

**EL EFECTO DE LA NO RESPUESTA PARCIAL EN EL ANÁLISIS DE
DATOS DE ENCUESTA: UNA COMPARACIÓN ENTRE LA
ELIMINACIÓN DE OBSERVACIONES Y LA IMPUTACIÓN MÚLTIPLE**

Laura Morales Diez de Ulzurrun

Centro de Estudios Avanzados en Ciencias Sociales (Instituto Juan March)

RESUMEN

Los problemas que la no respuesta parcial introduce en el análisis estadístico de los datos procedentes de encuestas han sido, habitualmente, ignorados por los científicos sociales. En términos generales, la práctica tradicional ha consistido en eliminar las observaciones incompletas del análisis estadístico. Sin embargo, esta solución puede sesgar de forma importante nuestras conclusiones y, en cualquier caso, desperdicia una cantidad importante de información. En este trabajo se realiza una revisión de los distintos métodos disponibles para analizar datos de encuesta con no respuesta parcial y se presenta una aplicación concreta. La parte central del artículo muestra los resultados que se obtienen al analizar la Encuesta Mundial de Valores de 1990 eliminando, por un lado, las observaciones incompletas y, por otro, realizando una imputación múltiple de los valores perdidos. De este modo, se ofrece una ilustración de las consecuencias técnicas y analíticas que tiene la adopción de uno u otro método de manipulación de una matriz de datos incompleta.

Palabras clave: datos de encuesta, imputación múltiple, no respuesta, participación política.

Introducción

Los usuarios de los datos de encuesta conocen el muy común problema de la no respuesta parcial. Independientemente de cuál sea el grado de no respuesta total en una determinada encuesta, siempre se produce una mayor o menor cantidad de no respuesta parcial. La práctica habitual a la hora de analizar los datos consiste en ignorar estos casos, utilizando los conocidos sistemas de eliminación por lista (*listwise deletion*) o por pares (*pairwise deletion*). Esto supone que aquellos individuos que no han contestado a alguna de las preguntas que se consideran en el análisis son ignorados en el procedimiento estadístico que se emplee. Como han apuntado numerosos autores, esta práctica puede tener consecuencias muy importantes para los resultados (Moser y Kalton, 1979; Little y Rubin, 1987; Brick y Kalton, 1996; Schafer, 1997; King y otros, 1998). Por un lado, nuestras estimaciones pueden estar sesgadas, ya que la eliminación de los que no responden supone asumir que la no respuesta se distribuye de forma aleatoria entre los distintos tipos de entrevistados. En muchos casos este supuesto no se sostiene y, por tanto, nuestras conclusiones serán erróneas. En la mejor de las situaciones, aquélla en la que la no respuesta sí se distribuye de forma aleatoria, estamos perdiendo una cantidad importante de información al eliminar también las respuestas que estos individuos dieron a otras preguntas del cuestionario. En este caso, nuestros procedimientos estadísticos perderán potencia, al disminuir el tamaño de la muestra analizada (véase, por ejemplo, Ares, 1999).

En las últimas décadas, la investigación estadística ha desarrollado una gran variedad de métodos para evitar los problemas derivados de la no respuesta parcial. En las próximas páginas se ofrece una breve descripción de los más conocidos, al mismo tiempo que se discuten sus ventajas e inconvenientes. Posteriormente, se explica con más detalle el método de la imputación múltiple, que está obteniendo un gran eco entre la comunidad estadística y, más recientemente, entre los científicos sociales por tratarse de un método más robusto y de aplicación general. La segunda parte del artículo ilustra los problemas derivados de la no respuesta parcial a través del análisis de la Encuesta Mundial de Valores de 1990 para la investigación del asociacionismo político en los países occidentales. Para ello, se contrastan los resultados de un modelo de regresión logística multinominal obtenidos, primero, tras la eliminación de las observaciones incompletas según lista (*listwise deletion*) y, después, mediante la imputación múltiple de los valores perdidos. De este modo, se hacen patentes las repercusiones analíticas y teóricas que se pueden derivar de los distintos métodos de manipulación de la no respuesta parcial.

El problema de la no respuesta parcial y los métodos de análisis disponibles

La no respuesta parcial es un tipo de error de observación que consiste en la ausencia de datos en una o varias celdas de una matriz de datos. Con mucha frecuencia, al realizar una encuesta, hay individuos que cooperan aceptando la entrevista que se les propone pero que no dan respuestas válidas a una o más de las preguntas del cuestionario. La no respuesta parcial puede, así, surgir porque el entrevistado no desee contestar a alguna pregunta que considere demasiado personal, porque no sepa la respuesta a dicha pregunta, o porque ofrezca una respuesta inconsistente con otras anteriores. También puede suceder que el entrevistador, por error, no formule la pregunta o que, durante el proceso de codificación de los datos, se introduzca un valor que pueda ser reconocido como inválido.

El fenómeno de la no respuesta parcial es extremadamente frecuente en los estudios por encuesta. Hay autores que estiman que alrededor de la mitad de los entrevistados en una encuesta tipo no contestan a alguna de las preguntas que se incluyen en el cuestionario (King y otros, 1998). Como es sabido, cualquier tipo de error de no respuesta tiene dos efectos sobre el análisis de los casos completos: por un lado, reduce el tamaño de la muestra; por el otro, introduce la posibilidad de sesgos en la estimación. Los efectos sobre el tamaño de la muestra son obvios: sólo dispondremos de las observaciones con valores válidos para todas y cada una de las variables de la encuesta. Los posibles sesgos de estimación dependerán de la proporción de observaciones con no respuesta parcial y del mecanismo de producción de la no respuesta (véanse Moser y Kalton, 1979; Little y Rubin, 1987; Schafer, 1997; King y otros, 1998). En general, los sesgos serán menores cuanto más aleatorio sea el proceso de no respuesta; es decir, cuando la no respuesta se distribuya de forma aleatoria entre los distintos entrevistados. Sin embargo, en la mayoría de los casos, esta circunstancia no se produce y existen relaciones de dependencia entre las características u opiniones de los entrevistados y su propensión a no responder a ciertas preguntas. Por tanto, en la mayoría de los casos, el riesgo de sesgo en la estimación es importante.

Numerosos autores han descrito los distintos métodos a disposición de los investigadores para compensar los diversos tipos de errores de no respuesta que se producen en las encuestas (Anderson, Basilevsky y Hum, 1983; Brick y Kalton, 1996). Aquí sólo se mencionarán los que resultan apropiados para el problema de la no respuesta parcial. Para ello, se recurre a la útil tipología de Little y Rubin (1987), adoptada en la mayor parte de la literatura especializada, y que versa sobre los distintos mecanismos de producción de casos incompletos. Estos autores distinguen entre mecanismos de pérdida de casos de forma completamente aleatoria (*missing completely at random*, MCAR), aquellas situaciones en las que la ausencia de información no está asociada a ninguna variable presente o ausente en la matriz de

datos; mecanismos de pérdida de casos de forma aleatoria (*missing at random*, MAR), cuando la ausencia de información está asociada a variables presentes en la matriz de datos, pero no lo está a variables ausentes; y mecanismos de pérdida de casos no ignorables (*non-ignorable missingness*, NI), que suponen la asociación entre la ausencia de información y variables no presentes en la matriz de datos.

