Metodología de la ECH no presencial

2020

INSTITUTO NACIONAL DE ESTADÍSTICA

Diego Aboal

Director Técnico

Federico Segui

Subdirector General

Elaboró este documento

Juan Pablo Ferreira

Muestrista

Índice general

1. Introducción

Como consecuencia de la declaración de la emergencia sanitaria producto de la aparición del primer caso de Covid-19, el INE suspendió el relevamiento en campo de la Encuesta Continua de Hogares a partir del 15 de marzo. Hasta ese momento, la ECH era de forma presencial, en donde los hogares que componían la muestra para cada uno de los meses, eran seleccionados de forma independiente. Es decir, la ECH hasta ese momento, era una encuesta de tipo cross-section[1]. Con el objetivo de poder seguir brindando los indicadores mensuales que produce el INE en base a la ECH, los cuales, se limitan únicamente a aquellos relacionados con el mercado de trabajo (actividad, empleo y desempleo) e ingresos de los hogares, es que se buscaron estrategias para continuar con la producción de estimaciones de calidad para los indicadores mencionados anteriormente, pero bajo una nueva modalidad de relevamiento no presencial.

Independientemente de la emergencia sanitaria, las encuestas a hogares y personas pueden ser llevadas a cabo tanto de forma presencial como de forma no presencial (telefónica y/o vía formulario web) siempre y cuando el marco de muestreo refleje lo mejor posible la población objetivo, es decir, no sufra de problemas de subcobertura y que la muestra resultante no contenga sesgos producto de la no respuesta.

La ECH ha sido históricamente llevada a cabo de forma presencial por dos motivos: i) es una encuesta multipropósito que releva varias dimensiones tanto de la vivienda, como del hogar y sus integrantes, lo que hacía inviable, por su extensión, el relevamiento de forma telefónica y ii) no existía ni existe en la actualidad un marco de muestreo de hogares con teléfonos, que permita hacer una encuesta cross section de forma no presencial que brinde estimación insesgadas.

La realización de la ECH de forma no presencial sólo podía llevarse a cabo si se cumplían al menos las siguientes condiciones:

1. Reducción del formulario actual de la ECH para acotar el tiempo de la entrevista.
2. Contar con al menos un subconjunto de la población (que sea un fiel reflejo de la misma) para poder seleccionar la muestra y que a su vez dicha subpoblación, contenga información de teléfonos de contacto, y características relevantes del hogar. Dichas características podrían ser un insumo imprescindible al momento de modelar la propensión de respuesta y así poder reducir posibles sesgos ocasionados por la no respuesta, la cual tiende a hacer más alta en relevamientos no presenciales.

El primer punto fue resuelto teniendo en cuenta que para la producción de las estadísticas mensuales del INE únicamente era necesario relevar los módulos de mercado laboral e ingresos y otras variables de control sobre la estructura del hogar (entradas y salidas de los integrantes) que releva de forma tradicional la ECH. El segundo punto fue resuelto eligiendo una submuestra utilizando como “marco de muestreo” un conjunto (pool) de las muestras efectivas de la ECH en los meses previos a la emergencia sanitaria.

2. ECH no presencial con panel rotativo

Una vez resuelta la reducción del formulario y el marco para seleccionar la muestra de la ECH no presencial, se decidió también abandonar las clásicas muestras cross-section y pasar a realizar una encuesta de tipo panel, las cuales son más eficientes (menores errores estándar) para las estimaciones de los cambios de los distintos indicadores de un periodo a

[1] En cada período se selecciona una muestra independiente y la inclusión de un hogar en la muestra en un período en particular no se ve influido por las muestras seleccionadas en períodos anteriores.

otro, en comparación con las encuestas cross-section. Aquí es necesario distinguir los distintos tipos de paneles, los cuales se pueden separar a grandes rasgos, en paneles puros y paneles rotativos.

En los *paneles puros*, la muestra de hogares es extraída por única vez al inicio del estudio y luego todos los hogares seleccionados serán entrevistados, a lo largo de la duración del panel. El principal problema de los paneles puros radica en que la no respuesta es acumulativa (atrición). La pérdida de hogares en la muestra se debe a varios factores como ser: mudanzas, migración, muerte y, sobre todo, el cansancio del hogar a responder a la encuesta en las sucesivas olas. En un panel puro, si no se renueva la muestra periódicamente (sorteando nacimientos), se produce el efecto cohorte, el cual tiene como implicancia que la muestra refleja a la población original pero no a la población en las sucesivas olas. En el caso de las encuestas continuas, los hogares deberían estar en la muestra de forma indefinida. Pero en la práctica, un límite de tiempo de permanencia es fijado.

Un *panel rotativo* es un compromiso entre un panel puro y muestras cross-section. Los paneles rotativos intentan captar parte de la reducción de los errores estándar de los paneles puros (producto de la correlación de los hogares en la muestra), mientras se reduce la carga excesiva a los hogares. Son utilizados frecuentemente cuando los objetivos principales son obtener estimaciones transversales (nivel) y estimaciones a corto plazo del cambio neto y bruto, por ejemplo, el cambio en la tasa de desempleo o en el ingreso promedio de los hogares de un mes a otro. La muestra en un período se compone por una serie de paneles o grupos de rotación (GR), en donde se debe definir un patrón de rotación para determinar las entradas y salidas de los hogares en la muestra. Existe una amplia gama de patrones de rotación y la elección final debe ser un compromiso entre: i) el mayor solapamiento posible para los periodos en donde los cambios son de mayor importancia y ii) preservar las tasas de respuesta y exigencias de las cargas a los hogares.

Teniendo en cuenta lo anterior y sumado al desconocimiento de la duración de la crisis sanitaria producto del Covid-19 se optó, para la ECH no presencial, un sistema de rotación (+3,-1), es decir, un hogar permanece tres meses en la muestra, para luego abandonar el panel de forma definitiva.

