Instituto Politécnico Nacional

Centro de Investigación en Cómputo

Python en el Ámbito Científico

Profesor: Alan Badillo Salas

Tarea 5

— Problemas Científicos con Python

Hoja de Actividades y Ejercicios

Julio 19, 2025.

Actividad 9

La regresión lineal simple, es un modelo potente que permite generalizar todas las posibles rectas que pasan por un conjunto de puntos, y determinar cuál es la que está más cercana de todos los puntos.

Una recta simple está determinada por dos parámetros β_0 y β_1 y dos variables x y y, bajo la relación:

$$y = \beta_0 + \beta_1 \cdot x \tag{1}$$

Donde x representa una variable conocida llamada la *covariable* o *variable predictiva* y y representa una variable desconocida llamda la *respuesta* o *variable de predicción*. Los parámetros $\beta_0, \beta_1 \in \mathbb{R}$ son valores de ajuste también llamados *parámetros de regresión* o *coeficientes de la regresión*.

Esto significa que podemos descubrir el valor de y para cada x que demos, si conocemos los parámetros β_0 y β_1.

El problema de regresión, consiste en encontrar o ajustar los mejor posible los parámetros desconocidos inicialmente β_0 y β_1, mediante los datos observados. Es decir, si se tienen n observaciones de x y y dadas por un conjunto de puntos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, entonces, podemos intentar ajustar β_0 y β_1 de tal manera que $y = \beta_0 + \beta_1 \cdot x$ tenga el mínimo error posible entre las observaciones y los valores evaluados.

En términos simples, buscaremos ajustar β_0 y β_1 dadas las observaciones de $x = \{x_1, x_2, \ldots x_n\}$ y $y = \{y_1, y_2, \ldots, y_n\}$.

Para encontrar β_0 y β_1 usaremos las ecuaciones derivadas de la multiplicación matricial usando las ecuaciones normales:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \cdot y_i \end{pmatrix} \tag{2}$$

O en términos simples:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n & S_x \\ S_x & S_{x^2} \end{pmatrix}^{-1} \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \tag{3}$$

Donde $S_x = \sum_{i=1}^{n} x_i$, $S_{x^2} = \sum_{i=1}^{n} x_i^2$, $S_y = \sum_{i=1}^{n} y_i$ y $S_{xy} = \sum_{i=1}^{n} x_i \cdot y_i$

Entonces, resolviendo la matriz inversa usando el método del determinante, tenemos que:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \frac{1}{nS_{x^2} - (S_x)^2} \begin{pmatrix} S_{x^2} S_y - S_x S_{xy} \\ -S_x S_y + n S_{xy} \end{pmatrix} \tag{4}$$

Por lo que usando la **Ec. (4)**, podemos determinar los valores de β_0 y β_1 como:

$$\beta_0 = \frac{S_{x^2} S_y - S_x S_{xy}}{n S_{x^2} - (S_x)^2}$$
$$\beta_1 = \frac{n S_{xy} - S_x S_y}{n S_{x^2} - (S_x)^2} \tag{5}$$

Donde $S_x = \Sigma(x)$, $S_{x^2} = \Sigma(x^2)$, $S_y = \Sigma(y)$ y $S_{xy} = \Sigma(x \cdot y)$, es decir, las sumas de los valores de x, x^2, $x \cdot y$ y y.

Para esta actividad usaremos la **Ec. (5)** para determinar los coeficientes de regresión y ver qué tan bien se ajustan los datos a la recta, así como determinar algunos valores de predicción.

Usaremos el conjunto de datos de la **Tabla 1**.

x	y
4	13.4
6	18.8
7	21.4
9	26.3
15	47.9
2	10.3
6	20.3
10	33.4

Tabla 1: Datos observados de x y y

Observa que los datos no necesitan están en orden, solo son observaciones y se pueden tener tantas observaciones n como se hayan obtenido. Por ejemplo, la x podría representar el precio del litro petróleo en dólares, mientras que la y podría representar el valor de una acción de la petrolera *Shell*. Cuando el litro de petróleo cuesta \$4 dólares, la acción vale cerca de \$13.4 dólares, mientras que si el litro de petróleo sube a \$15 dólares, la acción subirá de precio a los \$47.9 dólares. Indicando un comportamiento de alza en los precios con una correlación positiva.

Resuleve los pasos siguientes para calcular los coeficientes de regresión, visualizar qué también se ajustan las observaciones a la recta ideal del modelo de regresión lineal simple y cuáles son las predicciones para diferentes precios del litro de petróleo:

1. Grafica los puntos observados x y y usando un punto por cada pareja (x_i, y_i), no conectes los puntos para mejor visualización (deja solo la nube de puntos).

2. Calcula S_x como la suma de todas las x, es decir, $S_x = 4 + 6 + 7 + 9 + 15 + 2 + 6 + 10$.

3. Calcula S_{x^2} como la suma de todas las x elevada cada una al cuadrado, es decir, $S_{x^2} = 4^2 + 6^2 + 7^2 + \ldots = 16 + 36 + 49 + \ldots$.