En primer lugar, se puede realizar un *análisis de los datos completos*. Este método es uno de los más utilizados por los investigadores, por su simplicidad. En este caso, sólo se utilizan en la estimación de parámetros las observaciones para las que todas las variables de interés tienen un valor válido (eliminación por lista). Algunas ventajas de este tipo de análisis son su simplicidad y la posibilidad de comparar los estadísticos univariantes entre sí, dado que se estiman con las mismas observaciones. Sin embargo, los inconvenientes son muchos: los análisis perderán potencia (mayor cuantas más variables se introduzcan en el análisis) y existe el riesgo de que los estimadores estén sesgados si el proceso de no respuesta no es completamente aleatorio (MCAR). Finalmente, este método desperdicia una cantidad importante de información que sería aprovechable para la estimación de estadísticos univariantes, puesto que elimina todos los casos para los que se carece de algún valor. Por esta razón, con mucha frecuencia se emplea el *método de casos disponibles*, en el que se utilizan todas las observaciones que tienen valores válidos para las variables de interés en cada momento. La eliminación por pares (*pairwise deletion*) es una extensión de este modo de análisis.

Little y Rubin (1987) han señalado los inconvenientes de todos los métodos que consisten en eliminar las observaciones incompletas. Aunque estos métodos pueden ser aceptables cuando la proporción de casos incompletos es pequeña, en términos generales conducirán a estimaciones sesgadas, puesto que asumen que el proceso de pérdida de información es completamente aleatorio (MCAR). King y otros (1998) muestran cómo los métodos de eliminación de observaciones incompletas pueden generar sesgos graves, tales como la inexactitud en la estimación de la magnitud del parámetro, o la predicción incorrecta del signo de la asociación entre variables. La imprecisión en la estimación de parámetros se produce incluso en los casos en los que el supuesto de pérdida de información de forma completamente aleatoria (MCAR) es razonable.

Una aproximación distinta al problema de los casos incompletos consiste en imputar los valores perdidos, sustituyéndolos por valores estimados con algún procedimiento. El objetivo es, por tanto, reducir el sesgo en la estimación de parámetros que se introduce al eliminar los casos incompletos.

Se pueden distinguir varios tipos de imputación; el más sencillo de los cuales es el *deductivo*, que consiste en la asignación de valores a las celdas incompletas, por parte del investigador, tras deducir con un cierto grado de certidumbre los valores más plausibles. Sin embargo, la mayoría de los métodos de imputación se pueden expresar a través de la siguiente función:

$$y_{mi} = f(y_{nm}) + \varepsilon$$

donde y_{mi} es el valor imputado, y_{nm} representa las observaciones con valores válidos, y ε es un residuo aleatorio.

Así, se puede distinguir entre los métodos de imputación que fijan ε a 0 (deterministas) y aquéllos que no lo hacen (estocásticos). Los primeros producen estimadores más precisos de puntos como la media, pero no reflejan la variación residual y distorsionan la forma de la distribución de la variable (Schafer, 1997).

Uno de los métodos de imputación más conocidos es el de *sustitución por la media*, en el que los valores perdidos pasan a adoptar el valor medio de la variable en cuestión (más o menos un cierto residuo aleatorio, en la versión estocástica). Si la variable se distribuye de forma aproximadamente normal, la sustitución de los valores perdidos por la media no sesgará la estimación del parámetro univariante. Sin embargo, uno de los problemas fundamentales de este método es que distorsiona la forma de la distribución de la variable, lo que tiene consecuencias especialmente graves cuando se realizan análisis bivariantes –en concreto tablas de contingencia–, ya que se reduce artificialmente la varianza de la variable afectada y la covarianza entre variables.

Los problemas de la sustitución por la media condujeron a la adopción de métodos de sustitución aleatoria simple (*simple hot-deck substitution*), en los que los valores perdidos son reemplazados por valores aleatorios dentro del rango de la variable en cuestión. Estos métodos, relacionados con los de *bootstrap*, tienen la ventaja de mantener intacta la forma de la distribución empírica de la variable, aunque presentan el inconveniente de que distorsionan la relación con otras variables y, por tanto, son sólo útiles para los análisis univariantes (Brick y Kalton, 1996).

Con el fin de preservar intacta la asociación entre variables, se ha extendido el uso de métodos estocásticos que emplean variables auxiliares en la imputación de los valores perdidos. Los métodos de imputación por clases (*hot-deck imputation*), dividen a los individuos entrevistados en distintas clases, de acuerdo con los valores que adoptan en una serie de variables. Posteriormente, se imputan los valores perdidos a través de la asignación aleatoria de los valores correspondientes a individuos que pertenecen a la misma clase y que ofrecieron respuestas válidas a las variables a imputar. Una de las mayores desventajas de estos métodos es que las clases han de ser definidas en base a un número reducido de variables, con el fin de asegurar que habrá suficientes observaciones completas en todas las clases. Para minimizar estos problemas, se han desarrollado diferentes versiones de los métodos de imputación por clases, como los *métodos jerárquicos de imputación por clases*, o los *métodos de imputación por clases de vecino más próximo*.

Otra alternativa es la imputación con *métodos de regresión*. Estos métodos consisten en la imputación de los valores incompletos tras la estimación de un modelo de regresión con covariables que se consideran asociadas a la variable en cuestión.

Las versiones estocásticas añaden un factor residual que refleja la incertidumbre de la estimación. Algunos de los problemas que presentan estos métodos, están relacionados con su tendencia a inflar la asociación entre variables tras la imputación (fundamentalmente en las versiones deterministas), y la complejidad de su implementación en situaciones en las que la no respuesta está presente también entre las covariables del modelo de regresión. En estos casos, los modelos de ecuaciones estructurales que pueden estimarse son muy costosos en términos de computación y problemáticos cuando la no respuesta afecta a todas las variables (Acock, sin fecha).

Por último, cabe destacar los métodos de imputación múltiple, tratados en más detalle en próximas páginas. Las ventajas de la imputación múltiple de los valores perdidos son, fundamentalmente, la posibilidad de aplicar estos métodos de forma general a una amplia variedad de análisis (univariantes, bivariantes o multivariantes) y la incorporación en la estimación de los parámetros de la incertidumbre introducida por el uso de valores imputados.

En definitiva, existe una gran variedad de métodos de imputación que muestran distintas propiedades y grados de sofisticación. Sin embargo, es necesario ser conscientes de los problemas que cada uno de ellos presentan. Como muestran Little y Rubin (1987), el uso de métodos de imputación simplistas puede generar problemas de estimación tan o más graves como los derivados de la eliminación de los casos incompletos, ya que pueden distorsionar los estimadores, los errores típicos y los tests de hipótesis. Por otro lado, la mayoría de estos métodos no permite distinguir entre los valores imputados y los reales, inflando así artificialmente la precisión de los coeficientes obtenidos en nuestros análisis.

