ESTADÍSTICA ESPAÑOLA
Vol. 48, Núm. 162, 2006, págs. 241 a 270

Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones

por
JUAN GÓMEZ GARCÍA
Departamento de Métodos Cuantitativos para la Economía. Universidad de Murcia
JAVIER PALAREA ALBALADEJO
Departamento de Informática de Sistemas Universidad Católica. San Antonio
JOSEP ANTONI MARTÍN FERNÁNDEZ
Departament d'Informática i Matemática Aplicada. Universitat de Girona

RESUMEN

En la práctica estadística es frecuente encontrar muestras con datos que no han podido observarse. En este artículo se comparan mediante un ejercicio de simulación el rendimiento y las propiedades de distintas estrategias de inferencia a partir de muestras con datos faltantes según un patrón arbitrario. Se estudian desde métodos heurísticos hasta métodos basados en verosimilitudes, bajo distintos mecanismos para la no respuesta y con variables de características dispares. Se analiza el efecto sobre las estimaciones puntuales y la cobertura de los intervalos de confianza. Finalmente, se extraen conclusiones de utilidad para la práctica del análisis de datos.

Palabras clave: datos faltantes, imputación múltiple, inferencia estadística.

Clasificación AMS: 62-07, 62F99.

1. INTRODUCCIÓN Y OBJETIVOS

En el desarrollo teórico de la mayoría de técnicas y modelos estadísticos no se tienen en cuenta algunas cuestiones que surgen en su aplicación práctica, en concreto, un problema al que con seguridad se ha enfrentado cualquier analista de datos es el de los datos faltantes, también denominados perdidos o incompletos. Cuando se toma una muestra, en general con k variables, de tamaño n obtenemos una matriz de datos de dimensiones $n \times k$. Habitualmente esa matriz es incompleta en el sentido de que faltan datos sobre alguna o algunas de las variables para alguno o algunos de los casos, u observaciones, de la muestra. El estudio sistemático y la formalización de este problema desde un punto de vista probabilístico no se inicia hasta mediados de los años setenta, destacando principalmente el trabajo de Rubin (1976). Aún hoy, se tiende a infravalorar el efecto de eliminar de la matriz de datos aquellos casos con valores perdidos o a sustituirlos por valores que intuitivamente parecen razonables con el fin de eludir el problema y disfrutar de una nueva matriz completa sobre la cual aplicar los análisis pertinentes. De hecho, muchos de los programas informáticos de análisis de datos de uso generalizado incorporan dichas pseudo-soluciones en sus versiones estándar, de modo que son las empleadas por la mayor parte de los usuarios no especialistas.

Hasta hace relativamente poco, los únicos métodos generalmente utilizados para tratar el problema de los datos perdidos eran métodos como la eliminación del caso con valores perdidos, la sustitución/imputación de éstos por valores plausibles como la media de la variable o la predicción obtenida mediante regresión sobre las demás variables del vector, etc. Este tipo de métodos clásicos no suelen tener una base teórica sólida y, aunque fáciles de implementar y adecuados en situaciones concretas, presentan en general importantes inconvenientes y carencias, especialmente en contextos multivariantes. Los principales problemas inferenciales asociados son ineficiencia, aparición de sesgos, distorsión de la estructura de covarianzas; además de no incorporar la incertidumbre asociada a los datos faltantes.

Frente a estos métodos clásicos, en los últimos años, y de forma paralela a la formalización del problema de los datos faltantes, se han ido desarrollando métodos con una base teórica más sólida. Así, en Dempster, Laird y Rubin (1977) se establece una formulación general y rigurosa para la inferencia en presencia de datos faltantes mediante el algoritmo EM. Por otro lado, Rubin (1987) desarrolla una nueva metodología de propósito general, flexible y fundamentada que denomina imputación múltiple, y que salva muchos de los inconvenientes asociados al tratamiento tradicional de los datos faltantes.

1.1 Objetivos

Nuestro propósito en este artículo es analizar y comparar mediante un experimento de simulación el comportamiento de las principales estrategias existentes en la actualidad para el análisis estadístico de matrices de datos incompletos.

Este trabajo se inspira fundamentalmente en el artículo de Schafer y Graham (2002) donde se analiza un problema bidimensional con un patrón de no respuesta univariante. Nuestro objetivo es ampliar dicho estudio planteando una situación más compleja y general, y por ende, más realista, con un patrón de no respuesta arbitrario, considerando variables de comportamiento y rango no homogéneos, con distintos niveles de variabilidad y distintos grados de interrelación entre ellas.

Lo que pretendemos es que el investigador aplicado disponga de una herramienta más para poder elegir qué es lo que más le interesa, dadas las características de sus datos, a la vista de las ventajas e inconvenientes de cada estrategia. Mostraremos cómo afectan los distintos métodos a las medidas de variabilidad y correlación, qué métodos son a priori más fiables, cuales se muestran más robustos ante distintos mecanismos de no respuesta, etc.

En las secciones siguientes iremos describiendo los procedimientos analizados y la forma en la que se han implementado. Se compararán los resultados obtenidos y se extraerán conclusiones útiles desde un punto de vista práctico.

2. MECANISMOS DE NO RESPUESTA

Cuando en una muestra aparecen valores perdidos por razones fuera del control del investigador, es necesario establecer unos supuestos sobre el proceso que los ha generado. Estos supuestos serán en general no verificables, y por ello deberán hacerse explícitos y analizar la sensibilidad del procedimiento general de estimación frente a desviaciones de los mismos.

Si entendemos la presencia de valores perdidos como un fenómeno probabilístico necesitamos un mecanismo matemático que describa las leyes que rigen su aparición, y que capte aproximadamente las posibles relaciones entre la aparición de valores perdidos y los datos no observados en sí mismos.

Consideremos, con carácter general, un vector aleatorio X k-dimensional que genera los datos y un vector R, también k-dimensional, formado por variables aleatorias binarias tomando valores 0 ó 1 para indicar valor observado o no observado. Llamaremos *mecanismo de no respuesta* a la distribución de probabilidad de R. Asociada a una muestra del vector X tendremos una muestra de R, cuya forma dependerá de la complejidad del patrón de no respuesta. Por ejemplo, si el patrón

es univariante(1) tendremos una realización de una variable aleatoria binaria unidimensional indicando si un valor concreto es observado o perdido. Si el patrón es arbitrario(2), tendremos entonces una matriz de dimensiones $n \times k$ con elementos r_{ij} tomando valor 0, si x_{ij} es observado, ó 1, si x_{ij} es no observado.

Si denotamos mediante $\mathbf{X}$ a una muestra multidimensional de X podemos hacer una partición de forma que $\mathbf{X} = (X_{obs}, X_{per})$, donde X_{obs} y X_{per} denotan la parte observada y la parte no observada, o perdida, respectivamente. Se dice que los datos faltantes son de tipo MAR (*missing at random*) si la probabilidad de que un valor no se observe depende de los valores de los datos observados, pero no de los faltantes. Esto es, si $P[R \mid X_{obs}, X_{per}, \xi] = P[R \mid X_{obs}, \xi]$, siendo ξ un vector de parámetros desconocidos del mecanismo de no respuesta. Como indican los resultados de Rubin (1976), no es necesario que se satisfaga para todas las posibles realizaciones de *R*, basta con que se verifique en la muestra dada. Por otro lado, se dice que los datos faltantes son de tipo MCAR si $P[R \mid X_{obs}, X_{per}, \xi] = P[R \mid \xi]$. La hipótesis MCAR (*missing completely at random*) es más restrictiva ya que implica que R y X son independientes, algo difícil de mantener en muchas situaciones prácticas. Por último, se dice que los datos faltantes son de tipo NMAR (*not missing at random*) si el mecanismo de no respuesta depende del verdadero valor del dato perdido (es decir, depende de X_{per}), o de variables no observables. La hipótesis NMAR es la más general, pero al mismo tiempo es la más difícil de modelizar ya que exige la especificación de un modelo para R, por lo que es frecuente hablar de mecanismo de no respuesta *no ignorable*.

