Problema 1

1.- Modelo de regresión lineal simple:

X: Horas de estudio

Y: Calificacion obtenida

Yi = B0 + B1Xi + Ei

B0 es el intercepto ya sea calificacion esperada cuando las horas son 0

B1 es la pendiente ya sea el incremento promedio en la calificacion por cada hora adicional de estudio

Por ultimo la Ei es el error aleatorio

- Metrica de error

Suma de los Errores Cuadrados (SSE):

SSE = Sumatoria(Yi-Yi)²

Esta metrica penaliza mas los errores grandes y permite obtener una solucion matematica optima.

3.- Metodo de optimizacion

Minimos Cuadrados

B1 = Cov(X,Y) / Var (X)

B0 = Y - BiX

Calcula la pendiente como la razon entre la covarianza y la varianza de X lo que puedo decir que eso minimiza el error total.

Descripcion del comportamiento de la poblacion

2. Enfoque academico / estadistico

Puedo decir que el modelo muestra que existe una realcion lineal entre las horas de estudio y la calificacion.

Coeficiente B1: representa el cambio promedio esperado en la calif por cada hora adicional. Si su p - value es menor a 0.05 entonces existe evidencia estadisictica suficiente para afirmar que la relacion es significativa

Esto se debe a que entre mas horas estudia un alumno mejor calif tiende a conseguir u obtener.

No significa que estudiar garantice una calif exacta pero si que existe una tendencia clara que es que estudiar mas por asi decir generalmente ayuda a mejorar el resultado final.

Comparacion de p-value del coeficiente beta1 y p-value del estadístico F:

Esto se debe a que es una regresion lineal simple. El estadistico F del modelo evalua la misma hipotesis que el estadistico t del coef B1.

Matemáticamente:

F = t^2

H0: B1 = 0

Ya con esta explicacion puedo decir que hacen o generan el mismo p-value ya que esto solo ocurre cuando hay un unico predictor.

3.- Estudio de predicción

Unos cambios que realizaria yo creo que podrian ser:

- mostrar graficas claras y no solo inferencia.
- Detectar y manejar outliers
- Automatizar el caluclo de metricas

Dividir los datos en entrenamiento y prueba ya sea 70%-30%.

Estos cambios permitirian mostrar que tan bien predice el modelo en datos nuevos y no solo en los datos de training o entrenamiento.

Metricas para evaluar calidad del modelo

Yo usaria RMSE la cual su formula es raiz cuadrada de 1 / n + sumatoria de (Yi - Yi con signo arriba de la Yi segunda)², porque penaliza grandes errores.

Otra seria MAE = 1 / n * sumatoria de |Yi-Yi con signo arriba de la segunda|, porque es facil de interpretar.

Para finalziar seria R^2 porque mide que proporcion de la variabilidad es explicada por el modelo.

Un modelo no lineal mejora la prediccion?

Un modelo de segundo orden podria tal vez mejorar un poco si existe efecto de rendimientos decrecientes.

Modelo polinomial de tercer orden no es recomendable porque tiene alto riesgo sobreajuste y sus comportamientos no son realistas.

Modelo regresion spline podria tambien mejorar si la relacion cambia a pendiente en diferentes rangos de horas.

KNN con k = 3, probable sobreajuste y muy sensible al ruido.

