Badler, Clara
Alsina, Sara
Beltrán, Celina
Bussi, Javier
Puigsubirá, Cristina
Vitelleschi, Ma. Susana
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

LA ESTIMACIÓN MÁXIMO VEROSÍMIL COMO ALTERNATIVA PARA EL TRATAMIENTO DE LA FALTA DE INFORMACIÓN EN ENCUESTAS

1. INTRODUCCIÓN

Cuando un conjunto de datos presenta pérdidas parciales en algunas variables y se desea estimar el vector de medias y la matriz de covariancias poblacionales, un procedimiento habitual consiste en excluir las observaciones con datos faltantes. En algunos casos, tales como la estimación de la covariancia entre dos variables, generalmente se utiliza la información proveniente de todas las observaciones que no presenten valores perdidos con respecto a estas variables, sin tener en cuenta posibles pérdidas en aquéllas que no participan de dicha estimación en particular. Estos métodos de tratamiento de datos con información perdida, usualmente referidos como Casos Completos y Casos Disponibles respectivamente, se caracterizan por su simplicidad pero descartan la información de las unidades incompletas.

En aquellos casos donde el mecanismo de pérdida de la información es ignorable, una alternativa válida es la de realizar estimaciones máximo verosímiles utilizando la información tanto de las unidades completas como la de las unidades que tienen datos faltantes en alguna de las variables. La estimación de los parámetros se ve facilitada para el caso en que los datos presenten un esquema monótono de pérdida y se verifique el cumplimiento del supuesto de distribución normal multivariada. Dicha estimación puede ser realizada mediante la aplicación del operador matricial “Sweep”.

En este trabajo se aplica dicha metodología a información relacionada con la variable ingreso proveniente de la Encuesta Permanente de Hogares (EPH) para el Aglomerado Gran Rosario. Se comparan las estimaciones obtenidas con las resultantes de aplicar los métodos de Casos Completos y Casos Disponibles.

2. MATERIAL

Las variables en estudio pertenecen a la onda de mayo de 1998 de la EPH, Aglomerado Gran Rosario y corresponden a mujeres del bloque ocupados (n=232):

- Edad (Y_1)
- Total de Horas Trabajadas en La Semana de Referencia (Y_2)
- Años de Escolaridad (Y_3)
- Tiempo que Lleva en su Ocupación Principal, en Meses (Y_4)

- Ingreso Horario de la Ocupación Principal (Y_5)

La inclusión en el análisis de la variable Ingreso Horario de la Ocupación Principal se relaciona con la importancia del ingreso en los análisis económicos y al hecho que las variables utilizadas para su estimación presentan generalmente alta proporción de no respuesta. La elección de las restantes variables responde a las propuestas de diversos autores en cuanto al rol de las mismas en la explicación del comportamiento de los individuos.

Se trabaja además transformando las variables Y_4 e Y_5 de la siguiente manera:

- Raíz cúbica del tiempo que lleva en su Ocupación Principal (Y_4^*)
- Logaritmo natural del Ingreso Horario de la Ocupación Principal (Y^*_5)

3. METODOLOGÍA

3.1 Estimación máximo verosímil en distribuciones normales multivariadas con esquema monótono de pérdida de la información

En situaciones donde existe información faltante, la estimación de los parámetros de una determinada distribución a través del método de máxima verosimilitud puede ser compleja. Esto se debe al hecho que el logaritmo de la función de verosimilitud puede tomar una forma complicada sin un máximo fácilmente obtenible.

Se consideran los casos en donde las pérdidas de las observaciones son al azar, es decir que la probabilidad de que un valor se pierda no depende del valor en sí o a lo sumo depende de los valores que tomen otras variables. Estas formas de pérdida de la información, usualmente referidas como completamente al azar (MCAR) o al azar (MAR) respectivamente avalan el hecho que el mecanismo de pérdida de la información es ignorable y que la función de verosimilitud basada en los valores observados puede ser utilizada para hacer estimaciones.

