LEVANTAMIENTO VII RONDA DE LA ENCUESTA DE PROTECCIÓN SOCIAL
Informe sobre Imputación de Datos

Abril, 2020

Contenido

Índice de Tablas

Índice de Figuras

1. Consideraciones previas

Como toda encuesta compleja, la EPS VII Ronda ha requerido de ajustes posteriores al resultado del trabajo de campo. Estos ajustes se realizan para asegurar calidad, completitud y adecuación de la información recolectada a los criterios establecidos y a las buenas prácticas. Dado que las encuestas son el producto del trabajo de personas y el aporte de información de los entrevistados, se encuentran sujetas a la falibilidad habitual de cualquier acción humana.

En este informe se repasan las acciones que se llevaron a cabo sobre los tres componentes de la EPS VII Ronda (encuesta Presencial, encuesta de Re-Entrevista y encuesta de Continuidad) para realizar la imputación de datos faltantes producto de la no respuesta al ítem (No sabe / No responde), la detección de *outliers*, inconsistencias contextuales, o saltos de flujo erróneos.

Estos ajustes realizados a las bases de datos permiten obtener una encuesta que se comporta como una unidad informativa tal y como fue concebida.

Para resolver el problema de no respuesta al ítem, es decir, cuando las personas deciden no entregar información en alguna pregunta particular, se han realizado imputaciones siguiendo los procedimientos detallados en esta sección. Las razones para la no respuesta al ítem son múltiples, desde la dificultad para entender la pregunta, falta de concentración en el momento, dificultades para lograr condensar y resumir información, esfuerzo involucrado en la elaboración de la respuesta, problemas de memoria, etc. El resultado es información incompleta que habitualmente se codifica como “No Sabe / No Responde”. Sin embargo, esa información puede ser vital para entender el comportamiento de la población.

Es por ello que se recurre a las imputaciones, es decir, a obtener información a través de distintos métodos estadísticos para lograr completar el conjunto informativo generado por la encuesta y poder hacer inferencia estadística. Son diversos los métodos propuestos en la literatura, y todos tienen sus ventajas y desventajas; generalmente, éstas parten con los supuestos que se hacen para declarar la validez de una imputación.[1]

En la EPS, tradicionalmente, se han imputado variables como la edad (apoyado en registros administrativos), el nivel educativo, los ingresos laborales y los arriendos (reales o estimados) pagados por la vivienda principal del entrevistado. Esta última variable es de las más complicadas y con mayor tasa de no respuesta, ya que implica imaginar, si es que no se arrienda efectivamente la vivienda, cuál sería el precio pagado por arriendo de una vivienda de similares condiciones en el sector. Esto pone una presión importante en los entrevistados, que terminan no respondiendo a la pregunta. Históricamente, las tasas de no respuesta para este ítem de la EPS han estado alrededor del 20%.

[1] Existe una amplia literatura para recorrer. Algunos clásicos que se sugiere revisar incluyen Allison (2001), Carlin et al. (2003, 2008), Lee y Carlin (2010), Marchenko y Eddings (2011), Rubin (1987, 1996), Schafer (1997), Medina y Galván (2007), Little y Rubin (1987, 2002), Van Buuren (2017), entre otros referentes en la materia.

En este informe, se propone y muestra una forma de abordar la imputación de datos en base a buenas prácticas observadas en otras encuestas, y a la calidad de datos que la EPS ha logrado reunir. Se imputan cuatro variables, dependiendo del componente de la EPS VII Ronda de que se trate: nivel educativo máximo alcanzado, ingresos laborales promedio mensuales, gastos promedio mensuales del hogar y arriendo (real o supuesto) promedio mensual pagado por la vivienda habitada por el entrevistado.

A continuación, se presentan los supuestos detrás del comportamiento de la información faltante, para luego abordar la imputación de educación, ingresos, gastos y arriendo en cada encuesta.

1.1. Mecanismos de pérdida de datos

Como se mencionó anteriormente, las razones para no responder pueden ser muchas. Sin embargo, sin importar las razones particulares de no respuesta que se observen, ni el patrón que esta siga en el cuestionario, se debe precisar qué tipo pérdida de información se está observando, pues de ella dependerá la metodología más apropiada para imputar. En la literatura, se reconocen ampliamente tres casos (Wood et al., 2019):

1) Pérdida aleatoria ("*Missing at random*" o MAR), definida por Rubin (1976): este tipo de pérdida de información implica que el dato faltante no depende de variables no observadas, pero puede deberse (explicarse) por la información contenida en variables observadas. Por ejemplo, si los hombres son más o menos propensos a responder alguna pregunta sobre ingresos, o si tienden a exagerar o sub-dimensionar la respuesta, o si las mujeres son menos propensas a revelar su situación de salud en la encuesta que los hombres (no respuesta parcial), se puede decir que el mecanismo de pérdida de información que se observa es MAR.

2) Pérdida totalmente aleatoria ("*Missing completely at random*" o MCAR): en este caso, la pérdida de información no está relacionada con información observable o no observable. En particular, las razones son enteramente ajenas a la encuesta en sí. Por lo tanto, la información faltante (total o parcial) es simplemente un subconjunto de la información total que sí se logró recolectar.

3) Pérdida no aleatoria ("*Missing not at random*" o MNAR): este tipo de pérdida de información es estudiada por Little y Rubin (2014), quienes aseguran que cuando la información no observada que tiene el entrevistado es el factor que determina la información aportada por el entrevistado en la encuesta, entonces se debe modelar la pérdida de información (por ejemplo, se pueden utilizar imputaciones múltiples) para reemplazar esta información perdida y evitar sesgos en los resultados. Sin embargo, se debe tener conocimiento (o información) previa sobre la observación, algo que está disponible para todos los casos abordados en la EPS VII Ronda, al tratarse solamente de observaciones de panel (sin refresco).

Las variables que se imputan en esta ronda se encuentran en el cruce de las tres categorías. El nivel educativo no reportado puede deberse más a problemas de comprensión o deseo de no evidenciar

alguna situación (por ejemplo, analfabetismo o baja escolaridad). Adicionalmente, en la encuesta de Continuidad, para disminuir el tiempo de la entrevista, se decidió no preguntar sobre educación a los mayores de 45 años. Por lo tanto, toda esta información debe ser imputada forzosamente.

Por su parte, los ingresos generalmente están atados a múltiples sesgos en sus reportes, algunos de ellos convertidos en datos faltantes. Lo mismo sucede con los gastos de los hogares, en especial si el entrevistado no está involucrado habitualmente en el manejo presupuestario familiar. Finalmente, los arriendos, como se mencionó anteriormente, puede ser un dato faltante por diversas razones.

En todo caso, frente a esta incertidumbre acerca del mecanismo originador de los datos, conviene utilizar técnicas robustas a pérdidas de información MNAR, que sería el caso más nocivo. Es por ello que, en el caso de las variables de ingresos, gastos y arriendos, se emplean imputaciones múltiples, que ha han probado ser mecanismos muy resilientes (y, para muchos, los únicos adecuados actualmente) para controlar la pérdida de datos no aleatoria (van Buuren, 2017).

1.2. Imputación de nivel educacional

1.2.1. Imputación en la encuesta Presencial

El nivel educativo es una variable clave para entender el contexto socioeconómico de los entrevistados, y suele ser información que éstos entregan sin mayores inconvenientes. Sin embargo, en esta ronda, se observaron algunos casos de información faltante.

Dado que la variable educación típicamente se utiliza en niveles para realizar análisis, se decidió construir la variable de tramos de educación. Esta variable se construyó en base a las variables de curso y nivel máximo de educación alcanzado, según auto-reporte del entrevistado. La variable fue definida con 5 tramos educativos: Ninguna educación (0), Básica (1), Media (2), Técnica Superior (3) y Universitaria (4).

Toda la información no reportada (No sabe / No responde) se pasó a dato perdido. Esto resultó en 16 casos faltantes (0,2% de no respuesta).

Para completar los valores faltantes, se llevó a cabo una imputación *cold deck*, empleando información consistente reportada en rondas anteriores de la EPS. La consistencia se chequeó en cada caso para asegurarse de que la educación (curso y nivel) se mantenía o se incrementaba entre olas, siguiendo la lógica del proceso educativo.

En las Tablas 1 a 5 se muestra la distribución de frecuencias inicial, las donaciones de cada ola, desde la más reciente hasta la más antigua, y la distribución final.

Tabla 1: Encuesta Presencial – Tramos de Educación: situación inicial 16 datos faltantes

t_educ_impCD ── Máximo nivel educacional alcanzado - Imputación Cold Deck

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	88	1.13	1.13	1.13
	1 Básica	2008	25.74	25.80	26.93
	2 Media	3582	45.92	46.02	72.94
	3 Técnica Superior	939	12.04	12.06	85.01
	4 Universitaria	1167	14.96	14.99	100.00
	Total	7784	99.79	100.00	
Missing	.	16	0.21		
Total		7800	100.00		

Fuente: Elaboración propia.

Tabla 2: Encuesta Presencial – Donación de la EPS VI Ronda: 12 casos

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	1	6.25	8.33	8.33
	1 Básica	7	43.75	58.33	66.67
	2 Media	3	18.75	25.00	91.67
	3 Técnica Superior	1	6.25	8.33	100.00
	Total	12	75.00	100.00	
Missing	.	4	25.00		
Total		16	100.00		

Fuente: Elaboración propia.

Tabla 3: Encuesta Presencial – Donación EPS V Ronda: 1 caso

		Freq.	Percent	Valid	Cum.
Valid	6 Educación Media Científico-Humanista	1	25.00	100.00	100.00
Missing	.	3	75.00		
Total		4	100.00		

Fuente: Elaboración propia.

