TRATAMIENTO DE MANOVAS COMO CASOS PARTICULARES DE LOS MODELOS DE SELECCIÓN DE COVARIANZAS o MODELOS GAUSSIANOS MIXTOS

AUTORES:

Miguel Ángel Fajardo Caldera

fajardo@unex.es

Lidia Andrades Caldito

andrades@unex.es

Jesús Pérez Mayo

jperez@unex.es

Dpto. Economía Aplicada y OO.EE

Universidad de Extremadura

PALABRAS CLAVES:

Análisis Multivariante. Modelos Gaussianos Mixtos. MANOVA.

ÁREA TEMÁTICA:

7, Métodos cuantitativos

ABSTRACT:

Los modelos de ANOVA multivariante, (MANOVA); tanto heterocedástico como homocedástico, pueden ser tratados de una forma elegante y precisa a través de los Modelos de selección de covarianzas (Modelos Gaussianos Mixtos).

Estos modelos que constan de una parte discreta (multinomial) y otra continua (normal multivariante) pueden formularse a través de los parámetros canónicos. Aplicando un conjunto de restricciones a estos parámetros canónicos es posible analizar como casos particulares el MANOVA.

Una explicación de la teoría subyacente en estos modelos y una aplicación de estas técnicas configuran este trabajo.

1. Introducción

1.1 Los Modelos Mixtos

Se trata de una familia de modelos para variables discretas y continuas que combina y generaliza los modelos basados en las distribuciones Multinomial y Normal Multivariante.

Los modelos gráficos para variables discretas y continuas fueron introducidos por Lauritzen y Wermuth (1989), quienes describieron tanto los dirigidos como los no dirigidos. Los modelos gráficos no dirigidos, (modelos de interacción gráfica) fueron generalizados a una clase más amplia, los modelos de interacción jerárquicos, por Edwards (1990). Estos se construyen combinando los modelos Loglineal para variables discretas con los modelos gráficos Gaussianos para variables continuas, como a continuación podrá verse.

1.2 La Distribución Gaussiana Condicional

Supongamos que tenemos p variables discretas y q variables continuas, denotando el conjunto de variables por Δ y Γ respectivamente. Escribiremos las variables aleatorias correspondientes como *(I,Y),* y una observación cualquiera como *(i,y)*. Aquí, i será vector p-dimensional que contendrá los valores de la variable discreta, e y es el vector que contiene las variables continuas de dimensión q. Denotaremos por τ al conjunto de todos los i posibles.

Supongamos que la probabilidad de que I=i es p_i y que la distribución de Y dado $I=i$ es normal Multivariante $N(\mu_i, \Sigma_i)$ por lo que la media y la covarianza condicional pueden depender de i. Esto es conocido como la distribución Gaussiana condicional. La densidad puede escribirse como:

$$f(i,y) = p_i \left|2\pi\Sigma_i\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)\right\} \qquad (1)$$

Los parámetros $\{p_i, \mu_i, \Sigma_i\}_{\forall i \in \tau}$ son llamados parámetros momentos.

En general, solemos interesarnos en los modelos cuya covarianza es constante sobre i, por lo que $\Sigma_i = \Sigma$. A estos modelos se les denomina homogéneos. Como se verá más tarde, hay generalmente dos tipos de modelos gráficos correspondientes a un grafo dado: un modelo heterogéneo y otro homogéneo.

La ecuación (1) podemos escribirla del siguiente modo:

$$f(i,y) = \exp\left\{\alpha_i + \beta_i' y - \frac{1}{2} y'\Omega_i y\right\} \qquad (2)$$

donde α_i es un es un escalar, β_i es un vector de dimensión *px1*, y Ω_i es una matriz definida positiva, simétrica, de dimensión *pxp*. Estos son denominados parámetros canónicos. Para pasar de los parámetros momentos a los canónicos utilizamos las siguientes expresiones:

$$\Omega_i = \Sigma_i^{-1}, \qquad (3)$$

$$\beta_i = \Sigma_i^{-1}\mu_i, \qquad (4)$$

$$\alpha_i = \ln(p_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i - \frac{q}{2}\ln(2\pi), \qquad (5)$$

y

$$\Sigma_i = \Omega_i^{-1}, \qquad (6)$$

$$\mu_i = \Omega_i^{-1}\beta_i, \qquad (7)$$

$$p_i = (2\pi)^{q/2}|\Omega_i|^{-1/2}\exp\left\{\alpha_i + \frac{1}{2}\beta_i\Omega_i^{-1}\beta_i\right\} \qquad (8)$$

