Badler, Clara E.
Alsina, Sara M. [1]
Puigsubirá, Cristina R . [1]
Vitelleschi, María S.[1]

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE)

VARIABLES MIXTAS CON INFORMACIÓN FALTANTE Y/O CONFUSA. ESTIMACIÓN MÁXIMO VEROSÍMIL CON R

1. INTRODUCCIÓN

Es frecuente la presencia de información faltante y/o confusa en bases de datos con variables mixtas, afectando tanto las variables categóricas como a las continuas.

La mayoría de los métodos estadísticos multivariados han sido formulados enfatizando modelos para variables de un solo tipo, continuas o categóricas, y se le ha dedicado menos atención a los modelos con ambos tipos de variables simultáneamente.

Un modelo que incorpora ambos tipos de variables es el modelo de locación general, así llamado por Olkin y Tate, que permite realizar la estimación de los parámetros a partir de la información conjunta de todas las variables intervinientes.

Para el caso en que se desee estimar los parámetros de dicho modelo a partir de una base de datos con información faltante y/o confusa en ambos tipos de variables se requiere metodología específica.

En este trabajo se presenta un método de estimación máximo verosímil utilizado por Little y Schluchter para la estimación de los parámetros del modelo de locación general cuando ambos tipos de variables presentan información faltante y/o confusa bajo el supuesto que el mecanismo de pérdida es ignorable. Dicha metodología se aplica a una base de datos del área biológica.

2. MATERIAL

Se trabaja con una base de datos de niños y adolescentes diabéticos atendidos en establecimientos asistenciales públicos y privados de la ciudad de Rosario en el año 2003. Las variables utilizadas son:

• ”Familiar de primer orden diabético” (F): SI - NO;

• “Descenso de peso” (P): SI – NO;

• ”Nivel de glucosa” (G);

• “Dosis diaria de insulina” (D).

[1] Docente-Investigador e Investigador del Consejo de Investigaciones de la Universidad Nacional de Rosario.

El soporte informático utilizado es el paquete "mix" del programa R versión 2.3.1.

3. MÉTODOS

3.1 Modelo de Locación General

El modelo de locación general, así llamado por Olkin y Tate, está definido en términos de la distribución marginal de las variables categóricas y la distribución condicional de las variables continuas dadas las categóricas.

3.1.1 Modelo y estimadores de los parámetros

Sea la matriz de datos **M**=(**Y**, **X**) con n observaciones registradas para q variables categóricas (**Y**) y p variables continuas (**X**). La variable categórica j-ésima tiene l_j niveles, de modo que ellas definen una tabla de contingencia de dimensión q con $C = \prod_{j=1}^{q} l_j$ celdas.

Para el sujeto i, sea x_i un vector de variables continuas de dimensión (1xp) e y_i el vector de variables categóricas de dimensión (1xq). Se construye a partir de y_i un vector w_i de dimensión (1xC), tal que $w_i = E_C$ si el sujeto i pertenece a la celda c de la tabla de contingencia y E_C es un vector con 1 en la c-ésima componente y ceros en las restantes.

El modelo de locación general para datos completos especifica la distribución de (x_i,w_i) para i=1,...,n en términos de la distribución marginal de w_i y la distribución condicional de x_i dado w_i:

• Las w_i son variables aleatorias idéntica e independientemente distribuídas según una multinomial con probabilidades de celda $Pr(w_i = E_C) = \pi_C$ para c=1...C y $\sum_{c=1}^{C} \pi_c = 1$.

• Dado que $w_i = E_C$, $(x_i / w_i = E_C)$ se distribuye independientemente según una normal p-variada con vector de promedio $\mu_c = (\mu_{c1}, \mu_{c2},...,\mu_{cp})$ y matriz de covariancias $\boldsymbol{\Omega}$.

El modelo de locación general tiene $(C-1)+pC+\frac{1}{2}p(p+1)$ parámetros, $\boldsymbol{\theta} = (\Pi, \boldsymbol{\Gamma}, \boldsymbol{\Omega})$ siendo $\Pi = (\pi_1,...,\pi_c)$ el vector de probabilidades de celda y $\boldsymbol{\Gamma} = \{\mu_{cp}\}$ la matriz de promedios de celda.

