Regresión lineal general (Múltiple)

▼ 1. Introducción a la Regresión Lineal Múltiple.

Definición y propósito de la regresión lineal múltiple.

La regresión lineal múltiple es una extensión de la regresión lineal simple que permite analizar la relación entre una variable dependiente y múltiples variables independientes. En la regresión lineal simple, hemos estudiado la relación entre dos variables, mientras que en la regresión lineal múltiple, se considera la influencia simultánea de dos o más variables independientes sobre la variable dependiente.

Se aplica en situaciones en las que se sospecha que puede haber múltiples factores que pueden afectar a la variable dependiente y cuyos datos estén en nuestro DataFrame. Una parte importante del análisis de una Regresión Lineal Múltiple será decidir qué variables merecen ser tenidas en cuenta y cuáles no.

1. **Selección de Variables:**
2. **Diagnóstico del Modelo:**
    - Al igual que en la regresión lineal simple, se realizan diagnósticos del modelo para verificar supuestos, identificar problemas y mejorar la calidad del ajuste.

▼ 2. Aplicaciones prácticas en diversos campos.

- **Economía y Finanzas:** Predicción de precios de acciones, análisis de factores que afectan el crecimiento económico, etc.
- **Ciencias Sociales:** Estudio de factores que influyen en la satisfacción laboral, rendimiento académico, etc.
- **Ciencia de Datos:** Modelado predictivo en análisis de datos, como en el campo de aprendizaje automático.
- **Ingeniería:** Predicción de rendimiento de sistemas, análisis de variables en ingeniería de procesos, etc.

▼ 3. Ecuación de Regresión Lineal Múltiple.

La ecuación de regresión lineal múltiple toma la forma general:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

Y es la variable dependiente.

$X_1, X_2, \ldots, X_n$ son las variables independientes.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes que representan la relación entre cada variable independiente y la dependiente.

Cada coeficiente β en la ecuación de regresión indica la magnitud y dirección de la relación entre la variable dependiente y la correspondiente variable independiente, manteniendo el valor de las demás variables constante.

> ☝ La interpretación precisa de los coeficientes es fundamental para entender cómo cada variable contribuye al modelo.

ε es el término de error, que representa la variación no explicada por el modelo.

▼ 4. Supuestos de la Regresión Lineal Múltiple.

La regresión lineal múltiple se basa en varios supuestos que, cuando se cumplen, garantizan la validez y eficacia del modelo. Es importante tener en cuenta estos supuestos al interpretar los resultados de la regresión y al realizar inferencias sobre la población. Aquí están los supuestos básicos de la regresión lineal múltiple:

1. **Linealidad:**
    - Supone que la relación entre las variables independientes y la variable dependiente es lineal. Esto significa que los cambios en la variable dependiente son proporcionales a los cambios en las variables independientes.
    - Los coeficientes de la regresión son constantes para cualquier observacion. Esto significa que el impacto de un cambio en una variable independiente en la variable dependiente es constante en toda la gama de valores.
2. **Normalidad de Errores:**
    - Se asume que los errores ε se distribuyen normalmente. Esto es importante para realizar pruebas de hipótesis y para la construcción de intervalos de confianza. Sin embargo, para tamaños de muestra

grandes, la regresión es robusta a esta asunción debido al teorema del límite central.

3. **Homocedasticidad:**
    - La varianza de los errores debe ser constante para todas las combinaciones de valores de las variables independientes. En otras palabras, la dispersión de los errores debe ser uniforme a lo largo de toda la gama de valores predichos.
    - Esto significa que el error asociado con una observación no está relacionado con el error de otra observación. La autocorrelación de errores puede afectar la eficiencia de los estimadores.
4. **No Colinealidad Perfecta:**
    - Las variables independientes no deben estar perfectamente correlacionadas. La colinealidad perfecta (correlación igual a ±1 entre dos o más variables independientes) puede causar problemas en la estimación de los coeficientes.
        - La colinealidad perfecta impide identificar un modelo único, ya que varias configuraciones de coeficientes pueden producir la misma combinación lineal de variables.
        - Pequeñas variaciones en los datos pueden resultar en grandes cambios en los coeficientes estimados. Esto hace que los coeficientes sean inestables y difíciles de interpretar.
        - En el cálculo de los coeficientes mediante la inversión de matrices, la colinealidad perfecta implica que la matriz de variables independientes no sea invertible. Esto conlleva a problemas numéricos en la estimación de los coeficientes.

Es importante evaluar estos supuestos antes de interpretar los resultados de la regresión y, si es necesario, aplicar técnicas de corrección o transformación de datos para abordar posibles incumplimientos. Además, algunos de estos supuestos pueden ser relajados en ciertos casos, dependiendo del contexto y del tamaño de la muestra.

▼ 5. Cómo Abordar la Colinealidad.

Antes de realizar la regresión, es importante examinar la matriz de correlación entre las variables independientes para detectar cualquier signo de colinealidad.

1. **Identificación y Eliminación de Variables Redundantes:**

- Si se identifica colinealidad, se puede considerar la eliminación de una de las variables redundantes. Esto implica conservar solo las variables que aportan información única al modelo.

2. **Transformación de Variables:**

   Transformar las variables puede ser útil para reducir la colinealidad. Por ejemplo, tomar logaritmos, diferencias o combinaciones lineales específicas.