3. Tamaños de muestra mensuales y puesta en funcionamiento

El tamaño de muestra esperado por mes asciende a 4.500 casos aproximadamente y se compone por tres GR representativos de la población. En cada mes, un GR es reemplazado por otro nuevo GR una vez completado su permanencia de tres meses en la muestra. Esto da como resultado un solapamiento de 2/3 de un mes a otro, por lo cual, el diseño es eficiente para estimar los cambios de los principales indicadores de la ECH de un mes a otro. La rotación de la muestra luego de seis meses previene la carga excesiva a los hogares y minimiza la no respuesta.

Se debe tener en cuenta que el tamaño de muestra esperado para el mes de marzo (comienzo del panel) se situó en 3400 hogares. Esta reducción en el tamaño de muestra se debió al retraso en el comienzo del campo producto de los cambios necesarios para poder llevar a cabo esta nueva modalidad de relevamiento, mientras se persigue tener estimaciones oportunas en el tiempo. En los meses siguientes los tamaños de muestra esperados se situaron dentro del tamaño esperado fijado previamente (4500 aproximadamente).

En las siguientes secciones se presenta: el diseño de la muestra, la metodología del cálculo de los pesos muestrales y el método elegido para el cómputo de las estimaciones de los errores estándar. Los resultados obtenidos siguiendo la metodología que se detalla a continuación corresponden al primer mes de implementación del panel, es decir, marzo de

2020. En los meses siguientes se presentará junto con los boletines de mercado de trabajo e ingresos, un breve resumen de los resultados obtenidos sobre todo en lo que se refiere al modelado de la propensión de la no respuesta así como para las tasas de respuesta obtenidas a nivel de los estratos de diseño entre otras métricas, como ser, la atrición del panel.

4. Diseño muestral ECH no presencial

El diseño muestra de la *ECH no presencial* es aleatorio y en dos fases de selección.

La primera fase corresponde a la muestra de la ECH presencial de los meses de diciembre de 2019 a febrero de 2020, la cual, de aquí en más se denota como $S_{(1)}$. Los hogares (y las personas que lo componen) incluidos en $S_{(1)}$ son seleccionados bajo un diseño estratificado, aleatorio y en dos etapas de selección. En la primera etapa se seleccionan unidades primarias de muestreo (UPM) que corresponden a conglomerados de zonas censales, bajo un diseño con probabilidades proporcional al tamaño (PPS) utilizando como medida de tamaño (MOS) la cantidad de viviendas particulares según el Censo 2011. En la segunda etapa, dentro de cada UPM, se seleccionan cinco viviendas con igual probabilidad de selección.

En la segunda fase se seleccionó una muestra $S_{(2)}$ bajo un muestreo aleatorio estratificado simple dentro de aquellos hogares respondentes a la ECH presencial para el pool definido anteriormente y que a su vez tenían al menos un teléfono de contacto[2]. Los estratos de diseño de la segunda fase se construyen en base a varios niveles de información. Para Montevideo y zona metropolitana se utilizaron los estratos socioeconómicos de la ECH presencial. Para el resto del país, los estratos se conformaron como la interacción entre seis regiones geográficas (agrupaciones de departamentos) y dos estratos de urbanicidad (localidades de 5000 habitantes o más y localidades de menos de 5000 habitantes y zona rural). Las regiones geográficas utilizadas fueron las siguientes:

- Interior Norte (Artigas, Rivera, Cerro Largo, Treinta y Tres)
- Costa Este (Canelones, Maldonado, Rocha)
- Litoral Norte (Salto, Paysandú, Río Negro)
- Litoral Sur (Soriano, Colonia, San José)
- Centro Norte (Tacuarembó, Durazno)
- Centro Sur (Flores, Florida, Lavalleja)

4.1. Tamaño de muestra y asignación por estrato de diseño

El tamaño de muestra esperado para **marzo** se situó en 3.400 hogares aproximadamente. Dicho tamaño efectivo permite obtener estimaciones para los distintos indicadores de la ECH a nivel total y para las distintas aperturas que se realizan habitualmente. Luego, el tamaño de muestra esperado fue incrementado aproximadamente un 25% teniendo en cuenta la tasa de respuesta esperada, situando el tamaño de muestra teórico en 4250 casos aproximadamente.

[2] Dentro de los hogares respondentes en la muestra de la primera fase, aproximadamente el 6% de los mismos no tenían teléfonos de contacto. Teniendo en cuenta que el relevamiento en la segunda es telefónico, esto implicaba que para estos hogares de la primera fase, su probabilidad de inclusión en la segunda fase de muestreo fuera cero y por ende violará lo principios de una muestra representativa (todas las unidades tienen chance de ser seleccionada). La exclusión de los mismos podía llegar a generar sesgos en las estimaciones si los mismos tuvieran comportamientos distintos respecto a los hogares con contacto telefónicos, sobre todo en las variables de interés (e.g. mercado laboral e ingresos). Por lo tanto previo la exclusión de los mismos se realizó un análisis para detectar posibles comportamientos distintos, el cual, no detectó diferencias significativas entre ambos subgrupos (con y sin teléfono).

Posteriormente el tamaño de muestra teórico fue asignado a nivel de estrato de forma proporcional en base al tamaño del estrato en la población, el cual, proviene de estimaciones de la ECH presencial. La muestra teórica se dividió de forma aleatoria en sub muestras o réplicas. Las réplicas se fueron utilizando hasta alcanzar los tamaños de muestra esperados por estrato. Debido a que las réplicas son construidas al azar el uso o no de alguna no le quita aleatoriedad a la muestra.

