Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, mayo de 2018

Tarea 6: Selección de variables

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

En este trabajo se presentarán dos algoritmos de selección de variables que son el Backward y el StepAIC, donde el criterio de selección para el Backward será el R^2_{adj}, se tratará de encontrar el mejor modelo posible aplicando estos dos algoritmos para los 500 datos seleccionados de la base de datos 'Cadata', y se analizará este modelo final. Además, se presentaran una serie de ventajas y desventajas de cada uno de los métodos y se obtendrán conclusiones a partir de los resultados arrojados.

En la mayoría de los problemas de construcción de modelos el analista tiene un grupo de regresores candidatos y debe determinar el subconjunto real de regresores que debe usarse en el modelo. La definición de un subconjunto adecuado de regresores para el modelo es lo que se llama problema de selección de variables.

La construcción de un modelo de regresión que sólo incluya un subconjunto de los regresores disponibles implica dos objetivos contrapuestos: 1) Se desea que el modelo incluya tantos regresores como sea posible, para que el contenido de información en ellos pueda influir sobre el valor predicho de y. 2) Se desea que el modelo incluya los menos regresores que sea posible, porque la varianza de la predicción $\hat{y}$ aumenta a medida que aumenta la cantidad de regresores. También, mientras más regresores haya en un modelo, los costos de recolección de datos y los de mantenimiento de modelo serán mayores. Entonces, con la selección de variables se trata de encontrar el punto medio entre los dos objetivos anteriores, logrando hallar así la 'mejor' recta de regresión.

2. Descripción de criterios y algoritmos de selección

- $R^2_{Ajustado}$: Usando el método del $R^2_{Ajustado}$, se deben realizar todas las posibles combinaciones de modelos a partir de las variables que tenemos, y seleccionar el modelo que tenga mayor $R^2_{Ajustado}$. Este método surgió como una alternativa al método del coeficiente de determinación R^2, ya que el R^2 no tiene en cuenta la cantidad de parámetros del modelo, es decir, este siempre aumenta a medida que se ingresa una nueva variable al modelo.
  El $R^2_{Ajustado}$ se define como:

$$R^2_{Adj} = 1 - \frac{N-1}{N-k-1}(1-R^2)$$

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

Donde N es el tamaño de muestra y k es la cantidad de parámetros del modelo y $R^2 = \frac{SCR_D}{SCT_D} = \frac{\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$.

- Backward (selección hacia atrás):
- StepAIC (selección paso a paso con el criterio de información de Akakike (AIC)):

3. Análisis del modelo seleccionado

Los modelos seleccionados por el método del $R^2_{Ajustado}$ y por los algoritmos Backward y StepAIC fueron exactamente el mismo, es decir, los tres métodos nos arrojaron las mismas variables seleccionadas que fueron Ingreso mediano, Edad mediana, Total de habitaciones, Población y Hogares (solo eliminaron la variable Total de Dormitorios).
Ese modelo fue:

$$\hat{ValorMediano} = 52921.68807 + 24923.24378 IngresoMediano + 3484.16424 EdadMediana$$
$$+17.58804 TotalHabitaciones - 118.65234 Poblacion + 243.26500 Hogares$$

Para el respectivo análisis del modelo, tenemos las siguientes dos tablas:

	Estimación	Std. Error	Valor t	Pr(>|t|)
Intercepto	52921.688	17525.287	3.020	0.00266
Ingreso Mediano	24923.244	2547.044	9.745	<2e-16
Edad Mediana	3484.164	280.184	12.435	<2e-16
Total Habitaciones	17.588	6.632	2.652	0.00826
Población	-118.652	10.727	-11.061	<2e-16
Hogares	243.265	27.699	8.782	<2e-16

$\hat{\sigma}$	R^2	$R^2_{Ajustado}$	Estadístico F	P Valor
81990	0.5417	0.5371	116.8	<2.2e-16

Podemos ver en las dos tablas anteriores que todas las variables incluidas son significativas (el p valor de cada una de ellas es menor a 0.05) y por tanto, el modelo en general también es significativo. Además, tanto el R^2 como el $R^2_{Ajustado}$ son casi los mismos que los del modelo completo (0.5425 y 0.537 respectivamente), por lo tanto podríamos concluir que la inclusión de la variable Total de Dormitorios no daba un aporte adicional a la explicación del valor mediano de la vivienda. Adicionalmente, la desviación del modelo a los datos disminuyó un poco (de 82000 del modelo completo a 81990 en el modelo con selección de variables), por lo cuál se podría decir que la variable excluida en este problema en particular solo servía para aumentar los costos de recolección de datos y para aumentar la desviación del modelo y en general, las desviaciones de los valores ajustados.

4. Ventajas y desventajas de los algoritmos y métodos utilizados

5. Conclusiones

? ? ? ? ? ?