FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRACIÓN

LICENCIATURA EN ESTADÍSTICA

Tratamiento de la no respuesta en encuestas de panel en el caso de poblaciones finitas: "Las damas perdidas".

Margarita Antía – Ana Coimbra

Tutores: Juan José Goyeneche
Guillermo Zoppolo

Octubre de 2009

Índice general

1. Introducción a las encuestas de panel **9**

2. El mundo ideal del muestreo por paneles **13**

2.1. Definición del "mundo ideal" 13

2.2. Obtención de estimadores insesgados 14

2.2.1. Estimación transversal 14

2.2.2. Estimación longitudinal: cambios entre olas 15

2.3. Conclusiones 16

3. No Respuesta **19**

3.1. Generalidades de la no respuesta 19

3.2. No respuesta al ítem 21

3.3. No respuesta de la unidad en las olas 23

3.4. Tratamiento general de la no respuesta 25

4. Imputación para la no respuesta al ítem **31**

4.1. Introducción 31

4.2. Métodos para construir los valores imputados 33

4.2.1. Imputación por regresión 33

4.2.2. *"Ratio Imputation"* e imputación por la media de respondentes 34
4.2.3. Imputación por el Vecino más Cercano 35
4.2.4. Imputación Cold-Deck 36
4.2.5. Imputación Hot-Deck 36
4.2.6. Grupos de Imputación 38
4.2.7. Árboles de clasificación y regresión 38
4.2.8. Imputación Múltiple 38
4.2.9. Imputación especial: juicio de expertos o con datos históricos 39
4.3. Discusión 39

5. Calibración para la no respuesta de unidades 41
5.1. Estimación en presencia de no respuesta: la necesidad de calibrar . . 42
5.1.1. Sesgo de un estimador del total 42
5.1.2. Descomposición del error de estimación 43
5.2. Requisito indispensable: Información auxiliar 46
5.3. Estimador puntual bajo calibración 48
5.4. Conjuntos alternativos para los pesos calibrados 50
5.4.1. Pesos iniciales alternativos 50
5.4.2. Variables auxiliares alternativas: vector instrumento 51
5.4.3. Ponderadores calibrados alternativos 51
5.5. Análisis del sesgo por no respuesta en el marco de la calibración . . . 52
5.6. Varianza y su estimación 55
5.7. Ejemplos de Estimadores Calibrados 58
5.7.1. El vector auxiliar más simple 58
5.7.2. Post estratificación 59
5.7.3. Raking 60
5.8. Estimadores calibrados en encuestas de panel 62

6. Aplicación: Las damas perdidas **67**

6.1. Introducción 67

6.2. Diseño Muestral 68

6.2.1. Diseño Muestral (2001) 68

6.2.2. La segunda ola del panel: entrevistas 2008 71

6.3. Obtención de ponderadores 73

6.3.1. Ponderadores para Estimaciones Longitudinales 73

6.3.2. Ponderadores para Estimaciones Transversales 77

6.4. Sugerencias para una potencial tercera ola 86

7. Conclusiones **89**

Índice de cuadros

3.1. Patrones de Respuesta en paneles 24

6.1. Totales poblacionales y muestrales según Segmento, Zona y Hogar por Estrato 70

6.2. Cantidad de entrevistas efectivas en 2008 72

6.3. Totales Poblacionales, Muestrales y Expansores originales por Estrato 74

6.4. Totales Poblacionales de las Variables Auxiliares 75

6.5. Totales muestrales según Edad por Estrato y Nivel Educativo 76

6.7. Totales Poblacionales, Muestrales y Expansores Originales por Estrato 79

6.8. Totales Poblacionales de la Variable Auxiliar 79

6.9. Totales Muestrales por Nivel Educativo según Estrato 80

6.10. Totales Poblacionales, Muestrales y Expansores Originales por Estrato 81

6.11. Totales Poblacionales de las Variables Auxiliares 81

6.12. Totales Muestrales según Edad por Estrato y Nivel Educativo 83

Sumario

Un problema central en las *encuestas de panel* es la *mortalidad de unidades*, que en términos generales es un problema de *no respuesta*. En este sentido, el trabajo se enfoca en fundamentar el marco teórico apropiado para ponderar individuos cuando se está frente al problema de la no respuesta en general, y en particular, en las encuestas de panel.

Luego de una breve introducción sobre las generalidades de las encuestas de panel, en el Capítulo 2 se proponen *estimadores de cambios* entre *olas* en presencia de respuesta perfecta. Este supuesto es levantado en los capítulos siguientes para proponer métodos para el tratamiento de la no respuesta. En el Capítulo 3 se desarrollan las definiciones y el tratamiento general de la no respuesta y se expresan los motivos por los cuales se opta por la aplicación de un enfoque que combina la *imputación* como tratamiento de la no respuesta en los ítems y la *calibración* para compensar la no respuesta a la unidad en el marco de las encuestas de panel. En el Capítulo 4 se presentan distintos métodos de imputación y los motivos por los cuales algunos son preferibles a otros; y en el Capítulo 5 se define el estimador calibrado, y las condiciones que deben cumplirse para asegurar que el sesgo introducido por la no respuesta sea mínimo.

El Capítulo 6 incluye la aplicación directa que resulta del cálculo de los ponderadores calibrados en la "Encuesta sobre Situaciones Familiares y Desempeños Sociales en Montevideo y Área Metropolitana" llevada a cabo en los años 2001 y 2008 por las Facultades de Ciencias Económicas y de Administración y de Ciencias Sociales de la Universidad de la República.

Capítulo 1

Introducción a las encuestas de panel

Las encuestas de panel se refieren a estudios basados en observaciones repetidas efectuadas sobre las mismas unidades de muestreo: personas, hogares, empresas, etc.

La medición periódica de elementos permite realizar un *seguimiento* de la población objetivo, logrando captar su dinámica en el tiempo. El objetivo de medir cambios netos en la población a nivel macro no es específico de los paneles, ya que puede lograrse mediante la comparación de resultados de encuestas cross-section convencionales realizadas en distintos momentos del tiempo. La justificación de la utilización de encuestas de panel radica en el interés de medir *cambios individuales* o micro en poblaciones específicas. Los resultados particulares en cada instancia de medición (estimaciones transversales) pueden ser obtenidos sin perjuicio de lo anterior y, aunque no sea el objetivo principal de las encuestas de panel suelen ser de interés en si mismos.

Los distintos momentos del tiempo en los que las encuestas son llevadas a cabo se denominan "olas"; la duración del panel y el período entre olas son definidos en la etapa del diseño de la encuesta.

Wayne Fuller y Jay Breidt (1999) distinguen tres variaciones de panel: *Panel Puro, Panel Rotativo, Panel Suplementado.* El Panel Puro es aquel en que las mismas unidades son observadas en distintos momentos del tiempo. La muestra es extraída

por única vez al inicio del estudio y todas las unidades seleccionadas serán observadas a lo largo de la duración del panel; una unidad que no fue seleccionada al principio nunca pertenecerá al panel. Esta metodología es aplicada en la "*SIPP*" (*Survey of Income and Program Participation*) llevada a cabo por el *US Bureau of Census* donde se relevan los datos de una primera y única muestra de hogares durante 8 períodos. En la Encuesta de Rotación una unidad es observada en un conjunto parcial de momentos, pero no se observa en el resto del estudio. Varios ejemplos en los que se especifica este patrón de observación son encuestas mensuales llevadas a cabo por el *US Bureau of Census*: *"CPS"* (*US Current Population Survey*), "*Monthly Retail Trade Survey", "Monthly Wholesale Survey"* entre otras. En la *CPS* la muestra original de hogares es dividida en 8 submuestras llamados grupos de rotación. Cada mes un grupo de rotación se introduce en el estudio, se entrevista durante 4 meses consecutivos y se descarta temporalmente durante 8 meses. De esta manera, 2 meses consecutivos comparten siempre el 75 % de la muestra. El uso de este procedimiento en estudios mensuales (u otros estudios de carácter muy repetitivo) se basa en las ventajas que ofrece frente a métodos no rotativos: se evita sobrecargar a los respondentes logrando obtener mayor tasa de respuesta (las encuestas repetitivas tienen la característica de aburrir a los entrevistados causando abandono del panel por parte de los mismos y producen cambios en los patrones de respuesta); individualmente, pueden ofrecer una solución insesgada al problema de observaciones extremas (Woodruff (1963)), ya que estas no son observadas en todos los períodos. En los Paneles Suplementados (Split Panels) la muestra original de individuos se observa en todos los momentos, y además se observan otros individuos en algunos momentos particulares llamados individuos adicionales o suplemantarios. Un ejemplo es el *"Erotion Update Study"* llevado a cabo por el Departamento de Agricultura de Estados Unidos. En ciertos períodos la muestra original de 3000 segmentos de tierra se complementa con 1000 unidades de muestreo adicionales. Los procedimientos de estimación en paneles suplementados se desarrollan en Fuller y Breidt (1999).

