CAPÍTULO 6
SELECCIÓN DE VARIABLES EN REGRESIÓN

Edgar Acuña Fernández

Departamento de Matemáticas
Universidad de Puerto Rico
Recinto Universitario de Mayagüez

Selección de variables

También llamado selección de un subconjunto de predictoras es un procedimiento estadístico que es importante por diversas razones, entre estas están:

- No todas las variables predictoras tienen igual importancia (variables irrelevantes).
- Algunas variables pueden perjudicar la confiabilidad del modelo (variables redundantes).
- Computacionalmente es más fácil trabajar con un conjunto de variables predictoras pequeño.
- Es más económico recolectar información para un modelo con pocas variables.
- Si se reduce el número de variables entonces el modelo se hace más **parsimonioso**. Un modelo es **parsimonioso** si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles.

Metodos "Stepwise"

La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una sola variable predictora en cada paso de acuerdo a ciertos criterios.

El proceso secuencial termina cuando una regla de parada se satisface.

Tres algoritmos para seleccionar variables son:

"Eliminacion hacia atras"

"Selección hacia adelante"

"Selección paso a paso"

“Backward Elimination” (Eliminación hacia atrás).

- Se comienza con el modelo completo y en cada paso se va eliminando una variable.

- Si resultara que todas las variables predictoras son importantes, es decir, tienen *“p-value” pequeños* para la prueba t, entonces no se hace nada Y se concluye que el **mejor modelo es el que tiene todas las variables** predictoras disponibles.

“Backward Elimination” (Eliminación hacia atrás).

- En cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de estos requisitos equivalentes entre sí:
- Aquella variable que tiene el estadístico de *t (*en valor absoluto) más pequeño entre las variables incluidas aún en el modelo (o F parcial más pequeño).
- Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo.
- Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables que quedarían en el modelo.

“Forward Selection” (Selección hacia adelante).

- Se empieza con la regresión lineal simple que considera como variable predictora a aquella que esta más altamente correlacionada con la variable de respuesta.

- Si esta primera variable no es significativa entonces se considera el modelo y se para el proceso.

- Si hay variables que son significativas se añade al modelo la variable que reune cualquiera de estos requisitos equivalentes entre sí:

Requisitos equivalentes para que una variable sea considerado en el modelo

- Aquella variable que tiene el estadístico de *t* (en valor absoluto) más grande entre las variables no incluidas aún en el modelo. Es decir, la variable con el *F-parcial* más grande.

- Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo. Es decir, aquella variable que produce la mayor reducción en la suma de cuadrados del error.

- Aquella variable que tiene la correlación parcial más alta (en valor absoluto) con la variable de respuesta, tomando en cuenta las variables ya incluidas en el modelo.

Criterios de parada para el metodo forward

- Se llega a un modelo con un número prefijado p* de variables predictoras.
- El valor de la prueba de *F parcial* para cada una de las variables no incluidas aun en el modelo es menor que un número prefijado *F-in* (por lo general este valor es 4).

- Cuando el valor absoluto del estadistico de t es menor que la raíz cuadrada de *F-in* (por lo general, $|t|<2$).

- Si se prefija de antemano un nivel de significación dado α^* (digamos del 15%) para la prueba de *t* o de *F parcial* en cada paso, en este caso se termina el proceso cuando todos los *p-values* de la prueba *t* de las variables no incluidas aún son mayores que α^*.

“Stepwise Selection” (Selección Paso a Paso)

- Efroymson (1960), subsana el problema de anidamiento de los dos métodos anteriores.
- Se puede considerar como una modificación del método “Forward”. Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada. Aqui se usan ***F-out*** y ***F-in*** con $F\text{-}in \leq F\text{-}out$.
- El proceso termina cuando ninguna de las variables, que no han entrado aún, tienen importancia suficiente como para entrar al modelo.

Método de los mejores subconjuntos

Si el problema tiene un número pequeño de variables predictoras (no más de 8), se podrían calcular uno o dos criterios de selección para las 2^k regresiones posibles, luego se escogerían unos cuantos de estos modelos para un análisis más detallado y decidir sobre el mejor modelo.

Pero si el número de variables predictoras es grande surgen nuevos métodos que permiten acelerar la busqueda de los mejores subconjuntos de variables como :

- "Branch and Bound" (Ramificación y acotamiento)
- "Leaps and Bound" (Brincando y acotando) ,éste, es adoptado por la mayoría de los programas estadísticos.

