METODOLÓGICO

Documentos de trabajo

Métodos de Imputación VIII EPF: Gastos diarios e ingresos de la actividad laboral y jubilaciones

Autores:
Agustín Arce
Andrea Cárdenas
Verónica Canales
Klaus Lehmann

Nº7, marzo de 2019

Instituto Nacional de Estadísticas · Chile

INSTITUTO NACIONAL DE ESTADÍSTICAS

Morandé 801, Santiago de Chile

Teléfono: 562 3246 1000

ine@ine.cl

Facebook: @ChileINE

Twitter: @INE_Chile

Los Documentos de Trabajo del INE están dirigidos a investigadores, académicos, estudiantes y público especializado en materias económicas, y tienen como objetivo proporcionar un análisis exhaustivo sobre aspectos conceptuales, analíticos y metodológicos claves de los productos estadísticos que elabora la institución y, de esta forma, contribuir al intercambio de ideas entre los distintos componentes del Sistema Estadístico Nacional.

Las interpretaciones y opiniones que se expresan en los Documentos de Trabajo pertenecen en forma exclusiva a los autores y colaboradores y no reflejan necesariamente el punto de vista oficial del INE ni de la institución a la que pertenecen los colaboradores de los documentos. El uso de un lenguaje que no discrimine ni marque diferencias entre hombres y mujeres ha sido una preocupación en la elaboración de este documento. Sin embargo, y con el fin de evitar la sobrecarga gráfica que supondría utilizar en castellano “o/a” para marcar la existencia de ambos sexos, se ha optado por utilizar -en la mayor parte de los casos- el masculino genérico, en el entendido de que todas las menciones en tal género representan siempre a hombres y mujeres, abarcando claramente ambos sexos.

Métodos de Imputación VIII EPF: Gastos diarios e ingresos de la actividad laboral y jubilaciones

Resumen

El documento describe los procedimientos y resultados de la imputación de gastos e ingresos, realizados durante la VIII Encuesta de Presupuestos Familiares (EPF). Se exponen los principales elementos de la teoría de la no respuesta, así como los problemas asociados con la misma. Luego, se presenta la metodología utilizada para la imputación de gastos diarios en la Libreta de Gastos Individuales (LGI) y para los ingresos del trabajo y jubilaciones en la Libreta de Ingresos (LI). Los resultados dan cuenta de la capacidad de estos procedimientos en la disminución del sesgo en estimaciones claves, así como aprendizajes que pueden ser de relevancia para las discusiones sobre el problema de la no respuesta parcial en encuestas de hogares.

Palabras clave: *imputación, gastos diarios, ingresos, hot deck, factor de no respuesta, regresión de Heckman, imputación múltiple, media condicionada, no respuesta parcial.*

Abstract

The document describes the procedures and results of the imputation of expenses and income, made during the VIII Household Budget Survey (HBS). The main elements of the theory of non-response are exposed, as well as the problems associated with it. Then, is presented the imputation methodology for daily expenses in the Individual Expenses Questionnaire and the labour and pensions income in the Income Questionnaire. The results show the capacity of these procedures in the reduction of bias in key estimates, as well as learning that may be relevant for discussions about the problem of partial non-response in household surveys.

Keywords: *imputation, daily expenses, income, hot deck, non-response factor, Heckman regression, multiple imputation, conditioned mean, partial response.*

Índice general

Índice de cuadros

Índice de figuras

1 Introducción

El presente documento tiene por objetivo dar cuenta de las pruebas y análisis realizados por el Departamento de Presupuestos Familiares para la VIII EPF, destinados a evaluar diferentes métodos de imputación para ingresos y gastos diarios. Concretamente, se busca mitigar el sesgo que se podría estar ocasionando en las estimaciones de gastos e ingresos debido a la existencia de no respuesta parcial en la encuesta.

A lo largo del documento se evidencia que la no respuesta parcial en la VIII EPF, tanto en gastos como ingresos, no es aleatoria, sino que obedece a características individuales y del hogar al que pertenecen las personas. Lo anterior pone sobre relieve la importancia de incluir procesos de imputación en la encuesta, que intenten reducir el sesgo por no respuesta.

Para la imputación de gastos, los esfuerzos se centraron en realizar un tratamiento de la información faltante de gastos diarios, respecto a la cual se ha observado un aumento de la no respuesta entre la VII EPF y la VIII EPF. Dicho aumento, si no es abordado por medio de imputación, podría causar un sesgo importante por subestimación del gasto a nivel hogar.

Durante la VII EPF se realizó un tratamiento de imputación de los gastos diarios por medio de un Factor de no respuesta (FNR), que corrigió en parte la ausencia de gasto. En la búsqueda de la mejora continua de la producción estadística, se realizaron distintas evaluaciones comparativas entre el FNR y el método *hot deck*. Este último busca un donante similar en ciertas características dentro de los datos de la propia encuesta. Las pruebas realizadas permitieron determinar que el método *hot deck* logra reducir el sesgo por subestimación de forma más satisfactoria que el FNR.

En cuanto a los ingresos, se realizaron pruebas con distintos métodos de imputación para las partidas de trabajo dependiente, trabajo independiente y para los ingresos por jubilación de vejez. Estos tres componentes del ingreso fueron seleccionados para pasar por un proceso de imputación, debido a su mayor participación en el ingreso promedio de los hogares. Adicionalmente, estas partidas cuentan con patrones claros en cuanto a características sociodemográficas que los explican, lo que permite realizar una imputación considerando dichos factores.

En el marco de la VII EPF estos ingresos fueron imputados por medio del método de media condicional, el cual resolvió una parte importante de la falta de respuesta. Sin embargo, al igual que para los gastos diarios, se realizaron nuevas pruebas para la VIII EPF, que evaluaron el desempeño de cuatro métodos de imputación: media condicional, *hot deck*, regresión de Heckman e imputación múltiple. Uno de los resultados más relevantes es que el método *hot deck*, en cuanto al ingreso promedio, entrega resultados similares a la media condicional, no obstante, el primero logra reproducir de manera más exacta la distribución de los datos perdidos.

Este documento intenta avanzar hacia una mayor transparencia respecto al modo en el que se elaboran las estadísticas oficiales, poniendo a disposición de los usuarios el detalle de ciertos aspectos del procesamiento de los datos. Se espera que la documentación del proceso de imputación sea de utilidad, no solo para las siguientes versiones de la encuesta, sino que también para otras encuestas de hogares y oficinas estadísticas que se vean enfrentadas a desafíos similares.

2 Teoría de la no respuesta

En lo que respecta a encuestas, la calidad de los datos responde fundamentalmente a dos conceptos: 1) la **precisión**, es decir, la cercanía entre el valor de la estimación obtenida a partir de la encuesta y el valor que se busca medir; y 2) la **confiabilidad**, que refiere a la consistencia de los resultados tras la repetición de la medición (Eurostat, 2014). Ambos elementos se pueden ver afectados por errores que perjudican la calidad de los datos. Dichos errores pueden agruparse en dos tipos principales: **errores muestrales**, que corresponden a la diferencia entre la inferencia que se realiza a partir de una muestra y el valor poblacional; y **errores no muestrales**, que como su nombre lo indica, se producen por factores ajenos al muestreo y pueden ocurrir en las diferentes etapas de la producción estadística. De acuerdo a la clasificación propuesta por Eurostat, entre los errores no muestrales se pueden identificar: **errores de cobertura**, **errores de no respuesta**, **errores de medición** y **errores de procesamiento**. En el presente documento se revisarán los errores de no respuesta.

2.1 No respuesta a la unidad y no respuesta al ítem

El ideal en una encuesta es tener respuestas para todos sus ítems, es decir, obtener información completa a partir de las unidades seleccionadas en la muestra. Sin embargo, esto en la realidad no ocurre, siendo usual encontrar información faltante, ya sea en algunas preguntas o bien para el cuestionario completo. Esto se produce por diversos motivos, que van desde el desconocimiento por parte del entrevistado de la información que se le solicita hasta el rechazo o negativa de contestar por parte del mismo.

En términos generales, se puede distinguir entre no respuesta a la unidad (también llamada no respuesta total) y no respuesta al ítem (o parcial). La **no respuesta a la unidad** se produce cuando una unidad seleccionada (persona, hogar, empresa, entre otros) no provee la información solicitada o bien cuando reporta información, pero esta es insuficiente para ser considerada en el estudio, perdiéndose la totalidad de la información de una o varias unidades muestrales (Cobben, 2009; Medina & Galván, 2007).[1]

En relación con este último punto, en la VIII EPF se utilizaron criterios mínimos de calidad y cantidad que debía cumplir la información recopilada para ser considerada en los resultados finales de la encuesta. Para ello se utilizó un método de evaluación de mínimos de calidad (en adelante *grilla técnica*). La idea detrás de este método era poder evaluar e identificar los tipos de no respuesta y así distinguir entre niveles de información aceptables para clasificar a un hogar como entrevistado versus aquellos niveles no aceptables, que debían ser considerados como no respuesta a la unidad. Las entrevistas con información insuficiente (según los criterios establecidos) fueron clasificadas como *break-off* y pasaron a formar parte de los rechazos. La no respuesta a la unidad se modifica

[1] En la VIII EPF la no respuesta a la unidad se producía, entre otros motivos, cuando existía un rechazo explícito desde el inicio, y por lo tanto el hogar no contestaba ninguno de los cuestionarios de la encuesta, o bien cuando existía un rechazo implícito, es decir, cuando el hogar aceptaba contestar la encuesta, sin embargo, su información era insuficiente. Estos últimos casos se producían, ya sea porque la entrevista se interrumpía antes de ser concluida (es decir, no se completaban las 4 visitas contempladas en la aplicación de la encuesta) o porque la información proporcionada no cumplía con los estándares mínimos de calidad.

eliminando las observaciones (que quedan en esta categoría) y ajustando los factores de expansión (Medina & Galván, 2007).[2]

Por su parte, la **no respuesta al ítem** sucede cuando una unidad seleccionada responde, pero falta información para algunas de sus preguntas, ya sea porque se negó a responderlas o bien porque no contaba con la información requerida (Cobben, 2009; Medina & Galván, 2007). En estos casos existe falta de respuesta, pero los niveles de información son suficientes para ser considerados en la encuesta, de acuerdo a los criterios de calidad y cantidad exigidos por la *grilla técnica*.

En la VIII EPF se utilizaron seis instrumentos de recolección de datos o **libretas**,[3] las que comprendían distintos apartados que agrupaban preguntas vinculadas por temáticas (módulos). En relación con lo anterior, la no respuesta se observa en diferentes niveles: a nivel de pregunta, apartado o libreta. La falta de información en alguno de estos niveles se consideró como no respuesta al ítem, pudiendo existir desde la falta de respuesta a una pregunta, hasta la falta de respuesta de alguna libreta completa (cuestionario). De acuerdo a los criterios establecidos por la *grilla técnica* de la VIII versión de la encuesta, las únicas libretas que no podían faltar de forma completa eran la Libreta de Gastos Individuales (LGI) y la Libreta de Gastos del Hogar (LGH), ya que la ausencia de alguna de ellas dejaba a esa unidad muestral con un puntaje por debajo del mínimo exigido para ser considerado en los resultados finales. Esto implicaba caer en la clasificación de *break-off* y, por ende, en la categoría de no respuesta a la unidad.[4]

2.2 Problemas asociados con la no respuesta

La no respuesta es un aspecto importante a considerar en las encuestas, ya que puede producir efectos no deseados en los resultados. Dos de los más importantes son el aumento del error muestral y problemas de sesgo en las estimaciones (Cobben, 2009; Groves & Peytcheva, 2008).

En la literatura hay consenso sobre la importancia de medir y mitigar la información faltante (Cobben, 2009; Eustat, 2008; Groves, 2006), por lo que las oficinas estadísticas, en general, toman medidas para atenuar la no respuesta, sin embargo, no es posible eliminarla por completo. En aquellos casos donde persiste, los esfuerzos apuntan a corregirla mediante procedimientos estadísticos.

La no respuesta a la unidad usualmente se aborda mediante ajustes en los factores de expansión, mientras que la no respuesta al ítem se corrige mediante métodos de imputación. Para este último punto es importante tener presente que, bajo ninguna circunstancia, un dato imputado será mejor que un dato observado (Medina & Galván, 2007), pero ante la imposibilidad de eliminar la no

[2] Para el caso de la VIII EPF se realizaron en forma secuencial cuatro ajustes: 1) Ajuste por omisión de conglomerados (manzanas); 2) Ajuste por elegibilidad desconocida y no elegibilidad; 3) Ajuste por no respuesta; 4) Ajuste por stock poblacional. El **ajuste por no respuesta** busca redistribuir el peso de los factores de aquellas viviendas que no respondieron entre aquellas que sí lo hicieron. El ajuste de factores de expansión por no respuesta a la unidad se desarrolla con mayor profundidad en el documento Metodología VIII EPF, mientras que la no respuesta al ítem se aborda en el Informe de Calidad VIII EPF. Ambos documentos se encuentran disponibles en www.ine.cl/epf.

[3] Para mayores detalles sobre los seis cuestionarios de la encuesta revisar el documento Metodología VIII EPF, disponible en el mismo sitio mencionado anteriormente.

[4] Para más información sobre la *grilla técnica* revisar el documento Reclasificación de la no respuesta: distinción entre la no respuesta al ítem y no respuesta a la unidad, disponible en www.ine.cl/inicio/documentos-de-trabajo.

respuesta, la imputación contribuye a reducir la posibilidad de sesgos y a evitar los errores de interpretación.

Con el objeto de comprender cómo opera el sesgo de no respuesta, Bethlehem (1988) propone una expresión sencilla para un estimador de la media:

$$Sesgo(\overline{y_r}) = \frac{Cor(y, \rho)S(y)S(\rho)}{\overline{\rho}} \quad (1)$$

Donde ρ representa la probabilidad o propensión a responder e y es la variable objetivo de la encuesta, mientras que Cor y S corresponden a la correlación y a la desviación estándar respectivamente. Es fácil ver que entre más grande sea la correlación entre ambas variables, mayor será el sesgo. Dicho de otro modo, se producirá sesgo cuando la variable objetivo defiere entre quienes responden y quienes no responden (Groves & Couper, 1998).

Otro término relevante de la expresión anterior es el promedio de la propensión a responder ρ. Este valor puede ser estimado de manera insesgada a partir de la tasa de respuesta de la encuesta. Se deduce que la tasa de respuesta opera como una protección ante el riesgo de sesgo, ya que cuando existe correlación entre y y ρ, una tasa de respuesta elevada contribuye a mitigar el sesgo.

Un tercer factor relevante es la desviación estándar de la probabilidad de respuesta $S(\rho)$. Entre mayor dispersión tenga esta, mayor será el sesgo. Nótese que cuando todas las unidades muestrales tienen la misma probabilidad de responder, es decir, cuando $S(\rho) = 0$, no existe sesgo de no respuesta en el estimador de la media. Bethlehem se refiere a la propensión a responder como a una "fuerza" que altera el diseño muestral. El razonamiento es que el diseño muestral opera sobre la base de ciertas probabilidades de selección de las unidades, que no necesariamente son iguales, pero sí son conocidas. Esto último permite desarrollar los factores de expansión y finalmente hacer inferencia sobre la población objetivo. Ahora bien, cuando las unidades seleccionadas tienen distintas propensiones a responder, determinadas por la edad, el sexo, la escolaridad, el ingreso o por cualquier otra variable sociodemográfica, las probabilidades de selección que inicialmente tenía cada unidad se distorsionan, lo cual obliga a introducir ajustes.

Es importante señalar que la correlación entre y y ρ es un fenómeno complejo que no necesariamente afecta del mismo modo a todas las variables de interés de una encuesta. Adicionalmente, los componentes de la no respuesta (no contacto, rechazo u otros) no necesariamente se correlacionan de la misma manera con lo que la encuesta desea medir. A modo de ejemplo, podría darse el caso de que el rechazo se comporte de manera totalmente aleatoria, mientras que el no contacto muestre algún patrón sistemático que se relacione con la variable de interés

Sobre la base de lo dicho, es posible establecer algunas observaciones importantes:

- Una alta tasa de respuesta constituye una protección contra el sesgo. Es posible que la propensión a responder y la variable objetivo estén altamente correlacionadas, pero si la tasa de respuesta es elevada, el sesgo será mínimo. En ese sentido, la primera estrategia para reducir el sesgo siempre debe ser procurar que los informantes colaboren con el estudio durante el trabajo de campo.

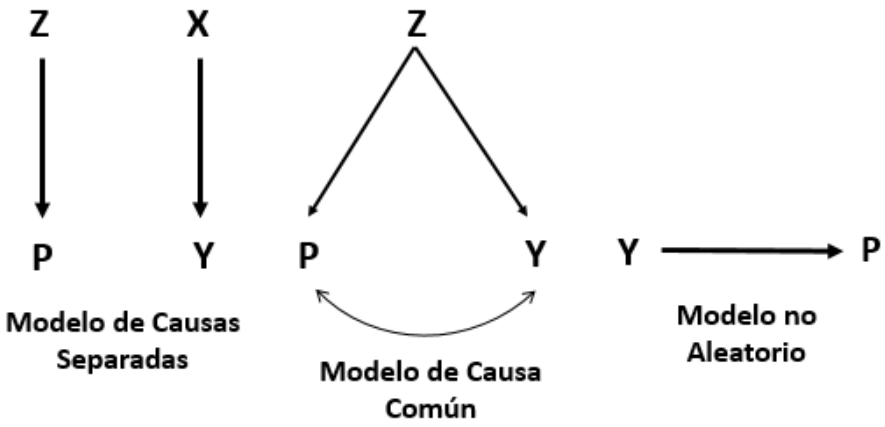
- La no respuesta no necesariamente es un problema desde el punto de vista del sesgo. De hecho, es posible obtener estimadores insesgados incluso en contextos de baja tasa de respuesta, si es que no existe correlación entre la variable objetivo y la propensión a responder.
- La no respuesta no necesariamente tiene el mismo efecto sobre todas las estimaciones de una encuesta. Por ejemplo, para una encuesta que pretenda estimar la media de dos variables, digamos X e Y, la no respuesta podría generar sesgo en $\overline{X}$ y no en $\overline{Y}$.

2.3 Mecanismos generadores de no respuesta

De acuerdo a (Medina & Galván, 2007), para elegir el método de imputación se debe tener en cuenta el comportamiento de los datos omitidos. La teoría de imputaciones se basa en el supuesto de aleatoriedad de la no respuesta, ya que, en caso de no ser aleatoria los resultados de la encuesta pueden presentar sesgos y, por lo tanto, la no respuesta no puede ser ignorada.

En 1976 Rubin introduce la noción de **mecanismo de no respuesta**, a partir del cual distingue tres tipos de datos faltantes, según el patrón de aleatoriedad que éstos sigan (ver figura 1). En primer lugar, se encuentran los datos faltantes tipo MCAR (*missing completely at random*), o lo que Groves y Peytecheva (2008) denominan "modelo de causas separadas". En este caso la propensión a responder P no depende de los datos observados X, sino de un vector Z. En ese sentido, las variables que explican la propensión a responder son distintas de las que explican la variable de interés Y. Esto ocurre cuando todas las unidades muestrales tienen la misma propensión a responder.

Figura 1: Mecanismo de no respuesta



Fuente: Traducción de Groves y Couper (1998), página 651.

En segundo lugar, Rubin distingue datos perdidos que siguen un mecanismo tipo MAR (*missing at random*). Esta situación es descrita por Groves y Peytecheva como "modelo de causa común". En este caso, tanto la variable de interés como la propensión a responder dependen del mismo vector Z. Este es el supuesto más común en lo que refiere a encuestas de hogares y ocurre cuando un mismo conjunto de variables explica tanto la no respuesta como la variable de interés.

En tercer lugar, se encuentran los datos perdidos tipo MNAR (*missing not at random*), donde la variable de interés determina la propensión a responder, por ende, la no respuesta adquiere un carácter no aleatorio (Gómez, Palarea, & Martín, 2006; Medina & Galván, 2007). En estos casos, el mismo fenómeno que se pretende medir influye en la propensión de las personas a responder.

La importancia de considerar los mecanismos de no respuesta radica en cada uno de ellos implica alternativas de imputación distintas. Para el caso de la VIII EPF se realizaron imputaciones para las variables de ingreso y de gasto. El cómo se seleccionaron los métodos y su aplicación es lo que se revisará a continuación en el presente documento.

3 Libreta de Gastos Individuales (LGI)

3.1 Contexto

3.1.1 Características generales de la Libreta de Gastos Individuales

La Libreta de Gastos Individuales (en adelante, LGI) es uno de los instrumentos de recolección de gastos de la VIII EPF. Esta libreta se entrega a cada integrante del hogar de 15 años o más y en ella se deben registrar los gastos personales realizados con frecuencia diaria.

Este es el único instrumento autoadministrado de la encuesta y se utiliza por los informantes como un cuadernillo en el que día a día registran los gastos realizados durante la quincena de aplicación. Los gastos deben ser registrados sin importar si estos se realizan para beneficio propio, para otro miembro del hogar o como regalo a una persona ajena al hogar.

Una característica importante de este instrumento es que posee una página para cada día de la quincena de colaboración, lo que hace posible identificar la falta de información para cada día por separado. Esta particularidad permite evaluar la no respuesta en las libretas parcialmente contestadas (con información en algunos días, pero no en todos) y diferenciarlas de las libretas rechazadas (aquellas en las que todos los días se encuentran sin registro).

3.1.2 Experiencia internacional respecto a la no respuesta en gastos diarios

En general, las encuestas de gastos e ingresos contemplan la captura de los gastos diarios, sin embargo, los países optan por diferentes modalidades respecto al instrumento a utilizar, lo que abre espacio a distintas estrategias para abordar la no respuesta.

A continuación, se presentan algunas experiencias internacionales que se han podido recabar en relación con la imputación de gastos diarios, de tal manera de entregar un panorama general de las prácticas de otras oficinas estadísticas. Cabe mencionar que el nivel de detalle con el que se aborda cada país depende de la información que fue posible obtener desde la documentación que cada oficina estadística pone a disposición.

- **Reino Unido**: cuenta con diarios individuales de registro para los gastos diarios. Respecto a estos, realiza imputación de aquellos rechazados, bajo la condición de contar con el diario del *main diary keeper*, correspondiente a la persona que realiza la mayor cantidad de compras del hogar. Si esta condición es satisfecha, se busca una persona de otro hogar, que guarde similitud en distintas variables sociodemográficas con el informante con el diario faltante (receptor). Al encontrar dicho donante, se copia toda la información de gasto de dicho diario hacia el receptor. El donante concreto que se selecciona en cada caso corresponde a quien logre mayor puntaje en una evaluación de similitud, que puntúa la coincidencia en las siguientes características: la edad del informante (8 puntos), relación de parentesco con la persona definida como representante del hogar (4 puntos), condición de actividad económica (2 puntos) y mes de aplicación de la encuesta (1 punto) (Bulman & Carrel, 2017).

- **Francia**: en cuanto a la recolección de gastos diarios, esta se realiza por medio de diarios individuales. Y la imputación de gastos faltantes se realiza por medio del método de *hot deck* aleatorio. Tanto para los montos faltantes de los cuestionarios a nivel hogar, como para los diarios de registro individual. En primer lugar, se particionan o agrupan a los donantes con características similares, para, posteriormente, dentro de cada grupo elegir de forma aleatoria al donante (INSEE, 2014).
- **Canadá**: para el registro de gasto diarios cuenta con un diario por hogar. Los diarios rechazados o que no cumplen con mínimos de calidad son excluidos de las estimaciones. Para el caso de la no respuesta parcial de los diarios de gasto se utilizan dos estrategias. A nivel de días sin registro, se utiliza un factor de ajuste por no respuesta, que consiste en un multiplicador que expande los montos de gasto registrados en diarios con días faltantes de registro. Dicho multiplicador es el cociente entre los días totales a registrar y los días con registro (Laperrière, 2015). En segundo lugar, se realizan imputaciones a nivel de registro, las cuales consisten en asignar un valor cuando existen montos faltantes o cuando se deben desglosar registros que se encuentran agregados. En estos casos, se utiliza una imputación por la metodología del vecino más cercano, que implica la utilización de los datos de otro respondiente con similares características y obedecen a aquellas variables correlacionadas con la variable a imputar, tales como el ingreso del hogar, tipo de vivienda, número de adultos y niños. Adicionalmente, se realizan imputaciones para mejorar el nivel de detalle en la codificación de los registros informados, cuando no cumplen con la especificidad requerida por la encuesta. En estos casos, la imputación se realiza a nivel de registro y se consideran características tales como el costo, ingreso del hogar, tamaño del mismo, provincia y barrio (Laperrière, 2015).
- **Estados Unidos**: cuenta con un instrumento autoadministrado llamado *Diary Survey*, que permite capturar los gastos diarios frecuentes durante dos semanas consecutivas en el hogar. De dicho instrumento se excluyen los gastos poco frecuentes y periódicos, que se capturan por medio de entrevista (DOL, 2011). Respecto a la imputación de los gastos diarios, si algún diario se encuentra vacío o provee un pequeño número de gastos durante la semana de referencia, se tratan como diarios no respondientes y se excluyen del procesamiento. La única imputación realizada respecto a los diarios es la de completar información categórica faltante de gastos (por ejemplo, la presentación en la que viene un bien), no así los montos faltantes.
- **Australia**: los gastos diarios se capturan por medio de diarios individuales de registro. Los datos faltantes en estos diarios fueron imputados en su totalidad, con excepción de los perceptores principales de ingresos. Para estos hogares, cualquier valor faltante fue imputado completándolo con el valor reportado por otro informante, al que se le llama donante. Los posibles donantes fueron aquellos respondientes con información completa. Para completar los registros faltantes se buscó dentro del *pool* de donantes a la persona con similitudes con el receptor de la donación en cuanto a ciertas características como la locación geográfica, el sexo, la edad, la condición de actividad económica y sus ingresos. Se indica que, dentro de lo posible, la información imputada es un *proxy* apropiado para la información faltante. En cuanto a la selección del donante concreto en cada caso entre los respondientes similares, este fue elegido de forma aleatoria desde el *pool* de individuos con información completa (ABS, 2017).

- **España**: para la captura de gastos diarios, España cuenta con las libretas individuales, donde los miembros del hogar de 14 años o más registran sus gastos durante la primera semana de aplicación, exceptuando al encargado de la administración del hogar, que realiza sus registros individuales en la libreta del hogar. Para el caso de las libretas individuales no recogidas o sin registros (libretas completamente rechazadas) se realiza una imputación en dos etapas. Primero, es imputado el nivel de gasto (monto total por libreta). Este proceso consiste en asignar como gasto total de la libreta rechazada el gasto promedio de las libretas efectivamente recogidas, considerando la creación de ciertos grupos, que funcionan como donantes. Dichos grupos obedecen a las siguientes variables: comunidad autónoma, trimestre, situación en la relación con la actividad y nivel de estudios. La segunda etapa consiste en la imputación de la estructura del gasto asociada al monto ya imputado en la primera etapa. En ella se definen grupos en función del sexo y el tramo etario (14 a 30, 31 a 60 y más de 60 años), por comunidad autónoma y trimestre[5], para luego calcular la estructura media de gasto a nivel de estos grupos construidos (INE España, 2016).
- **Brasil**: en este caso, la captura de gastos diarios se realiza por medio de diarios de gastos diarios a nivel hogar, mientras que el registro de gastos individuales se realiza por medio de recuerdo. A este respecto, no se encontró referencia a imputación de montos (gasto), sin embargo, la variable cantidad faltante o inválida es imputada a través de la metodología de *hot deck*. Para esto, se crea una matriz de similitudes formada por variables consideradas correlacionadas con la variable cantidad, de tal manera de agrupar posibles donantes en cuanto a similitudes en características de los informantes (donante y receptor). Luego de encontrar un grupo similar respecto a las variables sexo, grupo de edad, unidad geográfica y unidad de medida, se realiza una selección aleatoria del donante concreto, de tal manera de prevenir posibles distorsiones en la distribución de los valores de la variable cantidad consumida (IBGE, 2011).

3.1.3 Definición de la no respuesta en Libreta de Gastos Individuales (LGI)

La LGI es la libreta con mayor presencia de no respuesta parcial en la encuesta. Aquella ausencia de información se puede analizar: a nivel de días de registro y a nivel de cuestionario en su totalidad.

- **A nivel de día**

Para cada día en el que los informantes registran sus gastos diarios en la LGI, el trabajo de campo puede arrojar tres resultados: **1) días con gasto**, **2) días sin gasto** y **3) días sin registro**. El primer caso muestra un día que registra al menos un gasto realizado por el informante, mientras que el segundo caso corresponde a un día en el que el informante declara no haber realizado ningún gasto. Estos dos primeros casos corresponden a días con registro, es decir, se trata de registros válidos. En consecuencia, la no respuesta en la LGI corresponde a los días sin registro (tercer caso), es decir, a días en los que no se cuenta con información y que, por ende, no se tiene conocimiento de qué es lo que ocurrió (en cuanto a si se realizó o no gasto) y cuyo tratamiento se presenta más adelante.

[5]Se indica que estas variables fueron elegidas ya que explicaban en mejor medida la estructura del gasto total individual.

- **A nivel de libreta**

Considerando los tres tipos de días que se pueden obtener, es posible clasificar las libretas en:

1) Libretas rechazadas: corresponden a aquellas libretas para las cuales no se cuenta con ningún día de registro. Cabe mencionar que en este caso el término rechazo no alude únicamente a un deseo explícito del informante de no responder, pues la existencia de una libreta vacía puede deberse a otras causas.

2) Libretas parcialmente completadas: corresponden a las libretas que, teniendo uno o más días de registro, no se encuentran completas.

3) Libretas completas: corresponden a las libretas que cuentan con todos los días de registro de la quincena[6].

3.2 Análisis de la no respuesta

3.2.1 Características generales de la no respuesta en la LGI

Teniendo en consideración la tipología del apartado anterior, los datos del cuadro 1 muestran que para la VIII EPF, del total de libretas que debían ser respondidas, el 18% está en la categoría de rechazo, el 22,1% está parcialmente contestada y el 59,9% corresponde a libretas completas. En comparación con la VII EPF, se advierte un aumento relevante en el porcentaje de libretas rechazadas y una disminución de las libretas completas.

Cuadro 1: Completitud de la LGI en la VII y VIII EPF

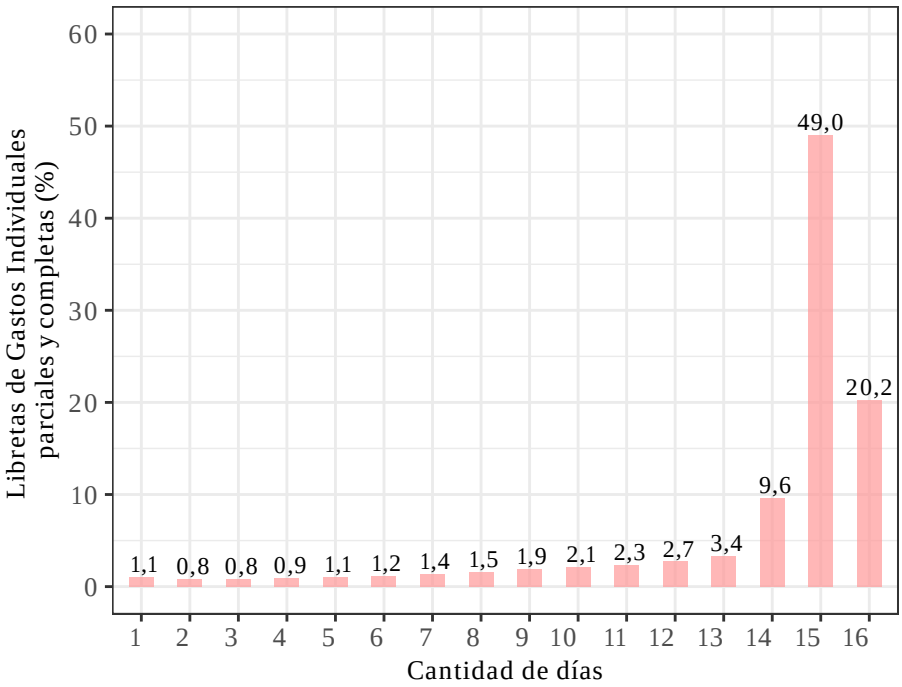
Completitud	Frecuencia VII EPF	Porcentaje VII EPF	Frecuencia VIII EPF	Porcentaje VIII EPF
Rechazadas	3.124	11,8	6.788	18,0
Parcialmente completadas	5.271	19,9	8.343	22,1
Completas	18.047	68,3	22.587	59,9

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

La figura 2 muestra el porcentaje de libretas, según el número de días con registro, pero solo para las libretas que tienen al menos un día de registro, es decir, se excluye el 18% de rechazo mostrado en el cuadro 1. Se advierte que aquellas libretas con 14 o más días de registro concentran aproximadamente el 79% del total. En contrapartida, las libretas con 3 o menos días de registro acumulan aproximadamente el 3%.

[6] Una quincena puede tener 14, 15 o 16 días, dependiendo del mes de trabajo de campo. Para más detalles respecto a este punto, revisar el documento de Metodología VIII EPF, disponible en www.ine.cl/epf.

Figura 2: Libretas, según días de registro, sobre parciales y completas



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Lo que se desprende de estos datos es que las personas que cooperan con la LGI, en general, lo hacen con un alto porcentaje de completitud en cuanto a días de registro. Si bien se observa un aumento en el porcentaje de libretas rechazadas respecto a la versión anterior de la encuesta, las personas que contestan, lo hacen con un alto grado de colaboración.

Esta información es relevante respecto a la discusión internacional sobre la utilización de una semana de registro versus dos. Realmente, no existe consenso en relación con este punto. Uno de los argumentos en contra de utilizar un período de 2 semanas dice relación con el agotamiento de los hogares, lo cual podría generar una disminución de la colaboración conforme pasan los días. Los datos del gráfico de ningún modo son concluyentes respecto a la decisión de utilizar una o dos semanas, sin embargo, aportan evidencia de que, al menos en términos de días de registro, el período de dos semanas no implica necesariamente una pérdida notoria de información.

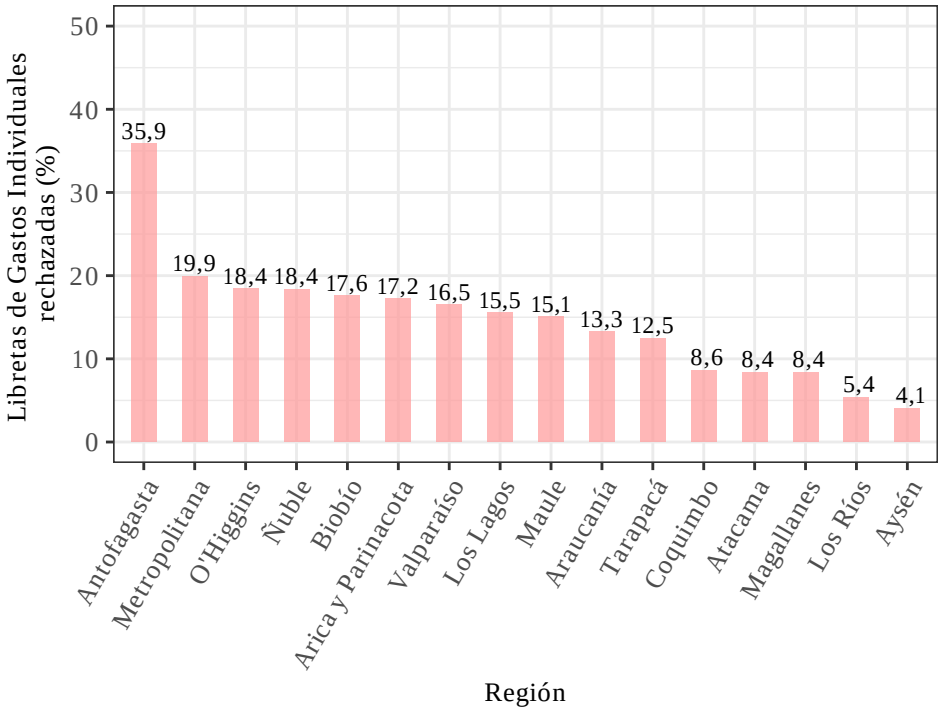
3.2.2 Variables relacionadas con la no respuesta en el registro de gastos diarios

En lo que refiere a encuestas de hogares, en general, se asume que la falta de respuesta no es completamente aleatoria, ya que ella está asociada a ciertas características sociodemográficas de las personas. Para comprender el fenómeno de la no respuesta y tomar una decisión respecto a su posterior tratamiento es importante conocer cuáles son los factores que se relacionan con ella. El presente apartado, entonces, tiene por objeto dar cuenta de cuáles son las variables que explican la no respuesta. Para ello, se presenta estadística descriptiva, correlaciones y, finalmente, un modelo de respuesta.

3.2.2.1 Estadística descriptiva

Una primera mirada posible respecto a la no respuesta en la LGI es identificar si esta tiene algún sesgo respecto a su distribución territorial. Los datos de la figura 3 muestran que el rechazo de la LGI no se distribuye homogéneamente a lo largo del país. De hecho, existe bastante variabilidad entre las regiones. Mientras en Antofagasta llega a un 36% aproximadamente, en la región de Aysén es de tan solo 4%. Respecto a la región Metropolitana cabe señalar que corresponde a la segunda región con mayor porcentaje de rechazo, alcanzando cerca de 20% del total. Un último aspecto que vale la pena mencionar es que las regiones del sur del país exhiben un menor porcentaje de rechazo que el resto del territorio, pues, como se ve en la figura 3, Magallanes, Los Ríos y Aysén ocupan los últimos tres lugares, con 8,39%, 5,37% y 4,07%, respetivamente. Estos resultados dan cuenta de algunas dificultades que enfrentó el estudio durante el trabajo de campo, que varían de forma importante entre regiones.