Un problema central en las encuestas de panel, independientemente de la variante elegida, es la mortalidad de unidades, que en términos generales es un problema de no respuesta. Este es un fenómeno presente en la mayoría de las encuestas por muestreo y es imprescindible su tratamiento para evitar sesgos en las estimaciones. La inclusión del factor "tiempo" en las encuestas de panel provoca un agravamien-

to del problema de no respuesta respecto a las encuestas cross-section, reflejado en reducciones considerables en el "tamaño de muestra" período a período debido a la movilidad, el fallecimiento y otros factores (como la pérdida de cooperación de unidades) que resultan en el "agotamiento" del panel. Otro efecto causado por la inclusión del factor "tiempo" es la potencial pérdida de representatividad de la muestra para inferir resultados transversales en olas posteriores a la primera.

Capítulo 2

El mundo ideal del muestreo por paneles

2.1. Definición del "mundo ideal"

Las condiciones ideales en cualquier tipo de encuesta por muestreo están regidas por la obtención de respuesta completa de las unidades muestreadas a partir de un marco muestral perfecto acorde a la población objetivo, sin presencia de errores de medición. En las encuestas por panel, dichas condiciones requieren el supuesto adicional de que la población objetivo sea fija en el tiempo. Adicionalmente se requiere obtener respuesta completa de todas las unidades en todas las olas. Bajo estos supuestos, en este capítulo se proponen estimadores para medir la evolución de cierta característica o variable de interés en un panel puro genérico.

Sea s una muestra aleatoria de la población U de N individuos, tomada bajo un diseño $p(s)$[1] de tamaño n_s. Al tratarse de un panel puro, la muestra será extraída una sola vez en el "momento cero", definido como el momento anterior a la primera etapa de entrevistas.

Las condiciones ideales implican el cumplimiento de:

[1] La elección del diseño a utilizar depende del objeto de estudio y no de la utilización de paneles.

- Los n_s individuos que pertenecen a la muestra inicial son entrevistados tanto en el primer momento como en los siguientes; mantienen durante todo el estudio el interés de pertenecer al panel, y no se registran bajas de ninguna unidad por motivo alguno (por ejemplo, no existe fallecimiento de unidades). Se obtienen así valores de la variable de interés Y para cada individuo en cada ola.
- La población U está fija en el tiempo.

2.2. Obtención de estimadores insesgados

La utilización de muestras por panel permitirá medir la evolución de la variable Y en el tiempo, además de obtener las estimaciones usuales de un muestreo transversal. La obtención de estimadores insesgados de totales para ambos casos es sencilla y directa cuando se está bajo las condiciones del "mundo ideal".

2.2.1. Estimación transversal

En cada momento del tiempo $(i = 1, 2, 3.....)$, se estimará el total poblacional de la variable Y (definido como $t_i = \sum_{k \in U} y_{ik}$, donde y_{ik} es el valor de la variable y en la i-ésima ola de entrevistas para el individuo k) mediante el estimador π (Horvitz Thompson (1952)):

$$\hat{t}_{\pi_i} = \sum_{k \in s} y_{ik} \frac{1}{\pi_k} = \sum_{k \in s} \overset{\surd}{y}_{ik} \tag{2.1}$$

Donde $\overset{\surd}{y}_{ik} = \frac{y_{ik}}{\pi_k}$ y $\pi_k = P(k \in s)$.

Luego se tiene que:

$$E_{p(s)}\left(\hat{t}_{\pi_i}\right) = t_i \tag{2.2}$$

$$Var_{p(s)}\left(\hat{t}_{\pi_i}\right) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \overset{\surd}{y}_{ik} \overset{\surd}{y}_{il} \tag{2.3}$$

Donde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ y $\pi_{kl} = P(k\ y\ l \in s)$. El estimador π de la varianza, $\hat{Var}_{p(s)}\left(\hat{t}_{\pi_i}\right) = \sum \sum_s \overset{\surd}{\Delta}_{kl} \overset{\surd}{y}_{ik} \overset{\surd}{y}_{il}$ es insesgado para estimar $Var_{p(s)}(\hat{t}_{\pi_i})$, siendo

$\overset{\checkmark}{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$.

Observación: $\sum_{k \in s} = \sum_{s}$; $\sum_{k \in s} \sum_{l \in s} = \sum_{s} \sum$ y $\sum_{\substack{k \in s \\ k \neq l}} \sum_{l \in s} = \sum_{\substack{s \\ k \neq l}} \sum$ de aquí en adelante.

Cada individuo es ponderado por el inverso de su probabilidad de selección en la muestra, $\frac{1}{\pi_k}$, que es constante para todo i, dado que bajo las condiciones ideales establecidas, una vez que la unidad pertenece a la muestra, será entrevistada en todas las olas.

2.2.2. Estimación longitudinal: cambios entre olas

La aplicación específica de las encuestas de panel consiste en la estimación del cambio de la variable de interés Y entre la ola j y la ola $j+h$, definido como:

$$\begin{aligned} A_{j,j+h} &= t_{j+h} - t_j \\ &= \sum_U y_{(j+h)\,k} - \sum_U y_{j\,k} \\ &= \sum_U \left[y_{(j+h)\,k} - y_{j\,k} \right] \\ &= \sum_U a_{(j,j+h)\,k} \end{aligned} \tag{2.4}$$

Siendo $\boldsymbol{j} = 1, 2, 3, 4, ..., \kappa$ y κ la cantidad de olas; con $h > 0$; $j + h \leqq \kappa$.

El cambio total entre olas se define entonces como la suma total de las diferencias de las mediciones de la variable de interés para cada uno de los elementos, y se estimará mediante:

$$\begin{aligned} \hat{A}_{j,j+h} &= (\hat{t_{j+h}} - t_j) \\ &= \sum_s \left(\frac{y_{(j+h)\,k} - y_{j\,k}}{\pi_k} \right) \\ &= \sum_s \frac{a_{(j,j+h)\,k}}{\pi_k} \end{aligned} \tag{2.5}$$

Los estimadores $\hat{A}_{j,j+h}$ tienen la forma de estimadores π (Horvitz-Thompson), por

lo tanto, también comparten sus propiedades. Esto es:

$$E_{p(s)}\left(\hat{A}_{j,j+h}\right) = A_{j,j+h} \tag{2.6}$$

$$Var_{p(s)}\left(\hat{A}_{j,j+h}\right) = \sum\sum_{U} \Delta_{kl}\, \overset{\vee}{a}_{(j,j+h)\,k}\, \overset{\vee}{a}_{(j,j+h)\,l} \tag{2.7}$$

$$\hat{Var}_{p(s)}\left(\hat{A}_{j,j+h}\right) = \sum\sum_{s} \overset{\vee}{\Delta}_{kl}\, \overset{\vee}{a}_{(j,j+h)\,k}\, \overset{\vee}{a}_{(j,j+h)\,l} \tag{2.8}$$

$$E\left[\hat{Var}_{p(s)}\left(\hat{A}_{j,j+h}\right)\right] = Var_{p(s)}\left(\hat{A}_{j,j+h}\right) \tag{2.9}$$

La varianza de $\hat{A}_{j,j+h}$ también puede expresarse como

$$Var_{p(s)}\left(\hat{A}_{j,j+h}\right) = Var(\hat{t}_j) + Var(\hat{t}_{j+h}) - 2Cov(\hat{t}_j, \hat{t}_{j+h}) \tag{2.10}$$

y dado que se estos totales están calculados en base a las mismas unidades, se espera los totales estimados en olas sucesivas estén correlacionados positivamente resultando en una varianza pequeña del estimador de diferencias.

Una alternativa a la utilización de paneles es la estimación de diferencias de la variable de interés utilizando totales estimados mediante encuestas cross-section en los momentos j y $j+h$. En este caso el estimador de diferencias también será insesgado pero su varianza será mayor que en el caso anterior, ya que la independencia entre muestras en una y otra instancia determina que el tercer término de (2.10) sea cero.

Este es un aspecto importante en la justificación de utilización de paneles frente a encuestas cross-section para medir cambios.

2.3. Conclusiones

Los estimadores propuestos en este capítulo tienen la cualidad de ser insesgados y de sencillo cálculo, ya que son estimadores π. Para su desarrollo, se partió de supuestos

muy restrictivos, rara vez presentes en la práctica, a saber: la existencia de *respuesta perfecta* y población fija en el tiempo, reflejados en ponderadores constantes en el tiempo para cada elemento e iguales al inverso de su probabilidad de inclusión en la muestra.

