Regresión Lineal Simple

📈 ¿Qué es la Regresión Lineal Simple?

La **regresión lineal simple** es un método estadístico que permite establecer una relación lineal entre dos variables: una variable independiente (X) y una variable dependiente (Y). Nos ayuda a predecir el valor de Y a partir de X.

🎯 Conceptos Fundamentales

Variable Independiente (X)

Es la variable explicativa o predictora. Se representa en el eje horizontal.

Ejemplo: Horas de estudio, temperatura, edad

Variable Dependiente (Y)

Es la variable respuesta que queremos predecir. Se representa en el eje vertical.

Ejemplo: Calificación, ventas, peso

📐 Ecuación de la Recta de Regresión

Forma general:

$$\hat{Y} = a + bX$$

Donde:

$\hat{Y}$ = Valor predicho de Y

a = Intercepto (ordenada al origen)

b = Pendiente de la recta

X = Valor de la variable independiente

🔢 Cálculo de los Coeficientes

Fórmula de la Pendiente (b):

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

Fórmula del Intercepto (a):

$$a = \bar{Y} - b\bar{X}$$

Donde:

- n = Número de observaciones
- $\bar{X}$ = Media de X
- $\bar{Y}$ = Media de Y
- $\sum XY$ = Suma de los productos XY
- $\sum X$ = Suma de todos los valores de X
- $\sum Y$ = Suma de todos los valores de Y
- $\sum X^2$ = Suma de los cuadrados de X

📊 Coeficiente de Correlación (r)

Mide la fuerza y dirección de la relación lineal entre X e Y:

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Interpretación de r:

$$-1 \leq r \leq 1$$

$r = 1$: Correlación positiva perfecta

$r = -1$: Correlación negativa perfecta

$r = 0$: No hay correlación lineal

$$|r| > 0.7 \text{ : Correlación fuerte}$$

$$0.3 < |r| < 0.7 \text{ : Correlación moderada}$$

$$|r| < 0.3 \text{ : Correlación débil}$$

📈 Coeficiente de Determinación (R^2)

Indica el porcentaje de variabilidad de Y explicado por X:

$$R^2 = r^2$$

Interpretación: Un $R^2 = 0.85$ significa que el 85% de la variación en Y es explicada por X.

📚 Ejemplo 1: Horas de Estudio vs Calificación ▲

🎯 Enunciado:

Un profesor quiere determinar la relación entre las horas de estudio (X) y la calificación final (Y) de 5 estudiantes.

📝 **Datos:**

Estudiante	Horas de Estudio (X)	Calificación (Y)
1	2	12
2	3	14
3	4	16
4	5	17
5	6	19

📊 Paso 1: Calcular las sumas necesarias

X	Y	X^2	Y^2	XY
2	12	4	144	24
3	14	9	196	42
4	16	16	256	64
5	17	25	289	85
6	19	36	361	114
Σ = 20	**Σ = 78**	**Σ = 90**	**Σ = 1246**	**Σ = 329**

📐 **Paso 2: Calcular la pendiente (b)**

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$b = \frac{5(329) - (20)(78)}{5(90) - (20)^2}$$

$$b = \frac{1645 - 1560}{450 - 400}$$

$$b = \frac{85}{50} = 1.7$$

📐 **Paso 3: Calcular el intercepto (a)**

$$\bar{X} = \frac{20}{5} = 4 \qquad \bar{Y} = \frac{78}{5} = 15.6$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 15.6 - 1.7(4)$$

$$a = 15.6 - 6.8 = 8.8$$

✅ **Ecuación de Regresión:**

$$\hat{Y} = 8.8 + 1.7X$$

Interpretación:

Por cada hora adicional de estudio, la calificación aumenta en 1.7 puntos

Un estudiante que no estudia nada tendría una calificación base de 8.8

📊 Paso 4: Calcular el coeficiente de correlación (r)

$$r = \frac{5(329) - (20)(78)}{\sqrt{[5(90) - 400][5(1246) - 6084]}}$$

$$r = \frac{85}{\sqrt{50 \times 146}}$$

$$r = \frac{85}{\sqrt{7300}} = \frac{85}{85.44} \approx 0.995$$

✅ **Coeficiente de Determinación:** $R^2 = (0.995)^2 = 0.99$

Interpretación: El 99% de la variación en las calificaciones se explica por las horas de estudio. Existe una correlación positiva casi perfecta.

🎯 Paso 5: Hacer predicciones

Pregunta: ¿Qué calificación se espera para un estudiante que estudia 7 horas?

$$\hat{Y} = 8.8 + 1.7(7) = 8.8 + 11.9 = 20.7$$

Respuesta: Se espera una calificación de aproximadamente 20.7 puntos.

💡 Supuestos de la Regresión Lineal

Linealidad: La relación entre X e Y debe ser lineal

Independencia: Las observaciones deben ser independientes entre sí

Homocedasticidad: La varianza de los errores debe ser constante

Normalidad: Los errores deben seguir una distribución normal

No multicolinealidad: En regresión múltiple, las variables independientes no deben estar altamente correlacionadas

⚠️ Errores Comunes

Extrapolar más allá del rango de datos observados

Confundir correlación con causalidad

No verificar los supuestos del modelo

Usar regresión lineal cuando la relación no es lineal

Ignorar valores atípicos (outliers) que pueden afectar significativamente el modelo