Rivero Rodríguez, Gonzalo (2011). *Análisis de datos incompletos en Ciencias Sociales*. Madrid: CIS.

Ha pasado un cuarto de siglo desde la publicación del ya clásico *Multiple Imputation for Non-Response in Surveys* de Donald B. Rubin y, a pesar de la notable popularidad que han adquirido los métodos de imputación en la literatura especializada, lo cierto es que la mayor parte de la investigación empírica basada en datos de encuesta sigue prestando una escasa atención al problema de los datos perdidos. Es justo reconocer que el problema de los datos perdidos no es exclusivo de la investigación por encuesta, ya que raramente es posible disponer de información sobre todas las unidades de análisis en una investigación (por ejemplo, las estadísticas nacionales no son siempre completas y las fuentes a veces no son comparables). No obstante, el problema adquiere una relevancia especial en el contexto de la investigación por encuesta porque, a diferencia de lo que ocurre generalmente con las macro magnitudes agregadas, la no respuesta en una encuesta suele estar relacionada con características relevantes de los encuestados. Sabemos, entre otras cosas, que quienes no proporcionan una respuesta sobre su intención de voto no votan igual que quienes lo hacen. Y también es probable que quienes no revelan sus ingresos estén más cerca de los extremos de la distribución. Por estos motivos, el peligro de analizar los datos de encuesta sin tener en cuenta los mecanismos que hay detrás de la pérdida de información dista de ser una cuestión de purismo metodológico y se convierte en una preocupación real de la investigación aplicada que puede sesgar de forma importante los resultados obtenidos.

La ubicuidad del problema de los datos perdidos en las encuestas contrasta con dos hechos bastante llamativos. Por una parte, la falta de un tratamiento sistemático de este problema en los textos metodológicos al uso, lo que hace que haya que buscar las referencias en los textos especializados ya clásicos de Rubin (1976; 1987) o el trabajo del politólogo norteamericano Gary King (2001). Y por otra parte, el hecho de que en la investigación empírica aplicada se obvia rutinariamente el problema de los datos perdidos, salvo en muy contadas excepciones. Sin duda, son éstas dos buenas razones que justifican la aparición del libro de Gonzalo Rivero, el cual permite al investigador aplicado entender los peligros del no tratamiento de la no respuesta y descubrir que las medidas remediales son relativamente asequibles desde el punto de vista estadístico. Adicionalmente, cabe señalar que la escasez de textos sobre el tema es más notable en la literatura en castellano, que hasta la publicación del presente libro no disponía de un texto de referencia para el tratamiento de los datos perdidos. En la literatura sobre metodología de encuesta existen excelentes referencias orientadas a solucionar o limitar el problema de la no respuesta, como los trabajos de Cea D'Ancona (2004) sobre la calidad de la encuesta, o los más específicos sobre la no respuesta de Díaz de Rada (2001) o Sánchez Carrión (2000). Pero a pesar de la mejora en la recolección de los datos, el problema del análisis los datos incompletos sigue estando presente y requiere de un tratamiento estadístico apropiado. Por último, el presente libro también aparece

oportunamente, tras la introducción de la suite de comandos *mi* en el paquete estadístico Stata a partir de la versión 11. Esta librería es, en gran parte, deudora del ado *ice* (Royston, 2004) pero da una solución nativa al problema de la imputación múltiple con una mejora notable en la rapidez de las imputaciones, además de proporcionar nuevas funcionalidades. Esto permite al autor introducir las aplicaciones prácticas usando un software de uso estandarizado en la investigación social, a través un estilo que es útil tanto para el investigador con más familiaridad con el análisis estadístico como para el investigador aplicado que busca soluciones concretas a problemas que se presentan en el desarrollo de una investigación.

