REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (7). Páginas 3–30.
Junio de 2009. ISSN: 1886-516X. D.L: SE-2927-06.
URL: http://www.upo.es/RevMetCuant/art25.pdf

Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus

Muñoz Rosas, Juan Francisco
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
Correo electrónico: jfmunoz@ugr.es

Álvarez Verdejo, Encarnación
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
Correo electrónico: encarniav@ugr.es

RESUMEN

La aparición de datos faltantes es un problema común en la mayoría de las encuestas llevadas a cabo en distintos ámbitos. Una técnica tradicional y muy conocida para el tratamiento de datos faltantes es la imputación. La mayoría de los estudios relacionados con los métodos de imputación se centran en el problema de la estimación de la media y su varianza y están basados en diseños muestrales simples tales como el muestreo aleatorio simple. En este trabajo se describen los métodos de imputación más conocidos y se plantean bajo el contexto de un diseño muestral general y para el caso de diferentes mecanismos de respuesta. Mediante estudios de simulación Monte Carlo basados en datos reales extraídos del ámbito de la economía y la empresa, analizamos las propiedades de varios métodos de imputación en la estimación de otros parámetros que también son utilizados con frecuencia en la práctica, como son las funciones de distribución y los cuantiles. Con el fin de que los métodos de imputación descritos en este trabajo se puedan implementar y usar con mayor facilidad, se proporcionan sus códigos en los lenguajes de programación R y Splus.

Palabras clave: información auxiliar; encuesta; probabilidades de inclusión; mecanismo de respuesta.
Clasificación JEL: C13; C15; C80.
2000MSC: 62D05.

Artículo recibido el 27 de noviembre de 2008 y aceptado el 9 de enero de 2009.

Imputation methods to handle the problem of missing data: an application using R/Splus

ABSTRACT

Missing values are a common problem in many sampling surveys, and imputation is usually employed to compensate for non-response. Most imputation methods are based upon the problem of the mean estimation and its variance, and they also assume simple sampling designs such as the simple random sampling without replacement. In this paper we describe some imputation methods and define them under a general sampling design. Different response mechanisms are also discussed. Assuming some populations based upon real data extracted from the context of the economy and business, Monte Carlo simulations are carried out to analyze the properties of the various imputation methods in the estimation of parameters such as distribution functions and quantiles. The various imputation methods are implemented using the popular statistical softwares R and Splus, and codes are here presented.

Keywords: auxiliary information; survey; inclusion probabilities; response mechanism.
JEL classification: C13; C15; C80.
2000MSC: 62D05.

1. Introducción

En el mundo económico y empresarial es conocida la utilización de las encuestas como herramientas para recoger la información necesaria para llevar a cabo estudios de muy diversa índole. La calidad de los resultados obtenidos y la posibilidad de generalización de los mismos dependen de la composición de la muestra, del tipo de encuesta y de la tasa de respuesta de la misma. A su vez, también se ha constatado como un problema importante la aparición de datos faltantes en determinados *item* del cuestionario. En otras palabras, la aparición de valores perdidos o la falta de respuesta es un problema común presente en cualquier estudio, especialmente en el ámbito de las ciencias sociales (véase Rubin, 1996).

Una técnica tradicional y muy conocida para el tratamiento de datos faltantes es la imputación. Las técnicas de imputación se pueden clasificar, en primer lugar, en dos grandes grupos: las técnicas de imputación simples y las de imputación múltiple.

Las técnicas simples de imputación han sido una de las herramientas más conocidas y aceptadas para el tratamiento de la falta de respuesta (véase Sedransk, 1985, Kalton y Kasprzyk, 1986 y Little y Rubin, 2002). Las técnicas simples de imputación presentan algunas ventajas frente a las técnicas de imputación múltiple. Por ejemplo, las técnicas simples tienen una implantación más sencilla sin que por el contrario sufran una importante pérdida de eficiencia en comparación con las técnicas de imputación múltiple. Por último, destacamos que las técnicas simples de imputación se pueden dividir en dos categorías: aleatorias y determinísticas.

El uso de la imputación puede provocar problemas serios de subestimación de la verdadera varianza cuando la proporción de datos faltantes es apreciable (Rao y Shao 1992). En general, un método de imputación aleatorio tiene la ventaja de añadir una mayor variabilidad a través de las imputaciones que un método determinístico de imputación; es decir, las técnicas simples determinísticas de imputación generalmente subestiman más las varianzas que las técnicas simples aleatorias de imputación. Sin embargo, las técnicas determinísticas proporcionan, en general, estadísticos más precisos que las técnicas aleatorias.

La imputación múltiple fue propuesta por Rubin (1978) como una alternativa a las técnicas simples de imputación. La imputación múltiple requiere la construcción de M (≥ 2) conjuntos de datos completos, los cuales se obtienen reemplazando cada dato faltante por M valores imputados, obtenidos mediante el mismo procedimiento de imputación. Aunque la imputación múltiple es una aproximación muy potente, sufre algunas limitaciones que no debemos

pasar por alto. Por ejemplo, Fay (1991) señala que la imputación múltiple puede conducir a estimadores de la varianza inconsistentes en el caso de encuestas multietápicas estratificadas. Rao y Shao (1992) afirman que las agencias estadísticas prefieren, en general, el uso de la imputación simple, debido especialmente a las dificultades operacionales que supone el mantenimiento de conjuntos de datos múltiples. Este problema se hace aún más grave en el caso de muestras con un tamaño muestral muy elevado, como por ejemplo en el caso de muestras nacionales. Rao (1996) también destaca algunas otras desventajas de la imputación múltiple en comparación con los métodos simples de imputación.

A pesar de la diversidad y variedad de técnicas de imputación que existen en la literatura, la mayoría de las aportaciones se han centrado en el impacto de las imputaciones realizadas sobre la estimación de la media poblacional y su varianza. Por ejemplo, Bello (1993) realizó un estudio de simulación para comparar varios métodos de imputación. Bello utilizó como criterio de comparación la precisión de los distintos métodos de imputación en la estimación de la media de la variable de interés y su varianza. Sin embargo, en la práctica existen otros parámetros, tales como la función de distribución y los cuantiles, que tienen una cantidad importante de aplicaciones, pero que no han sido estudiados en el contexto de datos faltantes y su tratamiento mediante imputación. Por otra parte, la mayoría de los estudios relacionados con métodos de imputación están basados en el muestreo aleatorio simple.

El objetivo de este trabajo es plantear las técnicas más conocidas de imputación simple bajo el esquema de un diseño general y analizarlas en el problema de la estimación de numerosos parámetros, incluyendo la media poblacional, la función de distribución y los cuantiles. Este estudio también analiza los distintos métodos de imputación, tanto para mecanismos de respuesta uniforme como para mecanismos de respuesta no uniforme. Además, se proporcionan las funciones o códigos en los lenguajes de programación R y Splus de los distintos métodos de imputación descritos en este trabajo, de modo que el lector pueda implementar y usar cada uno de los métodos descritos en este trabajo.

R es un entorno de análisis y programación estadística, compatible con el popular y comercial software Splus, que está atrayendo cada vez más a un alto número de usuarios. Una gran ventaja de R frente a otros entornos es que se trata de un lenguaje gratuito y disponible en la dirección

```
http://www.r-project.org                                    (1)
```

En realidad, R es un lenguaje que puede ser usado y distribuido libremente bajo los términos de Free Software Foundation's GNU General Public License en forma de código fuente, lo que implica que un gran número de personas

colaboran en su desarrollo y actualización. Aunque R es un lenguaje similar a otros lenguajes de programación muy conocidos y usados, como Fortran o C++, su nivel de ejecución, al igual que MATLAB (Merino y Vadillo, 2007), es muy superior, incluyendo numerosas operaciones con un solo comando.

Por su parte, Splus es un programa comercial distribuido por MATHSOFT Corporation, que incluye un interfaz bastante complejo y alta capacidad gráfica. A diferencia de R, Splus no es gratuito y tiene el inconveniente del coste de sus licencias.

En resumen, tanto R como Splus son entornos de programación con un lenguaje orientado a objetos y concebidos, originalmente, para ser utilizados en aplicaciones estadísticas. No obstante, en la actualidad, R y Splus son dos lenguajes poderosos y flexibles, que resultan suficientes para la resolución de la mayor parte de los problemas estadísticos habituales y sus aplicaciones en distintos ámbitos. En la página web oficial de R, dada en (1), pueden consultarse una serie de manuales actualizados y adecuados a todo tipo de usuarios. Además, podemos consultar Ihaka y Gentleman (1996), Arcos *et al.* (2004) y Arcos *et al.* (2005) como referencias más específicas. Todas las referencias anteriores podrían utilizarse para programar en Splus, puesto que en ambos lenguajes de programación, en general, se pueden utilizar los mismos comandos y funciones. No obstante, también podemos consultar Everitt (1994) y Krause y Olson (2005) como referencias más orientadas a Splus.

El presente artículo se estructura del siguiente modo. En la Sección 2 se describen los métodos simples de imputación más conocidos y usados. En la Sección 3 se llevan a cabo estudios de simulación Monte Carlo para estudiar el impacto de distintos métodos de imputación en la estimación de diferentes parámetros en el caso de un mecanismo de respuesta uniforme. Los estudios de simulación están basados en datos reales extraídos del ámbito de la Economía y la Empresa. Un mecanismo de respuesta uniforme es poco frecuente en la práctica; es decir, es bastante común encontrarse que las unidades muestrales fallan para proporcionar una respuesta con una determinada probabilidad. De este modo, en la Sección 4 se realizan algunas observaciones sobre los mecanismos de respuesta no uniforme. El comportamiento de los distintos métodos de imputación en el caso de mecanismos de respuesta no uniforme puede consultarse en la Sección 5. Este artículo también contiene dos apéndices. En el Apéndice A se incluyen los códigos en los lenguajes R y Splus de los distintos métodos de imputación descritos en este trabajo. Por su parte, en el Apéndice B mostramos, mediante un ejemplo, cómo podemos utilizar cualquiera de los mencionados métodos a partir de un vector con datos faltantes.

2. Métodos de imputación simple para un diseño muestral general

Consideremos una población finita $U = \{1, 2, \ldots, N\}$, con N unidades, de la cual se extrae, mediante un determinado diseño muestral, una muestra aleatoria s_n de tamaño n. Denotaremos como π_i a la probabilidad de que la unidad i pertenezca a la muestra s_n. Esta probabilidad π_i es conocida popularmente como probabilidad de inclusión de primer orden. El peso básico del diseño asociado a la i-ésima unidad vendrá expresado por $d_i = \pi_i^{-1}$. Por otra parte, y_i es el valor de la variable de interés y para la i-ésima unidad. En este trabajo también asumiremos que existe una variable auxiliar x asociada con la variable y. La extensión al caso de varias variables auxiliares de los distintos métodos de imputación discutidos en esta sección es un problema muy simple, que está cubierto por la literatura, y de ahí que nos centremos por simplicidad en el caso de una única variable auxiliar.