Los modelos de interacción jerárquica son construidos restringiendo los parámetros canónicos de un modo similar a los modelos loglineales. Es decir, los parámetros canónicos son desarrollados como sumas de términos de interacción, y los modelos son definidos estableciendo como nulos los términos de interacción de orden superior. Para facilitar su comprensión, veremos algunos ejemplos.

1.3 Fórmula de los Modelos Gaussianos Mixtos (MGM)

Continuaremos con el caso general. Supongamos que la fórmula de un modelo es:

$$d_1,...,d_r / l_1,...,l_s / q_1,...q_t. \qquad (9)$$

La primera parte se refiere a los generadores discretos especificados por la expansión de α_i. La segunda parte corresponde a los generadores lineales especificados por la expansión de β_i. Cada generador lineal contiene una variable continua. La expansión para β_i^γ para cualquier $\gamma \in \Gamma$ viene dada por los generadores lineales que contienen γ. La tercera la parte cuadrática proporciona la expansión para la matriz de covarianza inversa Ω_i. Cada generador cuadrático contendrá al menos una variable continua. La expansión para $\omega_i^{\gamma\zeta}$ para $\gamma,\zeta \in \Gamma$ viene dada por los generadores cuadráticos contenidos en γ,ζ.

Dos reglas de sintaxis restringen los valores posibles de la fórmula:

- El generador lineal no debe ser más largo que el generador discreto, es decir, por cada generador lineal l_j debe corresponder a un generador discreto d_k tal que $l_j \cap \Delta \subseteq d_k$.

Por ejemplo, A,B/ABX/AX no sería admisible ya que aparece un generador lineal ABX pero no aparece el generador discreto conteniendo AB.

- Los generadores cuadráticos, no podrán ser más largos que los generadores lineales correspondientes, es decir, para cada generador cuadrático q_j y cada variable continua $\gamma \in q_j$, debe existir el correspondiente generador lineal l_k tal que $(q_j \cap \Delta) \cup \{\gamma\} \subseteq d_k$.

  Por ejemplo, ABC/AX,BY,CZ/AXY,CZ no sería posible debido a que aparece un generador cuadrático AXY pero ningún generador lineal que contenga AY.

Estas reglas se establecen para asegurar que los modelos no varíen ante cambios de origen y escala de las variables continuas.

1.4 Relación entre la fórmula de un MGM y el grafo que lo representa

Para estudiar la correspondencia entre la fórmula de un modelo y el grafo a través del cual lo representamos, expandiremos la ecuación 2 del siguiente modo:

$$f(i,y)\exp\left\{\alpha_i + \sum_{\gamma\in\Gamma}\beta_i^{\gamma} y_{\gamma} - \frac{1}{2}\sum_{\gamma\in\Gamma}\sum_{\eta\in\Gamma}\omega_i^{\gamma\eta} y_{\gamma} y_{\eta}\right\} \qquad (10)$$

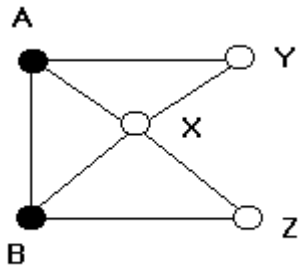
y después aplicaremos el criterio de factorización para analizar las independencias por pares de Markov del modelo dado.

Para las dos variables discretas del modelo, A y B, $A \perp B|(resto)$ cuando todas las interacciones que involucren a A y B sean siempre cero. Es decir, ninguna de las expansiones de α_i, β_i^{γ} o $\omega_i^{\gamma\eta}$, para cualquier $\gamma,\eta\in\Gamma$ puede contener la interacción AB. En términos de la fórmula del modelo, implica simplemente que si el generador discreto no contiene AB, por las reglas de sintaxis sabremos que ni los generadores cuadráticos ni los lineales podrán contener AB.

