}(y_j - \bar{y})^2 = 0.957$

$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = 0.597$

Cálculo de parámetros de las rectas:

$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{0.597}{0.889} = 0.672$

$b' = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{0.597}{0.957} = 0.624$

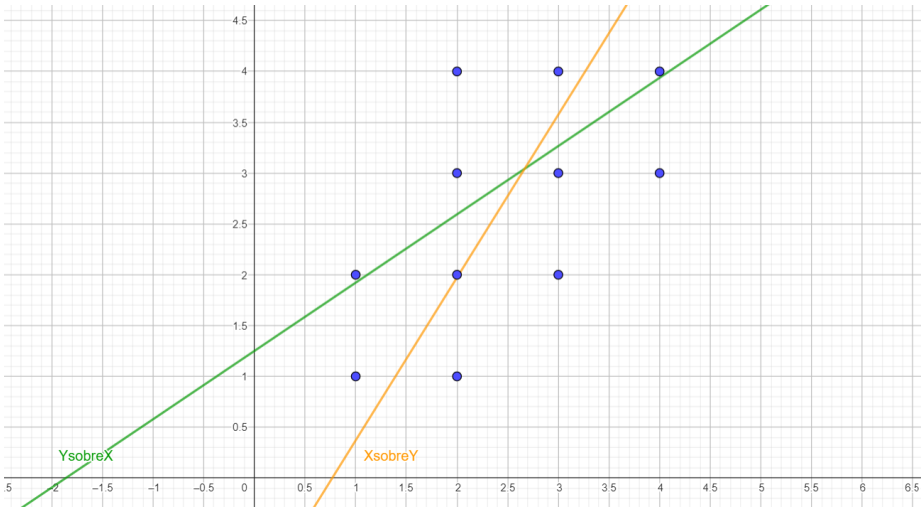
$a = \bar{y} - b\bar{x} = 3.042 - 0.672 \times 2.667 = 1.250$

$a' = \bar{x} - b'\bar{y} = 2.667 - 0.624 \times 3.042 = 0.769$

Por lo tanto, las rectas que nos quedan son:

Recta Y/X: $y = 1.25 + 0.672x$

Recta X/Y: $x = 0.769 + 0.624y$



b) Para ver si la relación lineal es fuerte, nos apoyamos en el coeficiente de correlación lineal:

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{0.597^2}{0.889 \times 0.957} = 0.419$$

Este valor se interpreta como que las rectas de regresión explican un 41.9 % de la distribución, lo cual indica que no es un buen modelo de aproximación y no sería apropiado suponer que la relación entre las variables X e Y es lineal.

c) Un puesto con 6 dependientes significa $y = 6$. Utilizamos la recta X/Y para obtener el valor esperado de x.

$$x = 0.769 + 0.624 \times 6 = 4.513 \; balanzas$$

Esta predicción no es fiable a pesar de no estar muy alejados del rango de y, puesto que se trata de un ajuste con una baja correlación lineal entre las variables.

1.9. Ejercicio 9:

Se eligen 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 X denota la edad del hombre e Y la de la mujer:

X/Y	(10,20]	(20,25]	(25,30]	(30,35]	(35,40]	$n_{i.}$
(15,18]	3	2	3	0	0	8
(18,21]	0	4	2	2	0	8
(21,24]	0	7	10	6	1	24
(24,27]	0	0	2	5	3	10
$n_{.j}$	3	13	17	13	4	50

Estudiar la interdependencia lineal entre ambas variables.
Calcularé para ello el correlación lineal que me indicará el grado de interdependencia de las 2 variables. Su expresión viene dada por:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_{xy} = \mu_{11} = \sum_{i=1}^{4}\sum_{j=1}^{5} f_{ij}(c_i - \bar{x})(c_j - \bar{y}) = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij}(c_i - \bar{x})(c_j - \bar{y})$$

Construiré una tabla más completa y calcularé las medias:

X/Y	(10,20]	(20,25]	(25,30]	(30,35]	(35,40]	$n_{i.}$	c_i
(15,18]	3	2	3	0	0	8	16.5
(18,21]	0	4	2	2	0	8	19.5
(21,24]	0	7	10	6	1	24	22.5
(24,27]	0	0	2	5	3	10	25.5
$n_{.j}$	3	13	17	13	4	50	
c_j	15	22.5	27.5	32.5	37.5		

$$\bar{x} = \frac{1}{n} c_i n_{i.} = \frac{16.5 \times 8 + 19.5 \times 8 + 22.5 \times 24 + 25.5 \times 10}{50} = 21.66$$

$$\bar{y} = \frac{1}{n} c_j n_{.j} = \frac{15 \times 3 + 22.5 \times 13 + 27.5 \times 17 + 32.5 \times 13 + 37.5 \times 4}{50} = 27.55$$

$$\sigma_{xy} = \frac{1}{50}(194.274 + 52.116 + 0.774 + 43.632 + 0.216 - 21.384 - 29.694 - 0.42 + 24.948 +$$

$$8.358 - 0.384 + 95.04 + 114.624) = 9.642$$

$$\sigma_x = \sqrt{\mu_{20}} = \sqrt{\frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij}(c_i - \bar{x})^2} = 2.88$$

$$\sigma_y = \sqrt{\mu_{02}} = \sqrt{\frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{5} n_{ij}(c_j - \bar{y})^2} = 5.51$$

$$r = \frac{9.642}{2.88 \times 5.51} = 0.6076$$

Tienen por tanto una interdependencia lineal del 60.76 %

1.10. Ejercicio 10:

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

$$x + 4y = 1$$

$$x + 5y = 2$$

Como no sabemos cual es la recta de regresión lineal Y /X y cual es la X/Y , supondremos que la primera es la que explica la variable x en función de la variable y (X/Y) y la segunda la que explica la variable y en función de la variable x y (Y/X).
Sabemos que

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

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

Por lo tanto, despejando de las ecuaciones tenemos:

$$x + 4y = 1 \;\Rightarrow\; x = 1 - 4y \;\Rightarrow\; \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2}\bar{y} = 1 \quad \frac{\sigma_{xy}}{\sigma_y^2} = -4$$

$$x + 5y = 2 \;\Rightarrow\; y = 0.4 - 0.2x \;\Rightarrow\; \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x} = 0.4 \quad \frac{\sigma_{xy}}{\sigma_x^2} = -0.2$$

De estos resultados saco que $\sigma_{xy} < 0 \Rightarrow r < 0$
Trabajando con la expresión de r^2 tenemos:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x} \times \frac{\sigma_{xy}}{\sigma_y} = (-4) \times (-0.2) = 0.8$$

Por lo tanto

$$r = -\sqrt{r^2} = -0.8944 \quad \Rightarrow \quad \text{Tienen una relación inversa del } 89.44\,\%$$

1.11. Ejercicio 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.

La recta de regresión X sobre Y tiene la forma $x = a + by$, con $a = \bar{x} + b\bar{y}$ y con $b = \frac{\sigma_{xy}}{\sigma_y^2}$, por lo que los datos que necesitamos son $\bar{x},\ \bar{y},\ \sigma_{xy}$ y σ_y^2

Primero, calculamos las medias. La media de x se obtiene de la tabla:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} n_{i.}x_i = \frac{3\times 5 + 5\times 1 + 8\times 2 + 9\times 1}{9} = 5$$

Sabiendo que ambas rectas de regresión pasan por el punto $(\bar{x}, \bar{y})$ deducimos que $\bar{y} = 5\bar{x} - 20 = 5\times 5 - 20 = 5$.

Teniendo ya la media de x, podemos calcular la varianza de x:

$$\sigma_x^2 = \frac{1}{9}\sum_{i=1}^{4} n_{i.}(x_i - \bar{x})^2 = 6$$

De la recta de regresión Y/X, sabemos que $5 = \frac{\sigma_{xy}}{\sigma_x^2}$ y por tanto:

$$\sigma_{xy} = 5\sigma_x^2 = 5\times 6 = 30$$

Haciendo uso del dato restante, calculamos la varianza de y:

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{4} n_{.j}(y_j - \bar{y})^2 = \frac{1}{9}\sum_{i=1}^{4} n_{.j}(y_j^2 - 2y_j\bar{y} + \bar{y}^2) =$$

$$= \frac{1}{9}(\sum_{i=1}^{4} n_{.j}y_j^2 - 2\sum_{i=1}^{4} n_{.j}y_j\bar{y} + \sum_{i=1}^{4}\bar{y}^2) = \frac{1}{9}\sum_{i=1}^{4} y_j^2 n_{.j} - \bar{y}^2 = \frac{3240}{9} - 25 = 335$$

Ya finalmente calculamos la recta de regresión:

$$b = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{30}{335} = 0.0896 \qquad a = \bar{x} - b\bar{y} = 5 - 0.0896\times 5 = 4.55$$

y = 4.55 + 0.09x

1.12. Ejercicio 12:

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 esos datos se han obtenido las siguientes medidas:

$$\sum y_i = 219.719 \qquad \sum {y_i}^2 = 2396.504 \qquad \sum x_i y_i = 349.486$$

$$\sum x_i = 31.470 \qquad \sum {x_i}^2 = 51.075 \qquad \sum {x_i}^2 y_i = 633.993$$

$$\sum {x_i}^4 = 182.977 \qquad \sum {x_i}^3 = 93.6$$

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 compuesto de 100 vuelos de media hora, 200 de una hora y 100 de dos horas? ¿Es fiable esta estimación?

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

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

Empezamos calculando los momentos que se puede ver, que son estas expresiones divididas entre el numero de vuelos estudiados: (cabe destacar que $n_{ij} = 1$ para cualesquiera i, j.

$$m_{10} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} x_i = \frac{1}{24} 31.470 = 1.31125$$

$$m_{20} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} x_i^2 = \frac{1}{24} 51.075 = 2.128125$$

$$m_{21} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} x_i^2 y_j = \frac{1}{24} 633.993 = 26.416375$$

$$m_{01} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} y_j = \frac{1}{24} 219.719 = 9.154958333$$

$$m_{02} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} y_j^2 = \frac{1}{24} 2396.504 = 99.85433333$$

$$m_{11} = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}}{n}x_i y_j = \frac{1}{24}349.486 = 14.56191667$$

$$m_{30} = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}}{n}x_i^3 = \frac{1}{24}93.6 = 4.0125$$

$$m_{40} = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}}{n}x_i^4 = \frac{1}{24}182.977 = 7.624041667$$

$$\sigma_{xy} = \mu_{11} = m11 - m10m01 = 14.56191667 - 1.31125 * 9.154958333 = 2.557477556$$

$$\sigma_x^2 = \sum_{i=1}^{k}\frac{n_{i\cdot}}{n}x_i^2 - x^2 = m20 - m10^2 = 2.128125 - 1.31125^2 = 0.4087484375$$

$$\sigma_y^2 = \sum_{j=1}^{p}\frac{n_{\cdot j}}{n}y_j^2 - y^2 = m02 - m01^2 = 99.85433333 - 9.154958333^2 = 16.041071251$$

a)Una recta de regresión lineal es d la forma $y = ax + b$ con

$$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{2.557477556}{0.4087484375} = 6.2568497427$$

$$b = y - \frac{\sigma_{xy}}{\sigma_x^2}x = 9.154958333 - 6.2568497427 * 1.31125 = 0.95066410788$$

entonces $y = 6.2568497427x + 0.95066410788$

$$400f(30) = 400 * 188.656156389 = 75462.4625556$$

$$200f(60) = 200 * 376.36164867 = 75272.329734$$

$$100f(120) = 100 * 751.772633232 = 75177.2633232$$

es bastante fiable pues estos resultados se parecen mucho, pero vamos a hacer el coeficiente de estimación:

$$r = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{2.557477556}{\sqrt{0.4087484375}\sqrt{16.041071251}} = 0,9987537291$$

los resultados son prácticamente idénticos como podemos observar.

