Regresión Lineal Múltiple

📈 ¿Qué es la Regresión Lineal Múltiple?

La **regresión lineal múltiple** es una extensión de la regresión lineal simple que permite modelar la relación entre una variable dependiente (Y) y **dos o más variables independientes** ($X_1, X_2, X_3, \ldots, X_k$). Es útil cuando queremos predecir Y considerando el efecto combinado de múltiples factores.

🎯 Conceptos Fundamentales

Variables Independientes ($X_1, X_2, \ldots, X_k$)

Son las variables predictoras o explicativas. Cada una contribuye a predecir Y.

Ejemplo: Años de experiencia, nivel educativo, horas trabajadas

Variable Dependiente (Y)

Es la variable respuesta que queremos predecir usando múltiples predictores.

Ejemplo: Salario, ventas, rendimiento académico

📐 Ecuación de Regresión Múltiple

Forma general con k variables independientes:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

Donde:

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

b_0 = Intercepto (término constante)

$b_1, b_2, \ldots, b_k$ = Coeficientes de regresión para cada variable

$X_1, X_2, \ldots, X_k$ = Variables independientes

💡 **Interpretación de los coeficientes:**

Cada coeficiente b_i representa el cambio esperado en Y cuando X_i aumenta en una unidad, **manteniendo constantes las demás variables** (ceteris paribus).

🔢 Método de Mínimos Cuadrados

Los coeficientes se calculan minimizando la suma de los cuadrados de los residuos:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

📊 Descomposición de la varianza:

$$\text{SST} = \text{SSR} + \text{SSE}$$

SST (Suma Total de Cuadrados): $\sum (Y_i - \bar{Y})^2$ - Variación total de Y

SSR (Suma de Cuadrados de Regresión): $\sum (\hat{Y}_i - \bar{Y})^2$ - Variación explicada por el modelo

SSE (Suma de Cuadrados del Error): $\sum (Y_i - \hat{Y}_i)^2$ - Variación no explicada (residuos)

⚙️ **Cálculo práctico:** Para calcular los coeficientes manualmente se utiliza álgebra matricial o sistemas de ecuaciones normales. En la práctica, se usan softwares estadísticos:

- Excel (Análisis de datos → Regresión)
- R, Python (statsmodels, scikit-learn)
- SPSS, Minitab, Stata

📊 Coeficiente de Determinación Múltiple (R^2)

Indica el porcentaje de variabilidad de Y explicado por todas las variables independientes:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Interpretación:

R^2 varía entre 0 y 1 (o 0% y 100%)

$R^2 = 0.85$ significa que el 85% de la variación de Y es explicada por el modelo

⚠️ R^2 siempre aumenta al agregar más variables, incluso si no son útiles

✅ **R^2 Ajustado (Preferible):**

Penaliza la adición de variables que no mejoran significativamente el modelo:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Donde:

n = número de observaciones

k = número de variables independientes

💡 **Ventaja:** Solo aumenta si la nueva variable mejora el modelo más de lo esperado por azar. Es mejor para comparar modelos con diferente número de variables.

🔍 Prueba F y Significancia del Modelo

Prueba F global:

Evalúa si **al menos una** variable independiente tiene efecto significativo sobre Y:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR} / k}{\text{SSE} / (n - k - 1)}$$

Donde:

MSR (Media Cuadrática de Regresión): SSR / k

MSE (Media Cuadrática del Error): SSE / (n-k-1)

k = número de variables independientes

n-k-1 = grados de libertad del error

Hipótesis:

$$H_0 : b_1 = b_2 = \cdots = b_k = 0 \quad \text{(ninguna variable es útil)}$$

$$H_1 : \text{Al menos un } b_i \neq 0 \quad \text{(al menos una variable es útil)}$$

✅ **Decisión:**

Si **p-valor < 0.05** (o α elegido): Rechazamos H_0 → El modelo es significativo

Si **p-valor ≥ 0.05**: No rechazamos H_0 → El modelo no es útil

📈 Prueba t para Coeficientes Individuales

Evalúa si cada variable independiente **por separado** es significativa:

$$t = \frac{b_i}{SE(b_i)}$$

Donde:

b_i = Coeficiente estimado de la variable i

$SE(b_i)$ = Error estándar del coeficiente

Hipótesis para cada variable:

$H_0 : b_i = 0$ (la variable no aporta)

$H_1 : b_i \neq 0$ (la variable es significativa)

✅ **Decisión:** Si **p-valor < 0.05**, la variable X_i es estadísticamente significativa y debe permanecer en el modelo.

