Herramientas computacionales - 2022 - I

Correlación & Causalidad

Nota: El proyecto se entrega con un formato tipo artículo en un máximo de 5 páginas (con texto y gráficas) antes de las 08:00 AM del día 27 de enero de 2022.

Bases de datos:

1. **Casos de Ébola, 2014 - 2016**.
2. **Vivienda en California**.
3. **Ejecuciones en Estados Unidos de América, 1976 - 2016**.

1. Teniendo en cuenta la base de datos asignada:
   a) Describa de manera breve el fenómeno que relaciona, las variables involucradas y formule una pregunta problema.
2. Elija las variables más relevantes a su criterio (explique el porqué) y muéstrelas en una gráfica.
3. Realice una clase que se llame THC_{corr} cuyas funciones le permitan calcular la correlación entre las dos variables seleccionadas. Recuerde que solo puede utilizar las librerías Numpy o Scipy para cálculos elementales (*e.g.,* sumas, restas, multiplicaciones, etc.).
4. Explique los resultados obtenidos.
5. Partiendo de los resultados anteriores, ajuste un polinomio (del grado que usted considere), genere la clase THC_{graf} que le permita mostrar el ajuste.
6. Explique y analice los resultados obtenidos, formule una predicción con base en el modelo matemático que construyó y explique si, ¿la correlación implica causalidad?

Recuerde:

- Entregar el código explicado en un cuaderno de programación respondiendo cada pregunta.
- Realizar una presentación de 10 diapositivas.

Anexo

1. Regresión polinómica

Se puede considerar una aproximación de la función mediante un polinomio de orden "**n**çomo:

$$y = \beta_0 + \beta_1 x + \beta_2^2 + \cdots + \beta_n x^n,$$

donde x^j con $j \in \{1, 2, \ldots n\}$, son conocidas como variables regresoras y **y** como variable respuesta.

El concepto básico es sencillo, se tiene una sola variable explicativa contínua, x, pero podemos ajustar potencias mayores de x, como x^2, x^3, etc, y añadirlas al modelo, junto a x, para describir diversos tipos de curvatura en la relación $y - x$.
Hay varias consideraciones importantes que se presentan cuando se ajusta un polinomio de una variable, algunas de ellas son:

- Orden del modelo: como regla general, se debe evitar el uso de polinomios de orden superior (con $k > 2$), a menos que se puedan justificar por razones ajenas a los datos. Un modelo de orden menor en una variable transformada casi siempre es preferible a un modelo de orden superior en la métrica original. El ajuste arbitrario de polinomios de orden superior es un grave abuso del análisis de regresión. Siempre se debe mantener un sentido de parsimonia, esto es, se debe usar el modelo más simple posible que sea consistente con los datos y el conocimiento del ambiente del problema.

- Extrapolación: puede ser peligrosa en extremo, por ejemplo, si se extrapola más allá del rango de los datos originales, la respuesta predicha se va hacia abajo; esto puede ser contrario al comportamiento real del sistema. En general, los modelos polinomiales pueden dirigirse hacia direcciones imprevistas e inadecuadas, tanto en la interpolación como en la extrapolación.

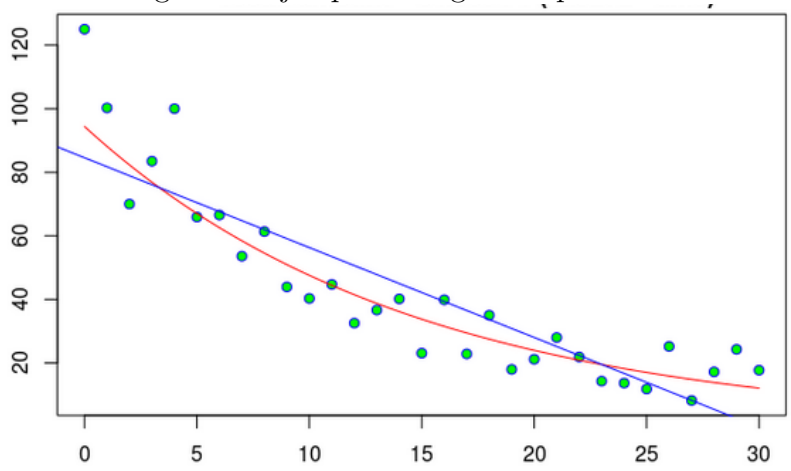
- Mal acondicionamiento: si los valores de x se limitan a un rango estrecho, puede haber mal acondicionamiento o multicolinealidad apreciables en las variables.

En la Figura 1, se puede visualizar un ejemplo de como se vería una regresión polinomial, de grado 1 (azul) y de grado 2 (roja).

2. Coeficiente de correlación de Pearson ρ_{xy}

Es una medida de dependencia lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida

Figura 1: Ejemplo de regresión polinomial.



de las variables.

De manera menos formal, se puede definir como un índice que mide el grado de relación de dos variables, siempre y cuando ambas sean cuantitativas y continuas expresado como:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}},$$

donde:

- σ_{xy} es la covarianza de (x, y).
- σ_x es la desviación estándar de x.
- σ_y es la desviación estándar de y.

Sin embargo, en términos computacionales no es accesible realizar cálculos de variables continuas, debido a eso se usa el estadístico muestral denotado por r_{xy}

$$r_{xy} = \frac{\sum_{i=1}^{N} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{n\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{n\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}$$

donde N es el número de muestras y r_{xy} el coeficiente de correlación de Pearson muestral.

2.1. Interpretación

El valor del índice está acotado entre -1 y 1, es decir, $r_{xy} \in [-1, 1]$

- Si $r_{xy} = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
- Si $r_{xy} = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante.
- Si $r_{xy} = 0$, entonces no existe relación lineal pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.
- Si $r_{xy} \geq 0{,}6$ Se considera una correlación positiva moderamente alta.
- Si $0 < r_{xy} < 0{,}6$ Se considera una correlación positiva moderamente baja.
- Si $r_{xy} \leq -0{,}6$ Se considera una correlación positiva moderamente alta.
- Si $-0{,}6 < r_{xy} < 0$ Se considera una correlación positiva moderamente baja.