Existen casos de encuestas por panel para los cuales estos supuestos pueden ser aproximadamente ciertos, por ejemplo, la medición del crecimiento de árboles en una plantación determinada suponiendo que no hay árboles que mueran. La muestra de árboles seleccionada se mantendría intacta a lo largo de todo el período de medición, haciendo que sea posible obtener datos en todas las etapas. Cuando el estudio refiere a personas u hogares, es inocente creer que en cada instancia serán medibles todas las variables en estudio para todas las unidades. Los fallecimientos, las mudanzas, y hasta la negativa del individuo seleccionado a contestar alguna pregunta en particular o todo el cuestionario, son hechos comunes y deben tomarse en cuenta al momento de realizar un análisis estadístico. La *no respuesta* existe, y es deber de los investigadores analizarla y planear estrategias que permitan que el sesgo introducido al estudio por esta causa sea el menor posible. Es por esto que el resto del trabajo estará enfocado en desarrollar y comparar distintas estrategias para el manejo de la no respuesta en este tipo particular de muestreo, con énfasis en la búsqueda de los ponderadores de individuos respondentes en cada ola que resulten más adecuados para minimizar el sesgo de dichas estimaciones. Lejos de planteos para analizar el *"mundo ideal"* del muestreo por paneles, en los capítulos que siguen se levantará el supuesto de *respuesta perfecta* para dar paso al estudio del *"mundo real"*.

Capítulo 3

No Respuesta

3.1. Generalidades de la no respuesta

Un problema cotidiano para los investigadores en cualquier análisis de encuestas por muestreo es la *no respuesta*, o sea, la imposibilidad de obtener toda o alguna información para una o más de las unidades seleccionadas en la muestra.

Las causas del faltante de información pueden ser de distinta naturaleza: negativa del individuo a colaborar en la encuesta, imposibilidad de localizar a la unidad seleccionada, formularios incompletos, etc. En base a esto pueden distinguirse distintos tipos de no respuesta: *no respuesta al ítem*, que refiere a faltantes en la respuesta para un ítem en particular del formulario debido a omisión (tanto del entrevistador como del entrevistado) o negativa del encuestado a contestar; y *no repuesta de la unidad*, que se da cuando la unidad seleccionada para ser entrevistada no es encontrada o se rehusa a participar en la encuesta.

Si la no respuesta (tanto de la unidad como del ítem) se presentara de manera completamente aleatoria, el único incoveniente al que se enfrenta el investigador resulta en la reducción del tamaño de muestra y su respectivo aumento en la varianza de las estimaciones, que podría ser fácilmente contrarrestado mediante un "sobremuestreo" (fijando un tamaño de muestra mayor en la etapa de diseño). De esta manera el único efecto negativo de la no respuesta sería el incremento en la carga administrativa y los costos de recolección de datos. En la práctica, la situación anterior sería una "feliz"

casualidad. Las unidades que no contestan "normalmente" difieren en algunas características de aquellas que sí lo hacen, y el sesgo introducido en las estimaciones por esta causa constituye el obstáculo más importante por corregir. Frente a la pérdida del insesgamiento de los estimadores, el incremento de su varianza es un disturbio menor: en presencia de sesgo significativo, un intervalo de confianza calculado estará centrado en un valor erróneo y no se logra el nivel de confianza requerido.

El sesgo por no respuesta solo puede definirse en relación a una estimación deseada, no es una propiedad inherente de la muestra de respondentes. En otras palabras, es posible que la misma muestra pueda ser insesgada respecto a una estimación pero sesgada respecto a otra. Esto dependerá de la asociación de las variables en las cuales se basa la estimación y la propensión de respuesta (desconocida) de las unidades de la muestra.

Existe una amplia literatura acerca de las características de los no respondentes. Los dos componentes principales de la no respuesta son los rechazos a contestar por parte de la unidad, y el no contacto de la misma. Tanto la propensión a contactar las unidades muestreadas como la intención de las mismas de participar, pueden estar afectadas por las características demográficas de la población. Groves y Couper (1998) afirman que las características demográficas que más influyen en la participación de los individuos en una encuesta son: la edad avanzada, los menores ingresos, el menor nivel educativo, ser soltero, pertenecer a una minoría étnica, pertenecer a hogares con alta movilidad y residir en áreas urbanas; mientras que los individuos que no pueden ser localizados comparten las características de tener edad avanzada, ser jóvenes, ser hombres, poseer mayores ingresos y/o tener empleo, pertenecer a hogares unipersonales o con alta movilidad, residir en hogares en zonas urbanas.

Aunque los mecanismos que llevan a la no respuesta en encuestas por paneles son similares a aquellos que operan en las escuestas cross-section, en las primeras estos mecanismos se profundizan debido a la reiteración de entrevistas y a la variación inevitable en el tiempo de la población. El mayor problema del muestreo por paneles radica en la no respuesta, ya que ésta afecta directamente a su objetivo. Estudiar la dinámica de una población sin realizar acciones para compensar los faltantes de información conduce a resultados incorrectos.

3.2. No respuesta al ítem

El origen de los datos faltantes permite distinguir tres formas de no respuesta al ítem:

1. Información no relevada: no se provee información por parte del respondente (por omisión del entrevistador, por desconocimiento de la respuesta por parte del encuestado, o simplemente porque el mismo se rehúsa a contestar). También se consideran faltantes aquellas respuestas que no tienen ningún valor informativo, por ejemplo: contestar "No sabe" en una encuesta de ingresos (aunque la misma respuesta en una encuesta de intención de voto sí tiene valor).

2. Información incosistente: la información que provee el respondente no puede usarse (la respuesta cae fuera del rango de respuestas posibles, no puede ser codificada, no es confiable, o no es coherente con respuestas anteriores, etc.)

3. Información extraviada: la información se pierde (ilegibilidad del formulario, pérdida de formularios en la etapa de entrada de datos)

Las primeras dos formas son originadas en la etapa de trabajo de campo, y la tercera resulta en errores en la fase de proceso o análisis de los mismos. La forma más problemática de la no respuesta al ítem es el caso en que el respondente no provee información, ya que pueden estar funcionando mecanismos de datos faltates diferentes. Los puntos 2 y 3 son de alguna manera evitables por parte del investigador dado que éste tiene control sobre ellos, mientras que en el punto 1, la obtención de respuestas depende mayoritariamente de la disposición de la unidad que es entrevistada.

Durante años los investigadores solían solucionar el faltante de información en los ítems restringiendo el análisis a aquellas unidades observadas en su totalidad. Sin embargo, existe la posibilidad de que hayan diferencias sistemáticas entre unidades que responden a un ítem particular, y aquellas que no lo hacen (Rubin, 1976). Un requisito para el tratamiento estadístico de datos faltantes es conocer un poco más acerca de "cómo" y "por qué" los faltantes ocurren: un faltante por omisión accidental difiere del faltante por negativa del encuestado a brindar información sobre ciertas

variables. Por lo tanto, el primer tema para centrar la atención es la distinción entre la aleatoriedad y la no aleatoriedad de los datos faltantes.

- Los datos son *faltantes completamente aleatorios* (MCAR: missing completely at random) si el faltante de respuesta está incorrelacionado con su valor desconocido, y también incorrelacionado con los valores de respuestas a otras preguntas (por ejemplo la omisión completamente involuntaria). Cuando los datos son MCAR, los valores faltantes son una muestra aleatoria de todos los valores, no existiendo diferencias sistemáticas entre respondentes y no respondentes. El análisis de datos proveerá estimaciones insesgadas, con el único inconveniente de un tamaño de muestra menor. Este es el caso de no respuesta *ignorable*.

- Los datos son *faltantes aleatorios* (MAR: missing at random) cuando están relacionados a los datos observados, pero no se relacionan al valor de la respuesta faltante en sí misma. Por ejemplo, si una persona no recuerda un evento debido a deficiencias en la memoria, el faltante puede estar relacionado con la edad del respondente, y no con el evento que no recuerda. Cuando los datos son de este tipo, el faltante es un proceso aleatorio condicional en los datos observados (en el ejemplo, los faltantes son una muestra aleatoria dentro de subgrupos formados por la edad).

- Los datos son *faltantes no aleatorios* (MNAR: missing not at random) en los casos que el faltante está relacionado con la propia pregunta, por ejemplo, cuando el respondente percibe que la verdadera respuesta es políticamente incorrecta, y se niega a contestar. Cuando los datos son faltantes no aleatorios suelen aparecer serios sesgos. En este caso, la no respuesta es no ignorable, debiéndose introducir en el estudio algún modelo de no respuesta, de manera que los sesgos generados sean lo menor posible.