Si A es discreto y X es continuo, y vemos que $A \perp X|(resto)$ significará que cualquier término que involucre la interacción de A con la de X deberá ser cero. Es decir, ninguna expansión de β_i^X ó $\omega_i^{X\eta}$ para cualquier $\eta\in\Gamma$ puede incluir un término de interacción que contenga a A. En relación a la fórmula del modelo, significa que ningún generador lineal puede contener AX, y por las reglas de sintaxis sabremos también que ningún generador cuadrático tampoco contendrá AX.

Para dos variables continuas, X e Y, $X \perp Y|(resto)$ conlleva que cualquier ω_i^{XY} será nulo. En términos de la fórmula del modelo significa que ningún generador cuadrático podrá contener XY.

Estos resultados se pueden derivar con facilidad del grafo de independencia construido a partir de la fórmula de un modelo. Por ejemplo, el grafo de $AB / AX, BX, AY, BZ / XY, XZ$ es:



Para realizar la operación inversa, es decir, encontrar la fórmula de un modelo gráfico a partir del grafo correspondiente G, necesitaremos identificar la máxima interacción que sea consistente con G. Supongamos que tenemos el grafo anterior, y necesitamos encontrar los cliques G_Δ, es decir, el subgrafo de G inducido por las variables discretas. En el grafo representado antes sería AB, luego la primera parte de la fórmula será AB.

Para la parte lineal de la fórmula, deberemos encontrar los cliques de $G_{\Delta\cup\{\gamma\}}$ para cada $\gamma \in \Gamma$. Observamos que estos serán los generadores ABX, AY y BZ, luego la parte lineal será $ABX, AY, BZ \cdot$

Finalmente para la parte cuadrática, esta dependerá en si estamos interesados en el modelo gráfico homogéneo o en el heterogéneo. Para el modelo homogéneo, necesitaremos identificar los cliques en G_Γ. En este ejemplo en concreto, los cliques de G_Γ son $\{X, Y\}$ y $\{X, Z\}$, por lo que nuestra fórmula tomaría la siguiente forma:

$$AB / AX, BX, AY, BZ / XY, XZ \cdot$$

Para el modelo heterogéneo, tendremos que encontrar los cliques de G que presenten alguna intersección con Γ. En nuestro ejemplo estos serían $\{A, X, Y\}, \{A, B, X\}$ y $\{B, X, Z\}$, por lo que la fórmula del modelo que obtendríamos sería:

$$AB / ABX, AY, BZ / AXY, ABX, BXZ \cdot$$

1.5 Estimación por máxima verosimilitud

Los modelos por su naturaleza necesitan datos. Supongamos que tenemos una muestra de N observaciones independientes e idénticamente distribuidas $\left(i^{(k)}, y^{(k)}\right)$ para $k = 1,...N$. Donde i representa las p dimensiones de la variable discreta con sus correspondientes niveles, e y es un vector q-dimensional. Sean $\left(n_j, t_j \bar{y}_j, SS_j, S_j\right)_{j\in I}$ las frecuencias observadas, las sumas totales de las variables, las medias y las sumas de cuadrados y productos sin corregir, variando los elementos para cada j, es decir,:

$$
\begin{aligned}
n_i &= \neq \left\{k : i^{(k)} = i\right\} \\
t_i &= \sum_{k:i^{(k)}=i} y^{(k)}, \\
\bar{y}_i &= t_i / n_i \ , \\
SS_i &= \sum_{k:i^{(k)}=i} y^{(k)} \left(y^{(k)}\right)', \\
S_i &= \sum_{k:i^{(k)}=i} \left(y^{(k)} - \bar{y}_i\right)\left(y^{(k)} - \bar{y}_i\right)' / n_i \\
&= \left(SS_i / n_i\right) - \bar{y}_i \ \bar{y}_i{}' .
\end{aligned}
$$

También necesitaremos una notación para las magnitudes marginales correspondientes. Para $a \subseteq \Delta$, denotaremos al elemento marginal correspondiente a i como i_a y del mismo modo para $d \subseteq \Gamma$, escribiremos el subvector de y como y^d. Igualmente, escribiremos las frecuencias marginales como $\left\{n_{i_a}\right\}_{i_a \in I_a}$, los totales marginales de las variables como $\left\{t_{i_a}^d\right\}_{i_a \in I_a}$, y las sumas de cuadrados y productos sin corregir como $\left\{SS_{i_a}^d\right\}_{i_a \in I_a}$.