Existe una alternativa de trabajo que permite la descomposición del logaritmo de dicha función en una suma de términos, hecho que facilita la maximización. En aquellos casos donde es posible hallar una descomposición que cumpla con estas características, se puede hallar la solución maximizando cada uno de los términos que conforman la suma.

Cuando se cuenta con un esquema de pérdida monótono donde se registran J variables, se debe notar que para cada unidad i, si el valor $y_{i,j+1}$ está observado, el valor y_{ij} también lo está; si $y_{i,j+1}$ no está observado, y_{ij} puede estar observado o no, para j=1,...,J-1. De esta manera la variable Y_j está más observada que la variableY_{j+1}.

En forma más general, el esquema monótono se puede definir para J bloques de variables Z_j , donde existen v_j variables por bloque con un mismo número de observaciones. En forma análoga resulta que las v_j variables del bloque Z_j están más observadas que las v_{j+1} variables del bloque Z_{j+1} para j=1,...,J-1. Se representa este esquema en la Figura 1.

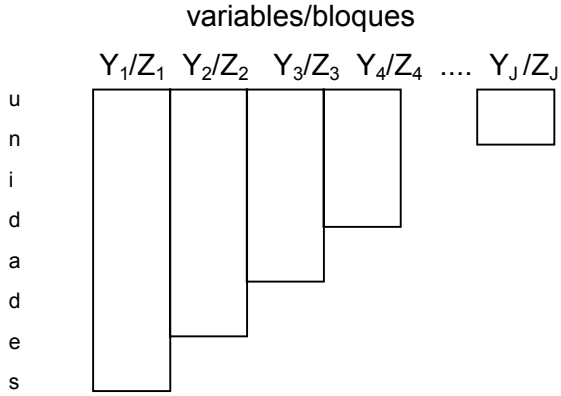

Figura 1. Representación de un esquema de pérdida monótono

La función de verosimilitud para una muestra proveniente de una distribución normal multivariada donde se cuenta con J variables y un esquema monótono de pérdida no provee una solución explícita para la estimación del vector de esperanzas μ y la matriz de covariancias poblacionales $\sum$.

Bajo este esquema y siendo n el tamaño muestral, existen m_J unidades completamente observadas. Consecuentemente existen m_{J-1} unidades con observaciones para todas las variables excepto Y_J , m_{J-2} unidades con observaciones para todas las variables excepto Y_J e Y_{J-1} , obteniéndose la caracterización de las restantes unidades de manera similar. La variable Y_1 se encuentra observada para las $n=m_1$ unidades.

Los estimadores máximo verosímiles de las medias y covariancias de todas las variables pueden ser obtenidos fácilmente utilizando el operador matricial “Sweep” (1).

El operador “Sweep” se define para matrices simétricas. Una matriz G de dimensión $p \times p$ es “barrida” por este operador a través de la fila y la columna k cuando es reemplazada por otra matriz simétrica H de igual dimensión $p \times p$ de la siguiente manera:

$$h_{kk} = -1/g_{kk}\,; h_{jk} = h_{kj} = g_{jk}/g_{kk} \quad k \neq j\,; \; h_{jl} = g_{jl} - g_{jk}g_{kl}/g_{kk} \quad k \neq j, k \neq l$$

La matriz H obtenida a través de esta operación se indica como SWP[k]G. El operador “Sweep” es conmutativo, es decir el orden en el cual un cierto número de barridos es computado no altera el resultado final, es decir: SWP[j,k]G=SWP[k,j]G

Existe además un operador inverso al operador “Sweep” que recibe el nombre de “Sweep Inverso” y se indica H=SWI[k]G, donde:

$$h_{kk} = -1/g_{kk}\,; h_{jk} = h_{kj} = -g_{jk}/g_{kk} \quad k \neq j\,; \; h_{jl} = g_{jl} - g_{jk}g_{kl}/g_{kk} \quad k \neq j, k \neq l$$

Es fácil demostrar que este operador es también conmutativo y que es el operador inverso del “Sweep”.

