Badler, Clara
Alsina, Sara
Puigsubirá, Cristina
Vitelleschi, Ma. Susana

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

IMPUTACIÓN MÚLTIPLE CON SAS® PARA ESTIMACIONES A PARTIR DE BASES DE DATOS CON INFORMACIÓN FALTANTE

1. INTRODUCCIÓN

La técnica de imputación múltiple constituye un aporte metodológico importante en el tratamiento de la información faltante, ya que permite incorporar un error aleatorio debido al proceso de imputación en la inferencia estadística, pudiendo ser aplicado a todo tipo de datos y en cualquier análisis estadístico.

La disponibilidad de los programas computacionales para implementarla no sólo facilita la construcción de los conjuntos completos a partir de las varias imputaciones, sino que además permite evaluar la variabilidad que la imputación incorpora en la estimación a través de los distintos métodos de imputación disponibles y del número de replicaciones.

Este trabajo tiene el objetivo de ilustrar la operatividad de la aplicación de la técnica de imputación múltiple para realizar estimaciones a partir de bases de datos con información faltante a través de los procedimientos específicos del programa SAS y la posibilidad de evaluar inmediatamente el efecto de las distintas opciones en los resultados, a través de la eficiencia de las estimaciones.

2. IMPUTACIÓN MÚLTIPLE

La técnica de imputación múltiple reemplaza cada valor faltante por un conjunto de posibles valores que representa la incertidumbre del verdadero valor a imputar. a partir de los varios conjuntos “completos” se puede aplicar el análisis estadístico estándar elegido y sus resultados son combinados para la estimación de los parámetros propuestos.

La misma comprende tres fases:

- **Imputación**: los valores perdidos son imputados m veces, generando m conjuntos de datos “completos”.

· **Análisis**: cada uno de los m conjuntos de datos "completos" es analizado a través de métodos estadísticos estándares.

· **Combinación**: los resultados obtenidos a partir de cada uno de los m conjuntos son combinados para realizar la inferencia.

El paso de imputación es el más crítico ya que depende de supuestos relativos a las características del mecanismo de las pérdidas para el modelo de imputación a aplicar.

La técnica de imputación múltiple puede ser aplicada a todo tipo de datos, en forma previa a cualquier análisis estadístico y permite incorporar información auxiliar en el modelo de imputación. Trabaja bajo el supuesto de que los datos son perdidos al azar (MAR), o sea que la probabilidad de que una observación está perdida depende de los valores observados pero no de los valores perdidos de la unidad correspondiente.

2.1. INFERENCIA A PARTIR DE LA APLICACIÓN DE IMPUTACIÓN MÚLTIPLE

Es de interés realizar estimaciones de los parámetros y obtener una medida de la variabilidad asociada a partir de los m conjuntos de datos.

Sea $\hat{\theta}_j \ (j = 1,..., m)$ el estimador del parámetro θ y su variancia asociada U_j (j = 1,...,m), calculados en cada uno de los m conjuntos completos. El estimador de θ a través de los m conjuntos se define como:

$$\bar{\hat{\theta}}_m = \sum_{j=1}^{m} \frac{\hat{\theta}j}{m}$$

Y la variancia asociada a dicho estimador como:

$$\hat{T}_m = \bar{\hat{U}}_m + (1 + m^{-1})\hat{B}_m$$

donde:

$$\bar{\hat{U}}_m = \sum_{j=1}^{m} \frac{\hat{U}_j}{m}$$

mide la variabilidad dentro de las imputaciones, y

$$\hat{B}_m = \sum_{j=1}^{m} (\hat{\theta}j - \bar{\hat{\theta}}_m)(\hat{\theta}j - \bar{\hat{\theta}}_m)' \Big/ (m-1)$$

refleja la variabilidad entre imputaciones.

Además, $(1+m^{-1})$ es el factor de ajuste por trabajar con un número finito de imputaciones.

Si el parámetro es un escalar la estimación por intervalos y los tests de hipótesis se basan en una distribución t-Student:

$$(\mathbf{q}-\hat{\bar{\mathbf{q}}}_m)\,\hat{T}_m^{-1/2} \sim t_v$$

donde los grados de libertad se calculan como:

$$v=(m-1)\left\{1+\left[\left(1+m^{-1}\right)\hat{B}_m \Big/ \hat{\bar{U}}_m\right]^{-1}\right\}^2$$

el cual está basado en la aproximación de Satterthwaite.