En los estudios relacionados con las encuestas por muestreo se asume, en general, que todas las respuestas en la muestra s_n son conocidas. Sin embargo, esta situación puede no presentarse en la práctica; es decir, es frecuente que en las encuestas se dispongan de valores faltantes por alguna determinada razón, como por ejemplo, la negativa del encuestado a dar la información requerida, la imposibilidad de contactar con el individuo encuestado, la pérdida casual de información, etc. De este modo, suponemos que r de los n valores de la variable y son observados (*respondientes*), mientras que el resto de $m = n - r$ valores de y corresponden a datos faltantes (*no respondientes*). Las muestras $s_r = \{i \in s_n \,|\, y_i \text{ es observado}\}$ y $s_m = \{i \in s_n \,|\, y_i \text{ no es observado}\}$ denotarán, por tanto, los conjuntos de respondientes y no respondientes asociados con la variable y. Cuando $i \in s_m$, el valor y_i necesita ser imputado, mientras las técnicas de imputación no serán necesarias en la muestra s_r. La proporción de datos faltantes se denotará por $p = m/n$. Por último, $Y_\alpha = \text{ínf}\{t : F(t) \geq \alpha\}$ denotará el cuantil poblacional de orden α de la variable y, donde

$$F(t) = \frac{1}{N} \sum_{i \in U} \Delta(t - y_i)$$

es la función de distribución poblacional, $\Delta(a) = 1$ si $a \geq 0$ y $\Delta(a) = 0$ en otro caso.

El objetivo de esta sección es plantear, para un diseño muestral general, las técnicas de imputación más conocidas y utilizadas, de modo que se puedan imputar los m valores faltantes de la variable y en la muestra s_n, y poder obtener la estimación de un determinado parámetro o realizar un análisis estadístico general usando los n valores de la muestra. Por ejemplo, el estimador de la media poblacional $\overline{Y} = N^{-1} \sum_{i \in U} y_i$ basado en los n valores de la muestra

está dado por

$$\overline{y}_I = \frac{1}{\sum_{i \in s_n} d_i} \sum_{i \in s_n} d_i \widetilde{y}_i, \tag{2}$$

donde $\widetilde{y}_i = y_i$ si $i \in s_r$, $\widetilde{y}_i = y_i^*$ si $i \in s_m$ y y_i^* es el valor imputado o donante para el dato faltante y_i.

El *método de imputación de la media* o *método de sustitución* consiste en utilizar la media muestral de los valores disponibles como donante en cada uno de los valores perdidos; es decir, los valores imputados por el método de la media están dados por $y_i^* = \overline{y}_r$, $i \in s_m$, donde

$$\overline{y}_r = \frac{1}{\sum_{i \in s_r} d_i} \sum_{i \in s_r} d_i y_i.$$

Este método es, sin duda, el más simple pero también el menos atractivo de los distintos métodos de imputación. La única ventaja de este método es que proporciona estimaciones insesgadas para la media poblacional, pero sólo en el caso de un mecanismo de respuesta uniforme. En el lado opuesto, este método de imputación distorsiona considerablemente la distribución de los datos, debido a la concentración de valores en torno a la media. Algunas consecuencias de este hecho, tal como se analiza en la Sección 3, son la presencia de sesgos muy elevados en la estimación de cuantiles y una considerable subestimación del verdadero valor de la varianza.

Una modificación del método de imputación anterior fue propuesta por Cohen (1996). Cohen propuso añadir más variabilidad a los valores imputados mediante el método de la media usando la variabilidad de los datos muestrales. Asumiendo un diseño muestral general, el método de Cohen consiste en imputar la mitad de los valores faltantes con los valores

$$\overline{y}_r + \sqrt{\frac{n+r-1}{r-1}} \widehat{\sigma}_r,$$

y la otra mitad de valores faltantes con los valores

$$\overline{y}_r - \sqrt{\frac{n+r-1}{r-1}} \widehat{\sigma}_r,$$

donde

$$\widehat{\sigma}_r^2 = \frac{1}{\sum_{i \in s_r} d_i} \sum_{i \in s_r} d_i (y_i - \overline{y}_r)^2. \tag{3}$$

El método NNI (acrónimo de *Nearest Neighbor Imputation*) es utilizado en numerosas encuestas llevadas a cabo por el Instituto de Estadística de Canadá y

algunas agencias nacionales de Estados Unidos. Este método utiliza el criterio del valor más proximo asociado a una variable auxiliar para proporcionar los valores imputados o donantes. En el caso de que la variable auxiliar disponga de varios valores equidistantes, se presenta el problema de la presencia de varios donantes para un mismo valor faltante. En esta situación, el criterio que se sigue es elegir aleatoriamente a un donante entre los diferentes candidatos. Una revisión más detallada del método NNI, así como numerosos resultados teóricos relacionados con este método, pueden consultarse en Chen y Shao (2000).

El *método de la razón* (véase Rao, 1996) es otro método de imputación determinística popularmente conocido y usado con bastante frecuencia en numerosos estudios. El método de la razón utiliza las cantidades

$$y_i^* = \frac{\overline{x}_r}{\overline{y}_r} x_i, \quad i \in s_m \tag{4}$$

como valores imputados, donde

$$\overline{x}_r = \frac{1}{\sum_{i \in s_r} d_i} \sum_{i \in s_r} d_i x_i.$$

Este método está basado en el hecho de que los valores definidos en (4) son los mejores predictores bajo un modelo de superpoblación que presente las siguientes características

$$E(y_i) = \beta x_i, \quad V(y_i) = \sigma^2 x_i. \tag{5}$$

Cuando el modelo (5) no sostenga, una alternativa es utilizar el *método de regresión* (véase Healy y Westmacott, 1956), el cual se basa en el modelo de superpoblación

$$E(y_i) = \alpha + \beta x_i, \quad V(y_i) = \sigma^2. \tag{6}$$

Asumiendo que el modelo (6) se ajusta razonablemente bien a los datos en estudio, el método de regresión utiliza como donantes los valores

$$y_i^* = \overline{y}_r + \widehat{\beta}(x_i - \overline{x}_r), \quad i \in s_m, \tag{7}$$

donde el estimador $\widehat{\beta}$ en la ecuación (7) viene dado por

$$\widehat{\beta} = \frac{\sum_{i \in s_r} d_i (x_i - \overline{x}_r)(y_i - \overline{y}_r)}{\sum_{i \in s_r} d_i (x_i - \overline{x}_r)^2}.$$

Notamos que los métodos de imputación anteriores se pueden clasificar como determinísticos, mientras que los métodos de imputación que describimos a continuación corresponden a métodos de imputación aleatoria.

El método de imputación conocido popularmente como *Random Hot Deck* (RHD) es uno de los métodos de imputación más utilizados en la práctica. Este método consiste en seleccionar mediante muestreo aleatorio simple con reemplazamiento m valores a partir de los r valores disponibles de la variable y. Este planteamiento, sin embargo, tan solo resulta apropiado cuando la muestra s_n es extraída bajo muestreo aleatorio simple. En el caso de un diseño muestral general, el método RHD necesita ser modificado para que los m valores seleccionados aleatoriamente tengan en cuenta el efecto del diseño muestral. De este modo, el método RHD puede usarse eficientemente en un diseño muestral general si la muestra de m valores es seleccionada con reemplazamiento y con probabilidades de selección

$$u_i = \frac{d_i}{\sum_{i \in s_r} d_i}, \quad i \in s_r.$$

Los métodos de razón y regresión descritos en (4) y (7) pueden obtener estimaciones que subestimen la verdadera varianza de la variable de interés. Por esta razón, es usual añadir pequeñas perturbaciones aleatorias a los valores predichos obtenidos en los métodos de razón y regresión. Estas perturbaciones aleatorias pueden generarse a partir de una distribución normal con media cero y varianza $\widehat{\sigma}_r^2$, donde $\widehat{\sigma}_r^2$ está definida en la ecuación (3).

Los métodos de imputación RHD y NNI tienen la ventaja frente al resto de métodos descritos en esta sección el utilizar como donantes a los propios valores de los respondientes; es decir, los métodos RHD y NNI utilizan valores observados para las imputaciones. Esta propiedad es especialmente atractiva en el caso de variables discretas, puesto que los donantes tomarán también valores discretos. En el lado opuesto, con los métodos de la media, Cohen, razón y regresión, los valores imputados no tienen por qué ser discretos.

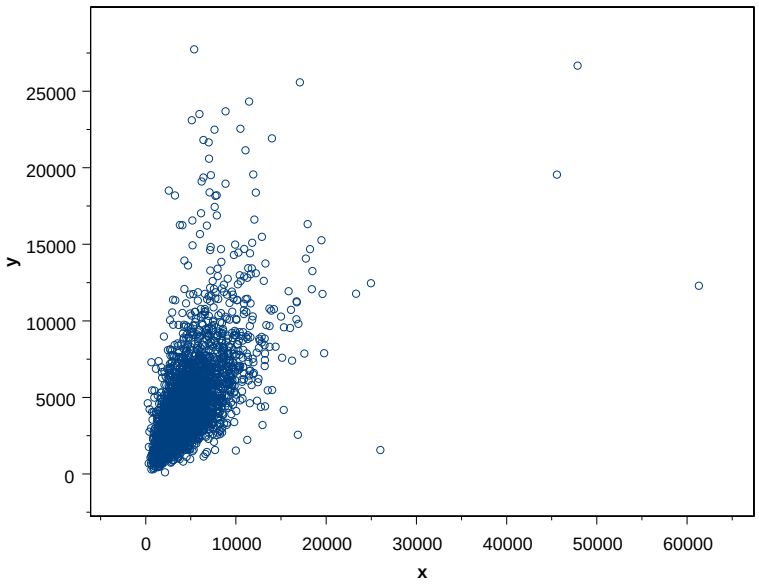
3. Comparación mediante estudios de simulación

En esta sección se comparan numéricamente los distintos métodos de imputación descritos en la sección anterior. Usaremos dos poblaciones reales para el estudio mediante simulación Monte Carlo de los distintos métodos de imputación en la estimación de varios parámetros que pueden presentarse en la práctica.

En primer lugar, los métodos de imputación se compararán con datos extraídos

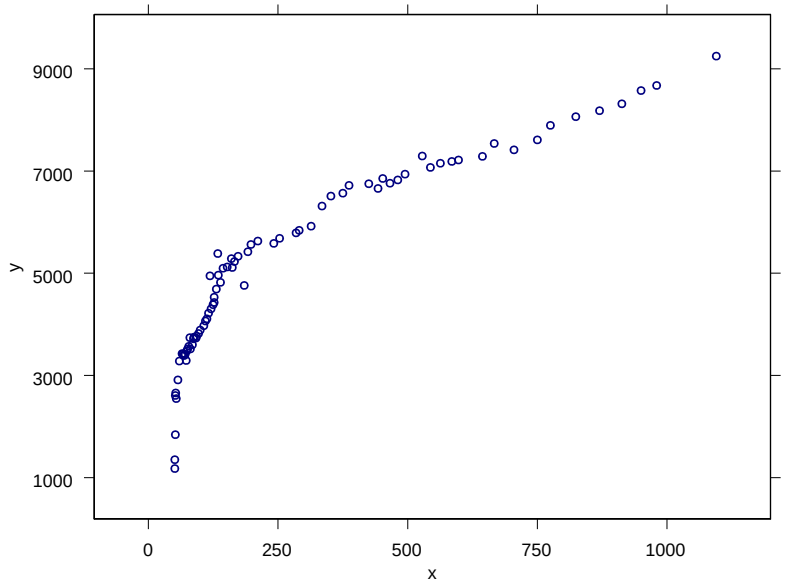
de la Encuesta Continua de Presupuestos Familiares (ECPF) del año 1997, que elabora el Instituto Nacional de Estadística (INE) de manera periódica. Esta población consta de N =3.114 individuos, donde la variable de interés son los gastos familiares, mientras que la variable auxiliar son los ingresos familiares. Notamos que, debido a la naturaleza de dichas variables, la presencia de datos faltantes es bastante común en este caso; es decir, la imputación resulta ser una solución práctica a este problema. La Figura 1 muestra la nube de puntos de las variables de la población ECPF. En segundo lugar, también analizamos las distintas técnicas de imputación en la población Factories, analizada en Murthy (1967) y Kuk y Mak (1993). Para esta población, la variable auxiliar es el número de trabajadores y la variable de interés el *output* de cada factoría. La nube de puntos de las variables de la población Factories puede consultarse en la Figura 2.

