Esquema de Selección de variables

- **Proceso de selección a pasos**

Backward. Eliminación hacia atrás

El proceso comienza con la inclusión de todas las variables explicativas en el modelo.

Se eliminan variables una a la vez. Se van eliminando las variables con menor poder predictivo. Las variables se eliminan del modelo siempre y cuando cumplan los criterios de salida.

La primera variable explicativa seleccionada para ser eliminada es la que presenta menor poder predictivo, es decir es la que tiene menor correlación con la variable de respuesta, dado que las demás variables explicativas están en el modelo.

El criterio de salida se fija determinando un nivel de significación o el valor del estadístico de prueba asociado a la significación de la prueba que se realiza en cada paso. El estadístico de prueba es el F[1]

El proceso continua hasta que en determinado momento la variable candidata a salir no cumple el criterio de salida, es decir, el valor del estadístico calculado para la misma excede el valor del F crítico asociado al nivel de significación predeterminado.

Ejemplo.

El modelo es $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$

La primera variable candidata a salir es aquella que aporta menos a la SCE del modelo, la que menos aporta al R^2 . Si cumple con el criterio de salida, la variable se va. Supongamos que la variable con menor poder predictivo es x_4 .

Si $F = \dfrac{SCE(x_4 / x_1, x_2, x_3)/1}{SCR/(n-k-1)} < F_{critico} = F_{1,n-k-1}$ no se rechaza $H_0)\beta_4 = 0$ y la variable x_4 es removida del modelo.

[1] El estadístico de prueba es un **F** parcial. Se demuestra que el ***F*** parcial es igual al estadístico de prueba ***t*** al cuadrado. En los software estadístico se utilizara indistintamente el ***F*** o el ***t***. El SPSS fija valores del F o niveles de significación. En el output sin embargo se puede ver el valor del estadístico t.

El modelo es ahora: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

La segunda variable candidata a salir es aquella que aporta menos a la SCE del modelo que ya no cuenta con x_4 , es la que menos aporta al R^2 . Si cumple con el criterio de salida, la variable se va. Supongamos que la variable con menor poder predictivo es x_2 .

Si $F = \frac{SCE(x_2 / x_1, x_3)/1}{SCR/(n-k-1)} < F_{critico} = F_{1,n-k-1}$ no se rechaza $H_0) \beta_2 = 0$ y la variable x_2 es removida.

Ahora, el modelo es $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$

La variable candidata a salir en esta instancia es x_3 pero $F = \frac{SCE(x_3 / x_1)/1}{SCR/(n-k-1)} > F_{critico} = F_{1,n-k-1}$ entonces se rechaza $H_0) \beta_3 = 0$, por lo que la variable no sale del modelo. El proceso se detiene aquí.

El modelo final es $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$

NOTA:

EL procedimiento instrumentado en el R NO HACE PRUEBA DE HIPOTESIS, utiliza estadístico AIC. ANALIZA LOS MODELOS CON MINIMO AIC. La variable candidata a salir es aquella que al ser eliminada genera un modelo con un mínimo AIC, el procedimiento se detiene cuando el modelo con mínimo AIC es aquel que se produce sin sacar variables.

***Forward*.** Selección progresiva de variables.

El proceso comienza con ninguna variable explicativa en el modelo.

Se agregan variables una a la vez. Se van incorporando las variables con mayor poder predictivo. Las variables ingresan al modelo siempre y cuando cumplan los criterios de entrada.

El criterio de entrada se fija determinando un nivel de significación o el valor del estadístico de prueba asociado a la prueba de significación que se realiza en cada paso. El estadístico de prueba es el F[2]

La primera variable explicativa seleccionada para entrar en el modelo es la que presenta mayor correlación con la variable de respuesta.

La segunda variable que se incluye en el modelo es aquella que, combinada con la primera - la que ya está en el modelo -, proporciona el mayor R^2 . Es la variable con mayor poder predictivo en un modelo que ya cuenta con variable explicativa aquella que ingresó en el primer paso. Es la variable que tiene mayor correlación parcial con la variable de respuesta, dado que está en el modelo la primer variable ingresada.