Cuando se tienen datos completos la verosimilitud para el modelo es:

$$l(\Gamma,\Omega,\Pi) = \sum_{i=1}^{n} \ln f(x_i / w_i, \Gamma, \Omega) + \sum_{i=1}^{n} \ln f(w_i / \Pi) =$$

$$= h(\Omega) - \frac{1}{2} tr(\Omega^{-1} \sum_{i=1}^{n} x_i^T x_i) + tr\Omega^{-1}\Gamma(\sum_{i=1}^{n} w_1^T x_i) + \sum_{c=1}^{C}\left[\left(\sum_{i=1}^{n} w_{ic}\right)\left(\ln \pi_c - \frac{1}{2}\mu_c \Omega^{-1} \mu_c^T\right)\right] \qquad (1)$$

donde w_{ic} es la componente c-ésima de w_i y $h(\Omega)$ es igual a $\left\{-\frac{1}{2} n[K \ln(2\pi)+Ln|\Omega|]\right\}$. Maximizando la ecuación anterior se obtienen los estimadores máximo verosímiles:

$$\hat{\Pi} = n^{-1}\sum_{i=1}^{n} w_i$$

$$\hat{\Gamma} = \left(\sum_{i=1}^{n} x_i^T w_i\right)\left(\sum_{i=1}^{n} w_i^T w_i\right)^{-1} \qquad (2)$$

$$\hat{\Omega} = n^{-1}\sum_{i=1}^{n}\left(x_i - w_i\hat{\Gamma}\right)^T (x_i - w_i\hat{\Gamma})$$

los cuales son, respectivamente, las proporciones de celda observadas, los promedios observados de celda y la matriz de covariancias amalgamada de la matriz **X** dentro de las celdas.

3.1. 2 Estimación de los parámetros con información faltante y/o confusa

Se considera el caso general en el cual un subconjunto arbitrario de Y_1, Y_2,....,Y_q y X_1, X_2,.....,X_p presenta información faltante y/o confusa (Figura 1).

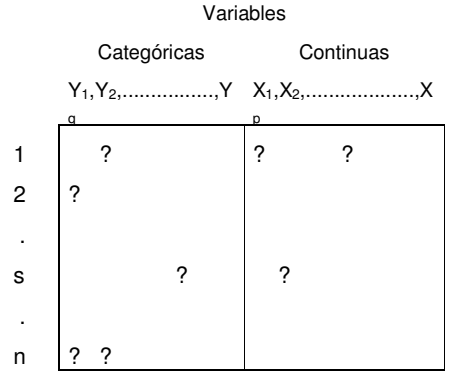

Figura 1. Matriz de datos con información faltante y/o confusa en ambos tipos de variables

Para la unidad i sea:

$x_{obs,i}$: vector de variables continuas observadas,

$x_{per,i}$: vector de variables continuas perdidas y

S_i: conjunto de celdas en la tabla de contingencia donde la unidad i podría estar presente dadas las variables categóricas observadas.

Para la estimación máximo verosímil de θ dado los datos $\{x_{obs,i}; S_i : i = 1,...,n\}$, una alternativa es la utilización del algoritmo EM.

La función de densidad (1) pertenece a la familia exponencial regular de estadísticas suficientes con datos completos $\Sigma x_i^T x_i$, $\Sigma w_i^T x_i$ y Σw_i que son, respectivamente, la suma de cuadrados y productos cruzados de las variables X, los totales de celdas de las variables X y los conteos de celdas. Ello posibilita la aplicación de la forma simplificada del algoritmo EM.

En la iteración t, del paso E, el algoritmo calcula los valores esperados de las estadísticas suficientes de los datos completos dados los datos $\{x_{obs,i}; S_i : i = 1,...,n\}$ y las estimaciones corrientes de los parámetros $\theta^{(t)} = (\Pi^{(t)}, \mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)})$. Las contribuciones del caso i son:

$$T_{1i}^{(t)} = E(x_i^T x_i / x_{obs,i}, S_i, \theta^{(t)}),$$

$$T_{2i}^{(t)} = E(w_i^T x_i / x_{obs,i}, S_i, \theta^{(t)}),$$

El paso M calcula las estimaciones máximo verosímiles de (2) con los datos completos mediante las estimaciones de las estadísticas suficientes de los datos completos obtenidas en el paso E:

$$\Gamma^{(t+1)} = D^{-1}(\sum_{i=1}^{n} T_{2i}^{(t)})$$

$$\Omega^{(t+1)} = n^{-1}(\sum_{i=1}^{n} T_{1i}^{(t)} - (\sum_{i=1}^{n} T_{2i}^{(t)})^{T} D^{-1}(\sum_{i=1}^{n} T_{2i}^{(t)})$$

$$\Pi^{(t+1)} = n^{-1}\sum_{i=1}^{n} T_{3i}^{(t)} ;$$

siendo **D** una matriz con elementos ΣT_{3i} a lo largo de la diagonal principal y el resto ceros. El algoritmo vuelve luego al paso E para recalcular $T^{(t)}_{1i}$, $T^{(t)}_{2i}$ y $T^{(t)}_{3i}$ con las nuevas estimaciones de los parámetros y continúa el ciclo de los pasos E y M hasta la convergencia.