Figura 3: Porcentaje de rechazo de la LGI, según región del país



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

El documento Informe de Calidad de la VIII EPF (INE Chile, 2018a) muestra que las dos regiones donde la encuesta obtuvo menores tasas de respuesta total a nivel de viviendas (Antofagasta y Metropolitana) coinciden con las regiones con mayor rechazo de la LGI. Pese a que el cálculo del rechazo de la LGI no depende de la tasa de respuesta total de la encuesta[7], esta última puede ser

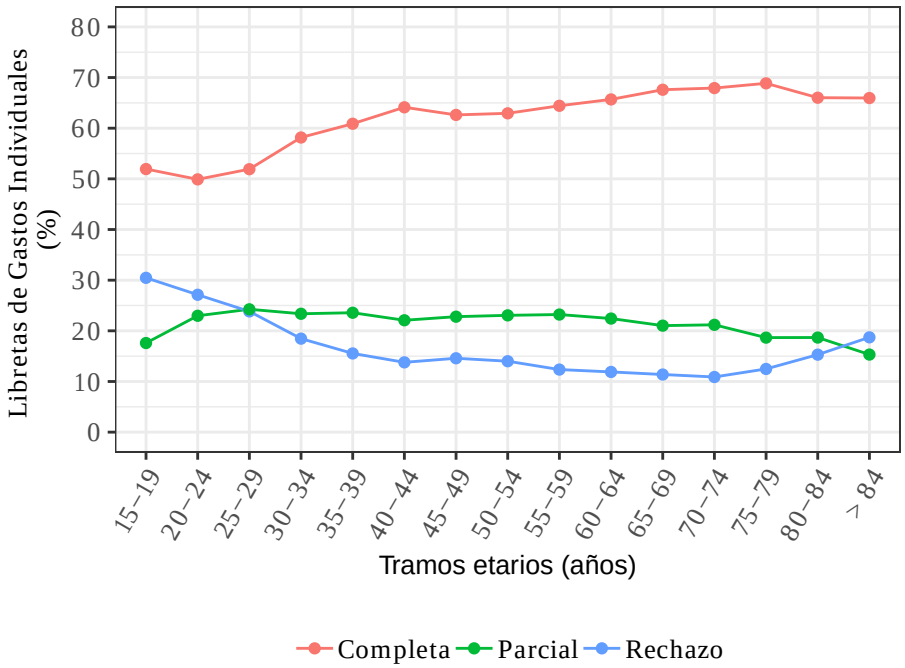
[7]La tasa de respuesta de la encuesta se obtiene dividiendo el total de viviendas entrevistadas con información suficiente por la suma de las viviendas elegibles y de elegibilidad desconocida. Por su parte, el porcentaje de rechazo de la LGI se calcula sobre los hogares que respondieron la encuesta y cumplieron con los criterios mínimos de calidad. En ese sentido, el cálculo de la primera no influye sobre el de la segunda.

interpretada como un indicador general de la dificultad del trabajo de campo en los distintos territorios. En ese sentido, Antofagasta y Santiago podrían describirse como regiones con mayor dificultad en cuanto a participación, lo cual queda reflejado tanto en la respuesta total de la encuesta como en la respuesta de la LGI.

La figura 4 muestra el rechazo, según tramo etario (línea azul). Se advierte un comportamiento en forma de "u", es decir, inicialmente, el rechazo cae conforme aumenta la edad de las personas, pero solo hasta el tramo de 40 a 44 años, tras el cual se observa una suerte de meseta, para finalmente comenzar a subir. Es importante constatar que, pese a la tendencia en "u", las personas de edad avanzada presentan un menor rechazo que los jóvenes. Así, mientras que para las personas en el tramo de 15-19 se advierte un rechazo superior al 30%, para las del último tramo (mayor o igual a 85) dicho valor es inferior al 20%.

Otro indicador relevante es el porcentaje de libretas contestadas completamente. Al respecto, los datos muestran que dicho porcentaje es creciente en la edad, es decir, las personas mayores de 30 años responden el instrumento de manera completa con mayor frecuencia que los jóvenes.

Figura 4: Completitud de la LGI, según tramo etario



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Al desagregar por sexo (cuadro 2), se observa que las mujeres tienen un porcentaje de libretas rechazadas marcadamente menor que el de los hombres (12% versus 25%). Además, muestran una mayor cooperación al momento de responder, ya que un 65% corresponde a libretas completas, mientras que este valor es solo de 53% para los hombres.

Cuadro 2: Completitud de la LGI, según sexo del informante

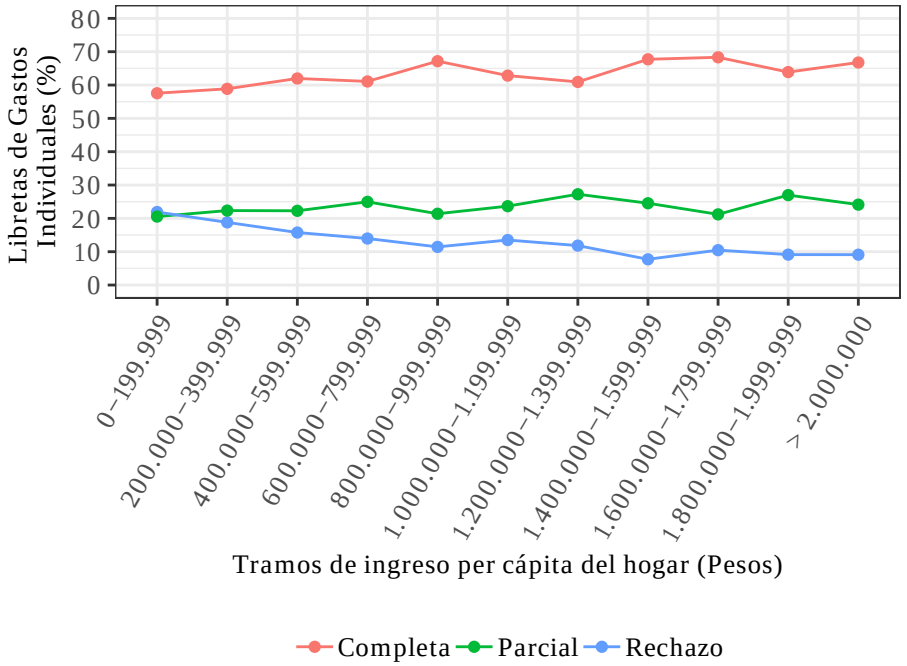
Sexo	Rechazadas	Parcialmente completadas	Completas
Hombre	25,1	21,5	53,4
Mujer	12,0	22,6	65,4

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

La figura 5 muestra la completitud de las libretas, según el ingreso *per cápita* de los hogares. Se observa que el porcentaje de libretas completas es creciente en el ingreso. Por su parte, el rechazo tiene un comportamiento inverso, es decir, cae a medida que aumenta el ingreso. Se advierte, entonces, que los hogares de más ingresos presentan una mayor colaboración al momento de responder la LGI que aquellos de menores ingresos. Esta situación es interesante, ya que la evidencia respecto a tasas de respuesta totales en encuestas de hogares indica que, en general, conseguir la colaboración de los hogares de ingresos altos es más difícil que para los hogares de ingresos bajos. Esta tendencia se verifica en la VIII EPF, ya que la tasa de respuesta en el estrato socioeconómico alto es más baja que en los estratos medio y bajo.

Considerando esta situación, lo que sugieren los resultados es que conseguir la colaboración de los hogares de ingresos altos constituye una tarea que implica mayores esfuerzos que en hogares con ingresos menores, sin embargo, cuando estos colaboran, lo hacen de mejor manera que los hogares de menores ingresos, por lo menos en cuanto a la LGI.

Figura 5: Completitud de la LGI, según tramo de ingreso per cápita del hogar

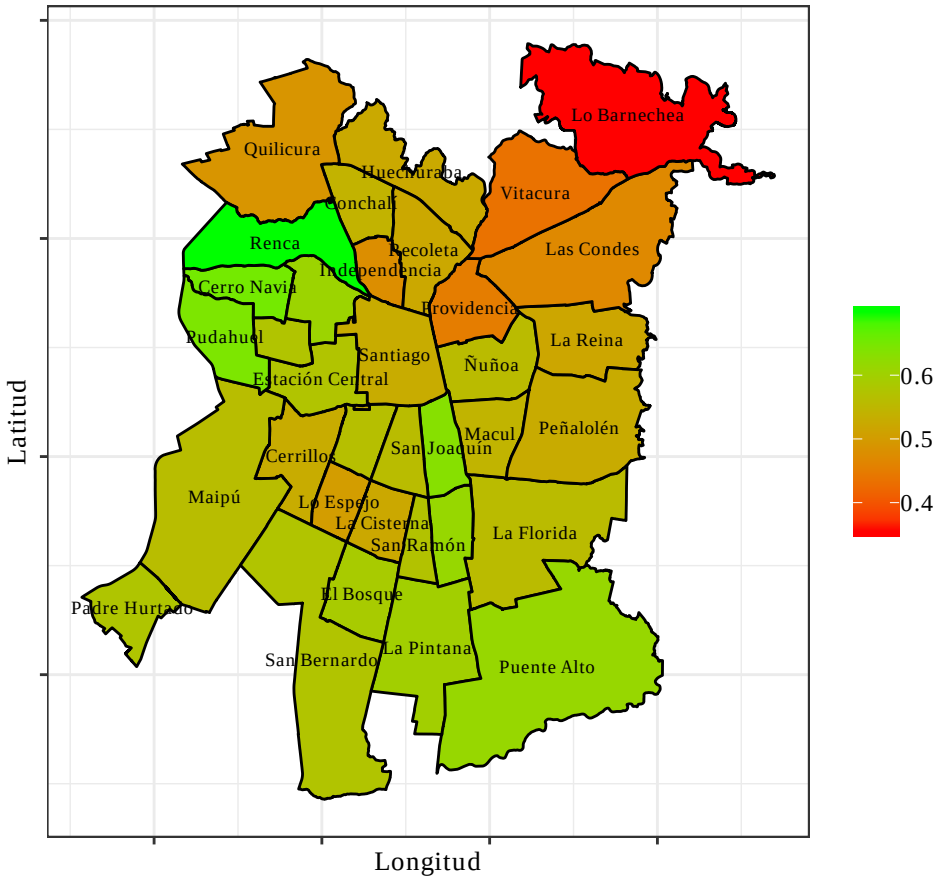


Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Al observar el comportamiento de la no respuesta en términos territoriales se corrobora la idea de que los hogares de mayores ingresos presentan mayor falta de respuesta total, pero que cuando deciden participar, lo hacen con menores niveles de no respuesta al ítem que los hogares de menores ingresos. Esto se ejemplifica en el mapa de la figura 6, el que muestra las tasas de respuesta a nivel comunal en el Gran Santiago. Las comunas, que en promedio, presentan mayores ingresos, ubicadas en el sector oriente de la capital, tienen menores tasas de respuesta total que otras comunas con ingresos en promedio menores. En esta zona es posible ubicar a comunas como Lo Barnechea, Vitacura, Las Condes y Providencia (tonos más cercanos al rojo).

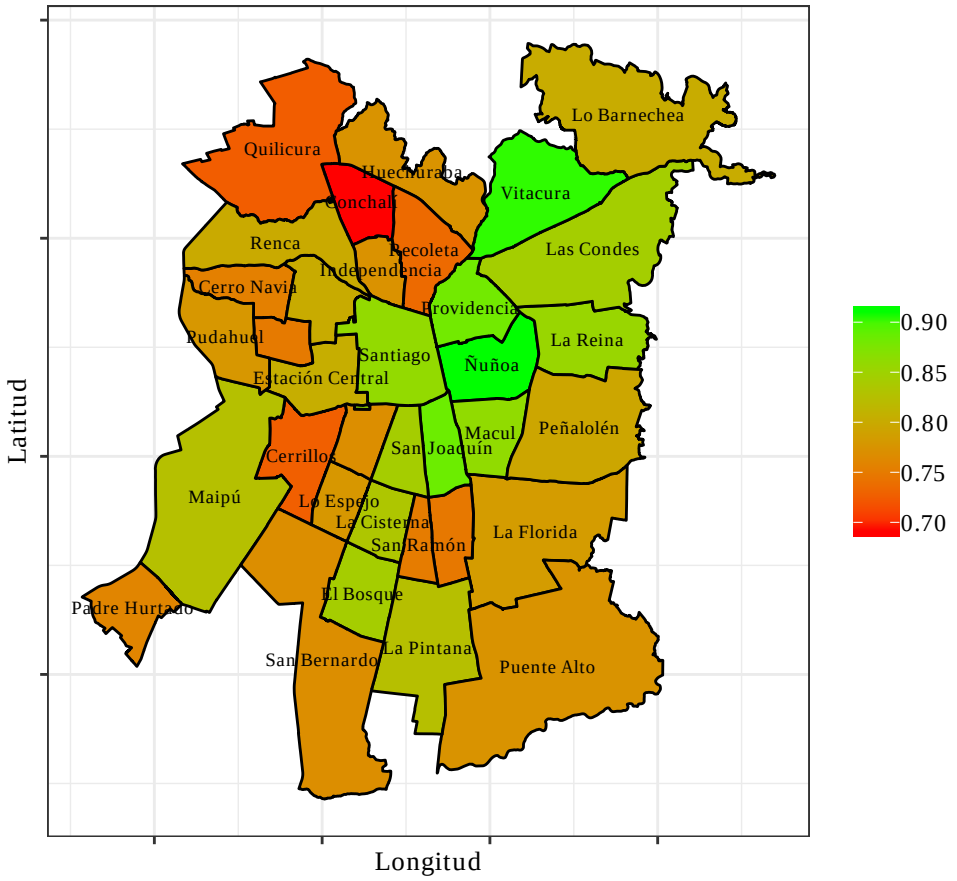
Al desplegar los mismos datos, pero ahora utilizando la tasa de respuesta para la LGI (figura 7), se advierte que en una buena parte de los casos la relación se invierte. Así, por ejemplo, comunas como Lo Barnechea, Vitacura y Las Condes, cuyas tasas de respuesta total son bajas, en lo que respecta a la LGI muestran un buen nivel de colaboración. Al contrario, comunas como Conchalí, Puente Alto y la Pintana, que presentan tasas de respuesta total elevadas, exhiben un menor nivel de cooperación en LGI.

Figura 6: Tasa de respuesta a nivel hogar por comuna en el Gran Santiago



Fuente:Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF) – Información cartográfica Precenso 2016

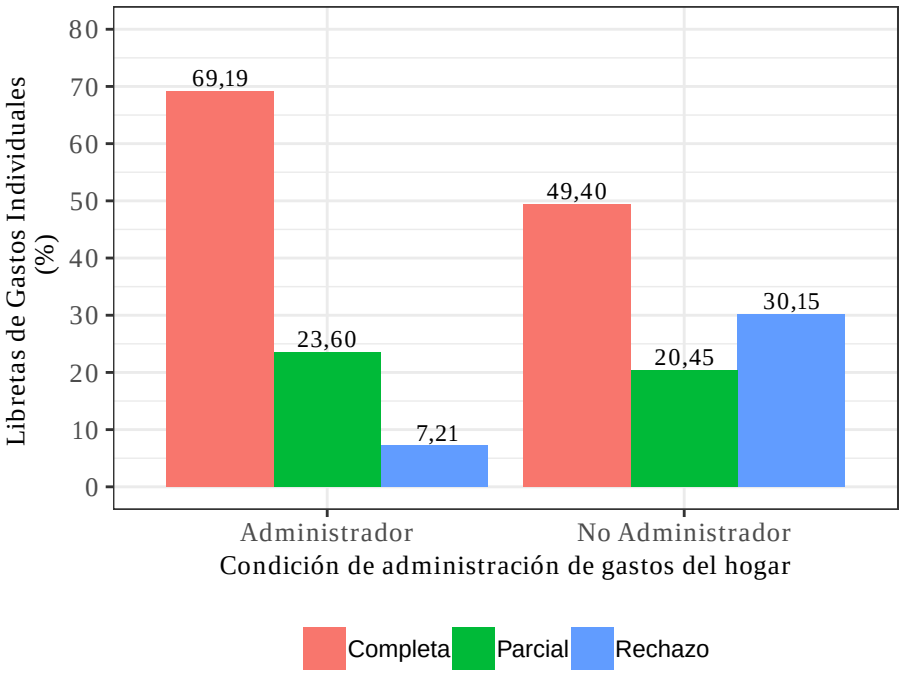
Figura 7: Tasa de respuesta a nivel persona por comuna para la LGI en el Gran Santiago



Fuente:Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF) – Información cartográfica Precenso 2016

Cuando se desagrega la no respuesta según condición de administrador de gasto (figura 8), se advierte que quienes son administradores presentan una respuesta marcadamente mayor que quienes no lo son. Así, por ejemplo, el 69% de los administradores de gasto responde la LGI de manera completa, mientras que dicho valor es de 49% para quienes no lo son. Del mismo modo, los administradores rechazan contestar la LGI en un porcentaje mucho menor que los no administradores (7% y 30%, respectivamente). Esta distribución de no respuesta parcial es sumamente relevante, ya que las personas encargadas de las compras del hogar, en promedio, registran más gasto que quienes no tienen esta función. Ello quiere decir que una baja tasa de respuesta de LGI en dicho grupo implicaría una subestimación del gasto, situación que afortunadamente no ocurre.

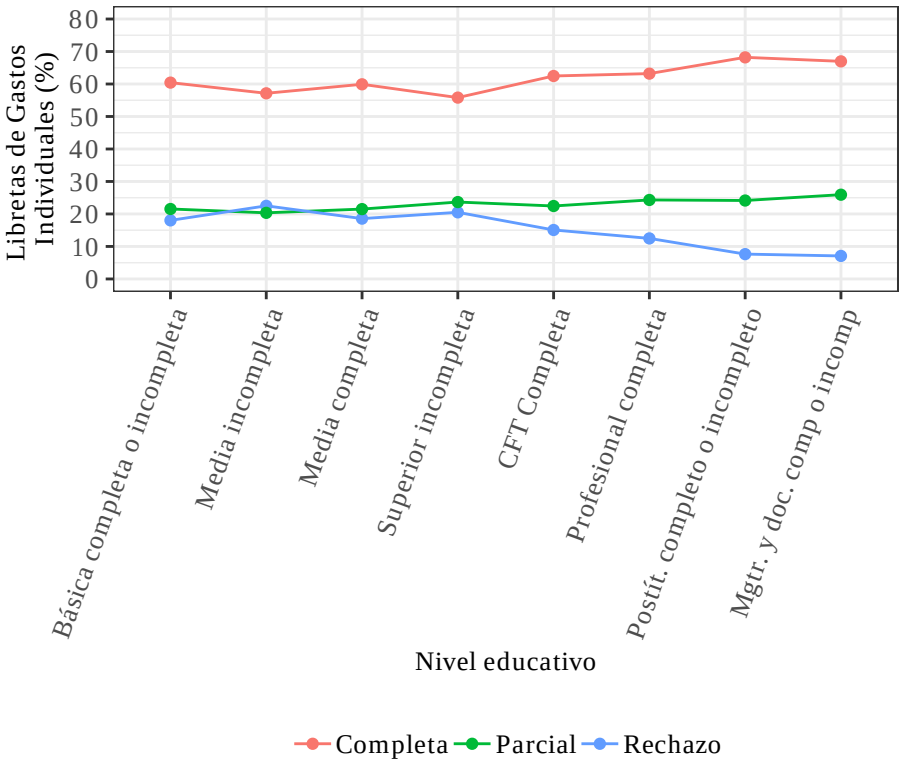
Figura 8: Completitud de la LGI, desagregado por condición de administrador de gastos



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Finalmente, vale la pena revisar cuál es la relación que existe entre nivel educativo y tasa de respuesta en la LGI. La figura 9 muestra que existe una correlación positiva entre el porcentaje de libretas completas y el nivel educativo. Por su parte, el rechazo cae a medida que sube el nivel educativo. En ese sentido, las personas con un mayor nivel de estudios cooperan más al momento de completar la LGI.

Figura 9: Completitud de la LGI, según nivel educativo



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

A partir de la estadística descriptiva revisada, es posible observar que la no respuesta en el caso de la LGI no se distribuye homogéneamente en la población. Al contrario, tiende a concentrarse más en ciertos grupos. Así, por ejemplo, los hombres responden menos que las mujeres; los jóvenes contestan menos que las personas de edad más avanzada y los administradores de gasto lo hacen más que quienes no lo son.

La estadística descriptiva es fundamental como punto de partida, pues muestra, en términos generales, cuál es la situación de la no respuesta. Sin embargo, no permite controlar adecuadamente por varias variables de manera simultánea. Es por ello que en el siguiente apartado se muestran algunos resultados de un modelo de respuesta.

3.2.2.2 Modelo de respuesta

Como se ha señalado en apartados anteriores, para que la no respuesta sea considerada un potencial problema, las unidades seleccionadas deben presentar distintas propensiones a responder. Si tanto las personas que responden como las que no tuviesen una misma propensión a responder, el escenario sería el de una no respuesta completamente aleatoria (MCAR), en cuyo caso el problema de sesgo desaparece. En ese sentido, es necesario identificar si efectivamente existen perfiles de respondientes con diferentes propensiones a responder.

Cabe mencionar que la propensión a responder no es un fenómeno medible directamente. De hecho, lo que realmente se observa es si el informante responde o no. Para este tipo de problemas, típicamente (Bethlehem, 2012; Olson, 2006; Valliant, Dever, & Kreuter, 2013) se recurre a modelos de regresión, que permitan estimar la probabilidad de ocurrencia de un evento, los cuales funcionan sobre la base de una variable latente, que en este caso refiere a la actitud o propensión de las personas a responder la LGI. Esta variable no observada se modela de manera lineal:

$$y_i^* = x_i^T \beta + e_i \tag{2}$$

Así, diremos que y_i toma valor 1 cuando la variable latente (propensión a responder) pasa cierto umbral y 0, cuando no. De manera estándar, dicho umbral es 0, de modo que y será igual a 1 cuando $y_i^* > 0$ y 0, cuando $y_i^* < 0$.

Considerando lo anterior, la probabilidad de que una persona conteste puede escribirse del siguiente modo

$$Pr(y_i = 1) = Pr(x_i^T \beta + e_i > 0) = Pr(e_i < x_i^T \beta) \tag{3}$$

Dado que la distribución del error es simétrica, es posible establecer la siguiente igualdad

$$Pr(e_i < x_i^T \beta) = F(x_i^T \beta) \tag{4}$$

Donde F es la función de distribución acumulada, que en este caso corresponde a la función logística, la cual recibe como argumento un modelo lineal, de modo que

$$y = \frac{1}{1 - e^{-f(x)}} \tag{5}$$

Finalmente, el modelo estimado, expresado en términos matriciales, es el siguiente:

$$Y = \alpha + X\beta + e \tag{6}$$

Dónde:

Y: Responde al menos un día en la LGI/Rechazo

X: Sexo, jefe de hogar, administrador de gastos, región, ingreso del hogar, escolaridad, n° personas en el hogar, edad.

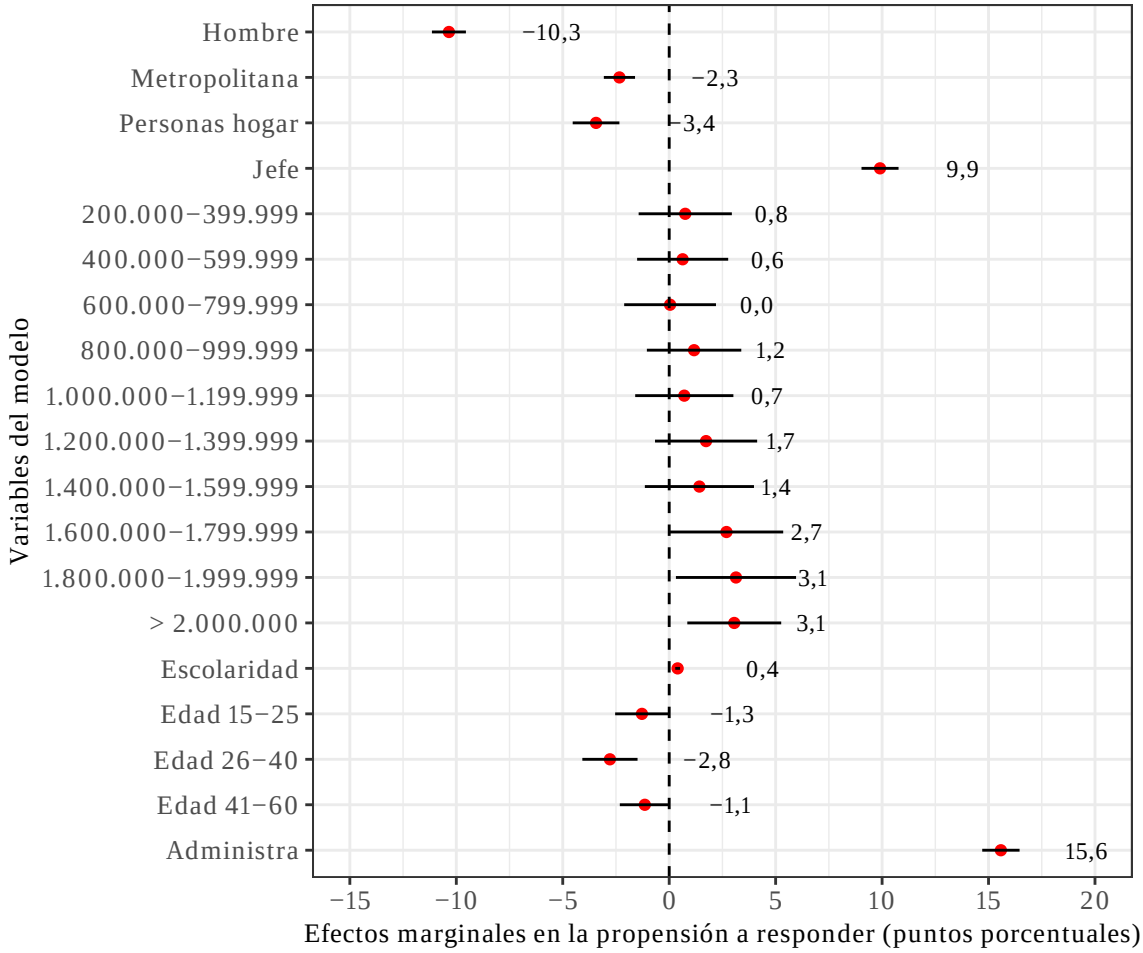
ϵ: Término de error

La figura 10 muestra los efectos marginales y los intervalos de confianza para cada una de las variables del modelo. La línea vertical achurada corresponde al 0, por ende, si un intervalo de confianza pasa por el 0, implica que el efecto de la variable no es significativo a un 95% de confianza. Teniendo esto en consideración, se observa que la mayoría de las variables utilizadas son significativas. Así, por ejemplo, ser hombre disminuye en 10 puntos porcentuales la probabilidad de responder la LGI, respecto a ser mujer. En el caso de la variable de administrador de gastos, se advierte que tener

dicha condición aumenta significativamente la probabilidad de responder la libreta en 15,6 puntos porcentuales en comparación con quienes no lo son. Por su parte, ser jefe de hogar tiene un efecto importante en la propensión a responder, con 9,9 puntos porcentuales, respecto a no serlo.

En relación con el ingreso del hogar, cabe señalar que las primeras categorías[8] no poseen efectos significativos, situación que cambia en la categoría 9, a partir de la cual los efectos marginales se tornan significativos. Esto se condice con la estadística descriptiva mostrada anteriormente y corrobora la idea de que los hogares de ingreso alto presentan una menor tasa de respuesta a nivel de encuesta, pero cuando acceden a participar muestran un mayor nivel de cooperación que los hogares de ingresos bajos, al menos en lo que respecta a la LGI.

Figura 10: Efectos marginales del modelo de respuesta. Regresión logística



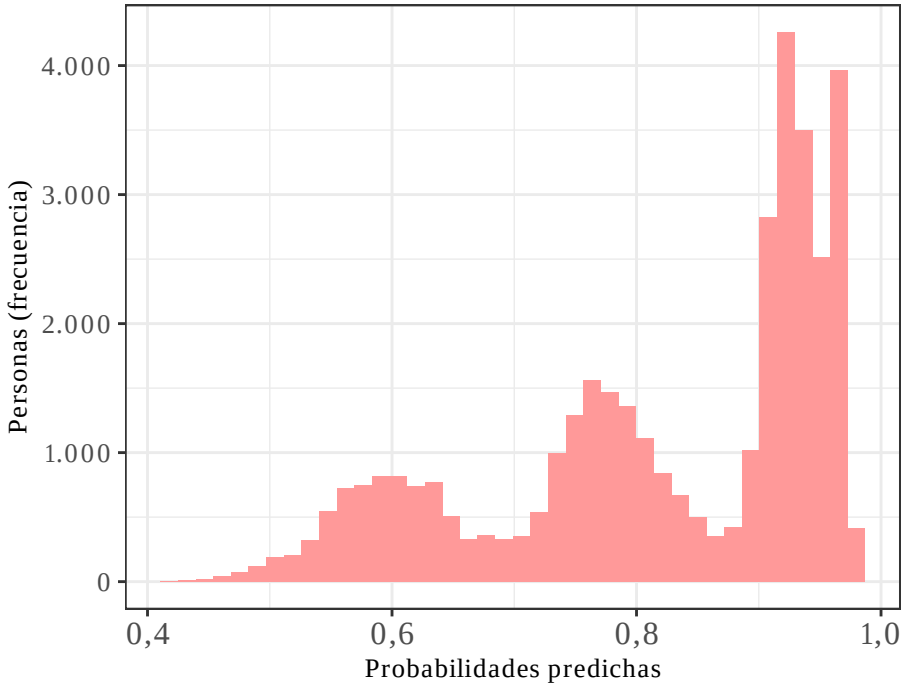
Fuente: Instituto Nacional de Estadísticas (INE) − VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

A partir del modelo estimado, es posible generar predicciones para cada una de las personas que participaron en la encuesta (tanto para los que respondieron como para los que no). La distribución

[8] La categoría base corresponde a los hogares con ingresos menores a 200.000 pesos.

de las probabilidades predichas (figura 11) muestra que, sin ser demasiado alta, existe cierta variabilidad en la propensión a responder. Así, en el extremo inferior se encuentra una persona con una probabilidad predicha de 0,42, mientras que en el extremo superior se ubica una con probabilidad de 0,98. Un aspecto positivo, desde el punto de vista del sesgo de no respuesta, es que la distribución se encuentra inclinada hacia la derecha, es decir, la mayor parte de las personas tiene una probabilidad relativamente alta de responder la LGI. El hecho de que la dispersión no sea tan elevada es algo deseable, ya que entre mayor sea esta, mayor será la distorsión introducida por la no respuesta respecto al diseño muestral inicial (Bethlehem, Cobben, & Schouten, 2008).

Figura 11: Distribución de probabilidades predichas por modelo de respuesta



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Con el objeto de ilustrar la existencia de distintos perfiles de respondientes, el cuadro 3 muestra predicciones para dos grupos[9] de respondientes. El primero, corresponde a un perfil de baja probabilidad de respuesta, pues está construido únicamente con categorías que en el modelo tienen signo negativo, lo cual da como resultado una probabilidad media de 0,53. El segundo grupo, construido solo con categorías que tienen signo positivo, se ubica en el otro extremo, con una probabilidad media de 0,97.

[9]Las variables del modelo que no aparecen en el cuadro están fijas en el promedio.

Cuadro 3: Perfiles de respondientes de la LGI

	Baja probabilidad	Alta probabilidad
	Hombre	Mujer
	No administrador	Administrador
	Región Metropolitana	No región Metropolitana
	Más de 4 personas	Hasta 4 personas
	No jefe	Jefe de hogar
Prob. de respuesta LGI	0,53	0,97

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Para que la no respuesta genere sesgo en la estimación de la media es preciso que esta se encuentre correlacionada con la variable objetivo, que en el caso de la EPF corresponde al gasto. A raíz de lo anterior, es interesante observar la relación que existe entre las probabilidades predichas y el gasto por LGI. El coeficiente de Pearson es de 0,38, valor que da cuenta de que efectivamente existe una asociación entre la propensión a responder y el gasto, lo cual indica que es relevante llevar a cabo algún procedimiento de imputación que intente corregir el posible sesgo por no respuesta.

Los resultados del modelo, sumados a la estadística descriptiva, sugieren que la no respuesta no se distribuye homogéneamente en la población, sino que, al contrario, se concentra en ciertos grupos, lo que quiere decir que no todas las personas tienen la misma propensión a responder. En ese sentido, la no respuesta está introduciendo una distorsión respecto al diseño muestral del estudio, ya que las probabilidades de selección que inicialmente se determinaron para cada una de las unidades están siendo modificadas por las propensiones que distintos grupos tienen a responder. Dentro de las variables más importantes respecto a dicha propensión están:

- Sexo de la persona
- Si la persona es jefa de hogar
- Si la persona es administradora de gasto

Los datos mostrados hasta el momento tienen una doble implicancia. En primer lugar, constituyen una señal de que la no respuesta es un fenómeno que debe ser atendido por medio de algún método de imputación, pues en caso de no hacerlo se corre el riesgo de introducir sesgo en las estimaciones. En segundo lugar, los resultados del modelo pueden convertirse en un insumo para una próxima recolección, pues dan cuenta de ciertos perfiles con menor probabilidad de responder, para los cuales es necesario aumentar los esfuerzos al momento de desplegar el trabajo de campo. Conocer de antemano cuáles son los perfiles de personas que presentan menor nivel de cooperación puede contribuir a mejorar la respuesta de la LGI en el futuro, lo que redunda en una disminución del sesgo de no respuesta.

3.2.3 Variables correlacionadas con el gasto

Para cualquier método de imputación es fundamental determinar cuáles son las variables relacionadas con la variable objetivo, ya que esta información es la que permite llevar a cabo procedimientos estadísticos que corrijan el sesgo de no respuesta. En el caso específico de la EPF,

para determinar cuáles son las variables correlacionadas con el gasto promedio mensual, en este apartado se presentan, en primer lugar, resultados de correlaciones entre algunas variables y el gasto capturado por LGI. En segundo lugar, se revisan resultados de un modelo de regresión.

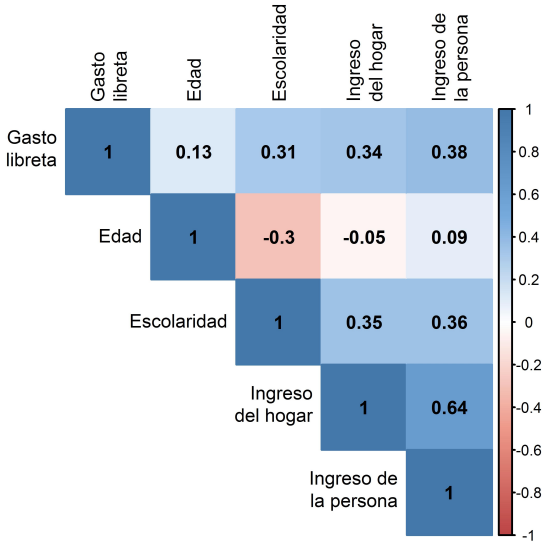
Es importante señalar, que si bien la variable objetivo del estudio es el gasto a nivel hogar y no a nivel persona, los análisis aquí presentados se realizaron a nivel de características individuales. El motivo de esto es que se parte del supuesto de que al menos una parte del gasto registrado en las LGI depende de las características particulares de las personas y no del hogar como un todo. Ello hace que tenga sentido incorporar, tanto en las correlaciones como en el modelo de regresión, información de las personas.

3.2.3.1 Correlaciones

Dado que algunas variables con las que interesa correlacionar el gasto por LGI son continuas y otras discretas, es necesario utilizar dos coeficientes de correlación diferentes para cada uno de los casos. Así, para las variables continuas se utiliza el coeficiente de correlación lineal de Pearson, mientras que para las variables dicotómicas se utiliza el coeficiente de punto biserial.

La figura 12 muestra la correlación lineal que existe entre el gasto por libreta y algunas variables continuas. La relación más fuerte se produce con el ingreso a nivel personal (0,38), seguido por el ingreso a nivel hogar (0.34). Respecto a los signos, son los esperados: el gasto por libreta crece a medida que aumentan la edad, la escolaridad y el ingreso (personal y del hogar).

Figura 12: Correlación con el gasto. Correlación de Pearson



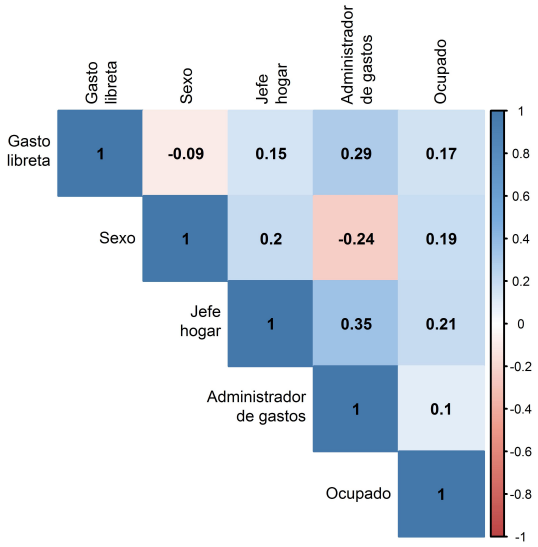
Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Para el caso de las variables dicotómicas, se utiliza la correlación de punto biserial. Respecto al tamaño de los coeficientes, se observa que el más grande corresponde a la relación entre gasto y

la condición de administrador de gasto, con un valor de 0,29 (figura 13). Esta relación se explica por el hecho de que quiénes ejercen la función de administrador, no solo realizan compras vinculadas a su propio consumo sino también al consumo del resto del hogar, por ende, estas personas registran un mayor monto que quienes no ejercen dicha función. Respecto a los signos, cabe mencionar que en el caso de la variable sexo se observa una relación negativa, es decir, ser hombre está vinculado a una disminución del gasto por libreta.

Teniendo en consideración los datos sobre correlaciones, se observa que las variables que tienen una mayor correlación con el gasto por libreta son el ingreso, la condición de administrador de gasto y la escolaridad. Esta información arroja evidencia respecto a cuáles son las variables importantes de considerar al momento de implementar un método de imputación.

Figura 13: Correlación con el gasto. Correlación biserial



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

3.2.3.2 Modelo de regresión para gastos diarios

Con el fin de apreciar el efecto aislado de cada una de las variables, a continuación se presentan los resultados de un modelo de regresión lineal. La variable dependiente corresponde al gasto por libreta y las variables independientes corresponden a una serie de características del hogar y de las personas. El modelo, expresado en términos matriciales es el siguiente:

$$Y = \alpha + X\beta + \epsilon \tag{7}$$

Dónde:

Y: Gasto por libreta en pesos

X: sexo, jefatura de hogar, administrador de gastos, macrozona[10], ingreso del hogar, situación en el empleo, escolaridad, cantidad de miembros del hogar en tramos, ocupación y edad.