4.2. Tasa de respuesta

El tamaño de muestra efectivo en marzo se situó en 3.373 casos aproximadamente lo que se traduce en una tasa de respuesta de aproximadamente el 79%. Las tasas de respuesta obtenidas fueron similares en todos los estratos a excepción del estrato (Centro-Norte). Esto se debió a dificultades propias de la puesta en práctica de una nueva forma de relevamiento y no a características propias de los hogares incluidos en dichos estrato que los hicieran menos propensos a responder.

5. Determinación de los pesos finales

El primer paso para la construcción de los pesos de los hogares (y sus integrantes) respondentes consiste en la determinación de los pesos de la primera fase de muestreo (ECH presencial). El peso final del hogar i en la primera fase viene dado por:

$$w_{(1)i} = d_{(1)i} \times \mathrm{nr}_{(1)i} \times g_{(1)i},$$

donde

$d_{(1),i}$ es el ponderador original del hogar i el cual es computado como el inverso de la probabilidad de selección en la primera fase siguiendo el diseño muestral de la ECH presencial (estratificado, por conglomerados y en dos etapas de selección).

$\mathrm{nr}_{(1)i}$ es el ajuste por no respuesta en la fase 1, el cual, se define como el inverso de la tasa de respuesta ponderada en el estrato de diseño de la primera fase.

$g_{(1)i}$ es el ajuste proveniente de la calibración utilizando el raking truncado exigiendo factores de ajuste iguales por hogar y utilizando como variables de control los conteos poblacionales provenientes de la proyecciones de población.

Una vez finalizado el relevamiento del panel telefónico (segunda fase), el **peso final** para el hogar i respondente viene determinado por:

$$w_{(2)i} = w_{(1)i} \times d_{(2)i|(1)} \times \hat{\phi}^{-1}_{(2)i|(1)} \times g_{(2)i|(1)},$$

donde

$d_{(2)i|(1)} = \pi^{-1}_{(2)ih|(1)}$ es el peso original de la fase 2 condicionado a la fase 1, el cual, viene determinado por el inverso de la inclusión del hogar en la fase 2 condicionado a que fue seleccionado (y respondió) en la fase 1. En este caso la probabilidad de inclusión en la fase 2 condicionada en la fase 1 del hogar i que pertenece al estrato h viene dada por $\pi_{(2)ih|(1)} = n_{(2)h}/n_{(1)h}$ donde $n_{(2)h}$ es el tamaño de muestra teórico en la fase 2 en el estrato h y $n_{(1)h}$ es la cantidad de hogares respondentes en la fase 1 que pertenecen al estrato de diseño h de la fase 2, el cual, es aleatorio.

$\hat{\phi}^{-1}_{(2)i|(1)}$ es el inverso de la propensión estimada del hogar i en la fase 2 responda a la encuesta panel (fase 2). La propensión de los hogares incluidos en la muestra teórica de la fase 2 se estimó (ver más adelante) utilizando información conocida de los hogares proveniente de la ECH presencial y por medio de algoritmos de aprendizaje automático, más

precisamente bosque aleatorios (random forest) y utilizando posteriormente clases de ajuste de propensión de no respuesta.

$g_{(2)i|(1)}$ es el factor de ajuste proveniente de la calibración utilizando al igual que en la calibración de los pesos de la fase 1 $w_{(1)}$ raking truncado exigiendo pesos iguales para cada uno de los integrantes del hogar y utilizando conteos provenientes de las proyecciones de población del INE.

A continuación se explica con mayor detalle los ajustes realizados por no respuesta y calibración a los pesos finales.

5.1. Ajuste por no respuesta

En este caso se modeló la probabilidad o propensión de los hogares de responder en base a un set de características conocidas tanto para los hogares que respondieron (R) como a los que no respondieron (NR). Para modelar la no respuesta se asume que el mecanismo es aleatorio (MAR), es decir, que la propensión de un hogar de responder la encuesta depende de variables x que son conocidas tanto para los hogares que respondieron (R) como a los que no respondieron (NR). Al ser la encuesta una submuestra de la ECH presencial se disponen una amplia gama de características tanto a nivel de la vivienda, hogar, del informante así como variables que pueden ayudar a maximizar la probabilidad de contacto por parte del INE (número de teléfonos disponibles).

Las características elegidas en un principio para modelar la propensión de responder del hogar son:

- Edad, sexo, nivel educativo y condición de actividad del informante.
- Ingreso del hogar.
- Cantidad de ocupados, desocupados y activos en el hogar.
- Cantidad de hombres y mujeres en el hogar.
- Cantidad de menores de 14 años en el hogar
- Cantidad de mayores de 60 años en el hogar.
- Departamento.
- Vivienda ubicada en un asentamiento.
- Vivienda rural.
- Estrato de la segunda fase de muestreo.
- Cantidad de teléfonos disponibles en el hogar
- Ratio entre la cantidad de teléfonos disponibles y el número de integrantes del hogar.

La estimación de la propensión de un hogar a responder puede ser llevada a cabo por métodos clásicos como ser modelos logit o probit, o por medio de árboles de regresión. El problema que tiene los algoritmos mencionados anteriormente (sobretodo el árbol de regresión) es que tienden a presentar sobreajuste (overfitting). Esto implica que el algoritmo tiende a ajustarse muy bien a la muestra en particular, pero pierde la capacidad de generalizar, es decir, el algoritmo puede llegar a tener un buen ajuste para la muestra en particular, pero quizás no logre representar de forma correcta el *mecanismo de no respuesta subyacente*.

