Estadística Descriptiva e Introducción a la Probabilidad

Relación de Ejercicios 2

Autores, por orden alfabético:
Shao Jie Hu Chen
Adrián Jaén Fuentes
Aarón Jerónimo Fernández
Noura Lachhab Bouhmadi
Laura Lázaro Soraluce

Doble Grado en Ingeniería Informática y Matemáticas

Lista de Problemas

Problema 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

1. Construir la tabla de frecuencias.
2. Calcular las puntuaciones medias obtenidas con cada dado y ver cuales son más homogéneas.
3. ¿Qué resultado del segundo dado es más frecuente cuando en el primero se obtiene un 3?
4. 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.

Antes de resolver el ejercicio crearemos un nuevo enunciado, pues no tiene sentido calcular las cosas que nos piden, cuando estas dependen de la aleatoriedad. En nuestro nuevo enunciado, la variable X será el número de personas que componen distintas familias de Granada, y la variable Y es el número de habitaciones que tienen en sus casas. Resolvemos ahora el ejercicio.

Apartado 1

Calculamos la tabla de frecuencia y además, los elementos que necesitaremos durante el ejercicio.

x_i	$n_{i.}$	$x_i n_{i.}$	$x_i^2 n_{i.}$
1	4	4	4
2	3	6	12
3	5	15	45
4	6	24	96
5	4	20	100
6	2	12	72
	24	81	329

y_j	$n_{.j}$	$y_j n_{.j}$	$y_j^2 n_{.j}$
1	6	6	6
2	6	12	24
3	2	6	18
4	2	8	32
5	3	15	75
6	5	30	180
	24	77	335

Apartado 2

Calculamos la media de nuestras variables:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_{i.} = \frac{81}{24} = 3{,}375 \approx 3 hijos \qquad \overline{y} = \frac{1}{n}\sum_{k=1}^{p} y_j n_{j.} = \frac{77}{24} = 3{,}2083 \approx 3 habitaciones$$

Ahora calculamos la desviación típica de cada variable para poder calcular el coeficiente de variación de Pearson y ver cuál es más homogénea:

$$\sigma_x = \sqrt{\frac{328}{24} - 3{,}375^2} = 1{,}5224 hijos \qquad \sigma_y = \sqrt{\frac{335}{24} - 3{,}2083^2} = 1{,}9144 habitaciones$$

$$CV_x = \frac{\sigma_x}{\overline{x}} = 0{,}4511 \qquad CV_y = \frac{\sigma_y}{\overline{y}} = 0{,}5967$$

Como vemos, los resultados de la primera variable son más homogéneos pues su coeficiente de variación de Pearson es menor.

Apartado 3

Mirando la tabla de resultados, vemos que cuando las familias tenían 3 habitaciones en sus casas, el número de hijos más frecuente es el 4, con 2 veces, luego $M_0 = 4$.

Apartado 4

Vamos a calcular la media de X/Y cuando las familias tienen dos o cinco habitaciones en sus casa. Mirando la tabla de resultados obtenemos:

x_i/Y	n_i	N_i
1	1	1
2	1	2
3	1	3
4	3	6
5	1	7
6	2	9

Como $n/2 = 4{,}5$ y la frecuencia absoluta acumulada inmediatamente superior a este número es $N_4 = 6$, deducimos que la mediana es $M_e = 4$.

Problema 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 medidas y las desviaciones típicas):

X/Y	160	162	164	166	168	170	$n_{i.}$	$n_{i.}x_i$	$n_{ij}x_i^2 - \bar{x}$
48	3	2	2	1	0	0	8	384	304,689
51	2	3	4	2	2	1	14	714	140,773
54	1	3	6	8	5	1	24	1296	0,702
57	0	0	1	2	8	3	14	798	412,045
60	0	0	0	2	4	4	10	600	339,772
$n_{.j}$	6	8	13	15	19	9	70	3792	
$n_{.j}y_j$	960	1296	2132	2490	3192	1530	11600		
$n_{.j}y_j^2 - \bar{y}^2$	195.899	110.35	38.191	1.227	99.29	165.328			

Apartado 1

Para hallar el peso medio, calculamos la media de X, $\bar{x}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i = \frac{3792}{70} = 54,171 kg.$$

Hacemos lo mismo con la variable Y para hallar la altura media, $\bar{y}$.

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j = \frac{11600}{70} = 165,714kg.$$

Para ver cual de las dos es mas representativas, utilizamos el coeficiente de variación de Pearson.
$C.V(X) = \frac{\sigma_x}{\bar{x}}$
$C.V(Y) = \frac{\sigma_y}{\bar{y}}$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2 - \bar{x}^2 = m_{20} - m_{10}^2 = 12,874kg^2.$$

$\sigma_x = 3,588$ kg.

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2 - \bar{y}^2 = m_{02} - m_{01}^2 = 8,813cm^2.$$

$\sigma_y = 2,969$ cm.

$C.V(X) = \frac{3,588}{54,171} = 0,066$
$C.V(Y) = \frac{2,969}{165,714} = 0,018$

Como vemos el coeficiente de variación de Pearson de la X es mayor, por tanto, es más representativa la altura media.

Apartado 2

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

Para hallar el porcentaje, dividimos el número total de individuos que cumplen las dos condiciones entre el número total de individuos estudiados.
$\frac{20}{70} * 100 = 28,571\,\%$

Apartado 3

X/Y	166	168	170	$n_{i.}$
48	1	0	0	1
51	2	2	1	5
54	8	5	1	14
57	2	8	3	13
60	2	4	4	10
$n_{.j}$	15	19	9	43

Para hallar este porcentaje, tomamos el total de individuos que miden más de 165cm y miramos cuantos de ellos pesan más de 52kg. Hacemos $n_{i3} + n_{i4} + n_{i5} = 14 + 13 + 10 = 37$ personas que pesan más de 52kg de las 43 que miden más de 165cm. Hacemos $\frac{37}{43} * 100 = 86,047\,\%$.

Apartado 4

En la tabla del apartado a, buscamos el mayor n_{ij} dentro de los que pesan entre 51 y 57kg. En este caso vemos que el mayor es $n_{34} = 8$. Esta es la moda, la altura más frecuente entre los individuos que pesan entre 51 y 57kg.