ϵ: Término de error

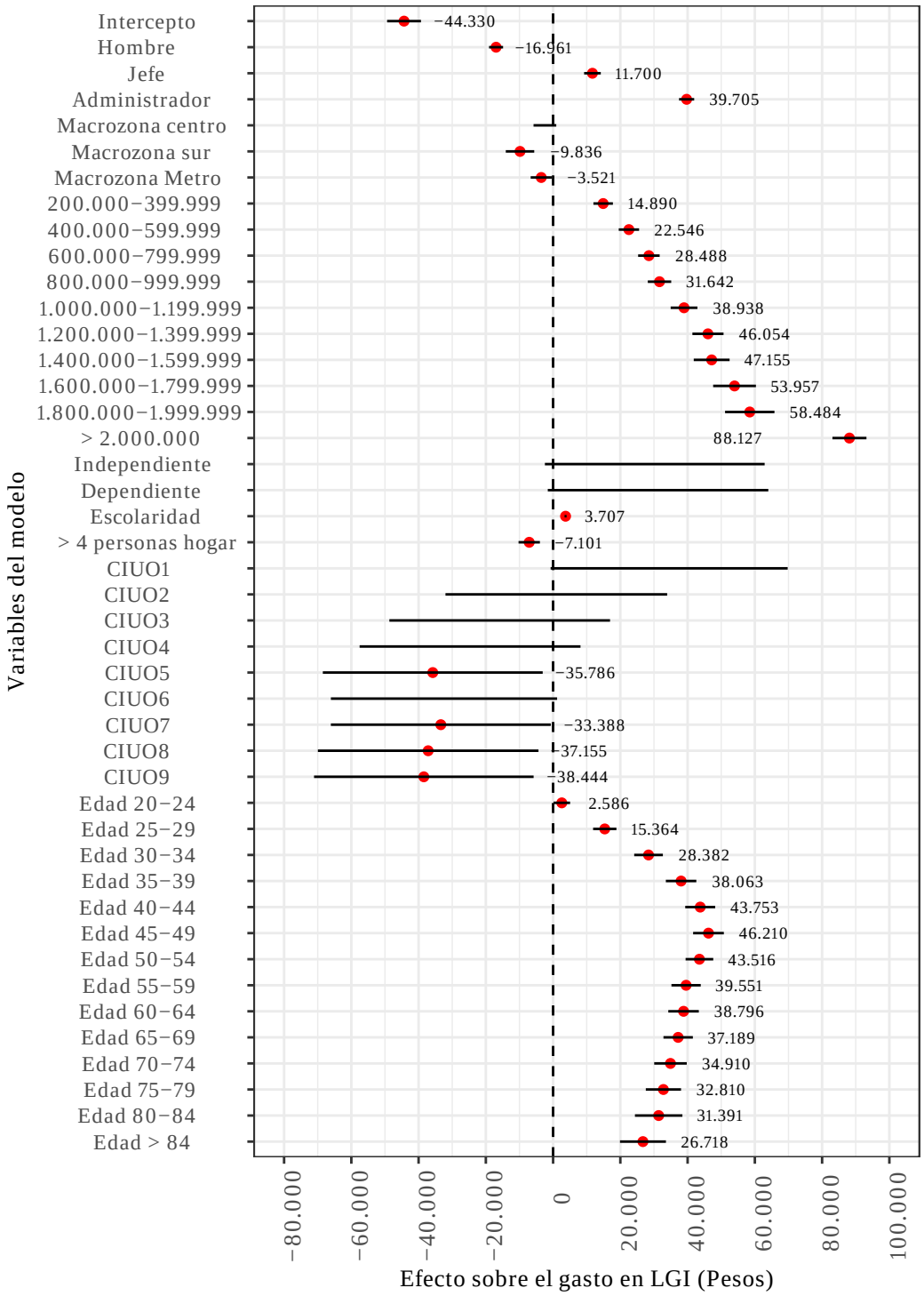
La figura 14 muestra el valor de los coeficientes y su significancia. Dado que el modelo contiene una gran cantidad de variables, para facilitar su lectura, los valores no significativos han sido excluidos (sin punto rojo). En términos generales, es posible señalar que los signos de los coeficientes son los esperados y coinciden con los de las correlaciones mostradas anteriormente. Así, por ejemplo, ser hombre genera una disminución de aproximadamente 17.000 pesos en el gasto, respecto a ser mujer. Ser administrador de gasto, al contrario, tiene un efecto positivo aumentando el gasto en 39.700 pesos aproximadamente. Otras variables que tienen signo positivo son escolaridad e ingreso. En relación con este último, se aprecia que todos los tramos tienen un coeficiente positivo y creciente, es decir, a medida que aumenta el ingreso del hogar, también lo hace el gasto por libreta.

Es interesante constatar que el número de personas por hogar se relaciona negativamente con el gasto por LGI. Esto puede deberse a que conforme crece el tamaño de los hogares, el gasto se va distribuyendo entre las personas. Esto tiene sentido a la luz del hecho de que no todo el gasto registrado en las libretas obedece a una lógica individual, ya que parte de este tiene como destino el consumo de todo el hogar, lo cual, dicho de paso, se condice con la noción de que los hogares constituyen unidades económicas y, en cuanto tal, no presentan un comportamiento puramente individual.

Finalmente, vale la pena mencionar la relación que existe entre la edad y el gasto por libreta. Se observa que hasta el tramo de edad 45-49 los coeficientes crecen progresivamente, para luego comenzar a caer hasta el tramo de edad más alto (> 84). En ese sentido, se advierte que el gasto en la LGI tiene un comportamiento cuadrático a lo largo de la vida. En términos conceptuales, ello quiere decir que conforme pasa el tiempo, las personas van aumentando su gasto en relación con el primer tramo etario (15-19), pero dicho aumento es cada vez menor, lo cual se vincula con las diferentes necesidades que las personas presentan a lo largo de la vida.

[10] Las macrozonas están conformadas por las siguientes conurbaciones: 1) norte: Arica, Gran Iquique, Antofagasta, Copiapó, Gran La Serena; 2) centro: Gran Valparaíso, Rancagua, Talca, Gran Concepción, Gran Chillán; 3) Sur: Gran Temuco, Valdivia, Puerto Montt, Coyhaique, Punta Arenas; 4) Gran Santiago.

Figura 14: Modelo MCO de gastos en la LGI



Fuente: Instituto Nacional de Estadísticas (INE) − VIII Encuesta de Presupuestos Familiares (EPF)

Sobre la base de los resultados obtenidos, se observa que las variables más relacionadas con el gasto son: Administrador de gastos, sexo, ingreso, edad y escolaridad. Esta información es relevante al

momento de construir el método de imputación. En el siguiente apartado se volverá sobre este punto.

3.3 Descripción de los métodos para gastos diarios

3.3.1 Experiencia de la VII EPF

Durante el proceso de producción de la VII EPF se realizó un estudio sobre el método a utilizar para la imputación de gastos diarios, lo cual quedó plasmado en el documento de trabajo "Métodos de imputación VII EPF: Gastos diarios e ingresos de la actividad principal y jubilaciones" (INE Chile, 2014).

Para enfrentar el problema de no respuesta parcial en la LGI, se estudiaron tres métodos de imputación:

- Factor de no respuesta (en adelante FNR)
- Ajuste por peso diario
- Media condicionada

Los resultados de aquel estudio muestran que los tres métodos presentan un resultado similar respecto al promedio de gasto, sin embargo, el análisis efectuado dio cuenta de que el FNR era el método que menos se alejaba del conjunto de libretas con registro completo. Considerando esto y su simplicidad de cálculo, se decidió utilizar el método de FNR para imputar las libretas parcialmente completadas. Cabe señalar que no se llevó a cabo ninguna imputación para las libretas rechazadas.

3.3.2 Métodos probados en la VIII EPF

Considerando la experiencia de la VII EPF, en el marco de la VIII versión de la encuesta se testeó el método de *hot deck* y el FNR. Uno de los objetivos fundamentales en el análisis de imputación en la VIII EPF fue detectar el sesgo generado por las libretas rechazadas, de manera de observar el efecto que tiene imputar dichas libretas en el gasto promedio de los hogares.

3.3.2.1 Factor de no Respuesta

El FNR es un método empleado en la EPF diseñado fundamentalmente para atender el problema de la no respuesta parcial de la LGI. Se basa en el supuesto de que los días con registro de una libreta pueden representar a los días que no tienen registro. El objetivo consiste es reponderar la información de gastos que tiene una libreta, para así completar la quincena.

El ajuste realizado depende del número de días de registro. Así, cuando el número de días es mayor o igual a 6 se aplica

$$FNR = \frac{dquincena_q}{dregistro_i} \tag{8}$$

Dónde

- $dquincena$: corresponde al número de días de la quincena q[11]
- $dregistro$: corresponde al número de días de registro de la persona i

Se observa que entre menor sea el número de días de registro, mayor será el factor de ajuste aplicado. Así, las libretas con pocos días de registro tendrán un ajuste mayor que aquellas cercanas a la completitud. A modo de ejemplo, una libreta con 10 días de registro en una quincena de 15 días tendrá un factor de 1,5. Por otro lado, una libreta que tenga 7 días de registro tendrá un factor de 2,1. En concreto, el ajuste realizado consiste en multiplicar todos los gastos de una libreta por el factor obtenido.

Por otro lado, cuando el número de días de registro es inferior a 6, en la VII EPF se decidió utilizar un factor igual a 2. Esto significa que las libretas con menos de 6 días quedarán incompletas. Si se piensa, por ejemplo, en una libreta con 5 días de registro, el ajuste debiese ser 3 (15/5). Sin embargo, los gastos solo son multiplicados por 2, lo cual puede ser interpretado como si la libreta solo quedara con 10 de los 15 días de la quincena. La decisión de fijar un umbral dice relación con el hecho de que no parece razonable suponer que muy pocos días de registro puedan representar a toda una quincena[12]. De hecho, es posible que completar una quincena a partir de pocos registros, en lugar de corregir la estructura de gastos, introduzca mayores distorsiones. En ese sentido, cuando una libreta tiene muy pocos días con registro, el FNR asume la pérdida de un cierto volumen de información, en pos de no realizar una imputación de mala calidad.

3.3.2.2 Método de imputación *hot deck*

Para intentar corregir el posible sesgo generado por la no respuesta, en la VIII EPF se utilizó el método de *hot deck*. Este método se basa en la idea de que un buen candidato (conocido en la literatura como donante) para completar un dato faltante es una unidad que comparte características con aquella a la cual le falta información (receptor). En el caso de la EPF, la idea que está detrás es que el gasto de las personas no es aleatorio, sino que está determinado por ciertas características socioeconómicas y/o demográficas. En ese sentido, para dos personas que compartan dichas características, lo esperable es que su consumo sea similar. El desafío consiste, entonces, en idear un procedimiento que permita encontrar a dos unidades que sean los más parecidas posibles.

El criterio para establecer la similitud y luego el mecanismo para dirimir entre dos posibles donantes son decisiones que deben ser abordadas por este método. Sin embargo, para estas cuestiones no existe una única respuesta. En realidad, son decisiones que deben considerar la naturaleza del problema específico que se está abordando. En ese sentido, el método de *hot deck* no debe entenderse como una estrategia rígida, sino más bien como una familia de técnicas basadas en la idea común de que un buen donante es aquel que comparte características con la unidad a la que le falta información.

[11] Este término no es constante debido a que una quincena puede tener 14, 15 o 16 días, dependiendo del mes de levantamiento. Para más detalle sobre este punto, revisar documento Metodología VIII EPF.

[12] El punto de corte de 6 días está asociado a la necesidad de resguardar que una libreta tenga la mayor cantidad posible de días de la semana.

Lo deseable dentro de este método es contar con un set de variables que permita encontrar a la unidad más parecida posible. Naturalmente, no es de interés lograr coincidencia en todas las características de las personas, sino únicamente en aquellas que tienen relación con el fenómeno estudiado. En el caso de la VIII EPF se construyeron agrupaciones de personas o *clusters* en base a un set de variables que debían cumplir con, al menos, una de las siguientes condiciones:

1) **Estar correlacionadas con el gasto**: estas variables se seleccionan a partir de los resultados de correlaciones bivariadas y del modelo de regresión expuesto previamente.

2) **Explicar la estructura del gasto**: es posible que algunas variables no correlacionen con el nivel del gasto, pero sí con su distribución. La idea subyacente es que algunas características de las personas no guardan necesariamente relación con el monto del gasto, pero sí con la decisión respecto a qué tipos de gastos se realizan. Un ejemplo de ello es qué proporción del gasto se destina a cada una de las 12 divisiones.

3) **Estar relacionadas con la propensión a responder**: para corregir el sesgo de no respuesta, es importante considerar la probabilidad de respuesta que tienen las personas. Es por ello que entre las variables seleccionadas están aquellas que en el modelo de respuesta, presentado en el apartado 3.2.2.2, tienen un efecto significativo. Adicionalmente, en el set de variables se incluyeron algunas que fueron utilizadas en el diseño muestral. El motivo de ello es no asumir que todas las unidades tienen la misma probabilidad de selección (Andridge & Little, 2009). Dado que, desde un inicio, el diseño establece diferentes probabilidades de selección para las unidades, el método de imputación debe intentar respetar dichas probabilidades.

Idealmente, el donante debiese escogerse dentro del grupo de personas que coincide con el receptor en todas las variables escogidas como relevantes, ya que ello permite encontrar al donante más parecido. Ahora bien, ello no siempre es posible. A raíz de ello, es preciso ir flexibilizando el criterio de similitud, hasta encontrar un donante.

Para determinar en qué orden se flexibilizan las variables, se utilizan los resultados de las correlaciones y del modelo de regresión mostrados en el apartado anterior. Así, aquellas variables no significativas o con una correlación baja con el gasto fueron identificadas como candidatos iniciales a flexibilizarse. Es preciso aclarar que relajar una variable puede implicar dos procedimientos. El primero de ellos guarda relación con una transformación de las variables, lo cual puede consistir en la construcción de tramos o en una ampliación de los mismos. Así, una variable que originalmente es continua, como la edad, es convertida a 18 tramos, para luego ser convertida nuevamente a una variable con 10 tramos. La segunda forma de flexibilizar una variable implica simplemente eliminarla del proceso.

Cabe mencionar que el primer mecanismo de flexibilización es preferible al segundo, ya que implica una menor pérdida de información. Es por ello que, al no ser encontrado un donante, el primer paso siempre consiste en generar tramos. Solo una vez hecho ello, se elimina la variable en cuestión.

Considerando estos criterios, se generaron 23 niveles de imputación que dan lugar a una matriz de transferencia. Tal como muestra el cuadro 4, el primer nivel está compuesto por todas las variables a utilizar, por lo que corresponde a la máxima exigencia.[13] Al contrario, el nivel 23 solo contiene 2

[13] Dado que el primer nivel incluye el identificador de la persona y el hogar, los donantes de este nivel corresponden

variables y representa el mínimo nivel de exigencia. A medida que disminuye la exigencia, algunas variables se van transformando y otras se van eliminado (cuadros 25 al 28 de los anexos). Cabe notar que las únicas 2 variables que nunca se relajan son ingreso del hogar y administrador de gastos.

Cuadro 4: Máximo nivel de exigencia para la imputación de gastos diarios

Variables
Submuestra
Condición socioeconómica
Sexo
Ingreso hogar
Escolaridad
Edad
Administrador de gasto
CIUO
CISE
Comuna
Manzana
Identificador del hogar (folio)
Número personas en el hogar
Identificador de la persona en el hogar

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

3.3.2.3 Tipos de donante

Hasta el momento, fuera de las variables de similitud, no se ha impuesto ninguna otra restricción sobre los potenciales donantes. Con el objeto de evitar que libretas con pocos días de registro entren al proceso como donantes, se establecieron dos situaciones que permitían a una LGI ser parte del *pool* de donantes:

- **LGI de la misma persona a la cual le falta información**: un día sin registro (dato faltante) puede ser completado con otro de la misma persona, si esta cuenta con información disponible del día de la semana correspondiente. En ese sentido, se genera una autodonación.
- **LGI con todos los días de registro**: corresponde a libretas completas, es decir, que tienen 14, 15 o 16 días de registro, dependiendo de la quincena.

Lo anterior quiere decir que un dato faltante se completa con información de la misma persona, si se cuenta con ella. Si la misma persona no cuenta con información, se recurre a un segundo donante, el cual, dado que cumple con la característica de tener todos los días con registro, le donará la información faltante, quedando la libreta completa.

Se desprende que el método da prioridad a la autodonación, pues se asume que el mejor donante para una persona es ella misma. A modo de ejemplo, si una libreta tiene un lunes sin registro, pero cuenta con otro lunes que sí tiene información, este último se utiliza para completar los datos que faltan. Ahora bien, si lo anterior no ocurre (por ejemplo, en el caso de una libreta rechazada) el método

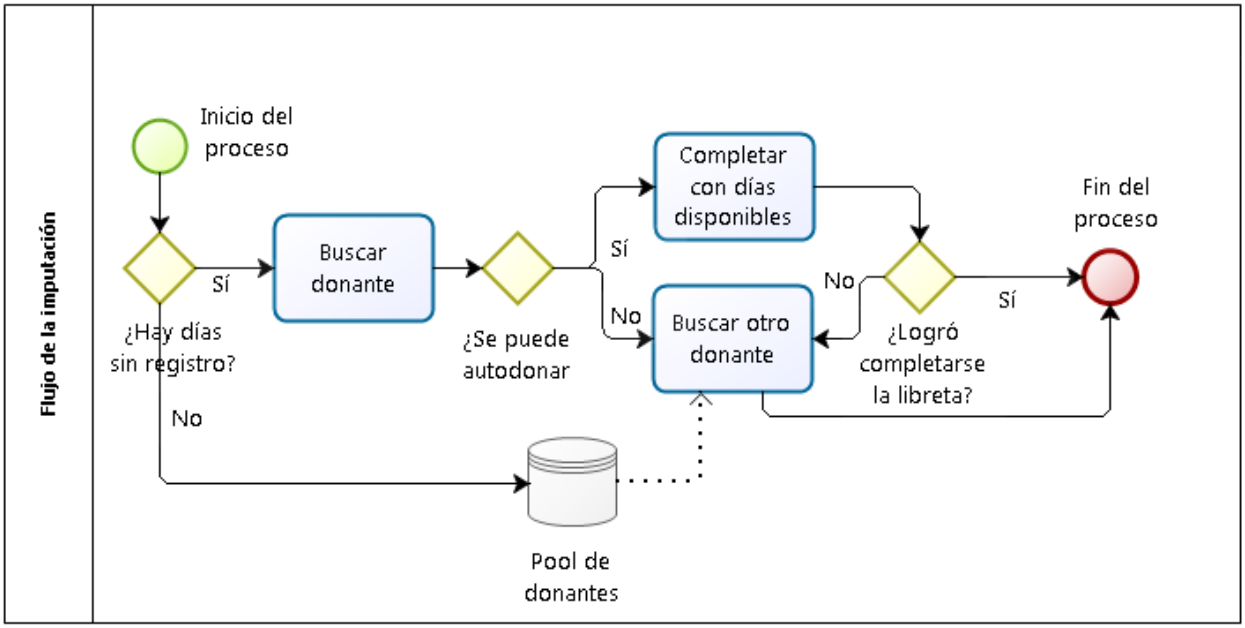
a la misma persona para la cual se busca completar información, lo cual quiere decir que es irrelevante agregar más variables. La decisión de incluir todas variables obedece a la necesidad de lograr mayor claridad en la exposición.

busca la persona más parecida para hacer la donación, siguiendo las reglas de flexibilización de los cuadros 25 al 28 de los anexos.

Según esta configuración, los datos de una libreta pueden estar compuestos como máximo por la información entregada por la misma persona, más los datos de una segunda. Si una libreta no se completa por medio de la autodonación, se buscará otra y con los datos de esta última la quincena quedará completa. Es relevante dejar en claro que la imputación implica que toda la información del donante será copiada en la libreta que tiene datos faltantes. En ese sentido, la descripción del gasto, la cantidad consumida y el valor se donan simultáneamente.

La figura 15 muestra en detalle cuál es el flujo del proceso de imputación.

Figura 15: Flujo del proceso de imputación



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

3.3.2.4 Selección del donante y día

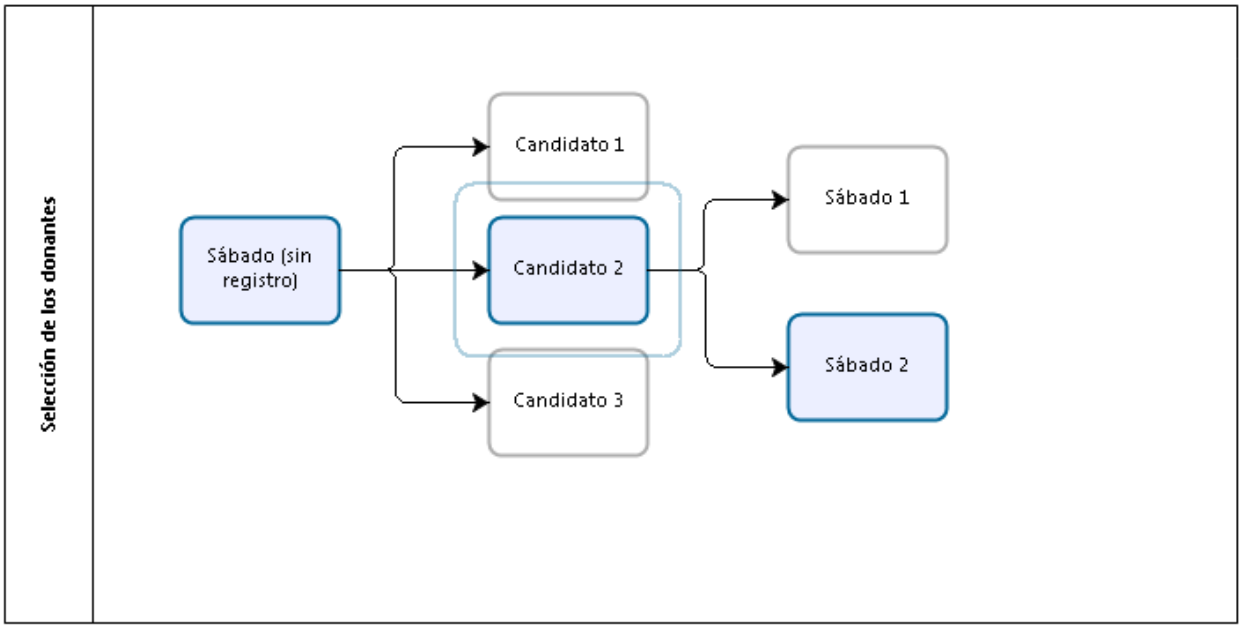
Por las características del método construido, es posible que para una LGI exista uno o más donantes. Ello obliga a tomar la decisión respecto a cuál de todos los posibles candidatos será el que se utilice finalmente para obtener los datos faltantes. Dicha selección se lleva a cabo de manera aleatoria con igual probabilidad para todas las unidades.

Una vez seleccionada la persona que hará la donación, los días faltantes son reemplazados por los del donante. Cabe señalar que el proceso asume que cada día de la semana tiene cierta especificidad en cuanto a los patrones de consumo, por lo que la donación se lleva a cabo considerando dicho supuesto. Así, por ejemplo, un día sábado será completado con los datos de otro sábado, ya sean de la misma o de otra persona. A raíz de ello puede ocurrir que para un día faltante existan hasta 3

días[14] disponibles para hacer la imputación. Cuando ello ocurre, se selecciona aleatoriamente a uno de dichos días.

La figura 16 muestra un ejemplo en el cual hay tres posibles libretas para hacer la donación, dentro de las cuales se selecciona aleatoriamente la segunda. Se observa que el donante cuenta con dos días disponibles para hacer la imputación, por lo cual se lleva a cabo una segunda selección aleatoria, que determina qué datos serán finalmente imputados.

Figura 16: Método de selección de los donantes



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Nota: En el segundo paso se selecciona de forma aleatoria el donante, para luego seleccionar el día de la semana correspondiente, también de forma aleatoria

3.4 Evaluación de los métodos

Para evaluar el desempeño que tiene cada uno de los métodos debe tenerse en consideración el escenario contra factual, lo cual equivale a hacerse la pregunta ¿qué hubiera ocurrido si quienes no respondieron, lo hubieran hecho? La mejor manera de evaluar el desempeño de cada uno de los métodos es contar con datos externos para todas aquellas personas que no respondieron la encuesta. Ello permitiría observar la diferencia que se produce entre los valores imputados y los "reales", para cada una de las técnicas aplicadas.

Existen algunos casos en los que se cuenta con información externa que permite observar el sesgo de no respuesta. Típicamente, los estudios de panel y las encuestas con muestras bifásicas contienen

[14] La estructura de las quincenas genera que estas pueden llegar a tener hasta 3 veces un mismo día de la semana (para quincenas de 16 días).

esta información. En el primer caso, se cuenta con mediciones de las unidades muestrales realizadas en versiones anteriores de la encuesta. En el segundo caso, la muestra se selecciona a partir de una muestra más grande, cuyo trabajo de campo ya fue realizado, por ende, previo al levantamiento se cuenta con información de los hogares. Lamentablemente, en muchas ocasiones no se cuenta con dicha información.

En el caso de la VIII EPF no existe información de gasto de los hogares que permita observar directamente el desempeño de cada uno de los métodos, sin embargo, sigue siendo necesario evaluar los resultados que estos tienen, para lo cual, usualmente, se recurre a ejercicios de simulación. La idea básica es generar un escenario de no respuesta a partir de los datos observados. Para ello, se toma como base las LGI completas, y se introducen valores perdidos, que luego son imputados con cada uno de los métodos. Una vez hecho esto, es posible comparar los valores imputados con los observados, de modo de evaluar la capacidad que tiene cada uno para enfrentar el problema de la no respuesta. En ese sentido, las simulaciones buscan emular la no respuesta, con el fin de estudiar el comportamiento de cada uno de los métodos.

3.4.1 Características de la simulación

Con el objeto de trabajar con supuestos un poco más realistas, algunos autores (Sukasih, Jang, Vartivarian, Cohen, & Zhang, 2009; West, 2009) sugieren no utilizar una respuesta completamente aleatoria, ya que es muy poco probable que esta tenga ese comportamiento, el cual no implicaría sesgo en el promedio de gasto por persona.[15] En vista de lo anterior, las simulaciones intentan recrear un escenario de no respuesta lo más realista posible, utilizando un mecanismo de no respuesta diferenciado para las **libretas rechazadas** y las libretas **parcialmente completadas**.

En el caso de las libretas rechazadas, se asume un mecanismo tipo MAR (por su nombre en inglés *missing at random*), para lo cual se requiere incorporar algún modelo, de modo de poder utilizar sus predicciones y así generar la no respuesta. Se utilizó el mismo modelo logístico presentado en el apartado sobre variables relacionadas con la no respuesta. En ese sentido, la probabilidad que tiene cada persona de responder está determinada por las predicciones de dicho modelo.

Respecto a las libretas contestadas parcialmente, el modelamiento es un poco más complicado y, de hecho, los modelos probados no fueron satisfactorios. Es así que, para este caso, se generó la no respuesta de manera un poco más simple. Se asumió un mecanismo completamente aleatorio, pero se tomó la precaución de mantener las proporciones de libretas según cantidad de días de registro observadas en la muestra. Esto quiere decir, por ejemplo, que si en la muestra se observó un 49% de libretas con 15 días de registro, en el conjunto de simulación se utilizó un porcentaje muy similar. Una vez establecidas estas proporciones, tanto las libretas asignadas a cada tramo de no respuesta, como los días borrados fueron seleccionados de manera completamente aleatoria.

Sobre la base de esta configuración, se elimina información de la base de datos, que luego es imputada, utilizando cada uno de los métodos. Ello permite observar la diferencia entre los valores observados e imputados, de modo de identificar cuál es el que tiene mejor rendimiento.

[15] Si la no respuesta fuera completamente aleatoria, pierde relevancia la imputación de libretas rechazadas, ya que bastaría únicamente con imputar las libretas parcialmente completadas (mediante *hot deck* o FNR) para corregir el sesgo de no respuesta en el promedio de gasto a nivel de LGI (a nivel hogar puede seguir existiendo sesgo).

Vale la pena señalar que producto del azar, cada vez que se haga el experimento de borrar e imputar datos, se obtendrán resultados diferentes[16]. Así, es posible que en algunos escenarios se obtenga una evaluación muy buena de los métodos y en otros, una muy mala. Con el objeto de que las conclusiones sean robustas al azar, usualmente se recurre a varias iteraciones. En este caso, se realizaron mil simulaciones, a partir de las cuales se construyen promedios y otros estadísticos agregados. Esto permite suavizar el efecto que tiene una sola simulación.

3.4.2 Resultados de las simulaciones

Los resultados de las simulaciones serán reportados a dos niveles: hogar (suma del gasto en LGI de sus integrantes) y persona (gasto por LGI). Si bien la EPF está diseñada para entregar información de gasto solo a nivel hogar, es conveniente observar los datos de las libretas antes de que estos sean agregados, pues ello permite comprender de mejor forma la manera en la que opera el sesgo de no respuesta.

El cuadro 5 muestra estadística descriptiva del gasto a nivel libreta para las mil simulaciones. Se aprecia que la media del gasto de las libretas después de borrar información (sin imputación) es menor al valor observado (185.471 pesos versus 191.036 pesos). El FNR, por su parte, sobreestima ligeramente el valor observado con un promedio de 196.490 pesos. Finalmente, el método de *hot deck* genera una media muy cercana al valor observado[17].

Es importante señalar que al graficar las distribuciones de medias obtenidas en cada uno de los métodos (figura 17), no se genera ninguna región en la que ambas se superpongan. Esto sugiere que, independiente de la aleatoriedad, los resultados siempre son distintos en cada uno de los métodos.

Cuadro 5: Resultados de las simulaciones (1000): gasto promedio a nivel libreta

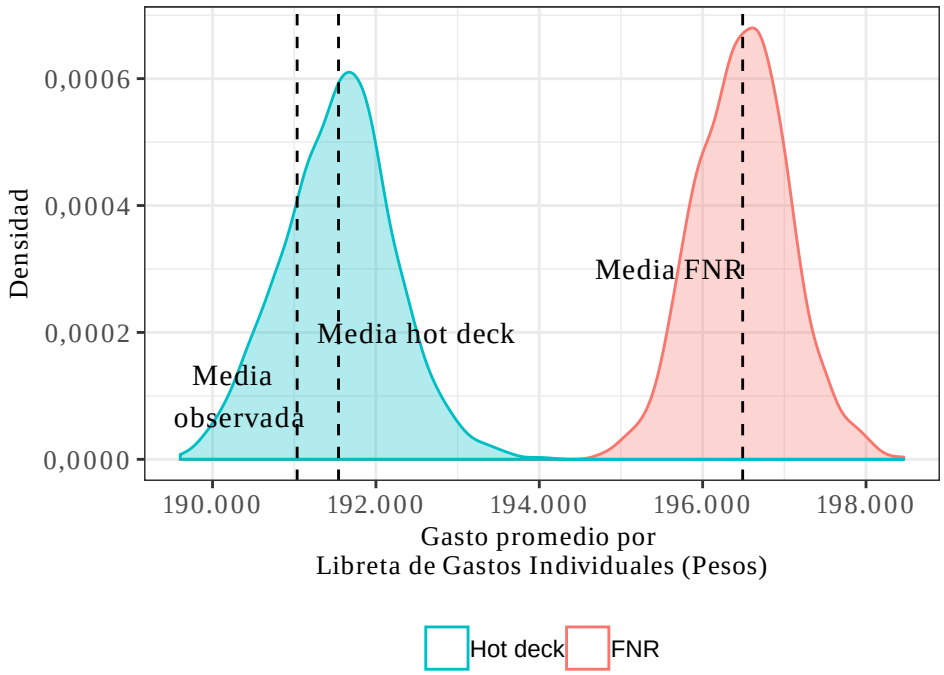
Estadísticos	Observado	Sin imputación	FNR	Hot deck
Media	191.036	185.471	196.490	191.542
Min	191.036	183.911	194.773	189.605
Max	191.036	187.151	198.462	194.017

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

[16] En la práctica, la implementación se lleva a cabo con semillas que van cambiando sobre la base de cierta secuencia. De este modo, es posible obtener el mismo resultado cada vez que se ejecuta la imputación.

[17] Es importante tener en consideración que los promedios están construidos sobre la base de conjuntos distintos de libretas. Esto se produce debido a que el método de *hot deck* incorpora las libretas rechazadas, mientras que el FNR, no.

Figura 17: Distribución de medias imputadas para la LGI



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Para comprender estos resultados es preciso tener en consideración que la no respuesta (libretas rechazadas) no se genera de manera completamente aleatoria, sino que en función de ciertas variables. De hecho, el modelo de respuesta tiene un sesgo a eliminar libretas de personas que tienen poco gasto, ya que como se ha señalado en otros apartados, quienes presentan un menor gasto, simultáneamente, tienen una menor propensión a responder. Dado que el FNR solo actúa sobre las libretas parcialmente completadas (y no sobre las rechazadas), la muestra resultante está compuesta por perfiles de personas con mayor gasto, lo cual tiende a sobreestimar el promedio observado. Por su parte, el método de *hot deck*, al incorporar todas las libretas (tanto las parcialmente completadas como las rechazadas), logra mantener en la muestra a aquellas personas que tienen una baja probabilidad de respuesta y, al mismo tiempo, una baja propensión a responder. Ello permite reducir el sesgo de no respuesta y obtener una media muy cercana a la observada.

En el cuadro 6 se muestran los mismos estadísticos anteriores, pero ahora referidos al gasto en LGI a nivel hogar. Nuevamente, se aprecia que el promedio, luego de haber borrado información, cae bastante. Ante esta situación, el FNR logra reducir en parte el sesgo generado por la no respuesta, aumentando el gasto promedio desde 286.090 pesos hasta 303.087 pesos. Finalmente, el método de *hot deck* alcanza una media bastante cercana al valor observado.

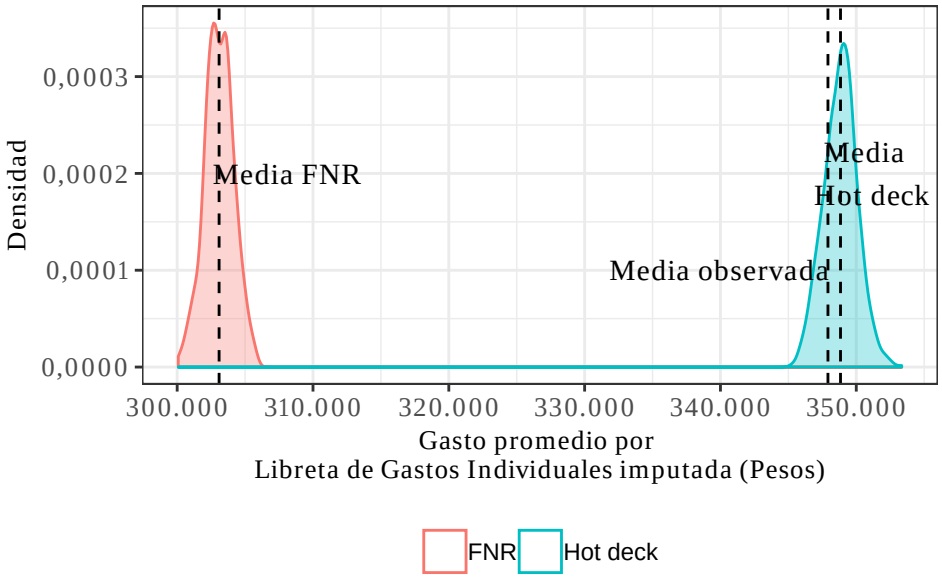
Cuadro 6: Resultados de las simulaciones (1000): gasto promedio a nivel hogar

Estadísticos	Observado	Sin imputación	FNR	Hot deck
Media	347.915	286.090	303.087	348.838
Min	347.915	283.089	300.081	345.309
Max	347.915	288.832	305.878	353.345

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Al observar la distribución de medias a nivel hogar (figura 18), se aprecia que los posibles resultados de cada uno de los métodos no se superponen y, de hecho, ambas distribuciones se encuentran bastante alejadas, lo que sugiere algo similar al caso anterior, es decir, independiente del componente aleatorio que introduce la simulación, ambos métodos siempre generarán diferentes resultados.

Figura 18: Distribución de medias imputadas para los hogares



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

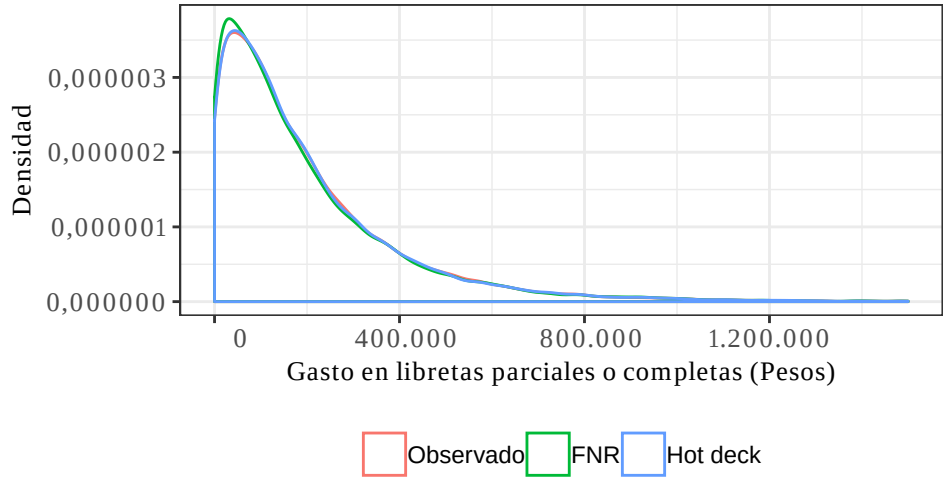
Dado que el FNR está diseñado para funcionar solo sobre las libretas que tienen respuesta parcial, vale la pena observar el comportamiento de los dos métodos en el conjunto de libretas que tiene al menos un día de registro (excluyendo las libretas rechazadas), ya que ello permite hacer una comparación sobre un mismo conjunto de libretas.

La figura 19 muestra las distribuciones de gasto generadas a partir de imputaciones por *hot deck* y FNR, para una de las mil simulaciones realizadas[18]. Se advierte que tanto la distribución *hot deck* (línea azul) como la del FNR (línea verde) se acercan bastante a la distribución observada (línea roja). Esto da cuenta de que en lo que respecta a la imputación de la no respuesta parcial de la LGI, ambos métodos presentan un comportamiento similar.

[18] Una simulación no es concluyente respecto al desempeño de los métodos, sin embargo, se llevó a cabo este mismo análisis gráfico en varias de ellas y los resultados no varían de manera importante.

No obstante lo anterior, es importante notar que el FNR tiende a alejarse del valor observado en algunas partes de la distribución, sobreestimando los gastos bajos y subestimando ligeramente aquellos ubicados en el rango que va de los 125.000 a los 500.000. Esto se debe, en gran medida, a que el FNR no logra completar las libretas que tienen menos de 6 días de registro, quedando estas con menos gasto del que deberían tener. En ese sentido, el método de *hot deck* nuevamente muestra un mejor desempeño que el FNR.