Cuando los datos son faltantes aleatorios es importante detectar qué variables se relacionan al faltante. Como ejemplo, la edad y nivel educativo suelen estar correlacionadas con la no respuesta a ciertos ítems, resultando en datos aleatorios, por lo que estas variables deberían ser incorporadas en el modelo de ajuste para lograr

superar el problema. Cuando los datos son faltantes no aleatorios, el proceso es más complicado. Se deben inspeccionar los patrones de faltantes. En base a este análisis se podría detectar si los mayores faltantes se dan en una variable poco relevante al estudio y decidir quitarla del mismo; o se pueden encontrar algunos individuos con muchos ítems sin responder, quitándolos también del análisis. De todas maneras, el caso general es que esto no suceda: los faltantes suelen estar "desparramados" por toda la matriz de datos. Será de interés analizar si estos faltantes forman algún patrón en especial, o sea, si pueden relacionarse con algunas de las variables observadas.

En la mayoría de los paquetes estadísticos se incluyen formas de inspeccionar estos patrones. De todas maneras, estas inspecciones no podrán asegurar si el faltante es independiente o no al valor desconocido de la variable. Para testear el supuesto de datos faltantes aleatorios es necesario contar con información adicional, que puede provenir, por ejemplo, de estudios previos.

3.3. No respuesta de la unidad en las olas

La no respuesta de unidades en la ola es una forma de no respuesta parcial particular al muestreo por paneles. Es común que algunas unidades del panel no provean datos en una o más de las olas de la encuesta. Algunos miembros de la muestra pueden abandonar la encuesta en cierta ola y perderse para el resto del estudio (*desertores*); mientras que otros pueden perderse en una ola, y volver al panel en alguna de las siguientes (*respondentes episódicos*). En el cuadro que sigue, se especifican todos los patrones de respuesta posibles para un panel de cinco olas.

Cuadro 3.1: Patrones de Respuesta en paneles

Patrón	Estado de respuesta	Ola 1	Ola 2	Ola 3	Ola 4	Ola 5
1	Respondentes	x	x	x	x	x
2	No	x	x	x	x	-
3	Respondentes	x	x	x	-	-
4	por	x	x	-	-	-
5	Desgaste	x	-	-	-	-
6	No	x	x	-	x	x
7	Respondentes	x	-	-	x	x
8	Episódicos	x	-	-	-	x

Ref: x: respuesta, -: no respuesta

En la tabla se puede ver que los *no respondentes totales* (no respondentes en la primera ola) no están representados por ningún patrón, ya que en la mayoría de las encuestas de panel, las unidades que no responden en la primera instancia no se siguen en las olas siguientes; posiblemente porque no se cuenta con información suficiente para poder hacerlo.

Es necesario que los investigadores fijen reglas para especificar cuándo deben cesar los intentos por seguir a los no respondentes (por ejemplo se abandona el seguimiento de las unidades después de 2 olas consecutivas de no respuesta). En base a estas reglas los patrones de respuesta serán específicos del estudio (en el caso del ejemplo, los patrones 7 y 8 de la tabla no existirían). En el cuadro también puede verse la distinción entre "no respuesta por desgaste" y "no respuesta episódica" dado que en general los métodos para compensarlas no son iguales, en los dos casos se cuenta con información auxiliar distinta. La no respuesta por desgaste refiere únicamente a los casos en los que una vez que la unidad no se reporta, tampoco lo hace para el resto del estudio: cada ola sucesiva suma un conjunto adicional de no respondentes a aquellos que ya eran no respondentes en olas anteriores. La información auxiliar disponible a ser usada para compensar las unidades perdidas en la ola es muy rica porque se cuenta con información de las instancias anteriores además de las variables utilizadas en el diseño. Esta es la principal diferencia frente a la no respuesta episódica, también definida por varios autores como "re-entry". En este caso, no siempre se cuenta con

información auxiliar proveniente de la ola anterior, y cada patrón de no respuesta episódica debe analizarse en particular para reconocer cual es la información que hay disponible. Por estos motivos suelen plantearse métodos de ajuste particulares para cada tipo de no respuesta.

3.4. Tratamiento general de la no respuesta

Dada la problemática que la no respuesta introduce al análisis de los datos, es de gran importancia invertir esfuerzo en intentar evitarla, o al menos, lograr minimizarla.

En general, la participación en el panel por parte de las unidades de la muestra está influenciada por factores exógenos a la encuesta como lo son el entorno social y los atributos psicológicos y sociales de los individuos (sentido de deber cívico e interés personal en el tópico de la encuesta); y factores sobre los cuales el investigador tiene algún tipo de control: protocolos de la encuesta (modo de coleccionar los datos, incentivos, carga) y la selección y entrenamiento de los encuestadores. Otro factor que puede incidir en la participación del encuestado es la reputación de la organización que lleva a cabo el estudio y su índole (pública, privada, otras). Si se logra que el respondente se interese en el tema o crea que se verá beneficiado en alguna manera por brindar información estará más interesado en participar. La teoría de "costo-oportunidad" establece que un miembro de la muestra analizará los costos y los beneficios de participar en la encuesta, y si los beneficios ofrecidos son mayores que los costos las unidades serán más propensas a participar. Como aspectos negativos influyen factores como el largo o la complejidad del cuestionario, preocupación acerca de la privacidad y confidencialidad de los datos proporcionados y una alta frecuencia de entrevistas en el período de duración del panel. Como aspecto positivo el respondente puede estar interesado en el tópico de la encuesta, pueden existir incentivos (económicos, por ejemplo) para que participe o puede tener un fuerte sentido de deber cívico. En una encuesta de panel el respondente usará su experiencia en las olas previas como guía para decidir si quiere o no seguir participando. Taylor y Brook (1997) encontraron que comunicarle explícitamente a los miembros de la muestra que se trata de una encuesta por panel (a partir de la segunda ola) resulta en una pequeña reducción en la tasa de respuesta de esa ola, pero la respuesta neta en las olas siguientes se verá mejorada.

La no respuesta también puede ser ocasionada cuando ciertas unidades de la muestra no pueden ser localizadas o contactadas. El eventual contacto se ve afectado por la estrategia que siga el encuestador, por ejemplo, si la unidad a ser entrevistada es el jefe de hogar y los intentos de contacto se realizan únicamente en días hábiles y durante horario laboral, es muy posible que esa entrevista falle.

En encuestas por panel a partir de la segunda ola surge una posible fuente adicional de no contacto: la movilidad. La posibilidad de localizar la nueva dirección del miembro de la muestra depende parcialmente del esfuerzo por parte del investigador para recolectar información en olas anteriores que permitan rastrearlo o re-contactarlo.

A pesar del esfuerzo de los investigadores para evitar la no respuesta, ésta existe y la pérdida de datos ocasionada por este fenómeno debe ser compensada para lograr estimaciones con sesgo reducido, ya sea en la etapa de recolección de datos como en la de obtención de resultados. En lo que sigue se señalan algunos métodos que se han utilizado para el tratamiento de la no respuesta referidos a la etapa de estimación.

Sustitución de no respondentes

Como principal ventaja de esta técnica se destaca su bajo costo y la obtención del tamaño de muestra planificado. Como principal desventaja, la sustitución de no respondentes no permite el cálculo de las probabilidades de inclusión de las unidades muestreadas y por tanto no permite una solución satisfactoria. Se asume que el tamaño de muestra real es la suma de las encuestas efectivamente realizadas más las sustituidas y se utilizan los procedimientos usuales para el caso de respuesta completa. De esta forma, la validez del procedimiento descansa en que la sustitución sea lo suficientemente meticulosa para garantizar que las unidades sustituidas brindarán las mismas respuestas que las originales. Aún en este caso, en las encuestas de panel, la sustitución no resulta ser una solución válida, ya que imposibilita las estimaciones de cambios entre olas para los individuos sustitutos y sustituídos.

Submuestreo de los no respondentes

Este mecanismo es inobjetable desde el punto de vista teórico y permite la estimación insesgada aún en el caso de que no todos las unidades de la muestra sean observadas. Se basa en tomar una nueva muestra de las unidades que no respondieron en una primera instancia; para esta segunda muestra se supone que no hay no respuesta y todos los renuentes a contestar en el primer intento sí colaboran en la segunda

instancia. El método guarda algún parecido con el muestreo en dos fases para estratificación, en tales circunstancias se obtienen estimadores insesgados para totales poblacionales y para las varianzas de estos. La debilidad de este método radica en el supuesto de respuesta perfecta de la segunda muestra; aún en caso de ser alcanzable, resultaría en costos de relevamiento muy altos.

