Badler, Clara E.
Alsina, Sara M. [1]
Puigsubirá, Cristina B. [1]
Vitelleschi, María S.[1]
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE)

TRATAMIENTO DE BASES DE DATOS CON INFORMACIÓN FALTANTE SEGÚN ANÁLISIS DE LAS PÉRDIDAS CON SPSS

INTRODUCCIÓN

El análisis estadístico de bases de datos provenientes de experimentos y de estudios observacionales resulta afectado cuando se presenta falta de información. La precisión de los resultados está condicionada a la proporción de unidades con pérdidas en una o más variables y a las características del mecanismo que las produce.

Actualmente, este problema es considerado fundamental para la inferencia dada su presencia permanente, hecho que continúa motivando la aparición de diferentes tratamientos para su solución. Como la elección de los mismos debe realizarse teniendo en cuenta el comportamiento de las pérdidas, el tipo de variables afectadas y el procedimiento de análisis que se desea aplicar, resulta de utilidad la simulación de situaciones que abarquen los distintos mecanismos.

Los programas de análisis estadístico no siempre contemplan la posibilidad de aplicar los procedimientos estadísticos incorporando las unidades con pérdidas, sino que suprimen los casos incompletos, afectando el análisis. Algunos presentan rutinas para realizar un tratamiento a la información incompleta previo a la aplicación de métodos estadísticos clásicos.

En este trabajo se presenta un módulo para el análisis de bases de datos con información faltante del software SPSS, programa ampliamente utilizado por usuarios de distintas disciplinas, con una aplicación a datos de la onda mayo 2003 de la Encuesta Permanente de Hogares (EPH) correspondiente al Aglomerado Gran Rosario, en el que se simulan pérdidas en algunas variables.

MATERIAL

* Datos básicos:

Las variables en estudio pertenecen a la onda mayo 2003 de la EPH, Aglomerado Gran Rosario y corresponden a individuos que en dicha encuesta declararon tener ingreso no nulo:

- Monto ingresos totales (P47T)

---

[1] Docente-investigador e Investigador del Consejo de Investigaciones de la Universidad Nacional de Rosario.

- Años cumplidos (H12)
- Cantidad de ocupaciones (P12)
- Estado ocupacional (ESTADO)
- Sexo (H13)
- Tipo de establecimiento (P18B)
- Asistencia escolar (P55)

Se trabaja con el logaritmo neperiano de la variable "Monto ingresos totales" para intentar corregir la asimetría que presenta su distribución (lnP47T).

* Soporte informático:

Se utiliza el módulo "Análisis de Datos Perdidos" del programa SPSS (v 9.0) y el programa Statistical Analysis System (SAS) (v. 8.1)

MÉTODOS

Simulación

Se simulan pérdidas en una variable cuantitativa y en otra cualitativa, según dos mecanismos de pérdida: perdidos completamente al azar y perdidos no al azar.

Mecanismos de pérdida

El proceso que produce o conduce a la pérdida de información en un relevamiento o experimento es denominado mecanismo de pérdida.

Es importante el acercamiento al conocimiento del mismo ya que cualquier análisis de datos depende de los supuestos sobre el mecanismo de pérdida, el cual debe ser explicitado.

La información incompleta en una variable puede presentarse en forma aleatoria, ligada a valores correspondientes a otra variable relacionada con la que presenta pérdidas o en categorías de valores de la propia variable, determinando en este último caso que los valores no observados sean diferentes a los observados. Dichos mecanismos de pérdida a partir de las características de la probabilidad de respuesta se pueden clasificar:

- Los datos están perdidos completamente al azar (MCAR): si la probabilidad de respuesta es independiente de las variables observadas y de las no observadas completamente. El mecanismo de pérdida es ignorable tanto para inferencias basadas en muestreo como en máxima verosimilitud.

- Los datos están perdidos al azar (MAR): si la probabilidad de respuesta es independiente de las variables no observadas completamente y no de las observadas. El mecanismo de pérdida es ignorable para inferencias basadas en máxima verosimilitud.

- Los datos no están perdidos al azar (MNAR): si la probabilidad de respuesta no es independiente de las variables no observadas completamente y posiblemente, también, de las observadas El mecanismo de pérdida es no ignorable.

Un acercamiento a la identificación del mecanismo de pérdida puede lograrse a partir de:

- Un análisis descriptivo univariado y multivariado de las variables completas y parcialmente observadas.

