

Instituto Politécnico Nacional
Centro de Investigación en Computo

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:

$$y = \beta_0 + \beta_1 \cdot x \quad (1)$$

Donde x representa una variable conocida llamada la *covariable* o *variable predictiva* y y representa una variable desconocida llamada 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 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 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 $x = \{x_1, x_2, \dots, x_n\}$ y $y = \{y_1, y_2, \dots, y_n\}$.

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

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

O en términos simples:

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

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:

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

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

$$\beta_0 = \frac{S_{x^2}S_y - S_xS_{xy}}{nS_{x^2} - (S_x)^2}$$

$$\beta_1 = \frac{nS_{xy} - S_xS_y}{nS_{x^2} - (S_x)^2}$$
(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 estar 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 de 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.

Resuelve 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, $S_{x^2} = 4^2 + 6^2 + 7^2 + \dots = 16 + 36 + 49 + \dots$.
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, $S_{xy} = 4 \cdot 13.4 + 6 \cdot 18.8 + 7 \cdot 21.4 + \dots = 53.5 + 112.84 + 150.12 + \dots$.
6. Determina el número de observaciones como n .
7. Calcula $\beta_0 = \frac{S_{x^2}S_y - S_xS_{xy}}{nS_{x^2} - (S_x)^2}$ sustituyendo los valores de $S_{x^2}, S_y, S_x, S_{xy}$.
8. Calcula $\beta_1 = \frac{nS_{xy} - S_xS_y}{nS_{x^2} - (S_x)^2}$ 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 `movies.csv`.

```
import pandas

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

movies.head()
```

- ¿Qué puedes observar en la columna **title**?
- ¿Qué puedes observar en la columna **genres**?

Hoja 22 de Ejercicios

Extrae el año de cada película codificada con el patrón regular donde hay más de un dígito del 0 al 9.

```
movies["Year"] = movies["title"].str.extract(r"\(([0-9]+\)\)")
movies["Year"] = movies["Year"].fillna("0").astype("int64")
movies.head()
```

- ¿Por qué a la columna **title** se le llama al método `str.extract`?
- ¿Por qué se usa la `r` antes del patrón `"[0-9]+"`?
- ¿Qué representa la extracción del patrón regular `"[0-9]+"`?
- ¿Por qué se reajusta la nueva columna **Year** con `fillna` y `astype`?
- ¿Qué hace el código en general?

Hoja 23 de Ejercicios

Obtén los géneros de todas las películas. Separa la columna **genres** por `|` y linealiza los valores con `reshape(-1)`, luego extrae el conjunto de todos los distintos omitiendo `None`.

Finalmente, para cada género en el conjunto de géneros, agrega una nueva columna con el género y si la película contiene ese género, haciendo una búsqueda del género. Convierte la búsqueda en entero, para tener 0 si la película no contiene el género y 1 si la película contiene ese género.

```
genres = set(movies["genres"].str\
              .split("|", expand=True).values.reshape(-1)) - {None}

for genre in genres:
    movies[f"genre_{genre}"] = movies["genres"].str\
                                .match(genre).astype(int)

movies.head()
```

- ¿Cuántas columnas se generaron?
- Para la película 2 de Jumanji (1995), ¿Qué generos tienen 1?

Hoja 24 de Ejercicios

Para las películas del año 2000 o superior, compara el porcentaje (media) de películas con el género de acción y películas con el género de comedia a 2 decimales.

```
movies[movies["Year"] >= 2000]\
    [["Year", "genre_Action", "genre_Comedy"]]\
    .groupby("Year").mean().round(2)
```

- ¿En qué años el porcentaje de películas de acción superan o son similares a las de comedia?
- ¿Qué género de películas es más visto del año 2000 al 2018, de acción o de comedia?
- ¿Qué género de películas es más visto del año 1990 a 1999, de acción o de comedia? Pista: Filtra las películas mayores a 1990 y menores al 2000

```
movies[(movies["Year"] >= 1990) & (movies["Year"] < 2000)]
```

Hoja 25 de Ejercicios

Visualiza la comparación de los géneros de acción y comedia del año 2000 o superior.

```
movies[movies["Year"] >= 2000]\
    [["Year", "genre_Action", "genre_Comedy"]]\
    .groupby("Year").mean().plot.bar()
```

- ¿En qué año las películas de acción superan por mucho a las de comedia? ¿Cómo explicarías este fenómeno?
- ¿Cómo se ve la comparación entre las películas de Aventura contra las de Drama?

- ¿En qué año las películas de Aventura superan a las de Drama y cómo se explicaría?

Reto de la semana

Crea una matriz que cuente cuántas películas hay en los años del 2000 en adelante para los géneros de Acción, Comedia, Aventura, Drama, Animación, Documental y Romance. Luego muestra un mapa de calor que visualice esta información (la matriz de conteos).

- ¿Cuál es el genero más alto y en qué año?
- ¿Qué géneros tienen valores más altos en todos los años?
- ¿Qué géneros tienen valores más bajos en todos los años?
- ¿Cuál género tenía valores altos excepto en el año 2018?