Las estimaciones máximo verosímiles se pueden obtener mediante el uso combinado de los operadores “Sweep” y “Sweep Inverso”. Se presenta el caso para un esquema monótono con tres bloques de variables. La extensión a más de tres bloques y a más variables por bloque es inmediata. Con el fin de obtener las estimaciones de los parámetros,

se deben encontrar los estimadores máximo verosímiles del vector de medias y de la matriz de covariancias del primer bloque que cuenta con dos variables, las cuales se encuentran completamente observadas. Estas estimaciones son simplemente el vector de medias aritméticas de orden 2×1 ($\hat{\mu}_1$) y la matriz de covariancias de orden 2×2 de Z_1 ($\hat{\Sigma}_{11}$), calculados a partir de la muestra.

A continuación se deben obtener los estimadores máximo verosímiles del intercepto, coeficientes y variancia de los residuos de la regresión de la variable del bloque Z_2 en las dos variables del bloque Z_1. Para ello se utilizan los procedimientos usuales de regresión múltiple obteniéndose:

- $\hat{\beta}_{20.1}$ estimación del intercepto.
- $\hat{\beta}_{21.1}$ un vector de dimensión 2×1 que contiene las estimaciones de los coeficientes de regresión de las variables del bloque Z_1.
- $\hat{\sigma}_{22.1}$ estimación de la variancia de los residuos.

El paso siguiente consiste en encontrar los estimadores máximo verosímiles correspondientes al intercepto, coeficientes de regresión y variancia residual de la regresión de la variable del bloque Z_3 en las tres variables correspondientes a los bloques Z_1 y Z_2. Estas estimaciones se pueden obtener mediante los procedimientos clásicos de regresión múltiple a partir de las unidades donde las cuatro variables pertenecientes a los tres bloques están completamente observadas. Las estimaciones se indican de la siguiente manera, obteniéndose las dimensiones de las matrices en forma análoga a la de la regresión previa: $\hat{\beta}_{30.12}, \hat{\beta}_{31.12}, \hat{\beta}_{32.12}$ y $\hat{\sigma}_{33.12}$.

Una vez obtenidas las estimaciones máximo verosímiles indicadas se debe calcular la matriz **A** que se obtiene de la siguiente manera:

$$A = SWP[\mathbf{1}]\begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

donde SWP[**1**] indica el barrido a través de las dos variables del bloque Z_1. La matriz A resultante es de dimensión 3×3.Luego es necesario calcular la matriz B, la cual resulta:

$$B = SWP[\mathbf{2}]\begin{bmatrix} a_{11} & A_{1.23} & \hat{\beta}_{20.1} \\ A_{1.23}^T & A_{23.23} & \hat{\beta}_{21.1} \\ \hat{\beta}_{20.1} & \hat{\beta}_{21.1}^T & \hat{\sigma}_{22.1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix}$$

donde: $A_{1.23} = \begin{bmatrix} a_{12} & a_{13} \end{bmatrix}$ $\qquad$ $A_{23.23} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$

y SWP[**2**] indica el barrido de la variable del bloque Z_2. La matriz B es de dimensión 4×4. Finalmente, la estimación máximo verosímil de la matriz de covariancias ampliada de las variables de los bloques Z_1, Z_2 y Z_3 está dada por:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = SWI[\mathbf{1},\mathbf{2}] \begin{bmatrix} c_{11} & C_{1.23} & c_{14} & \hat{\beta}_{30.12} \\ C_{1.23}^T & C_{23.23} & C_{23.4} & \hat{\beta}_{31.12} \\ c_{41} & C_{23.4}^T & c_{34} & \hat{\beta}_{32.12} \\ \hat{\beta}_{30.12} & \hat{\beta}_{31.12}^T & \hat{\beta}_{32.12} & \hat{\sigma}_{33.12} \end{bmatrix}$$

