UNIVERSITAS Miguel Hernández

Tema 1. Regresión Lineal. Ejemplos.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional y Ciencia de Datos para la Toma de Decisiones.

1. Carga y Preparación de los Datos

- Descargar el archivo `regresion.xlsx` que contiene los datos para la regresión.
- Leer el archivo Excel en un `DataFrame` de `pandas`.
- Visualizar las primeras filas del conjunto de datos.
- Definir las variables independientes X_1, X_2, X_3 y la variable dependiente Y.

2. Cálculo Manual de los Coeficientes

- Usar la fórmula matricial de Mínimos Cuadrados Ordinarios (MCO):

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Agregar una columna de unos a X para incluir el intercepto.
- Calcular los coeficientes β utilizando la fórmula.
- Almacenar las predicciones en un archivo Excel llamado `predicciones.xlsx`.

3. Evaluación Manual del Modelo

- Calcular los residuos:

$$\text{errores} = Y - \hat{Y}$$

- Calcular SSE (Suma de los Errores al Cuadrado) y SST (Suma Total de Cuadrados):

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

- Calcular el R^2 ajustado:

$$R^2_{\text{ajustado}} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

- Guardar los coeficientes calculados y los resultados.

4. Regresión con `scikit-learn`

- Definir X y Y nuevamente.
- Ajustar el modelo de regresión lineal con:

```
LinearRegression(fit_intercept=True)
```

- Obtener los coeficientes β y el intercepto con `scikit-learn`.
- Comparar los coeficientes con los obtenidos mediante la fórmula.

5. Evaluación del Modelo con `scikit-learn`

- Calcular las predicciones $\hat{Y}$ usando `modelo.predict(X)`.
- Calcular R^2 con `modelo.score(X, Y)`.
- Calcular el R^2 ajustado mediante la fórmula:

$$R^2_{\text{ajustado}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

6. Comparación de Resultados

- Comparar los coeficientes β, el R^2, y el R^2 ajustado obtenidos mediante la fórmula y usando `scikit-learn`.
- Reflexionar sobre las diferencias encontradas y su impacto en la precisión del modelo.

1. Carga de Datos y Análisis de Correlación

- Leer el archivo Excel `multicolinealidad.xlsx`.
- Mostrar las primeras filas del conjunto de datos para familiarizarse con ellos.
- Calcular la matriz de correlación entre las variables X_1, X_2, y X_3:

$$\text{Matriz de Correlación} = X.corr()$$

- Identificar la correlación entre las variables.

2. Análisis de la Multicolinealidad (VIF)

- Calcular el **VIF (Variance Inflation Factor)** para las variables independientes.
- Si el VIF es superior a 10, indica un alto grado de multicolinealidad.
- Usar la fórmula para calcular el VIF para cada variable:

$$\text{VIF} = \frac{1}{1 - R^2_{\text{variable}}}$$

- Mostrar los resultados del VIF para X_1, X_2, y X_3.

3. Regresión Lineal con Todas las Variables

- Definir X con las variables X_1, X_2, y X_3, y Y como la variable dependiente.
- Ajustar un modelo de regresión lineal múltiple utilizando `scikit-learn`.
- Obtener los coeficientes y el intercepto:

$$\text{Coeficientes} = \beta_1, \beta_2, \beta_3$$

- Calcular R^2 y R^2 ajustado para evaluar el modelo:

$$R^2_{\text{ajustado}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

4. Predicciones con Todas las Variables

- Calcular las predicciones $\hat{Y}$ del modelo utilizando X_1, X_2, X_3.
- Mostrar las primeras filas de Y_{real} vs Y_{predicho} y los errores.
- Realizar una gráfica de los valores reales frente a los predichos:

Gráfico: Y_{real} vs. Y_{predicho}

5. Regresión Lineal con Variables No Correlacionadas

- Definir X solo con X_2 y X_3, eliminando X_1 debido a su alta correlación con las otras variables.
- Ajustar un modelo de regresión lineal con estas dos variables.
- Obtener los coeficientes y el intercepto del modelo.
- Calcular el R^2 y R^2 ajustado para este nuevo modelo.