   Algunas técnicas comunes de transformación de variables que se utilizan para abordar la colinealidad incluyen:

   1. **Estandarización:**
      - La estandarización implica restar la media y dividir por la desviación estándar de cada variable. Esto no cambia la relación lineal entre las variables, pero puede ayudar a reducir la magnitud de las diferencias en las escalas entre ellas.
   2. **Transformaciones Logarítmicas o Exponenciales:**
      - Tomar el logaritmo o la exponencial de una variable puede ser útil si la relación entre las variables es no lineal. Esto puede ayudar a estabilizar la varianza y reducir la colinealidad.
   3. **Diferenciación:**
      - Calcular diferencias entre observaciones consecutivas de una variable puede ayudar a eliminar patrones de tendencia o estacionalidad, reduciendo así la colinealidad.
   4. **Creación de Variables Interactivas o Cuadráticas:**
      - Introducir términos de interacción o cuadráticos puede modificar la relación entre las variables y reducir la colinealidad. Por ejemplo, si hay una relación lineal entre X_1 y X_2, añadir un término $X_1 \times X_2$ podría ayudar.
   5. **Reducción de Dimensionalidad:**
      - Técnicas como el Análisis de Componentes Principales (PCA) o técnicas de reducción de dimensionalidad pueden ayudar a reducir la colinealidad al transformar las variables originales en un conjunto de variables no correlacionadas (componentes principales).

Es recomendable evaluar la efectividad de las transformaciones realizadas mediante la inspección de la matriz de correlación y otros diagnósticos de colinealidad después de aplicarlas. En algunos casos, puede ser necesario probar varias transformaciones para encontrar la que mejor aborda el problema de colinealidad.

3. **Regularización:**
    - Técnicas de regularización, como la regresión ridge o la regresión lasso, introducen penalizaciones en los coeficientes para reducir la multicolinealidad y estabilizar los resultados.

La regresión Ridge y la regresión Lasso

▼ 6. Selección de Variables.

- Métodos para seleccionar variables predictoras.

  La selección de variables predictoras es un paso crítico en el desarrollo de modelos predictivos. Se busca identificar las variables más relevantes y eliminar las redundantes o poco informativas.

  Éstos son los métodos más comunes para la selección de variables predictoras.

  1. **Métodos de Filtrado:**

     Estos métodos evalúan la relación de cada variable independiente con la variable dependiente antes de ajustar el modelo de regresión.

     - **Correlación:**
       - Selecciona variables basándose en su correlación con la variable dependiente. Puedes establecer un umbral de correlación para incluir solo aquellas variables que tienen una correlación significativa.
     - **Pruebas de Hipótesis Univariables:**
       - Utiliza pruebas estadísticas como la prueba t o la prueba F para evaluar la relación entre cada variable y la variable dependiente. Las variables con p-valores significativos se consideran relevantes.

  2. **Métodos Wrapper:**

     Estos métodos evalúan el rendimiento del modelo utilizando diferentes combinaciones de variables.

- **Selección hacia Adelante (Forward Selection):**
  - Comienza con un modelo sin variables y agrega iterativamente la variable más relevante en cada paso hasta que se cumple un criterio de parada.
- **Eliminación hacia Atrás (Backward Elimination):**
  - Comienza con un modelo que incluye todas las variables y elimina iterativamente la menos relevante en cada paso hasta que se cumple un criterio de parada.
- **Selección Paso a Paso (Stepwise Selection):**
  - Combina elementos de la selección hacia adelante y hacia atrás, agregando o eliminando variables en cada paso basándose en algún criterio.

3. **Métodos Embedded:**

   Estos métodos incorporan la selección de variables en el proceso de ajuste del modelo.

   - **Regularización (Ridge y Lasso):**
     - Ridge y Lasso penalizan los coeficientes de las variables, lo que puede llevar a la selección automática de variables. Lasso tiene la capacidad de llevar algunos coeficientes a cero, realizando así selección de variables.
   - **Importancia de Variables en Árboles de Decisión:**
     - Los modelos basados en árboles, como Random Forest o Gradient Boosted Trees, proporcionan medidas de importancia de variables, que se pueden utilizar para seleccionar las más importantes.

4. **Métodos de Evaluación de Importancia:**

   Estos métodos evalúan la importancia de cada variable después de ajustar el modelo.

   - **Importancia de Variables en Bosques Aleatorios:**
     - Los bosques aleatorios asignan importancias a las variables basándose en su contribución a la reducción de la impureza en los nodos del árbol.

- **Importancia de Variables en Modelos Lineales:**
    - Algunos métodos, como el análisis de varianza de la regresión (ANOVA) o el análisis de sensibilidad de los coeficientes, pueden proporcionar indicadores de la importancia relativa de las variables en modelos lineales.

▼ Consideraciones Importantes.

- **Validación Cruzada:**
    - Es importante realizar la selección de variables en conjunción con técnicas de validación cruzada para evitar el sobreajuste a un conjunto de datos específico.
- **Interpretación del Modelo:**
    - La selección de variables debe equilibrarse con la interpretación del modelo. Es crucial entender la relevancia práctica y teórica de las variables seleccionadas.
- **Enfoque Contextual:**
    - La elección del método de selección de variables depende del contexto del problema y de la naturaleza de los datos. No hay un método único que sea óptimo para todas las situaciones.

Al aplicar estos métodos, es recomendable tener en cuenta la complejidad del problema, la cantidad de datos disponibles y la interpretación deseada del modelo resultante.

7. Interpretación de Coeficientes.

- Interpretación práctica de los coeficientes en el contexto de variables múltiples.
- Coeficientes parciales y cómo afectan a la variable dependiente manteniendo otras constantes.

8. Diagnóstico del Modelo.

- Métodos para diagnosticar la calidad del modelo de regresión múltiple.
- Análisis de residuos y detección de puntos atípicos.
- Uso de gráficos y pruebas estadísticas.