HENRY

# Data Science

**Hands on - M4L2 | Aprendizaje supervisado I: Regresión**

→ soyhenry.com

# Objetivos

- Examinar los supuestos fundamentales de la regresión lineal, evaluando su impacto en la validez del modelo y su aplicación en datos reales.
- Interpretar los coeficientes de regresión, analizando su significado y relevancia en la relación entre las variables independientes y dependientes.
- Evaluar el rendimiento del modelo mediante el coeficiente de determinación y métricas de error, comparando regresión lineal simple, múltiple y polinómica.

# Agenda

COMENCEMOS →

#TEMAS

.01 Introducción a la regresión lineal: supuestos y definiciones

.02 Evaluación de rendimiento

.03 Regresión lineal polinómica y regresión lineal múltiple

.04 Homework

<-->

# ¿Qué vimos en la lecture?

<01>

# Introducción a la regresión lineal: supuestos y definiciones

# Idea geométrica de la regresión

La regresión lineal busca una **recta** que describa cómo cambia una variable dependiente (y) según una independiente (x).

**Fórmula:** $y = \beta_0 + \beta_1 x + \varepsilon$

- $\beta_0$ → intercepto (valor de y cuando x=0)
- $\beta_1$ → pendiente (cuánto cambia y por cada unidad de x)
  **Ejemplo CityScoot:** x = temperatura, y = cantidad de viajes.

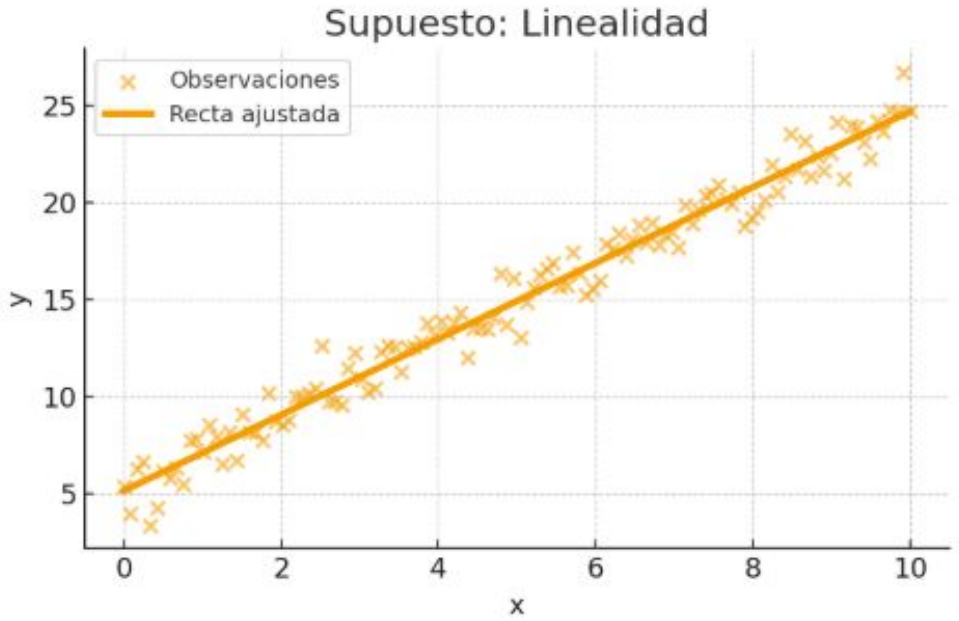
# Intuición geométrica

Cada punto representa un día con su temperatura y número de viajes.

La recta ajustada minimiza la **suma de errores al cuadrado**:

→ cuanto menor la distancia entre los puntos reales y la recta, mejor el modelo.

**Pendiente positiva:** más calor → más viajes.



# Supuestos del modelo

Para que el modelo sea válido, deben cumplirse ciertos supuestos:

- ★ **Linealidad:** la relación entre x e y se aproxima a una recta.
- ★ **Independencia de errores:** los residuos no deben tener patrón.
- ★ **Homocedasticidad:** varianza constante de errores (sin “embudo”).
- ★ **Normalidad:** los residuos siguen una distribución normal.
- ★ **Ausencia de multicolinealidad:** las x no deben aportar la misma información.

# Interpretación de coeficientes

**Ecuación:** rides = 250 + 18.5 × temp

- $\beta_0$ = 250 → viajes base aun con baja temperatura.
- $\beta_1$ = 18.5 → por cada grado más, +18.5 viajes.

**Sentido práctico:** cada coeficiente explica cuánto influye una variable en la demanda.

# Cálculo de $\beta_0$ y $\beta_1$

Ejemplo con 5 días:

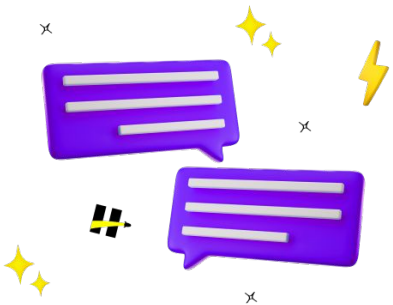
Temp = [10,15,20,25,30] → Viajes = [450,540,620,710,800]

Resultado: $\beta_1 = 17.4$ y $\beta_0 = 276$

**Modelo final:** rides = 276 + 17.4 × temp

→ Con 0 °C → 276 viajes esperados.

→ Por cada grado más → +17.4 viajes.



