Psicothema, 1997. Vol. 9, nº 1, pp. 187-197
ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

ESTIMACIÓN DE DATOS PERDIDOS POR MÁXIMA VEROSIMILITUD EN PATRONES «MISSING» ALEATORIOS (MAR) Y COMPLETAMENTE ALEATORIOS (MCAR) EN MODELOS ESTRUCTURALES

Concepción San Luis Costas, Juan A. Hernández Cabrera y Gustavo Ramírez Santana
Universidad de La Laguna

En las investigaciones del campo aplicado con técnicas multivariadas es muy frecuente encontrar matrices de datos con valores perdidos. Las estrategias más comúnmente utilizadas para reconducir este problema, utilizan los métodos listwise, pairwise y los de estimación de máxima verosimilitud. En este articulo se demuestra mediante las técnicas de simulación de Monte Carlo en el ámbito de los modelos estructurales, que independientemente del patrón de missing simulado (missing completamente aleatorio, monotónico o condicional) la estimación mediante el algoritmo de máxima verosimilitud EM arroja los mejores resultados, en cuanto a la precición de la estimación de los parámetros de los modelos, disminución de los errores típicos, y la posibilidad de encontrar soluciones adecuadas y convergentes en aquellos patrones de missing donde las estrategias MCAR (listwise y pairwise) son imposibles de utilizar.

Maximun likelihood missing values estimation in patterns of missing MAR and MCAR in structurals models. In the research's of the applied field is very common to find matrices of data with lost values. The main strategies used in order to fix this problem, are the methods listwise, pairwise and maximum likelihood estimates. This article shows through Monte Carlo simulation in the field of the structural models, that irrespective of the pattern of missing simulated (missing completely at random, monotonic missing or conditional missing) the estimates through the maximum likelihood algorithm EM throws the better results, concerning the biases in the estimate of the parameters of the models, decrease of the standard errors, and the possibility of finding convergent and adequate solutions in those patterns of missing where the strategies MCAR (listwise and pairwise) are impossible to use.

La inferencia estadística con datos perdidos es un problema muy importante de la investigación aplicada en general y de las investigaciones con modelos estructurales en particular. Básicamente son tres las dificultades fundamentales en el uso de matrices con datos perdidos. En primer lugar, si los casos con "missing" son diferentes a los casos completos, las estrategias comunes de

Correspondencia: Concepción San Luis Costas
Facultad de Psicología
Universidad de La Laguna. Campus de Guajara
Tenerife (Spain)

tratamiento de este problema presentan un importante sesgo. En segundo lugar, la existencia de datos perdidos generalmente implica una importante pérdida de información, por lo que las estimaciones de parámetros pueden ser ineficientes. Finalmente, las técnicas estadísticas disponibles están diseñadas para datos completos, por lo que, la sola presencia de datos perdidos perjudica notablemente el análisis (Roderick, Little & Schenker, 1995).

Antes de exponer los métodos disponibles para la estimación de casos perdidos, se hace necesario describir sucintamente los diferentes patrones de "missing" que pueden encontrarse en la investigación aplicada.

Patrones de missing.

Si una matriz es completa (sin casos perdidos), puede ser definida como una matriz $X=X_{ij}$ de orden *n x p*, de tal forma que X_{ij} es el valor de la variable *j, j=1... p* en el caso i, *i=1 ...n*. Si consideramos a la matriz $M=m_{ij}$ de orden n x p, como una matriz de indicadores de datos perdidos, de tal forma que $m_{ij}= 1$ si x_{ij} es un dato perdido y $m_{ij}=0$ si x_{ij} está presente. La matriz M describe el patrón de missing, y su media marginal de columna, puede ser interpretada como la probabilidad de que x_{ij} sea missing.

La determinación del patrón de missing presente en los datos, es una tarea de gran interés. Responder a preguntas del tipo. ¿Los sujetos que responden a una determinada variable son en realidad diferentes de los que no responden?. La ausencia de respuesta a una determinada variable, ¿es función de otra variable antecedente? (vg: a mayor nivel socioeconómico, menor índice de respuesta en la variable ingresos brutos anuales). En general, si podemos considerar que la matriz generada mediante los procedimientos listwise o pairwise es una matriz aleatoria de la matriz global, diremos que los datos presentan un patrón de missing completamente aleatorio (missing completely at random , MCAR), o lo que es lo mismo, diremos que el patrón de missing no es función de ninguna variable de la investigación. El patrón MCAR se define finalmente según Rubin (1976) como:

$$P(M|X,\phi) = P(M|\phi) \text{ para todo } X$$

Ec. 1

la distribución de missing dado X, depende exclusivamente del parámetro ϕ, que caracteriza a las respuestas. Esta premisa MCAR puede ser evaluada mediante el programa BMDPP8 (Dixon, 1988), que arroja valores t, para cada variable, bajo la hipótesis nula de un patrón univariado de missing completamente aleatorio. Sin embargo, si este patrón depende de otra variable y de ésta se dispone de respuesta tanto para los que responden como para los que no, el sesgo en la estimación de los datos perdidos mediante las estrategias anteriormente comentadas que exigen un patrón MCAR, puede ser controlado mediante un análisis que estratifica o ajusta la variable missing en función de la variable o variables antecedentes correlacionadas con la variable con ausencia de respuesta, de las cuales se dispone de datos para todos los sujetos de la muestra. Este patrón de missing se conoce como valores perdidos aleatorios MAR (missing at random). Se define funcionalmente como:

$$P(M|X_{observada}, X_{missing},\phi) = P(M|X_{observada}) \text{ para todo } X_{missing}$$

Ec. 2

Es decir, la distribución de los casos perdidos, dado X, depende exclusivamente de la variable $X_{observada}$ de la matriz de datos X.