El método de la imputación múltiple

En la última década, se ha desarrollado la investigación estadística en torno a otro método: la imputación múltiple. Tras la publicación de los distintos trabajos de Little y Rubin (Rubin y Schenker, 1986; Little y Rubin, 1987; Rubin, 1987), han aparecido otros muchos que profundizan en distintas versiones de imputación múltiple (Little y Rubin, 1989; Rubin y Schafer, 1990; Schafer, Khare y Ezzati-Rice, 1993; Rubin, 1996; Schafer, 1997; Schafer y Olsen, 1998; King y otros, 1998; Allison, 2000).

La imputación múltiple es una técnica de Monte Carlo en la que los valores perdidos son sustituidos por $m > 1$ valores simulados. De forma resumida, la imputación múltiple consiste en la imputación de los casos perdidos a través de la estimación de un modelo estocástico apropiado. Sin embargo, esta imputación se realiza M veces y se producen M archivos completos con los valores imputados. Posteriormente, se lleva a cabo el análisis estadístico ordinario con las M matrices de datos completas y se combinan los resultados con una serie de fórmulas específicas proporcionadas por Little y Rubin (1987).

El siguiente ejemplo puede ilustrar algunos de los aspectos más importantes del proceso de imputación. El recuadro 1 muestra una matriz de datos con 15 casos y 11

variables¹. Se puede observar que la matriz está incompleta, ya que algunas celdas no contienen valor alguno: en total, desconocemos 14 respuestas de un total de 165. Aunque la proporción de no respuesta parcial no es demasiado alta (aproximadamente un 8,5%), ésta se distribuye de tal modo que afecta a la mayoría de las observaciones (10 de 15). Si realizáramos un análisis multivariante con todas las variables empleando el método de la eliminación de observaciones incompletas según lista, perderíamos más del 66% de nuestros casos. Como se puede ver en este ejemplo, dado que en la mayoría de los casos sólo tenemos un par de valores perdidos para cada observación, estaríamos desperdiciando toda la información que las observaciones incompletas sí tienen en las restantes variables. Por tanto, además de obtener resultados probablemente sesgados, haríamos un uso muy ineficiente de los datos de encuesta de que disponemos. Como ya se ha mencionado, el objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de parámetros.

Recuadro 1: *Matriz original con valores perdidos.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	.	3	.
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	.	3	2
4	1	5	1	2	2	1	27	9	.	7	.
5	2	2	2	2	3	2	61	6	.	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	.	1	1	25	9	.	6	2
9	2	3	2	3	1	1	31	9	.	6	2
10	1	2	2	2	.	1	39	5	.	6	2
11	2	4	3	2	2	2	27	9	.	6	2
12	2	5	4	2	2	2	29	9	.	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	.	4	3

Para llevar a cabo la imputación múltiple de los valores perdidos, procederíamos del siguiente modo. En primer lugar, se seleccionan las variables que se emplearán en el modelo de imputación. Es imprescindible que todas las variables que se van a utilizar conjuntamente en posteriores análisis se incluyan en dicho modelo, pero se puede y se debe incluir también todas aquellas variables que puedan ayudar a estimar los valores perdidos². En segundo lugar, se decide el número de imputa-

¹ En realidad, se trata de una sección de la matriz de datos de la muestra belga de la Encuesta Mundial de Valores de 1990-93. Se muestran tan sólo los quince primeros casos y unas cuantas de las variables a modo de ilustración. Sin embargo, las imputaciones que aparecen más adelante se han realizado con el conjunto de la muestra y 15 variables.

² Así, si tenemos muchos valores perdidos en una variable como la de nivel educativo del entrevistado y disponemos en la misma encuesta de una variable de nivel educativo de los

ciones que se desea realizar. En general, entre 3 y 5 imputaciones son suficientes (véase Rubin, 1987: 114).

A continuación se muestran las matrices completas resultantes tras realizar 5 imputaciones de los valores perdidos de la matriz original, y se resaltan en negrilla los valores que se han imputado. Como se puede ver, tan sólo los valores que inicialmente se desconocían varían con cada imputación; como es natural, el modelo de imputación sólo afecta a las celdas inicialmente vacías.

Recuadro 2: *Archivo imputado n° 1.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	4	3	2
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	10	3	2
4	1	5	1	2	2	1	27	9	7	7	3
5	2	2	2	2	3	2	61	6	4	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	2	1	1	25	9	8	6	2
9	2	3	2	3	1	1	31	9	10	6	2
10	1	2	2	2	1	1	39	5	2	6	2
11	2	4	3	2	2	2	27	9	3	6	2
12	2	5	4	2	2	2	29	9	4	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	9	4	3

Recuadro 3: *Archivo imputado n° 2.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	2	3	2
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	8	3	2
4	1	5	1	2	2	1	27	9	9	7	2
5	2	2	2	2	3	2	61	6	3	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	2	1	1	25	9	9	6	2
9	2	3	2	3	1	1	31	9	9	6	2
10	1	2	2	2	1	1	39	5	9	6	2
11	2	4	3	2	2	2	27	9	9	6	2
12	2	5	4	2	2	2	29	9	10	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	10	4	3

padres, será conveniente introducir esta última en el modelo de imputación aunque no esté relacionada con nuestro modelo analítico.

Recuadro 4: *Archivo imputado n° 3.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	6	3	2
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	10	3	2
4	1	5	1	2	2	1	27	9	3	7	2
5	2	2	2	2	3	2	61	6	7	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	3	1	1	25	9	1	6	2
9	2	3	2	3	1	1	31	9	5	6	2
10	1	2	2	2	2	1	39	5	4	6	2
11	2	4	3	2	2	2	27	9	10	6	2
12	2	5	4	2	2	2	29	9	10	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	5	4	3

Recuadro 5: *Archivo imputado n° 4.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	4	3	2
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	5	3	2
4	1	5	1	2	2	1	27	9	9	7	2
5	2	2	2	2	3	2	61	6	8	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	2	1	1	25	9	3	6	2
9	2	3	2	3	1	1	31	9	8	6	2
10	1	2	2	2	2	1	39	5	10	6	2
11	2	4	3	2	2	2	27	9	4	6	2
12	2	5	4	2	2	2	29	9	9	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	10	4	3

Recuadro 6: *Archivo imputado n° 5.*

ID	V94	V147	V241	V249	V338	V353	V355	V356	V363	V368	V378
1	2	1	1	2	1	2	69	6	2	3	2
2	2	1	4	1	1	2	44	8	8	3	2
3	2	4	4	2	2	1	25	8	10	3	2
4	1	5	1	2	2	1	27	9	6	7	3
5	2	2	2	2	3	2	61	6	6	6	2
6	2	2	2	2	2	1	66	7	8	6	3
7	2	2	1	3	1	1	65	3	5	6	1
8	1	6	2	2	1	1	25	9	5	6	2
9	2	3	2	3	1	1	31	9	4	6	2
10	1	2	2	2	3	1	39	5	9	6	2
11	2	4	3	2	2	2	27	9	10	6	2
12	2	5	4	2	2	2	29	9	10	6	2
13	1	2	2	2	1	2	61	7	8	6	3
14	1	1	2	3	4	1	66	8	5	6	2
15	1	4	3	2	2	1	35	7	4	4	3

Si observamos con atención las cinco matrices imputadas podemos apreciar el funcionamiento del modelo de estimación. El valor de la variable v378 para el caso n° 1 ha sido estimado con una gran precisión por el modelo de imputación, ya que en las cinco imputaciones el valor predicho es el mismo. Sin embargo, los valores perdidos de la variable v363 se imputan con una mayor incertidumbre y varían bastante de una matriz imputada a otra. No obstante, para algunos individuos el grado de variación es mayor que para otros: compárese el caso n° 12 (con valores imputados 4, 10, 10, 9, y 10) con el caso n° 8 (valores imputados 8, 9, 1, 3, y 5). El modelo de imputación estima con mayor o menor precisión en función de la capacidad que tienen las variables introducidas de predecir los valores perdidos de las restantes variables. En este ejemplo, las restantes variables no tienen una gran capacidad predictiva de los valores de v363. De este modo, cuando la información de la que disponemos en la encuesta no nos permite predecir bien los valores perdidos de determinadas variables, la imputación de dichos valores variará de forma estocástica.