Apartado 5

Este apartado se resuelve de manera análoga al apartado a. Esta vez calculamos las medias y las varianzas de las x correspondientes a 164 y 168cm.

Para las medias, hacemos

X/Y	164	n_{i3}	$n_{i3}x_i$
48	2	2	38,456
51	4	4	7,673
54	6	6	15,649
57	1	1	21,298
60	0	0	0
		13	83.076

$$\bar{x}_{164} = \frac{1}{n_{,3}} \sum_{i=1}^{k} n_{i3}x_i = \frac{48 * 2 + 51 * 2 + 54 * 6 + 57}{2 + 4 + 6 + 1} = \frac{681}{13} = 52,385kg.$$

$$\sigma^2_{x_{164}} = \frac{1}{n_{,3}} \sum_{i=1}^{k} n_{i3}x_i^2 - \bar{x_{164}}^2 = m_{20} - m_{10}^2 = \frac{83,076}{13} = 6,39kg^2.$$

$\sigma_{x_{164}} = 2,528$ kg.

$C.V(X) = \frac{\sigma_{x_{164}}}{\bar{x_{164}}} = \frac{2,528}{52,385} = 0,0483$

X/Y	168	n_{i5}	$n_{i5}x_i^2 - \bar{x_i}^2$
48	0	0	0
51	2	2	54,309
54	5	5	24,443
57	8	8	4,98
60	4	4	57,426
		19	141,158

$$\bar{x}_{168} = \frac{1}{n_{,5}} \sum_{i=1}^{k} n_{i5}x_i = \frac{2 * 51 + 5 * 54 + 8 * 57 + 4 * 60}{2 + 5 + 8 + 4} = \frac{1068}{19} = 56,211kg.$$

$$\sigma^2_{x_{168}} = \frac{1}{n_{,5}} \sum_{i=1}^{k} n_{i5} x_i^2 - \bar{x_{168}}^2 = m_{20} - m_{10}^2 = \frac{141,158}{19} = 7,429 kg^2.$$

$\sigma_{x_{168}} = 2,726$ kg.

$C.V(Y) = \frac{\sigma_{x_{168}}}{\bar{x_{168}}} = \frac{2,726}{56,211} = 0,0485$

Es más representativo el peso de los individuos que miden 164 cm.

Problema 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:

$X \backslash Y$	1	2	3	4
1	7	0	0	0
2	10	2	0	0
3	11	5	1	0
4	10	6	6	0
5	8	6	4	2
6	1	2	3	1
7	1	0	0	1
8	0	0	1	1

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

Apartado 1

Esta será la tabla que utilicemos en el ejercicio, con todos los datos necesarios ya calculados.

X/Y	1	2	3	4	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum_{j=1}^{p} n_{ij}y_j$	$x_i \sum_{j=1}^{p} n_{ij}y_j$
1	7	0	0	0	7	7	7	7	7
2	10	2	0	0	12	24	48	14	28
3	11	5	1	0	17	51	153	24	72
4	10	6	6	0	22	88	352	40	160
5	8	6	4	2	20	100	500	40	200
6	1	2	3	1	7	42	252	18	108
7	1	0	0	1	2	14	98	5	35
8	0	0	1	1	2	16	128	7	56
$n_{.j}$	48	21	15	5	89	342	1538		666
$n_{.j}y_j$	48	42	45	20	155				
$n_{.j}y_j^2$	48	84	135	80	347				

Para calcular la recta de regresión de Y sobre X tenemos que tener en cuenta lo siguiente:

$$y = ax + b \qquad a = \frac{\sigma_{xy}}{\sigma_x^2} \qquad b = \overline{y} - a\overline{x}$$

Ahora haremos los cálculos necesarios:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_{i.} = \frac{342}{89} = 3{,}8427 \approx 4 personas \qquad \overline{y} = \frac{1}{n}\sum_{k=1}^{p} y_j n_{j.} = \frac{155}{89} = 1{,}7415 \approx 2 personas$$

$$\sigma_x^2 = m_{11} - m_{10}^2 = \frac{1539}{89} - \overline{x}^2 = 2{,}5146 \qquad \sigma_{xy} = m_{11} - m_{10}m_{01} = \frac{666}{89} - \overline{xy} = 0{,}791(1)$$

Teniendo en cuenta los resultados obtenidos:

$$a = \frac{\sigma_{xy}}{\sigma_x^2} = 0{,}315 \qquad b = \overline{y} - a\overline{x} = 0{,}531$$

Y, finalmente, nuestra recta de regresión es:

$$y = 0{,}315x + 0{,}531$$

Apartado 2

Para saber si es adecuada o no esta suposición, calcularemos el coeficiente de determinación lineal. Para ello necesitamos calcular lo siguiente:

$$\sigma y^2 = m_{02} - m_{01}^2 = \frac{347}{89} - \overline{y}^2 = 0{,}866$$

Por último, calculamos el coeficiente:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = 0{,}287$$

Esto significa que la recta que hemos calculado, explica el 28.7 % de nuestra distribución, con lo que deducimos que no es adecuado suponer una relación lineal.

Problema 4

Se realiza un estudio sobre la tensión de vapor de agua (Y, en ml. de Hg.) a distintas temperaturas (X, en ^{o}C). Efectuadas 21 medidas, los resultados son:

$X \backslash Y$	$(0.5, 1.5]$	$(1.5, 2.5]$	$(2.5, 5.5]$
$(1, 15]$	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?