Figura 1. Nube de puntos de las variables x e y de la población ECPF.



Para realizar el estudio de simulación Monte Carlo seleccionamos 1.000 muestras bajo muestreo estratificado con afijación uniforme en cada una de las dos poblaciones, siguiendo a Chambers y Dunstan (1986) y Rao *et al.* (1990). Al igual que en las referencias anteriores, se utilizó el criterio de equipartición basado en la variable auxiliar x para la formación de los estratos (véase también Silva y Skinner, 1995). En cada una de las muestras se seleccionaron de manera uniforme y aleatoria m datos de la variable y. Dichas unidades se consideraron como valores perdidos y se imputaron mediante las distintas técnicas de imputación descritas en este trabajo, de modo que a partir de los datos muestrales (incluyendo las imputaciones) se estimaron diferentes parámetros de interés. Se consideraron valores de m de forma que la propor-

Figura 2. Nube de puntos de las variables x e y de la población Factories.



ción de datos faltantes en cada una de las poblaciones fuese $p = 0{,}1, 0{,}3, 0{,}5$, siendo $p = m/n$. El comportamiento de las estimaciones realizadas se midió en términos de Sesgo Relativo (SR) y Error Cuadrático Médio Relativo (ECMR), donde SR y ECMR se definen como

$$SR = \frac{1}{\theta}\left[\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\theta}_i - \theta)\right] \quad ; \quad ECMR = \frac{1}{\theta}\left[\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\theta}_i - \theta)^2\right]^{\frac{1}{2}},$$

donde θ es el parámetro de interés y $\widehat{\theta}_i$ es el valor de un dado estimador $\widehat{\theta}$ para la i-ésima muestra simulada.

Los distintos parámetros utilizados para evaluar el comportamiento de los métodos de imputación fueron la media poblacional, $\overline{Y}$, la varianza de la variable de interés, $S_y^2 = N^{-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$, los cuartiles poblacionales (es decir, $Y_{0,25}$, $Y_{0,5}$ y $Y_{0,75}$) y, por último, la función de distribución evaluada en los anteriores cuartiles poblacionales.

Notamos que los parámetros anteriores son bastantes frecuentes en la práctica. Por ejemplo, la media muestral es el parámetro más común en cualquier estudio mediante encuestas por muestreo. La función de distribución posee propiedades interesantes, como el hecho de describir las características más importantes de una distribución, o que otras medidas tal como los cuantiles pueden obtenerse a partir de la función de distribución. Por último, los cuantiles son también requeridos a menudo en la práctica. Por ejemplo, los cuantiles son altamente usados en muchas encuestas para obtener medidas de pobreza,

tal como la proporción de bajos ingresos (Berger y Skinner, 2003).

Los métodos de imputación involucrados en el estudio de simulación Monte Carlo son los siguientes. Asumiendo el caso de ausencia de información auxiliar, consideramos el método de imputación de la media, el método de Cohen y el método RHD. Además, con el fin de analizar si la imputación produce beneficios en términos de estimación, obtendremos estimaciones de los distintos parámetros a partir de los datos disponibles, sin usar imputaciones; es decir, consideraremos $\overline{y}_r$ para estimar la media poblacional $\overline{Y}$,

$$\widehat{S}_r^2 = \frac{1}{\sum_{i \in s_r} d_i} \sum_{i \in s_r} d_i (y_i - \overline{y}_r)^2$$

para estimar S_y^2,

$$\widehat{Y}_r(\alpha) = \text{ínf}\{t : \widehat{F}_r(t) \geq \alpha\}$$

para estimar Y_α, con $\alpha = 0{,}25, 0{,}5, 0{,}75$ y

$$\widehat{F}_r(t) = \frac{1}{\sum_{i \in s_r} d_i} \sum_{i \in s_r} d_i \Delta(t - y_i)$$

para estimar $F(t)$, con $t = Y_{0{,}25}, Y_{0{,}5}, Y_{0{,}75}$. A este método lo denotaremos como SI (sin usar imputaciones) en las distintas tablas de esta sección. Asumiendo información auxiliar en la etapa de estimación, consideramos el método NNI, los métodos de razón y regresión, y los métodos de razón y regresión añadiéndoles una perturbación aleatoria para corregir el problema de la subestimación en las varianzas. Estos últimos métodos se denotarán, respectivamente, como Razón.PE y Reg.PE.

Las Tablas 1 y 2 muestran los resultados obtenidos del estudio de simulación. El método de Cohen proporciona estimaciones con sesgos muy altos, que desvirtúan considerablemente las tablas, y de ahí que este método esté omitido en las tablas.

De la Tabla 1 (población ECPF) puede observarse que el método de la media produce grandes sesgos en la estimación de los distintos parámetros. Especialmente, el método de media subestima la varianza. El método de regresión también produce, en general, grandes sesgos. Este hecho quizás se deba a que la población en estudio sea más apropiada para el método de razón que para el método de regresión. Los métodos de razón y regresión que incorporan una perturbación aleatoria también proporcionan estimaciones con sesgos altos, especialmente en el caso de una alta proporción de datos faltantes. El resto de métodos de imputación obtienen estimaciones con sesgos dentro de un rango razonable de valores. En lo que respecta a la eficiencia, puede comprobarse que el método de razón es el más eficiente en la mayoría de los casos.

En la Tabla 2 pueden consultarse los valores de SR y ECMR de los distintos

Tabla 1
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población ECPF. Las muestras fueron seleccionadas bajo muestreo estratificado aleatorio con afijación uniforme y tamaño $n = 150$. Los 3 estratos utilizados se formaron mediante el criterio de equipartición.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	5.7 (-0.1)	5.7 (-0.1)	5.9 (-0.3)	5.6 (-0.3)	5.6 (-0.2)	5.6 (-0.2)	5.8 (-0.3)	5.8 (-0.2)
	$Y_{0,25}$	7.4 (0.0)	9.2 (5.4)	7.7 (0.0)	7.4 (-0.2)	7.1 (0.1)	8.0 (3.3)	7.6 (-1.7)	7.6 (-0.9)
	$Y_{0,5}$	5.9 (-0.2)	10.1 (7.8)	6.3 (-0.3)	6.1 (-0.2)	5.8 (-0.1)	5.8 (0.9)	5.9 (0.4)	6.0 (0.7)
	$Y_{0,75}$	6.5 (-0.1)	8.0 (-4.9)	6.9 (-0.3)	6.6 (-0.2)	6.4 (-0.2)	6.5 (-1.2)	6.8 (1.6)	6.9 (1.6)
	$F(Y_{0,25})$	15.1 (0.0)	16.9 (-10.0)	15.8 (0.3)	15.4 (0.4)	14.7 (-0.2)	16.0 (-6.6)	14.9 (3.3)	14.7 (1.7)
	$F(Y_{0,5})$	8.7 (0.4)	12.5 (-9.6)	9.1 (0.6)	8.7 (0.5)	8.4 (0.2)	8.9 (-1.3)	8.2 (-0.4)	8.3 (-0.8)
	$F(Y_{0,75})$	5.0 (0.0)	5.6 (3.4)	5.2 (0.2)	5.0 (0.1)	4.9 (0.1)	5.0 (0.8)	5.1 (-1.2)	5.1 (-1.2)
	S_y^2	29.5 (-0.9)	28.7 (-10.8)	30.3 (-1.1)	30.2 (-1.4)	32.1 (-1.3)	29.9 (-6.4)	35.1 (7.9)	32.0 (3.0)
0.3	$\overline{Y}$	6.6 (0.0)	6.6 (0.0)	7.0 (-0.5)	6.6 (-0.7)	6.5 (0.0)	6.3 (0.0)	7.1 (0.0)	7.0 (0.1)
	$Y_{0,25}$	8.5 (0.0)	22.0 (20.0)	9.1 (-0.3)	8.9 (-0.3)	7.8 (0.9)	13.7 (10.1)	10.2 (-5.1)	9.5 (-2.7)
	$Y_{0,5}$	6.8 (-0.1)	19.4 (17.7)	7.5 (-0.4)	7.3 (-0.4)	6.5 (0.6)	7.2 (3.2)	7.4 (2.1)	7.9 (3.2)
	$Y_{0,75}$	7.5 (0.0)	17.0 (-15.9)	8.1 (-0.3)	7.9 (-0.4)	7.1 (0.1)	7.4 (-2.6)	9.9 (6.0)	9.7 (5.6)
	$F(Y_{0,25})$	18.0 (0.1)	32.5 (-30.0)	19.3 (0.9)	18.6 (0.8)	16.3 (-1.5)	26.6 (-19.9)	18.4 (9.0)	17.0 4.7
	$F(Y_{0,5})$	10.0 (0.4)	30.7 (-29.5)	10.9 (0.8)	10.6 (1.0)	9.1 (-0.5)	12.7 (-5.6)	9.2 (-2.3)	9.8 (-3.6)
	$F(Y_{0,75})$	5.6 (0.0)	10.8 (10.0)	6.1 (0.3)	5.9 (0.2)	5.5 (-0.1)	6.3 (2.2)	7.0 (-4.2)	7.1 (-4.1)
	S_y^2	35.1 (0.0)	38.8 (-30.0)	35.4 (-2.4)	34.0 (-3.6)	42.2 (-0.8)	35.6 (-16.8)	56.4 (27.7)	41.6 (11.8)
0.5	$\overline{Y}$	7.7 (0.0)	7.7 (0.0)	8.3 (-1.1)	7.8 (-1.0)	7.4 (0.0)	7.1 (0.1)	8.4 (0.0)	8.2 (0.1)
	$Y_{0,25}$	9.1 (0.1)	49.9 (48.4)	10.2 (-0.6)	9.6 (-0.6)	8.0 (1.0)	19.9 (15.9)	14.1 (-9.7)	11.8 (-5.0)
	$Y_{0,5}$	8.1 (0.3)	19.9 (17.7)	9.2 (-0.3)	8.4 (-0.2)	7.2 (1.1)	8.6 (4.7)	9.2 (3.9)	10.3 (5.8)
	$Y_{0,75}$	9.2 (0.3)	20.9 (-19.9)	9.9 (-0.6)	9.4 (-0.6)	7.8 (0.2)	8.5 (-3.9)	14.6 (10.4)	14.0 (9.6)
	$F(Y_{0,25})$	19.5 (-0.2)	51.0 (-50.1)	21.8 (1.6)	20.2 (1.3)	16.2 (-1.6)	37.7 (-31.8)	21.3 (14.7)	17.9 (7.4)
	$F(Y_{0,5})$	11.5 (0.1)	50.7 (-47.8)	12.9 (1.0)	12.0 (0.7)	10.1 (-1.1)	17.1 (-9.1)	10.0 (-3.9)	11.5 (-6.4)
	$F(Y_{0,75})$	6.9 (-0.1)	17.0 (16.6)	7.5 (0.4)	7.2 (0.5)	6.0 (-0.1)	7.8 (3.7)	9.5 (-6.8)	9.7 (-6.7)
	S_y^2	40.5 (-1.7)	54.7 (-50.8)	41.5 (-5.6)	42.0 (-5.7)	49.5 (0.1)	44.0 (-26.8)	79.6 (48.3)	56.1 (21.3)