Métodos basados en el enfoque de la "quasi-randomization"

Desde principios de los '80 comenzó a popularizarse un enfoque donde la no respuesta puede pensarse como un segundo mecanismo de selección que genera los datos efectivamente observados, una vez tomada la muestra. De esta manera, cada individuo en la población tiene una probabilidad de selección conocida y también una probabilidad de responder, condicional a haber sido seleccionado en la muestra. Si las probabilidades de responder fueran conocidas el problema estaría resuelto bajo la teoría del muestreo en dos fases. Por su propia naturaleza, las probabilidades de responder son desconocidas y muchos trabajos se han enfocado en la forma de tratar el problema suponiendo un modelo para estimar dichas probabilidades. Bajo este enfoque, condicional a la validez del modelo, se obtienen estimadores aproximadamente insesgados.

Los modelos ensayados para estimar las probabilidades de respuesta han sido variados y en general requieren del uso de algún tipo de información auxiliar que puede disponerse a nivel poblacional y a nivel de la muestra original (o sea, se recoge alguna información para respondentes y no respondentes). Uno de los más populares es el *modelo de respuestas homogéneas* donde los individuos de la muestra son clasificados en grupos exhaustivos y excluyentes donde se supone que las probabilidades de responder son homogéneas y que los individuos responden de manera independiente (Särndal y Swensson (1987) y Särndal et al (1992)). La principal crítica a estos enfoques es que es difícil defender un modelo como "mejor" que cualquier otro modelo competidor.

Calibración e Imputación

Son éstas las técnicas dominantes en la literatura actual (Särndal y Lundström (2005)). Usualmente la calibración predomina en el tratamiento para el caso de no respuesta de unidades; mientras que la imputación es más extensamente aplicada en los problemas de no respuesta de ítems. La primera es una estrategia global, tratan-

do todas las variables de forma simultánea, mientras que la segunda es particular, específica de cada variable. No obstante lo anterior, se han ensayado soluciones que aplican ambas técnicas en conjunto (Deville y Särndal (1994)). La decisión acerca de utilizar una u otra no es obvia, y depende de distintos factores como lo son: la cantidad y el número de olas en las que existen datos faltantes, el tipo de análisis a llevarse a cabo, la disponibilidad de variables auxiliares con poder predictivo de los valores faltantes y el costo de implementar los procedimientos.

La imputación es el procedimiento a través del cual los valores faltantes en una o más variables de estudio se completan con sustitutos. Los valores perdidos en la base de datos se reemplazan por los valores "plausibles" dando como resultado una matriz completa de valores.

Existen varios métodos de imputación que básicamente difieren en como definen "plausible", pero la mayoría coinciden en la necesidad de utilización de información auxiliar. Los valores imputados son por definición artificiales y contienen error similar al error de medida (definido como el error existente cuando el respondente provee un valor erróneo para el ítem), con la salvedad de que el primero ocurre por construcción, dado que el estadístico sabe que el valor insertado no es el real.

La calibración, o más precisamente el uso de estimadores calibrados, se basa fuertemente en el uso de información auxiliar tanto a nivel poblacional como a nivel de la muestra original. Su creciente popularidad puede explicarse porque no se basa en la especificación de un modelo de no respuesta, brinda un enfoque unificado dentro de la teoría del muestreo de poblaciones finitas, es computacionalmente sencilla de implementar y generaliza otras técnicas del tratamiento de la no respuesta como la post-estratificación, el raking y algunos casos de los ajustes basados en la teoría del muestreo en dos fases. Adicionalmente, al permitir el tratamiento de la no respuesta, también contribuye en la reducción de la varianza de los estimadores (Särndal y Lundström (2005)). La idea central de los estimadores calibrados es sencilla, consiste en modificar los ponderadores originales de la muestra minimizando alguna función de distancia entre dichos ponderadores y los ponderadores finales (o calibrados) y de manera que estos últimos estimen sin error algunas cantidades conocidas con los datos de los respondentes.

Si bien el uso de información auxiliar es imprescindible, en general siempre existe algún mínimo de información disponible a nivel de la población (o estimadores su-

ficientemente precisos), y a nivel muestral no resulta excesivamente costoso obtener alguna información auxiliar útil, tanto para respondentes como no respondentes.

Cambiar las ponderaciones de los individuos respondentes de la muestra e imputar se piensa que son métodos muy distintos de tratar los datos faltantes, pero en realidad, en análisis univariados están relacionados. En algunos esquemas de imputación, la celda faltante del no respondente se completa con la respuesta dada por el respondente más parecido: el respondente le dona la respuesta al no respondente parecido a él, lo que es equivalente a aumentar el peso de la respuesta del donante de cierto ítem.

En el caso de la calibración el conjunto de respondentes deberá representarse a sí mismo y al grupo de no respondentes, modificando sus ponderadores para lograrlo. De esta manera, las distribuciones basadas en la muestra de respondentes y los valores imputados serán iguales a las de los respondentes con pesos modificados, obteniéndose las mismas medidas de resumen. Este razonamiento esconde las diferencias que existen entre imputar y calibrar. No siempre calibrar implica duplicarle el peso al donante en una unidad, sino que estos nuevos pesos suelen estar fraccionados y distribuidos sobre toda la muestra de respondentes. Al hacer esto, se evita que aumente la varianza de los estimadores asociados a los respondentes. Con la imputación es menos sencillo evitar este aumento en la variación, pero puede reducirse usando imputación múltiple, o métodos que seleccionen al azar qué respondentes serán donantes de respuestas.

La imputación en la ola implica obtener valores de todas las variables de los individuos que no contestan teniendo como resultado una matriz completa de datos. A pesar de esto se pueden dar respuestas incompatibles, en el sentido de que para una misma unidad una variable es imputada sin tener en cuenta las demás, o sea sin considerar las relaciones existentes entre las variables del mismo individuo. Cuando la cantidad de preguntas en un cuestionario es grande la imputación en la ola genera una fabricación masiva de datos, y sus correspondientes distorsiones en las asociaciones entre variables. Dado que esta asociación es el principal objetivo del panel, el riesgo a su distorsión es el mayor motivo por el cual es preferible utilizar una estrategia global como lo es la reponderación cuando es la unidad la que no provee de respuesta.

No obstante a todo esto es necesario aplicar métodos de imputación para el tratamien-

to de no respuesta al ítem previo a la calibración de la no respuesta de unidades.

Capítulo 4

Imputación para la no respuesta al ítem

4.1. Introducción

La razón principal por la cual se realiza la imputación es la obtención de un conjunto de datos completo y consistente al cual se le pueda realizar análisis e inferencia estadística. La imputación de datos faltantes es la etapa final del proceso de depuración de datos (previo a la calibración por la pérdida de unidades) y consiste en reemplazar los valores faltantes por otros aceptables.

En décadas anteriores era habitual, a la hora de analizar los datos, ignorar aquellos registros que poseían algún valor faltante en alguna variable. Se empleaban los métodos de eliminación por lista (listwise deletion) o por pares (pairwise deletion). Esto suponía que aquellos individuos que no habían contestado a alguna de las preguntas del análisis eran ignorados provocando ciertos problemas en los resultados y pérdida de información valiosa.

Encontrar un buen método de imputación es una tarea importante ya que si se cometen errores en las imputaciones de datos individuales, estos pueden magnificarse al realizar estadísticas agregadas. Es por esto que es importante utilizar métodos de imputación que conserven ciertas características deseables de la variable a ser imputada, por ejemplo, no distorsionar las relaciones con el resto de las variables en

estudio, proporcionar valores válidos, etc. Los métodos de imputación más reconocidos están basados en información auxiliar, que puede ser información del marco, respuesta a otras preguntas y en el caso de paneles, la información obtenida en la ola anterior.

La construcción de valores imputados pueden clasificarse en tres categorías:

1. valores construídos con modelos estadísticos de predicción.
2. valores que han sido observados para elementos respondentes que son similares a aquellos que no responden.
3. valores construídos por la opinión de expertos.

Las primeras dos categorías utilizan modelos para producir valores sustitutos: la primera se basa en la relación asumida entre variables (predicción por regresión, por ejemplo) y la segunda usa métodos para clasificar qué elementos similares serán donantes de respuesta y cuales serán receptores. Los métodos de la tercera categoría se basan fuertemente en la destreza y conocimiento de expertos acerca del elemento en particular que requiere imputación.

Los valores imputados pueden ser *determinísticos* (cuando al repetir el procedimiento de imputación se obtiene el mismo valor imputado) o *aleatorios* (cuando la repetición del procedimiento genera valores imputados diferentes). La imputación por regresión es un ejemplo de regla determinística, mientras que un ejemplo de regla aleatoria es la imputación hot-deck.