Figura 19: Distribución de gasto en las libretas con respuesta parcial y completa (al menos un día de registro)



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Continuando con el análisis anterior, la figura 20 muestra distribuciones de medias obtenidas a partir de las mil iteraciones para las libretas que tienen al menos un día de registro. Dado que se está trabajando con un subconjunto de la muestra (libretas con al menos un día de registro), se genera variabilidad en los datos observados.[19]

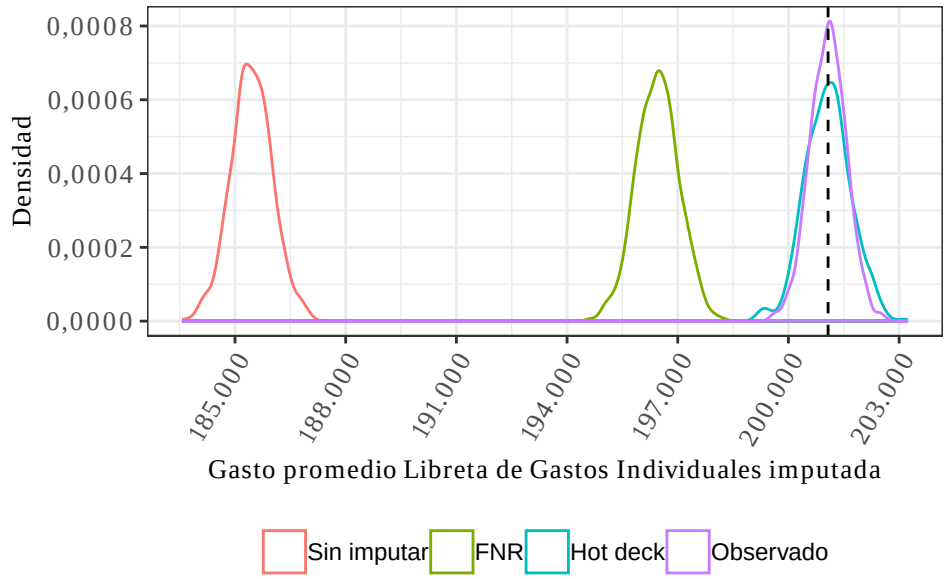
La curva en el extremo izquierdo corresponde a la distribución de gasto sin imputar (línea roja), lo cual es esperable, ya que una parte importante de la información de las libretas fue borrada. Por su parte, el FNR (línea verde), si bien se aleja un poco de la distribución observada, logra corregir gran parte del sesgo de no respuesta, desplazando la distribución hacia un mayor nivel de gasto de manera importante. En ese sentido, cabe destacar que, pese a su simpleza, este método no muestra un mal desempeño.

La curva correspondiente al método de *hot deck* (línea azul) es la que más se acerca a la distribución de medias de los datos observados. Vale la pena señalar que la distribución *hot deck*, en relación con la observada, es más abultada en los extremos y concentra menos valores en torno al promedio, lo cual implica un distanciamiento respecto a lo observado. Pese a ello, ambas distribuciones se encuentran

[19] Esto se produce porque en cada iteración las libretas seleccionadas para borrar parcial y completamente van cambiando, lo cual no sucede si se calculan promedios a partir de la muestra completa, como ocurre en los cuadros anteriores. Dicho de otro modo, al seleccionar las libretas que al menos tienen un día de registro, el conjunto resultante es distinto en cada simulación.

centradas en torno a un valor similar, lo cual es deseable desde el punto de vista de la corrección del sesgo.

Figura 20: Distribución de medias en las libretas con respuesta parcial y completa (al menos un día de registro)



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Una evaluación conjunta de los datos a nivel persona y hogar permite observar que la sobreestimación del FNR a nivel persona no logra compensar la disminución del gasto que se produce por la no inclusión de libretas rechazadas, generándose así una subestimación a nivel hogar. Esta situación da cuenta de que, incluso en un escenario en el cual las libretas rechazadas no introduzcan ningún sesgo en el promedio por persona, este se producirá a nivel hogar. El motivo de ello es que el número de libretas finales puede variar en función del método utilizado, controlándose así el efecto de incluir o no las libretas rechazadas, sin embargo, el número de hogares se mantiene constante, lo cual equivale a decir que para el caso del FNR todas las libretas rechazadas entran en el cálculo con un gasto igual a 0.

Dado que se cuenta con datos para mil simulaciones, es posible calcular la diferencia entre el promedio de los datos observados y el de los datos imputados, para cada una de las iteraciones (ecuación 9). Calculando una media de dichas diferencias, se obtiene un sesgo promedio (Sukasih et al., 2009). Esta medida indica cuánto se aleja la media de cada una de las distribuciones (FNR y *hot deck*) respecto del promedio observado.

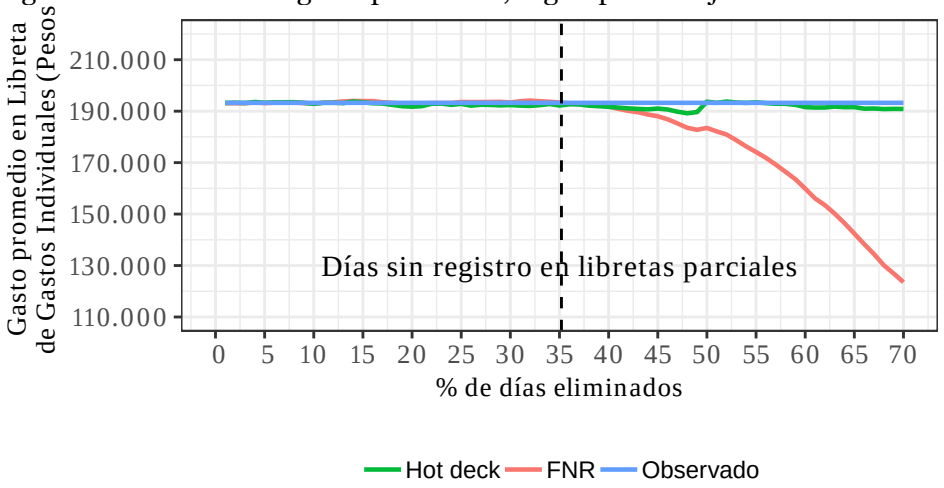
$$Sesgo = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\overline{X}_i} - \overline{X}) \tag{9}$$

Los resultados muestran que tanto a nivel hogar como a nivel libreta el FNR presenta mayor sesgo que el método de *hot deck*. A nivel de libreta el FNR presenta un sesgo positivo de 5.454 pesos, mientras

que el método de *hot deck* de 507 pesos. Por otro lado, a nivel hogar, para el FNR se obtiene un sesgo de -44.828 pesos, mientras que para *hot deck* dicho valor es de 923 pesos.

A continuación se realiza un último ejercicio en base a las simulaciones. Con el objeto de evaluar qué tan sensible es cada uno de los métodos al aumento de la no respuesta, se llevó a cabo el siguiente ejercicio. Se seleccionó aleatoriamente un 22% de las libretas[20] y se fueron eliminando días de registro, los cuales se imputaron con cada uno de los métodos. Se comenzó con un 1% de días borrados y se fue avanzando progresivamente hasta llegar a un 70%, lo cual permite observar cuál es el desempeño de cada uno de los métodos para cada nivel de no respuesta. Debe notarse que en los primeros niveles existe un número importante de libretas completas, el cual irá disminuyendo conforme se avanza en el porcentaje de días borrados. En contrapartida, el número de libretas parcialmente completado irá en aumento.

Figura 21: Promedio de gasto por libreta, según porcentaje de días borrados



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Los resultados de la figura 21 dan cuenta de que, hasta cierto nivel de no respuesta (cercano al 40%), los dos métodos son equivalentes y se ubican bastante cerca de la media observada. A partir de dicho punto, ambas curvas comienzan a distanciarse. Así, mientras el método de *hot deck* sigue oscilando en torno al valor observado, el FNR cae abruptamente. El comportamiento de este último es esperable, ya que desde cierto umbral, la proporción de libretas con muy pocos días de registro comienza a crecer aceleradamente y, como se ha señalado en el apartado sobre descripción de los métodos, cuando el número de días de registro es menor a 6, el FNR toma valor 2, lo cual implica que esas libretas, por construcción, quedan incompletas.

Respecto al comportamiento del método de *hot deck*, es notable la robustez que tiene al aumento de la no respuesta. Ello es importante, por cuanto arroja evidencia de que incluso en subpoblaciones que tengan un alto nivel de no respuesta parcial, el método puede seguir teniendo un buen desempeño.

[20] Este porcentaje corresponde a las libretas que presentan no respuesta parcial en la muestra.

A partir de los resultados de las simulaciones es posible concluir que en casi todas las pruebas realizadas el método de *hot deck* presentó un mejor desempeño que el FNR. A nivel hogar esta diferencia se explica principalmente por el hecho de que dicho método es capaz de atender al problema de las libretas rechazas, logrando corregir de mejor manera el sesgo de no respuesta. Un segundo aspecto relevante que sugieren los ejercicios de simulación es que la no respuesta muy probablemente genera sesgo en el promedio de gasto, lo cual refuerza aún más la idea de que es relevante incorporar algún procedimiento de imputación, que intente corregir el sesgo por no respuesta.

Sin perjuicio de lo anterior, es importante señalar que los ejercicios que excluyen a las libretas rechazadas dan cuenta de que el FNR es una estrategia de imputación adecuada para lidiar con la no respuesta parcial. Pese a que su desempeño es inferior al del método de *hot deck*, su simplicidad lo convierte en un buen candidato para lidiar con las libretas parcialmente completadas. En ese sentido, lo que hace preferible al *hot deck* por sobre el FNR es la capacidad que tiene el primero de imputar libretas completamente rechazadas.

3.4.3 Resultados con datos oficiales

En el presente apartado se describen los resultados de la imputación de gastos diarios para los datos oficiales de la VIII EPF. En primer lugar, se muestra un resumen de la imputación por medio del método de *hot deck*, para luego presentar algunos estadísticos que buscan comparar el FNR con este método, de modo de dar cuenta del efecto que tiene en el gasto promedio un cambio en la metodología de imputación de gastos diarios.

El cuadro 7 muestra la cantidad y porcentaje de libretas y días imputados en cada nivel de la matriz de transferencia (anexo cuadro 25), junto con el tamaño promedio de los *clusters* formados en cada nivel de imputación. Se observa que a medida que el nivel de imputación aumenta, el tamaño promedio de los *clusters* crece. Esto es esperable, ya que conforme se avanza en los niveles de imputación, la exigencia para encontrar donantes disminuye, lo cual implica que para un dato faltante sea posible encontrar una mayor cantidad de candidatos para efectuar la donación. Dado que en los primeros niveles el criterio de similitud para encontrar donantes es más exigente, las imputaciones allí realizadas son de mejor calidad que la de los siguientes niveles.

Teniendo en consideración lo anterior, se observa que el 36,13% de las libretas sujetas a imputación se completan en el nivel 1, lo que indica que una parte importante de las libretas son imputadas con información del mismo informante[21]. Por otro lado, los niveles más altos de imputación, donde el tamaño promedio de los *clusters* es mayor a diez, concentra un reducido número de imputaciones en libretas (4,08%) y días (5,66%). Eso quiere decir que, en general, las imputaciones se llevan a cabo en *clusters* pequeños, los cuales pueden ser asociados a imputaciones de buena calidad.

[21] El nivel 1 de la matriz de transferencia corresponde al más exigente, con las variables de búsqueda al nivel más de desagregado. El nivel más exigente implica la búsqueda del día de la semana faltante dentro de los días con respuesta del mismo informante. Véase anexo, cuadros 25 al 28 para el detalle de cada nivel de dicha matriz.

Cuadro 7: Niveles en los que se realizaron las imputaciones

Nivel de imputación	Tamaño promedio cluster	Número de libretas	Porcentaje libretas	Número de días	Porcentaje días
1	1,0000	5.464	36,13	17.232	11,65
2	1,0012	5	0,03	64	0,04
5	1,0018	5	0,03	75	0,05
6	1,0024	6	0,04	79	0,05
7	1,0078	60	0,40	803	0,54
8	1,0185	28	0,19	258	0,17
9	1,0205	3	0,02	25	0,02
10	1,0619	108	0,71	1.019	0,69
11	1,1092	122	0,81	1.146	0,77
12	1,1943	1.510	9,98	21.539	14,56
13	1,2589	673	4,45	9.032	6,11
14	1,5420	1.343	8,88	17.968	12,15
15	2,4421	2.707	17,90	36.401	24,61
16	2,8406	607	4,01	8.306	5,61
17	3,6229	729	4,82	10.074	6,81
18	7,8107	1.138	7,52	15.541	10,51
19	16,4134	421	2,78	5.692	3,85
20	91,7713	163	1,08	2.190	1,48
21	90,4508	30	0,20	437	0,30
22	673,5357	3	0,02	46	0,03

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Al fijar la atención en el cuadro 8, se observa que el 40,1% de las LGI tiene al menos un día imputado. Esta cifra debe ser analizada con precaución, ya que puede inducir a pensar que dicho valor corresponde al porcentaje de imputación realizado, lo cual no es correcto. En realidad, el 40,1% recién mencionado incluye tanto las libretas imputadas en su totalidad como aquellas imputadas parcialmente[22]. Es por ello que para tener un panorama más completo se debe observar el porcentaje que representan los **días** imputados dentro del total de días que debiese haber en la VIII EPF. Dicho porcentaje es 25,79%, bastante menor al expuesto a nivel de libretas. Sin embargo, sigue tratándose de una proporción relevante, en lo que respecta al efecto que podría tener la imputación sobre el gasto promedio de los hogares.

Cuadro 8: Resumen de las imputaciones

	Número de libretas	Porcentaje libretas	Número de días	Porcentaje días
Imputado	15.125	40,1	147.927	25,8
Total	37.718	100,0	573.637	100,0

Fuente: Instituto Nacional de Estadísticas - VIII Encuesta de Presupuesto Familiares (EPF)

En el cuadro 9 se muestra el efecto que tiene en el gasto promedio el hecho de reemplazar el FNR por el método de *hot deck*. Se advierte que este último genera un alza del gasto promedio a nivel hogar de 65.315 pesos. Este aumento se debe a que el número de hogares se mantiene con ambos

[22] Esto quiere decir que incluso las libretas que recibieron un día de registro se incluyen en el 40,1% mencionado.

métodos (3.373.786 hogares expandidos), sin embargo, *hot deck* aumenta el número de LGI con registro desde 7.043.928 a 8.656.977[23]. Así, la diferencia en el promedio entre los dos métodos se debe en gran medida a que el FNR de manera implícita imputa gasto cero a las libretas rechazadas, mientras que el método de *hot deck* en muchos casos imputa un valor distinto de cero.[24]

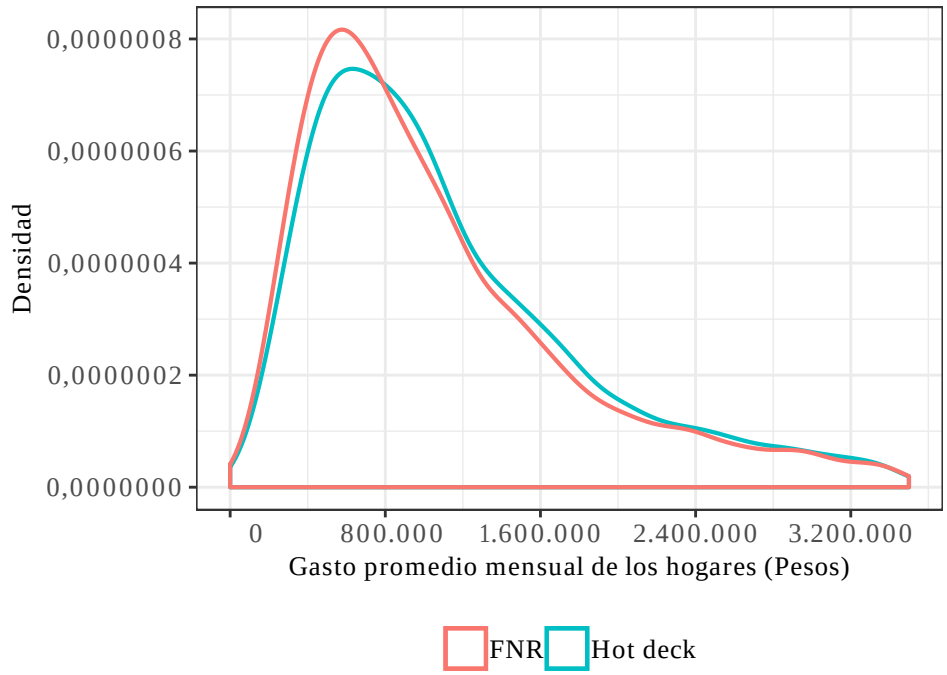
Cuadro 9: Comparación de método Hot deck y FNR, VIII EPF

Imputación	Gasto - LGI	Gasto - Hogar Total	Hogares FE	LGI FE	Hogares Muestrales	LGI Muestrales
HD - VIII EPF	180.390	1.121.925	3.373.786	8.656.977	15.239	37.705
FNR - VIII EPF	195.508	1.056.610	3.373.786	7.043.928	15.239	30.937

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

La figura 22 muestra una representación gráfica de cómo se modifica la distribución del gasto de los hogares al cambiar del método FNR a *hot deck*. Con FNR la parte baja de la distribución (izquierda) es más abultada, es decir, genera un mayor número de hogares con ingresos bajos. Por su parte, *hot deck* suaviza la distribución, traspasando casos desde la parte baja a la parte media.

Figura 22: Distribución del gasto de los hogares. FNR y Hot deck



Fuente: Instituto Nacional de Estadísticas – VIII Encuesta de Presupuestos Familiares (EPF)

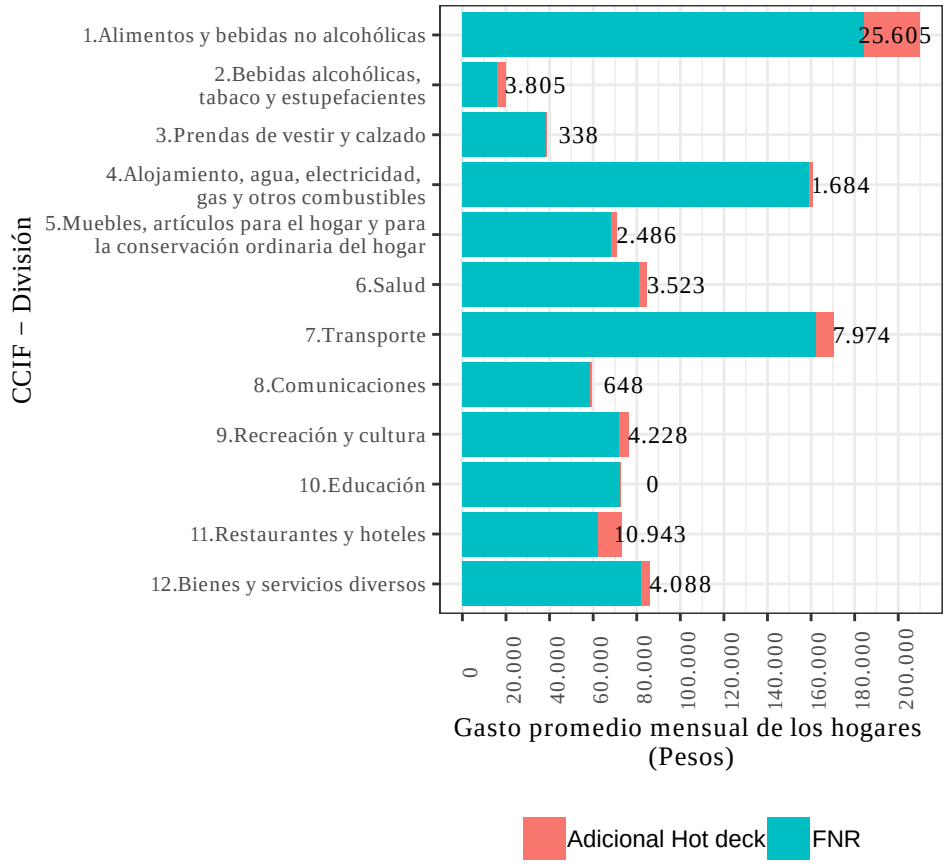
[23] El motivo es que el método de *hot deck* hace permanecer en la muestra las libretas que inicialmente habían sido rechazadas, mientras que el FNR las excluye.

[24] Es importante tener en consideración que la imputación por *hot deck* contempla la posibilidad de que existan libretas completadas con gasto cero, ya que una libreta puede estar respondida únicamente con días sin gasto, que se considera un registro válido.

Una vez identificado el incremento producido por la imputación en el gasto promedio, es importante analizar cómo se distribuye dicho incremento. La figura 23 muestra una comparación entre FNR y *hot deck* respecto al gasto promedio en cada una de las divisiones CCIF[25]. En relación con esto, las divisiones 1 (Alimentos y bebidas no alcohólicas), 11 (Restaurantes y hoteles) y 7 (Transporte) son las que exhiben las alzas más importantes. Esto se debe a que la Libreta de Gastos Individuales corresponde al instrumento de recolección donde se capturan en mayor medida este tipo de gastos[26].

El hecho de que el gasto no aumente en la misma proporción en todas las divisiones sugiere que no realizar una imputación para las LGI rechazadas no solo introduce un sesgo en el gasto promedio por hogar, sino también en el gasto por división.

Figura 23: Gasto promedio mensual de los hogares, según división de gasto



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Ahora bien, si se analizan las diferencias en la estructura de gastos de los hogares[27] entre ambos métodos de imputación (figura 24), se observa que si bien *hot deck* genera un aumento importante

[25]Clasificación del Consumo Individual por Finalidades

[26]En el caso de la división 11, que agrupa gastos en restaurantes y hoteles, la LGI captura principalmente lo concerniente a restaurantes y compra de alimentos para servir.

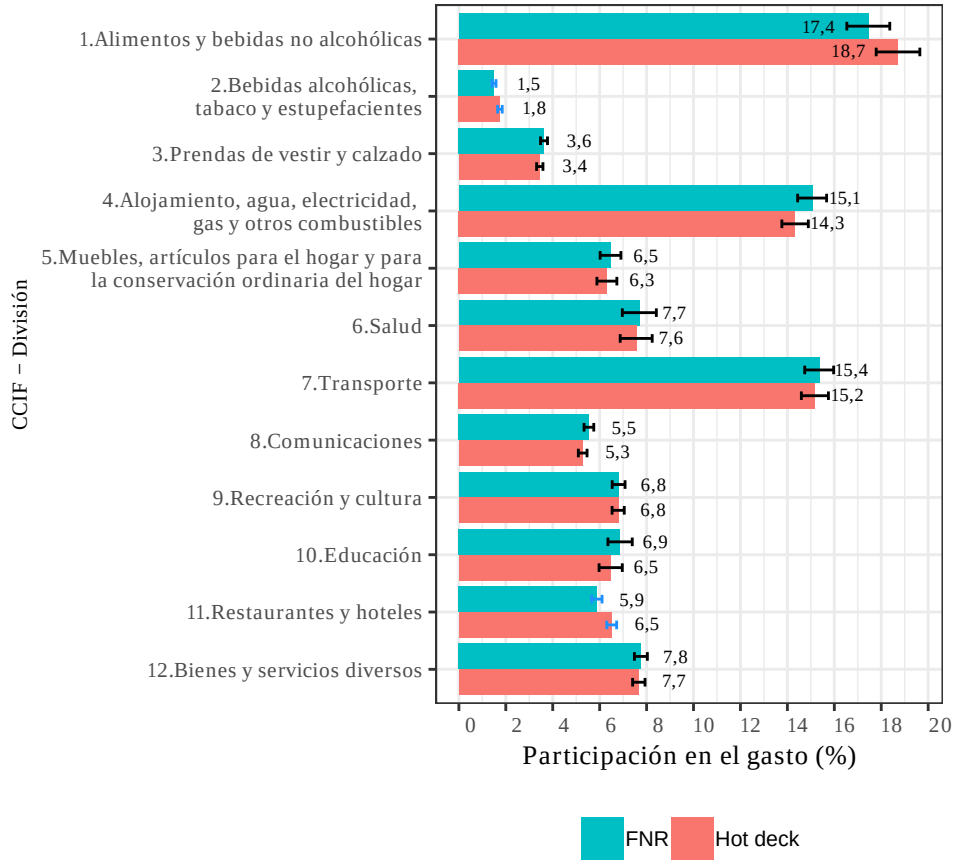
[27]Esto es, la proporción del gasto promedio mensual de los hogares destinada a cada división de gasto.

en el gasto de la división 1 (figura 23), en términos proporcionales dicha diferencia no es estadísticamente diferente entre los dos métodos. Lo mismo sucede con la división 7.

En el caso de la división 11, sí es posible observar un aumento estadísticamente significativo en su participación en la estructura de gastos. Por su parte, la división 2 (Bebidas alcohólicas y estupefacientes), aunque en la figura 23 no muestra una gran alza en el monto, es un efecto estadísticamente significativo en términos relativos. Esto quiere decir que, en lo que respecta a la participación de cada división en el gasto, solo en las divisiones 2 y 11 existe una diferencia estadísticamente significativa entre ambos métodos.

La participación de las demás divisiones de gasto en la estructura varía, pero no de forma significativa. Sin embargo, las estimaciones puntuales de estas sí se ven afectadas, experimentando una reducción en su participación, con excepción de la división 1, que aumenta, debido a la magnitud de captura de este tipo de gastos en la LGI.

Figura 24: Participación de cada división en el gasto promedio de los hogares



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

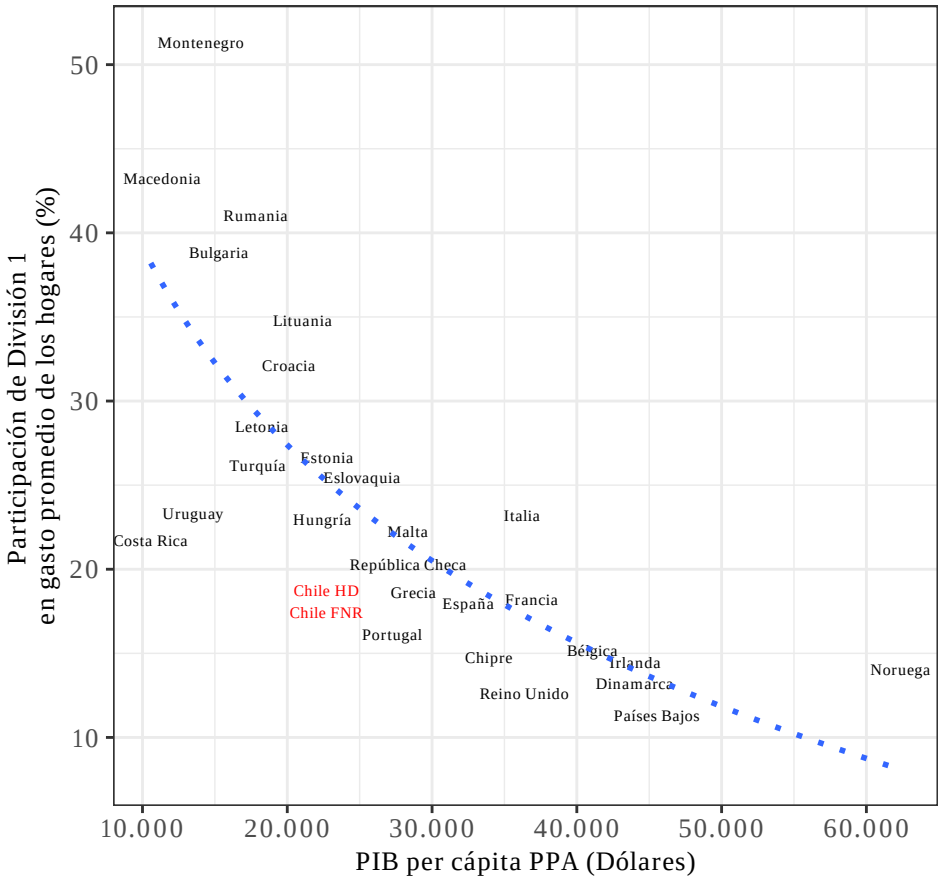
Por último, en la figura 25 muestra cómo se comporta la participación de la división 1 por país en función de su PIB per cápita PPA. Se hace evidente una correlación inversa, esto es, en países con ingreso per cápita mayor, la participación en el presupuesto familiar de los alimentos y bebidas no

alcohólicas es menor. Este patrón es esperable, por cuanto los hogares con más recursos, al tener cubiertas sus necesidades de alimentación, pueden destinar mayor proporción de su presupuesto a otras divisiones de gasto. Cabe destacar que lo anterior no implica que los países con mayores ingresos destinen un menor monto a la división uno, sino que, de un monto mayor, destinan un porcentaje menor.

Es interesante observar que la tendencia a la baja en la participación de la división 1, dado el ingreso *per cápita*, es decreciente. Esta reducción en la pendiente sugiere que mientras más alto sea el ingreso del país, un aumento en su ingreso tendrá un menor efecto sobre la participación de la división 1.

En la figura 25 utilizando los datos de la VIII EPF Chile, se ha incluido la participación de Chile realizando imputación por FNR (17,4%) y por *hot deck* (18,7%). A este respecto, es posible observar que con la imputación por *hot deck*, Chile se acerca un poco a la línea de tendencia (en comparación con FNR). Esta línea marca el nivel promedio esperado de participación de la división 1, según el ingreso *per cápita*[28]. Esto es relevante, ya que la EPF la participación de la división 1 en Chile está por debajo de lo que se esperaría para un país de su ingreso.

Figura 25: PIB per cápita PPA y participación de alimentos en el gasto promedio de los hogares



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF) – Eurostat, ENGIH 2005–2006, INEC 2006

[28] La línea de tendencia corresponde a una regresión logarítmica.

Considerando los ejercicios de simulación y los resultados con datos oficiales, en la VIII EPF se decidió imputar gastos diarios utilizando el método de *hot deck*. Esta decisión implicó un cambio metodológico importante respecto a la VII EPF, ya que se implementó la imputación de libretas rechazadas. Cabe destacar que, pese a que este es un cambio relevante, el efecto en los principales indicadores de gasto no es significativo, lo cual es deseable, ya que permite mantener un cierto nivel de comparabilidad en el tiempo. De hecho, la diferencia en el gasto promedio de los hogares entre FNR y *hot deck* es de 65.315 pesos, cifra que no es estadísticamente significativa. El efecto más acusado se observa en la división de alimentos y bebidas no alcohólicas, cuya diferencia es de 25.605, lo cual se traduce en un aumento en su participación de 1,3 puntos porcentuales. Como se ha señalado, ello es deseable, por cuanto acerca un poco a Chile al valor esperado según su ingreso *per cápita*.

4 Libreta de Ingresos (LI)

4.1 Contexto

4.1.1 Características generales de la Libreta de Ingresos (LI)

La Libreta de Ingresos es un cuestionario que responde al objetivo secundario de la EPF, es decir, permite identificar la estructura del ingreso total disponible de los hogares urbanos del país y algunas de sus conurbaciones.

En esta libreta se registra información respecto a los ingresos monetarios y no monetarios de todos los integrantes del hogar de 15 años o más. El período de referencia es, en la mayoría de los casos, el mes anterior al de la entrevista, registrándose los ingresos generados en dicho período, sin importar si estos realmente se recibieron durante ese mes (criterio de ingreso devengado).

El cuestionario de ingresos se organiza en los siguientes módulos:

- Ingresos del trabajo dependiente
- Ingresos del trabajo independiente
- Otros ingresos del trabajo
- Ingresos por jubilaciones y/o pensiones de vejez
- Ingresos por otras pensiones y transferencias recibidas
- Otros ingresos de carácter no habitual
- Ingresos de la propiedad y ganancias por tenencia de instrumentos financieros

4.1.2 Experiencia internacional y nacional en imputación de ingresos

Si bien se observa que las oficinas nacionales de estadísticas (en adelante, ONEs) atienden el problema de la no respuesta parcial de ingresos de distinta manera, en general todas desarrollan estrategias para mitigar los posibles efectos que esta situación puede tener en las estimaciones. Al mismo tiempo, a pesar de la diversidad de metodologías, es común el uso de algoritmos basados en la idea de *hot deck* y *matching*.

A continuación, se presentan algunas metodologías que forman parte de los procedimientos de imputación que las ONEs llevan a cabo en sus productos. Del mismo modo, se revisan recomendaciones realizadas por organismos como ONU y algunos trabajos académicos llevados a cabo en diferentes países.

Manual de Canberra (ONU): El Manual del Grupo de Canberra de las Naciones Unidas (ONU, 2011) presenta la heterogeneidad de estrategias para abordar esta problemática en las distintas regiones y países del mundo. Así por ejemplo, en la Unión Europea, algunos países como Irlanda, Malta, Luxemburgo, Países Bajos y Hungría decidieron eliminar los hogares que incluyeran al menos una persona que no reportara ingresos. En el caso de Bulgaria, Alemania, Grecia, Letonia, Portugal, Rumania y Eslovaquia se siguió una estrategia de aplicación de factores de ajuste a los ingresos del hogar. Finalmente, Bélgica, República Checa, Estonia, Francia, Chipre, Italia, Lituania, Austria, Polonia, España y el Reino Unido utilizaron procedimientos de imputación (ONU, 2011).

Australia: La Oficina de Estadísticas de Australia, en su encuesta de gastos del hogar (*Household Expenditure Survey*), realiza procedimientos de imputación de la no respuesta parcial, los que se basan en la idea de un mecanismo de *matching* de información basada en diversas características (como región, sexo, edad, estado en la fuerza laboral e ingreso) entre las personas que responden el cuestionario completo (donantes) y las personas con información faltante. Los donantes son escogidos aleatoriamente y, dependiendo de qué variables se quieren imputar, se asigna la información completa al set de preguntas faltantes (ABS, 2017).

Canadá: En la Encuesta Canadiense de Ingresos (*Canadian Income Survey, CIS*) también se utiliza un procedimiento basado en la idea de *matching* para imputar la mayoría de las variables de ingreso. Este método considera la selección de un donante sobre la base de un set de variables que se encuentran correlacionadas con los ítems de ingreso. Utilizando una función de distancia, el donante más cercano es elegido para imputar los casos faltantes (Statistics Canada, 2018)

México: En Rodríguez-Oreggia & López-Videla (2015) se realizan diversos ejercicios de imputación de ingresos utilizando la Encuesta Nacional de Ocupaciones y Empleo. En particular, se estudia el desempeño de dos metodologías y se presenta una corrección de estimaciones por remuestreo para las observaciones con ingreso reportado.

La metodología de Rodríguez-Oreggia & López-Videla (2015) también se basa en un mecanismo de *matching*. En particular, se utilizan dos variantes del método *hot deck*. Este método permite generar celdas con un vector de variables sociodemográficas, de manera tal de identificar individuos similares en estas dimensiones, para posteriormente imputar un ingreso a las observaciones con datos faltantes. Las dos variantes analizadas son el *hot deck* con imputación aleatoria, donde la asignación de individuos donantes a receptores se realiza al azar, y el *hot deck* con función de distancia. En este último caso, la asignación cumple el criterio de minimizar una medida de distancia (en particular, de Mahalanobis[29]) entre el donante y el receptor. Ambos métodos de imputación muestran un sesgo en las estimaciones de ingreso, lo que impacta a nivel de indicadores sociales, como una sobreestimación de la tasa de pobreza laboral.

Argentina: Donza (2013) realiza una exhaustiva revisión de las tendencias de la no respuesta de ingresos en la Encuesta Permanente de Hogares y de las alternativas que pueden tomarse para su corrección. Con respecto a los tratamientos que no consideran imputación se mencionan: análisis con datos completos o *listwise deletion*, donde se incluyen solo las unidades que poseen información completa en todas las variables; análisis con los datos disponibles o *pairwise deletion*, donde se consideran todos los datos de los que se dispone para cada variable analizada; y el ajuste de ponderadores.

Los procedimientos de imputación analizados en el estudio consideran los métodos de imputación por la media, la imputación deductiva, imputación *cold deck*, imputación *hot deck*, imputación por regresión, imputación mediante el método de regresión secuencial multivariante y la estimación por máxima verosimilitud. También se analizan métodos de imputación múltiple e imputación de modelos de Monte Carlo con Cadenas de Markov (MCMC).

[29] Algunos estudios (West et al., 1990) plantean también la posibilidad de utilizar la distancia euclidiana. La ventaja de la distancia de Mahalanobis es que su cálculo pondera por la matriz de varianzas y covarianzas del vector de covariables utilizadas en la imputación.

El autor expone las ventajas de la imputación múltiple, dada la mayor eficacia de los estimadores (pues se minimiza el error estándar), la mayor validez de las inferencias del modelo y la posibilidad de generar sensibilidad de estas inferencias. No obstante, dado que la imputación múltiple no genera una única respuesta y que en muestras grandes produce resultados similares, recomienda el uso del método de máxima verosimilitud.

Brasil: El Instituto Brasileño de Geografía y Estadísticas (IBGE) aplica métodos de imputación de ingresos en diferentes productos que capturan esta variable: Censo Demográfico, Encuesta Nacional por Muestra de Viviendas (*Pesquisa Nacional por Amostra de Domicílios, PNAD*) y Encuesta de Presupuestos Familiares (*Pesquisa de Orçamentos Familiares*). En general, el método más utilizado por los diversos productos corresponde a la imputación por *hot deck*, la que en todos los casos aparece como más próxima a los datos reales y armonizada entre encuestas (Souza, 2015).

Encuesta CASEN: la metodología aplicada para la imputación de datos faltantes en las preguntas de ingreso se basa en la imputación por medias, donde se asigna a cada dato faltante (receptor) el valor declarado en promedio por los casos más similares posibles (donantes). En el caso de los trabajadores ocupados, esta imputación se realizó separando a los trabajadores asalariados e independientes, y utilizando una batería de variables para establecer la similitud entre receptores y donantes (ubicación geográfica del hogar, tramo de edad, sexo, nivel educativo por tramos, categoría de la ocupación, entre otras). Para el caso de los perceptores de jubilaciones y pensiones, se agregan al análisis las variables nivel educativo alcanzado y parentesco respecto del jefe de hogar (Ministerio de Desarrollo Social, 2017).

Encuesta Suplementaria de Ingresos (ESI): el método de imputación aplicado en esta encuesta, producida por el INE, es una variante del método de medias condicionadas, en donde se identifica un grupo de individuos con información de ingresos (donantes) que tengan características similares a las del perceptor, imputándole la mediana de los ingresos del grupo que posee información (INE Chile, 2018).

4.1.3 Definición de la no respuesta en Libreta de Ingresos (LI)

Al igual que la no respuesta en la LGI, la no respuesta en la Libreta de Ingresos (en adelante, LI) forma parte de la no respuesta parcial de la encuesta. El año 2018 se publicó el Informe de Calidad de la VIII EPF (INE Chile, 2018a) donde la no respuesta parcial para cada uno de los cuestionarios fue una de las temáticas analizadas. Cabe aclarar que se observan diferencias al comparar los indicadores de ese informe con los de este documento, las que se producen por la manera distinta de operacionalizar el nivel de no respuesta parcial en cada uno. Si bien tanto los indicadores del Informe de Calidad como los de este documento buscan atender un mismo fenómeno - la no respuesta al ítem -, considerando que estos responden a distintos objetivos, la manera de aproximarse a este problema también presenta matices.

En el Informe de Calidad la no respuesta se definió para todos los cuestionarios a nivel de módulo. En este sentido, se consideró que una persona no había respondido un módulo cuando todas las preguntas asociadas a este se encontraban sin respuesta (no sabe o no responde). Esta definición de la no respuesta parcial permite contar con una medida comparable entre los más de 50 módulos

que tiene la encuesta y permite elaborar de manera sintética un panorama general de la no respuesta parcial en la EPF.