Sobre la hipótesis MAR descansan la mayoría de las técnicas actuales para datos faltantes, sin embargo no existen procedimientos generales para contrastarla sobre un conjunto de datos incompletos. Nos interesarán por lo tanto métodos que ofrezcan resultados robustos frente a posibles desviaciones. La sensibilidad de las respuestas obtenidas a partir de una muestra incompleta frente a supuestos débiles o injustificables es un problema básico asociado al análisis de datos incompletos, especialmente en el caso NMAR. En muchas aplicaciones lo prudente será considerar distintos modelos plausibles para el mecanismo de no respuesta y realizar un análisis de sensibilidad de las estimaciones. Aún así, como destacan Molenberghs *et al* (2001), esta estrategia puede llevar a conclusiones equivocadas. Podemos encontrar una revisión de las técnicas de análisis de sensibilidad en Serrat (2001), en el contexto del análisis de supervivencia. Por último, destacar el trabajo de

(1) Se habla de patrón univariante cuando los valores perdidos sólo aparecen en una de las variables del vector aleatorio.

(2) Se habla de patrón arbitrario o general cuando los valores perdidos pueden aparecer en cualquier variable y observación de la muestra.

Troxel *et al* (2004), el que se presenta un índice de sensibilidad a la no ignorabilidad, una medida del potencial impacto de la no ignorabilidad en un análisis.

3. PRIMEROS MÉTODOS HEURÍSTICOS

En esta sección nos ocuparemos de las soluciones habitualmente utilizadas en la práctica ante una matriz de datos con valores perdidos. Aunque intuitivamente pueden parecer soluciones razonables y cuando la cantidad de información perdida es pequeña, pueden funcionar relativamente bien en algunos casos, veremos en la sección 8 que no son procedimientos generalmente aceptables.

3.1 Análisis de casos completos

Dada una muestra $\mathbf{X}$ de k variables y n casos, supongamos que $n_{per} < n$ de estos casos presentan al menos un valor perdido para alguna de las k variables. Como su propio nombre indica, mediante este método se descartan los n_{per} casos y sólo se aplica la técnica o análisis sobre aquellos con valores observados para todas las variables. De este modo se pasa a trabajar con una muestra de datos completa de tamaño $n - n_{per}$. Las consecuencias de esta medida dependerán fundamentalmente de la cantidad de información que se pierda al descartar los casos con faltantes, del mecanismo según el cual faltan los datos y de la relación entre los casos completos e incompletos. La pérdida de información relevante se traducirá en sesgo y falta de precisión de las estimaciones si los faltantes no son una muestra aleatoria de la muestra completa, es decir, si no se verifica la hipótesis MCAR. Como ventajas destacaremos su simplicidad y el hecho de que todos los estadísticos se calculan utilizando el mismo tamaño muestral, lo que permite su comparación.

3.2 Análisis de casos disponibles

Si para el i-ésimo caso de una muestra se observan p de las k variables, al aplicar análisis de casos completos estamos perdiendo la información que sobre las k - p variables restantes contiene dicho caso. Una alternativa natural es utilizar para cada cálculo toda la información disponible en la muestra. Por ejemplo, a la hora de calcular la media o varianza de las variables X_i y X_j se utilizarán los n_i y n_j datos disponibles sobre cada una de ellas respectivamente. O bien, para calcular la covarianza entre las dos variables X_i y X_j se considerarán los casos h para los que los pares (x_{hi}, x_{hj}) son observados. Es obvio que ello implicará en general trabajar con distintos tamaños muestrales e incluso combinarlos en el cálculo de un mismo estadístico. Como puede verse en Little y Rubin (2002) es posible entonces que resulten correlaciones fuera del intervalo [-1,1] o matrices de correlaciones no definidas positivas, condición requerida por diversas técnicas multivariantes.

4. MÉTODOS DE IMPUTACIÓN SIMPLE

Los métodos de imputación pretenden solucionar el problema de los datos faltantes sustituyendo los mismos por valores estimados a partir de la información suministrada por la muestra. Con ello se consigue una matriz completa sobre la que realizar los análisis, salvando algunos de los problemas mencionados en la sección 3. Es precisamente la forma de estimar o predecir los valores perdidos lo que diferenciará unos métodos de otros. Nos centraremos en algunos de los más utilizados, concretamente en aquellos de aplicabilidad general y basados en un modelo estadístico explícito: imputación mediante la media, imputación mediante regresión e imputación mediante regresión estocástica.

4.1 Imputación mediante la media

Dada una variable X_i que presenta valores perdidos, mediante este método se reemplaza cada uno de ellos por $\bar{x}_i^{obs}$, la media de los valores observados de X_i. Aunque esta estrategia es sencilla y puede resultar intuitivamente satisfactoria, presenta un importante defecto, y es que, como veremos en la sección 8, tiende a subestimar la variabilidad real de la muestra al sustituir los faltantes por valores centrales de la distribución.

4.2 Imputación mediante regresión

Consideremos una variable X_i que presenta n_{per} valores perdidos y $n_i = n - n_{per}$ valores observados. Supongamos que las k-1 restantes variables X_j, con $j \neq i$, no presentan valores perdidos. Con este método se estima la regresión de la variable X_i sobre las variables X_j, $\forall j \neq i$, a partir de los n_i casos completos y se imputa cada valor perdido con la predicción dada por la ecuación de regresión estimada. Esto es, si para el caso l el valor x_{li} no se observa, entonces se imputa mediante:

$$\hat{x}_{li} = \hat{\beta}_{0\cdot obs} + \sum_{j \neq i} \hat{\beta}_{j\cdot obs} x_{lj} \qquad [4.1]$$

donde $\hat{\beta}_{0\cdot obs}$ y $\hat{\beta}_{j\cdot obs}$, $j \neq i$, representan los coeficientes de la regresión de X_i sobre X_j, $\forall j \neq i$, basada en las n_i observaciones completas. Frente a la imputación mediante la media, este método incorpora la información que sobre X_i contienen el resto de variables.

4.3 Imputación mediante regresión estocástica

Al imputar mediante regresión se está reemplazando el valor perdido por una media condicionada, por lo que, como destacábamos en el caso de imputación mediante la media, se tiende sistemáticamente a subestimar la variabilidad. Una sencilla alternativa para atenuar este efecto consiste en añadir al valor predicho por la regresión una perturbación aleatoria, con lo que se obtiene una realización de la distribución predictiva de los valores perdidos condicionada a los valores observados. Esto es, en vez de imputar mediante [4.1], utilizar:

$$\hat{x}_{li} = \hat{\beta}_{0\cdot obs} + \sum_{j \neq i} \hat{\beta}_{j\cdot obs} x_{lj} + \varepsilon_{li} \qquad [4.2]$$

donde $\varepsilon_{li} \sim N(0, \sigma^2_{resid})$, siendo σ^2_{resid} la varianza residual de la regresión de X_i sobre X_j, $\forall j \neq i$.

5. MÉTODOS BASADOS EN VEROSIMILITUDES

En esta sección nos centramos en métodos que se basan en funciones de verosimilitud, y que son por lo tanto métodos bajo los que subyace un modelo probabilístico. A continuación revisaremos el marco formal, debido a Rubin (1976), que da soporte a estos métodos y que se mantiene en la actualidad.

5.1 Un marco formal para la inferencia basada en muestras incompletas

Consideremos un fenómeno multivariante real cuyo comportamiento viene descrito por un vector aleatorio k-dimensional $X = (X_1,...,X_k) \in R^k$ con distribución de probabilidad $P[X;\theta]$, siendo θ el vector de parámetros desconocidos.

Cuando se dispone de una muestra completa de X, una amplia clase de métodos de inferencia se justifican en la interpretación de $P[X;\theta]$ como una función de verosimilitud que resume la evidencia que sobre θ hay en los datos. Pero en presencia de valores perdidos sólo disponemos de X_{obs} cuya distribución se obtiene como

$$P[X_{obs};\theta] = \int P[X;\theta] dX_{per} \qquad [5.1]$$

Si pretendemos hacer inferencia sobre θ a partir de la parte observada, es necesario comprobar que (5.1) es una verosimilitud adecuada. Rubin (1976) identifica

las condiciones para que así sea, estableciendo que basta con que se verifique la hipótesis MAR como comprobamos a continuación.