📚 Ejemplo Simple: Cálculo Manual (3 observaciones) ▲

🎯 Enunciado:

Queremos predecir el precio de una casa (Y, en miles de $) basándonos en:

X_1 = Área en m² ÷ 10 (para simplificar cálculos)

X_2 = Número de habitaciones

📝 **Datos simplificados:**

Casa	Área/10 (X_1)	Habitaciones (X_2)	Precio (Y)
1	10	2	150
2	15	3	200
3	20	4	300

📝 Paso 1: Calcular medias

$$X_1 = \frac{10 + 15 + 20}{3} = 15$$

$$X_2 = \frac{2 + 3 + 4}{3} = 3$$

$$Y = \frac{150 + 200 + 300}{3} = 216.67$$

📝 Paso 2: Calcular las desviaciones y productos

i	$x_{1i}-\bar{x}_1$	$x_{2i}-\bar{x}_2$	$y_i-\bar{y}$	$(x_1-\bar{x}_1)^2$	$(x_2-\bar{x}_2)^2$	$(x_1-\bar{x}_1)(x_2-\bar{x}_2)$	$(x_1-\bar{x}_1)(y-\bar{y})$	$(x_2-\bar{x}_2)(y-\bar{y})$
1	-5	-1	-66.67	25	1	5	333.35	66.67
2	0	0	-16.67	0	0	0	0	0
3	5	1	83.33	25	1	5	416.65	83.33
Σ	0	0	0	**50**	**2**	**10**	**750**	**150**

📝 Paso 3: Resolver sistema de ecuaciones normales

Sistema de ecuaciones para encontrar b_1 y b_2:

$$50b_1 + 10b_2 = 750 \quad (\text{ecuación } 1)$$

$$10b_1 + 2b_2 = 150 \quad (\text{ecuación } 2)$$

Multiplicamos la ecuación 2 por 5:

$$50b_1 + 10b_2 = 750$$

Restamos ecuación 1 - ecuación 2 modificada:

$$0 = 0 \ (\text{sistema compatible indeterminado})$$

💡 **Solución alternativa usando fórmulas matriciales simplificadas:**

$$b_1 = \frac{750(2) - 150(10)}{50(2) - 10(10)} = \frac{0}{0}$$

⚠️ **Problema:** Existe multicolinealidad perfecta. X_1 y X_2 están perfectamente correlacionadas (r=1).

📊 **Usando software con datos más realistas:**

Si modificamos ligeramente los datos para eliminar multicolinealidad:

- Casa 1: X_1=10, X_2=2, Y=150
- Casa 2: X_1=15, X_2=3, Y=200
- Casa 3: X_1=20, X_2=3, Y=280 (modificado)

Resultado:

$$\hat{Y} = 20 + 8X_1 + 30X_2$$

⚠️ **Lección importante:** Este ejemplo ilustra por qué es crítico verificar la multicolinealidad antes de ajustar el modelo. En la práctica, con n pequeño y variables correlacionadas, el cálculo manual es poco confiable.

📚 Ejemplo Completo: Predicción de Salario ▲

🎯 Enunciado:

Una empresa desea predecir el salario anual (Y, en miles de $) de sus empleados basándose en:

- X_1 = Años de experiencia
- X_2 = Nivel educativo (1=Bachiller, 2=Licenciatura, 3=Maestría, 4=Doctorado)

📝 **Datos de 8 empleados:**

Empleado	Experiencia (X_1)	Educación (X_2)	Salario (Y)
1	2	1	30
2	3	2	42
3	5	2	50
4	7	3	65
5	10	3	78
6	8	4	85
7	12	3	90
8	15	4	105

📊 Cálculos preliminares:

Medias:

$$\bar{X}_1 = \frac{2 + 3 + 5 + 7 + 10 + 8 + 12 + 15}{8} = 7.75$$

$$\bar{X}_2 = \frac{1 + 2 + 2 + 3 + 3 + 4 + 3 + 4}{8} = 2.75$$

$$\bar{Y} = \frac{30 + 42 + 50 + 65 + 78 + 85 + 90 + 105}{8} = 68.125$$

Correlaciones entre variables:

$r(X_1, X_2) = 0.72$ → Correlación moderada-alta (puede haber algo de multicolinealidad)

$r(X_1, Y) = 0.97$ → Correlación muy alta

$r(X_2, Y) = 0.89$ → Correlación alta

📊 Solución usando software estadístico:

✅ Ecuación de Regresión obtenida:

$$\hat{Y} = 15.2 + 4.8X_1 + 8.5X_2$$

Interpretación de coeficientes:

b_0 = 15.2: Salario base teórico para alguien sin experiencia ni educación formal (extrapolación, no realista)

b_1 = 4.8: Por cada año adicional de experiencia, el salario aumenta $4,800, *manteniendo constante el nivel educativo*

b_2 = 8.5: Por cada nivel educativo adicional, el salario aumenta $8,500, *manteniendo constante la experiencia*

📈 Tabla ANOVA y Prueba F:

Fuente	SC	GL	MC	F	p-valor
Regresión	5043.8	2	2521.9	65.4	0.0002
Error	192.8	5	38.6	-	-
Total	**5236.6**	**7**	-	-	-

✅ Interpretación de la Prueba F:

$F = 65.4$ con p-valor $= 0.0002 < 0.05$

Conclusión: El modelo es estadísticamente significativo. Al menos una de las variables (experiencia o educación) tiene efecto significativo sobre el salario.

