TRATAMIENTO DE MISSINGS EN ESTUDIOS DE COHORTE

TFM – Mar Harmut Prats
Máster en Bioinformática y Bioestadística (UOC)
JUNIO 2017

Tutora: Núria Pérez

INTRODUCCIÓN (I)

- La presencia en las bases de datos o de registros sin información (datos faltantes o *missings*) es muy frecuente en investigación.

- No tenerlo en cuenta puede generar situaciones no deseadas en la fase de inferencia estadística.

- La aplicación de procedimientos inapropiados de sustitución de información produce sesgos en el análisis, reduce la potencia estadística de los métodos estadísticos y le resta eficiencia a la fase de inferencia, pudiendo incluso invalidar las conclusiones de un estudio.

INTRODUCCIÓN (II)

- Algunas de las causas más comunes para la pérdida de datos:
  - Pérdida del seguimiento del paciente: por abandono del estudio, muerte del individuo, cambio de localización geográfica…
  - Fallo en los instrumentos de medida.
  - El paciente no conoce o no quiere contestar.

MÉTODOS DE IMPUTACIÓN (I)

- Los métodos de imputación se clasifican, según Little y Rubin (1987), en:
  - Análisis de datos completos (*listwise*)
  - Análisis de datos disponibles (*pairwise*)
  - Imputación por medias no condicionadas
  - Imputación por medias condicionadas mediante métodos de regresión
  - Máxima verosimilitud (MV)
  - Imputación múltiple (IM)
- Se considera que los algoritmos de máxima verosimilitud (MV) y de imputación múltiple (IM) son los más robustos.

MÉTODOS DE IMPUTACIÓN (II)

- El supuesto de que los datos faltantes siguen un patrón completamente aleatorio(*Missing Completely at Random, MCAR*) fue introducido por Rubin (1976) y es el supuesto que asume la mayoría de los métodos de imputación: la omisión no depende de los datos observados.
- Se afirma que un proceso de datos omitidos se genera en forma aleatoria (Missing at Random, MAR), si la distribución de los valores observados no depende del patrón de comportamiento de los registros sin información.
- Datos perdidos no ignorables o no debidos al azar (MNAR=missing not at random).

MÉTODOS DE IMPUTACIÓN (III)

- **Análisis de datos completos (listwise):** (MCAR) Trabajar únicamente con las observaciones que disponen de información completa para todas las variables.

- **Análisis de datos disponibles (pairwise):** (MCAR) Las observaciones que no tienen datos se eliminan, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados.

- **Imputación por medias no condicionadas:** (MCAR) La falsa creencia de que en una distribución de probabilidad normal el promedio de los datos es un buen estimador de las observaciones omitidas (atenúa la correlación con el resto de las variables y subestima la varianza)

MÉTODOS DE IMPUTACIÓN (IV)

- **Imputación por regresión:** (MCAR) Predecir el valor a través del estimador de la media condicionada.
- **Estimación por máxima verosimilitud:** (MAR) 1) Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud. 2) Utilizar los parámetros estimados para predecir los valores omitidos. 3) Sustituir los datos por las predicciones, y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa. (algoritmo Esperanza-Maximización).
- **Imputación Múltiple:** (MAR) IM utiliza métodos de simulación de Monte Carlo generando muestras aleatorias a partir de métodos bayesianos, y sustituye los datos faltantes a partir de un número (m>1) de simulaciones.

REGISTRO SORASE (I)

- Registro de Seguridad SORASE, llevado a cabo por el Grupo Internacional Ferrer, S.A., iniciado en 2013 en México.
- Registro de tipo longitudinal de una cohorte de pacientes tratada con un fármaco combinado (polypill), para la prevención secundaria de eventos cardiovasculares.
- Análisis centrado en cumplimiento de los pacientes con el tratamiento, valorado a través del cuestionario Morisky-Green.