Tabla 2
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población Factories. Las muestras fueron seleccionadas bajo muestreo estratificado aleatorio con afijación uniforme y tamaño $n = 100$. ** denota cantidades superiores a 100. Los 2 estratos utilizados se formaron mediante el criterio de equipartición.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	3.5 (-0.1)	3.5 (-0.1)	3.7 (-0.1)	3.3 (-0.1)	4.0 (0.0)	3.3 (-0.1)	4.2 (0.0)	4.0 (-0.2)
	$Y_{0,25}$	4.1 (-0.7)	5.3 (2.0)	4.3 (-0.6)	3.7 (-0.6)	5.1 (-3.6)	4.1 (1.2)	4.7 (-2.9)	4.2 (-1.0)
	$Y_{0,5}$	5.5 (-1.1)	4.0 (1.4)	5.8 (-1.1)	5.2 (-0.9)	6.4 (-2.7)	6.1 (-2.5)	6.2 (-2.5)	5.5 (-1.3)
	$Y_{0,75}$	4.0 (-0.5)	6.3 (-3.4)	4.4 (-0.6)	3.6 (-0.4)	3.8 (0.1)	4.1 (-1.0)	3.7 (0.1)	3.8 (-0.4)
	$F(Y_{0,25})$	17.7 (-0.2)	19.0 (-10.2)	18.5 (-0.2)	16.4 (-0.2)	21.6 (13.4)	19.7 (-8.3)	20.0 (10.4)	17.5 (1.6)
	$F(Y_{0,5})$	10.0 (0.3)	17.2 (-2.8)	10.5 (0.1)	9.5 (0.1)	9.7 (3.0)	9.9 (2.9)	9.7 (2.7)	9.7 (0.7)
	$F(Y_{0,75})$	5.6 (0.1)	6.1 (3.5)	5.9 (0.2)	5.2 (0.0)	5.1 (-0.6)	5.2 (1.0)	5.2 (-0.6)	5.3 (0.2)
	S_y^2	10.8 (-1.4)	14.9 (-11.3)	11.3 (-1.8)	10.1 (-1.3)	74.0 (61.3)	10.8 (-2.9)	85.5 (72.2)	15.8 (7.4)
0.3	$\overline{Y}$	4.1 (0.2)	4.1 (0.2)	4.3 (0.3)	3.2 (0.1)	5.5 (0.2)	3.3 (0.2)	5.7 (0.2)	3.8 (0.2)
	$Y_{0,25}$	4.9 (-0.3)	15.9 (12.9)	5.7 (-0.1)	3.9 (-0.6)	15.2 (-13.0)	5.5 (3.6)	10.4 (-8.9)	5.0 (-1.4)
	$Y_{0,5}$	6.2 (-0.5)	4.5 (1.9)	6.8 (-0.5)	4.9 (-0.6)	10.4 (-7.5)	8.3 (-5.8)	8.7 (-5.7)	5.4 (-1.1)
	$Y_{0,75}$	4.9 (-0.5)	14.2 (-12.4)	5.4 (-0.6)	3.8 (-0.3)	4.4 (1.0)	5.1 (-2.2)	4.4 (1.1)	4.2 (-0.3)
	$F(Y_{0,25})$	20.4 (-0.8)	33.7 (-30.4)	22.4 (-0.8)	17.0 (-0.6)	44.9 (41.4)	32.1 (-24.0)	36.0 (31.7)	19.1 (4.1)
	$F(Y_{0,5})$	11.6 (-0.5)	35.5 (-10.8)	12.6 (-0.9)	9.6 (-0.4)	12.3 (8.7)	12.3 (8.2)	11.3 (7.3)	9.9 (0.9)
	$F(Y_{0,75})$	6.6 (-0.2)	10.9 (9.9)	7.2 (-0.1)	5.4 (0.0)	5.3 (-1.7)	5.7 (2.5)	5.4 (-1.9)	5.6 (0.0)
	S_y^2	13.4 (-1.7)	32.6 (-31.1)	14.8 (-2.1)	11.1 (-1.6)	** (**)	13.4 (-5.7)	** (**)	31.5 (24.2)
0.5	$\overline{Y}$	4.9 (-0.1)	4.9 (-0.1)	5.3 (0.0)	3.4 (-0.2)	7.7 (0.6)	3.6 (0.0)	8.3 (0.7)	4.6 (0.1)
	$Y_{0,25}$	6.6 (-0.2)	35.8 (34.4)	7.9 (0.4)	4.1 (-0.9)	34.9 (-33.4)	6.5 (4.5)	25.7 (-23.1)	6.7 (-2.3)
	$Y_{0,5}$	8.0 (-1.4)	5.2 (1.5)	8.3 (-1.2)	5.4 (-1.0)	17.5 (-15.3)	11.2 (-9.6)	13.4 (-11.0)	6.1 (-1.6)
	$Y_{0,75}$	6.2 (-1.3)	23.1 (-22.7)	6.5 (-1.1)	3.9 (-0.4)	6.1 (2.8)	6.6 (-4.1)	6.1 (2.9)	4.6 (-0.5)
	$F(Y_{0,25})$	25.5 (1.1)	51.2 (-49.5)	28.4 (1.1)	18.8 (1.3)	71.8 (69.5)	45.3 (-34.5)	56.8 (53.7)	22.9 (8.1)
	$F(Y_{0,5})$	13.4 (0.2)	54.6 (-11.1)	15.1 (-0.5)	10.0 (0.0)	17.1 (14.8)	16.8 (14.0)	15.0 (12.3)	10.7 (1.6)
	$F(Y_{0,75})$	8.0 (0.2)	17.3 (16.8)	8.5 (0.2)	5.6 (0.0)	5.8 (-3.0)	6.7 (4.4)	6.0 (-3.3)	5.6 (0.4)
	S_y^2	16.1 (-2.1)	51.8 (-51.1)	17.4 (-2.5)	11.7 (-1.2)	** (**)	16.2 (-7.5)	** (**)	48.3 (41.3)

métodos de imputación para la población Factories. Puede observarse que el método de imputación de la media nuevamente obtiene estimaciones con sesgos elevados en la estimación de los distintos parámetros. En esta población, el método de razón también produce serios sesgos debido a que el modelo de superpoblación (5) no se ajusta bien a los datos en estudio. Cuando el coeficiente de correlación lineal entre y y x sea elevado y la variabilidad de la variable auxiliar sea pequeña, el método NNI tendrá un mejor comportamiento que otras técnicas de imputación. En efecto, a medida que disminuye la variabilidad de x, los valores x_i (el valor de x asociado al dato faltante y_i, con $i \in s_m$) y x_j (el valor de x en la muestra s_r que se encuentra más próximo a x_i) estarán, en general, más próximos. Si a esto unimos que la correlación lineal entre y y x es elevada, resulta evidente que el valor utilizado como donante (el valor de y en la muestra s_r asociado a x_j, es decir, y_j) reemplazará de forma más eficiente al dato faltante y_i. Este hecho explicaría que el método NNI sea más eficiente que otras técnicas de imputación en la población Factories, cuyas variables tienen una correlación lineal muy elevada. Debido a un mejor ajuste, el método de regresión generalmente es más eficiente que el método de razón.

4. Algunas indicaciones sobre el mecanismo de respuesta

Las secciones anteriores están basadas en el caso de un mecanismo de respuesta uniforme. Sin embargo, esta situación suele presentarse con poca frecuencia en la práctica. Los métodos de imputación tradicionales tratan la respuesta no uniforme formando clases de imputación y entonces asumen respuesta uniforme dentro de cada una de estas clases de imputación (Berger y Rao, 2006). De este modo, todas las unidades no tendrán la misma probabilidad de respuesta. Por ejemplo, el método RHD dentro de clases de imputación es uno de los métodos de imputación usados con mayor frecuencia para la imputación de datos faltantes en encuestas por muestreo de hogares (Brick y Kalton, 1996). Una amplia revisión sobre clases de imputación puede consultarse en Kalton (1983).

En esta sección se describen los mecanismos de respuesta no uniforme más comunes en la práctica, así como las características más importantes de cada uno de ellos. En la Sección 5 evaluaremos empíricamente los distintos métodos de imputación en la presencia de clases de imputación para el tratamiento de mecanismos de respuesta no uniforme. Esto nos ayudará a extraer algunas conclusiones importantes para el caso de mecanismos de respuesta no uniforme.

En términos generales, las clases de imputación se construyen usando una variable discreta cuyos valores son observados para todas las unidades muestrales. También, estratos o uniones de estratos son usados con frecuencia para la construcción de las clases de imputación. En el caso de estratos numerosos

y pequeños, las clases de imputación se construyen mediante estratificación a posteriori (Valliant, 1993) y/o combinación de estratos pequeños. En resumen, podemos decir que la información auxiliar también se utiliza para crear las clases de imputación y mejorar, de este modo, la estimación de parámetros cuando el mecanismo de respuesta no es uniforme.

En la práctica, son dos los mecanismos de respuesta no uniforme que pueden presentarse. Por un lado, el mecanismo de respuesta UCRM (acrónimo de *unconfounded response mechanism*) depende exclusivamente de las variables auxiliares, mientras que en el mecanismo de respuesta CRM (acrónimo de *confounded response mechanism*) la probabilidad de respuesta también depende de la variable objeto de estudio. Es importante conocer los riesgos derivados de un mecanismo de respuesta no uniforme, puesto que sesgos considerables pueden obtenerse con un mecanismo CRM. Dichos sesgos pueden reducirse si se conocen las características del mecanismo de respuesta. Desafortunadamente, el mecanismo de respuesta es desconocido en la mayoría de las aplicaciones reales. Otra opción para reducir el sesgo es el uso de una corrección en los datos imputados. En este sentido, Rancourt *et al.* (1994) discutieron varios estimadores de tipo razón (véase también Hu *et al.*, 2001) diseñados para corregir el sesgo en el caso de un mecanismo de respuesta CRM.

5. Comparación en presencia de mecanismos de respuesta no uniforme

En esta sección se comparan los distintos métodos de imputación mediante estudios de simulación Monte Carlo usando las poblaciones ECPF y Factories y considerando un número diferente de clases de imputación con el fin de poder analizar la ganancia que supone un incremento del número de clases en la estimación de los distintos parámetros. Las clases de imputación se obtuvieron mediante el criterio de equipartición, también utilizado para la formación de los estratos en la Sección 3.