b) Una parábola de la forma $y = a + bx + cx^2$, lo haremos obviamente por el método de mínimos cuadrados, debemos minimizar la siguiente función:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - f(x_i))^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - a - bx_i - cx_i^2)^2$$

hagamos la derivada parcial respecto de cada variable a, b y c, y igualamos a cero para averiguar la función mínima.
$\phi_{(a,b,c)} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - a - bx_i - cx_i^2)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - a - bx_i - cx_i^2)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j^2 - ay_j - bx_iy_j - cx_i^2y_j - ay_j + a^2 + abx_i + acx_i^2 - bx_iy_j + abx_i + b^2x_i^2 + bcx_i^3 - y_jcx_i^2 + acx_i^2 + bcx_i^3 + c^2x_i^4) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j^2 - 2ay_j + a^2 + 2abx_i - 2bx_iy_j - 2cx_i^2y_j + 2acx_i^2 + b^2x_i^2 + 2bcx_i^3 + c^2x_i^4)$

$$\frac{\partial\phi_{(a,b,c)}}{\partial a} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j^2 - 2y_j + 2a + 2bx_i - 2bx_iy_j - 2cx_i^2y_j + 2cx_i^2 + b^2x_i^2 + 2bcx_i^3 + c^2x_i^4)$$

$$\frac{\partial\phi_{(a,b,c)}}{\partial b} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j^2 - 2ay_j + a^2 + 2ax_i - 2x_iy_j - 2cx_i^2y_j + 2acx_i^2 + 2bx_i^2 + 2cx_i^3 + c^2x_i^4)$$

$$\frac{\partial\phi_{(a,b,c)}}{\partial c} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j^2 - 2ay_j + a^2 + 2abx_i - 2bx_iy_j - 2x_i^2y_j + 2ax_i^2 + b^2x_i^2 + 2bx_i^3 + 2cx_i^4)$$

Nos queda entonces el sistema de ecuaciones:

$$2m_{01} + 2a + 2bm_{10} + 2cm_{20} = 0a + 1,3113b + 2,1281c = 9,155$$

$$2m_{11} + 2am_{10} + 2bm_{20} + 2cm_{30} = 01,3113a + 2,1281b + 3,9c = 14,5619$$

$$2m_{21} + 2am_{20} + 2bm_{30} + 2cm_{20} \to 2,1281a + 3,9b + 7,624c = 26,4139$$

$$\left.\begin{array}{rcl} a + 1,3113b + 2,1281c &=& 9,155 \\ 1,3113a + 2,1281b + 3,9c &=& 14,5619 \\ 2,1281a + 3,9b + 7,624c &=& 26,4139 \end{array}\right\}$$

$$\left.\begin{array}{rcl} a + 1,3113b + 2,1281c &=& 9,155 \\ 1,3113a + 2,1281b + 3,9c &=& 14,5619 \\ 2,1281a + 3,9b + 7,624c &=& 26,4139 \end{array}\right\} \left\{\begin{array}{ll} a & = 0,7491 \\ b & = 6,6368 \\ c & = -0,1395 \end{array}\right.$$

Entonces $y = 0,7491 + 6,6368x - 0,1395x^2$.

$$400f(30) = 400 * 188.656156389 = 75462.4625556$$

$$200f(60) = 200 * 376.36164867 = 75272.329734$$

$$100f(120) = 100 * 751.772633232 = 75177.2633232$$

c) Es claro ver que en el segundo modelo lo único que se añade es un insignificante $c = -0,1395$, y el resto de la ecuación es casi idéntica, pero no ganamos casi nada de precisión a cambio de una gran complicación en los cálculos.

1.13. Ejercicio 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

para hacer ajuste de hipérbola equilátera cambiamos el dato X por z=1/x, así: $y = a\frac{1}{x} + b \rightarrow y = az + b$, y aplicamos regresión lineal.

	x	y	z
	10	50	0,1
	12,5	90	0,08
	20	160	0,05
	25	180	0,04
media:	16,875	120	0,0675
desv tipica	35,546875	2750	0,00056875
var	5,962120009	52,44044241	0,02384848004

Covarianza:

$$\sigma_{zy} = m11 - m10m01 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}z_i y_j - (zy) = 6,85 - 120 * 0,0675 = -1,25$$

