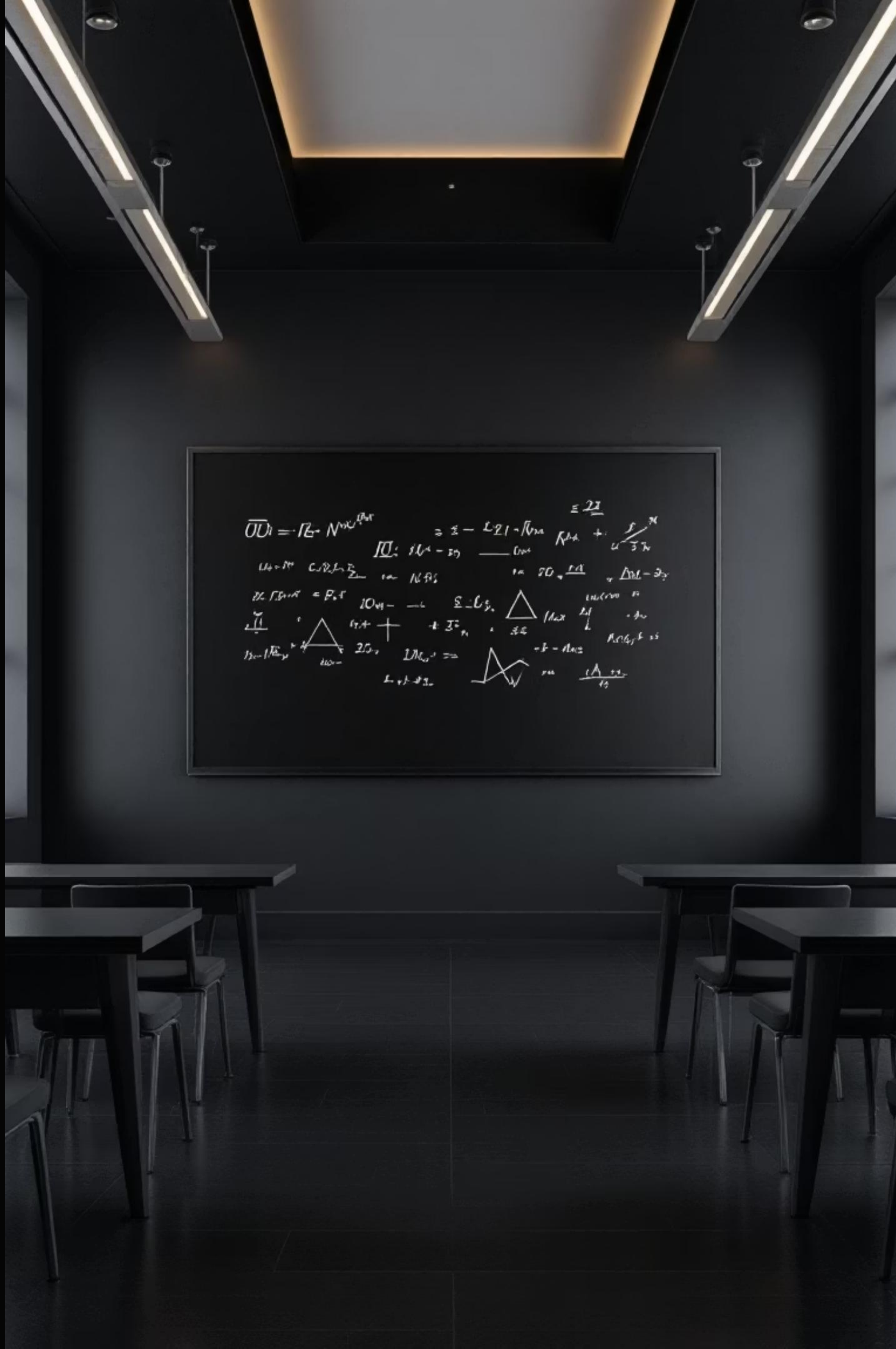
Modelos de Regresión I: Enfoque basado en Modelos Lineales

Comprender y aplicar la regresión lineal para resolver problemas de predicción numérica en el mundo real.

Gabriel Rengifo



¿Por qué necesitamos la regresión lineal?

En el mundo real enfrentamos constantemente preguntas de predicción numérica:

Mercado inmobiliario

¿Cuál será el precio de una casa basándose en su ubicación, tamaño y características?

Eficiencia energética

¿Cómo predecir el consumo de combustible de un vehículo según sus especificaciones?