KNN con K = 300, igulamente sobreajuste y demasiado suavizado ya eso hace que pierda estructura.

```
import numpy as np
import pandas as pd
from scipy import stats

# Cambia la ruta si es necesario
data = pd.read_csv("P1 Datos1(2).csv")

# Asegúrate que las columnas se llamen exactamente así:
# horas (X) y calificacion (Y)
X = data.iloc[:, 0].values   # Horas de estudio
Y = data.iloc[:, 1].values   # Calificación

n = len(X)

X_mean = np.mean(X)
Y_mean = np.mean(Y)

beta1 = np.sum((X - X_mean)*(Y - Y_mean)) / np.sum((X - X_mean)**2)
beta0 = Y_mean - beta1 * X_mean

Y_hat = beta0 + beta1 * X


SSE = np.sum((Y - Y_hat)**2)
MSE = SSE / (n - 2)
RMSE = np.sqrt(MSE)

SST = np.sum((Y - Y_mean)**2)
R2 = 1 - (SSE / SST)


SE_beta1 = np.sqrt(MSE / np.sum((X - X_mean)**2))

t_stat = beta1 / SE_beta1


p_value_t = 2 * (1 - stats.t.cdf(abs(t_stat), df=n-2))
SSR = SST - SSE
F_stat = (SSR / 1) / (SSE / (n - 2))

p_value_F = 1 - stats.f.cdf(F_stat, 1, n-2)

print("=== RESULTADOS REGRESIÓN LINEAL SIMPLE ===\n")
print(f"Beta0 (Intercepto): {beta0}")
print(f"Beta1 (Pendiente): {beta1}\n")

print(f"SSE: {SSE}")
print(f"MSE: {MSE}")
print(f"RMSE: {RMSE}")
print(f"R^2: {R2}\n")

print(f"t-stat beta1: {t_stat}")
print(f"p-value (t): {p_value_t}\n")

print(f"F-stat: {F_stat}")
print(f"p-value (F): {p_value_F}")
```

```
=== RESULTADOS REGRESIÓN LINEAL SIMPLE ===

Beta0 (Intercepto): 48.07005444261577
Beta1 (Pendiente): 4.257566455755204
```

```
SSE: 36399.39696370549
MSE: 73.09115856165761
RMSE: 8.549336732265118
R^2: 0.7015346576331934

t-stat beta1: 34.21308865733037
p-value (t): 0.0

F-stat: 1170.535435474348
p-value (F): 1.1102230246251565e-16
```

Problema 2

Multicolinealidad:

ft2: tamaño en ft cuadrados m2: tamaño en m cuadrados

Estas 2 variables representan exactamente la misma info en distintas unidades y eso hace que exista colinealidad entre ambas.

Llegue a esta conclusion ya que son conversaciones directas y matematicamente una es combinacion lineal exacta de la otra y esto puede generar o genera mas bien inflacion en err estandares, problemas de interpretacion, etc.

- Tamaño de muestra pequeño

Como dice en las instrucciones nada mas hay 50 observaciones. Con varias variables explicativas, el modelo puede sobreajustar o no generalizar correctamente ya que pues en si son pocas observaciones.

- Posible variable irrelevante

Variable fibra tiene p - value = 0.938 y t = -0.079 y esto indica que no es estadisticamente significativa.

- Posible mala seleccion de variables

Nada mas se usó solo 2 variables del sistema lo cual no sabemos cuales ni bajo que criterio y por esto puede haber: Omision de variable relevante y mala especificacion del modelo.

Ideas de como los resolveria

Multicolinealidad: simplemente eliminar unas de las variables y conservar metros cuadrados yua que es la que mas se usa en México. Esto mejoraria la estabilidad del modelo. Es la mejor solucion porque ambas contienen exacatamente la misma info.

- Solucion a variable irrelevante:

p - value = 0.938 > 0.05 no hay evidencia estadistica que afecte el precio. algunas opciones serian eliminarla del modelo y mantenerla solo si existe justificación para simplificar el modelo.

- Solucion al tamaño de muestra:

No se puede modificar el tamaño y lo que haria seria validacion cruzada, separar datos en training y test y analizar estabilidada de coeficientes.

- Solucion a posible mala especificacion:

Revisaria si se omitio baños, si existe interaccion entre variables y si la relacion es estrictamente lineal. Compararia modelos usando R al cuadrado ajustado, AIC y BIC.

Comentario sobre la calidad del modelo mostrado

- Metricas relevantes de la tabla: R al cuadrado = 0.977 R al cuadrado ajustado = 0.976 Estadisitco - F = 969.3 Prob F = 2.50e-38

Calidad global: R al cuadrado = 0.977 indica que el modelo explica 97,7% de la variabilidad del precio lo cual es muy alto. La prueba F tiene p-value 0 eso da que el modelo globalmente es significativo y pues en conclusion el modelo es fuerte por asi decir.