Entonces la pendiente de la recta sería:

$$\frac{\sigma_{zy}}{\sigma_z^2} = \frac{-1,25}{0,00056875} = -2197,802198$$

$y = az + b$, con $a = -2197,802198$ y
$b = \frac{\sigma_{zy}}{\sigma_z^2} y - z = \frac{-2197.802198}{0.00056875} 120 - 0.0675 = 268,3516484$
sustituimos, entonces
$y = -2197,802198z + 268,3516484 \rightarrow y = \frac{-2197,802198}{x} + 268,3516484$
Veamos ahora como de bueno es este ajuste, calculemos los residuos:

$$\sigma_{ry}^2 = \frac{1}{4}\sum_{i=1}^{4} n_i (f(x_i) - y))^2 = 2,237048666$$

Veamos entonces el coeficiente de determinación:

$$\frac{\sigma_{ey}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2} = 1 - \frac{2,237048666}{2750} = 0,9991865278$$

es prácticamente perfecto, casi igual a 1. Porque los residuos son casi despreciables

1.14. Ejercicio 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 comportamiento de Y por X mediante:

a) Relación lineal.

b) Hipérbola equilátera.

c) Curva potencial.

d) Curva exponencial.

¿Qué ajuste es más adecuado?

a) Para el ajuste lineal, el más sencillo, realizaremos un ajuste por mínimos cuadrados a la distribución dada realizando primero los cálculos de los coeficientes necesarios:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{6} n_{i.}x_i = \frac{9+10+12+15+22+32}{6} = 16.67$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{6} n_{.j}y_j = \frac{30+5+70+80+120+140}{6} = 81.67$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.}(x_i-\bar{x})^2 = 65.22$$

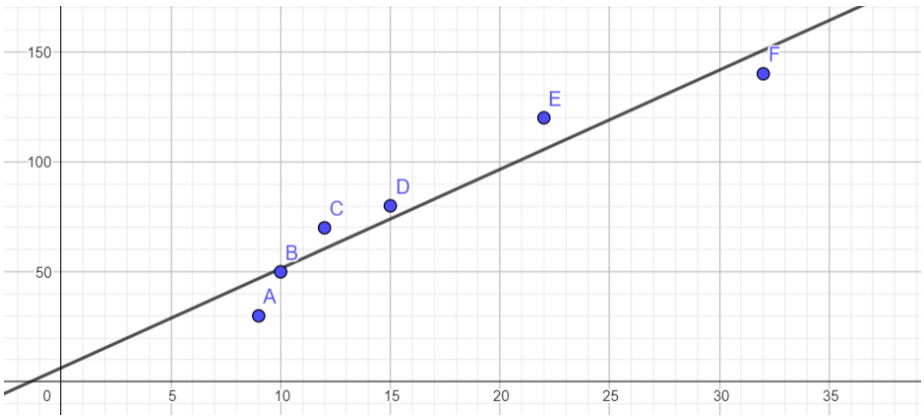
$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{6} n_{.j}(y_j-\bar{y})^2 = 1447.22$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}(x_i-\bar{x})(y_j-\bar{y}) = 293.89$$

Por tanto, los coeficientes de la recta de regresión $y = a + bx$ son:

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{293.89}{65.22} = 4.51$$

$$a = \bar{y} - b\bar{x} = 81.67 - 4.51 \times 16.67 = 6.49$$



Ajuste lineal ($y = 6.49 + 4.51x$)

b) Para el ajuste hiperbólico, realizaremos un cambio de variable $z = \frac{1}{x}$ de manera que la ecuación quede como $y = a + \frac{b}{x} = a + bz$. Así, el problema se reduce a hacer un ajuste lineal entre z e y para obtener los parámetros a y b.

Y	30	50	70	80	120	140
$z = \frac{1}{x}$	0.111	0.1	0.083	0.067	0.045	0.031

Ahora calculamos la media, varianza y covarianza:

$$\bar{z} = \bar{x} = \frac{1}{n}\sum_{i=1}^{6} n_{i.} z_i = \frac{0.111 + 0.1 + 0.083 + 0.067 + 0.045 + 0.031}{6} = 0.073$$

