

Assessment of the performance of imputation techniques in observational studies with two measurements

Urko Agirre

Unidad de Investigación, Hospital Galdakao-Usansolo
urko.aguirrelarracoechea@osakidetza.net

Primeras Jornadas Científicas de Estudiantes de la SEB

Índice

1 Introducción

2 Metodología

- Perfiles de valores faltantes
- Métodos de imputación

3 Aplicación

- Diseño de estudio
- Resultados

4 Estudio de simulación

- Diseño del estudio
- Criterios de evaluación
- Resultados

5 Conclusiones

6 Referencias

Introducción

- Los estudios observacionales basados en medidas repetidas son utilizados para evaluar la evolución de la variable de resultado de interés.
- Motivación de los estudios basados en la Calidad de Vida Relacionada con la Salud (CVRS): Determinar los factores predictores del cambio medio de las variables resultado principales.
- Debido al proceso del estudio, es muy común la presencia de pérdidas de datos a lo largo del estudio. En consecuencia, los resultados derivados del mismo, pueden ser sesgados.

Metodología. Perfiles de valores faltantes (I)

- Para cada sujeto ($i = 1, \dots, n$) : $\mathbf{Y} = \{Y_{ij}\}_{i=1, \dots, n}^{j=1, \dots, n_i}$
- $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ donde \mathbf{Y}_{obs} , es el vector de valores observados y \mathbf{Y}_{miss} el vector de valores faltantes.
- Definimos \mathbf{R} , variable indicadora de presencia de valores no observados:

$$\mathbf{R} = \begin{cases} 1 & \text{si } y_{ij} \text{ es observada} \\ 0 & \text{si } y_{ij} \text{ no es observada} \end{cases}$$

- La función densidad de la probabilidad conjunta se expresa como:

$$f(\mathbf{Y}, \mathbf{R} | \theta, \phi) = f(\mathbf{Y} | \theta) f(\mathbf{R} | \mathbf{Y}, \phi)$$

donde ϕ y θ son parámetros desconocidos.

Metodología. Perfiles de valores faltantes (II)

- Según la asociación con la variable faltante:

- Missing Completely At Random (MCAR)

$$f(\mathbf{Y}, \mathbf{R}|\theta, \phi) = f(\mathbf{Y}|\theta)f(\mathbf{R}|\phi)$$

- Missing At Random (MAR)

$$f(\mathbf{Y}, \mathbf{R}|\theta, \phi) = f(\mathbf{Y}|\theta)f(\mathbf{R}|\mathbf{Y}_{obs}, \phi)$$

- Missing Not At Random (MNAR)

$$f(\mathbf{Y}, \mathbf{R}|\theta, \phi) = f(\mathbf{Y}|\theta)f(\mathbf{R}|\mathbf{Y}, \phi)$$

- Según la distribución:

- Monótona
- Intermitente

Metodología. Métodos de imputación

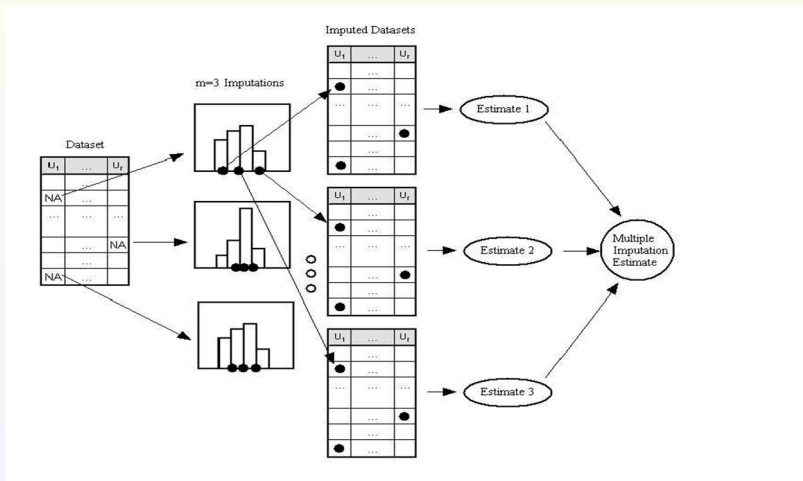
- Método más común:
 - Análisis de Datos Completos - *Complete Case* (CC).
- Alternativas al CC:
 - Análisis de Datos Disponibles - *Available Case* (AC)
 - Métodos de imputación:
 - Imputación por K vecinos más cercanos (K -NNI)
 - Markov Chain Monte Carlo (MCMC).
 - Propensity Score (PS)

Métodos de imputación.

Imputación por K vecinos más cercanos (K -NNI)

- Método basado en las similitudes entre los sujetos reclutados en el estudio.
- Considerando un conjunto de datos de valores faltantes en \mathbf{y}_i , $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, para cada variable y sujeto, y basándonos en una métrica, se seleccionan los K sujetos más cercanos a la variable.
- Una vez seleccionados los K casos, los valores no observados son reemplazados: 1) media si la variable imputada es continua; 2) moda para datos cualitativos.

