# Tema 2: Modelización estadística

1. Considera la siguiente tabla de frecuencias de la variable bidimensional $(X, Y)$ :

| $Y \backslash X$ | 0 | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|---|
| $[0,4]$ | 3 | 3 | 1 | 0 | 0 | 0 | 7 |
| $(4,6]$ | 3 | 4 | 2 | 0 | 0 | 0 | 9 |
| $(6,8]$ | 1 | 3 | 2 | 1 | 0 | 0 | 7 |
| $(8,12]$ | 0 | 1 | 1 | 2 | 3 | 2 | 9 |
| | 7 | 11 | 6 | 3 | 3 | 2 | 32 |

   Se pide:

   (a) Representar la distribución condicionada $(Y/1 \leq X \leq 3)$ y calcular el sesgo y la curtosis.

   (b) Calcular las rectas de regresión de $X$ sobre $Y$ y de $Y$ sobre $X$.

   (c) Calcular el coeficiente de correlación lineal y la raíz del error cuadrático medio del modelo lineal $Y/X$.

2. Demostrar la igualdad de las dos siguientes fórmulas que permiten calcular la covarianza:

$$Cov(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{p}(x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{p} x_i \cdot y_j \cdot f_{ij} - \bar{x}\bar{y}$$

3. Calcular la recta de regresión y determinar el grado de correlación lineal de la variable $X$ con cada una de las variables $Y_i,\ i = 1, 2$ que se presentan en la siguiente tabla:

| $X$ | $Y_1$ | $Y_2$ |
|---|---|---|
| 1 | 4 | 1 |
| 2 | 2 | 3 |
| 3 | 3 | 5 |
| 4 | 2 | 7 |
| 5 | 4 | 9 |

4. Los siguientes datos están tomados de un estudio sobre el flujo de tráfico a través de un túnel para vehículos. Las cifras son los valores promedio basados en las observaciones que se hicieron en 10 intervalos de 5 minutos.

| Densidad(veh/km) | 43 | 55 | 40 | 52 | 39 | 33 | 50 | 33 | 44 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|
| Velocidad(km/h) | 27′0 | 23′8 | 30′7 | 24′0 | 34′8 | 41′4 | 27′0 | 40′4 | 31′7 | 51′2 |

   Se pide:

   (a) Representar el diagrama de dispersión.

   (b) A la vista del diagrama, elegir el valor correcto de $r$ entre estos tres valores: 0'968, -0'968, -0'198.

   (c) Verificar la respuesta calculando $r$.

   (d) ¿Hay alguna evidencia real de que exista asociación entre la velocidad de los vehículos y la densidad?

5. Se dice que dos variables son linealmente incorreladas si $r = 0$. Demuestra que 2 variables son linealmente incorreladas si y solo si su covarianza es 0.

6. Las rectas $x - 2y = 4$ y $2x - 9y = 8$ son las rectas de regresión de una variable estadística bidimensional $(X, Y)$, con $N = 10$ y $\sigma_x^2 = 9$.

   (a) Hallar el coeficiente de correlación lineal, la varianza de $Y$ y la covarianza.

   (b) Si se descubre que uno de los puntos considerados, el $(2, -1)$, no debería haberse utilizado, hallar las nuevas rectas de regresión.

7. Consideremos los siguientes modelos de regresión:

$$y = a \cdot e^{bx} \quad , \quad y = a \cdot x^b \quad , \quad y = \frac{1}{a + b \cdot x}$$

   Determina las ecuaciones normales para cada uno de ellos.

8. Ajustar el modelo $y = a \cdot b^x$ a los siguientes datos:

| Variable $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Variable $Y$ | 3'0 | 4'5 | 7'0 | 10'0 | 15'0 |

9. Ajustar el modelo $y = a \cdot x^b$ a los siguientes datos:

| Variable $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Variable $Y$ | 0'5 | 2'0 | 4'5 | 8'0 | 12'5 |

10. Ajustar el modelo $y = \dfrac{1}{a + b \cdot x}$ a los siguientes datos:

| Variable $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Variable $Y$ | 1'00 | 0'50 | 0'33 | 0'25 | 0'20 |

11. Dados los puntos $(x, y)$: $(1, 1)$, $(2, 1)$, $(3, 2)$, $(4, 4)$ y $(5, 8)$, se pide:
    (a) Estudiar si resultaría conveniente realizar un ajuste lineal.
    (b) Ajustar una función del tipo $y = a \cdot b^x$.
    (c) ¿Qué modelo es el más adecuado para la predicción? Justifica la respuesta.
    (d) Utilizar los modelos para predecir y comparar los valores $y$ para $x = 6$ y $x = 10$.

12. Dada la tabla de doble entrada:

| $X \backslash Y$ | 0 | 1 | 2 |
|---|---|---|---|
| 20 | 2 | 0 | 0 |
| 30 | 1 | 3 | 2 |
| 40 | 1 | 3 | 2 |
| 50 | 2 | 0 | 0 |

    se pide
    (a) Ajustar un modelo lineal de regresión.
    (b) Calcular el coeficiente de correlación lineal y la covarianza.
    (c) Estudiar la dependencia e independencia de las distribuciones.
    (d) Ajustar una parábola de regresión y comparar la bondad del modelo con el caso lineal.

13. Algunas veces se requiere que la curva de regresión pase por el origen. En estos casos, elegimos modelos que no tengan término independiente, como en el siguiente ejercicio. Ajustar el modelo $E = aC$ a los siguientes datos obtenidos en un experimento para determinar la rigidez de un resorte. Se midió la extensión $(E)$ del resorte (a partir de su longitud natural) bajo la acción de diferentes cargas $(C)$:

| Carga (Newtons) | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| Extensión (mm) | 10 | 19 | 29 | 40 | 48 | 56 |

14. **Regresión múltiple.** En la tabla, $z$ representa una propiedad física particular de las barras de acero forjado, y $x$ e $y$ son los porcentajes de elementos $a$ y $b$ que se encuentran presentes en la aleación. Se escogieron cuatro niveles para $x$ y cuatro para $y$, lo que da 16 posibles combinaciones, y se registró experimentalmente un valor de $z$ para barra de cada tipo. (Este es un ejemplo de lo que se conoce como diseño factorial completo).

| $x$ | 5 | 5 | 5 | 5 | 10 | 10 | 10 | 10 | 15 | 15 | 15 | 15 | 20 | 20 | 20 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| $z$ | 28 | 30 | 48 | 74 | 29 | 50 | 57 | 42 | 20 | 24 | 31 | 47 | 9 | 18 | 22 | 31 |

    (a) Demostrar que las ecuaciones normales para el modelo lineal de regresión múltiple $z = a + b \cdot x + c \cdot y$ en forma matricial son
    $$\begin{pmatrix} n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum y & \sum xy & \sum y^2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum z \\ \sum xz \\ \sum yz \end{pmatrix}$$
    (b) Determina los coeficientes $a$, $b$ y $c$ usando los datos de la tabla.

15. Consideramos los datos $(1, 2, 1)$, $(1, 4, 3)$, $(2, 2, 4)$, $(2, 2, 5)$, $(2, 4, 3)$, $(1, 4, 3)$ y $(2, 4, 5)$ de una muestra de la variable $(x, y, z)$. Se pide:
    (a) Ajustar un plano de regresión $z = a \cdot x + b \cdot y$ a la nube de puntos.
    (b) Ajustar un modelo del tipo $z = a + b \cdot \ln(xy)$.
    (c) Determinar el modelo de regresión más apropiado.