El siguiente paso que seguiríamos sería el de llevar a cabo los análisis estadísticos (univariantes, bivariantes o multivariantes) necesarios para nuestra investigación. El análisis se realizaría, en nuestro caso, con cada una de las 5 matrices imputadas y, posteriormente, se combinarían los resultados de los 5 modelos con las fórmulas proporcionadas por Little y Rubin (1987). Se obtendría, por tanto, un único coeficiente (\bar{q}) que combina los m estimadores de punto (q_j) con la siguiente ecuación:

$$\bar{q} = \frac{1}{m} \sum_{j=1}^m q_j$$

De modo similar, se calcula el error típico del nuevo coeficiente como:

$$SE(q)^2 = \frac{1}{m} \sum_{j=1}^m SE(q_j)^2 + \left[\sum_{j=1}^m (q_j - \bar{q})^2 / (m-1) \right] \left(1 + \frac{1}{m} \right)$$

Así, la variación estocástica que se introduce con la imputación nos permite, en la posterior agregación de los parámetros estimados por nuestro análisis estadístico ordinario, incorporar la incertidumbre real de la imputación en los errores estándar calculados.

En definitiva, el aspecto crucial de la imputación múltiple reside en la definición del modelo de imputación y en el método de imputación. Es fundamental que el modelo empleado en la estimación de los valores perdidos contenga las variables que se emplearán posteriormente en los análisis estadísticos ordinarios, con el fin de preservar las relaciones entre variables. Por otro lado, cuanto mejor sea el modelo —en términos predictivos— menor será la variación de los valores imputados y más

precisos serán nuestros estimadores posteriores. El método de estimación de los valores imputados varía de unas aplicaciones a otras, y sus propiedades son también distintas (Allison, 2000).

En general, la imputación múltiple es una de las soluciones más adecuadas al problema de la no respuesta parcial, ya que es relativamente sencillo de aplicar (si se compara, por ejemplo, con los métodos de máxima verosimilitud) y de validez general. Aunque otros métodos basados en el algoritmo EM (*Expectation-Maximization*) pueden ser más eficientes —al no emplear técnicas de simulación—, su uso es más complejo y costoso en términos de tiempo (Schafer, 1997). Por otro lado, se ha demostrado que la eficacia de la imputación de valores mediante imputación múltiple es alta, incluso cuando se realiza sobre un número reducido de matrices completas (Rubin, 1987: 114); pues a no ser que la proporción de información perdida sea muy alta, la imputación de entre tres y cinco matrices completas será suficiente.

Actualmente, existen varias aplicaciones que permiten realizar imputaciones múltiples con distintos tipos de matrices de datos. Por un lado, se encuentran las aplicaciones exclusivamente dedicadas a la imputación y, por otro, las aplicaciones incorporadas en paquetes estadísticos más generales. Entre las primeras, destacan los programas AMELIA, EMCOV, MICE, NORM-CAT-MIX-PAN, y SOLAS (se puede encontrar una descripción breve de cada uno de ellos en <http://www.multiple-imputation.com>). Entre las segundas, destacan las incluidas últimamente en SAS (versiones 8.1 y 8.2).

Una de las versiones de imputación múltiple disponibles es la descrita por King y otros (1998), implementada con el programa gratuito *Amelia* (Honaker y otros, 1999) y disponible en <http://Gking.harvard.edu>. Esta aplicación emplea el algoritmo EMis (*Expectation-Maximization with importance sampling*), que se basa en el algoritmo EM modificado con una ronda de muestreos iterativos basados en la simulación. El usuario introduce un fichero con la matriz de datos incompleta, en alguno de los formatos aceptados. Posteriormente establece el algoritmo de imputación que desea utilizar —se puede elegir entre cuatro variantes—, indica el número de ficheros completos que desea imputar (el valor de m), y selecciona las opciones disponibles que permiten ajustar el modelo de estimación al tipo de datos (variables ordinales, categóricas, series temporales, etc.). Finalmente, *Amelia* produce M matrices de datos completos que habrán de ser empleadas sucesivamente en los análisis estadísticos ordinarios que el investigador desee realizar. Junto con el programa *Amelia*, se ofrece también un programa para *Stata*, llamado “Miest” (Scheve, 1999), que combina automáticamente los resultados de los M modelos. Además, el programa “Miest” ofrece otros datos que permiten evaluar la significación de los coeficientes (estadístico t , grados de libertad y probabilidad).

El programa *Amelia* ha sido empleado con datos de la Encuesta Mundial de Valores de 1990 con el fin de evaluar el impacto que el problema de la no respuesta parcial tiene sobre la estimación de un determinado modelo. A continuación se pasa a describir la encuesta y el modelo a analizar.

La Encuesta Mundial de Valores de 1990-93 y el análisis de la pertenencia a grupos políticos

La Encuesta Mundial de Valores (1990-1993) es un estudio internacional que supuso la realización de la misma encuesta en 45 sociedades (*World Values Study Group*, 1994). Esta encuesta describe, fundamentalmente, los valores y actitudes de los ciudadanos de 45 países del mundo, e incluye una amplia lista de preguntas relacionadas con diferentes cuestiones de interés para los científicos sociales. El cuestionario que se aplicó fue, sustancialmente, el mismo en todos los países, aunque en algunos casos se pueden encontrar excepciones importantes.

Esta encuesta fue empleada con el fin de estudiar el fenómeno del asociacionismo político en los países occidentales. Para ello, se seleccionaron las muestras correspondientes a diecisiete de éstos³. Todas ellas son muestras representativas a nivel nacional de los ciudadanos mayores de 18 años (exceptuando Bélgica e Irlanda con 17 años o más y Noruega con 19 años o más). Algunas de las muestras fueron seleccionadas por muestreo estratificado, otras mediante un muestreo aleatorio multi-etápico y, finalmente, otras por muestreo con cuotas por grupos sociales.

El análisis que aquí se describe tiene por objetivo estudiar las características socio-económicas, las actitudes políticas y los valores que están relacionados con la pertenencia a asociaciones de carácter político en los países occidentales. Se consideran asociaciones políticas aquellos grupos organizados de ciudadanos que persiguen bienes colectivos —ya sean bienes públicos puros u otros bienes colectivos— y que tengan como principal objetivo influir en la selección del personal gubernamental o en sus actividades, incluir temas en la agenda política o cambiar los valores y preferencias que guían el proceso de toma de decisiones.