Criterios para elegir el mejor modelo:

- El coeficiente de Determinación R^2
- El R^2 ajustado
- La varianza estimada del error (s^2).
- El C_p de Mallows.
- PRESS (Suma de cuadrados de Predicción)- Predicted R^2
- Validación Cruzada (CV)
- AIC
- BIC
- Validación Cruzada Generalizada (CGV)
- Otros Criterios

El coeficiente de Determinación R^2

- Se elige aquél modelo que tenga un R^2 bastante alto con el menor número de variables predictoras posibles.
- Se elige un modelo con k variables si al incluir una variable adicional el R^2 no se incrementa sustancialmente (5%).

Algunos problemas de este criterio

- Efecto de datos anormales.
- Un modelo con pocas variables siempre tendrá un R^2 menor o igual que un modelo que incluye un mayor número de variables.

El R^2 ajustado

Para subsanar la tendencia del R^2 se ha definido un ***R^2-ajustado*** de la siguiente manera:

$$R^2_{ajus} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2)$$

Donde, *p* es el número de parámetros en el modelo.

El modelo que se busca es aquel que tiene un ***R^2-ajustado*** alto con pocas variables.

Nota

El R^2 ajustado podría disminuir al incluirse una variable adicional en el modelo.

La varianza estimada del error (s^2).

El mejor modelo será aquel que tenga la varianza estimada (o desviación estándar) del error más pequeña.

El C_p de Mallows.

- Mallows (1973), el mejor modelo es áquel que no tiene ni mucha falta de ajuste (“underfitting”) ni mucho sobreajuste (“overfitting”) al ajustar los datos.
- *Falta de ajuste*, se da cuando el estimado del valor predicho de la variable de respuesta tiene **mucho sesgo y poca varianza**,
- *Sobreajuste*, se da cuando **la varianza** del estimado del valor predicho es bastante **alta**, pero el **sesgo es bajo**.

El C_p de Mallows.

- El **cuadrado medio del error** para un valor predicho sumando sobre todas las observaciones está dado por

$$\sum_{i=1}^{n} \frac{MSE(\hat{y}(x_i))}{\sigma^2} = \sum_{i=1}^{n} \frac{E[\hat{y}(x_i) - y(x_i)]^2}{\sigma^2} = \sum_{i=1}^{n} \frac{Var(\hat{y}(x_i)) + Sesgo^2(\hat{y}(x_i))}{\sigma^2}$$

Donde,

$$\sum_{i=1}^{n} \frac{Var(\hat{y}(x_i))}{\sigma^2} = p \quad \text{y}$$

$$\sum_{i=1}^{n} \frac{Sesgo^2(\hat{y}(x_i))}{\sigma^2} = (n-p)\left(\frac{E(s_p^2) - \sigma^2}{\sigma^2}\right)$$

Criterio de Mallows

- Se trata de encontrar un modelo donde *el sesgo* y *la varianza* sean moderados.
- El estadístico de *Mallows* está dado por:

$$C_p = p + (n-p)\frac{s_p^2}{s^2} - (n-p) = \frac{SSE_p}{s^2} - (n-2p)$$

SSE_p, es la suma de cuadrados del error del modelo que contiene ***p*** parámetros, incluyendo el intercepto, y

s^2, es la **varianza estimada** con el modelo completo.

un modelo con *p* parámetros es adecuado si

$E(SSE_p)=(n-p)\sigma^2$, luego, $E[SSE_p/s^2]$ es aproximadamente $(n-p)$. En consecuencia **$E(C_p)=p$**.

Para elegir el valor de *p* se acostumbra a plotear C_p versus p. Los valores p más adecuados serán aquellos cercanos a la intersección de la gráfica con la línea $C_p=p$

PRESS (Suma de cuadrados de Predicción)

- Allen (1974) es una combinación de todas las regresiones posibles, análisis de residuales y “leave-one-out” (validación cruzada).
- Supongamos que hay *p* parámetros en el modelo y que tenemos *n* observaciones disponibles para estimar los parámetros.
- En cada paso se deja de lado la *i-ésima* observacion del conjunto de datos y se calculan todas las regresiones posibles.
- Se calcula la predicción y el residual correpondiente para la observación que no fue incluida, el cual es llamado el residual **PRESS.**

PRESS (Suma de cuadrados de Predicción)

- La relación entre el residual PRESS y el residual usual esta dado por:

$$e_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}}$$

- donde h_{ii} representan los elementos de la diagonal de la matríz H

- La medida PRESS para el modelo de regresión que contiene p parámetros se define por:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \quad \text{o equivalentemente} \quad PRESS = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1 - h_{ii}}\right)^2$$

El mejor modelo es aquel que tiene el valor de PRESS más bajo.