Para el caso del proceso de imputación se requiere de una forma de cálculo que atienda la no respuesta parcial al nivel de determinadas preguntas del cuestionario. La imputación de ingresos busca responder a la falta de respuesta para las preguntas de ingresos de la ocupación principal y de las jubilaciones. En este sentido, la definición de la no respuesta se encuentra sujeta a la falta de respuesta en las preguntas que permitan construir estas partidas. A continuación, se detalla la forma de operacionalización de la no respuesta para cada partida de ingresos imputada en la encuesta. En cada uno de los casos se indica entre paréntesis el número de la pregunta al que corresponde cada ítem en la LI.

1) **Asalariados y honorarios**: falta de respuesta (simultánea) en las preguntas sobre el sueldo bruto del mes pasado (TA02), sueldo líquido mes pasado (TA09) y sueldo líquido mensual promedio de los últimos 12 meses para quienes tienen sueldo variable (TA10)
2) **Cuenta propia y profesionales independientes**: falta de respuesta (simultánea) en las preguntas sobre el monto neto de los ingresos del mes pasado (TI01), monto disponible de los ingresos del mes pasado (TI06) y monto disponible promedio mensual de los últimos 12 meses (TI08)
3) **Jubilados**: falta de respuesta (simultánea) en las preguntas sobre el monto bruto jubilación del mes pasado (JU02) y el monto líquido jubilación del mes pasado (JU03)

La selección de estas preguntas dice relación con los ítems mínimos para construir el ingreso de la ocupación principal o de las jubilaciones. Si todos estos faltan y la persona señala recibir esta partida, entonces se decide imputar, para este caso, la variable de ingresos correspondiente.

4.2 Análisis de la no respuesta en la LI

4.2.1 Características generales de la no respuesta en la LI

Las principales fuentes de ingresos de los hogares son aquellas que provienen del ejercicio de una ocupación y los ingresos de pensiones de vejez y jubilaciones. En términos agregados, estas categorías de ingreso representan el 87,6% del total de ingresos disponibles. Teniendo en cuenta el peso de estas categorías en la composición del ingreso de los hogares, los esfuerzos para lidiar con la falta de respuesta parcial se concentran también en estas fuentes.

Una de las dificultades que enfrentan las encuestas de caracterización socioeconómica, en general, es la captura de los ingresos de las personas. Particularmente, estudios que comparan los microdatos tributarios con los datos provenientes de las encuestas a hogares observan una tendencia a la subestimación de los ingresos más altos de la población por parte de estas últimas (Burdin et. al, 2014; Jiménez, 2015). En un contexto en donde la población colabora cada vez menos con las encuestas en general, resulta importante contemplar estrategias de mitigación frente a esta situación.

La estrategia de imputación de ingresos parte con la identificación de distintos grupos de trabajadores, que junto a los jubilados corresponden a su vez a categorías de ingresos utilizadas

para la imputación. Los grupos identificados son: asalariados, honorarios, trabajadores por cuenta propia, profesionales independientes. Las primeras dos categorías corresponden a trabajadores dependientes, mientras que la tercera y la cuarta, a independientes. Estos grupos ocupacionales, junto a los jubilados, constituyen las categorías de análisis empleadas para el procedimiento de imputación, las que serán denominadas grupos o categorías de ingreso. La consideración de estos grupos busca ganar precisión en el resultado de las imputaciones al especificar, de acuerdo a las características de estos grupos, la relación entre las variables explicativas y los ingresos de las personas. Aunque las variables predictoras puedan ser similares entre estos (ej. sexo, educación, edad, etc.), la relación entre estas variables y los ingresos es distinta para cada grupo. Así, por ejemplo, al comparar los grupos de trabajadores, fijando para todos algunas características comunes (área de residencia y sexo, por ejemplo), se observan diferencias en el promedio y dispersión de los ingresos de estos.

Cabe mencionar que las categorías de ingreso definidas para la imputación son diferentes a aquellas utilizadas para presentar los resultados de la VIII Encuesta de Presupuestos Familiares. Esto se debe a que la selección de los grupos para la imputación responde a la búsqueda de grupos que compartan características en un determinado mercado laboral y que se asocien con el ingreso, para, como se mencionó anteriormente, ganar precisión en los resultados de las imputaciones.

Cuadro 10: Estadísticas descriptivas por categoría de Ingresos - Mujeres Región Metropolitana

Partida	Promedio	Desviación	50%	Mínimo	Máximo
Asalariados	689.144,7	777.557,1	440.000	10.000	13.158.333
Honorarios	702.482,3	813.202,8	480.000	0	8.152.573
Cuenta propia	284.851,9	506.856,5	156.000	0	7.500.000
Prof. independientes	476.168,9	1.024.920,9	200.000	2.750	16.308.604
Jubilados	193.427,9	175.124,4	138.000	3.000	1.718.337

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

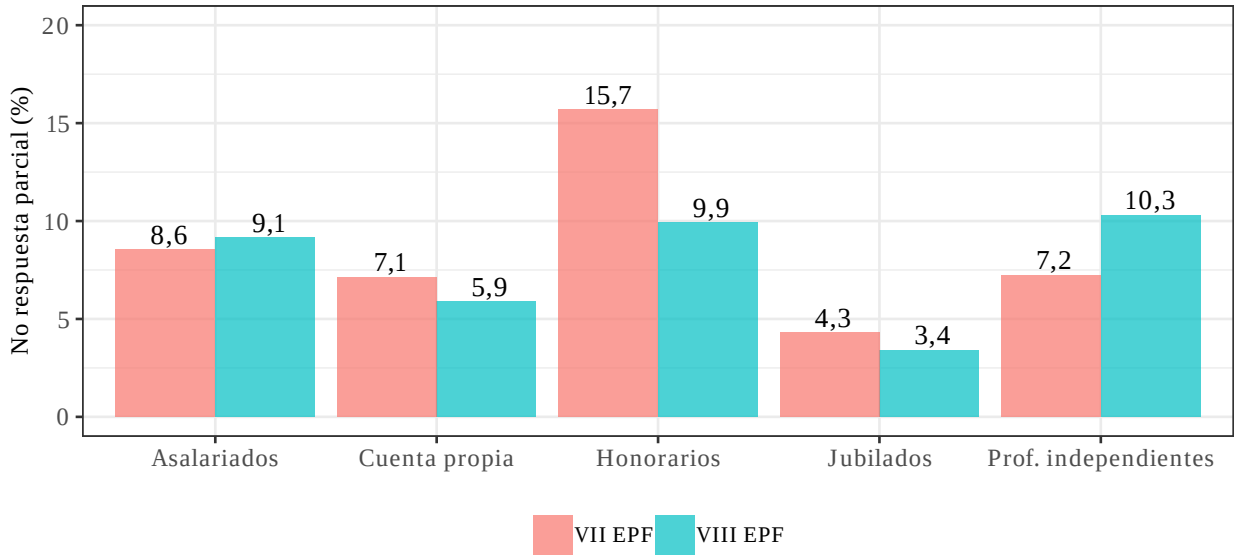
Aunque se observa una similitud entre los ingresos de los trabajadores asalariados y honorarios, estos grupos presentan diferencias respecto a la obligatoriedad de sus cotizaciones previsionales, lo que motiva que su proceso de imputación se realice de manera separada. La reforma previsional chilena establece que los trabajadores que entreguen boletas deberán comenzar a cotizar de manera obligatoria a partir del año 2018. Dado el período de referencia de la VIII EPF, las cotizaciones de los trabajadores a honorarios se realizan en un contexto en donde estas son voluntarias y por montos diferentes a los obligatorios para los trabajadores asalariados con contrato.

Para situar este proceso de imputación, a continuación, se analizan las características de la no respuesta parcial para las categorías de ingreso de interés. Esta caracterización considerará variables que eventualmente podrían estar relacionadas con la falta de respuesta en cada uno de los grupos de ingresos. El estudio de la no respuesta es un aspecto fundamental de todo proceso de imputación, toda vez que, como fue mencionado en el capítulo anterior, varios algoritmos para imputar datos faltantes asumen cierta distribución de la no respuesta (completamente aleatoria o condicionalmente aleatoria).

En comparación a otras variables que también se imputan para esta encuesta (como lo son los gastos individuales), la falta de respuesta para el ítem de ingresos de la ocupación es relativamente bajo

(ver figura 26). El grupo de los profesionales independientes es el que presenta el nivel de falta de respuesta más alto con un 10,3%. Le siguen el grupo de los trabajadores a honorarios (dependientes) con un 9,9%. Por su parte, el grupo de ingreso jubilados es el que presenta menor no respuesta parcial, con un 3,4%.

Figura 26: No respuesta parcial por categorías de ingreso



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

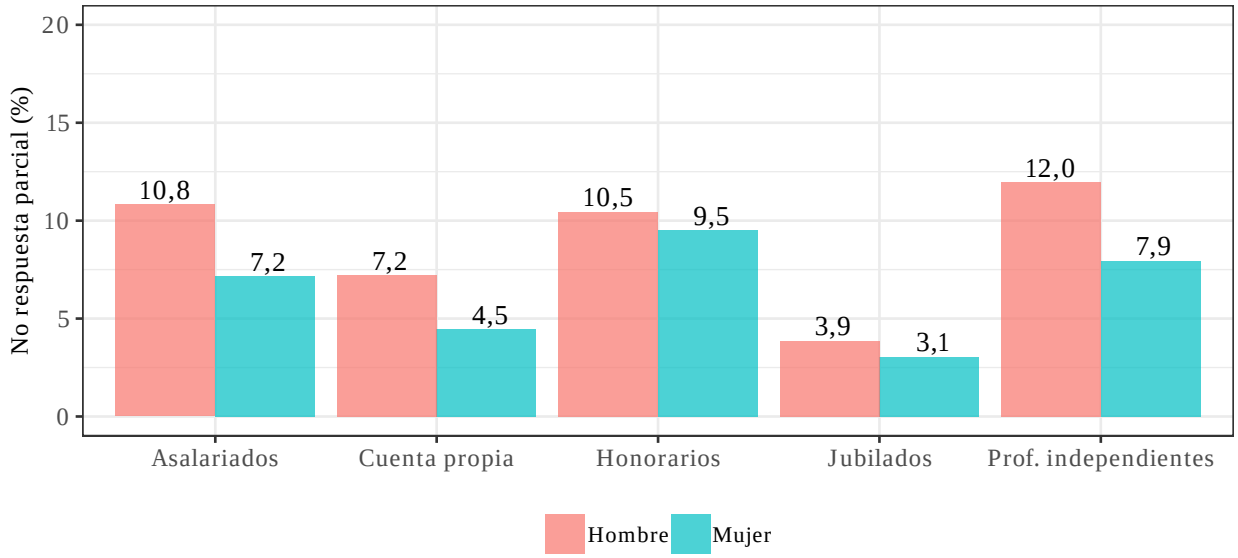
A pesar de que los grupos de trabajadores son distintos, no se observan diferencias importantes para la no respuesta en términos agregados. En los siguientes apartados se analizará si los distintos perfiles de trabajadores presentan también una propensión distinta a responder el módulo de ingresos de la encuesta.

4.2.2 Variables relacionadas con la no respuesta en ingresos

4.2.2.1 Estadística descriptiva

En las encuestas de hogares, la falta de respuesta, por lo general, se encuentra asociada a determinados rasgos sociodemográficos de la población. Así, por ejemplo, al desagregar la respuesta por sexo (figura 27), se observa que para la mayoría de las categorías de ingreso el rechazo es mayor por parte de los hombres. El grupo de trabajadores a honorarios es el único en donde esta tendencia se invierte.

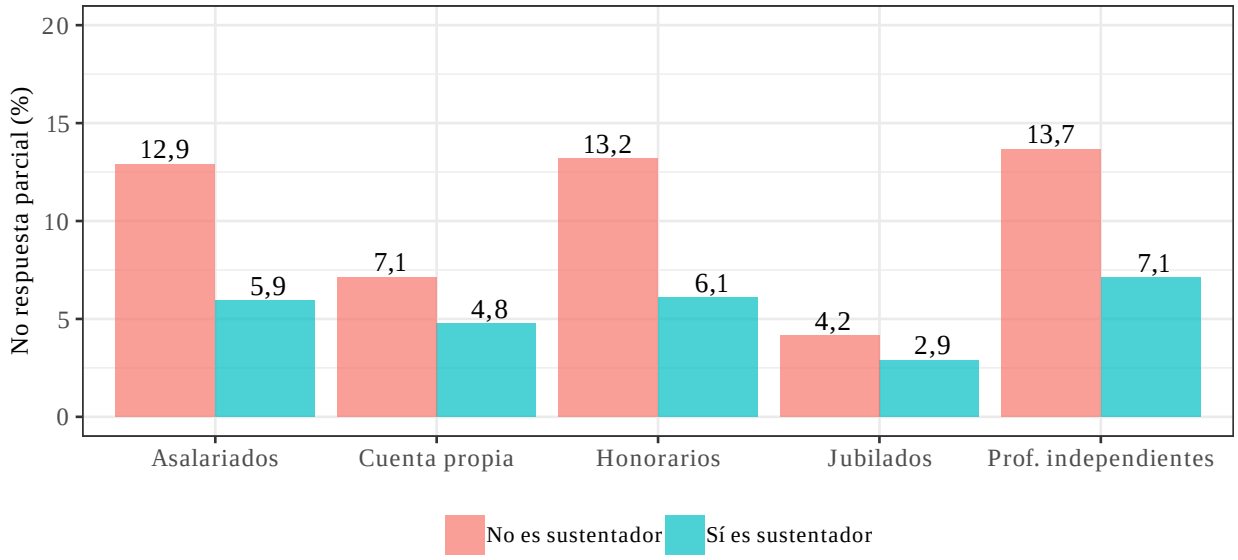
Figura 27: No respuesta parcial por categorías de ingresos y sexo



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

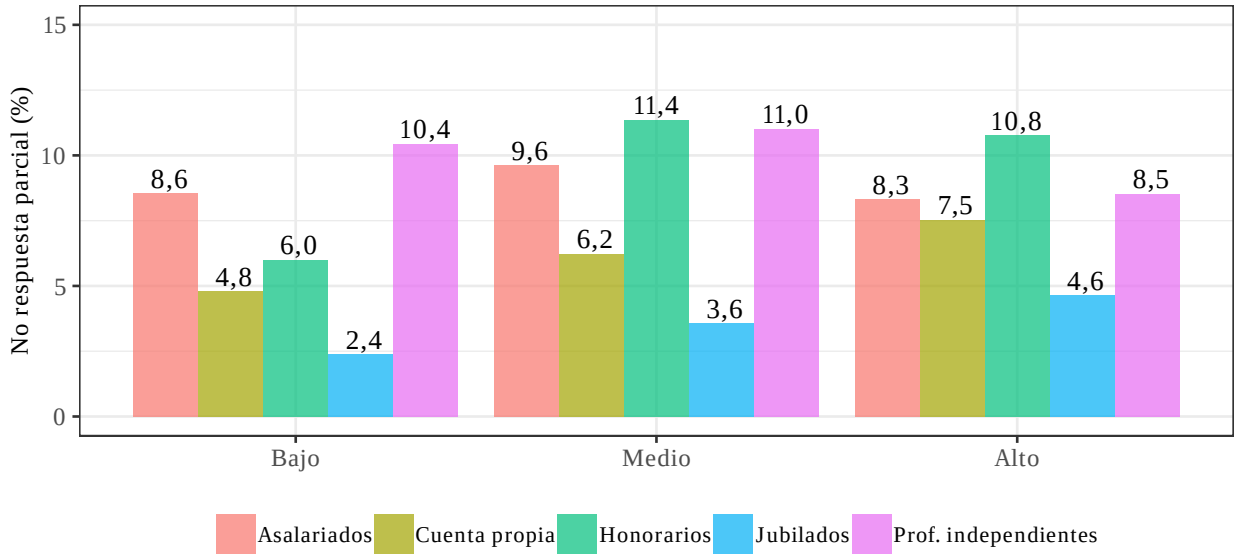
La encuesta consulta por la persona que más aporta al presupuesto del hogar, la que se denomina sustentadora principal. Para todos los grupos se observa que la falta de respuesta es menor para quienes aportan más al presupuesto de sus hogares (figura 28). En este sentido, quienes se identifican como sustentadores principales tienden a presentar un porcentaje menor de no respuesta.

Figura 28: No respuesta parcial por categoría de ingresos y condición de sustentador



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Figura 29: No respuesta parcial por categoría de ingresos y clasificación socioeconómica del marco



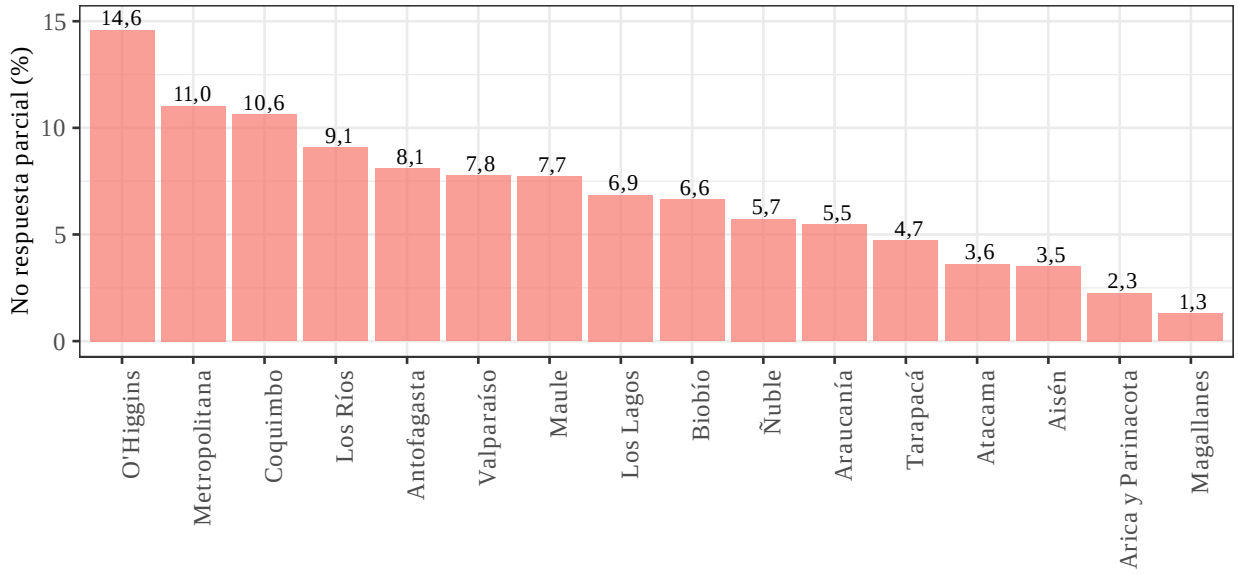
Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

De acuerdo a la clasificación socioeconómica del marco muestral[30], se observa que, en comparación al nivel más bajo, el grupo socioeconómico medio tiene un porcentaje de no respuesta superior para todas las categorías de ingresos. A la vez, si se compara el grupo alto con el bajo, se advierte una tendencia similar, con excepción de los trabajadores asalariados y el grupo de profesionales independientes, donde el porcentaje de no respuesta es mayor en el nivel socioeconómico bajo.

En términos geográficos, no se observa que las categorías de ingresos compartan una tendencia respecto a las regiones con mayores porcentajes de no respuesta. Así, por ejemplo, en el caso de los trabajadores asalariados, las regiones de O'Higgins y Metropolitana tienen el porcentaje de no respuesta más alto, mientras que Aysén, Arica y Parinacota y Magallanes presentan la menor tasa de no respuesta parcial. Si se analiza al grupo de trabajadores por cuenta propia, las regiones con mayor falta de respuesta son Maule, Los Lagos y Antofagasta.

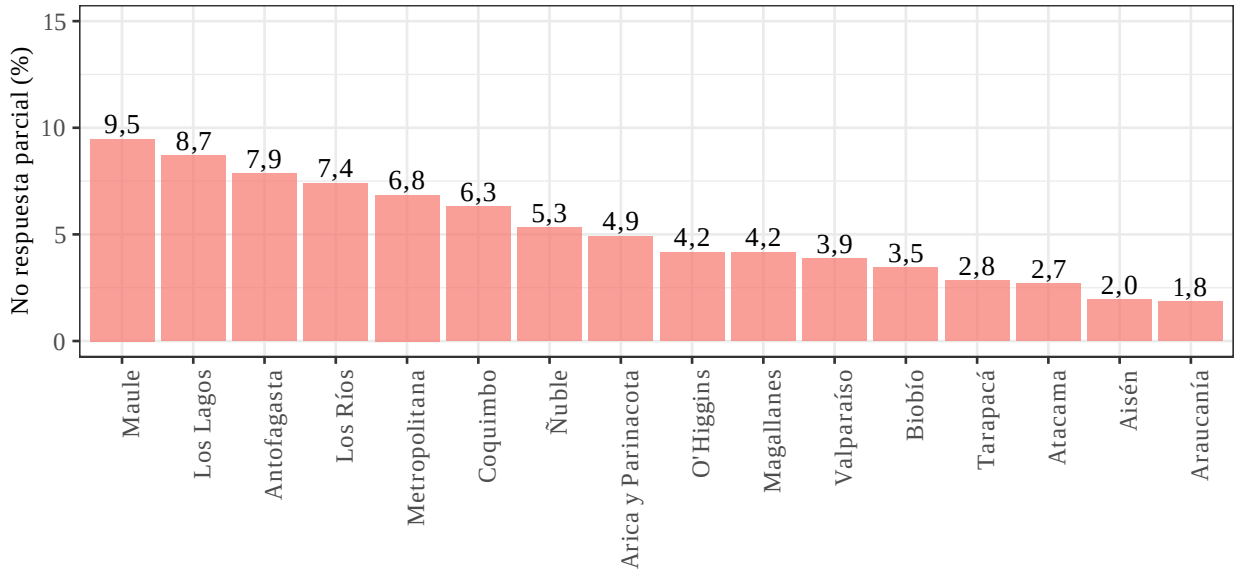
[30] La variable de clasificación socioeconómica (CSE) entrega una estratificación de los hogares realizada con los datos del CENSO 2002. Para mayor información sobre la construcción de esta variable consultar Guerrero (2003).

Figura 30: No respuesta parcial por categoría de ingresos y región - Asalariados



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Figura 31: No respuesta parcial por categoría de ingresos y región - Cuenta propia



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

4.2.2.2 Modelo de respuesta para ingresos del trabajo y de la jubilación

Para la mayoría de los métodos de imputación, las condiciones bajo las cuales las estimaciones resultan insesgadas son conocidas. Sin embargo, rara vez estas se presentan al momento de analizar la falta de respuesta. En la mayoría de los casos, la no respuesta se encuentra relacionada con características de la población que hacen que ciertos perfiles sean menos o más proclives a participar de una encuesta. En este contexto, es de suma relevancia identificar aquellas variables

que se encuentran asociadas a la propensión a responder, de manera de poder controlar por estas al momento de realizar la imputación de datos faltantes.

El análisis de estos patrones de no respuesta supone la elaboración de un modelo lineal que permita analizar estas relaciones. En particular, se asume la existencia de una variable latente R para el perceptor de ingresos j del hogar i como función de determinantes a nivel individuo y hogar. Esto se expresa como:

$$R_{i*j} = \alpha X_j + \beta Z_i + \epsilon_{ij}$$

En la práctica, la variable latente no es observada y el investigador solo observa si el perceptor de ingresos respondió o no, por lo que se define la variable binaria:

$$R_{ij} = \begin{cases} 1 & si \quad el \quad perceptor \quad responde \\ 0 & si \quad no \quad responde \end{cases}$$

Siguiendo este modelamiento, la estimación corresponde a una especificación *logit*, donde la propensión a responder es explicada por un conjunto de variables que caracterizan al perceptor de ingresos, resumidos en la matriz X_j, y otro conjunto de variables referidas al hogar, representadas en la matriz Z_i. X_j incluye la edad, la escolaridad, una variable dicotómica que toma valor 1 si el perceptor es hombre y 0 si es mujer, y otra variable dicotómica que toma valor 1 si el perceptor es sustentador principal y 0 si no. Z_j incluye el estrato socioeconómico del hogar (bajo, medio o alto), la variable macrozona (Norte, Centro, Sur y Metropolitana) y las distintas especificaciones incorporan variables sobre el informante idóneo de la libreta de ingresos, en particular, el sexo, la edad y si es sustentador principal.

Se probaron distintas especificaciones para cada una de las categorías de ingresos. Los modelos que aquí se presentan fueron elegidos utilizando distintos criterios estadísticos, en particular R cuadrado ajustado, los criterios de información de Akaike y Bayesiano y el factor de inflación de la varianza (con el objetivo de detectar la posible presencia de multicolinealidad entre las covariables).

La justificación para incluir un set de variables sobre el informante idóneo viene dada por el hecho de que la información de ingresos podría no ser recolectada directamente por el perceptor de estos[31]. De esta forma, existe una dificultad al momento de modelar la propensión a responder, puesto que en un número importante de casos la persona que responde no coincide con el perceptor directo de los ingresos consultados.

Para cada libreta solo se registra un informante idóneo por hogar, por lo que, en el caso que el encuestador haya podido consultar directamente al perceptor de ingresos y el informante idóneo registrado sea otro, esta información no queda reflejada en los instrumentos de recolección.

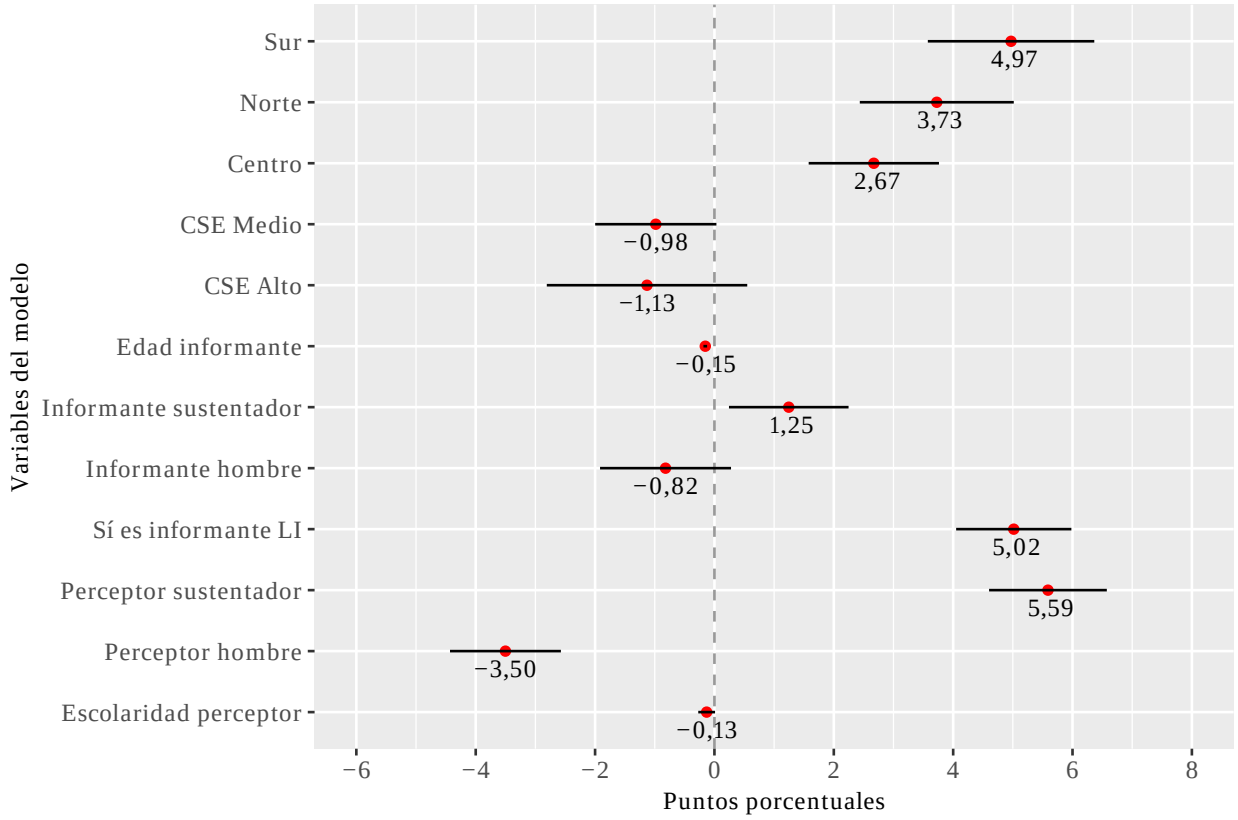
Como una medida estimada de esta coincidencia, el siguiente gráfico resume los porcentajes en que el informante idóneo corresponde a la misma persona que percibe por ingresos para cada grupo a imputar. Se recuerda que es una medida aproximada, puesto que, por el diseño del cuestionario, sólo es posible registrar un informante idóneo para cada libreta.

[31]Algo que sí ocurre en la Libreta de Gastos Individuales, donde es el informante quien directamente registra la información de sus gastos diarios.

Comenzando por los trabajadores asalariados, la figura 32 muestra los efectos marginales y los intervalos de confianza para cada una de las variables del modelo estimado. La línea achurada atraviesa al eje x en el 0, por lo que, si un intervalo de confianza cruza esta línea, la variable no tiene un efecto estadísticamente significativo al nivel de exigencia de 95%.

Se observa que residir en una zona distinta a la Región Metropolitana aumenta la probabilidad de responder el módulo de ingresos para los trabajadores asalariados. Así también, se observa que vivir en una manzana clasificada en un estrato socioeconómico medio disminuye la probabilidad de respuesta frente de quienes viven en una manzana de estrato bajo. En términos de las características de los perceptores de ingresos, ser informante directo de la libreta aumenta en 5 puntos porcentuales la probabilidad de responder el módulo. De la misma manera, si el perceptor se reconoce también como el sustentador del hogar también es más probable que este responda. Cabe destacar que las características del informante de la libreta también son relevantes en la explicación de la probabilidad de respuesta. Así, por ejemplo, si el informante es sustentador principal esto aumenta en 1,25 puntos porcentuales la probabilidad de contar con la información del perceptor de ingresos.

Figura 32: Modelo de respuesta. Regresión logística - Asalariados

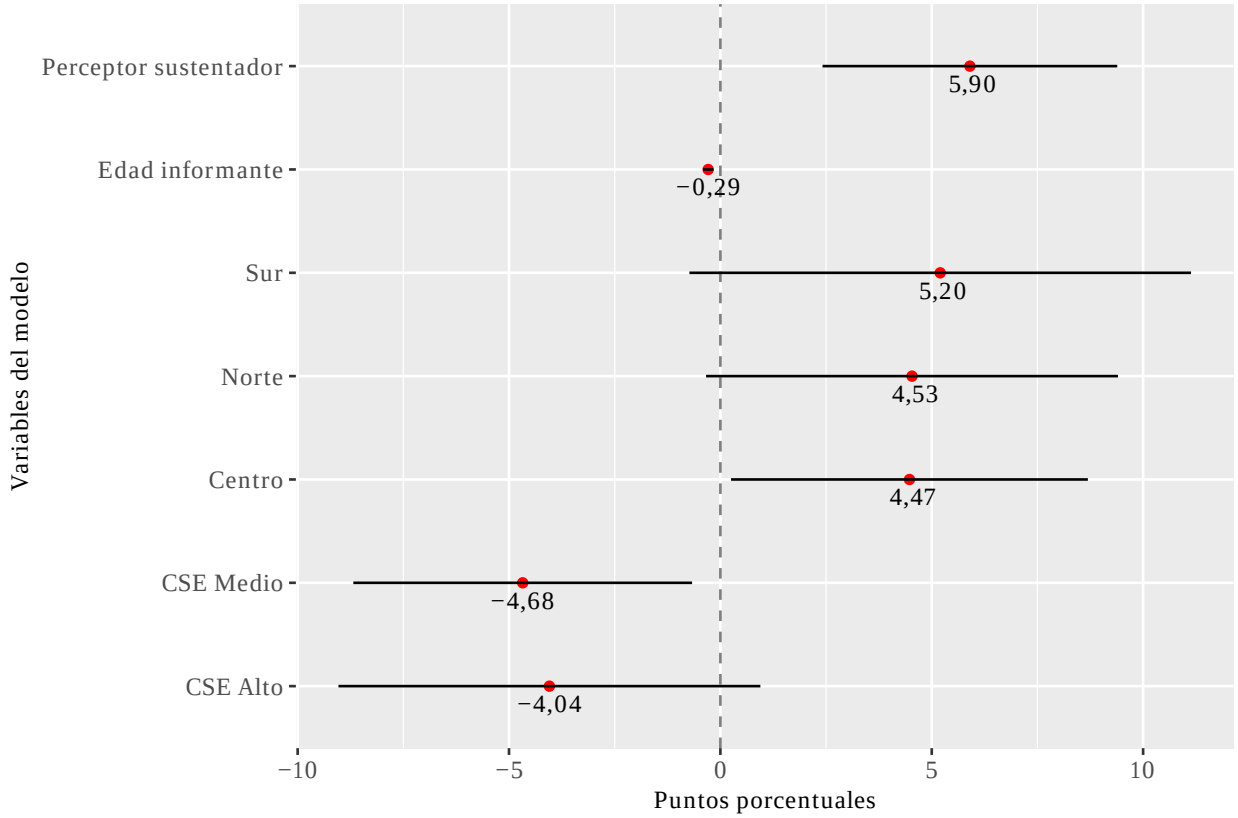


Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

En cuanto a los trabajadores a honorarios, la figura 33 grafica los efectos marginales para el modelo de no respuesta ajustado para este grupo de ocupados. A diferencia del modelo de trabajadores asalariados, el ajuste del modelo para este grupo es más modesto, tanto por la cantidad de

observaciones (1.036 versus 15.251 en el caso de los trabajadores asalariados) como por las variables disponibles para el ajuste del modelo. A diferencia del grupo anterior de trabajadores, las especificaciones que incluían las variables de sexo del informante o del perceptor no mejoran el ajuste del modelo de manera significativa, y la multicolinealidad entre las variables del informante idóneo y el perceptor llevaron a excluir algunas para la especificación elegida. No obstante, se observa que, al igual que en el caso de los trabajadores asalariados, los perceptores de ingresos a honorarios que son sustentadores principales tienen mayor probabilidad de responder el cuestionario de ingresos de la encuesta. Así también, los trabajadores a honorarios que residen en una manzana de estrato medio tienen una propensión a responder menor que aquellos de estrato bajo.

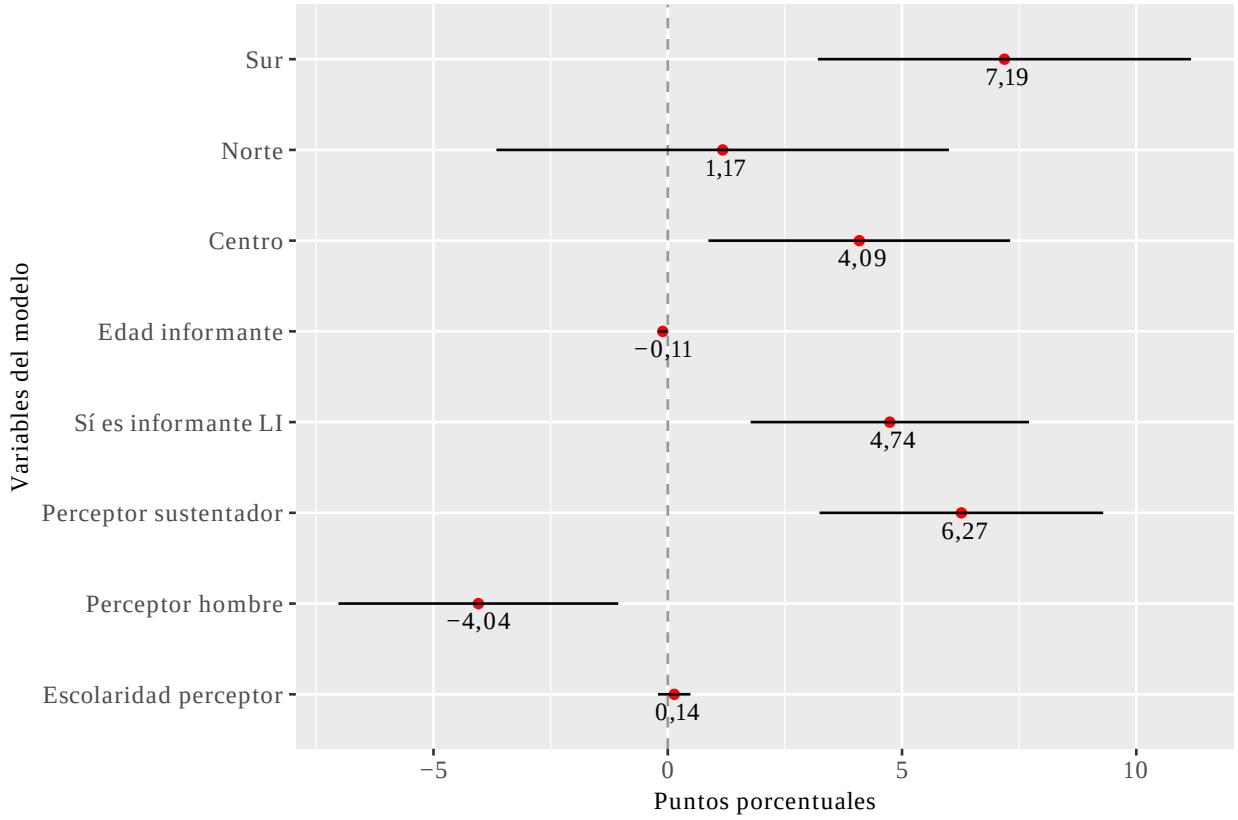
Figura 33: Modelo de respuesta. Regresión logística - Honorarios



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

Al revisar el modelo de respuesta para los profesionales independientes no se observan tendencias distintas a las que ya se han visto en los trabajadores asalariados y a honorarios. No obstante, una diferencia a destacar es que las personas que residen en la macrozona norte del país (regiones I, II, III, IV y XV) no presentan una diferencia significativa en su la propensión a responder respecto de aquellas que residen en la Región Metropolitana. Si se vuelve a la figura anterior (33), en el caso de los honorarios no se observaban diferencias significativas para las regiones del sur y el norte del país. Esto muestra que la no respuesta al ítem en el caso de ingresos no se distribuye de la misma manera en términos geográficos entre grupos de ingresos.