La mayor parte de la literatura especializada en participación política ha señalado la relevancia de los recursos socio-económicos a disposición de los individuos a la hora de determinar la distinta propensión de unos y otros a participar en política (Verba y Nie, 1972; Verba, Nie y Kim, 1978; Barnes y Kaase, 1979; Jennings y van Deth, 1989; Kaase, 1989; Parry, Moyser y Day, 1992; Verba, Schlozman y Brady, 1995). Muchos de estos estudios han comprobado la existencia de pautas comunes entre los distintos países en relación con la influencia de estas variables en el comportamiento participativo de los ciudadanos occidentales. Características sociales como el sexo, la edad, el nivel educativo o la posición en el mercado de trabajo son consideradas junto con la más obvia del nivel de ingresos familiares recursos de gran importancia. Factores como éstos suelen moldear las experiencias y la vida social de los ciudadanos de distintas formas, y suelen determinar las oportunidades de participación que se les presentan. En segundo lugar, la literatura académica también resalta la importancia de las actitudes políticas y los valores de los ciudada-

³ Los países son Alemania occidental, Austria, Bélgica, Canadá, Dinamarca, España, EE.UU., Finlandia, Francia, Gran Bretaña, Irlanda, Islandia, Italia, Noruega, Países Bajos, Portugal, y Suecia.

nos a la hora de explicar su participación política (Inglehart, 1991; Verba, Schlozman y Brady, 1995; Dekker, Koopmans y van den Broek, 1997). Entre las actitudes políticas que se tomarán en consideración están el interés por la política, la confianza interpersonal, y las actitudes hacia el cambio social. Con respecto a los valores, el índice de materialismo/postmaterialismo de Inglehart (1991) se empleará para distinguir entre individuos con valores materialistas, postmaterialistas y mixtos.

Por razones teóricas y analíticas, resulta útil distinguir entre grupos políticos de tipo tradicional y grupos políticos de nuevo tipo. Por tanto, la variable dependiente, la pertenencia a grupos políticos, se analiza como una variable categórica, en la que los individuos son clasificados en cuatro categorías:

- (1) aquéllos que no pertenecen a ninguna asociación de carácter político;
- (2) aquéllos que pertenecen sólo a grupos políticos de tipo tradicional —sindicatos, partidos políticos, asociaciones profesionales y grupos de acción comunitaria—;
- (3) aquéllos que pertenecen sólo a grupos políticos de nuevo tipo —grupos pro-derechos humanos o en ayuda al tercer mundo, grupos ecologistas o de conservación de la naturaleza, grupos de mujeres o feministas, movimiento pacifista, y grupos en defensa de los derechos de los animales—; y
- (4) aquéllos que pertenecen tanto a grupos políticos tradicionales como de nuevo tipo. La distribución de los entrevistados entre cada una de las categorías se muestra en la Tabla 1 (a y b).

Tabla 1a: *Distribución (%) de la variable "pertenencia a grupos políticos".*

Categorías	ALE	AUS	BEL	CAN	DIN	ESP	EEUU	FIN
Ninguno	62.9	64.1	61.5	62.2	35.7	91.9	62.1	46.6
Tradicional	21.3	24.3	15.4	20.5	46.0	4.8	19.6	40.8
Nuevo	10.1	6.8	12.9	7.7	4.7	1.9	7.7	4.1
Ambos	5.7	4.7	10.2	9.6	13.5	1.4	10.6	8.5
Nº de casos	2101	1460	2792	1730	1030	2637	1836	588

Tabla 1b: *Distribución (%) de la variable "pertenencia a grupos políticos".*

Categorías	FRA	GRB	IRL	ITA	ISL	NOR	PBA	POR	SUE
Ninguno	83.5	66.6	76.6	77.2	28.2	42.6	46.1	86.1	30.6
Tradicional	9.3	21.2	15.1	15	56.5	45.0	15.5	10.7	46.4
Nuevo	4.5	6.2	5.3	5	3.4	4.1	17.9	1.7	5.4
Ambos	2.7	5.9	3	2.9	11.8	8.2	20.4	1.5	17.6
Nº de casos	1002	1484	1000	2010	702	1239	1017	1185	1047

En estos casos, la técnica multivariable más apropiada es la regresión logística multinominal, que aquí se expresa como:

$$\ln \left[\frac{P_{ir}}{P_{i1}} \right] = \beta_{0r} + \sum \beta_{kr} x_{ik} \quad i = 1, \dots, n; k = 1, \dots, K; r = 2, 3, 4$$

donde $P_{ir} = \Pr(y_i = r)$, e y_i denota el valor de la variable definida con anterioridad para el individuo i ,

x_{ik} = características personales del individuo i ,

β_{kr} = el coeficiente de regresión asociado a la variable k para la categoría r ,

y $r = 1$ es la categoría de referencia (“no pertenece”).

Aunque el problema de la no respuesta parcial es común a cualquier análisis estadístico de los datos de una encuesta, cuando se emplea la regresión logística multinomial la investigadora puede encontrar dificultades añadidas. El hecho de que la variable dependiente esté compuesta por varias categorías y , por tanto, se estimen varias ecuaciones, hace que el número de observaciones disponibles para el análisis sea aún más importante de lo habitual. Si, además, tenemos en cuenta que en el caso que nos ocupa una de las categorías de la variable dependiente (“pertenece a ambos tipos de grupos”) contiene un número inferior de observaciones que las demás (véanse las tablas 1a y 1b), resulta imprescindible recuperar las observaciones perdidas por la no respuesta parcial.

Sin embargo, no se trata solamente de un problema técnico. Como han mostrado King y otros (1998), los resultados de modelos estadísticos realizados con las observaciones completas de una matriz de datos incompleta pueden variar sustancialmente de los que se obtendrían mediante la aplicación de los mismos modelos a las matrices imputadas. Por tanto, nuestras conclusiones teóricas pueden diferir de forma importante.

Con el fin de estudiar los efectos que se derivan, en esta investigación, del análisis de datos completos y del análisis de los datos imputados de forma múltiple se ha procedido de la siguiente forma. En primer lugar, se han estimado las diecisiete regresiones multinomiales con la matriz de casos completos; es decir, se han eliminado todas las observaciones para las que faltaba algún valor de las variables introducidas en el modelo (método de observaciones disponibles con eliminación según lista). En segundo lugar, se han imputado cinco matrices de datos completas para cada uno de los 17 países estudiados⁴. Posteriormente se han estimado los mismos modelos de regresión multinomial con los cinco archivos para cada país, combinando los resultados con el programa “Miest” en Stata versión 6 (©Stata Corporation). Tras obtener los coeficientes combinados se procedió a comparar los resultados obtenidos con el análisis de observaciones completas y con los datos imputados. Dada la complejidad de visualizar los resultados de 34 regresiones multinomiales

⁴ Se ha preferido realizar la imputación múltiple de los valores perdidos país por país en lugar de hacerlo con la matriz conjunta de los 17 países. Las razones son varias: en primer lugar, las muestras son de muy distintos tamaños (entre los 2792 casos de Bélgica y los 588 de Finlandia) y algunos países habrían tenido un peso desproporcionado en el modelo de imputación, ya que el programa Amelia no permite incluir variables de ponderación. En segundo lugar, resulta arriesgado asumir que el mismo modelo de imputación sería válido para los 17 países. La proporción de casos con valores perdidos varía enormemente entre unos y otros (entre el 4,1% de Islandia y el 72,2% de Bélgica), así como la asociación entre la no respuesta parcial y las características de los ciudadanos también difiere entre países.