3. 2 Estimación máximo verosímil de los parámetros del modelo de locación general con el programa R

En el paquete "mix" del programa R se dispone de funciones que posibilitan la aplicación de la metodología presentada:

• **prelim.mix (**argumentos**)**

Realiza un tratamiento preliminar sobre la matriz de datos incompletos mediante operaciones de agrupamiento y clasificación, creando los datos de entrada necesarios para la utilización de las funciones em.mix, imp.mix, etc..

Argumentos:

x: matriz de datos con valores faltantes, cuyas filas corresponden a las unidades y las columnas a las variables. Los valores faltantes son codificados con NA. Las variabales categóricas deben figurar en las primeras p columnas y estar codificadas con enteros positivos consecutivos mayores o iguales a 1.

p: número de variables categóricas.

La aplicación de "prelim.mix" proporciona distintos resultados del análisis de la matriz ingresada, que pueden ser requeridos especialmente. Por ejemplo: con "nmis*"* se visualiza una tabla conteniendo el número de valores faltantes para cada variable de la matriz **X**; con "r" se obtiene una tabla que informa sobre los diferentes esquemas de pérdida y la frecuencia con que se presentan, expresados con valores indicadores 1 y 0 para cada variable según la misma haya sido o no observada.

• **em.mix (**argumentos**)**

Calcula las estimaciones máximo verosímiles de los parámetros del modelo de locación general sin restricciones a partir de una matriz de datos incompletos mediante la aplicación del algoritmo EM.

Argumentos:

s: matriz de datos incompletos producida por la función prelim.mix.

start: permite, opcionalmente, establecer un valor inicial del parámetro. Si no se especifica, "em.mix" asigna un valor inicial apropiado.

maxits: número máximo de iteraciones a realizar. El algoritmo se detendrá si el parámetro aún no ha convergido luego de las iteraciones especificadas.

showits: si se especifica 'true', informa las iteraciones realizadas por el algoritmo EM de manera tal que el usuario pueda monitorear el progreso del mismo.

eps: criterio de convergencia opcional. El algoritmo se detiene cuando la diferencia máxima relativa en todos los parámetros entre una iteración y la siguiente es menor o igual al valor especificado.

Los resultados obtenidos mediante la aplicación de "em.mix" son las estimaciones máximo verosímiles de las probabilidades de celda, los promedios de celdas y las covariancias.

Si la tabla de datos completamente clasificada presenta celdas con conteos iguales a cero, la estimación máximo verosímil puede no ser única y el algoritmo puede converger a diferentes valores estacionarios dependiendo del valor inicial.

• **getparam.mix(**argumentos**)**

Permite visualizar los parámetros del modelo de locación general.

Argumentos:

s: matriz con datos incompletos creada por la función "prelim.mix".

theta: valor de los parámetros producidos por otra función, por ejemplo por "em.mix".

corr: si es igual a "FALSE" en la salida se obtienen las probabilidades de celda, una matriz de promedios de las celdas y una matriz de covariancias. Si es igual a "TRUE" la salida anterior cambia la matriz de covariancias por un vector con los desvíos estándares y la matriz de correlaciones.

• **imp.mix**

Esta función puede ser utilizada para crear imputaciones bajo el modelo de locación general. Debe instalarse una semilla generadora de números aleatorios mediante la función "rngseed" antes de ser utilizada.

Argumentos:

s: matriz de datos incompletos creada por la función "prelim.mix" .

theta: valor de los parámetros bajo el cual los valores se imputan aleatoriamente.

x: matriz de datos original usada para crear s.

4. RESULTADOS

A partir de la matriz de datos completos y con ambos tipos de variables, se generan pérdidas en las variables P, G y D según un mecanismo completamente al azar (MCAR) en un porcentaje, aproximadamente, del 10%, 15% y 20%, respectivamente. De esta manera en la variable P se genera una tercer categoría (NA).

Mediante la utilización del programa R se reestructura la base y se obtiene la matriz de datos incompleta y su descripción. Se detalla la cantidad de pérdida por variable y la tabla de esquemas de pérdida (Tablas 1 y 2).

Tabla 1. Total de pérdidas por variable.

Variables	F	P	G	D
Frecuencia	0	11	13	17

Tabla 2. Esquemas de pérdida según frecuencia y variable .

