UNIVERSITAS
Miguel Hernández

Tema 1. Regresión Lineal.

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.

Índice

1. Introducción
2. Fundamentos Lineal Múltiple
3. Violaciones de supuestos
4. Transformación de variables.

Objetivos.

- Comprender el concepto de **regresión lineal** como una técnica para modelar la relación entre una variable dependiente y una o más variables independientes.
- Estimar los **coeficientes de la regresión** usando el método de **Mínimos Cuadrados Ordinarios (MCO)**.
- Analizar la **bondad del ajuste** a través del coeficiente de determinación R^2 y el R^2 ajustado.
- Entender los **supuestos del modelo de regresión lineal** y cómo afectan la validez de los resultados.
- Identificar y resolver problemas comunes como la **multicolinealidad**, la **heterocedasticidad** y la **autocorrelación**.
- Aplicar la regresión lineal en distintos contextos para la **predicción y análisis** de relaciones entre variables.

Introducción

- La regresión lineal múltiple es una extensión de la regresión lineal simple.
- Se utiliza para modelar la relación entre una variable dependiente Y y varias variables independientes $X_1, X_2, \ldots, X_p$.
- Fórmula general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Ejemplo (relación del peso con la altura y el índice de masa muscular):

$$Y_{peso} = 0.9 X_{altura} + 2.3 X_{IndiceMasaMuscular} + \epsilon$$

Para una persona tal que $X_{altura} = 1.75m$ e $X_{IndiceMasaMuscular} = 22kg/m^2$, predeciremos un peso:

$$\hat{Y}_{peso} = 0.9 * 1.75 + 2.3 * 22 = 52.175kg$$

Supuestos de la Regresión Lineal

1. Linealidad: La relación entre Y y las X es lineal.
2. Independencia: Las observaciones son independientes.
3. Homocedasticidad: La varianza del error es constante.
4. No multicolinealidad: Las variables independientes no están altamente correlacionadas.
5. Normalidad: Los errores ϵ siguen una distribución normal.

Estimación de Parámetros

- El método de estimación más utilizado es el de Mínimos Cuadrados Ordinarios (MCO).
- La estimación de los coeficientes se realiza minimizando la suma de los errores al cuadrado:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Matricialmente:

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

Estimación de Parámetros (continuación)

- Para llegar a la estimación matricial de $\hat{\boldsymbol{\beta}}$, recordemos que el modelo de regresión lineal múltiple puede expresarse de forma matricial como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Donde:
  - $\mathbf{Y}$ es un vector de n observaciones de la variable dependiente.
  - $\mathbf{X}$ es la matriz de diseño de tamaño $n \times (p+1)$, que contiene las p variables independientes (con un término de sesgo o intercepto).
  - $\boldsymbol{\beta}$ es un vector de los coeficientes que queremos estimar.
  - $\boldsymbol{\epsilon}$ es un vector de los errores.

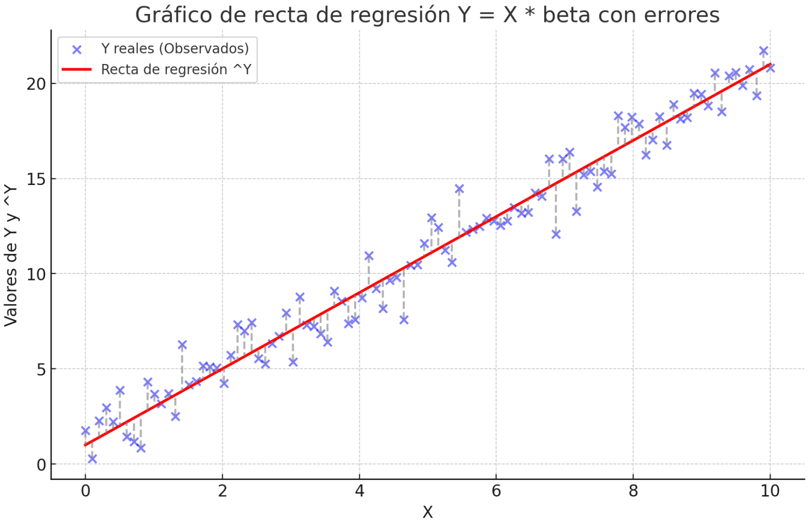
El enfoque de Mínimos Cuadrados Ordinarios (MCO)