Según se ha formalizado el problema de las muestras incompletas, es necesario especificar un modelo para X, $P[X;\theta]$, y un modelo para la no respuesta, $P[R \mid X_{obs}, X_{per}, \xi]$. Mediante el producto $P[R \mid X_{obs}, X_{per}, \xi]P[X;\theta]$ obtenemos la distribución conjunta $P[X,R;\theta,\xi]$. La verosimilitud basada en la parte observada puede expresarse como

$$P[X_{obs},R;\theta,\xi] = \int P[X,R;\theta,\xi]dX_{per} = \int P[R \mid X_{obs}, X_{per}, \xi]P[X;\theta]dX_{per} \qquad [5.2]$$

Bajo el supuesto MAR, [5.2] queda como

$$P[X_{obs},R;\theta,\xi] = P[R \mid X_{obs},\xi]\int P[X;\theta]dX_{per} = P[R \mid X_{obs},\xi]P[X_{obs};\theta] \qquad [5.3]$$

De modo que la verosimilitud [5.2] bajo el supuesto MAR queda factorizada en dos partes, una relativa al vector θ y otra relativa al vector ξ. Si además θ y ξ son *distinguibles*(3), entonces las inferencias sobre θ basadas en verosimilitudes no se verán afectadas por $P[R \mid X_{obs},\xi]$, esto es, el mecanismo de no respuesta puede ser ignorado y la función de verosimilitud L de θ será $L(\theta \mid X_{obs}) \propto P[X_{obs};\theta]$. Este resultado pone de relieve que bajo ignorabilidad podemos realizar inferencias sobre el vector de parámetros θ de la distribución de X a partir de la verosimilitud $L(\theta \mid X_{obs})$.

Por otro lado, desde una perspectiva bayesiana, todas las inferencias se basan en la distribución de probabilidad a posteriori de los parámetros desconocidos, que puede escribirse utilizando el Teorema de Bayes como

$$P[\theta,\xi \mid R, X_{obs}] = \frac{P[R, X_{obs} \mid \theta,\xi]\varphi(\theta,\xi)}{\int\int P[R, X_{obs} \mid \theta,\xi]\varphi(\theta,\xi)d\theta d\xi} \qquad [5.4]$$

donde φ denota la distribución a priori de (θ,ξ). Bajo el supuesto MAR, podemos sustituir [5.3] en [5.4], obteniendo que $P[\theta,\xi \mid R, X_{obs}]$ es proporcional a $P[R \mid X_{obs},\xi]P[X_{obs} \mid \theta]\varphi(\theta,\xi)$. Si además θ y ξ son distinguibles, entonces la distribución marginal a posteriori de θ queda como

(3) En la práctica este supuesto implica que ε proporciona poca información sobre θ, y viceversa.

$$P[\theta \mid X_{obs}, R] = \int P[\theta, \xi \mid R, X_{obs}]d\xi \;\alpha\; P[X_{obs} \mid \theta]\varphi_\theta(\theta)\int P[R \mid X_{obs}, \xi]\varphi_\xi(\xi)d\xi$$
$$\alpha\; L(\theta \mid X_{obs})\varphi_\theta(\theta)$$

Por lo tanto, bajo la hipótesis de ignorabilidad, toda la información sobre θ se recoge en la distribución a posteriori que ignora el mecanismo de no respuesta observado, $P[\theta \mid X_{obs}] \alpha L(\theta \mid X_{obs})\varphi_\theta(\theta)$. La necesidad de especificar una distribución a priori para los parámetros puede resultar algo subjetivo o artificial desde un punto de vista no bayesiano. Sin embargo, como se destaca en Gelman *et al* (1995), conforme aumenta el tamaño muestral la verosimilitud domina a la distribución a priori, y las respuestas bayesianas y verosímiles tienden a converger. Además, en la mayoría de los problemas, la especificación de una distribución a priori no informativa que refleje el estado de ignorancia sobre los parámetros resulta adecuada, y esta es la opción tomada para el presente trabajo.

5.2 El algoritmo EM

El algoritmo EM (Dempster, Laird y Rubin, 1977) es un algoritmo iterativo diseñado para la obtención de estimadores máximo-verosímiles (EMV) en problemas con muestras incompletas. Sea $\mathbf{X}$ = (X_{obs}, X_{per}) una muestra incompleta de $X \sim P[X;\theta]$ a partir de la cual se desea obtener el EMV de θ. Podemos factorizar $P[X;\theta]$ como

$$P[X;\theta] = P[X_{obs};\theta]P[X_{per} \mid X_{obs}, \theta],$$

de donde es sencillo deducir que

$$\log L(\theta \mid X_{obs}) = \log L(\theta \mid X) - \log P(X_{per} \mid X_{obs}, \theta), \qquad [5.5]$$

siendo $\log L(\theta \mid X_{obs})$ la log-verosimilitud para los datos observados y $\log L(\theta \mid X)$ la log-verosimilitud para los datos completos X. En presencia de datos faltantes, el objetivo es estimar θ mediante la maximización de $\log L(\theta \mid X_{obs})$ con respecto a θ dada X_{obs}. Como vemos a continuación, el algoritmo EM relaciona el EMV de θ a partir de $\log L(\theta \mid X_{obs})$ con el EMV de θ a partir de $\log L(\theta \mid X)$. Tomando esperanzas respecto a $P[X_{per} \mid X_{obs}, \theta]$ a ambos lados de [5.5] y dado un estimador $\theta^{(t)}$ de θ, tenemos que

$$\log L(\theta \mid X_{obs}) = Q(\theta;\theta^{(t)}) - H(\theta;\theta^{(t)}),$$

donde $\quad Q(\theta;\theta^{(t)}) = \int \log L(\theta \mid X)P[X_{per} \mid X_{obs}, \theta^{(t)}]dX_{per}$

y

$$H(\theta;\theta^{(t)}) = \int \log P[X_{per} \mid X_{obs},\theta] P[X_{per} \mid X_{obs},\theta^{(t)}] dX_{per}$$

El paso E (*expectation*) del algoritmo EM calcula $Q(\theta;\theta^{(t)})$, reemplazando los valores perdidos, o una función de ellos, por su esperanza condicionada dados X_{obs} y $\theta^{(t)}$. El paso M (*maximization*) simplemente determina el EMV $\theta^{(t+1)}$ que maximiza $Q(\theta;\theta^{(t)})$ como si no hubiera datos perdidos. Los pasos E y M se repiten alternativamente generando una sucesión de estimadores $\{\theta^{(t)}\}$. La diferencia en el valor de la log-verosimilitud $\log L(\theta \mid X_{obs})$ en dos iteraciones sucesivas viene dada por

$$\log L(\theta^{(t+1)} \mid X_{obs}) - \log L(\theta^{(t)} \mid X_{obs}) = Q(\theta^{(t+1)};\theta^{(t)}) - Q(\theta^{(t)};\theta^{(t)}) + H(\theta^{(t)};\theta^{(t)}) - H(\theta^{(t+1)};\theta^{(t)})$$

Como el estimador $\theta^{(t+1)}$ se escoge de manera que $Q(\theta^{(t+1)};\theta^{(t)}) \geq Q(\theta^{(t)};\theta^{(t)})$, y $H(\theta^{(t)};\theta^{(t)}) \geq H(\theta^{(t+1)};\theta^{(t)})$, lo cual se deduce de la desigualdad de Jensen y la concavidad de la función logarítmica, tenemos que $\log L(\theta \mid X_{obs})$ se va incrementando en cada iteración con lo que se converge hacia el EMV de θ.

En McLachlan y Krishnan (1996) o Little y Rubin (2002), pueden encontrarse resultados teóricos y condiciones acerca de la convergencia del algoritmo. Un criterio de convergencia habitual en la práctica consiste en detener el proceso cuando la diferencia entre dos estimaciones sucesivas de θ sea suficientemente pequeña.

Es sencillo emplear el algoritmo EM como método de imputación. Una vez que se ha producido la convergencia, basta con dar un nuevo paso E y obtener las esperanzas matemáticas de los valores no observados condicionadas a los valores observados dado el EMV del vector de parámetros θ.

Para profundizar en los detalles, la base teórica y extensiones del algoritmo EM nos remitimos a la monografía de McLachlan y Krishnan (1996).

