Aprendizaje Automático: Regresión

Máster en Inteligencia Artificial Avanzada y Generativa

David Rey Blanco - david.rey@mbitschool.com

mbit school

presentación

David Rey

Objetivos del bloque

- Introducción a los modelos de regresión lineal: simple y múltiple
- Utilidad de estos modelos
- Retos
- Regresión polinómica
- Regresión generalizada
- Árboles de regresión

Estructura de las sesiones

Sesión 1 (5h)	**Sesión 2** (5h)
Regresión simple	Regresión generalizada
Regresión Múltiple	Árboles de regresión
Evaluación del ajuste	Taller final
Otros tipos de regresión	

¿Qué es un modelo de regresión?

Un **modelo de regresión** es una herramienta estadística que permite **estimar la relación entre una variable dependiente y una o más variables explicativas**, con el objetivo de **predecir valores y comprender cómo cambian los resultados cuando varían los predictores**.

Ejemplo: *predicción del precio de una vivienda, temperatura dentro de una semana, etc …*

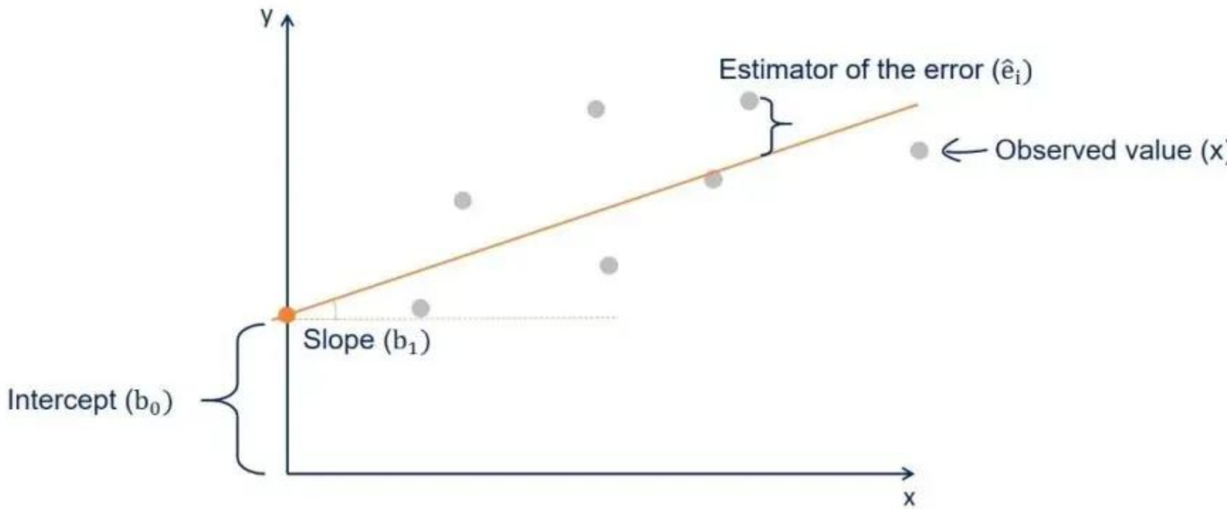
Sesión 1

Modelo lineal simple

- Ecuación: $y = \beta_0 + \beta_1 x + \varepsilon$
- Objetivo: estimar la relación lineal entre x e y.

Linear regression model. Geometrical representation

$$\hat{y}_i = b_0 + b_1 x_i$$



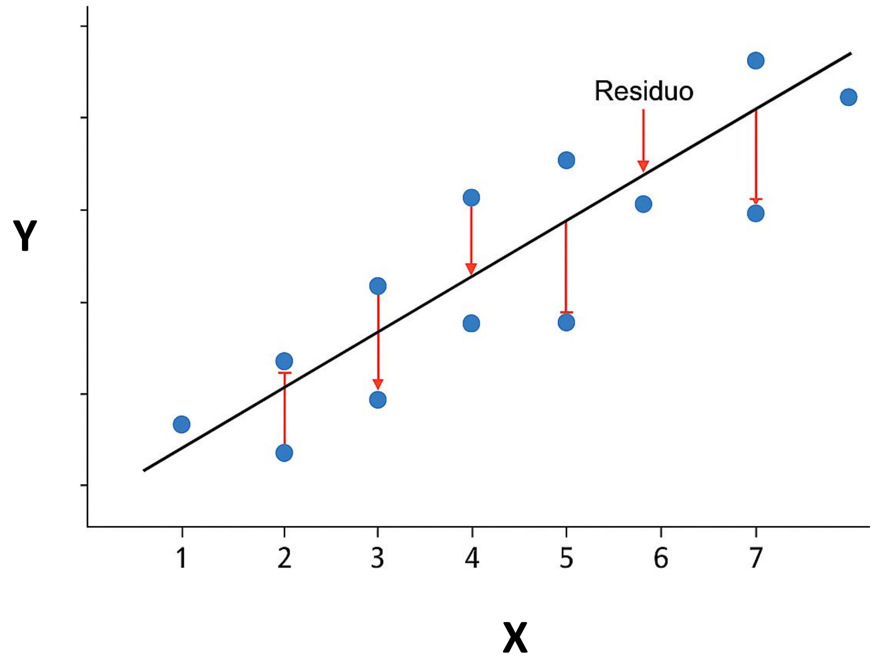
Regresión por mínimos cuadrados

Es una forma de resolver el problema es a través de la **regresión por mínimos cuadrados ordinarios (OLS)** busca ajustar una recta (o hiperplano) que **minimiza la suma de los errores al cuadrado** (entre los valores observados y los predichos.