Supongamos a continuación un modelo dado por la fórmula $d_1,...,d_r / l_1,...,l_S / q_1,...,q_t$. A partir de la expresión (2), se demuestra directamente el conjunto de estadísticos mínimo suficientes viene dado por

1. Un conjunto de frecuencias marginales $\left\{n_{i_a}\right\}_{i_a \in I_a}$ correspondiente a los generadores discretos, es decir, para $a = d_1,...,d_r$.
2. Un conjunto de totales marginales de las variables $\left\{t_{i_a}^d\right\}_{i_a \in I_a}$, correspondiente a los generadores lineales, es decir, para $a = l_j \cap \Delta, \gamma = l_j \Gamma, \quad \forall \quad \mathrm{j} = 1,...,\mathrm{s}.$
3. Un conjunto de sumas y cuadrados marginales sin corregir $\left\{SS_{i_a}^d\right\}_{i_a \in I_a}$ correspondientes a los generadores cuadráticos, es decir, para $a = q_j \cap \Delta,$ y $\mathrm{d} = \mathrm{q_j} \cap \Gamma$ para $j = 1,...,t$.

Como ya se vio, los modelos se construyen restringiendo los parámetros canónicos a través de sus expansiones factoriales. Dado un conjunto de datos, estimaremos por máxima verosimilitud los parámetros de un modelo sujeto a dichas restricciones. A partir de la teoría de la familia exponencial, sabemos que los (EMV estimadores de máxima verosimilitud) se obtienen igualando los estadísticos mínimo suficientes con sus valores esperados. Es decir, para $a = d_1,...,d_r$,

$$\left\{n_{i_a}\right\}_{i_a \in I_a} = \left\{m_{i_a}\right\}_{i_a \in I_a} \qquad (11)$$

para $a \cup \gamma = l_1,...,l_S$,

$$\left\{{}^{\gamma}_{i_a}\right\}_{i_a \in I_a} = \left\{\sum_{j:J_a = i_a} m_j \mu_j^{\gamma}\right\}_{i_a \in I_a} \qquad (12)$$

y para $a \cup d = q_1,...,q_t$,

$$\left\{SS^{d}_{i_a}\right\}_{i_a \in I_a} = \left\{\sum_{j:J_a = i_a} m_j \left[\Sigma_j^{dd} + \mu_j^{d}\left(\mu_j^{d}\right)'\right]\right\}_{i_a \in I_a} \qquad (13)$$

estas son conocidas como las ecuaciones de verosimilitud. Cuando los EMV existan, la solución será única[1] y satisfará todas las restricciones del modelo. Generalmente las ecuaciones deberán resolverse a través de un proceso iterativo. Se puede utilizar el algoritmo EPIM (escalamiento proporcional iterativo modificado) descrito por Frydenberg y Edwards (1989).

1.7 Contraste de la bondad de un modelo: La Desviance

La expresión de la desviance, utilizada para contrastar la bondad de los modelos estimados, se puede escribir como:

$$\ln f(i,y) = \ln p_i - q\ln(2\pi)/2 - \ln|\Sigma_i|/2 - (y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)/2$$

por lo que el logaritmo de la función de verosimilitud muestral $\left(i^{(k)}, y^{(k)}\right)$ k = 1,..., N es

$$l = \sum_i n_i \ln p_i - Nq\ln(\pi)/2 - \sum_i n_i \ln|\Sigma_i|/2 - \sum_{k=1}^{N}\left(y^{(k)} - \mu_{i^{(k)}}\right)'\Sigma_{i^{(k)}}^{-1}\left(y^{(k)} - \mu_{i^{(k)}}\right)/2$$

el último término puede simplificarse como

$$\sum_i \sum_{k:i^{(k)}=i}\left(y^{(k)} - \mu_{i^{(k)}}\right)'\Sigma_{i^{(k)}}^{-1}\left(y^{(k)} - \mu_{i^{(k)}}\right) = \sum_i \left\{n_i tr\left(S_i\Sigma_i^{-1}\right) + n_i\left(\bar{y}_i - \mu_i\right)'\Sigma_i^{-1}\left(\bar{y}_i - \mu_i\right)\right\}$$

