OCW-UPM Estadística para Ingeniería Civil y Medioambiental

Autores: E. M. García del Toro, C. Hermoso, E. J. Huertas

PROBLEMAS Y EJERCICIOS RESUELTOS

TEMA 2 - REGRESIÓN

Ejemplo resuelto 1

En una investigación sobre los factores que influyen en los rendimientos de las cosechas de cierta variedad de cereal, se midieron el peso de la cosecha (en toneladas métricas) y las precipitaciones totales acumuladas en el entorno de la plantación (en litros por metro cuadrado), durante los meses que van desde la siembra del cereal a la recogida del mismo.

$precipitaciones\,(l/m^2):$	24	25	19	25	27	18	20	22	28	32
$peso\,(tm):$	44	51	45	44	53	34	39	46	52	62

a) Determine la ecuación de regresión lineal que expresa el peso de la cosecha en función de las precipitaciones totales acumuladas en el entorno de la plantación.

b) Estime el peso de la cosecha para un valor de precipitaciones acumuladas de 24 l/m^2. ¿Cuánto vale el residuo o error de esta estimación?

c) Halle el valor del coeficiente de correlación e interprételo. ¿En qué porcentaje la variación en el peso de la cosecha es explicado por la variación en las precipitaciones totales acumuladas?

d) Este año, el pluviómetro que recoge los datos de precipitaciones se ha averiado, y se han obtenido 59 tm de cereal. ¿Qué valor de precipitaciones estima que ha habido en la zona?

Solución Ejemplo resuelto 1.

a) Como piden el peso de la cosecha **en función de** las precipitaciones totales acumuladas en el entorno de la plantación, tomamos $Y \equiv peso$ y $X \equiv precipitaciones$. Necesitamos la media y varianza de las X, la media de las Y y la covarianza S_{XY}. Si nos fijamos, las cantidades $(x_i - \bar{x})$ e $(y_i - \bar{y})$ son siempre las mismas, lo que permite ahorrar cálculos

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N=10} y_i = \frac{44+51+45+44+53+34+39+46+52+62}{10} = 47\ tm,$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N=10} x_i = \frac{24+25+19+24+27+18+20+22+28+32}{10} = 24\ l/m^2,$$

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{N=10} (x_i - \bar{x})^2 = \frac{0^2+1^2+(-5)^2+1^2+3^2+(-6)^2+(-4)^2+(-2)^2+4^2+8^2}{10} = 17.2,$$

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N=10} (x_i - \bar{x})(y_i - \bar{y}) = \frac{0\cdot(-3)+1\cdot4+(-5)\cdot(-2)+1\cdot(-3)+3\cdot6+(-6)\cdot(-13)+(-4)\cdot(-8)+(-2)\cdot(-1)+4\cdot5+8\cdot15}{10} = 28.1.$$

Con estos datos

$$b = \frac{S_{XY}}{S_X^2} = \frac{28.1}{17.2} = 1.6337, \qquad a = \bar{y} - b\bar{x} = 47 - 1.6337 \cdot 24 = 7.7912$$

y por tanto, la recta de regresión Y/X pedida es

$$Y(X) = 1.6337X + 7.7912$$

b) Usando la recta anterior con $X = 24$ tenemos el siguiente valor estimado (o predicho) del peso de la cosecha

$$Y(24) = 1.6337 \cdot 24 + 7.7912 = 47\,tm.$$

Como $X = 24\ l/m^2$ es un dato de las precipitaciones, podemos calcular su correspondiente residuo, que es

$$Y(24) - y_1 = 47 - 44 = 3\,tm$$

c)

$$S_Y^2 = \frac{1}{N}\sum_{i=1}^{N=10} (y_i - \bar{y})^2 = \frac{(-3)^2+4^2+(-2)^2+(-3)^2+6^2+(-13)^2+(-8)^2+(-1)^2+5^2+15^2}{10} = 55.8.$$

Coeficiente de correlación lineal de Pearson r: Interpretación: Existe una relación lineal positiva y relativamente fuerte entre las variables X e Y, dado que

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{28.1}{\sqrt{17.2} \cdot \sqrt{55.8}} = \frac{28.1}{4.1473 \cdot 7.4699} = 0.907$$