- El uso del test de Little para evaluar el supuesto MCAR.

Análisis de Datos Incompletos con SPSS

Este programa dispone de un módulo específico para el análisis y el tratamiento de la información incompleta, que ejecuta tres funciones:

- Describe esquemas de pérdida.
- Imputa los datos faltantes con valores estimados a través del método de regresión y el algoritmo EM. Permite disponer de las bases "completas".
- Estima algunos parámetros a partir de la aplicación de casos completos, casos disponibles, regresión y algoritmo EM.

Casos completos

Este procedimiento consiste en usar solamente las unidades que tienen información completa en todas las variables. Es simple y permite comparar estadísticas univariadas pero presenta la limitación que si los valores perdidos de una variable son los más altos y los más bajos, se distorsionan las distribuciones marginales de todas las variables y son sesgadas las estimaciones de los parámetros.

Casos disponibles

Dicho método consiste en incluir todos los casos que son observados en cada variable. Se presenta el problema que el tamaño de muestra varía de variable a variable de acuerdo al esquema de datos perdidos. Por lo tanto, los promedios y las variancias se calculan para los casos disponibles en cada variable y para las covariancias y correlaciones en base a todos los casos que no presenten datos faltantes para el par de variables implicado.

Regresión

En este método los valores faltantes son estimados por regresión lineal múltiple, debiéndose especificar las variables predictoras y la dependiente a considerar en el proceso; presenta opciones para las estimaciones con componentes aleatorias.Cada valor estimado por regresión es:

$$x_{ij}^{\mathrm{Re}} = x_{ij} \; si \; x_{ij} \; no \; es \; perdido$$

$$x_{ij}^{\mathrm{Re}} = x_{ij} \; estimado \; por \; regresion, si \; x_{ij} \; es \; perdido$$

Se puede disponer de la base de datos con los valores imputados por este método.

Algoritmo EM

Es un proceso iterativo que consiste en un paso E y un paso M y que permite encontrar los estimadores máximo verosímiles de los parámetros de interés. Consiste en:

- reemplazar los valores perdidos por los valores estimados;
- estimar los parámetros;
- re-estimar los valores perdidos asumiendo que son correctas las nuevas estimacio-

nes de los parámetros;

- re-estimar los parámetros y así sucesivamente seguir iterando hasta la convergencia.

En el paso E se calcula la esperanza condicional de los datos faltantes dados los datos observados y la estimación de los parámetros, luego estas esperanzas sustituyen a los datos faltantes.

El paso M realiza la estimación máximo-verosímil del parámetro de interés como si no existieran datos faltantes.

Este algoritmo converge confiablemente, su convergencia puede ser lenta cuando existe una gran proporción de datos faltantes. Para su aplicación requiere el cumplimiento del supuesto MAR.

El programa SPSS presenta tres opciones para especificar el supuesto de la distribución de la variable con falta de información; por defecto se supone que tiene un comportamiento normal. Se puede disponer de la base de datos con los valores imputados por este método.

Cuadros de resultados del módulo

La aplicación del módulo Análisis de Datos Perdidos permite obtener los siguientes cuadros, con variaciones según las opciones especificadas en cada uno:

Estadísticas Univariadas: de cada variable solicitada se obtiene el tamaño muestral, promedio y desvíos de variables cuantitativas, valores extremos, valores faltantes. Permite un primer análisis descriptivo, una apreciación de la magnitud de las pérdidas y la posible consideración de los valores extremos como información confusa.

Esquema de pérdida con todos los individuos de la base o sólo aquellos con pérdidas: permite apreciar la ubicación de las pérdidas según individuos y variables y la presencia de grupos de individuos o variables afectadas.

Patrones tabulados: se especifica la cantidad de individuos en los que al menos una variable no fue observada agrupándolos según las mismas; el número de individuos sin pérdidas, el número de individuos que se obtiene si la/s variable/s detallada/s que presenta/n pérdida son eliminadas; los promedios de las variables cuantitativas en cada uno de los casos detallados y las frecuencias de las variables categóricas solicitadas.Permite analizar cómo incide en los promedios y en las categorías el uso de sólo casos completos o el ir agregando individuos con pérdidas en una o más variables.

Porcentaje de discordancia en las variables indicadoras: SPSS crea internamente una variable indicadora de pérdidas. En función de la misma, en una matriz en cuyas filas y columnas se representan las mismas variables, en la diagonal principal se observa el porcentaje de pérdida de cada variable individual y fuera de la misma el porcentaje de individuos en los que se presenta la pérdida en una u otra variable pero no en ambas.