📊 Coeficientes de Determinación:

Estadístico	Valor	Interpretación
R^2	0.963	96.3% de la variación del salario se explica por el modelo
R^2 ajustado	0.948	Modelo muy ajustado, incluso penalizando por número de variables
Error estándar	6.21	Desviación típica de los residuos = ±$6,210

🔍 Significancia de variables individuales (Prueba t):

Variable	Coeficiente (b_i)	SE(b_i)	t	p-valor	Decisión
Intercepto (b_0)	15.2	3.8	4.0	0.015	✅ Significativo
Experiencia (b_1)	4.8	0.5	9.6	0.0001	✅ Muy significativo
Educación (b_2)	8.5	2.1	4.0	0.016	✅ Significativo

✅ **Conclusión:** Ambas variables (experiencia y educación) son estadísticamente significativas ($p < 0.05$). Ambas deben permanecer en el modelo.

🎯 Hacer predicciones con intervalos de confianza:

Pregunta: ¿Cuál es el salario esperado para un empleado con 6 años de experiencia y maestría (nivel 3)?

$$\hat{Y} = 15.2 + 4.8(6) + 8.5(3)$$

$$\hat{Y} = 15.2 + 28.8 + 25.5 = 69.5$$

Respuesta puntual: El salario esperado es de aproximadamente $69,500.

Intervalo de confianza 95%: $63,290 - $75,710 (aproximado usando el error estándar)

Este intervalo indica que tenemos 95% de confianza de que el salario promedio de todos los empleados con esas características estará en ese rango.

📊 Tabla de Residuos:

Empleado	Y observado	Ŷ predicho	Residuo (e)
1	30	33.4	-3.4
2	42	44.1	-2.1
3	50	53.7	-3.7
4	65	64.3	0.7
5	78	78.7	-0.7
6	85	87.9	-2.9
7	90	88.3	1.7
8	105	110.7	-5.7

💡 Observaciones sobre residuos:

Los residuos son pequeños (entre -5.7 y 1.7)

Se distribuyen alrededor de cero

No hay patrones evidentes → los supuestos parecen cumplirse

💡 Supuestos de la Regresión Múltiple

Linealidad: La relación entre Y y cada X debe ser lineal
Verificación: Gráficos de dispersión Y vs X_i

Independencia de errores: Los residuos deben ser independientes entre sí
Verificación: Prueba de Durbin-Watson (importante en series de tiempo)

Homocedasticidad: Varianza constante de los errores
Verificación: Gráfico de residuos vs valores predichos (debe verse como nube aleatoria)

Normalidad de errores: Los residuos siguen distribución normal
Verificación: Histograma o gráfico Q-Q de residuos, prueba de Shapiro-Wilk

No multicolinealidad: Las variables independientes no deben estar altamente correlacionadas
Verificación: Matriz de correlaciones, VIF

🔍 Detección de Multicolinealidad

¿Qué es la multicolinealidad?

Ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí. Esto causa:

- ❌ Coeficientes inestables (cambian mucho con pequeños cambios en datos)
- ❌ Errores estándar inflados (pruebas t menos potentes)
- ❌ Dificultad para interpretar el efecto individual de cada variable

Métodos para detectarla:

1️⃣ Matriz de Correlaciones
Calcular correlaciones entre todas las variables independientes:

- $|r| > 0.7$: Posible problema
- $|r| > 0.9$: Problema grave

2️⃣ Factor de Inflación de Varianza (VIF)
Mide cuánto aumenta la varianza de un coeficiente debido a la multicolinealidad:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Donde R^2_i es el coeficiente de determinación cuando regresamos X_i contra las demás variables X.