Cuando los grados de libertad del conjunto de datos completo (v_0) y la proporción de datos perdidos son pequeños, el cálculo de los grados de libertad del conjunto de datos "completos", v_m, puede resultar mucho más grande que v_0, lo cual es inapropiado. Barnard y Rubin recomiendan el uso de :

$$v_m^* = \left[\frac{1}{v_m}+\frac{1}{v_{\hat{obs}}}\right]^{-1}$$

donde:

$$v_{\hat{obs}} = \frac{v_0+1}{v_0+3}\,v_0\,(1-\mathbf{g})$$

y

$$\mathbf{g}=\frac{(1+m^{-1})\,B}{T}$$

Cuando el parámetro de interés es un vector de p componentes se utiliza la contrapartida multivariada de las expresiones previas.

Estas expresiones incorporan la variabilidad de las imputaciones y proveen estimadores consistentes de los parámetros y sus errores estándares, bajo el supuesto de que el modelo de imputación sea el correcto.

2.2 EFICIENCIA DE LA IMPUTACIÓN MÚLTIPLE

Rubin propone una medida de la eficiencia de la estimación basada en m imputaciones a través de la siguiente expresión:

$$\left(1+\frac{\hat{\lambda}}{m}\right)^{-1}$$

donde:

$$\hat{\lambda}=\frac{r+2/(v_m+3)}{r+1}$$

es la fracción de información faltante para la cantidad que está siendo estimada y cuantifica cuánto más precisa podría haber sido la estimación si no hubieran habido pérdidas, y

$$r=\frac{\left(1+m^{-1}\right)\hat{B}_m}{\hat{\bar{U}}_m}$$

es el incremento relativo en la variancia debido a la no respuesta, que se anula cuando no existe información perdida. Tanto $\hat{\lambda}$ como r son medidas utilizadas para diagnóstico ya que evalúan el grado de influencia de la información faltante en la estimación del parámetro.

La tabla 1 permite observar los valores de la eficiencia ante combinaciones de m y $\hat{\lambda}$.

Tabla 1: Eficiencia de la estimación a través de imputación múltiple según el número de imputaciones (m) y la fracción de información perdida ($\hat{\lambda}$)

m	$\hat{\lambda}$				
	0.10	0.20	0.30	0.40	0.70
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Se observa que en la mayoría de las situaciones no se justifica trabajar con un gran número de conjuntos imputados ya que la ganancia en eficiencia disminuye luego de los primeros valores de m.

Puede obtenerse una mayor eficiencia incrementando el número de imputaciones si los grados de libertad (v_m) son muy pequeños (menos de 10), en caso contrario no se logrará

ninguna ganancia en la precisión de las estimaciones haciendo crecer m. Dado que v_m depende de m y de r, si el valor de m es grande o el de r es pequeño, los grados de libertad serán grandes y la distribución se aproximará a la normal. Cuando hay mayor influencia de $\hat{B}_m$ con respecto a $\hat{\bar{U}}_m$, los grados de libertad se acercan al valor mínimo de (m-1), pero cuando ocurre lo inverso los mismos tienden a infinito.

3. IMPUTACIÓN MÚLTIPLE EN SAS

La implementación de esta técnica se ha visto facilitada varios años después de su original propuesta por el tardío desarrollo de distintas herramientas computacionales, constituyendo una activa área de investigación y produciendo un incremento notable de su utilización. Las mismas son de fácil acceso y se encuentran disponibles en diferentes publicaciones o comercializadas, facilitando su aplicación y la rápida evaluación de las distintas propuestas.

Entre los programas de análisis estadístico, SAS/STAT en su Versión 8.1, presenta el procedimiento PROC MI que permite obtener valores para las variables no observados en la base de datos multivariada incompleta, mediante la utilización de distintos métodos, creando m conjuntos imputados. Luego que cada uno de los m conjuntos de datos “completos” son analizados con métodos estadísticos estándares, otro procedimiento, PROC MIANALYZE, puede ser usado para generar inferencias válidas de los parámetros de interés, combinando los resultados a partir de los m conjuntos de datos “completos”.