Planificación operativa

¿Cuánto tiempo durará una misión naval considerando múltiples factores?

Utilizamos **regresión lineal** cuando nuestra variable objetivo es numérica y continua.

Fundamentos matemáticos

La regresión lineal establece una relación matemática entre una variable dependiente *y* y una o más variables independientes *x*.

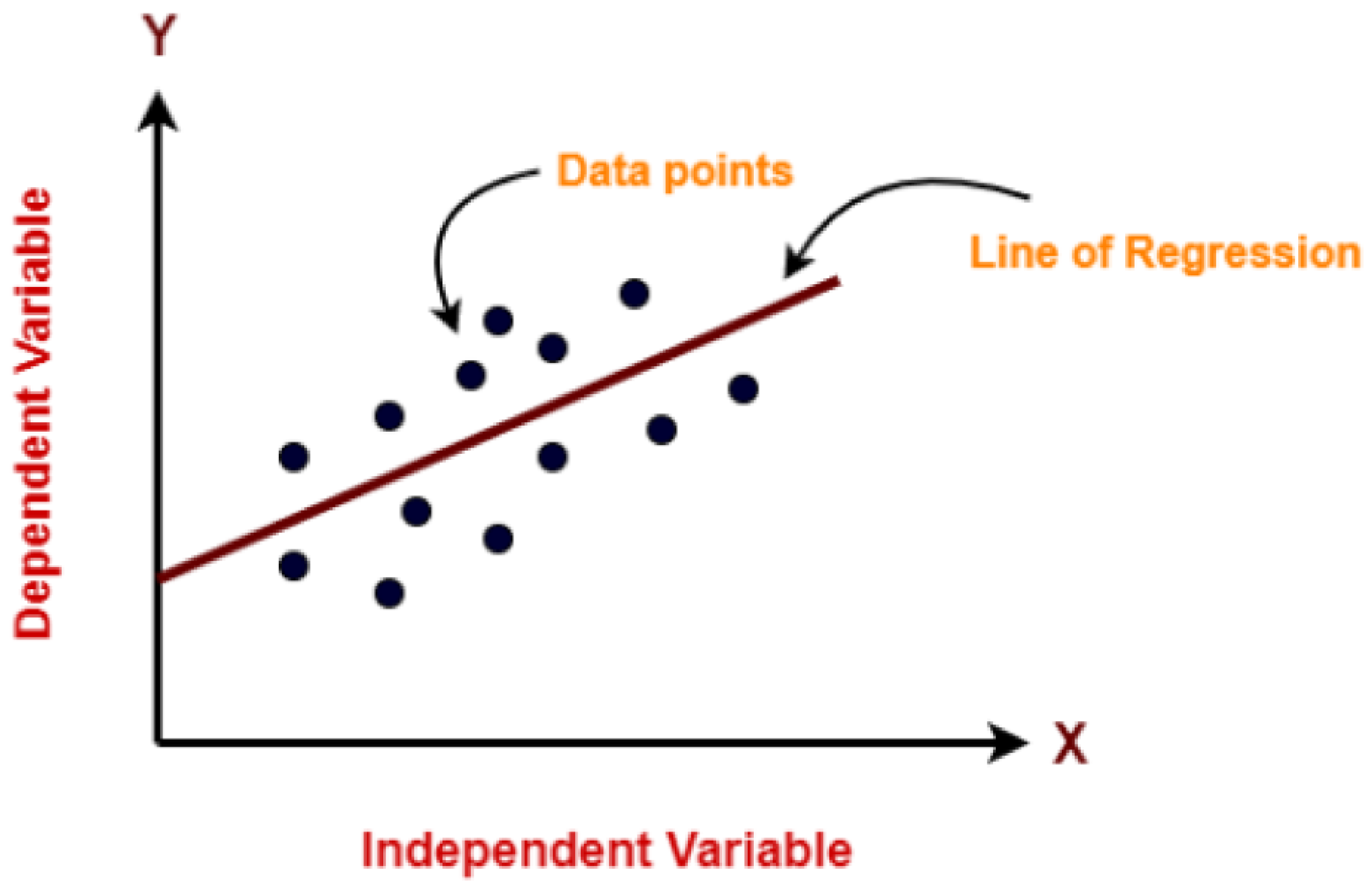
Fórmula general

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Componentes clave:

- **β (beta)**: coeficientes que determinan el peso de cada variable
- **ε (épsilon)**: término de error que captura la variabilidad no explicada
- **β_0**: intercepto o término constante

¿Qué es la Regresión?