6. Comparación de Resultados

- Comparar los resultados del modelo que incluye X_1 frente al modelo que excluye X_1.
- Comparar los valores de R^2 y R^2 ajustado:

$$\Delta R^2 = R^2_{\text{completo}} - R^2_{\text{reducido}}$$

- Evaluar el impacto de la multicolinealidad en los coeficientes del modelo y en la precisión de las predicciones.

1. Carga de Datos y Definición del Modelo

- Leer el archivo `heterocedasticidad.xlsx`.
- Definir las variables independientes X_1, X_2, X_3 y la variable dependiente Y.
- Ajustar un modelo de regresión lineal múltiple utilizando `scikit-learn`.
- Mostrar los coeficientes y el intercepto:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

2. Evaluación del Modelo

- Calcular R^2 para evaluar la bondad del ajuste del modelo:

$$R^2 = 1 - \frac{\text{Suma de los errores al cuadrado}}{\text{Suma total de cuadrados}}$$

- Calcular el R^2 ajustado para corregir la varianza explicada por el número de predictores:

$$R^2_{\text{ajustado}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

- Mostrar los valores de R^2 y R^2 ajustado.

3. Predicciones y Errores

- Calcular las predicciones $\hat{Y}$ del modelo.
- Definir los residuos errores $= Y - \hat{Y}$.
- Mostrar una tabla con las primeras filas de Y_{real}, Y_{predicho}, y los errores.

4. Gráficos de Residuos (Heterocedasticidad)

- Visualizar los residuos frente a cada variable independiente X_1, X_2, X_3.
- Si hay heterocedasticidad, los residuos mostrarán un patrón (dispersión no constante) al aumentar los valores de X.
- Generar un gráfico de dispersión para los residuos frente a cada variable X:

Gráfico: Residuos vs X_1, X_2, X_3

5. Subgráficos de Residuos en una Figura

- Crear una figura con múltiples subgráficos para mostrar los residuos frente a todas las variables X_1, X_2, y X_3 en una sola figura.
- Cada subplot debe contener un gráfico de residuos vs. una variable X.
- Ajustar el diseño para evitar solapamientos entre los subgráficos.

6. Evaluación de la Normalidad de los Residuos

- Generar un histograma de los residuos para evaluar si siguen una distribución normal.
- Crear un gráfico Q-Q (Quantile-Quantile) para comparar los residuos con una distribución normal teórica:

Gráfico: $Q - Q$ de residuos vs. distribución normal

- Si los puntos siguen una línea recta, los residuos son aproximadamente normales.

1. Generación de Residuos No Normales

- Generar residuos que no siguen una distribución normal, utilizando una distribución exponencial para simular datos sesgados.
- Ajustar una semilla aleatoria para reproducibilidad.
- La fórmula para generar residuos sesgados es:

$$\text{residuos} = \text{np.random.exponential(scale=1, size=100)} - 1$$

- Estos residuos no son simétricos, mostrando un sesgo positivo.

2. Visualización de Residuos con Histograma

- Crear un histograma para visualizar la distribución de los residuos generados.
- Si los residuos fueran normales, deberíamos observar una distribución simétrica en forma de campana.
- En este caso, la distribución será asimétrica (sesgo).

3. Gráfico Q-Q para Evaluar la Normalidad

- Utilizar un gráfico Q-Q (Quantile-Quantile) para comparar los residuos con una distribución normal teórica.
- Si los residuos son normales, los puntos deberían alinearse a lo largo de una línea recta.
- En este caso, los puntos no seguirán la línea, indicando que los residuos no son normales.

5. Interpretación de los Resultados

- El **histograma** muestra que los residuos tienen una distribución sesgada hacia la derecha.
- El **gráfico Q-Q** indica que los residuos no siguen una distribución normal, ya que los puntos no están alineados sobre la línea recta.
- La normalidad de los residuos es un supuesto clave en los modelos de regresión, y su incumplimiento puede afectar la validez de los resultados.