3.1. PROC MI

El PROC MI posibilita la realización de imputaciones a través de tres métodos. Para esquemas monótonos, el método paramétrico de regresión, bajo el supuesto de distribución normal multivariada o el “propensity scores” para el caso no paramétrico. Para esquemas arbitrarios, dispone del método “Markov Chain Monte Carlo” (MCMC), bajo el supuesto de distribución normal multivariada.

Otra alternativa para el manejo de datos con un esquema arbitrario de pérdida es usar el método MCMC para imputar valores de tal forma de obtener un esquema monótono y así poder utilizar métodos de imputación más flexibles.

En este procedimiento están disponibles las siguientes sentencias:

```
PROC MI <options>;
BY variables;
```

FREQ variable;
MULTINORMAL <options>;
VAR variables;

La sentencia PROC MI es la única requerida en este procedimiento. En ella hay distintas opciones disponibles:

- Número de imputaciones (NIMPU = número), por defecto cinco imputaciones.

- Creación de un archivo de salida (OUT = SAS-data-set), en el cual están los resultados de las imputaciones. El conjunto de datos incluye una variable de imputación, _IMPUTATION_ para identificar el número de la imputación.

- Especificación de un entero positivo (SEED = número) para el comienzo del número generador seudo aleatorio. Por defecto toma la hora indicada en el reloj de la computadora. Este valor es importante tenerlo en cuenta si se quieren reproducir en el futuro los mismos resultados.

- Opciones que permiten obtener valores imputados consistentes con los valores observados. La primera de ellas se refiere a acotar los valores para imputar, dándoles un valor máximo y un valor mínimo (MAXIMUM = números; MINIMUN = números). La segunda especifica las unidades de redondeo para los valores obtenidos para imputar (ROUND = números). Si hay varias variables a imputar y se especifica un solo número, éste se utiliza para todas ellas. En caso contrario se debe usar la sentencia VAR y especificar en ella los números que corresponden a cada variable.

- La sentencia BY que permite obtener el análisis por separado sobre las observacioines en grupos definidos por las variables especificadas.
- La sentencia FREQ que permite especificar el nombre de las variables que en el archivo de datos de entrada representen frecuencias de ocurrencia para otros valores en la observación.

- La sentencia VAR que lista las variables numéricas a ser utilizadas; si es omitida todas las variables numéricas que no estén en otra sentencia serán utilizadas.

- En la sentencia MULTINORMAL se dispone de los siguientes métodos para imputar:

METHOD=REGRESSION;

METHOD=PROPENSITY;

METHOD=MCMC;

Si ninguno de estos métodos es especificado, por defecto se utiliza el MCMC.

En la sentencia METHOD= PROPENSITY <(NGROUPS = número)>, se especifica el número de grupos basados en el propensity score; por defecto es 5.

En la sentencia METHOD=MCMC <(options)>, la opción NIMPU = número, permite requerir la cantidad de imputaciones; por defecto el procedimiento utiliza cadenas múltiples para crear cinco imputaciones. También por defecto el procedimiento utiliza las estadísticas a partir de casos disponibles como estimadores iniciales para el algoritmo EM. La tabla de valores iniciales muestra los valores para los promedios y covariancias utilizadas en cada imputación. Otras opciones disponibles para este procedimiento son:

- CHAIN=SINGLE/MÚLTIPLE, para especificar si se usa una cadena para todas las imputaciones o una para cada una;
- INITIAL=EM<(BOOSTRAP<=p>)>

  INITIAL=INPUT= SAS-data-set

  La sentencia INITIAL= EM usa, en el PROC MI, los promedios y las desviaciones

estándares de los casos disponibles como estimadores iniciales para el algoritmo EM . Las correlaciones son fijadas iguales a cero y los estimadores resultantes son usados para comenzar el MCMC.

- NBITER = número, especifica el número de iteraciones “burn” antes de la primera imputación en cada cadena, por defecto su valor es 50.
- NITER = número, especifica el número de iteraciones entre imputaciones en una cadena simple, por defecto su valor es 30.

Aunque el PROC MI para regresión o MCMC supone distribución normal multivariada, las inferencias por imputación múltiple podrían ser robustas para desviaciones de la normalidad si no es grande la cantidad de información perdida.