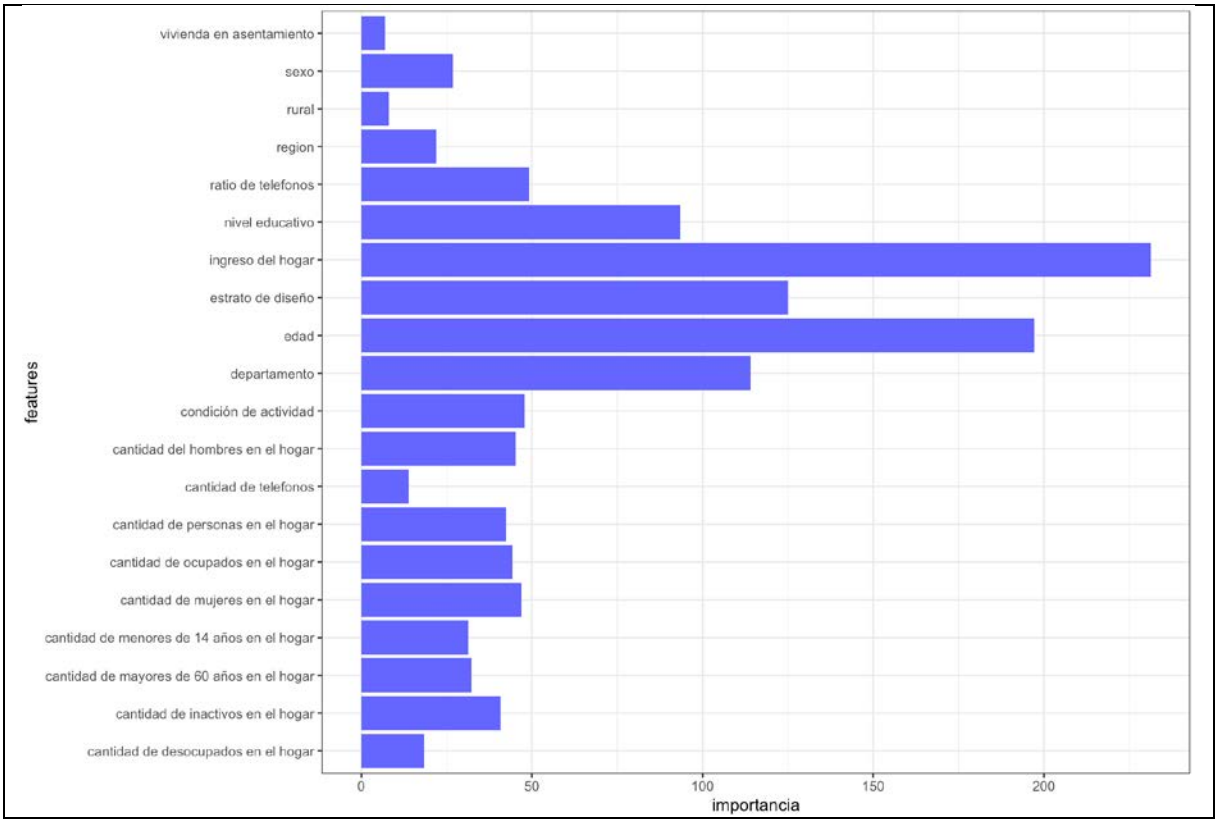
Es por lo anterior que para modelar la no respuesta se utilizó *random forest* el cual consiste en ajustar una gran cantidad de árboles de regresión simple (estimadores) y luego promediar los resultados de los mismos con el objetivo de obtener predicciones de las propensiones de responder que sean más robustas y con menor varianza. Para la construcción de cada uno de los árboles se selecciona una muestra bajo un muestreo aleatorio simple con reposición (muestras *bootstrap*) del mismo tamaño que la muestra original y se ajusta un árbol para cada una de las muestras bootstrap y luego se promedian las predicciones obtenidas en cada una de las muestras bootstrap. El problema del procedimiento anterior, es que puede existir una variable que sea muy predominantemente

para explicar la propensión a responder y por ende sea elegida siempre en cada uno de los árboles dando como resultado que todas las propensiones estimadas entre los diferentes árboles estén altamente correlacionadas. Para solucionar el problema anterior, dentro de cada una de las muestras bootstrap también se selecciona una submuestra al azar entre las características del hogar.

En la Figura 1 se presenta la importancia de las variables explicativas para estimar la propensión que tiene un hogar de responder en el mes de marzo

Como se puede apreciar en la figura 1, las variables más importantes son: el ingreso del hogar, la edad del informante, el estrato de diseño de la segunda fase, el departamento y nivel educativo del informante. Es importante aclarar que si bien el departamento y el estrato de diseño explican la propensión del hogar a responder no se deben a características propias del hogar si no a falencias en el relevamiento en donde los tamaños de muestra esperados para algunos estratos fueron menores a los esperados producto de la puesta en marcha del panel.

Figura 1 - Importancia de las variables para explicar la propensión de un hogar a responder

