Ejercicio 4: Explica lo siguiente.

a. ¿Qué supuestos del modelo de regresión lineal múltiple deben verificarse?

b. ¿Cómo se interpretan los intervalos de confianza? Si construimos un intervalo de confianza del 95% para un coeficiente βj

b. ¿cuál sería la lectura correcta o interpretación correcta sobre este intervalo?

c. Describe los métodos de selección de variables y sus ventajas y desventajas:

   Selección hacia adelante (forward)

   Selección hacia atrás (backward)

   selección por pasos (stepwise) y/o mejor subconjunto (best subset)

   Explica cómo se utilizan para elegir el modelo final.

a) ¿Qué supuestos del modelo de regresión lineal múltiple deben verificarse?

Los principales supuestos del modelo de regresión lineal múltiple que deben verificarse son los siguientes:

1. **Linealidad:** La relación entre las variables independientes y la variable dependiente debe ser lineal. Es decir:

   $Y = b_0 + b_1*X_1 + b_2*X_2 + \dots b_p*X_p$

   Si la relación de la variable dependiente (Y) con las variables independientes (X_i) no es lineal, la aproximación del modelo lineal será muy deficiente.

2. **Modelo de error**. El modelo de regresión lineal asume que tenemos muestras fijas de las X y la correspondiente Y está afectada además por un término de error:

   $Y_1 = b_0 + b_1*X_{11} + b_2*X_{12} + \dots b_p*X_{1p} + e_1$

   $Y_2 = b_0 + b_1*X_{11} + b_2*X_{12} + \dots b_p*X_{1p} + e_2$

   …

   $Y_n = b_0 + b_1*X_{11} + b_2*X_{12} + \dots b_p*X_{1p} + e_n$

   Donde los términos de error e_i siguen una distribución normal con media cero, varianza constante y covarianza $Cov(e_j, e_k) = 0$ ($j \neq k$). Esto significa que:

   - **Independencia de los errores**: Los residuos no deben estar correlacionados entre sí ($Cov(e_j, e_k) = 0$). En otras palabras, el término de error de una observación no contiene información útil para predecir el término de error de otra observación.
   - **Homocedasticidad**: La varianza de los errores debe ser constante para todos los valores de las variables independientes (X). Esto se ve claramente si tenemos una única X. Los valores probables de Y estarían en una franja constante alrededor de la línea recta calculada, independiente del valor de X, sea pequeño, intermedio o grande.
   - **Normalidad de los errores**: Los residuos deben distribuirse normalmente.

3. **No multicolinealidad perfecta**: Las variables independientes no deben estar perfectamente correlacionadas entre sí. Si están perfectamente correlacionadas, la matriz X'X no tiene inversa. Mientras mayor sea la correlación entre variables, el calculo de la inversa de X'X es numéricamente inestable.

b) ¿Cómo se interpretan los intervalos de confianza? Si construimos un intervalo de confianza del 95% para un coeficiente βj

Un intervalo de confianza del 95% para un coeficiente βj significa que si repitiéramos el estudio muchas veces con diferentes muestras de la misma población, el 95% de los intervalos construidos contendrían el verdadero valor del parámetro poblacional βj.

b) ¿cuál sería la lectura correcta o interpretación correcta sobre este intervalo?

Veamos un ejemplo con números. Si el intervalo de confianza del 95% para β_1 es (3.1, 3.5), esto significa que con el 95% de confianza, podemos afirmar que por cada unidad de aumento en X_1, la variable dependiente Y aumenta entre 3.1 y 3.5 unidades, manteniendo constantes las demás variables.

c) Describe los métodos de selección de variables y sus ventajas y desventajas:

Selección hacia adelante (forward)

Selección hacia atrás (backward)

selección por pasos (stepwise) y/o mejor subconjunto (best subset)

Explica cómo se utilizan para elegir el modelo final.

El objetivo principal es encontrar el subconjunto de variables predictoras que maximice la capacidad predictiva del modelo y minimice el error, evitando:

- Sobreajuste (Overfitting): Modelos demasiado complejos que aprenden el "ruido" de los datos en lugar de la tendencia general. Funcionan bien en los datos de entrenamiento pero mal en datos nuevos.
- Multicolinealidad: Variables redundantes que aportan la misma información.
- Costos de medición y complejidad: Incluir variables irrelevantes aumenta el costo de recolección de datos y la complejidad del modelo sin beneficio.

Veamos ahora los métodos más comunes:

1. Selección hacia adelante (Forward Selection)

Comienza con un modelo que no contiene ninguna variable (solo el intercepto, β_0). En cada paso, se añade la variable que proporciona la mejora estadísticamente más significativa al modelo (usando un criterio como el p-value), hasta que ninguna variable restante cumpla el criterio para entrar.

Proceso:

1. Empezar con $Y = b_0$
2. Para cada variable X_i no presente en el modelo, ajustar un nuevo modelo que la incluya y calcular un criterio de selección (ej., el p-value de la correspondiente b).