donde: $C_{1.23} = \begin{bmatrix} c_{12} & c_{13} \end{bmatrix}$ $\quad C_{23.23} = \begin{bmatrix} c_{22} & c_{23} \\ c_{32} & c_{33} \end{bmatrix}$ $\quad C_{23.4} = \begin{bmatrix} c_{24} \\ c_{34} \end{bmatrix}$

y SWI[**1**,**2**] indica el barrido inverso de las variables de los bloques Z_1 y Z_2. La matriz resultante es de dimensión 5×5, ya que es la matriz ampliada de las estimaciones máximo verosímiles de las medias y covariancias de un conjunto de cuatro variables.

3.2 Estimación a través de casos disponibles y casos completos

Con el fin de de realizar comparaciones con respecto al método de máxima verosimilitud cuando existe información perdida, se consideran las estimaciones obtenidas a partir de casos disponibles y casos completos.

Las estimaciones de μ y Σ a través de casos disponibles se calculan computando las medias aritméticas y las variancias muestrales utilizando todas las observaciones disponibles para cada variable. Para el cálculo de las covariancias, es necesario especificar una extensión natural del método de casos disponibles, que es el método de casos disponibles de a pares. El cálculo de la covariancia de dos variables Y_j e Y_k se basa en las i unidades para las cuales tanto y_{ij} e y_{ik} están observadas, de la siguiente manera:

$$s_{jk}^{(jk)} = \sum_{(jk)} (y_{ij} - \bar{y}_j^{(jk)})(y_{ik} - \bar{y}_k^{(jk)}) / (n^{(jk)} - 1)$$

donde $n^{(jk)}$ es el número de casos donde ambas variables están observadas, y tanto las medias aritméticas como la sumatoria están calculadas en base a esas $n^{(jk)}$ observaciones. Para el cómputo de las variancias resulta que $n^{(jk)} \equiv n^{(jj)} \equiv m_j$.

En la estimación de los parámetros μ y Σ a través de casos completos, tanto las medias como las covariancias se calculan utilizando exclusivamente aquellas unidades en donde todas las variables han sido observadas, es decir utilizando m_J unidades del total de las n unidades iniciales. El cálculo de la covariancia entre dos variables resulta:

$$s_{jk}^{(m_J)} = \sum_{(m_J)} (y_{ij} - \bar{y}_j^{(m_J)})(y_{ik} - \bar{y}_k^{(m_J)}) / (m_J - 1)$$

donde la sumatoria como las medias aritméticas están calculadas en base a las m_J observaciones.

3.3 Medidas de comparación de las estimaciones obtenidas a través de máxima verosimilitud con respecto a casos disponibles y a casos completos

La comparación se realiza en porcentaje de desviaciones estándares estimadas a través del método de máxima verosimilitud. Las medidas A y B comparan máxima verosimilitud con casos completos:

$$A = \frac{(\bar{y}_j^{(m_J)} - \hat{\mu}_j)}{\hat{\sigma}_j}.100 \qquad B = \frac{(s_j^{(m_J)} - \hat{\sigma}_j)}{\hat{\sigma}_j}.100$$

donde $\bar{y}_j^{(m_J)}$ y $\hat{\mu}_j$ son las estimaciones con casos completos y máximo verosímil respectivamente de la media de la variable Y_j; $s_j^{(m_J)}$ y $\hat{\sigma}_j$ son las estimaciones con casos completos y máximo verosímil respectivamente del desvío estándar de la variable Y_j.

Las medidas C y D comparan máxima verosimilitud con casos disponibles:

$$C = \frac{(\bar{y}^{(j)} - \hat{\mu}_j)}{\hat{\sigma}_j}.100 \qquad D = \frac{(s_j^{(j)} - \hat{\sigma}_j)}{\hat{\sigma}_j}.100$$

donde $\bar{y}^{(j)}$ y $s_j^{(j)}$ son las estimaciones con casos disponibles de la media y el desvío estándar respectivamente de la variable Y_j.