Construye una **línea o curva que pasa a través de todos los puntos de datos** en el gráfico de predicción objetivo de tal manera que la distancia vertical entre los puntos de datos y la curva de regresión es **mínima**.

Tipos de regresión lineal

Regresión Simple

Utiliza una sola variable independiente para predecir la variable dependiente. Es el caso más básico y fácil de visualizar.

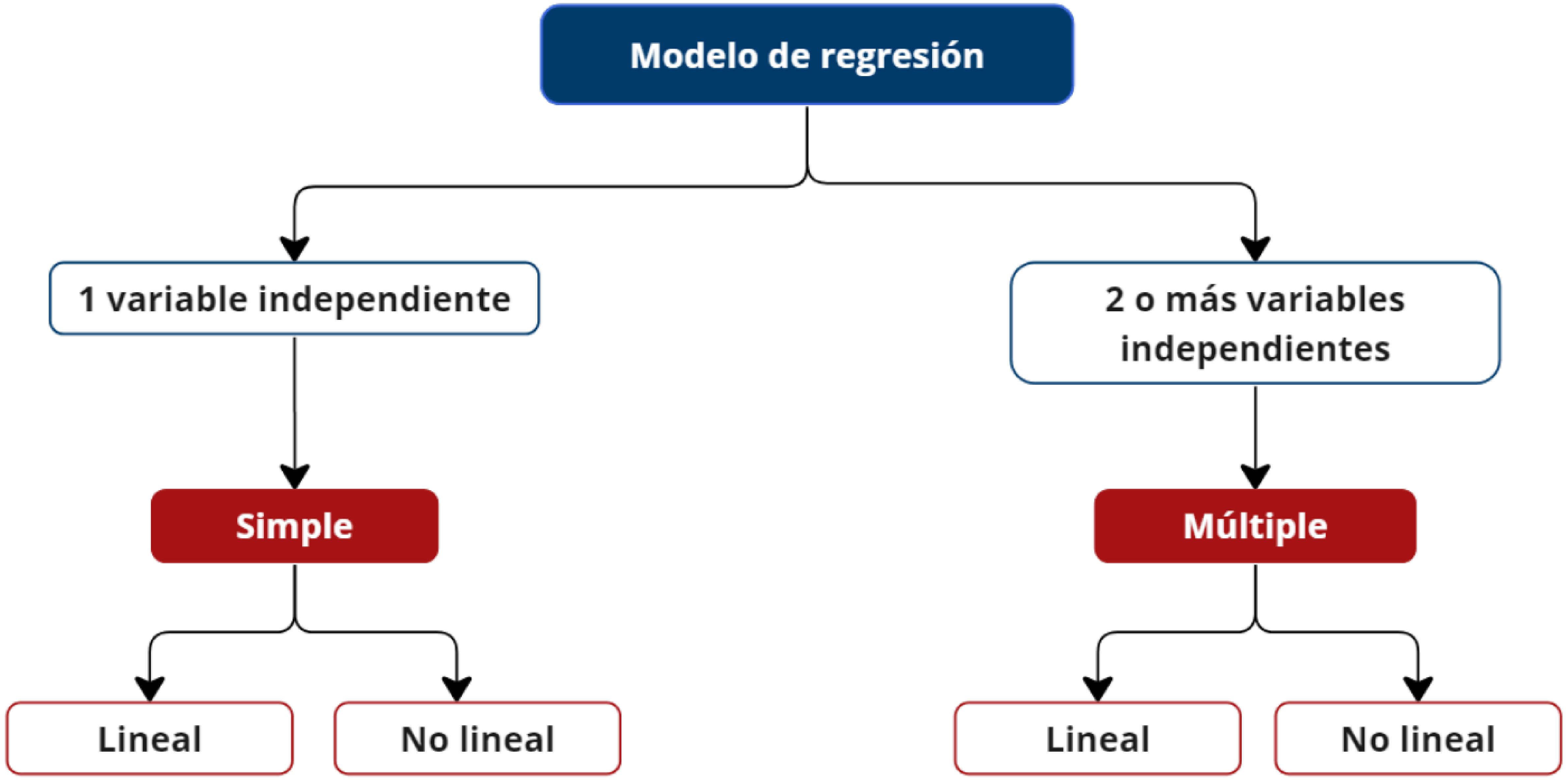
Ejemplo: Predecir el precio de una casa solo basándose en su tamaño en metros cuadrados.

Regresión Múltiple

Incorpora múltiples variables independientes, capturando relaciones más complejas y realistas del mundo real.