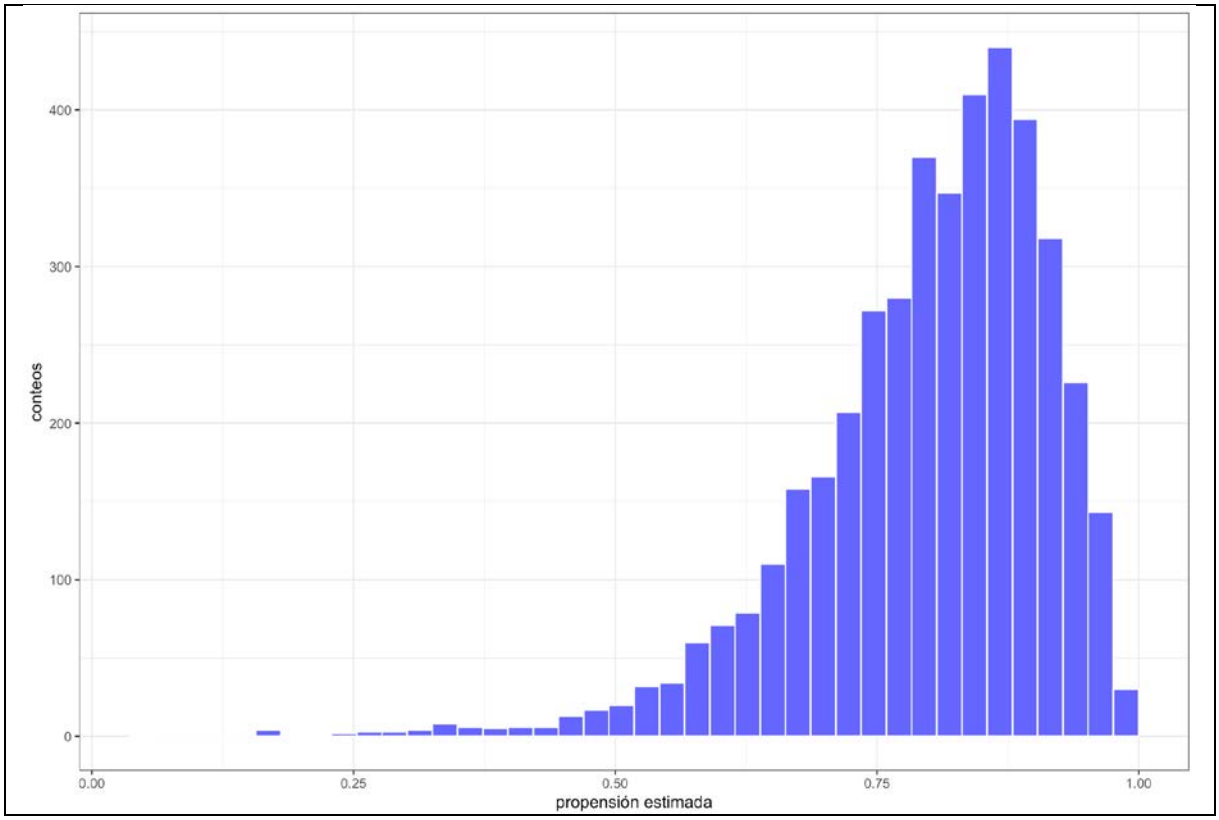


Fuente: INE, Encuesta Continua de Hogares no presencial.

Una vez implementando el algoritmo se estimaron las propensiones a responder la encuesta para todos los hogares de la muestra teórica de la segunda fase. En la figura 2 se aprecia la distribución de las propensiones estimadas $\hat{\phi}$

Como se puede apreciar en la figura 2 existen hogares cuya propensión estimada de responder es muy pequeña, esto implica que usar las propensiones a nivel de cada uno de los hogares (propensiones simples) como ajustes para la no respuesta si bien pueden reducir el sesgo ocasionado por la no respuesta, provocaría un aumento injustificado en los errores estándar (SE) de las estimaciones de los indicadores de interés, producto de un aumento excesivo en la variabilidad de los pesos final debido a la existencia de pesos extremos (aquellos con propensiones cercanas a cero).

Figura 2 - Distribución de las propensiones estimadas para cada hogar de la muestra



Fuente: INE, Encuesta Continua de Hogares no presencial.

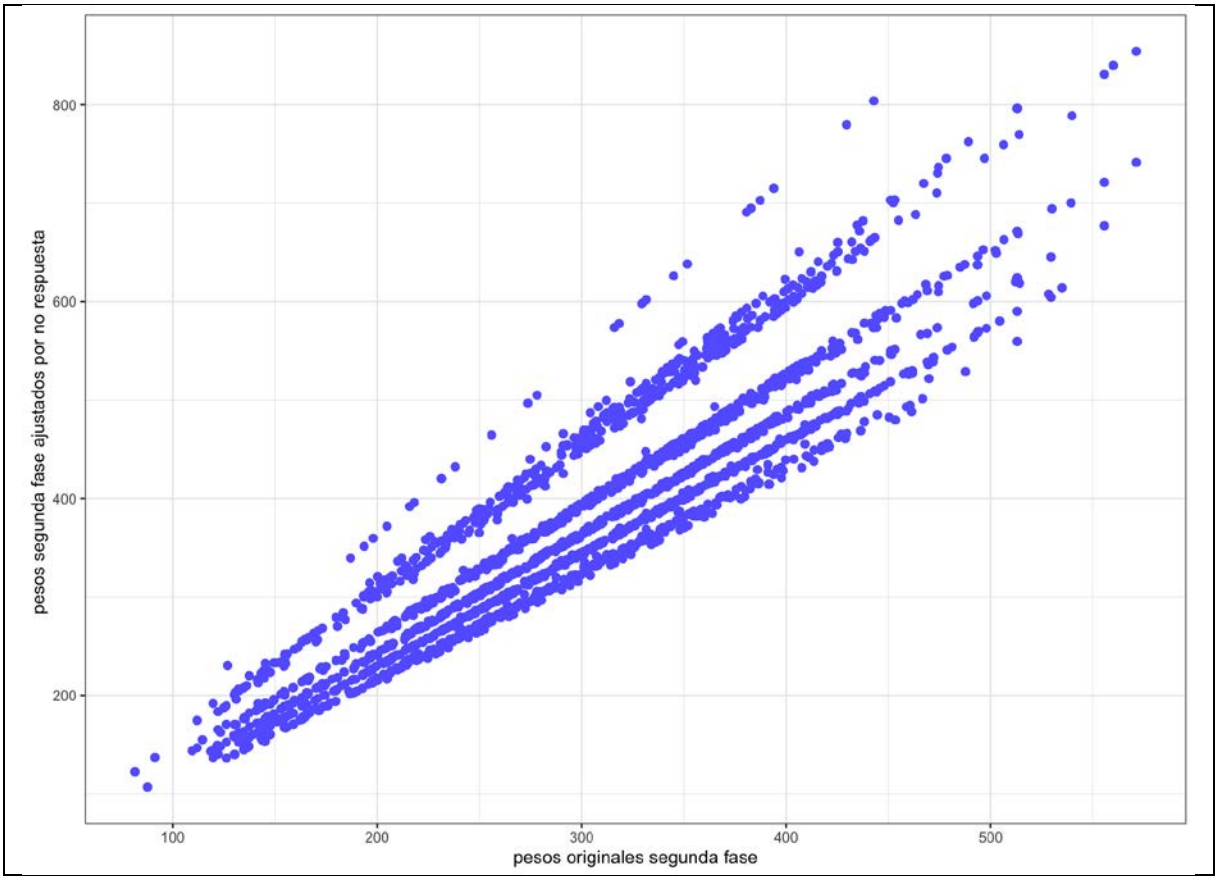
Teniendo en cuenta lo anterior el ajuste por no respuesta se realiza utilizando clases o post-estratos de no respuesta, los cuales son definidos en base a la distribución de las propensiones estimadas (a nivel de quintil) en combinación con los estratos de diseño de la muestra, en donde, se exigió que el post-estrato tuviera al menos 20 hogares (R+NR). Una vez conformado los post estratos, se realiza un ajuste único por no respuesta en base al inverso de la mediana de las propensiones estimadas dentro de cada post estrato.