5.3 El método de imputación múltiple

Mediante imputación múltiple se reemplaza cada valor perdido por un conjunto de valores simulados con el fin de incorporar a la estimación la incertidumbre debida a la presencia de datos faltantes. La referencia básica sobre imputación múltiple es Rubin (1987), aunque podemos encontrar una variedad de trabajos relevantes como por ejemplo Rubin (1996), Schafer (1997), Little y Rubin (2002) o Zhang (2003).

Esta metodología ha permanecido durante algunos años en un segundo plano por su limitada aplicabilidad, debido principalmente a la inexistencia de herramientas computacionales adecuadas para poder crear las imputaciones. El desarrollo tecnológico de las últimas décadas ha permitido la implementación de algoritmos y procedimientos de cálculo computacionalmente intensivos necesarios para dar solución a problemas intratables analíticamente. En concreto, durante la década de los 90 se han popularizado los algoritmos MCMC (Markov Chain Monte Carlo) (véase p. ej. Gilks, Richardson y Spiegelhalter, 1996 ó Palarea, 2003) que permiten una modelización estadística más compleja al tiempo que realista. Este tipo de algoritmos también han encontrado su aplicación en el ámbito de los datos faltantes, en concreto, su incorporación al contexto de la imputación múltiple (Schafer, 1997) ha convertido este procedimiento en un destacado método para el análisis de datos incompletos.

El método consta de tres etapas:

1. IMPUTACIÓN: en esta etapa cada valor perdido se reemplaza por un conjunto de m valores simulados a partir de la distribución predictiva de X_{per} dado un modelo de probabilidad para X y una distribución a priori para θ. Dicha distribución $P[X_{per} \mid X_{obs}]$ puede obtenerse como

$$P[X_{per} \mid X_{obs}] = \int P[X_{per} \mid X_{obs}, \theta] P[\theta \mid X_{obs}] d\theta \qquad [5.6]$$

En [5.6] se refleja tanto la incertidumbre sobre X_{per} dado el vector de parámetros θ, como la propia incertidumbre asociada a θ. Destacar que en las imputaciones así generadas no interviene R, se elude el mecanismo de no respuesta. En consecuencia, como estudiamos en la subsección 5.1, esta forma de proceder sólo será teóricamente apropiada bajo la hipótesis MAR.

En general, $P[\theta \mid X_{obs}]$ y los cálculos donde interviene resultan intratables analíticamente, especialmente en contextos multidimensionales. Es aquí donde intervienen de forma natural los algoritmos MCMC dentro de esta metodología. En concreto, se utiliza el algoritmo de Aumento de Datos (Tanner y Wong, 1987) adaptado a este contexto para simular valores de [5.6] con los que realizar las imputaciones. El algoritmo de Aumento de Datos responde al siguiente esquema iterativo:

Dado θ^n

Repetir

Generar $X_{per}^{n+1} \sim P[X_{per} \mid X_{obs}, \theta^n]$

$\theta^{n+1} \sim P[\theta \mid X_{obs}, X_{per}^{n+1}]$

Incrementar n

Puede demostrarse (véase p. ej. Robert y Casella, 1999) que este algoritmo genera una cadena de Markov que converge hacia la distribución $P[X_{per} \mid X_{obs}]$ y, en consecuencia, tras un número suficientemente grande de iteraciones se estarán generando valores de X_{per} condicionados a X_{obs} que se emplearan para reemplazar los valores perdidos.

2. ANÁLISIS: En esta etapa se aplica el análisis de datos deseado (regresión, análisis discriminante,...) sobre cada una de las matrices imputadas y se almacenan los resultados.

3. COMBINACIÓN: Finalmente, las estimaciones resultantes se combinan para obtener una única estimación global.

Sea Q un parámetro poblacional unidimensional de interés (ej. una media, una varianza, un coeficiente β de una regresión,...), sea $\hat{Q}$ su estimación puntual si no hubiera datos perdidos, denotando con U la varianza estimada de $\hat{Q}$. Tras analizar los datos imputados tenemos m estimaciones $\{\hat{Q}_1,...,\hat{Q}_m\}$ con varianzas estimadas asociadas $\{U_1,...,U_m\}$. Se derivan las siguientes reglas (Rubin, 1987) para combinar las m estimaciones: la estimación puntual para Q basada en imputación múltiple, $\overline{Q}_m$, y su varianza asociada, T_m, vendrán dadas por

$$\overline{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \quad \text{y} \quad T_m = \overline{U}_m + \left(1+\frac{1}{m}\right)B_m \qquad [5.7]$$

donde $\overline{U}_m = \sum_{i=1}^{m} U_i/m$ representa la *variabilidad intra-imputaciones* y $B_m = \sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})(\hat{Q}_i - \overline{Q})'/(m-1)$ la *variabilidad entre-imputaciones*. Sin datos perdidos, las estimaciones $\hat{Q}_i$, $i = 1,...,m$, serían idénticas, con lo que B = 0 y $T = \overline{U}$.

Pueden encontrarse resultados ampliados para combinar ratios de verosimilitudes o p-valores en Meng y Rubin (1992) y Schafer (1997), ajustes para m pequeña en Rubin (1987), y análisis sobre el comportamiento asintótico de los estimadores en Rubin (1996) y Robins y Wang (2000).

6. DISEÑO DEL EXPERIMENTO

En esta sección describiremos el procedimiento y criterios seguidos para generar las muestras con datos faltantes utilizadas en la comparativa de los distintos métodos.

6.1 El modelo para los datos

Se han generado por simulación 5000 muestras de tamaño 150 de un vector aleatorio (X_1,X_2,X_3) que se distribuye según una normal 3-dimensional con las siguientes características: vector de medias $\mu = [30\ 15\ 100]'$, vector de desviaciones típicas $\sigma = [4\ 0.5\ 30]'$, y matriz de correlaciones

$$\rho = \begin{pmatrix} 1 & 0.8 & -0.5 \\ 0.8 & 1 & -0.2 \\ -0.5 & -0.2 & 1 \end{pmatrix}$$

Efectivamente, nos hemos centrado en variables de tipo continuo y normalmente distribuidas y esto puede parecer, en principio, una restricción importante. No consideramos que sea así ya que en la mayoría de modelos y técnicas estadísticas utilizadas en la práctica, y en sus implementaciones en distintos paquetes informáticos, subyace el supuesto de normalidad. Por lo tanto, las conclusiones obtenidas serán aplicables a una gran parte de los problemas prácticos de análisis de datos.

La determinación de los parámetros del modelo responde al interés por generar valores de un vector aleatorio cuyas componentes presentaran distintos niveles de variabilidad (alta, media y baja) e interrelación entre sí (positiva, negativa; fuerte, media, débil), y que los valores se concentraran en distintos intervalos de la recta real. Si calculamos los coeficientes de variación (CV) de cada componente del vector tenemos que: $CV_{X_1} = 0.133$, $CV_{X_2} = 0.033$ y $CV_{X_3} = 0.3$.

6.2 Implementación de las hipótesis MCAR, MAR y NMAR

Una vez generadas las muestras nos disponemos a eliminar artificialmente valores de cada una de ellas según los distintos mecanismos de no respuesta descritos en la sección 2 y siguiendo un patrón arbitrario. Como resultado tendremos tres conjuntos de 5.000 muestras con datos faltantes, uno para cada una de las hipótesis MCAR, MAR y NMAR.

A la hora de determinar el número de valores a eliminar por muestra se ha tenido en cuenta que la proporción final de valores perdidos resultase realista, un tanto elevada para acentuar las diferencias entre los métodos, y evitando además muestras con un excesivo número de casos con muchos valores perdidos. Los porcentajes promedio de casos con al menos un valor perdido para cada una de las hipótesis son: 42.12% para las muestras con perdidos bajo la hipótesis MCAR, 41.99% para las muestras bajo MAR y 40.11% para las muestras bajo NMAR.