- El objetivo del MCO es encontrar el valor de $\boldsymbol{\beta}$ que minimice la suma de los errores cuadrados, es decir, el residuo entre los valores observados $\mathbf{Y}$ y los valores ajustados $\hat{\mathbf{Y}} = \mathbf{X}\beta$.
- El criterio de minimización se expresa como:

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

- Esta es la función que deseamos minimizar respecto a $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$

Minimización de la Suma de Cuadrados de los Errores



Minimización de la Suma de los Errores Cuadrados

- Expandimos el criterio de minimización:

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

- Al desarrollar:

$$S(\beta) = \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

- Ahora, derivamos esta expresión con respecto a β y la igualamos a cero para encontrar el valor que minimiza la función.

Derivación e Igualación a Cero

- La derivada de la función de los errores cuadrados respecto a $\boldsymbol{\beta}$ es:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta$$

- Igualamos a cero para minimizar:

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

- Simplificando:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$$

Solución para β

- Finalmente, para despejar $\boldsymbol{\beta}$, multiplicamos ambos lados por $(\mathbf{X}^T\mathbf{X})^{-1}$, asumiendo que $\mathbf{X}^T\mathbf{X}$ es invertible:

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

- Esta es la fórmula de Mínimos Cuadrados Ordinarios (MCO) que proporciona el estimador de $\boldsymbol{\beta}$.
- Este estimador minimiza la suma de los errores al cuadrado y proporciona la mejor aproximación lineal para los coeficientes $\boldsymbol{\beta}$.

Regresión Lineal directamente como aplicación de la fórmula MCO

```python
# Convertimos X e Y a matrices NumPy
X = np.array(X)
Y = np.array(Y)
# Agregamos la columna de 1s a X para el intercepto
# Si no se agregase no habria intercepto (b0) en la regresion
X_b = np.hstack([np.ones((X.shape[0], 1)), X])
# Formula MCO: (X'X)^-1 X'Y
X_t = X_b.T
beta = np.linalg.inv(X_t @ X_b) @ (X_t @ Y)
print(f"Coeficientes:\n{beta}")
```

Regresión Lineal con Python + scikit-learn

```python
from sklearn.linear_model import
    LinearRegression
# Crear el modelo de reg. lineal
modelo = LinearRegression(fit_intercept=True)

# Ajustar el modelo a los datos
modelo.fit(X, Y)

# Coeficientes del modelo
print(f"Coeficientes: {modelo.coef_}")
print(f"Intercepto: {modelo.intercept_}")
```

Bondad del Ajuste

- R^2 (coeficiente de determinación). Indica la proporción de variabilidad de Y explicada por el modelo.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- R^2 ajustado: Penaliza la inclusión de variables que no mejoran el modelo.

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

Obtención de los R^2 con Python + scikit-learn

```
# Obtener el R^2
r_2 = modelo.score(X, Y)
print(f"R^2: {r_2}")

# Calcular el R^2 ajustado
n = X.shape[0]  # No. de observaciones
p = X.shape[1]  # No. de predictores (
    variables independientes)

r_2_ajustado = 1 - (1 - r_2) * (n - 1) / (n -
    p - 1)
print(f"R^2 ajustado: {r_2_ajustado}")
```

Ejemplo 1 de regresión múltiple.

El fichero *regresion.py* de la carpeta de ejemplos *Regresión* lee el fichero *Regresion.xlsx*. A continuación realiza las siguientes operaciones:

- Obtiene 'manualmente' los coeficientes MCO del modelo.
- Obtiene 'manualmente' las predicciones y los errores en base a los coeficientes obtenidos.
- Obtiene 'manualmente' el R^2 y el R^2 ajustado.

Ejemplo 1 de regresión múltiple (continuación).

Al igual que manualmente, en *regresion.py* también se realizan las siguientes operaciones llamando a la librería scikit-learn:

- Obtiene con sklearn los coeficientes MCO del modelo.
- Obtiene con sklearn las predicciones y los errores en base al modelo obtenido
- Obtiene con sklearn el R^2 y el R^2 ajustado.

Multicolinealidad

- Ocurre cuando dos o más variables independientes están altamente correlacionadas.
- Esto genera inestabilidad en las estimaciones de los coeficientes.
- Soluciones:
  - Eliminar una de las variables correlacionadas.
  - Realizar análisis de componentes principales.