Siguiendo los estudios de Rancourt *et al.* (1994), se consideraron los siguientes mecanismos de respuesta no uniforme:

(M1) La probabilidad de no respuesta es una función decreciente de x_i dada por $\exp(-\gamma x_i)$; es decir, (M1) es un mecanismo de respuesta UCRM.
(M2) La probabilidad de no respuesta es una función creciente de x_i dada por $1-\exp(-\gamma x_i)$; es decir, (M2) es también un mecanismo de respuesta UCRM.
(M3) La probabilidad de no respuesta es una función decreciente de y_i dada por $\exp(-\gamma y_i)$; es decir, (M3) es un mecanismo de respuesta CRM.
(M4) La probabilidad de no respuesta es una función creciente de y_i dada por $1-\exp(-\gamma y_i)$; es decir, (M4) es también un mecanismo de respuesta CRM.

La constante γ se determina de modo que la proporción media de no respuesta sea igual a los valores $p = 0{,}1$, $p = 0{,}3$ y $p = 0{,}5$. Los distintos métodos de imputación se compararon en términos de SR y ECMR, obteniéndose los resultados mostrados en las Tablas 3 a 6.

Asumiendo el mecanismo de respuesta (M1) y dos clases de imputación, el comportamiento de los distintos métodos de imputación puede consultarse en la Tabla 3, en la cual puede observarse que los estimadores obtenidos a partir de los datos disponibles (sin usar imputación, es decir, método SI) proporcionan sesgos muy elevados, especialmente a medida que aumenta la proporción de datos faltantes. Este hecho se debe a que la falta de respuesta es no uniforme, puesto que en el caso de respuesta uniforme se pudo comprobar que este método de estimación obtenía sesgos dentro de un rango razonable. A pesar de la introducción de clases, podemos observar cómo el método de imputación de la media también genera estimaciones muy sesgadas, especialmente cuando el parámetro de interés no es la media poblacional. En lo que respecta al sesgo del resto de estimadores, podemos destacar una importante subestimación de los métodos de razón y regresión en la estimación de la varianza, así como importantes sesgos del método de razón cuando la proporción de datos faltantes es elevada. Por último, puede observarse que la incorporación de perturbaciones en los métodos de razón y regresión solventa el problema de la subestimación en la varianza, aunque la variabilidad añadida es tan elevada que produce el efecto contrario, es decir, las estimaciones ahora sobreestiman de manera importante la varianza poblacional de los datos en estudio. Este hecho, tal como se confirma en las líneas siguientes, se debe a los *outlier* o datos anómalos presentes en la variable de interés. Desde el punto de vista de la eficiencia, podemos comprobar que ambos métodos de razón y el método de regresión son generalmente los que obtienen las estimaciones más eficientes en la mayoría de los parámetros.

En la Tabla 4 se introducen dos clases más en comparación con la Tabla 3; es decir, en la Tabla 4 se aplican los distintos métodos de imputación dentro cada una de las cuatro clases creadas en cada muestra seleccionada. Aunque la introducción de más clases de imputación obtiene estimaciones tan eficientes como las obtenidas en el caso de dos clases de imputación, puede observarse que, en general, las estimaciones en el caso de cuatro clases están menos sesgadas. Por ejemplo, el método de razón que incorpora perturbaciones aleatorias obtiene en algunos parámetros (en especial la varianza) valores de SR superiores al 5 % en el caso de dos clases de imputación, mientras que en el caso de cuatro clases dichos sesgos son siempre inferiores al 5 %. En este sentido, también podemos destacar que el hecho de usar un número mayor de clases hacen que los outlier estén controlados y colocados en una única clase, lo que implica que los métodos basados en perturbaciones aleatorias sean más precisos y trabajen mejor. A partir de estas indicaciones, se deduce que para la población ECPF el método de razón con perturbaciones aleatorias y cuatro clases de imputación

Tabla 3
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población ECPF. Se han utilizado dos clases de imputación para tratar el mecanismo de respuesta (M1). Las muestras fueron seleccionadas bajo muestreo aleatorio simple y tamaño $n = 150$.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	7.0	5.7	5.8	5.4	5.3	5.3	5.5	5.5
		(4.2)	(1.7)	(1.7)	(0.2)	(-0.3)	(0.2)	(-0.3)	(0.2)
	$Y_{0,25}$	9.0	13.2	8.2	7.4	6.9	7.3	7.2	7.5
		(5.6)	(11.2)	(3.5)	(0.2)	(-0.3)	(1.7)	(0.4)	(1.4)
	$Y_{0,5}$	7.3	6.5	6.1	5.9	5.8	5.8	5.9	5.9
		(4.0)	(-3.0)	(1.1)	(-0.6)	(-1.1)	(-1.4)	(0.3)	(0.7)
	$Y_{0,75}$	7.3	6.1	6.2	6.2	6.0	6.0	6.0	6.1
		(3.8)	(-0.3)	(1.0)	(0.1)	(-1.1)	(-0.9)	(0.4)	(0.5)
	$F(Y_{0,25})$	18.5	24.3	17.1	15.9	14.6	16.7	14.3	15.0
		(-11.9)	(-20.7)	(-7.1)	(-0.4)	(0.6)	(-4.0)	(-0.6)	(-2.6)
	$F(Y_{0,5})$	10.1	9.7	9.0	8.4	8.4	8.5	8.3	8.6
		(-5.6)	(3.2)	(-2.0)	(0.3)	(1.2)	(1.6)	(-0.9)	(-1.3)
	$F(Y_{0,75})$	5.5	4.5	4.8	4.7	4.4	4.5	4.6	4.7
		(-2.7)	(0.2)	(-0.7)	(-0.1)	(0.8)	(0.6)	(-0.4)	(-0.4)
	S_y^2	30.5	27.7	28.7	28.2	27.7	27.8	29.0	28.7
		(5.3)	(-3.5)	(1.0)	(1.6)	(0.5)	(-0.9)	(5.0)	(3.5)
0.3	$\overline{Y}$	10.9	6.7	7.1	6.3	5.9	6.0	6.3	6.2
		(8.5)	(2.8)	(2.8)	(-0.1)	(-1.1)	(0.1)	(-1.1)	(0.0)
	$Y_{0,25}$	13.7	30.3	9.9	9.4	8.5	11.1	8.2	8.7
		(11.0)	(29.2)	(4.3)	(-0.3)	(3.0)	(6.7)	(-1.6)	(0.7)
	$Y_{0,5}$	11.3	9.4	7.2	6.8	6.2	6.3	6.8	7.1
		(8.9)	(-7.0)	(2.5)	(-0.2)	(1.3)	(0.4)	(1.9)	(2.6)
	$Y_{0,75}$	10.4	9.0	7.9	7.5	7.7	7.4	7.4	7.6
		(7.2)	(2.9)	(1.9)	(-0.3)	(-4.5)	(-2.2)	(1.9)	(2.1)
	$F(Y_{0,25})$	25.2	45.3	20.7	19.4	17.0	25.3	15.0	15.6
		(-20.0)	(-44.0)	(-9.3)	(0.3)	(-6.1)	(-15.7)	(2.8)	(-1.4)
	$F(Y_{0,5})$	15.2	17.3	10.8	9.8	9.9	10.6	9.2	9.9
		(-12.1)	(6.8)	(-3.8)	(0.0)	(-2.5)	(-1.1)	(-2.8)	(-4.0)
	$F(Y_{0,75})$	8.3	5.4	5.8	5.5	5.9	6.0	5.6	5.8
		(-5.9)	(-1.5)	(-1.3)	(0.3)	(3.5)	(1.8)	(-1.4)	(-1.5)
	S_y^2	36.8	30.1	32.6	30.4	27.7	28.6	31.6	31.1
		(12.3)	(-15.0)	(2.5)	(-0.1)	(-9.4)	(-11.5)	(8.2)	(5.8)
0.5	$\overline{Y}$	13.9	7.8	8.6	7.1	6.4	6.5	7.1	7.4
		(11.3)	(3.8)	(4.1)	(0.0)	(-1.6)	(0.3)	(-1.6)	(0.1)
	$Y_{0,25}$	15.5	35.1	11.9	11.3	11.8	15.3	10.8	10.7
		(11.7)	(33.2)	(4.9)	(0.3)	(7.2)	(11.4)	(-4.1)	(-1.1)
	$Y_{0,5}$	14.8	11.9	9.2	8.4	8.0	8.5	8.2	9.0
		(11.1)	(-9.1)	(3.4)	(0.1)	(4.0)	(2.6)	(2.4)	(3.6)
	$Y_{0,75}$	15.0	16.0	9.8	9.0	8.9	9.0	9.0	9.7
		(11.8)	(12.0)	(3.5)	(0.2)	(-6.2)	(-0.8)	(3.5)	(4.0)
	$F(Y_{0,25})$	29.1	62.0	24.4	23.9	21.5	34.7	18.1	18.1
		(-22.5)	(-61.3)	(-9.8)	(-0.2)	(-12.7)	(-25.8)	(6.9)	(2.3)
	$F(Y_{0,5})$	18.7	25.8	12.9	11.9	13.7	14.3	10.5	11.8
		(-15.2)	(6.3)	(-5.1)	(-0.4)	(-7.3)	(-4.6)	(-3.4)	(-4.8)
	$F(Y_{0,75})$	11.0	8.6	7.5	6.6	7.4	7.9	6.8	7.3
		(-8.5)	(-5.7)	(-2.9)	(-0.1)	(5.1)	(1.0)	(-2.6)	(-3.0)
	S_y^2	46.8	37.9	40.4	35.9	32.5	35.0	38.7	38.4
		(18.6)	(-28.0)	(5.6)	(-1.9)	(-21.1)	(-24.2)	(12.5)	(9.7)