El texto comienza con dos capítulo dedicados a fijar ideas. El primer capítulo es de motivación y en él se explica el significado de los valores perdidos en la investigación social y sus consecuencias. El segundo capítulo se dedica a introducir algunos conceptos básicos para el tratamiento del problema. Se definen los patrones de pérdida de datos propuestos por Rubin (1987): el mecanismo completamente aleatorio (MCAR), el mecanismo aleatorio (MAR) y el mecanismo no aleatorio (MNAR). A continuación, el texto va introduciendo los métodos de tratamiento de la no respuesta en orden de complejidad, de forma que los lectores poco avezados estadísticamente todavía obtendrán una buena guía para el tratamiento de situaciones sencillas de datos faltantes. En el tercer capítulo se abordan los métodos que podríamos llamar simples. Se presentan los métodos de relleno con referencia a la imputación no condicionada (en la que los valores perdidos son sustituidos por un valor fijo de la variable correspondiente) y la imputación condicionada (en la que los valores perdidos se rellenan teniendo en cuenta los valores del caso en otras variables). Un método más sofisticado que el relleno es el emparejamiento, en el que a los casos con datos faltantes se les asignan los valores de aquellos casos con los que comparten una serie de características. El método *hot-deck* se aplica para el caso de las variables discretas, mientras que para el caso de las variables continuas se aplican los métodos *predictive mean matching* y *propensity score matching*.

El capítulo 4 presenta los métodos de estimación con información completa. Este capítulo no es de lectura obligada para los lectores menos iniciados en el análisis estadístico pero proporciona una excelente explicación de la lógica del tratamiento de datos incompletos desde la óptica de la teoría de la máxima verosimilitud. El autor introduce el supuesto de ignorabilidad, que se da cuando los datos son MAR y además se cumple que los procesos que generan los datos y la pérdida de información son independientes. En ese caso se puede decir que el mecanismo de pérdida de datos es ignorable, lo cual implica que se puede aplicar el mecanismo generador de los datos a los datos faltantes, sin tener que especificar el mecanismo de pérdida de datos. A partir de aquí, se explica el funcionamiento del algoritmo EM aplicado a este tipo de situaciones, a través del cual se combina una fase de expectación (paso E) con otra maximización de la verosimilitud de los parámetros estimados (paso M). El autor aplica este procedimiento al caso sencillo de una tabla de contingencia, que sirve para ilustrar la complejidad de la estimación con información completa al tiempo que permite entender cuál es el objetivo final de toda técnica de tratamiento de datos perdidos: obtener (bajo condiciones realistas)

las mismas estimaciones que obtendríamos en el caso de tener información completa.

Los capítulos 5 y 6 constituyen el núcleo central del libro y contienen su principal aportación: el análisis de la imputación múltiple, un enfoque originalmente introducido por Rubin (1987) y que se ha convertido en el estándar en el tratamiento de datos perdidos. El capítulo 5 comienza presentando las intuiciones fundamentales detrás de la imputación múltiple para centrarse luego en los principales métodos que se agrupan bajo esta denominación. Y finaliza discutiendo algunas herramientas para analizar la sensitividad de las imputaciones obtenidas. Los métodos discutidos en el capítulo son los siguientes: el aumento de datos, el algoritmo EMis, las ecuaciones encadenadas y el método *hot-deck*. El aumento de datos consiste en la extracción de valores imputados para los casos faltantes a partir de la información sobre la distribución multivariante de las variables que intervienen en el análisis. El algoritmo procede en dos pasos, de forma que en el primer paso se obtienen los valores imputados tomando los parámetros de la distribución como dados. A partir de los valores imputados, se vuelven a estimar los parámetros de la distribución y así sucesivamente, hasta que se obtiene convergencia. El algoritmo EMis se basa en la semejanza entre el aumento de datos y el referido algoritmo EM, explotando la mayor simplicidad del segundo que mejora la rapidez en la obtención de las imputaciones. El método de ecuaciones encadenadas es el apropiado cuando la distribución conjunta de las variables no es multivariante, lo que puede ocurrir frecuentemente (por ejemplo, cuando se trabaja con variables continuas y nominales simultáneamente). En este caso, la imputación para cada variable se obtiene a partir de ecuaciones de predicción específicas para cada una de las variables de análisis. A diferencia del método *hot-deck* simple (explicado en el capítulo 3), el procedimiento *hot-deck* de imputación múltiple no es determinista sino que asigna una probabilidad a cada uno de los potenciales donantes para un caso con valor perdido.