REGISTRO SORASE (II)

- Se incluyeron en el estudio un total de 1286 pacientes, que fueron clasificados en función del riesgo cardiovascular y seguidos durante un periodo total de 12 meses.

	Basal		Mes 1		Mes 3		Mes 12	
	n	%	n	%	n	%	n	%
Con antecedente de evento CV	628	48,90%	569	48,20%	542	47,70%	464	47,00%
Alto riesgo CV	243	18,90%	226	19,20%	220	19,40%	182	18,40%
Riesgo intermedio-moderado	414	32,20%	385	32,60%	374	32,90%	341	34,50%
Total	1285	100,00%	1180	100,00%	1136	100,00%	987	100,00%

REGISTRO SORASE (III)

- Las variables utilizadas para el análisis son:

Nombre	Descripción
gr_riesg	Grupo de riesgo de ECV
edad	Edad en años
sexo	Sexo del paciente (1=Femenino; 2=Masculino)
morgreen	Puntuación Morisky-Green obtenida por el paciente en la visita basal
morgreenc	Puntuación Morisky-Green (visita basal) clasificados como: 0=no cumplidor; 1=cumplidor
morgreen1	Puntuación Morisky-Green obtenida por el paciente en la visita 1 mes
morgreenc1	Puntuación Morisky-Green (visita 1 mes) clasificados como: 0=no cumplidor; 1=cumplidor
morgreen3	Puntuación Morisky-Green obtenida por el paciente en la visita 3 meses
morgreenc3	Puntuación Morisky-Green (visita 3 meses) clasificados como: 0=no cumplidor; 1=cumplidor
morgreen12	Puntuación Morisky-Green obtenida por el paciente en la visita 12 meses
morgreenc12	Puntuación Morisky-Green (visita 12 meses) clasificados como: 0=no cumplidor; 1=cumplidor

CAUSA DE MISSINGS (I)

- Algunas de las razones por las cuales los pacientes discontinuaron el estudio fueron: razones económicas, abandono del estudio, reacciones adversas (dolor estomacal, irritación cutánea, etc.), fallecimiento del paciente, imposibilidad de re-contactar con el sujeto o el cambio de residencia.
- En los tres grupos de riesgo, la disminución de pacientes a lo largo de las visitas de seguimiento de 1 mes, 3 meses y 12 meses fue similar.

IDENTIFICACIÓN DE MISSINGS (I)

- Test de Little: p=0,382

Estadísticos univariados							
	N	Media	Desviación estándar	Perdidos		Número de extremos[a]	
				Recuento	Porcentaje	Menor	Mayor
edad	1286	57,4489	14,45310	0	,0	40	0
gr_riesg	1285	1,8335	,88531	1	,1	0	0
sexo	1286	1,53	,499	0	,0	0	0
morgreenc	1232			54	4,2		
morgreenc1	1147			139	10,8		
morgreenc3	1106			180	14,0		
morgreenc12	955			331	25,7		

IDENTIFICACIÓN DE MISSINGS (II)

- Modelo GEE:
  - cumplimiento a los 12 meses como variable dependiente
  - Covariables: sexo, grupo de riesgo y edad