Tabla 4
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población ECPF. Se han utilizado cuatro clases de imputación para tratar el mecanismo de respuesta (M1). Las muestras fueron seleccionadas bajo muestreo aleatorio simple y tamaño $n = 150$.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	6.9	5.4	5.6	5.5	5.4	5.4	5.5	5.5
		(3.8)	(0.5)	(0.5)	(-0.1)	(-0.3)	(-0.1)	(-0.3)	(0.0)
	$Y_{0,25}$	8.9	9.7	7.3	7.3	7.0	7.2	7.1	7.2
		(5.6)	(6.5)	(1.6)	(0.0)	(1.2)	(1.7)	(0.9)	(1.4)
	$Y_{0,5}$	7.4	5.7	5.8	5.7	5.4	5.5	5.6	5.8
		(4.4)	(-0.5)	(0.3)	(-0.2)	(-0.7)	(-0.8)	(0.6)	(0.7)
	$Y_{0,75}$	7.0	6.1	6.1	6.2	6.1	6.1	6.0	6.2
		(3.5)	(-1.1)	(0.1)	(-0.1)	(-1.1)	(-1.0)	(0.1)	(0.2)
	$F(Y_{0,25})$	18.0	24.8	15.5	15.0	14.7	16.3	14.0	14.5
		(-11.6)	(-15.0)	(-3.2)	(0.2)	(-2.1)	(-3.6)	(-1.3)	(-2.4)
	$F(Y_{0,5})$	10.3	8.2	8.4	8.3	8.3	8.3	8.3	8.4
		(-6.1)	(0.5)	(-0.8)	(-0.1)	(0.8)	(0.8)	(-1.2)	(-1.4)
	$F(Y_{0,75})$	5.6	4.6	4.8	4.8	4.6	4.6	4.7	4.8
		(-2.7)	(0.7)	(-0.2)	(0.0)	(0.7)	(0.7)	(-0.2)	(-0.2)
	S_y^2	29.4	27.4	27.5	27.6	27.2	27.3	27.7	27.7
		(2.1)	(-5.1)	(-2.0)	(-1.3)	(-3.1)	(-3.6)	(0.4)	(-0.4)
0.3	$\overline{Y}$	11.1	6.0	6.2	6.2	5.8	5.8	6.2	6.2
		(8.8)	(1.1)	(1.2)	(0.2)	(-0.3)	(0.1)	(-0.3)	(0.2)
	$Y_{0,25}$	13.8	14.8	9.0	9.1	10.7	10.8	8.5	8.6
		(11.0)	(9.3)	(1.7)	(0.0)	(6.1)	(6.2)	(0.4)	(1.3)
	$Y_{0,5}$	11.7	7.8	7.1	7.2	6.4	6.8	7.1	7.2
		(9.1)	(2.3)	(0.4)	(-0.2)	(0.9)	(0.7)	(1.8)	(1.8)
	$Y_{0,75}$	11.0	7.6	7.3	7.4	7.3	7.4	7.1	7.3
		(7.8)	(-3.2)	(0.8)	(0.5)	(-3.4)	(-3.0)	(1.3)	(1.6)
	$F(Y_{0,25})$	24.9	44.1	18.6	18.9	20.4	25.0	15.4	16.2
		(-19.8)	(-29.1)	(-3.7)	(-0.2)	(-11.6)	(-14.4)	(-0.7)	(-2.1)
	$F(Y_{0,5})$	15.5	12.1	10.1	10.1	11.0	11.2	9.9	10.2
		(-12.2)	(-2.9)	(-1.1)	(0.1)	(-2.2)	(-1.7)	(-2.9)	(-3.0)
	$F(Y_{0,75})$	8.6	6.1	5.6	5.6	5.7	5.7	5.5	5.6
		(-6.3)	(2.5)	(-0.5)	(-0.2)	(2.6)	(2.3)	(-0.9)	(-1.1)
	S_y^2	36.4	28.2	30.0	30.1	27.2	27.5	29.6	29.7
		(13.4)	(-11.9)	(1.8)	(1.3)	(-9.6)	(-9.8)	(4.3)	(4.0)
0.5	$\overline{Y}$	13.9	7.1	7.7	7.4	6.5	6.8	7.1	7.4
		(11.0)	(1.4)	(1.4)	(0.1)	(-0.6)	(0.1)	(-0.6)	(0.2)
	$Y_{0,25}$	15.2	16.8	10.5	11.1	15.9	15.0	9.9	10.0
		(11.5)	(9.0)	(1.8)	(0.5)	(11.8)	(10.3)	(-0.1)	(0.4)
	$Y_{0,5}$	14.6	11.6	8.6	8.8	8.5	9.0	8.4	8.8
		(11.3)	(6.2)	(0.9)	(-0.2)	(3.7)	(3.3)	(2.8)	(2.8)
	$Y_{0,75}$	15.0	10.6	9.3	9.1	8.8	9.3	8.8	9.5
		(11.5)	(-4.5)	(1.3)	(0.4)	(-4.2)	(-3.0)	(2.3)	(2.9)
	$F(Y_{0,25})$	28.4	57.3	22.8	22.9	27.7	34.8	17.7	18.3
		(-21.8)	(-37.9)	(-3.3)	(-0.3)	(-20.5)	(-23.2)	(0.8)	(0.2)
	$F(Y_{0,5})$	18.4	17.8	11.9	12.2	15.2	15.0	11.5	11.7
		(-14.6)	(-7.8)	(-1.3)	(-0.2)	(-7.1)	(-5.7)	(-4.0)	(-4.0)
	$F(Y_{0,75})$	11.0	8.4	6.8	6.8	7.8	7.5	6.8	7.0
		(-8.4)	(3.0)	(-1.0)	(-0.3)	(3.6)	(2.4)	(-1.7)	(-2.1)
	S_y^2	48.4	36.2	39.3	37.7	34.5	34.6	37.5	38.5
		(19.2)	(-24.0)	(2.8)	(-1.1)	(-22.8)	(-21.8)	(4.9)	(6.2)

sería la técnica de imputación apropiada que estima los distintos parámetros de manera más eficiente y con sesgos, en términos absolutos, más pequeños.

Notamos que la ganancia más importante, en términos de SR y ECMR, se produce cuando pasamos de muestras sin clases de imputación a dos clases de imputación. A medida que incorporamos clases de imputación, la eficiencia de las distintas estimaciones permanece aproximadamente constante, mientras que los sesgos se van reduciendo paulatinamente. No obstante, la reducción en el sesgo se mantiene hasta el uso de cinco clases de imputación, obteniéndose incluso resultados negativos cuando el número de clases de imputación es demasiado elevado. Esto se debe a la existencia de clases con pocos valores muestrales que producen, para esa clase, estimaciones poco fiables.

Las Tablas 5 y 6 presentan los valores de SR y ECMR en la población Factories cuando el mecanismo de respuesta es (M1) y se usan, respectivamente, dos y cuatro clases de imputación. Las conclusiones que se derivan de estas tablas son similares a las que ya hemos comentado en las Tablas 3 y 4. No obstante, destacamos que si en el supuesto de respuesta uniforme el método NNI era el más eficiente en la mayoría de los casos debido a la alta correlación entre las variables, cuando el mecanismo de respuesta es no uniforme y se utilizan cuatro clases de imputación, los métodos de regresión con perturbaciones y NNI son los más eficientes de entre los distintos métodos comparados en el estudio; es decir, el método de regresión se vuelve tan eficiente como el método NNI con la introducción de clases de imputación.

Simulaciones basadas en el mecanismo de respuesta (M2) fueron también analizadas en los estudios de simulación, obteniéndose conclusiones similares a las expuestas para el mecanismo (M1), y de aquí que esta información esté omitida. Asumiendo los mecanismos de respuesta (M3) y (M4), los distintos métodos de imputación mostraron sesgos elevados debido a que el mecanismo de respuesta es de tipo CRM. En este caso, los distintos métodos de imputación necesitarían un ajuste para corregir el sesgo causado por este tipo de mecanismo no uniforme. En este sentido, en Rancourt *et al.* (1994) y Hu *et al.* (2001) se discuten algunos métodos para el problema de imputación en el caso de mecanismo de respuesta CRM. El estudio de estos métodos de imputación ajustados no es el objetivo de este trabajo y de aquí que esta información esté también omitida.

En resumen, se ha constatado que el mecanismo de respuesta es un factor muy importante en el problema de la imputación para el tratamiento de datos faltantes. Si bien la no utilización de imputación puede producir estimaciones eficientes y poco sesgadas en un mecanismo de respuesta uniforme, esta metodología resulta poco apropiada en el caso de mecanismos de respuesta no uniforme. Por otro lado, hemos comprobado que el uso de clases de imputación proporcionan estimaciones mas eficientes y menos sesgadas, espe-

Tabla 5
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población Factories. Se han utilizado dos clases de imputación para tratar el mecanismo de respuesta (M1). Las muestras fueron seleccionadas bajo muestreo aleatorio simple y tamaño $n = 100$.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	5.1 (3.9)	3.7 (1.9)	3.8 (1.9)	3.1 (0.0)	3.4 (-1.5)	3.1 (0.5)	3.4 (-1.5)	3.3 (0.5)
	$Y_{0,25}$	6.8 (4.2)	9.3 (7.0)	5.4 (2.3)	3.5 (-1.2)	4.5 (-2.9)	3.7 (-1.2)	4.1 (-2.3)	4.0 (0.2)
	$Y_{0,5}$	5.5 (3.3)	5.3 (-1.8)	4.6 (1.2)	4.9 (-1.2)	5.1 (-1.4)	5.2 (-1.6)	4.9 (-1.1)	4.6 (-0.5)
	$Y_{0,75}$	3.3 (1.6)	3.6 (-0.7)	3.5 (-0.6)	3.6 (-0.7)	3.6 (-0.6)	3.6 (-0.7)	3.5 (-0.6)	3.5 (-0.6)
	$F(Y_{0,25})$	24.8 (-19.0)	30.7 (-27.1)	23.0 (-15.0)	15.9 (-0.4)	16.3 (7.1)	17.1 (0.8)	15.5 (4.6)	17.3 (-5.7)
	$F(Y_{0,5})$	14.0 (-10.1)	8.8 (0.9)	11.6 (-5.7)	8.9 (-0.1)	8.8 (0.4)	8.8 (0.8)	8.9 (-0.2)	9.3 (-1.3)
	$F(Y_{0,75})$	6.6 (-3.5)	5.0 (0.1)	5.0 (0.1)	5.0 (0.1)	5.0 (0.1)	5.0 (0.1)	5.0 (0.0)	5.0 (0.1)
	S_y^2	11.2 (-4.4)	14.8 (-11.1)	12.7 (-7.2)	10.3 (0.1)	14.6 (10.6)	11.2 (-5.0)	18.8 (14.6)	11.1 (-1.1)
0.3	$\overline{Y}$	11.4 (10.7)	5.1 (3.9)	5.2 (3.9)	3.0 (0.2)	4.9 (-3.7)	3.2 (0.8)	5.1 (-3.7)	3.4 (0.8)
	$Y_{0,25}$	17.8 (15.0)	21.1 (20.6)	10.2 (6.5)	3.6 (-1.2)	11.7 (-10.0)	3.8 (-0.7)	9.4 (-8.0)	5.0 (0.2)
	$Y_{0,5}$	15.1 (13.3)	10.4 (-8.6)	6.2 (4.1)	4.8 (-0.8)	6.7 (-3.7)	7.2 (-4.7)	5.6 (-1.8)	4.7 (0.1)
	$Y_{0,75}$	6.0 (5.5)	3.6 (-0.3)	3.1 (-0.2)	3.3 (-0.4)	3.2 (0.3)	3.1 (-0.3)	3.1 (0.6)	2.9 (0.2)
	$F(Y_{0,25})$	40.6 (-37.4)	57.3 (-56.2)	31.9 (-24.6)	16.5 (-0.3)	30.9 (27.4)	21.3 (-0.9)	26.4 (22.1)	18.3 (-4.3)
	$F(Y_{0,5})$	29.8 (-28.0)	11.7 (9.0)	17.7 (-13.2)	8.9 (-0.8)	8.2 (3.2)	9.5 (5.4)	7.9 (1.0)	8.9 (-2.1)
	$F(Y_{0,75})$	15.4 (-13.7)	5.2 (-0.5)	5.1 (-0.4)	5.0 (-0.1)	5.4 (-1.5)	5.1 (-0.4)	5.7 (-1.9)	5.3 (-1.1)
	S_y^2	12.6 (-5.1)	24.9 (-22.9)	16.8 (-11.1)	10.9 (0.3)	34.6 (32.7)	13.4 (-8.8)	47.1 (44.5)	13.6 (3.3)
0.5	$\overline{Y}$	15.9 (15.3)	5.6 (4.1)	6.0 (4.2)	3.2 (0.1)	6.5 (-5.0)	3.3 (0.5)	6.6 (-4.9)	3.6 (0.5)
	$Y_{0,25}$	24.9 (21.6)	21.3 (20.4)	12.3 (6.8)	3.9 (-1.3)	21.1 (-19.0)	4.7 (-0.7)	17.8 (-15.7)	6.0 (-0.9)
	$Y_{0,5}$	23.8 (22.0)	12.9 (-12.1)	7.4 (3.8)	5.6 (-1.5)	10.8 (-8.6)	10.1 (-8.3)	8.4 (-5.6)	5.3 (-1.0)
	$Y_{0,75}$	8.9 (8.4)	5.3 (0.8)	3.9 (0.3)	3.6 (-0.6)	4.2 (0.6)	3.2 (-0.1)	3.9 (1.3)	3.3 (0.7)
	$F(Y_{0,25})$	46.9 (-43.1)	72.3 (-71.2)	34.9 (-22.1)	19.7 (0.2)	45.6 (41.9)	27.7 (-1.5)	39.4 (35.8)	19.9 (-0.3)
	$F(Y_{0,5})$	38.3 (-36.5)	22.0 (20.6)	20.3 (-13.3)	9.8 (-0.1)	12.1 (9.3)	13.6 (10.8)	9.8 (5.8)	8.9 (0.1)
	$F(Y_{0,75})$	24.5 (-23.2)	5.8 (-2.4)	5.5 (-1.4)	5.0 (-0.2)	5.6 (-2.1)	5.3 (-1.2)	6.2 (-3.1)	5.7 (-2.2)
	S_y^2	15.0 (-0.6)	29.0 (-26.3)	18.3 (-7.7)	12.9 (-0.3)	56.2 (54.1)	15.2 (-9.4)	76.6 (73.5)	19.5 (9.7)

Tabla 6
Valores de ECMR×100 (y SR×100) asociados a distintos métodos de imputación en la población Factories. Se han utilizado cuatro clases de imputación para tratar el mecanismo de respuesta (M1). Las muestras fueron seleccionadas bajo muestreo aleatorio simple y tamaño $n = 100$.