- Ajusta la línea que mejor representa la relación entre variables.
- Cada punto tiene un “error” vertical respecto a la línea.
- OLS elige los coeficientes que minimizan el **cuadrado** de esos errores.

β_1 = cov(x,y) / var(x)

β_0 = mean(y) - β_1 x mean(x)



Regresión múltiple

- Ecuación: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{\square} x_{\square} + \varepsilon$
- Permite múltiples predictores.

$$B = (X^T X)^{-1} X^T y$$

Interpretación de los coeficientes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{\square} x_{\square} + \varepsilon$$

- Cada coeficiente $\beta_{\square}$ mide **el cambio esperado en la variable respuesta** y **cuando** x_j **aumenta en una unidad**, manteniendo las demás variables constantes.
- En escala normal (no estandarizada), los coeficientes están en las **unidades originales** de las variables.

🧮 Ejemplo

Y = 25.000 (β_0) + 800 x Superficie en metros cuadrados (β_1) + 12.000 x Habitaciones (β_2)

- Si la **superficie** aumenta en 1 m², el **precio** aumenta en promedio **800 €** (manteniendo las demás variables fijas).
- Cada **habitación adicional** incrementa el precio en **12 000 €**, en promedio.

⚠️ Consideraciones importantes

- **Escala de las variables** afecta la magnitud de los coeficientes.
- No compara importancia entre variables con diferentes unidades.
- Para comparar impacto relativo → usar **coeficientes estandarizados** (modelo con variables escaladas).

Evaluación de modelos de regresión

- Métricas principales:
- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 (Coeficiente de determinación)

Validación del modelo

- Train/Test Split
- Cross-validation (K-fold)
- Importancia del escalado de variables

Supuestos del modelo lineal

- 1. Linealidad
- 2. Independencia de errores
- 3. Homocedasticidad
- 4. Normalidad de errores
- 5. Independencia de variables

En la práctica, casi nunca se cumplen

Los datos del mundo real son **ruidosos, incompletos y heterogéneos**, por lo que:

Supuesto	En el mundo real	Consecuencia
Linealidad	Relación parcialmente lineal o con umbrales	Sesgo sistemático
Independencia	Observaciones agrupadas (por barrio, cliente…)	Errores correlacionados
Homocedasticidad	Varianza mayor en valores extremos	Inferencias no fiables
Normalidad	Residuos asimétricos o con colas pesadas	Intervalos y tests menos válidos
No colinealidad	Variables socioeconómicas correlacionadas	Coeficientes inestables

Si es tan complicado

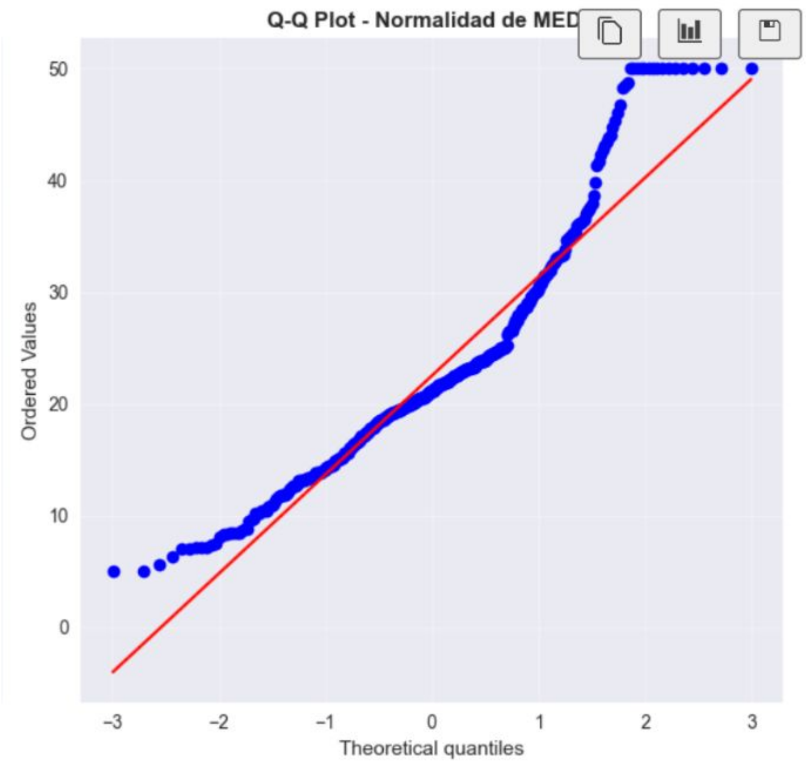
- Es **interpretable, rápida y robusta** si se usa con cuidado.
- Muchas violaciones **no invalidan la predicción**, pero **sí la interpretación.**
- Se pueden **corregir o mitigar**:

Diagnóstico de supuestos

- Análisis de residuos
- Gráficos de residuos vs predichos
- Multicolinealidad (VIF)
- Correlación entre variables

Análisis de la regresión

El paquete **statsmodels** nos da herramientas de diagnóstico del ajuste



Ejemplo

```
dataset = pd.read_csv('../data/50_Startups.csv')
```

✓ 0.0s

```
dataset.head()
```

✓ 0.0s

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

```
from sklearn.linear_model import LinearRegression
regression = LinearRegression()
regression.fit(X_train, y_train)
```

✓ 0.0s

Interpretación de los coeficientes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{\square} x_{\square} + \varepsilon$$

¿Donde está esto?