Las comparaciones de las estimaciones de las covariancias máximo verosímiles se realizan en porcentaje de covariancias estimadas por máxima verosimilitud:

$$CC = \frac{(s_{jk}^{(m_J)} - \hat{\sigma}_{jk})}{\hat{\sigma}_{jk}}.100 \qquad CD = \frac{(s_{jk}^{(jk)} - \hat{\sigma}_{jk})}{\hat{\sigma}_{jk}}.100$$

donde $s_{jk}^{(m_J)}, \hat{\sigma}_{jk}$ y $s_{jk}^{(jk)}$ son las estimaciones con casos completos, máximo verosímil y con casos disponibles respectivamente de la covariancia entre la variable Y_j y la variable Y_k.

4. RESULTADOS

En primer lugar se analiza el comportamiento de las variables Y_1, Y_2, Y_3, Y_4 e Y_5^* para estudiar el cumplimiento del supuesto de normalidad conjunta de las mismas. Al no verificarse, se busca una trasformación para Y_3 e Y_4. Para Y_4 se halla que la raíz cúbica (Y^*_4) presenta distribución normal marginal, no siendo posible una transformación adecuada para Y_3, excluyéndola. Se continúa el análisis con las variables Y_1, Y_2, Y_4^* e Y_5^*.

Se analiza el esquema de pérdida del nuevo conjunto de variables. Los *g*rupos de valores observados para cada variable se indican con 1 y los no observados con 0, y con

letras los diferentes esquemas de pérdida que se presentan según si las variables están observadas o no (Tabla 1).

Tabla 1: Esquema de pérdida observado

Variables Esquemas	Y_1	Y_2	Y_4^*	Y_5^*	Cant. de unidades
A	1	1	1	1	199
B	1	1	0	1*	2
C	1	1	1	0	27
D	1	1	0	0	4

Al descartar las dos observaciones marcadas con 1*, se obtiene un esquema de pérdida monótono en el cual las variables Y_1 e Y_2 constituyen un bloque de variables completamente observadas y la variable Y_4^* se encuentra más observada que Y_5^*. En la Tabla 2 se muestra la cantidad de unidades observadas para cada variable.

Tabla 2: Cantidad de observaciones por variable en el esquema monótono

Variable	m_j
Y_1	232
Y_2	232
Y_4^*	226
Y_5^*	199

Con el fin de determinar si el mecanismo de pérdida puede ser considerado ignorable, se aplica el test de Little, no rechazándose la hipótesis que los datos faltantes son perdidos completamente al azar (MCAR). Las estimaciones máximo verosímiles se llevan a cabo mediante el operador "Sweep". Su implementación se realiza mediante un programa especialmente desarrollado a tal fin en S-PLUS. Previamente es necesario realizar los siguientes cálculos:

1. Para las variables observadas en forma completa:

   $\hat{\mu}_1$=41,736 $\hat{\sigma}_{11}$ =101,728 ; $\hat{\mu}_2$ =35,028 $\hat{\sigma}_{22}$ =471,322 $\hat{\sigma}_{12}$ =-0,8084

2. De la regresión de Y_4* sobre Y_1 e Y_2 se obtiene:

   $\hat{\beta}_{20.1}$=2.085 $\hat{\beta}_{21.1}^T$ =(0.044688 0.001181) $\hat{\sigma}_{22.1}$=2.2807

3. De la regresión de Y_5^* sobre las restantes variables se obtiene:

   $\hat{\beta}_{30.12}$ =1.6054 $\hat{\beta}_{31.12}^T$ =(-0.006 -0,0223) $\hat{\beta}_{32.12}$ =0.137236 $\hat{\sigma}_{33.12}$ =0.2925

Se presentan las estimaciones para las medias y desvíos obtenidos por máxima verosimilitud y las comparaciones con las estimaciones alternativas (Tabla 3). Se observa que existen diferencias entre casos completos y disponibles con respecto a máxima verosimilitud.