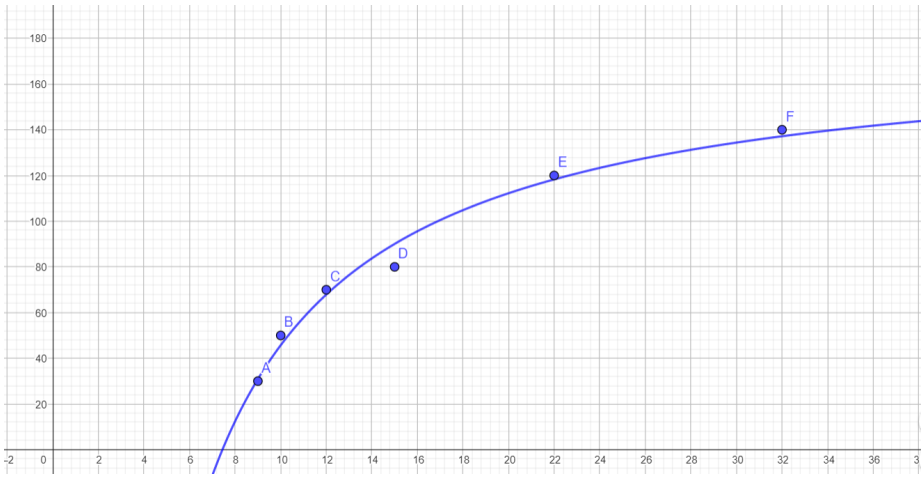
$$\sigma_z^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.}(z_i - \bar{z})^2 = 0.000805$$

$$\sigma_{zy} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}(z_i - \bar{z})(y_j - \bar{y}) = -1.071$$

Calculamos los parámetros de la hipérbola:

$$b = \frac{\sigma_{zy}}{\sigma_z^2} = \frac{-1.071}{0.000805} = -1330.27$$

$$a = \bar{y} - b\bar{z} = 81.67 - (-1330.43) \times 0.073 = 178.73$$



Ajuste hiperbólico ($y = 178.73 - \frac{1330.27}{x}$)

c) Para el ajuste potencial es necesario tomar logaritmos a ambos lados de la ecuación. De manera que la ecuación $y = ax^b$ se nos queda como $ln(y) = ln(a) + b\ ln(x)$.
Ahora tomamos la ecuación $Y = A + b\ X$, donde $Y = ln(y)$, $A = ln(a)$ y $X = ln(X)$, con lo que el problema se nos queda en un ajuste lineal.

$Y = ln(y)$	3.4	3.91	4.24	4.38	4.79	4.94
$X = ln(x)$	2.2	2.3	2.485	2.71	3.09	3.47

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{6} n_{i.}X_i = \frac{2.2 + 2.3 + 2.485 + 2.71 + 3.09 + 3.47}{6} = 2.71$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{6} n_{.j}y_j = \frac{3.4 + 3.91 + 4.24 + 4.38 + 4.79 + 4.94}{6} = 4.28$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{6} n_{i.}(X_i - \bar{X})^2 = 1.196$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{6} n_{.j}(Y_j - \bar{Y})^2 = 1.61$$

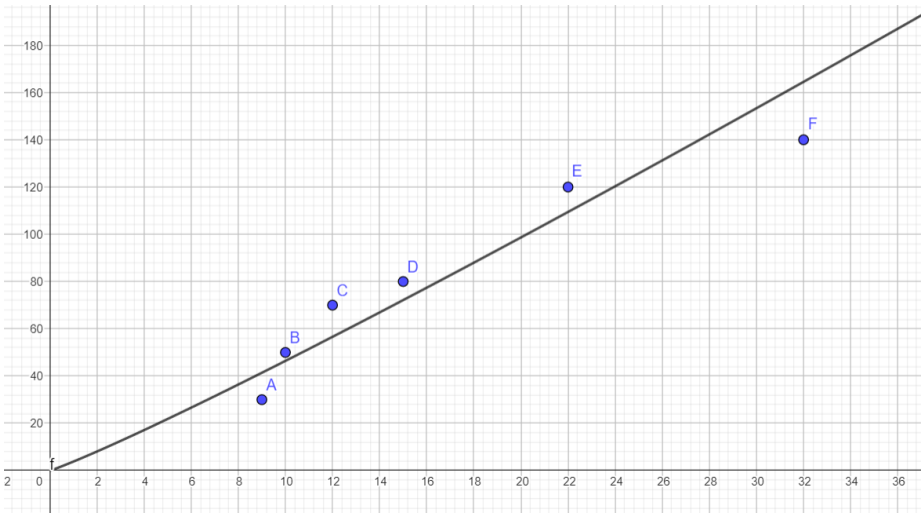
$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}(X_i - \bar{X})(Y_j - \bar{Y}) = 1.3$$