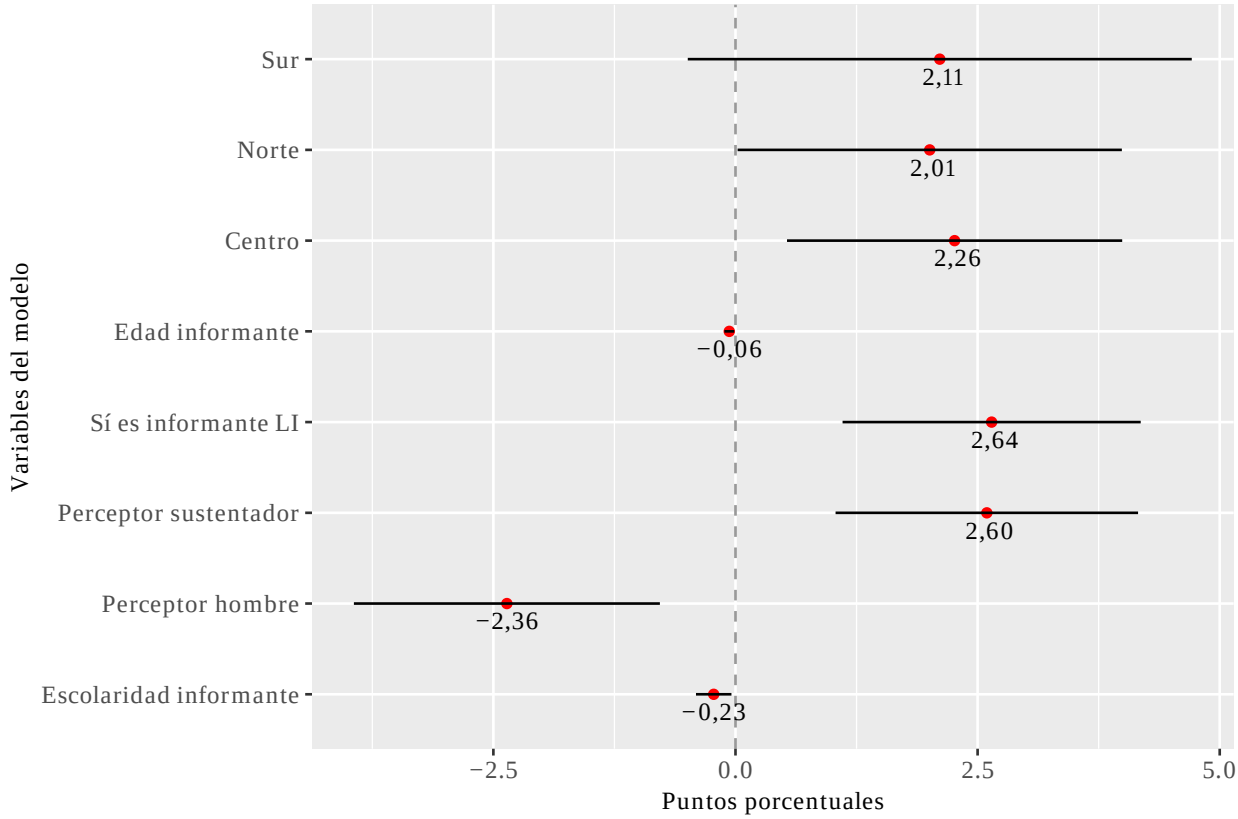
Figura 34: Modelo de respuesta. Regresión logística - Profesionales independientes



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

Para el último grupo de ocupados, los trabajadores por cuenta propia, el modelo también tiene un ajuste menor. Sin embargo, la tendencia es similar a la observada en los modelos anteriores. Ser informante de la libreta y ser identificado como el sustentador económico del hogar aumenta en aproximadamente 2 puntos porcentuales la probabilidad de responder el módulo de ingresos del trabajo para este grupo de ocupados. A su vez, ser hombre disminuye la probabilidad de respuesta de este grupo de trabajadores.

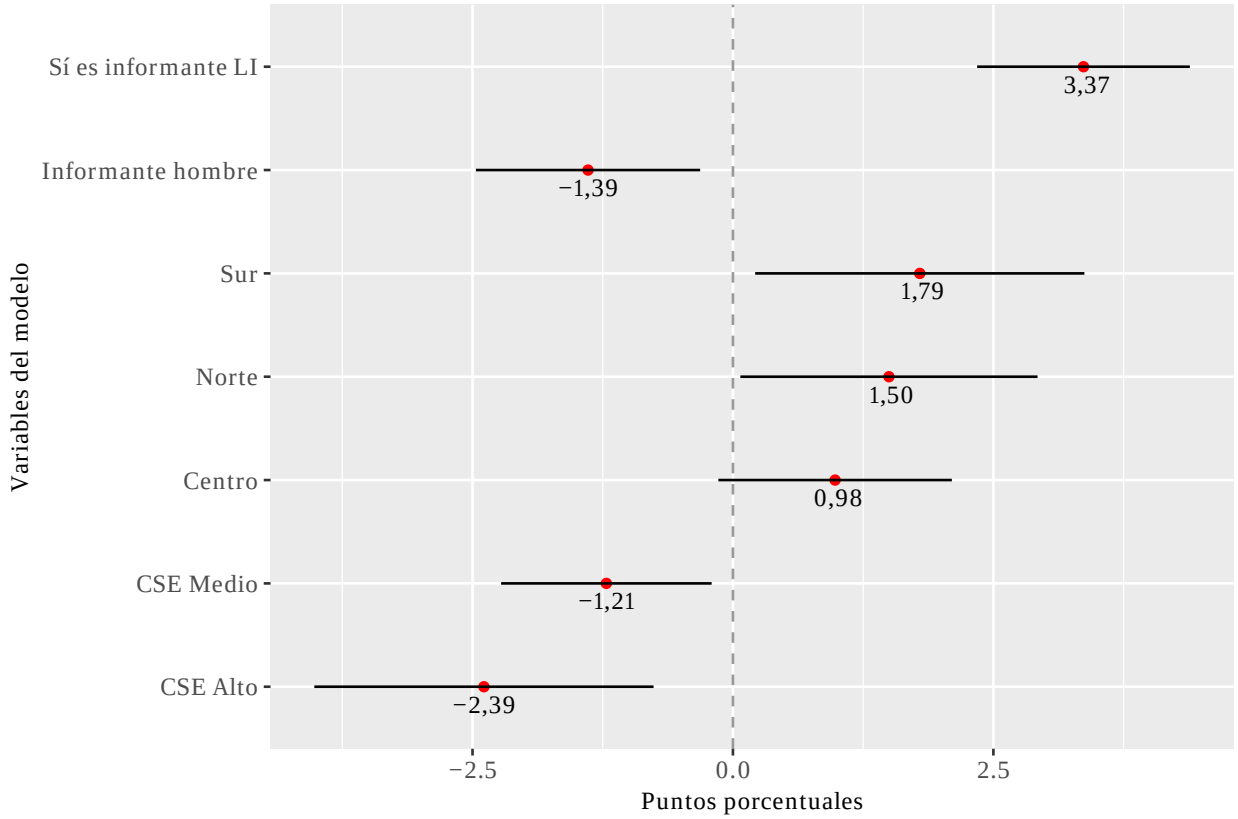
Figura 35: Modelo de respuesta. Regresión logística - Cuenta propia



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

Finalmente, el modelo de respuesta para el grupo de los jubilados es en el cual se consigue el ajuste más modesto, particularmente porque es el grupo que presenta la no respuesta más baja entre todos los grupos de ingreso analizadas. A pesar de esto, el modelo ajustado para los jubilados presenta tendencias muy similares a las que ya observadas en los grupos de ocupados. Los hombres tienen menor probabilidad de responder y quienes residen en la Región Metropolitana también tienen un perfil de respuesta de baja probabilidad. Una diferencia a destacar con el resto de los modelos es que, en el caso de los jubilados, vivir en una manzana clasificada como socioeconómicamente alta, disminuye las probabilidades de respuesta en comparación a los que residen en una manzana de estrato bajo. En los grupos de ocupados, cuando se presentaba una diferencia significativa para la variable de clasificación socioeconómica, solo era en la comparación entre estrato medio y bajo.

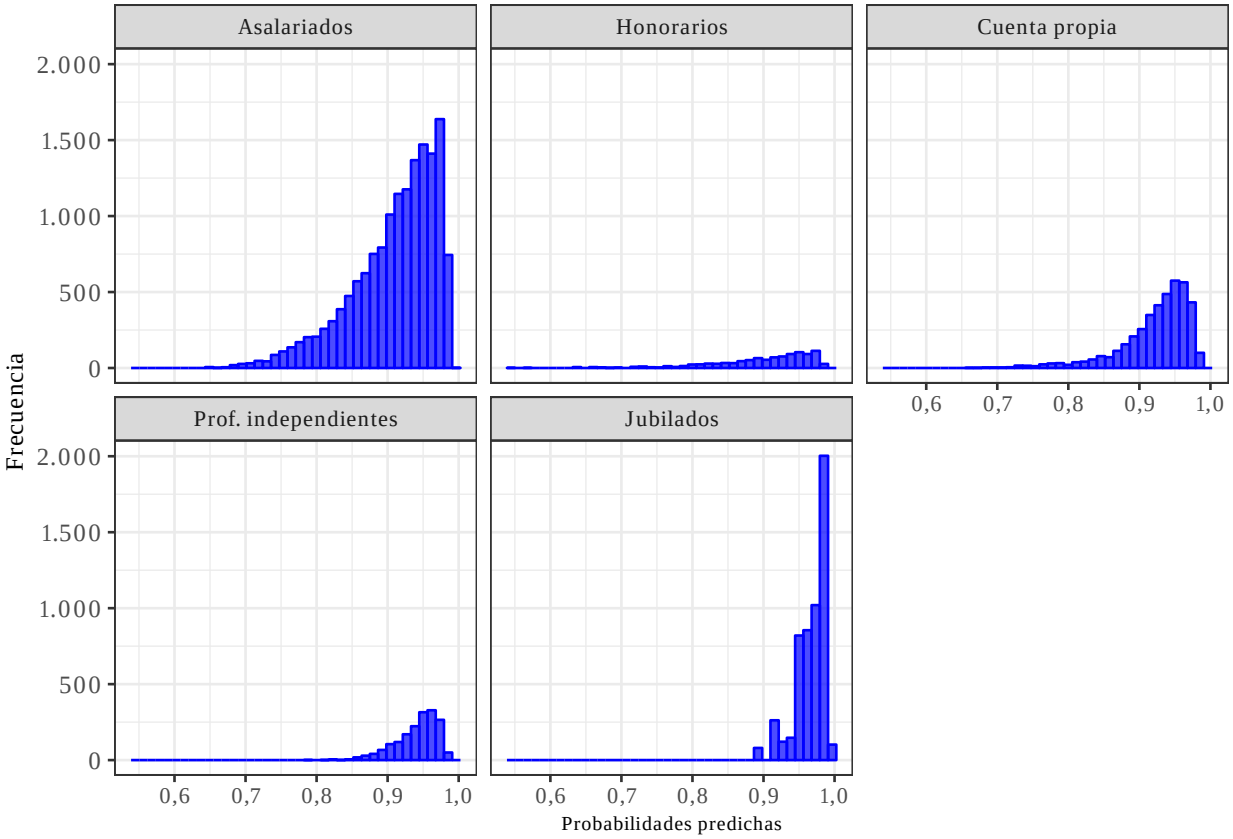
Figura 36: Modelo de respuesta. Regresión logística - Jubilados



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)
Nota: Los efectos marginales han sido multiplicados por 100 para facilitar su lectura

A partir de los modelos es posible generar predicciones para cada uno de las personas que perciben ingresos en el período de referencia de la encuesta (tanto aquellos que responden como quienes no). Se observa que la distribución de las probabilidades predichas (figura 37) se encuentra inclinada a la derecha para todas las categorías de ingresos, con lo que la mayor parte de las personas tienen una probabilidad alta de responder. La dispersión de esta probabilidad también es baja, fluctuando en el caso de los trabajadores asalariados (el grupo más grande de ocupados) entre 0,65 en el extremo inferior y 0,99 en el superior. Tal como fue mencionado en el análisis de la no respuesta para gastos diarios, el hecho de que la dispersión no sea tan elevada es algo deseable (Bethlehem et al., 2008).

Figura 37: Distribución de probabilidades predichas por modelo de respuesta



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Si se analiza la correlación de la no respuesta con el ingreso, se observa que para todos los casos esta es relativamente baja. Como ya se ha señalado, aun cuando la no respuesta presente niveles altos, si esta no se encuentra correlacionada con la variable objetivo no se produciría un sesgo en las estimaciones. La existencia de correlación, para el caso de ingresos, entrega más argumentos para señalar que es necesario utilizar un método de imputación para mitigar, en parte, el efecto de la falta de respuesta en ingresos.

Cuadro 11: Correlación entre ingresos y probabilidades de respuesta

Partida	Correlación
Asalariados	0,129
Honorarios	0,102
Prof. independientes	0,128
Cuenta propia	-0,0607
Jubilados	-0,035

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Como conclusión preliminar de estos resultados, es posible señalar que la no respuesta en ingresos no es un fenómeno que se distribuye de manera aleatoria en la población. Más bien, esta depende de características relacionadas tanto con los perceptores, los informantes de la libreta, así como también

con características del hogar. Entre las variables más relevantes y que se repiten entre los distintos grupos de ingreso se encuentran:

- Sexo
- Sustentador del hogar
- Ser informante directo de sus ingresos

Mientras que las dos primeras pueden ser consideradas en los métodos de imputación, no resulta congruente incluir la condición de ser informante de la libreta de ingresos como parte de los algoritmos para predecir los valores faltantes de ingresos. No obstante, esta información puede ser de suma relevancia para los futuros procesos de recolección del estudio, especialmente en lo relativo a esta libreta.

4.2.3 Variables relacionadas con el ingreso

4.2.3.1 Selección de variables a estudiar

Para realizar la imputación de ingresos, con cualquier método que quiera utilizarse, es necesario contar con variables que se relacionen y expliquen dicho ingreso. La VIII EPF cuenta con una gran cantidad de variables posibles, por lo tanto, se debe definir cuáles serán seleccionadas. Para esto, se estudia la correlación de estas variables con el ingreso y su significancia a la hora de ser utilizadas como regresores para su predicción.

Como se mencionó, la selección de variables a estudiar para la imputación de ingreso se realiza utilizando dos criterios. Por un lado, la correlación y la significancia[32] de la variable como predictor del ingreso, a las cuales se les denomina **variables empíricas**. Por otro lado, la revisión bibliográfica sobre variables predictoras del ingreso que han sido investigadas, cuya importancia ha sido demostrada para este tipo de procedimientos. A este último conjunto de variables se les denomina **variables teóricas**.

4.2.3.1.1 Variables teóricas

Se han incluido variables recogidas de diferentes estudios, que han demostrado su importancia para predecir el ingreso de las personas. En el caso de los ingresos laborales, se incorporan las variables de edad (como *proxy* de experiencia), edad al cuadrado y escolaridad, basado en la teoría del capital humano, la teoría del ciclo de vida (Mincer, 1974) y más recientemente de la señalización (Spence, 1973; Weiss, 1995). La primera teoría muestra que un año más de educación tiene un impacto positivo en el salario, así como también la experiencia genera un aumento en el ingreso. Se incorpora la variable edad al cuadrado, ya que dicha experiencia tiene retornos marginales decrecientes. Por su parte, la teoría de la señalización indica que el mercado laboral responde a señales de características no observables de los trabajadores (productividad, absentismo laboral, entre otras) y una de las señales entregadas son los grados educacionales, por lo que, al alcanzar un grado académico mayor, aumenta el ingreso laboral. La variable de sexo se incorpora para

[32] Se seleccionaron variables con *p-value* menor a 0,05

visibilizar la importante diferencia salarial entre hombres y mujeres. Finalmente, se incluye la clasificación socioeconómica[33] y variables geográficas que tienen un componente teórico, ya que, en el diseño muestral la probabilidad de selección está condicionada por estas variables (Andridge & Little, 2009) y, por otra parte, la distribución de ingresos muestra una correlación con estas, por lo que es relevante su inclusión como control.

Cuadro 12: Variables teóricas seleccionadas para el proceso de imputación de ingresos

Variables teóricas
Edad
Edad^2
Escolaridad
Sexo
Clasificación socioeconómica
Variables geográficas

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

4.2.3.1.2 Variables empíricas

La VIII EPF cuenta con diferentes variables que caracterizan al perceptor de ingresos, siendo el cuestionario de Registro de Personas del Hogar (RPH) el que recopila dichas características. Entre estas se encuentran la edad, el estado civil, el sexo, el nivel educativo y la condición de actividad económica, entre otras.

Además, se consideran variables generadas en el procesamiento de los datos, como el código de ocupaciones de la Clasificación Internacional Uniforme de Ocupaciones (CIUO) y la presencia de menores de 6 o 15 años en el hogar[34].

Para seleccionar las variables empíricas a utilizar, se realizan regresiones lineales y, mediante estas, se escoge el modelo que mejor explica el ingreso, considerando la significancia de las variables. Las pruebas incluyen diferentes combinaciones de variables y se opta por aquel modelo que, considerando las variables definidas previamente como teóricas, incorpore otras características disponibles en la base de datos (variables empíricas) que sean estadísticamente significativas como variables explicativas.

A continuación se presenta la regresión general utilizada para la selección de variables.

$$Ing.Laboral = \alpha X + \beta Z + \epsilon$$

[33] La variable de clasificación socioeconómica (CSE) entrega una estratificación de los hogares realizada con los datos del CENSO 2002. Para su construcción se utiliza el método PRINCALS y se incluye variables socioeconómicas, de educación y ocupación del jefe de hogar, entre otras. Para mayor información sobre la construcción de esta variable consultar Guerrero (2003).

[34] para la construcción de estas variables se toma como referencia el documento "Ciclo vital de la familia y género" realizado por el departamento de estudios del Ministerio de Desarrollo Social. Para mayor información revisar (Jiménez, Ramírez, & Pizarro, 2008).

Dónde:

X= variables empíricas que explican el ingreso del trabajo

Z= variables teóricas que explican el ingreso del trabajo

Luego de realizar las diferentes pruebas que incluían distintos grupos de variables, fueron seleccionadas las siguientes:

Cuadro 13: Variables empíricas seleccionadas para el proceso de imputación de ingresos

Variables empíricas
Clasificación Internacional Uniforme de Ocupaciones (CIUO)
Clasificación Internacional de Situación de Empleo (CISE)
Sustentador principal
Presencia de menores de 6 o 15 años en el hogar
Situación previsional

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

4.2.3.2 Correlaciones

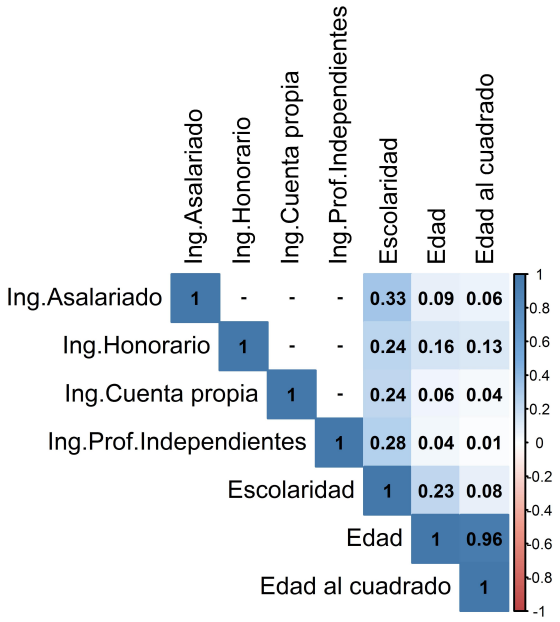
El estudio de las correlaciones entre el ingreso y las variables seleccionadas se utiliza para establecer la magnitud de la relación entre dichas variables y, así, generar un orden e identificar cuáles pueden eliminarse o agruparse en categorías según su grado de correlación. En particular, este procedimiento es relevante cuando se prueban métodos de imputación que impliquen generar *clusters* y niveles.

Al igual que en el análisis para la imputación de gastos, dada la naturaleza de las variables, se aplican coeficientes de correlación diferentes para las variables continuas y para las variables categóricas. En el caso de las variables continuas se usa el coeficiente de correlación lineal de Pearson y para las variables categóricas se construyen variables dicotómicas que representan cada una de las categorías, para luego utilizar el coeficiente de punto biserial.

A continuación, se muestran ejemplos de la correlación entre los ingresos a imputar[35] y las variables edad, edad al cuadrado y escolaridad.

[35] Como ya se ha señalado, los ingresos a imputar son: ingresos del trabajo asalariado, ingresos del trabajo a honorarios, ingresos del trabajo por cuenta propia, ingresos de profesionales independientes e ingresos de jubilaciones.

Figura 38: Correlación entre ingresos del trabajo y escolaridad, edad y edad al cuadrado. Correlación de Pearson



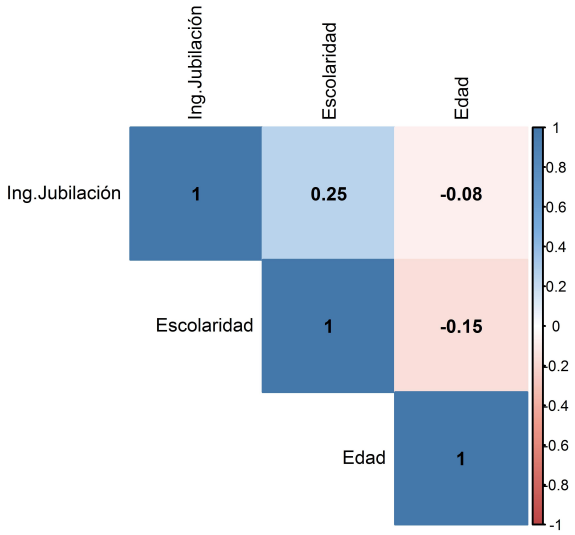
Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

En la figura 38 se observa una correlación positiva entre las tres variables presentadas y las categorías de ingreso. La variable años de escolaridad es aquella que presenta la relación de mayor intensidad y los signos de los coeficientes son los esperados. A mayor cantidad de años de educación, el pago del mercado es mayor. La edad, por su parte, funciona como *proxy* de la experiencia, por lo que también presenta un coeficiente positivo. Como es de esperarse, la variable edad al cuadrado muestra un coeficiente positivo, pero menor al de la edad, ya que esta variable da cuenta de los rendimientos marginales decrecientes de la experiencia.

Para todos los grupos de ingreso, la escolaridad es la variable que presenta el coeficiente de correlación más alto, llegando a su máximo nivel en el caso de los asalariados. La edad muestra valores pequeños, pero para el grupo de los trabajadores a honorarios se observa una mayor correlación que en el resto de los casos.

La similitud de resultados apoya el uso de estas variables en todos los tipos de ingresos, tal como es indicado en los estudios revisados.

Figura 39: Correlación entre ingreso de jubilaciones, escolaridad y edad. Correlación de Pearson



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

En el caso del ingreso por jubilaciones se utilizan las variables de escolaridad y edad, ya que son las variables continuas que se tienen a disposición para caracterizar a quienes reciban jubilación.

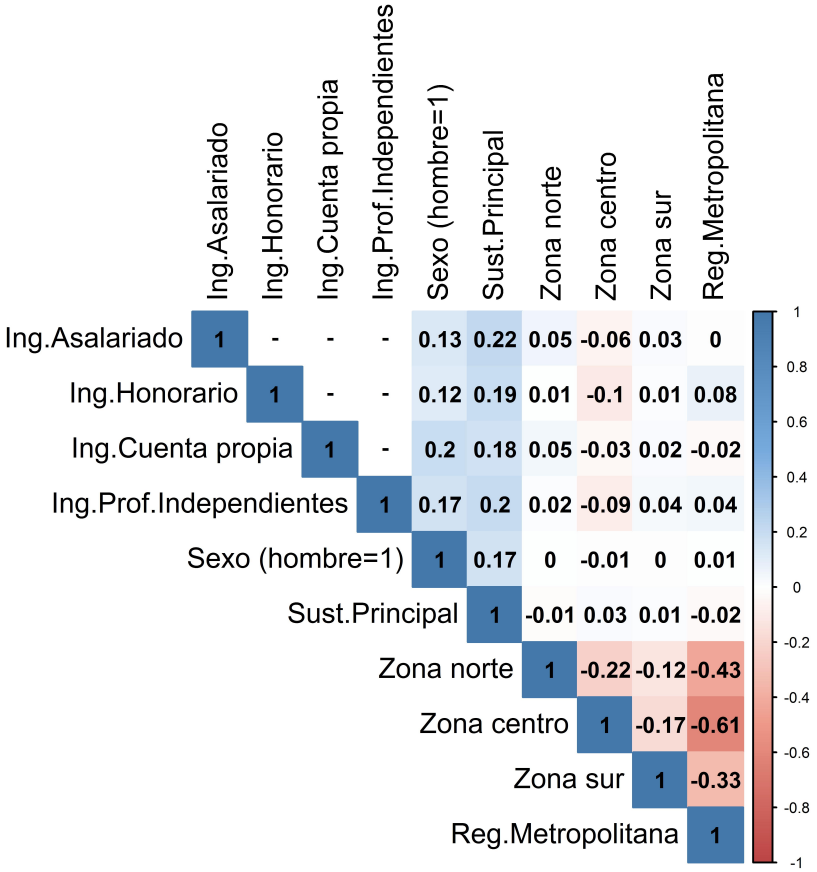
La escolaridad presenta la relación más fuerte con un coeficiente de 0,25, lo que indica una relación positiva entre los años de escolaridad y el monto del ingreso por jubilaciones. El signo del coeficiente es esperable, ya que educación correlaciona con ingreso y, por lo tanto, posiblemente también con mayor monto de cotización. La edad tiene un coeficiente negativo y de menor valor.

Es importante recalcar que es necesario contar con esta información, ya que será utilizada en los modelos de imputación. La correlación será utilizada para ordenar las variables, por lo que debe ser analizada tanto para el grupo de variables teóricas, como para el grupo de variables empíricas.

A continuación, se revisa la correlación de los ingresos con las variables categóricas. Cabe señalar que, a diferencia de las variables continuas, los resultados ahora se muestran diferenciando entre ingresos del trabajo e ingresos de la jubilación. El motivo de esto es que las variables utilizadas son diferentes en cada uno de los casos.

Ingresos del trabajo

Figura 40: Correlación entre ingresos del trabajo y sexo, sustentador principal y macrozona. Correlación biserial



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Las variables sexo y sustentador principal presentan una correlación positiva, por lo que ambas están relacionadas con un mayor ingreso del trabajo. Es importante notar que la variable sexo es aquella con mayor valor en el caso de los ingresos por cuenta propia y la segunda de mayor valor para el resto de los ingresos.

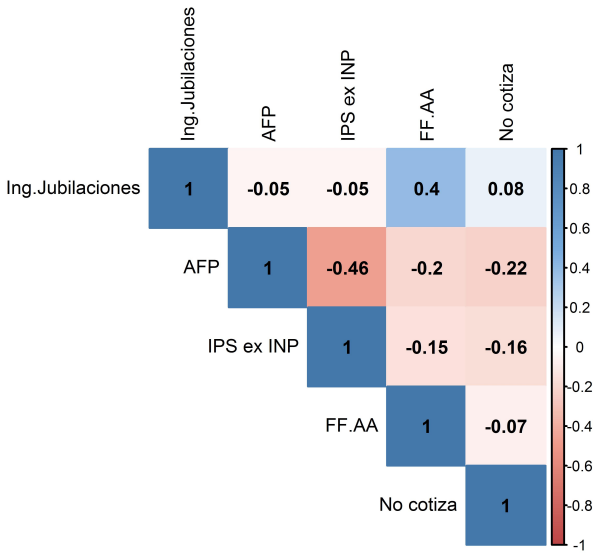
Las variables de macrozona presentan una correlación baja y que difiere en signo según la zona geográfica, siendo positiva para zona sur y zona norte, y negativa para la zona centro. En el caso de la Región Metropolitana el valor del coeficiente es cercano a cero, y cero en el caso de los asalariados. Respecto al signo, este es positivo, a excepción de los trabajadores por cuenta propia.

Además de las variables presentadas en la figura 40, se analizó la correlación de los ingresos del trabajo con otras variables, como la clasificación socioeconómica (CSE), la clasificación de empleo (CISE), la clasificación de ocupaciones (CIUO), entre otras. Esta información se encuentra en el anexo incorporado al final del documento.

Jubilados

Para los ingresos por jubilaciones y pensiones de vejez, el ser hombre (0,27) y sustentador principal (0,22) muestran coeficientes positivos. Por su parte, la correlación con la clasificación socioeconómica es la esperada, ya que conforme se avanza en los estratos, el valor de la correlación aumenta. En el caso de la clasificación socioeconómica baja, el valor es -0,18, mientras que para el grupo clasificado como medio es -0,1 y finalmente para los residentes en manzanas clasificadas como alta es 0,36.

Figura 41: correlación entre ingreso de jubilaciones y variables de sistema previsional. Correlación biserial



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Los valores de la correlación con la variable indicadora del sistema previsional en el que se cotiza muestran resultados llamativos. El hecho de pertenecer al sistema de AFP o al IPS (ex INP) se correlaciona de manera negativa con el ingreso de jubilación. Por el contrario, el haber cotizado en algún sistema de las fuerzas armadas o simplemente no haber cotizado, se correlaciona de forma positiva con en el ingreso. Vale la pena subrayar que para quienes cotizaron en sistemas de las fuerzas armadas, la correlación se hace más fuerte, con un coeficiente de 0,4. Cabe recordar que dicho ingreso incluye también las pensiones básicas solidarias y los aportes previsionales solidarios.

4.2.3.3 Modelos de ingreso

Como se mencionó en el apartado anterior, para seleccionar las **variables empíricas** a utilizar, se realizaron regresiones lineales y mediante estas se busca el modelo que explicase de mejor manera el ingreso, considerando la significancia de las variables. Se realizaron pruebas con diferentes combinaciones de variables y se optó por aquel modelo que, considerando las variables definidas

previamente como teóricas, incorporara otras características disponibles en la base de datos y que fuesen estadísticamente significativas como variables explicativas.

En el cuadro 14 se presentan los modelos para cada categoría de ingreso. Es necesario precisar que incorporan de manera forzosa las **variables teóricas** mencionadas con anterioridad, ya que su inclusión en los modelos de imputación no dependerá del nivel de correlación ni de la significancia que presente como variable explicativa, pero es necesario visualizar la interacción que dichas variables tienen con otras que sean candidatas a ser incluidas.

Como se observa, para todos los modelos de ingreso que se presentan, la gran mayoría de las variables muestra coeficientes estadísticamente significativos. Aquellas variables que no cuentan con un p-valor menor a 0,05 son la clasificación socioeconómica en el caso del ingreso de honorarios y macrozona en el caso de cuenta propia y jubilados. Sin embargo, estas variables pertenecen a la categoría de **variables teóricas**, por lo que su inclusión se relaciona con los estudios mencionados en el comienzo del apartado.

Es posible encontrar tendencias en los efectos de ciertas variables que son compartidas por todos los modelos. La escolaridad, por ejemplo, presenta un coeficiente positivo en todos los casos, lo que va en línea con lo indicado por los estudios, es decir, a mayor escolaridad, mayor ingreso laboral. La edad, por otra parte, presenta un efecto positivo en todos los casos y los coeficientes negativos de la edad al cuadrado van en línea con el impacto de esta variable de suavizar el efecto de la edad como *proxy* de la experiencia.

En todos los casos, ser hombre implica un aumento del ingreso, sobresaliendo el coeficiente del ingreso por cuenta propia con un valor de 0,55. En el caso de la macrozona se encuentra que pertenecer a la zona centro o zona sur impacta negativamente el ingreso laboral, no así el ingreso por jubilaciones.

La variable CIUO entrega diferentes resultados, pero se observa claramente que desde el CIUO 21 en adelante los coeficientes son negativos para todos los casos y todos los ingresos, a excepción de CIUO 22 (Profesionales de las ciencias biológicas, la medicina y la salud) para los trabajadores por cuenta propia. Estos resultados conversan con la construcción del clasificador, ya que el orden de los códigos se hace respecto a las competencias y estudios necesarios para desempeñar la ocupación, y por lo tanto, se espera que a medida que aumenta el valor del código, el ingreso disminuya.

Ser sustentador principal del hogar implica un mayor ingreso tanto para los ingresos laborales como para los jubilados. Es necesario tener en cuenta que para la VIII EPF el sustentador principal se define como quien aporta la mayor parte del presupuesto familiar, y no necesariamente quien tiene mayor ingreso.

Por último, pertenecer a un sistema previsional distinto a la AFP (IPS, FF.AA. u otro), en el caso de los jubilados, afecta positivamente al ingreso. Cabe destacar que la pertenencia al sistema de Fuerzas Armadas es el que implica una mayor jubilación.

Cuadro 14: Modelos de ingresos para asalariados, honorarios, cuenta propia, profesionales independientes y jubilados

	Variable dependiente				
	Log.Natural del ingreso laboral				
	Asalariados	Honorarios	C.Propia	P.Independientes	Jubilados
	(1)	(2)	(3)	(4)	(5)
CSE medio	0,06*** (0,01)	−0,03 (0,05)	0,14*** (0,03)	0,10* (0,06)	0,13*** (0,02)
CSE alto	0,42*** (0,02)	0,11 (0,08)	0,58*** (0,07)	0,55*** (0,09)	0,65*** (0,04)
Escolaridad	0,02*** (0,00)	0,02** (0,01)	0,02*** (0,01)	0,02* (0,01)	0,01*** (0,00)
Edad	0,07*** (0,00)	0,09*** (0,01)	0,09*** (0,01)	0,10*** (0,01)	0,00*** (0,00)
Edad al cuadrado	−0,00*** (0,00)	−0,00*** (0,00)	−0,00*** (0,00)	−0,00*** (0,00)	
Sexo	0,25*** (0,01)	0,14*** (0,05)	0,55*** (0,04)	0,44*** (0,07)	0,31*** (0,02)
Norte	0,04** (0,02)	−0,10* (0,06)	−0,06 (0,04)	0,11 (0,08)	0,03 (0,03)
Centro	−0,18*** (0,01)	−0,34*** (0,06)	−0,15*** (0,04)	−0,27*** (0,06)	0,05** (0,02)
Sur	−0,05** (0,02)	−0,06 (0,07)	−0,13** (0,06)	−0,12 (0,09)	0,09*** (0,03)
CIUO 12	0,13 (0,16)	−0,05 (0,22)			
CIUO 13	0,00 (0,17)	−0,82 (0,54)	−0,33** (0,13)	−0,59** (0,29)	
CIUO 21	−0,24 (0,16)	−0,60*** (0,13)	−0,38* (0,20)	−0,35 (0,24)	
CIUO 22	−0,32** (0,16)	−0,45*** (0,15)	0,44 (0,30)	−0,12 (0,25)	
CIUO 23	−0,72*** (0,16)	−1,11*** (0,12)	−0,90*** (0,32)	−0,92*** (0,27)	
CIUO 24	−0,39** (0,16)	−0,67*** (0,12)	−0,12 (0,18)	−0,52** (0,23)	
CIUO 31	−0,73*** (0,16)	−1,13*** (0,18)	−1,29*** (0,28)	−0,93*** (0,27)	
CIUO 32	−0,95*** (0,16)	−1,11*** (0,16)	−1,57*** (0,31)	−1,01*** (0,26)	
CIUO 33	−1,19*** (0,16)	−1,60*** (0,17)	−2,46*** (0,38)	−1,81*** (0,28)	
CIUO 34	−0,78*** (0,16)	−1,03*** (0,14)	−0,36*** (0,14)	−0,68*** (0,23)	
CIUO 41	−1,08*** (0,16)	−1,22*** (0,15)	−0,57** (0,26)	−1,51*** (0,30)	
CIUO 42	−1,20*** (0,16)	−1,36*** (0,17)	−0,87** (0,41)	−1,15* (0,62)	
CIUO 51	−1,44*** (0,16)	−1,44*** (0,15)	−1,10*** (0,15)	−1,59*** (0,25)	
CIUO 52	−1,38*** (0,16)	−1,26*** (0,16)	−0,74*** (0,13)	−1,18*** (0,25)	
CIUO 61	−1,54*** (0,17)	−1,46*** (0,23)	−1,19*** (0,16)	−1,35*** (0,31)	
CIUO 71	−1,19*** (0,16)	−1,07*** (0,17)	−1,03*** (0,14)	−1,17*** (0,24)	
CIUO 72	−1,10*** (0,16)	−1,01*** (0,19)	−0,95*** (0,15)	−1,03*** (0,24)	
CIUO 73	−1,51*** (0,18)	−1,25*** (0,15)	−1,44*** (0,17)	−1,87*** (0,32)	
CIUO 74	−1,52*** (0,16)	−1,24*** (0,27)	−1,48*** (0,14)	−1,84*** (0,28)	
CIUO 81	−0,77*** (0,17)	−0,20* (0,12)		−1,15*** (0,21)	
CIUO 82	−1,25*** (0,16)	−0,86*** (0,14)	−0,87*** (0,24)	−0,87 (0,58)	
CIUO83	−1,21*** (0,16)	−1,15*** (0,18)	−0,72*** (0,14)	−1,21*** (0,25)	
CIUO 91	−1,52*** (0,16)	−1,68*** (0,16)	−1,41*** (0,14)	−1,77*** (0,25)	
CIUO 92	−1,59*** (0,18)	−1,46*** (0,25)	−1,44*** (0,27)	−0,85*** (0,28)	
CIUO 93	−1,46*** (0,16)	−1,56*** (0,20)	−1,50*** (0,16)	−1,69*** (0,26)	
Sustentador principal	0,24*** (0,01)	0,22*** (0,05)	0,36*** (0,03)	0,48*** (0,05)	0,22*** (0,02)
CISE: asalariado sec.Pub.	0,23*** (0,01)				
CISE:serv.D puertas adentro	−0,09 (0,09)				
CISE:serv.D puertas afuera	−0,44*** (0,03)				
Presencia de menores de 6 años en el hogar	0,04*** (0,01)				
CISE: cuenta propia			−0,65*** (0,06)	−0,48*** (0,11)	
Presencia de menores de 15 años en el hogar			0,10*** (0,03)	0,12** (0,05)	
Sist.Previsional IPS					0,13*** (0,02)
Sist.Previsional FF.AA					0,94*** (0,04)
Otro sist.Previsional					0,25*** (0,04)
Sin sist.Previsional					−0,12*** (0,02)
Constante	12,31*** (0,17)	11,85*** (0,30)	11,31*** (0,23)	11,04*** (0,38)	11,26*** (0,09)

Nota *p<0,1; **p<0,05; ***p<0,01

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

4.3 Descripción de los métodos

4.3.1 Experiencia de la VII EPF

Durante la VII EPF, se realizó un estudio sobre el método a utilizar para la imputación de ingresos, lo cual quedó plasmado en el documento de trabajo "**Métodos de imputación VII EPF: Gastos diarios e ingresos de la actividad principal y jubilaciones**" (INE Chile, 2014).

La no respuesta en ingresos para la VII EPF se presenta en el cuadro 15.

