PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. I - N°2 Pgs. 105-110 Lima – Perú Dic. 1998

SELECCIÓN DEL CONJUNTO ÓPTIMO DE VARIABLES EXPLICATIVAS DE UNA VARIABLE RESPUESTA

Rosa María Inga Santivañez
Universidad Particular Cayetano Heredia
Universidad Nacional Mayor de San Marcos

RESUMEN

Cuando uno analiza datos categóricos es interesante estudiar el caso en que una de ellas es la variable respuesta de las otras variables explicativas y es interesante seleccionar el conjunto óptimo de variables explicativas.

A continuación se presenta un procedimiento basado en el criterio de Información de Akaike, mediante el cual se puede determinar el conjunto óptimo de variables explicativas de una variable respuesta en dos situaciones:

i) Cuando el número de variables explicativas es razonablemente manejable.

ii) Cuando el número de variables explicativas es demasiado grande.

NOTACIÓN

Los datos vienen dados en una tabla de contingencia multidimensional con k factores de clasificación I_1, I_2,..., I_k, donde I_1 es una variable respuesta e I_2, I_3,...,I_k son variables explicativas de la variable respuesta.

Donde:

I_j tiene C_{I_j} categorías.

$n(i_1, i_2,..., i_k)$: la frecuencia de la celda $(i_1, i_2,..., i_k)$

$$\sum_{i_1=1}^{C_{I_1}} \sum_{i_2=1}^{C_{I_2}} \sum_{i_3=1}^{C_{I_3}} n(i_1, i_2,..., i_k) = N$$

$p(i_1, i_2,..., i_k)$: probabilidad de ocurrencia de la celda $(i_1, i_2,..., i_k)$

$$\sum_{i_1=1}^{C_{I_1}} \sum_{i_2=1}^{C_{I_2}} \sum_{i_3=1}^{C_{I_3}} p(i_1, i_2, ..., i_k) = 1$$

PRIMER CASO: CUANDO EL NUMERO DE VARIABLES EXPLICATIVAS ES RAZONABLE

El siguiente modelo representa la asociación entre la variable I_1 y un subconjunto de variables explicativas E.

MODELO:
$$p(i_1, I) = \frac{p(i_1, E).p(I)}{p(E)} \qquad (1)$$

donde $I = \{I_2, ..., I_k\}$ con $p(I) = p(i_2, ..., i_k)$ y E es un subconjunto de I.

Las variables que aparecen en el denominador de los diferentes modelos, E, constituyen los candidatos para ser la combinación óptima de variables explicativas.

El estadístico AIC (Criterio de información de Akaike) reducido del modelo (1) será,

$$AIC^* = -\sum_{i_1,E} n(i_1, E).\log\left(\frac{n(i_1, E)}{n(E)}\right) + 2\left[\left(C_{I_1} C_E - 1\right) - \left(C_E - 1\right)\right] \qquad (2)$$

PROCEDIMIENTO

1) Calcular los AIC^* de todos los modelos de la forma(1).
2) Aplicar el criterio MAIC (Mínimo AIC) a cada grupo de modelos I_1 frente a (k-1) variables explicativas.

   I_1 frente a (k-2) variables explicativas.

   .

   .

   .

   I_1 frente a una variable explicativa.

   Así el investigador seleccionará de entre los modelos seleccionados de cada grupo, al mejor modelo siguiendo el Principio de Parsimonia, es decir se trataría de llegar a un

compromiso entre el modelo que más explique y el más sencillo. Luego el conjunto "E" del modelo seleccionado es el conjunto óptimo de variables explicativas.

SEGUNDO CASO: CUANDO EL NUMERO DE VARIABLES EXPLICATIVAS ES DEMASIADO GRANDE

Cuando el número de variables explicativas es demasiado grande el procedimiento para seleccionar el conjunto optimo de variables explicativas consta de dos etapas.

PRIMERA ETAPA: Preselección de variables explicativas.

1) Formulación de los modelos MODELO(I_1, I_j):

$$p(i_1, i_2, ..., i_k) = \frac{p(i_i, i_j).p(i_2, ..., i_k)}{p(i_j)} \quad (3)$$

para j = 1,..., k.

2) Se calcula los AIC* reducidos AIC , asociado a cada MODELO(I_1, I_J)

$$AIC^*(I_1, I_j) = -2\sum_{i_1, i_j} n(i_1, i_j).\log\left(\frac{N.n(i_1, i_j)}{n(i_1).n(i_j)}\right) + 2(C_{I_1} - 1)(C_{I_j} - 1) \quad (4)$$

para j = 2,..., k

3) Para efectuar la preselección se ordenan los AIC* de menor a mayor, los primeros son las variables explicativas más significativas, estas variables constituyen el conjunto de variables explicativas preseleccionadas.

