UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS
Departamento de Estadística e Investigación Operativa

TESIS DOCTORAL

El criterio de información de Akaike en el análisis de datos categorizados

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Rosa María Inga Santiváñez

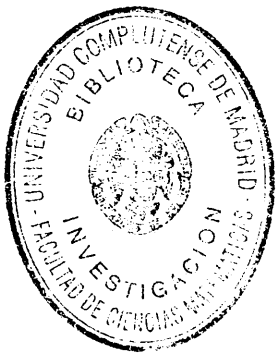
DIRECTORES:

Leandro Pardo Llorente
Domingo Morales González

Madrid, 2015

UNIVERSIDAD COMPLUTENSE DE MADRID
Facultad de Matemáticas
Departamento de Estadística e I.O.

EL CRITERIO DE INFORMACION DE AKAIKE EN EL ANALISIS DE DATOS CATEGORIZADOS



Rosa María Inga Santiváñez

Madrid, 1993

Colección Tesis Doctorales. N.º 181/93

Edita e imprime la Editorial de la Universidad
Complutense de Madrid. Servicio de Reprografía.
Escuela de Estomatología. Ciudad Universitaria.
Madrid, 1993.
Ricoh 3700
Depósito Legal: M-30752-1993

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE MATEMATICAS

Departamento de Estadística e I.O.

EL CRITERIO DE INFORMACION DE AKAIKE EN EL ANALISIS DE DATOS CATEGORIZADOS

ROSA MARIA INGA SANTIVAÑEZ

EL CRITERIO DE INFORMACION DE AKAIKE

EN EL ANALISIS DE DATOS CATEGORIZADOS

ROSA MARIA INGA SANTIVAÑEZ

Memoria para optar al grado de Doctor en Ciencias Matemáticas, realizada bajo la dirección de los doctores:

Dr. D. Leandro **PARDO LLORENTE**,

Dr. D. Domingo **MORALES GONZALES**.

Madrid, Mayo de 1992

A mis padres Nestor e Hilda.

INDICE

INTRODUCCION

Una gran cantidad de datos biólogicos, sociales, etc., suelen venir dados en forma de tablas de clasificación cruzada de frecuencias, comunmente conocidas como tablas de contingencia. Las observaciones de una muestra en tal circunstancia se clasifican de acuerdo a cada una de las variables categóricas o conjunto de categorías tales como sexo (hombre, mujer), edad (joven, mediana edad, viejo), o especies; es decir, los datos sor categorizados.

Cuando se observan simultáneamente muchas categorías (o factores) se dice que forman una tabla de contingencia multidimensional. Tales tablas presentan problemas para su análisis e interpretación, estos problemas han ocupado un lugar destacado dentro de la Estadistica desde la aparición del primer articulo sobre el test de tablas 2 x 2 x 2 presentado por Bartlett (1935).

Muchos han sido los estadisticos presentados para el estudio y análisis de ajuste de los posibles modelos que podrían seguir datos categorizados. Algunos de estos estadisticos son:

El estadistico Ji-Cuadrado de Pearson,

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

donde O_i es la frecuencia observada de la i-ésima celda, E_i es la frecuencia esperada de la i-ésima celda, i=1,...,K.

El estadistico G basado en la cantidad media de información de Kullback-Leibler (G = 2I, ver Kullback (1959)),

$$G = 2 \sum_{i=1}^{K} O_i \log \left(\frac{O_i}{E_i} \right)$$

El estadistico $\left(2\ n\ I^{\lambda}\right)$ dada por Cressie y Read (1984),

donde $\left(2\ n\ I^{\lambda}\right)$ toma la expresión

$$2\ n\ I^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{K} O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right] .$$

Para $\lambda = 1$ se obtiene el estadístico χ^2 y cuando λ tiende a cero se obtiene el estadístico G. Read y Cressie (1989) realizan una discusión interesante de estos estadísticos y establecieron que el estadístico G es preferible al Ji-Cuadrado de Pearson, lo cual coincide con las investigaciones de Ku y kullback (1974), William (1976) y otros.

A la hora de tomar una decisión en base a los estadísticos G o Ji-Cuadrado es necesario fijar un nivel de significación α. Ahora bien cabe preguntarse si es posible construir un estadístico de forma que, a la vista de los datos observados y sin necesidad de fijar un nivel de significación, permita discernir entre varios modelos.

Akaike (1973) propuso un estadístico con estas características, el cual se basa en la idea de minimizar la medida de discriminación esperada de Kullback-Leibler, este estadístico viene dado por

$$\text{AIC} = -2 \begin{pmatrix} \text{El máximo del logaritmo} \\ \text{de la verosimilitud del} \\ \text{modelo} \end{pmatrix} + 2 \begin{pmatrix} \text{Número de parámetros} \\ \text{libres del modelo} \end{pmatrix}$$

Además, Akaike propuso un criterio para seleccionar el mejor modelo, el cual consiste en lo siguiente: elegir, como mejor modelo, aquel que tenga el mínimo AIC (MAIC).

Generalmente, existen dos tipos diferentes de análisis con datos categorizables: uno va dirigido a evaluar la dependencia o relación entre los factores (variables) de una tabla de contingencia y el otro va dirigido al estudio de una variable respuesta a través de sus variables explicativas. El objetivo en el segundo caso es obtener el conjunto óptimo de variables explicativas. Este es uno de los motivos por el cual hemos creído conveniente analizar ambos casos en el presente trabajo.

El trabajo consta de cuatro capítulos que pasamos a describir.

En el capítulo primero se desarrolla el estadístico asociado al criterio de información de Akaike (AIC) para el estudio de dos tipos de modelos: modelo no restringido (no tiene restricción en los parámetros) y modelo restringido (con restricción en los parámetros). Además, se da una justificación rigurosa del criterio MAIC que propuso Akaike para seleccionar el mejor modelo. También se describen los resultados que obtuvimos en el estudio del error del AIC. Los resultados obtenidos en este capítulo se utilizarán a lo largo de la presente memoria.

En el segundo capítulo se presenta un método para realizar el análisis de tablas de contingencia de tres factores de clasificación. De esta forma se puede efectuar el estudio de independencia, homogeneidad e interacción entre los factores de clasificación. Como estos estudios se reducen al contraste de parámetros de una distribución multinomial, el método que proponemos en este capítulo es para el análisis de estos últimos contrastes.

En primer lugar se consideran las relaciones entre los diferentes contrastes dadas por Kullback (1959). Luego se establecen los modelos asociados a las diferentes hipótesis y se calcula el AIC asociado a cada modelo. Finalmente se aplica el criterio MAIC para seleccionar el mejor modelo y así tomar una decisión.

El método que se propone permite realizar un análisis más detallado de los datos, pues se pueden detectar las verdaderas causas de la independencia, homogeneidad e interacción de los factores de clasificación.

En el tercer capítulo se presenta un procedimiento basado en el criterio MAIC, 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.

En el cuarto capítulo se proporciona una discusión de los métodos introducidos para analizar datos categorizados. Se consideran dos casos: i) El análisis de tablas de contingencia (Capítulo II). ii) La selección del conjunto óptimo de variables explicativas de una variable respuesta (Capítulo III).

Además se presentarán los programas en Basic de los métodos propuestos y dos aplicaciones, las cuales son: i) El criterio de información de Akaike para el análisis de tablas de contingencia aplicado a datos del paro en España en 1990. ii) El criterio de información de Akaike para la selección del conjunto óptimo de variables explicativas de una variable respuesta aplicado al análisis de la fecundidad en España.

Por último deseo expresar mi sincero agradecimiento a los directores de esta memoria el Dr. D. Leandro Pardo Llorente y el Dr. D. Domingo Morales Gonzáles por la segura y objetiva orientación durante la elavoración de este trabajo, y por la gran dedicación y amistad que me brindaron.

CAPITULO I

"EL CRITERIO DE INFORMACION DE AKAIKE"

I.0.- SUMARIO

En este capítulo se desarrolla el estadístico asociado al Criterio de Información de Akaike (AIC) para el estudio de dos tipos de modelos: modelo no restringido (cuando no existen restricciones en los parametros del modelo) y modelo restringido (cuando existen restricciones en los parámetros del modelo).

Además se presenta el criterio de selección que propuso Akaike para determinar el mejor entre varios modelos propuestos. Dicho criterio esta basado en el mínimo AIC (MAIC).

Si bien los resultados básicos que se presentan son una recopilación de los trabajos de Akaike (1973, 1977, 1991) y Sakamoto, Ishiguro y kitagawa (1986) en este capítulo se proporciona una justificacion rigurosa del criterio que propuso Akaike para seleccionar el mejor modelo.

Sakamoto, Ishiguro y Kitagawa (1986) estudiaron el error del AIC cuando $\theta_k^{\bullet}$ es $(\theta_1^{\bullet}, \ldots, \theta_k^{\bullet}, 0, \ldots, 0)$ y la matriz de información de Fisher, $J_{\bullet}$, es la matriz unidad. Nosotros complementamos el estudio del error del AIC analizando los siguientes casos: 1) $\theta_k^{\bullet} = (\theta_1^{\bullet}, \ldots, \theta_k^{\bullet}, 0, \ldots, 0)$ y $J_{\bullet}$ es una matriz definida positiva, 2) $\theta_k^{\bullet} = (\theta_1^{\bullet}, \ldots, \theta_k^{\bullet}, 0_{k+1}, \ldots, 0_K)$ y $J_{\bullet}$ es la matriz unidad, 3) $\theta_k^{\bullet} = (\theta_1^{\bullet}, \ldots, \theta_k^{\bullet}, 0_{k+1}, \ldots, 0_K)$ y $J_{\bullet}$ es una matriz definida positiva. Estos resultados se presentan en la Sección I.2.

Los resultados obtenidos en este capítulo se utilizaran posteriormente en el estudio de algunos problemas:

a) Tablas de contingencia de tres factores de clasificacion: independencia, homogeneidad e interaccion entre los factores. Estos problemas seran tratados en el Capítulo II.

b) Análisis de una variable respuesta a través de variables explicativas, donde el objetivo es la selección del conjunto óptimo de variables explicativas de una variable respuesta. Este problema se tratará en el Capitulo III.

I.1 PRINCIPIO DE MAXIMIZACION DE LA INFORMACION DE AKAIKE EN EL CASO DE MODELO NO RESTRINGIDO

En este apartado se introducirá la notación que se utilizará a lo largo de la memoria asi como algunas definiciones y desarrollos esenciales a lo largo de la misma.

Considérese el espacio estadistico $(\mathcal{X}, \beta_{\mathcal{X}}, P_\theta)_{\theta \in \Theta_K}$ y y sea $f(x/\theta)$ con $\theta = (\theta_1,\ldots,\theta_K) \in \Theta_K$ la densidad de P_θ con con respecto a una medida σ-finita υ. Sea $X = (X_1,\ldots,X_n)$ una muestra aleatoria simple y sea $\theta^* = (\theta_1^*,\ldots,\theta_K^*) \in \Theta_K$ el valor del vector de parámetros que el estadistico propone como verdadero. Sea $Z = (Z_1,\ldots,Z_n)$ una variable aleatoria con la misma distribución que X, luego la función de densidad de Z dado θ^* es $f(z_1,\ldots,z_n/\theta^*)$. La función de densidad de Z dado θ es $f(z_1,\ldots,z_n/\theta)$. Si $\theta(z_1,\ldots,z_n)$ es un estimador de θ, la función de densidad estimada de X será $f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))$ con esta notación el logaritmo de la función de verosimilitud viene dado por

$$\ell(\theta) = \log f(x_1,\ldots,x_n/\theta) = \sum_{i=1}^{n} \log f(x_i/\theta) \qquad (1.1)$$

y el logaritmo de la función de verosimilitud estimada por

$$\ell(\theta(z_1,\ldots,z_n)) = \sum_{i=1}^{n} \log f(x_i/\theta(z_1,\ldots,z_n)) \ . \qquad (1.2)$$

El estimador de máxima verosimilitud se denotará por $\hat{\theta}_K(z_1,\ldots,z_n)$ que bajo condiciones de regularidad (ver Lindgren (1976))

$$\sqrt{n}\left(\hat{\theta}_K(z_1,\ldots,z_n) - \theta^*\right) \xrightarrow{L} N\left(0,\ J_*^{-1}\right) \tag{1.3}$$

donde J_* es la matriz de información de Fisher.

Además, por $\ell^*(\theta)$, $\ell^*(\theta(z_1,\ldots,z_n))$, $\ell_n^*(K)$ y $\ell_n^{**}(K)$ se denotarán respectivamente a las esperanzas

$$\ell^*(\theta) = E_{X/\theta^*}\left(\ell(\theta)\right) = n\, E_{X_i/\theta^*}\left(\log f(x_i/\theta)\right) \tag{1.4}$$

$$\begin{aligned}\ell^*(\theta(z_1,\ldots,z_n)) &= E_{X/\theta^*}\left(\ell(\theta(z_1,\ldots,z_n))\right)\\ &= n\, E_{X_i/\theta^*}\left(\log f(x_i/\theta(z_1,\ldots,z_n))\right)\\ &= n\int_{\mathcal{X}} \log f(x_i/\theta(z_1,\ldots,z_n))\, f(x_i/\theta^*)\, dx_i\end{aligned} \tag{1.5}$$

$$\begin{aligned}\ell_n^*(K) &= E_{Z/\theta^*}\left(\ell^*(\hat{\theta}_K(z_1,\ldots,z_n))\right)\\ &= \int_{\mathcal{X}^n}\left[\int_{\mathcal{X}^n} \log f(x_1,\ldots,x_n/\hat{\theta}_K(z_1,\ldots,z_n))\, f(x_1,\ldots,x_n/\theta^*)\, dx_1..dx_n\right]^*\\ &\quad * f(z_1,\ldots,z_n/\theta^*)\, dz_1\ldots dz_n\end{aligned} \tag{1.6}$$

$$\ell_n^{**}(K) = E_{X/\theta^*}\left[\ell(\hat{\theta}_K(x_1,\ldots,x_n))\right]$$

$$= \int_{\mathcal{X}^n} \log f(x_1,\ldots,x_n/\hat{\theta}_K(x_1,\ldots,x_n))\, f(x_1,\ldots,x_n/\theta^*)\, dx_1\ldots dx_n \qquad (1.7)$$

donde con el objeto de simplificar la notación se ha escrito dx_i y dz_i en lugar de $d\nu(x_i)$ y $d\nu(z_i)$ respectivamente.

Supongamos que se desea saber si un vecto- de parámetros $\theta = (\theta_1,\ldots,\theta_K)$, $\theta \in \Theta_K$ se aproxima a θ^*. Este problema se podrá resolver mediante el siguiente contraste de hipótesis

$$H_0 : f(x_1,\ldots,x_n/\theta) = f(x_1,\ldots,x_n/\theta^*)$$

frente a

$$H_1 : f(x_1,\ldots,x_n/\theta) \neq f(x_1,\ldots,x_n/\theta^*) \qquad (1.8)$$

es decir,

$H_0 : \theta = \theta^*$ frente a $H_1 : \theta \neq \theta^*$

lo cual equivale a analizar los siguientes modelos que numeramos según el número de parámetros desconocidos

MODELO(0) : $f(x_1,\ldots,x_n/\theta^*)$, $\theta^* = (\theta_1^*,\ldots,\theta_K^*)$

MODELO(K) : $f(x_1,\ldots,x_n/\theta)$, $\theta = (\theta_1,\ldots,\theta_K) \in \Theta_K$

El MODELO(K), se denomina modelo no restringido ya que no existen restricciones en los parámetros, y por tanto el numero de parámetros libres del modelo es K.

Si $\theta(z_1,\ldots,z_n)$ es el resultado de un proceso de estimación previo, la función de verosimilitud del MODELO(K) será $f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))$.

Akaike en (1973) propuso un procedimiento para seleccionar el mejor MODELO(K) en el caso de que se tuvieran varios modelos MODELO(K). Este procedimiento consiste en primero calcular el AIC(K) (es decir, el estadistico del Criterio de Información de Akaike del MODELO(K)) para cada MODELO(K), donde

$$AIC(K) = -2\,\ell(\hat{\theta}_K(z_1,\ldots,z_n)) + 2K \qquad (1.9)$$

$$AIC(K) = -2\left(\begin{array}{l}\text{El máximo del logaritmo}\\ \text{de la verosimilitud del}\\ \text{modelo}\end{array}\right) + 2\left(\begin{array}{l}\text{Número de parámetros}\\ \text{libres del modelo}\end{array}\right)$$

y luego elegir como el mejor modelo, aquel que tenga el mínimo AIC(K) (MAIC).

En esta sección se presentará una justificación del criterio que propuso Akaike para seleccionar el mejor MODELO(K).

Una medida natural para contrastar las hipótesis dadas en (1.8) es la medida de discriminación de Kullback-Leibler (ver Kullback (1959)), la cual viene dada por

$$I\left(f(x_1,\ldots,x_n/\theta^*);f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))\right) =$$

$$= \int_{\mathcal{X}^n} f(x_1,\ldots,x_n/\theta^*)\,\log\left(\frac{f(x_1,\ldots,x_n/\theta^*)}{f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))}\right) dx_1\ldots dx_n$$

$$= \int_{\mathcal{X}^n} \log f(x_1,\ldots,x_n/\theta^*)\, f(x_1,\ldots,x_n/\theta^*)\, dx_1\ldots dx_n$$

$$- \int_{\mathcal{X}^n} \log f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))\, f(x_1,\ldots,x_n/\theta^*)\, dx_1\ldots dx_n$$

$$= E_{X/\theta^*}\left[\log f(x_1,\ldots,x_n/\theta^*)\right]$$

$$- E_{X/\theta^*}\left[\log f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))\right]$$

$$= \ell^*(\theta^*) - \ell^*(\theta(z_1,\ldots,z_n)) \ . \qquad (1.10)$$

El criterio ideal sería minimizar uniformemente en $(Z_1,\ldots,Z_n)$ la medida de discriminación de Kullback-Leibler.

Basándose en esta idea, Akaike en 1977 propuso el siguiente principio de minimización: minimizar la medida de discriminación esperada de Kullback-Leibler

$$\underset{\theta(z_1,\ldots,z_n)}{\text{Min}} \ E_{Z/\theta^*}\left[I\left(f(x_1,\ldots,x_n/\theta^*);f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))\right)\right] \qquad (1.11)$$

$$= \underset{\theta(z_1,\ldots,z_n)}{\text{Min}} \ E_{Z/\theta^*}\left[\ell^*(\theta^*) - \ell^*(\theta(z_1,\ldots,z_n))\right] \qquad \text{de (1.10)}$$

$$= \underset{\theta(z_1,\ldots,z_n)}{\text{Min}} \left[\ell^*(\theta^*) - E_{Z/\theta^*}\left[\ell^*(\theta(z_1,\ldots,z_n))\right]\right] \ .$$

Como $\ell^*(\theta^*)$ es una constante, minimizar $\left[\ell^*(\theta^*) - E_{Z/\theta^*}\left[\ell^*(\theta(z_1,\ldots,z_n))\right]\right]$ equivale

$$\underset{\theta(z_1,\ldots,z_n)}{\text{Max}} \ E_{Z/\theta^*}\left[\ell^*(\theta(z_1,\ldots,z_n))\right] \ . \qquad (1.12)$$

Analicemos esta última expresión.

Si $\hat{\theta}_K(z_1,\ldots,z_n)$ es un estimador máxima verosimilitud de de θ obtenido a partir de $(z_1,\ldots,z_n)$ entonces

$$\log f(x_1,\ldots,x_n/\theta) \leq \log f(x_1,\ldots,x_n/\hat{\theta}_K(z_1,\ldots,z_n)) \quad \forall\, x_1,\ldots,x_n.$$

Sea $\theta(z_1,\ldots,z_n)$ un estimador de θ obtenido a partir de $z_1,\ldots,z_n$, entonces

$$\ell(\theta(z_1,\ldots,z_n)) = \log f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))$$

$$\leq \log f(x_1,\ldots,x_n/\hat{\theta}_K(z_1,\ldots,z_n))$$

$$\forall\, z_1,\ldots,z_n.$$

Tomando esperanza en X/θ^*, se tiene

$$E_{X/\theta^*}\left(\ell(\theta(z_1,\ldots,z_n))\right) \leq E_{X/\theta^*}\left(\log f(x_1,\ldots,x_n/\hat{\theta}_k(z_1,\ldots,z_n))\right)$$

de (1.5) se tiene que

$$\ell^*(\theta(z_1,\ldots,z_n)) \leq \ell^*(\hat{\theta}_K(z_1,\ldots,z_n)) .$$

Volviendo a tomar esperanza en Z/θ^*, se llega a

$$E_{Z/\theta^*}\left(\ell^*(\theta(z_1,\ldots,z_n))\right) \leq E_{Z/\theta^*}\left(\ell^*(\hat{\theta}_K(z_1,\ldots,z_n))\right) .$$

De (1.6) se tiene que

$$E_{Z/\theta^*}\left(\ell^*(\theta(z_1,\ldots,z_n))\right) \leq \ell_n^*(K) ,$$

luego

$$\underset{\theta(z_1,\ldots,z_n)}{\text{Max}} \; E_{Z/\theta^*}\left(\ell^*(\theta(z_1,\ldots,z_n))\right) \leq \ell_n^*(k) . \qquad (1.13)$$

Por otro lado si se efectúa el desarrollo de Taylor de $\ell^{\bullet}(\theta) = n\, E_{X_i/\theta^{\bullet}}\left[\log f(x_i/\theta)\right]$ alrededor de $\theta^{\bullet}$, se tiene

$$\ell^{\bullet}(\theta) = \ell^{\bullet}(\theta^{\bullet}) + n\,[\theta - \theta^{\bullet}]\, E_{X_i/\theta^{\bullet}}\left[\frac{d}{d\theta}\log f(x_i/\theta)\right]_{\theta^{\bullet}} +$$

$$+ \frac{1}{2}\, n\,[\theta - \theta^{\bullet}]\, E_{X_i/\theta^{\bullet}}\left[\frac{d^2}{d\theta^2}\log f(x_i/\theta)\right]_{\theta^{\bullet}} [\theta - \theta^{\bullet}]^t + R_n ,$$

donde $R_n \xrightarrow[n\to\infty]{P} 0$.

(1.14)

Es inmediato comprobar que

$$E_{X_i/\theta^{\bullet}}\left[\frac{d}{d\theta}\log f(x_i/\theta)\right]_{\theta^{\bullet}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} ,$$

por tanto el segundo término de (1.14) se anula.

Además,

$$J_{\bullet} = -E_{X_i/\theta^{\bullet}}\left[\frac{d^2}{d\theta^2}\log f(x_i/\theta)\right]_{\theta^{\bullet}}$$

luego (1.4) se puede expresar como

$$\ell^{\bullet}(\theta) = \ell^{\bullet}(\theta^{\bullet}) - \frac{1}{2}\sqrt{n}\,[\theta - \theta^{\bullet}]\, J_{\bullet}\, \sqrt{n}\,[\theta - \theta^{\bullet}]^t + R_n .$$

(1.15)

Haciendo $\theta = \hat{\theta}_K(z_1, \ldots, z_n)$ en (1.15) se tiene

$$\ell^{\bullet}(\hat{\theta}_K(z_1, \ldots, z_n)) =$$

$$= \ell^{*}(\theta^{*}) - \frac{1}{2}\sqrt{n}\,[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]\,J_{*}\,\sqrt{n}\,[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]^{t} + R_n \,. \tag{1.16}$$

Tomando esperanza en Z/θ^{*} (1.16) y aplicando (1.6) se llega

$$\ell_n^{*}(K) = E_{Z/\theta^{*}}\left(\ell^{*}(\hat{\theta}_K(z_1,\ldots,z_n))\right) =$$

$$= E_{Z/\theta^{*}}\left(\ell^{*}(\theta^{*})\right) - \frac{1}{2}\,E_{Z/\theta^{*}}\left(\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]J_{*}\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]^{t}\right)$$

$$+ E_{Z/\theta^{*}}\left(R_n\right) \,. \tag{1.17}$$

Como $R_n \xrightarrow[n\to\infty]{P} 0$ y

$$\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]J_{*}\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]^{t} \xrightarrow{L} \chi^2_K \,, \tag{1.18}$$

se tiene para n grande

$$E_{Z/\theta^{*}}\left(\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]J_{*}\sqrt{n}[\hat{\theta}_K(z_1,\ldots,z_n)-\theta^{*}]^{t}\right) = K \,. \tag{1.19}$$

Además como $\ell^{*}(\theta^{*})$ es una constante, se tiene que $E_{Z/\theta^{*}}\left(\ell^{*}(\theta^{*})\right) = \ell^{*}(\theta^{*})$. Aplicando este último resultado y (1.19) en (1.17) se tiene para n grande que

$$\ell_n^{*}(K) = \ell^{*}(\theta^{*}) - \frac{K}{2} \,. \tag{1.20}$$

Como $K/2 > 0$, entonces

$$\ell_n^*(K) \leq \ell^*(\theta^*) \ . \qquad (1.21)$$

De otro lado, si se efectúa el desarrollo de Taylor del logaritmo de la función de verosimilitud $\ell(\theta)$, alrededor del estimador de máxima verosimilitud $\hat{\theta}_K(x_1,\ldots,x_n)$ se obtiene

$$\ell(\theta) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) + [\theta-\hat{\theta}_K(x_1,\ldots,x_n)]\left(\frac{d}{d\theta}\ell(\theta)\right)_{\hat{\theta}_K(x_1,\ldots,x_n)} +$$

$$+ \frac{1}{2}[\theta-\hat{\theta}_K(x_1,\ldots,x_n)]\left(\frac{d^2}{d\theta^2}\ell(\theta)\right)_{\hat{\theta}_K(x_1,\ldots,x_n)}[\theta-\hat{\theta}_K(x_1,\ldots,x_n)]^t + R_n$$

donde $R_n \xrightarrow[n\to\infty]{P} 0$.

Multiplicando y dividiendo por n el tercer término del lado derecho

$$\ell(\theta) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) + [\theta-\hat{\theta}_K(x_1,\ldots,x_n)]\left(\frac{d}{d\theta}\ell(\theta)\right)_{\hat{\theta}_k(x_1,\ldots,x_n)} +$$

$$+ \frac{1}{2}\sqrt{n}[\theta-\hat{\theta}_k(x_1,\ldots,x_n)]\frac{1}{n}\left(\frac{d^2}{d\theta^2}\ell(\theta)\right)_{\hat{\theta}_K(x_1,\ldots,x_n)}\sqrt{n}\,[\theta-\hat{\theta}_k(x_1,\ldots,x_n)]^t +$$

$$+ R_n \ . \qquad (1.22)$$

Como $\hat{\theta}_K(x_1,\ldots,x_n)$ es un estimador de máxima verosimilitud de θ, entonces $\left(\frac{d}{d\theta}\ell(\theta)\right)_{\hat{\theta}_K(x_1,\ldots,x_n)} = 0$, luego el segundo término de (1.22) se anula.

De otro lado por la Ley de los Grandes Números se tiene

$$\frac{1}{n}\left(\frac{d^2}{d\theta^2}\,\ell(\theta)\right)_{\theta^*} = \frac{1}{n}\left(\frac{d^2}{d\theta^2}\sum_{i=1}^{n}\log f(x_i/\theta)\right)_{\theta^*} =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(x_i/\theta)\right)_{\theta^*} \xrightarrow[n\to\infty]{} E_{X_i/\theta^*}\left(\frac{d^2}{d\theta^2}\log f(x_i/\theta)\right)_{\theta^*} = -J_{\theta^*} .$$

Como $\hat{\theta}_K(x_1,\ldots,x_n) \longrightarrow \theta^*$ cuando $n \to \infty$, se concluye que

$$\frac{1}{n}\left(\frac{d^2}{d\theta^2}\,\ell(\theta)\right)_{\hat{\theta}_K(x_1,\ldots,x_n)} \longrightarrow -J_{\theta^*} \quad \text{cuando } n \to \infty .$$

Luego para n suficientemente grande se tiene que (1.22) será

$$\ell(\theta) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{1}{2}\sqrt{n}[\theta-\hat{\theta}_K(x_1,\ldots,x_n)]J_{\theta^*}\sqrt{n}[\theta-\hat{\theta}_K(x_1,\ldots,x_n)]^t + R_n .$$

(1.23)

Tomando $\theta = \theta^*$ en la ecuación anterior y reordenando se llega a

$$\ell(\theta^*) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{1}{2}\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^*]J_{\theta^*}\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^*]^t + R_n ,$$

y tomando esperanza en X/θ^* se tiene

$$E_{X/\theta^*}\left(\ell(\theta^*)\right) = E_{X/\theta^*}\left(\ell(\hat{\theta}_K(x_1,\ldots,x_n))\right) - \frac{1}{2}E_{X/\theta^*}\left(\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^*]J_{\theta^*}\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^*]^t\right) + E_{X/\theta^*}\left(R_n\right).$$

(1.24)

De (1.4) se tiene

$$\ell^{*}(\theta^{*}) = n\, E_{X_i/\theta^{*}}\left(\log f(x_i/\theta^{*})\right) = E_{X/\theta^{*}}\left(\sum_{i=1}^{n} \log f(x_i/\theta^{*})\right)$$

$$= E_{X/\theta^{*}}\left(\ell(\theta^{*})\right) . \qquad (1.25)$$

Además de (1.18) se tiene para n grande que

$$E_{X/\theta^{*}}\left(\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^{*}]\,J_{*}\sqrt{n}[\hat{\theta}_K(x_1,\ldots,x_n)-\theta^{*}]^{t}\right) = K . \qquad (1.26)$$

Luego aplicando (1.7), (1.25) y (1.26) en (1.24) se tiene para n grande que

$$\ell^{*}(\theta^{*}) = \ell_n^{**}(K) - \frac{K}{2} . \qquad (1.27)$$

Sustituyendo (1.27) en (1.20) se llega a

$$\ell_n^{*}(K) = \ell_n^{**}(K) - K .$$

Por tanto

$$\ell_n^{**}(K) = \ell_n^{*}(K) + K \quad ; \text{ es decir,}$$

$$\ell_n^{**}(K) = E_{X/\theta^{*}}\left(\ell(\hat{\theta}_K(x_1,\ldots,x_n))\right) = \ell_n^{*}(K) + K . \qquad (1.28)$$

Como $K > 0$, se tiene que

$$\ell_n^{**}(K) \geq \ell_n^{*}(K) . \qquad (1.29)$$

De (1.13), (1.21), (1.27) y (1.29), se obtiene

$$\underset{\theta(z_1,\ldots z_n)}{\text{Max}} \; E_{Z/\theta^*}\left(\ell^*(\theta(z_1,\ldots,z_n))\right) \le \ell_n^*(K) \le \ell^*(\theta^*) \le \ell_n^{**}(K) .$$

Se observa que $\ell_n^*(K)$ está más próximo de $\ell^*(\theta^*)$ que es en verdad lo que tratamos de hallar. Asi pues, $\ell_n^*(K)$ es una medida que esta cerca de $\ell^*(\theta^*)$ y evalua el MODELO(K). De esta forma en lugar de seleccionar el MODELO(K) que maximice en $\theta(.)$ la expresión

$$E_{Z/\theta^*}\left(\ell^*(\theta(z_1,\ldots,z_n))\right)$$

se puede usar un criterio "análogo" que consistiria en seleccionar el modelo que maximice $\ell_n^*(K)$. Ahora bien, como $\ell_n^*(K)$ no es siempre analiticamente calculable, Akaike propuso evaluar AIC(K) que es un estimador centrado de $-2\,\ell_n^*(K)$ y elegir el MODELO(K) que minimice el valor de AIC(K) observado.

Esta conclusión se puede ratificar con el siguiente razonamiento. Por la desigualdad de Gibbs se tiene que para todo conjunto de valores $(z_1,\ldots,z_n)$ prefijado.

$\ell^*(\theta(z_1,\ldots,z_n)) \le \ell^*(\theta^*)$ con igualdad si y sólo si $f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n)) = f(x_1,\ldots,x_n/\theta^*)$, donde $\theta(z_1,\ldots,z_n)$ es un estimador de θ.

Se observa que mientras más grande sea $\ell^*(\theta(z_1,\ldots,z_n))$ mejor será la aproximación de la distribución $f(x_1,\ldots,x_n/\theta(z_1,\ldots,z_n))$ a la verdadera distribución $f(x_1,\ldots,x_n/\theta^*)$.

La bondad del ajuste del MODELO(K) se puede evaluar por $\ell^*(\hat{\theta}_K(z_1,\ldots,z_n))$, donde $\hat{\theta}_K(z_1,\ldots,z_n)$ es el estimador de máxima verosimilitud de θ. Sin embargo, esta cantidad depende de la realización z de Z. No obstante si se calcula la esperanza en Z/θ^*

de $\ell^*(\hat{\theta}_K(z_1,\ldots,z_n))$; es decir,

$$\ell_n^*(K) = E_{Z/\theta^*}\left(\ell^*(\hat{\theta}_K(z_1,\ldots,z_n))\right) ,$$

se obtiene una medida que no depende de una realización particular. Por lo tanto evaluaremos el MODELO(K) mediante $\ell_n^*(K)$.

A continuación se procede a estimar $\ell_n^*(K)$: de (1.28) se observa que $\ell(\hat{\theta}_K(x_1,\ldots,x_n))$ es un estimador sesgado de $\ell_n^*(K)$ con sesgo K, si se corrige ese sesgo se obtiene el estimador

$$T = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - K$$

el cual es un estimador insesgado de $\ell_n^*(K)$.

El estadistico asociado al criterio de información de Akaike (1973) del MODELO(K) es -2 T, es decir

$$AIC(K) = -2\,\ell(\hat{\theta}_K(x_1,\ldots,x_n)) + 2K$$

AIC(K) es un estimador insesgado de $-2\,\ell_n^*(K)$ y esto viene del hecho

$$\begin{aligned} E_{X/\theta^*}\left(AIC(K)\right) &= E_{X/\theta^*}\left(-2\,\ell(\hat{\theta}_K(x_1,\ldots,x_n)) + 2K\right) \\ &= -2\,E_{X/\theta^*}\left(\ell(\hat{\theta}_K(x_1,\ldots,x_n))\right) + 2K \\ &= -2\,\ell_n^{**}(K) + 2K \qquad \text{de (1.7)} \\ &= -2\left(\ell_n^*(K) + K\right) + 2K \qquad \text{de (1.28)} \\ &= -2\,\ell_n^*(K) . \end{aligned}$$

Si se tuvieran varios modelos MODELO(K), el criterio que propuso Akaike para elegir el mejor modelo MODELO(K) es el siguiente: primero calcular el AIC(K) para cada MODELO(K) y luego

seleccionar aquel modelo que tenga el mínimo AIC(K).

Obsérvese que al tomar el mínimo AIC(K), en realidad se está tomando el modelo que tenga el mayor $\ell(\hat{\theta}_K(x_1,\ldots,x_n))$, lo cual implicará que se seleccionará el modelo cuyo $\ell(\hat{\theta}_K(x_1,\ldots,x_n))$ esté más próximo de $\ell^*(\theta^*)$, que es lo que se busca. Esta elección pretende forzar la proximidad de $\ell_n^*(K)$ a $\ell^*(\theta^*)$, y ésto a su vez posibilita que la divergencia o la medida de discriminación esperada de Kullback-Leibler sea pequeña.

I.2.- CRITERIO DE INFORMACION DE AKAIKE EN EL CASO DE MODELO RESTRINGIDO Y EL ERROR DEL AIC

En esta sección se presenta la obtención del estadístico asociado al Criterio de Información de Akaike para el caso en el que existan restricciones en los parámetros del modelo.

Aquí se utilizan las mismas suposiciones definidas anteriormente, pero se trabajará en el caso específico en que el espacio paramétrico esté restringido, es decir,

$$\Theta_k = \left\{\theta \in \Theta_K \;/\; \theta_{k+1} = O_{k+1},\ldots,\theta_K = O_K\right\} \qquad , \; 1 \le k \le K$$

$$\text{y} \quad \theta = (\theta_1,\ldots,\theta_k,O_{k+1},\ldots,O_K) \in \Theta_k \qquad , \; 1 \le k \le K$$

donde O_j es un valor dado del parámetro θ_j para $j = k+1,\ldots,K$.