Existen dos enfoques principales de imputación: "imputación completa" e "imputación combinada con calibración" o "enfoque combinado". Ambos enfoques resultan en matrices completas. El enfoque de imputación completa implica el uso de imputación tanto para compensar la no respuesta en el ítem como también la no respuesta a la unidad. La matriz de datos resultante tendrá tantas filas como individuos participantes en la encuesta y tantas columnas como variables, será de dimensión $(n \times J)$. El enfoque combinado propone imputar para obtener respuesta completa solamente para las unidades que han contestado al menos a una variable de la encuesta (no respuesta en el ítem). La matriz rectangular resultante tendrá, al igual que el caso anterior, tantas columnas como variables, pero tendrá tantas filas como individuos

respondentes ($m \,\mathrm{x}\, J, \quad m < n$). En base a esta matriz se calibrará para compensar las unidades no respondentes.

Las variables de la encuesta generalmente están afectadas tanto por la no respuesta a la unidad como la no respuesta al ítem, por lo que los valores y_{ik} de la i-ésima variable de estudio necesarios para construir estimadores de totales estarán disponibles sólo para los $k \in r_i \subset r \subset s$, donde r es el conjunto de respondentes de la muestra s y r_i es el conjunto de respondentes de la i-ésima variable de estudio. La idea detrás del enfonque combinado es imputar aquellos valores pertenecientes al conjunto $r - r_i$, de modo de lograr una matriz rectangular que tendrá tantas filas como individuos respondentes, y tantas columnas como variables. La matriz resultante contendrá los elementos $\{y_{i\cdot k} : k \in r, \}$:

$$y_{i\cdot k} = \begin{cases} y_{ik} & k \in r_i \\ \hat{y}_{ik} & k \in r - r_i \end{cases} \tag{4.1}$$

De esta manera, el total poblacional de interés se estima utilizando los valores $y_{i\cdot k}$ obtenidos.

4.2. Métodos para construir los valores imputados

Cualquiera sea el enfoque de imputación utilizado, es necesario definir qué reglas se utilizarán para construir los valores imputados $\hat{y}_k$, por simplicidad $\hat{y}_k$ hace referencia a $\hat{y}_{ik}$ la i-ésima variable de estudio. En esta sección se pondrá más énfasis en el enfoque combinado, aunque los métodos desarrollados podrán también ser utilizados en el enfoque de imputación completa.

4.2.1. Imputación por regresión

La imputación por regresión es un método de imputación determinística, ya que al repetir el procedimiento se obtienen los mismos valores imputados de y, siendo éstos:

$$\hat{y}_k = \mathbf{x}_k' \hat{\beta}_i \tag{4.2}$$

donde

$$\hat{\beta}_i = \left(\sum_{r_i} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{r_i} a_k \mathbf{x}_k y_k \tag{4.3}$$

es el vector de coeficientes que resulta de la regresión múltiple de los datos y_k con la información $\mathbf{x}_k$ disponible para todos los elementos del conjunto de respuesta al ítem; ponderados por los valores a_k.

En el caso especial de regresión lineal simple con constante, el vector auxiliar de imputación es $\mathbf{x}_k = (1,\, x_k)'$, y los valores imputados son

$$\hat{y}_k = \bar{y}_{r_i;a} - (x_k - \bar{x}_{r_i;a})\, B_{r_i;a} \tag{4.4}$$

donde $\bar{y}_{r_i;a} = \frac{\sum_{r_i} a_k y_k}{\sum_{r_i} a_k}$, $\bar{x}_{r_i;a} = \frac{\sum_{r_i} a_k x_k}{\sum_{r_i} a_k}$, y $B_{r_i;a} = \frac{\sum_{r_i} a_k \left(x_k - \bar{x}_{r_i;a}\right)\left(y_k - \bar{y}_{r_i;a}\right)}{\sum_{r_i} a_k \left(x_k - \bar{x}_{r_i;a}\right)^2}$.

Este método determinístico puede transformarse en estocástico si se adiciona aleatoriamente un término de error. Esto alterará el aspecto del conjunto completo de datos en la dirección de una variabilidad más natural.

$$\hat{y}_k = \mathbf{x}_k' \hat{\beta}_i + e_{0k} \tag{4.5}$$

Donde e_{0k} es seleccionado aleatoriamente del conjunto de residuos $\{e_k : k \epsilon r\}$ tal que $e_k = y_k - \mathbf{x}_k' \hat{\beta}_i$.

Si se practica esta técnica unicamente para obtener estimaciones puntuales, la varianza del estimador imputado aumentará. El uso más importante de esta técnica es la estimación de la varianza de los estimadores ya que el conjunto completo de datos se acercará a contener la "cantidad natural de variación".

4.2.2. *"Ratio Imputation"* e imputación por la media de respondentes

Éstos son casos particulares de imputación por regresión, cuando la variable auxiliar de imputación x es uni-dimensional y siempre positiva, y los ponderadores son

$a_k = \frac{1}{x_k}$. Los valores imputados se convierten en:

$$\hat{y}_k = x_k \hat{\beta}_i = x_k \frac{\sum_{r_i} y_k}{\sum_{r_i} x_k} \tag{4.6}$$

La aplicación de esta técnica a estudios por paneles consiste en considerar como variable auxiliar únicamente los valores de la variable en la ola anterior. En este tipo de encuestas la mayoría de los ítems se repiten en todas las olas. Cuando las respuestas a un ítem repetido están muy correlacionadas en el tiempo, la respuesta en una ola tendrá un gran poder predictivo del faltante en la ola siguiente. Sin embargo, la alta correlación de los respodentes no garantiza que los valores de los no respondentes sean bien predichos. El "ratio" $\hat{\beta}_i$ mide el cambio en la variable de interés de una ola a la siguiente.

Cuando se fija $a_k = 1$ para todos los elementos, los valores imputados resultan ser:

$$\hat{y}_k = \bar{y}_{r_i} \tag{4.7}$$

Esto es, los valores en los ítems faltantes se reemplazan por el valor de la media de los respondentes en dicha variable (imputación por la media de respondentes). A pesar de ser una de las opciones incluidas por defecto como tratamiento de valores faltantes en la mayoría de los paquetes estadísticos, y de ser uno de los métodos de producción de valores sustitutos más utilizado, su uso no es recomendado ya que es un método que subestima la varianza muestral de la variable con valores imputados.

4.2.3. Imputación por el Vecino más Cercano

Los valores creados por este método son valores observados en otros elementos respondentes. Es un método determinístico, y se clasifica dentro de los métodos basados en donantes. Los valores imputados $\hat{y}_k$ han sido observados, por lo que se salva la desventaja de crear valores "imposibles de existir". Sea $l(k)$ el elemento (observado) donante de respuesta para el individuo no respondente k, se verifica:

$$\hat{y}_k = y_{l(k)} \tag{4.8}$$

La decisión acerca de qué elemento donará respuesta a cada uno de los no respondentes recae en la minimización de distancias, y la justificación estadística para esto es que se espera que dos elementos muy cercanos también presenten valores cercanos en la variable a imputar. Para cada elemento no respondente k, el valor a ser imputado resulta del siguiente procedimiento:

- Se calculan las distancias D_{lk} (de acuerdo a alguna medida de distancia) entre k y todos los elementos $l \in r_i$, en base a los valores de los ítems no faltantes.
- El individuo l que donará la respuesta es aquel que haga mínima la distancia D_{lk}.
- Se le asigna como valor imputado del elemento no respondente el valor que presenta el donante en la variable a imputar.

4.2.4. Imputación Cold-Deck

Se define un registro donante por estrato como "registro tipo" en base a fuentes de información externas: datos históricos, distribuciones de frecuencias, etc. El método asigna a los campos a imputar de todos los registros candidatos los valores del registro donante correspondiente al mismo estrato. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible. A partir de este método se originó el procedimiento hot-deck.

4.2.5. Imputación Hot-Deck

De igual manera al método del vecino más cercano, esta forma de imputación está basada en donantes de respuesta, pero la diferencia radica en que la imputación hot-deck es un método aleatorio, y el anterior es determinístico. El valor imputado del elemento k será también:

$$\hat{y}_k = y_{l(k)} \tag{4.9}$$

donde $l(k)$ es un donante seleccionado de manera aleatoria entre un conjunto posible de elementos donantes. El conjunto completo de datos tendrá una apariencia bastante

natural, aunque puede diferir considerablemente del conjunto de datos obtenido mediante respuesta completa. El motivo: aunque la selección del donante sea aleatoria, el conjunto de potenciales donantes está compuesto necesariamente por elementos respondentes, que pueden diferir sustancialmente de aquellos que no responden.

Un procedimiento propuesto por Kalton y Lepkowski, 1989, para imputar la no respuesta al ítem en un panel de dos etapas, usando el valor de la variable en una ola para imputar el faltante en la otra es el siguiente:

- Se categorizan los elementos de la muestra en distintas celdas formadas con la información auxiliar de la que se dispone. En ambas olas la categorización es la misma.