SEGUNDA ETAPA: Tomando como base las variables preseleccionadas se procede

1) Formulación de los modelos

MODELO(I_1, F): $$p(I^*) = \frac{p(I_1, F).p(I_2, ..., I_k)}{p(F)} \quad (5)$$

Donde F es un conjunto de variables explicativas preseleccionadas y $I^*=\{I_1, I_2, ..., I_k\}$

2) Se calcula los AIC* del MODELO(I_1, F)

$$AIC^*(I_1,F) = -2\sum_{i_1,F} n(i_1,F).\log\left(\frac{N.n(i_1,F)}{n(i_1).n(F)}\right) + 2(C_{I_1}-1)(C_F-1) \qquad (6)$$

3) Se aplica el criterio MAIC a cada grupo de modelos.

   I_1 frente a todas las variables explicativas.

   .

   .

   .

   I_1 frente a una variable explicativa preseleccionada.

   Así el investigador seleccionará entre los modelos seleccionados de cada grupo, al mejor modelo siguiendo el Principio de Parsimonia.

APLICACIÓN:

Para ilustrar el método de selección de conjunto optimo de variables explicativas en el caso de que el número de variables explicativas es demasiado grande, se presenta el estudio de la fecundidad.

ESTUDIO DE LA FECUNDIDAD

Los datos proceden de la "Encuesta de Fecundidad 1985" del Instituto Nacional de Estadística de España.

VARIABLE RESPUESTA

I_1: Número de hijos nacidos vivos.

VARIABLES EXPLICATIVAS

I_2 : Estado civil.

I_3 : Historia de su actividad laboral.

I_4 : Tamaño de municipio de residencia.

I_5 : El número de hermanos nacidos vivos.

I_6 : Creencia y practica religiosa.

I_7 : Edad actual.

I_8 : Nivel de instrucción.

I_9 : Tipo de municipio en el que paso la primera infancia.

I_{10} : Relación con la actividad económica.

Aplicamos la Primera Etapa y se obtuvo lo siguiente:

PRESELECCIÓN

TABLA 1. Los AIC de los modelos en orden decreciente.

# MODELO	MODELO(I_1, I_j) I_1 : VARIABLE RESPUESTA I_J : VARIABLE EXPLICATIVA	AIC
1	MODELO(I_1, I_2)	-6219994*
6	MODELO(I_1, I_7)	-4601772*
9	MODELO(I_1, I_{10})	-3033884*
7	MODELO(I_1, I_8)	-1924524*
4	MODELO(I_1, I_5)	-414946
5	MODELO(I_1, I_6)	-318936
8	MODELO(I_1, I_9)	-246772
2	MODELO(I_1, I_3)	- 95338
3	MODELO(I_1, I_4)	-24382

Observamos que las variables más significativas son I_2, I_7, I_{10}, I_8 estas constituirían el conjunto de variables explicativas preseleccionadas y luego entre ellas se elegirá el conjunto óptimo.

Pero debido a la limitada información con que se cuenta, se opto por analizar solo los modelos que se pueden formular con la información de la "Encuesta de Fecundidad 1985".

Aplicamos la Segunda Etapa, así se obtuvo lo siguiente.

TABLA 2. Los AIC de los modelos

	# MODELO	MODELO(V.RES;{V.EXPLICS.}	AIC
MODELOS CON DOS VARIABLES EXPLICATIVAS	1	MODELO(I_1; I_2, I_3)	1.601671E+07
	2	MODELO(I_1; I_2, I_4)	1.604693E+07
	3	MODELO(I_1; I_2, I_5)	1.579953E+07
	4	MODELO(I_1; I_2, I_6)	1.584052E+07
	5	MODELO(I_1; I_2, I_7)	1.386274E+07*
	6	MODELO(I_1; I_2, I_8)	1.535024E+07
	7	MODELO(I_1; I_2, I_9)	1.600509E+07
	8	MODELO(I_1; I_2, I_{10})	1.563808E+07
MODELOS CON UNA VARIABLE EXPLICATIVA	9	MODELO(I_1; I_2)	1.607806E+07*
	10	MODELO(I_1; I_3)	2.220269E+07
	11	MODELO(I_1; I_4)	2.227367E+07
	12	MODELO(I_1; I_5)	2.188311E+07
	13	MODELO(I_1; I_6)	2.197912E+07
	14	MODELO(I_1; I_7)	1.769615E+07
	15	MODELO(I_1; I_8)	2.037356E+07
	16	MODELO(I_1; I_9)	2.20513E+07
	17	MODELO(I_1; I_{10})	1.926415E+07
	18	MODELO(I_1; { })	2.229806E+07

De donde siguiendo el Principio de Parsimonia el conjunto de variables explicativas seleccionadas son el estado civil y la edad actual.

BIBLIOGRAFÍA

1. Sakamoto, H.; Akaike, H. (1982). "Efficiente use of Akaike's information criterion for model selection in high dimentional contingenccy table analysis". Metron 40, 257 – 275.
2. Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. (1986). "Akaike information statistics". KTK Scientific Publishers / Tokyo.