PROBLEMAS DE ESTADISTICA

5. REGRESION LINEAL

5–1. Tomamos un grupo de observación de 10 muestras de agua en las que medimos la concentración de nitratos en mg/l y la temperatura en ^{0}C, obteniéndose los resultados:

mg/l	70	65	80	60	75	85	70	65	80	85
^{0}C	36.5	36.5	37.0	36.0	37.0	37.5	37.0	36.0	37.5	37.0

Hallar las ecuaciones lineales que representan este fenómeno y hacer su representación gráfica. A la vista de dichas ecuaciones, ¿qué temperatura media tendría una muestra con concentración de 72 mg/l?

5–2. En la tabla siguiente se dan, para diferentes países, datos sobre la esperanza de vida (en años) (V), el número medio de habitantes por cada aparato de televisión (N_T), y el número medio de habitantes por cada médico (N_M).

País	V	N_T	N_M	País	V	N_T	N_M
Alemania	76.0	2.6	346	Indonesia	61.0	24.0	7427
Argentina	70.5	4.0	370	Irán	64.5	23.0	2992
Bangla Desh	53.5	315.0	6166	Italia	78.5	3.8	233
Birmania	54.5	592.0	3485	Japón	79.0	1.8	609
Brasil	65.0	4.0	684	Kenia	61.0	96.0	7615
Canada	76.5	1.7	449	Marruecos	64.5	21.0	4873
Corea del Norte	70.0	90.0	370	Méjico	72.0	6.6	600
Corea del Sur	70.0	4.9	1066	Pakistán	56.5	73.0	2364
China	70.0	8.0	643	Perú	64.5	14.0	1016
Egipto	60.5	15.0	616	Polonia	73.0	3.9	480
España	78.5	2.6	275	Rumanía	72.0	6.0	559
Estados Unidos	75.5	1.3	404	Rusia	69.0	3.2	259
Etiopía	51.5	503.0	36660	Sudán	53.0	23.0	12550
Francia	78.0	2.6	403	Tailandia	68.5	11.0	4883
Gran Bretaña	76.0	3.0	611	Turquía	70.0	5.0	1189
India	57.5	44.0	2471	Vietnan	65.0	29.0	3096

a) Calcular la recta de regresión lineal entre la esperanza de vida y el número de habitantes por televisión (se sugiere convertir estos datos a escala logarítmica), es decir: $V = a + b \log N_T$. Calcular el coeficiente de correlación.

b) Calcular la recta de regresión lineal entre la esperanza de vida y el número de habitantes por médico (convertir también estos datos a escala logarítmica), es decir: $V = a + b \log N_M$. Calcular el coeficiente de correlación.

c) Interpretar y discutir los resultados anteriores.

5–3. Se ha ajustado una recta de regresión de la variable y sobre x, resultando:

$$y + 5 = 0.6x$$

Teniendo en cuenta que, para un grupo de 20 observaciones, las variables x e y presentan, respectivamente, medias aritméticas de 50 y 25, y varianzas de 0.64 y 0.36, calcular el coeficiente de correlación.

5–4. En 1929 Hubble presentó las primeras medidas sobre distancias y velocidades radiales de galaxias, demostrando que había una correlación positiva entre ambas magnitudes y que, por lo tanto, el Universo se está expandiendo. La siguiente tabla lista parte de los datos originales de Hubble:

d (Mpc)	0.5	0.5	0.8	0.9	0.9	1.0	1.1	1.1	1.4	1.7	2.0	2.0
v (km/s)	290	270	300	650	150	920	450	500	500	960	800	1090

a) Calcular la media y la varianza de las variables distancia y velocidad. Calcular asímismo la covarianza.

b) Usando los resultados del apartado anterior, calcular y representar la recta de regresión de v sobre d (En este caso, al coeficiente de regresión se le conoce como constante de Hubble).

c) A partir de los resultados anteriores, calcular el coeficiente de correlación lineal y la desviación típica residual. Discutir los resultados.