Seguidamente haremos un breve recorrido por las estrategias mas comunes, utilizadas por los investigadores del campo aplica-

do en el tratamiento de matrices de datos con valores perdidos.

Análisis de casos completos

En este tipo de análisis, el investigador simplemente elimina aquellos casos que presentan datos perdidos en las variables que vayan a ser utilizadas. Es el sistema estandarizado en la mayoría de los paquetes estadísticos comerciales, y se conoce con el nombre de "listwise". Presenta como ventaja fundamental su facilidad de implementación, así como la obtención de estimadores válidos, siempre que el patrón de missing sea completamente aleatorio (MCAR). En otras palabras, si podemos considerar la muestra de datos incompletos como una muestra aleatoria de la muestra global (sin datos perdidos) (Rubin, 1976). Sin embargo, esta estrategia de análisis pierde una importante cantidad de información directamente proporcional al número de "missing". La solución a este problema de pérdida de información suele radicar en eliminar aquellas variables con un mayor porcentaje de valores missing. En cualquier caso, si el patrón de missing no es MCAR, el tamaño del sesgo depende, entre otros aspectos, del grado de asociación entre la variable missing y otras variables de la investigación, de la cantidad de datos perdidos así como de las características intrínsecas del análisis que se esté llevando a cabo.

Análisis de las respuestas disponibles

El método conocido como "pairwise", constituye otra estrategia muy utilizada. Esta metodología de análisis forma una matriz de varianzas y covarianzas utilizando para ello todos los datos disponibles . De esta forma, los elementos de la matriz de momentos resultante surgen, como es obvio, de diferentes tamaños muestrales, lo cual confiere a esta estrategia un inconveniente fundamental derivado del hecho de que la matriz así obtenida, es frecuentemente no positiva definida, lo que la invalida para ser usada en técnicas estadísticas que requieran la inversión de la matriz de momentos.

Sustitución de los valores perdidos por el valor medio de la variable

Otra estrategia muy común en presencia de matrices de datos, donde la metodología listwise conduce a matrices de varianzas y covarianzas con muy pocos sujetos, consiste en sustituir el valor perdido por la media de la variable que corresponda. Sin embargo, este método presenta más inconvenientes que ventajas, dado que se produce una disminución artificial de la varianza de la variable que se ha imputado, sesgándose por tanto las asociaciones entre las mismas, dando lugar a estimaciones erróneas (Browne, 1982, 1984).

Estimación de Máxima verosimilitud con datos perdidos

La distribución normal multivariada es una premisa básica en la mayoría de las técnicas estadísticas multivariadas, y especialmente en todas aquellas que realizan la estimación de los parámetros de los modelos mediante máxima verosimilitud. Este método de estimación en el entorno de datos perdidos, requiere la especificación de un modelo de la distribución de X y M:

$$P(X, M|\theta, \psi) = P(X|\theta)\ P(M|X, \psi)$$

Ec. 3

donde $P(X,\theta)$ representa el modelo de la distribución de la matriz de datos X en ausencia de datos perdidos, $P(M|X, \psi)$ el modelo para los datos perdidos y θ y ψ son parámetros desconocidos. El interés de la estimación, se centra generalmente en la esti-

mación de los parámetros del vector θ, considerando a los parámetros del patrón de missing (ψ) como ignorables. En este sentido, se asume que las filas de X siguen una distribución normal, con media μ y matriz de varianzas y covarianzas $\Sigma(\theta(\mu,\Sigma))$. Las estimaciones de máxima verosimilitud de θ son los valores que maximizan la Ec. 3. Dado que el patrón de missing se asume MAR y, por tanto, ignorable sólo el término P(X,θ) de la Ec. 3 contribuye a la estimación ML de θ. Consiguientemente esta estimación es realizada sin incluir el modelo que explica el mecanismo subyacente a los datos perdidos. En este sentido, la probabilidad al ignorar el mecanismo missing es la probabilidad de θ en función de la densidad marginal de $X_{observada}$, ignorando la contribución de M al modelo. Rubin (1976) indica que el mecanismo de los datos perdidos es ignorable sí:

a.-) θ y ψ son parámetros distintos, o sea no funcionalmente relacionados.

b.-) El patrón de missing es MAR; es decir, el método de estimación ML con mecanismo missing ignorable, hace depender la estimación de los datos perdidos de las puntuaciones observadas de X.