• En una población de tamaño n = 21 medidas se ha observado dos variables estadísticas, X = temperatura en ^{o}C e Y = vapor de agua en ml de Hg, las cuales han presentado k = 3, p = 3 modalidades distintas, con distribución de frecuencia conjunta (x_i, y_j), $n_{ij} i = 1, ..., 3$ $j = 1, ..,3$

• Empezamos rellenando nuestra tabla:

X Y	(0,5, 1,5]	(1,5, 2,5]	(2,5, 5,5]	$n_{i.}$	$c_{i.}$	$n_{i.}c_{i.}$	$n_{i.}c_{i.}^2$	$x_{i.}\sum_{j=i}^{p} n_{ij}y_j$
(1,15]	4	2	0	6	8	48	384	64
(15,25]	1	4	2	7	20	140	2800	340
(25,30]	0	3	5	8	27.5	220	6050	715
$n_{.j}$	5	9	7	21		408	9234	1119
$c_{.j}$	1	2	4					
$n_{.j}c_{.j}$	5	18	28	51				
$n_{.j}c_{.j}^2$	5	36	112	153				

Apartado 1

Calculamos la recta de regresión de Y sobre X:

• Para ello primero calculamos las medias de X e Y:

$\bar{x} = \frac{48+140+220}{21} = \frac{408}{21} = 19{,}428$ °C

$\bar{y} = \frac{5+18+28}{21} = \frac{51}{21} = 2{,}428$ ml de Hg

• Tenemos que calcular también las varianzas y la covarianza:

$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2 - \bar{x}^2 = \frac{9234}{21} - 19{,}4286^2 = 62{,}2438\ C^2$

$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2 - \bar{y}^2 = \frac{153}{21} - 2{,}428^2 = 1{,}3876\ (mldeHg)^2$

$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_iy_j - \bar{x}\bar{y} = \frac{1}{21}{,}1119 - 47{,}1843 = 6{,}1014$

•Ahora que tenemos todos los datos necesarios podemos ya calcular los coeficientes de la recta de regresión :

$y = ax + b$

$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{6{,}1014}{62{,}2438} = 0,098$

$b = \bar{y} - a.\bar{x} = 2,4286 - 0,098{,}19,4286 = 0,5241$

$y = 0,098x + 0,5241$

• Por ultimo calculamos el coeficiente de correlación lineal:

$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = \frac{7{,}1667^2}{72{,}5568{,}1{,}583} = 0,431055$

El resultado que hemos obtenido nos indica que la recta de regresión de Y sobre X nos da una información de menos del 45 % de la variabilidad de Y, luego no seria buena idea suponer una relación lineal entre las variables ya que la bondad de la función es baja. Pero como podemos ver, el coeficiente de correlación lineal $r = \sqrt{r^2} = 0,6565$ nos indica que hay un significante grado de correlación lineal directa entre las variables.

Problema 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	$n_{i.}$
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

Y es independiente estadísticamente de X, si $f_j^i \equiv f_{j/i}$, es decir, si $\frac{n_{1j}}{n1.} = \frac{n2j}{n2.} = ... = \frac{n_{kj}}{n_{k.}}\ \forall j = 1, 2, ..., p$. En este caso vemos que se cumple. Por ejemplo, para j=1, tenemos: $\frac{2}{30} = \frac{1}{15} = \frac{3}{45} = \frac{4}{60}$, y pasa lo mismo con para j=2,3,4,5.
Así vemos que Y es independiente de X y en este caso, X también es independiente de Y. Aplicando el mismo criterio: $f_i^j \equiv f_{i/j}$, es decir, $\frac{n_{i1}}{n,1} = ... = \frac{n_{ip}}{n.p}$. Tenemos para i=1: $\frac{2}{10} = \frac{4}{20} = \frac{6}{30} = \frac{10}{50} = \frac{8}{40}$, y pasa igual para i=2,3,4.

Como las variables X e Y son independientes, no tiene sentido estudiar la curva de regresión, y sabemos que la covarianza va a ser 0: $\sigma_{xy} = 0$.

X/Y	1	2	3	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
-1	0	1	0	1	-1	1
0	1	0	1	2	0	0
1	0	1	0	1	1	1
$n_{.j}$	1	2	1	4	0	2
$n_{.j}y_j$	1	4	3	8		
$n_{.j}y_j^2 - \bar{y}^2$	1	8	9	18		

Sabemos que estas variables no pueden ser independientes ya que si hay algún $n_{ij} = 0$, no se va a dar la igualdad $\frac{n1j}{n1.} = ... = \frac{n_{kj}}{nk.}$ a no ser que $n_{ij} = 0\ \forall i, j$. Esto nos lleva a una contradicción, pues eso sería una variable que no ha presentado distribución de frecuencias.
Sabemos también que X no depende funcionalmente de Y, pues tenemos que $n_{i2} \neq 0$ para más de un i. Por esto mismo, sabemos que Y no depende funcionalmente de X, pues se da $n_{2j} \neq 0$ para más de un j.

Calculamos la covarianza: σ_{xy}.

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

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

$\sigma_x^2 = m_{20} - m_{10}^2 = m_{20} = \frac{2}{4} = 0,5$.
$\sigma_y^2 = m_{02} - m_{01}^2 = \frac{18}{4} - 2^2 = 4,5 - 4 = 0,5$.

$$\sigma_{xy} = m_{11} - m_{10}m_{01} = m_{11} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} x_i y_j n_{ij} = \frac{1*2*-1 + 1*2*1}{4} = \frac{-2+2}{4} = 0$$

Calculamos los puntos por los que pasa la curva de regresión de X/Y: $(\bar{x_j}, y_j)$

$\bar{x_1} = \frac{0}{1} = 0$
$\bar{x_2} = \frac{1*-1+1*1}{2} = 0$
$\bar{x_3} = \frac{0}{1} = 0$
Por lo tanto, los puntos son: (0,1) (0,2) (0,3)
Calculamos ahora los puntos por los que pasa la curva de regresión de Y/X: $(x_i, \bar{y_i})$
$\bar{y_1} = \frac{1*2}{1} = 2$
$\bar{y_2} = \frac{1*1+1*3}{2} = 2$
$\bar{y_3} = \frac{1*2}{1} = 2$
Por lo tanto, los puntos son: (-1,2) (0,2) (1,2)

Problema 6

Dada la siguiente distribución bidimensional:

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

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

X/Y	1	2	3	4	$n_i.$	$n_i. \cdot x_i$	$n_i. \cdot \mathrm{x}_i^2$	$x_i \sum_{j=1}^{p} n_{ij} y_j$
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
$n._j \cdot y_j$	7	8	12	20	47			
$n._j \cdot y_j^2$	7	16	36	80	139			

Apartado 1

¿Son estadísticamente independientes X e Y?