Por tanto, los coeficientes de la recta de regresión $y = a + bx$ son:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{1.3}{1.196} = 1.088$$

$$A = \bar{Y} - b\bar{X} = 4.28 - 4.28 \times 2.71 = 1.33$$

Teniendo en cuenta que $a = e^A = e^1.33 = 3.79$ el ajuste queda así:



Ajuste potencial ($y = 3.79\ x^{1.088}$)

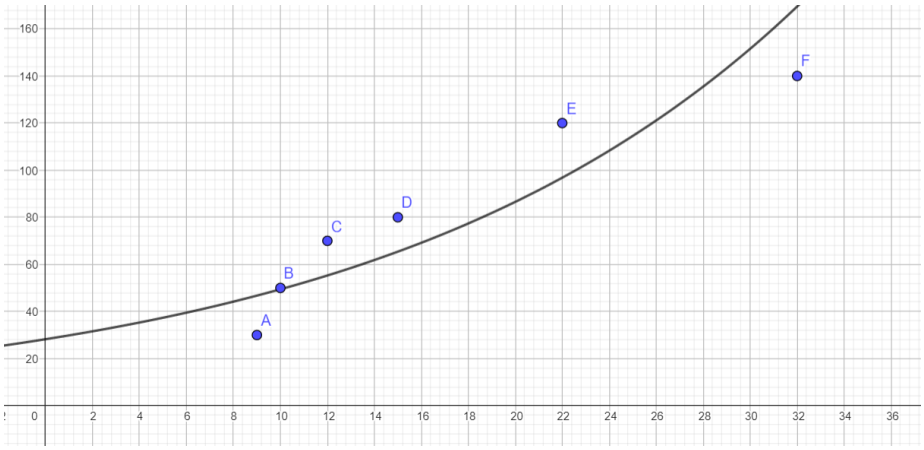
d) Al igual que para el ajuste potencial, para la curva exponencial tomaremos logaritmo a ambos lados de la ecuación de manera que nos queda $Y = A + bx$, donde $Y = ln(y)$ y $A = ln(a)$. Todos los coeficientes necesarios ya están calculados para este ajuste, salvo σ_{xY}.

$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij}(x_i - \bar{X})(Y_j - \bar{Y}) = 3.67$$

$$b = \frac{\sigma_{xY}}{\sigma_x^2} = \frac{3.67}{65.22} = 0.056$$

$$A = \bar{Y} - b\bar{x} = 4.28 - 0.056 \times 16.67 = 3.34$$

Teniendo en cuenta que $a = e^A = e^{3.34} = 28.25$ el ajuste queda así:



Ajuste exponencial ($y = 28.25\ e^{0.056x}$)

Para ver qué ajuste es mejor de todos, nos apoyaremos en los respectivos coeficientes de correlación.

$$Lineal : R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{293.89^2}{65.22 \times 1447.22} = 0.915$$

$$Hiperbólico : R^2 = \frac{\sigma_{zy}^2}{\sigma_z^2 \sigma_y^2} = \frac{(-1.07)^2}{0.000805 \times 1447.22} = 0.983$$

Para los modelos potencial y exponencial no podemos usar la fórmula de R^2 pues no son lineales en los parámetros. En su lugar tendremos que usar las respectivas varianzas residuales:

$$Potencial : \sigma_{ry} = \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} n_{i.} (f(x_i) - y_j)^2 = 182.5$$

$$\eta^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2} = 1 - \frac{182.5}{1447.22} = 0.874$$

$$Exponencial : \sigma_{ry} = \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} n_{i.} (f(x_i) - y_j)^2 = 361.75$$

$$\eta^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2} = 1 - \frac{361.75}{1447.22} = 0.75$$

Con esto concluimos que el mejor ajuste de todos es el hiperbólico, con un coeficiente de correlación de 0.983, lo cual ya podíamos intuirlo al ver que la gráfica pasaba muy cerca de todos los puntos.
Por contra, el peor ajuste de todos es el exponencial, con un coeficiente de correlación de 0.75, que también podíamos intuirlo, pues la concavidad con la que están dispuestos es opuesta a la concavidad de una función exponencial.