El porcentaje de variación nos lo da el coeficiente de determinación R^2 multiplicado por 100

$$R^2 = r^2 = 0.907^2 = 0.823 \rightarrow 82.3\,\%$$

La variación del peso de la cosecha es explicada por la variación de las precipitaciones totales acumuladas en un $82.3\,\%$.

d) X/Y

$$X = \tilde{b}Y + \tilde{a}, \quad \tilde{b} = \frac{S_{XY}}{S_Y^2} = \frac{28.1}{55.8} = 0.5036, \quad \tilde{a} = \bar{x} - \tilde{b}\bar{y} = 24 - 0.5036 \cdot 47 = 0.3308.$$

$$X = 0.5036Y + 0.3308.$$

$$X(Y = 59) = 0.5036 \cdot 59 + 0.3308 = 30.043\,l/m^2$$

Ejemplo resuelto 2.

En una investigación sobre los factores que influyen en la presencia de zooplancton en un lago, se midieron la densidad de zooplancton en 10 puntos diferentes del lago (en número de organismos por mililitro) y la temperatura del agua en ese punto (en grados Celsius):

$temperatura\,(^oC):$	20	21	15	21	23	14	16	18	24	28
$densidad\,(organismos/ml):$	40	47	41	40	49	30	35	42	48	58

a) Determine la ecuación de regresión lineal que expresa la densidad de zooplancton en función de la temperatura del agua en el punto donde se extrae la muestra.

b) Estime la densidad de zooplancton para una temperatura de 20 ^oC. ¿Cuánto vale el residuo o error de esta estimación?

c) Halle el valor del coeficiente de correlación e interprételo. ¿En qué porcentaje la variación en la densidad de zooplancton es explicada por la variación en la temperatura?

d) Se realiza la extracción de una muestra y se obtiene en ella una densidad de $55\,organismos/ml$, pero resulta que el termómetro que recoge los datos de temperatura se ha averiado. ¿A qué temperatura estimas que se encontraba el agua en ese punto?

Solución Ejemplo resuelto 2.

a) Como piden la densidad de zooplancton **en función de** la temperatura del agua, hacemos la recta de regresión Y/X y tomamos $Y \equiv densidad$ y $X \equiv temperatura$. Necesitamos la media y varianza de las X, la media de las Y y la covarianza. Si nos fijamos, las cantidades $(x_i - \bar{x})$ e $(y_i - \bar{y})$ son siempre las mismas, lo que permite ahorrar cálculos

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N=10} y_i = \frac{40+47+41+40+49+30+35+42+48+58}{10} = 43\ org/ml,$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N=10} x_i = \frac{20+21+15+21+23+14+16+18+24+28}{10} = 20\ ^oC,$$

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{N=10} (x_i - \bar{x})^2 = \frac{0^2+1^2+(-5)^2+1^2+3^2+(-6)^2+(-4)^2+(-2)^2+4^2+8^2}{10} = 17.2,$$

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N=10} (x_i - \bar{x})(y_i - \bar{y}) = \frac{0\cdot(-3)+1\cdot 4+(-5)\cdot(-2)+1\cdot(-3)+3\cdot 6+(-6)\cdot(-13)+(-4)\cdot(-8)+(-2)\cdot(-1)+4\cdot 5+8\cdot 15}{10} = 28.1.$$

Con estos datos

$$b = \frac{S_{XY}}{S_X^2} = \frac{28.1}{17.2} = 1.6337, \qquad a = \bar{y} - b\bar{x} = 43 - 1.6337 \cdot 20 = 10.326$$

luego $Y = 1.6337X + 10.326$ es la recta de regresión pedida.

b) Usando la recta anterior con $X = 20$ tenemos un valor estimado de $Y = 1.6337 \cdot 20 + 10.326 = 43\,org/ml$. El residuo es, por tanto $43 - 40 = 3\,org/ml$.

c) Necesitamos S_Y. Ahorramos cálculos sabiendo que las $(y_i - \bar{y})$ son las del apartado **a)** que usamos en la covarianza

$$S_Y^2 = \frac{1}{N}\sum_{i=1}^{N=10} (y_i - \bar{y})^2 = \frac{(-3)^2+4^2+(-2)^2+(-3)^2+6^2+(-13)^2+(-8)^2+(-1)^2+5^2+15^2}{10} = 55.8.$$