Tabla 4: Encuesta Presencial – Donación de la EPS IV Ronda: 2 casos

		Freq.	Percent	Valid	Cum.
Valid	3 preparatoria (s. antiguo)	1	33.33	50.00	50.00
	4 básica	1	33.33	50.00	100.00
	Total	2	66.67	100.00	
Missing	.	1	33.33		
Total		3	100.00		

Fuente: Elaboración propia.

Tabla 5: Encuesta Presencial – Donación EPS III Ronda: 1 caso

		Freq.	Percent	Valid	Cum.
Valid	6 Medica Cientifico-Humanista	1	100.00	100.00	100.00

Fuente: Elaboración propia.

El proceso de imputación *cold deck* fue suficiente para cubrir los déficits de información del tramo educacional. Así, de acuerdo a la Tabla 14, la variable de tramo educativo se encuentra completa para esta encuesta.

Tabla 6: Encuesta Presencial – Tramos de Educación: Situación final sin datos faltantes

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	89	1.14	1.14	1.14
	1 Básica	2017	25.86	25.86	27.00
	2 Media	3587	45.99	45.99	72.99
	3 Técnica Superior	940	12.05	12.05	85.04
	4 Universitaria	1167	14.96	14.96	100.00
	Total	7800	100.00	100.00	

Fuente: Elaboración propia.

1.2.2.Imputación en la encuesta de Continuidad

Para quienes respondieron esta encuesta, existen dos razones para la pérdida de información sobre el nivel educativo: por un lado, en la encuesta sólo se preguntó el nivel educativo a aquellas personas que tenían edad menor o igual a 45 años. Por otro lado, dentro del grupo de entrevistados que debía responder a estas preguntas, algunos declararon no saber o prefirieron no responder a las mismas. Estas pérdidas de información implican no tener disponible una variable clave en el análisis del comportamiento de las personas. Por esa razón, se decidió construir la variable de tramos de educación.

Al igual que en la encuesta Presencial, la variable de tramos de educación se construyó en base a las variables de curso y nivel máximo de educación alcanzado, según auto-reporte del entrevistado, con los 5 tramos ya descritos.

La información de auto-reporte proviene:

- de la misma encuesta, para los menores de 46 años
- de olas anteriores, después de verificar su consistencia, para los entrevistados mayores de 45 años.

En la Tabla 7, se observa la situación para la muestra completa y sólo para aquellos que debieron haber respondido en la encuesta. Para la muestra completa, hay una pérdida de 2.630 casos (52,28% de los casos). Si sólo se considera a los que debían responder esta información en la encuesta, existen 5 casos perdidos, es decir, un 0,21% de la muestra. Estos casos aparecen como resultado de convertir toda la información no reportada (No sabe / No responde) a datos perdido (*missing*, en terminología habitual en encuestas).

Tabla 7: Encuesta de Continuidad – Tramos de Educación: situación inicial de datos faltantes

Muestra completa: 2.630 casos faltantes

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	3	0.06	0.12	0.12
	1 Básica	176	3.50	7.33	7.46
	2 Media	993	19.74	41.36	48.81
	3 Técnica Superior	422	8.39	17.58	66.39
	4 Universitaria	807	16.04	33.61	100.00
	Total	2401	47.72	100.00	
Missing	.	2630	52.28		
Total		5031	100.00		

Muestra sólo menores de 46 años: 5 casos faltantes

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	3	0.12	0.12	0.12
	1 Básica	176	7.32	7.33	7.46
	2 Media	993	41.27	41.36	48.81
	3 Técnica Superior	422	17.54	17.58	66.39
	4 Universitaria	807	33.54	33.61	100.00
	Total	2401	99.79	100.00	
Missing	.	5	0.21		
Total		2406	100.00		

Fuente: Elaboración propia.

Para completar los valores faltantes, se llevó a cabo una imputación *cold deck*, empleando información consistente reportada en rondas anteriores de la EPS. La consistencia se chequeó en cada caso para asegurarse que la educación (curso y nivel) se mantenía o se incrementaba entre olas, siguiendo la lógica del proceso educativo.

En las Tablas 9 a 13 se muestran las donaciones de cada ola, desde las más recientes hasta las más antiguas.

Tabla 8: Encuesta de Continuidad – Donación de la EPS VI Ronda: 1.757 casos

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	36	1.37	2.05	2.05
	1 Básica	577	21.94	32.84	34.89
	2 Media	834	31.71	47.47	82.36
	3 Técnica Superior	129	4.90	7.34	89.70
	4 Universitaria	181	6.88	10.30	100.00
	Total	1757	66.81	100.00	
Missing	.	873	33.19		
Total		2630	100.00		

Fuente: Elaboración propia.

Tabla 9: Encuesta de Continuidad – Donación EPS V Ronda: 302 casos

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	1	0.11	0.33	0.33
	1 Básica	73	8.36	24.17	24.50
	2 Media	160	18.33	52.98	77.48
	3 Técnica Superior	31	3.55	10.26	87.75
	4 Universitaria	37	4.24	12.25	100.00
	Total	302	34.59	100.00	
Missing	.	571	65.41		
Total		873	100.00		

Fuente: Elaboración propia.

Tabla 10: Encuesta de Continuidad – Donación de la EPS IV Ronda: 284 casos

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	2	0.35	0.70	0.70
	1 Básica	48	8.41	16.90	17.61
	2 Media	134	23.47	47.18	64.79
	3 Técnica Superior	34	5.95	11.97	76.76
	4 Universitaria	66	11.56	23.24	100.00
	Total	284	49.74	100.00	
Missing	.	287	50.26		
Total		571	100.00		

Fuente: Elaboración propia.

Tabla 11: Encuesta de Continuidad – Donación EPS III Ronda: 150 casos

		Freq.	Percent	Valid	Cum.
Valid	1 Básica	16	5.57	10.67	10.67
	2 Media	69	24.04	46.00	56.67
	3 Técnica Superior	26	9.06	17.33	74.00
	4 Universitaria	39	13.59	26.00	100.00
	Total	150	52.26	100.00	
Missing	.	137	47.74		
Total		287	100.00		

Fuente: Elaboración propia.

Tabla 12: Encuesta de Continuidad – Donación de la EPS II Ronda: 60 casos

		Freq.	Percent	Valid	Cum.
Valid	1 Básica	10	7.30	16.67	16.67
	2 Media	31	22.63	51.67	68.33
	3 Técnica Superior	4	2.92	6.67	75.00
	4 Universitaria	15	10.95	25.00	100.00
	Total	60	43.80	100.00	
Missing	.	77	56.20		
Total		137	100.00		

Fuente: Elaboración propia.

Tabla 13: Encuesta de Continuidad – Donación de la EPS I Ronda: 77 casos

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	1	1.30	1.30	1.30
	1 Básica	11	14.29	14.29	15.58
	2 Media	33	42.86	42.86	58.44
	3 Técnica Superior	10	12.99	12.99	71.43
	4 Universitaria	22	28.57	28.57	100.00
	Total	77	100.00	100.00	

Fuente: Elaboración propia.

El proceso de imputación *cold deck* fue suficiente para cubrir los déficits de información del tramo educacional. Así, de acuerdo a la Tabla 14, la variable de tramo educativo se encuentra completa para esta encuesta.

Tabla 14: Encuesta de Continuidad – Tramos de Educación: Situación final sin datos faltantes

		Freq.	Percent	Valid	Cum.
Valid	0 Ninguno	43	0.85	0.85	0.85
	1 Básica	911	18.11	18.11	18.96
	2 Media	2254	44.80	44.80	63.76
	3 Técnica Superior	656	13.04	13.04	76.80
	4 Universitaria	1167	23.20	23.20	100.00
	Total	5031	100.00	100.00	

Fuente: Elaboración propia.

1.3. Imputación de variables de ingreso y gasto

1.3.1. Imputación en la encuesta Presencial

Las variables de ingreso y gasto presentan, habitualmente, mayores tasas de no respuesta al ítem. Adicionalmente, algunos de los valores reportados aparecen como valores extremos (*outliers*) o sospechosos de ser el resultado de ingreso erróneo del dato en los sistemas de levantamiento de encuestas. Para evitar que los datos ausentes o aquellos erróneamente reportados generen distorsiones en la medición, se recurre a la imputación para obtener variables con mejor comportamiento estadístico, que no distorsionan la distribución original de los datos, al tiempo que mejoran la inferencia estadística.

En la encuesta levantada presencialmente, se revisaron las siguientes variables:

- Ingreso laboral promedio mensual del último trabajo: variable b12 en montos, y b12_t en tramos
- Arriendo estimado mensual: variable d8_a en montos, y d8_t en tramos

En la Tabla 15, se muestran los datos faltantes para cada variable que resultan de no respuesta y de determinación de *outliers* o datos no fiables.

Tabla 15. Datos faltantes de ingresos y arriendos en la encuesta Presencial

Datos faltantes para Arriendo Mensual (7.800 obs.)

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
arriendo	1,519		6,281	105	10000	4500000
arriendo_t	7,099		701	11	1	11
arriendo_ed	818		6,982	108	10000	4500000

Datos faltantes para Ingreso Promedio Mensual del entrevistado ocupado (4.393 obs.)

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
ing_lab	265		4,128	324	20000	7000000
ing_lab_t	4,151		242	13	1	88

Fuente: Elaboración propia.

En el caso de los arriendos, la pérdida de información alcanzó al 19,5% de la muestra (1.519 casos sobre 7.800), mientras que los ingresos laborales presentaron una pérdida de información de 6,0% (265 casos sobre 4.393). Se consideró que todos aquellos que declaraban tener ocupación y no eran de la categoría "trabajo de familiar no remunerado", debían tener respuesta en este ítem.

Para realizar la imputación se adoptó el método de imputaciones múltiples con ecuaciones encadenadas, que es bastante robusto a MNAR y otros mecanismos más complejos de pérdida de datos.[2] Las imputaciones múltiples, si bien pueden aparecer como una estrategia complicada, son una buena forma de resolver el problema de datos faltantes generando la menor distorsión posible. Además, han demostrado ser muy eficientes en muchos entornos, mejorando la calidad de los estimadores sin poner en riesgo el sesgo.