Para que se diera la independencia estadística, se tendría que cumplir que las frecuencias de X condicionadas a Y son las mismas para valor de x_i.

$$\frac{n_{i1}}{n._1} = \frac{n_{i2}}{n._2} = \cdots = \frac{n_{ij}}{n._j} = \cdots = \frac{n_{ip}}{n._p} \qquad \forall i = 1, 2, \ldots, k.$$

Pero como en algunos casos $n_{ij} = 0$, no se puede dar la independencia ya que no todos son nulos.

Apartado 2

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

La curva de regresión tipo 1 de X sobre Y vendrá dada por los puntos $(x_i, \overline{y_i}), i = 1, \dots k$.

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

La curva de regresión tipo 1 de Y sobre X vendrá dada por los puntos $(y_j, \overline{x_j}), i = 1, \dots k$.

$$\{(1, 14{,}571), (2, 10{,}5), (3, 12), (4, 12{,}8)\}$$

Apartado 3

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

Con los datos de la tabla:

$$\overline{x} = \frac{256}{20} = 12{,}8 \qquad \overline{y} = \frac{27}{20} = 2{,}35$$

$$\sigma_x^2 = m_{20} - m_{10}^2 = 4{,}16 \qquad \sigma_y^2 = m_{02} - m_{01}^2 = 1{,}4275$$

Calculemos la varianza de los residuos para Y/X:

$$\sigma_{ry}^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} \cdot (y_j - f(x_i))^2 = \frac{13{,}25}{20} = 0{,}6625$$

Para X/Y:

$$\sigma_{ry}^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} \cdot (x_i - f(y_j))^2 = \frac{37{,}51}{20} = 1{,}875$$

Podemos ahora calcular el coeficiente de correlación:

$$\eta_{\frac{y}{x}}^2 = 1 - \frac{\sigma_{ry}}{\sigma_y^2} = 0{,}536 \qquad \eta_{\frac{x}{y}}^2 = 1 - \frac{\sigma_{rx}}{\sigma_x^2} = 0{,}549$$

Con lo que podemos ver que la bondad está por debajo del 55 %, por lo que la curva no es demasiado precisa.

Apartado 4

¿Están X e Y correladas linealmente? Dar las expresiones de las rectas de regresión.

Calculemos el coeficiente de correlación lineal:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = -0{,}32$$

Podemos observar una baja correlación lineal.

Calculamos las rectas:

Y sobre X :

$$y = ax + b$$

$$a = \frac{\sigma_{xy}}{\sigma_x^2} = -0{,}187 \qquad b = \overline{y} - a\overline{x} = 4{,}75$$

Recta de Y sobre X $\equiv y = -0{,}187x + 4{,}75$.

X sobre Y :

$$x = ay + b$$

$$a = \frac{\sigma_{xy}}{\sigma_y^2} = -0{,}55 \qquad b = \overline{x} - a\overline{y} = 14$$

Recta de X sobre Y $\equiv x = -0{,}55y + 14$

Problema 7

Para cada una de las distribuciones: ¿Dependen funcionalmente X de Y o Y de X? Calcular las curvas de regresión y comentar los resultados.

Distribución A:

X/Y	10	15	20	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2 - \bar{x}^2$
1	0	2	0	2	2	4,084
2	1	0	0	1	2	0,184
3	0	0	3	3	9	0,978
4	0	1	0	1	4	2,468
$n_{.j}$	1	3	3	7	17	7,714
$n_{.j}y_j$	10	45	60	115		
$n_{.j}y_j^2 - \bar{y}^2$	41,332	6,126	38,256	85,714		

Y depende funcionalmente de X ya que para cada x_i hay un solo $y_j \neq 0$. Sin embargo, X no depende funcionalmente de Y.

Calculamos la curva de regresión de X/Y:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i = \frac{17}{7} = 2,429$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j = \frac{115}{7} = 16,429$$

$$\sigma_y^2 = m_{02} - m_{01}^2 = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j^2 - \bar{y}^2 = \frac{85,714}{7} = 12,245$$

$\sigma_{xy} = m_{11} - m_{10}m_{01} = \frac{2*1*15+1*2*10+3*20*3+1*15*4}{7} - 2,429 * 16,429 = 1,523$

Hallamos la curva de regresión de X/Y: x=ay+b.
$a = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{1,523}{12,245} = 0,124$
$b = \bar{x} - a\bar{y} = 2,429 - 0,124 * 16,429 = 0,392$
$x = 0,125y + 0,375$

Distribución B:

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

X depende funcionalmente de Y y viceversa ya que para cada y_j hay un solo $x_i \neq 0$ y para cada x_i un solo $y_j \neq 0$

Distribución C:

X/Y	10	15	20	25	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2 - \bar{x}^2$
1	0	3	0	1	4	4	2,039
2	0	0	1	0	1	2	0,082
3	2	0	0	0	2	6	3,308
$n_{.j}$	2	3	1	1	7	12	5,429
$n_{.j}y_j$	20	45	20	25	110		

X depende funcionalmente de Y, pues para cada y_j hay un solo $x_i \neq 0$. Calculamos la curva de regresión de Y/X:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i.} x_i = \frac{12}{7} = 1,714$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j = \frac{110}{7} = 15,714$$

$$\sigma_x^2 = m_{20} - m_{10}^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.} x_i^2 - \bar{x}^2 = \frac{5,429}{7} = 0,776$$

$\sigma_{xy} = m_{11} - m_{10}m_{01} = \frac{3*15*1+1*1*25+1*2*20+2*3*10}{7} - 1,714 * 15,714 = 24,286 - 26,934 = -2,648$

Hallamos la curva de regresión de X/Y: =ay+b.
$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-2,648}{0,776c} = -3,412$
$b = \bar{y} - a\bar{x} = 15,714 + 3,412 * 1,714 = 21,562$
$y = -3,412x + 21,562$

Problema 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:

$X\backslash Y$	1	2	3	4
1	1	2	0	0
2	1	2	3	1
3	0	1	2	6
4	0	0	2	3

1. Determinar las rectas de regresión.
2. ¿Es apropiado suponer que existe una relación lineal entre las variables?
3. Predecir, a partir de los resultados, el número de balanzas que puede esperarse en un puesto con seis dependientes. ¿Es fiable esta predicción?

