UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE ESTADÍSTICA
PROGRAMA ACADÉMICO DE ESTADÍSTICA

Javier Olaya

TAREA 6

Estadística Aplicada II - 2018

Selección de Variables

En los problemas de regresión se tiene generalmente un grupo de variables candidatas para predecir o describir el comportamiento de la variable de respuesta. Sin embargo, no todas las variables contribuyen de igual manera a explicar el comportamiento de la variable de respuesta, por lo que mantener variables innecesarias en el análisis solamente hace el modelo más complejo, no necesariamente mejor. En este caso se acostumbra a evaluar estadísticamente qué variables deben incluirse en el modelo. De esta manera se alcanzarán estimaciones y predicciones más precisas, por el mecanismo de ajustar un modelo basado solamente en los términos relevantes

En la literatura se han propuesto varios criterios para encontrar el mejor subconjunto de variables que deben ingresarse en el modelo, basados esencialmente en el principio de parsimonia, que sugiere seleccionar el modelo con la menor Suma de Cuadrados de los Errores, utilizando la menor cantidad de parámetros posible. Algunos de estos criterios son el estadístico F, el Cuadrado Medio del Error, el C_p de Mallows, el R^2 y el R^2 ajustado (R^2_{adj}) y los criterios de información de Akaike (AIC) y Bayesiano (BIC), entre otros.

Al mismo tiempo, se han propuesto diversos algoritmos de selección que se han incorporado al software estadístico y están disponibles en todos los paquetes estadísticos que se ocupan del problema de la selección de variables. Los más comunes son el método *forward*, el método de eliminación *backward* y el método *stepwise*. Finalmente, en R se han incorporado dos soluciones interesantes: selección usando bosques aleatorios (RF: Random Forests) (paquete VSURF) y selección paso a paso con el criterio de información de Akakike (procedimiento StepAIC).

Esta tarea busca estudiar los métodos de selección de variables a través de su aplicación a un problema real.

A cada grupo se le asignará un criterio y un algoritmo clásico y uno moderno de selección, para elegir el mejor subconjunto de variables en el problema de la base de datos *cadata*.

Se sugiere que el informe escrito contenga los siguientes elementos:

- Describir detalladamente el criterio y los algoritmos de selección
- Presentar un análisis del modelo final
- Una enumeración sustentada de las ventajas y desventajas de los métodos y algoritmos utilizados
- Conclusiones