[2] En Chile, esta metodología se emplea en la Encuesta Financiera de Hogares del Banco Central de Chile.

En sí, la imputación múltiple es una metodología de simulación que permite completar múltiples veces los sets de datos. Siguiendo las reglas de Rubin (1987, 1996), es posible obtener estimadores robustos e insesgados. Se debe tener en cuenta que la imputación múltiple es un procedimiento sencillo de reemplazo de información faltante por múltiples versiones de un mismo dato. Dado esto, se puede partir de la imputación tradicional, replicarla y aplicar las reglas de Rubin para obtener mejores estimadores. Las reglas de Rubin también ayudan a controlar la incertidumbre causada por los datos faltantes, por lo que el conjunto de datos también mejora.

Sin embargo, al contrario de otros métodos como *cold deck*, *hot deck*, medias predictivas, etc., esta metodología no está interesada en un valor particular del dato faltante, sino más bien en su distribución y cómo ésta contribuye a explicar la distribución conjunta de la población. Entonces, el foco no está en una imputación, sino en el uso de múltiples imputaciones para resolver un conjunto predictivo del comportamiento poblacional. Puesto en términos prácticos, no nos interesa, por ejemplo, el salario promedio mensual no declarado de un entrevistado, sino el promedio y la varianza de salarios de los entrevistados de la encuesta.

Para imputar los salarios y el arriendo, se usó la variante de ecuaciones encadenadas (conocida como MICE Multiple Imputation by Chained Equations) que permite retroalimentar, iteración tras iteración, el sistema de imputación, apoyando cada nueva imputación en los valores obtenidos anteriormente. Si bien es intensiva en cómputo, tiene la ventaja de ser muy flexible y generar estimadores eficientes (Carlin et al., 2003).

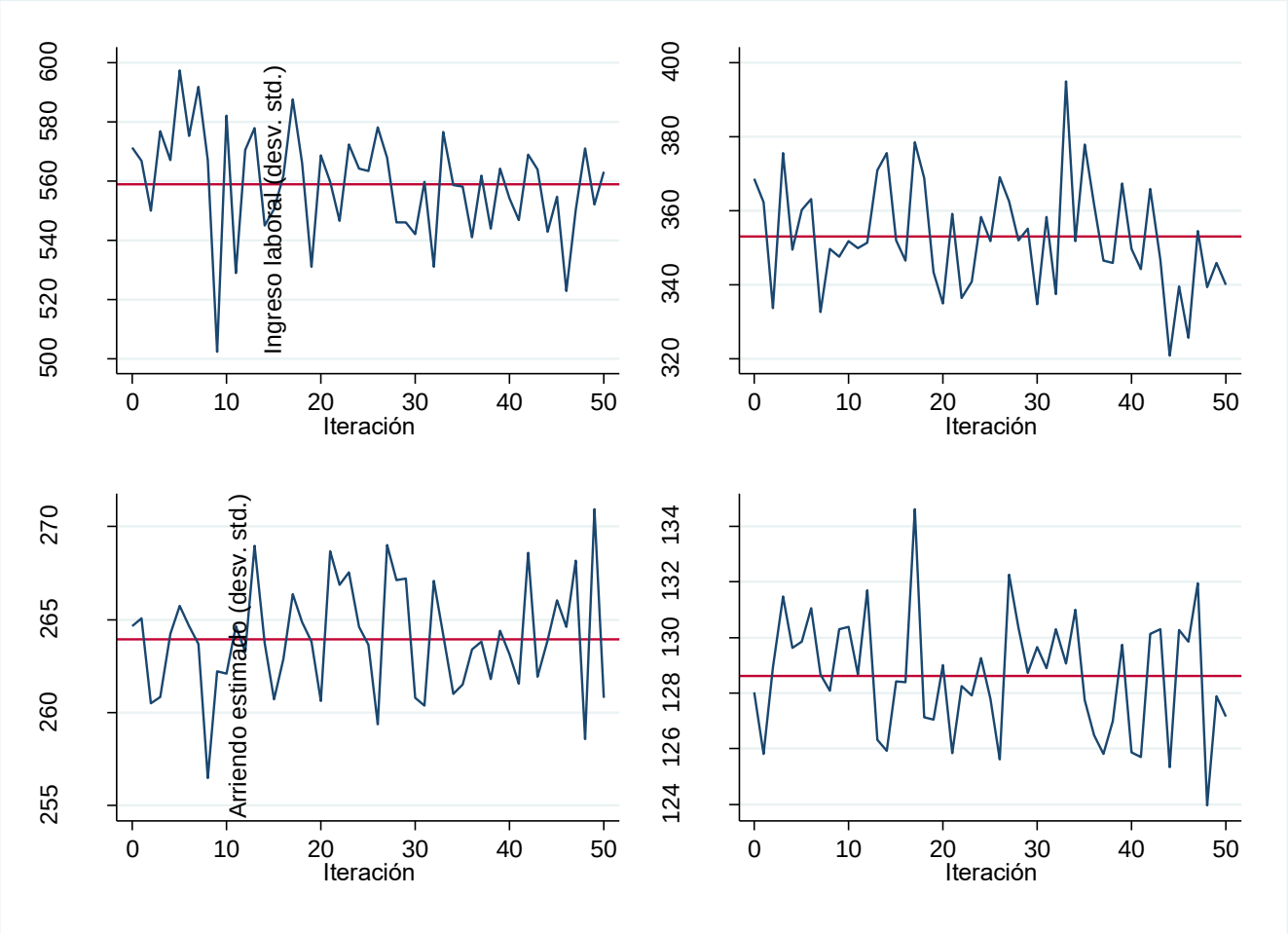
Los modelos que se emplearon incluyeron variables socio-demográficas (sexo, edad, educación, composición del hogar), del mercado laboral (CIUO, CIIU), previsionales, propiedad de la vivienda, y participación en la EPS (para controlar por posibles sesgos MNAR). Dadas las características de las dos variables imputadas, se utilizaron regresiones censuradas que establecieron las siguientes cotas:

$$20.000 < ingreso\ laboral < 7.000.000$$

$$10.000 < arriendo\ mensual < 4.500.000$$

Dado que la imputación múltiple con ecuaciones encadenadas es un proceso iterativo cuando la pérdida de datos no es monótona, se debe chequear la convergencia de las series encadenadas para asegurar que no haya comportamientos anómalos (por ejemplo, divergencia persistente de la media de las imputaciones, o cambios bruscos en su varianza). Para ello, se realizó un ejercicio de imputación simulada con 50 réplicas cuyos resultados, en términos de media y desviación standard de las imputaciones de ambas variables, se muestran en la Figura 1.

Figura 1. Convergencia de imputaciones de ingreso laboral y arriendo, encuesta Presencial



Fuente: Elaboración propia.

Si bien se ha discutido mucho el tema, no hay un acuerdo sobre el número de réplicas a utilizar en la imputación. En general, éstas van entre 5 y 20 como sugerencia, aunque algunos establecen que "más es mejor", y que debería haber tantas réplicas como celdas individuales faltantes se deban imputar. Sin embargo, los tiempos de cómputo establecen límites razonables a éstas. En el caso de la encuesta Presencial, siguiendo las sugerencias más habituales, se incluyeron 6 réplicas.[3]

Los datos completos, que se encuentran en la base de datos de imputaciones, quedaron configurados según se muestra en la Tabla 16.

[3] Agregar más replicas sólo requiere cambiar un parámetro en el modelo y correr nuevamente el software, por lo que es trivial incluir más casos. En la Encuesta Financiera de Hogares del Banco Central de Chile, por ejemplo, se emplean 30 réplicas, aunque no se ha podido justificar adecuadamente la razón de esa cantidad.