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.102e+05	1757.977	62.700	0.000	1.07e+05	1.14e+05
x1	3.858e+04	2997.976	12.870	0.000	3.25e+04	4.47e+04
x2	777.8871	1977.177	0.393	0.697	-3265.894	4821.668
x3	4047.7194	2817.646	1.437	0.162	-1715.014	9810.453
x4	185.2116	1982.934	0.093	0.926	-3870.343	4240.766
x5	147.2751	1944.330	0.076	0.940	-3829.326	4123.876

Omnibus:	12.556	Durbin-Watson:	2.432
Prob(Omnibus):	0.002	Jarque-Bera (JB):	14.600
Skew:	-1.017	Prob(JB):	0.000676
Kurtosis:	5.423	Cond. No.	3.14

Interpretación de los coeficientes

	coef	std_error	t	p_value	ci_lower	ci_upper
const	110225.320571	1757.977368	62.700079	1.607698e-32	106629.853149	113820.787994
x1	38584.381135	2997.976241	12.870142	1.623668e-13	32452.831261	44715.931009
x2	777.887099	1977.176879	0.393433	6.968772e-01	-3265.893661	4821.667860
x3	4047.719444	2817.646393	1.436560	1.615443e-01	-1715.014480	9810.453367
x4	185.211597	1982.933570	0.093403	9.262259e-01	-3870.342919	4240.766112
x5	147.275100	1944.330004	0.075746	9.401414e-01	-3829.326257	4123.876458

Error estándar

Rango en el intervalo de confianza del 95%

Significancia estadística (cuanto más pequeño mejor)

Conclusión: Solo la variable **`R&D Spend (x1)`** es estadísticamente significativa ($p < 0.05$).

Las demás variables **no muestran evidencia de efecto significativo** sobre la variable objetivo.

Interpretación de los coeficientes - tests

🔷 Omnibus y Prob(Omnibus)

- Evalúan la **normalidad de los residuos** del modelo (combinan los tests de Skewness y Kurtosis).
- **Hipótesis nula (H_0):** los residuos siguen una distribución normal.
- **Interpretación:**
  - Si *Prob(Omnibus)* < 0.05 → se **rechaza la normalidad** (los residuos no son normales).
  - Si *Prob(Omnibus)* ≥ 0.05 → no hay evidencia para rechazar la normalidad.

🔷 Durbin–Watson

- Mide la **autocorrelación** de los residuos (dependencia entre errores consecutivos).
- Valores posibles: de 0 a 4.
  - ≈ 2 → **sin autocorrelación** (ideal).
  - < 2 → **autocorrelación positiva**.
  - 2 → **autocorrelación negativa**.
- Importante especialmente en **series temporales**.

🔷 Jarque–Bera (JB) y Prob(JB)

- Otro test de **normalidad de los residuos**, basado en su **asimetría (Skew)** y **curtosis (Kurtosis)**.
- **Hipótesis nula (H_0):** los residuos son normales.
- **Interpretación:**
  - *Prob(JB)* < 0.05 → se rechaza la normalidad.
  - *Prob(JB)* ≥ 0.05 → no se rechaza la normalidad.

Interpretación de los coeficientes - tests

🔷 Skew (Asimetría)

- Indica si la distribución de los residuos está **sesgada**.
    - Skew = 0 → distribución simétrica.
    - Skew < 0 → sesgo hacia la izquierda.
    - Skew > 0 → sesgo hacia la derecha.

🔷 Kurtosis

- Mide el **grado de concentración** o "altura" de la distribución.
    - Kurtosis = 3 → distribución normal.
    - Kurtosis > 3 → **leptocúrtica** (colas más pesadas).
    - Kurtosis < 3 → **platicúrtica** (colas más ligeras).

🔷 Cond. No. (Número de Condición)

- Evalúa posibles problemas de **multicolinealidad** entre las variables independientes.
- Valores altos (> 30) pueden indicar **colinealidad severa**, que afecta la estabilidad de los coeficientes.

Interpretación de los coeficientes - tests

Omnibus:	12.556	Durbin-Watson:	2.432
Prob(Omnibus):	0.002	Jarque-Bera (JB):	14.600
Skew:	-1.017	Prob(JB):	0.000676
Kurtosis:	5.423	Cond. No.	3.14

En este caso:

- *Prob(Omnibus)* y *Prob(JB)* son **menores a 0.05** → los residuos **no son perfectamente normales**.
- *Durbin–Watson ≈ 2.4* → **no hay autocorrelación significativa**.
- *Kurtosis > 3* y *Skew < 0* → la distribución de los residuos es **asimétrica y con colas pesadas**.
- *Cond. No. = 3.14* → no hay evidencia de multicolinealidad.

La evaluación de la autocorrelación es principalmente útil en modelos autorregresivos de series temporales

Colinealidad

Problema de la colinealidad

Un modelo puede predecir bien (bajo error global) pero aún así tener **coeficientes que no representan relaciones reales** entre variables

- **Coeficientes inestables:** pequeños cambios en los datos → grandes cambios en los betas.
- **Signos y magnitudes incoherentes:** una variable puede parecer negativa cuando en realidad tiene efecto positivo.**Buen ajuste global (R^2 alto)**, pero 👉 **coeficientes no interpretables individualmente.**
- **Errores estándar inflados:** menor significación estadística.

