U N I V E R S I D A D D E G U A D A L A J A R A

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERÍA

DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Seminario de Solución de Problemas de Sistemas Basados en Conocimiento

Práctica No. 4

Regresión Lineal Múltiple

Nombre: Hurtado González Edgar Arturo
Código: 212597894

Introducción

La regresión lineal múltiple es una técnica estadística que se usa para modelar la relación entre una variable dependiente y dos o más variables independientes. A diferencia de la regresión lineal simple que involucra solo dos variables (una dependiente y una independiente), la regresión lineal múltiple maneja múltiples variables independientes.

Desarrollo – Regresión Multilineal

En esencia, la regresión lineal múltiple encuentra la mejor ecuación lineal que describe la relación entre las variables independientes y la variable dependiente. El objetivo es encontrar coeficientes para cada variable independiente que minimicen la diferencia entre los valores predichos por el modelo y los valores reales observados.

La ecuación para un modelo de regresión lineal múltiple con 'p' variables independientes se ve así:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Y es la variable Dependiente.
- $X_1, X_2, \ldots, X_p$ son las 'p' variables independientes.
- $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes que representan el cambio en 'Y' por cada unidad de cambio en las variables independientes, manteniendo las otras variables constantes.
- ϵ es el término de error que representa la diferencia entre los valores observados y los valores predichos por el modelo.

El análisis de regresión múltiple implica evaluar la significancia estadística de cada coeficiente, la bondad de ajuste del modelo (cuánto se ajustan los valores predichos a los valores observados), la multicolinealidad (la interrelación entre las variables independientes), entre otros aspectos.

Se utilizan métodos como el método de mínimos cuadrados para estimar los coeficientes, y se pueden realizar pruebas de hipótesis y análisis de residuos para evaluar la idoneidad del modelo.

La regresión lineal múltiple es ampliamente utilizada en campos como la economía, la ciencia de datos, la investigación médica y muchas otras áreas donde se necesita comprender y predecir relaciones complejas entre múltiples variables.

La regresión lineal múltiple es una herramienta poderosa para entender y predecir relaciones complejas entre variables. Aquí hay algunos conceptos clave para profundizar en esta técnica:

Supuestos de la regresión lineal múltiple:

1. *Linealidad:* La relación entre las variables dependientes e independientes debe ser lineal. Esto significa que los cambios en la variable dependiente deben estar linealmente relacionados con los cambios en las variables independientes.
2. *Homocedasticidad*: La varianza de los errores debe ser constante a lo largo de todas las combinaciones de valores de las variables independientes. En otras palabras, los residuos (diferencia entre los valores observados y predichos) deben tener una dispersión constante.
3. *Independencia de errores:* Los errores o residuos del modelo no deben mostrar patrones discernibles ni estar correlacionados entre sí. Cada error debe ser independiente de los otros.
4. *Normalidad de errores:* Los errores deben seguir una distribución normal. Aunque este no es un requisito estricto para todos los tamaños de muestra, un gran número de observaciones tiende a seguir la distribución normal debido al teorema del límite central.

Evaluación del modelo:

- *R cuadrado (R^2):* Esta métrica indica qué tan bien el modelo explica la variabilidad de los datos. Mide la proporción de la variación en la variable dependiente que es explicada por las variables independientes. Un valor cercano a 1 indica un buen ajuste, mientras que cercano a 0 indica un ajuste deficiente.
- *P-valores y significancia de coeficientes:* Se analizan los p-valores asociados con los coeficientes para determinar si las variables independientes tienen un efecto significativo sobre la variable dependiente.

Multicolinealidad:

Se refiere a la alta correlación entre dos o más variables independientes en el modelo. Puede afectar la precisión de los coeficientes estimados y dificultar la interpretación del impacto individual de cada variable. Técnicas como el análisis de correlación y el factor de inflación de la varianza (VIF) se utilizan para detectar y abordar este problema.

Pasos para construir un modelo de regresión lineal múltiple:

1. *Recopilación y preparación de datos:* Seleccionar las variables relevantes y limpiar los datos de posibles errores o valores atípicos.
2. *Selección del modelo:* Identificar qué variables incluir en el modelo y cómo pueden interactuar entre sí.
3. *Ajuste del modelo:* Estimar los coeficientes utilizando métodos como mínimos cuadrados ordinarios u otros algoritmos de optimización.
4. *Evaluación del modelo:* Analizar la significancia de los coeficientes, la bondad de ajuste (R^2), la multicolinealidad y otros diagnósticos de residuos.
5. *Validación y ajuste:* Probar el modelo en datos nuevos y realizar ajustes si es necesario para mejorar su desempeño predictivo.