En aquellos casos en los que hay coincidencias en las estimaciones, éstas se deben a que fueron computadas por razones metodológicas a partir del mismo número de observaciones.

Tabla 3: Estimaciones máximo verosímiles de la media y el desvío estándar y comparaciones con estimaciones a partir de casos completos y casos disponibles

Varia	Max. Verosimil		Casos Completos				Casos Disponibles			
-ble	Media	DS	Media	DS	A	B	Media	DS	C	D
Y_1	41.736	10.064	41.058	10.092	-6.737	0.278	41.736	10.086	0.000	0.219
Y_2	35.028	21.710	35.462	20.530	1.999	-5.435	35.028	21.757	0.000	0.216
Y_4^*	3.991	1.576	3.963	1.556	-1.777	-1.269	3.987	1.580	-0.254	0.254
Y^*_5	1.123	0.752	1.113	0.734	-1.330	-2.394	1.113	0.734	-1.330	-2.394

Se presentan las diferencias entre las estimaciones de las covariancias máximo verosímiles y las estimaciones obtenidas a partir de casos completos y casos disponibles (Tabla 4).

Tabla 4: Medidas de comparación de las estimaciones de las covariancias

Covariancia entre	CC	CD
$Y_1 - Y_2$	808.416	1.856
$Y_1 - Y^*_5$	341.935	341.935
$Y_1 - Y_4^*$	-5.657	1.326
$Y_2 - Y^*_5$	-11.361	-11.361
$Y_2 - Y_4^*$	49.328	-35.317
$Y^*_5 - Y_4^*$	-4.305	-4.305

En casos completos se observan, en general, diferencias mayores al 10%. Existen dos situaciones extremas donde la diferencia es mayor al 100%.

Las estimaciones a través de casos disponibles presentan diferencias mayores al 10 % en la mitad de los casos. Estas estimaciones están más cercanas a las máximo verosímiles que las correspondientes a casos completos. Es importante destacar la mayor cercanía de casos disponibles para la covariancia entre Y_1 e Y_2.

Al igual que en casos completos, la covariancia entre Y_1 e Y^*_5 es muy diferente debido a la eliminación de observaciones para lograr monotonía, generando esta diferencia extrema.

5. DISCUSIÓN

Cuando se dispone de un conjunto de datos que presenta pérdidas parciales en algunas variables:

- la estimación máximo verosímil del vector de medias y la matriz de covariancias se ve facilitada cuando el mecanismo de pérdida es ignorable y el esquema es monótono.
- si los datos se distribuyen conjuntamente normal se puede utilizar el operador "sweep" para dichas estimaciones.
- la estimaciones obtenidas por máxima verosimilitud difieren de las obtenidas en casos completos y casos disponibles.

Es preferible aplicar el método de máxima verosimilitud ante la presencia de un esquema monótono de pérdida ya que utiliza mayor cantidad de información que los métodos de casos disponibles y casos completos y su aplicación es relativamente sencilla.

Es de interés en futuros trabajos aplicar el método de máxima verosimilitud a datos con esquemas de pérdidas diferentes y evaluar su eficiencia.

Bibliografía

Dempster, A. P.. (1969). "Elements of Continuous Multivariate Analysis". Reading, MA: Addison-Wesley.

Di Paola, R.; Bergés, M.; Rodríguez, E.(1997)."Diferencias de ingreso entre jefes de familia en la ciudad de Mar del Plata.Un enfoque de la teoría del capital humano"Anales de la Asociación Argentina de Economía Política..

Paz, J. (2000). "En cuánto y por qué difieren las remuneraciones en la Argentina" Anales de la Asociación Argentina de Economía Política.

Griliches, Z. (1997). "Estimating the returns to schooling: some economic problems" Econometrica,45,January,1997.

Little, R. J. (1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values". Journal of the Royal Statistical Society. Vol. 83, N° 404.

Little, R. J. and D. B. Rubin. (1987). "Statistical Analysis with Missing Data". John Wiley & Sons. New York.