Esta es una de las particularidades más interesantes de la estimación ML, ya que supera con creces a la estimación basada en el patrón MCAR. El mismo autor, declara que el método ML ignorable es preferible en todos los casos al resto de los métodos presentados y en muchas ocasiones al método ML con patrón de missing definido y por tanto no ignorable ya que: a) la especificación de un modelo adecuado al mecanismo missing presente en los datos, es frecuentemente una tarea imposible. b) Aún cuando a ciencia cierta el mecanismo missing sea no ignorable, el método ML ignorable, puede ser superior a un mecanismo no ignorable mal especificado.

La estimación ML con patrón de missing ignorable más frecuentemente utilizada es el algoritmo EM (Expected-Maximization) (Dempsted, Laird & Rubin, 1977) que maximiza la siguiente función de probabilidad para estimar la matriz de varianzas y covarianzas así como el vector de medias a partir de matrices de datos incompletas.

$$L(\theta \mid X_{obs}) = \int p(X_{obs}, X_{miss} \mid \theta)\, dX_{miss}$$

Ec. 4

Sea

$$P(\theta \mid X_{observada}, X_{missing})$$

la probabilidad de (basada en los datos completos $X=(X_{observada}, X_{missing})$. Así en el método de máxima verosimilitud EM, θ^t es la estimación de θ en la iteración t del algoritmo. La iteración t+1 consiste en un primer paso de esperanza (Expected) y otro de maximización (Maximization). El paso E toma la esperanza de

$$P(\theta \mid X_{observada}, X_{missing})$$

en función de la distribución condicional de $X_{missing}$ dado $X_{observada}$, evaluada en $\theta = \theta^t$. En la práctica el paso E puede ser considerado como un procedimiento de predicción de datos perdidos por el método de regresión iterativa. De hecho, este paso predice los valores perdidos a través de la regresión de las variables missing sobre las variables observadas para cada sujeto de la muestra, con coeficientes β basados en la estimación de esos parámetros en la iteración t. El paso M estima la matriz de varianzas y covarianzas así como el vector de medias, a partir del relleno de los datos missing realizados en el paso E anterior, es decir maximizando el logaritmo de la función (Orchad & Woodbury, 1972, Little & Rubin, 1987, Dixon, 1988, Schoemberg, 1988). Este méto-

do asume una distribución normal multivariada de las variables implicadas. Si ésta no fuese una premisa realista por la naturaleza no normal de los datos, Little y Smith (1987) describen una variación del método EM, denominada ER que utiliza la distancia de Mahalanobis para ponderar a la baja la influencia de los valores extremos en la estimación. Esta variación del algoritmo EM es útil cuando EM no encuentra convergencia.

Una vez que se ha estimado el vector de medias y la matriz de varianzas y covarianzas mediante ML, es posible "imputar" los datos perdidos para cada caso utilizando el valor esperado de las observaciones dada la matriz de varianzas y covarianzas y el vector de medias ML. La técnica de imputación, es similar a la generación de puntuaciones factoriales del análisis de componentes principales o ejes principales. Este método, sin embargo, no va a generar una matriz de datos completa con varianzas y covarianzas idénticas a la estimada. Es exactamente el mismo problema que se encuentra cuando se computan las puntuaciones factoriales, dado que la matriz de covarianzas de las puntuaciones factoriales puede no ser la misma que la matriz teórica de los auténticos factores. La solución evidente a este problema, se encuentra en solicitar múltiples imputaciones de los datos. En este sentido Rubin y Schenker (1986, 1987) encuentran que un número de imputaciones igual a 3, es para la mayoría de las ocasiones el mejor, dado que conduce con una mayor probabilidad a los valores reales de los datos perdidos. En cualquier caso, de llevarse a cabo la triple imputación de los datos perdidos, la matriz de datos aparecerá triplicada para cada caso. Su análisis posterior con cualquiera de las técnicas estadísticas disponibles, requerirá la ponderación de cada caso por 1/3. Una vez realizada esta ponderación la matriz de datos puede ser analizada como una matriz completa normal, aunque los errores típicos estimados en cualquiera de las técnicas habrán de ser multiplicados por la raíz cuadrada del número de imputaciones realizadas para obtener así el auténtico valor del error típico estimado en cada caso.

La existencia de datos perdidos es, tal y como hemos indicado, un problema frecuente en la investigación aplicada. En este trabajo, pretendemos evaluar mediante simulación de Monte Carlo, la eficacia de las distintas estrategias examinadas para reconducir el problema de los datos perdidos, y específicamente en el ámbito de los modelos de estructura de covarianza. Esta es una técnica estadística muy difundida, donde es muy frecuente el uso de matrices de varianzas y covarianzas listwise como input, aún cuando el patrón de missing no sea MCAR, lo cual ocurre la mayor parte de las ocasiones.