Para implementar la hipótesis MCAR simplemente se han eliminado valores de cada matriz de datos original de forma aleatoria. Para la hipótesis MAR, a partir de las muestras originales, se ha seguido el siguiente criterio: eliminar x_{i1} si $x_{i2} > 18$ ó $x_{i3} > 130$, eliminar x_{i2} si $x_{i1} < 27$ y $x_{i3} < 90$ y eliminar x_{i3} si $x_{i1} > 33$ ó $x_{i2} < 14$. Para la hipótesis NMAR, el criterio ha sido: eliminar x_{i1} si $x_{i1} > 33.9$, eliminar x_{i2} si $x_{i2} < 14.5$, eliminar x_{i3} si $80 < x_{i3} < 90$. Estos criterios se basan en los valores cuartiles para cada muestra, aunque ajustados para alcanzar los porcentajes de perdidos recogidos en el párrafo anterior. Hay que tener en cuenta que las situaciones MCAR, MAR y NMAR aquí consideradas son, por decirlo de alguna manera, muy "puras", en la práctica es de esperar que se presenten situaciones combinadas y que los efectos sean menos radicales, pero ahora mismo lo que nos interesa es acentuar las diferencias.

6.3 Implementación de los métodos de inferencia para muestras incompletas

Para implementar los distintos procedimientos y automatizar el proceso de aplicación a cada una de las muestras han sido programadas varias rutinas y funciones utilizando el paquete S-PLUS.

La aplicación de análisis de casos completos, análisis de casos disponibles e imputación mediante la media se ha efectuado según lo descrito en las subsecciones 3.1, 3.2 y 4.1.

Respecto a los métodos de regresión, descritos en las subsecciones 4.2 y 4.3, tenemos que para cada matriz de datos se han estimado ecuaciones de regresión lineal de cada variable sobre el resto, imputando los faltantes en un caso con la predicción obtenida a partir de los valores observados para dicha caso(4). Para aplicar regresión estocástica se ha añadido a los valores imputados una perturbación aleatoria normal de media cero y varianza igual a la varianza de los residuos de las regresiones sobre casos completos.

El algoritmo EM, tal y como se ha descrito en la subsección 5.2, converge hacia las estimaciones máximo-verosímiles de los parámetros. A partir de estas estimaciones para cada muestra, hemos ejecutado una iteración más del algoritmo con el fin de imputar los datos faltantes mediante las esperanzas condicionadas.

El método de imputación múltiple (subsección 5.3) requiere un tratamiento más detallado. Para simular valores de la distribución predictiva [5.6] recurrimos al algoritmo de Aumento de Datos, el cual genera una cadena de Markov que converge a [5.6] tras un número suficientemente grande de iteraciones. No existen por el

(4) En ningún caso se han empleado valores imputados de una variable como predictores para las demás, siempre se parte de la matriz original.

momento reglas generalmente aplicables para determinar cuándo se alcanza la convergencia y cuándo parar la cadena, por lo que se suele recurrir al análisis y observación de la senda muestral de la cadena mediante técnicas para series temporales, análisis exploratorio de datos, etc. (véase p. ej. Mengersen y Robert, 1999). En nuestro caso, dado que resultaría prácticamente imposible analizar la convergencia del algoritmo para cada una de las 15000 muestras, se ha optado por una estrategia conservadora. Siguiendo las recomendaciones de Schafer (1997), para cada muestra se toma como referencia el número de iteraciones necesarias para la convergencia del algoritmo EM, en nuestro caso multiplicado por 100, y como punto de partida del algoritmo MCMC las estimaciones máximo-verosímiles obtenidas. Aún así la distribución normal no presenta modas múltiples ni regiones especialmente singulares, por lo que no se esperan problemas de convergencia.

Una vez alcanzada la convergencia se toman de la cadena los valores con los que se imputan los datos faltantes. Se extraen m conjuntos de valores simulados para generar las m matrices imputadas que requiere el método. Es sencillo comprobar que el valor m no necesita ser muy elevado para obtener estimaciones eficientes, suelen bastar 4 ó 5 simulaciones. Nosotros realizaremos 5 imputaciones para cada valor perdido. Dada la estructura de dependencias de una cadena de Markov, para asegurar la independencia de los valores simulados éstos se extraerán de la cadena cada cierto número de transiciones. Para fijar este número se suelen utilizar las funciones de autocorrelación (FAC). De nuevo, tras observar algunas de las FAC, adoptamos una postura conservadora y tomamos valores cada 1000 iteraciones. Finalmente, obtenemos las estimaciones mediante las ecuaciones [5.7] a partir de las 5 matrices imputadas para cada muestra.

7. INFERENCIA A PARTIR DE LAS MUESTRAS SIMULADAS

Se han obtenido las conocidas estimaciones máximo-verosímiles del vector de medias, de la matriz de covarianzas y de la matriz de correlaciones. Además, se ha calculado la cobertura(5) real de los intervalos de confianza para la media, las desviaciones típicas y las correlaciones.

Para medir la discrepancia entre las estimaciones puntuales obtenidas con cada método y el verdadero valor de los parámetros se ha calculado la raíz del error cuadrático medio (RECM) a lo largo de las 5000 estimaciones para cada parámetro y para cada mecanismo de no respuesta como $\sqrt{1/t \sum_{i=1}^{t} (\hat{\theta}_i - \theta)^2}$, siendo θ y $\hat{\theta}$, el parámetro y su estimación respectivamente, y t el número de muestras.

(5) Porcentaje de intervalos de confianza que contienen el verdadero valor del parámetro.

Respecto a los intervalos de confianza, se han utilizado los conocidos intervalos para la media con varianza desconocida basado en la distribución t de Student y para la desviación típica con media desconocida basado en la distribución χ^2 (pueden encontrarse en p. ej. Casella y Berger, 2001), todos ellos a un nivel de confianza $(1-\alpha)$ de 0.95. Como intervalo de confianza para la correlación lineal, ρ, se ha considerado aquel basado en la transformación de Fisher $z = \tanh^{-1}(r)$ con $z \pm 1.96(n-3)^{-1/2}$ a un nivel $1-\alpha = 0.95$, siendo r la correlación muestral. Dados los límites del intervalo de la transformación de Fisher, basta aplicar la función tangente hiperbólica sobre ellos para obtener los límites del intervalo para ρ.

En el caso particular del método de imputación múltiple debemos obtener intervalos de confianza combinados a partir de las m matrices imputadas, de forma que la variabilidad debida a la presencia de datos faltantes quede incorporada correctamente. Estos intervalos se basan en una distribución t de Student con v grados de libertad y se construyen a partir de las *m* estimaciones puntuales y sus respectivas varianzas, combinadas según [5.7], y con

$$v = (m-1)\left[1 + \left(\overline{U}_m / \left(1 + m^{-1}\right)B_m\right)\right]^2$$

El intervalo de confianza resultante para una cantidad Q es $\overline{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$ con un nivel de confianza, $(1-\alpha)$, el cual se deduce de la teoría sobre la distribución normal.

Para construir los intervalos de confianza combinados necesitamos las varianzas de los estimadores. Con el estimador media muestral, $\overline{x}$, de la media poblacional, μ, esto no es un problema, sin embargo para el estimador varianza muestral, , s^2de la varianza poblacional, α^2, la cuestión no es tan inmediata. Siguiendo a Schafer (1997) podemos obtener imputaciones múltiples válidas utilizando estimadores máximo-verosímiles y sus varianzas asintóticas. Así, la varianza asintótica de s^2 basada en la teoría de la distribución normal viene dada por la expresión $2\left(\sigma^2\right)^2/(n-1)$ (véase p. ej. Vélez y García, 1993). Para estimar σ^2 en esta expresión utilizaremos s^2 (en su versión insesgada). Otra posibilidad, más adecuada si el ajuste de los datos a la distribución normal no fuera bueno, sería aproximar la varianza de s^2 mediante métodos de remuestreo (véase p. ej. Casella y Berger, 2002). Para las correlaciones, tomamos como estimador $\hat{Q} = \tanh^{-1}(r)$ y como varianza del estimador $U = (n-3)^{-1}$, deshaciendo después la transformación para obtener el intervalo combinado para ρ.

Una vez construidos los intervalos obtenemos la cobertura real a lo largo de cada conjunto de muestras MCAR, MAR y NMAR, y la comparamos con la cobertura nominal (en este caso, el 95%). Hemos calculado también las amplitudes medias

de los intervalos, aunque las diferencias observadas no son relevantes y decidimos no incluirlas en este trabajo.