Un caso particular de este espacio paramétrico restringido es el dado por

$$\Theta_k = \left\{\theta \in \Theta_K \;/\; \theta_{k+1} = \theta_{k+2} = \ldots = \theta_K = 0\right\} \qquad , \; 1 \le k \le K$$

luego $\theta = (\theta_1,\ldots,\theta_k,0,\ldots,0) \in \Theta_k$ es el vector de parámetros.

Y $\theta_k^* = (\theta_1^*,\ldots,\theta_k^*,0,\ldots,0) \in \Theta_k$ es el valor del vector de

parámetros que el estadístico propone como verdadero.

Sea $\hat{\theta}_k(z_1,\ldots,z_n)$ el estimador de máxima verosimilitud de θ_k^*, es decir

$$\ell(\hat{\theta}_k(z_1,\ldots,z_n)) = \underset{\theta \in \Theta_k}{\text{Max}} \ \ell(\theta) .$$

El contraste que interesa efectuar es

$$H_0 : f(x_1,\ldots,x_n/\theta) = f(x_1,\ldots,x_n/\theta_k^*) \quad , \ \theta_k^* \in \Theta_k \ , \ \theta \in \Theta_k \ , \ 1 \le k \le K$$

frente a

$$H_1 : f(x_1,\ldots,x_n/\theta) \neq f(x_1,\ldots,x_n/\theta_k^*) \quad , \ \theta_k^* \in \Theta_k \ , \ \theta \in \Theta_k \ , \ 1 \le k \le K$$

Los modelos asociados a este contraste de hipótesis son

MODELO(0) : $f(x_1,\ldots,x_n/\theta_k^*) \quad , \ \theta_k^* \in \Theta_k$

MODELO(k) : $f(x_1,\ldots,x_n/\theta) \quad , \ \theta \in \Theta_k$

A continuación se va a deducir el AIC(k) del MODELO(k). La expresión $\ell^*(\theta)$ dada en (1.15) se puede escribir como

$$\ell^*(\theta) = \ell^*(\theta^*) - \frac{n}{2} \|\theta - \theta^*\|^2 + R_n \ , \qquad (1.30)$$

donde

$$\|\theta\|^2 = \theta \, J_* \, \theta^t \qquad (1.31)$$

Si se toma $\theta^* = \theta_k^*$ en (1.30) se tiene

$$\ell^*(\theta) = \ell^*(\theta_k^*) - \frac{n}{2} \|\theta - \theta_k^*\|^2 + R_n \quad . \tag{1.32}$$

De otro lado, se tiene que

$\hat{\theta}_k(z_1,\ldots,z_n) = (\hat{\theta}_1(z_1,\ldots,z_n),\ldots,\hat{\theta}_k(z_1,\ldots,z_n),0,\ldots,0)$
es un estimador de máxima verosimilitud de $\theta_k^* = (\theta_1^*,\ldots,\theta_k^*,0,\ldots,0)$
luego se tiene de (1.3) que

$$\sqrt{n}\ (\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*) \xrightarrow{L} N\left(0,\ J_{*k}^{-1}\right) \quad ,$$

donde J_{*k} es la matriz de informacion de Fisher asociaca al subespacio paramétrico Θ_k.

Luego

$$\sqrt{n}\ [\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*]\ J_{*k}\ \sqrt{n}\ [\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*]^t \xrightarrow{L} \chi_k^2 \quad .$$

Utilizando la notación dada en (1.31) en la expresión anterior, se tiene

$$n\ \|\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*\|^2 \xrightarrow{L} \chi_k^2 \quad . \tag{1.33}$$

Por tanto

$$E_{Z/\theta_k^*}\left(n\ \|\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*\|^2\right) = k \quad ,$$

de donde se obtiene

$$E_{Z/\theta_k^*}\left(\|\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^*\|^2\right) = \frac{k}{n} \quad . \tag{1.34}$$

Si se toma $\theta = \hat{\theta}_k(z_1,\ldots,z_n)$ en (1.32) se tiene

$$\ell^{*}(\hat{\theta}_k(z_1,\ldots,z_n)) = \ell^{*}(\theta_k^{*}) - \frac{n}{2}\,\|\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^{*}\|^2 + R_n \ .$$

Tomando esperanza en Z/θ_k^{*} a $\ell^{*}(\hat{\theta}_k(z_1,\ldots,z_n))$ y aplicando (1.6) se tiene

$$\ell_n^{*}(k) = E_{Z/\theta_k^{*}}\left(\ell^{*}(\hat{\theta}_k(z_1,\ldots,z_n))\right)$$

$$= E_{Z/\theta_k^{*}}\left(\ell^{*}(\theta_k^{*})\right) - \frac{n}{2}\,E_{Z/\theta_k^{*}}\left(\|\hat{\theta}_k(z_1,\ldots,z_n) - \theta_k^{*}\|^2\right) + E_{Z/\theta_k^{*}}\left(R_n\right).$$

Aplicando (1.34) y como $\ell^{*}(\theta_k^{*})$ es constante, se tiene para n grande que la expresión anterior será

$$\ell_n^{*}(k) = \ell^{*}(\theta_k^{*}) - \frac{k}{2} \ . \qquad (1.35)$$

De otro lado, si en la expresión de $\ell(\theta)$ dado en (1.23) se considera la notación dada en (1.31), se tiene

$$\ell(\theta) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{n}{2}\,\|\theta - \hat{\theta}_K(x_1,\ldots,x_n)\|^2 + R_n \ , \qquad (1.36)$$

donde $R_n \xrightarrow[n\to\infty]{P} 0$.

Si se toma $\hat{\theta}_K(x_1,\ldots,x_n) = \hat{\theta}_k(x_1,\ldots,x_n)$ en (1.36) se llega a

$$\ell(\theta) = \ell(\hat{\theta}_k(x_1,\ldots,x_n)) - \frac{n}{2}\,\|\theta - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 + R_n$$

y si en esta última ecuación se hace $\theta = \theta_k^{*}$, se tiene

$$\ell(\theta_k^{*}) = \ell(\hat{\theta}_k(x_1,\ldots,x_n)) - \frac{n}{2}\,\|\hat{\theta}_k(x_1,\ldots,x_n) - \theta_k^{*}\|^2 + R_n \ .$$

Tomando esperanza en X/θ_k^* a $\ell(\theta_k^*)$

$$E_{X/\theta_k^*}\left(\ell(\theta_k^*)\right) = E_{X/\theta_k^*}\left(\ell(\hat{\theta}_k(x_1,\ldots,x_n))\right)$$

$$-\frac{1}{2}\, E_{X/\theta_k^*}\left(n\,\|\theta_k(x_1,\ldots,x_n)-\theta_k^*\|^2\right) + E_{X/\theta_k^*}\left(R_n\right) .$$

Aplicando (1.34) y para n suficientemente grande se tiene que la expresión anterior será

$$E_{X/\theta_k^*}\left(\ell(\theta_k^*)\right) = E_{X/\theta_k^*}\left(\ell(\hat{\theta}_k(x_1,\ldots,x_n))\right) - \frac{k}{2} . \qquad (1.37)$$

De (1.4) se tiene

$$\ell^*(\theta_k^*) = E_{X/\theta_k^*}\left(\ell(\theta_k^*)\right) .$$

Aplicando este resultado y (1.7) en (1.37), se obtiene

$$\ell^*(\theta_k^*) = \ell_n^{**}(k) - \frac{k}{2} . \qquad (1.38)$$

Sustituyendo (1.38) en (1.35) se tiene

$$\ell_n^*(k) = \ell_n^{**}(k) - k ,$$

luego

$$\ell_n^{**}(k) = \ell_n^*(k) + k , \qquad (1.39)$$

o lo que es lo mismo

$$E_{X/\theta_k^*}\left(\ell(\hat{\theta}_k(x_1,\ldots,x_n))\right) = \ell_n^*(k) + k .$$

Por tanto $\ell(\hat{\theta}_k(x_1,\ldots,x_n))$ es un estimador sesgado de $\ell_n^*(k)$ con sesgo k. Si se corrige ese sesgo se obtiene

$$S = \ell(\hat{\theta}_k(x_1,\ldots,x_n)) - k \ ,$$

el cual es un estimador insesgado de $\ell_n^*(k)$.

El estadistico asociado al criterio de información de Akaike del MODELO(k) es -2 S; es decir,

$$AIC(k) = -2\ \ell(\hat{\theta}_k(x_1,\ldots,x_n)) + 2\ k \quad , \ k = 1,\ldots,K \ . \qquad (1.40)$$

El AIC(k) es un estimador insesgado de $-2\ \ell_n^*(k)$, ya que

$$E_{X/\theta_k^*}\left(AIC(k)\right) = E_{X/\theta_k^*}\left(-2\ S\right)$$

$$= E_{X/\theta_k^*}\left(-2\ \ell(\hat{\theta}_k(x_1,\ldots,x_n)) + 2\ k\right)$$

$$= -2\ E_{X/\theta_k^*}\left(\ell(\hat{\theta}_k(x_1,\ldots,x_n))\right) + 2\ k$$

$$= -2\ \ell_n^{**}(k) + 2\ k \qquad \text{de (1.7)}$$

$$= -2\ \left(\ell_n^*(k) + k\right) + 2\ k \qquad \text{de (1.39)}$$

$$= -2\ \ell_n^*(k) \ .$$

Si se tienen varios modelos MODELO(k) el criterio para la selección del mejor modelo es el mismo que el dado en la sección anterior; es decir, elegir el modelo con el mínimo AIC.

Cabe señalar que se obtienen los mismos resultados anteriores, si se hubiera considerado el MODELO(k) definido sobre el espacio paramétrico restringido

$$\Theta_k = \left\{ \theta \in \Theta_K \;/\; \theta_{k+1} = 0_{k+1}, \ldots, \theta_K = 0_K \right\} \quad , \; 1 \le k \le K \; ,$$

para la obtención del estadístico asociado al criterio de información de Akaike.

A continuación presentamos el error del AIC(k) del MODELO(k).

ERROR DEL AIC(k)

En esta sección se ha visto que AIC(k) es un estimador insesgado de $-2\,\ell_n^*(k)$, por tanto $-\frac{1}{2}$ AIC(k) es un estimador de $\ell_n^*(k)$. A continuación se analizara el error de estimación de AIC(k).

En primer lugar se deduciran unas relaciones que serán necesarias para la descomposición de $-\frac{1}{2}$ AIC(k).

Tomando $\theta = \theta_k^*$ en (1.30) se tiene

$$\ell^*(\theta_k^*) = \ell^*(\theta^*) - \frac{n}{2} \|\theta_k^* - \theta^*\|^2 + R_n \; , \qquad (1.41)$$

donde $R_n \xrightarrow[n\to\infty]{P} 0$.

Sustituyendo (1.41) en (1.32), se obtiene

$$\ell^*(\theta) = \ell^*(\theta^*) - \frac{n}{2} \|\theta_k^* - \theta^*\|^2 - \frac{n}{2} \|\theta - \theta_k^*\|^2 + R_n \; .$$

Remplazando (1.30) en esta última ecuación y para n grande, se llega a

$$\|\theta - \theta^*\|^2 = \|\theta - \theta_k^*\|^2 + \|\theta_k^* - \theta^*\|^2 \; . \qquad (1.42)$$

Por otro lado, si se toma $\theta = \hat{\theta}_k(x_1, \ldots, x_n)$ en (1.36), se tiene

$$\ell(\hat{\theta}_k(x_1,\ldots,x_n)) = \ell(\hat{\theta}_K(x_1,\ldots,x_n))$$

$$- \frac{n}{2} \|\hat{\theta}_k(x_1,\ldots,x_n) - \hat{\theta}_K(x_1,\ldots,x_n)\|^2 + R_n ,$$

donde $R_n \xrightarrow[n\to\infty]{P} 0$.

Por tanto, se obtiene

$$\ell(\hat{\theta}_K(x_1,\ldots,x_n)) = \ell(\hat{\theta}_k(x_1,\ldots,x_n))$$

$$+ \frac{n}{2} \|\hat{\theta}_k(x_1,\ldots,x_n) - \hat{\theta}_K(x_1,\ldots,x_n)\|^2 - R_n .$$

(1.43)

Tomando $\theta^* = \hat{\theta}_K(x_1,\ldots,x_n)$, $\theta_k^* = \hat{\theta}_k(x_1,\ldots,x_n)$, $\theta = \theta_k^*$ en (1.42) y aplicando el resultado obtenido en (1.43), se tiene

$$\ell(\hat{\theta}_K(x_1,\ldots,x_n)) = \ell(\hat{\theta}_k(x_1,\ldots,x_n))$$

$$+\frac{n}{2}\left(\|\theta_k^*-\hat{\theta}_K(x_1,\ldots,x_n)\|^2-\|\theta_k^*-\hat{\theta}_k(x_1,\ldots,x_n)\|^2\right) - R_n,$$

de donde se obtiene

$$\ell(\hat{\theta}_k(x_1,\ldots,x_n)) = \ell(\hat{\theta}_K(x_1,\ldots,x_n))$$

$$-\frac{n}{2}\left(\|\theta_k^*-\hat{\theta}_K(x_1,\ldots,x_n)\|^2-\|\theta_k^*-\hat{\theta}_k(x_1,\ldots,x_n)\|^2\right) + R_n$$

$$=\ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{n}{2}\|(\theta_k^*-\theta^*)+(\theta^*-\hat{\theta}_K(x_1,\ldots,x_n))\|^2$$

$$+ \frac{n}{2} \|\theta_k^* - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 + R_n$$

$$\ell(\hat{\theta}_k(x_1,\ldots,x_n)) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{n}{2}\|\theta^* - \hat{\theta}_K(x_1,\ldots,x_n)\|^2$$

$$- \frac{n}{2}\|\theta_k^* - \theta^*\|^2 - n[\theta_k^* - \theta^*]J_*[\theta^* - \hat{\theta}_K(x_1,\ldots,x_n)]^t$$

$$+ \frac{n}{2}\|\theta_k^* - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 + R_n \ . \qquad (1.44)$$

Por otro lado, si se toma $\theta = \theta^*$ en (1.23) y utilizando la notación dada en (1.31), se llega a

$$\ell(\theta^*) = \ell(\hat{\theta}_K(x_1,\ldots,x_n)) - \frac{n}{2}\|\theta^* - \hat{\theta}_K(x_1,\ldots,x_n)\|^2 + R_n \ .$$

Despejando $\ell(\hat{\theta}_K(x_1,\ldots,x_n))$ y sustituyendo en (1.44), se tiene

$$\ell(\hat{\theta}_k(x_1,\ldots,x_n)) = \ell(\theta^*) - \frac{n}{2}\|\theta_k^* - \theta^*\|^2 + \frac{n}{2}\|\theta_k^* - \hat{\theta}_k(x_1,\ldots,x_n)\|^2$$

$$- n\,[\theta_k^* - \theta^*]\,J_*\,[\theta^* - \hat{\theta}_K(x_1,\ldots,x_n)]^t \ .$$

Sumando y restando $\ell^*(\theta^*)$ en la ecuación anterior y aplicando (1.41) para un n suficientemente grande, se obtiene

$$\ell(\hat{\theta}_k(x_1,\ldots,x_n)) = \ell^*(\theta_k^*) + \ell(\theta^*) - \ell^*(\theta^*)$$

$$-n\,[\theta_k^* - \theta^*]\,J_*\,[\theta^* - \hat{\theta}_K(x_1,\ldots,x_n)]^t$$

$$+ \frac{n}{2}\|\theta_k^* - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 \ . \qquad (1.45)$$

Aplicando (1.35) en (1.45) y aplicando a su vez el resultado obtenido en (1.40), se llega a

$$-\frac{1}{2}\,AIC(k) = \ell_n^*(k) + \left[\ell(\theta^*) - \ell^*(\theta^*)\right]$$

$$+ \left[-n\,[\theta_k^* - \theta^*]\,J_*\,[\theta^* - \hat{\theta}_k(x_1,\ldots,x_n)]^t\right.$$

$$\left. + \frac{1}{2}\left(n\,\|\theta_k^* - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 - k\right)\right] . \qquad (1.46)$$

Desde que la diferencia $[\ell(\theta^*) - \ell^*(\theta^*)]$ no depende de k, se le denomina Error Común. Como el tercer sumando depende de k, a este término se le denomina Error Individual.

Luego (1.46) puede ser expresado de la siguiente manera

$$-\frac{1}{2}\,AIC(k) = \begin{pmatrix}\text{Media del logaritmo}\\ \text{de verosimilitud}\\ \text{esperado}\end{pmatrix} + \begin{pmatrix}\text{Error}\\ \text{Común}\end{pmatrix} + \begin{pmatrix}\text{Error}\\ \text{Individual}\end{pmatrix}$$

Los términos de error dados en (1.46) pueden ser estudiados en las siguientes situaciones:

i) $\theta_k^* = (\theta_1^*,\ldots,\theta_k^*,0,\ldots,0)$ Y J_* ES LA MATRIZ IDENTIDAD (I_k)

Error Común $= \ell(\theta^*) - \ell^*(\theta^*)$

luego su esperanza

$$E_{X/\theta^*}(\text{Error Común}) = E_{X/\theta^*}\left[\ell(\theta^*) - \ell^*(\theta^*)\right]$$

$$= \ell^*(\theta^*) - \ell^*(\theta^*) \qquad \text{de (1.4)}$$

$$= 0 .$$

Luego el error común se distribuye asintóticamente como una

variable aleatoria de media cero.

En cuanto al error individual se tiene

$$\text{Error Individual} = -\, n\,[\theta_k^{\bullet} - \theta^{\bullet}]\, J_{\bullet}\, [\theta^{\bullet} - \hat{\theta}_K(x_1,\ldots,x_n)]^t + \frac{1}{2}\left(n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 - k\right) . \tag{1.47}$$

De (1.33) se tiene $n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2$ se distribuye según una Ji Cuadrado con k grados de libertad, luego

$$E_{X/\theta_k^{\bullet}}\left(\frac{1}{2}\left[n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 - k\right]\right) =$$

$$= \frac{1}{2}\left[E_{X/\theta_k^{\bullet}}\left(n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2\right) - k\right] = 0 \tag{1.48}$$

$$V_{X/\theta_k^{\bullet}}\left(\frac{1}{2}\,n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2 - k\right) =$$

$$= \frac{1}{4}\,V_{X/\theta_k^{\bullet}}\left(n\,\|\theta_k^{\bullet} - \hat{\theta}_k(x_1,\ldots,x_n)\|^2\right) = \frac{k}{2} . \tag{1.49}$$

Por otro lado se tiene que el primer componente del error individual se puede expresar de la siguiente manera

$$-n[\theta_k^{\bullet}-\theta^{\bullet}]J_{\bullet}[\theta^{\bullet}-\hat{\theta}_K(x_1,\ldots,x_n)]^t = n\sum_{j=k+1}^{K}\theta_j^{\bullet}[\theta_j^{\bullet}-\hat{\theta}_j(x_1,\ldots,x_n)] ,$$

donde $J_{\bullet} = I_K$.

De donde aplicando (1.3), es decir, $\hat{\theta}_K(x_1,\ldots,x_n)$ se distribuye según $N\left(\theta^{\bullet}, \frac{1}{n}\,I_K\right)$, se obtiene

$$E_{X/\theta^*}\left[n\sum_{j=k+1}^{K}\theta_j^*[\theta_j^* - \hat{\theta}_j(x_1,\ldots,x_n)]\right] =$$

$$= n\sum_{j=k+1}^{K}\theta_j^*\left[\theta_j^* - E_{X/\theta^*}\left(\hat{\theta}_j(x_1,\ldots,x_n)\right)\right] = 0 \qquad (1.50)$$

$$V_{X/\theta^*}\left[n\sum_{j=k+1}^{K}\theta_j^*[\theta_j^* - \hat{\theta}_j(x_1,\ldots,x_n)]\right] =$$

$$= n^2\sum_{j=k+1}^{K}(\theta_j^*)^2\, V_{X/\theta^*}\left(\hat{\theta}_j(x_1,\ldots,x_n)\right)$$

$$= n^2\sum_{j=k+1}^{K}(\theta_j^*)^2\,\frac{1}{n}$$

$$= n\sum_{j=k+1}^{K}(\theta_j^*)^2 \,. \qquad (1.51)$$

Luego de (1.48) y (1.50) se tiene que el error individual dado en (1.47) se distribuye asintóticamente como una variable aleatoria de media cero.

De (1.49) y (1.51) se tiene que la varianza del error individual dado en (1.47) es

$$n\sum_{j=k+1}^{K}(\theta_j^*)^2 + \frac{k}{2} \,.$$

Los resultados presentados en este primer caso fueron obtenidos por Sakamoto, Ishiguro y Kitagawa (1986). Los resultados que presentamos a continuación los obtuvimos en el estudio del

comportamiento de los términos de error en otras situaciones.

ii) $\theta_k^{\bullet} = (\theta_1^{\bullet}, \ldots, \theta_k^{\bullet}, 0, \ldots, 0)$ Y $J_{\bullet}$ ES UNA MATRIZ DEFINIDA POSITIVA

$E_{X/\theta^{\bullet}}$ (Error Común) = 0 .

Por tanto el error común se distribuye asintóticamente como una variable aleatoria de media cero.

De otro lado se tiene para el error individual lo siguiente:

$$- n \, [\theta_k^{\bullet} - \theta^{\bullet}] \, J_{\bullet} \, [\theta^{\bullet} - \hat{\theta}_K(x_1, \ldots, x_n)]^t =$$

$$= n \sum_{l=1}^{K} [\theta_l^{\bullet} - \hat{\theta}_l(x_1, \ldots, x_n)] \sum_{j=K+1}^{K} \theta_j^{\bullet} \left[-E_{X_l/\theta^{\bullet}} \left(\frac{d^2 \log f(x_l/\theta)}{d\theta_j \, d\theta_l} \right)_{\theta^{\bullet}} \right]$$

$$= n \sum_{l=1}^{K} [\theta_l^{\bullet} - \hat{\theta}_l(x_1, \ldots, x_n)] \, C(l) \, ,$$

donde $C(l) = \sum_{j=K+1}^{K} \theta_j^{\bullet} \left[- E_{X_l/\theta^{\bullet}} \left(\frac{d^2 \log f(x_l/\theta)}{d\theta_j \, d\theta_l} \right)_{\theta^{\bullet}} \right]$ y

$$J_{\bullet} = -E_{X_i/\theta^{\bullet}}\left[\begin{array}{ccc} \dfrac{d^2 \log f(x_i/\theta)}{d\theta_1^2} & \cdots & \dfrac{d^2 \log f(x_i/\theta)}{d\theta_1\, d\theta_K} \\ \vdots & & \vdots \\ \dfrac{d^2 \log f(x_i/\theta)}{d\theta_2\, d\theta_1} \quad \dfrac{d^2 \log f(x_i/\theta)}{d\theta_2^2} & & \\ \vdots & & \vdots \\ \dfrac{d^2 \log f(x_i/\theta)}{d\theta_K\, d\theta_1} & \cdots & \dfrac{d^2 \log f(x_i/\theta)}{d\theta_K^2} \end{array}\right]_{\theta^{\bullet}} \tag{1.52}$$

De (1.3), $\hat{\theta}_K(x_1,\ldots,x_n)$ se distribuye según $N\left(\theta^{\bullet}, \frac{1}{n} J_{\bullet}^{-1}\right)$, se tiene

$$E_{X/\theta^{\bullet}}\left(-n\,[\theta_K^{\bullet} - \theta^{\bullet}]\; J_{\bullet}\; [\theta^{\bullet} - \hat{\theta}_K(x_1,\ldots,x_n)]^t\right) = 0 \tag{1.53}$$

y

$$V_{X/\theta^{\bullet}}\left(-n\,[\theta_K^{\bullet} - \theta^{\bullet}]\; J_{\bullet}\; [\theta^{\bullet} - \hat{\theta}_K(x_1,\ldots,x_n)]^t\right) =$$

$$= n\left(\sum_{l=1}^{K} [C(l)]^2\, a_{ll} + \sum_{l_1 \neq l_2} C(l_1)\, C(l_2)\, a_{l_1 l_2}\right), \tag{1.54}$$

donde: a_{ll} es el (l,l)-ésimo elemento de la matriz $J_{\bullet}^{-1}$

$a_{l_1 l_2}$ es el (l_1,l_2)-ésimo elemento de la matriz $J_{\bullet}^{-1}$.

De (1.48) y (1.53) tenemos que el error individual se distribuye asintóticamente como una variable aleatoria de media cero.

De (1.49) y (1.54) se tiene que la varianza del error individual es

$$n\left(\sum_{i=1}^{K}[C(i)]^2 a_{ii} + \sum_{i_1 \neq i_2} C(i_1)\, C(i_2)\, a_{i_1 i_2}\right) + \frac{k}{2}\,.$$

iii) $\theta_k^* = (\theta_1^*, \ldots, \theta_k^*, O_{k+1}, \ldots, O_K)$ Y $J_* = I_K$

El término correspondiente al error común se distribuye asintóticamente como una variable aleatoria de media cero.

En cuanto al error individual se tiene lo siguiente:

$$-n\,[\theta_k^* - \theta^*]\, J_*\, [\theta^* - \hat{\theta}_K(x_1, \ldots, x_n)]^t =$$

$$= -n \sum_{j=k+1}^{K} (O_j - \theta_j^*)[\theta_j^* - \hat{\theta}_j(x_1, \ldots, x_n)]\,,$$

luego de (1.3), $\hat{\theta}_K(x_1, \ldots, x_n)$ se distribuye según $N\left(\theta^*, \frac{1}{n} I_K\right)$, se obtiene que

$$E_{X/\theta^*}\left(-n[\theta_k^* - \theta^*]\, J_*\, [\theta^* - \hat{\theta}_K(x_1, \ldots, x_n)]^t\right) = 0 \qquad (1.55)$$

y

$$V_{X/\theta^*}\left(-n\,[\theta_k^* - \theta^*]\, J_*\, [\theta^* - \hat{\theta}_K(x_1, \ldots, x_n)]^t\right) = n \sum_{j=k+1}^{K} (O_j - \theta_j^*)^2\,. \qquad (1.56)$$

De (1.48) y (1.55) se obtiene que el error individual se distribuye asintóticamente como una variable aleatoria de media cero.

De (1.49) y (1.56) se tiene que la varianza del error individual es

$$n \sum_{j=k+1}^{K} (O_j - \theta_j^*)^2 + \frac{k}{2} .$$

iv) $\theta_k^* = (\theta_1^*, \ldots, \theta_k^*, O_{k+1}, \ldots, O_K)$ Y J_* ES UNA MATRIZ DEFINIDA POSITIVA.

El error común se distribuye asintóticamente como una variable aleatoria de media cero.

En cuanto al error individual se tiene lo siguiente

$$-n\,[\theta_k^* - \theta^*]\; J_*\; [\theta^* - \hat{\theta}_K(x_1, \ldots, x_n)]^t =$$

$$= -n \sum_{i=1}^{K} [\theta_i^* - \hat{\theta}_i(x_1, \ldots, x_n)] \sum_{j=k+1}^{K} (O_j - \theta_j^*) \left[-E_{X_i/\theta^*} \left(\frac{d^2 \log f(x_i/\theta)}{d\theta_j\, d\theta_i} \right)_{\theta^*} \right]$$

$$= -n \sum_{i=1}^{k} [\theta_i^* - \hat{\theta}_i(x_1, \ldots, x_n)]\; D(i) ,$$

donde

$$D(i) = \sum_{j=k+1}^{K} (O_j - \theta_j^*) \left[- E_{X_i/\theta^*} \left(\frac{d^2 \log f(x_i/\theta)}{d\theta_j\, d\theta_i} \right)_{\theta^*} \right]$$

y J_* está definida en (1.52).

De (1.3), $\hat{\theta}_K(x_1, \ldots, x_n)$ se distribuye segun $N\left(\theta^*, \frac{1}{n} J_*^{-1}\right)$, se tiene

$$E_{X/\theta^*} \left(-n[\theta_k^* - \theta^*]\; J_*\; [\theta^* - \hat{\theta}_K(x_1, \ldots, x_n)]^t \right) = 0 \qquad (1.57)$$

y

$$V_{X/\theta^{\bullet}}\left(-n[\theta_k^{\bullet} - \theta^{\bullet}]\, J_{\bullet}\, [\theta^{\bullet} - \hat{\theta}_K(x_1,\ldots,x_n)]^t\right) =$$

$$= n\left(\sum_{i=1}^{K} [D(i)]^2\, a_{ii} + \sum_{i_1 \neq i_2} D(i_1)\, D(i_2)\, a_{i_1 i_2}\right) \qquad (1.58)$$

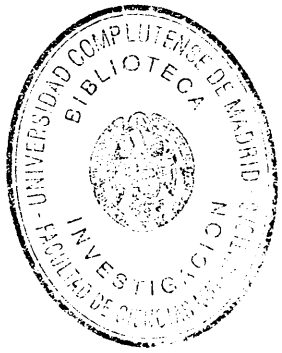
donde: a_{ii} es el (i,i)-ésimo elemento de la matriz $J_{\bullet}^{-1}$

$a_{i_1 i_2}$ es el (i_1,i_2)-ésimo elemento de la matriz $J_{\bullet}^{-1}$.

De (1.48) y (1.57) se tiene que el error individual se distribuye asintóticamente como una variable aleatoria de media cero.

Y de (1.49) y (1.58) se tiene que la varianza del error individual es

$$n\left(\sum_{i=1}^{K} [D(i)]^2\, a_{ii} + \sum_{i_1 \neq i_2} D(i_1)\, D(i_2)\, a_{i_1 i_2}\right) + \frac{k}{2} .$$



CAPITULO II

"EL CRITERIO DE INFORMACION DE AKAIKE PARA EL ANALISIS DE TABLAS DE CONTINGENCIA"

II.0 SUMARIO

El objetivo de este capitulo es el análisis de tablas de contingencia.

Es por ello que se presenta un método basado en el criterio MAIC para el estudio de independencia, homogeneidad e interacción entre los tres factores de clasificación, los cuales constituyen los puntos principales en el análisis de una tabla de contingencia.

En el estudio de la independencia de los tres factores de clasificación es interesante estudiar cuales son las causas que originan la dependencia o independencia de los factores de clasificación, este análisis detallado se puede realizar siguiendo las pautas sobre independencia dadas por Kullback en 1959.

Para el estudio de la homogeneidad de tablas de contingencia de dos factores de clasificación es importante examinar cuales son las causas que originan la homogeneidad, este análisis se puede realizar siguiendo las pautas dadas por Kullback sobre homogeneidad, que están basadas en el análisis de la homogeneidad parcial y la homogeneidad condicional.

Como el análisis de la Homogeneidad Condicional (D/C) nos lleva al análisis de las interacciones entre los factores, ésto permite realizar un análisis más detallado de la tabla de contingencia.

También es de gran interés el estudio de la interacción entre los tres factores de clasificación. Kullback (1959) sugiere, para ello, la descomposición algebraica del componente de Homogeneidad Condicional (D/C). El estadístico que surge para examinar dicha interacción sigue una función densidad de Wittaker (ver Inga (1990)), pero su cálculo es engorroso.

Bishop (1975) sugiere realizar este análisis mediante el uso de modelos log lineal y efectuar el ajuste del modelo mediante el estadístico G, es decir, dos veces la Información de Kullback-Leibler, el cual sigue una distribución asintótica Ji-Cuadrado. Por lo tanto, fijando previamente un nivel de

significación se podrá tomar una decisión acerca del modelo.

En el presente capítulo se presenta un método alternativo para efectuar este análisis y los anteriormente mencionados.

Para poder tomar una decisión se propone calcular el estadístico AIC (Criterio de Información de Akaike) para cada modelo asociado a una hipótesis y luego utilizar el criterio de selección MAIC (Mínimo AIC) para seleccionar el mejor modelo, el cual indicará la hipótesis a aceptar (ver Akaike (1973), Akaike (1977), Sakamoto, Ishiguro y Kitagawa (1986)).

Este método, como se presentó en el Capítulo I, permite seleccionar el mejor modelo sin necesidad de fijar un nivel de significación. Esta es una ventaja sobre los métodos tradicionales.

En la Sección II.1 se mostrará el planteamiento del problema para el análisis de tablas de contingencia, así como la notación que se utilizará en el capítulo. En las secciones II.2 a II.4 se presenta la metodología para el análisis de independencia, homogeneidad e interacción de los factores de clasificación.

La metodología desarrollada en este capítulo para el análisis de tablas de contingencia de tres factores de clasificación basado en el AIC, generalizan los resultados obtenidos por Sakamoto, Ishiguro y Kitagawa (1986), Sakamoto, Akaike(1978) y Sakamoto (1982) sobre el análisis de independencia y homogeneidad en tablas de contingencia ya que estos únicamente se ocupan de dos factores de clasificación.

Una ventaja importante que posee el método que se propone es que no sólo da los estadísticos AIC para todas las hipótesis planteadas en una tabla de tres factores, si no que además proporciona las relaciones que existen entre las diferentes hipótesis, lo cual permite realizar un análisis más detallado de los datos, pues se pueden detectar las verdaderas causas de la independencia, homogeneidad e interacción de los factores de clasificación.

II.1 PLANTEAMIENTO DEL PROBLEMA PARA EL ANALISIS DE TABLAS DE CONTINGENCIA

Sea (W,Y,Z) una variable aleatoria tridimensional con función de distribución conjunta $G(w,y,z)$. En el modelo de independencia se trata, en base a una muestra aleatoria simple $(w_1,y_1,z_1),\ldots,(w_n,y_n,z_n)$, de contrastar

$$G(w,y,z) = G_1(w)\, G_2(y)\, G_3(z) \qquad \forall\, w,y,z$$

donde G_1, G_2 y G_3 son las funciones de distribución marginal correspondientes a W, Y y Z respectivamente.

Si se divide el recorrido de la variable W en r conjuntos de Borel $F_1,\ldots,F_r$, el de Y en c conjuntos de Borel $C_1,\ldots,C_c$ y el de la Z en d conjuntos de Borel $D_1,\ldots,D_d$, se escribe

$$p_{ijk} = P_G\left(F_i \times C_j \times D_k\right), \quad p_{i..} = P_{G_1}\left(F_i\right), \quad p_{.j.} = P_{G_2}\left(C_j\right),$$

$$p_{..k} = P_{G_3}\left(D_k\right),$$

y se define la variable aleatoria X_{ijk}, como el número de valores de la muestra que caen en $F_i \times C_j \times D_k$. Los datos muestrales de la población (W,Y,Z) se pueden representar mediante una tabla de contingencia de tres factores de clasificación $(r \times c \times d)$, con las probabilidades y categorías sujetas a la distribución multinomial; es decir

$$P\left(X_{111} = x_{111},\ldots, X_{rcd} = x_{rcd}\right) = \frac{N!}{x_{111}!\ldots x_{rcd}!}\, p_{111}^{x_{111}} \cdots p_{rcd}^{x_{rcd}}.$$

Así, analizar si las variables aleatorias W, Y, Z son independientes, es equivalente a estudiar en la tabla de contingencia, si los tres factores de clasificación son independientes. Es decir, el problema consiste en contrastar

$H_0 : p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k}$, $i=1,\ldots,r$ $j=1,\ldots,c$ $k=1,\ldots,d$

frente a

$H_1 : p_{ijk} \neq p_{i..}\, p_{.j.}\, p_{..k}$, para algún (i,j,k) .

En lo que se refiere a la homogeneidad de varias muestras: sean r muestras aleatorias independientes $X^{(1)},\ldots,X^{(r)}$ de tamaño $N_1,\ldots,N_r$.

$X^{(1)} = \left(\left(y_1^{(1)}, z_1^{(1)}\right),\ldots,\left(y_{N_1}^{(1)}, z_{N_1}^{(1)}\right)\right) \equiv$ (Inicialmente se supone que su procedencia es de una población con funcion de distribución G_1 desconocida)

.

.

.

$X^{(r)} = \left(\left(y_1^{(r)}, z_1^{(r)}\right),\ldots,\left(y_{N_r}^{(r)}, z_{N_r}^{(r)}\right)\right) \equiv$ (Inicialmente se supone que su procedencia es de una población con función de distribucion G_r desconocida)

y se desea saber si éstas proceden de una misma población con función de distribución G desconocida .