- Se imputa el valor faltante en la segunda ola de cada elemento k aplicando el método hot-deck para las celdas: a un individuo con faltante en la ola 2 de cierta variable se le asignará el valor en la ola 2 de aquel individuo que haya sido clasificado en la misma celda en la ola anterior.

Generalidades de los procedimientos desarrollados

En los métodos desarrollados hasta ahora se realizó la distinción entre aquellos que producen valores imputados determinísticos y aleatorios, y aquellos que producen valores artificiales o se basan en valores observados en otros elementos. En particular, la cercanía de los valores imputados a los reales no observados se basan, para los procedimientos de imputación por regresión y del vecino más cercano, en el supuesto de existencia de una relación fuerte entre los valores de la variable a imputar y el vector auxiliar de imputación. Los procedimientos hot-deck e imputación por la media de respondentes "casi" no utilizan información auxiliar. Estos métodos deberían considerarse únicamente como último recurso en ausencia de información auxiliar para imputar, ya que se corre el riesgo de que los valores imputados no sean "sustitutos cercanos" de los reales. Si no es posible la aplicación de otros procedimientos, éstos al menos lograrán el objetivo de obtener una matriz de datos rectangular.

4.2.6. Grupos de Imputación

Al separar la muestra s en g grupos no solapados s_g , con $g = 1, ..., G$, es posible imputar cada subgrupo por algún método descrito anteriormente (no necesariamente el mismo para todos los grupos). Se pueden distinguir dos razones para utilizar más de un grupo de imputación: pueden existir relaciones diferentes en distintos subgrupos de la muestra o las variables auxiliares necesarias para algún método de imputación pueden no estar disponibles para toda la muestra. Esto puede forzar una jerarquía de métodos de imputación de manera de utilizar los métodos más fuertes como primera opción en uno o más grupos para abarcar gran parte de los elementos no respondentes y luego aplicar progresivamente métodos más débiles a los grupos restantes.

4.2.7. Árboles de clasificación y regresión

Los árboles de clasificación y regresión generan subgrupos de población que contienen elementos homogéneos dentro de ellos y heterogéneos entre distintos subgrupos con respecto a la variable a discriminar, en esta situación dicha variable será la variable que se desea imputar. La idea básica de un modelo de imputación basado en árboles es muy simple: dada una variable de respuesta categórica o continua cuyo valor es faltante y una serie de variables explicativas, el método emplea en primer lugar los registros con valor conocido en la variable respuesta y con dichos registros se construye el árbol que explica su distribución en función de las variables explicativas. Los nodos terminales de este árbol son tratados como clases de imputación. De esta forma, cada registro con valor faltante en la variable respuesta llega a un determinado nodo terminal en función de los valores que posea en las variables explicativas empleadas en la construcción del árbol. Los métodos a aplicar una vez clasificada la unidad en un nodo terminal pueden ser muy diversos: imputación a la categoría más probable, imputación aleatoria en función de la distribución de frecuencias de dicho nodo, imputación hot-deck, imputación al vecino más próximo, entre otras.

4.2.8. Imputación Múltiple

La imputación múltiple propuesta por Rubin (1987) es una técnica en la que los valores faltantes son sustituidos por $m > 1$ valores simulados. Consiste en la imputación

de los casos perdidos a través de la estimación de un modelo aleatorio apropiado realizada m veces y, como resultado, se obtienen m conjuntos de datos completos con los valores imputados. Posteriormente, se lleva a cabo el análisis estadístico en cada una de las m matrices de datos completas y se combinan los resultados con una serie de fórmulas específicas proporcionadas por Little y Rubin (1987). Observando las distintas matrices generadas tras la imputación múltiple se puede tener una idea respecto a la precisión del método de imputación, si no se observan grandes variaciones entre los valores imputados de las distintas matrices se tiene una gran precisión de las estimaciones.

El objetivo de la imputación múltiple es hacer un uso eficiente de la información disponible, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de parámetros. La mayor dificultad de la aplicación de este método de imputación radica en la generación del modelo del cual son simulados los m conjuntos de datos.

4.2.9. Imputación especial: juicio de expertos o con datos históricos

La imputación especial se convierte en una necesidad cuando se cuenta con elementos grandes, influyentes y/o únicos para los cuales no existen "elementos similares" o grupos de referencia. Se reserva a un pequeño número de casos, donde el deseo de proveer el mejor valor imputado posible solo se podrá realizar al estudiar el elemento con detalle, acudiendo a técnicos u otras fuentes.

4.3. Discusión

La mayor preocupación acerca de la imputación es que su uso puede distorsionar las asociaciones entre la variable imputada y las otras variables del conjunto de datos. Si la variable y tiene algunos valores imputados y en el análisis se estima su asociación con x, la asociación entre x e y se verá atenuada hacia cero, a menos que la variable x haya sido utilizada en el proceso de imputación. El mayor desafío de la imputación es mantener intactas todas las asociaciones de potencial interés para el analista en encuestas con gran número de variables. Este reto se vuelve aún más grande en encuestas de panel, dado que los analistas estarán particularmente interesados en las

asociaciones de las variables tanto dentro de la ola como entre olas de la encuesta. Otro aspecto del desafío de imputar es manejar respuestas faltantes para un rango amplio de variables distintas, con distintos patrones de faltantes, para diferentes respondentes (el *"problema del gruyerito"*). Esta es la situación usual en la práctica y se suma al desafío de llevar a cabo imputaciones de manera de no distorsionar las asociaciones. Esta posible distorsión de asociaciones puede llevar a los analistas a preferir descartar la información obtenida de unidades con faltantes en algún ítem y tratarla como una unidad no respondente a la hora de calibrar, que arriesgar la distorsión de asociaciones que puede surgir de una imputación masiva para todos los ítems en la ola faltante.

Capítulo 5

Calibración para la no respuesta de unidades

La teoría de inferencia basada en el diseño muestral, también llamada teoría de aleatorización, es aplicable cuando en la encuesta existe respuesta completa. La aleatoriedad viene dada por el mecanismo de selección reflejado en el diseño muestral, controlado por el investigador; las probabilidades de inclusión (π_k, generadas por el diseño) juegan un rol decisivo en la inferencia estadística en presencia de respuesta completa.

Cuando la no respuesta entra en juego es conveniente pensar que cada individuo tiene su propia probabilidad de respuesta, y a diferencia de la etapa de muestreo, la etapa de respuesta está más allá del control del estadístico: ocurre con probabilidades desconocidas.

Los valores de la variable de estudio y son observados sólo para los elementos k de un subconjunto de la muestra s, llamado conjunto de respondentes r. Cualquiera sea el diseño utilizado, el estimador del total $t_y = \sum_U y_k$ será sesgado (a menos que la no respuesta ocurra de manera totalmente aleatoria).

En el desarrollo de este capítulo no se tomará en cuenta la presencia de no respuesta al ítem, ya que previo a la calibración se suponen aplicados algunos de los métodos de imputación presentados en el capítulo anterior. Durante el desarrollo del procedimiento de calibración, y_k hace referencia a $y_{i \cdot k}$ definido en (4.3).

5.1. Estimación en presencia de no respuesta: la necesidad de calibrar

5.1.1. Sesgo de un estimador del total

Como herramienta para construir un estimador en presencia de no respuesta, debe asumirse un *modelo de respuesta,* que no es más que un conjunto de supuestos sobre la verdadera *distribución de respuesta*, $q(r/s)$, la distribución desconocida de todos los conjuntos posibles de respondentes r dada la muestra s.

Se asume que cada elemento seleccionado en la muestra tiene una probabilidad de responder θ_k, y las unidades responden de manera independiente. De esta manera, el modelo de respuesta asumido es:

$$P(k \in r/s) = \theta_k \quad \forall\, k \in U; \qquad P(k\,\&\,l \in r/s) = \theta_k\theta_l \quad \forall\, k,l \in U \tag{5.1}$$

Suponiendo que el estimador del total poblacional de y es $\hat{t}_y = N\tilde{y}_s = N\,\frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k}$, en presencia de no respuesta solo puede ser calculado en base al conjunto de respondentes r, por lo que se transforma en $\hat{t}_{y1} = N\tilde{y}_r = N\,\frac{\sum_r y_k/\pi_k}{\sum_r 1/\pi_k}$.