Para cada método requerido se genera una salida con información sobre el método y las opciones utilizadas en el procedimiento de imputación; la descripción del esquema de pérdida con sus correspondientes frecuencias, porcentajes y promedios de cada variables por grupo de unidades igualmente observados.

Mediante el PROC PRINT del archivo de salida se pueden visualizar los valores imputados para cada variable.

3.1.1. MÉTODO DE IMPUTACIÓN POR REGRESIÓN

Se ajusta un modelo de regresión para cada variable con valores faltantes, utilizando como covariables las variables previas en el esquema monótono. Basándose en el modelo resultante, se simula un nuevo modelo de regresión, el cual es utilizado para imputar los valores faltantes de cada variable. Como el conjunto de datos presenta un esquema monótono de pérdida, el proceso se repite secuencialmente para las variables con valores faltantes.

Para una variable Y_j con valores faltantes, el modelo:

$$E(Y_j) = \beta_o + \beta_1 Y_1 + \beta_2 Y_2 + ...+ \beta_{(j-1)} Y_{(j-1)}$$

se ajusta con sólo los valores observados.

Para cada imputación, los nuevos parámetros (β_{*0}, β_{*1}, ..., $\beta_{*(j-1)}$) son simulados a través de (β_0, β_1, ..., $\beta_{(j-1)}$). Los valores perdidos son reemplazados por:

$$\beta_{*o} + \beta_{*1} y_1 + \beta_{*2} y_2 + ... + \beta_{*(j-1)} y_{(j-1)} + z_i \sigma_{*J}$$

donde y_1, y_2, ..., $y_{(j-1)}$ son los valores de las covariables de las primeras (j-1) variables y z_i es el valor resultante de la simulación de un desvío normal.

3.1.2. METODO “PROPENSITY SCORE”

El “propensity score” es la probabilidad condicional de asignación de un tratamiento particular dado un vector de covariables observables.

En éste método es generado un “propensity score” para cada variable con información faltante para estimar la probabilidad que la observación ha sido perdida. Luego, las observaciones son agrupadas en relación a este “propensity score” y se aplica para cada grupo una imputación con aproximación bootstrap bayesiana.

Bajo un esquema monótono el método sigue los siguientes pasos para imputar los valores faltantes de cada variable Y_j :

1. Crea una variable indicadora R_j que toma el valor cero si existe algún dato faltante en la variable Y_j y 1 en otro caso.

2. Ajusta un modelo de regresión logit:

$$\text{logit } (\pi_j) = \beta_o + \beta_1 Y_1 + \beta_2 Y_2 + ...+ \beta_{(j-1)} Y_{(j-1)}$$

donde $\pi_j = Pr(R_j =0/ Y_1,Y_2, ...,Y_{(j-1)})$ y $\text{logit } (\pi) = \log(\pi/(1-\pi))$.

3. Crea un "propensity score" para cada observación para estimar la probabilidad de que ella esté perdida.

4. Divide las observaciones en un número fijo de grupos basado en estos "propensity score".

5. Aplica para cada grupo una imputación con aproximación bootstrap bayesiana. En el grupo k-ésimo, se denota con Y_{obs} a las n_1 observaciones sin valores faltantes para Y_j y con Y_{per} a las n_0 observaciones con información faltante. La imputación con aproximación bootstrap bayesiana primero extrae aleatoriamente con reemplazo las n_1 observaciones a partir de Y_{obs} para crear el nuevo conjunto de datos Y^*_{obs}. Este es un procedimiento no paramétrico análogo al de extraer los parámetros a partir de la distribución a posteriori predictiva de los datos perdidos. El proceso, luego, extrae aleatoriamente con reemplazo los n_0 valores de Y_{per} a partir de Y^*_{obs}.

Se repite el proceso secuencialmente para cada variable con información faltante. Este método es efectivo para inferencias a partir de distribuciones individuales de variables imputadas, tal como un análisis univariado pero no lo es para análisis que involucran relaciones entre variables, tal como un análisis de covariancia.

3.1.3. METODO MCMC

El mismo consta de una colección de técnicas para generar extracciones seudo-aleatorias a partir de distribuciones multidimensionales.

En este método se construye una cadena de Markov lo suficientemente extensa para que la distribución de los elementos se estabilicen en una distribución común. Esta distribución estacionaria es la de interés.

