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

Badler, C.; Alsina, S.;F.; Puigsubirá, C.; Vitelleschi, M.S.

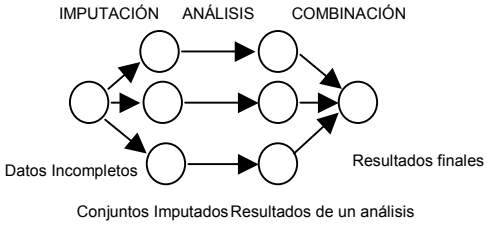
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística.Universidad Nacional de Rosario.

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.

IMPUTACIÓN MÚLTIPLE

- Reemplaza cada valor faltante por un conjunto de posibles valores que respresenta la incertidumbre del verdadero valor a imputar.
- Aplicable a todo tipo de datos y para cualquier análisis estadístico estándar.
- Trabaja bajo el supuesto MAR.



INFERENCIA A PARTIR DE IMPUTACIÓN MÚLTIPLE

$\hat{\theta}_j \ (j = 1,\ldots,m)$ estimador del parámetro θ

$\bar{\hat{\theta}}_m = \sum_{j=1}^{m} \frac{\hat{\theta}_j}{m}$ estimador de θ a través de los m conjuntos

$\hat{T}_m = \hat{\bar{U}}_m + (1+m^{-1})\hat{B}_m$ variancia asociada a dicho estimador , donde:

$\hat{\bar{U}}_m = \sum_{j=1}^{m} \frac{\hat{U}_j}{m}$ mide la variabilidad dentro de las imputaciones y

$\hat{B}_m = \sum_{j=1}^{m} (\hat{\theta}_j - \bar{\hat{\theta}}_m)(\hat{\theta}_j - \bar{\hat{\theta}}_m)' / (m-1)$ refleja la variabilidad entre imputaciones

$(1+m^{-1})$ 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: $(\theta - \bar{\hat{\theta}}_m)\hat{T}_m^{-1/2} \sim t_v$
donde los grados de libertad se basan en la aproximación de Satterthwaite:

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

Cuando los grados de libertad del conjunto de datos completo (v_0) y la proporción de datos perdidos son pequeños, Barnard y Rubin proponen alternativa de cálculo.

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.

Eficiencia a partir de Imputación Múltiple

Rubin propone una medida de la eficienciade la estimación basada en m imputaciones: $\left(1 + \frac{\hat{\lambda}}{m}\right)^{-1}$ 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.

$r = \frac{(1+m^{-1})\hat{B}_m}{\hat{\bar{U}}_m}$ es el incremento relativo en la variancia debido a la no respuesta, que se anula cuando no existe información perdida.

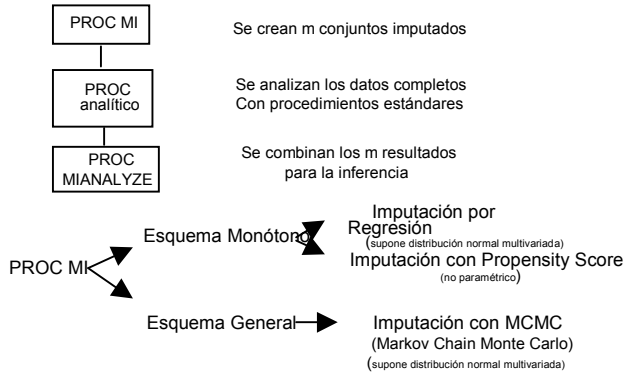
Tanto $\hat{\lambda}$ como r son medidas utilizadas para diagnóstico ya que evalúan el grado de influencia de la información faltante en la estimación del parámetro.

Tabla 1: Eficiencia de la estimación a través de imputación múltiple según el número de imputaciones (m) y la fracción de información perdida ( )

m	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

La ganancia en eficiencia disminuye luego de los primeros valores de m.

IMPUTACIÓN MÚLTIPLE EN SAS (V.8.1)



APLICACIÓN

• Material: sub-base de datos provenientes de la onda octubre 2001 de la Encuesta Permanente de Hogares para el Aglomerado Gran Rosario. Variables:

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"

Pérdidas simuladas en la variable p22t, en porcentaje aproximado a 50% con el objeto de evidenciar su efecto en las estimaciones, bajo mecanismo MAR (p15t < 37)

• Análisis propuesto: estimación del modelo de regresión $p47t = \beta_o + \beta_1\ p15t + \beta_2\ p22t + \varepsilon$

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

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;
```

Cada método fue aplicado con distintos valores de m a partir de m=3, disponiéndose así también de de una herramienta práctica para elegir el valor definitivo de m, a través de la estabilización en el valor de las estimaciones.

Se obtienen las estimaciones de los coeficientes de regresión de cada variable del modelo y sus desvíos a partir de cada método de imputación. Se les asocia r , y la medida de la eficiencia como elementos para una primera evaluación de los resultados.

Tabla 2: Estimación de los coeficientes de regresión y eficiencia asociada según método de imputación y cantidad de imputaciones

Método de Imputación	Variables	Estimación de los coeficientes de regresión		r				Eficiencia de la estimación	
		m=5	m=20	m=5	m=20	m=5	m=20	m=5	m=20
REGRESSION	p15t	0.473 (0.725)	6.400 (0.726)	0.064	0.062	0.062	0.059	0.987	0.997
	p22t	0.948 (0.147)	0.952 (0.145)	0.095	0.069	0.090	0.065	0.982	0.996
PROPENSITY SCORE	p15t	5.600 (0.726)	5.605 (0.726)	0.005	0.007	0.005	0.007	0.999	0.999
	p22t	1.040 (0.150)	0.151 (0.137)	0.051	0.076	0.050	0.071	0.990	0.996
MCMC	p15t	7.292 (0.752)	7.160 (0.774)	0.149	0.222	0.137	0.184	0.973	0.990
	p22t	0.790 (0.193)	0.828 (0.149)	1.706	0.563	0.687	0.368	0.879	0.981

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.

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.
• 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.