8. RESULTADOS

8.1 Estimaciones puntuales e intervalos de confianza

La tabla 1 recoge de forma sintética las estimaciones puntuales y la cobertura (en %) de los intervalos de confianza calculados con cada método(6). En general, se observa que las estimaciones puntuales y la cobertura van empeorando desde la situación MCAR a la NMAR, y van mejorando desde ACC hasta IMU o IEM. Para cualquiera de los tipos de parámetros y de mecanismos de no respuesta, los métodos que peores resultados arrojan son ACC, ACD y IME. En la situación MCAR las diferencias entre los métodos son menos acusadas, especialmente en el caso de las medias.

Los datos faltantes provocan un efecto de subcobertura de los intervalos de confianza respecto al nivel nominal del 95%, aunque mucho menor, o casi nulo, cuando se verifica la hipótesis MCAR. Bajo las hipótesis MAR o NMAR la situación empeora, y en varios casos de manera notable llegando a coberturas nulas. Como ocurre con las estimaciones puntuales, el grupo formado por los métodos IMU y IEM es el que mejores resultados proporcionan, mientras que el método ACC es el que globalmente presenta un peor comportamiento.

A continuación, aumentamos el nivel de detalle analizando los resultados para cada tipo de parámetro y variable.

Estimación e intervalos de confianza para μ_1, μ_2 y μ_3

En la situación MCAR presentan un sesgo casi nulo con cualquiera de los métodos, aunque también hay que tener en cuenta la varianza de los estimadores. Por ejemplo, la subestimación de las desviaciones típicas obtenidas con los métodos IME, IRE y IEM, llevará a sobrevalorar la precisión de sus estimadores para las medias. Bajo la hipótesis MAR los métodos IMU, IEM, IRE y IRS siguen proporcionando una estimación promedio ajustada, mientras que con el resto el sesgo se hace patente. En la situación NMAR todos los métodos producen estimaciones sesgadas, siendo de nuevo IMU, IEM, IRE y IRS los que ofrecen los mejores

(6) Se han utilizado las siguientes abreviaturas: ACC (análisis de casos completos), ACD (análisis de casos disponible), IME (imputación mediante la media), IRE (imputación mediante regresión), IRS (imputación mediante regresión estocástica), IEM (imputación mediante algoritmo EM), IMU (imputación múltiple).

resultados, especialmente IMU y IEM. Se observa también una relación directa entre el coeficiente de variación de las variables y la disparidad entre las distintas estimaciones.

En cuanto a los intervalos de confianza, destaca el mal comportamiento de IME que ya bajo la hipótesis MCAR da lugar coberturas reales entorno al 90%, lo cual es achacable a la falsa recuperación del tamaño muestral. En las situaciones MAR y NMAR es el método IMU el que se muestra más robusto, mientras que sólo bajo MAR los métodos IEM, IRS y IRE proporcionan también coberturas por encima del 90% para μ_1 y μ_2.

Para la media μ_3 se obtienen los peores resultados, sin embargo, en contra de lo esperado, observamos que las coberturas aumentan al pasar de la hipótesis MAR a la NMAR, cuestión ésta que analizaremos más adelante.

Estimación e intervalos de confianza para σ_1, σ_2, y σ_3,

La presencia de datos faltantes da lugar en general a una subestimación de la variabilidad que se acentúa de MCAR a NMAR. Como ocurría con las medias, en el supuesto MCAR las estimaciones promedio son bastante similares, aunque ya se desmarcan negativamente los métodos IME, IRE y IEM. Destaca por un lado el nefasto comportamiento de ACC bajo MAR y NMAR y, por otro, la robustez y buenas estimaciones de IMU. Los métodos IEM y IRS presentan un comportamiento intermedio y muy similar. Para σ_3 y bajo NMAR ocurre de nuevo algo inesperado: la RECM disminuye al pasar de MAR a NMAR.

En cuanto a los intervalos de confianza, los resultados son bastante parecidos a los obtenidos para las medias.

Estimación e intervalos de confianza para ρ_{12}, ρ_{13} y ρ_{23},

En general todos los métodos sobreestiman la correlación entre las variables, lo que se observa sobre todo para ρ_{13} y ρ_{23}. Como en los casos anteriores, en la situación MCAR todas las estimaciones promedio se sitúan en un entorno muy cercano al verdadero valor de las correlaciones. Los peores métodos en general son ACC, ACD y IME, especialmente al pasar de la situación MAR a la NMAR. En los casos MAR y NMAR son IMU, IEM, IRE y IRS los que mejor estiman las correlaciones. De nuevo, cuando en el cálculo interviene la variable X_3, se observa una mejora de las estimaciones al pasar de la situación MAR a la NMAR, salvo con el método ACC en ρ_{23}.

La disparidad de los resultados respecto a los intervalos de confianza es mayor que la observada para las medias y desviaciones típicas, y sigue siendo IME el peor situado en todos los casos arrojando ya una cobertura del 6.5% para ρ_{12} bajo MCAR. El método IMU se sitúa siempre cerca de la cobertura nominal, no viéndose

apenas afectado por la subcobertura. En un escalón más abajo se sitúan los métodos IRS, IRE, IEM.

A lo largo de esta sección hemos hecho referencia al inesperado comportamiento de las distintas estimaciones cuando el parámetro en cuestión se refería a la variable X_3 o bien ésta intervenía en su cálculo. Al describir el diseño del experimento de simulación veíamos que para imponer la hipótesis NMAR se eliminaban de cada una de las muestras los valores de X_3 en el intervalo (80;90). Dado que $\mu_3 = 100$ y $\sigma_3 = 30$ y, dicho intervalo se sitúa en la zona central de la distribución con mayor masa de probabilidad. Sin embargo, para X_1 y X_2 se eliminaban valores situados principalmente en las colas de la distribución. Al eliminar valores sólo del centro, hasta cierto punto se preservan las características de variabilidad de la distribución, y esto afecta positivamente a las estimaciones por cualquiera de los métodos, especialmente a varianzas y correlaciones. Además los métodos de imputación generarán con mayor probabilidad valores cercanos a los valores no observados. Por lo tanto, lo que está ocurriendo es que la imputación de los valores de X_3 está siendo muy buena y, por ello, la diferencia entre las estimaciones y los valores reales disminuye cuando el cálculo se ve afectado por X_3. En cuanto a los métodos que no reemplazan los valores perdidos (ACC y ACD), las estimaciones mejoran porque la parte perdida es la menos influyente en el cálculo de los estadísticos.

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continúa)

			MCAR		
	REAL		EST.	RECM	COB
ACC	mu X1	30	30,0098	0,4332	94,64
	mu X2	15	15,0010	0,0536	95,08
	mu X3	100	99,9535	3,2488	95,04
	sigma X1	4	3,9814	0,3016	95,08
	sigma X2	0,5	0,4976	0,0380	95,1
	sigma X3	30	29,8741	2,2790	95,56
	rho 12	0,8	0,7975	0,0396	94,96
	rho 13	-0,5	-0,4964	0,0823	94,76
	rho 23	-0,2	-0,1967	0,1042	95
ACD	mu X1	30	30,0059	0,3618	94,74
	mu X2	15	15,0001	0,0443	95,32
	mu X3	100	99,9545	2,6932	94,78
	sigma X1	4	3,9877	0,2509	95,34
	sigma X2	0,5	0,4986	0,0316	94,88
	sigma X3	30	29,8975	1,9162	95,16
	rho 12	0,8	0,7969	0,0568	94,92
	rho 13	-0,5	-0,4970	0,0810	95,08
	rho 23	-0,2	-0,1977	0,0957	95,12

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continuación)