En una población de tamaño n = 24 se ha observado dos variables estadísticas, X = número de balanzas e Y = número de dependientes, las cuales han presentado k = 4, p = 4 modalidades distintas, con distribución de frecuencia conjunta (x_i, y_j), n_{ij} $i = 1, ..., 3$ $j = 1, .., 3$

• Empezamos rellenando nuestra tabla:

X Y	1	2	3	4	$n_{i.}$	$n_{i.}x_{i.}$	$n_{i.}x_{i.}^2$	$x_{i.}\sum_{j=i}^{p} n_{ij}y_j$
1	1	2	0	0	3	3	3	5
2	1	2	3	1	7	14	28	36
3	0	1	2	6	9	27	81	96
4	0	0	2	3	5	20	80	72
$n_{.j}$	2	5	7	10	24	64	192	209
$n_{.j}y_{.j}$	2	10	21	40	73			
$n_{.j}y_{.j}^2$	2	20	63	160	245			

Apartado 1

• Para ello primero calculamos las medias de X e Y:

$\bar{x} = \frac{3+14+27+20}{24} = \frac{64}{24} = 2,6667$ balanzas

$\bar{y} = \frac{2+10+21+40}{24} = \frac{73}{24} = 3,0417$ dependientes

• Tenemos que calcular también las varianzas y la covarianza:

$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2 - \bar{x}^2 = \frac{192}{24} - 2,6667^2 = 0,8887\ balanzas^2$

$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2 - \bar{y}^2 = \frac{245}{24} - 3,0417^2 = 0,9564\ dependientes^2$

$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_iy_j - \bar{x}\bar{y} = \frac{1}{24},209 - 8,113 = 0,5970$

•Ahora que tenemos todos los datos necesarios podemos ya calcular los coeficientes de las rectas de regresión :

Para Y / X:
$y = ax + b$

$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{0{,}5970}{0{,}8887} = 0,6718$

$b = \bar{y} - a.\bar{x} = 3,0417 - \frac{0{,}5970}{0{,}8887}{,}2,6667 = 1,2502$

La recta de regresión de Y sobre X es $y = 0,6718 + 1,2502$

Para X / Y:
$x = ay + b$

$a = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{0{,}5970}{0{,}9564} = 0,6242$

$b = \bar{x} - a.\bar{y} = 2,6667 - \frac{0{,}5970}{0{,}9564}{,}3,0417 = 0,768$

La recta de regresión de X sobre Y es $x = 0,6242y + 0,768$

Apartado 2

• Para ello tenemos que calcular el coeficiente de correlación lineal:

$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{0{,}5970^2}{0{,}8887{,}0{,}9564} = 0,4193$

•El resultado que hemos obtenido nos indica que la recta de regresión de Y sobre X nos da una información de menos del 42 % de la variabilidad de Y, luego no seria buena idea suponer una relación lineal entre las variables ya que la bondad de la función es relativamente baja. Tampoco, el coeficiente de correlación lineal $r = \sqrt{r^2} = 0,6475$ nos indica que existe una buena correlación lineal directa entre las variables.

Apartado 3

$x = 0,6242y + 0,786 = 0,6242{,}6 + 0,786 = 4,5312 \approx 5$ balanzas. Si bien la recta de regresión proporcionaría una estimación de la cantidad de balanzas que cabría esperar para 6 dependiente, el resultado no sería fiable, puesto que la recta de regresión tendría únicamente validez para el intervalo muestral estudiado (en este caso, de 1 a 4 dependientes). Más allá de estos límites, no se sabe si la nube de puntos seguiría comportándose como los datos estudiados.

Problema 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:

$X \backslash Y$	$(10, 20]$	$(20, 25]$	$(25, 30]$	$(30, 35]$	$(35, 40]$
$(15, 18]$	3	2	3	0	0
$(18, 21]$	0	4	2	2	0
$(21, 24]$	0	7	10	6	1
$(24, 27]$	0	0	2	5	3

Apartado 1

Estudiar la interdependencia lineal entre ambas variables.

En una población de n familias se han observado la edad de marido (variable X) y la edad de la mujer (variable Y), presentando 4 modalidades para la variable X y 5 modalidades para la variable Y. La distribución conjunta bidimensional aparece en la siguiente tabla:

	c_j	15	22,5	27,5	32,5	37,5					
c_i	$X \backslash Y$	(10,20]	(20,25]	(25,30]	(30,35]	(35,40]	$n_{i.}$	$n_{i.}c_i$	$n_{i.}c_i^2$	$\sum n_{ij}c_j$	$c_i \sum n_{ij}c_j$
16,5	(15,18]	3	2	3	0	0	8	132	2178	172,5	2846,25
19,5	(18,21]	0	4	2	2	0	8	156	3042	210	4095
22,5	(21,24]	0	7	10	6	1	24	540	12150	665	14962,5
25,5	(24,27]	0	0	2	5	3	10	255	6502,5	330	8415
	$n_{.j}$	3	13	17	13	4	50	1083	23872,5		30318,75
	$n_{.j}c_j$	45	292,5	467,5	422,5	150	1377,5				
	$n_{.j}c_j^2$	675	6581,25	12856,25	13731,25	5625	39468,75				

Puesto que cada modalidad se presenta en forma de intervalo, conviene calcular la marca de clase para realizar cálculos. Para determinar la recta de regresión X = f(Y), hemos de realizar los siguientes cálculos:

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

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j$$

$$m_{20} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2$$

$$m_{02} = \frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2$$

$$m_{11} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_iy_j$$

$$\sigma_x^2 = m_{20} - \overline{x}^2$$

$$\sigma_y^2 = m_{02} - \overline{y}^2$$

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

Teniendo en cuenta los datos de la tabla presentada, tenemos los siguientes resultados: $\overline{x} = 21,66$ años, $\overline{y} = 27,55$ años, $m_{20} = 477,45$ años^2, $m_{02} = 789,375$ años^2, $m_{11} = 606,375$ años^2 $\sigma_x^2 = 8,2944$ años^2, $\sigma_y^2 = 30,3725$ años^2, $\sigma_{xy} = 9,642$ años^2.