Se considera una partición $\{E_{jk}\}$, $j=1,\ldots,c$ $k=1,\ldots,d$, del soporte de G y se definen las r variables aleatorias (c x d)-dimensionales siguientes: $\left(X_{111},\ldots,X_{1jk},\ldots,X_{1cd}\right),\ldots\ldots,$

$\left(X_{i11},\ldots,X_{ijk},\ldots,X_{icd}\right),\ldots,\left(X_{r11},\ldots,X_{rjk},\ldots,X_{rcd}\right)$, donde,

X_{ijk} = Número de elementos de la i-ésima muestra que caen en E_{jk}.

Evidentemente,

$$P\left(X_{111}=x_{111},\ldots,X_{1cd}=x_{1cd}\right) = \frac{N_1!}{x_{111}!\ldots x_{1cd}!}\, p_{11/1}^{x_{111}}\ldots p_{cd/1}^{x_{1cd}}$$

donde $p_{jk/1} = P_{G_1}\left(E_{jk}\right)$, $j=1,\ldots,c$, $k=1,\ldots,d$

.

.

.

$$P\left(X_{r11}=x_{r11},\ldots,X_{rcd}=x_{rcd}\right) = \frac{N_r!}{x_{r11}!\ldots x_{rcd}!}\, p_{11/r}^{x_{111}}\ldots p_{cd/r}^{x_{rcd}}$$

donde $p_{jk/r} = P_{G_r}\left(E_{jk}\right)$, $j=1,\ldots,c$, $k=1,\ldots,d$.

El problema de homogeneidad de las r muestras quedará resuelto al contrastar H_0 frente a H_1, donde,

H_0 : $p_{jk/i} = p_{.jk}$, para todo $i=1,\ldots,r$, $j=1,\ldots,c$, $k=1,\ldots,d$

$$p_{.jk} = P_G\left(E_{jk}\right)$$

frente a

H_1 : $p_{jk/i} \neq p_{.jk}$, para algún (i,j,k) .

Obsérvese que cada muestra $X^{(i)}$, $i = 1,\ldots,r$ se puede representar en una tabla de contingencia bidimensional (c x d).

Por lo tanto, este contraste de hipótesis es equivalente a contrastar que las r muestras independientes de una tabla (c x d) son homogéneas.

En lo que sigue se utilizará la siguiente notación :

$X_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{d} X_{ijk}$, $X_{ij.} = \sum_{k=1}^{d} X_{ijk}$ y de forma análoga $X_{.j.}, X_{..k}$,

$X_{i.k}$ y $X_{.jk}$. Además observese que

$$\sum_{i=1}^{r} X_{i..} = \sum_{j=1}^{c} X_{.j.} = \sum_{k=1}^{d} X_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij.} = \sum_{i=1}^{r}\sum_{k=1}^{d} X_{i.k} = \sum_{j=1}^{c}\sum_{k=1}^{d} X_{.jk} = N .$$

II.2 TEST DE INDEPENDENCIA ENTRE LOS TRES FACTORES DE CLASIFICACION

Considerese una tabla de contingencia de tres factores de clasificación, en la cual se quiere contrastar la hipótesis nula

$$H_0 : p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k} , \quad i=1,\ldots,r \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d$$

frente a

$$H_1 : p_{ijk} \neq p_{i..}\, p_{.j.}\, p_{..k} , \quad \text{para al menos un } (i,j,k) .$$

(2.1)

Para efectuar el contraste de hipótesis dado en (2.1) en esta memoria se propone el siguiente procedimiento basado en el Criterio de Información de Akaike.

El contraste de hipótesis dado en (2.1) se puede formular mediante la comparación de los modelos MODELO(0) y MODELO(1) los cuales se obtienen de la siguiente manera:

Suponiendo que los tres factores de clasificación son independientes se puede construir el modelo,

$$\text{MODELO(0)}: \ p_{ijk} = \theta_{i..}\, \theta_{.j.}\, \theta_{..k} , \ i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d$$

donde

$$\theta_{i..} = \sum_{j=1}^{c} \sum_{k=1}^{d} p_{ijk} , \qquad \theta_{.j.} = \sum_{i=1}^{r} \sum_{k=1}^{d} p_{ijk} , \qquad \theta_{..k} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ijk} ,$$

$$\sum_{i=1}^{r} \theta_{i..} = 1 , \qquad \sum_{j=1}^{c} \theta_{.j.} = 1 , \qquad \sum_{k=1}^{d} \theta_{..k} = 1 .$$

(2.2)

El otro modelo se obtiene suponiendo que los tres factores de clasificación no son independientes, y es.

MODELO(1): $p_{ijk} = \theta_{ijk}$, $\quad i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d$

donde, $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{ijk} = 1$.

(2.3)

Por lo expuesto en la Sección II.1 se tiene

$$P\left(X_{ijk} = x_{ijk},\ i=1,\ldots,r,\ j=1,\ldots,c,\ k=1,\ldots,d\right) =$$

$$= \frac{N!}{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{d} x_{ijk}!}\ \prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{d}\left(p_{ijk}\right)^{x_{ijk}} ,$$

luego el logaritmo de la función de verosimilitud será

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\ \log p_{ijk} \ . \qquad (2.4)$$

donde $\quad K_1 = \log\left(\frac{N!}{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{d} x_{ijk}!}\right)$.

En base a esta expresión, se obtienen los estadísticos AIC (Criterio de Información de Akaike) de los modelos (2.2) y (2.3). Estos se presentan en el siguiente teorema.

TEOREMA 2.1

Supuesto que los tres factores de clasificación son independientes (MODELO(0)), entonces

$$AIC(0) = -2\left(K_1 + \sum_{i=1}^{r} x_{i\cdot\cdot}\ \log x_{i\cdot\cdot} + \sum_{j=1}^{c} x_{\cdot j\cdot}\ \log x_{\cdot j\cdot} + \right.$$

$$\left. + \sum_{k=1}^{d} x_{\cdot\cdot k}\ \log x_{\cdot\cdot k} - 3N \log N\right) + 2(r + c + d - 3)$$

Supuesto que los tres factores no son independientes (MODELO(1)), entonces

$$AIC(1) = -2\left[K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right] + 2(rcd - 1) .$$

Demostración

Bajo el MODELO(0) dado en (2.2), se tiene que el logaritmo de la función de verosimilitud dado en (2.4) será,

$$\ell = K_1 + \left[\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log (\theta_{i..}\ \theta_{.j.}\ \theta_{..k})\right] . \qquad (2.5)$$

Además sabemos que

$$\sum_{i=1}^{r} \theta_{i..} = 1 \quad \Rightarrow \quad \theta_{r..} = 1 - \sum_{i=1}^{r-1} \theta_{i..}$$

$$\sum_{j=1}^{c} \theta_{.j.} = 1 \quad \Rightarrow \quad \theta_{.c.} = 1 - \sum_{j=1}^{c-1} \theta_{.j.}$$

$$\sum_{k=1}^{d} \theta_{..k} = 1 \quad \Rightarrow \quad \theta_{..d} = 1 - \sum_{k=1}^{d-1} \theta_{..k} .$$

Por tanto

$$\ell = K_1 + \sum_{i=1}^{r-1} \log \theta_{i..} \left[\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\right] + \log \left[1 - \sum_{i=1}^{r-1} \theta_{i..}\right]\left[\sum_{j=1}^{c}\sum_{k=1}^{d} x_{rjk}\right]$$

$$+ \sum_{j=1}^{c-1} \log \theta_{.j.} \left[\sum_{i=1}^{r}\sum_{k=1}^{d} x_{ijk}\right] + \log \left[1 - \sum_{j=1}^{c-1} \theta_{.j.}\right]\left[\sum_{i=1}^{r}\sum_{k=1}^{d} x_{ick}\right] +$$

$$+ \sum_{k=1}^{d-1} \log \theta_{..k} \left[\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ijk}\right] + \log \left[1 - \sum_{k=1}^{d-1} \theta_{..k}\right]\left[\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ijd}\right] , \qquad (2.6)$$

a partir de esta última ecuación obtenemos los estimadores de máxima verosimilitud de $\theta_{i..}$, $\theta_{.j.}$, $\theta_{..k}$ que vienen dada por:

$$\hat{\theta}_{i..} = \frac{x_{i..}}{N}, \qquad i = 1,\ldots,d \tag{2.7}$$

$$\hat{\theta}_{.j.} = \frac{x_{.j.}}{N}, \qquad j = 1,\ldots,c \tag{2.8}$$

$$\hat{\theta}_{..k} = \frac{x_{..k}}{N}, \qquad k = 1,\ldots,d . \tag{2.9}$$

Sustituyendo los valores $\hat{\theta}_{i..}$, $\hat{\theta}_{.j.}$, $\hat{\theta}_{..k}$ en (2.5) se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log\left(\frac{x_{i..}}{N}\,\frac{x_{.j.}}{N}\,\frac{x_{..k}}{N}\right),$$

luego el AIC del MODELO(0) siguiendo lo expuesto en la sección I.2 del capitulo I será

$$AIC(0) = -2\left(K_1 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \right.$$

$$\left. + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 3N \log N\right) + 2[(r-1)+(c-1)+(d-1)] . \tag{2.10}$$

Bajo el MODELO(1) dado en (2.3) se tiene que el logaritmo de la de la función de verosimilitud dada en (2.4) es

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{ijk} . \tag{2.11}$$

Y como $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{ijk} = 1$, entonces

$$\theta_{rcd} = 1 - \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d-1} \theta_{ijk} - \sum_{i=1}^{r}\sum_{j=1}^{c-1} \theta_{ijd} .$$

Aplicando esta relación en (2.11) y reordenado se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d-1} x_{ijk} \log \theta_{ijk} + \sum_{i=1}^{r}\sum_{j=1}^{c-1} x_{ijd} \log \theta_{ijd} +$$

$$+ x_{rcd} \log \left(1 - \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d-1} \theta_{ijk} - \sum_{i=1}^{r}\sum_{j=1}^{c-1} \theta_{ijd} \right) . \qquad (2.12)$$

Derivando (2.12) con respecto a θ_{ijk} e igualando a cero, para $i=1,\ldots,r$, $j=1,\ldots,r$, $k=1,\ldots,d$ se obtiene,

$$\hat{\theta}_{ijk} = \frac{x_{ijk}}{N} , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d .$$

Sustituyendo $\hat{\theta}_{ijk}$ en (2.11) se llega a

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \left(\frac{x_{ijk}}{N} \right) ,$$

luego el AIC del MODELO(1) será

$$AIC(1) = -2 \left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N \right) + 2(rcd - 1). \qquad (2.13)$$

OBSERVACION

Se seleccionará como mejor modelo entre el MODELO(0) y el MODELO(1), al modelo correspondiente al mínimo AIC (MAIC).

Los AIC de estos modelos fueron dados en (2.10) y (2.13). Así aplicando el criterio MAIC a estos modelos se tiene:

Si AIC(0) < AIC(1) , entonces el MODELO(0) se elegirá como el mejor modelo, lo cual indicará que los tres factores son independientes.

Si AIC(1) < AIC(0) , entonces el MODELO(1) se elegirá como

mejor modelo, lo cual indicará que los tres factores no son independientes.

Obsérvese que, a efectos de comparación se puede ignorar la constante $-2\ K_1$ en (2.10) y (2.13). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{*}(0)$ y $AIC^{*}(1)$, donde

$$AIC^{*}(0) = -2\left(\sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \sum_{k=1}^{d} x_{..k} \log x_{..k} \right.$$

$$\left. -3\ N \log N \right) + 2(r + c + d - 3)$$

$$AIC^{*}(1) = -2\left(\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N \right) + 2(rcd - 1) \ .$$

(2.14)

Es decir, siguiendo el criterio MAIC (Mínimo AIC) se adopta el modelo de independencia de los tres factores si AIC(0) es menor que AIC(1), en caso contrario se adopta el MODELO(1), es decir, que los tres factores son dependientes.

Esto indica que se escoge el modelo por el signo de [AIC(1) - AIC(0)], es decir, si es positivo deberemos elegir el MODELO(0), y si es negativo el MODELO(1)

La siguiente proposición pone de manifiesto la relación existente entre la diferencia [AIC(1) - AIC(0)] y la cantidad de información de Kullback-Leibler.

PROPOSICION 2.1

Bajo las hipótesis dadas en (2.1) se tiene

$$AIC(1) - AIC(0) = -2\ \hat{I}\left(H_1 : H_0(F \times C \times D)\right) + 2(rcd-r-c-d+2)$$

donde $\hat{I}(\ )$ es el estadístico de Kullback-Leibler para contrastar la hipótesis nula de independencia entre los tres factores

Demostración

De las expresiones obtenidas para el AIC(1) y AIC(0), se tiene

$$AIC(1) - AIC(0) =$$

$$= -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1) +$$

$$+2\left(K_1 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \right.$$

$$\left. + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 3N \log N\right) - 2(r + c + d - 3)$$

$$= -2\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{N^2 x_{ijk}}{x_{i..} x_{.j.} x_{..k}}\right) + 2(rcd-r-c-d+2)$$

$$= -2\,\hat{I}(H_1 : H_0(F \times C \times D)) + 2(rcd - r - c - d + 2)$$

donde $2\,\hat{I}\left(H_1 : H_0(F \times C \times D)\right)$ es la cantidad de información de Kullback-Leibler, donde se han sustituido los parámetros por sus correspondientes estimadores de máxima verosimilitud en $H_0(F \times C \times D)$.

Además $2\,\hat{I}\left(H_1 : H_0(F \times C \times D)\right)$ sigue, bajo $H_0(F \times C \times D)$, asintóticamente una distribución Ji-Cuadrado con (rcd-r-c-d+2) grados de libertad; y por tanto puede ser usada para contrastar dicha hipótesis.

∎

A continuacion se analiza este mismo problema de

independencia entre los tres factores desde otra perspectiva.

Los tres factores son independientes si y solamente si el factor Fila es independiente del par (Columna, Profundidad) y el factor Columna es independiente del factor Profundidad, ya que,

$$p_{ijk} = p_{i..}\, p_{.jk} \quad y \quad p_{.jk} = p_{.j.}\, p_{..k} \Leftrightarrow p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k} .$$

En lo que sigue utilizaremos la notación

$$H_0(F \times C \times D) \Leftrightarrow H_0(F \times CD) \cap H_0(C \times D) .$$

Para analizar la independencia entre el factor Fila y el par (Columna, Profundidad), es decir $H_0(F \times (C,D))$, se efectúa el siguiente contraste de hipótesis,

$$H_0 : \; p_{ijk} = p_{i..}\, p_{.jk} , \quad i=1,\ldots,r \;,\; j=1,\ldots,c \;,\; k=1,\ldots,d$$

$$\sum_{i=1}^{r} p_{i..} = \sum_{j=1}^{c} \sum_{k=1}^{d} p_{.jk} = 1$$

frente a

$$H_1 : \; p_{ijk} \neq p_{i..}\, p_{.jk} , \quad \text{para por lo menos algún } (i,j,k) .$$

(2.15)

De manera similar al caso anterior, se efectúa el contraste de hipótesis a través del análisis de los siguientes modelos:

$$\text{MODELO(0)}: \quad p_{ijk} = \theta_{i..}\, \theta_{.jk}$$

donde

$$\theta_{i..} = \sum_{j=1}^{c} \sum_{k=1}^{d} p_{ijk} \;,\quad \theta_{.jk} = \sum_{i=1}^{r} p_{ijk} \;,\quad \sum_{i=1}^{r} \theta_{i..} = 1 \;,\quad \sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1 .$$

(2.16)

Obsérvese que el modelo está construido bajo la suposición

de que el factor Fila es independiente del par (Columna, Profundidad).

MODELO(1): $p_{ijk} = \theta_{ijk}$, $i=1,\ldots,r$, $j=1,\ldots,c$, $k=1,\ldots,d$

donde $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{ijk} = 1$

(2.17)

en este caso se supone que el factor Fila no es independiente del par (Columna, Profundidad).

Los estadísticos AIC se presentan a continuación

TEOREMA 2.2

Supuesto que el factor Fila es independiente del par (Columna, Profundidad) (MODELO(0)), entonces

$$AIC(0) = -2\left(K_1 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - 2N \log N\right) + 2(r + cd - 2) .$$

Supuesto que el factor Fila no es independiente del par (Columna, Profundidad) (MODELO(1)), entonces

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1).$$

Demostración

Bajo el MODELO(0) dado en (2.16) se tiene que el logaritmo de la función de verosimilitud dado en (2.4) será

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log (\theta_{i..}\ \theta_{.jk}) \qquad (2.18)$$

y como

$$\sum_{i=1}^{r} \theta_{i..} = 1 \quad \Rightarrow \quad \theta_{r..} = 1 - \sum_{i=1}^{r-1} \theta_{i..}$$

$$\sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1 \quad \Rightarrow \quad \theta_{.cd} = 1 - \sum_{j=1}^{c} \sum_{k=1}^{d-1} \theta_{.jk} - \sum_{j=1}^{c-1} \theta_{.jd} ,$$

luego el logaritmo de la función de verosimilitud será

$$\ell = K_1 + \sum_{i=1}^{r-1} x_{i..} \log x_{i..} + x_{r..} \log \left(1 - \sum_{i=1}^{r-1} \theta_{i..} \right) +$$

$$+ \sum_{j=1}^{c} \sum_{k=1}^{d-1} x_{.jk} \log \theta_{.jk} + \sum_{j=1}^{c-1} x_{.jd} \log \theta_{.jd} +$$

$$+ x_{.cd} \log \left(1 - \sum_{j=1}^{c} \sum_{k=1}^{d-1} \theta_{.jk} - \sum_{j=1}^{c-1} \theta_{.jd} \right) . \qquad (2.19)$$

Siendo inmediato obtener los estimadores de máxima verosimilitud,

$$\hat{\theta}_{i..} = \frac{x_{i..}}{N} , \qquad i=1,\ldots,r$$

$$\hat{\theta}_{.jk} = \frac{x_{.jk}}{N} , \qquad j=1,\ldots,c \ , \ k=1,\ldots,d .$$

Sustituyendo los valores $\hat{\theta}_{i..}$, $\hat{\theta}_{.jk}$ obtenidos en (2.18) se tiene

$$\ell = K_1 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log x_{.jk} - 2 N \log N ,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2 \left(K_1 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log x_{.jk} - \right.$$

$$\left. - 2N \log N \right) + 2 \left[(r-1) + (cd-1) \right] . \qquad (2.20)$$

Por otro lado de (2.17), (2.4) y siguiendo un procedimiento similar al caso anterior se llega a que el AIC del MODELO(1) es

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd-1) . \qquad (2.21)$$

OBSERVACION

Siguiendo el criterio MAIC de selección del modelo tenemos:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor modelo, lo que indicará que el factor Fila es independiente del par de factores Columna y Profundidad (donde los AIC están dados en (2.20) y (2.21)).

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo cual indica que el factor Fila no es independiente del par de factores Columna y Profundidad.

Observese que, a efectos de comparación, se puede ignorar la constante $-2\ K_1$ en (2.20) y (2.21). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left(\sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - 2N \log N\right) +$$

$$+ 2(r + cd - 2)$$

$$AIC^{\bullet}(1) = -2\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1) . \qquad (2.22)$$

Para analizar la independencia entre el factor Columna y el factor Profundidad, esto es $H_0(C \times D)$ se realiza el siguiente

contraste de hipótesis

$$H_0 : p_{.jk} = p_{.j.}\, p_{..k} , \qquad j=1,\ldots,c \ , \ k=1,\ldots d$$

frente a

$$H_1 : p_{jk} \neq p_{.j.}\, p_{..k} , \qquad \text{para por lo menos un } (j,k) .$$

(2.23)

De manera similar al caso anterior, se observa que los datos siguen una distribución Multinomial

$$P\left(X_{.jk} = x_{.jk} \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d\right) =$$

$$= \frac{N\,!}{\prod_{j=1}^{c}\prod_{k=1}^{d} x_{.jk}\,!} \prod_{j=1}^{c}\prod_{k=1}^{d}\left(p_{.jk}\right)^{x_{.jk}} .$$

de ahí que el logaritmo de la función de verosimilitud es

$$\ell = K_2 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log p_{.jk} , \qquad (2.24)$$

donde $K_2 = \log \dfrac{N!}{\prod_{j=1}^{c}\prod_{k=1}^{d} x_{.jk}\,!}$.

Y los modelos son los que se indican a continuación:

$$\text{MODELO(0)}: \quad p_{.jk} = \theta_{.j.}\, \theta_{..k} \ , \quad j=1,\ldots,c \ , \ k=1,\ldots,d$$

$$\text{donde} \quad \sum_{j=1}^{c} \theta_{.j.} = 1 \ , \qquad \sum_{k=1}^{d} \theta_{..k} = 1$$

(2.25)

en este caso se supone que los factores Columna y Profundidad son independientes.

$$\text{MODELO(1)}: \quad p_{.jk} = \theta_{.jk} \ , \qquad j=1,\ldots,c \ , \ k=1,\ldots,d \qquad (2.26)$$

donde $\sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1$

en este caso se supone que los factores Columna y Profundidad no son independientes.

En base a (2.24) se obtienen los estadisticos AIC(0) y AIC(1) de los modelos MODELO(0) y MODELO(1) dados en (2.25) y (2.26) respectivamente. Estos estadisticos son presentados en el siguiente teorema.

TEOREMA 2.3

Supuesto que los factores Columna y Profundidad son independientes (MODELO(0)), entonces

$$AIC(0) = -2\left(K_2 + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 2N \log N\right) +$$

$$+ 2(c + d - 2) .$$

Supuesto que los factores Columna y Profundidad no son independientes (MODELO(1)), entonces

$$AIC(1) = -2\left(K_2 + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log x_{.jk} - N \log N\right) + 2(cd - 1) .$$

Demostración

Bajo el MODELO(0) dado en (2.25) se tiene que el logaritmo de la función de verosimilitud dado en (2.24) será

$$\ell = K_2 + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log (\theta_{.j.} \theta_{..k}) , \qquad (2.27)$$

donde

$$\sum_{j=1}^{c} \theta_{.j.} = 1 \quad , \quad \sum_{k=1}^{d} \theta_{..k} = 1 .$$

Obteniendose

$$\hat{\theta}_{.j.} = \frac{x_{.j.}}{N} , \qquad j=1,\ldots,c \tag{2.28}$$

$$\hat{\theta}_{..k} = \frac{x_{..k}}{N} , \qquad k=1,\ldots,d . \tag{2.29}$$

Reemplazando (2.28) y (2.29) en (2.27) se tiene

$$\ell = K_2 + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 2N \log N ,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2\left(K_2 + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 2N \log N\right) +$$

$$+ 2[(c - 1) + (d - 1)] . \tag{2.30}$$

Bajo el MODELO(1) dado en (2.26) se tiene que el logaritmo de la función de verosimilitud dado en (2.24) será

$$\ell = K_2 + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log \theta_{.jk} , \tag{2.31}$$

donde

$$\sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1 .$$

Obteniendose

$$\hat{\theta}_{.jk} = \frac{x_{.jk}}{N} , \qquad j=1,\ldots,c , \quad k=1,\ldots,d , \tag{2.32}$$

reemplazando el valor $\hat{\theta}_{.jk}$ en (2.31) se tiene

$$\ell = K_2 + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{\cdot jk} \log x_{\cdot jk} - N \log N ,$$

luego el AIC del MODELO(1) será

$$AIC(1) = -2\left(K_2 + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{\cdot jk} \log x_{\cdot jk} - N \log N \right) + 2(cd - 1) . \qquad (2.33)$$

OBSERVACION

Siguiendo el criterio MAIC y considerando los AIC(0) y AIC(1) dados en (2.30) y (2.33), se tiene:

Si AIC(0) < AIC(1), entonces, el modelo indicará que los factores Columna y Profundidad son independientes.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo que indicará que los factores Columna y Profundidad no son independientes.

Observese que, a efectos de comparación, se puede ignorar la constante $-2\ K_2$ en (2.30) y (2.33). En tal caso, las comparaciones anteriores se podrian realizar con los estadisticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left(\sum_{j=1}^{c} x_{\cdot j\cdot} \log x_{\cdot j\cdot} + \sum_{k=1}^{d} x_{\cdot\cdot k} \log x_{\cdot\cdot k} - 2N \log N \right) +$$

$$+ 2(c + d - 2)$$

$$AIC^{\bullet}(1) = -2\left(\sum_{j=1}^{c} \sum_{k=1}^{c} x_{\cdot jk} \log x_{\cdot jk} - N \log N \right) + 2(cd - 1) . \qquad (2.34)$$

Se pueden efectuar otros análisis con respecto a la

independencia observando que:

i) $H_0(F \times C \times D) \Leftrightarrow H_0(C \times FD) \cap H_0(F \times D)$

pues,

$p_{ijk} = p_{.j.}\, p_{i.k}$ y $p_{i.k} = p_{i..}\, p_{..k} \Leftrightarrow p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k}$

ii) $H_0(F \times C \times D) \Leftrightarrow H_0(D \times FC) \cap H_0(F \times C)$

pues,

$p_{ijk} = p_{ij.}\, p_{..k}$ y $p_{ij.} = p_{i..}\, p_{.j.} \Leftrightarrow p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k}$

iii) $H_0(F \times CD) \Leftrightarrow H_0(F \times C/D) \cap H_0(F \times D)$.

En este tercer caso se esta afirmando que el factor Filla es independiente del par de factores Columna y Profundidad, si y solamente si, los factores Fila y Columna son condicionalmentte independientes dado el factor Profundidad y si los factores Fila y Profundidad son independientes, ya que,

$$p_{ijk} = \frac{p_{i.k}\, p_{.jk}}{p_{..k}} \quad y \quad p_{i.k} = p_{i..}\, p_{..k} \Leftrightarrow p_{ijk} = p_{i..}\, p_{.jk} .$$

Para analizar si los factores Fila y Columna soon condicionalmente independientes dado el factor Profundidad, sse efectúa el siguiente contraste de hipótesis

$$H_0 : \; p_{ijk} = \frac{p_{i.k}\, p_{.jk}}{p_{..k}} , \qquad i=1,\ldots,r \;,\; j=1,\ldots,c \;,\; k=1,\ldots,d$$

frente a

$$H_1 : \; p_{ijk} \neq \frac{p_{i.k}\, p_{.jk}}{p_{..k}} , \qquad \text{para por lo menos algún } (i,j,k) .$$

(2.35)

Procediendo de forma análoga al primer caso estudiado se observa que los modelos son:

$$\text{MODELO(0):}\quad p_{ijk} = \frac{\theta_{i.k}\,\theta_{.jk}}{\theta_{..k}}\,, \qquad i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d$$

$$\text{donde}\quad \sum_{i=1}^{r}\sum_{k=1}^{d}\theta_{i.k} = 1\,, \qquad \sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{.jk} = 1\,, \qquad \sum_{k=1}^{d}\theta_{..k} = 1 \tag{2.36}$$

en este caso se supone independencia condicional entre los factores Fila y Columna dado el factor profundidad.

$$\text{MODELO(1):}\quad p_{ijk} = \theta_{ijk}\,, \qquad i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d$$

$$\text{donde}\quad \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{ijk} = 1 \tag{2.37}$$

en este caso se supone que los factores Fila y Columna no son condicionalmente independientes dado el factor Profundidad.

Como se vio en el primer caso tratado los datos siguen una distribución Multinomial con el logaritmo de la función de verosimilitud dado en (2.4). En base a esto se calculan los estadísticos AIC de cada uno de los modelos dados en (2.36), (2.37). Los cuales son presentados en el siguiente teorema.

TEOREMA 2.4

Supuesto independencia condicional entre los factores Fila y Columna dado el factor Profundidad (MODELO(0)), entonces

$$AIC(0) = -2\left\{K_1 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k}\log x_{i.k} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk}\log x_{.jk} - \right.$$

$$\left. - \sum_{k=1}^{d} x_{..k}\log x_{..k} - N\log N\right\} + 2(rd + cd - d - 1)\,.$$

Supuesto que no existe independencia condicional entre los factores Fila y Columna dada el factor Profundidad (MODELO(1)), entonces

$$AIC(1) = -2\left\{K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right\} + 2(rcd - 1).$$

Demostración

Bajo el MODELO(0) dado en (2.36), se obtiene que el logaritmo de la función de verosimilitud (2.4) es

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{\theta_{i.k}\,\theta_{.jk}}{\theta_{..k}}, \qquad (2.38)$$

donde

$$\sum_{i=1}^{r}\sum_{k=1}^{d} \theta_{i.k} = 1, \qquad \sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{.jk} = 1, \qquad \sum_{k=1}^{d} \theta_{..k} = 1.$$

Obteniendose en este caso

$$\hat{\theta}_{i.k} = \frac{x_{i.k}}{N}, \qquad i=1,\ldots,r,\ k=1,\ldots,d \qquad (2.39)$$

$$\hat{\theta}_{.jk} = \frac{x_{.jk}}{N}, \qquad j=1,\ldots,c,\ k=1,\ldots,d \qquad (2.40)$$

$$\hat{\theta}_{..k} = \frac{x_{..k}}{N}, \qquad k=1,\ldots,d. \qquad (2.41)$$

Sustituyendo (2.39), (2.40) y (2.41) en (2.38) se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} -$$

$$- \sum_{k=1}^{d} x_{..k} \log x_{.k} - N \log N,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2\left[K_1 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - \right.$$

$$\left. - \sum_{k=1}^{d} x_{..k} \log x_{..k} - N \log N\right] + 2(rd + cd - d - 1) \,. \tag{2.42}$$

Bajo el MODELO(1) dado en (2.37) se tiene que (2.4) será

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{ijk} \,.$$

procediendo de forma similar al caso anterior se obtiene que el estimador de máxima verosimilitud de θ_{ijk} es

$$\hat{\theta}_{ijk} = \frac{x_{ijk}}{N} \,.$$

De donde se obtiene que

$$AIC(1) = -2\left[K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right] + 2(rcd-1) \,. \tag{2.43}$$

OBSERVACION

Siguiendo al criterio de selección MAIC para los modelos (2.36), (2.37) y considerando sus respectivos AIC dados en (2.42) (2.43) se tiene:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor modelo, lo cual indicará que existe independencia condicional entre los factores Fila y Columna dado el factor Profundidad.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, esto indicará que no existe independencia condicional entre los factores Fila y Columna dado el factor Profundidad.

Observese que, a efectos de comparación, se puede ignorar

la constante $-2\ K_1$ en (2.42) y (2.43). En tal caso, las comparaciones se podrian realizar con los estadisticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left(\sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log x_{i.k} + \sum_{j=1}^{c} \sum_{k=1}^{d} x_{.jk} \log x_{.jk} - \right.$$

$$\left. - \sum_{k=1}^{d} x_{..k} \log x_{..k} - N \log N \right) + 2(rd + cd - d - 1)$$

$$AIC^{\bullet}(1) = -2\left(\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N \right) + 2(rcd - 1) .$$

(2.44)

Para analizar si los factores Fila y Profundidad son independientes se efectúa el siguiente contraste de hipótesis

$$H_0 : \ p_{i.k} = p_{i..}\ p_{..k} , \qquad i=1,\ldots,r \ , \ k=1,\ldots,d$$

frente a

$$H_1 : \ p_{i.k} \neq p_{i..}\ p_{..k} , \qquad \text{para por lo menos un } (i,k) .$$

(2.45)

Procediendo de manera análoga al caso de estudio de independencia entre el factor Columna y Profundidad, se tiene que los modelos son:

MODELO(0): $p_{i.k} = \theta_{i..}\ \theta_{..k}$

donde $\sum_{i=1}^{r} \theta_{i..} = 1 , \qquad \sum_{k=1}^{d} \theta_{..k} = 1$

(2.46)

en este caso se supone que los factores Fila y Profundidad son independientes.

MODELO(1): $p_{i.k} = \theta_{i.k}$

donde $\sum_{i=1}^{r} \sum_{k=1}^{d} \theta_{i.k} = 1$ (2.47)

en este caso se supone que los factores Fila y Profundidad no son independientes.

Los estadisticos AIC de los modelos MODELO(0) Y MODELO(1) dados en (2.46) y (2.47) son:

$$AIC(0) = -2\left[K_3 + \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 2N \log N\right] +$$

$$+ 2(r + d - 2)$$

$$AIC(1) = -2\left[K_3 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log x_{i.k} - N \log N\right] + 2(rd - 1)$$

donde $K_3 = \log \dfrac{N!}{\prod_{i=1}^{r} \prod_{k=1}^{d} x_{i.k}!}$.

OBSERVACION

El criterio de selección MAIC entre el MODELO(0) y el MODELO(1) dados en (2.46) y (2.47) es el siguiente:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es elegido como mejor modelo, lo cual indicará que los factores Fila y Profundidad son independientes.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es elegido como mejor modelo, lo cual indicará que los factores Fila y Columna no son independientes.

Observese que, a efectos de comparación, se puede ignorar la constante $-2\ K_3$ en los estadisticos AIC. En tal caso, las comparaciones anteriores se podrían realizar con los estadisticos $AIC^{*}(0)$ y $AIC^{*}(1)$, donde

$$AIC^{\bullet}(0) = -2\left\{ \sum_{i=1}^{r} x_{i..} \log x_{i..} + \sum_{k=1}^{d} x_{..k} \log x_{..k} - 2N \log N \right\} +$$

$$+ 2(r + d - 2)$$

$$AIC^{\bullet}(1) = -2\left\{ \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log x_{i.k} - N \log N \right\} + 2(rd - 1) . \qquad (2.48)$$

II.3 TEST DE HOMOGENEIDAD

Por lo expuesto en la introducción de este capitulo el contrastar si r muestras $X^{(i)}$, i=1,...,r proceden de una misma población es equivalente a contrastar que las r muestras independientes de una tabla (cxd) son homogéneas, esto es, que las r muestras de poblaciones multinomiales con (cxd) categorias son homogéneas.

Se puede abordar este problema considerando estas r muestras como una tabla de tres factores de clasificación (rxcxd) y en esta situación el contraste de hipótesis viene dado por

$$H_0 : \frac{p_{ijk}}{p_{i..}} = p_{.jk} , \qquad j=1,...,c , \ k=1,...,d$$

frente a

$$H_1 : \frac{p_{ijk}}{p_{i..}} \neq p_{.jk} , \qquad \text{para algún } (i,j,k) . \qquad (2.49)$$

Como $p_{jk/i}$ es la probabilidad condicional de la ocurrencia en la (j,k)-ésima categoria de los factores Columna y Profundidad dado la i-ésima categoria del factor Fila, es decir,

$$p_{jk/i} = \frac{p_{ijk}}{p_{i..}} , \quad \text{tal que, } \sum_{j=1}^{c} \sum_{k=1}^{d} p_{jk/i} = 1 , \quad \text{para } i=1,...,r$$

los modelos pueden formularse de la siguiente manera

MODELO(0): $p_{jk/i} = \theta_{.jk}$, $j=1,\ldots,c$, $k=1,\ldots,d$

donde $\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{.jk} = 1$

(2.50)

en este caso se supone que las r muestras de poblaciones multinomiales con (cxd) categorías son homogéneas

MODELO(1): $p_{jk/i} = \theta_{jk/i}$, $i=1,\ldots,r$

donde $\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{jk/i} = 1$, para $i=1,\ldots,r$

(2.51)

en este caso se supone que las r muestras de poblaciones multinomiales con (cxd) categorías no son homogeneas.

Por otro lado se tiene que la función de probabilidad de los datos es,

$$P\left(X_{ijk} = x_{ijk} \text{ , } i=1,\ldots,r \text{ , } j=1,\ldots,c \text{ , } k=1,\ldots,d\right) =$$

$$= \prod_{i=1}^{r} \left(\frac{x_{i..}!}{\prod_{j=1}^{c}\prod_{k=1}^{d} x_{ijk}!} \prod_{j=1}^{c}\prod_{k=1}^{d} \left(p_{jk/i}\right)^{x_{ijk}} \right) ,$$

luego el logaritmo de la función de verosimilitud será

$$\ell = K_4 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log p_{jk/i} \text{ ,} \qquad (2.52)$$

donde $K_4 = \log \dfrac{\prod_{i=1}^{r} x_{i..}!}{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{d} x_{ijk}!}$.

Los AIC de estos modelos se obtienen en el siguiente

teorema.