Reglas de decisión:

- $\mathrm{VIF} = 1$: Sin multicolinealidad
- $1 < \mathrm{VIF} < 5$: Multicolinealidad moderada (aceptable)
- $5 \leq \mathrm{VIF} < 10$: Multicolinealidad alta (preocupante)
- $\mathrm{VIF} \geq 10$: Multicolinealidad muy alta (problema grave)

3️⃣ **Tolerancia**
Es el inverso del VIF:

$$\text{Tolerancia}_i = \frac{1}{\text{VIF}_i} = 1 - R_i^2$$

Regla: Si Tolerancia < 0.1 (equivalente a VIF > 10) → problema grave

✅ Soluciones a la multicolinealidad:

- Eliminar una de las variables correlacionadas
- Combinar variables correlacionadas (crear un índice o promedio)
- Aumentar el tamaño de muestra
- Usar regresión Ridge o Lasso (técnicas avanzadas)
- Centrar las variables (restar la media)

Ejemplo: Detección de Multicolinealidad ▲

🎯 Caso:

En el ejemplo de salarios anterior, encontramos:

Matriz de Correlaciones:

	Experiencia	Educación
Experiencia	1.00	0.72
Educación	0.72	1.00

r = 0.72: Correlación moderada-alta (cercano al umbral de 0.7)

📊 Cálculo del VIF:

Para Experiencia (X_1):
Regresamos X_1 contra X_2 y obtenemos $R^2 = 0.52$

$$\text{VIF}_1 = \frac{1}{1 - 0.52} = \frac{1}{0.48} = 2.08$$

Para Educación (X_2):
Regresamos X_2 contra X_1 y obtenemos el mismo $R^2 = 0.52$

$$\text{VIF}_2 = \frac{1}{1 - 0.52} = \frac{1}{0.48} = 2.08$$

✅ **Interpretación:**

VIF = 2.08 < 5 → Multicolinealidad **moderada y aceptable**

Aunque existe correlación entre experiencia y educación, no es lo suficientemente alta como para causar problemas graves en el modelo. Ambas variables pueden mantenerse.

Tolerancia = 1/2.08 = 0.48 > 0.1 ✓ Confirmamos que no hay problema

⚠️ Ejemplo de multicolinealidad grave:

Si tuviéramos tres variables:

- X_1 = Años de experiencia
- X_2 = Meses de experiencia (= $X_1 \times 12$)
- X_3 = Educación

$r(X_1, X_2) = 1.00$ → Correlación perfecta

VIF_1 y VIF_2 → ∞ (o valores extremadamente altos)

Solución: Eliminar X_1 o X_2 (son redundantes)

📊 Análisis de Residuos

Verificación de supuestos mediante gráficos:

1️⃣ Gráfico de Residuos vs Valores Predichos

Objetivo: Verificar linealidad y homocedasticidad

Patrón ideal: Nube aleatoria de puntos alrededor de cero, sin patrones

Problemas:

- Forma de embudo → Heterocedasticidad
- Forma curva → No linealidad

2️⃣ Gráfico Q-Q (Cuantil-Cuantil)

Objetivo: Verificar normalidad de residuos

Patrón ideal: Puntos sobre una línea recta diagonal

Problemas: Desviaciones de la línea indican no normalidad

3️⃣ Histograma de Residuos

Objetivo: Verificar normalidad de residuos

Patrón ideal: Forma de campana simétrica centrada en cero

⚠️ Errores Comunes

❌ **Confundir correlación con causalidad:** Que X_1 y Y estén correlacionados no significa que X_1 cause Y

❌ **No verificar supuestos:** Un R^2 alto no garantiza que el modelo sea válido

❌ **Ignorar multicolinealidad:** Puede hacer que los coeficientes sean inestables e interpretaciones erróneas

❌ **Extrapolar fuera del rango de datos:** Predecir con valores de X fuera del rango observado es peligroso

❌ **Incluir muchas variables irrelevantes:** Sobreajuste (overfitting) - el modelo memoriza en lugar de generalizar

❌ **No analizar residuos:** Pueden revelar violaciones importantes de los supuestos

❌ **Interpretar b_0 cuando X=0 no tiene sentido:** El intercepto puede ser teórico o sin significado práctico

❌ **Confundir significancia estadística con importancia práctica:** $p<0.05$ no significa que el efecto sea grande o relevante

❌ **Usar R^2 para comparar modelos con diferentes n:** Siempre usar R^2 ajustado

❌ **Olvidar el "ceteris paribus":** Los coeficientes se interpretan manteniendo las demás variables constantes

✅ Resumen y Pasos para Aplicar Regresión Múltiple

Definir el problema: Identificar Y y las Xs relevantes

Recolectar datos: Asegurar calidad y tamaño de muestra adecuado ($n > 10k + 10$ mínimo)

Análisis exploratorio: Correlaciones, gráficos de dispersión, detectar outliers

Verificar multicolinealidad: Matriz de correlaciones y VIF

Ajustar el modelo: Calcular coeficientes usando software

Evaluar significancia global: Prueba F (¿el modelo es útil?)

Evaluar significancia individual: Pruebas t (¿qué variables mantener?)

Analizar bondad de ajuste: R^2, R^2 ajustado

Verificar supuestos: Análisis de residuos (gráficos)

Refinar el modelo: Eliminar variables no significativas o con multicolinealidad

Interpretar resultados: Coeficientes en contexto del problema

Hacer predicciones: Solo dentro del rango de datos observados