(las 17 obtenidas con eliminación de observaciones incompletas y las 17 obtenidas mediante la agregación de los resultados con imputación múltiple), se ha creado la Tabla 2 (a y b) con el fin de resumirlos.

Tabla 2a: *Cambios en la significación de las variables del modelo tras imputar los valores perdidos.*

Variables	ALE	AUS	BEL	CAN	DIN	ESP	EEUU	FIN
Sexo					+	+		+
Edad								
Ingresos			+		•			
Educación			+					
Estado civil	+	+			+	+		+
Frecuencia de práctica religiosa								
Situación laboral								
Tamaño del lugar de residencia								
Confianza interpersonal						+		
Interés en la política								
Eficacia política		+	+				+	
Índice de valores		+				+		
Actitudes cambio		+			+	+		
% de casos recuperados	21,7	30,9	72,2	27	24,7	5,8	20,1	9,7

• Variable significativa en el modelo con casos completos que pasa a no serlo con datos imputados

+ Variable no significativa en el modelo con casos completos que pasa a serlo con datos imputados

Así, en esta tabla, podemos observar que los modelos estimados con las observaciones completas (mediante eliminación por lista) y los estimados con los valores imputados difieren de manera importante, excepto en el caso de Canadá, para el que el modelo no cambia sustancialmente. En la mayoría de los casos, el uso de la matriz de observaciones completas conducía a no rechazar la hipótesis nula en falso y, por tanto, en un buen número de ocasiones se habría llegado a la conclusión de que determinadas variables no mantenían una asociación significativa con la pertenencia

a grupos políticos⁵. Esto sucede, en especial, con variables como el estado civil, el género y la eficacia política. En otros casos, la eliminación de los datos incompletos conducía en algunos países a rechazar la hipótesis nula en falso. Así pasa, sobre todo, con el índice de valores de Inglehart y el indicador de las actitudes hacia el cambio social.

Tabla 2b: *Cambios en la significación de las variables del modelo tras imputar los valores perdidos.*

Variables	FRA	GRB	IRL	ITA	ISL	NOR	PBA	POR	SUE
Sexo			+						
Edad									
Ingresos									+
Educación						+			
Estado civil	+	•							+
Frecuencia de práctica religiosa							+	•	
Situación laboral									
Tamaño del lugar de residencia		+							•
Confianza interpersonal								+	
Interés en la política									
Eficacia política									•
Índice de valores				•	•	+			+
Actitudes cambio	+	•		•	•				
% de casos recuperados	20,6	33	14,5	38,3	4,1	19,4	32,3	13,5	19

• Variable significativa en el modelo con casos completos que pasa a no serlo con datos imputados

+ Variable no significativa en el modelo con casos completos que pasa a serlo con datos imputados

En esta misma tabla podemos también observar la proporción de observaciones que se han podido recuperar para el análisis tras imputar los valores perdidos. Destaca, especialmente, el caso de Bélgica, en el que la alta proporción de no respuesta a la variable de ingresos hace que se pierda más del 70% de la muestra. También es alta la tasa de casos perdidos en Austria, Italia, Gran Bretaña y los Países Bajos. Sin

⁵ Para la elaboración de la Tabla 2 (a y b), se ha considerado que una variable es significativa cuando muestra un efecto significativo sobre la probabilidad condicional de pertenecer a grupos políticos en alguna de las tres ecuaciones del modelo multinomial estimado.

embargo, es importante destacar que, incluso en los casos en los que la proporción de no respuesta parcial sobre el total de la muestra es reducida (Islandia y España), los resultados cambian sustancialmente cuando se comparan los modelos estimados con ambos métodos de manipulación de los valores perdidos.

A modo de ilustración, la Tabla 3 (*a* y *b*) compara los coeficientes de las ecuaciones del modelo multinomial con uno y otro método de tratamiento de la no respuesta para las muestras de Bélgica e Islandia. Se han escogido estos dos países porque representan los dos casos con mayor y menor tasa de observaciones con no respuesta parcial, respectivamente. Esta tabla permite ver con mayor claridad el tipo de cambios que se producen en la estimación de parámetros cuando, en lugar de eliminar las observaciones con valores perdidos, se emplea la imputación múltiple de la información no disponible. Como se puede observar, en muchos casos el tamaño de los coeficientes varía de forma considerable (en algunos casos, la ratio entre ambos llega a ser igual a dos), y lo mismo sucede con el valor del estadístico *t*. Las implicaciones analíticas se perciben también mejor: por ejemplo, en el caso de Bélgica, la eliminación de casos incompletos nos conduciría a concluir que la situación laboral es prácticamente irrelevante a la hora de explicar la pertenencia a grupos políticos en sus distintas modalidades; sin embargo, tras la imputación de la información perdida, afirmaríamos que la situación laboral es especialmente relevante para comprender la pertenencia a grupos de nuevo tipo, ya que las personas laboralmente inactivas y las desempleadas tienden en mayor medida a pertenecer a este tipo de grupos políticos que las personas con un empleo a tiempo completo. Del mismo modo, los resultados con imputación múltiple en el caso de Islandia nos harían conceder un efecto mucho más moderado (aunque todavía importante) al interés de los ciudadanos en la política a la hora de explicar su pertenencia a grupos políticos; hasta el punto de que no tendría un efecto significativo sobre la pertenencia a grupos de tipo tradicional. En general, como cabía esperar dada la diferencia en la proporción de casos perdidos, los cambios en la estimación de parámetros son menores en el caso islandés que en el belga.

Por último, la Tabla 4 resume las principales consecuencias interpretativas que uno y otro método tienen sobre los resultados de este modelo. En especial, destacan las diferentes conclusiones a las que llegaríamos sobre el efecto del género y el estado civil de los ciudadanos de estos 17 países occidentales en su probabilidad de pertenecer a grupos políticos. Ambas variables parecen poco relevantes cuando eliminamos las observaciones incompletas, pero resultan significativas en una gran mayoría de países cuando utilizamos el método de imputación múltiple. Del mismo modo, la interpretación del modelo varía de forma crucial en los casos de Austria, España y Bélgica, modificando mucho nuestras conclusiones sobre este fenómeno de participación política y los factores individuales asociados a él en estos países.