En el ámbito de los modelos estructurales, se han propuesto otras técnicas para solucionar el problema de los missing, la primera de ellas lleva a cabo la estimación simultánea del modelo a partir de dos grupos (Baker & Fulker, 1983; Allison 1987), el primero de ellos contiene la matriz de varianzas y covarianzas y el vector de medias de los datos sin missing, mientras que el segundo contiene las mismas matrices para los datos con missing, con ceros en los parámetros relativos a las variables con datos perdidos. El problema fundamental de esta estrategia es doble, por un lado, si existen muchas variables con missing será necesario reparametrizar el modelo adecuadamente, lo cual no es una tarea fácil y, por otro, será necesaria una buena aproximación a los parámetros de comienzo para evitar así los problemas de convergencia y de soluciones inapropiadas por estimación de varianzas negativas. La otra estrategia consiste en incorporar los valores perdidos a la función de discrepancia a minimizar, así como al cómputo del vector de gradientes y matriz de segundas derivadas parciales (Lee, 1986).

Desgraciadamente, tal incorporación no está actualmente disponible en los paquetes comerciales y exige un trabajo tedioso y complicado por parte del analista de datos.

Método

La presente investigación se realizó a partir de un modelo estructural de 11 variables observables y 5 latentes (3 exógenas y 2 endógenas). En este modelo de la Figura 1 existen 33 parámetros a estimar. Por tanto, es un modelo con (11*(11+1)/2)-33=33 grados de libertad. A partir de la matriz de varianzas y covarianzas poblacional correspondiente al modelo de la figura 1, se generaron 500 muestras de tamaño 300 en 11 variables utilizando el algoritmo de Fleishman (1978) y Vale and Maurelli (1983) según un programa GAUSS (Hernández, J., San Luis, C. & Sánchez Bruno, 1995). A cada una de estas muestras se le aplicó 3 patrones de missing distintos, con un 20% de datos perdidos en cada uno de ellos.

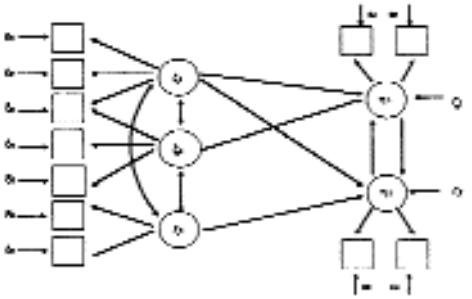

Figura 1. Modelo estructural de 11 variables observables, 5 factores y 33 parámetros a estimar.

En la Figura 2, pueden verse los distintos patrones aplicados. El primer patrón es completamente aleatorio (MCAR), el segundo es monotónico creciente, es decir a medida que aumentamos el número de la variable observable, disminuye el número de missing por variable. El tercer patrón, de missing condicional, hace inviable la estimación de matrices de varianzas y covarianzas por los métodos listwise y pairwise, dado que si para los sujetos i=1 hasta 10 hay missing en la primera variable, para esos mismos sujetos en las siguientes variables los casos están completos. Una vez aplicadas las tres "máscaras missing" a cada una de las muestras, éstas eran analizadas una a una mediante un paquete de modelos estructurales creado a tal efecto en lenguaje GAUSS (Aptech Systems, 1995) (Hernández, J. Ramírez, G. & Sánchez, A, 1995), primero de forma completa (muestra completa) y luego cada una de las "nuevas" muestras con el patrón de missing simulado, utilizando como matriz de momentos de entrada la matriz de los datos sin missing, la matriz listwise y la matriz de varianzas y covarianzas estimada según el algoritmo EM implementado en el módulo MISS del paquete GAUSS (Schoenberg, 1988).

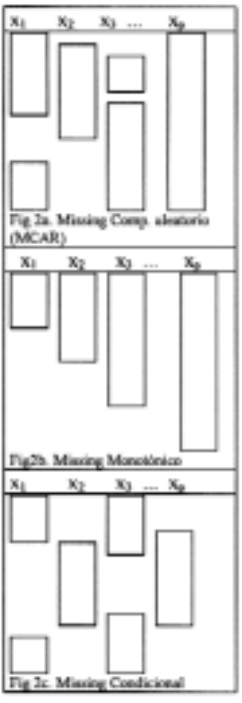

Figura 2. Patrones de Missing simulados.

Resultados

En la Tabla 1 se presentan el valor medio y desviación típica del mínimo de la función de discrepancia (cuyo producto por el tamaño de la muestra da lugar al estadístico χ^2) el estadístico χ^2, los índices de ajuste GFI y AGFI (para la estimación ML) y el error cuadrático medio para cada una de las condiciones simuladas. En esta tabla puede verse en primer lugar, que el valor más pequeño de la función de discrepancia, se sitúa como cabría esperar, en la estimación de la matriz de datos sin missing. Sin embargo, este valor ha sido incluido por un interés meramente comparativo con el mínimo de las distintas estrategias utilizadas para solucionar el problema de los datos perdidos en los tres patrones de missing simulados (completamente aleatorio, monotónico y condicional). En este sentido, puede verse que el mínimo de la estrategia listwise, es siempre mayor que el encontrado en la estimación ML independientemente del patrón de missing simulado. Por otra parte, puede observarse que, como cabría esperar, no existe solución listwise alguna para el patrón de missing condicional, consiguiéndose sin embargo el 100% de las soluciones a través de las 500 muestras al utilizar la estimación ML. En la segunda línea de esta tabla, se encuentra el error cuadrático medio para cada una de las condiciones. Nuevamente, el valor más bajo se sitúa en la estimación sin missing, si comparamos este valor con el resto de los errores, vemos que en todos los casos la estimación mediante matrices listwise genera un error considerablemente mayor que el encontrado en la estimación de máxima verosimilitud, los cuales se encuentran muy próximos tanto al valor medio como a la desviación típica de la estimación sin missing.