Por lo que una expresión alternativa del logaritmo de verosimilitud es:

$$l = \sum_i n_i \ln p_i - Nq\ln(2\pi)/2 - \sum_i n_i \ln|\Sigma_i|/2 - \sum_i n_i tr\left(S_i\Sigma_i^{-1}\right)/2 - \sum_i n_i\left(\bar{y}_i - \mu_i\right)'\Sigma_i^{-1}\left(\bar{y}_i - \mu_i\right)/2$$

[1] Las condiciones generales de existencia son complejas incluso para el caso discreto (Glonek, Darroch, y Speed, 1988)

Para el modelo saturado heterogéneo, a partir de las ecuaciones de verosimilitud (11-13) obtenemos que $\hat{p}_i = n_i/N$, $(\hat{m}_i = N\hat{p}_i)$, $\hat{\mu}_i = \bar{y}_i$, y $\hat{\Sigma}_i = S_i$, por lo que el máximo del logaritmo de verosimilitud es

$$\hat{l}_f = \sum_i n_i \ln(n_i/N) - Nq\ln(2\pi)/2 - \sum_i n_i \ln|S_i|/2 - Nq/2 \qquad (14)$$

Por lo que si un modelo tiene EMV $\hat{p}_i, \hat{\mu}_i, \hat{\Sigma}_i$, su desvianza respecto al modelo saturado homogéneo será:

$$2\sum_i n_i \ln(n_i/\hat{m}_i) - \sum_i n_i \ln\left|S_i\hat{\Sigma}_i^{-1}\right| + \sum_i n_i \left\{tr\left(S_i\hat{\Sigma}_i^{-1}\right) - q\right\} + \sum_i n_i (\bar{y}_i - \hat{\mu}_i)'\hat{\Sigma}_i^{-1}(\bar{y}_i - \hat{\mu}_i)$$

El modelo saturado homogéneo tiene estimadores

$$\hat{p}_i = n_i/N,\ (\hat{m}_i = N\hat{p}_i),\ \hat{\mu}_i = \bar{y}_i,\quad \hat{\Sigma} = \hat{\Sigma}_i = S,$$

$$\text{donde } S = \sum_i n_i S_i/N,$$

por lo que el logaritmo de máxima verosimilitud para este modelo será

$$\hat{l}_f^h = \sum_i n_i \ln(n_i/N) - Nq\ln(2\pi)/2 - N\ln|S|/2 - Nq/2 \qquad (15)$$

y la desviazna de un modelos homogéneo con estimadores de máxima verosimilitud $\hat{p}_i, \hat{\mu}_i$, y $\hat{\Sigma}$ con respecto al modelo saturado homogéneo se simplifica a la siguiente expresión

$$2\sum_i n_i \ln(n_i/\hat{m}_i) - N\ln\left|S\hat{\Sigma}^{-1}\right| + N\left\{tr\left(S\hat{\Sigma}^{-1}\right) - q\right\} + \sum_i n_i (\bar{y}_i - \hat{\mu}_i)'\hat{\Sigma}^{-1}(\bar{y}_i - \hat{\mu}_i)$$

Para dos modelos $M_0 \subseteq M_1$, la diferencia de desvianzas sigue bajo M_0 una distribución asintótica χ^2 con unos grados de libertad dados por la diferencia entre los dos modelos del número de parámetros no restringidos.