TEOREMA 2.5

Suponiendo que las r muestras de poblaciones multinomiales con (cxd) categorías son homogeneas (MODELO(0)), entonces

$$AIC(0) = -2\left\{K_4 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - N \log N\right\} + 2(cd - 1) .$$

Suponiendo que las r muestras no son homogeneas (MODELO(1)), entonces

$$AIC(1) = -2\left\{K_4 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r} x_{i..} \log x_{i..}\right\} +$$

$$+ 2[r(cd - 1)] .$$

Demostración

Bajo el MODELO(0) dado en (2.50) se obtiene que el logaritmo de la función de verosimilitud dada en (2.52) es

$$\ell = K_4 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \theta_{.jk} , \qquad (2.53)$$

donde $\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{.jk} = 1$.

De donde se tiene

$$\hat{\theta}_{.jk} = \frac{x_{.jk}}{N} , \qquad j=1,\ldots,c \ , \ k=1,\ldots,d . \qquad (2.54)$$

Sustituyendo $\hat{\theta}_{.jk}$ en (2.53) se llega a

$$\ell = K_4 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \frac{x_{.jk}}{N} ,$$

luego el AIC del MODELO(0) es

$$AIC(0)= -2\left(K_4 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - N \log N\right) + 2(cd - 1) . \qquad (2.55)$$

Bajo el MODELO(1) dado en (2.51) es inmediato que el logaritmo de la función de verosimilitud dado en (2.52) es

$$\ell = K_4 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{jk/i} , \qquad (2.56)$$

donde $\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{jk/i} = 1$, para $i=1,\ldots,r$.

Obteniendose

$$\hat{\theta}_{jk/i} = \frac{x_{ijk}}{x_{i..}} , \qquad i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d ,$$

reemplazando $\hat{\theta}_{jk/i}$ en (2.56) se tiene

$$\ell = K_4 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{x_{i..}} ,$$

luego el AIC del MODELO(1) es

$$AIC(1) = -2\left(K_4 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r} x_{i..} \log x_{i..}\right) +$$

$$+ 2[r\,(\,cd - 1)] . \qquad (2.57)$$

OBSERVACION

Aplicando el criterio de seleccion MAIC a los modelos (2.50), (2.51) y considerando los AIC dados en (2.55) y (2.57) se tiene:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor modelo, lo cual indica que las r muestras de poblaciones multinomiales con (cxd) categorias son homogéneas.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo cual indica que las r muestras de poblaciones multinomiales con (cxd) categorias no son homogéneas.

Observese que, a efecto de comparación, se puede ignorar la constante $-2\,K_4$ en (2.55) y (2.57). En tal caso, las comparaciones anteriores se podrian realizar con los estadisticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left\{\sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk}\log x_{.jk} - N\log N\right\} + 2(cd-1)$$

$$AIC^{\bullet}(1) = -2\left\{\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log x_{ijk} - \sum_{i=1}^{r} x_{i..}\log x_{i..}\right\} +$$

$$+ 2[r\,(cd-1)]\,. \qquad (2.58)$$

Podemos realizar otros análisis dentro de este mismo problema de Homogeneidad (C, D), puesto que,

Homogeneidad (C, D) $\Leftrightarrow$ Homogeneidad (D/C) $\cap$ Homogeneidad (C)

ya que,

$$\frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}} \quad y \quad \frac{p_{ij.}}{p_{i..}} = p_{.j.}\,, \quad i=1,\ldots,r\,,\ j=1,\ldots,c\,,\ k=1,\ldots,d$$

$$\Leftrightarrow \quad \frac{p_{ijk}}{p_{i..}} = p_{.jk}\,, \qquad i=1,\ldots,r\,,\ j=1,\ldots,c\,,\ k=1,\ldots,d\,.$$

Para analizar la homogeneidad de las r muestras del factor Columna, se realiza el siguiente contraste de hipótesis

$$H_0 : \frac{p_{ij.}}{p_{i..}} = p_{.j.} , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c$$

frente a

$$H_1 : \frac{p_{ij.}}{p_{i..}} \neq p_{.j.} , \qquad \text{para por lo menos un } (i,j) .$$

(2.59)

Como $p_{j/i} = \frac{p_{ij.}}{p_{i..}}$, los modelos pueden formularse de la siguiente manera:

MODELO(0): $p_{j/i} = \theta_{.j.}$

donde

$$\sum_{j=1}^{c} \theta_{.j.} = 1$$

(2.60)

en este caso se supone que existe homogeneidad del factor Columna.

MODELO(1): $p_{j/i} = \theta_{j/i}$

donde $\sum_{j=1}^{c} \theta_{j/i} = 1 , \quad i=1,\ldots,r$

(2.61)

en este caso se supone que no existe homogeneidad del factor Columna.

Y los datos siguen la función de probabilidad

$$P\left(X_{ij.} = x_{ij.} \quad , \; i=1,\ldots,r \; , \; j=1,\ldots,c\right) =$$

$$= \prod_{i=1}^{r} \left(\frac{x_{i..}!}{\prod_{j=1}^{c} x_{ij.}!} \prod_{j=1}^{c} \left(p_{j/i}\right)^{x_{ij.}} \right) ,$$

luego el logaritmo de la función de verosimilitud es

$$\ell = K_S + \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij.} \log p_{j/i} \; , \qquad (2.62)$$

donde $$K_S = \log \frac{\prod_{i=1}^{r} x_{i..}!}{\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij.}!} .$$

Los estadisticos AIC de los modelos (2.60) y (2.61) son dados en el siguiente teorema.

TEOREMA 2.6

Suponiendo homogeneidad del factor Columna (MODELO(0)), entonces

$$AIC(0) = -2\left(K_S + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} - N \log N\right) + 2(c-1) .$$

Suponiendo que no existe homogeneidad del factor Columna (MODELO(1)), entonces

$$AIC(1) = -2\left(K_S + \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij.} \log x_{ij.} - \sum_{i=1}^{r} x_{i..} \log x_{i..}\right) +$$

$$+ 2[r\,(c-1)] .$$

Demostración

Bajo el MODELO(0) dado en (2.60) el logaritmo de la funcion de verosimilitud toma la expresión

$$\ell = K_S + \sum_{j=1}^{c} x_{.j.} \log \theta_{.j.} \quad , \qquad (2.63)$$

donde $\sum_{j=1}^{c} \theta_{.j.} = 1$.

De donde se obtiene

$$\hat{\theta}_{.j.} = \frac{x_{.j.}}{N} \quad , \qquad j=1,\ldots,c \ .$$

Sustituyendo $\hat{\theta}_{.j.}$ en (2.63) se tiene

$$\ell = K_S + \sum_{j=1}^{c} x_{.j.} \log \frac{x_{.j.}}{N} \quad ,$$

luego el AIC del MODELO(0) es

$$AIC(0) = -2\left[K_S + \sum_{j=1}^{c} x_{.j.} \log x_{.j.} - N \log N\right] + 2(c-1) \ . \qquad (2.64)$$

Procediendo de la misma forma, se obtiene que el AIC del MODELO(1) es

$$AIC(1) = -2\left[K_S + \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij.} \log x_{ij.} - \sum_{i=1}^{r} x_{i..} \log x_{i..}\right] +$$

$$+ 2[r\,(c-1)] \ . \qquad (2.65)$$

∎

OBSERVACION

Aplicando el criterio de selección MAIC a los modelos (2.60) y (2.61), cuyos estadísticos AIC están dados en (2.64) y (2.65) se tiene:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor

modelo, lo cual indica que las r muestras de poblaciones multinomiales con c categorías son homogéneas (o existe homogeneidad del factor Columna).

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo cual indica que las r muestras de poblaciones multinomiales con c categorías no son homogéneas.

Observese que, a efecto de comparación, se puede ignorar la constante $-2\,K_5$ en (2.64) y (2.65). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{*}(0)$ y $AIC^{*}(1)$, donde

$$AIC^{*}(0) = -2\left(\sum_{j=1}^{c} x_{.j.} \log x_{.j.} - N \log N \right) + 2(c-1)$$

$$AIC^{*}(1) = -2\left(\sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij.} \log x_{ij.} - \sum_{i=1}^{r} x_{i..} \log x_{i..} \right) +$$

$$+ 2[r\,(c-1)] \,. \qquad (2.66)$$

Para analizar la Homogeneidad Condicional del factor Profundidad dado el factor Columna, se efectúa el siguiente contraste de hipótesis

$$H_0 : \frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}} , \qquad i=1,\ldots,r \;,\; j=1,\ldots,c \;,\; k=1,\ldots,d$$

frente a

$$H_1 : \frac{p_{ijk}}{p_{ij.}} \neq \frac{p_{.jk}}{p_{.j.}} , \qquad \text{para algún } (i,j,k) \,. \qquad (2.67)$$

Como $p_{k/ij} = \dfrac{p_{ijk}}{p_{ij.}}$ y $p_{k/j} = \dfrac{p_{.jk}}{p_{.j.}}$, los modelos pueden formularse de la siguiente manera:

MODELO(0): $p_{k/ij} = \theta_{k/j}$

donde $\sum_{k=1}^{d} \theta_{k/j} = 1 \ , \qquad j=1,\ldots,c$

(2.68)

en este caso se supone que existe homogeneidad condicional del factor Profundidad dado el factor Columna.

MODELO(1): $p_{k/ij} = \theta_{k/ij}$

donde $\sum_{k=1}^{d} \theta_{k/ij} \ , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c$

(2.69)

en este caso se supone que no existe homogeneidad condicional del factor Profundidad dado el factor Columna.

Por otro lado se tiene que la función de probabilidad que siguen los datos

$$P\left(X_{ijk} = x_{ijk} \ , \ i=1,\ldots,r \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d\right) =$$

$$= \prod_{i=1}^{r} \left(\frac{x_{i..}!}{\prod_{j=1}^{c} \prod_{k=1}^{d} x_{ijk}!} \prod_{j=1}^{c} \prod_{k=1}^{d} \left(p_{k/ij}\right)^{x_{ijk}} \right) .$$

luego el logaritmo de la función de verosimilitud será

$$\ell = K_6 + \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log p_{k/ij} \ , \qquad (2.70)$$

donde $K_6 = \log \dfrac{\prod_{i=1}^{r} x_{i..}!}{\prod_{i=1}^{r} \prod_{j=1}^{c} \prod_{k=1}^{d} x_{ijk}!}$,

a partir de (2.70) se obtienen los AIC de los modelos (2.68) y (2.69), los cuales se indican en el siguiente teorema.

TEOREMA 2.7

Suponiendo homogeneidad del factor Profundidad dado el factor Columna (MODELO(0)), entonces

$$AIC(0) = -2\left[K_6 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - \sum_{j=1}^{c} x_{.j.} \log x_{.j.}\right] +$$

$$+ 2[c\,(d-1)] \,.$$

Suponiendo que no existe homogeneidad del factor Profundidad dado el factor Columna (MODELO(1)), entonces

$$AIC(1) = -2\left[K_6 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log x_{ij.}\right] +$$

$$+ 2[rc\,(d-1)] \,.$$

Demostración

Bajo el MODELO(0) dado en (2.68) se tiene que el logaritmo de la función de verosimilitud dado en (2.70) es

$$\ell = K_6 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \theta_{k/j} \,, \qquad (2.71)$$

donde $\sum_{k=1}^{d} \theta_{k/j} = 1 \,, \qquad j=1,\ldots,c \,.$

De donde se obtiene

$$\hat{\theta}_{k/j} = \frac{x_{.jk}}{x_{.j.}} \,, \qquad j=1,\ldots,c \,,\ k=1,\ldots,d \,.$$

Sustituyendo $\hat{\theta}_{k/j}$ en (2.71) se tiene

$$\ell = K_6 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \frac{x_{.jk}}{x_{.j.}} \,,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2\left[K_6 + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - \sum_{j=1}^{c} x_{.j.} \log x_{.j.}\right] + 2[c\,(d-1)]\,. \tag{2.72}$$

Bajo el MODELO(1) dado en (2.69) se tiene que el logaritmo de la función de verosimilitud dada en (2.70) es

$$\ell = K_6 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{k/ij}\,, \tag{2.73}$$

donde $\sum_{k=1}^{d} \theta_{k/ij} = 1\,, \qquad i=1,\ldots,r\,,\ j=1,\ldots,c\,.$

De donde se obtiene

$$\hat{\theta}_{k/ij} = \frac{x_{ijk}}{x_{ij.}}\,, \qquad i=1,\ldots,r\,,\ j=1,\ldots,c\,,\ k=1,\ldots,d\,.$$

Sustituyendo $\hat{\theta}_{k/ij}$ en (2.73) se tiene

$$\ell = K_6 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{x_{ij.}}\,,$$

luego el AIC del MODELO(1) es

$$AIC(1) = -2\left[K_6 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log x_{ij.}\right] + 2[rc\,(d-1)]\,. \tag{2.74}$$

OBSERVACION

Aplicando el criterio de selección MAIC a los modelos (2.68) y (2.69), cuyos estadísticos AIC están dados en (2.72) y (2.74) se tiene:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor modelo, lo cual indica que existe homogeneidad condicional del factor Profundidad dado el factor Columna.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo cual indica que no existe homogeneidad condicional del factor Profundidad dado el factor Columna.

Observese que, a efectos de comparación, se puede ignorar la constante $-2\,K_6$ en (2.72) y (2.74). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{*}(0)$ y $AIC^{*}(1)$, donde

$$AIC^{*}(0) = -2\left[\sum_{j=1}^{c}\sum_{k=1}^{d} x_{\cdot jk}\,\log x_{\cdot jk} - \sum_{j=1}^{c} x_{\cdot j\cdot}\,\log x_{\cdot j\cdot}\right] + 2[c(d-1)]$$

$$AIC^{*}(1) = -2\left[\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\,\log x_{ijk} - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij\cdot}\,\log x_{ij\cdot}\right] +$$

$$+ 2[rc\,(d-1)]\,. \qquad (2.75)$$

Se puede realizar un análisis más detallado analizando si las clasificaciones (o categorías) del factor Profundidad son condicionalmente homogéneas dada la j-ésima clasificación del factor Columna. Para ello el contraste de hipótesis correspondiente será

$$H_0 : \frac{p_{ijk}}{p_{ij\cdot}} = \frac{p_{\cdot jk}}{p_{\cdot j\cdot}}\,, \qquad i=1,\ldots,r\,,\ k=1,\ldots,d\,,\ \sum_{k=1}^{d} p_{\cdot jk} = p_{\cdot j\cdot}$$

frente a

$$H_1 : \frac{p_{ijk}}{p_{ij\cdot}} \neq \frac{p_{\cdot jk}}{p_{\cdot j\cdot}}\,, \qquad i=1,\ldots,r\,,\ \sum_{k=1}^{d} p_{ijk} = p_{ij\cdot}\,. \qquad (2.76)$$

Y procediendo de forma análoga a los casos anteriores, se obtiene el siguiente resultado.

TEOREMA 2.8

Suponiendo homogeneidad condicional del factor Profundidad dado la j-ésima clasificación del factor Columna (MODELO(0)), entonces

$$AIC(0) = -2\left[K_7 + \sum_{k=1}^{d} x_{.jk} \log x_{.jk} - x_{.j.} \log x_{.j.}\right] + 2(d - 1) .$$

Suponiendo que no existe homogeneidad condicional del factor Profundidad dado la j-ésima clasificación del factor Columna (MODELO(1)), entonces

$$AIC(1) = -2\left[K_7 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r} x_{ij.} \log x_{ij.}\right] +$$

$$+ 2[r\,(d - 1)] ,$$

donde $$K_7 = \log \frac{\prod_{i=1}^{r} x_{i..}!}{\prod_{i=1}^{r} \prod_{k=1}^{d} x_{ijk}!} .$$

OBSERVACION

Siguiendo el criterio de selección MAIC se tiene:

Si AIC(0) < AIC(1), entonces, el MODELO(0) es el mejor modelo, lo cual indica que existe homogeneidad condicional del factor Profundidad dada la j-ésima clasificación del factor Columna.

Si AIC(1) < AIC(0), entonces, el MODELO(1) es el mejor modelo, lo cual indica que no existe homogeneidad condicional del factor Profundidad dado la j-ésima clasificación del factor Columna.

Observese que, a efectos de comparación, se puede ignorar la constante $-2K_7$ en los estadísticos AIC. En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left(\sum_{k=1}^{d} x_{.jk} \log x_{.jk} - x_{.j.} \log x_{.j.} \right) + 2(d-1)$$

$$AIC^{\bullet}(1) = -2\left(\sum_{i=1}^{r} \sum_{k=1}^{d} x_{ijk} \log x_{ijk} - \sum_{i=1}^{r} x_{ij.} \log x_{ij.} \right) + 2[r(d-1)].$$

II.4 TEST DE INTERACCION

Se puede analizar aún más la Homogeneidad Condicional (D/C) puesto que

H_0(Homogeneidad Condicional (D/C)) $\Leftrightarrow$

H_0(Interacción (FD)) $\cap$ H_0(Interacción (FD, C))

ya que,

$$p_{i.k} = \sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} \qquad y \qquad p_{ijk} = \frac{p_{i.k}\, p_{ij.}\, p_{.jk}}{\left(\sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} \right) p_{.j.}}$$

$$\Leftrightarrow \frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}}, \qquad i=1,\ldots,r \;,\; j=1,\ldots,c \;,\; k=1,\ldots,d \;.$$

Para analizar la interacción entre el factor Fila y Profundidad, se efectúa el siguiente contraste de hipótesis

$$H_0 : p_{i.k} = \sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} , \quad i=1,\ldots,r \ , \ k=1,\ldots,d$$

$$\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij.} = 1 \ , \ \sum_{j=1}^{c} \sum_{k=1}^{d} p_{.jk} = 1 \ , \ \sum_{j=1}^{c} p_{.j.} = 1$$

frente a

$$H_1 : p_{i.k} \neq \sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} , \quad \text{para algún } (i,k) .$$

(2.77)

Los modelos serán:

$$\text{MODELO(0)}: \quad p_{i.k} = \sum_{j=1}^{c} \frac{\theta_{ij.}\, \theta_{.jk}}{\theta_{.j.}}$$

$$\text{donde} \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \theta_{ij.} = 1 \ , \ \sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1 \ , \ \sum_{j=1}^{c} \theta_{.j.} = 1$$

(2.78)

en este caso se supone que existe interacción entre los factores Fila y Profundidad.

$$\text{MODELO(1)}: \quad p_{i.k} = \theta_{i.k}$$

$$\text{donde} \quad \sum_{i=1}^{r} \sum_{k=1}^{d} \theta_{i.k} = 1$$

(2.79)

en este caso se supone que no existe interacción entre los factores Fila y Profundidad.

De otro lado se tiene que los datos siguen la función de probabilidad

$$P\left(X_{i.k} = x_{i.k} \ , \ i=1,\ldots,r \ , \ k=1,\ldots,d\right) =$$

$$= \frac{N!}{\prod_{i=1}^{r}\prod_{j=1}^{d} x_{i.k}!} \prod_{i=1}^{r}\prod_{k=1}^{d} \left(p_{i.k}\right)^{x_{i.k}} ,$$

luego el logaritmo de la función de verosimilitud será

$$\ell = K_8 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log p_{i.k} \ , \qquad (2.80)$$

donde $K_8 = \log \dfrac{N!}{\prod_{i=1}^{r}\prod_{k=1}^{d} x_{i.k}!}$,

a partir de (2.80) se obtienen los AIC de los modelos (2.78) y (2.79). Estos son presentados en el siguiente teorema.

TEOREMA 2.9

Suponiendo que existe interacción entre los factores Fila y Profundidad (MODELO(0)), entonces

$$AIC(0) = -2\left[K_8 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log\left(\sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}}\right) - N \log N\right] +$$

$$+ 2(r + d - 2) \ .$$

Suponiendo que no existe interacción entre los factores Fila y Profundidad (MODELO(1)), entonces

$$AIC(1) = -2\left[K_8 + \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} - N \log N\right] + 2(rd - 1) \ .$$

Demostración

Bajo el MODELO(0) dado en (2.78), se tiene que el logaritmo de la

función de verosimilitud dada en (2.80) será

$$\ell = K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \left(\sum_{j=1}^{c} \frac{\theta_{ij.}\, \theta_{.jk}}{\theta_{.j.}} \right) , \qquad (2.81)$$

donde $\sum_{i=1}^{r} \sum_{j=1}^{c} \theta_{ij.} = 1$, $\sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1$, $\sum_{j=1}^{c} \theta_{.j.} = 1$.

De donde se obtiene que el estimador de máxima verosimilitud de

$\sum_{j=1}^{c} \frac{\theta_{ij.}\, \theta_{.jk}}{\theta_{.j.}}$ es $\frac{1}{N} \sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}}$.

El cual al reemplazarlo en (2.81) da

$$\ell = K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \left(\frac{1}{N} \sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}} \right) .$$

Luego el AIC del MODELO(0) es

$$AIC(0) = -2 \left[K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \left(\sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}} \right) - N \log N \right] +$$

$$+ 2(r + d - 2) . \qquad (2.82)$$

Bajo el MODELO(1) dado en (2.79) el logaritmo de la función de verosimilitud es

$$\ell = K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \theta_{i.k} , \qquad (2.83)$$

donde $\sum_{i=1}^{r} \sum_{k=1}^{d} \theta_{i.k} = 1$.

Obteniendose

$$\hat{\theta}_{i.k} = \frac{x_{i.k}}{N} , \qquad i=1,\ldots,r \ , \ k=1,\ldots,d \ .$$

Sustituyendo $\hat{\theta}_{i.k}$ en (2.83) se tiene

$$\ell = K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \frac{x_{i.k}}{N} \ .$$

Luego el AIC del MODELO(1) es

$$AIC(1) = -2\left(K_8 + \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log x_{i.k} - N \log N\right) + 2(rd - 1) \ . \qquad (2.84)$$

OBSERVACION

Aplicando el criterio de selección MAIC a los modelos (2.78), (2.79) y considerando sus estadisticos AIC dados en (2.82) y (2.84) se tiene

Si AIC(0) < AIC(1), entonces el MODELO(0) es el mejor modelo, lo cual indica que existe interacción entre los factores Fila y Profundidad.

Si AIC(1) < AIC(0), entonces el MODELO(1) es el mejor modelo, lo cual indica que no existe interacción entre los factores Fila y Profundidad.

Observese que, a efectos de comparación, se puede ignorar la constante $-2\ K_8$ en (2.82) y (2.84). En tal caso, las comparaciones anteriores se podrian realizar con los estadisticos $AIC^{*}(0)$ y $AIC^{*}(1)$, donde

$$AIC^{*}(0) = -2\left(\sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log \left(\sum_{j=1}^{c} \frac{x_{ij.}\ x_{.jk}}{x_{.j.}}\right) - N \log N\right) + 2(r+d-2)$$

$$AIC^{*}(1) = -2-\left(\sum_{i=1}^{r} \sum_{k=1}^{d} x_{i.k} \log x_{i.k} - N \log N\right) + 2(rd - 1) \ . \qquad (2.85)$$

Para analizar la interacción entre el factor Columna y el par de factores Fila y Profundidad se efectúa el siguiente contraste de hipótesis

$$H_0 : p_{ijk} = \frac{p_{i.k}\, p_{ij.}\, p_{.jk}}{\left(\sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} \right) p_{.j.}} , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c \ , \quad k=1,\ldots,d$$

frente a

$$H_1 : p_{ijk} \neq \frac{p_{i.k}\, p_{ij.}\, p_{.jk}}{\left(\sum_{j=1}^{c} \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} \right) p_{.j.}} , \qquad \text{para algún } (i,j,k) \ .$$

(2.86)

Los modelos serán:

$$\text{MODELO(0)}: \quad p_{ijk} = \frac{\theta_{i.k}\, \theta_{ij.}\, \theta_{.jk}}{\left(\sum_{j=1}^{c} \frac{\theta_{ij.}\, \theta_{.jk}}{\theta_{.j.}} \right) \theta_{.j.}}$$

$$\text{donde} \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \theta_{ij.} = 1 \ , \qquad \sum_{i=1}^{r} \sum_{k=1}^{d} \theta_{i.k} = 1 \ ,$$

$$\sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{.jk} = 1 \ , \qquad \sum_{j=1}^{c} \theta_{..j} = 1$$

(2.87)

en este caso se supone que existe interacción entre el factor Columna y el par de factores Fila y Columna.

$$\text{MODELO(1)}: \quad p_{ijk} = \theta_{ijk}$$

$$\text{donde} \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} \theta_{ijk} = 1$$

(2.88)

en este caso se supone que no existe interacción entre el factor Columna y el par de factores Fila y Columna.

Aquí los datos siguen la función de probabilidad y logaritmo de la función de verosimilitud dado en (2.4) a partir de esto, se obtienen los estadísticos AIC de los modelos (2.87) y (2.88), los cuales se presentan en el siguiente teorema.

TEOREMA 2.10

Suponiendo que existe interacción entre el factor Columna y el par de factores Fila y Profundidad (MODELO(0)), entonces

$$AIC(0) = -2\left[K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{i.k}\, x_{ij.}\, x_{.jk}}{\left(\sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}}\right) x_{.j.}} - N \log N\right]$$

$$+ 2(rc + rd + cd - r - c - d).$$

Suponiendo que no existe interacción entre el factor Columna y el par de factores Fila y Profundidad (MODELO(1)), entonces

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1).$$

Demostración

Bajo el MODELO(0) dado en (2.87), el logaritmo de la función de verosimilitud es

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{\theta_{i.k}\, \theta_{ij.}\, \theta_{.jk}}{\left(\sum_{j=1}^{c} \frac{\theta_{ij.}\, \theta_{.jk}}{\theta_{.j.}}\right) \theta_{.j.}}. \qquad (2.89)$$

De donde se obtiene que el estimador de máxima verosimilitud de

$$\frac{\theta_{i.k}\,\theta_{ij.}\,\theta_{.jk}}{\left(\sum_{j=1}^{c}\frac{\theta_{ij.}\,\theta_{.jk}}{\theta_{.j.}}\right)\theta_{.j.}} \quad \text{es} \quad \frac{x_{i.k}\,x_{ij.}\,x_{.jk}}{N\left(\sum_{j=1}^{c}\frac{x_{ij.}\,x_{.jk}}{x_{.j.}}\right)x_{.j.}} .$$

Sustituyendo este resultado en (2.89) se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{i.k}\,x_{ij.}\,x_{.jk}}{\left(\sum_{j=1}^{c}\frac{x_{ij.}\,x_{.jk}}{x_{.j.}}\right)x_{.j.}} - N \log N ,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2\left[K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{i.k}\,x_{ij.}\,x_{.jk}}{\left(\sum_{j=1}^{c}\frac{x_{ij.}\,x_{.jk}}{x_{.j.}}\right)x_{.j.}} - N \log N\right]$$

$$+ 2(rc + rd + cd - r - c - d) . \qquad (2.90)$$

Bajo el MODELO(1), se tiene que el logaritmo de la función de verosimilitud dado en (2.4) es

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{ijk} , \qquad (2.91)$$

donde $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{ijk} = 1 .$

De donde se obtiene

$$\hat{\theta}_{ijk} = \frac{x_{ijk}}{N} , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d .$$

Sustituyendo $\hat{\theta}_{ijk}$ en (2.91) se tiene

$$\ell = K_1 + \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{N} ,$$

luego el AIC del MODELO(1) será

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N \right) + 2(rcd - 1) \quad (2.92)$$

■

OBSERVACION

Siguiendo el criterio de selección MAIC para los modelos (2.87), (2.88) y considerando los AIC dados en (2.90) y (2.92) se tiene:

Si AIC(0) < AIC(1), entonces el MODELO(0) se elegirá como el mejor modelo, lo cual indica que existe interacción entre el factor Columna y el par de factores Fila y Profundidad.

Si AIC(1) < AIC(0), entonces se elegirá el MODELO(1) como el mejor modelo, lo cual indica que no existe interacción entre el factor Columna y el par de factores Fila y Profundidad.

Observese que, a efecto de comparación, se puede ignorar la constante $-2\,K_1$ en (2.90) y (2.92). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

$$AIC^{\bullet}(0) = -2\left[\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{i.k}\, x_{ij.}\, x_{.jk}}{\left(\sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}} \right) x_{...}} - N \log N \right] +$$

$$+ 2(rc + rd + cd - r - c - d)$$

$$AIC^{\bullet}(1) = -2\left(\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log x_{i.k} - N \log N \right) + 2(rcd - 1) \quad (2.93)$$

Kullback (1959), propone que el componente de Homogeneidad Condicional del factor Profundidad dado el factor Columna se puede descomponer algebraicamente como

$$2\hat{I}\left(H_1 : H_0(\text{Homogeneidad Condicional (D/C)}\right) =$$

$$= 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{.j.}}{x_{ij.}\,x_{.jk}}$$

$$= 2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k}\log\frac{N\,x_{i.k}}{x_{i..}\,x_{..k}} +$$

$$+ 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{i..}\,x_{.j.}\,x_{..k}}{N\,x_{ij.}\,x_{i.k}\,x_{.jk}}$$

$$= 2\hat{I}\left(H_1 : H_0(\text{Homogeneidad D})\right) + 2\hat{I}\left(H_1 : H_0(\text{Interacción FCD})\right). \qquad (2.94)$$

De donde se observa lo siguiente:

i) $H_0(\text{Homogeneidad D}) \cap H_0(\text{Interacción FCD})$

$\Rightarrow H_0(\text{homogeneidad Condicional (D/C)})$

ya que,

$$\frac{p_{i.k}}{p_{i..}} = p_{..k} \quad \text{y} \quad p_{ijk} = \frac{p_{ij.}\,p_{.jk}\,p_{i.k}}{p_{i..}\,p_{.j.}\,p_{..k}} \Rightarrow \frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}}$$

ii) $H_0(\text{Homogeneidad Condicional (D/C)}) \cap H_0(\text{Interacción FCD})$

$\Rightarrow H_0(\text{Homogeneidad D})$

ya que,

$$\frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}} \quad y \quad p_{ijk} = \frac{p_{ij.}\ p_{.jk}\ p_{i.k}}{p_{i..}\ p_{.j.}\ p_{..k}} \quad \Rightarrow \quad \frac{p_{i.k}}{p_{i..}} = p_{..k}$$

iii) H_0(Homogeneidad Condicional (D/C)) $\cap$ H_0(Homogeneidad D)

$\Rightarrow$ H_0(Interacción FCD)

ya que,

$$\frac{p_{ijk}}{p_{ij.}} = \frac{p_{.jk}}{p_{.j.}} \quad y \quad \frac{p_{i.k}}{p_{i..}} = p_{..k} \quad \Rightarrow \quad p_{ijk} = \frac{p_{ij.}\ p_{.jk}\ p_{i.k}}{p_{i..}\ p_{.j.}\ p_{..k}}$$

De (2.94) se observa que el componente de Homogeneidad Condicional (D/C) puede ser descompuesto en dos componentes, el componente de Homogeneidad D y el componente de Interacción FCD.

No siempre el componente de Homogeneidad D es menor que el componente de Homogeneidad Condicional (D/C) Si la componente de Homogeneidad D es mayor que el componente de homogeneidad (D/C), entonces la Interacción FCD es negativa y por tanto no sigue una distribución Ji-Cuadrado (ya que esa diferencia es la diferencia de dos Ji-Cuadrados y esa diferencia puede ser negativa).

Inga (1990) obtiene que su función de densidad suponiendo independencia entre las variables es la función de Wittaker. Sin embargo su uso para el análisis de interacción FCD es muy engorroso.

En base a la relación que existe entre la cantidad de información de Kullback-Leibler y el Criterio de Información de Akaike expuesta en la Proposición 2.1, se propone abordar el análisis de interacción entre los tres factores mediante el Criterio de Información de Akaike de la siguiente manera:

El analizar la hipótesis nula

H_0 : Existe interaccion FCD

equivale a contrastar

$$H_0 : \; p_{ijk} = \frac{p_{ij.}\; p_{.jk}\; p_{i.k}}{p_{i..}\; p_{.j.}\; p_{..k}} , \quad i=1,\ldots,r \; , \; j=1,\ldots,c \; , \; k=1,\ldots,d$$

frente a

$$H_1 : \; p_{ijk} \neq \frac{p_{ij.}\; p_{.jk}\; p_{i.k}}{p_{i..}\; p_{.j.}\; p_{..k}} , \quad \text{para algún } (i,j,k) .$$

(2.95)

Los modelos serán

$$\text{MODELO(0)}: \quad p_{ijk} = \frac{\theta_{ij.}\; \theta_{.jk}\; \theta_{i.k}}{\theta_{i..}\; \theta_{.j.}\; \theta_{..k}}$$

(2.96)

donde

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\theta_{ij.} = 1 , \quad \sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{.jk} = 1 , \quad \sum_{i=1}^{r}\sum_{k=1}^{d}\theta_{i.k} = 1 ,$$

$$\sum_{i=1}^{r}\theta_{i..} = 1 , \quad \sum_{j=1}^{c}\theta_{.j.} = 1 , \quad \sum_{k=1}^{d}\theta_{..k} = 1$$

en este caso se supone que existe interacción entre los tres factores.

$$\text{MODELO(1)}: \quad p_{ijk} = \theta_{ijk}$$

$$\text{donde} \quad \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d}\theta_{ijk} = 1$$

(2.97)

en este caso se supone que no existe interacción entre los tres factores.

Por otro lado, los datos siguen la función de probabilidad y el logaritmo de la función de verosimilitud dada en (2.4), a partir del cual se obtienen los estadísticos AIC de los modelos (2.96) y (2.97), los cuales se presentan en el siguiente teorema.

TEOREMA 2.11

Suponiendo que existe interacción entre los tres factores (MODELO(0)), entonces

$$AIC(0) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log x_{ij.} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} + \right.$$

$$+ \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} - \sum_{i=1}^{r} x_{i..} \log x_{i..} -$$

$$\left. - \sum_{j=1}^{c} x_{.j.} \log x_{.j.} - \sum_{k=1}^{d} x_{..k} \log x_{..k}\right) +$$

$$+ 2[r(c-1) + c(d-1) + d(r-1)] .$$

Suponiendo que no existe interaccion entre los tres factores (MODELO(1)), entonces

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1).$$

Demostración

Bajo el MODELO(0) dado en (2.96) se tiene que el logaritmo de la función de verosimilitud dada en (2.4) es

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{\theta_{ij.}\ \theta_{.jk}\ \theta_{i.k}}{\theta_{i..}\ \theta_{.j.}\ \theta_{..k}} . \qquad (2.98)$$

De donde se obtiene que el estimador de $\dfrac{\theta_{ij.}\ \theta_{.jk}\ \theta_{i.k}}{\theta_{i..}\ \theta_{.j.}\ \theta_{..k}}$ es $\dfrac{x_{ij.}\ x_{.jk}\ x_{i.k}}{x_{i..}\ x_{.j.}\ x_{..k}}$.

Sustituyendo este resultado en (2.98) se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ij.}\, x_{.jk}\, x_{i.k}}{x_{i..}\, x_{.j.}\, x_{..k}} ,$$

luego el AIC del MODELO(0) será

$$AIC(0) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log x_{ij.} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} + \right.$$

$$+ \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} - \sum_{i=1}^{r} x_{i..} \log x_{i..} - \sum_{j=1}^{c} x_{.j.} \log x_{.j.}$$

$$\left. - \sum_{k=1}^{d} x_{..k} \log x_{..k} \right) + 2[r(c-1) + c(d-1) + d(r-1)] . \qquad (2.99)$$

Bajo el MODELO(1) dado de (2.97) se tiene que el logaritmo de la función de verosimilitud dada en (2.4) será

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \theta_{ijk} , \qquad (2.100)$$

donde $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} \theta_{ijk} = 1$.

Obteniendose

$$\hat{\theta}_{ijk} = \frac{x_{ijk}}{N} , \qquad i=1,\ldots,r \ , \ j=1,\ldots,c \ , \ k=1,\ldots,d .$$

Sustituyendo $\hat{\theta}_{ijk}$ en (2.100) se tiene

$$\ell = K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{N} .$$

luego el AIC del MODELO(1) sera

$$AIC(1) = -2\left(K_1 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1).$$

(2.101)

OBSERVACION

Aplicando el criterio de selección MAIC a los modelos (2.96), (2.97) y considerando los estadísticos AIC dados en (2.99) y (2.101) se tiene:

Si AIC(0) < AIC(1), entonces el MODELO(0) será elegido como el mejor modelo, lo cual indica que existe interacción entre los tres factores.

Si AIC(1) < AIC(0), entonces el MODELO(1) será elegido como el mejor modelo, lo cual indica que no existe interacción entre los tres factores.

