PROYECTO

Término II · 2017 – 2018

APROXIMACIÓN POR MÍNIMOS CUADRADOS

En las ciencias aparecen a menudo fórmulas, por ejemplo, la fórmula del caudal estudiado en ciencias físicas:

$$Q = S.V$$

establece que, dado un fluido que atravieza un conducto de sección S con superficie en unidades cuadradas, con una velocidad V en unidades lineales por segundo, produce un flujo de volumen de S.V unidades cúbicas por segundo. Si el vector velocidad forma un ángulo respecto al vector normal del plano, la fómula suele expresarse como:

$$Q = S.V\,\mathrm{Cos}(\theta)\,,$$

de donde deriva la expresión más general del producto interno entre el vector velocidad del fluido y el vector normal al plano:

$$Q = \left\langle \vec{S}, \vec{V} \right\rangle$$

Lo que no es evidente es la cantidad de trabajo que el científico tiene que realizar para obtener una fórmula como esta. Las mediciones y observaciones experimentales raramente se comportan de manera exacta según una fórmula, y uno de los recursos que queda al experimentador es lograr hallar la fórmula que mejor se aproxime al comportamiento dado por las mediciones.

Por ejemplo, en 2001 se efectuaron varias mediciones del caudal que atravezaba la sección transversal del canal de Lido di Venezia, Italia. La Figura 1 muestra una de estas mediciones del 8 de marzo de 2001. Los colores indican la magnitud de la velocidad de corriente de agua, variando desde casi 0 m/s en la zona periférica, hasta alcanzar casi los 2 m/s en la zona más profunda del canal. La medición se efectuó desde una barca portanto un instrumento especializado que no podía medir muy cerca del fondo del canal. Las fluctuaciones de la barca con el oleaje volvían poco confiables las mediciones cercanas a la superficie; mientras en la zona confiable la velocidad no era constante:

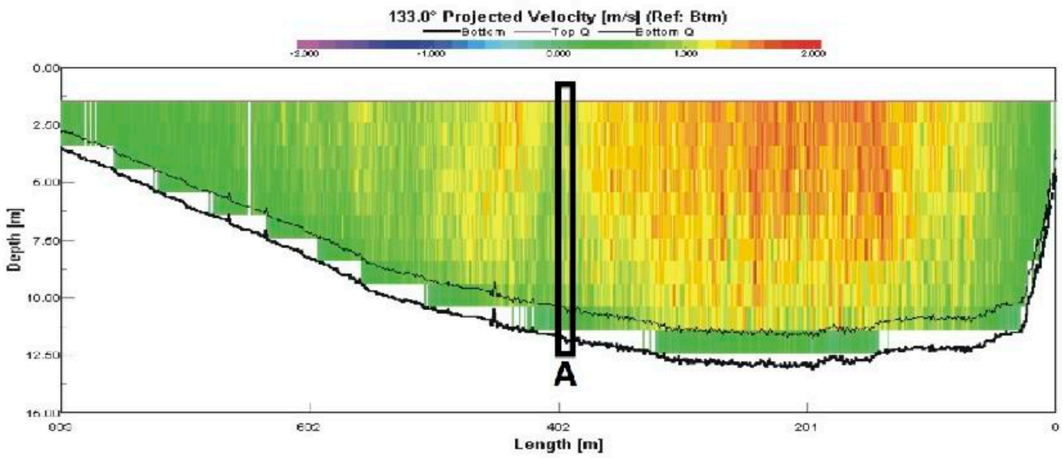


Figura 1: Campo de velocidades en el canal de Lido di Venezia, 2001.

El caudal tuvo que expresarse como una sumatoria de caudales elementales con muchos errores experimentales:

$$\hat{Q}_T = \sum_i \langle \vec{S}_i, \vec{V}_i \rangle$$

Además, la idea final era poder calcular el caudal total sin necesidad de la barca sino utilizando la velocidad medida por un solo correntómetro localizado en el punto A, que solo medía la velocidad promedio de la columna de agua ubicada encima del mismo. ¿Es suficiente este instrumento para calcular todo el caudal del canal? Los resultados mostraron que sí. Los datos del caudal total medido por la barca, al compararse con las mediciones simultáneas del correntómetro del fondo del canal, mostraron una relación de marcado carácter lineal, lo que permitió calcular coeficientes de proporcionalidad y finalmente prescindir de la barca para *estimar* el caudal del canal.

