APROXIMACIÓN POR MÍNIMOS CUADRADOS

"No manipules tus datos, pues ellos podrían estar correctos"
– Wilbur Wright

Introducción

En la investigación científica es esencial estudiar las relaciones entre distintos fenómenos con el objetivo de descubrir conexiones de causa-efecto; así como la relación entre diferentes parámetros o indicadores para determinar si se puede conocer el estado de un sistema según una variable relacionada al mismo. Esto conlleva una apreciable cantidad de trabajo a realizar: Mediciones y observaciones experimentales que raramente se comportan de manera exacta según una fórmula. Uno de los recursos del experimentador es lograr hallar la fórmula que mejor se aproxime al comportamiento mostrado por las mediciones.

Por ejemplo, en 2015 se publicó un estudio sobre el efecto tóxico que tiene la concentración de nitratos en los cultivos del camarón de cola blanca (*Litopenaeus Vannamei*), sobre el crecimiento de los mismos [2]. Se determinó que la tasa de crecimiento semanal G (en gr/semana) decrece a medida que se incrementa la concentración de nitratos N (en mg/L), según la expresión:

$$G = 0,874 - 0,0007 \times N$$

La Figura 1 muestra uno de los resultados de ese estudio:

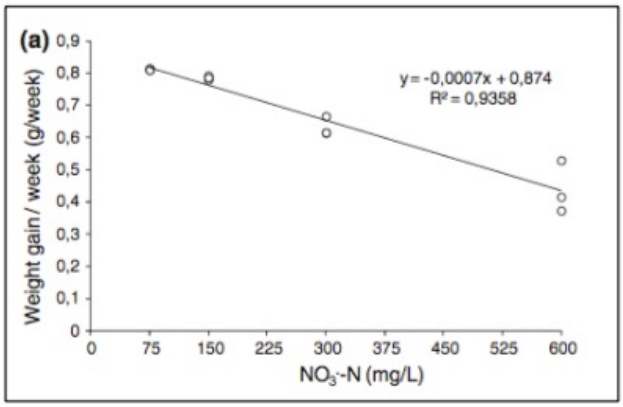

Figura 1: Decrecimiento de la ganancia de peso por semana, según la concentración de nitratos en las piscinas de cultivo de camarón de cola blanca, *L. Vannamei* [2].

En Figura 1, los pequeños círculos indican los datos reales medidos, mientras que la línea recta corresponde a la función lineal que mejor se aproxima a los datos medidos, no necesariamente que contenga dichos puntos.

En muchas ocasiones resolver un problema en estadística puede ser reducido a resolver un sistema de ecuaciones lineales. Esto es cierto en el caso del método de mínimos cuadrados para estimar los parámetros de un modelo lineal. Y el concepto de existencia o no de la solución se convierte en determinar cuáles funciones paramétricas son estimables o no.

Regresión Lineal

Suponga que se dispone de dos secuencias de mediciones **x** e **y**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

con correspondencia entre mediciones (a la i-ésima medida x_i le corresponde la medición y_i). Suponga que hay evidencia experimental de una relación lineal entre los vectores dados, de modo que se desea encontrar una expresión de tipo $y = mx + b$ (lineal), cuya gráfica represente mejor al conjunto de puntos (x_i, y_i) de las mediciones tomadas.

Si la relación fuera exacta, y las mediciones perfectas y sin errores, se cumpliría que:

$$\begin{aligned} y_1 &= b + mx_1 \\ y_2 &= b + mx_2 \\ &\vdots \\ y_n &= b + mx_n \end{aligned}$$

pero en la vida real esto nunca se cumple, por lo cual, al evaluar los x_i en la recta $y = mx + b$ se producen como resultado $\hat{y}_i$ "teóricos" o "predichos", que no son iguales a los y_i "medidos":

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{bmatrix} b \\ m \end{bmatrix}$$

O, $\hat{\mathbf{y}} = \mathbf{Au}$ en su forma matricial. ¿Cuáles son los valores de los parámetros m y b que producen la menor distancia entre los y_i y los $\hat{y}_i$? Es decir, que minimizan la distancia entre $\mathbf{y}$ y $\hat{\mathbf{y}}$, $\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|$.

La expresión $\hat{\mathbf{y}} = \mathbf{Au}$ corresponde a un sistema de ecuaciones lineales, con $A_{n\times 2}$, y donde $\hat{\mathbf{y}}$ es un vector de $\mathbb{R}^n$ que pertenece a la imagen de A. Así también, Im(A) es un subespacio de dimensión menor o igual que 2, pues A solo tiene dos columnas. Según el teorema de la Aproximación de la Norma [1], la distancia $\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|$ es mínima cuando $\mathbf{y} - \hat{\mathbf{y}}$ es perpendicular a Im(A).