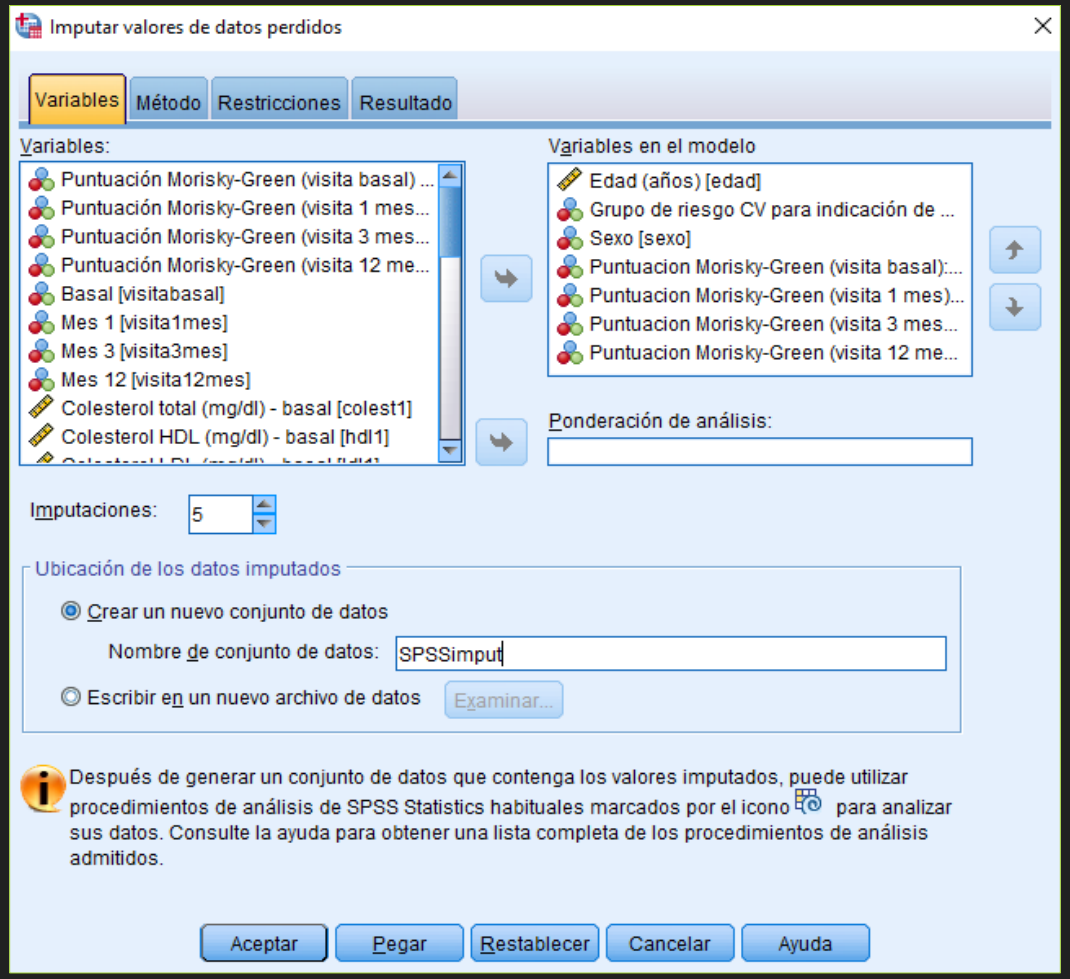
Variables en la ecuación							
		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	,830	,070	139,083	1	,000	2,293

- De acuerdo a los resultados del test de Little y del GEE, procedemos con el método de la imputación múltiple y generaremos un modelo de regresión.

IMPUTACIÓN MÚLTIPLE (I)

- El objetivo de la imputación múltiple es generar valores posibles para los valores perdidos, creando así varios conjuntos de datos "completos".
- En SPSS:
  - Analizar > Imputación múltiple > Imputar valores de datos perdidos.
  - Nº imputaciones 5.

IMPUTACIÓN MÚLTIPLE (II)



IMPUTACIÓN MÚLTIPLE (III)

- Generamos un dataset nuevo (SPSSimput) que contiene la variable “Imputation_”, la cual consiste en los números de 0 a 5, referidos a la sesión de imputación en particular (Imputación = 0 se refiere al archivo de datos original).