Observese que, a efecto de comparación, se puede ignorar la constante $-2\ K_1$ en (2.99) y (2.101). En tal caso, las comparaciones anteriores se podrían realizar con los estadísticos $AIC^{\bullet}(0)$ y $AIC^{\bullet}(1)$, donde

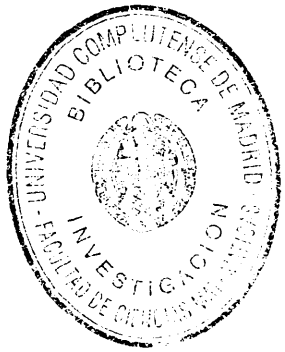
$$AIC^{\bullet}(0) = -2\left(\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log x_{ij.} + \sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log x_{.jk} + \right.$$

$$+ \sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log x_{i.k} - \sum_{i=1}^{r} x_{i..} \log x_{i..} - \sum_{j=1}^{c} x_{.j.} \log x_{.j.}$$

$$\left. - \sum_{k=1}^{d} x_{..k} \log x_{..k}\right) + 2[r(c-1) + c(d-1) + d(r-1)]$$

$$AIC^{\bullet}(1) = -2\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log x_{ijk} - N \log N\right) + 2(rcd - 1) .$$

(2.102)



CAPITULO III

"SELECCION DEL CONJUNTO OPTIMO DE VARIABLES EXPLICATIVAS DE UNA VARIABLE RESPUESTA"

III.0 SUMARIO

Resulta de gran interés en algunos estudios estadísticos el análisis de una variable respuesta a través de variables explicativas (ver Sakamoto y Akaike (1982), Sakamoto, Ishiguro y Kitagawa (1989)). Esto motivó el estudio de este problema y es por éllo que en el presente capítulo se expone un procedimiento basado en el criterio MAIC (Mínimo AIC), 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.

Este último caso se presenta cuando se trabaja con cuestionarios que tienen muchas variables explicativas para estudiar una variable respuesta.

Este tipo de información generaría una tabla de contingencia de varias variables con muchas categorías, lo cual daría lugar a muchas celdas vacías. Esto hace necesario reordenar la información para su posterior análisis.

Los procedimientos para hallar el conjunto óptimo de variables explicativas en estas dos situaciones se exponen en las Secciones III.1 y III.2 respectivamente.

La notación que se empleará a lo largo del capítulo será la siguientes:

Los datos vienen dados en una tabla de contingencia multidimensional con K factores de clasificación $I_1, I_2, \ldots, I_K$, donde I_1 es la variable respuesta e $I_2, I_3, \ldots, I_K$ son variables explicativas de la variable respuesta.

Además se considera que el factor I_j tiene C_{I_j} categorías,

para j=1,...K; e i_j representa uno de los valores $1,2,\ldots,C_{I_j}$. Siendo, $n(i_1,i_2,\ldots,i_K)$ la frecuencia de la celda $(i_1,i_2,\ldots,i_K)$ donde

$$\sum_{i_1=1}^{C_{I_1}} \sum_{i_2=1}^{C_{I_2}} \cdots \sum_{i_K=1}^{C_{I_K}} n(i_1,i_2,\ldots,i_K) = N .$$

Por tanto,

$$n(i_1,i_2,\ldots,i_{K-1}) = \sum_{i_K=1}^{C_{I_K}} n(i_1,\ldots,i_{K-1},i_K) .$$

Por otro lado, por $p(i_1,i_2,\ldots,i_K)$ se denota la probabilidad de ocurrencia de la celda $(i_1,i_2,\ldots,i_K)$ con

$$\sum_{i_1=1}^{C_{I_1}} \sum_{i_2=1}^{C_{I_2}} \cdots \sum_{i_K=1}^{C_{I_K}} p(i_1,i_2,\ldots,i_K) = 1$$

de donde,

$$p(i_1,i_2,\ldots,i_{K-1}) = \sum_{i_K=1}^{C_{I_K}} p(i_1,i_2,\ldots,i_{K-1},i_K) .$$

Finalmente, por facilidad de notación

$$\sum_{i_1=1}^{C_{i_1}} \sum_{i_2=1}^{C_{i_2}} \cdots \sum_{i_K=1}^{C_{i_K}} \quad \text{equivale a escribir} \quad \sum_{i_1,i_2,\ldots,i_K} .$$

III.1 PRIMER CASO: CUANDO EL NUMERO DE VARIABLES EXPLICATIVAS ES RAZONABLE

Las interrelaciones entre muchas variables se pueden entender y comunicar más facilmente si se pueden caracterizar por un modelo de asociación. Los modelos multiplicativos forman una clase de tales patrones de asociación.

Existen algunas ventajas si se consideran solamente modelos multiplicativos en el estudio de bondad de ajuste de modelos para un conjunto de datos. Por ejemplo, sus interpretaciones son relativamente fáciles, porque en cada modelo se pueden distinguir por un lado, las variables que pertenecen a un conjunto de variables dadas y por otro, cuales de ellas se pueden separar.

La aplicación de los modelos multiplicativos para el análisis de tablas de contingencia ha sido discutido por Darroch (1962), Bishop (1969, 1975), Goodman (1970), Wermuth (1976, [22], [23]).

Como el objetivo es seleccionar un conjunto óptimo de variables explicativas de una variable respuesta, es conveniente hacer uso de los modelos multiplicativos para formular todos los modelos de asociación entre la variable respuesta $\{I_1\}$ y todos los posibles sub-conjuntos de variables explicativas $\{I_2,\ldots,I_k\}$, el investigador seleccionará el mejor modelo, mediante el criterio de selección MAIC y siguiendo el "principio de parsimonia". Es razonable pensar que el modelo seleccionado indique el conjunto óptimo de variables explicativas.

A continuación se presenta el procedimiento a seguir para la obtención del conjunto de variables explicativas.

Los modelos multiplicativos que se pueden establecer,

i) MODELO(0,1): $p(i_1,i_2,\ldots,i_K) = p(i_1,i_2,\ldots,i_K)$.

ii) Los modelos multiplicativos que miden la independencia de la variable respuesta $\{I_1\}$ y alguna variable explicativa $\{I_j\}$, $j=1,\ldots,K$ dado el resto de variables explicativas.

Para seguir cierto orden, en este tipo de modelo se presenta la independencia condicional entre las variables $\{I_1\}$ e $\{I_K\}$ dado $\{I_2,\ldots,I_{K-1}\}$, luego la independencia condicional entre $\{I_1\}$ e $\{I_{K-1}\}$ dado $\{I_2,\ldots,I_{K-2},I_K\}$ y así sucesivamente hasta la independencia condicional entre $\{I_1\}$ e $\{I_2\}$ dado $\{I_3,\ldots,I_K\}$.

Como se tiene K-1 variables explicativas, se pueden obtener C_1^{K-1} = K-1 modelos de este tipo. Estos modelos se presentan a continuación:

El modelo que expresa la independencia condicional entre $\{I_1\}$ e $\{I_K\}$ dado $\{I_2,\ldots,I_{K-1}\}$ es

MODELO(1,1):

$$p(i_1,i_2,\ldots,i_{K-1},i_K) = \frac{p(i_1,i_2,\ldots,i_{K-1})\ p(i_2,\ldots,i_{K-1},i_K)}{p(i_2,\ldots,i_{K-1})}\ .$$

El modelo que expresa la independencia condicional entre $\{I_1\}$ e $\{I_{K-1}\}$ dado $\{I_2,\ldots,I_{K-2},I_K\}$ es

MODELO(1,2):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2,\ldots,i_{K-2},i_K)\ p(i_2,\ldots,i_{K-1},i_K)}{p(i_2,\ldots,i_{K-2},i_K)}\ .$$

De manera similar se obtienen los otros modelos

MODELO(1,3):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2,\ldots,i_{K-3},i_{K-1},i_K)\ p(i_2,\ldots,i_{K-1},i_K)}{p(i_2,\ldots,i_{K-3},i_{K-1},i_K)}\ .$$

MODELO$\left(1, C_1^{K-1}\right)$:

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_3,\ldots,i_K)\; p(i_2,i_3,\ldots,i_K)}{p(i_3,\ldots,i_K)} .$$

iii) Los modelos multiplicativos que miden la independencia condicional entre $\{I_1\}$ e $\{I_i, I_j\}$, con $i \neq j$ $i,j=1,\ldots,K$, dado las variables explicativas restantes.

El número de modelos de este tipo que se pueden formular son C_2^{K-1}. Estos modelos se presentan seguidamente:

Asi, el modelo que expresa la independencia condicional entre $\{I_1\}$ e $\{I_{K-1}, I_K\}$ dado $\{I_2,\ldots,I_{K-2}\}$ es

MODELO(2,1):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2,\ldots,i_{K-2})\; p(i_2,\ldots,i_{K-2},i_{K-1},i_K)}{p(i_2,\ldots,i_{K-2})} .$$

El modelo que expresa la independencia condicional entre $\{I_1\}$ e $\{I_{K-2}, I_{K-1}\}$ dado $\{I_2,\ldots,I_{K-3},I_K\}$ es

MODELO(2,2):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2,\ldots,i_{K-3},i_K)\; p(i_2,\ldots,i_{K-1},i_K)}{p(i_2,\ldots,i_{K-3},i_K)} .$$

De manera similar se formulan los otros modelos

MODELO(2,3):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2,\ldots,i_{K-4},i_{K-1},i_K)\; p(i_2,\ldots,i_{K-1},i_K)}{p(i_2,\ldots,i_{K-4},i_{K-1},i_K)} .$$

.

.

.

$$\text{MODELO}\left(2, C_2^{K-1}\right):$$

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_4,\ldots,i_K)\; p(i_2,i_3,i_4,\ldots,i_K)}{p(i_4,\ldots,i_K)} .$$

De forma similar se formulan los modelos restantes hasta llegar por último a los modelos que expresan la independencia condicional entre $\{I_1\}$ y (K-2) variables explicativas dado la restante variable explicativa.

El número de modelos de este tipo que se pueden formular es C_{K-2}^{K-1}. Estos modelos se presentan a continuación:

El modelo que expresa la independencia condicional entre $\{I_1\}$ e $\{I_3,\ldots,I_K\}$ dado $\{I_2\}$ es

MODELO(K-2,1):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_2)\; p(i_2,i_3,\ldots,i_K)}{p(i_2)} .$$

De manera similar se obtienen los otros modelos

MODELO(K-2,2):

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_3)\; p(i_2,i_3,i_4,\ldots,i_K)}{p(i_3)} ,$$

.

.

.

$$\text{MODELO}\left(K-2, C_{K-2}^{K-1}\right):$$

$$p(i_1,\ldots,i_K) = \frac{p(i_1,i_K)\; p(i_2,\ldots,i_{K-1},i_K)}{p(i_K)} .$$

Finalmente tenemos que el modelo que expresa la

independencia entre $\{I_1\}$ e $\{I_2,\ldots,I_K\}$ es

$\text{MODELO}\left(K\text{-}1, C_{K-1}^{K-1}\right)$:

$p(i_1,\ldots,i_K) = p(i_1)\ p(i_2,\ldots,i_K)$.

Los modelos anteriores pueden ser denotados por el

$$\text{MODELO}(l,m)$$

El MODELO(l,m) representa la asociación entre la variable respuesta I_1 y un conjunto de "l " variables explicativas y esta asociación se expresa a través de la independencia condicional entre la variable respuesta I_1 y el conjunto de l variables explicativas dado las variables explicativas restantes.

Y "m" denota que el modelo MODELO(l,m) es el m-ésimo modelo de ese tipo de asociación.

Luego los modelos presentados anteriormente a excepción del $\text{MODELO}\left(K\text{-}1, C_{K-1}^{K-1}\right)$ pueden ser expresados de la siguiente manera,

$$\text{MODELO:}\quad p(i_1,I) = \frac{p(i_1,E)\ p(I)}{p(E)} \qquad (3.1)$$

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

Como todos los modelos que expresan la independencia condicional entre la variable $\{I_1\}$ e $\{I\text{-}E\}$ dado . E, están representados por el modelo (3.1), se observa que dicho modelo proporciona una descripción perfecta entre la variable respuesta y las variables explicativas.

Además, se nota que las variables que aparecen en el denominador de los diferentes modelos, es decir E, constituyen los posibles candidatos para ser la combinación optima de variables explicativas de una variable respuesta.

El criterio que se sigue para seleccionar dicha combinación

óptima es el criterio MAIC. Para aplicar este criterio se procede primero a hallar el AIC asociado al modelo (3.1). Los datos siguen una distribución Multinomial dada por

$$P = \frac{N!}{\prod_{i_1,\ldots,i_K} n(i_1,\ldots,i_K)!} \prod_{i_1,\ldots,i_K} \left(p(i_1,\ldots,i_K)\right)^{n(i_1,\ldots,i_k)} ,$$

donde $\prod_{i_1,\ldots,i_K}$ significa $\prod_{i_1=1}^{C_{I_1}} \prod_{i_2=1}^{C_{I_2}} \cdots \prod_{i_K=1}^{C_{I_K}}$.

De donde se obtiene que el logaritmo de la función de verosimilitud será

$$\ell = K_1 + \sum_{i_1,\ldots,i_K} n(i_1,\ldots,i_K) \log p(i_1,\ldots,i_K) , \qquad (3.2)$$

donde $K_1 = \log \frac{N!}{\prod_{i_1,\ldots,i_K} n(i_1,\ldots,i_K)!}$.

Bajo el modelo dado en (3.1) se obtiene que el logaritmo de la función de verosimilitud dada en (3.2) será

$$\ell = K_1 + \sum_{i_1,I} n(i_1,I) \log \frac{p(i_1,E)\, p(I)}{p(E)} , \qquad (3.3)$$

donde

$$\sum_{I} p(I) = 1 \quad , \quad \sum_{i_1,E} p(i_1,E) = 1 \quad , \quad \sum_{E} p(E) = 1 .$$

Los estimadores de máxima verosimilitud de $p(i_1,E)$, $p(I)$ y $p(E)$ son:

$$\hat{p}(i_1,E) = \frac{n(i_1,E)}{N}$$

$$\hat{p}(I) = \frac{n(I)}{N}$$

$$\hat{p}(E) = \frac{n(E)}{N} .$$

Sustituyendo los valores de $\hat{p}(i_1,E)$, $\hat{p}(I)$ y $\hat{p}(E)$ en (3.3) se obtiene

$$\ell = K_1 + \sum_{i_1,I} n(i_1,I) \log \frac{n(i_1,E)\, n(I)}{N\, n(E)} ,$$

luego el AIC asociado al modelo (3.1) es

$$AIC = -2\left[K_1 + \left(\sum_{i_1,I} n(i_1,I) \log \frac{n(i_1,E)\, n(I)}{N\, n(E)}\right)\right] +$$

$$+ 2\left[(C_{I_1} C_E - 1) + (C_I - 1) - (C_E - 1)\right]$$

donde C_E y C_I denotan el número de categorías de los correspondientes conjuntos de variables. Además en el caso del $MODELO\left(K-1, C_{K-1}^{K-1}\right)$, $n(E) = N$ y $C_E = 1$ en la expresión anterior.

El AIC se puede reescribir en la forma,

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

$$+ \left(-2 K_1 - 2 \sum_{I} n(I) \log\left(\frac{n(I)}{N}\right) + 2 (C_I - 1)\right) . \qquad (3.4)$$

Se observa que los elementos del segundo corchete de (3.4) constituyen una constante común de los AIC de todos los modelos, por lo que la podemos omitir para efectos de comparación de los modelos. Así, el estadístico AIC "reducido" del modelo (3.1) será

$$AIC^{*} = -2 \sum_{i_1,E} n(i_1,E) \log\left(\frac{n(i_1,E)}{n(E)}\right) + 2\left[(C_{I_1} C_E - 1) - (C_E - 1)\right]. \qquad (3.5)$$

En base a (3.5) obtenemos el conjunto óptimo de variables explicativas de la variable respuesta de la siguiente manera:

Considerando (3.5) se calcula el AIC^{*} de todos los modelos de la forma (3.1). Luego se selecciona aplicando el criterio MAIC el mejor modelo dentro de 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.

Se puede observar que con los test clásicos es imposible comparar el MODELO(1,1) con el MODELO(1,2) en cambio mediante el estadístico AIC y el criterio MAIC podemos efectuar dicha comparación.

Por consiguiente, el procedimiento presentado en esta sección, basado en el AIC y en el criterio de selección MAIC,

proporciona una herramienta muy útil para obtener el conjunto óptimo de variables explicativas de la variable respuesta.

III.2 SEGUNDO CASO: CUANDO EL NUMERO DE VARIABLES EXPLICATIVAS ES DEMASIADO GRANDE

Generalmente los datos investigados en un cuestionario contienen cientos de respuestas, ésto origina que el número de variables y el número de categorías aumente. Por consiguiente el número de celdas de la tabla de contingencia correspondiente a los datos se ve incrementado considerablemente, por lo que es muy probable que la tabla contenga muchas celdas con frecuencia cero. Además, origina que el número de modelos a ser analizados crezca demasiado.

La existencia de muchas celdas vacías hace necesario reordenar la información para su posterior análisis.

A continuación se presenta un método para seleccionar el conjunto óptimo de variables explicativas cuando se está en la situación en la que el número de variables explicativas es demasiado grande.

Este método está basado en el estadístico AIC y en el criterio MAIC y se aplica en dos etapas:

i) En la primera, se efectúa una preselección de variables explicativas.

ii) En la segunda, tomando como base las variables explicativas preseleccionadas en la primera etapa se efectúa una nueva selección para obtener así el conjunto óptimo de variables explicativas.

Este método se presenta a continuación:

Considérese una tabla de contingencia de K factores de clasificación $I_1, I_2, \ldots, I_K$, donde I_1 es la variable respuesta e $I = \{I_2, \ldots, I_K\}$ es un conjunto de variables explicativas.

Sean E y F sub-conjuntos de $I^{\bullet} = \{I_1, I_2, \ldots, I_K\}$ tales que E y F sean disjuntos.

Para efectos de este estudio, se considera $E = \{I_1\}$ y F como un sub-conjunto de $I = \{I_2, \ldots, I_K\}$, donde P(E) y P(F) son las probabilidades de E y F respectivamente.

Para medir la fuerza de dependencia entre E y F, usaremos como criterio que una dependencia fuerte entre E y F equivale a afirmar que dado F, I-F y E son independientes; es decir, conocido F, I-F no proporciona ninguna explicación adicional sobre E. Se define el modelo

$$\text{MODELO(E,F):} \quad P(I^{\bullet}) = \frac{P(E,F)\, P(I^{\bullet} - E)}{P(F)} . \qquad (3.6)$$

A continuación se va a obtener su estadistico AIC.

Como los datos siguen una distribución Multinomial con el logaritmo de la función de verosimilitud dada en (3.2).

Se tiene que bajo el modelo (3.6), el logaritmo de la función de verosimilitud dada en (3.2) será

$$\ell = K_1 + \sum_{I^{\bullet}} n(I^{\bullet}) \log \frac{P(E,F)\, P(I^{\bullet} - E)}{P(F)} , \qquad (3.7)$$

donde $K_1 = \log \dfrac{N!}{\prod_{I^{\bullet}} n(I^{\bullet})!}$.

De (3.7) se tiene que los estimadores de máxima verosimilitud de P(E,F) y $P(I^{\bullet} - E)$, P(F) son

$$\hat{P}(E,F) = \frac{n(E,F)}{N}$$

$$\hat{P}(I^{\bullet} - E) = \frac{n(I^{\bullet} - E)}{N}$$

$$\hat{P}(F) = \frac{n(F)}{N}$$

reemplazando estos valores en (3.7) se obtiene

$$\ell = K_1 + \sum_{I^\bullet} n(I^\bullet) \log \frac{n(E,F)\, n(I^\bullet - E)}{N\, n(F)} .$$

Luego el AIC asociado viene dado por:

$$AIC(E,F) = -2 \left[K_1 + \sum_{I^\bullet} n(I^\bullet) \log \frac{n(E,F)\, n(I^\bullet - E)}{N\, n(F)} \right] +$$

$$+ 2 \left[\left(C_E\, C_F - 1 \right) + \left(C_{(I^\bullet - E)} - 1 \right) - \left(C_F - 1 \right) \right] \qquad (3.8)$$

donde, C_E , C_F , $C_{(I^\bullet - E)}$ denotan el número de categorías de los correspondientes conjuntos de variables.

Si sumamos y restamos en (3.8) la expresión

$$2 \sum_{I^\bullet} n(I^\bullet) \log \left(\frac{n(E)}{N} \right) + 2 \left[C_E - 1 \right]$$

y reordenando se tiene

$$AIC(E,F) = \left(-2 \sum_{E,F} n(E,F) \log \left(\frac{N\, n(E,F)}{n(E)\, n(F)} \right) + 2 \left(C_E - 1 \right) \left(C_F - 1 \right) \right) +$$

$$+ \left(-2\, K_1 - 2 \sum_{I^\bullet} n(I^\bullet) \log \left(\frac{n(I^\bullet - E)\, n(E)}{N^2} \right) \right) +$$

$$+ 2\left[\left(C_{(I^{*}-E)} - 1\right) + \left(C_{E} - 1\right)\right]\Bigg) \; . \tag{3.9}$$

Se observa que el segundo corchete de (3.9) es una constante común en todos los AIC(E,F) de los modelos, por lo cual se omite a efectos de comparación.

Así, el AIC simplificado es

$$AIC^{*}(E,F) = -2 \sum_{E,F} n(E,F) \log \left(\frac{N\, n(E,F)}{n(E)\, n(F)}\right) + 2\left(C_{E} - 1\right)\left(C_{F} - 1\right) \tag{3.10}$$

y se supone que $n(\phi) = N$, $C_{\phi} = 1$.

Se observa que el estadístico dado en (3.10) mide la fuerza de dependencia entre E y F. Además, permite efectuar la comparación de varios modelos del tipo (3.6) sin utilizar la tabla de contingencia completa.

El método que se presenta a continuación para obtener el conjunto óptimo de variables explicativas está basado en el modelo (3.6) y en su AIC^{*} asociado dado en (3.10).

A continuación se pasa a describir el método:

En la primera etapa se realizará una preselección, para éllo primero se medirá al grado de dependencia entre la variable respuesta $\{I_1\}$ y cada una de las variables explicativas. Para ésto se considera $E = \{I_1\}$ y $F = \{I_j\}$, para algún $j=2,\ldots,k$ en el modelo (3.6). Y para medir el grado de dependencia entre E y F se utiliza el estadístico AIC^{*} dado en (3.10) en cada uno de los modelos formulados.

Así, se obtiene

$$AIC^{*}(I_1, I_j) = -2 \sum_{i_1, i_j} n(i_1, i_j) \log \frac{N\, n(i_1, i_j)}{n(i_1)\, n(i_j)} + 2\left(C_{I_1} - 1\right)\left(C_{I_j} - 1\right) \qquad (3.11)$$

para j=2,...,K.

Se ordenan , posteriormente, los AIC^{*} de los (K-1) modelos de menor a mayor y siguiendo el criterio MAIC los primeros AIC indicarán las variables explicativas más relacionadas con la variable respuesta. Esto permitirá elegir las variables explicativas más significativas.

En segunda etapa, se toma como base el conjunto de variables explicativas preseleccionadas para obtener el conjunto óptimo de variables explicativas. El procedimiento que se sigue para éllo es similar al primer caso y es el siguiente:

Se calcula el AIC^{*} dado en (3.10) de todos los modelos $MODELO(I_1, F)$ dado en (3.6), donde F son todos los posibles subconjuntos del conjunto de variables explicativas preseleccionadas. Aplicando el criterio MAIC seleccionamos el mejor modelo de cada grupo de modelos: I_1 frente a todas las variables explicativas preseleccionadas, ..., I_1 frente a una variable explicativa preseleccionada. El investigador seleccionará entre los modelos seleccionados de cada grupo, al mejor modelo siguiendo el "principio de parsimonia".

El conjunto F del modelo seleccionado será el conjunto óptimo de variables explicativas de la variable respuesta I_1.

Otra forma de elegir el conjunto óptimo de variables explicativas preseleccionadas es la siguiente:

Se calcula el AIC^{*} dado en (3.5) de todos los modelos $MODELO(I_1, E)$, donde E son todos los subconjuntos del conjunto de variables explicativas preseleccionadas y se procede a seleccionar al mejor modelo de forma similar al procedimiento antes descrito. Así, el conjunto E del modelo seleccionado será el conjunto óptimo de variables explicativas de la variable respuesta I_1.

CAPITULO IV

"DISCUSION DEL CRITERIO DE INFORMACION DE AKAIKE EN EL ANALISIS DATOS CATEGORIZADOS. PROGRAMAS DE LOS METODOS EXPUESTOS. APLICACIONES"

IV.3.- PRESENTACION DE LOS PROBLEMAS Y APLICACION DE LOS METODOS

A. ESTUDIO DEL PARO EN ESPAÑA

B. ESTUDIO DE LA FECUNDIDAD

IV.1 SUMARIO

En la Sección IV.1 de este capitulo, se presenta una discusión del criterio de información de Akaike en el análisis de datos categorizados y en la Sección IV.2 los programas en Basic de los métodos presentados en los capitulos II y III. Finalmente en la Sección IV.3 se consideran unos datos reales y se les estudia y analiza en base a los métodos expuestos en esta memoria.

IV.1 DISCUSION DEL CRITERIO DE INFORMACION DE AKAIKE EN EL ANALISIS DE DATOS CATEGORIZADOS

El análisis de datos categorizados ha sido objeto de una especial atención en los últimos años y ha dado lugar a que aparezcan nuevos procedimientos para su análisis. De entre ellos sobresale, el método propuesto por Akaike (1973), el cual está basado en el estadistico AIC y el criterio MAIC. La importancia de este criterio radica en que a la vista de los resultados permite seleccionar el modelo que mejor justifique los datos sin necesidad de fijar un nivel de significación, esto es una ventaja sobre los métodos tradicionales como el test Ji-Cuadrado, los modelos log-lineal (Bishop, Fienberg y Holland (1975)), el método de ajuste iterativo de Wermuth (Wermuth (1976) [23]) e inclusive sobre el estadistico G de Kullback, es decir, la medida de discriminación de Kullback-Leibler (Kullback (1959)), etc.

Otra ventaja que tiene el procedimiento propuesto por Akaike es que se pueden efectuar comparaciones entre una gran diversidad de modelos, cosa que con los test clásicos no se puede realizar. Akaike y Sakamoto (1978) probaron que el procedimiento basado en el AIC y el criterio MAIC es más eficiente que el test Ji-Cuadrado.

En el Capitulo I se vio que el estadistico AIC y el criterio MAIC surge de la idea de minimizar la medida de

discriminación esperada de Kullback-Leibler entre la verdadera distribución y la hipótetica.

En esta memoria se ha presentado un método basado en el AIC y el criterio MAIC para el análisis de datos categorizados para:

i) El análisis de tablas de contingencia.

ii) La obtención del conjunto óptimo de variables explicativas de una variable respuesta.

Como ya se ha indicado repetidamente a lo largo de la memoria para el análisis de tablas de contingencia existen varios métodos como el test Ji-Cuadrado, G de Kullback, los modelos log lineal, el proceso iterativo de Wermuth, el estadistico AIC con el criterio MAIC, y es sabido que el método basado en el estadistico G de Kullback es preferible al test Ji-Cuadrado y otros test clásicos en el análisis de tablas de contingencia para examinar la independencia, homogeneidad e interacción. Sin embargo, presenta un problema al analizar la interacción entre las tres variables, pues el estadistico que surge sigue una distribución de Wittaker, lo que hace que su cálculo sea engorroso.

Bishop, Fienberg y Holland (1975) propusieron un método alternativo que son los modelos log lineal para el análisis de independencia e interacción entre los tres factores de clasificación.

Wermuth (1976) presenta un método para el análisis de independencia basado en un proceso de ajuste iterativo, éste no necesariamente selecciona el modelo con los mismos patrones de asociación que podría formular el investigador, es decir, que este método podría llevar a una interpretación equivocada.

En el Capítulo II se presentó un método para el análisis de tablas de contingencia de tres factores de clasificación basado en el estadistico AIC, criterio MAIC y en las relaciones entre las diferentes hipótesis formuladas para el análisis de una tabla de contingencia dadas por Kullback (1959). Una ventaja importante que posee el método que se propone es que no sólo da los estadisticos AIC para todas las hipótesis planteadas en una tabla de tres factores, sino que además proporciona las relaciones que existen entre las diferentes hipótesis, lo cual permite realizar un análisis más detallado de los datos, pues se pueden detectar las

verdaderas causas de la independencia, homogeneidad e interacción de los tres factores de clasificación.

En el Capítulo III, se presentó un método para obtener un conjunto óptimo de variables explicativas de una variable respuesta en los siguientes términos:

Una vez formulados todos los modelos de asociación entre la variable respuesta y todos los subconjuntos de variables explicativas, se calcula el AIC de cada modelo y se aplica el criterio MAIC y el "principio de parsimonia" para obtener el mejor modelo, el modelo elegido indicará el conjunto óptimo de variables explicativas. Es interesante observar, que la comparación entre modelos tan diversos no se podría efectuar con los métodos tradicionales

En el análisis de este tipo de problemas se presentan dos casos, el primero cuando el número de variables explicativas es razonablemente manejable y otra cuando el número de variables explicativas es demasiado grande, en este último caso se deberá efectuar una preselección de variables explicativas antes de pasar a la selección del conjunto óptimo de estás. El tratamiento de estos casos se han analizado de forma detallada en el Capítulo III.

En el caso de que la tabla de contingencia presenta muchas celdas vacías, se deberá reordenar para su posterior análisis. En esta reordenación (es decir, en la unión de categorías de los factores de clasificación) intervendría el criterio de investigador. Si la variable es cuantitativa, la ordenación también se puede efectuar a través del proceso de reordenación que presenta Sakamoto (1977).

IV.2 PROGRAMAS DE LOS METODOS EXPUESTOS

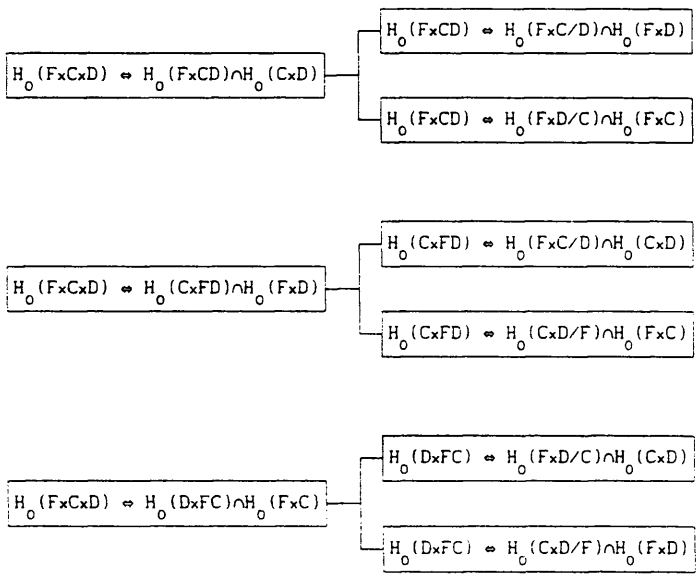
En esta sección se exponen los programas en Basic de los métodos expuestos en los capítulos II y III. Estos, permiten analizar tablas de contingencia de tres factores de clasificación y seleccionar el conjunto óptimo de variables explicativas de una variable respuesta.

Los programas se presentan a continuación.

A. PROGRAMA QUE ANALIZA TABLAS DE CONTINGENCIA DE TRES FACTORES DE CLASIFICACION MEDIANTE EL CRITERIO MAIC (MINIMO AIC)

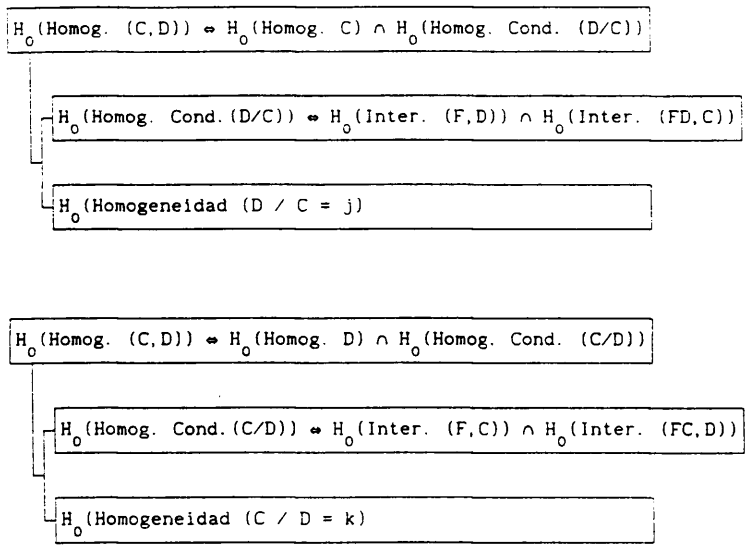
Este programa analiza la independencia, homogeneidad e interacción entre los factores. En cuanto a la independencia de factores se realizan todas las hipótesis que de ella se desprenden para realizar un análisis minucioso de las causas. Las hipótesis que se analizan y sus relaciones son presentadas en el grafico 1.

GRAFICO 1.- ANALISIS DE LA INDEPENDENCIA H_0(F x C x D)



En lo que respecta al análisis de homogeneidad de los factores se realizan todas las hipótesis que de ella se desprenden a fin de realizar un análisis pormenorizado de las causas. Las hipótesis que se realizan y sus relaciones son presentadas en el grafico 2.

GRAFICO 2.- ANALISIS DE LA HOMOGENEIDAD H_0(HOMOGENEIDAD (C,D))



donde:

Homog. : Homogeneidad

Homog. Cond. : Homogeneidad Condicional

Inter. : Interaccion

Por último se analiza la interacción entre los tres factores de clasificación.

Los datos de entrada que necesita este programa son:

X(I,J,K): Matriz de datos
N1: Número de categorias del factor Fila (1^{o})
N2: Número de categorias del factor Columna (2^{o})
N3: Número de categorias del factor Profundidad (3^{o})
L: Opción (indica que hipotesis se desea contrastar)

En base a estos datos de entrada el programa calcula los estadisticos:

A0: Estadistico asociado al MODELO(0)
A1: Estadistico asociado al MODELO(1)

correspondientes al contraste de hipótesis que se desea contrastar.