Tabla 3a: *Resultados para Bélgica del modelo multinominal con los métodos de eliminación de observaciones incompletas (según lista) e imputación múltiple.*

Variables	BÉLGICA					
	Eliminación casos incompletos			Imputación de valores perdidos		
	Sólo trad.	Sólo nuev.	Ambos	Sólo trad.	Sólo nuev.	Ambos
Sexo (Mujer)	-.76 (-4.2)	.862 (4.3)	.332 (1.72)	-.43 (-2.7)	.773 (4.76)	.144 (0.84)
Edad	.008 (1.26)	-.001 (-.2)	-.002 (-.2)	.006 (1.13)	.002 (0.32)	.001 (0.2)
Ingresos (en deciles)	.068 (1.91)	.055 (1.35)	.047 (1.16)	.069 (1.88)	.053 (1.03)	.017 (0.44)
Estudios(edad finalizó)	.072 (1.84)	.012 (0.29)	.066 (1.45)	.118 (2.83)	.032 (0.76)	.061 (1.27)
No vive en pareja	-.48 (-2.5)	-.26 (-1.2)	-.54 (-2.4)	-.51 (-2.9)	-.17 (-.96)	-.84 (-3.9)
Religiosidad (de + a -)	.024 (0.7)	-.15 (-4.0)	-.12 (-3.1)	-.01 (-0.4)	-.1 (-3.43)	-.12 (-3.2)
Situación laboral						
Trabaja <30h	-.26 (-0.9)	-.25 (-0.7)	.376 (1.39)	-.2 (-0.97)	.037 (0.14)	.09 (0.37)
Inactivo	-.59 (-2.6)	.322 (1.33)	-.04 (-0.2)	-.57 (-3.1)	.522 (2.38)	-.22 (-1.0)
Desempleado	.303 (0.98)	.264 (0.61)	.207 (0.5)	.208 (0.71)	.705 (1.98)	.305 (0.83)
Tamaño localidad	-.15 (-3.0)	-.09 (-1.6)	-.12 (-2.2)	-.1 (-2.68)	.013 (0.31)	-.06 (-1.3)
Confianza interpersonal						
Confía	.05 (0.29)	.383 (2.03)	.416 (2.15)	-.1 (-0.6)	.44 (2.78)	.532 (2.93)
No sabe	.214 (0.7)	-1.1 (-1.8)	.477 (1.4)	.196 (0.78)	-.78 (-2)	.589 (1.94)
Interés en la política						
Muy interesado	.586 (1.82)	0.8 (2.02)	1.86 (5.6)	1.04 (3.51)	1.15 (3.33)	1.96 (6.6)
Algo interesado	.852 (4.02)	.665 (2.59)	.985 (3.75)	.674 (3.55)	.442 (1.95)	.971 (4.13)
No muy int.	-.05 (-0.2)	.724 (3.31)	.473 (1.8)	-.00 (-.02)	.588 (3.25)	.406 (1.81)
Eficacia política	.128 (1.86)	-.17 (-2.0)	.11 (1.38)	.06 (0.942)	-.02 (-0.2)	.125 (1.72)
Índice de valores						
Materialista	-.16 (-0.7)	-.48 (-1.9)	-.34 (-1.1)	-.1 (-0.54)	-.58 (-2.7)	-.32 (-1.2)
Postmaterialista	.391 (2.03)	.235 (1.01)	.959 (4.6)	.436 (2.55)	.024 (0.12)	.739 (3.85)
Actitudes cambio						
Radical	-.04 (-0.8)	.252 (0.5)	.069 (0.15)	.027 (0.07)	.19 (0.42)	.31 (0.743)
Conservador	.099 (0.5)	.168 (0.79)	-.62 (-2.3)	.146 (0.81)	.13 (0.677)	-.1 (-0.42)
Constante	-1.8 (-3.1)	-1.3 (-1.9)	-2.1 (-3.0)	-2.1 (-3.9)	-2.7 (-4.5)	-2.4 (-3.8)

Se presentan los coeficientes de regresión y, entre paréntesis, los valores del estadístico t. En negrilla se resaltan los coeficientes significativos para $p < 0.05$.

Tabla 3b: Resultados para Islandia del modelo multinomial con los métodos de eliminación de observaciones incompletas (según lista) e imputación múltiple.

Variables	ISLANDIA					
	Eliminación casos incompletos			Imputación de valores perdidos		
	Sólo trad.	Sólo nuev	Ambos	Sólo trad.	Sólo nuev	Ambos
Sexo (Mujer)	.31 (1.19)	.4 (0.57)	1.46 (3.7)	-.22 (-1.0)	1.02 (1.75)	.91 (2.78)
Edad	-.007 (-.8)	.026 (1.14)	.019 (1.44)	-.01 (-1.9)	.023 (1.5)	.002 (0.24)
Ingresos (en deciles)	*	*	*	*	*	*
Estudios(edad finalizó)	-.01 (-0.2)	-.01 (-.09)	.109 (1.42)	-.003 (-.1)	.106 (1.18)	.076 (1.24)
No vive en pareja	-.47 (-1.8)	.904 (1.23)	-.71 (-1.7)	-.35 (-1.7)	-.08 (-.14)	-.61 (-1.8)
Religiosidad (de + a -)	.016 (0.25)	-.16 (-0.9)	-.21 (-2.2)	-.008 (-.2)	-.03 (-0.3)	-.21 (-2.7)
Situación laboral						
Trabaja <30h	-.57 (-1.5)	.913 (1.1)	-.38 (-0.7)	-.45 (-1.2)	.496 (0.58)	-.48 (-.99)
Inactivo	-.12 (-0.9)	-.31 (0)a	-.53 (-0.3)	-.85 (-0.7)	-.29 (-1.5)a	-.18 (-0.1)
Desempleado	*	*	*	*	*	*
Tamaño localidad	*	*	*	*	*	*
Confianza interpersonal						
Confía	-.07 (-0.3)	.143 (0.22)	.563 (1.52)	-.04 (-0.2)	.119 (0.23)	.713 (2.39)
No sabe	-.84 (-1.6)	-.88 (-0.6)	.032 (0.04)	-.54 (-1.2)	.206 (0.3)	.032 (0.04)
Interés en la política						
Muy interesado	1.13 (2.2)	3.88 (2.9)	3.15 (3.43)	.603 (1.52)	2.29 (2.6)	2.4 (3.19)
Algo interesado	.652 (1.90)	-.08 (-.05)	2.16 (2.66)	.481 (1.69)	.54 (0.584)	2.1 (3.12)
No muy int.	.588 (1.78)	1.65 (1.44)	1.25 (1.51)	.307 (1.15)	1.09 (1.36)	1.38 (2.05)
Eficacia política	-.18 (-2)	-.21 (-0.9)	.006 (0.04)	-.17 (-2.2)	-.25 (-1.3)	.047 (0.38)
Índice de valores						
Materialista	-.3 (-1.13)	1.2 (1.74)	.438 (1.08)	-.2 (-0.93)	.589 (1.2)	.48 (1.47)
Postmaterialista	.542 (1.27)	-.12 (-0.1)	.987 (1.82)	.432 (1.26)	-.31 (-0.3)	.835 (1.86)
Actitudes cambio						
Radical	-1.15 (-2)	-.31 (0)a	-.43 (-0.4)	-.83 (-1.6)	.55 (0.497)	.252 (0.37)
Conservador	-.15 (-0.5)	-.73 (-0.7)	.058 (0.12)	-.33 (-1.3)	-.21 (-0.3)	-.09 (-0.2)
Constante	1.48 (1.86)	-.4 (-1.73)	-3.9 (-2.8)	1.88 (2.84)	-4.9 (-2.9)	-3.1 (-2.5)

Se presentan los coeficientes de regresión y, entre paréntesis, los valores del estadístico t. En negrilla se resaltan los coeficientes significativos para $p < 0.05$.