			MCAR		
	REAL		EST.	RECM	COB
IME	mu X1	30	30,0059	0,3618	89,56
	mu X2	15	15,0001	0,0443	90,08
	mu X3	100	99,9545	2,6932	89,7
	sigma X1	4	3,6378	0,4283	65,4
	sigma X2	0,5	0,4549	0,0536	65,42
	sigma X3	30	27,2742	3,2368	65,2
	rho 12	0,8	0,6632	0,1450	6,5
	rho 13	-0,5	-0,437	0,1097	74,58
	rho 23	-0,2	-0,1646	0,0872	92,5
IRE	mu X1	30	30,0043	0,3399	92,82
	mu X2	15	15,0003	0,0423	93,38
	mu X3	100	99,9695	2,6368	91,48
	sigma X1	4	3,8734	0,2730	90
	sigma X2	0,5	0,4808	0,0360	88,34
	sigma X3	30	28,1023	2,6495	78,9
	rho 12	0,8	0,8321	0,0444	71,8
	rho 13	-0,5	-0,5343	0,0806	84,54
	rho 23	-0,2	-0,2225	0,0980	87,74
IRS	mu X1	30	30,0027	0,3494	93,3
	mu X2	15	15,0004	0,0434	93,64
	mu X3	100	99,9652	2,7885	91,9
	sigma X1	4	3,9880	0,2506	92,68
	sigma X2	0,5	0,4985	0,0317	92,92
	sigma X3	30	29,9054	2,0165	90,94
	rho 12	0,8	0,7797	0,0445	85,14
	rho 13	-0,5	-0,4887	0,0762	89,4
	rho 23	-0,2	-0,2026	0,0932	90,14

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continuación)

			MCAR		
	REAL		EST.	RECM	COB
IEM	mu X1	30	30,0041	0,3398	92,94
	mu X2	15	15,0003	0,0422	93,74
	mu X3	100	99,9699	2,6267	91,74
	sigma X1	4	3,8822	0,2695	90,66
	sigma X2	0,5	0,4819	0,0355	89,24
	sigma X3	30	28,1591	2,6123	79,66
	rho 12	0,8	0,8324	0,0445	71,52
	rho 13	-0,5	-0,5346	0,0805	84,4
	rho 23	-0,2	-0,2228	0,0977	88,12
IMU	mu X1	30	30,0041	0,3431	94,56
	mu X2	15	15,0003	0,0425	95,16
	mu X3	100	99,9803	2,6550	94,98
	sigma X1	4	3,9925	0,2428	94,5
	sigma X2	0,5	0,4992	0,0307	94,88
	sigma X3	30	29,9511	1,9210	93,98
	rho 12	0,8	0,7976	0,0340	95,42
	rho 13	-0,5	-0,4966	0,0711	94,76
	rho 23	-0,2	-0,1975	0,0889	94,86

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continuación)

			MAR		
	REAL		EST.	RECM	COB
ACC	mu X1	30	29,1990	0,8455	14,7
	mu X2	15	14,8997	0,1080	30,86
	mu X3	100	96,8425	3,9038	73,42
	sigma X1	4	2,4985	1,5146	0
	sigma X2	0,5	0,3726	0,1301	2,36
	sigma X3	30	21,3359	8,8268	0,92
	rho 12	0,8	0,6455	0,1665	15,12
	rho 13	-0,5	0,3618	0,1606	69,3
	rho 23	-0,2	-0,0205	0,2075	61,84
ACD	mu X1	30	30,5746	0,6693	61,6
	mu X2	15	15,0200	0,0449	92,42
	mu X3	100	105,5787	6,1874	45,48
	sigma X1	4	3,8015	0,3092	89,64
	sigma X2	0,5	0,4902	0,0306	94,04
	sigma X3	30	28,1068	2,6803	85,9
	rho 12	0,8	0,7711	0,0555	94,6
	rho 13	-0,5	-0,0080	0,4965	32,94
	rho 23	-0,2	-0,0482	0,1686	64,88
IME	mu X1	30	30,5746	0,6693	49,28
	mu X2	15	15,0200	0,0449	91,48
	mu X3	100	105,5787	6,1874	27
	sigma X1	4	3,4850	0,5620	38,92
	sigma X2	0,5	0,4827	0,0335	90,88
	sigma X3	30	24,3159	5,9444	9,68
	rho 12	0,8	0,6920	0,1173	17,68
	rho 13	-0,5	0,0072	0,4954	0
	rho 23	-0,2	-0,0408	0,1709	49,96
IRE	mu X1	30	30,0378	0,3448	92,72
	mu X2	15	15,0015	0,0405	95,06
	mu X3	100	99,6688	2,9379	86,68
	sigma X1	4	3,8460	0,2874	88,62
	sigma X2	0,5	0,4948	0,0294	94,74
	sigma X3	30	27,2309	3,5118	60,74
	rho 12	0,8	0,8086	0,0334	89,82
	rho 13	-0,5	-0,5437	0,1008	72,46
	rho 23	-0,2	-0,2244	0,1051	85,18

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continuación)

			MAR		
	REAL		EST.	RECM	COB
IRS	mu X1	30	30,0384	0,3508	92,42
	mu X2	15	15,0015	0,0406	95,32
	mu X3	100	99,6721	3,0369	87,82
	sigma X1	4	3,9062	0,2672	91,18
	sigma X2	0,5	0,4970	0,0296	94,54
	sigma X3	30	28,9085	2,5143	82,04
	rho 12	0,8	0,7926	0,0358	91,28
	rho 13	-0,5	-0,5035	0,0944	79,22
	rho 23	-0,2	-0,2100	0,1032	85,8
IEM	mu X1	30	30,0762	0,3683	90,64
	mu X2	15	15,0011	0,0407	94,98
	mu X3	100	100,7889	3,5473	78,96
	sigma X1	4	3,8834	0,2730	90,44
	sigma X2	0,5	0,4953	0,0296	94,4
	sigma X3	30	26,9314	3,9005	53,5
	rho 12	0,8	0,8222	0,0386	81,28
	rho 13	-0,5	-0,4802	0,1715	56,86
	rho 23	-0,2	-0,1802	0,1370	75,5
IMU	mu X1	30	30,0852	0,3723	92,3
	mu X2	15	15,0011	0,0407	94,32
	mu X3	100	100,8939	3,6247	91,62
	sigma X1	4	3,9552	0,2490	93,84
	sigma X2	0,5	0,4977	0,0295	93,7
	sigma X3	30	29,7900	2,3502	92,54
	rho 12	0,8	0,8002	0,0326	94,94
	rho 13	-0,5	-0,4205	0,1791	91,18
	rho 23	-0,2	-0,1584	0,1324	93,12

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Continuación)

			NMAR		
	REAL		EST.	RECM	COB
ACC	mu X1	30	29,5794	0,5028	67,64
	mu X2	15	15,0311	0,0464	86,12
	mu X3	100	105,7037	6,5093	54,24
	sigma X1	4	2,5800	1,4309	0
	sigma X2	0,5	0,3205	0,1809	0
	sigma X3	30	29,0504	2,3819	93,52
	rho 12	0,8	0,5711	0,2383	0,96
	rho 13	-0,5	-0,4290	0,1108	88,06
	rho 23	-0,2	0,0145	0,2378	47,3
ACD	mu X1	30	28,8175	1,2158	0,82
	mu X2	15	15,1433	0,1475	1,4
	mu X3	100	101,9564	3,3828	88,26
	sigma X1	4	3,1472	0,8763	3,92
	sigma X2	0,5	0,3964	0,1065	5
	sigma X3	30	31,2980	2,2819	88,78
	rho 12	0,8	0,2970	0,5082	0,46
	rho 13	-0,5	-0,4019	0,1253	81,72
	rho 23	-0,2	-0,1660	0,0997	93,88
IME	mu X1	30	28,8175	1,2158	0,38
	mu X2	15	15,1433	0,1475	0,54
	mu X3	100	101,9564	3,3828	83,2
	sigma X1	4	2,8742	1,1421	0,04
	sigma X2	0,5	0,3634	0,1386	0,02
	sigma X3	30	29,3921	1,9183	92,14
	rho 12	0,8	0,2403	0,5631	0
	rho 13	-0,5	-0,3467	0,1676	37,04
	rho 23	-0,2	-0,1422	0,0989	89,1
IRE	mu X1	30	29,5463	0,5374	62,74
	mu X2	15	15,0669	0,0756	49,64
	mu X3	100	101,7860	3,2405	85,74
	sigma X1	4	3,3932	0,6498	25,02
	sigma X2	0,5	0,4131	0,0912	13,24
	sigma X3	30	29,7503	1,8546	93,2
	rho 12	0,8	0,7739	0,0531	79,04
	rho 13	-0,5	-0,5197	0,0730	88,62
	rho 23	-0,2	-0,2041	0,0924	90,16