El ajuste por no respuesta para el hogar respondente i que pertenece al post-estrato de no respuesta g viene dado como:

$$\hat{\phi}_{(2)i|(1)}^{-1} = \frac{1}{\text{mediana}[\phi_{(2)i|(1)}^{-1}]_{i \in s_{(2),g}}}$$

La elección del uso de la mediana de la propensiones estimadas dentro del post-estrato en vez del uso de la media (ponderada o no) se debe a que existían post-estratos en donde la distribución de dichas propensiones presentan una alta asimetría negativa. En la figura 3 se presentan los pesos originales de la segunda fase ($w_{(1)i} \times d_{(2)i|(1)}$) respecto a los pesos ajustados por no respuesta ($w_{(1)i} \times d_{(2)i|(1)} \times \hat{\phi}_{(2)i|(1)}^{-1}$).

Figura 3 - Pesos originales de la segunda fase respecto a los pesos ajustados por no respuesta



Fuente: INE, Encuesta Continua de Hogares no presencial.

5.2. Ajuste por calibración

Como último paso, los pesos fueron ajustados utilizando técnicas de calibración de forma que la muestra "expandida" coincida con información conocida de la estructura de la población. El objetivo principal del ajuste por calibración es: i) reducir los SEs de las estimaciones si las variables utilizadas para la calibración se encuentran correlacionadas con las variables de interés de la ECH, ii) reducción del posible sesgo en las estimaciones producido por la no respuesta y iii) brindar comparabilidad con estimaciones provenientes de otras fuentes.

Como ya se mencionó, la información utilizada para la calibración de los pesos de la ECH proviene de las proyecciones de población para el año 2020. Los conteos utilizados son los mismos que se han aplicado en la calibración de la ECH presencial y se describen a continuación:

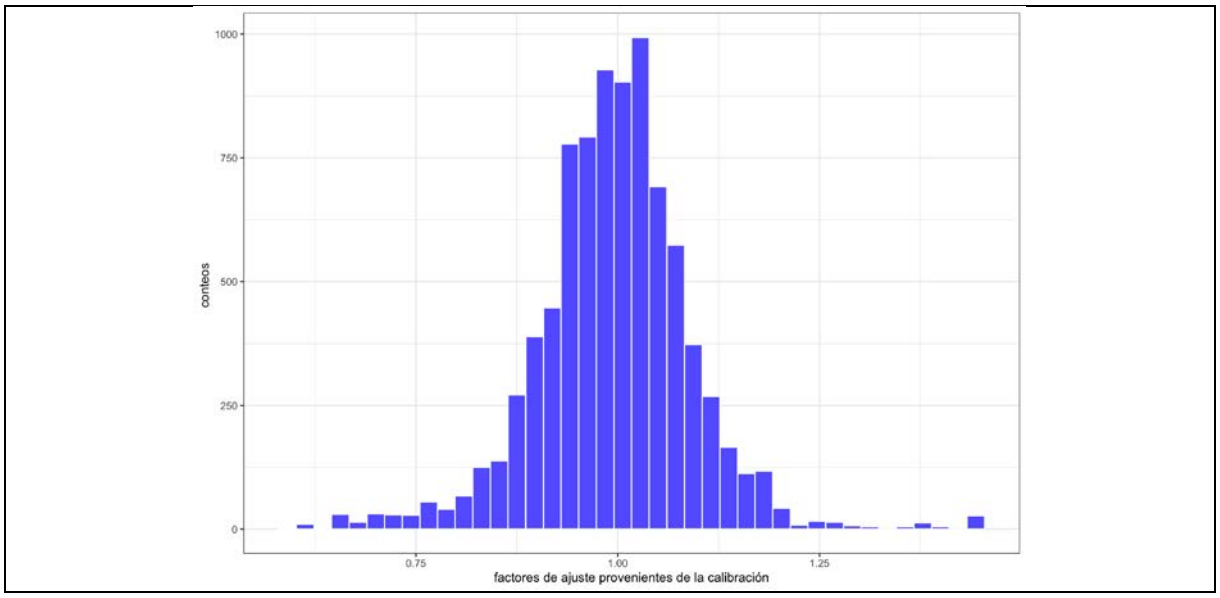
- Total de personas por departamento
- Total de personas a nivel país por sexo y cinco tramos de edad (0 a 14 años, 15 a 29 años, 30 a 49 años, 50 a 64 años y 65 años o más).

A su vez, se agregó como restricción adicional que la ponderación (participación) de los hogares y las personas estuviera balanceado para la semana de referencia de la encuesta, de forma de tener controlado los cambios en las tendencias de los indicadores de mercado de trabajo e ingresos de los hogares, una vez decretada la alerta sanitaria en el país.