Tabla 16. Encuesta Presencial: resultado final del proceso de imputación

```
Multivariate imputation                     Imputations =        5
Chained equations                                 added =        5
Imputed: m=1 through m=5                        updated =        0

Initialization: monotone                     Iterations =       50
                                                burn-in =       10

Conditional imputation:
  ing_lab_new: no incomplete out-of-sample observations

      ing_lab_new: interval regression
     arriendo_new: interval regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
ing_lab_new	7535	265	265	7800
arriendo_new	6281	1519	1519	7800

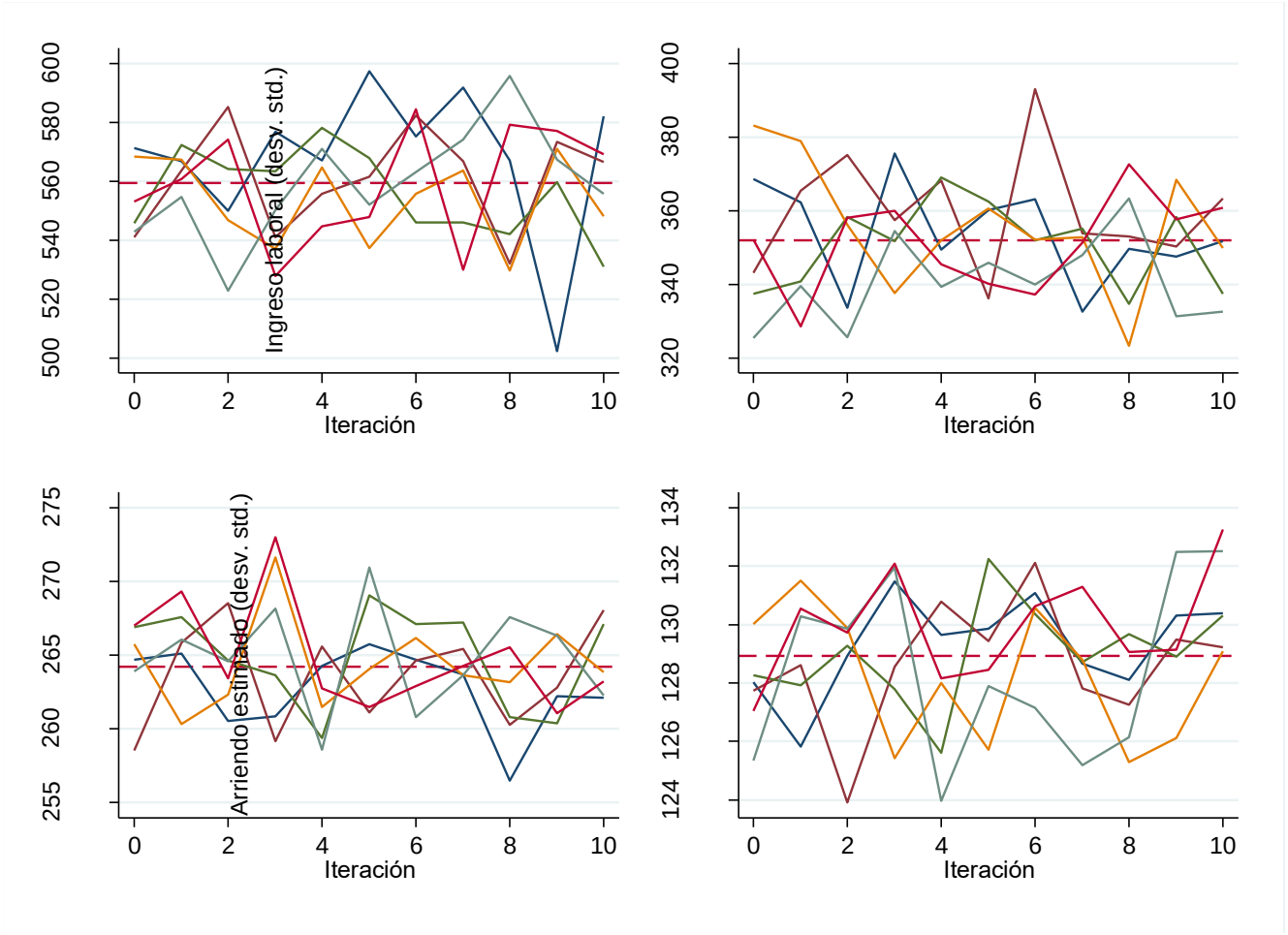
(complete + incomplete = total; imputed is the minimum across *m* of the number of filled-in observations.)

Nota: El reporte se encuentra expresado en idioma inglés, ya que es la salida original del *software* empleado para realizar las imputaciones, Stata 14.

Fuente: Elaboración propia.

El comportamiento de las cadenas de imputaciones, en donde se puede ver la convergencia de cada una de las réplicas imputadas, se puede observar en la Figura 2.

Figura 2.Comportamiento de la convergencia de las cadenas de imputación

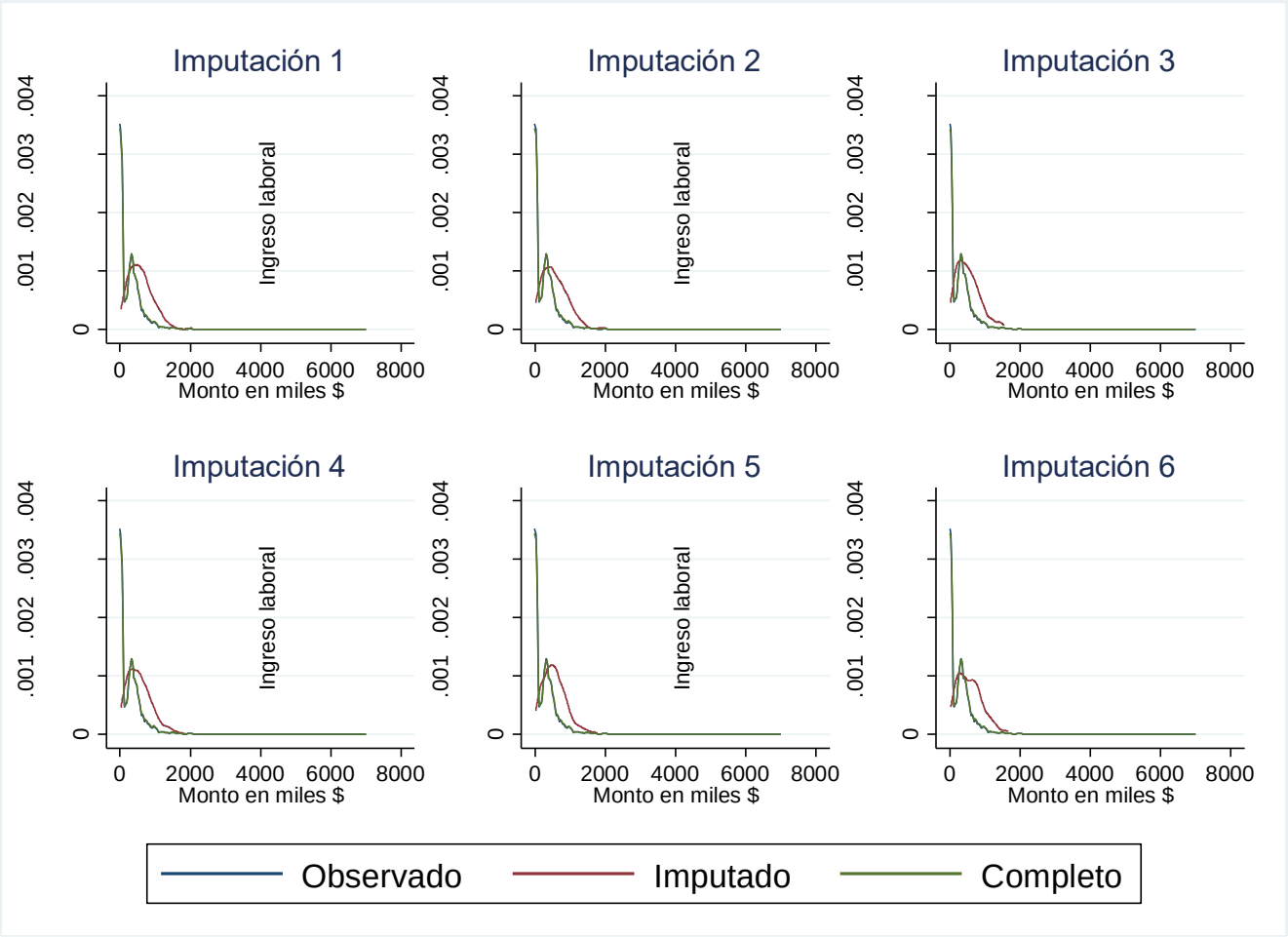


Fuente: Elaboración propia.

1.3.1.1. Imputación de ingreso promedio mensual laboral

Para el caso de los ingresos mensuales, el comportamiento de las distribuciones original, imputada, y completada para cada una de las iteraciones, tiene el e aspecto mostrado en la Figura 3.

Figura 3. Comportamiento de las imputaciones – Ingreso Laboral Promedio Mensual

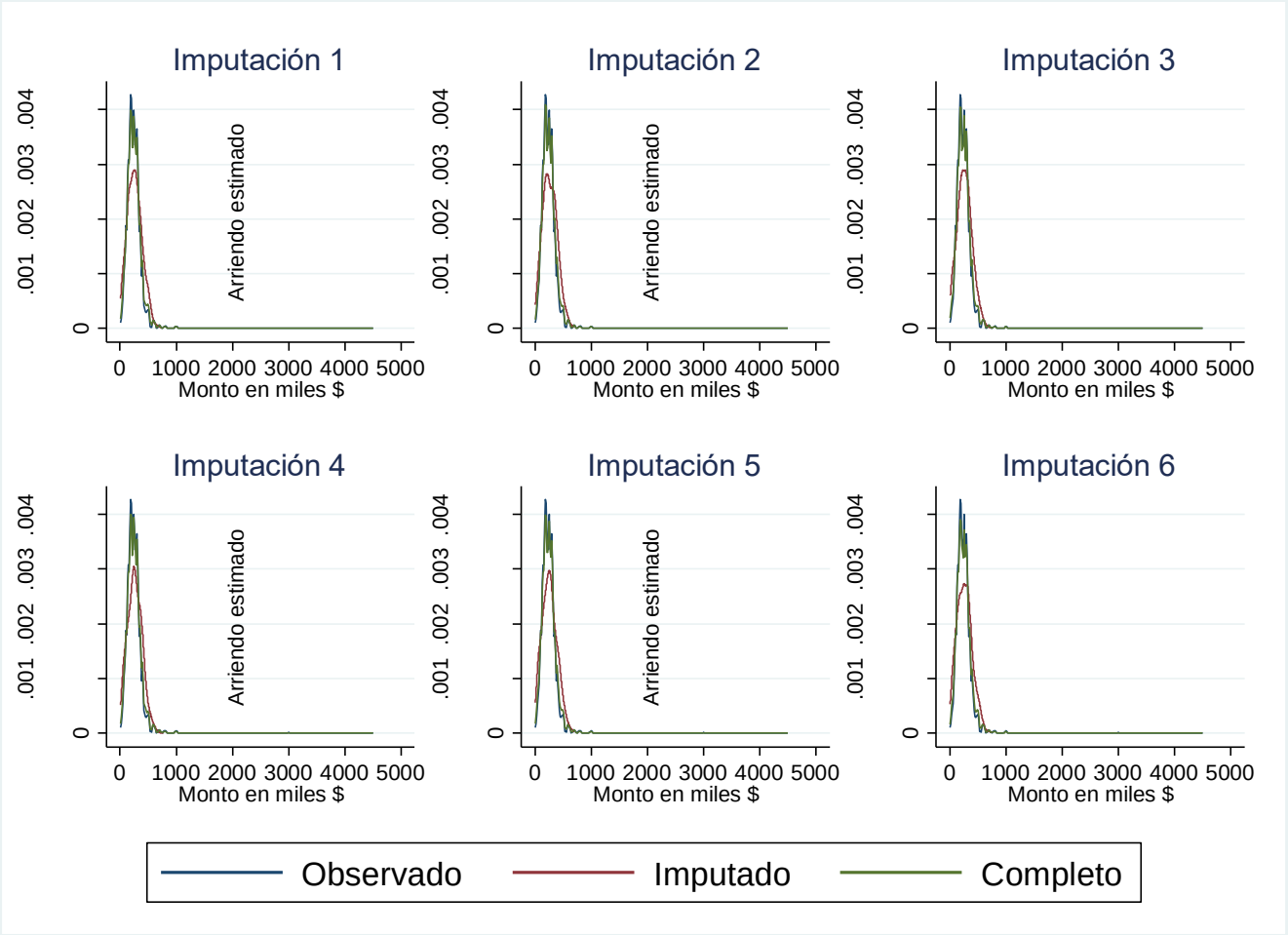


Fuente: Elaboración propia.

1.3.1.2. Imputación de arriendo

En el caso de los arriendos, las distribuciones se comportan según se muestra en la **Figura 4**.

Figura 4. Comportamiento de las imputaciones – Arriendo estimado de la vivienda particular



Fuente: Elaboración propia.

Para mostrar el efecto que tiene la incorporación de información imputada, se puede observar la **Tabla 17**. Allí, se muestra la distribución del ingreso promedio mensual y del arriendo estimado mensual obtenido con y sin ponderaciones (es decir, dependiendo de si tomamos en cuenta o no los factores de expansión de corte transversal), e incorporando o no información imputada. Nótese, por ejemplo, que los montos de ingreso mensual promedio son mayores cuando se aplican ponderaciones e imputaciones, pues una de las razones de pérdida de información es que las personas tienden a ocultar sus ingresos si estos son altos. Este fenómeno recibe el nombre de reversión a la media: las personas tienden a reportar un valor promedio para alguien en su condición, antes que el verdadero valor de la variable en cuestión.

Tabla 17. Comparación de los efectos de las ponderaciones e imputaciones – Ingreso laboral promedio mensual en la Encuesta Presencial

Método | Variable	N. Obs	Media	Error Std.	[Intervalo Conf. 95%]	
Sin Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	7535	**256.0418**	*4.399891*	247.4167	264.6668
Arriendo de vivienda (similar)	6281	**248.0221**	*1.828724*	244.4372	251.6071
Sin Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	7800	**266.3252**	*4.406151*	257.6869	274.9635
Arriendo de vivienda (similar)	7800	**251.2158**	*1.698628*	247.8781	254.5535
Con Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	7535	**306.411**	*9.792791*	286.9151	325.907
Arriendo de vivienda (similar)	6281	**256.7023**	*6.042754*	244.6721	268.7325
Con Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	7800	**319.1525**	*9.79611*	299.6352	338.6697
Arriendo de vivienda (similar)	7800	**259.14**	*5.928841*	247.3266	270.9535

Fuente: Elaboración propia.

1.3.2.Imputación en la encuesta de Re-Entrevista

En la encuesta de Re-Entrevista, se imputó el ingreso promedio mensual proveniente del trabajo. En la **Tabla 18**, se muestra que hay una pérdida de 6,7% de las respuestas (70 casos sobre 1.046).

Tabla 18.Datos faltantes de ingresos laborales en la encuesta de Re-Entrevista

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
ing_lab_1000	70		976	172	20	9000

Fuente: Elaboración propia.

El proceso de imputación MICE se realizó con 6 réplicas. Dado que se imputó una única variable, el proceso de imputación no requirió iteraciones porque la pérdida de datos es monótona. La **Tabla 19** muestra el resultado de estas imputaciones múltiples.

Tabla 19. Encuesta de Re-Entrevista: resultado final del proceso de imputación