Tabla 1

VALORES REALES, ESTIMACIONES PUNTUALES Y COBERTURA REAL DE LOS INTERVALOS DE CONFIANZA

(Conclusión)

			NMAR		
	REAL		EST.	RECM	COB
IRS	mu X1	30	29,5471	0,5398	63,9
	mu X2	15	15,0667	0,0759	52
	mu X3	100	101,7774	3,3146	86,04
	sigma X1	4	3,4733	0,5799	37,62
	sigma X2	0,5	0,4247	0,0806	24,5
	sigma X3	30	31,0357	2,1734	88,68
	rho 12	0,8	0,7366	0,0823	55,6
	rho 13	-0,5	-0,4863	0,0731	91,06
	rho 23	-0,2	-0,1903	0,0918	90,84
IEM	mu X1	30	29,6023	0,4939	70,98
	mu X2	15	15,0595	0,0695	60,08
	mu X3	100	101,3981	3,0624	88,1
	sigma X1	4	3,4821	0,5759	39,22
	sigma X2	0,5	0,4249	0,0810	25,96
	sigma X3	30	29,7835	1,8544	93,24
	rho 12	0,8	0,8059	0,0428	80,58
	rho 13	-0,5	-0,5204	0,0734	88,54
	rho 23	-0,2	-0,2203	0,0942	89,18
IMU	mu X1	30	29,6007	0,4969	75,02
	mu X2	15	15,0596	0,0698	66,58
	mu X3	100	101,4071	3,0857	91,18
	sigma X1	4	3,5774	0,4948	56,68
	sigma X2	0,5	0,4387	0,0687	45,4
	sigma X3	30	31,2122	2,2311	95,02
	rho 12	0,8	0,7577	0,0643	87,3
	rho 13	-0,5	-0,4911	0,0696	95,1
	rho 23	-0,2	-0,2008	0,0872	94,86

CONCLUSIONES

En este trabajo se plantea y ejecuta un experimento de simulación con el fin de contrastar el rendimiento de las principales estrategias frente a problemas de

inferencia estadística con datos faltantes bajo distintos mecanismos de no respuesta.

Se ha comprobado que, globalmente, los resultados van empeorando desde la situación MCAR a la NMAR, y van mejorando desde el análisis de casos completos hasta la imputación múltiple o el algoritmo EM.

En general, los métodos para datos faltantes estudiados provocan la aparición de sesgos, subestimación de las varianzas, sobreestimación de las correlaciones y subcobertura de los intervalos de confianza. Estos efectos se acentúan de forma creciente al pasar a las situaciones MAR y NMAR. Es entonces cuando los métodos basados en verosimilitudes muestran un comportamiento más estable y robusto, especialmente el método de imputación múltiple.

El análisis de casos completos o casos disponibles, y la imputación mediante la media son procedimientos nada recomendables para la inferencia, con un comportamiento muy inestable y unas estimaciones sólo aceptables bajo la hipótesis MCAR, especialmente en lo que se refiere a la cobertura de los intervalos de confianza.

Hemos comprobado que para variables con alto coeficiente de variación, las estimaciones de los distintos métodos resultan ser más dispares y las cotas de error mayores. Por otro lado, cuando los valores perdidos se sitúan principalmente en el centro de la distribución, sus efectos nocivos son bastante más débiles.

En resumen, ante un problema de datos faltantes hay que tener en cuenta lo que se pretende estimar, la cantidad de datos faltantes, el conocimiento sobre cómo y donde faltan los datos, las características de las variables, y el esfuerzo que se esté dispuesto a realizar para obtener unas mejores estimaciones. En este trabajo se pone de manifiesto que las soluciones heurísticas habituales son garantía de una inferencia pobre e imprecisa y que los métodos basados en verosimilitudes ofrecen la mejor alternativa. Aún así, en determinadas situaciones, este tipo de métodos pueden suponer para el analista un esfuerzo y un gasto computacional que no quede compensado por sus bondades para la inferencia.

En futuros trabajos se profundizará en el análisis de estos procedimientos desde una perspectiva multivariante, y se abordará el estudio de métodos y modelos no sustentados en la hipótesis MAR con el fin de modelizar situaciones NMAR de forma más adecuada.

REFERENCIAS

CASELLA, G. Y BERGER, R.L., (2001), «Statistical Inference», 2ª edición, Duxbury.

DEMPSTER, A.P., LAIRD, N.M., Y RUBIN, D.B., (1977), «Maximum likelihood estimation from incomplete data via the EM algorithm (with discusion)», *J. of the Royal Statistical Society*, 39, 1-38.

GELMAN, A., CARLIN, J.B., STERN, H.S. Y RUBIN, D.B., (1995), «Bayesian Data Analysis», Chapman & Hall.

GILKS, W.R., RICHARSON, S. Y SPIEGELHALTER, D.J., (1996), «Markov Chain Monte Carlo in Practice», Chapman & Hall.

LITTLE, R.J.A. Y RUBIN, D.B., (2002), «Statistical Analysis with Missing Data», Wiley & Sons.

MCLACHLAN, G.J. Y KRISHNAN, T., (1996), «The EM algorithm and extensions», Wiley.

MENG, X.L. Y RUBIN, D.B., (1992), «Performing likelihood ratio tests with multiply imputed data sets», *Biometrika*, 79, 103-111.

MENGERSEN, K. Y ROBERT, C.P., (1999), «MCMC convergence diagnostics: a revieWWW», en Bayesian Statistics, eds. J. Berger, J. Bernardo, A.P. Dawid y A.F.M. Smith, Oxford Sciencies Publications.

MOLENBERGHS, G., KENWARD, M.G. Y GOETGHEBEUR, E., (2001), «Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case», *Applied Statistics*, 50 (1), 15-29.

PALAREA, J., (2003), «Algoritmos Monte Carlo basados en cadenas de Markov. Aplicación a la inferencia mediante imputación múltiple», Trabajo de investigación de tercer ciclo, Universidad de Murcia.

ROBERT, C.P. Y CASELLA, G., (1999), «Monte Carlo Statistical Methods», Springer.

ROBINS, J.M. Y WANG, N., (2000), «Inference for imputation estimators», *Biometrika*, 87, 113-124.

RUBIN, D.B., (1976), «Inference and missing data», *Biometrika*, 63, 581-592.

RUBIN, D.B., (1987), «Multiple Imputation for Nonresponse in Surveys», Wiley & Sons.

RUBIN, D.B., (1996), «Multiple imputation after 18+ years», *J. of the American Statistical Association*, 91, 473-489.

SERRAT, C., (2001), «Study and validation of data structures with missing values. Application to survival analysis», doctoral thesis, Universitat Politècnica de Catalunya, Barcelona.

SCHAFER, J.L., (1997), «Analysis of Incomplete Multivariate Data», Chapman & Hall.

SCHAFER, J.L. Y GRAHAM, J. W., (2002), «Missing data: our view of the state of the art», *Psychological Methods*, 7, 147-177.

TANNER, M. Y WONG, W., (1987), «The calculation of posterior distributions by data augmentation», *J. of the American Statistical Association*, 82, 528-550.

TROXEL, A.B., MA, G. Y HEITJAN, D.F., (2004), «An index of local sensitivity to nonignorability», *Statistica Sinica*, 14, 1221-1237.

VÉLEZ, R. Y GARCÍA, A., (1993), «Principios de inferencia estadística», UNED.

ZHANG, P., (2003), «Multiple imputation: theory and method», *International Statistical Review*, 71, 3, 581-592.

STATISTICAL INFERENCE METHODS WITH MISSING DATA. SIMULATION STUDY OF THE EFFECTS ON THE ESTIMATES.

ABSTRACT

In practice, data analysts frequently deal with non-observed data. In this paper we compare by means of a simulation study the performance and properties of different statistical procedures for missing data with an arbitrary pattern for non-response. We study from heuristic methods to model-based methods, under different missingness mechanisms and considering heterogenous variables. We analyze their effect on point estimates and on the coverage of confidence intervals. Finally, recommendations for practical data analysis are obtained.

Key words: missing data, multiple imputation, statistical inference.

AMS classification: 62-07, 62F99.