Con respecto al estadístico χ^2, vemos que se encuentra muy próximo al valor esperado de 33 para la media y 8 de desviación típica para la estimación sin missing. Sin embargo, la estimación a partir de una matriz listwise, genera valores superiores al esperado tanto para la media como para la desviación típica tanto en el patrón MCAR como para el patrón monotónico. Aunque hay que considerar que la estrategia listwise, para ambos patrones, genera matrices de varianzas y covarianzas a partir de 86 y 109 casos completos, respectivamente. Con respecto a la estrategia ML, si consideramos al número de sujetos de la muestra como de 300 (sin missing), evidentemente arrojaría valores del estadístico χ^2 muy superiores al esperado a pesar de presentar un error cuadrático medio y un mínimo de la función de discrepancia mucho menores que los valores encontrados en la estimación a partir de las matrices listwise. Por este motivo, se ha escogido como indicadores de ajuste mas adecuado, los índices GFI y AGFI independientes del tamaño muestral. Estos indicadores, evidencian valores medios de ajuste óptimos con una gran estabilidad como se evidencia en la escasa desviación típica de los mismos, independientemente del patrón de missing investigado, en clara concordancia con los errores cuadráticos medios encontrados.

Tabla 1
Media y desviación típica del mínimo de la función de discrepancia, error cuadrático medio en la estimación de los 33 parámetros del modelo estructural χ^2, GFI y AGFI

		Patrón de Missing				
		MCAR		Missing Monotónico		Missing C.
	Sin Missing	Listwise	ML	Listwise	ML	ML
Mínimo F.D.						
$\bar{x}$	.1142	.4106	.1939	.3264	.2402	.3882
σ	.0277	.0961	.0454	.0840	.0586	.1101
Error						
$\bar{x}$	.0140	.326	.0178	.289	.0215	.0201
σ	.0042	.0389	.0073	.0393	.0108	.0069
x^2						
$\bar{x}$	34.28	35.31		35.58		
σ	8.333	8.270		9.159		
GFI						
$\bar{x}$			.9917		.9899	.9837
σ			.0022		.0024	.0046
AGFI						
$\bar{x}$			0.9917		.9899	.9837
σ			.0022		.0024	.0046

En la Tabla 2, se presentan los valores medios de los 33 parámetros estimados para cada una de las condiciones simuladas. En general, si los comparamos con los encontrados para la estimación sin missing, vemos que presentan valores muy próximos a los de referencia. Aunque como es obvio, en el patrón condicional sólo encontramos los referentes a las soluciones provenientes de la estimación de la matriz de varianzas y covarianzas mediante ML.

Tabla 2
Media del parámetro estimado a través de los patrones de missing simulados

		Patrón de Missing				
		MCAR		Missing Monotónico		Missing C.
	Sin Missing	Listwise	ML	Listwise	ML	ML
λ_{21}	1.00619	1.00450	1.00692	1.00612	1.00367	1.01506
λ_{42}	.99636	.99522	.99630	.99677	.99594	.98574
λ_{21}	1.09257	1.09965	1.08949	1.09623	1.09579	1.09878
λ_{31}	.51458	.51484	.50956	.51200	.51830	.51726
λ_{32}	.70919	.71157	.71162	.71476	.71144	.71001
λ_{52}	.99032	.98456	.98695	.98781	.98642	.98408
λ_{73}	1.06118	1.06262	1.06284	1.06789	1.06609	1.06513
β_{21}	.60822	.60524	.60463	.59963	.61133	.59735
β_{12}	.23785	.24108	.23966	.24151	.24037	.23689
γ_{11}	.25048	.25160	.24634	.24457	.25451	.24992
γ_{21}	.48259	.36274	.44031	.43494	.44044	.47381
γ_{12}	−.48945	−.44988	−.47169	−.48042	−.47903	−.48694
γ_{23}	.64658	.70355	.66452	.66817	.67229	.65491
ϕ_{11}	.13148	.13121	.13313	.13252	.12806	.05747
ϕ_{12}	.08244	.11007	.09353	.09430	.09134	.11328
ϕ_{13}	.12970	.47478	.15774	.18054	.17141	.12762
ϕ_{22}	.75514	.73346	.73202	.72471	.72193	.70464
ϕ_{32}	.49331	.50342	.49270	.49299	.48737	.47605
ϕ_{33}	.82027	.84439	.81364	.81908	.81006	.80247
ψ_{11}	.39663	.40381	.39547	.39210	.39280	.38527
ψ_{21}	.19214	.19636	.19714	.19002	.19179	.18729
ψ_{22}	.84098	.84781	.83383	.83027	.82915	.81938
δ_1	.07203	.07101	.07081	.07021	.06895	.15479
δ_2	.06039	.05980	.05939	.06026	.06023	.13512
δ_3	.02213	.02236	.02192	.02228	.02255	.01166
δ_4	.02906	.02947	.02930	.02838	.02844	.03893
δ_5	.27034	.26486	.27003	.27264	.27207	.26666
δ_6	.13482	.13026	.13425	.12971	.13472	.13174
δ_7	.04043	.03887	.03943	.03913	.04044	.03999
ε_1	.18551	.17970	.18353	.18434	.18303	.18042
ε_2	.20564	.20602	.20610	.20583	.20518	.20015
ε_3	.16384	.16074	.16349	.16650	.16523	.16266
ε_4	.05723	.05763	.05284	.05579	.05709	.06087