2. MGM y MANOVA

Vistos los MGM, analizaremos su relación con los MANOVA (Modelos de análisis multivariante de la varianza). Para simplificar la exposición, presentaremos el caso de un MGM con una variable discreta y otra continua. Es decir, el caso en el que *p=q=1,* es decir. Sea *Δ={A}* y *Γ={Y}.* La función de densidad la escribiremos como:

$$f(i,y) = p_i(2\pi\sigma_i)^{-1/2}(y-\mu_i)^2/\sigma_i$$
$$= \exp\left\{\alpha_i + \beta_i y - \frac{1}{2}\omega_i y^2\right\} \qquad (16)$$

Sustituyendo los parámetros canónicos con los términos de interacción desarrollados, podemos escribir la expresión (16) como:

$$f(i,y) = \exp\left\{(u+u_i^A) + (v+v_i^A)y - \frac{1}{2}(\omega+\omega_i^A)y^2\right\}$$

El parámetro canónico cuadrático es $\omega_i = \omega + \omega_i^A$. Por lo que $\sigma_i = \omega_i^{-1}$ los elementos de la matriz de varianza serán constantes cuando $\omega_i^A = 0$.

El parámetro canónico lineal es $\beta_i = v + v_i^A$. Como $\mu_i = \omega_i^{-1}$, los elementos del vector de medias serán constantes cuando $\omega_i^A = v_i^A = 0$. También utilizando el criterio de factorización[2] podemos ver que $A \perp Y$ si y sólo si $\omega_i^A = v_i^A = 0$.

El parámetro canónico discreto es $\alpha_i = u + u_i^A$, se trata de el efecto principal del modelo Loglineal y no puede simplificarse, es decir no podemos establecer u_i^A.

Podremos considerar tres modelos posibles. El modelo más simple de independencia marginal, formado asumiendo $\omega_i^A = v_i^A = 0$. La función de densidad que lo representa será:

$$f(i,y) = p_i(2\pi\sigma)^{-1/2}\exp\left\{-\frac{1}{2}(y-\mu)^2 \Big/ \sigma^2\right\},$$

La fórmula de este modelo es *A/Y/Y*.

El segundo modelo, formado asumiendo $\omega_i^A = 0$, permite variar el vector de medias, pero obliga a las varianzas a ser homogéneas. La función de densidad será:

$$f(i,y) = p_i(2\pi\sigma)^{-1/2}\exp\left\{-\frac{1}{2}(y-\mu_i)^2 \Big/ \sigma^2\right\},$$

y la fórmula del modelo será *A/AY/AY*.

Como muestran estos ejemplos, la fórmula de un modelo para los modelos mixtos, esta formada por tres partes separadas por salsees (/). Las tres partes especifican las interacciones expandidas de α_i, β_i y Ω_i, respectivamente. De este modo, en el segundo y tercer modelo antes vistos, la segunda parte de la fórmula era *AY*; lo que significa que el elemento β_i

[2] $f_{XYZ}(x,y,z) = h(x,z)k(y,z)$

correspondiente a Y posee un a expansión con fórmula A, es decir, $\beta_i = \nu + \nu_i^A$. En el modelo $A/AY/Y$, el término Y indica que el parámetro canónico cuadrático ω_i^{YY} tiene una fórmula nula, es constante en sus elementos: $\omega_i^{YY} = \omega$.

Estos tres modelos están muy relacionados con los modelos ANOVA de una vía: la única diferencia está en la estructura, los elementos se asumen aleatorios. Si consideramos fijo A, tendremos la estructura del ANOVA de una vía, y el primer modelo denota homogeneidad contra independencia.

La generalización a los MANOVAs es inmediata.

2.1 Resolución de un caso práctico:

Del Anuario Económico de España (2001), donde se recoge información sobre cada uno de los municipios de más de 1000 habitantes existentes en España, hemos tomado una muestra aleatoria de 242 municipios menores de 15000 habitantes, de los que hemos seleccionado las variables :nivel económico (A), población (Z), índice industrial (X) e índice comercial (Y).

Para la variable nivel económico hemos considerado cuatro niveles de los diez que se definen en el anuario, y que se corresponden con la renta familiar disponible por habitante (Ptas.) del año 1999.