```
Multivariate imputation                      Imputations =        6
Chained equations                                  added =        6
Imputed: m=1 through m=6                         updated =        0

Initialization: monotone                      Iterations =        0
                                                 burn-in =        0

Conditional imputation:
  ing_lab_new: no incomplete out-of-sample observations

      ing_lab_new: interval regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
ing_lab_new	2012	70	70	2082

(complete + incomplete = total; imputed is the minimum across *m* of the number of filled-in observations.)

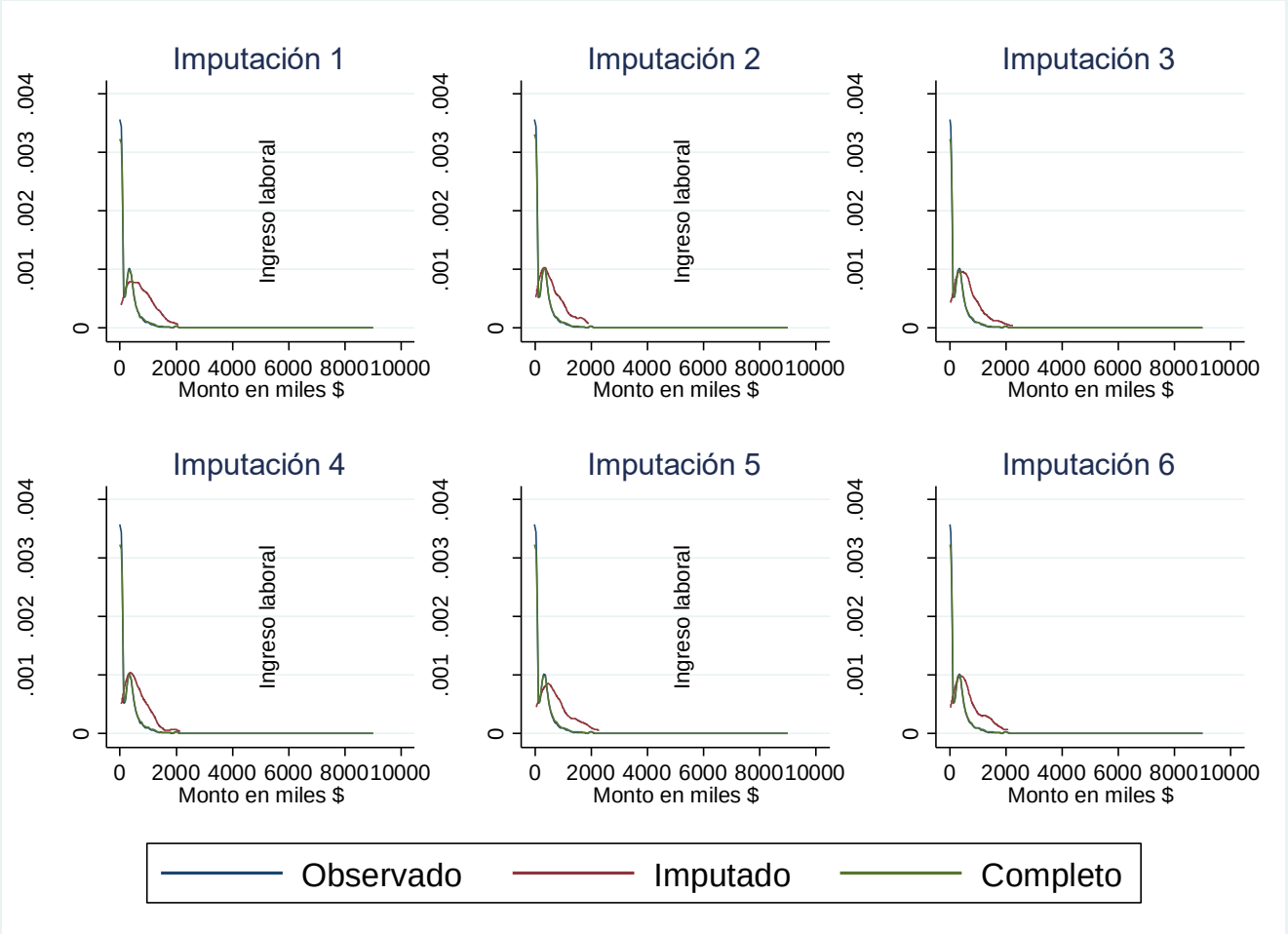
Nota: el reporte se encuentra expresado en idioma inglés, ya que es la salida original del *software* empleado para realizar las imputaciones, Stata 14.

Fuente: Elaboración propia.

1.3.2.1. Imputación de ingreso laboral mensual promedio

Las distribuciones de las imputaciones en las 6 réplicas se pueden observar en la **Figura 5**.

Figura 5. Comportamiento de las imputaciones – Ingreso Laboral Promedio Mensual, Encuesta de Re-Entrevista



Fuente: Elaboración propia.

La **Tabla 20** muestra el comportamiento de los ingresos dependiendo de si se aplican los factores de expansión (ponderaciones) y si se incluyen las imputaciones en los cálculos de la distribución de la variable.

Tabla 20. Comparación de los efectos de las ponderaciones e imputaciones – Ingreso laboral promedio mensual en la Encuesta de Re-Entrevista

Método | Variable	N. Obs	Media	Error Std.	[Intervalo Conf. 95%]	
Sin Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	2012	**234.3031**	*10.43964*	213.8294	254.7767
Sin Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	2082	**248.9334**	*10.5938*	228.1518	269.7151
Con Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	2012	**255.463**	*11.70699*	232.1514	278.7745
Con Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	2082	**279.3634**	*17.40004*	243.8448	314.8819

Fuente: Elaboración propia.

1.3.3.Imputación en la encuesta de Continuidad

La encuesta de Continuidad presentó datos faltantes en tres variables de monto importantes: ingreso laboral del entrevistado, ingreso promedio mensual del hogar y gasto mensual promedio del hogar. Las tasas de datos faltantes fueron de 7,6% para el ingreso laboral, 8,7% para el ingreso mensual promedio del hogar, y tan solo 0,6% para el gasto del hogar. Esto evidencia la sensibilidad de los entrevistados a entregar información acerca de ingresos versus gastos y, además, la dificultad para calcular el gasto del hogar, que requiere recorrer los gastos de varios ítems incluidos en la encuesta.