- Relevancia individual de variables

m2:

coef = 0.9267 p-value. = 0.000

Muy significativo lo cual tiene fuerte impacto en el precio.

fibra:

coef = -0.3345 p-value = 0.938

No es significativo y no aporta por asi decri a la evidencia estadistica.

Falta alguna metrica?

Si, RMSE, VIF (colinealidad) y Division entrenamiento/prueba porque R al cuadrado no garantiza buena prediccion futura.

Interpretacion de coeficientes

Modelo estimado:

precio = 533.8875 + 0.9267(m al cuadrado) - 0.3345(fibra), trabajando con un 95% de confianza.

Intercepto:

Si una casa tuviera 0 m cuadrados y sin fibra el precio seria 533.89, interpretacion matematica y no practica.

Coeficiente de m al cuadrado:

Por cada metro cuadrado adicional el precio aumenta en prom 0.9267 manteniendo constante la variable fibra. El intervalo de confianza 95% es [0.884,0.970] y como no incluye el 0 el efecto es significativo.

Coeficiente fibra:

Manteniendo constante el tamaño, tener fibra cambiai el precio en -0.3345 pero no es estadisticamente significativo, el intervalo [-8.898, 8.229] incluye el 0 y no podemos concluir que tenga efecto real.

Problema 3

Combinar: Eliminacion hacia atras y criterio rapido tipo fast forward.

En lugar de comenar sin variables e ir agregando podemos comenzar con todas las variables y vamos eliminando asi rapido las menos relevantes usando un criterio eficiente por asi decir.

- Idea general del metodo:

Comenzar con las 50 variables, ajustaar un modelo completo, evaluar relevancia de cada variable, elminar rapido la peor variable segun un criterio estadistico, etc. El metodo busca ser computacionalmente mas eficiente que backward clásico.

- Criterio de eliminacion:

En este podemos usar: p-value mayor a 005, o mayor incremente en AIC/BIC o menor contribucion absoluta al modelo. Para hacer esto rapido se puede elminar directamente la peor variable en cada iteracion sin reevaluar combinacion multiples.

- Pseudocodigo

Inicio

1.- Cargar datos desde "Pq Datos3.csv" Leer matriz X (50 variables). Leer vector y (variable objetivo). n: numero de observaciones. p: 50.

2.- Definir: alpha: 0.05. Variables seleccionadas: {1,2,3,..,50}. Mejorar: Verdadero.

3.- Ajustar modelo completo: Modelo: Regresionlineal(X[Variables Seleccionadas], y).

4.- Mientras mejorar sea verdadeor hacer: Obtener p-values de cada varaibles en el modelo actual. Encontrar: j: indice con mayor p-value. Si p-value(j) > alpha entonces: Eliminar j de Variables Seleccionadas. Ajustar nuevo modelo con las variables restantes. Sino: Mejorar: Falso. Fin mientas.

5.- Devolver Variables Seleccionadas.

Fin.

- Version mas rapida con AIC

Puede ser como:

4.- En cada iteracion: Para cada variable j en Variables Seleccionadas: Calcular AIC sin variable j. Elminar la variable cuya eliminacion reduzca mas el AIC. Si ningun AIC mejora: detener.

Condiciones de paro posibles: Todas las variables restantes tienen p-value < alpha, AIC no mejora, se alcanza lmiite de iteraciones, etc.

- Generaria la misam solucion que seleccion rapdia hacia adelante

No, seleccion hacia adelante parte sin variables, solo agrega variables relevantes que individualmente mejoran el modelo.

Seleccion hacia atras: inicia con todas y evalua variables entre todas las demas.

Por correlaciones entre variables, interacciones ocultas y supresion estadistica puede pasar que una variable no sea significativa sola pero si sea importante cuando otras no estan presnetes.