Dos variables muy correlacionadas (ej. superficie y número de habitaciones):

- Ambas explican precio, pero **se solapan** en información.
- El modelo no puede “decidir” cuál tiene el efecto real.

Detección y solución

- **Matriz de correlación** → detectar correlaciones altas ($|r| > 0.8$).
- **VIF (Variance Inflation Factor)** → medir la inflación de varianza.
  - VIF > 5 o 10 → posible colinealidad.
- **Soluciones:** eliminar variables correladas, regularizar, eliminar colinealidad con PCA

VIF

El **VIF (Factor de Inflación de la Varianza)** cuantifica cuánto se **incrementa la varianza de un coeficiente** debido a la **colinealidad** con otras variables. Indica **cuánto "se hincha" la incertidumbre** en la estimación de B_j al no ser independiente de los demás predictores.

$$\textbf{VIF} = 1 / (1 - R_j^2)$$

donde R^2 es el coeficiente de determinación al **ajustar un modelo de regresión de** X_j sobre **todas las demás variables** (es decir la capacidad de reconstruir esta variable en función del resto de variables).

Cuanto mayor es R_j^2 (menor es el numerado = más cercano a cero) y por tanto mayor inflación causado por esa variable:

- 1 => No hay colinealidad
- 1-5 => Colinealidad reducida, generalmente aceptable
- 5-10 => Colinealidad elevada, revisar / transformar las variables
- >10 => Colinealidad excesiva, recomendable eliminar o combinar variables

Selección de variables

`SequentialFeatureSelector` Es un **método “wrapper” de selección de variables** que construye modelos repetidamente para *añadir o quitar* variables en pasos, buscando el subconjunto que **optimiza una métrica** (accuracy, R^2, RMSE negativo, ROC-AUC, etc.) mediante **validación cruzada**.

- forward=True → *Forward Selection*: empieza sin variables y va **añadiendo** la que más mejora la métrica.
- forward=False → *Backward Elimination*: empieza con todas y va **eliminando** la menos útil.
- floating=True → variantes *SFFS/SBFS* (permite dar pasos hacia atrás/adelante intercalados para escapar de elecciones miopes).
- Controlas cuántas variables quieres con k_features (p. ej., 5, (3, 8) para buscar el mejor tamaño entre 3 y 8, o "best").

Transformación logarítmica

Caso 1: Log-transformación de la variable dependiente

$$\log(y) = \beta_0 + \beta_1 x_1 + \varepsilon$$

- El coeficiente $\beta_1$representa **cambios porcentuales en** y.
- Si x_1 aumenta en una unidad, y cambia en promedio un 100 \times \beta_j**%**.

Ejemplo:

$$\log(\text{precio}) = 9.2 + 0.04\ \text{Superficie}$$

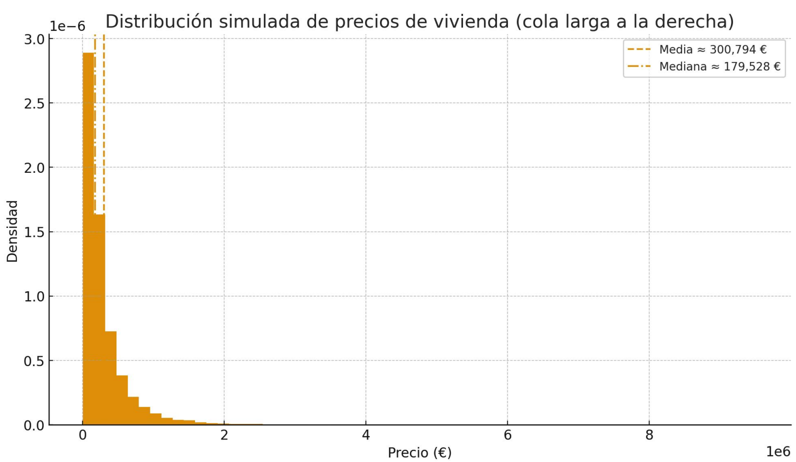
→ Cada m² adicional se asocia con un **aumento del 4 %** en el precio promedio.

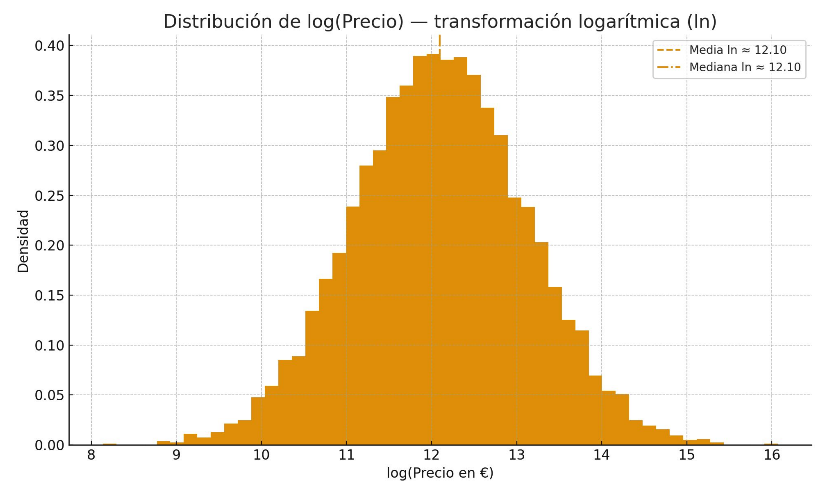
Interpretación de los coeficientes escala logarítmica

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

🧩 Contexto

- Cuando aplicamos una **transformación logarítmica** a la variable dependiente o a los predictores, **la interpretación de los coeficientes cambia**.
- Es útil cuando los datos presentan **distribuciones sesgadas** o **relaciones no lineales**.