En la **Tabla 21**, se muestra el comportamiento de los datos faltantes en la muestra de 5.031 encuestados.

Tabla 21.Datos faltantes en la encuesta de Continuidad

Datos faltantes de monto de gasto mensual promedio del hogar: (5.031 obs.)

Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
d41_6_ajus~d	4,407		624	223	0	9000000
gasto_1000	30		5,001	>500	28	9000

Datos faltantes de monto de ingreso promedio mensual del hogar: (5.031 obs.)

			Obs<.			
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
ing_hog_1000	437		4,594	342	20	8000

Datos faltantes de monto de ingreso mensual salarial promedio: (3.257 obs.)

			Obs<.			
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
ing_lab_1000	248		3,009	236	20	8500

Nota: Las variables de monto se encuentran escaladas a miles de pesos. La variable "d41_6_ajus~d" es una variable auxiliar de la encuesta para ajuste de monto de gasto total del hogar, y se encuentra reportada sin escalar.

Fuente: Elaboración propia.

Tal como con las otras encuestas de la EPS VII Ronda, la imputación se realizó mediante MICE, empleándose 6 réplicas. Las tres variables se imputaron en conjunto con regresiones censuradas en ambas colas:

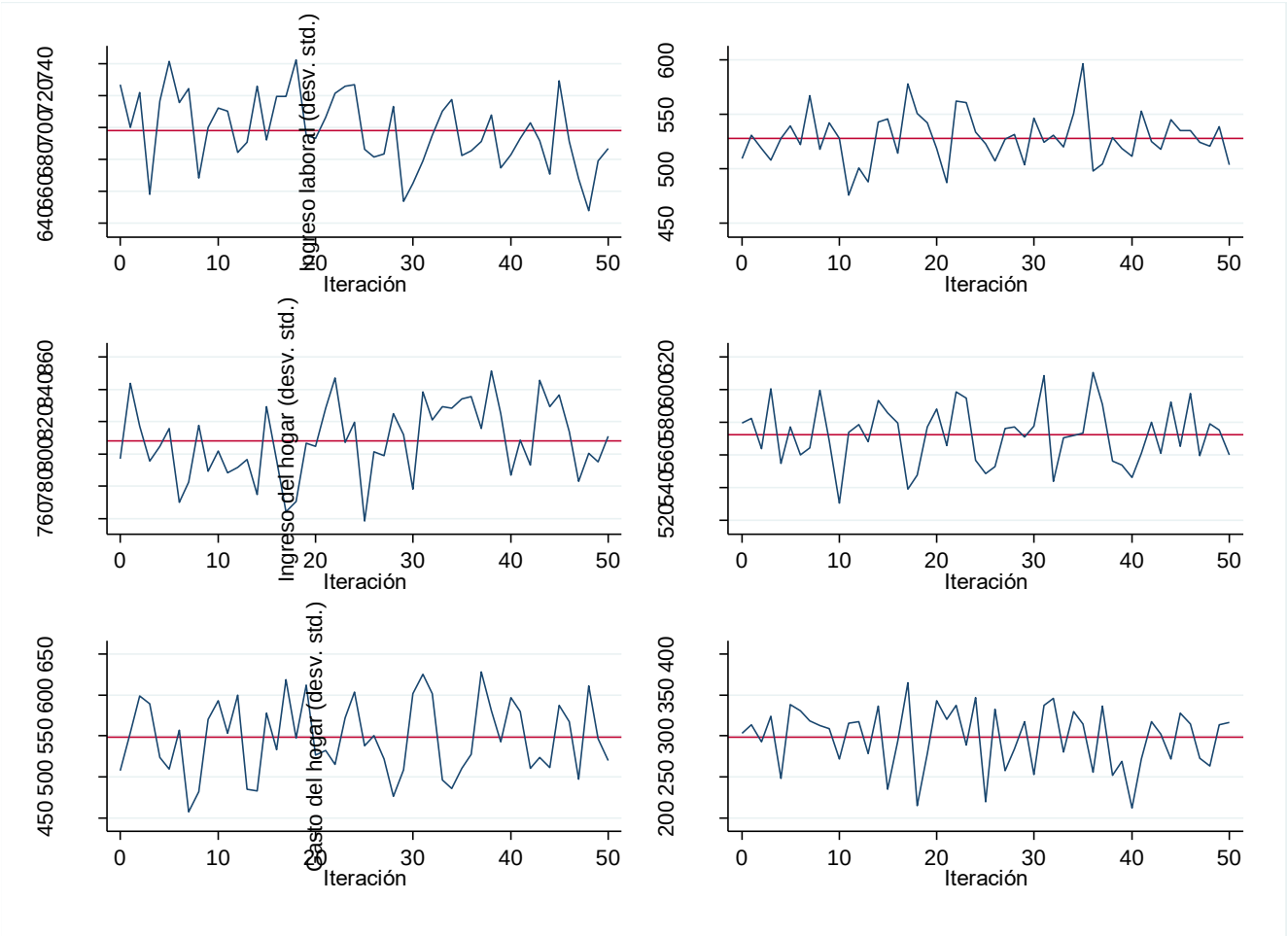
$$20.000 < ingreso\ laboral\ mensual < 10.000.000$$

$$10.000 < ingreso\ mensual\ del\ hogar < 10.000.000$$

$$10.000 < gasto\ mensual\ del\ hogar < 10.000.000$$

Para testear los modelos de imputación, se verificó el comportamiento de la convergencia en media y varianza de las cadenas de imputación, como se muestra en la **Figura 6**.

Figura 6. Comportamiento de la convergencia de las cadenas de imputación



Fuente: Elaboración propia.

La **Tabla 22** muestra el resultado del proceso de imputación para 6 réplicas.

Tabla 22. Encuesta de Re-Entrevista: resultado final del proceso de imputación

```
Multivariate imputation                      Imputations =        6
Chained equations                                  added =        6
Imputed: m=1 through m=6                         updated =        0

Initialization: monotone                      Iterations =       60
                                                 burn-in =       10

Conditional imputation:
  ing_lab_new: no incomplete out-of-sample observations

        ing_lab_new: interval regression
        ing_hog_new: interval regression
          gasto_new: interval regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
ing_lab_new	4783	248	248	5031
ing_hog_new	4594	437	437	5031
gasto_new	5001	30	30	5031

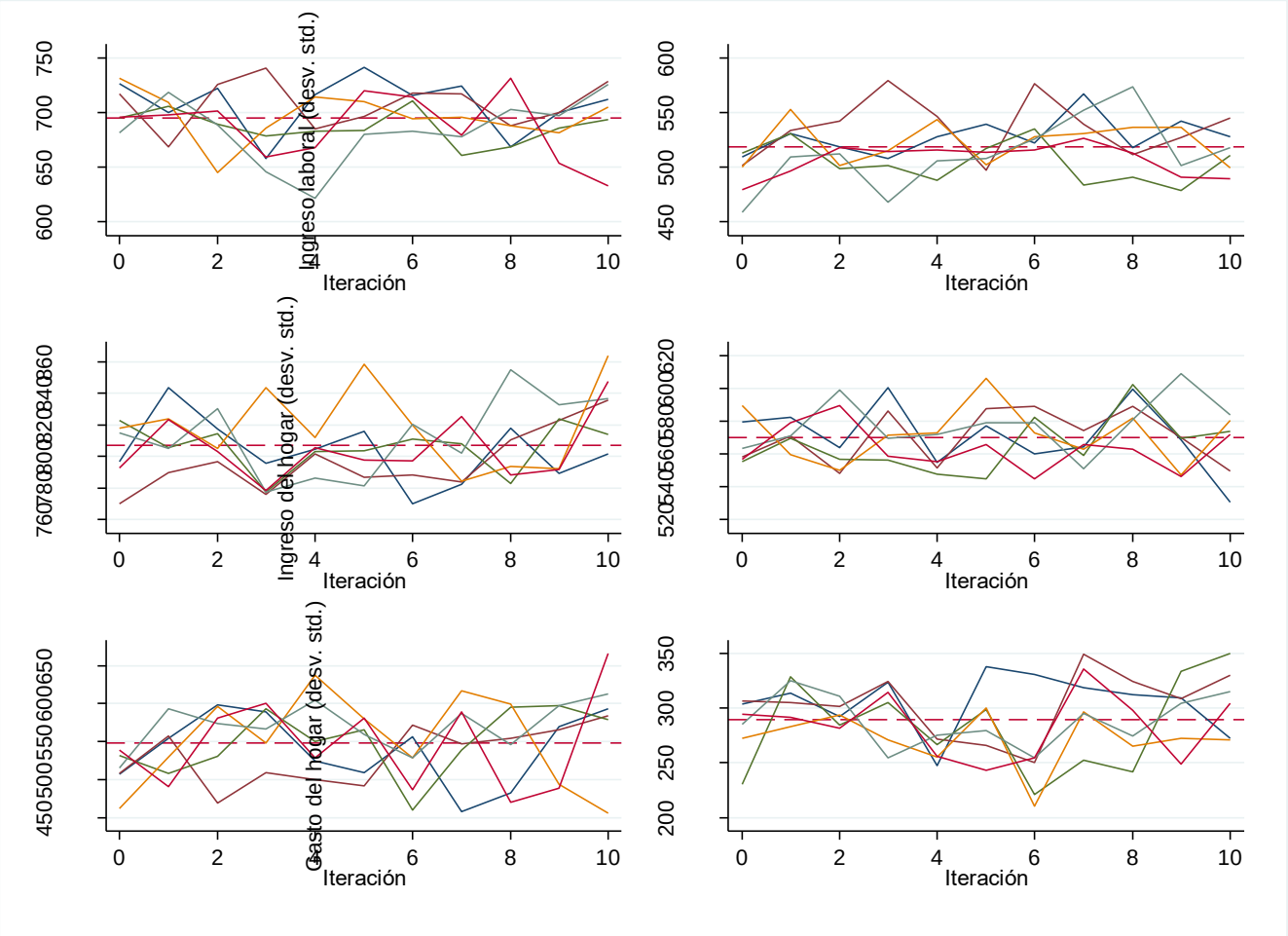
```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Nota: El reporte se encuentra expresado en idioma inglés, ya que es la salida original del *software* empleado para realizar las imputaciones, Stata 14.