IMPUTACIÓN MÚLTIPLE (IV)

- Procedemos con un modelo de regresión logística binaria:

Codificaciones de variables categóricas				
Número de imputación			Codificación de parámetro	
			-1	-2
5	Grupo de riesgo CV	Con antecedente de evento CV	0	0
		Alto riesgo CV	1	0
		Riesgo intermedio-moderado	0	1
	Sexo	Femenino	0	
		Masculino	1	
	Puntuación Morisky-Green (visita 3 meses): cumplidores	No cumplidores	0	
		Cumplidores	1	
	Puntuación Morisky-Green (visita 1 mes): cumplidores	No cumplidores	0	
		Cumplidores	1	
	Puntuación Morisky-Green (visita basal): cumplidores	No cumplidores	0	
		Cumplidores	1	

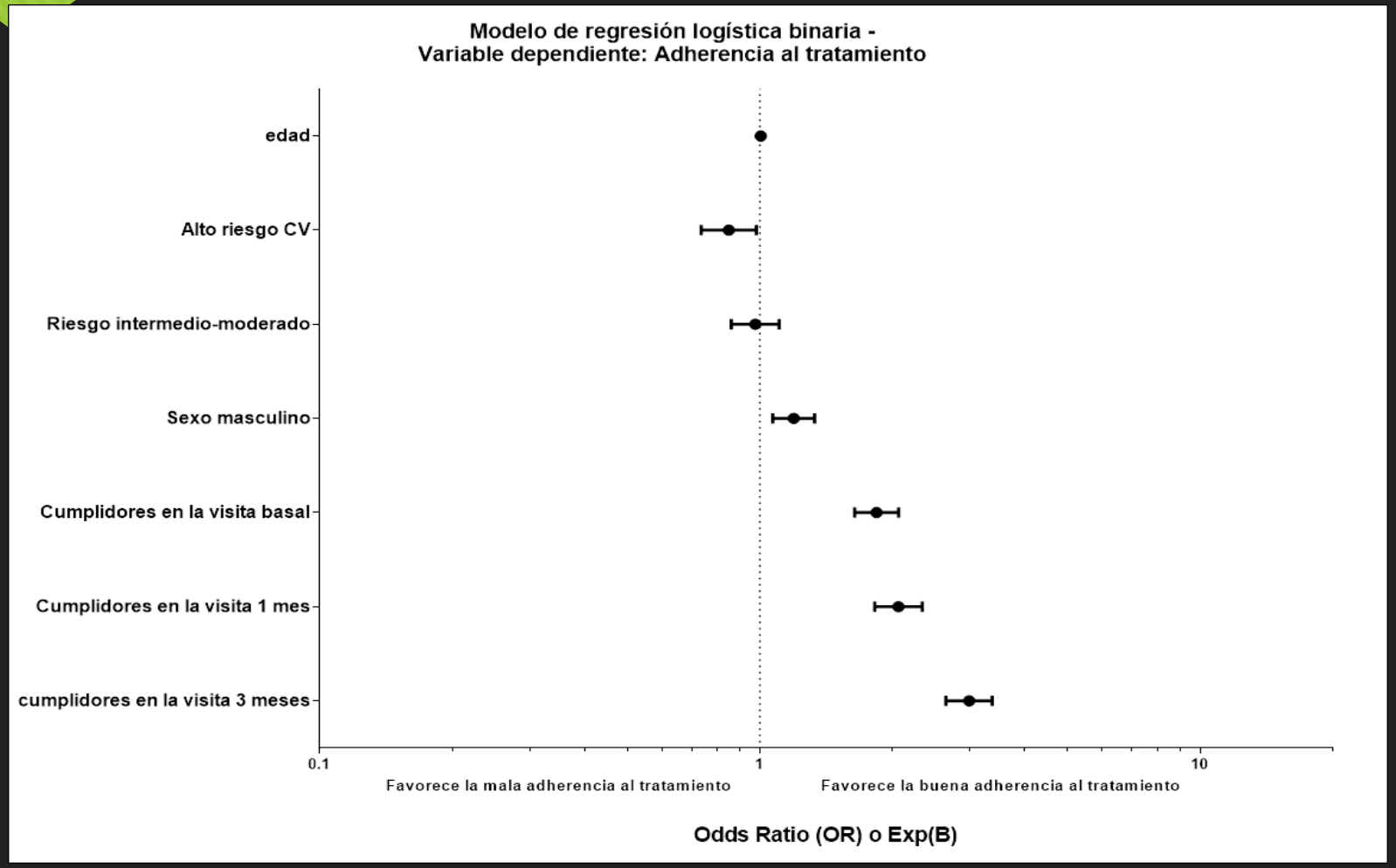
IMPUTACIÓN MÚLTIPLE (V)

