

Regresión Lineal

Minería de Datos (CC3074) - 2026

Regresión Lineal

Semestre 01, 2026

El problema

- Queremos entender la relación entre dos variables.
- ¿Cómo cambia Y cuando cambia X?
- ¿Podemos predecir Y usando X?

Ejemplo real

Altura del hijo vs altura del padre.

Pregunta:

¿Existe relación?

¿Podemos predecir la altura del hijo?

Intuición

Imaginemos una nube de puntos.

La regresión busca:

- La mejor recta que pase por los datos.
- Que minimice el error. (La parte de Y que el modelo no puede explicar usando X)
- Que permita hacer predicciones.

¿Qué significa “mejor recta”?

No es la que se ve bonita.

Es la que minimiza:

Suma de errores al cuadrado.

Minimiza distancia vertical entre puntos y recta.

El modelo

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde:

- $\beta_0 \rightarrow$ intercepto
- $\beta_1 \rightarrow$ pendiente
- $\varepsilon \rightarrow$ error

Interpretación de la pendiente

Si $\beta_1 = 2$:

Por cada unidad que aumenta X, Y aumenta en promedio 2 unidades.

Eso es interpretación.

Interpretación del Intercepto (β_0)

¿Qué es β_0 ?

- Es el valor esperado de Y cuando $X = 0$
- Es el punto donde la recta cruza el eje Y
- No siempre tiene interpretación práctica

Ejemplo: Costo de Producción (tiene sentido)

Modelo:

$$\text{Costo} = 500 + 20X$$

Donde:

- X = cantidad producida
- Y = costo total

Interpretación

Si $X = 0$:

$$\text{Costo} = 500$$

Significa:

Existe un costo fijo de 500.

Aquí β_0 tiene significado real.

Ejemplo: Altura del hijo

Modelo:

$$\text{Altura_hijo} = 40 + 0.6 \cdot \text{Altura_padre}$$

Interpretación

Si $\text{Altura_padre} = 0$:

$$\text{Altura_hijo} = 40 \text{ cm}$$

Pero nadie mide 0 cm.

El intercepto es solo matemático.

Regresión y correlación

Si la correlación es alta:

La pendiente será más pronunciada.

Si la correlación es cercana a 0:

La recta será casi horizontal.

Validación

Nunca evaluar en datos de entrenamiento.

Dividir:

- Train
- Test

Medir desempeño en test.

Métricas de Evaluación en Regresión

¿Por qué necesitamos métricas?

- La recta siempre se puede ajustar
- Pero necesitamos saber qué tan buena es
- Queremos medir el error del modelo
- Siempre evaluar en datos que no se usaron para entrenar

MAE

Error Absoluto Medio

¿Qué mide?

- El error promedio del modelo
- Cuánto nos equivocamos en promedio
- Está en las mismas unidades que la variable

Si $MAE = 2$

→ en promedio nos equivocamos 2 unidades.

MSE

Error Cuadrático Medio

¿Qué mide?

- Penaliza más los errores grandes
- Si el modelo comete errores muy grandes, MSE aumenta mucho
- Es sensible a outliers

Sirve cuando queremos castigar errores grandes.

RMSE

Raíz del Error Cuadrático Medio

¿Qué mide?

- Similar al MSE
- Pero vuelve a las mismas unidades de la variable
- Muy usado en práctica

Representa el error típico del modelo.

R²

Coeficiente de Determinación

¿Qué significa?

- Indica qué porcentaje de la variabilidad explica el modelo
- Va de 0 a 1
- Más cercano a 1 → mejor ajuste

Ejemplo:

$$R^2 = 0.80$$

→ el modelo explica el 80% del comportamiento observado.

Comparación rápida

- MAE → error promedio simple
- RMSE → penaliza errores grandes

- $R^2 \rightarrow$ qué tanto explica el modelo

Supuestos

Linealidad

- La relación entre X y Y debe ser aproximadamente lineal
- La recta debe capturar la tendencia real
- Si la relación es curva, el modelo será incorrecto

Validación

- Gráfico de dispersión
- No hay una forma específica
- Correlación (pero con cuidado)

Independencia

Definición

- Los errores no deben estar correlacionados entre sí
- Un error no debe depender del anterior

No usar cuando hablamos de

- Series de tiempo
- Datos secuenciales
- Procesos donde hay memoria

Homocedasticidad

Varianza constante

Definición

- La variabilidad del error debe ser constante
- No debe aumentar ni disminuir con X

¿Cómo se ve cuando falla?

- Forma de embudo en el gráfico de residuos
- Error pequeño en valores bajos y grande en valores altos

heterocedasticidad.

Normalidad de los errores

Definición

- Los errores deben distribuirse aproximadamente normal
- Importante para inferencia estadística
- No es estrictamente necesario para predicción

Afecta intervalos de confianza y pruebas de hipótesis.

¿Cómo evaluarlo?

- Gráfico Q-Q
- Histograma de residuos
- Prueba de normalidad

Desviaciones fuertes afectan inferencia.

¿Qué pasa si se violan los supuestos?

- Coeficientes pueden seguir siendo estimables
- Pero la inferencia puede ser incorrecta
- Intervalos de confianza poco confiables
- Pruebas estadísticas inválidas

¿Qué puede salir mal?

- Relación no lineal
- Outliers
- Heterocedasticidad (varianza no constante)
- Sobreajuste

"Transformaciones"

- Transformaciones (log, polinomios)
- Escalamiento

Puede mejorar el modelo.