En la Tabla 3, vemos los errores típicos empíricos (desviación típica cada parámetro estimado en las 500 resplicaciones) y los errores típicos estimados (media de los errores típicos estimados para cada parámetro) a través de los distintos patrones de missing investigados. Si observamos los errores típicos empíricos y estimados para las muestras sin missing, encontramos que ambos son bajos y coinciden (los errores típicos están correctamente estimados). Sin embargo, si los comparamos con los errores típicos de la estimación a partir de listwise en el patrón aleatorio y monotónico vemos que, aunque tanto los errores típicos como los estimados coinciden, éstos son considerablemente mayores conduciendo, por tanto, a valores t de significación de cada parámetro menores a los esperados.

Con respecto a la estimación de máxima verosimilitud, vemos que todos los errores típicos empíricos son claramente menores a los obtenidos con listwise, lo cual concuerda con una estimación mas certera de los parámetros del modelo. Sin embargo, si observamos los errores típicos estimados, vemos que en general éstos son infraestimados, conduciendo a valores t de significación superiores a los que correspondería. Dicha situación, evidentemente, se debe al hecho de que la estimación de los errores típicos en un modelo estructural se lleva a cabo a partir del producto del inverso del tamaño muestral declarado por la raíz cuadrada de los elementos de la diagonal de la matriz hessiana (segundas derivadas parciales) en el mínimo de la función de discrepancia

$$\left(ET = \frac{1}{N}\sqrt{\text{diag}(H^{-1})} \right).$$

Para todas las estimaciones de ML se ha incluido como tamaño muestral el de la muestra sin missing (N=300). Con la intención de comprobar esta hipótesis, repetimos

nuevamente el análisis para la estimación de máxima verosimilitud, en los tres patrones missing incluyendo como tamaño muestral N-20% de los casos (porcentaje de missing de la muestra). En la Tabla 4, puede verse que existe una mayor coincidencia entre los errores típicos estimados y empíricos, lo cual demuestra lo acertado de la solución de disminuir el valor del tamaño muestral eliminando del mismo el porcentaje de missing de la muestra.

Conclusiones

A la luz de la claramente mayor eficacia de la estimación de máxima verosimilitud de las matrices de varianzas y covarianzas (utilizadas en todos las técnicas estadísticas multivariadas), la conclusión obvia de esta investigación recae en el hecho de recomendar la utilización de esta técnica para estimar la matriz de momentos siempre que el investigador se encuentre ante matrices de

Tabla 3
Desviación típica del parámetro estimado y media de error típico estimado a través de los tres patrones de missing y de las tres estimaciones de la matriz de varianzas y covarianzas

