Regresión Lineal Simple para Modelado Predictivo - Día 4

Change Status

pending 40 min

Learning Objectives

1. Estimar parámetros de regresión lineal e interpretar coeficientes correctamente
2. Evaluar bondad de ajuste con métricas apropiadas (R^2, RMSE)
3. Diagnosticar problemas comunes en modelos lineales y validar supuestos

Activities and Learning

Task 1: Fundamentos Matemáticos de la Regresión Lineal (10 minutos)

La regresión lineal modela la relación entre una variable dependiente (Y) y variables independientes (X) asumiendo una relación lineal.

El Modelo Matemático

Ecuación básica: $Y = \beta_0 + \beta_1 X + \varepsilon$

Donde:

Y: Variable dependiente (lo que queremos predecir)
X: Variable independiente (predictor)
β_0: Intercepto (valor de Y cuando X = 0)
β_1: Pendiente (cambio en Y por unidad de X)
ε: Error residual (variabilidad no explicada)
Método de Mínimos Cuadrados Ordinarios (OLS)

OLS encuentra los coeficientes que minimizan la suma de errores cuadráticos:

Fórmula de β_1: $\beta_1 = \Sigma((x_i - \bar{x})(y_i - \bar{y})) / \Sigma((x_i - \bar{x})^2)$ **Fórmula de β_0**: $\beta_0 = \bar{y} - \beta_1\bar{x}$

Ventajas de OLS:

Propiedad BLUE: Mejor estimador lineal insesgado
Mínima varianza: Entre todos los estimadores lineales insesgados
Fácil de interpretar: Coeficientes tienen significado intuitivo

Task 2: Estimación e Interpretación de Coeficientes (10 minutos)

Los coeficientes de regresión deben interpretarse en contexto, considerando unidades y significancia estadística.

Interpretación del Intercepto (β_0)

Significado: Valor esperado de Y cuando X = 0. **Consideraciones**:

Significativo solo si X=0 es plausible: No interpretar intercepto si X=0 está fuera del rango de datos
Escalado: Afectado por unidades de medición
Contexto: Puede no tener significado práctico
Interpretación de la Pendiente (β_1)

Significado: Cambio esperado en Y por cada unidad adicional de X. **Ejemplos**:

Ventas = 1000 + 50 × (gasto_publicidad) → Cada $1 extra en publicidad aumenta ventas en $50
Satisfacción = 3.2 + 0.8 × (calidad_servicio) → Cada punto de calidad aumenta satisfacción en 0.8 puntos

Consideraciones importantes:

Unidades: Interpretación depende de unidades de X e Y
Rango de validez: Coeficiente válido solo dentro del rango observado de X
Ceteris paribus: Asume otras variables constantes
Significancia Estadística

Prueba t para coeficientes: H_0: $\beta_1 = 0$ (variable no contribuye al modelo)

```
# En statsmodels
import statsmodels.formula.api as smf

modelo = smf.ols('ventas ~ gasto_publicidad', data=df).fit()
print(modelo.summary())

# Interpretar:
# coef: valor del coeficiente
```

```
# std err: error estándar
# t: estadístico t
# P>|t|: valor p
# [0.025, 0.975]: intervalo de confianza 95%
```

Task 3: Evaluación de Bondad de Ajuste (10 minutos)

Métricas cuantitativas evalúan qué tan bien el modelo explica los datos y su capacidad predictiva.

Coeficiente de Determinación (R^2)

Fórmula: $R^2 = 1 - (SS_{res} / SS_{tot})$

SS_res: Suma de cuadrados residuales (errores del modelo)
SS_tot: Suma de cuadrados totales (variabilidad total)

Interpretación:

$R^2 = 1.0$: Modelo perfecto (explica 100% de variabilidad)
$R^2 = 0.0$: Modelo no explica variabilidad (tan bueno como la media)
R^2 negativo: Modelo peor que usar la media

Limitaciones:

Aumenta con más variables: R^2 nunca disminuye al añadir predictores
No mide calidad predictiva: Alto R^2 no garantiza buenas predicciones
Sensibilidad a outliers: Puntos extremos pueden inflar R^2 artificialmente
Error Cuadrático Medio (RMSE) y Error Absoluto Medio (MAE)

RMSE: $\sqrt{(\Sigma(y_i - \hat{y}_i)^2 / n)}$ - penaliza errores grandes **MAE**: $\Sigma|y_i - \hat{y}_i| / n$ - más robusto a outliers

Interpretación práctica:

RMSE en unidades de Y: Error típico del modelo en las mismas unidades que Y
Comparación con variabilidad: RMSE vs desviación estándar de Y
R^2 Ajustado

Fórmula: $R^2\_ajustado = 1 - [(1 - R^2)(n - 1)/(n - p - 1)]$

Donde:

n: número de observaciones
p: número de predictores

Ventajas:

Penaliza complejidad: Disminuye al añadir predictores que no mejoran el modelo
Mejor para comparación: Entre modelos con diferente número de variables

Task 4: Validación de Supuestos y Diagnóstico (10 minutos)

Los modelos de regresión lineal tienen supuestos que deben verificarse para garantizar validez de inferencias.

Supuestos Clave

Linealidad: Relación entre X e Y es lineal. **Verificación**: Gráfico de residuos vs valores ajustados.

Independencia: Errores no están correlacionados. **Verificación**: Gráfico de residuos vs orden, prueba de Durbin-Watson.

Homocedasticidad: Varianza de errores es constante. **Verificación**: Gráfico de residuos vs valores ajustados, prueba de Breusch-Pagan.

Normalidad de residuos: Errores siguen distribución normal. **Verificación**: Q-Q plot, prueba de Shapiro-Wilk.

Diagnóstico Visual

```
import matplotlib.pyplot as plt
import scipy.stats as stats

# Crear gráficos de diagnóstico
fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(12, 10))

# 1. Residuos vs Valores ajustados (linealidad + homocedasticidad)
ax1.scatter(modelo.fittedvalues, modelo.resid)
ax1.axhline(y=0, color='red', linestyle='--')
ax1.set_xlabel('Valores Ajustados')
ax1.set_ylabel('Residuos')
ax1.set_title('Residuos vs Valores Ajustados')

# 2. Q-Q plot (normalidad)
stats.probplot(modelo.resid, dist="norm", plot=ax2)
ax2.set_title('Q-Q Plot de Residuos')

# 3. Histograma de residuos (normalidad)
ax3.hist(modelo.resid, bins=20, alpha=0.7)
ax3.set_xlabel('Residuos')
ax3.set_ylabel('Frecuencia')
ax3.set_title('Distribución de Residuos')
```

```
# 4. Residuos vs Orden (independencia)
ax4.scatter(range(len(modelo.resid)), modelo.resid)
ax4.axhline(y=0, color='red', linestyle='--')
ax4.set_xlabel('Orden de Observación')
ax4.set_ylabel('Residuos')
ax4.set_title('Residuos vs Orden')

plt.tight_layout()
plt.show()
```