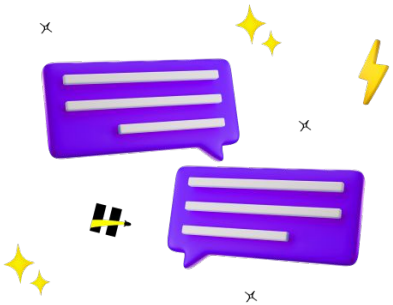
<02>

# Evaluación de rendimiento

# Evaluar el rendimiento del modelo

**Coeficiente de determinación ($R^2$):** mide cuánta variabilidad de y explica el modelo.

- $R^2 = 1$ → ajuste perfecto.
- $R^2 \approx 0$ → modelo casi no explica nada.



# Más allá del $R^2$

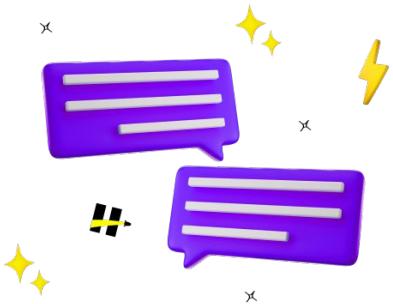
$R^2$ no mide el **error real**, solo el grado de explicación.
Por eso se complementa con:

- **MAE (Error Absoluto Medio)**

- **RMSE (Raíz del Error Cuadrático Medio)**
  Ambas expresan el error en las mismas unidades que y (ej. “viajes”).

# MAE y RMSE

- **MAE:** promedio del error absoluto.
  → “Cuánto se equivoca el modelo, en promedio.”
- **RMSE:** similar, pero penaliza más los errores grandes.
  → “Sensibilidad ante grandes desvíos.”
- **CityScoot:**
  MAE = 18 → se equivoca en promedio 18 viajes.
  RMSE = 18.44 → penaliza los errores grandes.



# Comparación de métricas

| Métrica | Qué mide | Cuándo usarla |
|---|---|---|
| **$R^2$** | % de variabilidad explicada | Comparar poder explicativo |
| **MAE** | Error medio en unidades reales | Fácil de interpretar |
| **RMSE** | Error penalizado por magnitud | Detectar errores grandes |

<03>

# Regresión lineal polinómica y regresión lineal múltiple

# Regresión polinómica

A veces la relación **no es lineal.**
Ejemplo CityScoot:

- A medida que sube la temperatura, la demanda crece, pero cae con calor extremo (> 35 °C).

**Solución:** agregar un término cuadrático o cúbico (temp², temp³).
→ Captura **curvas** de comportamiento real.

# Regresión lineal múltiple

La demanda depende de más factores:

**Modelo:** rides = $\beta_0 + \beta_1 temp + \beta_2 rain + \beta_3 marketing + \beta_4 price$

Permite medir efectos simultáneos:

- +500 USD en marketing → +40 viajes.
- +5 mm de lluvia → −30 viajes.

  Refleja mejor la complejidad del fenómeno real.

# Regularización Ridge y Lasso

Más variables = riesgo de **sobreajuste** y **multicolinealidad.**

- **Ridge:** reduce coeficientes → más estables.
- **Lasso:** puede llevar algunos coeficientes a **0** → selección automática de variables.

**Ejemplo CityScoot:**
Ridge estabiliza predicciones con variables redundantes.
Lasso simplifica el modelo, dejando solo las variables más relevantes.

<DATA SCIENCE/>

# Vayamos a la práctica

## Homework

# Consigna

La empresa ficticia **RetailBoost** desea comprender mejor el comportamiento de sus clientes. Para ello, se dispone de un dataset con información demográfica y de consumo. El objetivo es **predecir el monto de gasto mensual de los clientes** en la tienda online, a partir de variables como edad, ingresos, historial de visitas y otros factores.

Tu tarea será:

1. Realizar un **EDA exploratorio** sobre el dataset.
2. Ajustar al menos **dos modelos de regresión** (ej. regresión lineal simple, múltiple o polinómica).
3. Evaluar el rendimiento de los modelos utilizando métricas como **$R^2$, MAE y RMSE**.
4. Interpretar los resultados, explicando qué variables parecen influir más en el gasto mensual.

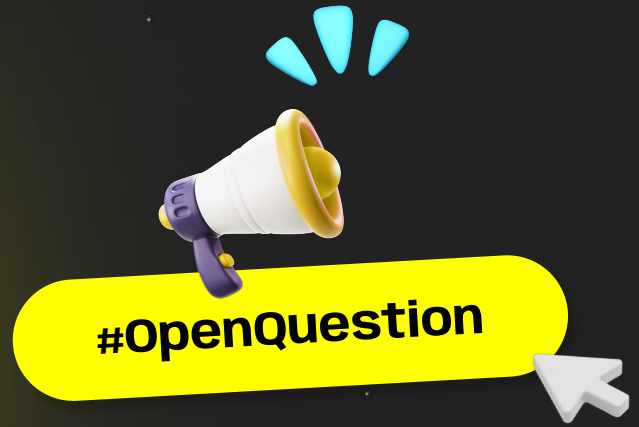
# Tareas a realizar

1. Importar y explorar correctamente el dataset.
2. Mostrar la distribución de la variable objetivo (`monthly_spent`) y de las variables predictoras principales.
3. Ajustar al menos **un modelo de regresión lineal simple** y **uno múltiple**.
4. Calcular las métricas $R^2$, MAE y RMSE en un conjunto de prueba.
   Explicar de forma breve qué tan bien se ajusta el modelo y si los errores son razonables.

# Extra credits

- Probar una **regresión polinómica** para capturar relaciones no lineales.
- Visualizar **y vs. y_pred** y el gráfico de **residuales** para verificar los supuestos de la regresión.
- Comparar el rendimiento entre distintos modelos en una tabla resumen.

HENRY




## ¿Preguntas?

HENRY

# ¡Muchas gracias!

→ soyhenry.com