El listado del programa se presenta a continuación.

```
10 REM ************************************************************
20 REM ** PROGRAMA QUE ANALIZA TABLAS DE CONTINGENCIA DE 3 FACTORES **
30 REM ** DE CLASIFICACION MEDIANTE EL CRITERIO MAIC (MINIMO AIC)   **
40 REM ************************************************************
50 REM * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
60 REM ************************************************************
70 REM **                  DICCIONARIO DE VARIABLES                **
80 REM **                  ---------- -- --------                  **
90 REM ************************************************************
100 REM **                                                         **
110 REM ** X(I,J,K): MATRIZ DE DATOS                               **
120 REM ** N1: NUMERO DE CATEGORIAS DEL FACTOR FILA (1°)           **
130 REM ** N2: NUMERO DE CATEGORIAS DEL FACTOR CCLUMNA (2°)        **
140 REM ** N3: NUMERO DE CATEGORIAS DEL FACTOR PROFUNDIDAD (3°)    **
150 REM **                                                         **
160 REM ** L:  OPCION (INDICA LA OPCION QUE SE DESEA CONTRASTAR)   **
170 REM **      DONDE:                                             **
180 REM **                                                         **
190 REM **      ┌──────────────────────────────────────────┐       **
200 REM **      │  L  │        HIPOTESIS NULA              │       **
210 REM **      ├─────┼────────────────────────────────────┤       **
220 REM **      │  1  │ LOS TRES FACTORES SON TINDEPENDIENTES │    **
230 REM **      │  2  │ HOMOGENEIDAD (C, D)                │       **
240 REM **      │  3  │ INTERACCION FCD                    │       **
250 REM **      │  4  │ TERMINA LA EJECUCION DEL PROGRAMA  │       **
260 REM **      └─────┴────────────────────────────────────┘       **
270 REM **                                                         **
280 REM ** ESTADISTICOS ASOCIADOS A LAS HIPOTESIS QUE SE DESEAN   **
290 REM ** CONTRASTAR:                                            **
300 REM **                                                         **
310 REM ** A0: ESTADISTICO ASOCIADO AL MODELO(0)                  **
320 REM ** A1: ESTADISTICO ASOCIADO AL MODELO(1)                  **
330 REM **                                                         **
340 REM ************************************************************
350 REM * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
360 REM ************************************************************
370 DIM X(12,12,12), X1(12,12), X2(12,12), X3(12,12)
380 DIM X12(12), X13(12), X23(12), Y(12,12), Y1(12,12)
```

```
390 S1= 0
400 S2= 0
410 S3= 0
420 S4= 0
430 S5= 0
440 S6= 0
450 S7= 0
460 S8= 0
470 S9= 0
480 S10= 0
490 REM ************************************************************
500 REM *              LECTURA DE LOS DATOS DEL PROBLEMA             *
510 REM ************************************************************
520 PRINT "         INGRESO DE DATOS"
530 PRINT "         ******* ** *****"
540 PRINT
550 INPUT "NUMERO DE CATEGORIAS DEL 1º FACTOR= ";N1
560 INPUT "NUMERO DE CATEGORIAS DEL 2º FACTOR= ";N2
570 INPUT "NUMERO DE CATEGORIAS DEL 3º FACTOR= ";N3
580 PRINT
590 PRINT "INTRODUCCION DE LA MATRIZ DE DATOS"
600 PRINT "************ ** ** ****** ** *****"
610 PRINT
620 FOR I=1 TO N1
630 X23(I)= 0
640 FOR J=1 TO N2
650 X3(I,J)= 0
660 FOR K=1 TO N3
670 PRINT "X(";I;",";J;",";K;")=";
680 INPUT X(I,J,K)
690 S1= S1 + X(I,J,K)
700 IF X(I,J,K)= 0 THEN GOTO 720
710 S2= S2 + X(I,J,K)*LOG(X(I,J,K))
720 X3(I,J)= X3(I,J) + X(I,J,K)
730 NEXT K
740 IF X3(I,J)= 0 THEN GOTO 760
750 S3= S3 + X3(I,J)*LOG(X3(I,J))
760 X23(I)= X23(I) + X3(I,J)
```

```
770 NEXT J
780 IF X23(I)= 0 THEN GOTO 800
790 S4= S4 + X23(I)*LOG(X23(I))
800 NEXT I
810 S= S1*LOG(S1)
820 FOR J=1 TO N2
830 X13(J)= 0
840 FOR K=1 TO N3
850 X1(J,K)= 0
860 FOR I=1 TO N1
870 X1(J,K)= X1(J,K) + X(I,J,K)
880 NEXT I
890 IF X1(J,K)= 0 THEN GOTO 910
900 S7= S7 + X1(J,K)*LOG(X1(J,K))
910 X13(J)= X13(J) + X1(J,K)
920 NEXT K
930 IF X13(J)= 0 THEN GOTO 950
940 S8= S8 + X13(J)*LOG(X13(J))
950 NEXT J
960 FOR K=1 TO N3
970 X12(K)= 0
980 FOR I=1 TO N1
990 X2(I,K)= 0
1000 FOR J=1 TO N2
1010 X2(I,K)= X2(I,K) + X(I,J,K)
1020 NEXT J
1030 IF X2(I,K)= 0 THEN GOTO 1050
1040 S5= S5 + X2(I,K)*LOG(X2(I,K))
1050 X12(K)= X12(K) + X2(I,K)
1060 NEXT I
1070 IF X12(K)= 0 THEN GOTO 1090
1080 S6 = S6 + X12(K)*LOG(X12(K))
1090 NEXT K
1100 FOR I=1 TO N1
1110 FOR K=1 TO N3
1120 Y(I,K)= 0
1130 FOR J=1 TO N2
1140 Y(I,K)= Y(I,K) + (X3(I,J)*X1(J,K)/X13(J))
```

```
1150 NEXT J
1160 IF Y(I,K)= 0 THEN GOTO 1180
1170 S9= S9 + X2(I,K)*LOG(Y(I,K))
1180 NEXT K
1190 NEXT I
1200 FOR I=1 TO N1
1210 FOR J=1 TO N2
1220 Y1(I,J)= 0
1230 FOR K=1 TO N3
1240 Y1(I,J)= Y1(I,J) + (X2(I,K)*X1(J,K)/X12(K))
1250 NEXT K
1260 IF Y1(I,J)= 0 THEN GOTO 1280
1270 S10= S10 + X3(I,J)*LOG(Y1(I,J))
1280 NEXT J
1290 NEXT I
1300 REM ************************************************************
1310 REM *                    SELECCION DE LA OPCION                  **
1320 REM ************************************************************
1330 PRINT "OPCIONES   HIPOTESIS"
1340 PRINT "********   *********"
1350 PRINT
1360 PRINT "     1       Ho(F x C x D)"
1370 PRINT "     2       Ho(HOMOGENEIDAD(C, D))"
1380 PRINT "     3       Ho(INTERACCION FCD)"
1390 PRINT "     4       TERMINA"
1400 INPUT "OPCION = ";L
1410 IF L>2 THEN GOTO 1440
1420 IF L=2 THEN GOSUB 4230
1430 GOSUB 1460
1440 IF L=4 THEN END
1450 GOSUB 6380
1460 REM ************************************************************
1470 REM * SUBRUTINA QUE ANALIZA LA HIPOTESIS CORRESPONDIENTE A L=1:  *
1480 REM * LOS TRES FACTORES SON INDEPENDIENTES                       *
1490 REM ************************************************************
1500 A0= -2*(S4 + S8 +S6 - 3*S) + 2*(N1 +N2 +N3 -3)
1510 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
1520 LPRINT "RESULTADO DE ANALIZAR Ho(F x C x D)"
```

```
1530 LPRINT "AIC(0) = ";A0
1540 LPRINT "AIC(1) = ";A1
1550 IF A0 > A1 THEN GOTO 1590
1560 LPRINT "DECISION: ACEPTAR Ho(F x C x D)"
1570 LPRINT
1580 GOTO 1610
1590 LPRINT "DECISION: RECHAZAR Ho(F x C x D)"
1600 LPRINT
1610 REM ************************************************************
1620 REM * ANALISIS  Ho(F x C x D) <=> Ho(F x CD) Y Ho(C x D)        *
1630 REM ************************************************************
1640 LPRINT "   Ho(F x C x D) <=> Ho(F x CD) Y Ho(C x D)"
1650 LPRINT
1660 REM * ANALISIS  Ho(C x D) *
1670 A0= -2*(S8 + S6 -2*S) + 2*(N2 + N3 -2)
1680 A1= -2*(S7 - S) + 2*(N2*N3 -1)
1690 LPRINT "   RESULTADO DE ANALIZAR Ho(C x D)"
1700 LPRINT "   AIC(0) = ";A0
1710 LPRINT "   AIC(1) = ";A1
1720 IF A0 > A1 THEN GOTO 1760
1730 LPRINT "   DECISION: ACEPTAR Ho(C x D)"
1740 LPRINT
1750 GOTO 1780
1760 LPRINT "   DECISION: RECHAZAR Ho(C x D)"
1770 LPRINT
1780 REM * ANALISIS  Ho(F x CD) *
1790 A0= -2*(S4 + S7 - 2*S) + 2*(N2*N3 + N1 -2)
1800 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
1810 LPRINT "   RESULTADO DE ANALIZAR Ho(F x CD)"
1820 LPRINT "   AIC(0) = ";A0
1830 LPRINT "   AIC(1) = ";A1
1840 IF A0 > A1 THEN GOTO 1880
1850 LPRINT "   DECISION: ACEPTAR Ho(F x CD)"
1860 LPRINT
1870 GOTO 1900
1880 LPRINT "   DECISION: RECHAZAR Ho(F x CD)"
1890 LPRINT
```

```
1900 REM ************************************************************
1910 REM * ANALISIS  Ho(F x CD) <=> Ho(F x C / D) Y Ho(F x D)          *
1920 REM ************************************************************
1930 LPRINT "      Ho(F x CD) <=> Ho(F x C/D) Y Ho(F x D)"
1940 LPRINT
1950 REM * ANALISIS Ho(F x C/D) *
1960 A0 = -2*(S5 + S7 - S6 - S) + 2*(N1*N3 + N2*N3 - N3 -1)
1970 A1 = -2*(S2 - S) + 2*(N1*N2*N3 - 1)
1980 LPRINT "      RESULTADO DE ANALIZAR Ho(F x C/D)"
1990 LPRINT "      AIC(0) = ";A0
2000 LPRINT "      AIC(1) = ";A1
2010 IF A0 > A1 THEN GOTO 2050
2020 LPRINT "      DECISION: ACEPTAR Ho(F x C/D)"
2030 LPRINT
2040 GOTO 2070
2050 LPRINT "      DECISION: RECHAZAR Ho(F x C/D)"
2060 LPRINT
2070 REM * ANALISIS Ho(F x D) *
2080 A0= -2*(S4 + S6 -2*S) + 2*(N1 + N3 -2)
2090 A1= -2*(S5 -S) + 2*(N1*N3 -1)
2100 LPRINT "      RESULTADO DE ANALIZAR Ho(F x D)"
2110 LPRINT "      AIC(0) = ";A0
2120 LPRINT "      AIC(1) = ";A1
2130 IF A0 > A1 THEN GOTO 2170
2140 LPRINT "      DECISION: ACEPTAR Ho(F x D)"
2150 LPRINT
2160 GOTO 2190
2170 LPRINT "      DECISION: RECHAZAR Ho(F x D)"
2180 LPRINT
2190 REM ************************************************************
2200 REM * ANALISIS  Ho(F x CD) <=> Ho(F x D / C) Y Ho(F x C)          *
2210 REM ************************************************************
2220 LPRINT "      Ho(F x CD) <=> Ho(F x D/C) Y Ho(F x C)"
2230 LPRINT
2240 REM * ANALISIS  Ho(F x D/C) *
2250 A0= -2*(S3 + S7 - S8 - S) + 2*(N1*N2 + N2*N3 - N2 -1)
2260 A1= -2*(S2 - S) + 2*(N1*N2*N3 - 1)
2270 LPRINT "      RESULTADO DE ANALIZAR Ho(F x D/C)"
```

```
2280 LPRINT "      AIC(0) = ";A0
2290 LPRINT "      AIC(1) = ";A1
2300 IF A0 > A1 THEN GOTO 2340
2310 LPRINT "      DECISION: ACEPTAR Ho(F x D/C)"
2320 LPRINT
2330 GOTO 2360
2340 LPRINT "      DECISION: RECHAZAR Ho(F x D/C)"
2350 LPRINT
2360 REM * ANALISIS Ho(F x C) *
2370 A0= -2*(S4 + S8 - 2*S) + 2*(N1 + N2 -2)
2380 A1= -2*(S3 - S) + 2*(N1*N2 -1)
2390 LPRINT "      RESULTADO DE ANALIZAR Ho(F x C)"
2400 LPRINT "      AIC(0) = ";A0
2410 LPRINT "      AIC(1) = ";A1
2420 IF A0 > A1 THEN GOTO 2460
2430 LPRINT "      DECISION: ACEPTAR Ho(F x C)"
2440 LPRINT
2450 GOTO 2480
2460 LPRINT "      DECISION: RECHAZAR Ho(F x C)"
2470 LPRINT
2480 REM ************************************************************
2490 REM * ANALISIS  Ho(F x C x D) <=> Ho(C x FD) Y Ho(F x D)        *
2500 REM ************************************************************
2510 LPRINT "   Ho(F x C x D) <=> Ho(C x FD) Y Ho(F x D)"
2520 LPRINT
2530 REM * ANALISIS  Ho(F x D) *
2540 A0= -2*(S4 + S6 - 2*S) + 2*(N1 + N3 -2)
2550 A1= -2*(S5 -S) + 2*(N1*N3 -1)
2560 LPRINT "   RESULTADO DE ANALIZAR Ho(F x D)"
2570 LPRINT "   AIC(0) = ";A0
2580 LPRINT "   AIC(1) = ";A1
2590 IF A0 > A1 THEN GOTO 2630
2600 LPRINT "   DECISION: ACEPTAR Ho(F x D)"
2610 LPRINT
2620 GOTO 2650
2630 LPRINT "   DECISION: RECHAZAR Ho(F x D)"
2640 LPRINT
2650 REM * ANALISIS  Ho(C x FD) *
```

```
2660 A0= -2*(S8 + S5 - 2*S) + 2*(N1*N3 + N2 -2)
2670 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
2680 LPRINT "   RESULTADO DE ANALIZAR Ho(C x FD)"
2690 LPRINT "   AIC(0) = ";A0
2700 LPRINT "   AIC(1) = ";A1
2710 IF A0 > A1 THEN GOTO 2750
2720 LPRINT "   DECISION: ACEPTAR Ho(C x FD)"
2730 LPRINT
2740 GOTO 2770
2750 LPRINT "   DECISION: RECHAZAR Ho(C x FD)"
2760 LPRINT
2770 REM ************************************************************
2780 REM * ANALISIS Ho(C x FD) <=> Ho(F x C / D) Y Ho(C x D)          *
2790 REM ************************************************************
2800 LPRINT "      Ho(C x FD) <=> Ho(F x C/D) Y Ho(C x D)"
2810 LPRINT
2820 REM * ANALISIS  Ho(F x C/D) *
2830 A0= -2*(S5 + S7 - S6 - S) + 2*(N1*N3 + N2*N3 - N3 -1)
2840 A1= -2*(S2 - S) + 2*(N1*N2*N3 -1)
2850 LPRINT "      RESULTADO DE ANALIZAR Ho(F x C/D)"
2860 LPRINT "      AIC(0) = ";A0
2870 LPRINT "      AIC(1) = ";A1
2880 IF A0 > A1 THEN GOTO 2920
2890 LPRINT "      DECISION: ACEPTAR Ho(F x C/D)"
2900 LPRINT
2910 GOTO 2940
2920 LPRINT "      DECISION: RECHAZAR Ho(F x C/D)"
2930 LPRINT
2940 REM * ANALISIS  Ho(C x D) *
2950 A0= -2*(S8 + S6 - 2*S) + 2*(N2 + N3 -2)
2960 A1= -2*(S7 -S) + 2*(N2*N3 -1)
2970 LPRINT "      RESULTADO DE ANALIZAR Ho(C x D)"
2980 LPRINT "      AIC(0) = ";A0
2990 LPRINT "      AIC(1) = ";A1
3000 IF A0 > A1 THEN GOTO 3040
3010 LPRINT "      DECISION: ACEPTAR Ho(C x D)"
3020 LPRINT
3030 GOTO 3060
```

```
3040 LPRINT "      DECISION: RECHAZAR Ho(C x D)"
3050 LPRINT
3060 REM ************************************************************
3070 REM * ANALISIS  Ho(C X FD) <=> Ho(C x D / F) Y Ho(F x C)        *
3080 REM ************************************************************
3090 LPRINT "      Ho(C x FD) <=> Ho(C x D/F) Y Ho(F x C)"
3100 LPRINT
3110 REM * ANALISIS  Ho(C x D/F) *
3120 A0= -2*(S5 + S3 - S4 - S) + 2*(N1*N3 + N1*N2 - N1 -1)
3130 A1= -2*(S2 - S) + 2*(N1*N2*N3 -1)
3140 LPRINT "      RESULTADO DE ANALIZAR Ho(C x D/F)"
3150 LPRINT "      AIC(0) = ";A0
3160 LPRINT "      AIC(1) = ";A1
3170 IF A0 > A1 THEN GOTO 3210
3180 LPRINT "      DECISION: ACEPTAR Ho(C x D/F)"
3190 LPRINT
3200 GOTO 3230
3210 LPRINT "      DECISION: RECHAZAR Ho(C x D/F)"
3220 LPRINT
3230 REM * ANALISIS  Ho(F x C) *
3240 A0= -2*(S4 + S8 - 2*S) + 2*(N1 + N2 -2)
3250 A1= -2*(S3 - S) + 2*(N1*N2 -1)
3260 LPRINT "      RESULTADO DE ANALIZAR Ho(F x C)"
3270 LPRINT "      AIC(0) = ";A0
3280 LPRINT "      AIC(1) = ";A1
3290 IF A0 > A1 THEN GOTO 3330
3300 LPRINT "      DECISION: ACEPTAR Ho(F x C)"
3310 LPRINT
3320 GOTO 3350
3330 LPRINT "      DECISION: RECHAZAR Ho(F x C)"
3340 LPRINT
3350 REM ************************************************************
3360 REM *   ANALISIS Ho(F x C x D) <=> Ho(D x FC) Y Ho(F x C)       *
3370 REM ************************************************************
3380 LPRINT "   Ho(F x C x D) <=> Ho(D x FC) Y Ho(F x C)"
3390 LPRINT
3400 REM * ANALISIS  Ho(F x C) *
3410 A0= -2*(S4 + S8 - 2*S) + 2*(N1 + N2 -2)
```

```
3420 A1= -2*(S3 - S) + 2*(N1*N2 -1)
3430 LPRINT "   RESULTADO DE ANALIZAR Ho(F x C)"
3440 LPRINT "   AIC(0) = ";A0
3450 LPRINT "   AIC(1) = ";A1
3460 IF A0 > A1 THEN GOTO 3500
3470 LPRINT "   DECISION: ACEPTAR Ho(F x C)"
3480 LPRINT
3490 GOTO 3520
3500 LPRINT "   DECISION: RECHAZAR Ho(F x C)"
3510 LPRINT
3520 REM * ANALISIS Ho(D x FC) *
3530 A0= -2*(S3 +S6 - 2*S) + 2*(N1*N2 + N3 -2)
3540 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
3550 LPRINT "   RESULTADO DE ANALIZAR Ho(D x FC)"
3560 LPRINT "   AIC(0) = ";A0
3570 LPRINT "   AIC(1) = ";A1
3580 IF A0 > A1 THEN GOTO 3620
3590 LPRINT "   DECISION: ACEPTAR Ho(D x FC)"
3600 LPRINT
3610 GOTO 3640
3620 LPRINT "   DECISION: RECHAZAR Ho(D x FC)"
3630 LPRINT
3640 REM ************************************************************
3650 REM *      ANALISIS Ho(D x FC) <=> Ho(F x D / C) Y Ho(C x D)     *
3660 REM ************************************************************
3670 LPRINT "      Ho(D x FC) <=> Ho(F x D/C) Y Ho(C x D)"
3680 LPRINT
3690 REM * ANALISIS  Ho(F x D/C) *
3700 A0 = -2*(S3 + S7 - S8 - S) + 2*(N2*(N1+N3-1) -1)
3710 A1= -2*(S2 - S) + 2*(N1*N2*N3 -1)
3720 LPRINT "      RESULTADO DE ANALIZAR Ho(F x D/C)"
3730 LPRINT "      AIC(0) = ";A0
3740 LPRINT "      AIC(1) = ";A1
3750 IF A0 > A1 THEN GOTO 3790
3760 LPRINT "      DECISION: ACEPTAR Ho(F x D/C)"
3770 LPRINT
3780 GOTO 3810
3790 LPRINT "      DECISION: RECHAZAR Ho(F x D/C)"
```

```
3800 LPRINT
3810 REM * ANALISIS  Ho(C x D) *
3820 A0= -2*(S8 + S6 - 2*S) + 2*(N2 + N3 -2)
3830 A1= -2*(S7 - S) + 2*(N2*N3 -1)
3840 LPRINT "       RESULTADO DE ANALIZAR Ho(C x D)"
3850 LPRINT "       AIC(0) = ";A0
3860 LPRINT "       AIC(1) = ";A1
3870 IF A0 > A1 THEN GOTO 3910
3880 LPRINT "       DECISION: ACEPTAR Ho(C x D)"
3890 LPRINT
3900 GOTO 3930
3910 LPRINT "       DECISION: RECHAZAR Ho(C x D)"
3920 LPRINT
3930 REM ************************************************************
3940 REM *     ANALISIS Ho(D x FC) <=> Ho(C x D / f) Y Ho(F x D)      *
3950 REM ************************************************************
3960 LPRINT "       Ho(D x FC) <=> Ho(C x D/F) Y Ho(F x D)"
3970 LPRINT
3980 REM * ANALISIS  Ho(C x D/F) *
3990 A0= -2*(S3 + S5 - S4 - S) + 2*(N1*(N2 + N3 -1) -1)
4000 A1= -2*(S2 - S) + 2*(N1*N2*N3 -1)
4010 LPRINT "       RESULTADO DE ANALIZAR Ho(C x D/F)"
4020 LPRINT "       AIC(0) = ";A0
4030 LPRINT "       AIC(1) = ";A1
4040 IF A0 > A1 THEN GOTO 4080
4050 LPRINT "       DECISION: ACEPTAR Ho(C x D/F)"
4060 LPRINT
4070 GOTO 4100
4080 LPRINT "       DECISION: RECHAZAR Ho(C x D/F)"
4090 LPRINT
4100 REM * ANALISIS  Ho(F x D) *
4110 A0= -2*(S4 +S6 -2*S) + 2*(N1 +N3 -2)
4120 A1= -2*(S5 - S) + 2*(N1*N3 -1)
4130 LPRINT "       RESULTADO DE ANALIZAR Ho(F x D)"
4140 LPRINT "       AIC(0) = ";A0
4150 LPRINT "       AIC(1) = ";A1
4160 IF A0 > A1 THEN GOTO 4200
4170 LPRINT "       DECISION: ACEPTAR Ho(F x D)"
```

```
4180 LPRINT
4190 GOTO 4220
4200 LPRINT "      DECISION: RECHAZAR Ho(F x D)"
4210 LPRINT
4220 RETURN 1330
4230 REM ************************************************************
4240 REM * SUBRUTINA QUE ANALIZA LA HIPOTESIS CORRESPONDIENTE A L=2:  *
4250 REM * EXISTE HOMOGENEIDAD (C, D)                                 *
4260 REM ************************************************************
4270 A0= -2*(S7 - S) + 2*(N2*N3 -1)
4280 A1= -2*(S2 - S4) + 2*N1*(N2*N3 -1)
4290 LPRINT "RESULTADO DE ANALIZAR Ho(HOMOGENEIDAD (C, D))"
4300 LPRINT "AIC(0) = ";A0
4310 LPRINT "AIC(1) = ";A1
4320 IF A0 > A1 THEN GOTO 4360
4330 LPRINT "DECISION: ACEPTAR Ho(HOMOGENEIDAD (C,D))"
4340 LPRINT
4350 GOTO 4380
4360 LPRINT "DECISION: RECHAZAR Ho(HOMOGENEIDAD (C,D))"
4370 LPRINT
4380 REM ************************************************************
4390 REM *   ANALISIS Ho(HOMOG(C,D)) <=> Ho(HOMOG C) Y Ho(HOMOG (D/C))*
4400 REM ************************************************************
4410 LPRINT "   Ho(HOMOG(C,D)) <=> Ho(HOMOG C) Y Ho(HOMOG (D/C))"
4420 LPRINT
4430 REM *   ANALISIS Ho(HOMOGENEIDAD C) *
4440 A0= -2*(S8 -S) + 2*(N2 -1)
4450 A1= -2*(S3 - S4) + 2*N1*(N2 -1)
4460 LPRINT "   RESULTADO DE ANALIZAR Ho(HOMOGENEIDAD C)"
4470 LPRINT "   AIC(0) = ";A0
4480 LPRINT "   AIC(1) = ";A1
4490 IF A0 > A1 THEN GOTO 4530
4500 LPRINT "   DECISION: ACEPTAR Ho(HOMOGENEIDAD C)"
4510 LPRINT
4520 GOTO 4550
4530 LPRINT "   DECISION: RECHAZAR Ho(HOMOGENEIDAD C)"
4540 LPRINT
4550 REM *   ANALISIS  Ho(HOMOG (D/C)) *
```

```
4560 A0= -2*(S7 - S8) + 2*N2*(N3 -1)
4570 A1= -2*(S2 - S3) + 2*N1*N2*(N3 -1)
4580 LPRINT "   RESULTADO DE ANALIZA Ho(HOMOG COND (D/C))"
4590 LPRINT "   AIC(0) = ";A0
4600 LPRINT "   AIC(1) = ";A1
4610 IF A0 > A1 THEN GOTO 4650
4620 LPRINT "   DECISION: ACEPTAR Ho(HOMOG COND (D/C))"
4630 LPRINT
4640 GOTO 4670
4650 LPRINT "   DECISION: RECHAZAR Ho(HOMOG COND (D/C))"
4660 LPRINT
4670 REM ************************************************************
4680 REM * ANALISIS Ho(HOMOG(D/C)) <=> Ho(INTER FD) Y Ho(INTER (FD,C))*
4690 REM ************************************************************
4700 LPRINT "      Ho(HOMOG(D/C)) <=> Ho(INTER FD) Y Ho(INTER(FD,C))"
4710 LPRINT
4720 REM *      ANALISIS  Ho(INTERACCION FD) *
4730 A0= -2*(S9 -S) + 2*(N1 + N3 -2)
4740 A1= -2*(S5 - S) + 2*(N1*N3 - 1)
4750 LPRINT "      RESULTADO DE ANALIZAR Ho(INTERACCION FD)"
4760 LPRINT "      AIC(0) = ";A0
4770 LPRINT "      AIC(1) = ";A1
4780 IF A0 > A1 THEN GOTO 4820
4790 LPRINT "      DECISION: ACEPTAR Ho(INTERACCION FD)"
4800 LPRINT
4810 GOTO 4840
4820 LPRINT "      DECISION: RECHAZAR Ho(INTERACCION FD)"
4830 LPRINT
4840 REM *      ANALISIS Ho(INTERACCION (FD, C)) *
4850 A0= -2*(S5+S3+S7-S9-S8-S) + 2*(N1*N2 +N1*N3 +N2*N3 -N1 -N2 -N3)
4860 A1= -2*(S2 - S) + 2*(N1*N2*N3 -1)
4870 LPRINT "      RESULTADO DE ANALIZAR Ho(INTERACCION (FD,C))"
4880 LPRINT "      AIC(0) = ";A0
4890 LPRINT "      AIC(1) = ";A1
4900 IF A0 > A1 THEN GOTO 4940
4910 LPRINT "      DECISION: ACEPTAR Ho(INTERACCION (FD,C))"
4920 LPRINT
4930 GOTO 4970
```

```
4940 LPRINT "      DECISION: RECHAZAR Ho(INTERACCION (FD,C))"
4950 LPRINT
4960 REM ************************************************************
4970 REM *        ANALISIS DE Ho(HOMOGENEIDAD CONDICIONAL (D/C=j)     *
4980 REM ************************************************************
4990 PRINT "(DESEA REALIZAR EL ANALISIS Ho(HOMOG COND (D/C=j))?"
5000 PRINT " SI ( 1)           NO (2)"
5010 INPUT " OPCION ELEGIDA : ",L1
5020 IF L1 > 1 THEN GOTO 5370
5030 PRINT "CONSIDERAR  j <= ";N2
5040 INPUT "CATEGORIA A ANALIZAR  j = ";M
5050 IF M > N2 THEN GOTO 5030
5060 S11=0
5070 S12= 0
5080 S13=0
5090 FOR I=1 TO N1
5100 IF X3(I,M) = 0 THEN GOTO 5120
5110 S11= S11 + X3(I,M)*LOG(X3(I,M))
5120 FOR K=1 TO N3
5130 IF X(I,M,K) = 0 THEN GOTO 5150
5140 S12= S12 + X(I,M,K)*LOG(X(I,M,K))
5150 NEXT K
5160 NEXT I
5170 FOR K=1 TO N3
5180 IF X1(M,K) = 0 THEN GOTO 5200
5190 S13= S13 + X1(M,K)*LOG(X1(M,K))
5200 NEXT K
5210 S14= X13(M)*LOG(X13(M))
5220 A0= -2*(S13 - S14) + 2*(N3 -1)
5230 A1= -2*(S12 - S11) + 2*N1*(N3 -1)
5240 LPRINT "      RESULTADO DE ANALIZAR Ho(HOMOG COND (D/C=";M;"))"
5250 LPRINT "      AIC(0) = ";A0
5260 LPRINT "      AIC(1) = ";A1
5270 IF A0 > A1 THEN GOTO 5310
5280 LPRINT "      DECISION: ACEPTAR Ho(HOMOG COND (D/C=";M;"))"
5290 LPRINT
5300 GOTO 5330
5310 LPRINT "      DECISION: RECHAZAR Ho(HOMOG COND (D/C=";M;"))"
```

```
5320 LPRINT
5330 PRINT "(DESEA REALIZAR OTRO ANALISIS Ho(HOMOG COND(D/C=j))?"
5340 PRINT " SI ( 3 )            NO ( 4 )"
5350 INPUT " OPCION ELEGIDA : ";L1
5360 IF L1 < 4 THEN GOTO 5030
5370 REM ************************************************************
5380 REM *   ANALISIS Ho(HOMOG(C,D)) <=> Ho(HOMOG D) Y Ho(HOMOG (C/D))*
5390 REM ************************************************************
5400 LPRINT "   Ho(HOMOG(C,D)) <=> Ho(HOMOG D) Y Ho(HOMOG (C/D))"
5410 LPRINT                                          ,
5420 REM *   ANALISIS  Ho(HOMOGENEIDAD D) *
5430 A0= -2*(S6 -S) + 2*(N3 -1)
5440 A1= -2*(S5 - S4) + 2*N1*(N3 - 1)
5450 LPRINT "   RESULTADO DE ANALIZAR Ho(HOMOGENEIDAD D)"
5460 LPRINT "   AIC(0) = ";A0
5470 LPRINT "   AIC(1) = ";A1
5480 IF A0 > A1 THEN GOTO 5520
5490 LPRINT "   DECISION: ACEPTAR Ho(HOMOGENEIDAD D)"
5500 LPRINT
5510 GOTO 5540
5520 LPRINT "   DECISION: RECHAZAR Ho(HOMOGENEIDAD D)"
5530 PRINT
5540 REM *   ANALISIS  Ho(HOMOG COND (C/D)) *
5550 A0= -2*(S7 - S6) + 2*N3*(N2 -1)
5560 A1= -2*(S2 - S5) + 2*N1*N3*(N2 - 1)
5570 LPRINT "   RESULTADO DE ANALIZAR Ho(HOMOG COND (C/D))"
5580 LPRINT "   AIC(0) = ";A0
5590 LPRINT "   AIC(1) = ";A1
5600 IF A0 > A1 THEN GOTO 5640
5610 LPRINT "   DECISION: ACEPTAR Ho(HOMOG COND (C/D))"
5620 LPRINT
5630 GOTO 5660
5640 LPRINT "   DECISION: RECHAZAR Ho(HOMOG COND (C/D))"
5650 LPRINT
5660 REM ************************************************************
5670 REM * ANALISIS Ho(HOMOG(C/D)) <=> Ho(INTER FC) Y Ho(INTER(FC,D)) *
5680 REM ************************************************************
5690 LPRINT "      Ho(HOMOG(C/D) <=> Ho(INTER FC) Y Ho(INTER(FC,D))"
```

```
5700 LPRINT
5710 REM *      ANALISIS  Ho(INTERACCION FC) *
5720 A0= -2*(S10 -S) + 2*(N1 + N2 -2)
5730 A1= -2*(S3 -S) + 2*(N1*N2 - 1)
5740 LPRINT
5750 LPRINT "      RESULTADO DE ANALIZAR Ho(INTERACCION FC)"
5760 LPRINT "      AIC(0) = ";A0
5770 LPRINT "      AIC(1) = ";A1
5780 IF A0 > A1 THEN GOTO 5820
5790 LPRINT "      DECISION: ACEPTAR Ho(INTERACCION FC)"
5800 LPRINT
5810 GOTO 5840
5820 LPRINT "      DECISION: RECHAZAR Ho(INTERACCION FC)"
5830 LPRINT
5840 REM *      ANALISIS  Ho(INTERACCION (FC , D)) *
5850 A0= -2*(S5+S3+S7-S10-S6-S) +2*(N2*N3 +N1*N3 +N1*N2 -N1 -N2 -N3)
5860 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
5870 LPRINT "      RESULTADO DE ANALIZAR Ho(INTERACCION (FC, D))"
5880 LPRINT "      AIC(0) = ";A0
5890 LPRINT "      AIC(1) = ";A1
5900 IF A0 > A1 THEN GOTO 5940
5910 LPRINT "      DECISION: ACEPTAR Ho(INTERACCION (FC, D))"
5920 LPRINT
5930 GOTO 5970
5940 LPRINT "      DECISION: RECHAZAR Ho(INTERACCION (FC, D))"
5950 LPRINT
5960 REM ************************************************************
5970 REM *      ANALISIS Ho(HOMOGENEIDAD CONDICIONAL (C/D=k))         *
5980 REM ************************************************************
5990 PRINT "(DESEA REALIZAR EL ANALISIS Ho(HOMOG COND (C/D=k))?"
6000 PRINT " SI ( 1 )          NO (2)"
6010 INPUT "OPCION = ";L2
6020 IF L2 > 1 THEN GOTO 6370
6030 PRINT "CONSIDERAR  k<= ";N3
6040 INPUT "CATEGORIA A ANALIZAR k = ";L3
6050 IF L3 > N3 THEN GOTO 6030
6060 S15=0
6070 S16=0
```

```
6080 S17=0
6090 FOR I=1 TO N1
6100 IF X2(I,L3) = 0 THEN GOTO 6120
6110 S15= S15 + X2(I,L3)*LOG(X2(I,L3))
6120 FOR J=1 TO N2
6130 IF X(I,J,L3) = 0 THEN GOTO 6150
6140 S16 = S16 + X(I,J,L3)*LOG(X(I,J,L3))
6150 NEXT J
6160 NEXT I
6170 FOR J=1 TO N2
6180 IF X1(J,L3) = 0 THEN GOTO 6200
6190 S17= S17 + X1(J,L3)*LOG(X1(J,L3))
6200 NEXT J
6210 S18= X12(L3)*LOG(X12(L3))
6220 A0= -2*(S17 - S18) + 2*(N2 -1)
6230 A1= -2*(S16 - S15) + 2*N1*(N2 - 1)
6240 LPRINT "      RESULTADO DE ANALIZAR Ho(HOMOG COND(C/D=";L3;")}"
6250 LPRINT "      AIC(0) = ";A0
6260 LPRINT "      AIC(1) = ";A1
6270 IF A0 > A1 THEN GOTO 6310
6280 LPRINT "      DECISION: ACEPTAR Ho(HOMOG COND (C/D=";L3;"))"
6290 LPRINT
6300 GOTO 6330
6310 LPRINT "      DECISION: RECHAZAR Ho(HOMOG COND (C/D=";L3;"))"
6320 LPRINT
6330 PRINT "(DESEA REALIZAR OTRO ANALISIS Ho(HOMOG COND (C/D=k))?"
6340 PRINT " SI ( 3 )          NO ( 4 )"
6350 INPUT " OPCION ELEGIDA : ";L2
6360 IF L2 < 4 THEN GOTO 6030
6370 RETURN 1330
6380 REM ************************************************************
6390 REM * SUBRUTINA QUE ANALIZA LA HIPOTESIS CORRESPONDIENTE A L=3:  *
6400 REM * EXISTE INTERACCION ENTRE LOS TRES FACTORES DE CLASIFICACION*
6410 REM ************************************************************
6420 A0= -2*(S3+S7+S5-S4-S8-S6) +2*(N1*(N2-1) +N2*(N3-1) +N3*(N1-1))
6430 A1= -2*(S2 -S) + 2*(N1*N2*N3 -1)
6440 LPRINT "RESULTADO DE ANALIZAR Ho(INTERACCION FCD)"
6450 LPRINT "AIC(0) = ";A0
```

```
6460 LPRINT "AIC(1) = ";A1
6470 IF A0 > A1 THEN GOTO 6510
6480 LPRINT "DECISION: ACEPTAR Ho(INTERACCION FCD)"
6490 LPRINT
6500 GOTO 6530
6510 LPRINT "DECISION: RECHAZAR Ho(INTERACCION FCD)"
6520 LPRINT
6530 RETURN 1330
```

B. PROGRAMA QUE SELECCIONA EL CONJUNTO OPTIMO DE VARIABLES EXPLICATIVAS DE UNA VARIABLE RESPUESTA

El programa que se incluye en este apartado permite seleccionar el conjunto óptimo de variables explicativas de una variable respuesta en dos casos: cuando el número de variables explicativas es menor o igual a tres y cuando el número de variables explicativas es mayor que tres. En este último caso el programa realiza una preselección de variables explicativas y luego procede a obtener el conjunto óptimo de variables explicativas del conjunto de variables explicativas preseleccionadas.