El Predicted R^2 se define por (1-PRESS/SST)*100%

Validación Cruzada (CV)

- Stone (1974) Se estima el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una *muestra de prueba* que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una *muestra de entrenamiento* que sirve para construir el modelo.
- Por lo general se usan 10 partes y eso es llamado una “10 fold cross-validation” , ó *n* partes y en ese caso es llamado el método “leave-one-out”(dejar uno afuera).

Cálculo del error por validación cruzada usando K-partes

- Esta dado por:

$$CV = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i}(y_j - \hat{y}_j^{(-i)})^2}{n}$$

- $\hat{y}_j^{(-i)}$ representa el valor predicho para la *j-ésima* observación de la parte Ni usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte.
- El mejor modelo es aquel que tiene el *error de validación cruzada promedio* más pequeño.

En el caso de "*leave-one-out*" el error de predicción promedio es PRESS/n.

Criterio de información de Akaike AIC

Akaike (1973) basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta Y usando el modelo reducido y bajo el modelo completo. Se define como:

AIC= -2*máximo de la log likelihhod +2p

Donde, *p* es el número de parametros del modelo.

En particular para el caso de regresión, asumiendo que la varianza de las y's es estimada por SSE/n, la fórmula anterior se reduce a:

$$AIC = n\log[SSEp/n] + 2p$$

Un buen modelo es aquel con **bajo AIC**.

BIC

Schwarz (1978), y está basado en argumentos bayesianos.

Se define por:

$$BIC=nlog[SSEp/n]+2plog(n)$$

Observación

Los criterios AIC y Cp de Mallows tienden a dar modelos óptimos más grandes que el criterio BIC.

Validación Cruzada Generalizada (CGV)

Golub,Heath and Whaba (1979) Dado que el cálculo de validación cruzada "leave-one out" es computacionalmente pesado, el GCV es una aproximación al "leave-one-out", que puede ser calculado más rápidamente.

Se define por

$$GCV = \frac{SSE_p}{[n-1-tr(H_p)]^2}$$

donde

H_p, es la matriz *HAT* para el modelo que incluye *p* variables.

El modelo óptimo será aquel que incluye las *p* variables predictoras que hacen que ***GCV sea mínimo***.

Otros Criterios

Otros criterios para la selección de variables en regresión son:

MDL: Longitud de Descripción Mínima (Rissanen, 1978).

RIC: Criterio de Inflación del Riesgo (Foster y George, 1994)

CIC: Criterio de Inflación del Covarianza (Tibshirani and Knigth, 1999)

Bootstrapping (Efron, 1983)

El pequeño Boostrapping (Breiman, 1992)

La Garrote (Breiman, 1995)

El Lasso (Tibshirani, 1996)

Recomendaciones para elegir el mejor modelo

En cualquier problema las variables predictoras pueden ser clasificadas en 3 grupos:

a) Las que *son importantes*.

b) Las que uno *no está seguro de su importancia*.

c) Las que *no son relevantes* para explicar el comportamiento de la variable de respuesta.

Lo que se recomienda es eliminar las variables tipo c) eligiendo un buen subconjunto de variables predictoras usando para ello los criterios Cp, AIC o BIC y luego aplicar "stepwise" para descartar las variables tipo b) y quedarnos con las variables tipo a) que son las que son interesantes.

Otros métodos de Selección de variables

Métodos Bayesianos

Mitchel y Beauchamp (JASA, 1988)

Supongamos que ya se tiene un conjunto de buenos modelos. La idea se basa en asignar probabilidades a priori a los coeficientes de cada uno de estos modelos que incluyen solo un subconjunto de predictoras e igualmente se asignan probabilidades a priori a cada uno de los modelos.

Finalmente se elige como mejor modelo aquel que tiene la probabilidad posterior más alta con respecto a la variable de respuesta.

Otros métodos de Selección de variables

Algoritmo Genéticos

En este caso el problema de selección de variables es considerado como un problema de optimización con respecto al número de variables predictoras que deben incluirse en el modelo.

Luego el problema de optimización es resuelto usando algoritmos Genéticos.