Transformación logarítmica

Caso 2: Log-transformación de una variable explicativa

$$y = \beta_0 + \beta_1 \log(x_1) + \varepsilon$$

- Si x aumenta un **1 %**, y cambia en promedio 0.01 * β_1 unidades.
- Interpreta la relación como **elasticidad parcial**.

Ejemplo:

Precio = 20,000 + 12.000 log(Ingresos)

→ Un aumento del **1 %** en los ingresos promedio se asocia con un incremento de **120 €** en el precio.

Transformación logarítmica

Caso 3: Log-log model

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \varepsilon$$

- **β_1** se interpreta directamente como una **elasticidad**:
  Si x aumenta 1 %, y cambia en promedio **β_1**%
- Muy usado en modelos económicos (por ejemplo, precio–ingreso).

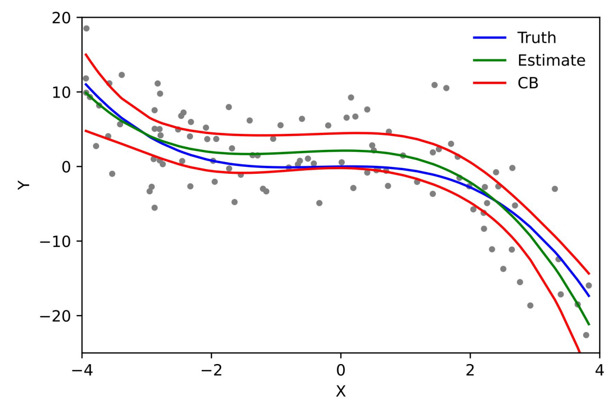
Regularización y sobreajuste

- Problema: el modelo puede aprender ruido del entrenamiento.
- Solución:
  - penalizar la complejidad
  - eliminar el impacto de los outliers

Otras regresiones: polinómica

La **regresión polinómica** es una extensión de la regresión lineal que permite modelar relaciones no lineales.

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$



Características

- Captura curvaturas y patrones complejos
- Lineal en los parámetros (no en x)
- El grado del polinomio controla la flexibilidad

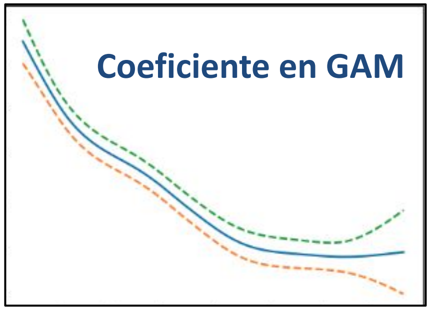
Temas a tener en cuenta

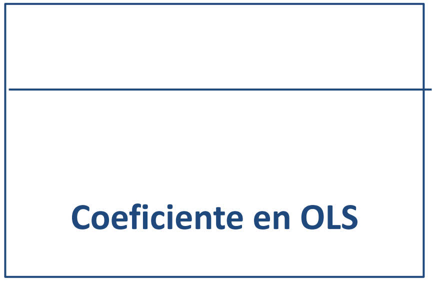
- Grados altos → riesgo de sobreajuste (overfitting)
- Usar validación cruzada y regularización
- Escalar variables para estabilidad numérica

Otros tipos de regresión lineal (GAM)

Recordamos lo que es una regresión lineal “ordinaria”

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$$





- Cada predictor x_i tiene un efecto lineal sobre y.
- Funciona bien para relaciones en línea recta,
- Falla cuando los efectos son no lineales.

Generalized Additive Model (GAM):