			Patrón de Missing									
			MCAR				**Missing Monotónico**				**M. Condicional**	
	Sin Missing		**Listwise**		**ML**		**Listwise**		**ML**		**ML**	
	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$
λ_{21}	.0239	.0226	.0394	.0418	.0276	.0226	.0391	.0378	.0381	.0227	.0579	.0341
λ_{42}	.0127	.0136	.0244	.0258	.0151	.0136	.0224	.0228	.0199	.0137	.0227	.0135
λ_{21}	.0486	.0497	.1012	.0926	.0608	.0495	.0838	.829	.0617	.0497	.0633	.0497
λ_{31}	.0382	.0381	.0735	.0715	.0493	.0379	.0672	.0642	0.479	.0381	0.482	.0381
λ_{32}	.0383	.0373	.0768	.0692	.0498	.0370	.0654	.0626	.0458	.0373	.0465	.0374
λ_{52}	.0434	.0419	.0785	.0771	.0536	.0418	.0681	.0695	.0459	.0418	.0490	.0422
λ_{73}	.0375	.0365	.0756	.0683	.0483	.0364	.0655	.0620	.0446	.0370	.0664	.0363
β_{21}	.0386	.0368	.0761	.0709	.0421	.0368	.0617	.0630	.0533	.0374	.0499	.0342
β_{12}	.0445	.0439	.093	.0838	.0497	.0437	.0710	.0746	.0601	.0440	.0547	.0398
γ_{11}	.0407	.0403	0.745	0.075	.0446	.0401	.0685	.0682	.0577	.0404	.0504	.0367
γ_{21}	.1129	.1238	.7704	.8270	.1433	.1262	.7419	.6649	.2654	.1666	.2103	.1462
γ_{12}	.0472	.0495	.2033	.0415	.0560	.0503	.2144	.2040	.0872	.0585	.0681	.0544
λ_{23}	.0636	.0708	.4186	.4752	.0777	.0719	.4092	.3665	.1333	.0896	.1081	.0802
ϕ_{21}	.0206	.0208	.0410	.0395	.0228	.0207	.0361	.0354	.0312	.0211	.0255	.0182
ϕ_{12}	.0324	.0357	.1747	.1979	.0397	.0362	.1799	.1677	.0714	.0463	.0507	.0318
ϕ_{13}	.0589	.0639	.5267	.7551	.0789	.0677	.1331	.3779	.2624	.1345	.2884	.1232
ϕ_{22}	.0774	.0798	.1463	.1491	.0847	.0794	.1334	.1325	.0907	.0798	.0840	.0780
λ_{32}	.0590	.0601	.1132	.1137	.0614	.0599	.1031	.1000	.0644	.0600	.0589	.0582
λ_{33}	.0780	.0821	.1528	.1562	.0820	.0819	.1369	.1368	.0849	.0819	.0893	.0795
ψ_{11}	.0529	.0557	.1018	.1047	.0569	.0555	.0951	.0918	.0605	.0553	.0613	.0544
ψ_{21}	.0483	.0525	.0942	.0995	.0510	.0524	.0856	.0866	.0546	.0521	.0538	.0510
λ_{22}	.0841	.0824	.1522	.1544	.0883	.0822	.1436	.1356	.0824	.0817	.1054	.0803
δ_1	.0094	.0089	.0173	.0163	.0109	.0088	.0164	.0145	.0161	.0087	.0261	.0145
δ_2	.0087	.0084	.0153	.0153	.0102	.0102	.0083	.0136	.0137	.0082	.0151	.0134
δ_3	.0044	.0043	.0086	.0080	.0059	.0043	.0082	.0071	.0077	.0043	.0056	.0034
δ_4	.0046	.0045	.0083	.0084	.0057	.0045	.0084	.00744	.0078	.0045	.0074	.0044
δ_5	.0263	.0272	.0502	.0495	.0327	.0269	.0478	.0448	.0352	.0272	.0356	.0265
δ_6	.0203	.0207	.0433	.0376	.0277	.0203	.0352	.0334	.0272	.0202	.0317	.0194
δ_7	.0087	.0091	.0170	.0167	.0125	.0090	.0161	.0150	.0124	.0090	.0131	.0085
ε_1	.0220	.0216	.0374	.0394	.0272	.0214	.0341	.0352	.0242	.0214	.0274	.0210
ε_2	.0215	.0226	.0406	.0417	.0273	.0224	.0399	.0372	.0271	.0224	.0258	.0219
ε_3	.0212	.0208	.0416	.0381	.0252	.0206	.0367	.0343	.0281	.0208	.0385	.0196
ε_4	.0193	.0185	.0335	.0340	.0233	.0184	.0332	.0307	.0279	.0185	.0357	.0171

datos con valores perdidos independientemente de que el patrón sea MCAR o MAR. Tal recomendación se sustenta en el hecho de que aunque la estrategia listwise es suficientemente eficiente en lo que a la estimación de los parámetros se refiere, en patrones missing completamente aleatorios y monotónicos, no lo es tanto en el estadístico de ajuste y en los errores típicos que son claramente más elevados que los de la muestra sin missing, lo que conducirá frecuentemente a la eliminación de parámetros "aparentemente no significativos" del modelo investigado. Por otra parte, el número de soluciones convergentes y adecuadas con esta estrategia es claramente menor al conseguido con la estimación ML. Cuando el patrón de missing es MAR o el número de casos perdidos muy elevado, puede producirse un sesgo en la estimación de los parámetros ya que la matriz muestral listwise no es una muestra aleatoria de la matriz de datos sin missing , o la imposibilidad de estimar el modelo dado que la matriz listwise contiene muy pocos casos. Tal y como hemos podido comprobar, en todas las ocasiones la estimación de máxima verosimilitud fue claramente superior a la realizada a partir de la matriz listwise, y esta estrategia fue imposible de utilizar cuando el patrón de missing era condicional. Hay que indicar, sin embargo, que la estimación ML en este patrón, aunque exitosa en las 500 muestras utilizadas, requirió de un número muy elevado de iteraciones (aproximadamente 200), dado que se utilizó como matriz de comienzo para iterar una matriz identidad de orden p x p (11 x 11).

En el caso de que se necesite disponer de los valores perdidos, y no solamente del vector de medias y de la matriz de varianzas y covarianzas, puede realizarse la triple imputación de los datos perdidos, una vez estimadas las matrices de momentos anteriores por ML, realizando posteriormente la ponderación de los casos por 1/3 para poder llevar a cabo de esta forma los análisis multivariados clásicos con normalidad.

