REGRESIÓN LINEAL MÚLTIPLE

SELECCIÓN DE VARIABLES Y CONSTRUCCION DEL MODELO

METODOS COMPUTACIONALES.

TODAS LAS REGRESIONES POSIBLES

Procedimiento

Ajuste todas las ecuaciones de regresión, que tengan un regresor candidato, dos regresores candidato, etc. Esas ecuaciones se evalúan de acuerdo con algún criterio adecuado y se selecciona el "mejor" modelo de regresión.

Si se supone que el término de ordenada al origen /30 se induye en todas las ecuaciones, entonces, si hay K regresores candidato, hay 2^K ecuaciones en total por estimar y examinar.

Se ve que la cantidad de ecuaciones por examinar aumenta con rapidez a medida que aumenta la cantidad de regresores candidato. Resumen para analizar los criterios.

Cantidad de regresores en el modelo	p	Regresores en el modelo	$SS_{Res}(p)$	R_p^2	$R_{Aj,p}^2$	$MS_{Res}(p)$	C_p

INTRODUCCIÓN ADELANTE (FORWARD)

Este método comienza con la suposición que no hay variables regresoras en el modelo, solamente el intercepto y se van incluyendo los regresores uno a uno hasta obtener la ecuación definitiva.

El procedimiento puede resumirse así:

La primera variable a entrar en el modelo es aquella que tiene la mayor correlación con la variable respuesta.

Suponga que este regresor es X_1 , y calculamos la regresión simple $y = f(X_1)$, se espera que esta regresión produzca el mayor valor de la estadística parcial al docimar la significancia de la regresión. Este regresor entrará al modelo si la estadística excede el valor de F preseleccionado, llamado F_{IN} , valor de entrada.

Calculamos luego el coeficiente de correlación parcial entre las variables restantes (**X_2, …, X_K) y la variable Y**,

El segundo regresor a entrar en el modelo será aquel que tiene la mayor correlación con Y después de ajustar y por el efecto del regresor que se introdujo **X_1. Dichas correlaciones son las correlaciones** parciales

Supongamos que el regresor X_2 es el regresor con la mayor correlación parcial con Y, esto implica que la estadística parcial F con mayor valor es:

$$F=\frac{SS_R(X_2/X_1)}{MS_{RES}(X_1,X_2)}$$

Si este valor F excede el valor de F_{IN} entonces X_2 es incluida en el modelo.

En general, a cada paso el regresor que tiene la mayor correlación parcial con Y, es incorporado en el modelo si su estadístico F parcial excede el valor de F_{IN} (valor de entrada) . El procedimiento termina cuando la estadística F parcial en un paso particular no exceda F_{IN} o cuando el último candidato regresor entra al modelo.

b. Eliminación hacia atrás (BACKWARD).

En la eliminación hacia atrás se trata de determinar un buen modelo trabajando en dirección contraria.

Esto es:

Se comienza con un modelo que incluya todos los K regresores candidatos. $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$

Luego se pasa a calcular la estadística parcial F para cada regresor, como si fuera la última variable que entro al modelo.

La mínima de las estadísticas parciales F se compara con un valor preseleccionado denominado $Fout$, entonces:

Si la estadística parcial F es menor que el valor preseleccionado $Fout$ se pasa a eliminar este regresor del modelo y se ajusta un modelo de regresión con K – 1 regresores.

Se vuelven a calcular las estadísticas parciales F para el nuevo modelo, y se repite el procedimiento ya mencionado.

El algoritmo de eliminación hacia atrás termina cuando el valor mínimo de F parcial no es menor que $Fout$, el valor preseleccionado de corte.

c. Regresión por segmentos (paso a paso ó STEPWISE).

La regresión por segmentos es una modificación de la selección hacia delante, en el cual a cada paso se reevalúan todos los regresores que habían entrado antes al modelo, mediante sus estadísticas parciales F.

Un regresor agregado en una etapa anterior puede volverse redundante, debido a las relaciones entre él y los regresores que ya estén en la ecuación.

Si la estadística parcial F de una variable es menor que F_{OUT}, esa variable se elimina del modelo.

En este método se requieren dos valores de corte, F_{IN} y F_{OUT}, algunos analistas prefieren definir F_{IN} = F_{OUT} , aunque eso no es necesario, con frecuencia se opta por F_{IN} > F_{OUT}, con lo que se hace algo más difícil agregar un regresor que eliminar.

El método termina cuando ya no hay variables candidatas a ser incluidas o a ser eliminadas.