Así,

nivel 1 : 1000000 - 1100000
nivel 2 : 1100000 - 1300000
nivel 3 : 1300000 - 1500000
nivel 4 : más de 1500000

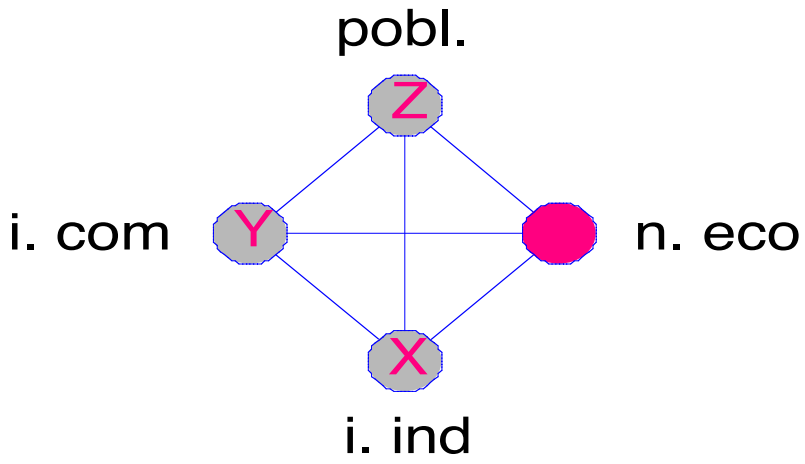
La variable población se ha considerado la definida por la correspondiente al padrón de 1 de enero de 1999.

La variable índice industrial, es un índice comparativo de la importancia de la industria (incluida la construcción) de cada municipio, referido a 1999.Este índice se elabora en función del impuesto de actividades económicas (IAE) correspondiente a las actividades industriales.

La variable índice comercial, es un índice comparativo de la importancia del comercio (comercio mayorista y minorista) de cada municipio, referido a 1999. Este índice se elabora en función del impuesto de actividades económicas (IAE) correspondiente a las actividades del comercio mayorista y comercio minorista conjuntamente.

Las relaciones entre las variables consideradas anteriormente se analizan a través del modelo saturado, a un nivel crítico del 0.05, mediante el modelo inicial

dado por : *A/AX;Y;AZ/AXYZ* ; cuyo grafo correspondiente viene dado por :



ARCOS	TEST ESTADISTICO	G.L.	P
[AX]	155.6516	12	0.000
[AY]	102.2967	12	0.000
[AZ]	88.6975	12	0.000
[XY]	32.9335	4	0.000
[XZ]	30.0077	4	0.000
[YZ]	228.9661	4	0.000

Todos las líneas son altamente significativas ; por lo que se acepta el modelo MANOVA no homogéneo, (distintas matrices de covarianzas).

El contraste entre los modelos saturados homogéneo y heterogéneo da los siguientes resultados :

TEST :

HO : A/AX,AY,AZ/XYZ

H1 : A/AX,AY,AZ/AXYZ

LR : 214.2866 g.l. : 18 P : 0.0000, rechazamos la hipótesis HO.

Si consideramos ahora como variable independientes (fijas) a la variables A y Z y a las variables dependientes X e Y ; obtenemos los siguientes sistemas de ecuaciones condicionadas :

A= 1 ; E[X] = -0.033 + 0.001 Z Var (X) = 40.778 Cov(X,Y)= 8.015

E[Y] = -2.316 + 0.002 Z Var (Y) = 14.475

A= 2 ; E[X] = -6.172 + 0.001 Z Var (X) = 541.966

Cov(X,Y)= 9.422

E[Y] = -2.850 + 0.002 Z Var (Y) = 15.001

A= 3 ; E[X] = -1.122 + 0.004 Z Var (X) = 268.483

Cov(X,Y)=-26.093

E[Y] = -2.316 + 0.002 Z Var (Y) = 26.163

A= 4 ; E[X] = -1.037 + 0.006 Z Var (X) = 489.593

Cov(X,Y)= 131.080

E[Y] = 2.297 + 0.002 Z Var (Y) = 88.840

Como podemos observar por las estimaciones realizadas anteriormente, a menor nivel de renta familiar disponible por habitante (niveles 1 y 2) el tamaño poblacional tiene el doble de peso en el índice comercial que en el índice industrial ; mientras que conforme va subiendo la renta familiar disponible (niveles 3 y 4) el tamaño poblacional tiene un peso doble y triple para el índice industrial que para el comercial. Por todo ello podríamos indicar que en niveles de renta familiar altos, el factor población afecta más positivamente a la industria que al comercio, al revés de aquellos municipios de renta baja, donde el comercio es más importante que su industria.