Métodos de imputación: Imputación Múltiple (MI) -I



Métodos de imputación: Imputación Múltiple (MI) - II

Métodos de imputación múltiple: **MCMC y PS**

MCMC:

- Bajo un patronaje general de pérdidas de seguimiento (sean monótonas o no), el método utiliza cadenas de Markov para la imputación de los datos faltantes.

$$\theta^{(t)}, Y_{obs} \rightarrow \{Y_{miss}^{(1)}, \theta^{(1)}; Y_{miss}^2, \theta^{(2)}; \dots\}$$

- Este proceso se repite hasta que se produzca una convergencia \rightarrow el conjunto de datos no tiene valores perdidos.

Métodos de imputación: Imputación Múltiple (MI) - III

PS:Escenario Valores Faltantes.

- Se calcula el PS para la variable respuesta con observaciones faltantes (Y_i). Se halla la probabilidad de presencia de no ser observado:

$$\text{logit}(Pr(\mathbf{R}_i|\mathbf{X}_i, \beta)) = (1, \mathbf{X}_i)' \beta$$

con $\mathbf{R}_i = 1$ si Y_i no está observado, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ es el conjunto de covariables para la observación i -ésima y β , los coeficientes para la matriz de diseno \mathbf{X}_i .

- Basados en los valores del propensity score, se dividen las observaciones en un número concreto de grupos.
- A cada subgrupo, se aplica una técnica de remuestreo (*bootstrapping*) para poder realizar la imputación.

Aplicación: Calidad de Vida Relacionada con la Salud (CVRS) en pacientes Enfermedad Pulmonar Obstructiva Crónica (EPOC)

- Se ha seleccionado una submuestra de pacientes con EPOC ($n = 400$), que completaron el cuestionario correspondiente de CVRS al **inicio** y **un año** después.
- Variable resultado (**Y**): CVRS medido a través del Total Score del St. George Questionnaire (0-100). A mayor puntuación, peor estado de CVRS.
- Covariable (**X**): HADO (severidad de la enfermedad de la EPOC). Rango: 0 (peor)-12 (mejor).
- Objetivo: Determinar el efecto de la covariable **X** en el cambio de la variable resultado → Nos centramos en la estimación de la interacción entre el efecto Tiempo y la covariable **X** (Time*HADO) .

Resultados derivados del análisis de datos en la base de datos

Without imputation			AC (original analysis)		CC			
	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value		
Intercept	98.408 (6.004)	<0.001	97.271 (6.391)	<0.001				
Age	-0.266 (0.080)	0.001	-0.244 (0.086)	0.005				
HADO	-6.853 (0.349)	<0.001	-6.852(0.377)	<0.001				
Time×Age effect (after one year)	-0.022 (0.073)	0.764	-0.030(0.074)	0.682				
Time×HADO effect (after one year)	1.023 (0.321)	0.002	1.022 (0.326)	0.002				
With imputation (K-NNI)			1-NNI		5-NNI		7-NNI	
	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value
Intercept	98.408 (6.036)	<0.001	98.408 (5.948)	<0.001	98.408 (6.894)	<0.001		
Age	-0.266 (0.080)	0.001	-0.266(0.079)	0.001	-0.266(0.091)	0.004		
HADO	-6.853 (0.351)	<0.001	-6.853(0.346)	<0.001	-6.835(0.401)	<0.001		
Time×Age effect (after one year)	-0.009 (0.069)	0.895	-0.018(0.066)	0.789	-0.303(0.094)	0.001		
Time×HADO effect (after one year)	0.960 (0.301)	0.002	1.065 (0.291)	<0.001	2.388 (0.413)	<0.001		
With imputation (MI methods)			PS		MCMC			
	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value	$\hat{\beta}$ (s.e.)	p-value
Intercept	98.408 (6.299)	<0.001	98.408 (6.026)					
Age	-0.266 (0.084)	0.002	-0.266 (0.080)					
HADO	-6.853 (0.366)	<0.001	-6.853(0.350)					
Time×Age effect (after one year)	0.040(0.090)	0.660	-0.028(0.073)					
Time×HADO effect	1.793 (0.407)	<0.001	1.036 (0.320)					

CC: Complete Case. AC: Available Case.

(1,5,7)-NNI: 1,5, 7-Nearest Neighbour Imputation.

PS: Propensity score. MCMC: Markov Chain Monte Carlo.

$\hat{\beta}$ (s.e.) : Beta regression coefficient (standard error).

Models adjusted by previous hospital admissions.