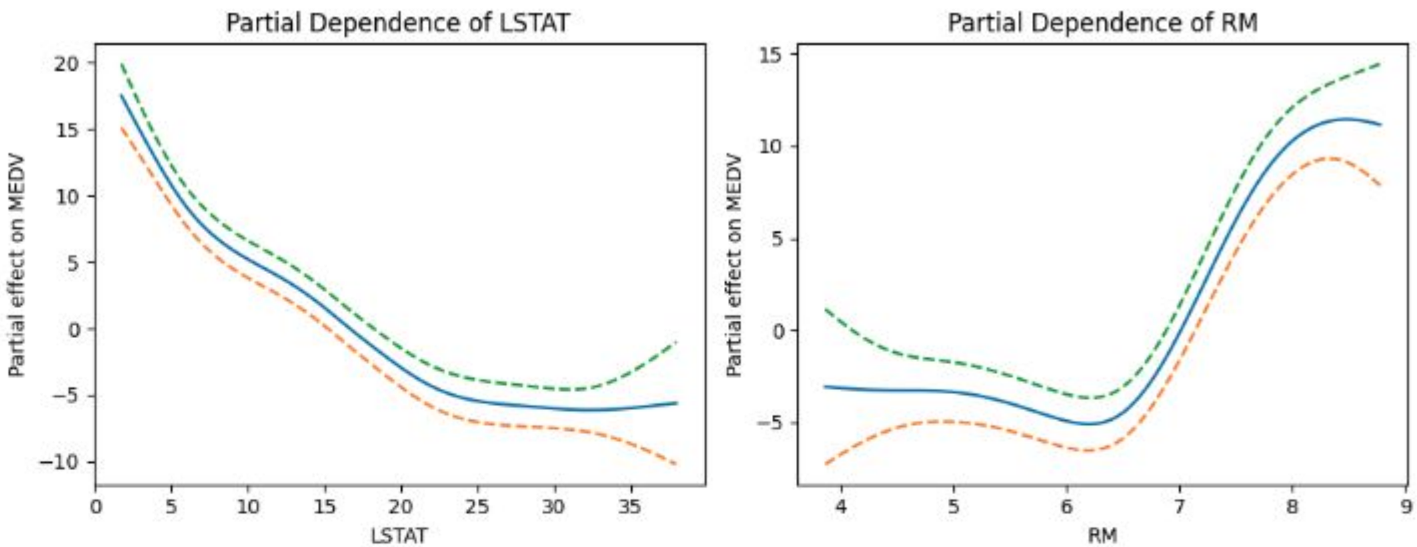
$$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots$$

✅ Cada $f_i(x_i)$ es una **función de suavizado**.

✅ Captura **tendencias no lineales** manteniendo la **interpretabilidad** del modelo.

Coeficientes como funciones (GAM)

Los modelos lineales ordinarios asumen un **efecto constante** de las variables para todos sus valores (lo que no es cierto). Podemos verlo en la influencia de las características en el precio de la vivienda.



Datos: Boston Housing

- LSTAT = porcentaje de población con bajo nivel socioeconómico.
- Se observa una **relación negativa no lineal**: a medida que LSTAT aumenta, el efecto sobre el precio disminuye fuertemente.
- **A mayor proporción de población de bajos ingresos, menor valor medio de las viviendas**.

- RM = número promedio de habitaciones por vivienda.
- Se aprecia una **relación positiva no lineal**: A medida que RM aumenta, el efecto sobre MEDV también aumenta.
- **Las casas con más habitaciones tienden a tener un valor medio más alto**.

Receta a la hora de modelizar

- Exploración de datos y limpieza
- Transformación de variables
- Construcción del modelo
- Evaluación
- Refinamiento:

Sesión 2

Modelos lineales generalizados (GLM)

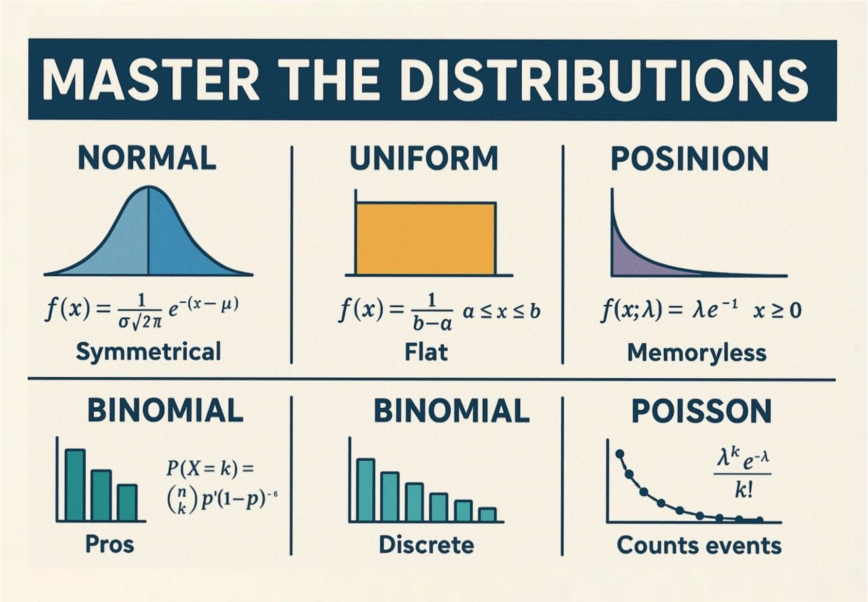
Los **GLM (Generalized Linear Models)** amplían la regresión lineal clásica para poder modelar **variables dependientes que no son continuas y normales**, como conteos, proporciones o datos binarios.

$$g(E[Y]) = X\beta$$

- **Y:** variable respuesta (dependiente).
- **E[Y]: media esperada** de la respuesta.
- **g(.): función de enlace**, que conecta la media E[Y] con la combinación lineal de predictores.
- **Xβ:** combinación lineal de las variables explicativas.