Pruebas t con variancias separadas: una forma de chequear si los valores faltantes de una variable son perdidos completamente al azar es a través de un test t-Student para dos muestras. Se compara para cada variable cuantitativa sin pérdida, los promedios de los grupos definidos por la variable indicadora (presente o perdido). Se detalla el valor de la estadística t, los grados de libertad, la cantidad de datos con valores perdidos y observados y los promedios de ambos grupos. De esta manera se obtiene una forma de acercamiento al mecanismo de pérdida MCAR.

Tablas de contingencia de la variable indicadora frente a las variables categóri-

cas: para cada variable categórica se obtiene una tabla en la que, para cada una de sus categorías se especifica la frecuencia y el % de valores no faltantes y faltantes. Se trabaja sobre el total de individuos sin pérdida en la variable que presenta pérdidas.

Estadísticas según lista (casos completos): proporciona el promedio, desvío, covariancias y correlaciones de las variables cuantitativas sobre los individuos sin pérdidas en cada una.

Estadísticas según pareja (casos disponibles): proporciona la cantidad de individuos para cada par de variables, promedio, desvío, covariancias y correlaciones de las variables cuantitativas cuando está presente la otra variable. Permite visualizar la frecuencia en que los pares de variables faltan en forma conjunta y la incidencia en los promedios y desvíos de una variable cuando son suprimidos los individuos incompletamente observados.

Estadísticas estimadas por EM: se obtienen las estimaciones del promedio, covariancias y correlaciones a partir de la aplicación del algoritmo EM según la especificación realizada sobre el supuesto de distribución de la variable afectada y el número máximo de iteraciones, y los resultados de la prueba de Little para evaluar el supuesto MCAR.

Estadísticas estimadas por regresión: se obtienen las estimaciones del promedio, covariancias y correlaciones utilizando regresión lineal múltiple según la opción realizada para el ajuste de la estimación mediante la incorporación de una componte aleatoria. Se puede establecer el número máximo de variables predictoras.

Resumen de las medias y desvíos típicos estimados: brinda las estimaciones de los promedios y desvíos estándar de las variables solicitadas a partir de los métodos aplicados; las opciones cubre casos completos (según lista), casos disponibles (todos los valores), EM y regresión.

RESULTADOS

Se presentan algunos resultados siguiendo una secuencia de análisis similar al adoptado por un usuario que debe aplicar un análisis estadístico a partir de una base de datos con información faltante y obtener una estimación del ingreso promedio de los individuos que declaran ingreso no nulo.

Las pérdidas han sido generadas en las variables “Monto ingresos totales“ y “Estado ocupacional “ en un porcentaje aproximado al 25% de los individuos, a partir de dos mecanismos de pérdida: MCAR (se generan las pérdidas en ambas variables con una semilla aleatoria fijando el porcentaje requerido de pérdida); MNAR (se generan pérdidas en forma aleatoria en los valores altos de la variable “Monto ingresos totales” y en la categoría ocupados de la variable “Estado ocupacional”).

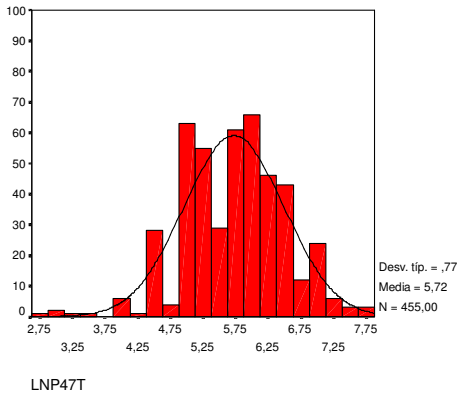
La rutina de análisis se reitera en cada una de las bases así obtenidas, con el propósito de observar la incidencia de los distintos mecanismos en la estructura de los datos y en las estimaciones.