		Sin usar x			Usando x				
p	Parámetro	SI	Media	RHD	NNI	Razón	Reg	Razón.PE	Reg.PE
0.1	$\overline{Y}$	5.1 (3.8)	3.2 (0.6)	3.2 (0.6)	3.1 (0.0)	3.1 (-0.4)	3.1 (0.0)	3.1 (-0.4)	3.1 (0.0)
	$Y_{0,25}$	6.4 (3.8)	4.3 (-2.9)	3.7 (0.1)	3.4 (-1.2)	3.8 (-2.0)	3.8 (-1.8)	3.7 (-1.5)	3.6 (-1.0)
	$Y_{0,5}$	5.5 (3.3)	5.0 (-0.9)	4.8 (-1.0)	4.8 (-1.1)	4.8 (-1.3)	4.7 (-1.2)	4.8 (-1.0)	4.7 (-0.9)
	$Y_{0,75}$	3.5 (1.7)	3.9 (-0.6)	3.8 (-0.6)	3.9 (-0.6)	3.9 (-0.6)	3.9 (-0.6)	3.8 (-0.6)	3.8 (-0.6)
	$F(Y_{0,25})$	23.4 (-18.1)	20.4 (7.5)	18.1 (-6.8)	14.9 (0.1)	14.7 (3.7)	14.7 (2.8)	14.5 (1.3)	14.9 (-0.9)
	$F(Y_{0,5})$	13.9 (-10.0)	8.9 (-0.8)	8.9 (-0.4)	8.9 (-0.1)	8.8 (0.2)	8.9 (0.2)	8.8 (-0.3)	9.0 (-0.6)
	$F(Y_{0,75})$	6.9 (-3.8)	5.2 (-0.1)	5.2 (-0.1)	5.2 (-0.1)	5.2 (-0.1)	5.2 (-0.1)	5.2 (-0.2)	5.2 (-0.1)
	S_y^2	11.1 (-3.7)	11.4 (-4.8)	11.6 (-3.0)	10.3 (0.5)	10.3 (2.6)	10.4 (-0.8)	11.6 (4.3)	11.0 (1.1)
0.3	$\overline{Y}$	11.3 (10.7)	3.4 (1.2)	3.5 (1.2)	3.1 (0.1)	3.3 (-1.1)	3.1 (0.2)	3.4 (-1.1)	3.2 (0.2)
	$Y_{0,25}$	17.8 (15.0)	5.8 (-4.0)	5.0 (0.7)	3.6 (-1.1)	5.7 (-4.5)	4.8 (-2.5)	5.2 (-3.6)	4.4 (-0.9)
	$Y_{0,5}$	15.0 (13.1)	6.3 (1.5)	5.3 (-0.1)	4.8 (-0.9)	5.5 (-2.7)	4.2 (-1.0)	5.4 (-1.8)	4.4 (0.0)
	$Y_{0,75}$	6.0 (5.5)	3.7 (-0.4)	3.5 (-0.4)	3.5 (-0.5)	3.4 (0.0)	3.5 (-0.5)	3.3 (0.1)	3.4 (-0.3)
	$F(Y_{0,25})$	40.2 (-37.1)	43.8 (18.2)	23.5 (-9.5)	16.2 (-0.5)	20.0 (14.1)	16.1 (6.2)	17.4 (9.6)	14.6 (-0.3)
	$F(Y_{0,5})$	29.8 (-28.0)	10.4 (-5.0)	9.5 (-1.9)	9.1 (-0.7)	8.9 (3.2)	9.1 (0.2)	8.4 (1.1)	9.3 (-2.0)
	$F(Y_{0,75})$	15.5 (-13.9)	5.1 (-0.6)	5.1 (-0.4)	5.0 (-0.2)	5.2 (-1.1)	5.0 (-0.3)	5.3 (-1.2)	5.0 (-0.4)
	S_y^2	12.5 (-4.5)	13.6 (-8.2)	13.4 (-3.8)	11.2 (0.5)	13.2 (8.5)	10.8 (-1.6)	17.8 (13.3)	13.1 (3.0)
0.5	$\overline{Y}$	15.8 (15.1)	3.6 (1.0)	3.7 (1.0)	3.3 (0.0)	4.0 (-1.7)	3.2 (0.0)	4.1 (-1.8)	3.3 (0.0)
	$Y_{0,25}$	25.1 (22.0)	7.1 (-4.3)	6.2 (0.8)	3.9 (-1.3)	8.2 (-6.7)	6.0 (-3.4)	7.8 (-5.9)	5.5 (-1.4)
	$Y_{0,5}$	23.7 (21.9)	8.2 (3.2)	5.8 (-0.2)	5.6 (-1.6)	7.8 (-5.4)	4.1 (-1.1)	7.3 (-4.3)	4.7 (-0.1)
	$Y_{0,75}$	8.7 (8.2)	4.7 (-0.2)	3.9 (-0.4)	3.9 (-0.8)	4.0 (0.2)	3.8 (-0.7)	3.8 (0.3)	3.4 (-0.5)
	$F(Y_{0,25})$	47.8 (-44.1)	61.3 (21.8)	28.6 (-8.4)	18.9 (0.0)	28.7 (22.5)	18.6 (9.1)	24.6 (17.5)	16.2 (1.6)
	$F(Y_{0,5})$	38.4 (-36.6)	11.6 (-7.2)	10.1 (-1.6)	10.3 (0.1)	12.2 (7.8)	10.5 (1.0)	10.3 (5.0)	9.8 (-1.2)
	$F(Y_{0,75})$	24.2 (-22.7)	5.7 (-2.0)	5.5 (-0.5)	5.2 (0.0)	5.4 (-1.5)	5.2 (-0.5)	5.5 (-1.7)	5.2 (-0.4)
	S_y^2	15.2 (-1.3)	16.2 (-9.3)	16.1 (-2.7)	13.2 (-0.8)	19.1 (15.2)	12.5 (-2.6)	26.3 (22.2)	16.2 (4.4)

cialmente en el caso de mecanismos de respuesta no uniforme. Con el fin de reducir en la medida de lo posible los sesgos, recomendamos el uso de cuatro o más clases de imputación, si bien debemos de controlar, en cualquier caso, que no existen clases con pocas unidades, puesto que esto podría producir resultados poco satisfactorios. Los métodos NNI, razón con perturbaciones y regresión con perturbaciones se han mostrado como los más eficientes en el caso de mecanismos de respuesta no uniforme, aunque debemos realizar un análisis previo para utilizar de entre ellos el más apropiado. Por ejemplo, el método NNI resulta apropiado cuando la correlación entre las variables es muy elevada. Para la elección entre los métodos de razón y regresión tendremos que estudiar el modelo ajustado entre las variables y analizar cuál de los dos métodos se adapta mejor a dicho modelo.

APÉNDICE A. Implementación de métodos de imputación mediante R/Splus

En este apéndice se proporcionan las funciones o códigos en el entorno de los lenguajes de programación estadística R y Splus para la implementación de los métodos de imputación descritos en este trabajo. Notamos que la idea de proporcionar códigos en R/Splus fue también seguida en Wu (2005) para la implementación del reciente método de verosimilitud empírica en el contexto del muestreo en poblaciones finitas.

En primer lugar, describimos la función SEPARA.muestras que separa la muestra inicial s_n en las muestras s_r y s_m. Además, esta función devuelve otros objetos, tales como el valor de m, necesarios para el uso del resto de funciones descritas en este apéndice. Notamos que los lenguajes de programación R y Splus utilizan el comando "NA" para indicar que un determinado valor no está disponible. La función SEPARA.muestras dispone de tres argumentos, los cuales se detallan a continuación:

1. muestray: valores de la variable y en la muestra s_n que tiene datos faltantes.
2. muestrax: valores de la variable x en la muestra s_n.
3. Pi: probabilidades de inclusión de primer orden asociadas a las unidades de la muestra s_n.

```
SEPARA.muestras <- function(muestray, muestrax, Pi)
{
POS.faltantes   <-  is.na(muestray)
POS.disponibles <- !POS.faltantes
datosy.r        <- muestray[POS.disponibles]
datosx.r        <- muestrax[POS.disponibles]
Pi.r            <- Pi[POS.disponibles]
datosx.m        <- muestrax[POS.faltantes]
m               <- length(datosx.m)
list(datosy.r=datosy.r,  datosx.r=datosx.r,  Pi.r=Pi.r,
     m=m,  datosx.m=datosx.m,  POS.faltantes=POS.faltantes)
}
```

A continuación se describe la función METODO.media que permite implementar el método de imputación de la media. La función METODO.media da como salida los m donantes que sustituirán los m datos faltantes en la variable y. Esta función dispone de los siguientes argumentos, los cuales se pueden obterner a partir de la función SEPARA.muestras:

1. datosy.r: valores de la variable y en la muestra s_r.
2. Pi.r: probabilidades de inclusión de primer orden asociadas a las unidades de la muestra s_r.
3. m: número de datos faltantes en la variable y.

```
METODO.media <- function(datosy.r, Pi.r, m)
{
Pesos          <- 1/Pi.r
N.est          <- sum(Pesos)
Media          <- (1/N.est)*sum(Pesos*datosy.r)
DONANTES.media <- rep(Media,m)
DONANTES.media
}
```

El siguiente método de imputación que hemos implementado en R es el método de Cohen. Los argumentos de esta función son los mismos que los descritos para la función METODO.media.

```
METODO.cohen <- function(datosy.r, Pi.r, m)
{
r              <- length(Pi.r)
n              <- r + m
Pesos          <- 1/Pi.r
N.est          <- sum(Pesos)
Media          <- (1/N.est)*sum(Pesos*datosy.r)
Dr             <- sqrt((1/N.est)*sum(Pesos*(datosy.r-Media)^2))
m1             <- round(m/2)
m2             <- m - m1
Raiz           <- sqrt(n+r+1)/sqrt(r-1)
DONANTES.cohen <- c(rep(Media+Raiz*Dr,m1),rep(Media-Raiz*Dr,m2))
DONANTES.cohen
}
```