MCMC ha sido aplicado como un método para explorar las distribuciones a posteriori en la inferencia bayesiana, es decir, se simula la distribución conjunta a posteriori de los parámetros desconocidos y se obtienen estimaciones, basadas en simulaciones de los parámetros a posteriori, que son de interés.

Asumiendo que los datos siguen una distribución normal multivariada el método consta de la repetición de dos pasos:

1. Imputación: a partir de la estimación del vector de promedios y la matriz de covariancias se simulan, independientemente, valores perdidos para cada una de las observaciones. Es decir, si se denota con $Y_{i(per)}$ las variables con valores faltantes y con

$Y_{i(obs)}$ las variables con valores observados para la unidad i-ésima, en este paso se extraen los valores para $Y_{i(per)}$ a partir de la distribución condicional de $Y_{i(per)}$ dada $Y_{i(obs)}$.

2. A posteriori: se simulan el vector de promedios y la matriz de covariancias poblacionales a posteriori a partir de la muestra "completa". Estos nuevos estimadores son usados en el paso 1. Se puede usar como información a priori, por ejemplo, la matriz de covariancias ya que es útil para estabilizar la inferencia acerca del vector de promedios para una matriz de covariancias singular.

Estos dos pasos son iterados hasta lograr resultados confiables para un conjunto de datos imputados en forma múltiple. El objetivo es que las iteraciones converjan a una distribución estacionaria y luego simular una extracción aproximadamente independiente de los valores faltantes.

3.2. PROCEDIMIENTO MIANALYZE

El PROC MIANALYZE combina los resultados del análisis estadístico aplicado en cada uno de los m conjuntos "completos" creados por imputación, para generar una única estimación de los parámetros, mediante las siguientes sentencias:

```
PROC MIANALYZE <opciones>
BY variables;
VAR variables;
```

Este procedimiento sólo requiere para funcionar su nombre y el de los parámetros cuyas estimaciones deberá combinar para obtener una única estimación. Estos últimos se deben especificar a través de la sentencia VAR donde se detallan los nombres que los parámetros han recibido en el archivo de salida con estructura SAS de las m replicaciones del análisis estadístico requerido.

Si no se especifica un archivo de entrada, el programa utilizará el conjunto SAS creado más recientemente. Para especificarlo se pueden utilizar distintas opciones, que deben estar de acuerdo con la forma en que el programa archiva los resultados del análisis estadístico previamente aplicado a los m conjuntos "completos", las cuales son:

- DATA = SAS-data-set, nombra una estructura SAS de datos para ser analizada a través del PROC MIANALYZE. El archivo de datos de entrada debe tener un TYPE de COV, CORR o EST. Las estimaciones de los parámetros y sus matrices de covariancias asociadas, de cada conjunto imputado, son leídos de un único archivo ingresado.

- PARMS= SAS-data-set y COVB= SAS-data-set, nombran y proveen respectivamente las estimaciones de los parámetros y la matriz de covariancias asociada en conjuntos separados;
- PARMS= SAS-data-set y XPXI= SAS-data-set, donde el primero nombra y provee las estimaciones de los parámetros y los errores estándares asociados y el segundo las estimaciones de los parámetros y las matrices inversas X'X asociadas. A partir de ellos el procedimiento calcula las matrices de covariancias.

Si se ha realizado un análisis de regresión mediante un PROC REG y se ha creado un OUTEST = , el cual contiene las estimaciones de los parámetros y la matriz de covariancias, para leerlo se podría utilizar la opción DATA=.

En todos los casos citados el procedimiento brinda los mismos resultados para cada uno de los parámetros estimados, en la "Tabla de información sobre la variancia de imputación múltiple": la variancia intra ($\hat{\bar{U}}_m$), entre ($\hat{B}_m$) y Total ($\hat{T}_m$), los grados de libertad para la variancia total (v), el incremento relativo en la variancia (r) y la fracción de información faltante ($\hat{\lambda}$), y en la "Tabla de Estimación de Parámetros con imputación múltiple", la estimación de cada parámetro ($\hat{\bar{\theta}}_m$), su desvío estándar, los límites del intervalo de confianza del 95%, los grados de libertad y los resultados del test t con la probabilidad asociada a la hipótesis de igualdad del parámetro a mu0 según el valor especificado en la opción MU0=números, para cada variable.