Base de datos con pérdidas generadas según mecanismo MCAR

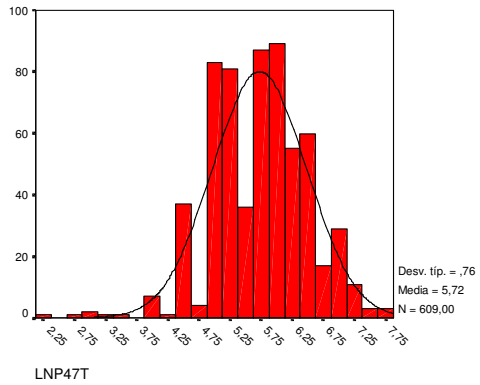
A través de la construcción del histograma para la variable LNP47T de la base de datos con pérdidas generadas según el mecanismo MCAR y de la original, se observa que la primera constituye una submuestra aleatoria de la original.

Gráfico 1: Histograma de la variable LNP47T en las bases con pérdidas MCAR y en la original

MCAR



DATOS ORIGINALES



En el Cuadro 1 se observa que los tamaños muestrales correspondientes a cada variable varían según la cantidad de individuos afectados por las pérdidas en cada una de ellas: 608 individuos integran la base, 455 registran valores en lnP47T y 456 en ESTADO, sobre ellos son calculadas las estadísticas. La cantidad de valores perdidos afectan en un 25,2% y en 25%, respectivamente, a las variables lnP47T y ESTADO:

Cuadro 1

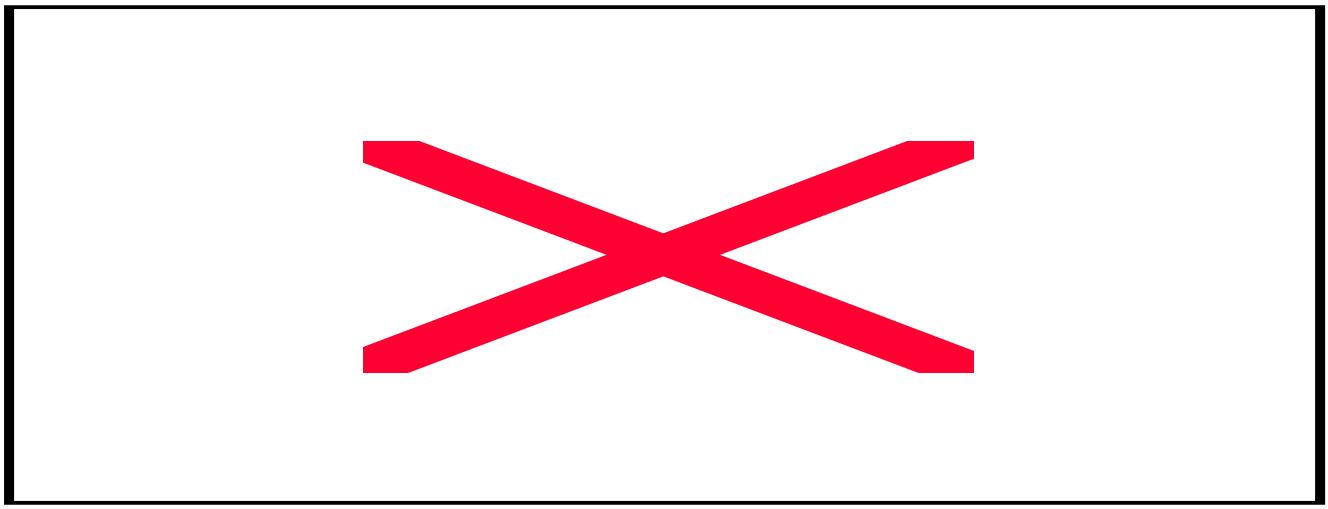
Estadísticos univariados

	N	Media	Desviación típ.	Perdidos		Nº de extremos [a]	
				Recuento	Porcentaje	Bajos	Altos
LNP47T	455	5,7208	,7666	153	25,2	5	3
P12	608	,7747	,5232	0	,0	0	4
H12	608	46,1776	18,7107	0	,0	0	0
H13	608			0	,0		
ESTADO	456			152	25,0		
P18B	608			0	,0		
P55	608			0	,0		

a. Número de casos fuera del rango (C1 - 1.5*AIC, C3 + 1.5*AIC).

