La selección de características, también conocida como selección de variables, selección de atributos o selección de subconjuntos de variables, es el proceso de seleccionar un subconjunto de características relevantes (variables, predictores) para su uso en la construcción del modelo (regresión lineal).

¿Por qué? La premisa central cuando se utiliza una técnica de selección de características es que los datos contienen algunas características que son redundantes o irrelevantes.

• simplificación de modelos para facilitar su interpretación por parte de investigadores / usuarios,

• tiempos de entrenamiento más cortos,

• para evitar la maldición de la dimensionalidad,

• mejorar la compatibilidad de los datos con una clase de modelo de aprendizaje,

• codificar simetrías inherentes presentes en el espacio de entrada.

La elección de la métrica de evaluación influye mucho en el algoritmo, y son estas métricas de evaluación las que distinguen entre las tres categorías principales de algoritmos de selección de características:

Los métodos de envoltura utilizan un modelo predictivo para puntuar subconjuntos de características. Cada nuevo subconjunto se utiliza para entrenar un modelo, que se prueba en un conjunto de reserva

Los métodos de filtro utilizan una medida proxy en lugar de la tasa de error para puntuar un subconjunto de características

Los métodos integrados son un grupo general de técnicas que realizan la selección de características como parte del proceso de construcción del modelo.

Dentro de estas tres categorías existen múltiples y diferentes algoritmos.

La elección de los criterios de optimización es difícil ya que hay múltiples objetivos en una tarea de selección de características. Muchos criterios comunes incorporan una medida de precisión, penalizada por el número de características seleccionadas. Los ejemplos incluyen el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC).

Los algoritmos de selección Tambien se clasifican en:

* forward
* backward
* stepwise

En general, si se incluyen cada vez más variables en un modelo de regresión, el ajuste a los datos mejora, aumenta la cantidad de par ́ametros a estimar pero disminuye su precisi ́on individual (mayor varianza) y por tanto la de la funci ́on de regresi ́on estimada, se produce un sobreajuste.

Por otra parte, algunas variables predictoras pueden  
perjudicar la confiabilidad del modelo, especialmente si est ́an correlacionadas con otras. De  
esta manera, el objetivo de los m ́etodos de selecci ́on de variables es buscar un modelo que  
se ajuste bien a los datos y que a la vez sea posible buscar un equilibrio entre bondad de  
ajuste y sencillez

Hay dos razones  
por las que el m ́etodo de m ́ınimos cuadrados podr ́ıa no ser adecuado para estimar modelos  
con variables no relevantes

* Precisi ́on de la predicci ́on: Las estimaciones de los par ́ametros por m ́ınimos cua-  
  drados tienen bajo sesgo pero gran varianza. La precisi ́on de la predicci ́on a veces se  
  puede mejorar mediante la reducci ́on o ajuste a cero de algunos coeficientes.
* Interpretaci ́on: Con un gran n ́umero de predictores a menudo nos gustar ́ıa deter-  
  minar un subconjunto m ́as peque ̃no que exiba los efectos m ́as Fuertes

**Problema inicial**: tienes un conjunto de *n* variables independientes que podrias usar para explicar y predecir una variable dependiente con un modelo de regresión. ¿que conjunto de variables usar?​

Con el método forward parte de un modelo sencillo y se van agregando variables basado en algún criterio, hasta que ninguna mejora el modelo {por que una ves que la agregas no la puedes quitar}. Es decir, en cada etapa se introduce la variable **más** **significativa** hasta una cierta **regla de paro.**​

Es un algoritmo de búsqueda por la cual se sigue una heurística consistente en elegir la opción óptima en cada paso local con la esperanza de llegar a una solución óptima, {básicamente busca un máximo o un mínimo}. {Conocido como un algoritmo voraz o greedy.}

Oah pues ya tienes dos parametros que considerar:

1-. Determiner si es necesaria.

2.- determiner si es suficiente.

Tienes los mismos supuestos que para una regression lineal, multiple o logística.

El algoritmo no especifica medir normalidad ni heterocedasticidad pero supongo que puedes hacerlo junto al revisar el criterio de bayes o Akaike.

Cuando el ajuste de modelos es posible aumentar la probabilidad mediante la adición de parámetros, pero si lo hace puede resultar en sobreajuste. Tanto el BIC y AIC resuelven este problema mediante la introducción de un término de penalización para el número de parámetros en el modelo, el término de penalización es mayor en el BIC que en el AIC.

**Bayes y Akaike son una penalidad a cada variable agregada**

**Bayes: nombrado y derivado desde una probabilidad bayesiana**

**Akaike: nombrado por Akaike y derivado desde una pespectiva de probabilidad frequeantista**



BIC = -2\*(max loglikelihood) + d\*log(n)

n o N es tamaño de la nuestra de entrenamiento.

K o d, es el total de parametros. K = número de variables dependientes + 1. (o intercepto y error)

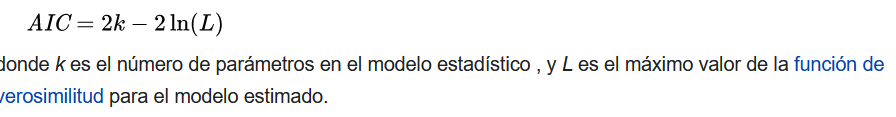
L sombrero o max loglikelihood es el valor log-likelihood de ols summary

Log-Likelihood: -50932

(-2\*-50932) + (7\*ln(6772)

-2\*-50932+7\*np.log(6772)

Akaike es



# Problemas:

1.- estamos usando pruebas estadísticas estándar que asumen una sola prueba de un modelo preespecificado y no son apropiadas cuando se usa una secuencia de pasos para elegir las variables explicativas.

2.- Los errores estándar de las estimaciones de los coeficientes se subestiman, lo que hace que los intervalos de confianza sean demasiado estrechos, las estadísticas t demasiado altas y los valores p demasiado bajos, lo que conduce a un sobreajuste.

3.- are not invariant to inconsequential linear transformations. Como en la ecuación del gasto del consumidor de Milton fierdman

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6

4-. No considera todas las posibles combinaciones

* Proporcionará una ventaja computacional sobre los métodos que consideran todas estas combinaciones. Inicialmente era algo Bueno por la potencia computacional de aquella epoca. Tampoco quieres la combinación de 100 variables.
* no está garantizado seleccionar la mejor combinación posible de variables.

5.- tus valores de coeficientes, intervalos de confianza, valores p y r squares

The significance values in your output are based on fitting a single model. Therefore, the significance values are generally invalid when a stepwise method (stepwise, forward, or backward) is used.



6-. No considera la relación causal entre variables.

Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing.

*Regression modeling strategies – Frank Harrell*

# Ventajas

1.-es facil de aplicar

2.- mejora el modelo generalmente

3.- es un metodo simple de interpreter

4-. Es objetivo y reproducible

Referencias

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6>

<https://quantifyinghealth.com/stepwise-selection/>

https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=regression-linear-variable-selection-methods