Los datos de entrada que necesita este programa son:

M1: Número de variables explicativas. Si M1 ≥ 4 realiza la preselección antes de realizar la selección del conjunto óptimo.

Si M1 = 2, los datos de entrada que necesita son:

Z(I,J,K): Matriz de datos.
N1: Número de categorías de la variable respuesta.
N2: Número de categorías de la primera variable explicativa
N3: Número de categorías de la segunda variable explicativa

A partir de lo cual se calcula:

A1: Estadístico AIC del
MODELO(0,1): $p(i,j,k) = p(i,j,k)$

A2: Estadístico AIC del
MODELO(1,1): $p(i,j,k) = p(i,j)*p(j,k)/p(j)$

A3: Estadístico AIC del
MODELO(1,2): $p(i,j,k) = p(i,k)*p(j,k)/p(k)$

A4: Estadistico AIC del
MODELO(2,1): p(i,j,k) = p(i)•p(j,k)

Si M1 = 3, los datos de entrada que necesita son:

Y(I,J,K,L): Matriz de datos
N1: Número de categorias de la variable respuesta
N2: Número de categorias de la primera variable explicativa
N3: Número de categorias de la segunda variable explicativa
N4: Número de categorias de la tercera variable explicativa

A partir de lo cual se calcula

C1: Estadistico AIC del
MODELO(0,1): p(i,j,k,l) = p(i,j,k,l)
C2: Estadistico AIC del
MODELO(1,1): p(i,j,k,l) = p(i,j,k)•p(j,k,l)/p(j,k)
C3: Estadistico AIC del
MODELO(1,2): p(i,j,k,l) = p(i,j,l)•p(j,k,l)/p(j,l)
C4: Estadistico AIC del
MODELO(1,3): p(i,j,k,l) = p(i,k,l)•p(j,k,l)/p(k,l)
C5: Estadistico AIC del
MODELO(2,1): p(i,j,k,l) = p(i,j)•p(j,k,l)/p(j)
C6: Estadistico AIC del
MODELO(2,2): p(i,j,k,l) = p(i,k)•p(j,k,l)/p(k)
C7: Estadistico AIC del
MODELO(2,3): p(i,j,k,l) = p(i,l)•p(j,k,l)/p(l)
C8: Estadistico AIC del
MODELO(3,1): p(i,j,k,l) = p(i)•p(j,k,l)

Si M1 > 3, se realiza primero la preselección y para realizar ésto se introduce la variable respuesta con cada una de las variables explicativas, asi los datos que se introducen son:

X(I,J): Matriz de datos

N1: Número de categorias de la variables respuesta

N2: Número de categorias de la variable explicativa

A partir de lo cual se calcula

A: Estadistico AIC de cada modelo

Con estos datos se preselecciona las variables explicativas más significativas para luego proceder a la selección del conjunto óptimo de variables explicativas dentro del conjunto de variables explicativas preseleccionadas.

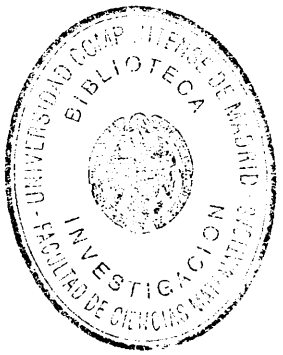
Su listado se presenta a continuacion.

```
10 REM ************************************************************
20 REM * PROGRAMA: OBTIENE EL CONJUNTO OPTIMO DE VARIABLES         *
30 REM *           EXPLICATIVAS DE UNA VARIABLE RESPUESTA.         *
40 REM ************************************************************
50 INPUT "NUMERO DE VARIABLES EXPLICATIVAS= ";L
60 IF L < 4 THEN GOTO 950
70 REM ************************************************************
80 REM  * PROGRAMA: CALCULA EL AIC DE LOS MODELOS                   *
90 REM  *           MODELO(I1 , Ij) , J=2,....L                     *
100 REM *           I1 : VARIABLE RESPUESTA                         *
110 REM *           Ij : VARIABLE EXPLICATIVA                       *
120 REM ************************************************************
130 REM *           LECTURA DE DATOS                                *
140 REM ************************************************************
150 DIM X(7,7), X1(7), X2(7)
160 DIM B(L,2), F(L)
170 PRINT "*************************************************************"
180 PRINT "* SE VA A EFECTUAR PRESELECCION DE VARIABLES EXPLICATIVAS.*"
190 PRINT "* CADA INGRESO DE DATO CONSTA DE LA VARIABLE RESPUESTA Y  *"
200 PRINT "* UNA VARIABLE EXPLICATIVA.                               *"
210 PRINT "*************************************************************"
220 PRINT "          INGRESO DE DATOS"
230 PRINT "          ******* ** *****"
240 PRINT
250 LPRINT "*****************************************************"
260 LPRINT "* # MODELO * MODELO(I1,Ij)              *    AIC     *"
270 LPRINT "*          * I1: VARIABLE RESPUESTA    *            *"
280 LPRINT "*          * Ij: VARIABLE EXPLICATIVA  *            *"
290 LPRINT "*****************************************************"
300 FOR N=1 TO L
310 B(N,1)= N
320 INPUT "NUMERO DE CATEGORIAS DE LA VARIABLE RESPUESTA = ";N1
330 INPUT "NUMERO DE CATEGORIAS DE LA VARIABLE EXPLICATIVA = ";N2
340 PRINT "INTRODUCCION DE LA MATRIZ DE DATOS"
350 PRINT "************ ** ** ****** ** *****"
360 S1= 0
370 S2= 0
380 S3= 0
```

```
390 S4= 0
400 FOR I=1 TO N1
410 X2(I)= 0
420 FOR J=1 TO N2
430 PRINT "X(";I;",";J;") = ";
440 INPUT X(I,J)
450 S1= S1 + X(I,J)
460 IF X(I,J)= 0 THEN GOTO 480
470 S2= S2 + X(I,J)*LOG(X(I,J))
480 X2(I)= X2(I) + X(I,J)
490 NEXT J
500 IF X2(I)= 0 THEN GOTO 520
510 S3= S3 + X2(I)*LOG(X2(I))
520 NEXT I
530 S5= S1*LOG(S1)
540 FOR J=1 TO N2
550 X1(J)= 0
560 FOR I=1 TO N1
570 X1(J)= X1(J) + X(I,J)
580 NEXT I
590 IF X1(J)= 0 THEN GOTO 610
600 S4= S4 + X1(J)*LOG(X1(J))
610 NEXT J
620 M= N + 1
630 A= -2*(S2 +S5 -S3 -S4) + 2*(N1 -1)*(N2 -1)
640 B(N,2)= A
650 LPRINT "*    ";N;"         MODELO(I1, I";M;")          ";A
660 NEXT N
670 LPRINT "*****************************************************"
680 REM ***************************************************
690 REM *   REORDENACION DE LOS AIC                       *
700 REM ***************************************************
710 C=0
720 FOR N=1 TO L-1
730 IF B(N,2) <= B(N+1,2) THEN GOTO 810
740 U= B(N,2)
750 D= B(N,1)
760 B(N,2)= B(N+1,2)
```

```
770 B(N,1)= B(N+1,1)
780 B(N+1,2)= U
790 B(N+1,1)= D
800 C=1
810 NEXT N
820 IF C <> 0 THEN GOTO 710
830 LPRINT "*********************************************"
840 LPRINT "* LOS AIC DE LOS MODELOS EN ORDEN CRECIENTE *"
850 LPRINT "*********************************************"
860 LPRINT "* # MODELO *             A I C              *"
870 LPRINT "*********************************************"
880 FOR N= 1 TO L
890 LPRINT "*   ";B(N,1);"               ";B(N,2)
900 NEXT N
910 LPRINT "*********************************************"
920 PRINT " NUMERO DE VARIABLES EXPLICATIVAS PRE SELECCIONADAS?"
930 INPUT " (# = 2 O 3)          # = ";M1
940 GOTO 960
950 M1 = L
960 IF M1 = 3 THEN GOTO 1940
970 REM  ************************************************************
980 REM  * PROGRAMA: SELECCIONA EL CONJUNTO OPTIMO DE VARIABLES      *
990 REM  * EXPLICATIVAS, CUANDO EL # DE VARIABLES EXPLICATIVAS ES 2  *
1000 REM ************************************************************
1010 S6= 0
1020 S7= 0
1030 S8= 0
1040 S9= 0
1050 S10= 0
1060 S11= 0
1070 S12= 0
1080 S13= 0
1090 REM ************************************************************
1100 REM *    LECTURA DE LA TABLA DE CONTINGENCIA DE TRES FACTORES   *
1110 REM ************************************************************
1120 PRINT "         INGRESO DE DATOS"
1130 PRINT "         ******* ** *****"
1140 PRINT
```

```
1150 INPUT "NUMERO DE CATEGORIAS DEL 1º FACTOR (V. RESPUESTA) = ";N1
1160 INPUT "NUMERO DE CATEGORIAS DEL 2º FACTOR (V. EXPLICATIVA)= ";N2
1170 INPUT "NUMERO DE CATEGORIAS DEL 3º FACTOR (V. EXPLICATIVA)= ";N3
1180 DIM Z(N1,N2,N3), Z1(N2,N3), Z2(N1,N3), Z3(N1,N2)
1190 DIM Z12(N3), Z13(N2), Z23(N1)
1200 PRINT
1210 PRINT "INTRODUCCION DE LA MATRIZ DE DATOS"
1220 PRINT "************ ** ** ****** ** *****"
1230 PRINT
1240 FOR I=1 TO N1
1250 Z23(I)= 0
1260 FOR J=1 TO N2
1270 Z3(I,J)= 0
1280 FOR K=1 TO N3
1290 PRINT "Z(";I;",";J;",";K;")= ";
1300 INPUT Z(I,J,K)
1310 S6= S6 + Z(I,J,K)
1320 IF Z(I,J,K)= 0 THEN GOTO 1340
1330 S7= S7 + Z(I,J,K)*LOG(Z(I,J,K))
1340 Z3(I,J)= Z3(I,J) + Z(I,J,K)
1350 NEXT K
1360 IF Z3(I,J)= 0 THEN GOTO 1380
1370 S8= S8 + Z3(I,J)*LOG(Z3(I,J))
1380 Z23(I)= Z23(I) + Z3(I,J)
1390 NEXT J
1400 IF Z23(I)= 0 THEN GOTO 1420
1410 S9= S9 + Z23(I)*LOG(Z23(I))
1420 NEXT I
1430 S14= S6*LOG(S6)
1440 FOR J=1 TO N2
1450 Z13(J)= 0
1460 FOR K=1 TO N3
1470 Z1(J,K)= 0
1480 FOR I=1 TO N1
1490 Z1(J,K)= Z1(J,K) + Z(I,J,K)
1500 NEXT I
1510 IF Z1(J,K)= 0 THEN GOTO 1530
1520 S12= S12 + Z1(J,K)*LOG(Z1(J,K))
```



```
1530 Z13(J)= Z13(J) + Z1(J, K)
1540 NEXT K
1550 IF Z13(J)= 0 THEN GOTO 1570
1560 S13= S13 + Z13(J)*LOG(Z13(J))
1570 NEXT J
1580 FOR K=1 TO N3
1590 Z12(K)= 0
1600 FOR I=1 TO N1
1610 Z2(I,K)= 0
1620 FOR J=1 TO N2
1630 Z2(I,K)= Z2(I,K) + Z(I,J,K)
1640 NEXT J
1650 IF Z2(I,K)= 0 THEN GOTO 1670
1660 S10= S10 +Z2(I,K)*LOG(Z2(I,K))
1670 Z12(K)= Z12(K) +Z2(I,K)
1680 NEXT I
1690 IF Z12(K)= 0 THEN GOTO 1710
1700 S11= S11 +Z12(K)*LOG(Z12(K))
1710 NEXT K
1720 A1= -2*(S7 - S12) + 2*N2*N3*(N1 - 1)
1730 A2= -2*(S8 - S13) + 2*N2*(N1 -1)
1740 A3= -2*(S10 - S11) + 2*N3*(N1 - 1)
1750 A4= -2*(S9 - S14) + 2*(N1 - 1)
1760 LPRINT "*****************************************************"
1770 LPRINT "*   MODELO    * HIPOTESIS ASOCIADA AL MODELO *   AIC    *"
1780 LPRINT "*             * E = {VARIABLES EXPLICATIVAS} *          *"
1790 LPRINT "*****************************************************"
1800 LPRINT "* MODELO(0,1) * p(i,j,k)= p(i,j,k)           * ";A1
1810 LPRINT "*             * E = { I2, I3 }               *          *"
1820 LPRINT "*             *                              *          *"
1830 LPRINT "* MODELO(1,1) * p(i,j,k)= p(i,j)*p(j,k)/p(j) * ";A2
1840 LPRINT "*             * E = { I2 }                   *          *"
1850 LPRINT "*             *                              *          *"
1860 LPRINT "* MODELO(1,2) * p(i,j,k)= p(i,k)*p(j,k)/p(k) * ";A3
1870 LPRINT "*             * E = { I3 }                   *          *"
1880 LPRINT "*             *                              *          *"
1890 LPRINT "* MODELO(2,1) * p(i,j,k)= p(i)*p(j,k)        * ";A4
1900 LPRINT "*             * E = { }                      *          *"
```

```
1910 LPRINT "*              *                              *          *"
1920 LPRINT "*************************************************************"
1930 END
1940 REM ***********************************************************
1950 REM * PROGRAMA: SELECCIONA EL CONJUNTO OPTIMO DE VARIABLES       *
1960 REM * EXPLICATIVAS, CUANDO EL # DE VARIABLES EXPLICATIVAS ES 3.  *
1970 REM ***********************************************************
1980 S15= 0
1990 S16= 0
2000 S17= 0
2010 S18= 0
2020 S19= 0
2030 S20= 0
2040 S21= 0
2050 S22= 0
2060 S23= 0
2070 S24= 0
2080 S25= 0
2090 S26= 0
2100 S27= 0
2110 S28= 0
2120 S29= 0
2130 S30= 0
2140 REM ***********************************************************
2150 REM *   LECTURA DE LA TABLA DE CONTINGENCIA DE CUATRO FACTORES   *
2160 REM ***********************************************************
2170 PRINT "          INGRESO DE DATOS"
2180 PRINT "          ******* ** *****"
2190 PRINT
2200 INPUT "NUMERO DE CATEGORIAS DEL 1º FACTOR (V. RESPUESTA) = ";N1
2210 INPUT "NUMERO DE CATEGORIAS DEL 2º FACTOR (V. EXPLICATIVA)= ";N2
2220 INPUT "NUMERO DE CATEGORIAS DEL 3º FACTOR (V. EXPLICATIVA)= ";N3
2230 INPUT "NUMERO DE CATEGORIAS DEL 4º FACTOR (V. EXPLICATIVA)= ";N4
2240 DIM Y(N1,N2,N3,N4), Y1(N2,N3,N4), Y2(N1,N3,N4), Y3(N1,N2,N4)
2250 DIM Y4(N1,N2,N3), Y12(N3,N4), Y13(N2,N4), Y14(N2,N3), Y23(N1,N4)
2260 DIM Y24(N1,N3), Y34(N1,N2), Y123(N4), Y124(N3), Y134(N2), Y234(N1)
2270 PRINT
2280 PRINT "INTRODUCCION DE LA MATRIZ DE DATOS"
```

```
2290 PRINT "************ ** ** ****** ** *****"
2300 PRINT
2310 FOR I=1 TO N1
2320 Y234(I)= 0
2330 FOR J=1 TO N2
2340 Y34(I,J)= 0
2350 FOR K=1 TO N3
2360 Y4(I,J,K)= 0
2370 FOR L=1 TO N4
2380 PRINT "Y(";I;",";J;",";K;",";L;")= ";
2390 INPUT Y(I,J,K,L)
2400 S15= S15 + Y(I,J,K,L)
2410 IF Y(I,J,K,L)= 0 THEN GOTO 2430
2420 S16= S16 + Y(I,J,K,L)*LOG(Y(I,J,K,L))
2430 Y4(I,J,K)= Y4(I,J,K) + Y(I,J,K,L)
2440 NEXT L
2450 IF Y4(I,J,K)= 0 THEN GOTO 2470
2460 S17= S17 + Y4(I,J,K)*LOG(Y4(I,J,K))
2470 Y34(I,J)= Y34(I,J) + Y4(I,J,K)
2480 NEXT K
2490 IF Y34(I,J)= 0 THEN GOTO 2510
2500 S18= S18 + Y34(I,J)*LOG(Y34(I,J))
2510 Y234(I)= Y234(I) + Y34(I,J)
2520 NEXT J
2530 IF Y234(I)= 0 THEN GOTO 2550
2540 S19= S19 + Y234(I)*LOG(Y234(I))
2550 NEXT I
2560 S31= S15*LOG(S15)
2570 FOR J=1 TO N2
2580 Y134(J)= 0
2590 FOR K=1 TO N3
2600 Y14(J,K)= 0
2610 FOR L=1 TO N4
2620 Y1(J,K,L)= 0
2630 FOR I=1 TO N1
2640 Y1(J,K,L)= Y1(J,K,L) + Y(I,J,K,L)
2650 NEXT I
2660 IF Y1(J,K,L)= 0 THEN GOTO 2680
```

```
2670 S20= S20 + Y1(J,K,L)*LOG(Y1(J,K,L))
2680 Y14(J,K)= Y14(J,K) + Y1(J,K,L)
2690 NEXT L
2700 IF Y14(J,K)= 0 THEN GOTO 2720
2710 S21= S21 + Y14(J,K)*LOG(Y14(J,K))
2720 Y134(J)= Y134(J) + Y14(J,K)
2730 NEXT K
2740 IF Y134(J)= 0 THEN GOTO 2760
2750 S22= S22 + Y134(J)*LOG(Y134(J))
2760 NEXT J
2770 FOR K=1 TO N3
2780 Y124(K)= 0
2790 FOR L=1 TO N4
2800 Y12(K,L)= 0
2810 FOR I=1 TO N1
2820 Y2(I,K,L)= 0
2830 FOR J=1 TO N2
2840 Y2(I,K,L)= Y2(I,K,L) + Y(I,J,K,L)
2850 NEXT J
2860 IF Y2(I,K,L)= 0 THEN GOTO 2880
2870 S23= S23 + Y2(I,K,L)*LOG(Y2(I,K,L))
2880 Y12(K,L)= Y12(K,L) + Y2(I,K,L)
2890 NEXT I
2900 IF Y12(K,L)= 0 THEN GOTO 2920
2910 S24= S24 + Y12(K,L)*LOG(Y12(K,L))
2920 Y124(K)= Y124(K) + Y12(K,L)
2930 NEXT L
2940 IF Y124(K)= 0 THEN GOTO 2960
2950 S25= S25 + Y124(K)*LOG(Y124(K))
2960 NEXT K
2970 FOR L=1 TO N4
2980 Y123(L)= 0
2990 FOR J=1 TO N2
3000 Y13(J,L)= 0
3010 FOR I=1 TO N1
3020 Y3(I,J,L)= 0
3030 FOR K=1 TO N3
3040 Y3(I,J,L)= Y3(I,J,L) + Y(I,J,K,L)
```

```
3050 NEXT K
3060 IF Y3(I,J,L)= 0 THEN GOTO 3080
3070 S26= S26 +Y3(I,J,L)*LOG(Y3(I,J,L))
3080 Y13(J,L)= Y13(J,L) + Y3(I,J,L)
3090 NEXT I
3100 IF Y13(J,L)= 0 THEN GOTO 3120
3110 S27= S27 + Y13(J,L)*LOG(Y13(J,L))
3120 Y123(L)= Y123(L) + Y13(J,L)
3130 NEXT J
3140 IF Y123(L)= 0 THEN GOTO 3160
3150 S28= S28 + Y123(L)*LOG(Y123(L))
3160 NEXT L
3170 FOR I=1 TO N1
3180 FOR K=1 TO N3
3190 Y24(I,K)= 0
3200 FOR J=1 TO N2
3210 Y24(I,K)= Y24(I,K) + Y4(I,J,k)
3220 NEXT J
3230 IF Y24(I,K)= 0 THEN GOTO 3250
3240 S29= S29 + Y24(I,K)*LOG(Y24(I,K))
3250 NEXT K
3260 NEXT I
3270 FOR I=1 TO N1
3280 FOR L=1 TO N4
3290 Y23(I,L)= 0
3300 FOR J=1 TO N2
3310 Y23(I,L)= Y23(I,L) + Y3(I,J,L)
3320 NEXT J
3330 IF Y23(I,L)= 0 THEN GOTO 3350
3340 S30= S30 + Y23(I,L)*LOG(Y23(I,L))
3350 NEXT L
3360 NEXT I
3370 C1= -2*(S16 - S20) + 2*N2*N3*N4*(N1 - 1)
3380 C2= -2*(S17 - S21) + 2*N2*N3*(N1 -1)
3390 C3= -2*(S26 - S27) + 2*N2*N4*(N1 -1)
3400 C4= -2*(S23 - S24) + 2*N3*N4*(N1 - 1)
3410 C5= -2*(S18 - S22) + 2*N2*(N1 - 1)
3420 C6= -2*(S29 - S25) + 2*N3*(N1 - 1)
```

```
3430 C7= -2*(S30 - S28) + 2*N4*(N1 - 1)
3440 C8= -2*(S19 - S31) + 2*(N1 - 1)
3450 LPRINT "*************************************************************"
3460 LPRINT "*            LOS  MODELOS  Y  SUS  AIC                      *"
3470 LPRINT "*            E = { VARIABLES EXPLICATIVAS }                 *"
3480 LPRINT "*************************************************************"
3490 LPRINT "* MODELO(0,1): p(i,j,k,l)=p(i,j,k,l)                        *"
3500 LPRINT "* E= {I2, I3, I4}        AIC = ";C1
3510 LPRINT "*                                                           *"
3520 LPRINT "* MODELO(1,1): p(i,j,k,l)= p(i,j,k)*p(j,k,l)/p(j,k)        *"
3530 LPRINT "* E= {I2, I3}            AIC = ";C2
3540 LPRINT "*                                                           *"
3550 LPRINT "* MODELO(1,2): p(i,j,k,l)= p(i,j,l)*p(j,k,l)/p(j,l)        *"
3560 LPRINT "* E= {I2,I4}             AIC = ";C3
3570 LPRINT "*                                                           *"
3580 LPRINT "* MODELO(1,3): p(i,j,k,l)= p(i,k,l)*p(j,k,l)/p(k,l)        *"
3590 LPRINT "* E= {I3,I4}             AIC = ";C4
3600 LPRINT "*                                                           *"
3610 LPRINT "* MODELO(2,1): p(i,j,k,l)= p(i,j)*p(j,k,l)/p(j)            *"
3620 LPRINT "* E= { I2 }              AIC = ";C5
3630 LPRINT "*                                                           *"
3640 LPRINT "* MODELO(2,2): p(i,j,k,l)= p(i,k)*p(j,k,l)/p(k)            *"
3650 LPRINT "* E= { I3 }              AIC = ";C6
3660 LPRINT "*                                                           *"
3670 LPRINT "* MODELO(2,3): p(i,j,k,l)= p(i,l)*p(j,k,l)/p(l)            *"
3680 LPRINT "* E= { I4 }              AIC = ";C7
3690 LPRINT "*                                                           *"
3700 LPRINT "* MODELO(3,1): p(i,j,k,l)= p(i)*p(j,k,l)                   *"
3710 LPRINT "* E= { }                 AIC = ";C8
3720 LPRINT "*                                                           *"
3730 LPRINT "*************************************************************"
3740 END
```

IV.3 PRESENTACION DE LOS PROBLEMAS Y APLICACION DE LOS METODOS

En esta sección se aplican los métodos expuestos en los capítulos II y III en dos situaciones prácticas. El criterio de información de Akaike para el análisis de tablas de contingencia se aplica para el análisis del paro en España en 1990 y el criterio de información de Akaike para la selección del conjunto óptimo de variables explicativas de una variable respuesta se aplica para el análisis de la fecundidad en España.

A. ESTUDIO DEL PARO EN ESPAÑA

El objetivo de nuestro estudio es analizar la influencia del sexo, la edad y el sector económico en el paro. A tal fin, se considera como variable indicadora "el número de parados".

Los datos de este estudio se han obtenido de la "Encuesta de Población Activa. Principales resultados. Octubre, Noviembre y Diciembre 1990", del Instituto Nacional de Estadística de España.

En este estudio se considera como parado a aquella persona de 16 o más años que busca su primer empleo o que ha trabajado antes y actualmente no trabaja.

La información del número de parados es trimestral, es por ello que se consideró el promedio de los cuatro trimestres como el número de parados del año. El número de parados está clasificado según el sector económico, sexo y edad.

SECTOR ECONOMICO, el sector económico comprende las siguientes sectores:

1) Agrario
2) Industria
3) Construcción
4) Servicios
5) No clasificable

Cada uno de los sectores anteriores comprende las siguientes ramas:

El sector Agrario tiene las ramas: agricultura, ganaderia y

servicios agrarios, silvicultura , caza y pesca.

El sector Industria comprende las ramas: minas de carbón; extracción de petroleo y refineria; electricidad, gas y agua; extracción de minerales; metálicas básicas; productos mineros no metálicos; industria quimica; transformados metálicos; maquinaria y equipos mecanicos, maquinaria y material eléctrico; material electronico y maquinas de oficina; vehiculos, automóviles y otro material de transporte; instrumentos de precisión y óptica; alimentos, bebidas y tabaco; industria textil, industria de cuero, calzado, vestido y otras confecciones, madera y corcho; papel y artes gráficas; transformación del caucho y plástico; otras industrias manufactureras.

El sector Construcción comprende sólo la rama construcción.

El sector Servicios comprende las ramas: comercio al por menor y resto de comercio; hosteleria; reparaciones; transporte por ferrocarril y otros transportes terrestres; transporte maritimo y aereos; actividades anexas y comunicación; banca, seguro e inmobiliaria; servicio a empresas y alquiler; administración pública y diplomática; saneamiento y similares; educación e investigación; sanidad y veterinaria; servicios sociales, recreativos y culturales; servicios personales; servicios doméstico.

SEXO: 1) Hombre
2) Mujer

EDAD: 1) 16 - 19 años
2) 20 - 24 años
3) 25 - 54 años
4) 55 y más años

Los datos se presentan en la siguiente tabla

TABLA 1.- NUMERO DE PARADOS CLASIFICADOS SEGUN SECTOR ECONOMICO, SEXO Y EDAD (1990)

SECTOR ECONOMICO	HOMBRE				MUJER			
	16-19	20-24	25-54	55 y+	16-19	20-24	25-54	55 y+
S1	12425	24525	71475	22125	8950	16075	40375	4525
S2	15975	41525	90400	19100	14575	39200	62050	4475
S3	10225	35975	134000	24825	400	2825	4450	375
S4	24700	70175	163550	18725	39275	107450	221900	11125
S5	79300	121900	152675	32575	113900	198900	369650	14575

donde:

S1: Agrario

S2: Industria

S3: Construccion

S4: Servicios

S5: No clasificables

En base a los datos de esta tabla se estudiará: la existencia de interacción, independencia total o parcial entre los factores Sector económico, Sexo y Edad; si el comportamiento del conjunto de los factores Sexo y Edad ha sido homogéneo a través de los diferentes Sectores económicos y la existencia de homogeneidad de la Edad condicionado al factor Sexo. Existencia de homogeneidad del Sexo dado el factor Edad.

A continuación se realizará el análisis de la influencia del sexo, la edad y el sector económico en el número de parados. Para ésto se aplica el criterio de información de Akaike para el análisis de tablas de contingencia para ver si los factores Sexo, Edad y Sector económico son independientes, o si existe algún tipo de homogeneidad e interaccion entre los factores.

Los resultados obtenidos aplicando este método son:

INDEPENDENCIA

Hipótesis nula: H_0(Sector económico x Sexo x Edad)
Estadísticos: AIC(0) = 15636808
AIC(1) = 15128198
Conclusión: Se rechaza H_0(Sector económico x Sexo x Edad), es decir, que los factores Sexo, Edad y Sector económico no son independientes.

A continuación se analiza que causa la cependencia entre los tres factores. Al ser,

H_0(Sector económico x Sexo x Edad)
$\Leftrightarrow H_0$(Sector económico x (Sexo, Edad)) $\cap$ H_0(Sexo x Edad)

Analizando esos dos componentes se obtuvo:

Hipótesis nula: H_0(Sexo x Edad)
Estadísticos: AIC(0) = 8880970
AIC(1) = 8822022
Coclusión: Se rechaza H_0(Sexo x Edad), es decir, los factores Sexo y Edad no son independientes.

Hipótesis nula: H_0(Sector económico x (Sexo, Edad))
Estadísticos: AIC(0) = 15577860
AIC(1) = 15128198
Conclusión: Se rechaza H_0(Sector económico x(Sexo, Edad)), es decir, no existe independencia parcial entre el factor Sector económico y el par (Sexo, Edad).

Además analizando el componente H_0(Sector económico x (Sexo,Edad)) ya que,

H_0(Sector económico x (Sexo, Edad))
$\Leftrightarrow H_0$(Sector económico x Sexo/ Edad) $\cap$ H_0(Sector económico x Edad)

se obtuvo:

Hipótesis nula: H_0(Sector económico x Sexo/ Edad)

Estadísticos: AIC(0) = 15478726

AIC(1) = 15128198

Conclusión: Se rehaza H_0(Sector económico x Sexo/ Edad), es decir, que no existe independencia condicional entre los factores Sector económico y Sexo dado el factor Edad.

Hipótesis nula: H_0(Sector económico x Edad)

Estadísticos: AIC(0) = 12257406

AIC(1) = 12158272

Conclusión: Se rechaza H_0(Sector económico x Edad), es decir, los factores Sector económico y Edad no son independientes.

Se realiza otros análisis teniendo en cuenta que

H_0(Sector económico x (Sexo, Edad))

$\Leftrightarrow H_0$(Sector económico x Edad/ Sexo) $\cap H_0$(Sector económico x Sexo)

analizando estos componentes se obtuvo

Hipótesis nula: H_0(Sector económico x Edad/ Sexo)

Estadísticos: AIC(0) = 15223250

AIC(1) = 15128198

Conclusión: Se rechaza H_0(Sector económico x Edad/ Sexo), es decir, que no existe independencia condicional entre los factores Sector económico y Edad dado el factor Sexo.

Hipótesis nula: H_0(Sector económico x Sexo)

Estadísticos: AIC(0) = 10135240

AIC(1) = 9780630

Conclusión: Se rechaza H_0(Sector económico x Sexo), es decir, los factores Sector económico y Sexo no son independientes.

De otro lado el componente H_0(Sector económico x Sexo x Edad) se

puede analizar de la siguiente manera

H_0(Sector económico x Sexo x Edad) $\Leftrightarrow$

H_0(Sexo x (Sector económico, Edad)) $\cap$ H_0(Sector económico x Edad)$^{\bullet}$

donde ($\bullet$) indica que esa hipótesis fue analizada anteriormente.

Los resultados obtenidos son:

Hipótesis nula:	H_0(Sexo x (Sector económico, Edad))
Estadísticos:	AIC(0) = 15537674
	AIC(1) = 15128198
Conclusión:	Se rechaza H_0(Sexo x(Sector económico, Edad)), es decir, no existe independencia parcial entre el factor Sexo y el par (Sector económico, Edad).

Y como,

H_0(Sexo x (Sector económico, Edad))

$\Leftrightarrow$ H_0(Sector económico x Sexo/ Edad)$^{\bullet}$ $\cap$ H_0(Sexo x Edad)$^{\bullet}$

De otro lado se tiene también que

H_0(Sexo x (Sector económico, Edad))

$\Leftrightarrow$ H_0(Sexo x Edad/ Sector económico) $\cap$ H_0(Sector económico x Sexo)$^{\bullet}$

para los cuales se obtuvo:

Hipótesis nula:	H_0(Sexo x Edad/ Sector económico)
Estadísticos:	AIC(0) = 15183064
	AIC(1) = 15128198
Conclusión:	Se rechaza H_0(Sexo x Edad/ Sector económico), es decir, no existe independencia condicional entre los factores Sexo y Edad dado el factor Sector económico.

El componente H_0(Sector económico x Sexo x Edad) también

puede ser analizado de la siguiente manera

H_0(Sector económico x Sexo x Edad) $\Leftrightarrow$

H_0(Edad x (Sector económico, Sexo)) $\cap$ H_0(Sector económico x Sexo)$^{\bullet}$

se efectuó este análisis y los resultados que se obtuvo fueron:

Hipótesis nula: H_0(Edad x (Sector económico, Sexo))
Estadísticos: AIC(0) = 15282198
AIC(1) = 15128198
Conclusión: Se rechaza H_0(Edad x (Sector económico, Sexo)), es decir, no existe independencia parcial entre el factor Edad y el par (Sector económico, Sexo).

Además se analizó

H_0(Edad x (Sector económico, Sexo))

$\Leftrightarrow$ H_0(Sector económico x Edad/ Sexo)$^{\bullet}$ $\cap$ H_0(Sexo x Edad)$^{\bullet}$

También se analizó

H_0(Edad x (Sector económico, Sexo))

$\Leftrightarrow$ H_0(Sexo x Edad/Sector económico)$^{\bullet}$ $\cap$ H_0(Sector económico x Edad)$^{\bullet}$

HOMOGENEIDAD

Hipótesis nula: H_0(Homogeneidad (Sexo, Edad))
Estadísticos: AIC(0) = 8822022
AIC(1) = 8372360
Conclusión: Se rechaza H_0(Homogeneidad (Sexo, Edad)), es decir, el comportamiento conjunto de los factores Sexo y Edad no es homogéneo en los diferentes Sectores económicos.

Además, se analizó cuales eran las causas de la no homogeneidad del (Sexo, Edad), para ello se consideró lo

siguiente:

Como,

H_0(Homogeneidad (Sexo, Edad)) ⇔

H_0(Homogeneidad (Sexo)) ∩ H_0(Homogeneidad Condicional (Edad/Sexo))

de donde se obtuvo

Hipótesis nula: H_0(Homogeneidad (Sexo))
Estadísticos: AIC(0) = 3379402
AIC(1) = 3024792
Conclusión: Se rechaza H_0(Homogeneidad (Sexo)), es decir, el factor Sexo no mantiene un comportamiento homogéneo en los diferentes Sectores económicos.

Hipotesis nula: H_0(Homogeneidad Condicional (Edad/ Sexo))
Estadisticos: AIC(0) = 5442620
AIC(1) = 5347568
Conclusion: Se rechaza H_0(Homog. Cond. (Edad/ Sexo)), es decir, el factor Edad no se mantiene homogéneo en los diferentes Sectores económicos dado el factor Sexo.

Además se axaminó este último componente para cada sexo y se hallo lo siguiente

Hipótesis nula: H_0(Homogeneidad Condicional (Edad/ Sexo= Hombre))
Estadísticos: AIC(0) = 2737728
AIC(1) = 2657710
Conclusión: Se rechaza H_0(Homog. Cond. (Edad/ Sexo= Hombre)), es decir, el factor Edad no tiene un comportamiento homogéneo en todos los Sectores económicos dado que el sexo es hombre. Se puede observar que en el grupo de los hombres, la distribución de la proporción de parados según la edad no sigue un comportamiento homogéneo en los diferentes sectores.

Hipótesis nula: H_0(Homogeneidad Condicional (Edad/ Sexo= Mujer))

Estadísticos: AIC(0) = 2704894

AIC(1) = 2689858

Conclusión: Se rechaza H_0(Homog. Cond. (Edad/ Sexo= Mujer)), es decir, el factor Edad no tiene un comportamiento homogéneo en todos los Sectores económicos dado que el sexo es mujer. Se puede observar que en el grupo de mujeres la distribución de la proporción de parados según la edad no tiene un comportamiento homogeneo en los diferentes sectores.

También se realizó el estudio de los siguiente:

H_0(Homogeneidad (Edad/ Sexo))

$\Leftrightarrow$ H_0(Interacción (Sector económico, Edad))

$\cap$ H_0(Interacción ((Sector económico, Edad), Sexo))

y se obtuvo lo siguiente

Hipótesis nula: H_0(Interacción (Sector económico, Edad))

Estadísticos: AIC(0) = 12232488

AIC(1) = 12158272

Conclusión: Se rechaza H_0(Inter. (Sector económico, Edad)), es decir, no existe interacción entre los factores Sector económico y Edad.