En el Cuadro 2 se puede observar, por ejemplo, que el promedio de edad es más alto para el conjunto de individuos con pérdida en ambas variables (31 individuos) y más bajo para el conjunto con pérdidas sólo en la variable LNP47T (122 individuos), ambas comparadas con el promedio a partir de los casos completos (334 individuos). Resulta de utilidad la consideración de la columna b del Cuadro 2, en la que se especifica el número de casos completos si las variables con pérdidas en ese patrón (marcadas con X) no se considera: son 334 individuos los que presentan información completa en todas las variables, si se incorporan los que presentan faltas sólo en la variable ESTADO, suman 455 y así sucesivamente:

Cuadro 2



A partir de la consideración de los grupos formados por los valores de la variable indicadora (creada por el programa), asociada a los individuos con y sin pérdidas en la variable LNP47T y en la variable ESTADO, no se rechaza la igualdad de los promedios de la edad en ambos grupos para las dos variables, constituyendo un primer acercamiento a la determinación de un mecanismo de pérdida MCAR (Cuadro 3):

Cuadro 3

Pruebas T con varianzas separadas[a]

		LNP47T	P12	H12
LNP47T	t	,	,6	,0
	gl	,	245,6	254,0
	nº presente	455	455	455
	nº perdido	0	153	153
	Media(Presentes)	5,7208	,7824	46,1604
	Media(Perdidos)	,	,7516	46,2288
ESTADO	t	-1,2	,1	-,5
	gl	233,4	260,5	269,4
	nº presente	334	456	456
	nº perdido	121	152	152
	Media(Presentes)	5,6952	,7763	45,9737
	Media(Perdidos)	5,7914	,7697	46,7895

Para cada variable cuantitativa, los pares de grupos están formados por variables indicador (presente, perdido).

a. Las variables indicador con menos del 5% de los valores perdidos no se muestran.

Al aplicar el de test de Little no se rechaza la hipótesis que el mecanismo de pérdida en

ambas variables es MCAR (χ^2=.0506, gl=2 p>0.05).

Se obtiene la estimación del promedio del monto del ingreso total y los respectivos desvíos a partir de la aplicación de casos completos (según lista), casos disponibles (todos los valores), regresión y algoritmo EM (Cuadro 4):

Cuadro 4

Resumen de las medias estimadas

	LNP47T	P12	H12
Según lista	5,7208	,7824	46,1604
Todos los valores	5,7208	,7747	46,1776
EM	5,7169	,7747	46,1776
Regresión	5,7356	,7830	46,1892

Resumen de las desviaciones típicas estimadas

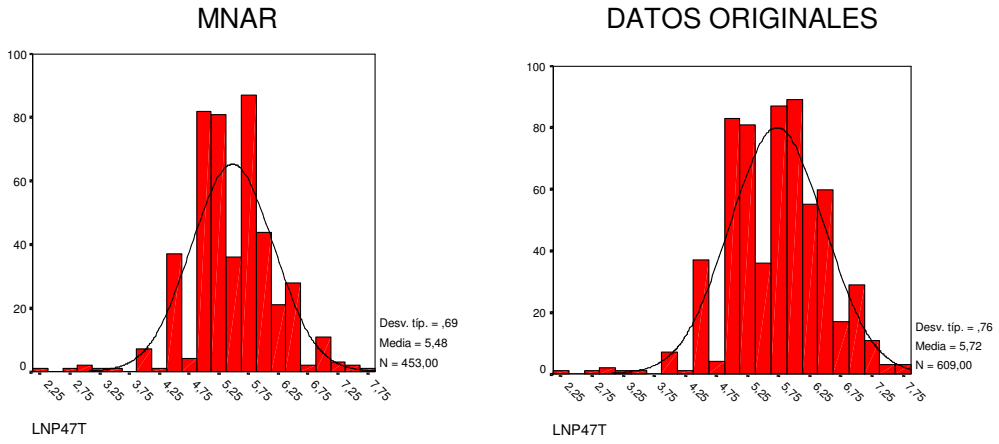
	LNP47T	P12	H12
Según lista	,7666	,5130	18,5605
Todos los valores	,7666	,5232	18,7107
EM	,7702	,5232	18,7107
Regresión	,7551	,5139	18,4778

Comparando los valores de las estimaciones a partir de los cuatro métodos no se observan diferencias notables. Este hecho llevaría a pensar que los individuos con pérdidas no presentarían características muy diferentes a los completamente observados.

Base de datos con pérdidas generadas según mecanismo MNAR

A través de la construcción del histograma para la variable LNP47T de la base de datos con pérdidas generadas según el mecanismo MNAR y de la original, se observa que la primera no constituye una submuestra aleatoria de la original.

Gráfico 2: Histograma de la variable LNP47T en las bases con pérdidas MNAR y en la original



En la base con pérdidas generadas con el mecanismo MNAR el porcentaje de individuos con falta de información en LNP47T es de 25,5 % y de 25 % en ESTADO, con 453 y 456 individuos respectivamente.