Cuadro 15: No respuesta parcial por categoría de ingreso

Ingresos	NR
Asalariados	8,56
Honorarios	15,71
Cuenta propia	7,74
Profesionales independientes	7,25
Jubilados	4,32

Fuente: Instituto Nacional de Estadísticas (INE) - VII Encuesta de Presupuestos Familiares (EPF)

Para enfrentar el problema de no respuesta en ingresos, se estudiaron cuatro métodos de imputación de datos:

- *Hot deck* media (media condicionada)
- Heckman
- Imputación múltiple con regresión
- Máxima verosimilitud con EM

Todos estos métodos serán descritos en el siguiente apartado del documento, a excepción del método de máxima verosimilitud, ya que en el caso de la VIII EPF no ha sido incluido en las pruebas.

Para la prueba de métodos se seleccionaron variables correlacionadas con el ingreso y también aquellas que en la teoría indicaban relación. Finalmente, se utilizaron las variables de sexo, estrato socioeconómico, desagregación geográfica, edad y ocupación.

Luego de imputar, se compararon distribuciones de datos observados e imputados, cantidad de datos imputados y estadísticos descriptivos de las variables imputadas.

Considerando la distribución de los datos, los estadísticos descriptivos (en particular aquellos de dispersión) y la facilidad de aplicación y explicación del método, se optó por imputar a través de media condicionada.

4.3.2 Métodos probados en la VIII EPF

En el caso de la imputación de ingresos, se decidió estudiar cuatro métodos de imputación. A diferencia de las pruebas realizadas en el contexto de la VII EPF, no se utiliza el método de máxima verosimilitud y se incluye el método *hot deck* aleatorio. La selección de los métodos a probar

responde a estrategias ya utilizadas en el Instituto Nacional de Estadísticas de Chile y en otras instituciones, junto con ser indicados en la literatura como los más comúnmente aplicados.

Como se ha mencionado, la VIII EPF imputa los ingresos asociados al trabajo en la ocupación principal y jubilaciones, ya que dichos ingresos representan el porcentaje mayoritario del ingreso del hogar. En todos los casos se generan 5 bases de datos que solo incluyen a quienes perciben cada tipo de ingreso, excluyendo a aquellos que no indican recibirlo.

4.3.2.1 Regresión de Heckman

Este método completa los datos faltantes utilizando como base las observaciones completas a través de un modelo de regresión en dos etapas. La primera ecuación del modelo es una "ecuación de selección", la cual, a través de un modelo *probit*, estima la probabilidad de selección o participación de un individuo. De dicha estimación se obtiene un estadístico conocido como la razón inversa de Mills, el cual captura la magnitud del sesgo de selección. Posteriormente, se incorpora la razón inversa de Mills como un regresor al modelo original, donde se imputan los datos faltantes y luego se estima por MCO la variable objetivo (Heckman, 1979).

Este modelo requiere incorporar el supuesto de que los errores cuentan con una distribución normal y que tanto los datos observados como los faltantes se distribuyen de la misma manera.

Ecuación de selección o participación

$$Participa = \alpha_0 + \alpha_1 Edad + \alpha_2 experiencia + \alpha_3 experiencia^2 + \alpha_k ...$$

Ecuación de ingreso

$$Ln(ingreso) = \beta_0 + \beta_1 educacin + \beta_2 experiencia + \beta_3 experiencia^2 + \beta_k ... + \beta(k + 1^{\lambda})$$

Como ecuación de selección, se utiliza la no respuesta al ingreso, considerando las variables escogidas en el apartado de variables relacionadas con la no respuesta. Entre ellas es posible encontrar sexo, sustentador principal del hogar, macrozona, escolaridad, entre otras, dependiendo del tipo de ingreso. Como ecuación objetivo, se estima el ingreso y se incluye como variables independientes la razón inversa de Mills (calculada de la primera ecuación) y variables asociadas al ingreso, como escolaridad, edad, edad al cuadrado como proxy de experiencia, zona, CIUO, entre otras, que fueron definidas en el apartado anterior.

A continuación se presentan los modelos para cada categoría de ingreso.

Modelos de selección

Cuadro 16: Variables utilizadas en las ecuaciones de selección para el modelo de regresión de Heckman

Variables	Asalariados	Honorarios	C.Propia	P.Independientes	Jubilados
Sexo	X		X	X	
Edad					
Sustentador principal	X	X	X	X	
Escolaridad	X			X	
CSE	X	X			X
Informante directo	X		X	X	X
Sexo informante	X				X
Edad informante	X	X	X	X	
Sustentador principal informante	X				
Escolaridad informante			X		
Macrozona	X	X	X	X	X

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Regresión de ingresos

Cuadro 17: Variables utilizadas en las ecuaciones de ingreso para el modelo de regresión de Heckman

Variables	Asalariados	Honorarios	C.Propia	P.Independientes	Jubilados
Sexo	X	X	X	X	X
Edad	X	X	X	X	X
Edad^2	X	X	X	X	
Sustentador principal	X	X	X	X	X
Presencia de menores de 6 años en el hogar	X				
Presencia de menores de 15 años en el hogar			X	X	
Escolaridad	X	X	X	X	X
Macrozona	X	X	X	X	X
CSE	X	X	X	X	X
CIUO	X	X	X	X	
CISE	X		X	X	
Sistema previsional					X

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Como fue mencionado, en el cuadro 17 se presentan las variables incluidas en los modelos de selección para la imputación de cada categoría de ingreso, es decir, variables explicativas de la no respuesta parcial para cada grupo, presentadas en el apartado de variables relacionadas con la no respuesta. Dado que los individuos que reciben distintos tipos de ingresos cuentan con diferentes características y comportamientos, viéndose, a la vez, afectados por variables sociodemográficas de distinta forma, los modelos de selección varían entre cada grupo. De igual manera, la selección de las variables a utilizar como regresores al imputar el ingreso, se realiza de acuerdo al nivel de significancia para explicar dichos ingresos y, finalmente, son ordenadas según su nivel de correlación, como fue explicado anteriormente en el apartado de variables relacionadas con el ingreso.

4.3.2.2 Hot deck

El método de imputación *hot deck* utiliza las observaciones con información completa presentes en la base de datos para completar los valores faltantes. La elección del donante de la información se hace de manera aleatoria dentro de un grupo de individuos que comparte determinadas características con el receptor.

La imputación a través de *hot deck* requiere, como primer paso, seleccionar las variables que determinarán la similitud de los donantes con el receptor. Se buscan variables que se relacionen con aquella que presenta datos faltantes. Posteriormente, se generan diferentes niveles de similitud, agrupando o eliminando ciertas variables, para así ampliar la posibilidad de cada individuo de encontrar uno o más donantes. Finalmente, para cada receptor se escoge de manera aleatoria un individuo del primer nivel donde se ha encontrado donantes y se le duplica la información de dicho donante (Andridge & Little, 2010).

Para la construcción de las matrices de niveles de imputación, se utilizan las variables seleccionadas anteriormente en el modelo de ingresos.

Variables modelo Hot deck

Cuadro 18: Variables utilizadas para la imputación por Hot deck

Variables	Asalariados	Honorarios	C.Propia	P.Independientes	Jubilados
Sexo	X	X	X	X	X
Escolaridad*	X	X	X	X	X
Edad*	X	X	X	X	X
Sustentador principal	X	X	X	X	X
CSE	X	X	X	X	X
Presencia de menores de 6 años en el hogar	X				
Presencia de menores de 15 años en el hogar			X	X	
CIUO*	X	X	X	X	
CISE*	X		X	X	
Sistema previsional					X
Variables geográficas*	X	X	X	X	X

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)
* Variables que cuentan con diferentes niveles de agregación

Como ya fue mencionado, las variables presentadas en el cuadro 18 fueron incorporadas por su nivel de correlación con el ingreso a imputar. Cuando ya se han seleccionado las variables a utilizar, se comienzan a generar los distintos niveles de imputación. Las variables escogidas se agrupan o eliminan para la construcción del siguiente nivel según su correlación, privilegiando mantener las variables teóricas. Para ir generando nuevos niveles, se tomaron en consideración las siguientes reglas:

- Se ordenan las variables según su nivel de correlación con el ingreso, de manera descendente.
- Se agrupa o “relaja” una variable por nivel. Se parte por aquella con menor correlación y que pertenezca al grupo de variables empíricas mencionadas anteriormente.

- Después de agrupar o “relajar” todas las variables, se elimina aquella variable empírica con menor correlación.
- Se repite el proceso con las variables teóricas.

Dada la cantidad de variables y las reglas establecidas para la creación de niveles, se genera un gran número de niveles para cada tipo ingreso, sin embargo no todos son utilizados.

Cuadro 19: Niveles de imputación generados y niveles de imputación utilizados

Ingresos	N.Creados	N.Utilizados
Asalariados	54	35
Honorarios	45	19
Cuenta propia	54	32
Profesionales independientes	55	30
Jubilados	27	16

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

La metodología de generación de matrices de imputación de la VIII EPF difiere de la utilizada en la VII EPF. En particular, para la VIII EPF se sistematiza la manera en que las variables son agrupadas y/o eliminadas, haciendo uso de las correlaciones para ordenarlas y darles prioridad. Así, se define claramente cómo se debe pasar al siguiente nivel. Lo anterior genera que, a medida que se agrupan las variables, el tamaño de los *clusters* crezca, pero al eliminar una variable y devolver a su estado inicial a aquellas que se mantienen, el tamaño del *cluster* vuelve a disminuir.

Finalmente, con este método de imputación todos los casos con datos faltantes pueden ser imputados.

4.3.2.3 Método de media condicional

De manera similar al *hot deck*, este método de imputación utiliza las observaciones con datos íntegros para imputar la información faltante, pero se diferencia en la elección del donante, ya que no utiliza un individuo por nivel, sino la media de todos aquellos que se encuentran en un mismo nivel.

Al igual que en el método anterior, se seleccionan las características que determinan la similitud de los donantes con el receptor y la elección de estas variables depende también de la relación de éstas con la variable a imputar. En este caso, se han utilizado las mismas variables para ambos métodos.

Se generan los diferentes niveles de imputación, agrupando o eliminando las variables escogidas. Posteriormente, se agrupan los donantes en los niveles de imputación creados y al receptor se le imputa la media de todo el grupo de donantes del primer nivel donde ha sido clasificado.

Este método realiza los mismos supuestos que el anterior respecto a la no respuesta aleatoria y el comportamiento de quienes responden y de quienes no responden.

Para la prueba con los datos de la VIII EPF, al igual que *hot deck*, todos los casos con datos faltantes son imputados.

4.3.2.4 Imputación múltiple

El método de imputación múltiple se basa en la repetición del proceso de imputación con algún mecanismo seleccionado, para luego combinar los resultados generados por esta serie de imputaciones. Para comenzar la imputación, se debe determinar el modelo a utilizar para imputar, como por ejemplo, el método de regresión o el *propensity score*, entre otros. Luego de que se ha iterado y se cuenta con suficientes repeticiones de la imputación, es necesario analizar el resultado. Para esto, se deben combinar todas las imputaciones realizadas para cada dato faltante por un solo valor, esto puede hacerse a través de la media, por ejemplo, u algún otro estadístico que se determine.

Los supuestos requeridos para este método están asociados al modelo escogido, sin embargo, dado que se utilizan los datos completos como base, es necesario asumir similitud de comportamiento entre quienes responden y quienes no lo hacen.

Para la prueba que se realiza con los datos de la VIII EPF, se escoge como método de imputación el método de regresión con una ecuación tipo Mincer y se realizan 30 repeticiones de cada imputación. Posteriormente se combinan las 30 repeticiones, calculando la media de éstas para cada individuo con datos faltantes, y se reemplaza en el valor perdido.

Cuadro 20: Variables modelo de regresión para imputación múltiple

Variables	Asalariados	Honorarios	C.Propia	P.Independientes	Jubilados
Sexo	X	X	X	X	X
Edad	X	X	X	X	X
Edad^2	X	X	X	X	
Sustentador principal	X	X	X	X	X
Presencia de menores de 6 años en el hogar	X				
Presencia de menores de 15 años en el hogar			X	X	
Escolaridad	X	X	X	X	X
Macrozona	X	X	X	X	X
CSE	X	X	X	X	X
CIUO	X	X	X	X	
CISE	X		X	X	
Sistema previsional					X

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Este método no consigue imputar todos los datos faltantes de ingreso, ya que cuando alguno de los regresores tiene un valor perdido, la imputación no se lleva a cabo.

4.4 Evaluación de los métodos

4.4.1 Características de la simulación

De la misma forma que se ha realizado con la imputación de gastos individuales, para evaluar el desempeño de los diferentes métodos estudiados, se utilizan simulaciones para poder tener en

consideración un escenario comparativo. Esto, debido a que, al igual que los datos de gastos individuales, la VIII EPF tampoco cuenta con información adicional de ingresos que permita evaluar los métodos.

De esta manera, en primer lugar se seleccionan todas las observaciones con información completa, según cada tipo de ingreso. Luego, a dichas observaciones se le incorporan valores perdidos, para poder simular la imputación bajo los distintos métodos y, finalmente poder comparar los valores imputados y los valores reales de estas observaciones.

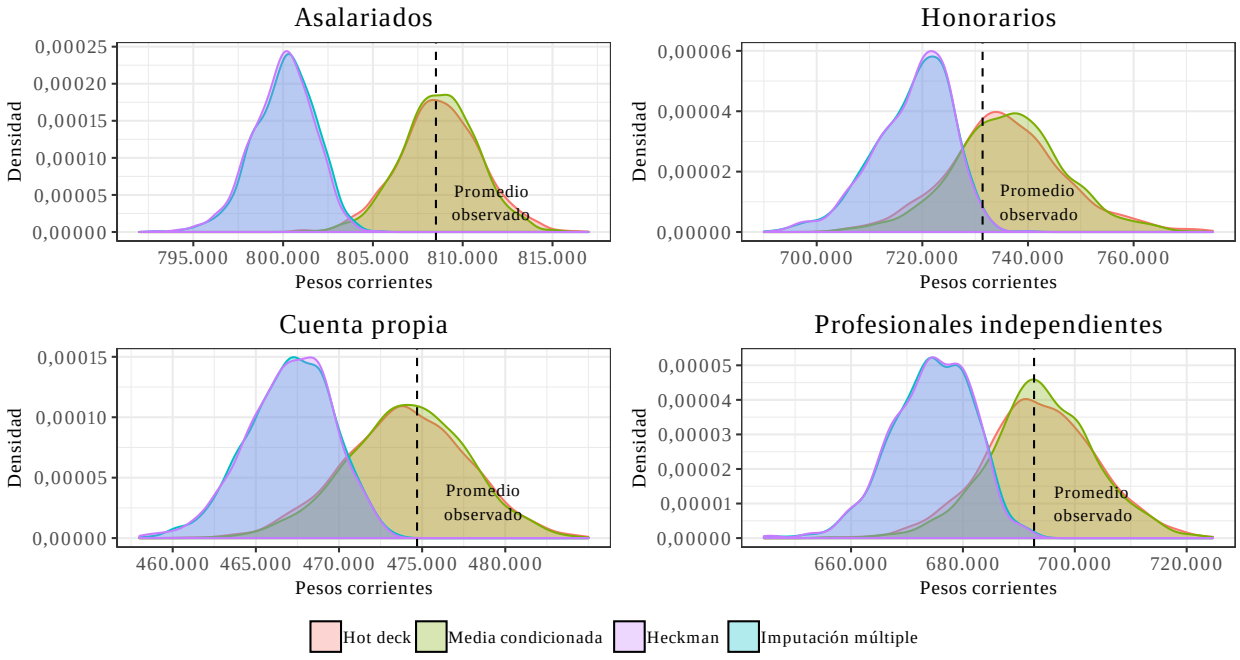
En el caso de la imputación de ingresos, y como ha sido sugerido anteriormente, las simulaciones asumen un mecanismo de no respuesta MAR (*missing at random*) para así realizar supuestos que se acerquen más a la realidad. Para generar la no respuesta y crear valores perdidos en las bases de datos, se utilizan los mismos modelos presentados en el apartado sobre variables relacionadas con la no respuesta.

Dado que la realización de experimentos puede generar diferentes resultados, se utilizan 1000 iteraciones, para así calcular promedios y suavizar el efecto de resultados anómalos producto del azar. Para esta serie de iteraciones, se utiliza una semilla para poder obtener los mismos resultados cada vez que se ejecute el ejercicio.

4.4.2 Resultados de las simulaciones

A continuación, se presenta la distribución de los promedios de los datos imputados para los diferentes grupos de trabajadores y para los jubilados.

Figura 42: Distribución de promedios imputados de los ingresos del trabajo, 1000 simulaciones



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

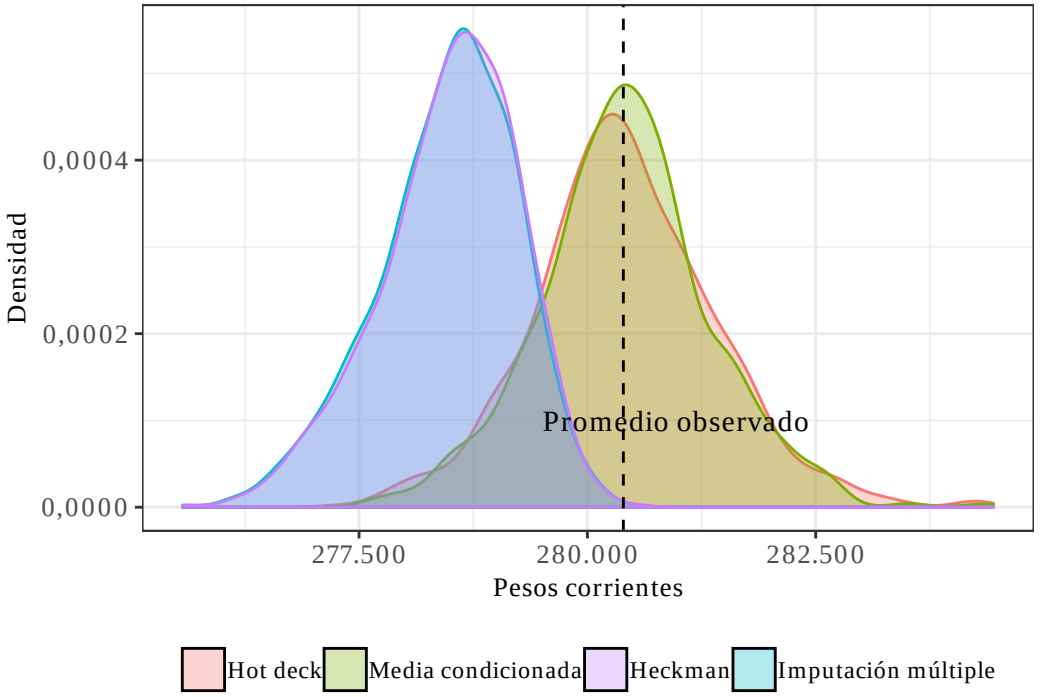
La figura 4.4.2 muestra una clara diferencia entre la distribución de las simulaciones. Se distinguen dos grupos, de acuerdo al comportamiento de los métodos de imputación. Por un lado, los métodos de imputación múltiple y regresión de Heckman y, por otro, *hot deck* y media condicionada.

Respecto al primer grupo de métodos (imputación múltiple y regresión de Heckman), se observa que sus distribuciones se alejan del promedio observado y solo la cola superior se sobrepone con la distribución de los otros dos métodos. En el caso de los asalariados la sobreposición es leve, lo que es relevante, ya que son el grupo de mayor tamaño. Para el resto de los tipos de ingresos la coincidencia es mayor, lo que significa que hay más casos donde es posible que todos los métodos den el mismo resultado. Además, cabe destacar que tanto imputación múltiple como regresión de Heckman utilizan la misma ecuación de ingresos. El último de estos métodos incorpora el sesgo de selección, lo que pareciera no tener un efecto importante a la hora de realizar las pruebas, ya que ambos métodos se comportan de manera muy similar.

En relación con el segundo grupo de métodos (*hot deck* y media condicionada), se advierte que se sobreponen a lo largo de toda la distribución. Ambos métodos están centrados en torno al promedio observado. Dada la construcción del método de media condicionada (utilización del promedio de ingresos de un cluster para imputar), en este caso se produce una mayor densidad cerca del promedio observado.

Para las cuatro categorías de ingreso, el comportamiento de los métodos es similar. Se debe notar que la escala utilizada en el eje horizontal es pequeña y esto debe ser considerado a la hora de evaluar las diferencias. Asimismo, debe tenerse en consideración que la escala no es igual para todos los gráficos.

Figura 43: Distribución de promedios imputados del ingreso bruto de las jubilaciones, 1000 simulaciones



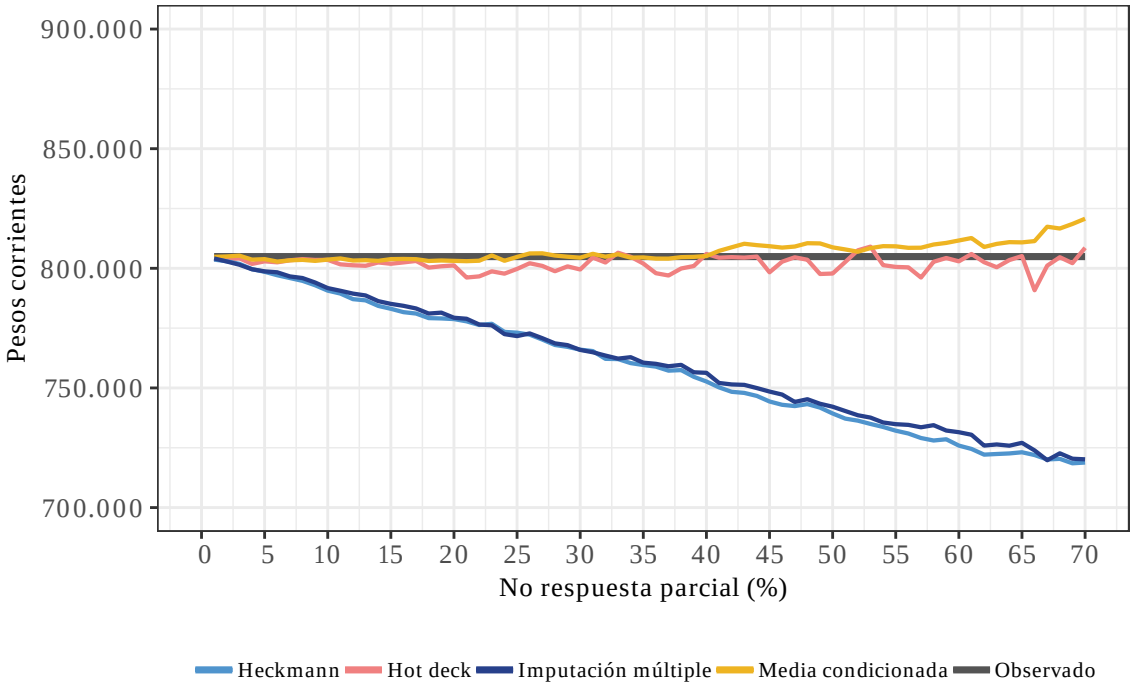
Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares

Finalmente, los resultados de las pruebas para los ingresos de jubilaciones muestran el mismo comportamiento observado anteriormente. La imputación múltiple y la regresión de Heckman exhiben semejanzas en la distribución y se encuentran más alejados del promedio observado. La media condicionada y *hot deck*, por su parte, se desempeñan de forma muy similar y se acercan de mejor modo al promedio observado.

Es interesante destacar que los dos métodos paramétricos probados (imputación múltiple y regresión de Heckman) se comportan de manera muy similar y ambos se alejan más del promedio observado que los métodos no paramétricos. Cabe señalar que imputación múltiple y regresión de Heckman permiten identificar de mejor forma la contribución de cada variable y, por ende, privilegiar, eliminar o agrupar variables para enriquecer el modelo. Sin embargo, logran resultados mucho menos satisfactorios que los métodos no paramétricos.

A continuación se presenta un último ejercicio, donde se aumenta progresivamente la tasa de no respuesta parcial. De esta manera, se logra observar cómo se comportan todos los métodos en diferentes escenarios. Es importante contar con los resultados de esta prueba, ya que es necesario evaluar cómo se comportarían los modelos con una menor tasa de respuesta.

Figura 44: Comparación media imputada a distintos niveles de no respuesta parcial



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Como es posible observar, los métodos de media condicionada y *hot deck* se mantienen cercanos a la media observada. Incluso cuando se enfrentan a un 70% de no respuesta, estos resultados apoyan la evidencia encontrada en las comparaciones realizadas anteriormente, donde ambos métodos muestran distribuciones que se acercan al promedio observado.

Por otra parte, los métodos paramétricos, es decir, imputación múltiple y regresión de Heckman se alejan progresivamente de la media observada. Este ejercicio es relevante, ya que el estudio de diferentes métodos de imputación es probado con los datos de la VIII EPF, pero debe ser un insumo para futuras versiones de la encuesta, en las cuales eventualmente podría ser necesario enfrentar una mayor incidencia de la no respuesta parcial.

El propósito de los ejercicios presentados ha sido evaluar y seleccionar un método de imputación para la no respuesta parcial en las principales categorías de ingresos. Particularmente, se evaluaron métodos para la imputación de los ingresos provenientes de la ocupación principal y de jubilación y pensiones de vejez.

Los resultados de las simulaciones muestran que, si se compara el desempeño de los métodos entre las distintas categorías, estas tienen un rendimiento similar, permitiendo elegir un método de imputación común para estas. Antes de la realización de estos ejercicios no resultaba evidente que los métodos tuvieran un desempeño similar entre los distintos grupos de ingreso. En primer lugar, el porcentaje de no respuesta parcial difiere de manera importante entre las categorías de ingreso, siendo tan bajo como 3,4% para el caso de los jubilados y tan alto como 9,9%, para el caso de los trabajadores a honorarios. Por otro lado, al modelar la propensión a responder para cada grupo, se observa que, aunque se repiten ciertos patrones en las variables que permiten explicar la respuesta (como por ejemplo la importancia de que el perceptor sea informante directo de sus datos de ingreso), cada partida presenta particularidades que hablan de la existencia de distintos mecanismos generadores de la falta de datos de cada grupo.

A pesar de estas diferencias, los resultados de las simulaciones ubican a *hot deck* y la media condicionada como los mejores algoritmos para mitigar el efecto de la no respuesta parcial en ingresos. Tal como ya fue descrito en la sección de gastos diarios de este documento, si se considera el sesgo promedio como una medida para identificar cuánto se aleja el promedio de ingresos tras el proceso de imputación respecto del promedio observado, se observa que este resulta más bajo para *hot deck* y media condicionada en todas las categorías de ingreso analizadas.

$$Sesgo = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\overline{X}}_i - \overline{X})$$

Cuadro 21: Sesgo promedio por método de imputación

Partida	Hot deck	Media condicionada	Heckman	Imputación múltiple
Asalariados	237	203	-8.527	-8.393
Honorarios	4.588	4.995	-13.146	-13.157
Cuenta propia	-1.657	-1.691	-26.507	-24.403
Profesionales independientes	-1.840	-1.807	-45.636	-37.682
Jubilados	22	22	-1.901	-1.928

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Ahora bien, al comparar *hot deck* y media condicionada, no se observan diferencias importantes en términos estadísticos.[36] Ambos métodos tienen rendimientos similares en las simulaciones de las diferentes categorías de ingresos, sin observarse diferencias sustantivas en sus resultados. En términos de sesgo promedio, sus valores son prácticamente idénticos, así como también en relación al promedio imputado. La diferencia más importante que se observa entre estos tiene que ver con el cambio que se introduce a la varianza de la distribución de los ingresos. Al revisar los resultados de las simulaciones, para 4 de las 5 categorías de ingreso, el método de media condicionada reduce en mayor medida la desviación estándar de la distribución de los ingresos respecto a la desviación efectivamente observada. En este sentido, este algoritmo estaría reduciendo artificialmente la varianza de las estimaciones, lo que amenaza con invalidar los cálculos para el intervalo de confianza de estas, así como también de sus test de hipótesis. La siguiente tabla resume la desviación estándar observada para los datos de las simulaciones en cada grupo de ingresos, junto con la desviación promedio de las mil simulaciones para el método *hot deck* y la media condicionada.

Cuadro 22: Desviación promedio por método de imputación

Partida	Observada	Hot deck	Media condicionada
Asalariados	808.509	808.746	808.712
Honorarios	731.407	735.995	736.402
Cuenta propia	472.873	471.216	471.182
Profesionales independientes	688.058	686.218	686.251
Jubilados	280.394	280.416	280.416

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Aunque ambos métodos entregan soluciones, en principio, muy parecidas para el problema de la no respuesta parcial, *hot deck* preserva de mejor manera la distribución de los datos, haciéndolo preferible como método de imputación para la VIII EPF. El objetivo secundario de la encuesta es identificar la estructura de ingreso de los hogares urbanos. Para cumplir con este objetivo no es suficiente con representar de manera confiable el promedio de los ingresos, indicador que predicen con casi idéntica fidelidad *hot deck* y media condicionada, sino que también es relevante para la encuesta que el método sea capaz de reproducir la distribución de los ingresos y sus distintas categorías.

4.4.3 Resultados con datos oficiales

A continuación se presentan los resultados de la imputación de ingresos con los datos oficiales de la encuesta. Con el objetivo de evaluar el efecto que puede tener el cambio de metodología respecto a lo realizado en la VII EPF, se comparan las estimaciones de ingresos considerando *hot deck* y media condicionada como métodos de imputación.

El cuadro 32 presenta el porcentaje de casos imputados de acuerdo al nivel de la matriz, tanto para media condicionada como para *hot deck*, considerando al grupo de trabajadores asalariados[37]. Es

[36] El método de media condicionada es levemente más preciso en este ejercicio que *hot deck*.

[37] Se ha decidido presentar este grupo pues corresponde al grupo de trabajadores más numeroso.

posible apreciar que en el séptimo nivel ya se ha imputado más del 55% de los casos.

Al comparar los resultados de este cuadro con los presentados para la imputación de gastos 7, se observan diferencias en el tamaño de los *clusters* a medida que se avanza en los niveles de la matriz. Aunque la tendencia para ingresos es de aumento del tamaño de los cluster para los niveles más altos, esta no es monotónica como en el caso de gastos diarios. Esta diferencia se explica por la forma en que se construyen los niveles de la matriz para la imputación de ingresos. En esta, cuando una variable era eliminada para la conformación de los *clusters*, el resto de las variables vuelven a tener el nivel de desagregación que tuvieron inicialmente. En este sentido, por construcción, la eliminación de una variable en la matriz de transferencia implica una reducción en el tamaño promedio de los *clusters*.

Cuadro 23: Niveles en los que se realizaron las imputaciones

Nivel de imputación	Tamaño promedio cluster	Número de observaciones	Porcentaje de libretas
1	1,00	2	0,14
2	1,02	19	1,36
3	1,16	154	11,03
4	1,18	23	1,65
5	1,23	27	1,93
6	2,36	413	29,58
7	3,17	161	11,53
8	5,46	139	9,96
9	5,48	1	0,07
10	7,75	163	11,68
11	8,13	15	1,07
12	15,83	103	7,38
13	1,37	1	0,07
14	3,12	12	0,86
15	4,47	6	0,43
16	8,64	6	0,43
18	12,42	12	0,86
20	26,07	15	1,07
21	1,50	5	0,36
22	3,92	12	0,86
23	5,74	3	0,21
24	11,46	4	0,29
25	16,92	10	0,72
27	37,52	6	0,43
29	5,96	12	0,86
30	9,25	2	0,14
31	20,26	5	0,36
32	30,30	7	0,50
34	67,95	7	0,50
35	7,34	26	1,86
36	37,05	12	0,86
37	62,65	4	0,29
38	153,39	1	0,07
39	250,21	1	0,07
47	540,35	7	0,50

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

El cuadro 24 presenta las estimaciones de ingresos para las distintas categorías imputadas, junto con sus medidas de precisión. Para guiar la lectura de este cuadro conviene notar que:

1) Para cada ingreso, primero se presentan las estimaciones para los datos efectivamente observados sin imputar en aquellos casos en los que falta respuesta. El grupo de trabajadores dependientes agrupa los ingresos de los trabajadores asalariados y honorarios, y el grupo de trabajadores independientes, a los cuenta propia y profesionales independientes.
2) Posteriormente, se presentan los resultados de las estimaciones considerando los datos observados junto con los imputados por *hot deck*.
3) Finalmente, se presentan los resultados de las estimaciones con datos observados y datos imputados con el método de media condicionada.

Todas las estimaciones se realizaron considerando el diseño complejo de la encuesta.

Cuadro 24: Estadísticos descriptivos por categoría de ingreso y método de imputación

Partida	Promedio	Error estándar	CI Inf	CI Sup	CV
Trabajadores dependientes (observado)	**816.914**	**28.850**	**760.093**	**873.736**	**3,532**
Hot deck	819.313	29.862	760.497	878.129	3,645
Media condicionada	817.314	29.298	759.610	875.018	3,585
Asalariados (observado)	**821.999**	**29.758**	**763.389**	**880.610**	**3,620**
Hot deck	824.853	30.765	764.259	885.447	3,730
Media condicionada	822.857	30.190	763.396	882.317	3,669
Honorarios (observado)	**736.971**	**39.113**	**659.936**	**814.007**	**5,307**
Hot deck	733.082	37.323	659.572	806.592	5,091
Media condicionada	731.032	37.926	656.334	805.731	5,188
Trabajadores independientes (observado)	**545.042**	**30.408**	**485.151**	**604.933**	**5,579**
Hot deck	546.974	29.164	489.533	604.416	5,332
Media condicionada	545.594	29.057	488.363	602.824	5,326
Cuenta propia (observado)	**464.554**	**28.010**	**409.385**	**519.722**	**6,030**
Hot deck	467.237	27.443	413.186	521.288	5,873
Media condicionada	467.741	27.248	414.074	521.407	5,825
Profesionales independientes (observado)	**727.398**	**56.381**	**616.352**	**838.444**	**7,751**
Hot deck	718.079	51.942	615.775	820.383	7,233
Media condicionada	712.656	52.022	610.195	815.118	7,300
Jubilados (observado)	**275.095**	**8.180**	**258.983**	**291.206**	**2,974**
Hot deck	275.666	8.090	259.732	291.601	2,935
Media condicionada	276.393	8.112	260.415	292.371	2,935

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

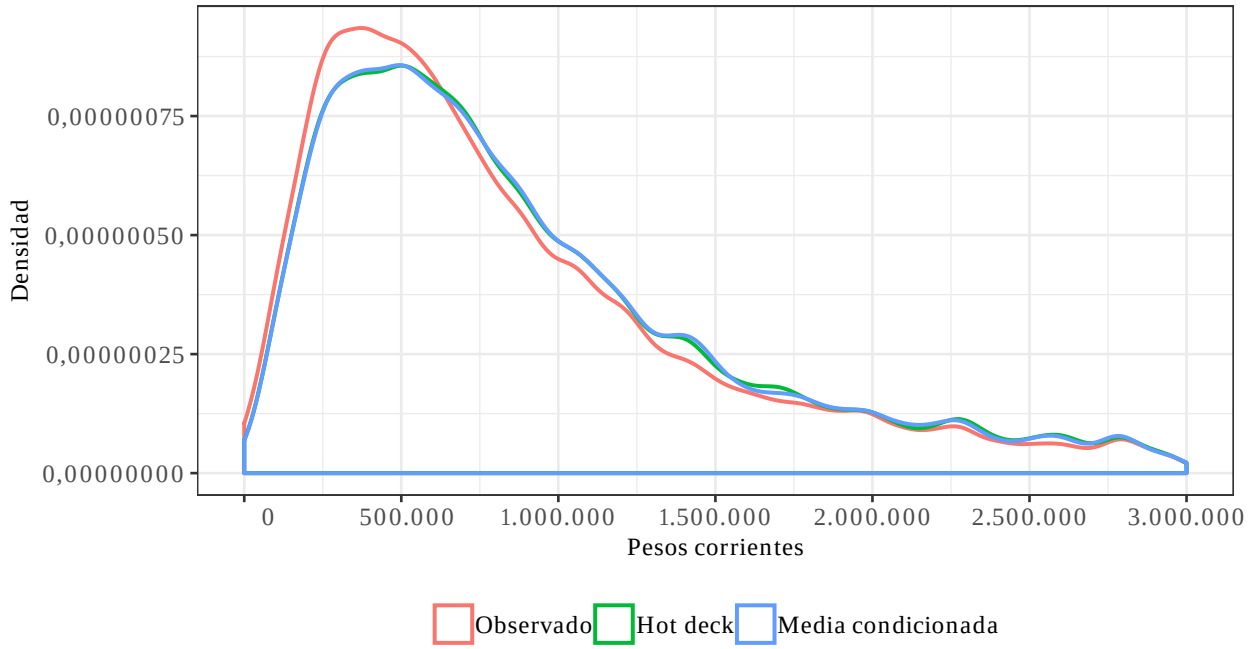
Tal como ya había sido observado en las simulaciones, los métodos *hot deck* y media condicionada entregan resultados bastante similares para los distintos grupos de ingreso. En el caso de los trabajadores asalariados, honorarios y profesionales independientes, las estimaciones para las que se corrigió la no respuesta utilizando *hot deck* son levemente superiores a las de media condicionada, contrario a lo que sucede con los ingresos de los trabajadores por cuenta propia y los ingresos por jubilaciones. En ningún caso estas diferencias son estadísticamente significativas.

Al analizar los coeficientes de variación, se observa que, al mirar los valores agregados para el grupo de trabajadores dependientes e independientes, el método *hot deck* tiende a presentar valores

levemente superiores a los que se observa para las estimaciones corregidas utilizando media condicionada. Estos resultados se encuentran en línea con lo observado en las simulaciones, donde el método *hot deck* tiende a preservar de mejor manera la distribución de los datos.

El gráfico 45 describe la distribución del ingreso disponible de los hogares, de acuerdo a una función de densidad de Kernel. En este se incluye la distribución de los datos observados sin corrección por algún método de imputación, la distribución para los datos corregidos por *hot deck* y la que se observaría si se usara el método de media condicionada. Se constata que ambos métodos suavizan la distribución de ingresos en una misma dirección, traspasando casos desde la parte baja de la distribución hacia la media. Esta tendencia es similar a la que se observa para la imputación de gastos diarios cuando se compara la distribución de acuerdo a la corrección con el método FNR y *hot deck* (figura 22).