En el capítulo 6 se presentan dos ejemplos de la investigación empírica que ilustran el uso de la imputación múltiple a través de los métodos de aumento de datos y de ecuaciones encadenadas. Ambos ejemplos son presentados de forma sencilla pero conectada con la teoría relevante y, a través de ellos, el autor aborda los problemas prácticos que aparecen en la aplicación de los procedimientos de imputación múltiple, tales como los diagnósticos de la imputación. Conviene destacar que los ejemplos están seleccionados con gran realismo y serán de gran interés especialmente para quienes trabajan en el campo del comportamiento electoral. Finalmente, el libro termina con un capítulo dedicado al tratamiento de datos perdidos en R. El capítulo es breve y tiene un carácter puramente instrumental para aquellos que se decanten por el uso de este paquete estadístico, cuya librería *mice* ofrece herramientas comparables a las de otros paquetes estadísticos de uso comercial.

Como conclusión final, el texto es de un excelente nivel técnico y pedagógico y de gran utilidad para investigadores avanzados y menos iniciados. El autor mantiene

un buen equilibrio entre un nivel intermedio/avanzado, que presenta con rigor las herramientas para el tratamiento de datos perdidos, y un nivel más asequible para aquellos investigadores que pretenden solucionar las cuestiones prácticas que plantea la información incompleta en las encuestas y que, como bien muestra el autor, son problemas que no pueden ser obviados en la investigación aplicada. Es posible que los lectores más avanzados echen en falta una discusión de las posibles limitaciones de la imputación múltiple, dado que es importante tener en cuenta las incógnitas que existen todavía en la teoría estadística que está detrás de la imputación múltiple. No obstante, ha de reconocerse que este tipo de discusiones no son propias de una colección esencialmente aplicada como la de Cuadernos Metodológicos del CIS. Y el texto cumple sobradamente el objetivo de difusión de la metodología a una audiencia amplia, presentando de forma recurrente ejemplos realistas para ilustrar cada uno de los procedimientos y su tratamiento con Stata. Es de esperar que el texto sirva para mostrar los problemas de los datos incompletos en la investigación por encuesta y anime al uso de técnicas remediales, especialmente los métodos de imputación múltiple. Más allá de la todavía escasa penetración de estas técnicas en la investigación aplicada, ha de reconocerse con Rubin (1996) que la imputación múltiple es un método bastante sencillo de implementar en comparación con otros. Y la incorporación de este método a paquetes estadísticos de referencia como Stata puede augurar un tratamiento sistemático más adecuado de estos problemas.

Referencias

Cea D'Ancona, M. A. (2004). *Métodos de Encuesta. Teoría y Práctica, Errores y Mejora*. Madrid: Síntesis.

Díaz de Rada, V. (2001). *Problemas Originados por la No Respuesta en Investigación Social: Definición, Control y Tratamiento*. Pamplona: Universidad Pública de Navarra.

King, G., J. Honaker, A. Joseph y K. Scheve (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation", *American Political Science Review*, 95(1): 49-69.

Royston (2004). "Multiple Imputation of Missing Values", *The Stata Journal*, 4: 227-241.

Rubin, D. B. (1976). "Inference and Missing Data", *Biometrika*, 63(3): 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Willey.

Rubin, D. B. (1996). "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association*, 91: 473-489.

Sánchez Carrión, J. J. (2000). *La Bondad de la Encuesta. El Caso de la No Respuesta*. Madrid: Alianza.

Antonio Manuel Jaime Castillo
Universidad de Málaga