Luego (interpretación) existe una relación lineal positiva y relativamente fuerte entre las variables X e Y, dado que

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{-28.7}{\sqrt{55.8} \cdot \sqrt{17.2}} = -0.9264.$$

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{28.1}{\sqrt{17.2} \cdot \sqrt{55.8}} = 0.907.$$

El porcentaje de variación nos lo da el coeficiente de determinación R^2 multiplicado por 100. Así, como $R^2 = r^2 = 0.907^2 = 0.823$, la variación de la densidad de zooplancton es explicada por la variación de las temperaturas en un $82.3\,\%$.

d) Necesitamos la recta de regresión X/Y, que es $X = \tilde{b}Y + \tilde{a}$, pero todo lo necesario lo hemos calculado ya antes

$$\tilde{b} = \frac{S_{XY}}{S_Y^2} = \frac{28.1}{55.8} = 0.5036, \qquad \tilde{a} = \bar{x} - \tilde{b}\bar{y} = 20 - 0.5036 \cdot 43 = -1.6548.$$

Luego $X = 0.5036Y - 1.6548$ y por tanto $X(Y = 55) = 0.5036 \cdot 55 - 1.6548 = 26.043\,^oC$ es la estimación pedida.

La recta de regresión que hemos estudiado en primer lugar se denomina *recta de regresión Y/X* (lo anterior se lee: "recta de regresión de Y sobre X") y hemos visto que está definida por la ecuación $Y = a + bX$, donde

$$a = \bar{y} - b\bar{x}, \quad b = \frac{S_{XY}}{S_X^2}.$$

Vamos a ver que existe otra recta de regresión semejante, pero en donde representamos los valores de X en función de los valores de Y.

Ejercicio propuesto 1.- En un análisis de las aguas de cierto lago, se obtienen los siguientes valores de *concentración de sólidos suspendidos* (medida en mg/l) y de *turbidez* del agua (medida en Unidades Nefelométricas de turbidez, o Nephelometric Turbidity Unit (NTU))

turbidez (NTU)	95	100	102	104	100	98	96	100	110	99
sólidos suspendidos (mg/l)	85	94	84	88	85	92	76	90	102	89

a) Utilizando el método de los mínimos cuadrados, obtenga la ecuación de la recta que expresa la concentración de sólidos suspendidos en función de la turbidez del agua.

b) ¿Cuál es el valor esperado de la concentración de sólidos suspendidos para una turbidez de 95 NTU? ¿Cuánto vale el residuo de esta estimación?

c) ¿En qué porcentaje la variación en la concentración de sólidos suspendidos es explicada por la variación en la turbidez del agua?

d) Si en cierto punto del lago la concentración de sólidos suspendidos presenta un valor de 86 mg/l, ¿qué valor de turbidez puede esperarse en esa zona?

Ejercicio propuesto 2.- Se administra un larvicida volátil en una balsa de riego para cultivos destinados a consumo humano. A continuación se determinan las concentraciones de larvicida (en $\mu g/ml$) en relación al paso del tiempo (en horas) y los resultados son los siguientes:

tiempo (h)	1	1.5	2	3	6	15
concentración ($\mu g/ml$)	11.8	11.0	10.9	10.1	9.6	5.7

a) Dibuja el diagrama de dispersión (nube de puntos) de los datos anteriores.

b) Determina, a partir de la forma de la nube, si el modelo de regresión lineal es el adecuado. En caso afirmativo, obtén la expresión matemática que relaciona la concentración de larvicida en el agua en función del tiempo.

c) Estima el valor de la concentración a las 9 horas.

d) Calcula el coeficiente de correlación e interprétalo.

e) Se sabe que el agua tratada con este larvicida puede ser destinada a riego siempre que no contenga una concentración superior a 2.5 $\mu g/ml$. En base a la información de que dispones, ¿crees que sería adecuado regar con el agua de la balsa pasadas 24 horas desde la administración del larvicida?