4. Calcula S_y como la suma de todas las y, es decir, $S_y = 13.4 + 18.8 + 21.4 + 26.3 + 47.9 + 10.3 + 20.3 + 33.4$.

5. Calcula S_{xy} como la suma de todas las $x \cdot y$, es decir, $S_{xy} = 4 \cdot 13.4 + 6 \cdot 18.8 + 7 \cdot 21.4 + \ldots = 53.5 + 112.84 + 150.12 + \ldots$.

6. Determina el número de observaciones como n

7. Calcula $\beta_0 = \frac{S_{x^2} S_y - S_x S_{xy}}{n S_{x^2} - (S_x)^2}$ sustituyendo los valores de $S_{x^2}, S_y, S_x, S_{xy}$

8. Calcula $\beta_1 = \frac{n S_{xy} - S_x S_y}{n S_{x^2} - (S_x)^2}$ sustituyendo los valores de $S_{xy}, S_x, S_y, S_{x^2}$

9. Verifica que $\beta_0 \approx 2$ y $\beta_1 \approx 3$, sino repite los cálculos.

10. Establece el modelo de regresión lineal simple como $y = \beta_0 + \beta_1 \cdot x$

11. Calcula las siguientes predicciones para x

    - Para $x = 0$, ¿Cuánto vale y?
    - Para $x = 5$, ¿Cuánto vale y?
    - Para $x = 10$, ¿Cuánto vale y? ¿Se parece al dato real?
    - Para $x = 15$, ¿Cuánto vale y? ¿Se parece al dato real?
    - Para $x = 20$, ¿Cuánto vale y?

12. Sobre la gráfica de puntos, dibuja con otro color los puntos anteriores calculados para las predicciones y únelos con una recta

13. ¿Se pegan los puntos a la recta de predicción?

14. Visualmente, ¿Cuánto vale una acción (y) si el precio del litro de petróleo (x) fuera de \$8 dólares?

15. Visualmente, ¿Cuánto vale una acción (y) si el precio del litro de petróleo (x) fuera de \$12 dólares?

16. ¿Qué tipos de observaciones x y y se te haría interesante analizar? ¿Qué representa la x? ¿Qué representa la y? ¿Cuándo el valor de x aumenta se espera que el valor de y aumente o disminuya?

Actividad 10

Responde las siguientes preguntas acerca de los problemas vistos en clase, indica la opción más adecuada:

1. ¿Qué es un indicador?
    - (a) Es un valor que se construye a partir de los datos observados, por ejemplo, la diferencia entre estatura y peso o el cociente entre el peso y la estatura al cuadrado.
    - (b) Es un índice que indica cuál es la posición del valor medio y sirve para saber dónde está el dato central.
    - (c) Es el número de experimentos necesarios para determinar si la muestra es representativa de la población.
2. ¿Cómo se construye el índice de Gini y qué representa?
    - (a) Se suma la diferencia de la proporción poblacional y la proporción de los recursos $y_i - x_i$ y representa un nivel sobre la desigualdad de los recursos que está entre 0 y 1.
    - (b) Se suma la multiplicación de la proporción poblacional $x_{i+1} - x_i$ por la proporción de los recursos $y_{i+1} + y_i$, si el valor es cercano a 0 la desigualdad entre los recursos es nula y si el valor es cercano a 1 la desigualdad entre los recursos es absoluta.
    - (c) Se suman las proporciones poblacionales x_i y se dividen entre la suma de la proporción de los recursos y_i, y representa un valor entre 0 y 1 que indica si hay poca o mucha desigualdad.
3. ¿Cómo podemos agregar una nueva columna **C** que contenga la suma de las columnas **A** y **B** en una Tabla de Datos de *pandas* (*DataFrame*)?
    - (a) Usando `datos["C"] = datos["A"] + datos["B"]`
    - (b) Usando `datos.C = datos.A + datos.B`
    - (c) Usando `datos$C <- datos$A + datos$B`

Hoja 21 de Ejercicios

Carga el conjunto de datos llamado `student_habits_performace.csv`.

```
import pandas

students = pandas.read_csv("https://github.com/dragonnomada/" +
                           "ipn-cic-pycien-junio-2025/" +
                           "raw/refs/heads/main/datasets/" +
                           "student_habits_performance.csv")

students.head()
```

- ¿Cuántas columnas observas?
- ¿Cuál será la columna de principal interés?

Hoja 22 de Ejercicios

Analiza y visualiza la serie de datos de las edades de los estudiantes.

```
from matplotlib.pyplot import subplots

age = students["age"]

print(age.describe())

figure, axis = subplots(nrows=2,
                gridspec_kw={"height_ratios": [1,3]})

age.plot.box(ax=axis[0], vert=False)
age.plot.hist(ax=axis[1], density=True, bins=8)
age.plot.density(ax=axis[1])
```

- ¿Cuál es la media de edades?
- ¿Cuál es la mediana de edades?
- ¿Cuál es la edad mínima y máxima?
- Según la caja, ¿Hacia donde hay más edades, arriba o debajo de 20 años?
- ¿Se podría decir que la distribución de edades es uniforme?