A través de las opciones del PROC MIANALYZE se puede, además, especificar:

- ALPHA= p, el nivel de significación si debe ser distinto de 0.05.
- EDF= números, los grados de libertad para la estimación del parámetro a partir de los datos completos, con el objeto de calcular grados de libertad ajustados, según la expresión de Barnard y Rubin.
- MU0= números, los promedios bajo la hipótesis nula en el test t, acompañado de una sentencia VAR para asignarlos a cada variable; en el caso de un sólo número se asigna a todas por igual.
- MULT o MULTIVARIATE, para requerir las medidas de la inferencia multivariada para las variables en conjunto.

4. APLICACIÓN

Se trata de reproducir el proceder de un usuario o analista que debe aplicar un determinado análisis estadístico a una base de datos con información faltante y elige la técnica de imputación múltiple para su tratamiento, ya que es una de las más recomendadas a través de la bibliografía actualizada.

Al mismo se le presentan distintas posibilidades para la aplicación de esta técnica a través del programa SAS. Ante los requerimientos de las sentencias, deberá tomar decisiones sobre el método de imputación a aplicar y las opciones que cada uno ofrece, aunque sea por defecto. Para su aplicación se requiere el conocimiento de la distribución conjunta de las variables y la forma del esquema de pérdida.

Aunque algunos de los métodos de imputación a aplicar se basen en el supuesto de distribución normal multivariada de las variables y la técnica de imputación múltiple supone que la característica del mecanismo de pérdida es MAR, al no verificarse estos supuestos, probablemente se trabaje apartándose de los mismos. De todas maneras a partir de experiencias de distintos autores la precisión de las estimaciones no siempre resulta afectada en estas situaciones.

4.1. ANÁLISIS DE REGRESIÓN A PARTIR DE IMPUTACIÓN MÚLTIPLE CON SAS

Se presenta una aplicación a partir de una sub-base de datos provenientes de la onda octubre 2001 de la Encuesta Permanente de Hogares para el Aglomerado Gran Rosario.

El análisis propuesto consiste en la estimación del siguiente modelo de regresión:

$$p47t = \beta_o + \beta_1\, p15t + \beta_2\, p22t + \varepsilon$$

donde:

p47t: ”Ingreso total”

p15t: “Total de horas trabajadas más horas extras en la semana de referencia”

p22t: “Cuanto tiempo hace que está en esa ocupación”

En un enfoque experimental, se simulan pérdidas en la variable p22t, en un porcentaje cercano al 50% con el objeto de evidenciar su efecto en las estimaciones, bajo el mecanismo MAR dependiendo si la variable p15t es menor que 37.

Mediante el PROC MI y PROC MIANALYZE de SAS se aplica la técnica de imputación múltiple a través de los tres métodos disponibles.

Se presenta la estructura general del programa utilizado para uno de los métodos.

```
proc mi data=mar2002 out=miout nimpute=20 seed=1234;
multinormal method=mcmc;
var p47t p15t p22t;
proc print data=miout;
run;
proc reg data=miout outest=outreg covout;
model p47t= p15t p22t;
by _Imputation_;run;proc print data=outreg;run;
proc print data=outreg(obs=8);
var _Imputation_ _Type_ _Name_
Intercept p15t p22t;run;
proc mianalyze data=outreg mult;
var Intercept p15t p22t;run;
```

En cuanto a la cantidad de imputaciones solicitadas en cada procedimiento (NIMPU=), cada método fue aplicado con distintos valores de m a partir de m=3. De esta manera, se dispone de una herramienta práctica para elegir el valor definitivo de m, a través de la observación de la estabilización del valor de las estimaciones. De todas maneras, en las Tablas 2 y 3 se incluye la información para m=5 y m=20 como forma de visualizar la utilización de una medida de la eficiencia en las estimaciones, también utilizada para la elección definitiva de m.

En las Tablas 2 y 3 se presentan para cada método de imputación las estimaciones de los coeficientes de regresión correspondiente a cada variable del modelo, sus desvíos, $\hat{\lambda}$, r y la medida de eficiencia como elementos para una evaluación de los resultados.