La regresión lineal múltiple es una técnica fundamental en el análisis estadístico y se emplea ampliamente en la investigación científica, la predicción y la toma de decisiones en una amplia gama de campos.

El código hecho en esta ocasión, realiza un análisis de regresión lineal múltiple utilizando el descenso de gradiente para encontrar los coeficientes óptimos para predecir el índice de rendimiento de los estudiantes en función de varias características. Aquí está la lógica paso a paso:

1. *Importación de bibliotecas:* Se importan las bibliotecas necesarias: pandas para manejar datos, numpy para cálculos numéricos, matplotlib para visualización y train_test_split de sklearn para dividir los datos en conjuntos de entrenamiento y prueba.

2. *Definición de funciones*: Se definen funciones para el cálculo del costo, coeficiente de determinación (R^2), error cuadrático medio (MSE) y el descenso de gradiente.

3. *Preprocesamiento de datos:* Se lee un archivo CSV que contiene datos de rendimiento estudiantil. Se copia el DataFrame original y se normalizan algunas columnas numéricas. Las columnas categóricas se transforman a valores binarios.

4. *Configuración de datos:* Se seleccionan las características ('Hours Studied', 'Previous Scores', etc.) y se preparan para el entrenamiento del modelo. Se agrega una columna de unos para el término de sesgo (bias).

5. *División de datos:* Se dividen los datos en conjuntos de entrenamiento y prueba utilizando train_test_split.

6. *Configuración de parámetros:* Se definen la tasa de aprendizaje (cx), el número de iteraciones para el descenso de gradiente y se inicializan los coeficientes (theta).

7. *Descenso de gradiente*: Se llama a la función descenso_gradiente con los datos de entrenamiento para encontrar los coeficientes óptimos (theta) que minimizan el costo.

8. *Predicción y evaluación:* Se utilizan los coeficientes encontrados para predecir los valores del conjunto de prueba. Además, se calculan y muestran métricas como R^2 y se generan gráficos de dispersión y las líneas de regresión para visualizar la relación entre las características seleccionadas y el índice de rendimiento.

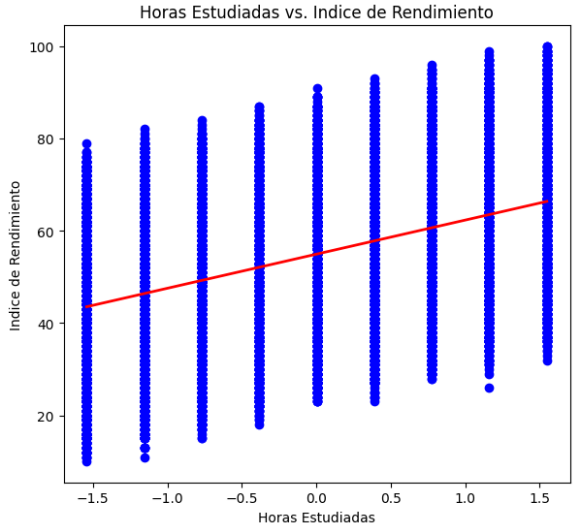
Conclusión

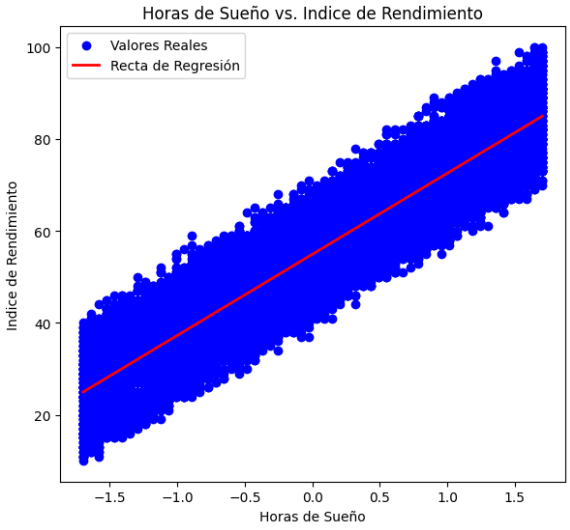
En conclusión, este código implementa un modelo de regresión lineal múltiple utilizando el descenso de gradiente para predecir el rendimiento estudiantil basado en diferentes características, y luego visualiza la relación entre algunas de esas características y el rendimiento.

Aborda desde el procesamiento de datos hasta la evaluación y visualización del modelo de regresión lineal múltiple, brindando una comprensión tanto de las relaciones entre las características y el rendimiento como de la capacidad predictiva del modelo.

Resultados

Regresión Multilineal





Referencias:

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Recuperado de https://www.discoveringstatistics.com/wp-content/uploads/2017/09/Discovering-Statistics-Using-IBM-SPSS-Statistics.pdf

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Pearson Education.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. Wiley.