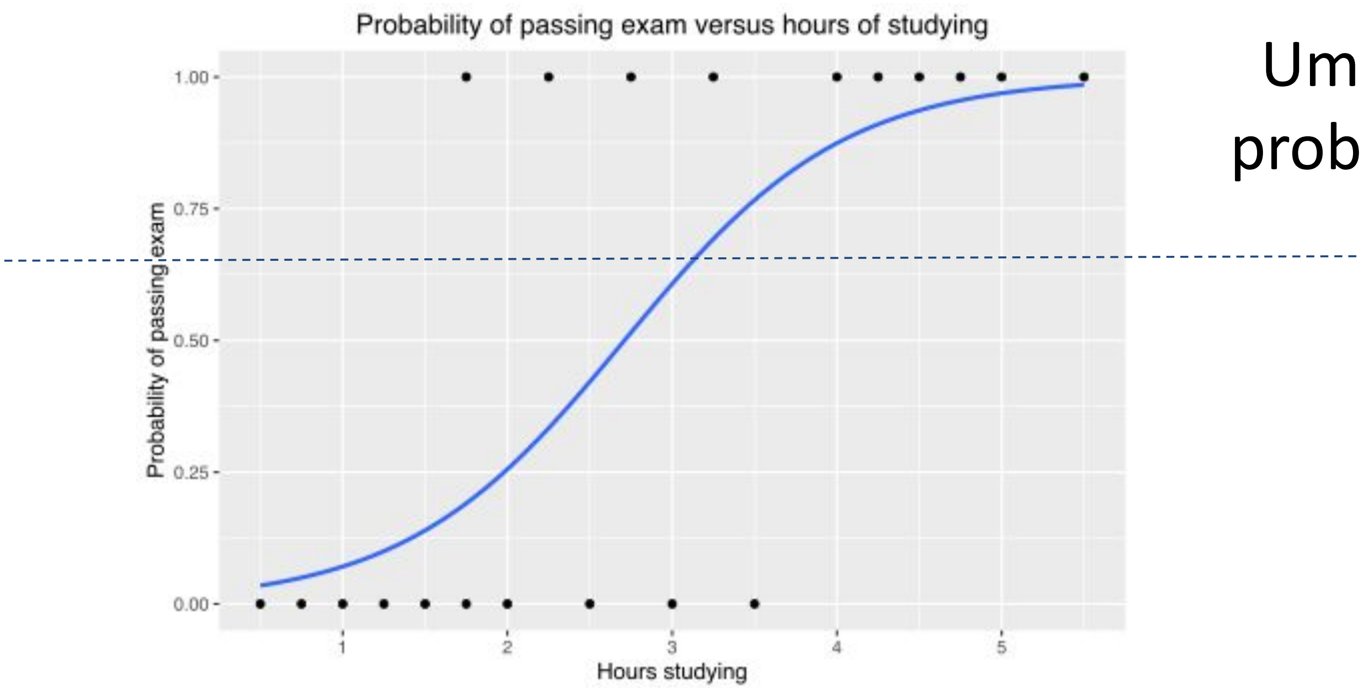
Tipos funciones de enlace

- Gaussiana (regresión lineal): Estima un valor continuo
- Binomial (logística) - Estima la probabilidad de un suceso binario
- Poisson (conteos) - Estima el **número esperado de ocurrencias de un evento** en un intervalo



Ejemplo GLM: Regresión logística

- Modelo: $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$
- Predice probabilidad de un suceso



Umbral de probabilidad

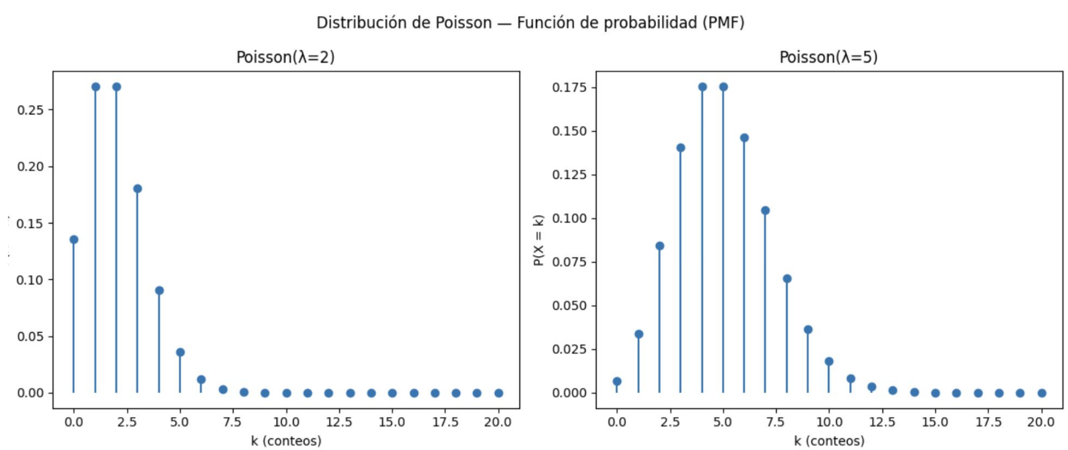
Ejemplo GLM: Regresión Poisson

La **distribución de Poisson** modela el **número de eventos** que ocurren en un intervalo fijo de tiempo o espacio, bajo estas condiciones:

- Los eventos ocurren **independientemente** entre sí.
- La **tasa media** de ocurrencia por unidad (λ, “lambda”) es **constante**.
- La probabilidad de más de un evento en un intervalo muy pequeño es despreciable