3. Añadir la variable con el p-value más pequeño (si es menor que un p-value predefinido, por ejemplo, de 0.05).
4. Repetir los pasos 2 y 3 con las variables restantes. En cada paso, se evalúa añadir una variable al modelo actual.
5. Detenerse cuando ninguna variable restante tenga un p-value por debajo del umbral de entrada.

Ventajas:

- Computacionalmente eficiente, especialmente con un número muy grande de variables predictoras (p >> n).

Desventajas:

- No garantiza encontrar el mejor modelo posible. Una variable añadida temprano puede volverse redundante después de añadir otras. Este método no remueve variables del modelo.

2. Selección hacia atrás (Backward Elimination)

Comienza con el modelo completo que incluye todas las variables disponibles. En cada paso, elimina la variable menos significativa (la con el p-value más alto), hasta que todas las variables restantes en el modelo sean estadísticamente significativas.

Proceso:

1. Empezar con el Modelo Completo que incluye todas las variables.
2. Ajustar el modelo y calcular los p-values para todos los coeficientes.
3. Eliminar la variable con el p-value más grande (si es mayor que un p-value predefinido, por ejemplo, 0.10).
4. Ajustar el nuevo modelo sin esa variable y repetir el paso 2.
5. Detenerse cuando todos los p-values de las variables en el modelo sean menores que el p-value predefinido.

Ventajas:

- Suele funcionar mejor que *Forward* cuando el conjunto de variables inicial tiene muchas correlaciones.
- Considera el efecto conjunto de las variables desde el principio.

Desventajas:

- Requiere ajustar el modelo completo al inicio, lo que puede ser imposible si el número de variables (p) es mayor que el número de observaciones (n).
- Computacionalmente más intensivo que *Forward* para conjuntos de datos muy grandes con muchas variables.
- Al igual que *Forward*, no garantiza el mejor modelo.

3. Selección por Pasos (Stepwise)

Es una combinación híbrida de *Forward* y *Backward*. Permite añadir variables en cada paso (como *Forward*), pero también revisa y elimina variables que pueden haber dejado de ser significativas después de añadir nuevas ones.

Proceso:

1. Similar a *Forward Selection*, se comienza con un modelo vacío y se añade la variable más significativa.
2. Después de añadir una nueva variable, el método verifica todas las variables incluidas previamente en el modelo.
3. Elimina cualquier variable que, debido a la nueva adición, ya no cumpla con el criterio de significancia (p-value de salida).
4. Solo después de este paso de "purgar" variables, procede a añadir la siguiente variable más significativa.
5. El proceso se repite hasta que ningún nuevo ingreso o salida de variables cumpla los criterios.

Ventajas:

- Más robusto que *Forward* o *Backward* por sí solos. Mitiga el principal defecto de *Forward* al permitir eliminar variables que se volvieron redundantes.
- Es el método más popular en la práctica por su equilibrio entre eficiencia y efectividad.

Desventajas:

- Aún así no garantiza encontrar el modelo óptimo global.
- Más complejo computacionalmente que *Forward*, pero generalmente manejable.

4. Mejor Subconjunto (Best Subset Selection)

Este método evalúa exhaustivamente todos los modelos posibles que se pueden construir con las p variables predictoras. Para cada k (de 0 a p), encuentra el mejor modelo que contiene exactamente k variables.

Proceso:

1. Ajusta el modelo nulo.
2. Para k = 1, 2, ..., p:
    - Ajusta todos los modelos posibles que contengan exactamente k variables.
    - Selecciona el mejor de estos modelos (usando por ejemplo R2_adj).
3. Elige el mejor modelo general entre M0, M1, ..., Mp, comparando los criterios de los mejores modelos de cada tamaño.

Ventajas:

- Garantiza encontrar el mejor modelo para cada tamaño de modelo (número de variables k).

Desventajas:

- Computacionalmente costoso. Para p variables, hay 2**p modelos posibles.
- Mayor riesgo de sobreajuste debido a la búsqueda exhaustiva en un espacio de modelos tan grande.

Explica cómo se utilizan para elegir el modelo final.

Se recomienda utilizar *Stepwise* (o *Forward/Backward* si hay limitaciones computacionales) para encontrar unos pocos modelos candidatos prometedores.

La elección final se hace usando un criterio de evaluación como el R2_ajustado que considera el número de variables. Se busca maximizar este criterio.

Normalmente se utiliza la técnica de validación cruzada (Cross-Validation). Se divide la muestra en una parte de entrenamiento y otra de validación. El mejor modelo es el que tiene el menor error de predicción en los conjuntos de validación. Esto evalúa directamente la capacidad de generalización del modelo a datos nuevos.

El objetivo es elegir el modelo que tenga el mejor equilibrio entre poder predictivo (bajo error de validación) y simplicidad.