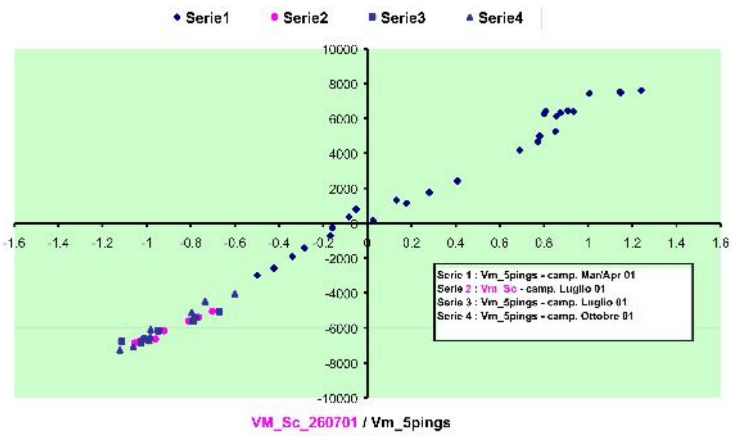


Figura 2: Relación lineal entre el caudal (eje y, en m^3/s) del canal Lido di Venezia, y la velocidad en el punto A (Eje x, en m/s).

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 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 el 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:

$$y_1 = b + mx_1, \quad y_2 = b + mx_2, \quad \dots \quad y_n = b + mx_n$$

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 $\hat{y}_i$ que no son iguales a los y_i:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{bmatrix} b \\ m \end{bmatrix} \quad \text{o} \quad \hat{\mathbf{y}} = A\mathbf{u}.$$

¿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 $\hat{\mathbf{y}}$, $\|\mathbf{y}-\hat{\mathbf{y}}\|$.

La expresión $\hat{\mathbf{y}} = A\mathbf{u}$ 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, $\mathrm{Im}(A)$ 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, la distancia $\|\mathbf{y}-\hat{\mathbf{y}}\|$ es mínima cuando $\mathbf{y}-\hat{\mathbf{y}}$ es perpendicular a $\mathrm{Im}(A)$.

Tarea 1: Enuncie y demuestre el Teorema de la Aproximación de la Norma.

Sea $\bar{\mathbf{u}}$ el vector minimiza $\|\mathbf{y}-\hat{\mathbf{y}}\|$, es decir, se cumple que:

$$A\mathbf{u} \perp (\mathbf{y} - A\bar{\mathbf{u}}),$$

o:

$$\langle A\mathbf{u}, (\mathbf{y} - A\bar{\mathbf{u}}) \rangle = 0$$

Aplicando las propiedades del producto interno:

$$\langle A\mathbf{u}, \mathbf{y} \rangle - \langle A\mathbf{u}, A\bar{\mathbf{u}} \rangle = 0$$

$$\mathbf{u}^T (A^T\mathbf{y} - A^T A\bar{\mathbf{u}}) = 0$$

lo cual se cumple para cualquier $\mathbf{u}$, solo si $A^T\mathbf{y} = A^T A\bar{\mathbf{u}}$. Y el vector $\bar{\mathbf{u}}$ es:

$$\bar{\mathbf{u}} = (A^T A)^{-1} A^T \mathbf{y}$$

Tarea 2: Justificar por qué se afirma que necesariamente $A^T\mathbf{y} = A^T A\bar{\mathbf{u}}$.

> **Teorema 1:** La matriz $A^T A$ es invertible si los n puntos de datos (x_i, y_i) no son colineales.

Tarea 3: Demostrar el Teorema 1.

ꕥ

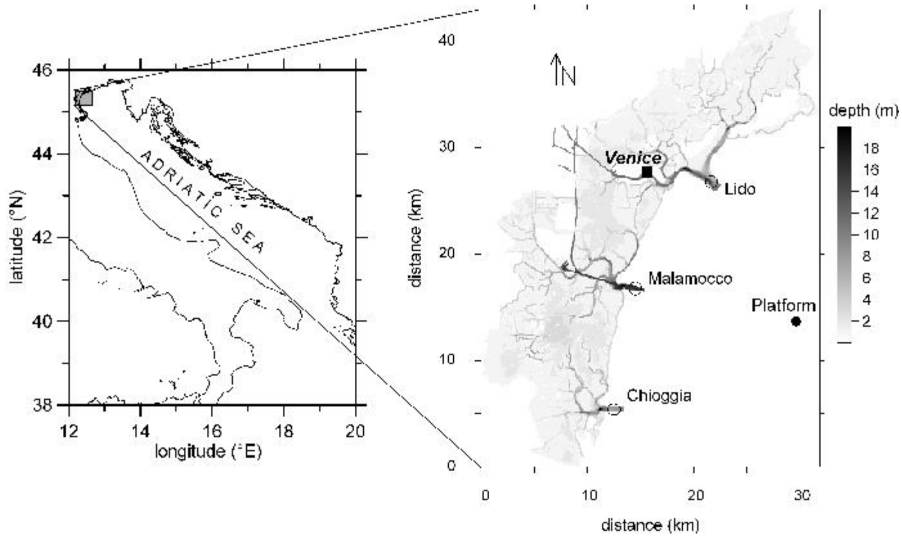
Tarea 4: Deducir la formulación para la regresión cuadrática, cuando se quiere relacionar dos secuencias de mediciones **x** e **y** mediante la expresión $y_i = c + bx_i + ax_i^2$. ¿Qué forma tienen en este caso las matrices **A** y **u** en la ecuación $\hat{\mathbf{y}} = A\mathbf{u}$? ¿Cuál es la expresión para el vector $\bar{\mathbf{u}}$ que minimiza la distancia entre $\mathbf{y}$ e $\hat{\mathbf{y}}$?