Ejercicio propuesto 3

a) Demuestra que la ecuación $Y = a + bX$ puede reescribirse de la siguiente forma:

$$(Y - \bar{y}) = \frac{S_{XY}}{S_X^2}(X - \bar{x}), \tag{1}$$

siendo ambas expresiones totalmente equivalentes. La primera es la *ecuación explicita* de la recta, y (1) es conocida como *ecuación punto-pendiente* de la recta.

b) Calcula la expresión de los coeficientes $\tilde{a}$ y $\tilde{b}$ para la *recta de regresión X/Y* (es decir, para la recta de regresión de X sobre Y), cuya ecuación explícita es $X = \tilde{a} + \tilde{b}Y$.

c) Obtén la ecuación punto-pendiente de la recta de regresión X/Y.

Indicaciones y soluciones: Antes de mirar la solución, se sugiere que el alumno intente realizar previamente el ejercicio.

a) Solución. Dado que $a = \bar{y} - b\bar{x}$, y $b = S_{XY}/S_X^2$ tenemos

$$Y = \bar{y} - \frac{S_{XY}}{S_X^2}\bar{x} + \frac{S_{XY}}{S_X^2}X.$$

Sacando factor común para los dos últimos sumandos de la derecha, y pasando $\bar{y}$ al otro lado del igual, concluimos

$$(Y - \bar{y}) = \frac{S_{XY}}{S_X^2}(X - \bar{x}).$$

b) Solución: La recta es $X(Y) = \tilde{a} + \tilde{b}Y$ donde

$$\tilde{a} = \bar{x} - \tilde{b}\bar{y}, \quad \tilde{b} = \frac{S_{XY}}{S_Y^2}.$$

c) Solución:

$$(X - \bar{x}) = \frac{S_{XY}}{S_Y^2}(Y - \bar{y})$$

Ejercicio propuesto 4

En la Figura 1 aparecen nubes de puntos de diversas formas. Da un valor estimado del coeficiente de correlación lineal de Pearson ρ (en ocasiones se denota también como r) para cada uno de los dibujos. Recuerda que ρ es un número sin unidades (número puro) cuyos valores se encuentran siempre en el intervalo $-1 \leq \rho \leq 1$.

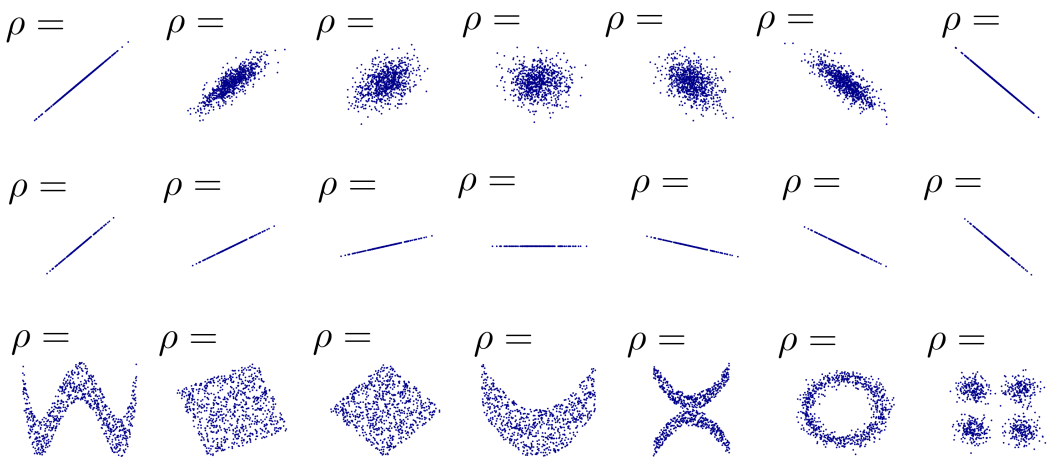

Figure 1: El autor de la imagen es Denis Boigelot para Wikipedia©

Solución: Puedes consultar la respuesta en la página web:

https://en.wikipedia.org/wiki/Correlation_and_dependence

donde aparecen los mismos dibujos del enunciado con su correspondiente coeficiente de correlación lineal de Pearson.