Hoja 23 de Ejercicios

Contrasta las horas de estudio contra las horas viendo netflix y separa los datos por los que sacaron más de 70 en el examen y los que no sacaron más de 70 en su examen.

```
import seaborn

seaborn.jointplot(students,
    x="study_hours_per_day", y="netflix_hours",
    hue=students["exam_score"] >= 70)
```

- ¿Cómo se ven los puntos después antes de 4 horas de estudio?
- ¿Cómo se ven los puntos después después de 4 horas de estudio?
- ¿Qué proporción visual de alumnos sacaría más de 70 puntos en su examen si estudia menos de 2 horas?
- ¿Qué proporción visual de alumnos sacaría más de 70 puntos en su examen si estudia más de 6 horas?
- ¿Se puede confirmar que al ver más netflix reprueban más personas?
- Observando las distribuciones horizontales (horas de estudio), ¿Dónde está la máxima área de confusión?
- Observando las distribuciones verticales (horas viendo netflix), ¿Cuál es el porcentaje aproximado que sacaría más de 70 al ver menos de 2 horas de netflix?

Hoja 24 de Ejercicios

Para las columnas categóricas, verifica la creencia sobre si la puntación promedio del examen es mayor o similar en alguna de las categorías. Define una lista de las columnas categóricas y recorre cada columna. Luego agrupa los estudiantes mediante las categorías de esa columna y selecciona la puntuación de exam, finalmente aplica el promedio a la puntación del examen e imprime el reporte. Esto generará en cada impresión un reporte de la puntuación promedio del examen para cada categoría.

```
columnas = [
    "gender",
    "part_time_job",
    "diet_quality",
    "internet_quality",
    "exercise_frequency",
    "parental_education_level",
    "extracurricular_participation",
    "mental_health_rating"
]

for columna in columnas:
    print(students.groupby(columna)["exam_score"].mean())
    print("-" * 40)
```

- ¿Se puede sospechar que la calificación promedio de las mujeres es mayor a la de los hombres o es similar?
- ¿Se puede sospechar que los estudiantes que no tienen un trabajo de medio tiempo, tienen un mejor puntaje en su examen?
- ¿Se puede sospechar que la puntuación promedio en el examen de los estudiantes que tienen mejor alimentación (Fair) es mayor al de estudiantes que tienen peor alimentación (Poor)?
- ¿Se puede sospechar que los alumnos que tienen buen internet (Good) son peores que los que tienen mal internet (Poor)? ¿Cómo explicarías esto?
- ¿Se puede sospechar que la frecuencia de ejercicios aumenta la puntuación promedio del examen en los estudiantes? ¿Cuántos puntos de diferencia habría entre un estudiante que no hace ejercicio contra uno que hace ejercicio 6 días a la semana?
- ¿Se puede sospechar que los estudiantes con padres con licenciatura son mejores que los que tienen padres con maestría? ¿Cómo explicarías esto?
- ¿Se puede sospechar que los estudiantes con actividades extracurriculares son mejores que los que no hacen?
- ¿Se puede sospechar que la salud mental de los estudiantes influye en la puntuación promedio de sus exámenes? ¿Cuánta diferencia hay entre un estudiante con salud mental de 1 y uno con salud mental de 10?

Hoja 25 de Ejercicios

Genera un mapa de calor sobre los estudiantes comparando el trabajo de medio tiempo y las actividades extracurriculares, usando la puntuación del examen promedio como objetivo.

```
reporte = students.groupby([
    "part_time_job", "extracurricular_participation"
])["exam_score"].mean().unstack()

print(reporte)

seaborn.heatmap(reporte, cmap="coolwarm")
```

- ¿En qué combinación hay una mayor puntuación promedio del examen (más rojo)?
- ¿En qué combinación hay una menor puntuación promedio del examen (más azul)?

- ¿Cómo explicas que los estudiantes que trabajan de medio tiempo y hacen actividades extracurricules tengan peor puntuación promedio del examen que los estudiantes que no tienen trabajo de medio tiempo y si hacen activades extracurriculares?

Reto de la semana

En este reto tendrás que investigar cómo construir nuevas series de datos a partir de la combinación de otras series de datos existentes.

1. Crea una serie y que sea la suma de la horas de netflix más las horas en redes sociales menos las horas de estudio
2. Crea una serie x que sea la frecuencia de ejercicio entre 7, más las horas de sueño entre 10, más la calificación de salud mental entre 10
3. Crea una serie z que sean las puntuaciones del examen mayores o iguales a 70
4. Dibuja una gráfica *jointplot* usando $x = x$, $y = y$ y $hue = z$ de tipo *hist*
5. Dibuja una gráfica *scatterplot* usando $x = y$, $y = exam_{score}$ y $hue = exam_{score}$
6. Dibuja una gráfica *scatterplot* usando $x = y \cdot (3 - x)^3$, $y = exam_{score}$ y $hue = exam_{score}$

* Para las últimas gráficas tienes que usar `students["exam_score"]` como $exam_{score}$