Detección de multicolinealidad

Las técnicas más habituales para detectar la multicolinealidad son:

- **Matriz de correlaciones de las variables independientes.** En general se considera problemática aquella correlación superior a 0.8 o 0.9.
- **Factor de Inflación de la Varianza (VIF).** Mide cuánto aumenta la varianza de un coeficiente estimado debido a la multicolinealidad. Un valor de VIF mayor a 10 se considera indicativo de multicolinealidad alta.
- **Autovalores**. Una diferencia de autovalores mayor de 30 o 100, resulta problemática.
- **Signos o magnitudes inusuales en los coeficientes**. Cuando hay multicolinealidad, los coeficientes estimados pueden volverse inestables, cambiar de signo inesperadamente, o tomar valores extremadamente altos o bajos.

En esta asignatura nos ceñiremos a emplear las dos primeras técnicas: matriz de correlaciones y VIF.

Detección de multicolinealidad: matriz de correlaciones.

- En general se considera problemática aquella correlación superior a 0.8 o 0.9.
- Resolveremos la multicolinealidad eliminando las variables altamente correladas.

	X1	**X2**	**X3**
X1	1.000000	0.999394	0.190840
X2	0.999394	1.000000	0.190352
X3	0.190840	0.190352	1.000000

```
correlaciones = X.corr()
```

Detección de multicolinealidad: Factor de Inflación de la Varianza (VIF).

- En general se considera problemáticas aquellas variables con un VIF superior a 10.
- Resolveremos la multicolinealidad eliminando paso a paso e iterativamente las variables con un VIF más alto y superior a 10.