Tarea 5: Un fabricante compra grandes cantidades de refacciones para las máquinas de su empresa. Encuentra que el costo depende de la cantidad de cajas de refacciones que compra, y que el costo por unidad comprada disminuye según aumenta la cantidad comprada. Si la relación entre cajas compradas y el costo total es de tipo cuadrático, encuéntrese la función de costo total, basado en los datos de la tabla siguiente:

Número de cajas	Costo total
10	$150
30	$260
50	$325
100	$500
175	$670

CAMPAÑAS DE MEDICIÓN DE FLUJO DE AGUA EN MALAMOCCO, ITALIA

El canal de Malamocco es la principal vía de entrada al puerto industrial de la ciudad de Venecia. Dada su importancia comercial, se decidió estudiar el flujo de agua a través del canal mediante una campaña de mediciones del flujo de volumen de agua. Esta campaña se llevó a cabo varios días a lo largo de 2008.



El flujo de volumen en m^3/s se midió en 172 ocasiones, mediante una barca a través del canal con un medidor acústico de flujo. Al mismo tiempo, un correntómetro situado en el fondo del canal midió la velocidad de la corriente de agua, en m/s, a lo largo del mismo. Aplique los métodos de aproximación por mínimos cuadrados para obtener una expresión que permita calcular el flujo de volumen $\mathbf{Q_T}$ en m^3/s, a partir de las mediciones **V** de la corriente de agua en m/s. Provea las respuestas requeridas en Entregables.
NOTA: Los datos se proveen en el archivo adjunto:
Proyecto_Regresión_Venecia_Datos2008.xls

REFERENCIAS:

- Grossman, S. I., Flores Godoy, J.: Álgebra Lineal. McGraw-Hill Ed., México

(2012).
- Gačić, M., Kovačević, V., Mancero Mosquera, I., Mazzoldi, A., Cosoli, S: Water fluxes between the Venice Lagoon and the Adriatic Sea; en: Flooding and environment challenges for Venice and its lagoon, Fletcher & Spencer (Ed.). Cambridge University Press, Cambridge (2005).

Entregables:

- Se debe presentar dos secciones. La primera parte consta de la resolución de las tareas planteadas, respondiendo a las preguntas dadas.
- En la segunda sección debe constar exclusivamente el análisis de regresión con los datos provenientes de Venecia, proveyendo respuestas a las siguientes cuestiones:

1. Elabore un gráfico de los datos provistos, y úselo para justificar su decisión sobre la elección de un modelo lineal o uno cuadrático.
2. A partir de los cálculos hechos, elabore una tabla con los valores estimados de los parámetros del vector **u** correspondiente.
3. Elabore un gráfico de la función elegida para aproximar los puntos de los datos.
4. Durante tormentas fuertes, la velocidad del agua a través del canal puede alcanzar 1,4 m/s, ¿Cuál es el volumen de agua que fluye por el canal bajo esas condiciones? Justifique su respuesta.
5. La barca también mide el área del canal (en m^2), repita el análisis pero ahora entre las variables Flujo Total y Área.
6. Compare los resultados del ítem 5 con los del ítem 2, y decida cuál variable es el mejor predictor del Flujo Total, si la Velocidad de corriente o el Área.
7. Si leyó o consultó un libro o artículo, debe poner al final del documento una sección Bibliografía o Referencias y hacer una lista de cada trabajo consultado, incluyendo título, autor(es), Capítulo/Volumen, páginas.

NOTA: Es lícito apoyarse en la tecnología, pero si necesita utilizar un software o calculadora (Matlab®, Phyton, etc), o algún sitio de resolución de matrices online (Online Matrix Calculator de blue-bit), debe ser indicado en el documento a entregar, planteando la fórmula teórica, e indicando si se utilizó para resolver esa ecuación; y repitiendo la indicación para cada una de las ecuaciones así resueltas.