Variables en la ecuación										
Número de imputación			B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
									Inferior	Superior
5	Paso 1	edad	0,005	0,002	6,947	1	0,008	1,005	1,001	1,009
		gr_riesg			5,041	2	0,080			
		gr_riesg(1)	-0,162	0,074	4,881	1	0,027	0,85	0,736	0,982
		gr_riesg(2)	-0,024	0,064	0,14	1	0,708	0,976	0,861	1,107
		sexo(1)	0,177	0,056	10,083	1	0,001	1,193	1,07	1,331
		morgreenc(1)	0,61	0,059	108,234	1	0,000	1,841	1,641	2,065
		morgreenc1(1)	0,725	0,063	130,603	1	0,000	2,064	1,823	2,337
		morgreenc3(1)	1,094	0,062	315,183	1	0,000	2,987	2,647	3,371
		Constante	-1,036	0,134	59,625	1	0,000	0,355		

- La edad del paciente, como el grado de riesgo cardiovascular (alto riesgo CV) y la adherencia durante el estudio, influyen en la adherencia final al tratamiento ($p<0.05$ y $Wald<1.96$)

?

IMPUTACIÓN MÚLTIPLE (VI)



CONCLUSIONES (I)

- Respecto al procedimiento a la hora de llevar a cabo el análisis de datasets que contengan *missings*:
  - En primer lugar, es conveniente llevar a cabo un análisis de *missings*, de manera que si el porcentaje de los desaparecidos es bajo (menos del 5%), podemos asumir que las inferencias no se verán afectadas por algún tipo de sesgo. En caso de que un porcentaje de datos faltantes superior al 5%, es necesario investigar la estructura de estos datos para confirmar si son MCAR, MAR o MNAR.
  - Una alternativa a tratar de determinar la estructura de los datos consiste en analizar las diferentes técnicas de datos faltantes para manejarlos y evaluar si los diferentes enfoques son válidos.
- Como recomendación para futuros análisis de estudios de tipo longitudinal o de cohorte, una vez depurada la base de datos, comprobar las posibles metodologías de análisis, ya que en función de cuál sea nuestra variable dependiente o de respuesta, así como la distribución y estructura de los datos debemos ser cuidadosos en la selección de la metodología más conveniente.

CONCLUSIONES (II)

- Concluimos que, efectivamente existe una buena adhesión al tratamiento con la polypill para la prevención secundaria de eventos cardiovasculares, mantenida a lo largo del tiempo.
- Adicionalmente, se plantea, para un análisis posterior la implementación del algoritmo PMM para imputar variables cuantitativas que no están normalmente distribuidas.

BIBLIOGRAFÍA (I)

- P. D. Allison, *Missing Data*, no. n.{º} 136. SAGE Publications, 2002.
- E. Lazcano-Ponce, E. Fernández, E. Salazar-Martínez, and M. Hernández-Avila, “Estudios de cohorte. Metodología, sesgos y aplicación,” *Salud Publica Mex.*, vol. 42, no. 3, pp. 230–241, Jun. 2000.
- Karahalios, L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson, “A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures.,” *BMC Med. Res. Methodol.*, vol. 12, p. 96, Jul. 2012.
- Jarrín, R. Geskus, S. Pérez-Hoyos, and J. del Amo, “Principales métodos de análisis en los estudios de cohortes de sujetos diagnosticados de infección por el virus de la inmunodeficiencia humana (VIH),” *Enferm. Infecc. Microbiol. Clin.*, vol. 28, no. 5, pp. 298–303, May 2010.
- “OMS | Enfermedades cardiovasculares,” *WHO*, 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/es/. [Accessed: 15-Mar-2017].
- “Alianzas estratégicas para la innovación | Ferrer Corporate.” [Online]. Available: http://www.ferrer.com/innovacion/alianzas-estrategicas-para-la-innovacion. [Accessed: 17-Mar-2017].