Variable	VIF
X1	835.729467
X2	835.585720
X3	1.034520

Detección de multicolinealidad: Factor de Inflación de la Varianza (VIF).

```python
from statsmodels.stats.outliers_influence import variance_inflation_factor
# Calcular el VIF para cada variable independiente
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

Heterocedasticidad

- Ocurre cuando la varianza de los errores no es constante.
- Esto puede invalidar las pruebas de significancia estadística así como proporcionar predicciones de baja calidad.
- Detección visual de la heterocedasticidad de los errores: representación de cada X vs ϵ.
- Test de Breusch-Pagan para detectar heterocedasticidad.
- Soluciones:
  - **Transformación de las variables** (logaritmos, raíces cuadradas).
  - Uso de regresión robusta (alternativa a MCO).

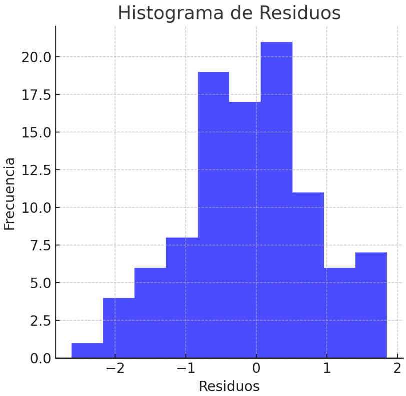
Normalidad de los Residuos

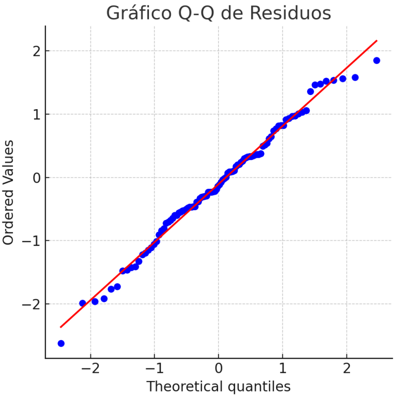
- La normalidad de los residuos es uno de los supuestos fundamentales en la regresión lineal.
- Se refiere a que los errores (residuos) deben seguir una distribución normal $\mathcal{N}(0, \sigma^2)$.
- La normalidad de los residuos es importante porque:
  - Permite realizar pruebas de significancia válidas (por ejemplo, los test t o F).
  - Asegura la eficiencia de los estimadores obtenidos.
- Si los residuos no son normales, las pruebas de hipótesis podrían no ser válidas y los intervalos de confianza podrían estar sesgados.

Normalidad en los Residuos: Detección

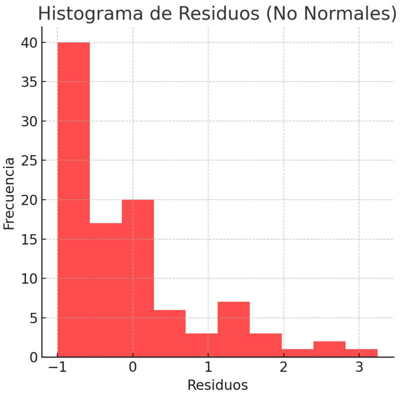
- Existen varias herramientas gráficas y pruebas estadísticas para verificar la normalidad de los residuos:
  - Histograma de residuos: Un histograma de los residuos debe tener forma de campana (distribución normal).
  - Gráfico Q-Q (Quantile-Quantile): Compara los cuantiles de los residuos con los de una distribución normal. Si los puntos siguen una línea recta, los residuos son normales.
  - Pruebas estadísticas:
    - Prueba de Shapiro-Wilk: Contrasta la hipótesis de que los residuos provienen de una distribución normal.
    - Prueba de Kolmogorov-Smirnov: Otra prueba que compara la distribución de los residuos con la normal.
- Si los residuos muestran una desviación significativa de la normalidad, es necesario tomar medidas para corregir este problema.

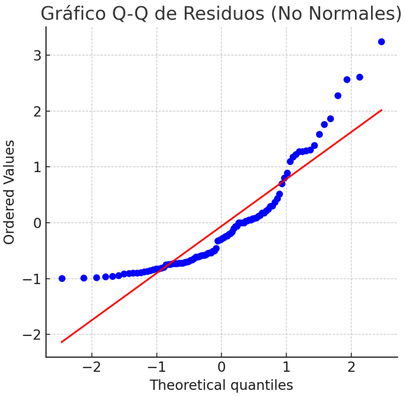
Normalidad en los Residuos: Detección





Normalidad en los Residuos: Detección





Normalidad en los Residuos: Soluciones

- Si los residuos no son normales, existen varias estrategias para corregir el problema:
  - Transformaciones de las variables:
  - Transformación Box-Cox: Encuentra automáticamente la mejor transformación para los datos.
  - Eliminar o ajustar outliers: Los valores atípicos pueden generar residuos no normales. Identifícalos (por ejemplo con el diagrama de caja-bigotes) y evalúa su impacto.
  - Regresión robusta (alternativa a MCO).
- Estas estrategias mejoran la normalidad de los residuos y permiten realizar inferencias más precisas en el modelo.

Autocorrelación

- Ocurre cuando los errores no son independientes entre sí.
- Test de Durbin-Watson para detectar autocorrelación.
- Soluciones:
  - Incluir variables retardadas.
  - Modelos ARIMA para series temporales.

No será estudiado este problema en la presente asignatura.

Transformaciones de las Variables Independientes

- Las transformaciones pueden ayudar a resolver problemas como la heterocedasticidad, la no linealidad o valores atípicos.
- Transformaciones comunes:
  - **Logaritmo natural ($\log X$)**: Común cuando las relaciones no son lineales o cuando la varianza crece con los valores de las X.

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon$$

  - **Raíz cuadrada ($\sqrt{X}$)**: Utilizada para reducir la variabilidad en los valores altos de X.

$$Y = \beta_0 + \beta_1 \sqrt{X} + \epsilon$$

Transformaciones de las Variables Independientes

- Transformaciones comunes (continuación):
  - **Inversa $(1/X)$**: Útil cuando Y se aproxima a un límite asintótico a medida que X crece.

$$Y = \beta_0 + \beta_1 \frac{1}{X} + \epsilon$$

  - **Polinomios $(X^2, X^3, \ldots)$**: Permite modelar relaciones no lineales entre X e Y.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Estas transformaciones pueden mejorar el ajuste y los supuestos del modelo.

Conclusiones

- La regresión lineal múltiple es una herramienta muy importante, pero debe utilizarse con cuidado.
- Es importante verificar los supuestos del modelo y tratar los problemas como la multicolinealidad y la heterocedasticidad.
- Un buen análisis de regresión requiere una evaluación tanto estadística como conceptual del modelo.
- Las transformaciones de las variables pueden ser útiles para mejorar el ajuste y cumplir con los supuestos del modelo.

UNIVERSITAS
Miguel Hernández