* Variable no disponible o sin observaciones para esta categoría

a Este parámetro ha sido estimado con muy pocos grados de libertad

Tabla 4: *Consecuencias sobre la interpretación de los resultados del modelo.*

Variables	Significativa en...	
	Eliminación casos incompletos	Imputación de valores perdidos
Sexo	12 países	16 de los 17 países
Edad	16 de los 17 países	Sin cambios
Ingresos	12 países	13 países
Estudios	13 países	15 países
Estado civil	4 países	10 países
Religiosidad	10 países	10 países (con cambios)
Situación laboral	16 de los 17 países	Sin cambios
Tamaño localidad	11 países	11 países (con cambios)
Confianza interpersonal	7 países	9 países
Interés en la política	16 de los 17 países	Sin cambios
Eficacia política	11 países	13 países
Índice de valores	10 países	12 países
Actitudes cambio social	6 países	7 países (con cambios)
Países		
Austria, España	Actitudes no muy relevantes	(casi) Todas las actitudes significativas
Bélgica	Recursos económicos no relevantes	Ingresos y educación son significativas

A modo de conclusión

Prácticamente todas las encuestas presentan un problema, de mayor o menor magnitud, de no respuesta parcial. Los científicos sociales han tendido a ignorar esta dificultad empleando métodos simples de análisis que se basan en la eliminación de las observaciones con valores perdidos en alguna de las variables introducidas en el análisis estadístico. La literatura estadística ha demostrado en numerosas ocasiones los inconvenientes de emplear esta solución: la pérdida de potencia estadística de la muestra y, lo que es peor, los sesgos derivados de la no aleatoriedad completa del proceso de pérdida de información. Hoy en día existen numerosas técnicas alternativas, todas ellas con ventajas e inconvenientes que deben ser valorados en función de la proporción de datos incompletos y del mecanismo de producción de los mismos. En términos generales, la técnica de la imputación múltiple está ganando partidarios entre la comunidad estadística —y poco a poco, también entre la comunidad de científicos sociales— por la simplicidad de su aplicación y su carácter general.

En estas páginas se han comparado los resultados de un mismo modelo de regresión logística multinominal obtenidos, primero, con el procedimiento de eliminación por lista de los casos incompletos y, después, con imputación múltiple. Para ello, se han utilizado las muestras nacionales de 17 países occidentales incluidos en la Encuesta Mundial de Valores de 1990-1993 y el programa de imputación múltiple Amelia creado por Honaker y otros (1999). Los resultados muestran que, en efecto, la eliminación de los casos incompletos nos conduciría a conclusiones que no parecen sostenerse cuando se aprovecha la información disponible en las observaciones incompletas. En concreto, hay variables que sistemáticamente pasan a ser estadísticamente significativas cuando se utilizan los datos provenientes de la imputación. En algunos países, incluso las probabilidades estimadas varían de forma considerable, según se adopte uno u otro método. Y en todos los casos el modelo de regresión multinominal será más estable y preciso, al estar basado en un tamaño de muestra mayor.

Referencias

- Acock, A. C. (sin fecha). *Working with Missing Data*. Trabajo inédito, URL: <http://www.orst.edu/instruct/hdfs632/missingdata.html> (05/07/2000)
- Ares, V. M. (1999) La prueba de significación de la hipótesis “cero” en las investigaciones por encuesta. *Metodología de Encuestas*, 1(1), 47-68.
- Allison, P. D. (2000). Multiple Imputation for Missing Data. A Cautionary Tale. *Sociological Methods & Research*, 28(3), 301-309.
- Anderson, A. B.; Basilevsky, A.; y Hum, D. P. J. (1983). Missing Data: a Review of the Literature. En P. H. Rossi; J. D. Wright; y A. B. Anderson (Eds.), *Handbook of Survey Research*. London: Academic Press, 415-494.
- Barnes, S. H.; y Kaase, M. (1979). *Political Action. Mass Participation in five western democracies*. Beverly Hills: Sage.
- Brick, J. M.; y Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.
- Dekker, P.; Koopmans, R.; y van den Broek, A. (1997). Voluntary associations, social movements and individual political behaviour in Western Europe. En J. W. van Deth (Ed.), *Private groups and Public life*. London: Routledge, 220-240.
- Honaker, J.; Joseph, A.; King, G.; Scheve, K.; y Singh, N. (1999). *Amelia: a Program for Missing Data*. Documento electrónico, URL: <http://GKing.harvard.edu> (09/07/1999).
- Inglehart, R. (1991). *El cambio cultural en las sociedades industriales avanzadas*. Madrid: CIS/Siglo XXI.
- Jennings, M. K.; y van Deth, J. W. (1989). *Continuities in Political Action*. Berlin: Walter de Gruyter.
- Kaase, M. (1989). Mass Participation. En M. K. Jennings; y J. W. van Deth (Eds.), *Continuities in Political Action*. Berlin/New York: Walter De Gruyter, 23-64.

- King, G.; Honaker, J.; Joseph, A.; y Scheve, K. (1998). *Listwise Deletion is Evil: what to Do about Missing Data in Political Science*. Documento inédito, URL: <http://GKing.harvard.edu> (19/08/1998).
- Little, R. J. A.; y Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, 18, 292-326.
- Moser, C.; y Kalton, G. (1979). *Survey Methods in Social Relations*. Aldershot: Gower.
- Parry, G.; Moyser, G.; y Day, N. (1992). *Political Participation and Democracy in Britain*. Cambridge: Cambridge University Press.
- Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
- Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91, 54-63.
- Rubin, D. B.; y Schafer, J. L. (1990). Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data. *The Statistical Computing Section of the American Statistical Association*, 83- 88.
- Rubin, D. B.; y Schenker, N. (1986). Multiple Imputation for Interval Estimation from Single Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81(394), 366-374.
- Schafer, J. L. (1997). *Analysis and Simulation of Incomplete Multivariate Data: Algorithms and Examples*. London: Chapman and Hall.
- Schafer, J. L.; Khare, M.; y Ezzati-Rice, T. M. (1993). Multiple Imputation of Missing Data in NHANES III. *The Annual Research Conference*, Washington D.C.: Bureau of the Census.
- Schafer, J. L.; y Olsen, M. K. (1998). *Multiple Imputation for Multivariate Missing-Data Problems: a Data Analyst's Perspective*. Trabajo inédito, URL: <http://www.stat.psu.edu/jls> (09/03/1998)
- Scheve, K. (1999). *Stata procedures for multiple imputation analysis (version 1.15)* [computer program]. URL : <http://Gking.harvard.edu> (6/02/1999).
- Verba, S.; y Nie, N. H. (1972). *Participation in America: political democracy and social equality*. New York: Harper & Row.
- Verba, S.; Nie, N. H.; y Kim, J. (1978). *Participation and Political Equality. A seven-nation comparison*. New York: Cambridge University Press.
- Verba, S.; Schlozman, K. L.; y Brady, H. E. (1995). *Voice and Equality. Civic Voluntarism in American Politics*. New York: Harper & Row.
- World Values Study Group. (1994). *World Values Survey, 1981-1984 and 1990-1993 [Computer file]*. [ICPSR Version]. Ann Arbor, MI: Institute for Social Research (producer). Inter-University Consortium for Political and Social Research (distributor)