Para la calibración se utilizó el raking truncado e imponiendo como restricción adicional que los pesos sean iguales para todos integrantes del hogar, siguiendo la metodología utilizada

para la calibración de la ENGIH 2016 - 2017[3]. En la figura 4 se presenta la distribución de los factores de ajuste proveniente de la calibración $g_{(2)i|(1)}$, donde se puede apreciar que los factores de ajustes de la calibración presentan una distribución simétrica. Esto implica que no existe un desbalance (significativo) entre la estructura de la población que estima la encuesta respecto a las proyecciones de población.

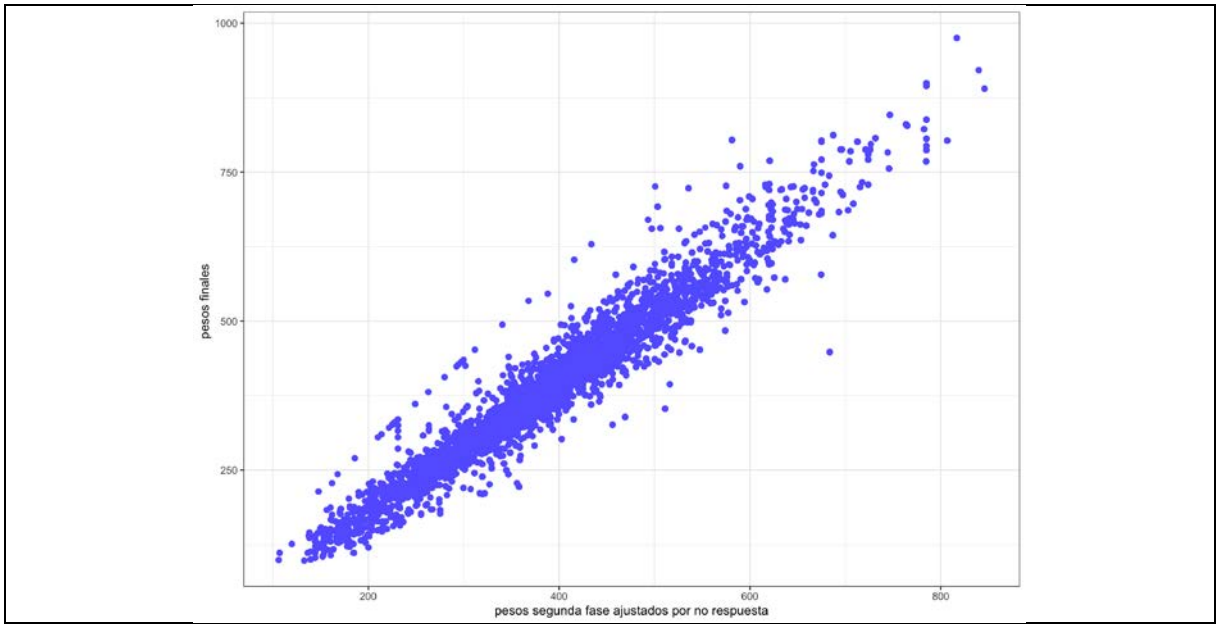
Figura 4. Distribución de los factores de ajuste provenientes de la calibración $g_{(2)i|(1)}$



Fuente: INE, Encuesta Continua de Hogares no presencial.

En la figura 5 se presentan los pesos ajustados por respuesta $w_{(1)i} \times d_{(2)i|(1)} \times \hat{\phi}^{-1}_{(2)i|(1)}$ respecto a los pesos finales.

Figura 5 - Pesos ajustados por no respuesta respecto a los pesos finales



Fuente: INE, Encuesta Continua de Hogares no presencial.

[3] http://www.ine.gub.uy/c/document_library/get_file?uuid=3b50400d-c42f-44e7-83a6-339f30798068&groupId=10181

Truncado y distribución de los pesos

En una última instancia y teniendo en cuenta la distribución obtenida de los pesos calibrados, se definió una cota superior, la cual se denota como U de forma de no permitir pesos extremos que puedan a llegar a influir en el análisis y/o producir estimaciones inestables producto de un aumento innecesario en los SEs. Una vez analizada la distribución de los pesos, el límite superior se fijó en $U = 780$.

El recorte y distribución de los pesos se realizada de la siguiente manera:

1. Se define el peso recortado como:
$$w_{(2)i,\text{trim}} = U \text{ si } w_{(2)i} > U \text{ y } w_{(2)i,\text{trim}} = w_{(2)i} \text{ si } w_{(2)i} < U$$
2. Se computa la suma
$$K = \sum_{i \in r_{(2)}} \text{abs}[w_{(2)i} - w_{(2)i,\text{trim}}]$$
donde $r_{(2)}$ es la muestra final.
3. Posteriormente, se distribuye K entre los pesos no recortados.
4. Se repiten los pasos 1 al 3 hasta que no haya pesos mayores al límite fijado U

5.3. Estimadores y sus precisiones

Las estimaciones que produce la ECH son totales y ratios entre totales (e.g. tasas de mercado laboral y la media del ingreso de los hogares[4]). El insumo principal para computar las estimaciones puntuales de los distintos estimadores son los pesos muestrales $w_{(2)i}$ descritos anteriormente.

El estimador del total de una variable cualquiera y viene dado por:
$$\hat{t} = \sum_{i \in r_{(2)}} w_{(2)i} \times y_i$$
donde y_i es el valor que toma la variable y en la unidad (hogar o persona) i y $r_{(2)}$ es la muestra final, es decir, los hogares respondentes.

El estimador del ratio o cociente entre dos variables cualquiera y y z viene dado por:
$$\hat{R} = \frac{\sum_{i \in r_{(2)}} w_{(2)i} \times y_i}{\sum_{i \in r_{(2)}} w_{(2)i} \times z_i}$$

6. Estimación de los errores estándar (SE)

El INE siguiendo las recomendaciones internacionales, computa intervalos de Confianza (IC) para las estimaciones de los principales indicadores que produce, de forma de brindar a los usuarios, una medida de calidad de las mismas y en base a ellos puedan interpretar los resultados de forma correcta.