El promedio de edad es más bajo para el conjunto de individuos con pérdida en ambas variables (67 individuos) y más alto para el conjunto con pérdidas sólo en la variable LNP47T (88 individuos), ambas comparadas con el promedio a partir de los casos completos (368 individuos). Ello debería ser tenido en cuenta al utilizar las distintas variables en los modelos de imputación (Cuadro 5):

Cuadro 5

Patrones tabulados

Número de caso	Patrones perdidos[a]							Completo si...[b]	H12[c]	P12[c]	ESTADO[d]			P18B[d]					P55[d]			H13[d]	
	H12	P12	P18B	P55	H13	ESTADO	LNP47T				1,00	2,00	3,00	,00	1,00	2,00	3,00	9,00	1,00	2,00	3,00	1,00	2,00
368								368	45,8278	,7351	236	17	115	132	64	170	1	1	24	331	13	181	187
85						X		453	39,3412	1,0235	0	0	0	0	21	63	1	0	5	77	3	38	47
67						X	X	608	41,4179	1,1194	0	0	0	0	12	55	0	0	2	65	0	44	23
88							X	456	51,6023	,7159	57	1	30	31	7	49	1	0	3	85	0	58	30

Los patrones con menos del 1% casos (6 o menos) no se muestran.

a. Las variables se ordenan según los patrones perdidos.

b. Número de casos completos si las variables perdidas en ese patrón (marcado con X) no se utilizan.

c. Medias en cada patrón único

d. Distribución de frecuencia en cada patrón único

Al aplicar el test de Little condujo al rechazo de la hipótesis que el mecanismo de pérdida en ambas variables es MCAR (χ^2=18,165; gl=2 p<0.05).

Se obtiene la estimación del promedio del monto del ingreso total y los respectivos desvíos a partir de la aplicación de casos completos (según lista), casos disponibles (todos los valores), regresión y algoritmo EM (Cuadro 6):

Cuadro 6

Resumen de las medias estimadas

	LNP47T	H12	P12
Según lista	5,4802	45,8278	,7351
Todos los valores	5,4802	46,1776	,7747
EM	5,5026	46,1776	,7747
Regresión	5,4571	44,8030	,7807

Resumen de las desviaciones típicas estimadas

	LNP47T	H12	P12
Según lista	,6909	19,5703	,5071
Todos los valores	,6909	18,7107	,5232
EM	,6932	18,7107	,5232
Regresión	,6585	18,9189	,4808

Comparando los valores de las estimaciones a partir de los cuatro métodos se observan diferencias, con un desvío mayor para el método EM.

DISCUSIÓN

El uso del programa SPSS constituye una herramienta eficiente en la elección y aplicación de un tratamiento para la información faltante, presente en bases de datos resultantes de proyectos multivariados planeados para analizar fenómenos reales de las distintas disciplinas técnicas y científicas.

Dado que dicha elección está condicionada a la característica del mecanismo que produjo la pérdida, a la proporción de las unidades que las contienen y a los objetivos que se plantean, el SPSS posee, en el módulo correspondiente, los elementos que permiten lograrla.

La simulación realizada en este trabajo posibilitó visualizar estas características bajo la óptica de dos diferentes escenarios.

REFERENCIAS BIBLIOGRÁFICAS

Badler,C.; Alsina,S.; Beltrán,C.;Puigsubirá, C.; Vitelleschi, M..(2000). "Simulación de pérdida de información generada por distintos mecanismos en datos provenientes de la Encuesta Permanente de Hogares, para la evaluación del supuesto MCAR". Cuadernos del IITAE Nº 7. Escuela de Estadística. Fac. Ciencias Económicas y Estadística. UNR.

Badler, C.; Alsina, S.; Puigsubirá, C.; Vitelleschi, M. S. (2002). "Imputación con SAS® para estimaciones a partir de bases de datos con información faltante".

www.fcecon/unr.edu.ar/scyt/jor/jor2002. Abril 2003.

Little , R.; Rubin, D. (1987) "Statistical Analysis with Missing Data". J. Wiley & Sons.

Little, R. J. (1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values". Journal of the Royal Statistical Society. Vol. 83, N° 404.

SPSS (1997) SPSS Missing Values Análisis TM 7.5. SPSS Inc.

SSPS; http://www.spss.com, julio 2003