Ejemplo: Predecir la tarifa del Titanic considerando clase, edad, sexo, número de familiares a bordo, puerto de embarque, etc.

Tipos de modelos de Regresión



Regresión lineal simple

Intercepto

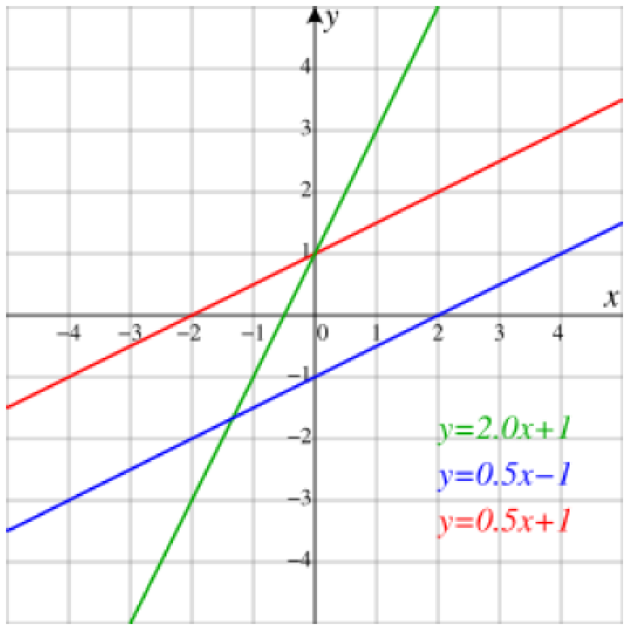
Pendiente

Error aleatorio

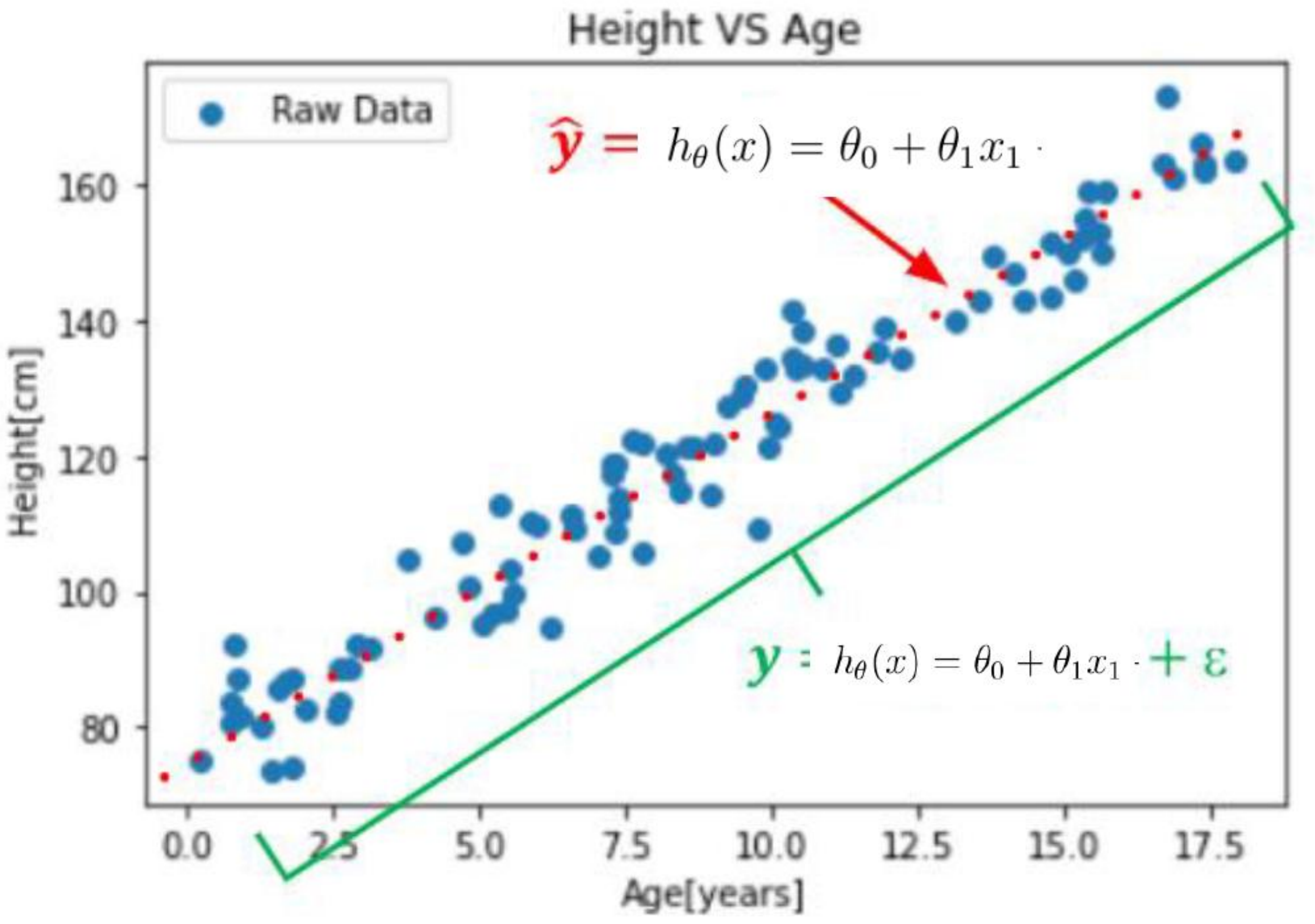
$$h(x) = \theta_0 + \theta_1 \cdot x + \epsilon$$