?

BIBLIOGRAFÍA (II)

- “La polypill cardiovascular, desarrollada por la colaboración CNIC-Ferrer, aumenta la adhesión al tratamiento en esta patología - Atención Especializada - Elmedicointeractivo.com.” [Online]. Available: http://www.elmedicointeractivo.com/articulo/atencion-especializada/polypill-cardiovascular-desarrollada-colaboracion-cnic-ferrer-aumenta-adhesion-tratamiento-patologia/20161115131644107170.html. [Accessed: 17-Mar-2017].
- M. Castro, C. Tutores, F. Gude Sampedro, A. Pérez, G. Máster, and T. Estadísticas, “IMPUTAC ON DE DATOS FALTANTES EN UN MODELO DE TIEMPO DE FALLO ACELERADO,” 2014.
- I. Corporation, “IBM SPSS Missing Values 20.”
- J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art.,” *Psychol. Methods*, vol. 7, no. 2, pp. 147–77, Jun. 2002.
- R. J. Little *et al.*, “The Prevention and Treatment of Missing Data in Clinical Trials,” *N. Engl. J. Med.*, vol. 367, no. 14, pp. 1355–1360, Oct. 2012.
- C. K. Enders, *Applied Missing Data Analysis*. Guilford Press, 2010.
- “American Heart Association - Building healthier lives, free of cardiovascular diseases and stroke.” [Online]. Available: http://www.heart.org/HEARTORG/. [Accessed: 18-May-2017].
- D. B. Rubin, “Inference and Missing Data,” *Biometrika*, vol. 63, no. 3, p. 581, Dec. 1976.
- M. Szklo, “Population-based Cohort Studies,” vol. 20, no. 1.

?

BIBLIOGRAFÍA (III)

- “Qué es la enfermedad cardiovascular: MedlinePlus enciclopedia médica.” [Online]. Available: https://medlineplus.gov/spanish/ency/patientinstructions/000759.htm. [Accessed: 18-May-2017].
- J. M. Baena-Díez, M. Vidal-Solsona, A. O. Byram, I. González-Casafont, G. Ledesma-Ulloa, and N. Martí-Sans, “Epidemiología de las enfermedades cardiovasculares en atención primaria. Estudio Cohorte Zona Franca de Barcelona,” *Rev. Española Cardiol.*, vol. 63, no. 11, pp. 1261–1269, Nov. 2010.
- L. A. Cupples, R. B. D’Agostino, K. Anderson, and W. B. Kannel, “Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study.,” *Stat. Med.*, vol. 7, no. 1–2, pp. 205–22.
- W. Vach and M. Blettner, “Missing Data in Epidemiologic Studies,” in *Encyclopedia of Biostatistics*, Chichester, UK: John Wiley & Sons, Ltd, 2005.
- J. Twisk and W. de Vente, “Attrition in longitudinal studies. How to deal with missing data.,” *J. Clin. Epidemiol.*, vol. 55, no. 4, pp. 329–37, Apr. 2002.
- A. Karahalios, L. Baglietto, K. J. Lee, D. R. English, J. B. Carlin, and J. A. Simpson, “EMERGING THEMES IN EPIDEMIOLOGY The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study.”
- J. Scheffer, “Dealing with Missing Data,” *Res. Lett. Inf. Math. Sci*, vol. 3, pp. 153–160, 2002.

?