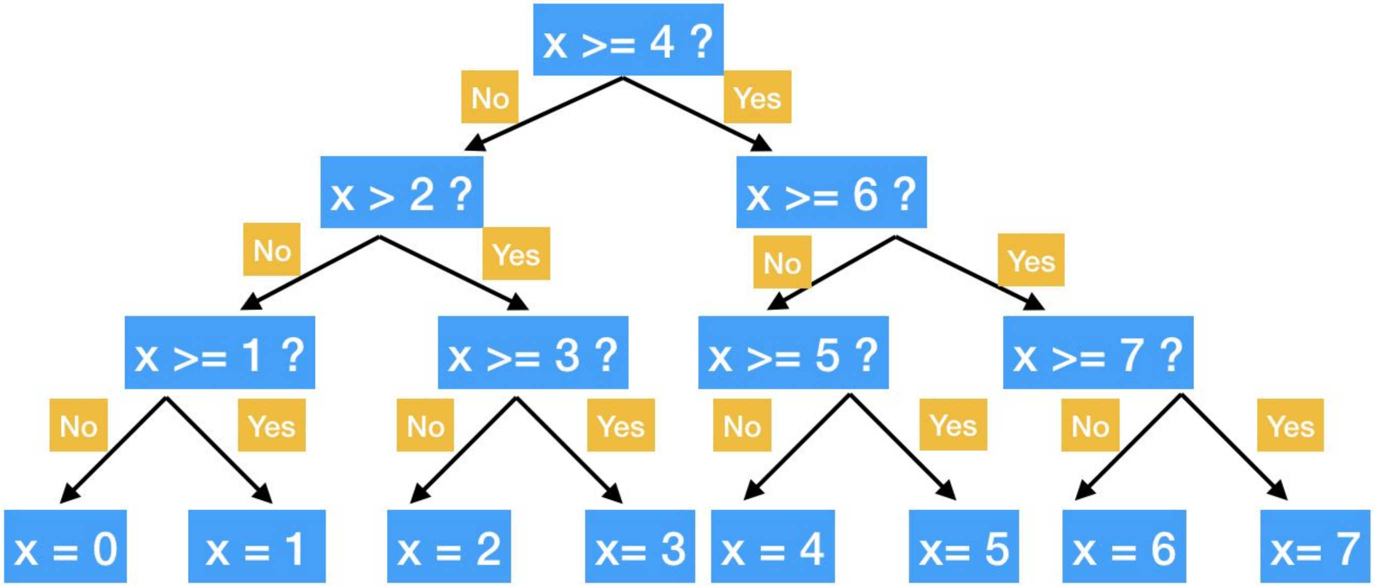
Procesos que modela:

- Llegadas de clientes a una cola (por minuto).
- Número de llamadas a un call center por hora.
- Conteo de defectos en una longitud de material.
- Casos de un suceso raro por unidad (p. ej., mutaciones por Mb)



Árboles de regresión

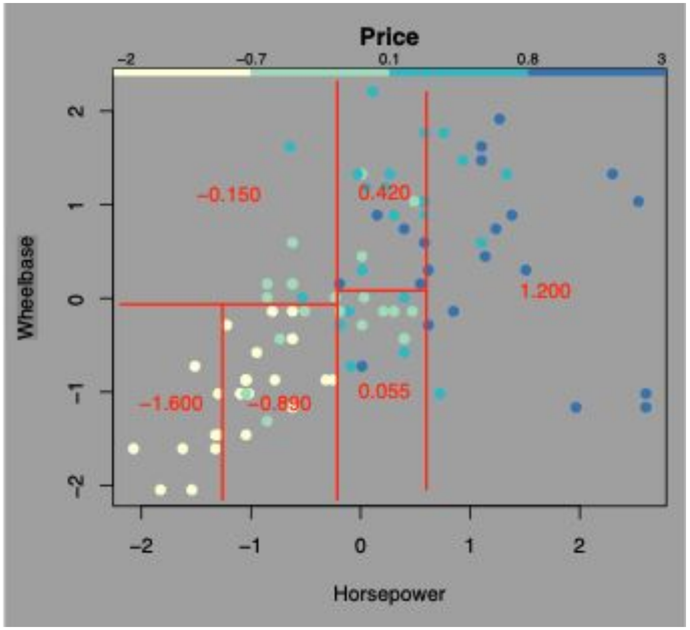
Regresión simple



Árboles de regresión

Idea principal:

- Cada división reduce la variabilidad dentro de los grupos.
- El objetivo es formar nodos lo más puros posible (valores de Y similares dentro de cada región).
- Resultado: una función por tramos constantes que aproxima Y de manera flexible y no lineal.



Reglas prácticas: Feature Engineering

- Transformar variables para mejorar el modelo:
- Escalado
- Variables polinomiales
- Interacciones
- One-Hot Encoding

Reglas prácticas: Selección de variables

- SelectKBest
- Recursive Feature Elimination (RFE)
- Importancia de coeficientes

Repaso final

Sesión 1 (5h)	**Sesión 2** (5h)
Regresión simple	Regresión generalizada
Regresión Múltiple	Árboles de regresión
Evaluación del ajuste	Taller final
Otros tipos de regresión	

¡Gracias!

David Rey Blanco
david.rey@mbitschool.com

Regularizaciones: Ridge y Lasso

- Lasso: penalización L1 (Penaliza el la **suma del valor absoluto** de los coeficientes (favorece coeficientes menores)
- Ridge: penalización L2 (Penaliza el la **suma de los cuadrados** de los coeficientes (favorece coeficientes menores)
- ElasticNet: combinación de ambas

Amplia el objetivo a minimizar:

- Suma de errores cuadráticos + $\lambda \times$ Penalización