Variable Dependiente (respuesta - y)

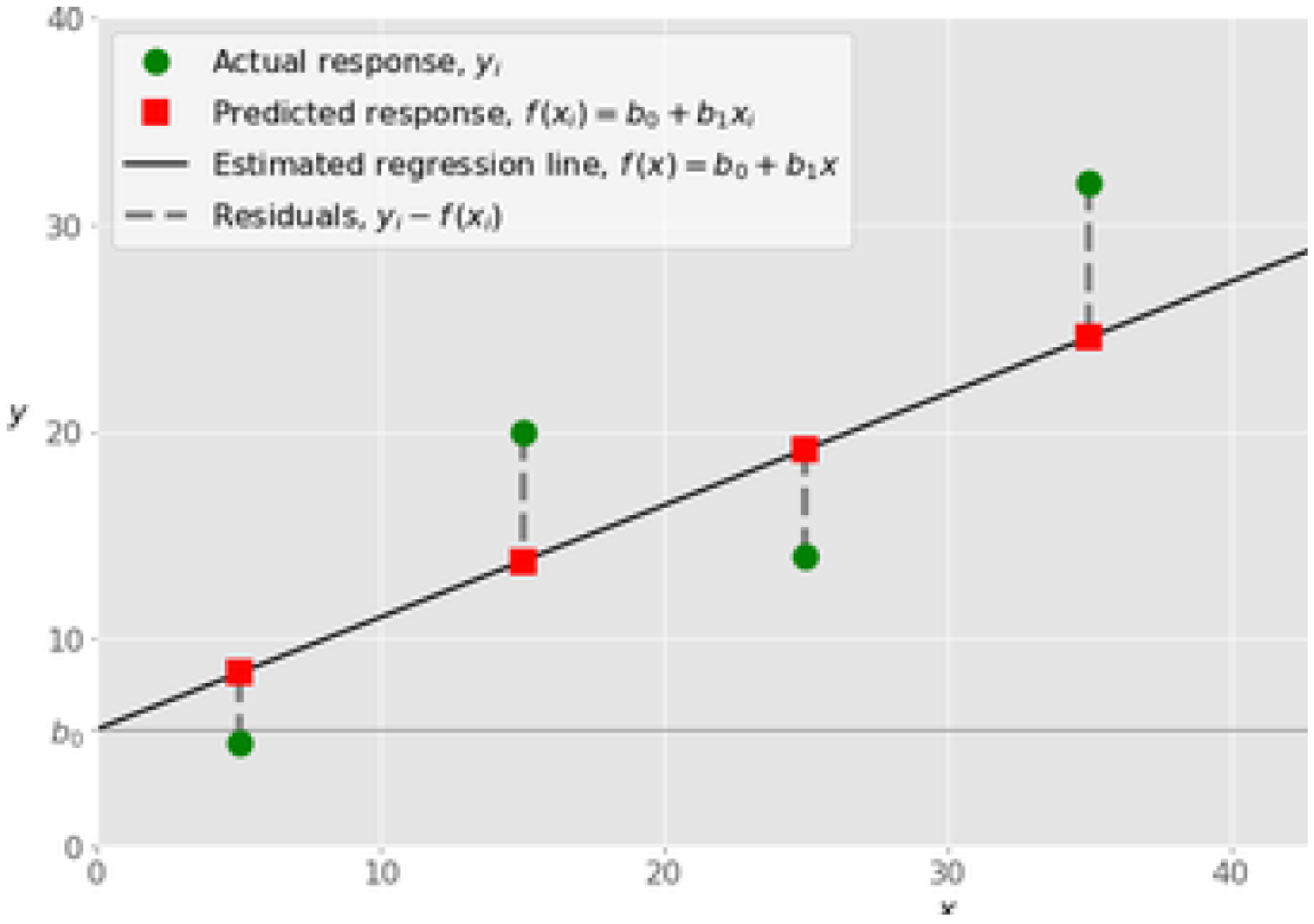
Variable Independiente (Explicatoria)