Para la recta $x = ay + b$ que minimice los cuadrados de los errores, hemos de calcular los siguientes valores:

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

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

Sustituyendo en los datos obtenidos, obtenemos: $a = 0,3174$ y $b = 12,9146$ años. Por tanto, la recta de regresión es:

$$x = 0,3174y + 12,9146$$

La interdependencia de ambas variables viene determinada por el coeficiente $R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$. En este caso, su

valor es $R^2 = 0,369$. Ello representa que el $36,9\,\%$ de la variabilidad de la edad en los varones viene explicado por la recta de regresión, mientras que el $63,1\,\%$ restante viene determinado por otras causas.

Problema 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 ecuación es para X/Y y la segunda para Y/X:

$y = \frac{-1}{5}x + \frac{2}{5}$

$x = -4y + 1$

•Las pendientes tienen que tener el mismo signo o sino no serían rectas de regresión, en este caso las dos pendientes son negativos.
Sabemos que :

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

Luego :

$a.a' = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = r^2 = \frac{4}{5}$

Entonces esta bien porque sabemos que $0 <= r^2 <= 1$:

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

Problema 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.

Apartado 1

Calculamos la media de X con los datos que nos dan:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i.} x_i = \frac{45}{9} = 5$$

Además, como sabemos que se cumple que $\overline{y} = a\overline{x} + b$, obtenemos que:

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

Ahora empezamos a calcular la recta de Y sobre X. Para ello debemos tener en cuenta que:

$$x = a'y + b' \qquad a' = \frac{\sigma_{xy}}{\sigma_y^2} \qquad b' = \overline{x} - a'\overline{y}$$

Ahora haremos los cálculos necesarios:

$$\sigma_y^2 = m_{11} - m_{10}^2 = \frac{3240}{9} - \overline{y}^2 = 335 \qquad \sigma_x^2 = m_{11} - m_{10}^2 = \frac{279}{9} - \overline{y}^2 = 6$$

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

Terminamos de calcular la recta:

$$a' = \frac{\sigma_{xy}}{\sigma_y^2} = 0{,}0896 \qquad b' = \overline{x} - a'\overline{y} = 4{,}552$$

Luego,

$$x = a'y + b' \Rightarrow x = 0{,}0896y + 4{,}552$$

Ahora veremos la bondad de los ajustes lineales. Para ello calculamos el coeficiente de determinación lineal:

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

Esto significa que la recta que hemos calculado explica el 44.8 % de nuestra distribución, por lo que podemos decir que no es bueno suponer una relación lineal.

Problema 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 estos datos se han obtenido las siguientes medidas:

$$\sum y_i = 219.719 \quad \sum y_i^2 = 2396.504 \quad \sum x_i y_i = 349.486$$

$$\sum x_i = 31.470 \quad \sum x_i^2 = 51.075 \quad \sum x_i^2 y_i = 633.993$$

$$\sum x_i^4 = 182.977 \quad \sum x_i^3 = 93.6$$

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).

1. 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?
2. Ajustar un modelo del tipo $Y = a+bX+cX^2$. ¿Qué consumo total se estimaría para el mismo programa de vuelos del apartado a)?
3. ¿Cuál de los dos modelos se ajusta mejor? Razonar la respuesta.

En el desarrollo de este ejercicio se va a considerar que a cada tiempo de trayecto de vuelo del avión va asociado una única cantidad de combustible asociada a dicho trayecto (se despreciarán las ligeras variaciones que se puedan dar tanto en el despegue como en el descenso de la aeronave). Por tanto, se presupone que $n_i = 1$, $\forall i = 1, 2, 3, ..., n$.

Apartado 1

Los cálculos que se han de hacer serían:

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

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j$$

$$m_{20} = \frac{1}{n}\sum_{i=1}^{k} n_{i.} x_i^2$$

$$m_{02} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j^2$$

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

$$\sigma_x^2 = m_{20} - \overline{x}^2$$

$$\sigma_y^2 = m_{02} - \overline{y}^2$$

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

Pero como se verifica que $n_i = 1$, $\forall i = 1, 2, 3, ..., n$, bastaría con calcular:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} y_j$$

$$m_{20} = \frac{1}{n}\sum_{i=1}^{k} x_i^2$$

$$m_{02} = \frac{1}{n}\sum_{j=1}^{p} y_j^2$$

$$m_{11} = \frac{1}{n}\sum_{i=1}^{k} x_i y_i$$

Así llegamos a las siguientes conclusiones: $\overline{x} = 1,31125$ h, $\overline{y} = 9,15496$ miles de libras, $m_{20} = 2,128125$ h^2, $m_{02} = 99,8543$ miles de libras^2, $m_{11} = 14,5619$, $\sigma_x^2 = 0,4087$ h^2, $\sigma_y^2 = 16,041$ miles de libras^2, $\sigma_{xy} = 2,5575$ años^2.

Para la recta $x = ay + b$ que minimice los cuadrados de los errores, hemos de calcular los siguientes valores:

$$a = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$b = \overline{y} - a\overline{x}$$

Sustituyendo en los datos obtenidos, obtenemos: $a = 6,2576$ miles de libras/h y $b = 0,9497$ miles de libras. Por tanto, la recta de regresión es:

$$y(x) = 6,2576x + 0,9497$$

donde x se expresa en horas e y se expresa en miles de libras de combustible.

Para determinar si el ajuste realizado es fiable, hemos de calcular $R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$. En este caso, su valor es $R^2 = 0,997689$. Ello representa que la estimación es muy fiable, siendo correcta en más de un 99 % de las situaciones.

El itinerario se presenta en la siguiente tabla:

Tiempo (en h)	Combustible / trayecto (miles de libras)	n.º trayectos	Combustible (miles de libras)
0,5	4,0785	100	407,85
1	7,2073	200	1441,46
2	13,4649	100	1346,49
			3195,8

Se emplea un total de 3195,8 miles de libras de combustible para realizar el itinerario.