El proceso continua hasta que en determinado momento la variable candidata a entrar no cumple el criterio de entrada. El valor del estadístico calculado para la misma no excede el valor del Fcritico asociado al nivel de significación predeterminado.

Ejemplo.

Se quiere explicar $Y,$ y *las variables explicativas candidatas* son x_1, x_2, x_3, x_4

En el primer momento el modelo no contiene ninguna variable explicativa. La primera variable candidata a entrar es aquella con mayor poder predictivo, la que presenta mayor correlación con la variable a explicar y cumple el criterio de entrada.

Supongamos que la variable con mayor poder predictivo es x_3. Si $F = \frac{SCE(x_3)/1}{SCR/(n-k-1)} > F_{critico} = F_{1,n-k-1}$ entonces se rechaza $H_0)\beta_3 = 0$ por lo que la variable entra en el modelo.

El modelo es ahora $Y_i = \beta_0 + \beta_3 x_{i3} + \varepsilon_i$

La variable candidata a entrar en esta instancia es x_1, es aquella con mayor correlación parcial con la variable de respuesta en un modelo que ya contiene a x_3.

[2] El estadístico de prueba es un **F** parcial. Se demuestra que el ***F*** parcial es igual al estadístico de prueba ***t*** al cuadrado. En los software estadístico se utilizara indistintamente el ***F*** o el ***t***. El SPSS fija valores del F o niveles de significación. En el output sin embargo se puede ver el valor del estadístico t.

Si $F = \frac{SCE(x_1 / x_3)/1}{SCR/(n-k-1)} > F_{critico} = F_{1,n-k-1}$ se rechaza $H_0)\beta_1 = 0$, por lo que la variable entra al modelo.

El modelo ahora es $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$

La variable candidata a entrar en esta instancia es aquella con mayor correlación parcial con la variable de respuesta en un modelo que ya tiene a, x_1 y x_3. Si $F = \frac{SCE(x_2 / x_3, x_1)/1}{SCR/(n-k-1)} < F_{critico} = F_{1,n-k-1}$ entonces no se rechaza $H_0)\beta_2 = 0$ por lo que la variable no entra al modelo.

El proceso se detiene y el modelo final es: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$

NOTA:

EL procedimiento instrumentado en el R NO HACE PRUEBA DE HIPOTESIS, utiliza estadístico AIC. ANALIZA LOS MODELOS CON MINIMO AIC. La variable candidata a entrar es aquella que al ser incluida genera un modelo con un mínimo AIC, el procedimiento se detiene cuando el modelo con mínimo AIC es aquel que se produce sin incorporar variables.

Stepwise

Es una combinación de los dos procesos anteriores. El proceso comienza como un proceso *Forward,* pero siempre analiza si al entrar variables las que ya estaban en el modelo pierden significación acorde al criterio de salida fijado, de ser así la variable es eliminada.

El proceso termina cuando ninguna variable cumple criterio de entrada o de salida.

- **Selección de subconjunto de variables**

Implica analizar los modelos con todas las combinaciones posibles de variables explicativas y analizar indicadores que permitan elegir el mejor modelo, por ejemplo se puede analizar el R_a^2 ajustado de los distintos modelos y analizar los incrementos

relativos del mismo en modelos con distinta cantidad de variables. Para modelos con igual cantidad de variables se puede analizar el R^2 .

Ejemplo

Se quiere explicar $Y,$ y *las variables explicativas candidatas* son x_1, x_2, x_3, x_4

Se analizan los modelos:

Modelos	AIC	R_{a^2}
Con una variable		
x_3		
x_1		
x_2		
x_4		
Con dos variables		
x_3, x_1		
x_3, x_2		
x_3, x_4		
x_1, x_2		
x_1, x_4		
x_2, x_4		
Con tres variables		
x_1, x_2, x_3		
x_1, x_2, x_4		
x_2, x_3, x_4		
x_1, x_3, x_4		
Con cuatro variables		
x_1, x_2, x_3, x_4		