Fuente: Elaboración propia.

El comportamiento de las cadenas de imputaciones para cada una de las réplicas se puede observar en la **Figura 7**.

Figura 7. Comportamiento de la convergencia de las cadenas de imputación

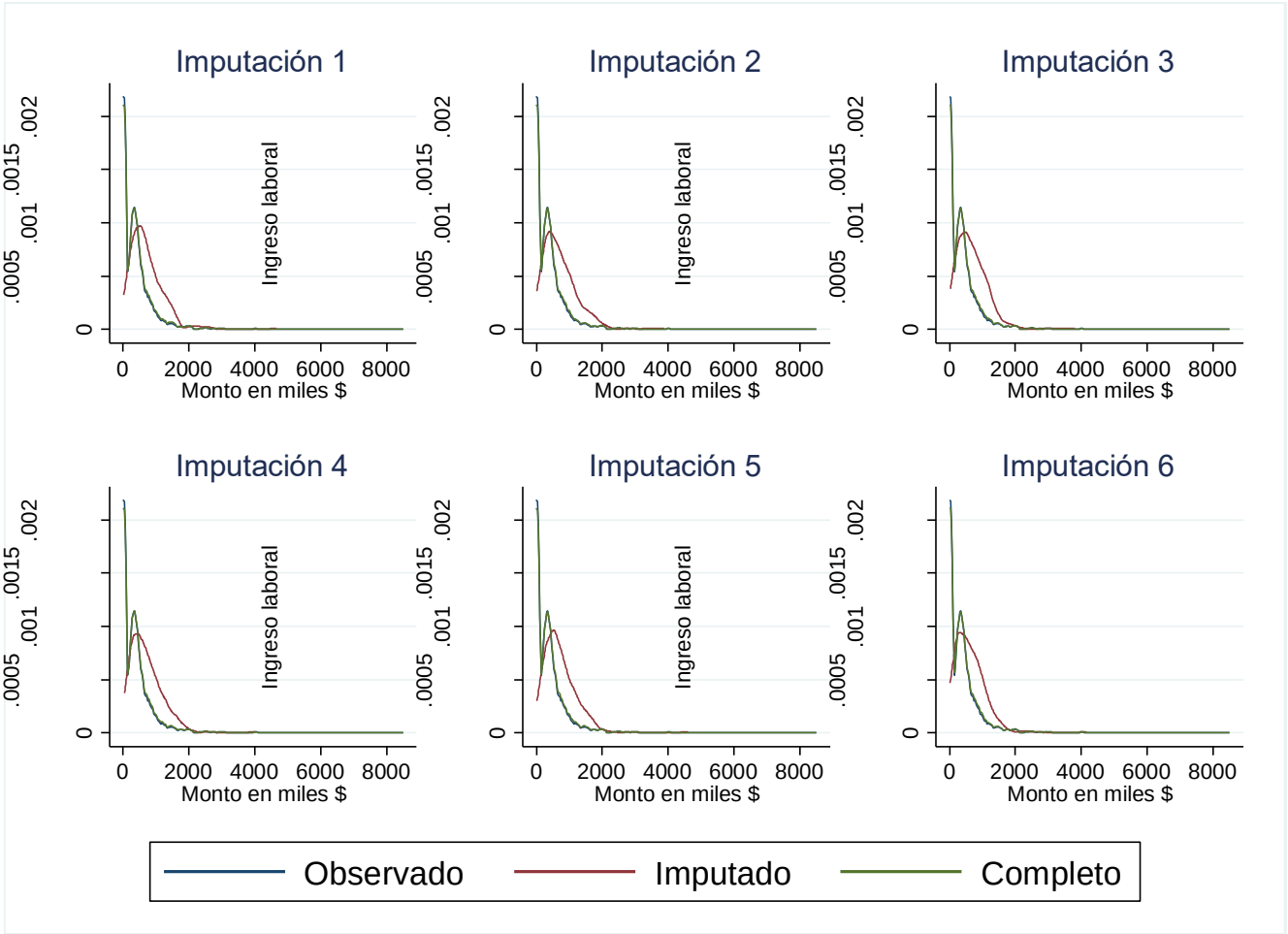


Fuente: Elaboración propia.

1.3.3.1. Ingreso laboral promedio mensual

Para el caso de los ingresos laborales mensuales, el comportamiento de las distribuciones original, imputada, y completada para cada una de las iteraciones, tiene el siguiente aspecto:

Figura 8. Comportamiento de las imputaciones – Ingreso Laboral Promedio Mensual, Encuesta de Continuidad

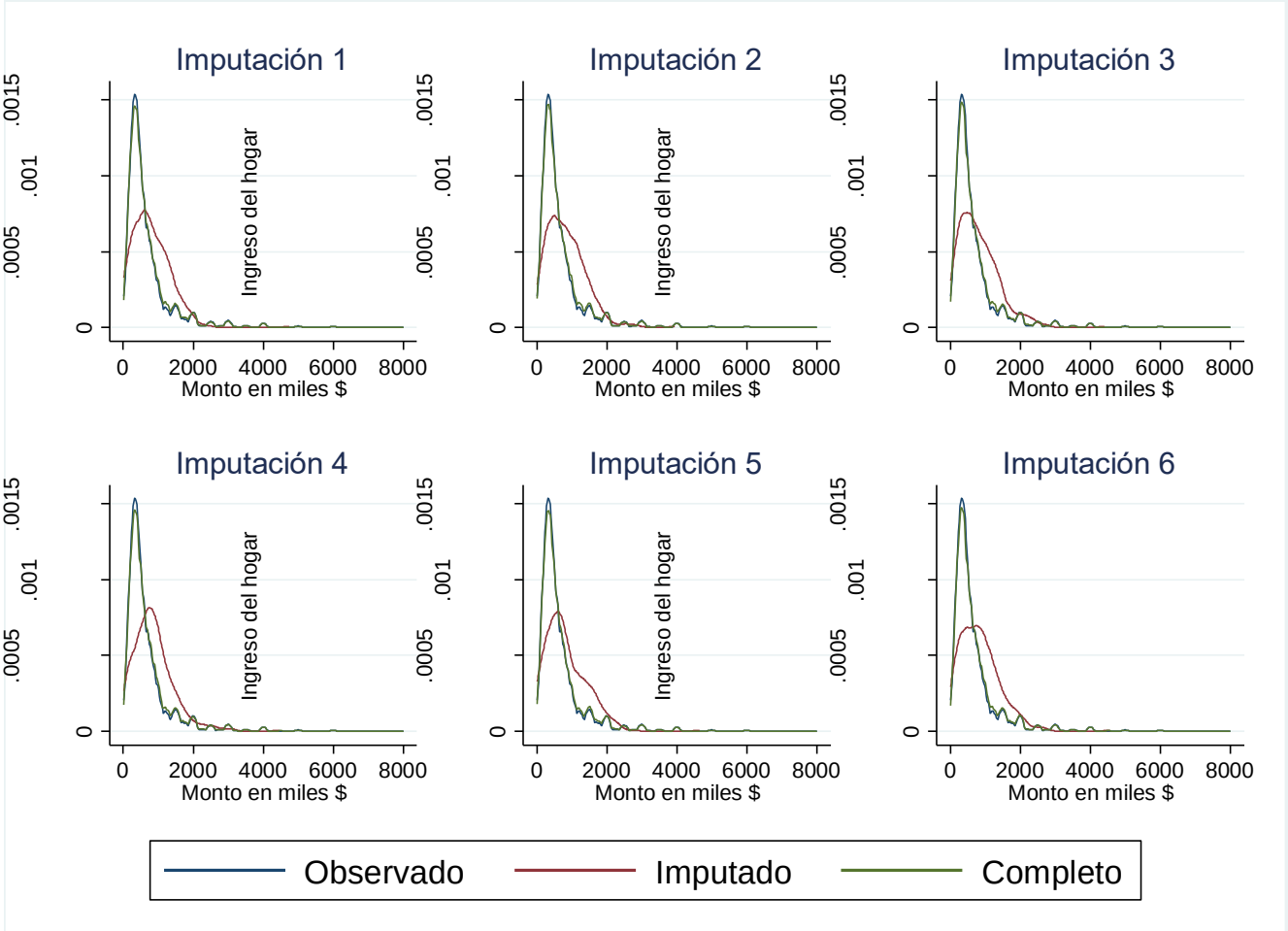


Fuente: Elaboración propia.

1.3.3.2. Imputación de ingreso promedio mensual del hogar

En el caso del ingreso mensual promedio del hogar, las distribuciones de las muestras observada, imputada y completa se comportan según se muestra en la **Figura 9**.