Apartado 2

Con la filosofía de minimizar el cuadrado de los errores para los parámetros de la curva $y = a + bx + cx^2$, tenemos que para calcularlos se ha de resolver el siguiente sistema de ecuaciones:

$$a + m_{10}b + m_{20}c = m_{01}$$
$$m_{10}a + m_{20}b + m_{30}c = m_{11}$$
$$`m_{20}a + m_{30}b + m_{40}c = m_{21}$$

Teniendo en cuenta $m_{30} = 3,9$, $m_{40} = 7,624$, $m_{21} = 26,4164$, además de los resultados ya calculados previamente, obtenemos:

$$a = 0,8086$$
$$b = 6,53$$
$$c = -0,1006$$

La ecuación de la curva quedaría $y = 0,8086 + 6,53x - 0,1006x^2$. La estimación que realizaría esta regresión al itinerario sería:

Tiempo (en h)	Combustible / trayecto (miles de libras)	n.º de trayectos	Combustible (miles de libras)
0,5	4,04845	100	404,845
1	7,238	200	1447,6
2	13,4662	100	1346,62
			3199,065

Apartado 3

No se proporciona suficiente información en el enunciado para determinar cuál de las dos regresiones se ajusta mejor a los datos experimentales. En principio, la regresión lineal minimiza bien los cuadrados de los

errores (ajuste casi perfecto, con dependencia lineal), mientras que de la parábola no se puede determinar su coeficiente de correlación.

Problema 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

Como tenemos una hipérbola equilátera, comenzamos haciendo un cambio de variable;

$$Z = \frac{1}{X}$$

Z/Y	50	90	160	180	$n_i.$	$n_i.x_i$	$n_i.x_i^2$	$x_i \sum_{j=1}^{p} n_{ij}y_j$
0.1	1	0	0	0	1	0.1	0.01	5
0.08	0	1	0	0	1	0.08	0.0064	7.2
0.05	0	0	1	0	1	0.05	0.0025	8
0.05	0	0	0	1	1	0.04	0.0016	7.2
$n._j$	1	1	1	1	4	0.27	0.0205	27.4
$n._j yj$	50	90	160	180	480			
$n._j y_j^2$	2500	8100	25600	32400	68600			

$$\overline{x} = \frac{0{,}27}{4} = 0{,}0675 \qquad \overline{y} = \frac{480}{4} = 120$$

$$\sigma_x^2 = m_{20} - m_{10}^2 = 0{,}0005687 \qquad \sigma_y^2 = m_{02} - m_{01}^2 = 2750$$

$$\frac{1}{n}\sum_{i=1}^{k} x_i \sum_{j=1}^{p} n_{ij}y_j = -1{,}25$$

$$a = \frac{\sigma_{xy}}{\sigma_x^2} = -2197{,}802 \qquad b = \overline{y} - a\overline{x} = 268{,}352$$

Por tanto tenemos que, deshaciendo el cambio de variable, la recta de regresión de Y sobre X es

$$y = -2197{,}802 \cdot \frac{1}{x} + 268{,}352$$

Para calcular la bondad tenemos que obtener la varianza de los residuos

$$\sigma_{ry}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\cdot(y_j - f(x_i))^2 = \frac{13{,}989}{4} = 2{,}747$$

Por tanto,

$$\eta_{\frac{y}{x}}^2 = 1 - \frac{\sigma_{ry}}{\sigma_y^2} == 0{,}9990$$

Por lo que observamos la distribución está definida en un 99 % por la recta. Esto se debe a que existe una dependencia funcional ya que en cada fila existe un único elemento no nulo.

Problema 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:

1. Relación lineal.
2. Hipérbola equilátera.
3. Curva potencial.
4. Curva exponencial.

¿Qué ajuste es más adecuado?

Apartado 1

Entre 6 familias que asisten a espectáculos se han observados, de cada familia, la renta familiar (la variable X, en cientos de euros) y el gasto general a los espectáculos (la variable Y, en euros). Ambas variables han presentado 6 modalidades, recogidos en la siguiente tabla bidimensional:

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

Respecto a cada ajuste, se han de realizar las siguientes operaciones:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} y_j$$

$$m_{20} = \frac{1}{n}\sum_{i=1}^{k} x_i^2$$

$$m_{02} = \frac{1}{n}\sum_{j=1}^{p} y_j^2$$

$$m_{11} = \frac{1}{n}\sum_{i=1}^{k} x_i y_i$$

$$\sigma_x^2 = m_{20} - \overline{x}^2$$

$$\sigma_y^2 = m_{02} - \overline{y}^2$$

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

Para el cálculo de la razón de correlación de $Y\backslash X$, como sabemos que a cada modalidad y_i le corresponde un único x_i (y viceversa), deducimos que, en este caso:

$$\sigma_{exp}^2 = \sum_i \sum_i f_{ij}(y^* - \overline{y})^2 = \frac{1}{n}\sum_i y_i^2 + \overline{y}^2 - 2\overline{y}^2 \sum_i y_i$$

$$\sigma_{res}^2 = \sum_i \sum_i f_{ij}(y_i^* - y_i)^2$$

De forma general, tenemos:

$$\eta_{Y\backslash X}^2 = \frac{\sigma_{exp}^2}{\sigma_y^2} = 1 - \frac{\sigma_{des}^2}{\sigma y^2}$$

Apartado 2

La tabla obtenida sería la siguiente:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
9	30	81	900	270
10	50	100	2500	500
12	70	144	4900	840
15	80	225	6400	1200
22	120	484	14400	2640
32	140	1024	19600	4480
100	490	2058	48700	9930

Se pretende buscar la recta $y = ax + b$ que minimice los cuadrados de los errores observados (y, por tanto, mejor se ajuste a los datos observados). Derivando parcialmente respecto a a y b, obtenemos:

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

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

Sustituyendo por los valores obtenidos, tenemos que:

$$y = 4,5060x + 6,5673$$

Nótese $\eta^2_{Y\setminus X} = R^2 = \dfrac{\sigma^2_{xy}}{\sigma^2_x\sigma^2_y} = 0,915$.