Tabla 2: Estimación de los coeficientes de regresión y eficiencia asociada según método de imputación (m=5)

Método de Imputación	Variables	Estimación de los coeficientes de regresión	r	$\hat{\lambda}$	Eficiencia de la estimación
REGRESSION	p15t	6,473 (0,725)	0.064	0.062	0.987
	p22t	0.948 (0.147)	0.095	0.090	0.982
PROPENSITY SCORE	p15t	5.600 (0.726)	0.005	0.005	0.999
	p22t	1.040 (0.150)	0.051	0.050	0.990
MCMC	p15t	7.292 (0.752)	0.149	0.137	0.973
	p22t	0.790 (0.193)	1.706	0.687	0.879

Tabla 3: Estimación de los coeficientes de regresión y eficiencia asociada según método de imputación (m=20)

Método de Imputación	Variables	Estimación de los coeficientes de regresión	r	$\hat{\lambda}$	Eficiencia de la estimación
REGRESIÓN	p15t	6,400 (0,726)	0.062	0.059	0.997
	p22t	0.952 (0.145)	0.069	0.065	0.996
PROPENSITY SCORE	p15t	5.605 (0.726)	0.007	0.007	0.999
	p22t	0.151 (0.137)	0.076	0.071	0.996
MCMC	p15t	7.160 (0.774)	0.222	0.184	0.990
	p22t	0.828 (0.149)	0.563	0.368	0.981

Se observa que, bajo un mecanismo de pérdida MAR y para todos los métodos de imputación aplicados, la eficiencia de la estimación es alta para m=5, manteniéndose constante para casi todos los casos cuando se aumenta el número de imputaciones a m=20. No se justificaría trabajar con un gran número de conjuntos imputados dado que al ser los grados de libertad todos mayores que 10 y el incremento relativo de la variancia (r) muy bajo no se lograría ninguna ganancia en la precisión.

5. DISCUSIÓN

Si el usuario o analista debe aplicar un análisis estadístico sobre bases de datos con información faltante, que le requieren la estimación de distintos parámetros, dispone de una herramienta operativa en el programa SAS para incorporar los valores desconocidos mediante la aplicación de la técnica de imputación múltiple. El procedimiento proporciona distintos métodos de imputación, dependiendo su elección del comportamiento de las variables y del esquema de pérdida y, además, brinda medidas para evaluar la eficiencia de las estimaciones.

A pesar que algunos métodos de imputación que proporciona el programa SAS se basan en el supuesto de la distribución conjunta normal de las variables y la técnica de imputación múltiple supone que la característica del mecanismo de pérdida es MAR, en la experiencia de distintos autores, la precisión de las estimaciones no siempre es afectada ante la no verificación de estos supuestos.

El usuario deberá profundizar la evaluación de los resultados a través del análisis de las medidas estadísticas asociadas a la estimación.

Bibliografía

Allison, P.D.. (2000). “Multiple imputation for missing data: A cautionary tale”. Sociological Methods and Research, vol. 28, pp. 301-309.

Barnard, J and Rubin D.. (1999). “Small-sample degrees of freedom with multiple imputation”. Biometrika, vol. 86, pp. 948-955.

Horton, N. and Lipsitz, S.. (2001). ”Multiple imputation in practice: comparison of software packages for regression models with missing variables”. American Statistician Association, vol.55, Nº 3.

Rubin, D.. (1987). “*Multiple imputation for nonresponse in surveys*”. John Wiley & Sons.

Rubin, D. B. (1996). “Multiple imputation after 18+ years”. Journal of the American Statistical Association, vol. 91, Nº 434, pp. 473-489.

SAS® Institute Inc., Statistics and Operations Research. “What´s new in data analysis. multiple imputation for missing data”, 2002. http://www.sas.com/rnd/app/da/new/dami.htm , (Mayo 2002)

Schaffer, J. (1997). *“Analysis of incomplete multivariate data”*. Chapman and Hall.

Rubin, D. and Schenker, J.. (2000). “Multiple imputation for missing data problems”. Course for the Joint Statistical Meeting. Dallas 1998. http:// www. stat. psu. edu/ ~jls/aug98.pdf , (Mayo 2002)

Yuan,Y.C..(2001).“Multiple imputation for missing data: concepts and new development”. SUGI Proceedings. http://www.ats.ucla.edu/stat/sas/library/ default.htm, (Octubre 2001).