Regresión lineal simple



Regresión lineal simple - Residuo



- La distancia entre los datos y la curva construida.
- Indica si el modelo ha capturado la relación entre los predictores y la variable objetivo.

Residuo (e) =
valor observado de salida - valor predicho

$$e = y - \hat{y}$$

Los modelos de regresión buscan minimizar el valor de **e** para el conjunto de predictores de entrenamiento.

Supuestos fundamentales

Para que la regresión lineal funcione correctamente, debe cumplir cinco supuestos críticos:

01

Linealidad

La relación entre variables independientes y dependiente debe ser lineal.

02

Independencia

Los errores de las observaciones deben ser independientes entre sí.

03

Homocedasticidad

La varianza de los errores debe ser constante a lo largo de todas las observaciones.

04

Normalidad

Los errores deben seguir una distribución normal.

05

No multicolinealidad

Las variables independientes no deben estar altamente correlacionadas entre sí.

Estimación de parámetros

Método de Mínimos Cuadrados Ordinarios (OLS)

Este método encuentra los coeficientes β que minimizan la suma de los errores al cuadrado:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Intuición: Buscamos la "mejor recta" que se ajusta a nuestros datos, minimizando las distancias entre los puntos reales y la línea predicha.

1. **Datos de entrada**
   Variables X y objetivo Y
2. **Optimización OLS**
   Minimizar errores cuadráticos
3. **Coeficientes β**
   Parámetros del modelo

Métricas de evaluación

Para determinar qué tan bien funciona nuestro modelo, utilizamos métricas específicas:

Coeficiente R^2 (R-cuadrado)

Indica qué proporción de la varianza en la variable dependiente es explicada por el modelo. Valores cercanos a 1 indican mejor ajuste.

RMSE (Error Cuadrático Medio)

Mide qué tan lejos están las predicciones de los valores reales. Valores más bajos indican mejor precisión.

MAE (Error Absoluto Medio)

Calcula el promedio de los errores absolutos, proporcionando una medida interpretable del error típico.

Técnicas de regularización

Cuando tenemos demasiadas variables, el modelo puede sufrir sobreajuste. La regularización añade penalizaciones para controlarlo:

1. **Problema**
   Exceso de variables conduce al sobreajuste
2. **Solución**
   Aplicar técnicas de regularización
3. **Resultado**
   Modelo más generalizable

Ridge (L2)

Penaliza coeficientes grandes, reduciendo su magnitud pero manteniéndolos todos en el modelo.

Lasso (L1)

Fuerza algunos coeficientes a ser exactamente cero, realizando selección automática de variables.

Elastic Net

Combina las penalizaciones L1 y L2, aprovechando las ventajas de ambas técnicas.

Ejemplo práctico: Dataset Titanic

Objetivo

Predecir la tarifa (Fare) que pagaron los pasajeros del Titanic basándose en sus características.

Preparación de datos

Dividir el dataset en conjuntos de entrenamiento (80%) y prueba (20%) para validación robusta.

Entrenamiento del modelo

Ajustar regresión lineal múltiple usando variables como clase, edad, sexo, número de familiares, puerto de embarque.

Evaluación de rendimiento

Medir performance usando R^2, RMSE y MAE en el conjunto de prueba.

Comparación de técnicas

Contrastar resultados con modelos Ridge y Lasso para evaluar beneficios de regularización.

Puntos clave para recordar

Base fundamental

Los modelos lineales constituyen la piedra angular del aprendizaje supervisado y son esenciales para comprender técnicas más avanzadas.

Interpretabilidad

Son fáciles de interpretar y sirven como excelente línea base para comparar modelos más complejos.

Control de sobreajuste

Las técnicas de regularización (Ridge, Lasso, Elastic Net) son herramientas poderosas para evitar el sobreajuste.

Benchmarking

En la práctica profesional, siempre debemos comparar contra modelos más sofisticados como Random Forest o Redes Neuronales.