Hipótesis nula: H_0(Interacción ((Sector económico, Edad), Sexo))

Estadísticos: AIC(0) = 15149034

AIC(1) = 15128198

Conclusión: Se rechaza H_0(I.(Sector económico, Edad), Sexo)), es decir, no existe interacción entre el factor Sexo y el par de factores Sector económico y Edad.

El componente de homogeneidad H_0(Homogeneidad (Sexo, Edad)) también fue analizado de la siguiente manera

H_0(Homogeneidad (Sexo, Edad)) ⇔

H_0(Homogeneidad (Edad)) ∩ H_0(Homogeneidad Condicional (Sexo/Edad))

y se obtuvieron los siguientes resultados

Hipótesis nula: H_0(Homogeneidad (Edad))
Estadísticos: AIC(0) = 5501568
AIC(1) = 5402434
Conclusión: Se rechaza H_0(Homogeneidad (Edad)), es decir, el factor Edad no mantiene un comportamiento homogéneo en los diferentes Sectores económicos. Se puede observar que la proporción de parados según edad varia en los diferentes sectores económicos. Así por ejemplo, en el sector Construcción la proporción de parados que tienen entre 25 y 54 años es 65% mientras que en el sector Industrial es del 53%.

Hipótesis nula: H_0(Homogeneidad Condicional (Sexo/ Edad))
Estadísticos: AIC(0) = 3320454
AIC(1) = 2969926
Conclusión: Se rechaza H_0(Homog. Cond. (Sexo/ Edad)), es decir, el factor Sexo no tiene un comportamiento homogéneo en los diferentes Sectores económicos.

Además se examinó este último componente para cada grupo de edad y se hallo lo siguiente

Hipótesis nula: H_0(Homog. Cond. (Sexo/ Edad = (16 - 19)))
Estadísticos: AIC(0) = 439510.4
AIC(1) = 421710.4
Conclusión: Se rechaza H_0(Homog. Cond. (Sexo/ Edad=(16-19))), es decir, el factor Sexo no tiene un comportamiento homogéneo en los diferentes sectores económicos dado el grupo de edad de 16 a 19 años. Se puede observar que en el grupo de parados que tienen una edad entre 16 y 19 años en el sector Agrario el 58% de los parados son

hombres y el 42% son mujeres, en el sector Industrial el 52% de los parados son hombres y el 48% son mujeres, en el sector Servicios el 39% de los parados son hombres y el 61% son mujeres, sin embargo en el sector Construcción el 96% de los parados son hombres y el 4% son mujeres.

Hipótesis nula: H_0(Homog. Cond. (Sexo/ Edad = (20 - 24)))
Estadísticos: AIC(0) = 905416.6
AIC(1) = 851024.8
Conclusión: Se rechaza H_0(Homog. Cond. (Sexo/ Edad=(20-24))), es decir, el factor Sexo no sigue un comportamiento homogéneo en los diferentes sectores económicos dado el grupo de edad de 20 a 24 años.

Hipótesis nula: H_0(Homog. Cond. (Sexo/ Edad = (25 - 54)))
Estadísticos: AIC(0) = 1811084
AIC(1) = 1548348
Conclusión: Se rechaza H_0(Homog. Cond. (Sexo/ Edad=(25-54))), es decir, el factor Sexo no sigue un comportamiento homogéneo en los diferentes sectores económicos dado el grupo de edad de 25 a 54 años.

Hipótesis nula: H_0(Homog. Cond. (Sexo/ Edad = (55 y más años)))
Estadísticos: AIC(0) = 164442.2
AIC(1) = 148843.6
Conclusión: Se rechaza H_0(Homog. Cod. (Sexo/Edad=(55 y más)))
es decir, el factor Sexo no sigue un comportamiento homogéneo en los diferentes sectores económicos dado el grupo de edad de 55 y más años.

También se realizó el estudio del componente de Homogeneidad Condicional del factor Sexo dado el factor Edad de la siguiente manera

H_0(Homogeneidad Condicional (Sexo/ Edad))

$\Leftrightarrow$ H_0(Interacción (Sector económico, Sexo))

$\cap$ H_0(Interacción ((Sector económico, Sexo), Edad))

Hipótesis nula: H_0(Interacción (Sector económico, Sexo))
Estadísticos: AIC(0) = 10102588
AIC(1) = 9780630
Conclusión: Se rechaza H_0(Interac. (Sector económico, Sexo)), es decir, no existe interacción entre los factores Sector económico y Sexo.

Hipótesis nula: H_0(Interacción ((Sector económico, Sexo), Edad))
Estadísticos: AIC(0) = 15156768
AIC(1) = 15128198
Conclusión: Se rechaza H_0(I. ((Sector económico, Sexo), Edad)) es decir, no existe interacción entre el factor Edad y el par de factores (Sector económico, Sexo).

INTERACCION

Hipótesis nula: H_0(Interacción (Sector económico, Sexo, Edad))
Estadísticos: AIC(0) = 15124116
AIC(1) = 15128198
Conclusión: Se acepta H_0(Int. (Sector económico, Sexo, Edad)) es decir, existe interacción entre los factores Sector económico, Sexo y Edad.

Los datos de la tabla 1 también se analizaron mediante la medida de discriminación de Kullback-Leibler (G), y los modelos log lineal. Antes de presentar los resultados, previamente se describen estos métodos.

El estadístico G de Kullback, descrito en el capítulo I, es dos veces la información de Kullback-Leibler y sigue asintóticamente una distribución Ji- Cuadrado (ver Kullback 1959).

Los estadísticos G de Kullback correspondientes a las hipótesis antes analizadas con el estadístico AIC (Criterio de Información de Akaike) son presentados en la siguiente tabla.

TABLA 2.- ESTADISTICOS G DE KULLBACK

HIPOTESIS	ESTADISTICO G	GRADOS DE LIBERTAD
H_0(FxCxD)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{N^2 x_{ijk}}{x_{i..} x_{.j.} x_{..k}}$	rcd-r-c-d+2
H_0(CxD)	$2\sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \frac{N x_{.jk}}{x_{.j.} x_{..k}}$	(c-1)(d-1)
H_0(FxCD)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{i..} x_{.jk}}$	(r-1)(cd-1)
H_0(FxC/D)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk} x_{..k}}{x_{i.k} x_{.jk}}$	d(r-1)(c-1)
H_0(FxD)	$2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log \frac{N x_{i.k}}{x_{i..} x_{..k}}$	(r-1)(d-1)
H_0(FxD/C)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk} x_{.j.}}{x_{ij.} x_{.jk}}$	c(r-1)(d-1)
H_0(FxC)	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log \frac{N x_{ij.}}{x_{i..} x_{.j.}}$	(r-1)(c-1)
H_0(CxFD)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{.j.} x_{i.k}}$	(c-1)(rd-1)

HIPOTESIS	ESTADISTICO G	GRADOS DE LIBERTAD
H_0(CxD/F)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{i..}}{x_{ij.}\,x_{i.k}}$	r(c-1)(d-1)
H_0(DxFC)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{N\,x_{ijk}}{x_{..k}\,x_{ij.}}$	(d-1)(rc-1)
Homg. (C,D)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{N\,x_{ijk}}{x_{i..}\,x_{.jk}}$	(r-1)(cd-1)
Homog. C	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.}\log\frac{N\,x_{ij.}}{x_{i..}\,x_{.j.}}$	(r-1)(c-1)
Homog. D	$2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k}\log\frac{N\,x_{i.k}}{x_{i..}\,x_{..k}}$	(r-1)(d-1)
H.C. (D/C)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{.j.}}{x_{ij.}\,x_{.jk}}$	c(r-1)(d-1)
HC. (D/C=j)	$2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{.j.}}{x_{ij.}\,x_{.jk}}$	(r-1)(d-1)
H.C. (C/D)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk}\log\frac{x_{ijk}\,x_{..k}}{x_{i.k}\,x_{.jk}}$	d(r-1)(c-1)

HIPOTESIS	ESTADISTICO G	GRADOS DE LIBERTAD
HC. (C/D=k)	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ijk} \log \frac{x_{ijk}\, x_{..k}}{x_{i.k}\, x_{.jk}}$	(r-1)(c-1)
Inter. FD	$2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{i.k} \log \frac{x_{i.k}}{y_{i.k}}$	(r-1)(d-1)
	donde: $y_{i.k} = \sum_{j=1}^{c} \frac{x_{ij.}\, x_{.jk}}{x_{.j.}}$	
Int. (FD,C)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}\, y_{i.k}\, x_{.j.}}{x_{i.k}\, x_{ij.}\, x_{.jk}}$	(r-1)x(c-1)x (d-1)
Inter. FC	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log \frac{x_{ij.}}{y_{ij.}}$	(r-1)(c-1)
	donde: $y_{ij.} = \sum_{k=1}^{d} \frac{x_{i.k}\, x_{.jk}}{x_{..k}}$	
Int. (FC,D)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}\, y_{ij.}\, x_{..k}}{x_{i.k}\, x_{ij.}\, x_{.jk}}$	(r-1)x(c-1)x (d-1)

donde:

Homog.: Homogeneidad

H. C. : Homogeneidad Condicional

Inter.: Interaccion

El procedimiento basado en los modelos log lineal para el análisis de interacción entre los tres factores de clasificación (ver Bishop, Fienberg, Holland (1975)) es el siguiente:

La hipótesis nula en este caso sería

H_0: No existe interacción entre los tres factores de clasificación

lo que equivale en los modelos log lineal a la hipótesis nula

$$H_0: \quad \mu_{123(ijk)} = 0 , \quad i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots d$$

donde: $\mu_{123(ijk)}$ es el efecto de interacción de la categoría i del factor Fila, la categoría j del factor Columna y la categoría k del factor Profundidad.

Luego, el modelo log lineal asociado a esa hipótesis es,

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}$$

donde:

m_{ijk} : Frecuencia esperada en la celda (i,j,k), $m_{ijk} = N\, p_{ijk}$.

μ : La media de los logaritmos de las frecuencias esperadas.

$\mu_{1(i)}$: Efecto de la categoría i del factor Fila.

$\mu_{2(j)}$: Efecto de la categoría j del factor Columna.

$\mu_{3(k)}$: Efecto de la categoría k del factor Profundidad.

$\mu_{12(ij)}$: Efecto de interacción de la categoría i del factor Fila y la categoría j del factor Columna.

$\mu_{13(ik)}$: Efecto de interacción de la categoría i del factor Fila y la categoría k del factor Profundidad.

$\mu_{23(jk)}$: Efecto de interacción de la categoría j del factor Columna y la categoría k del factor Profundidad.

Y se efectúa el procedimiento de ajuste iterativo para el modelo $\mu_{123(ijk)} = 0$, $i=1,\ldots,r \quad j=1,\ldots,c \quad k=1,\ldots,d$.

Este procedimiento iterativo se inicia con un conjunto de valores preliminares $\hat{m}^{(0)}_{ijk} = 1$ para cada celda. Luego se procede con un ciclo de tres pasos:

Primer paso:

$$\hat{m}^{(1)}_{ijk} = \hat{m}^{(0)}_{ijk} \frac{x_{ij.}}{\hat{m}^{(0)}_{ij.}}$$

Segundo paso:

$$\hat{m}^{(2)}_{ijk} = \hat{m}^{(1)}_{ijk} \frac{x_{i.k}}{\hat{m}^{(1)}_{i.k}}$$

Tercer paso:

$$\hat{m}^{(3)}_{ijk} = \hat{m}^{(2)}_{ijk} \frac{x_{.jk}}{\hat{m}^{(2)}_{.jk}}$$

Repetimos el ciclo de tres pasos hasta que la convergencia con la aproximación deseada se alcance. La regla de parada del proceso considerada consiste en parar cuando después de un ciclo completo observamos que

$$| \hat{m}^{(3r)}_{ijk} - \hat{m}^{(3r-3)}_{ijk} | < 0.001 \quad \text{para todo i, j, k.}$$

la matriz resultante contiene los valores estimados, $\hat{m}_{ijk}$, obtenidos bajo la hipótesis nula.

Y se contrasta la bondad de ajuste del modelo asociado a la hipótesis nula, que no existe interacción entre los tres factores de clasificación, mediante el estadistico G del Kullback

$$G = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{\hat{m}_{ijk}}$$

el cual sigue una distribución Ji-Cuadrado con (r-1)(c-1)(d-1) grados de libertad.

Los resultados obtenidos del análisis de la Tabla 1 a través del estadistico G de Kullback y el modelo log lineal son los siguientes

TABLA 3.- RESULTADOS OBTENIDOS CON EL ESTADISTICO G DE KULLBACK Y EL MODELO LOG LINEAL

HIPOTESIS	G	GRADOS DE LIBERTAD (υ)	$\chi^2_{\upsilon,\ 0.05}$	DECISION
H_0(F x C x D)	508672.	31	43.773	Rechazar
H_0(C x D)	58954.	3	7.815	Rechazar
H_0(F x CD)	449718.	28	41.337	Rechazar
H_0(F x C/D)	350560.	16	26.296	Rechazar
H_0(F x D)	99158.	12	21.026	Rechazar
H_0(F x D/C)	95100.	24	36.415	Rechazar
H_0(F x C)	354618.	4	9.488	Rechazar
H_0(C x FD)	409514.	19	30.144	Rechazar
H_0(C x D/F)	54896.	15	24.996	Rechazar
H_0(D x FC)	154054.	27	40.113	Rechazar
Homogeneidad (C,D)	449718.	28	41.337	Rechazar
Homogeneidad C	354618.	4	9.488	Rechazar
Homogeneidad D	99158.	12	21.026	Rechazar
Homog. Cond. (D/C)	95100.	24	36.415	Rechazar
Homog. Cond. (D/C= Hombre)	80042.	12	21.026	Rechazar
Homog. Cond. (D/C= Mujer)	15060.	12	21.026	Rechazar

HIPOTESIS	G	GRADOS DE LIBERTAD (υ)	$\chi^2_{\upsilon,\ 0.05}$	DECISION
Homog. Cond. (C/D)	350560.	16	26.296	Rechazar
Homog. Cond. (C/D=(16-19))	17808.	4	9.488	Rechazar
Homog. Cond. (C/D=(20-24))	54399.8	4	9.488	Rechazar
Homog. Cond. (C/D=(25-54))	262744.	4	9.488	Rechazar
Homog. Cond. (C/D=(55 y +))	15606.6	4	9.488	Rechazar
Interacción FD	74240.	12	21.026	Rechazar
Interacción (FD, C)	20860.	12	21.026	Rechazar
Interacción FC	321966.	4	9.488	Rechazar
Interacción (FC, D)	28594.	12	21.026	Rechazar
H_0: No existe interacción entre los 3 factores	15093.36*	12	21.026	Rechazar

* Modelos Log Lineal

donde:

F: Sector Economico

C: Sexo

D: Edad

Podemos observar que las conclusiones que se llega mediante el AIC de Akaike coinciden con las obtenidas mediante el estadístico G de Kullback y los modelos log lineal.

B. ESTUDIO DE LA FECUNDIDAD

El objetivo de este estudio es analizar la fecundidad y para ello se ha tomado en cuenta la encuesta de fecundidad realizada en los meses de Mayo y Junio de 1985 por el Instituto Nacional de Estadística.

La unidad de análisis en esta investigación son todas las mujeres de 18 a 49 años independiente de su estado civil. La encuesta cubrió todo el territorio nacional incluidos Ceuta y Melilla y se ha realizado en el periodo comprendido entre el 15 de Mayo y el 30 de Junio de 1985.

Como los objetivos de la encuesta eran varios, siendo el principal el estudio de la fecundidad es por ello que el cuestionario era extenso, constando de 136 preguntas.

Como se observa el estudio del análisis de la fecundidad puede ser tratado siguiendo el Criterio de Información de Akaike para la selección del conjunto óptimo de variables explicativas.

Como el objetivo es analizar la fecundidad, se considero como variable indicadora de la fecundidad "el número de hijos nacidos vivos". Es por ello que se tomó como variable respuesta "el número de hijos nacidos vivos".

Como el número de variables explicativas era grande se procedió primeramente a efectuar la preselección de variables explicativas. Cabe decir que el número de variables explicativas consideradas está limitado por la información que presenta el informe "Encuesta de fecundidad 1985" del Instituto Nacional de Estadística. La información restringida que presenta este informe limitó nuestro análisis, ya que sólo contamos con tablas de doble entrada.

Las variables que consideramos en el estudio de la fecundidad son:

VARIABLE RESPUESTA

I1: Número de hijos nacidos vivos. Con categorías:

1.- Ninguno

2.- Uno

3.- Dos
4.- Tres o más

VARIABLES EXPLICATIVAS

I2: Estado civil. Con categorías:
1.- Mujeres alguna vez casadas
2.- Solteras

I3: Historia de su actividad laboral. Con categorías:
1.- Nunca ha trabajado
2.- Ha trabajado alguna vez

I4: Tamaño de municipio de residencia. Con categorías:
1.- Hasta 10,000 habitantes
2.- De 10,001 a 50,000 habitantes
3.- De 50,001 a 500,000 habitantes
4.- De más de 500,000 habitantes

I5: El número de hermanos nacidos vivos. Con categorías:
1.- Ninguno
2.- Uno
3.- Dos
4.- Tres
5.- Cuatro
6.- Cinco o más

I6: Creencia y práctica religiosa. Con categorías:
1.- No creyente
2.- Católica no practicante
3.- Católica practicante
4.- De otra religión
5.- No sabe, no contesta

I7: Edad actual. Con categorías:
1.- De 18 a 19 años
2.- De 20 a 24 años

3.- De 25 a 29 años
4.- De 30 a 34 años
5.- De 35 a 39 años
6.- De 40 a 44 años
7.- De 45 a 49 años

I8: Nivel de instrucción. Con categorías:
1.- Analfabeta
2.- Sin estudios
3.- Primarios
4.- Bachiller elemental o equivalente
5.- Bachiller superior o equivalente
6.- Nivel anterior al superior
7.- Estudios superiores

I9: Tipo de municipio en que paso la primera infancia. Con categorías:
1.- Medio rural
2.- Medio urbano; municipio menor de 100,000 habitantes
3.- Medio urbano; municipio mayor de 100,000 habitantes

I10: Relación con la actividad económica. Con categorías:
1.- Trabaja al menos un tercio de la jornada normal
2.- Trabaja menos de un tercio de la jornada normal
3.- Parada que busca empleo habiendo trabajado antes
4.- Busca su primer empleo
5.- Estudiante o escolar
6.- Labores del hogar
7.- Otras

Y los datos son los siguientes

TABLA 4.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y SU ESTADO CIVIL

NUMERO DE HIJOS NACIDOS VIVOS	ESTADO CIVIL	
	MUJERES ALGUNA VEZ CASADAS	SOLTERAS
0	633431	2251501
1	1262022	49165
2	2084674	4307
3 o más	1973792	3444

TABLA 5.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y SU HISTORIA DE LA ACTIVIDAD LABORAL.

NUMERO DE HIJOS NACIDOS VIVOS	HISTORIA DE SU ACTIVIDAD LABORAL	
	NUNCA HA TRABAJADO	HA TRABAJADO ALGUNA VEZ
0	911413	1973519
1	269818	1041369
2	448389	1640592
3 o más	568566	1408670

TABLA 6.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y EL TAMAÑO DEL MUNICIPIO DE RESIDENCIA

NUMERO DE HIJOS NACIDOS VIVOS	NUMERO DE HERMANOS NACIDOS VIVOS			
	H1	H2	H3	H4
0	723859	616519	927332	617222
1	337745	301937	420749	250756
2	478667	455189	752811	402314
3 o mas	499398	466714	671587	339537

donde:

H1: Hasta 10,000 habitantes

H2: De 10,001 a 50,000 habitantes

H3: De 50,001 a 500,000 habitantes

H4: De mas de 500,000 habitantes

TABLA 7.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y EL NUMERO DE HERMANOS NACIDOS VIVOS

# DE HIJOS NACIDOS VIVOS	NUMERO DE HERMANOS NACIDOS VIVOS					
	0	1	2	3	4	5 o más
0	148652	679841	701291	560565	309699	484884
1	98901	278585	258082	235625	163804	276190
2	121146	358564	437475	358274	274546	538976
3 o más	84732	252026	290526	288251	290763	770938

TABLA 8.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, SU CREENCIA Y PRACTICA RELIGIOSA

# HIJOS NACIDOS VIVOS	CREENCIA Y PRACTICA RELIGIOSA				
	R1	R2	R3	R4	R5
0	136967	1368236	1317957	25383	36389
1	45963	629517	599086	13472	23149
2	33236	800127	1218705	17640	19273
3 o más	20960	590142	1332536	14519	19079

donde:

R1: No creyente

R2: Catolica no practicante

R3: Catolica practicante

R4: De otra religion

R5: No sabe, no contesta

TABLA 9.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y SU EDAD ACTUAL

# HIJOS NACIDOS VIVOS	EDAD ACTUAL						
	18-19	20-24	25-29	30-34	35-39	40-44	45-49
0	580371	1171266	518198	194862	159997	131362	128876
1	39740	308875	438844	229659	127022	76969	90079
2	4508	92737	348668	528989	463713	353085	297281
3 o más	1089	8079	135120	297050	476348	532758	526732

TABLA 10.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y SU NIVEL DE INSTRUCCION

# HIJOS NACIDOS VIVOS	NIVEL DE INSTRUCCION						
	N1	N2	N3	N4	N5	N6	N7
0	19880	126702	531965	913438	922495	248132	122320
1	13344	106480	476324	389858	212864	80148	32169
2	34884	312212	997591	415575	172170	107373	49176
3 o más	137323	428076	962241	254966	112122	60331	22177

donde:

N1: Analfabeta

N2: Sin estudios

N3: Primarios

N4: Bachiller elemental o equivalente

N5: Bachiller superior o equivalente

N6: Nivel anterior al superior

N7: Estudios superiores

TABLA 11.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y EL TIPO DE MUNICIPIO EN QUE PASO LA PRIMERA INFANCIA

NUMERO DE HIJOS NACIDOS VIVOS	TIPO DE MUNICIPIO EN QUE PASO LA 1º INFANCIA		
	MEDIO RURAL	MEDIO URBANO; MUNICIPIO MENOR DE 100000 HAB.	MEDIO URBANO; MUNICIPIO MAYOR DE 100000 HAB.
0	1211918	630141	1042873
1	649162	302302	359723
2	1181725	422437	484819
3 o más	1198637	407405	371194

TABLA 12.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS Y LA RELACION CON LA ACTIVIDAD ECONOMICA

# HIJOS NACIDOS VIVOS	RELACION CON LA ACTIVIDAD ECONOMICA						
	E1	E2	E3	E4	E5	E6	E7
0	1302937	103303	247717	189154	563267	457711	20843
1	491813	46248	63766	10567	1671	690757	6365
2	568351	73295	37218	7741	5390	1388933	8053
3 o más	409105	85933	39116	1720	0	1435933	5429

donde:

E1: Trabaja al menos un tercio de la jornada normal

E2: Trabaja menos de un tercio de la jornada normal

E3: Parada que busca empleo habiendo trabajado antes

E4: Busca su primer empleo

E5: Estudiante o escolar

E6: Labores del hogar

E7: Otras

Aplicando el método de preselección de variables explicativas basado en el Criterio de información de Akaike se obtuvo los siguientes resultados.

TABLA 13.- LOS AIC DE LOS MODELOS EN ORDEN CRECIENTE

# MODELO	MODELO(I1,IJ) I1: VARIABLE RESPUESTA IJ: VARIABLE EXPLICATIVA	AIC
1	MODELO(I1,I2)	-6219994
6	MODELO(I1,I7)	-4601772
9	MODELO(I1,I10)	-3033884
7	MODELO(I1,I8)	-1924524
4	MODELO(I1,I5)	-414946
5	MODELO(I1,I6)	-318936
8	MODELO(I1,I9)	-246772
2	MODELO(I1,I3)	-95338
3	MODELO(I1,I4)	-24382

Observamos que las variables en orden de significación son I2, I7, I10, I8, I5, I6, I9, I3, I4. Sobresaliendo I2, I7, I10, I8 luego estas variables constituyen el conjunto de variables explicativas preseleccionadas.

El siguiente paso para seleccionar el conjunto óptimo de variables explicativas sería calcular los AIC de todos los modelos que se pueden formular con las variables explicativas preseleccionadas, pero debido a la limitada información con que se cuenta se ha optado por analizar solo los modelos que se pueden formular con la información que proporciona el informe "Encuesta de fecundidad 1985".

La información es la siguiente

TABLA 14.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, HISTORIA DE LA ACTIVIDAD LABORAL

NUMERO DE HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADA		MUJERES SOLTERAS	
	NUNCA HA TRABAJADO	HA TRABAJADO ALGUNA VEZ	NUNCA HA TRABAJADO	HA TRABAJADO ALGUNA VEZ
0	108002	525429	803411	1448090
1	259004	1003018	10814	38351
2	447342	1637332	1047	3260
3 o más	568093	1405699	473	2971

TABLA 15.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, TAMAÑO DE MUNICIPIO

# HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS				MUJERES SOLTERAS			
	H1	H2	H3	H4	H1	H2	H3	H4
0	190579	148838	205718	88296	533280	467681	721614	528926
1	325469	294917	403883	237753	12276	7020	16866	13003
2	477240	455189	750994	401251	1427	0	1817	1063
3 o más	499398	465590	670467	338337	0	1124	1120	1200

donde:

H1: Hasta 10000 habitantes

H2: De 10001 a 50000 habitantes

H3: De 50001 a 500000 habitantes

H4: Mas de 500000 habitantes

TABLA 16.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, EL NUMERO DE HERMANOS NACIDOS VIVOS

# DE HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS					
	0	1	2	3	4	5 o más
0	41775	162201	146093	105645	64128	113589
1	90706	268507	253231	225069	161044	263465
2	121146	358564	436706	357212	274546	536500
3 o más	84732	252026	289327	288251	290763	768693

# DE HIJOS NACIDOS VIVOS	MUJERES SOLTERAS					
	0	1	2	3	4	5 o más
0	106877	517640	555198	454920	245571	371295
1	8195	10078	4851	10556	2760	12725
2	0	0	769	1062	0	2476
3 o más	0	0	1199	0	0	2245

TABLA 17.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, CREENCIA Y PRACTICA RELIGIOSA

# HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS				
	R1	R2	R3	R4	R5
0	38674	303379	279311	5756	6311
1	42365	607451	577235	13472	21499
2	33236	797248	1217277	17640	19273
3 o más	19957	588826	1331412	14519	19078

# HIJOS NACIDOS VIVOS	MUJERES SOLTERAS				
	R1	R2	R3	R4	R5
0	98293	1064857	1038646	19627	30078
1	3598	22066	21851	0	1650
2	0	2879	1428	0	0
3 o más	1003	1316	1124	0	1

donde:

R1: No creyente

R2: Catolica no practicante

R3: Catolica practicante

R4: De otra religion

R5: No sabe, no contesta

TABLA 18.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, EDAD

# HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS						
	18-19	20-24	25-29	30-34	35-39	40-44	45-49
0	20911	237778	204446	49896	46390	37784	36226
1	35671	298604	429160	218435	119663	73348	87141
2	3461	90905	348668	528989	463713	353085	295853
3 o más	616	8079	133921	296398	476348	532758	525672

# HIJOS NACIDOS VIVOS	MUJERES SOLTERAS						
	18-19	20-24	25-29	30-34	35-39	40-44	45-49
0	559460	933488	313752	144966	113607	93578	92650
1	4069	10271	9684	11224	7359	3620	2938
2	1047	1832	0	0	0	0	1428
3 o más	473	0	1199	652	0	0	1120

TABLA 19.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, NIVEL DE INSTRUCCION

# HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS						
	N1	N2	N3	N4	N5	N6	N7
0	1514	37625	188851	198645	124431	49371	32994
1	13344	98862	461689	376488	202774	76696	32169
2	34884	310784	996544	413743	172170	107373	49176
3 o más	137206	428076	960040	253841	112121	60331	22177

# HIJOS NACIDOS VIVOS	MUJERES SOLTERAS						
	N1	N2	N3	N4	N5	N6	N7
0	18366	89077	343114	714793	798064	198761	89326
1	0	7618	14635	13370	10090	3452	0
2	0	1428	1047	1832	0	0	0
3 o más	117	0	2201	1125	1	0	0

donde:

N1: Analfabeta

N2: Sin estudios

N3: Primario

N4: Bachiller elemental

N5: Bachiller superior

N6: Nivel anterior al superior

N7: Estudios superiores

TABLA 20.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, TIPO DE MUNICIPIO EN QUE PASO LA PRIMERA INFANCIA

# DE HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS			MUJERES SOLTERAS		
	T1	T2	T3	T1	T2	T3
0	301896	153841	177694	910022	476300	865179
1	622577	295565	343880	26585	6737	15843
2	1180297	421390	482987	1428	1047	1832
3 o más	1198164	407287	368341	473	118	2853

donde:

T1: Medio rural

T2: Medio urbano; municipio menor de 100000 habitantes

T3: Medio urbano; municipio mayor de 100000 habitantes

TABLA 21.- DISTRIBUCION DE MUJERES SEGUN EL NUMERO DE HIJOS NACIDOS VIVOS, ESTADO CIVIL, RELACION CON LA ACTIVIDAD ECONOMICA

# HIJOS NACIDOS VIVOS	MUJERES ALGUNA VEZ CASADAS						
	E1	E2	E3	E4	E5	E6	E7
0	300991	26419	44750	6079	15721	238461	1010
1	465726	43009	56642	8677	1671	679932	6365
2	566153	73295	37218	7741	5390	1386824	8053
3 o más	408454	85816	39116	1720	0	1433257	5429

# HIJOS NACIDOS VIVOS	MUJERES SOLTERAS						
	E1	E2	E3	E4	E5	E6	E7
0	1001946	76884	202967	183075	547546	219250	19833
1	26087	3239	7124	1890	0	10825	0
2	2198	0	0	0	0	2109	0
3 o más	651	117	0	0	0	2676	0

donde:

E1: Trabaja al menos un tercio de la jornada normal

E2: Trabaja menos de un tercio de la jornada normal

E3: Parada que busca empleo habiendo trabajado antes

E4: Busca su primer empleo

E5: Estudiante o escolar

E6: Labores del hogar

E7: Otras

Aplicamos el Criterio de Información de Akaike a todos los modelos que podemos plantear con esta información. Así se obtuvo

TABLA 22.- LOS AIC DE LOS MODELOS

# MODELO	MODELO(V. RESPUESTA;{V. EXPLICATIVAS})	AIC
1	MODELO(I1;I2,I3)	1.601671E+07
2	MODELO(I1;I2,I4)	1.604693E+07
3	MODELO(I1;I2,I5)	1.579953E+07
4	MODELO(I1;I2,I6)	1.584052E+07
5	MODELO(I1;I2,I7)	1.386274E+07 •
6	MODELO(I1;I2,I8)	1.535024E+07
7	MODELO(I1;I2,I9)	1.600509E+07
8	MODELO(I1;I2,I10)	1.563808E+07
9	MODELO(I1;I2)	1.607806E+07 •
10	MODELO(I1;I3)	2.220269E+07
11	MODELO(I1;I4)	2.227367E+07
12	MODELO(I1;I5)	2.188311E+07
13	MODELO(I1;I6)	2.197912E+07
14	MODELO(I1;I7)	1.769615E+07
15	MODELO(I1;I8)	2.037356E+07
16	MODELO(I1;I9)	2.20513E+07
17	MODELO(I1;I10)	1.926415E+07
18	MODELO(I1;{ })	2.229806E+07

Observamos que el mejor modelo o modelo MAIC, entre los modelos con dos variables explicativas es el MODELO(I1;I2,I7), este resultado era de esperar ya que las variables I2 e I7 eran las más significativas.

Por otro lado, el mejor modelo o modelo MAIC, entre los modelos con una variable explicativa es el MODELO(I1;I2), este resultado también era de esperar puesto que la variable I2 era la más significativa.

Entre estos dos modelos se eligio el MODELO(I1;I2,I7) ya que tiene el menor AIC, luego el conjunto óptimo de variables explicativas de la variable respuesta "número de hijos nacidos vivos" esta formado por las variables explicativas "estado civil", y "edad actual".

BIBLIOGRAFIA

[1] **AKAIKE, H. (1973)** : "*Information theory and an extension of the maximum likelihood principle*". 2nd International Symposium on Information Theory. B. N. Petrov and F. Csáki eds., Akademiai Kiado, Budapest. 267 - 281.

[2] **AKAIKE, H. (1977)** : "*On entropy maximization principle*". North-Holland Publishing Company, 27 - 41.

[3] **AKAIKE, H. (1991)** : "*Statistical inference and measurement of entropy*". En edición.

[4] **BERTLET, M. S. (1935)** : "*Contingency table interactions*". J. Roy. Statist. Soc. Suppl. 2, 248 - 252.

[5] **BISHOP, Y.M. (1969)** : "*Full contingency tables, logits and split contingency tables*". Biometrics 25, 383 - 400.

[6] **BISHOP, Y.M.; FIENBERG, S.; HOLLAND, P. (1975)** : "*Discrete multivariate analysis: Theory and Practice*". Cambridge, Massachusetts. The MIT Press.

[7] **BOLTZMANN, L. (1877)** : "*Uber die Beziehung zwischen dem zweiten Hauptzatz der mechanischen Warmetheorie und der WahrscheinlichKeitrechnung respective den Satzen uber das Warmegleichkeitrechnung respective den Satzen uber das Warmeglichgewicht*". Wiener Berichte, 76, 373 - 435.

[8] **CRESSIE, N., READ, T. (1988)** : "*Goodness-of-fit Statistics for Discrete Multivariate Data*". Springer - Verlag New York, Berlin, Heidelberg, London, Paris, Tokyo.

[9] **CRESSIE, N.; READ, T. (1984)** : "*Multinomial goodness-of-fit tests*". Journal of the Royal Statistical Society Series B, 46, 440 - 464.

[10] **DARROCH, J. N. (1962)** : "*Interactions in multifactor contingency tables*". Journal Royal Statistical Society Series B, 24, 251 - 263.

[11] **GOODMAN, L. A. (1970)** : "*The multivariate analysis of qualitative data interactions among multiple classifications*". J. Amer. Statist. Assoc. 65, 226-256.

[12] **INGA, R.M. (1990)** : "*La informacion de Kullback en el analisis de datos categorizables*". Universidad Complutense de Madrid.

[13] **KU, H. H.; KULLBACK, S. (1974)** : "*Log linear models in contingency table analysis*". Amer. Statistician, 28, 115 - 122.

[14] **KULLBACK, S. (1959)** : "*Information theory and statistics*". John Wiley and Sons, Inc. London.

[15] **LINDGREN, B. (1976)** : "*Statistical theory*". MacMillan Publishing Co., Inc.

[16] **READ, T. R. C. (1984)** : "*Small-sample comparisons for the power divergence goodnesss-of-fit statistics* ". Journal of the American Statistical Association 79, 929 - 935.

[17] **SAKAMOTO, Y. (1977)** : "*A model for the optimal pooling of categories of the predictor in a contingency table*". Research Memo., No. 119. The Institute of Statistical Mathematics. Tokyo.

[18] **SAKAMOTO, Y., AKAIKE, H. (1978)** : "*Analysis of cross classified data by* AIC". Ann. Inst. Statist. Math. Vol. 30 B, No. 1, 185 - 197.

[19] **SAKAMOTO, Y.; AKAIKE, H. (1978)** : "*Robot data screening of cross-classified data by an information criterion*". Proc. International Conference on Cybernetics and Society IEEE, New York, 398 - 403.

[20] **SAKAMOTO, Y. (1982)** : "*Efficient use of Akaike's information criterion for model selection in high dimentional contingency table analysis*". Metron. 40, 257 - 275.

[21] **SAKAMOTO, Y.; ISHIGURO, M.; KITAGAWA,G. (1986)** : "*Akaike information criterion statistics*". KTK Scientific Publishers / Tokyo.

[22] **WERMUTH, N. (1976)** : "*Analogies between multiplicative models in contingency tables and covariance selection*". Biometrics, 32, 95 - 108.

[23] **WERMUTH, N. (1976)** : "*Model search among multiplicative models*". Biometrics, 32, 253 - 263.