El insumo principal para el cómputo de los IC es el error estándar del estimador. Por lo tanto, el objetivo es encontrar una estimación del SE que refleje todas las fuentes de variabilidad del estimador, las cuales vienen determinadas por el diseño muestral, así como

[4] La media es una caso particular de un ratio en donde la variable del denominador es la estimación del tamaño de la población.

por los distintos ajustes que son llevados a cabo para obtener los pesos finales. Los ajustes por NR tienden a incrementar los SEs de las estimaciones, mientras que los ajustes por calibración tienden a reducirlos, siempre y cuando, las variables utilizadas para la calibración expliquen de alguna forma la variabilidad de las variables de interés.

Cuando la muestra es seleccionada bajo un muestreo complejo generalmente no existen estimadores insesgados de los SEs y en la práctica, suele recurrirse a métodos que aproximan o que capturan los principales componentes de los SEs. Estos métodos de aproximación como ser, el método del último conglomerado junto con linearización de taylor o métodos de remuestreo (e.g. Bootstrap y Jackknife) son utilizados cuando la muestra de la encuesta es llevada a cabo por muestreos aleatorios, por conglomerados y en dos o más etapas de selección, en donde los conglomerados o Unidades Primarias de Muestreo (UPM) son seleccionadas (generalmente) con distintas probabilidades de selección y sin reposición. Cualquiera de estos métodos hace dos supuestos: i) la mayor variabilidad de las estimaciones se encuentra en la primera etapa de muestreo y ii) la tasa de muestreo es despreciable y por ende se asume que el muestreo es con reposición.

Por otro lado, en la actualidad existe muy poco avance en la literatura para poder encontrar una aproximación de los SE en diseños en dos fases y el poco avance a la fecha, se limita a diseños sencillos (e.g. muestreo aleatorio simple en ambas fases) o diseños complejos, en donde existen variaciones del método de Jackknife que pueden ser utilizados siempre y cuando las tasas de muestreo en ambas fases, sean despreciables (cercanas a cero).

Para el caso de las estimaciones de los SEs de la ECH no presencial, el problema principal se encuentra en que la primera fase sigue un diseño de muestreo complejo (estratificado, por conglomerados y en dos etapas de selección) y en la segunda fase se sigue un diseño relativamente sencillo (muestreo estratificado simple directo) pero con una tasa de muestreo no despreciable. Es por esto que se debió tomar una decisión sobre el método a utilizar para poder aproximar de mejor forma los SEs de las estimaciones. Se optó por utilizar el método del último conglomerado junto con linealización de Taylor, en donde se tiene en cuenta únicamente las UPM seleccionadas en la segunda fase de muestreo y se ignora la posible reducción de los SE producto de la calibración. Lo anterior intenta de alguna forma balancear el impacto de ambas fuentes de variabilidad, en donde posiblemente haya una subestimación de los SEs producto de ignorar parcialmente la primera fase de muestreo, mientras que ignorar los efectos de calibración produce generalmente una sobre estimación de los SEs.

Por lo tanto, para el caso de la ECH no presencial, la estimación del SE cuadrado (varianza) de la estimación del total de una variable cualquiera y viene dada por:

$$\widehat{\mathrm{SE}}^2(\hat{t}) = \sum_{h=1}^{H} \frac{1}{m_{(2)h}(m_{(2)h}-1)} \times \sum_{j \in r_{(2)h}} (\hat{t}_{jh} \times m_{(2)h} - \hat{t}_h)^2$$

donde:

$m_{(2)h}$ es la cantidad de UPM seleccionadas en la primera fase y que quedaron seleccionadas en el estrato h de la segunda fase de muestreo

$\hat{t}_{jh} = \sum_{i \in r_{(2)jh}} w_{(2)i} \times y_i$ es la estimación del total de y en la UMP j perteneciente al estrato h de la segunda fase de muestreo.

$\hat{t}_h = \sum_{h=1}^{H} \hat{t}_{jh}$ es la estimación del total de la variable y en el estrato h de la segunda fase de muestreo.

La estimación del cuadrado del SE de la estimación de un ratio R viene dada por:

$$\widehat{\mathrm{SE}}^2(\hat{R}) = \frac{1}{\hat{t}_z} \times \sum_{h=1}^{H} \frac{1}{m_{(2)h}(m_{(2)h}-1)} \times \sum_{j \in r_{(2)h}} (\hat{t}_{r,jh} \times m_{(2)h} - \hat{t}_{r,h})^2$$

donde

$\hat{t}_z = \sum_{i \in r(2)} w_{(2)i} \times z_i$ es la estimación del total de la variable z

$\hat{t}_{r,jh} = \sum_{r_{(2)jh}} w_{(2)i} \times r_i$ es la estimación del total de r en la UMP j perteneciente al estrato h de la segunda fase, donde $r_i = y_i - \hat{R} z_i$.

$\hat{t}_{r,h} = \sum_{h=1}^{H} \hat{t}_{r,jh}$ es la estimación del total de la variable r en el estrato h de las segunda fase.

Una vez computados los SEs para los distintos indicadores de la ECH no presencial, se computan los IC teniendo en cuenta que la distribución de los estimadores es aproximadamente normal y por lo tanto el IC para cualquier parámetro θ al 95% queda definido como:

$$\hat{\theta} \pm 1.96 \times \widehat{\mathrm{SE}}(\hat{\theta})$$

donde 1.96 es el valor de una normal estándar que acumula un 0.975 de probabilidad.

Como es habitual los IC para los principales indicadores se encuentran contenidos en el Boletín Técnico del mes de referencia.