Apartado 3

Linealizamos empleando el cambio de variable $z = \frac{1}{x}$ y obtenemos la siguiente tabla:

z_i	y_i	z_i^2	y_i^2	z_iy_i
0,1111	30	0,0123	900	3,3333
0,1000	50	0,0100	2.500	5,0000
0,0833	70	0,0069	4.900	5,8333
0,0667	80	0,0044	6.400	5,3333
0,0455	120	0,0021	14.400	5,4545
0,0313	140	0,0010	19.600	4,3750
0,4378	490,0000	0,0368	48.700,0000	29,3295

Se pretende buscar la curva $y = \frac{a}{x} + b$ que minimice los cuadrados de los errores observados (y, por tanto, mejor se ajuste a los datos observados). De la tabla podemos obtener: $\overline{z} = 0,07297$ cientos de euros, $\overline{y} = 81,6667$ euros, $m_{20} = 0,0061$ cientos de euros^2, $m_{02} = 8116,6667$ euros^2, $\sigma_z^2 = 0,0008$ cientos de euros^2, $\sigma_y^2 = 1447,2168$ euros^2, $m_{11} = 4,8883$, $\sigma_{xy} = -1,0709$. Derivamos parcialmente respecto a a y b y obtenemos:

$$y = \frac{-1338,625}{x} + 179,3462$$

Para obtener la razón de correlación de $Y\setminus X$, realizamos la siguiente tabla:

x_i	y_i	y_i*	y_i*^2
9	30	30,6101	936,9775
10	50	45,4837	2.068,7670
12	70	67,7941	4.596,0423
15	80	90,1045	8.118,8269
22	120	118,4996	14.042,1574
32	140	137,5142	18.910,1466
16,6667	81,6667	490,0062	48.672,9177

Notemos que se verifica que $\sigma^2_{exp} = 1442,5398$. Por tanto, tenemos que $\eta^2_{Y\setminus X} = 0,9968$ (se trata de un buen ajuste).

Apartado 4

Se pretende buscar la curva $y = ax^b$ que minimice los cuadrados de los errores observados (y, por tanto, mejor se ajuste a los datos observados). Linealizamos tomando logaritmos en ambos miembros, de donde $\log y = \log a + b \log x$. La tabla sobre el que hemos de operar es la siguiente:

$\log x_i$	$\log y_i$	$\log {x_i}^2$	$\log {y_i}^2$	$\log x_i \log y_i$
0,9542	1,4771	0,9106	2,1819	1,4095
1,0000	1,6990	1,0000	2,8865	1,6990
1,0792	1,8451	1,1646	3,4044	1,9912
1,1761	1,9031	1,3832	3,6218	2,2382
1,3424	2,0792	1,8021	4,3230	2,7911
1,5051	2,1461	2,2655	4,6059	3,2302
7,0571	11,1496	8,5260	21,0234	13,3593

Datos que se pueden obtener de esta tabla:

$\overline{x}$	$\overline{y}$	m_{20}	m_{02}	m_{11}	$\sigma {x}^2$	$\sigma {y}^2$	σxy	b	$\log a$
1,1762	1,8583	1,4210	3,5039	2,2265	0,0376	0,0507	0,0409	1,0877	0,5789

Minimizando por mínimos cuadrados, tenemos que: $\log a = 0,5789$ y $b = 1,0877$. Por tanto, la curva potencial que mejor se ajusta a los datos experimentales es:

$$y = 3,7923x^{1,0877}$$

Para obtener la razón de correlación de $Y\setminus X$, realizamos la siguiente tabla:

x_i	y_i	y_i*	$(y_i^* - y_i)^2$
9	30	41,3861	129,6428
10	50	46,4116	12,8768
12	70	56,5919	179,7766
15	80	72,1384	61,8040
22	120	109,4186	111,9656
32	140	164,4733	598,9410
		490,4199	182,5011

Deducimos que se verifica $\sigma^2_{res} = 182,5011$. Por tanto, tenemos que $\eta^2_{Y\setminus X} = 1 - 0,1261 = 0,8739$.

Apartado 5

Se pretende buscar la curva $y = ab^x$ que minimice los cuadrados de los errores observados (y, por tanto, mejor se ajuste a los datos observados). Linealizamos aplicando tomando logaritmos en ambos miembros, de donde $\log y = \log a + x \log b$. La tabla sobre el que hemos de operar es la siguiente:

x_i	$\log y_i$	x_i^2	$\log {y_i}^2$	$x_i \log y_i$
9,0000	1,4771	81,0000	2,1819	13,2941
10,0000	1,6990	100,0000	2,8865	16,9897
12,0000	1,8451	144,0000	3,4044	22,1412
15,0000	1,9031	225,0000	3,6218	28,5463
22,0000	2,0792	484,0000	4,3230	45,7420
32,0000	2,1461	1.024,0000	4,6059	68,6761
100,0000	11,1496	2.058,0000	21,0234	195,3894

$\overline{x}$	$\overline{y}$	m_{20}	m_{02}	m_{11}	σx^2	σy^2	σxy	$\log b$	$\log a$
16,6667	1,8583	343,0000	3,5039	32,5649	65,2222	0,0507	1,5938	0,0244	1,4510

Tras aplicar mínimos cuadrados, obtenemos que $\log a = 1,4510$ y $\log b = 0,0244$. La curva que buscamos es:

$$y = 28,2488 \cdot 1,0578^x$$

Calculemos la bondad del ajuste. Nos atenemos a la siguiente tabla:

x_i	y_i	y_i*	$(y_i^* - y_i)^2$
9	30	46,8723	284,673912469069
10	50	49,5853	0,171982508254753
12	70	55,4915	210,49670214935
15	80	65,6957	204,613807613501
22	120	97,4079	510,402889459716
32	140	170,9866	960,16629128634
		486,0392	2170,52558548623

Deducimos que $\sigma^2_{res} = 361,753$. Por tanto, tenemos que $\eta^2_{Y\setminus X} = 1 - 0,2499 = 0,7501$.

Apartado 6

De las cuatro regresiones presentada, aquella que presenta mejor ajuste sería la hipérbola equilátera, pues un mayor porcentaje de datos observados son explicados a partir de la curva de regresión.