Variables / **Frecuencia**	F	P	G	D
23	1	1	1	1
6	1	0	1	1
11	1	1	0	1
10	1	1	1	0
5	1	0	1	0
2	1	1	0	0

Se observa en la tabla 2 que hay 23 individuos completamente observados, 6 con pérdidas sólo en la variable P y así sucesivamente.

Las tablas de contingencia para los individuos completa y parcialmente clasificados para la variable P, presentan para las dos categorías de F y las tres categorías de P, además de los conteos por celda, los promedios ($\bar{x}$) y las cantidades de observaciones (n) para cada variable continua G y D (Tabla 3).

Tabla 3. Individuos completa y parcialmente clasificados según P y F y promedios de celda de las variables G y D.

(a) Individuos completamente clasificados

		P=1		P=2	
		F=1	F=2	F=1	F=2
		Conteo de celda		Conteo de celda	
		9	3	27	7
G	$\bar{x}$	388.11	542.5	587.23	483
	n	9	2	17	5
D	$\bar{x}$	1.016	0.62	0.94	1.10
	n	6	3	21	4

(b) Individuos parcialmente clasificados

		P=NA	
		F=1	F=2
		Conteo de celda	
		9	2
G	$\bar{x}$	585.33	574
	n	9	2
D	$\bar{x}$	0.52	1.20
	n	5	1

Se observa que hay al menos un individuo con una o más variables continuas presentes en cada una de las cuatro celdas de la tabla completamente clasificada y pueden estimarse los ochos promedios de celda.

Las estimaciones máximo verosímiles de los parámetros del modelo de locación general, probabilidades de celda, promedios de celda, desvíos estándares y correlación, son calculados con el programa R, mediante la aplicación del algoritmo EM (Tablas 4 y 5).

Tabla 4. Estimaciones máximo verosímiles de probabilidades de celda y promedios de celda.

Variables		**Probabilidades de celda**	**Promedio de celda**	
F	**P**		**G**	**I**
Si	Si	0.18	396.43	1.05
Si	No	0.61	586.87	0.94
No	Si	0.06	494.35	0.64
No	No	0.15	513.39	0.98

Tabla 5. Estimaciones máximo verosímiles de las desviaciones estándares y la correlación.

Desvíos estándares		**Correlación (G, I)**
G	**I**	
198.52	0.35	0.536

Una rutina tipo en el programa R para la aplicación de la metodología, toma la forma:

```
xperd<-read.table("c:/DATOSper.txt",header=T)
xperd
library(mix)
xperdmat<-as.matrix(xperd)
xperdmat
s<- prelim.mix(xperdmat,2)
s
thetahat<-em.mix(s)
getparam.mix(s,thetahat,corr=TRUE)
rngseed(1234567)
newtheta <- da.mix(s,thetahat,steps=100)
ximp <- imp.mix(s, newtheta, xperd)
ximp
```

Si es de interés tener la base de datos completa para luego aplicarle algún procedimiento estadístico estándar, mediante la última sentencia se obtiene la base de datos imputada según los valores de los parámetros estimados.

5. DISCUSIÓN

• El modelo de locación general, combinado con el algoritmo EM, provee una herramienta para el análisis de bases de datos con información faltante y/o confusa en las variables categóricas y continuas.

• Cuando el tamaño de muestra no es apreciablemente mayor que el número de celdas, se puede considerar el modelo de locación general con restricciones sobre el espacio paramétrico.

• Los procedimientos y funciones incorporados en los softwares de acceso general favorecen su aplicación.

• La generación de pérdida en una etapa experimental, posibilita disponer de la información con las características que requiere la metodología.

6. REFERENCIAS BIBLIOGRÁFICAS

• Little, R. and Rubin, D.. (2002). "*Statistical Analysis with Missing Data*". Second Edition. John Wiley and Sons, New York. • Little, R. and Schluchter, M.. (1985)."Maximum likelihood estimation for mixed continious and categorical data with missing values". Biometrika, 72, 492-512.

• Olkin, I. and Tate, R.. (1961). "Multivariate correlation models with mixed discrete and continuous variables". Annals of Mathematical Statistics, 32, 448-465.

• R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

• Original by Joseph L. Schafer. (2006). mix: Estimation/múltiple Imputation for Mixed Categorical and Continuous Data. R package version 1.0-5. http://www.stat.psu.edu/~ jls/misoftwa.html

• Schafer, J.. (1997). "*Analysis of Incomplete Multivariate Data*". Chapman and Hall, Londres.