Estudio de simulación: Diseño

Pasos del estudio de simulación

- **Pérdidas del seguimiento:** distintos patrones de valores ausentes (MCAR, MAR, MNAR) y porcentajes (10 % and 30 %).
- **Métodos de Imputación:** K-NN, PS and MCMC. Además de ello, se han realizado también los análisis CC and AC.
- **Modelización:** Modelos lineales mixtos para estimar el coeficiente de la interacción del efecto Tiempo y la covariable (Time*HADO).

Estudio de simulación: Criterios de evaluación

Criterios de evaluación del rendimiento de los métodos

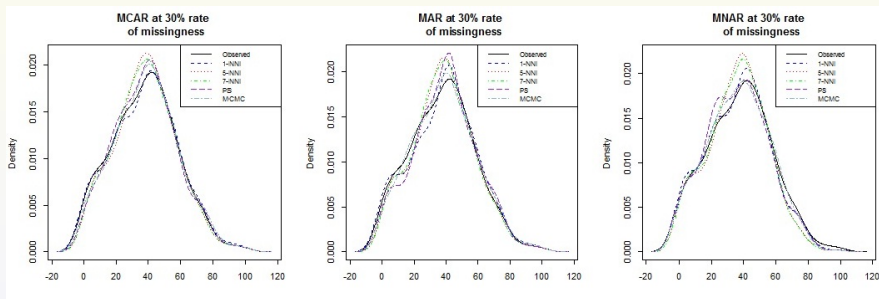
- Cálculo del sesgo relativo y estandarizado.
 - Sesgo relativo (SR): $\frac{\bar{\hat{\beta}} - \beta}{\beta}$
 - Sesgo estandarizado (SE): $\frac{\bar{\hat{\beta}} - \beta}{SE(\hat{\beta})}$
 - impacto: $|SE| > 50 \%$
- Gráficos de densidad.

Resultados (I): Sesgo estandarizado y relativo en $N=1000$ simulaciones

Missingness rate	Imputation method	MCAR		MAR		MNAR	
		SR	SE	SR	SE	SR	SE
10 %	CC	10.631	24.888	3.751	8.781	6.969	16.316
	AC	17.505	40.9824	-9.213	-21.570	4.758	11.139
	1-NNI	13.842	32.405	16.994	39.785	8.729	20.435
	5-NNI	18.727	43.842	11.676	27.334	14.561	34.089
	7-NNI	21.744	50.905	11.085	25.952	16.212	37.954
	PS	70.187	164.317	252.857	591.974	16.212	37.954
	MCMC	18.575	43.487	-3.882	-9.089	6.004	14.057
30 %	CC	14.772	34.582	18.878	44.195	27.289	63.887
	AC	22.050	51.622	21.657	50.702	25.839	60.492
	1-NNI	23.205	54.325	59.393	139.047	25.839	60.492
	5-NNI	23.146	54.187	63.539	148.753	83.273	194.954
	7-NNI	26.522	62.094	71.140	166.550	82.067	192.131
	PS	172.699	404.312	252.857	591.974	225.271	527.389
	MCMC	19.665	46.038	22.056	51.636	28.457	66.621

CC: Complete Case. AC: Available Case. (1,5,7)-NNI: 1,5, 7-Nearest Neighbour Imputation. PS: Propensity score. MCMC: Markov Chain Monte Carlo. SE: Sesgo estandarizado. SR: Sesgo relativo

Resultados (II): Funciones de densidad



Conclusiones

- **PS: NO es un método adecuado y recomendable para realizar imputaciones:**
 - A la hora de realizar los grupos, éstos no son homogéneos.
 - No contempla la asociación entre variables a la hora de imputar los valores faltantes en la variables resultado **Y**. Solamente utiliza la información de la asociación de las variables con la presencia de valores faltantes.
- Bajo MCAR, parece que no es necesario realizar imputaciones. El análisis de datos CC o AC muestran valores de sesgos más bajos.
- Cuando el patrón de valores faltantes es MAR, los resultados sugieren utilizar el método MCMC.
- En el caso MNAR, se recomienda realizar análisis de sensibilidad más exhaustivos.

Referencias

- Little, R.J.A., Rubin, D.B. (2002) Statistical analysis with missing data. New York, Ed. Wiley. Second Edition.
- Molenberghs, G., Kenward, M.G. (2007) Missing Data in Clinical studies. West Sussex, England, John Wiley Sons
- Janssen, K.J.M., Donders, A., Harrel, F., Vergouwe, Y., Chen, Q., Grobbee, D., Moons, K (2010): Missing covariate data in medical research: To impute is better than to ignore. J Clin Epidemiol. 63, 721-727.
- Esteban C., Quintana JM., Aburto M., Moraza J., Capelastegui A (2006). A simple score for assessing stable chronic obstructive pulmonary disease.Q J Med., 99, 751-759.

Mila esker!!

Muchas gracias!!