5–5. La tabla siguiente da la temperatura media y la precipitación en una ciudad durante los meses de julio de los años 1975–1984. Hallar el coeficiente de correlación y su nivel de significación. Interpretar el resultado.

Año	T(^{0}C)	P (l)
1975	25.6	158.2
1976	22.1	92.5
1977	24.2	86.9
1978	22.6	72.1
1979	24.1	46.5
1980	23.1	71.6
1981	23.9	102.6
1982	24.1	65.0
1983	23.2	30.0
1984	21.3	106.4

5–6. En el estudio de la correlación lineal entre dos variables, ¿cuál ha de ser el tamaño mínimo de la muestra para que un coeficiente de correlación medido de -0.25 implique que existe una correlación significativa con $\alpha = 0.01$?

5–7. En la tabla siguiente figuran la edad de la madre (x) y el peso del niño al nacer (y) para una muestra de 200 nacimientos agrupados en clases de edad de 2 años y en clases de pesos de 200 grs. Calcular el coeficiente de correlación ¿Es significativo? ¿A qué nivel? ¿Cómo se interpreta el valor obtenido?

$y \backslash x$	16.5	18.5	20.5	22.5	24.5	26.5	28.5	30.5	32.5	34.5	36.5	38.5	40.5	42.5	44.5
2100						3									
2300		2		1				1	1						
2500		2	2	1	2		1				1				
2700			2	3	2	1	1	1	2	1					
2900		2	3	3	4	2	1	1							
3100	2	1	9	2	1	3	2	4	3		2		1		
3300		2	9	7	6	4	6	1	1	1	3			1	
3500	1	5	2	4	3	6	4	3	3		2	2	1		
3700			3	3	2	2	3		1		1	1		1	
3900		1	1		2	1		1	2	4				1	1
4100			1		3	1			1		2	1			
4300					2		1			1					
4500					1					1					
4700															
4900															
5100									1						

5–8. A una serie de 25 datos de una variable bidimensional se han ajustado las siguientes recta de regresión de y sobre x, y de x sobre y

$$y = 4 + 0.9\,x$$

$$x = 1 + 0.3\,y$$

¿Puede decirse, con un nivel de significación $\alpha = 0.05$, que existe correlación entre ambas variables?

5–9. Sea el modelo de regresión lineal simpe $y = -31.19 + 1.46\,x$ calculado a partir de 5 parejas de valores (x_i, y_i). Las sumas de cuadrados correspondientes son

$$\begin{array}{lccr} \text{Variación explicada} & \text{VE} & = & 463.3 \\ \text{Variación no explicada} & \text{VNE} & = & 40.7 \end{array}$$

a) Calcule el coeficiente de correlación.

b) Calcule la pendiente de la recta de regresión de x sobre y ($x = a + b_{x_y}\,y$).

5–10. En la estación antártica de Halley Bay se determina la media mensual de la cantidad de ozono integrada en toda la columna atmosférica en unidades Dobson. En la tabla tenemos las cantidades registradas en los meses de octubre en el periodo 1975–1984

año	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
y	306	281	248	282	260	223	245	237	207	195

Utilizando $x =$año-1974, calcule

a) Coeficiente de correlación entre las variables x e y.

b) Recta de regresión lineal de y sobre x.

c) Determine si el coeficiente de correlación es significativamente distinto de cero con una significación del 5%.

d) ¿Para qué año nuestro modelo de regresión predice que desaparezca por primera vez todo el ozono en octubre sobre la estación de Halley Bay?

5–11. La tabla siguiente da los tiempos (en segundos) empleados en un ralley por doce coches en recorrer dos tramos de 1.5 km cada uno

tramo A: x	65	63	67	64	68	62	70	66	68	67	69	71
tramo B: y	68	66	68	65	69	66	68	65	71	67	68	70

a) Construya un diagrama de dispersión y las rectas de regresión para los tiempos de la tabla.

b) Calcule el coeficiente de correlación y la varianza residual para la variable y.

c) Decida si existe correlación entre los tiempos invertidos en recorrer cada tramo (con niveles de significación del 5% y del 1%.