Sea $\bar{\mathbf{u}}$ el vector que minimiza $\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|$, es decir, cuando se cumple que $A\mathbf{u} \perp \mathbf{y} - A\bar{\mathbf{u}}$. Esto necesariamente conlleva que $A^T\mathbf{y} = A^TA\bar{\mathbf{u}}$, en consecuencia, la solución del vector que minimiza la distancia $\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|$ es: $\bar{\mathbf{u}} = (A^TA)^{-1}A^T\mathbf{y}$.

Entregables

En la parte conceptual, cada estudiante debe poder responder a las siguientes cuestiones:

1. Demuestre que el sistema AX=B es consistente si y solo si $C_B \subset C_A$

2. Demuestre que el sistema AX=B es consistente si y solo si toda columna de B pertenece a C_A
3. Demuestre que el sistema AX=B es consistente si y solo si $C_{(A|B)} = C_A$
4. Demuestre que el sistema AX=B es consistente si y solo si el rango de la matriz aumentada (A | B) es igual al rango de A
5. Enuncie y demuestre el Teorema de la Aproximación de la Norma.
6. Demostrar que el núcleo de una matriz es el complemento ortogonal de R_A
7. Deducir la fórmula de la solución $\bar{\mathbf{u}} = (A^T A)^{-1} A^T \mathbf{y}$
8. Compare la deducción de la misma fórmula $\bar{\mathbf{u}} = (A^T A)^{-1} A^T \mathbf{y}$ pero utilizando cálculo y las derivadas, y establezca igualdades o diferencias entre los resultados.
9. Justificar por qué se afirma que necesariamente $A^T \mathbf{y} = A^T A \bar{\mathbf{u}}$.
10. Deducir la formulación para la regresión con polinomios de orden 2. Si se quiere relacionar dos secuencias de mediciones **x** e **y** mediante la expresión $y_i = c + bx_i + ax_i^2$. ¿Qué forma tienen en este caso las matrices **A** y **u** en la ecuación $\hat{\mathbf{y}} = \mathbf{Au}$? ¿Cuál es la expresión para el vector **u** que minimiza la distancia entre **y** e $\hat{\mathbf{y}}$?
11. Deducir la formulación para la regresión con polinomios de orden mayor a 2.
12. Analice el tipo de matriz al que pertenece A^TA ¿Cuándo es A^TA una matriz invertible?
13. ¿Como se tratarían otros modelos funcionales entre x y y?
14. ¿Como se generalizaría la formula para cuando se trata de explicar Y como una función de varias variables $x_1, x_2, ..x_N$?

En la parte aplicativa, debe constar exclusivamente el análisis del caso de estudio dado, las conexiones que existe entre las variables, proveyendo respuestas a las siguientes cuestiones:

- Elaborar el gráfico de los datos originales que se desean relacionar (en un plano **y** versus **x**).
- Decidir cuál es la mejor relación polinomial (lineal, cuadrática, cúbica, etc) entre estas variables.
- A partir de los cálculos hechos, provea una tabla con los coeficientes del polinomio que mejor se ajusta a los datos dados.
- Grafique el polinomio hallado.

Además, hay una parte interpretativa de los resultados, que puede ser muy complicado en problemas reales, se trata del fundamental problema filosófico de la correlación y la causa-efecto:

- Investigar acerca de la frase "correlación no implica causación", que acompaña a los científicos durante toda su vida como admonición. Explique en sus propias palabras y mediante ejemplos los dos casos: la correlación y la causa-efecto.

NOTA: Es lícito apoyarse en la tecnología: si utiliza un software o calculadora (Matlab®, Python, Excel, etc), o algún sitio web de resolución de matrices (Geogebra, etc), debe ser indicado en el documento: planteando la fórmula teórica, indicando lo que se utilizó para resolver esa ecuación y escribiendo el resultado directamente. Así, para cada una de las ecuaciones resueltas. Si consultó un libro o artículo, se debe incluir en una sección Bibliografía o Referencias del documento.

CASO DE ESTUDIO:

- Grupo 1: Aeronáutica
- Grupo 2: Biología
- Grupo 3: Oceanografía
- Grupo 4: Meteorología
- Grupo 5: Ecología
- Grupo 6: Economía
- Grupo 7: Agronomía

A cada grupo se le pedirá sustentar 3 de los ítems planteados en Entregables.