Tabla 4
Desviación típica y media del parámetro estimado y del error típico estimado respectivamente, para la estimación de la matriz de varianzas y covarianzas mediante ML, considerando el tamaño muestral como N-20%

	Patrón de Missing con estimación ML					
	MCAR		Missing Monotónico		Missing C.	
	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$
λ_{21}	.02666	.02954	.03830	.02923	.05877	.04538
λ_{42}	.01668	.01804	.02139	.01792	.02439	.01777
λ_{21}	.05941	.06482	.06718	.06565	.06141	.06549
λ_{31}	.04453	.04959	.04616	.05042	.04757	.04997
λ_{32}	.04910	.04898	.04819	.04929	.04646	.04855
λ_{52}	.04935	.05505	.04765	.05510	.04934	.05474
λ_{73}	.04624	.04818	.04462	.04834	.07035	.04812
β_{21}	.04099	.04877	.05656	.04782	.05326	.04467
β_{12}	.05229	.05744	.07063	.04724	.05287	.05183
γ_{11}	.04865	.05313	.06423	.05271	.05173	.04744
γ_{21}	.16950	.19152	.24200	.20791	.17187	.17063
γ_{12}	.06760	.07180	.08237	.07898	.06451	.06619
γ_{23}	.09027	.10698	.13444	.11700	.09456	.09545
ϕ_{11}	.02337	.02752	.03500	.02666	.02627	.02354
ϕ_{12}	.05067	.05497	.07324	.05877	.03671	.03602
ϕ_{13}	.12367	.12223	.25893	.17836	.10947	.10408
ϕ_{22}	.09051	.10518	.09424	.10420	.08587	.10171
ϕ_{32}	.06929	.07864	.06711	.07798	.06362	.07646
ϕ_{33}	.09224	.10664	.08646	.10621	.08714	.10509
ψ_{11}	.06237	.07292	.06018	.07236	.05987	.07096
ψ_{21}	.05440	.06833	.05630	.06798	.05803	.06705
ψ_{22}	.08749	.10715	.07921	.10677	.09978	.10520
δ_1	.01085	.01154	.01470	.01131	.02667	.01946
δ_2	.00997	.01087	.01302	.01076	.01603	.01787
δ_3	.00597	.00566	.00674	.00562	.00501	.00450
δ_4	.00631	.00599	.00710	.00588	.00789	.00591
δ_5	.03542	.03571	.03490	.03575	.03067	.03469
δ_6	.02686	.02710	.02793	.02702	.02998	.02586
δ_7	.01349	.01189	.01156	.01194	.01329	.01130
ε_1	.02632	.02809	.02504	.02803	.02652	.02755
ε_2	.02797	.02951	.02538	.02940	.02522	.02868
ε_3	.02753	.02717	.02628	.02711	.04028	.02605
ε_4	.022472	.02429	.0207	.02415	.03974	.02296

Referencias

Allison, P.D. (1987). *Estimation of linear models with incomplete data*. In C.C. Clogg, ed., Sociological Methodology, 1987. Washington, D.C.: American Sociological Association, (pp. 71-103).

Aptech Systems, Inc (1995). Gauss. The Gauss System Version 3.2. Washington.

Baker, L.A. and Fulker, D.W. (1983). Incomplete covariance matrices and LISREL. *Data Lanalyst, 1*, 3-5.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 7, 62-83.*

Dempsted, A.P, Laird, N.M. and Rubin, D.B. (1977). Maximun likelihood from incomplete data via the EM algorithm. *Journal of the American Statistical Association, 81*, 29-41

Dixon, W.J., ed. (1988). BMDP Statistical Software, Los Angeles: University of California Press.

Fleishman, A.(1978). A method for simulating non-normal distributions. *Psychometrika, 43, 4, 521-531.*

Hernández, J.; San Luis, C. y Sanchez, J. (1995). Un programa GAUSS para simular distribuciones no normales multivariadas. *Psicothema, 7, 427-434.*

Hernández, J. Ramírez, G. & Sánchez, A, (1995). A High-level language program to obtain the Bootstrap corrected Adf test statistic. *Behavior Research Methods Instruments, & Computer*. (En prensa).

Lee, S.Y, (1986). Estimation for structural equation models with missing data. *Psychometrika, 51*, 93-99.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.

Little, R.J.A. and Schenker, N. (1995). *Missing Data. Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp 39-75), New York: Arminger, Clifford, Clogg and Sobel. Plenum Press.

Little, R.J.A. and Smith, P.J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association, 82*, 58-68.

Orchad, T. and Woodbury, M.A.(1972). A missing information principle: theory and applications, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 697-715.

Rubin, (1976). Inference and missing data. *Biometrika, 70*, 41-55.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366-374.

Rubin, D.B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics, 3*, 375-387.

Schoenberg, R. (1988), MISS: A Program for Missing Data, in GAUSS Programming Language, Aptech Systems Inc., P.O. Box 6487, Kent, WA 98064.

Vale, D., & Maurelli, V. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48, 3, 465-471.*

Aceptado el 3 de mayo de 1996