Figura 9. Comportamiento de las imputaciones – Ingreso mensual del hogar, Encuesta de Continuidad

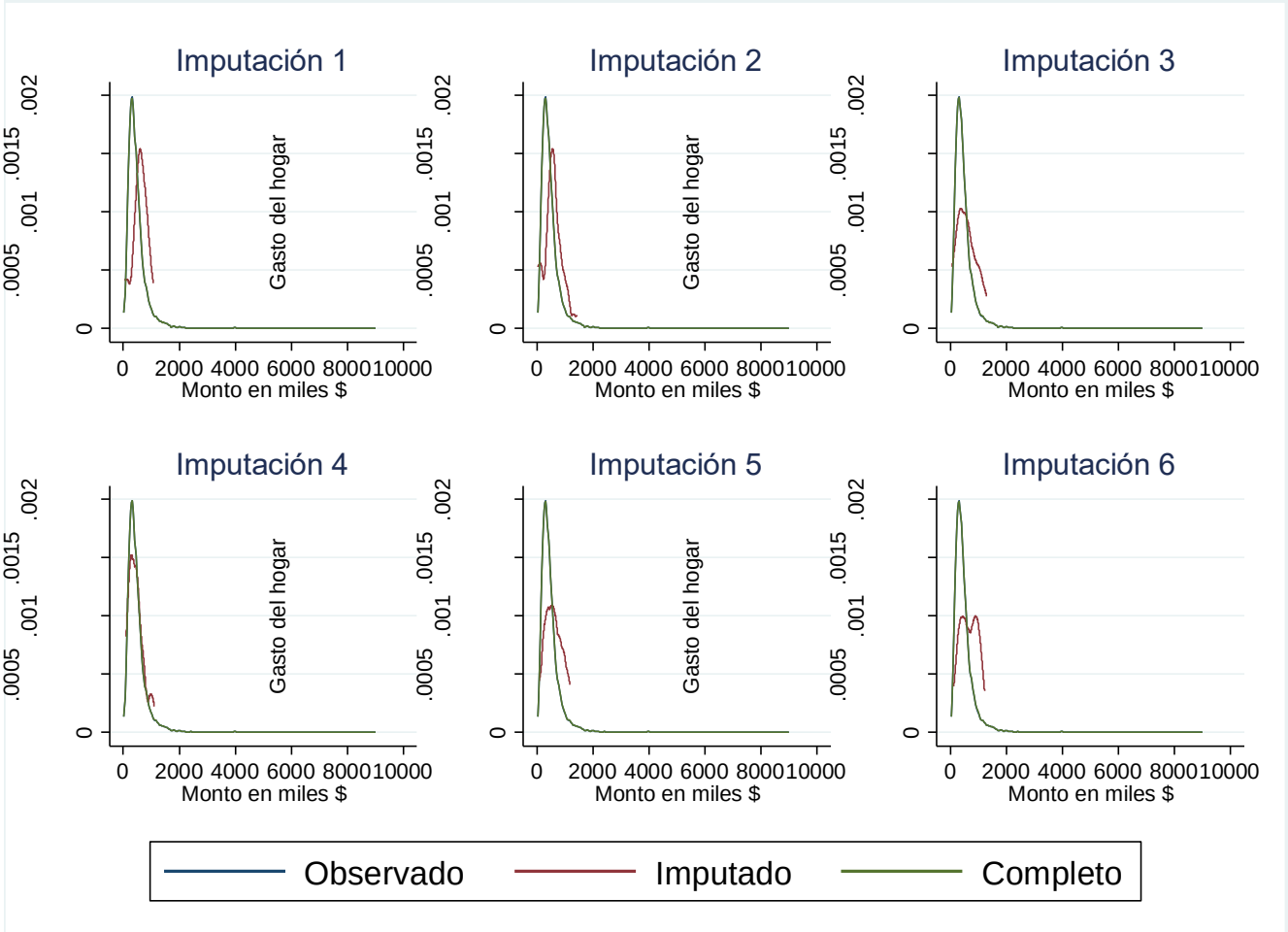


Fuente: Elaboración propia.

1.3.3.3. Imputación de gasto mensual del hogar

Por último, la **Figura 10** muestra el comportamiento de las cadenas de imputaciones del gasto mensual promedio del hogar.

Figura 10. Comportamiento de las imputaciones – Gasto mensual promedio del hogar, encuesta de Continuidad



Fuente: Elaboración propia.

Para mostrar el efecto que tiene la incorporación de información imputada, se puede observar la **Tabla 23. Comparación de los efectos de las ponderaciones e imputaciones – Ingreso laboral mensual, ingreso mensual del hogar, y gasto mensual del hogar en la Encuesta** Allí, se muestra la distribución del ingreso promedio mensual y del arriendo estimado mensual obtenido con y sin ponderaciones (es decir, dependiendo de si tomamos en cuenta o no los factores de expansión de corte transversal), e incorporando o no información imputada. Nótese, por ejemplo, que los montos de ingreso mensual promedio son mayores cuando se aplican ponderaciones e imputaciones, pues una de las razones de

pérdida de información es que las personas tienden a ocultar sus ingresos si estos son altos. Este fenómeno recibe el nombre de reversión a la media: las personas tienden a reportar un valor promedio para alguien en su condición, antes que el verdadero valor de la variable en cuestión.

Tabla 23. Comparación de los efectos de las ponderaciones e imputaciones – Ingreso laboral mensual, ingreso mensual del hogar, y gasto mensual del hogar en la Encuesta de Continuidad

Método | Variable	N. Obs	Media	Error Std.	[Intervalo Conf. 95%]	
Sin Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	4783	**357.0149**	*7.936438*	341.4559	372.574
Ingreso promedio mensual del hogar	4594	**669.6605**	*10.38746*	649.2961	690.0249
Gasto promedio mensual del hogar	5001	**469.9493**	*5.023564*	460.1009	479.7977
Sin Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	5031	**373.9147**	*8.007602*	358.2048	389.6246
Ingreso promedio mensual del hogar	5031	**683.8838**	*10.01658*	664.2375	703.5302
Gasto promedio mensual del hogar	5031	**470.6161**	*5.026525*	460.7617	480.4704
Con Ponderaciones, Sin Imputaciones					
Ingreso promedio mensual laboral	4715	**369.5309**	*15.92129*	337.8403	401.2214
Ingreso promedio mensual del hogar	4594	**699.5026**	*28.88026*	642.0179	756.9874
Gasto promedio mensual del hogar	5001	**498.3106**	*15.26459*	467.9272	528.694
Con Ponderaciones, Con Imputaciones					
Ingreso promedio mensual laboral	5031	**392.2015**	*16.71162*	358.8973	425.5057
Ingreso promedio mensual del hogar	5031	**716.9525**	*29.16189*	658.8756	775.0295
Gasto promedio mensual del hogar	5031	**498.9545**	*15.25725*	468.5736	529.3355

Fuente: Elaboración propia.

1.4. Consideraciones finales sobre imputaciones

En esta propuesta de imputaciones, se decidió abrir el camino a las imputaciones múltiples como técnica más robusta y cada vez más extendida en el tratamiento moderno de datos. El uso de esta técnica se ha hecho más sencillo dado que la mayoría de los *softwares* estadísticos incorporan procedimientos para procesar el conjunto informativo. Sin embargo, la intensidad computacional sigue siendo un tema a tener en cuenta si se lo compara con otros métodos uni- y multivariados de imputación única, como *hot deck* o medias predictivas.

La ganancia con esta metodología se encuentra en la inferencia. Mejores y más robustos indicadores permiten tener una mejor aproximación a los fenómenos, sin depender de los valores específicos de las imputaciones.

Referencias

Allison, P. D. 2001. Missing Data. Thousand Oaks, CA: Sage.

Card, D. y N. A. Smith (2015). "Automated Coding of Open-Ended Survey Responses," draft paper.

Carlin, J. B., J. C. Galati, y P. Royston. 2008. A new framework for managing and analyzing multiply imputed data in Stata. Stata Journal 8: 49—67.

Carlin, J. B., N. Li, P. Greenwood, y C. Coffey. 2003. Tools for analyzing multiple imputed datasets. Stata Journal 3: 226—244.

Dillman, D. A., & Christian, L. M. (2005). "Survey mode as a source of instability in responses across surveys," Field Methods, 17(30), 30-52.

Emde, M., & Fuchs, M. (2012). "Using adaptive questionnaire design in open-ended questions: A field experiment," Paper presented at the American Association for Public Opinion Research (AAPOR) 67th Annual Conference, San Diego, USA.

Giorgetti, D., I. Prodanof y F. Sebastiani (2002). "Automated Coding of Open-ended Surveys: Technical and Ethical Issues," International Conference on Universal Knowledge and Language. ICUKL-2002, Goa, India, 25-29 November 2002.

Kelley, K. B. Clark, V. Brown y J. Sitzia (2003), "Good practice in the conduct and reporting of survey research," International Journal for Quality in Health Care, Vol. 15(3):261–266

Lee, K. J., and J. B. Carlin. 2010. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology 171: 624—632.

Little, R. J., y D. Rubin (1987). Statistical analysis with missing data, New York, Wiley.

Little, R. J., y D. Rubin (2002), Statistical analysis with missing data (Sec. Ed.), New York, Wiley.

Little, Roderick y Donald Rubin (2014). *Statistical analysis with missing data*. John Wiley & Sons:Boston, MA.

Marchenko, Y. V., and W. D. Eddings. 2011. A note on how to perform multiple-imputation diagnostics in Stata. Webpage: http://www.stata.com/users/ymarchenko/midiagnote.pdf

Medina, F. y M. Galván (2007). Imputación de datos: teoría y práctica. Serie Estudios Estadísticos y Prospectivos, No. 54, CEPAL.

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: 473—489.

Rubin, Donald (1976). Inference and missing data. Biometrika, Vol. 581–592.

Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC.

Schonlau M. y M. P. Couper (2016). "Semi-automated categorization of open-ended questions," Survey Research Methods (2016), Vol. 10(2):143-152.

Van Buuren, S. (2017). Flexible imputation of missing data. (Sec. Ed.) Boca Raton, FL: CRC Press

Wood, Josepihne, Gregory Matthews, Jennifer Pellowski y Ofer Harel (2019). “Comparing Different Planned Missingness Designs in Longitudinal Studies.” Sankhya B, Vol. 81:226–250.

Züll, C. (2016). “Open-Ended Questions,” GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_002.