Bibliografía

BIBLIOGRAFÍA

Dawid, A.P. (1.979) Conditional independence in Statistical Theory (with discussion). J. Roy. Statist. Soc. Bb, 41,1,1-31.

Dawid, A.P. (1.980) Conditional Independence for Statistical Operations. Ann. Statist.,8, 598-617.

Dempster, A.P. (1.972) Covariance Selection. Biometrics 28, 157-175.

Edwards, D.E. (1.987) A guide to MIM. Research Report 87/1. Statistical Research Unit, University of Copenhagen.

Edwards, D.E. (1.995) Introduction to Graphical Modelling. Springer-Verlag New York, Inc.

Frydenberg, M. and Edwards, D.E (1.989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. Comput. Statist. Data Anal. To appear.

Goldberger, A. G. and Duncan, O.D. (1.973). Structural Equations Models in the social sciences . Seminar Press: New York.

Hawkins, D.M. and Eplett, W.J. (1.982). The Cholesky factorisation of the inverse correlation matrix in multiple regression. Technometrics, 24, 191-198.

Jöreskog, K.G (1.981) Analysis of Covariance Structures. Scan. J. Statist. 8, 65-92.

Knuinman, A. (1.978). Covariance Selection. Suppl. Adv. Appl. Prob., 10, 123-130.

Kullback, S. (1.967). A lower bound for discrimination information in terms of variation. IEEE Trans.

Lauritzen, S.L and Wermuth, N (1.989). Graphical Models for associations between variables, some of which are qualitative and some are quantitative. Ann. Statist. 17, 31-54.

Leimer, H-G (1.993). Optimal decomposition by clique separators. Discrete match. 113, 99-123.

Pearl, J. & Wermuth, N. (1.994) When can association graphs admit a causal interpretation? In P. Choseman and Woldford (eds) Models and data, Artificial Intelligence and Statistics IV New York.

Speed, T, Pand Kiiveri, H. T, (1.986) Gaussian Markov distributions over finite graphs. Ann. Statist. 14, 138-150.

Wermuth, N. (1.980) Linear recursive equations, covariance selection and path analysis. J. amer. Statist. Assoc. 75, 963-972.

Wermuth, N. (1.988) Introduction to the use of graphical chain models. Unpublished tutorial notes to Compstat 88.

Wermuth, N. and Lauritzen, S.L. (1.990). On substantive research hypotheses, conditional independence graphs and graphical chain models. J. Roy. Statit. Soc. B, 52, 1, 21-50.

Whittaker, J. Iliakopoulis, T. and Smith, P. (1.988). Graphical modelling with large number of variables: a comparison with principal components. In Edwards, D. G. and N. E. Raun (Eds) Compstat 88. Physica Verlag: Heidelberg. 73-80.

Whittaker, J. (1.990) Graphical Models in Applied Multivariate Statistics. Department of Mathmatics University of Lancaster. UK. John Wiley & Sons Ltd.

Wright, S. (1.954). The interpretation of multivariante systems. In Kempthorne, O. Et al. (Eds) Statistic and Mathematics Biology. Pp. 11-23. Iowa State University Press: Ames.

Pearl, J. (1.993) Aspects of graphical models connected with causality. Bull. Int. Stat. Inst., Proceedings 49th Session 1: 391-403.

Smith, P. W. F. (1.992). Assessing the power of model selection procedures used when graphical modelling. In Dodge, Y. and Whittaker, J (Eds) Computational Statistic, Proceedings, Vol I p.275-280. Physica-Verlag, Heidelberg.

Whittaker, J. (1.990) Graphical Models in Applied Multivariate Statistic. Wiley.