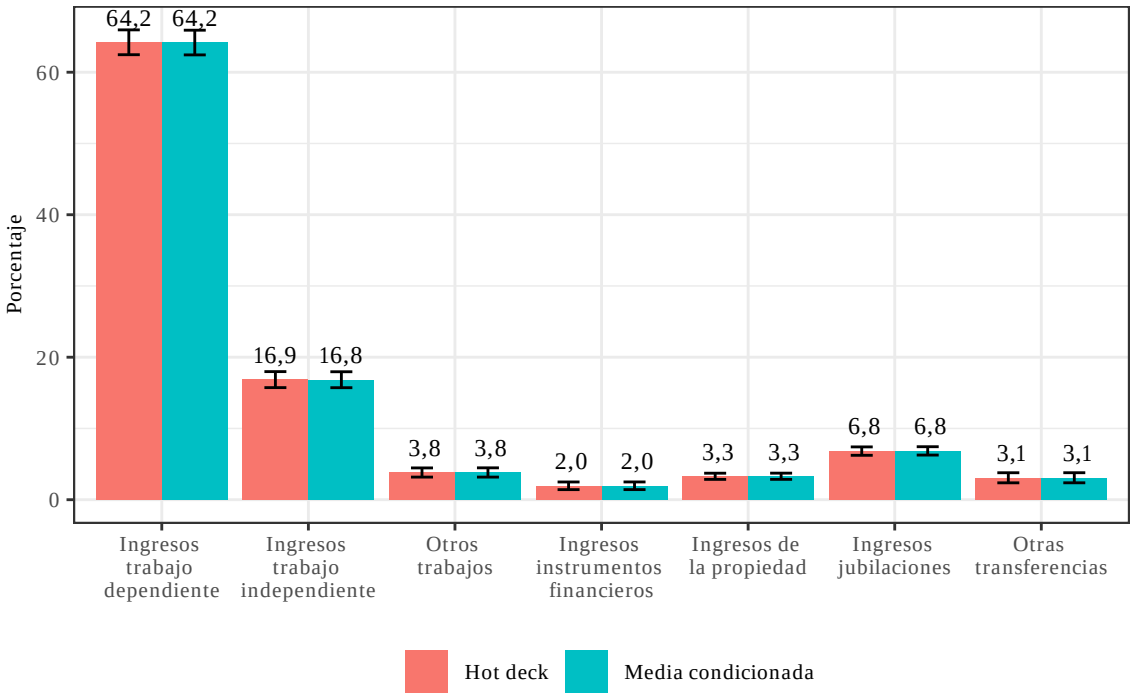
Figura 45: Distribución del ingreso disponible de los hogares



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

Finalmente, si se compara la estructura de ingresos, tampoco se observan diferencias importantes entre los métodos *hot deck* y media condicionada. En este sentido, los resultados más importantes se concentran en que el método *hot deck* entrega una solución a la falta de respuesta que no subestima la varianza en las categorías de ingreso imputadas.

Figura 46: Estructura del ingreso disponible de los hogares



Fuente: Instituto Nacional de Estadísticas (INE) – VIII Encuesta de Presupuestos Familiares (EPF)

5 Conclusiones

En este documento, se presentan los diferentes métodos de imputación que se testearon al momento de decidir cómo lidiar con el problema de la no respuesta parcial en la VIII EPF, en lo relativo a ingresos y gastos diarios. De este modo, se busca dar a conocer a los usuarios el detalle de la metodología seleccionada y los motivos que llevaron a preferir un método por sobre los otros.

Además de avanzar en una mayor transparencia respecto al modo en el que se elaboran las estadísticas oficiales, el presente documento intenta aportar a la discusión respecto al tratamiento de la no respuesta parcial, que, pese a ser un problema persistente para las encuestas de hogares, no ha sido abordado de manera amplia en la literatura chilena. En ese sentido, el documento busca perfilarse como una continuación de la publicación realizada sobre imputaciones en el marco de la VII EPF[38], de modo de ir generando un conocimiento acumulativo respecto al tratamiento de la no respuesta parcial, que entregue herramientas para la mejora continua de los productos estadísticos del Instituto Nacional de Estadísticas y otras instituciones productoras de estadísticas.

Respecto a gastos diarios, el principal desafío que debía abordar el equipo de la VIII EPF era la imputación de Libretas de Gastos Individuales completamente rechazadas. La metodología que se utilizó para imputar gastos diarios, tanto en la VI como en la VII EPF, fue el método de Factor de no Respuesta. Dicho método, como se ha mostrado en este documento, resuelve de un modo relativamente satisfactorio el problema de las libretas parcialmente completadas, pero no es capaz de abordar de manera adecuada el problema de las libretas completamente rechazadas.

Considerando el aumento de las libretas rechazadas y las prácticas internacionales respecto a este tema, el equipo técnico estudió el efecto de imputar libretas completamente rechazadas mediante el método de *hot deck*, el cual, como ya se ha señalado, fue la estrategia finalmente adoptada. Los motivos que se tenían para explorar este método en profundidad eran múltiples. En primer lugar, el estudio de imputación realizado durante la VII EPF sugiere que esta era una estrategia plausible para imputar libretas completas, a diferencia de otros métodos usados para imputar gastos diarios, como el ajuste por peso diario y el FNR. En segundo lugar, el método de *hot deck* es una estrategia utilizada en otros países para imputar gastos e ingresos, por lo cual existe documentación que muestra cuáles son sus resultados. En tercer lugar, *hot deck* es una herramienta flexible, que permite adecuaciones para cada contexto. Finalmente, cabe señalar que *hot deck* es un método relativamente sencillo de comunicar, lo cual no deja de ser relevante, considerando el carácter público que tiene la información producida por el INE.

Los resultados de las simulaciones muestran que *hot deck*, gracias a la imputación de libretas rechazadas, logra reducir el sesgo en el gasto promedio de mejor manera que el FNR. Asimismo, los ejercicios dan cuenta de que, incluso en contextos de alta no respuesta parcial, el método de *hot deck* muestra un desempeño aceptable, lo cual lo hace robusto a ciertas subpoblaciones con bajos niveles de colaboración en el reporte de sus gastos diarios.

En relación con los datos oficiales, el promedio de gastos en el método de *hot deck* supera al de FNR en aproximadamente 65.000 pesos. Este dato es positivo en tres sentidos. En primer lugar, porque existe evidencia de que, en general, las personas no reportan todos sus gastos, lo que implica

[38] Disponible en el sitio web www.ine.cl/epf

una subestimación del mismo ante la cual el método contribuye a su corrección. En segundo lugar, porque es indudable que la no imputación de libretas rechazadas en el contexto del FNR genera una subestimación del gasto promedio por hogar. En tercer lugar, porque si bien existe un efecto del método *hot deck* en el gasto promedio, este no es demasiado grande, lo cual permite mantener un cierto nivel de comparabilidad en el tiempo entre las distintas versiones de la encuesta.

En relación con la distribución del gasto, el efecto más importante se produce en la división de alimentos y bebidas no alcohólicas, cuya participación aumenta desde 17,4% hasta 18,7%, al transitar desde FNR a *hot deck*. Este cambio también fue identificado como algo deseable, por cuanto la evidencia a nivel internacional respecto a la proporción de gasto que destinan los países en relación con su ingreso, da cuenta de que Chile se ubica muy por debajo de lo esperado, lo cual hace pensar en una subestimación de este tipo de gastos. En ese sentido, *hot deck* mueve la participación del gasto en alimentos en la dirección esperada. Un segundo efecto relevante producido por *hot deck* en la estructura de gastos es el aumento que tiene la participación de estupefacientes y bebidas alcohólicas. Al respecto, se evidencia un aumento que, aunque leve, es estadísticamente significativo, lo cual es deseable, ya que la información de Cuentas Nacionales muestra que, tanto en Chile como en otros países, las encuestas de presupuestos familiares subestiman el gasto en alcohol y estupefacientes. Esto muy probablemente se deba a la escasa disposición de parte de las personas a revelar este tipo de consumo.

Respecto a la imputación de ingresos, se planteó el objetivo de actualizar el método de imputación hacia un mecanismo que cumpliera con hacerse cargo completamente de la no respuesta, que mejorara los modelos de no respuesta y de ingresos, que respetara la distribución de los datos y que permitiera sistematizar los pasos, para así poder replicar el proceso en toda ocasión. En la VII EPF se utilizó el método de media condicionada, que sin bien se hace cargo de la no respuesta y es posible de sistematizar, genera alteraciones en la distribución, debido a su tendencia a converger al promedio.

La realización de simulaciones y ejercicios probando los métodos imputación múltiple, regresión de Heckman, *hot deck* y media condicionada desembocó en la selección del método *hot deck* como mecanismo de imputación preferible. Esto se justifica principalmente en tres razones: las estimaciones puntuales generadas en las simulaciones son casi idénticas a los datos observados; si se compara con los métodos paramétricos, preserva de mejor forma de distribución de los datos y, finalmente, es un método simple, sistematizable, explicable y replicable sin mayor complejidad.

Los resultados de las simulaciones entregaron gráficamente evidencia respecto a cómo afectaban la distribución de los datos cada uno de los métodos y se evidenció la superioridad de los métodos *hot deck* y media condicionada. Por otra parte, el cálculo del sesgo muestra que en estos dos mecanismos el sesgo fluctúa entre 481 pesos y 5.267 pesos, versus el sesgo mostrado por imputación múltiple y regresión de Heckman, que va desde 10.660 pesos hasta los 45.636 pesos.

Al imputar los casos perdidos de la base de la VIII EPF con ambos métodos, los resultados de la estimación son muy similares entre ellos. Ante esto, se debe considerar que uno de los objetivos de la encuesta es presentar la estructura de ingresos de los hogares y, por lo tanto, preservar la distribución se torna fundamental. Siguiendo este principio, *hot deck* se comporta de mejor forma, ya que no tiende a concentrar la distribución y achatarla de manera artificial, como se ve en los coeficientes de variación, que tienden a ser menores en el caso de la media condicionada.

Finalmente, vale la pena mencionar algunos desafíos para una próxima versión de la encuesta:

- El método de *hot deck* es una buena estrategia para llevar a cabo una imputación aproximadamente insesgada, lo cual es de gran relevancia, ya que el insesgamiento es la propiedad más importante de un estimador. Una vez resuelto ese problema, un método de imputación debiese hacerse cargo del aumento de la varianza generado por el efecto de imputar. En ese sentido, incluir una estrategia que incorpore la pérdida de precisión generada por la imputación es un desafío que debiese ser atendido en una próxima versión de la encuesta. Al respecto, existen trabajos que proponen estrategias para considerar el aumento de la varianza, en un contexto de imputación por *hot deck* (Righi, Falorsi, & Fasulo, 2014; Righi et al., 2014), que posiblemente podrían implementarse en una próxima versión. La importancia de llevar a cabo este ajuste está vinculada con la necesidad de no sobreestimar la precisión de las estimaciones de gasto e ingreso.

- En la imputación de gastos diarios se optó por una definición de donante basada en la idea de información completa, es decir, solo aquellas libretas con todos los días de registro fueron parte del *pool* de donantes. En el marco del estudio de imputaciones se hicieron pruebas con otras definiciones de donantes y los resultados no cambian de manera significativa, sin embargo, si el porcentaje de libretas rechazadas sigue aumentando en el futuro, es posible que la definición de información completa implique una disminución del conjunto de donantes. Esto implica reevaluar dicha definición, lo cual supone un estudio del efecto que ello tiene en el gasto promedio.

- En el caso de la Libreta de Gastos Individuales, un posible desafío para una nueva versión es testear otros métodos para imputar libretas rechazadas. Si bien el método *hot deck* en los ejercicios de simulación logra corregir de manera satisfactoria el sesgo en el promedio, no es posible asegurar que en el futuro esto siga ocurriendo, lo cual hace necesario explorar nuevos métodos. Asimismo, queda la puerta abierta para el desarrollo de una estrategia que combine más de un método al momento de imputar.

- En los ejercicios de simulación, para el caso de gastos diarios, fue posible observar con bastante detalle lo que ocurre en términos del gasto a nivel agregado. Respecto a la distribución del gasto no se desarrolló un ejercicio con la misma profundidad, lo cual sugiere la posibilidad de realizar nuevos estudios que permitan evaluar el desempeño de los métodos en lo relativo a la estructura del gasto.

- En el caso de la Libreta de Ingresos, el informante no necesariamente coincide con la persona respecto a la cual se solicita información, cuestión que no ocurre en la LGI (por ser un instrumento autoadministrado). Esto supone un desafío para la elaboración de modelos que expliquen el fenómeno de la no respuesta, ya que se hace necesario separar el efecto del informante de aquel relacionado con las características de las personas sobre las que efectivamente se solicita información.

- En el caso de ingresos, se hace relevante reflexionar acerca de la separación entre las distintas categorías analíticas (grupos de ingreso) que se establecieron para hacer la imputación, ya que probablemente el desarrollo del mercado del trabajo hará irrelevantes ciertas distinciones que hasta hace algunos años tenían sentido. Especialmente, vale la pena revisar la separación entre las personas a honorarios y asalariados. Hasta hace un par de años la no

obligatoriedad de cotizar de los primeros implicaba una diferencia importante entre ambas categorías, sin embargo, los cambios legales (principalmente enfocados en las cotizaciones previsionales y de salud) aplicados el último tiempo han generado que las personas que se desempeñan entregando boletas de honorarios se asemejen cada vez más a los asalariados y, por lo tanto, se debiese estudiar si esta separación tiene sentido en el mercado laboral actual. Es necesario también revisar si la separación entre trabajadores por cuenta propia y profesionales independientes se mantiene vigente. Esta distinción se basa en la consideración de dos tipos de independientes: aquellos que cuentan con su propio negocio y del cual pueden hacer retiros, y el resto de los independientes, que principalmente se dedica al área de servicios. Hoy en día, el mercado de las pequeñas empresas, emprendimientos y comercio vía internet ha crecido de forma exponencial; es posible encontrarnos con infinidad de nuevos perfiles de trabajadores independientes(INE Chile, 2018b), por lo que vale la pena investigar esta área del mercado laboral para futuros ejercicios.

- Para la Libreta de Ingresos solo se ha explorado la imputación de la no respuesta parcial y particularmente de los ingresos laborales y de jubilación. La selección de los grupos de ingresos a imputar se debe principalmente a que representan el mayor porcentaje del ingreso del hogar, pero es importante evaluar si es relevante y posible realizar imputaciones de otros tipos de ingresos, para contar con estimaciones del ingreso del hogar más precisas. Asimismo, vale la pena el estudio de la no respuesta total y las oportunidades existentes para la imputación del ingreso del hogar en aquellos casos.

6 Referencias

ABS. (2017). *Household Expenditure Survey, Australia: Summary of Results, 2015-16*. Statistics Australian Bureau. Recuperado de http://abs.gov.au/ausstats/abs@.nsf/PrintAllPreparePage?

Andridge, R., & Little, R. (2009). The Use of Sample Weights in Hot Deck Imputation. *Journal of official statistics*, *25*(1), 21-36. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117228/

Andridge, R., & Little, R. (2010). *A Review of Hot Deck Imputation for Survey Non-response*.

Bethlehem, J. (2012). Using reponse probabilities for assesing representativity. Statistics Netherlands.

Bethlehem, J., Cobben, F., & Schouten, B. (2008). Indicators for the Representativeness of Survey Response. En. Canada.

Bulman, J., & Carrel, O. (2017). *Living Costs and Food Survey technical report for survey year April 2015 to March 2016*. Office for National Statistics UK.

Cobben, F. (2009). *Nonresponse in sample surveys, Methods for Analysis and Adjustment*. Statistics Netherlands.

DOL. (2011). *Consumer Expenditure Survey Anthology*. U.S. Bureau of Labor Statistics.

Eurostat. (2014). *European Statistical System Handbook for Quality Reports*. Recuperado de http: //ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eustat. (2008). *Estudio y ajuste de la no respuesta en las encuestas a hogares*.

Gómez, J., Palarea, J., & Martín, J. (2006). Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. *Estadística Española*, *48*(162), 241 a 270.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Groves, R., & Couper, M. (1998). *Nonresponse in Household Interview Surveys*. Wiley.

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167-189. https://doi.org/10.1093/poq/nfn011

Guerrero, M. Á. (2003). *Método PRINCALS para la clasificación socioeconómica del CENSO 2002*.

Heckman, J. J. (1979). *Sample selection bias as specification error*.

IBGE. (2011). *Pesquisa de Orçamentos Familiares 2008-2009: Análise do Consumo Alimentar Pessoal no Brasil*. Instituto Brasileiro de Geografi a e Estatística.

INE Chile. (2014). *Métodos de imputación VII EPF: Gastos diarios e ingresos de la actividad principal y jubilaciones*.

INE Chile. (2018a). *Informe de Calidad VIII EPF*. Recuperado de http://www.ine.cl/docs/default-source/ingresos-y-gastos/epf/viii-epf/documentacion/informe_de_calidad_viii_epf.pdf?sfvrsn=6

INE Chile. (2018b). Informe de resultados: El microemprendimiento en Chile: Quinta encuesta de microemprendimiento. Recuperado de https: //www.economia.gob.cl/wp-content/uploads/2018/02/Newsletter-Microemprendedor-EME5.pdf

INE España. (2016). *Encuesta de Presupuestos Familiares Metodología año 2016*. Instituto Nacional de Estadística.

INSEE. (2014). *Sources et méthodes Enquête Budget de famille 2011*. Institut national de la statistique et des études économiques.

Jiménez, M., Ramírez, M. de la L., & Pizarro, M. (2008). Ciclo Vital de la Familia. Transformaciones en la estructura familiar en Chile, Casen 1990-2006. Ministerio de Planifación.

Laperrière, C. (2015). *Data Collection Using a Diary: The Experience of the Survey of Household Spending*. Statistical Society of Canada Annual Meeting.

Medina, F., & Galván, M. (2007). *Imputación de datos: teoría y práctica*. CEPAL.

Mincer, J. (1974). *Schooling, experience, and earnings*. New York: National Bureau of Economic Research; distributed by Columbia University Press.

Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *University of Nebraska, Sociology Department, Faculty Publications*, *19*.

Righi, P., Falorsi, S., & Fasulo, A. (2014). Methods for variance estimation under random hot deck imputation in business surveys. *RIVISTA DI STATISTICA UFFICIALE*, *1*(2), 45-64.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355. https: //doi.org/10.2307/1882010

Sukasih, A., Jang, D., Vartivarian, S., Cohen, S., & Zhang, F. (2009). A simulation Study to Compare Weighting Methods for Nonresponses in the National Survey of Recent College Graduates. *Survey Research Methods*. Recuperado de file:///C:/Users/klehmann/Downloads/simulationstudy%20(1).pdf

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer.

Weiss, A. (1995). Human Capital vs. Signalling Explanations of Wages. *Journal of Economic Perspectives*, *9*(4), 133-154. https://doi.org/10.1257/jep.9.4.133

West, B. (2009). A Simulation Study of Alternative Weighting Class Adjustments for Nonresponse when Estimating a Population Mean from Complex Sample Survey Data. *Survey Research Methods*. Recuperado de https://pdfs.semanticscholar.org/4b4c/73f53e064a5902527a6dc64f479eb0df6d43.pdf

7 Anexos

Cuadro 25: Matriz de transferencia para la imputación de gastos diarios (parte 1)

Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5	Nivel 6
Submuestra	Submuestra	Submuestra	Submuestra	Submuestra	Submuestra
CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad
Edad	Edad	Edad	Edad	Edad	Edad
CIUO	CIUO	CIUO	CIUO	CIUO	CIUO
CISE	CISE	CISE	CISE	CISE	CISE
Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos
Comuna	Comuna	Comuna	Comuna	Comuna	Comuna
Manzana	Manzana	Manzana	Manzana	Manzana	
Folio	Folio	Folio	Folio		
N personas en el hogar	N personas en el hogar	N personas en el hogar			
Identificador persona					

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 26: Matriz de transferencia para la imputación de gastos diarios (parte 2)

Nivel 7	Nivel 8	Nivel 9	Nivel 10	Nivel 11	Nivel 12
Submuestra	Mes	Mes	Mes	Mes	Temporada1
CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Nivel educacional	Nivel educacional
Edad	Edad	Edad	Edad en tramos1	Edad en tramos1	Edad en tramos1
CIUO	CIUO	CIUO	CIUO	CIUO	CIUO
CISE	CISE	CISE	CISE	CISE	CISE
Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos
Región	Región	Macrozona	Macrozona	Macrozona	Macrozona

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 27: Matriz de transferencia para la imputación de gastos diarios (parte 3)

Nivel 13	Nivel 14	Nivel 15	Nivel 16	Nivel 17	Nivel 18
Temporada2					
CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar
Nivel educacional	Nivel educacional	Nivel educacional	Nivel educacional	Nivel educacional	Nivel educacional
Edad en tramos1	Edad en tramos1	Edad en tramos2	Edad en tramos2	Edad en tramos2	Edad en tramos2
CIUO	CIUO				
CISE	CISE	CISE	Ocupado		
Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos
Macrozona	Macrozona	Macrozona	Macrozona	Macrozona	

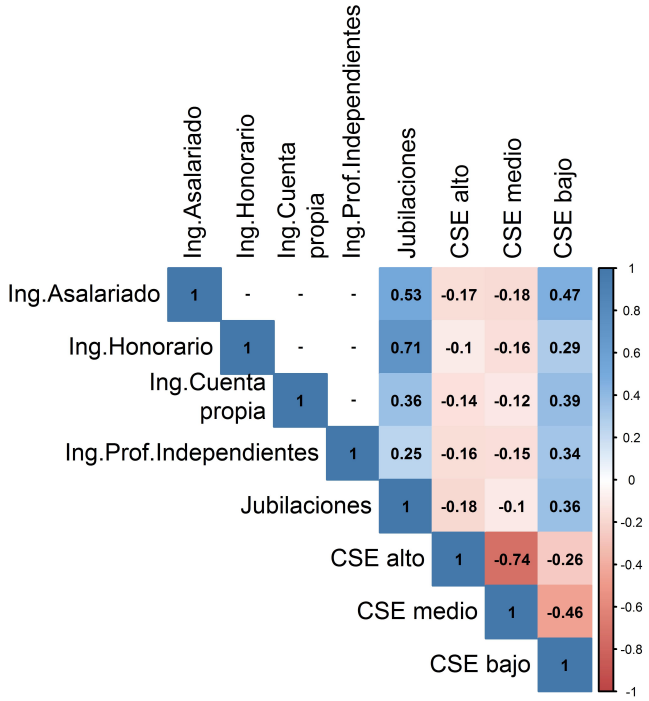
Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 28: Matriz de transferencia para la imputación de gastos diarios (parte 3)

Nivel 19	Nivel 20	Nivel 21	Nivel 22	Nivel 23
Sexo	Sexo	Sexo	Sexo	
Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar	Ingreso hogar
Nivel educacional	Nivel educacional			
Edad en tramos2		Edad en tramos2		
Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos	Administrador de gastos

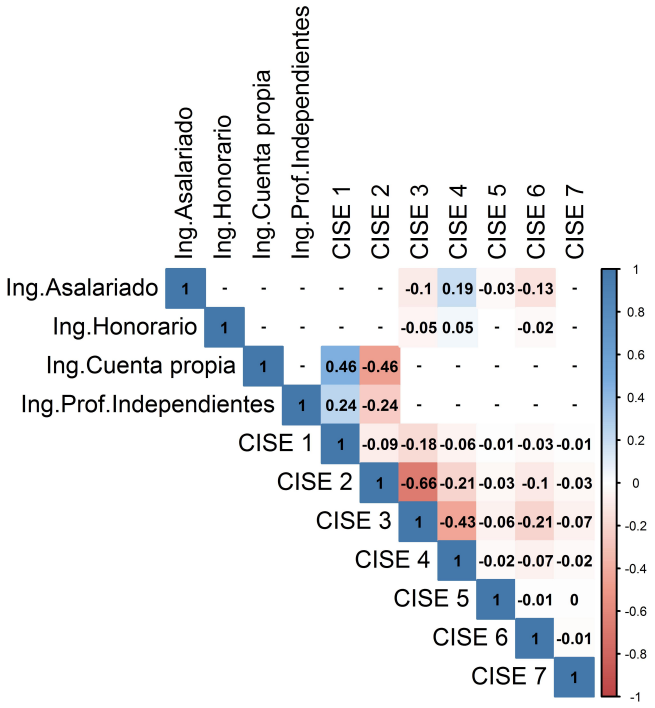
Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Figura 47: Correlación entre ingresos del trabajo y jubilaciones, y CSE



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Figura 48: Correlación entre ingresos del trabajo y CISE



Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

CISE 1: Patrón o empleador

CISE 2: Trabajador independiente o por cuenta propia

CISE 3: Asalariado del sector privado

CISE 4: Asalariado del sector público

CISE 5: Servicio doméstico puertas afuera

CISE 6: Servicio doméstico puertas adentro

CISE 7: Familiar o personal no remunerado

Cuadro 29: Niveles en los que se realizaron las imputaciones de honorarios

Nivel de imputación	Tamaño promedio cluster	Número de observaciones	Porcentaje de libretas
4	1,03	1	0,97
6	1,27	12	11,65
7	1,41	6	5,83
8	1,76	15	14,56
9	2,33	15	14,56
10	2,47	5	4,85
11	5,53	20	19,42
13	1,41	1	0,97
15	2,43	2	1,94
16	3,51	2	1,94
18	9,35	2	1,94
19	1,43	2	1,94
20	3,50	6	5,83
21	5,52	2	1,94
22	10,00	3	2,91
23	20,69	4	3,88
26	5,98	3	2,91
38	22,46	1	0,97
43	132,40	1	0,97

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 30: Niveles en los que se realizaron las imputaciones de cuenta propia

Nivel de imputación	Tamaño promedio cluster	Número de observaciones	Porcentaje de libretas
3	1,05	9	3,72
5	1,07	3	1,24
6	1,73	59	24,38
7	1,92	11	4,55
8	2,59	23	9,50
9	3,37	26	10,74
10	3,49	3	1,24
12	4,73	36	14,88
14	2,06	3	1,24
15	2,43	1	0,41
16	3,74	2	0,83
17	5,06	8	3,31
18	5,26	1	0,41
20	7,41	3	1,24
22	2,82	4	1,65
24	5,90	2	0,83
25	8,32	7	2,89
26	8,64	1	0,41
28	12,46	6	2,48
30	2,98	6	2,48
31	3,71	1	0,41
32	6,37	1	0,41
33	9,03	4	1,65
35	13,70	2	0,83
36	2,18	4	1,65
37	11,77	9	3,72
38	15,75	2	0,83
39	30,64	1	0,41
42	3,32	1	0,41
43	21,73	1	0,41
48	115,72	1	0,41
51	243,40	1	0,41

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 31: Niveles en los que se realizaron las imputaciones de profesionales independientes

Nivel de imputación	Tamaño promedio cluster	Número de observaciones	Porcentaje de libretas
2	1,01	2	1,11
6	1,28	21	11,67
7	1,37	12	6,67
8	1,60	18	10,00
10	2,20	23	12,78
11	3,14	26	14,44
12	3,57	1	0,56
13	1,05	2	1,11
14	1,37	1	0,56
15	1,52	1	0,56
16	2,01	2	1,11
18	3,08	2	1,11
19	4,77	13	7,22
25	3,30	2	1,11
26	5,24	1	0,56
28	1,10	1	0,56
29	1,66	7	3,89
31	2,94	2	1,11
32	5,04	1	0,56
33	8,36	4	2,22
35	1,47	10	5,56
36	4,71	15	8,33
37	6,32	3	1,67
38	12,10	2	1,11
39	23,46	1	0,56
42	8,13	1	0,56
45	45,18	1	0,56
47	42,43	2	1,11
50	89,35	1	0,56
53	173,75	2	1,11

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 32: Niveles en los que se realizaron las imputaciones de jubilados

Nivel de imputación	Tamaño promedio cluster	Número de observaciones	Porcentaje de libretas
2	1,02	1	0,54
3	1,03	3	1,62
4	1,36	38	20,54
5	4,50	78	42,16
6	5,07	7	3,78
7	6,25	11	5,95
8	1,25	24	12,97
9	1,39	2	1,08
10	1,69	1	0,54
11	6,08	7	3,78
12	1,03	2	1,08
13	21,44	2	1,08
15	3,04	2	1,08
16	16,03	3	1,62
18	61,62	1	0,54
20	54,46	3	1,62

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 33: Matriz de transferencia para la imputación de ingresos asalariados (parte 1)

Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5	Nivel 6	Nivel 7
CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad
CSE Sustentador principal	CSE Sustentador principal	CSE Sustentador principal	CSE Sustentador principal	CSE Sustentador principal	CSE Sustentador principal	CSE Sustentador principal
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
CISE	CISE	CISE	CISE	CISE	CISE	CISE
Edad	Edad	Edad	Edad	Edad	Tramo edad 1	Tramo edad 2
Manzana	Comuna	Region	Macrozona	Región Metropolitana	Región Metropolitana	Región Metropolitana
Presencia de menores de 6 años	Presencia de menores de 6 años	Presencia de menores de 6 años	Presencia de menores de 6 años	Presencia de menores de 6 años	Presencia de menores de 6 años	Presencia de menores de 6 años

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 34: Matriz de transferencia para la imputación de ingresos asalariados (parte 2)

Nivel 8	Nivel 9	Nivel 10	Nivel 11	Nivel 12	Nivel 13	Nivel 14
CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE Tramo edad 3 Región Metropolitana Presencia de menores de 6 años	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana Presencia de menores de 6 años	CIUO dos dígitos Rango escolaridad 1 CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana Presencia de menores de 6 años	CIUO dos dígitos Rango escolaridad 2 CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana Presencia de menores de 6 años	CIUO un dígito Rango escolaridad 2 CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana Presencia de menores de 6 años	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE Edad Región Metropolitana	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE Tramo edad 1 Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 35: Matriz de transferencia para la imputación de ingresos asalariados (parte 3)

Nivel 15	Nivel 16	Nivel 18	Nivel 20	Nivel 21	Nivel 22	Nivel 23
CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE Tramo edad 2 Región Metropolitana	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo CISE Tramo edad 3 Región Metropolitana	CIUO dos dígitos Rango escolaridad 1 CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana	CIUO un dígito Rango escolaridad 2 CSE Sustentador principal Sexo CISE agrupado Tramo edad 3 Región Metropolitana	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo Edad Región Metropolitana	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo Tramo edad 1 Región Metropolitana	CIUO dos dígitos Escolaridad CSE Sustentador principal Sexo Tramo edad 2 Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 36: Matriz de transferencia para la imputación de ingresos asalariados (parte 4)

Nivel 24	Nivel 25	Nivel 27	Nivel 29	Nivel 30	Nivel 31	Nivel 32
CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos
Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Sustentador principal	Sustentador principal	Sustentador principal	Sexo	Sexo	Sexo	Sexo
Sexo	Sexo	Sexo	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3
Tramo edad 3	Tramo edad 3	Tramo edad 3				
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 37: Matriz de transferencia para la imputación de ingresos asalariados (parte 5)

Nivel 34	Nivel 35	Nivel 36	Nivel 37	Nivel 38	Nivel 39	Nivel 47
CIUO un dígito						
Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1	Escolaridad
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Tramo edad 3	Edad	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3	
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 38: Matriz de transferencia para la imputación de ingresos honorarios (parte 1)

Nivel 4	Nivel 6	Nivel 7	Nivel 8	Nivel 9	Nivel 10	Nivel 11
CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito
Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Rango escolaridad 2
Edad	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3	Tramo edad 3	Tramo edad 3
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Macrozona	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 39: Matriz de transferencia para la imputación de ingresos honorarios (parte 2)

Nivel 13	Nivel 15	Nivel 16	Nivel 18	Nivel 19	Nivel 20	Nivel 21
CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito			
Escolaridad	Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad
Tramo edad 1	Tramo edad 3	Tramo edad 3	Tramo edad 3	Edad	Tramo edad 1	Tramo edad 2
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 40: Matriz de transferencia para la imputación de ingresos honorarios (parte 3)

Nivel 22	Nivel 23	Nivel 26	Nivel 38	Nivel 43
Escolaridad	Rango escolaridad 1	Escolaridad	Escolaridad	Escolaridad
Tramo edad 3	Tramo edad 3	Tramo edad 1	Tramo edad 1	
CSE	CSE	CSE		
Sexo	Sexo	Sexo		
Región Metropolitana	Región Metropolitana			

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 41: Matriz de transferencia para la imputación de ingresos por cuenta propia (parte 1)

Nivel 3	Nivel 5	Nivel 6	Nivel 7	Nivel 8	Nivel 9	Nivel 10
CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos
CISE	CISE	CISE	CISE	CISE	CISE	CISE
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1	Rango escolaridad 2
Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Edad	Edad	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3	Tramo edad 3
Región	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana
Presencia de menores de 15 años	Presencia de menores de 15 años	Presencia de menores de 15 años	Presencia de menores de 15 años	Presencia de menores de 15 años	Presencia de menores de 15 años	Presencia de menores de 15 años

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 42: Matriz de transferencia para la imputación de ingresos por cuenta propia (parte 2)

Nivel 12	Nivel 14	Nivel 15	Nivel 16	Nivel 17	Nivel 18	Nivel 20
CIUO un dígito	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito
CISE agrupado	CISE	CISE	CISE	CISE	CISE	CISE agrupado
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Rango escolaridad 2
Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Tramo edad 3	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3	Tramo edad 3	Tramo edad 3
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana
Presencia de menores de 15 años						

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 43: Matriz de transferencia para la imputación de ingresos por cuenta propia (parte 3)

Nivel 22	Nivel 24	Nivel 25	Nivel 26	Nivel 28	Nivel 30	Nivel 31
CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito	CIUO dos dígitos	CIUO dos dígitos
CISE	CISE	CISE	CISE	CISE agrupado		
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Escolaridad	Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Rango escolaridad 2	Escolaridad	Escolaridad
CSE	CSE	CSE	CSE	CSE	CSE	CSE
Tramo edad 1	Tramo edad 3	Tramo edad 3	Tramo edad 3	Tramo edad 3	Tramo edad 1	Tramo edad 2
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 44: Matriz de transferencia para la imputación de ingresos por cuenta propia (parte 4)

Nivel 32	Nivel 33	Nivel 35	Nivel 36	Nivel 37	Nivel 38
CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito			
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Escolaridad	Rango escolaridad 1	Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad
CSE	CSE	CSE	CSE	CSE	CSE
Tramo edad 3	Tramo edad 3	Tramo edad 3	Edad	Tramo edad 1	Tramo edad 2
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 45: Matriz de transferencia para la imputación de ingresos por cuenta propia (parte 5)

Nivel 39	Nivel 42	Nivel 43	Nivel 48	Nivel 51
Sexo Escolaridad CSE Tramo edad 3 Región Metropolitana	Sexo Escolaridad CSE Edad	Sexo Escolaridad CSE Tramo edad 1	Sexo Escolaridad CSE	Sexo Escolaridad

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 46: Matriz de transferencia para la imputación de ingresos de profesionales independientes (parte 1)

Nivel 2	Nivel 6	Nivel 7	Nivel 8	Nivel 10	Nivel 11	Nivel 12
Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Presencia de menores de 15 años Edad Comuna	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Presencia de menores de 15 años Tramo edad 1 Región Metropolitana	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Presencia de menores de 15 años Tramo edad 2 Región Metropolitana	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Presencia de menores de 15 años Tramo edad 3 Región Metropolitana	Escolaridad CIUO un dígito Sexo CSE Sustentador principal CISE agrupado Presencia de menores de 15 años Tramo edad 3 Región Metropolitana	Rango escolaridad 1 CIUO un dígito Sexo CSE Sustentador principal CISE agrupado Presencia de menores de 15 años Tramo edad 3 Región Metropolitana	Rango escolaridad 2 CIUO un dígito Sexo CSE Sustentador principal CISE agrupado Presencia de menores de 15 años Tramo edad 3 Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 47: Matriz de transferencia para la imputación de ingresos de profesionales independientes (parte 2)

Nivel 13	Nivel 14	Nivel 15	Nivel 16	Nivel 18	Nivel 19	Nivel 25
Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Edad Región Metropolitana	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Tramo edad 1 Región Metropolitana	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Tramo edad 2 Región Metropolitana	Escolaridad CIUO dos dígitos Sexo CSE Sustentador principal CISE Tramo edad 3 Región Metropolitana	Escolaridad CIUO un dígito Sexo CSE Sustentador principal CISE agrupado Tramo edad 3 Región Metropolitana	Rango escolaridad 1 CIUO un dígito Sexo CSE Sustentador principal CISE agrupado Tramo edad 3 Región Metropolitana	Escolaridad CIUO un dígito Sexo CSE Sustentador principal Tramo edad 3 Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 48: Matriz de transferencia para la imputación de ingresos de profesionales independientes (parte 3)

Nivel 26	Nivel 28	Nivel 29	Nivel 31	Nivel 32	Nivel 33
Rango escolaridad 1	Escolaridad	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1
CIUO un dígito	CIUO dos dígitos	CIUO dos dígitos	CIUO dos dígitos	CIUO un dígito	CIUO un dígito
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
CSE	CSE	CSE	CSE	CSE	CSE
Sustentador principal					
Tramo edad 3	Edad	Tramo edad 1	Tramo edad 3	Tramo edad 3	Tramo edad 3
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 49: Matriz de transferencia para la imputación de ingresos de profesionales independientes (parte 4)

Nivel 35	Nivel 36	Nivel 37	Nivel 38	Nivel 39
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Rango escolaridad 1
Sexo	Sexo	Sexo	Sexo	Sexo
CSE	CSE	CSE	CSE	CSE
Edad	Tramo edad 1	Tramo edad 2	Tramo edad 3	Tramo edad 3
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 50: Matriz de transferencia para la imputación de ingresos de profesionales independientes (parte 5)

Nivel 42	Nivel 45	Nivel 47	Nivel 50	Nivel 53
Escolaridad	Rango escolaridad 1	Escolaridad	Escolaridad	Escolaridad
Sexo	Sexo	Sexo	Sexo	
CSE	CSE	CSE		
Edad	Tramo edad 3			

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 51: Matriz de transferencia para la imputación de ingresos de jubilaciones (parte 1)

Nivel 3	Nivel 4	Nivel 5	Nivel 6	Nivel 7
CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Escolaridad
Sistema previsional	Sistema previsional	Sistema previsional	Sistema previsional	Sistema previsional
Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal	Sustentador principal
Manzana	Comuna	Región	Macrozona	Región Metropolitana
Tramo edad 2	Tramo edad 2	Tramo edad 2	Tramo edad 2	Tramo edad 2

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 52: Matriz de transferencia para la imputación de ingresos de jubilaciones (parte 2)

Nivel 8	Nivel 9	Nivel 10	Nivel 11	Nivel 12
CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo
Rango escolaridad 1	Rango escolaridad 2	Escolaridad	Escolaridad	Escolaridad
Sistema previsional	Sistema previsional	Sistema previsional	Sistema previsional	Sistema previsional
Sustentador principal	Sustentador principal			
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana
Tramo edad 2	Tramo edad 2	Edad	Tramo edad 1	Tramo edad 2

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)

Cuadro 53: Matriz de transferencia para la imputación de ingresos de jubilaciones (parte 3)

Nivel 13	Nivel 15	Nivel 16	Nivel 18	Nivel 20
CSE	CSE	CSE	CSE	CSE
Sexo	Sexo	Sexo	Sexo	Sexo
Rango escolaridad 1	Escolaridad	Escolaridad	Rango escolaridad 1	Escolaridad
Sistema previsional				
Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana	Región Metropolitana
Tramo edad 2	Edad	Tramo edad 1	Tramo edad 2	

Fuente: Instituto Nacional de Estadísticas (INE) - VIII Encuesta de Presupuestos Familiares (EPF)