La función METODO.NNI da como salida los m valores imputados mediante el método de imputación NNI. Los argumentos requeridos en esta función son los siguientes:

1. datosy.r: valores de la variable y en la muestra s_r.
2. datosx.r: valores de la variable x en la muestra s_r.
3. datosx.m: valores de la variable x en la muestra s_m.
4. m: número de datos faltantes en la variable y.

```
METODO.NNI <- function(datosy.r, datosx.r, datosx.m, m)
{
DONANTES.NNI  <- c()
for (j in 1:m)
  {
  Diferencias <- abs(datosx.m[j] - datosx.r)
  Dif.min     <- min(Diferencias)
  POS.min     <- Dif.min==Diferencias
  DONANTES    <- datosy.r[POS.min]
  Num.T       <- sum(POS.min)
  if (Num.T==1) DONANTES.NNI  <- c(DONANTES.NNI, DONANTES)
  else          DONANTES.NNI  <- c(DONANTES.NNI, sample(DONANTES,1))
  }
```

```
DONANTES.NNI
}
```

Imputaciones mediante los métodos de la razón y regresión pueden realizarse, respectivamente, mediante las funciones METODO.razon y METODO.regresion. Por otra parte, con las funciones METODO.razon.aleatorio y METODO.regresion.aleatorio se obtienen los donantes mediante los métodos de razón y regresión que añaden perturbaciones aleatorias. Los argumentos de estas funciones ya han sido descritos en funciones anteriores.

```
METODO.razon <- function(datosy.r, datosx.r, Pi.r, datosx.m)
{
Pesos          <- 1/Pi.r
N.est          <- sum(Pesos)
MediaY         <- (1/N.est)*sum(Pesos*datosy.r)
MediaX         <- (1/N.est)*sum(Pesos*datosx.r)
DONANTES.razon <- (MediaY/MediaX)*datosx.m
DONANTES.razon
}

METODO.regresion <- function(datosy.r, datosx.r, Pi.r, datosx.m)
{
Pesos          <- 1/Pi.r
N.est          <- sum(Pesos)
MediaY         <- (1/N.est)*sum(Pesos*datosy.r)
MediaX         <- (1/N.est)*sum(Pesos*datosx.r)
Beta.est       <- sum(Pesos*(datosx.r-MediaX)*(datosy.r-MediaY))/sum(Pesos*(datosx.r-MediaX)^2)
DONANTES.reg   <- (MediaY/MediaX)*datosx.m
DONANTES.reg
}

METODO.razon.aleatorio <- function(datosy.r, datosx.r, Pi.r, datosx.m)
{
Pesos           <- 1/Pi.r
N.est           <- sum(Pesos)
Media           <- (1/N.est)*sum(Pesos*datosy.r)
Desviacion      <- sqrt((1/N.est)*sum(Pesos*(datosy.r-Media)^2))
DONANTES.razonA <- METODO.razon(datosy.r, datosx.r, Pi.r, datosx.m) + rnorm(m,0,Desviacion)
DONANTES.razonA
}

METODO.regresion.aleatorio <- function(datosy.r, datosx.r, Pi.r, datosx.m)
{
Pesos          <- 1/Pi.r
N.est          <- sum(Pesos)
Media          <- (1/N.est)*sum(Pesos*datosy.r)
Desviacion     <- sqrt((1/N.est)*sum(Pesos*(datosy.r-Media)^2))
DONANTES.regA  <- METODO.regresion(datosy.r, datosx.r, Pi.r, datosx.m) + rnorm(m,0,Desviacion)
DONANTES.regA
}
```

La función METODO.RHD implementa el método de imputación RHD. Los argumentos de esta función coinciden con los argumentos de las funciones METODO.media y METODO.cohen.

```
METODO.RHD <- function(datosy.r, Pi.r, m)
{
Pesos        <- 1/Pi.r
Prob         <- Pesos/sum(Pesos)
DONANTES.RHD <- sample(datosy.r, m,replace=T, prob=Prob)
DONANTES.RHD
}
```

Por último, describimos la función REALIZA.imputacion, que devuelve todos los valores de la variable y en la muestra s_n; es decir, esta función asigna los valores devueltos por cada una de las funciones de imputación descritas a las posiciones donde se ha producido la no respuesta.

```
REALIZA.imputacion  <- function(muestray, DONANTES, POS.faltantes)
{
muestray[POS.faltantes]  <- DONANTES
muestray
}
```

APÉNDICE B. Ejemplo para el uso de las funciones descritas

En este apéndice se presenta un ejemplo que describe cómo utilizar las funciones descritas en el Apéndice A, dadas una muestra con unidades faltantes y las probabilidades de inclusión de cada una las unidades de la mencionada muestra. Por simplicidad, la muestra utilizada en este ejemplo tiene tamaño $n = 10$ y ha sido seleccionada mediante muestreo aleatorio simple de la población Factories descrita en la Sección 3. Los datos seleccionados fueron los siguientes:

Muestra de y : 7152, 5630, 6752, 6660, 4762, 3821, 7416, 5562, 6567, 5286.

Muestra de x : 563, 211, 425, 443, 185, 97, 705, 198, 375, 160.

Además, a partir de las $n = 10$ unidades que componen la muestra, se seleccionaron aleatoriamente $m = 3$ unidades, las cuales jugarn el papel de datos faltantes. Las unidades seleccionadas como datos faltantes fueron las que ocupan las posiciones 2, 4 y 7; es decir, los valores 5630, 6660 y 7416 serán tratados como datos faltantes. Las instrucciones que podemos seguir para imputar estos datos faltantes mediante el método NNI, por ejemplo, son las siguientes:

```
## Introducimos los datos muestrales de ambas variables, incluyendo los datos faltantes:
muestray <- c(7152, NA, 6752, NA, 4762,  3821, NA, 5562, 6567, 5286)
muestrax <- c(563, 211, 425, 443, 185, 97, 705, 198, 375, 160)

## Introducimos las probabilidades de inclusión:
Pi <- rep(1/10,10)

## Separamos la muestra s_n y obtenemos el resto de información necesaria:
SALIDA <- SEPARA.muestras(muestray, muestrax, Pi)

## Aplicamos el método NNI para obtener los donantes:
DONANTES.NNI <- METODO.NNI(SALIDA$datosy.r, SALIDA$datosx.r, SALIDA$datosx.m, SALIDA$m)

## Reemplazamos los datos faltantes por los donantes obtenidos:
REALIZA.imputacion(muestray, DONANTES.NNI, SALIDA$POS.faltantes)
```

Los valores devueltos por la última instrucción son

7152, 5562, 6752, 6752, 4762, 3821, 7152, 5562, 6567, 5286.

Estas cantidades corresponden a los valores muestrales de y después de sustituir los datos faltantes por sus correspondientes imputaciones.

Todas las funciones descritas en los Apéndices A y B están diseñadas para el caso de un diseño muestral general sin clases de imputación. En la presencia de clases de imputación, el uso de las funciones es bastante simple. En este caso tendremos que utilizar las mencionadas funciones dentro de cada una de las clases de imputación. No obstante, se puede solicitar información a los autores sobre las funciones descritas, u otras que sean requeridas, mediante correo electrónico.

Referencias

[1] Arcos, A., Gámiz, M.L., González, A., Martínez, M.D. y Rueda, M.M. (2004). *Programación en R. Aplicaciones al muestreo*. Ed. Los autores. ISBN: 84-609-3077-7. Depósito legal: GR-1880-2004.

[2] Arcos, A., Gámiz, M.L., González, A., Martínez, M.D., Muñoz, J.F., Román, Y. y Rueda, M.M. (2005). *Estadística Computacional con SPSS y R*. Ed. Los autores. ISBN: 84-689-5347-4. Depósito legal: GR-2110-2005.

[3] Bello, A.L. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Comunication in Statistics*, **22** 823–877.

[4] Berger, Y.G. y Rao, J.N.K. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, **68** 531–547.

[5] Berger, Y.G. y Skinner, C.J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series B*, **52** 457–468.

[6] Brick, J.M. y Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, **5** 215–238.

[7] Chambers, R.L. y Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73** 597–604.

[8] Chen, J. y Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, **16** 113–131.

[9] Cohen, M.P. (1996). A new approach to imputation. *American Statistical Association Proceding of the Section on Survey Research Methods* 293–298.

[10] Everitt, B.S. (1994). *A handbook of Statistical Analysis using S-Plus*. Chapman and Hall, New York.

[11] Fay, R.E. (1991). A design-based perspective on missing data variance. In Proc. Seventh Annual Res. Conf., Washington, D.C.: U.S. Bureau of the Census. 429-440.

[12] Hu, M., Salvucci, S. y Lee, R. (2001). *A Study of Imputation Algorithms*. Working Paper No. 200117. Washington DC: U.S. Department of Education, National Center for Education Statistics, 2001. 27 Stata Statistical Software.

[13] Healy, M.J.R. y Westmacott, M. (1956). Missing values in experments analysed on automatic computers. *Appled Statistics*, **5** 203–206.

[14] Ihaka, R. y Gentleman, R. (1996). R: a Language for Data Analysis and Graphics. *Journal of Computational and Grtaphical Statistics*, **5** 299–314.

[15] Kalton, G. (1983). *Compensating for missing data*. Ann Arbor: Institute for Social Research, University of Michigan.

[16] Kalton, G. y Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology* **12** 1–16.

[17] Krause, A. y Olson, M. (2005). *The basic of S-Plus. Fourth Edition*. Springer.

[18] Kuk, A.Y.C. y Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, **51** 261-269.

[19] Little, R.J.A. y Rubin, D.B. (2002). *Statistical analysis with missing data. 2^{nd} edition.* New York: John Wiley & Sons, Inc.

[20] Merino, M. y Vadillo, F. (2007). Matemática financiera con MATLAB©. *Métodos Cuantitativos para la Economía y la Empresa*, **4** 35–55.

[21] Murthy, M.N. (1967). *Sampling theory and method*. Calcutta: Statistical Publishing Society.

[22] Rancourt, E., Lee, H. y Särndal, C.E. (1994). Bias correction for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, **20** 137–147.

[23] Rao, J.N.K. (1996). On variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association*, **91** 499–520.

[24] Rao, J.N.K., Kovar, J.G. y Mantel, H.J. (1990). On estimating distribution function and quantiles from survey data using auxiliary information. *Biometrika*, **77** 365–375.

[25] Rao, J.N.K. y Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot-Deck Imputation. *Biometrika*, **79** 811–822.

[26] Rubin, D.B. (1978). Multiple imputations in sample surveys. A phenomenological bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association. 20–34.

[27] Rubin, D.B. (1996). Mutiple imputation after 18+ years. *Journal of the American Statistical Association*, **91** 473–489.

[28] Särndal, C.E., Swensson, B. y Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[29] Sedransk, J. (1985). The objetive and practice of imputation. In *Proc. First Annual Res. Conf.*, Washington, D.C.: Bureau of the Cencus. 445–452.

[30] Silva P.L.D. y Skinner C.J. (1995). Estimating distribution function with auxiliary information using poststratification. Journal of Official Statistics, **11** 277–294.

[31] Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, **88** 89–96.

[32] Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood methods in survey sampling. *Survey Methodology*, **31** 239–243.