Este estimador no es insesgado para estimar t_y. Bajo el modelo de respuesta (5.1) se tiene que:

$$B(\hat{t}_{y1}) = E(\hat{t}_{y1}) - t_y \tag{5.2}$$

$$\begin{aligned} E(\hat{t}_{y1}) = E_{pq}(\hat{t}_{y1}/s) &\doteq N\,\frac{E_{pq}(\sum_r \frac{y_k}{\pi_k}/s)}{E_{pq}(\sum_r \frac{1}{\pi_k}/s)} \\ &= N\,\frac{E_p(\sum_s \frac{y_k\theta_k}{\pi_k})}{E_p(\sum_s \frac{\theta_k}{\pi_k})} = N\,\frac{\sum_U \theta_k y_k}{\sum_U \theta_k} = N\bar{y}_{U;\theta} \end{aligned} \tag{5.3}$$

Siendo E_p la esperanza bajo el diseño $p(s)$, E_q la esperanza bajo $q(r/s)$, donde $q(r/s)$ es la distribución de todos los conjuntos de respondentes r posibles dada la muestra s y E_{pq} la esperanza considerando ambas fuentes de aleatoriedad conjuntamente.

Así el sesgo puede aproximarse:

$$\begin{aligned} B(\hat{t}_{y1}) &\doteq N\bar{y}_{U;\theta} - N\bar{y}_U = N(\bar{y}_{U;\theta} - \bar{y}_U) \\ &= \frac{N}{N\bar{\theta}_U}\left(\sum_U \theta_k y_k - N\bar{\theta}_U\bar{y}_U\right) \\ &= (N-1)\,\frac{S_{y_{U;\theta}}}{\bar{\theta}_U} \\ &= (N-1)\,\frac{S_{y_{U;\theta}}}{\bar{\theta}_U}\,\frac{S_{\theta U}S_{y_U}\bar{y}_U}{S_{\theta U}S_{y_U}\bar{y}_U} \\ &= (N-1)\bar{y}_U R_{y\theta U}\,cv_{\theta U}\,cv_{yU} \\ &\doteq R_{y\theta U}\,cv_{\theta U}\,cv_{yU}\,t_y \end{aligned} \tag{5.4}$$

Luego el sesgo relativo:

$$RB(\hat{t}_{y1}) = \frac{B(\hat{t}_{y1})}{t_y} \doteq R_{y\theta U}\,cv_{\theta U}\,cv_{yU} \tag{5.5}$$

donde $R_{y\theta U}$ es la correlación entre la variable de estudio y y la probabilidad de respuesta θ en la población U; $cv_{\theta U}$ y cv_{yU} son los coeficientes de variación de θ y y. Esta expresión muestra que la correlación $R_{y\theta U}$ constituye un factor clave en el sesgo: cuanto mayor sea la correlación entre la variable de estudio y y la probabilidad de respuesta θ mayor será el sesgo relativo del estimador.

5.1.2. Descomposición del error de estimación

Sea $\hat{t}_y$ el estimador de t_y cuando hay respuesta completa, o sea cuando $r = s$ (este estimador puede ser cualquier estimador insesgado o aproximadamente insesgado bajo repetidas muestras s, extraídas de U); sea $\hat{t}_{y_{NR}}$ el estimador de t_y en presencia de no respuesta. El *error de estimación* de $\hat{t}_{y_{NR}}$ puede expresarse mediante un término que representa el *error muestral* y otro que representa el *error por no respuesta*.

$$Error = \hat{t}_{y_{NR}} - t_y = (\hat{t}_y - t_y) + (\hat{t}_{y_{NR}} - \hat{t}_y) \tag{5.6}$$

Siendo

- $\hat{t}_y - t_y$ el error muestral (el error que surge por elegir y observar una muestra, en vez de observar toda la población).
- $\hat{t}_{y_{NR}} - \hat{t}_y$ el error por no respuesta (error que surge por la no existencia de respuesta completa).

La tendencia central del estimador $\hat{t}_{y_{NR}}$ se puede determinar por su valor esperado y su precisión por su error cuadrático medio. El *sesgo total de* $\hat{t}_{y_{NR}}$ se obtiene calculando el valor esperado bajo los mecanismos de selección y de respuesta de los dos componentes de error previamente definidos:

$$\begin{aligned} B_{pq}(\hat{t}_{y_{NR}}) &= B_{SAM} + B_{NR} \\ &= E(\hat{t}_{y_{NR}} - \hat{t}_y) + E(\hat{t}_y - t_y) \end{aligned} \tag{5.7}$$

El *sesgo muestral*, o error muestral esperado es:

$$B_{SAM} = E_p(\hat{t}_y - t_y) = E_p(\hat{t}_y) - t_y \tag{5.8}$$

El *sesgo de no respuesta* de $\hat{t}_{y_{NR}}$:

$$\begin{aligned} B_{NR} &= E(\hat{t}_{y_{NR}} - \hat{t}_y) \\ &= E_p E_q(\hat{t}_{y_{NR}} - \hat{t}_y/s) \\ &= E_p\left[E_q(\hat{t}_{y_{NR}}/s) - \hat{t}_y\right] \\ &= E_{pq}(\hat{t}_{y_{NR}}/s) - E_p(\hat{t}_y) \\ &= E_p(B_{NR/s}) \end{aligned} \tag{5.9}$$

donde $B_{NR/s} = E_q(\hat{t}_{y_{NR}}/s) - \hat{t}_y$.

El término B_{SAM} (sesgo muestral: esperanza bajo p del error muestral) es cero o irrelevante para la mayoría de los propósitos prácticos, por lo tanto el sesgo de $\hat{t}_{y_{NR}}$ se convierte casi enteramente en el sesgo por no respuesta desconocido.

En la práctica es virtualmente imposible obtener la magnitud de $B_{pq}(\hat{t}_{y_{NR}})$ porque la distribución de la no respuesta $q(r/s)$ nunca se conoce de manera exacta. Sin embargo, frecuentemente no se realizan ajustes por no respuesta dado que se asume (mal o bien) que este sesgo es suficientemente chico. Muy a menudo este supuesto no tiene justificación y resulta en intervalos de confianza no válidos.

La varianza de $\hat{t}_{y_{NR}}$ (también puede expresarse en términos del componente muestral y el componente de no respuesta):

$$\begin{aligned} V_{pq}(\hat{t}_{y_{NR}}) &= E_{pq}[\hat{t}_{y_{NR}} - E_{pq}(\hat{t}_{y_{NR}})]^2 \\ &= V_p E_q(\hat{t}_{y_{NR}}/s) + E_p V_q(\hat{t}_{y_{NR}}/s) \\ &= V_p(B_{NR/s} + \hat{t}_y) + E_p V_q(\hat{t}_{y_{NR}}/s) \\ &= V_{SAM} + V_{NR} \end{aligned} \tag{5.10}$$

donde V_{SAM} es la varianza muestral (varianza de $\hat{t}_y$ bajo todas las posibles muestras s que pueden ser extraídas bajo el diseño muestral; no depende de la distribución de respuesta) y V_{NR} la varianza por no respuesta (involucra el promedio de la varianza del estimador bajo todas las muestras s y todos los conjuntos de respuesta r, más un componente de la variación del sesgo por no respuesta condicional a la muestra s).

Si $B_{NR/s} \doteq 0$ entonces:

$$\begin{aligned} V_{SAM} &= V_p E_q(\hat{t}_{y_{NR}}/s) \\ &= V_p(B_{NR/s} + \hat{t}_y) \doteq V_p(\hat{t}_y) \end{aligned} \tag{5.11}$$

por lo que la varianza de $\hat{t}_{y_{NR}}$ se convierte en $V_{pq}(\hat{t}_{y_{NR}}) \doteq V_p(\hat{t}_y) + E_p V_q(\hat{t}_{y_{NR}}/s)$.

Si $B_{NR/s} \neq 0$,

$$\begin{aligned} V_{SAM} &= V_p(B_{NR/s} + \hat{t}_y) \\ &= V_p(B_{NR/s}) + V_p(\hat{t}_y) + 2Cov_p(\hat{t}_y,\, B_{NR/s}) \end{aligned} \tag{5.12}$$

entonces la varianza de $\hat{t}_{y_{NR}}$ se convierte en $V_{pq}(\hat{t}_{y_{NR}}) = V_p(B_{NR/s}) + V_p(\hat{t}_y) + 2Cov_p(\hat{t}_y,\, B_{NR/s}) + E_p V_q(\hat{t}_{y_{NR}}/s)$.

Esto es, si el sesgo por no respuesta condicional a la muestra s no puede despreciarse, en $V_{pq}(\hat{t}_{y_{NR}})$ aparecen dos términos condicionales $V_p(B_{NR/s}) + 2Cov_p(\hat{t}_y,\, B_{NR/s})$.

La ecuación (5.10) afirma que aún frente a la circunstancia ideal de que la no respuesta no causa sesgo la misma tendrá un efecto de incrementar la varianza en comparación al caso de respuesta completa, donde la $V_{pq}(\hat{t}_{y_{NR}}) = V_{SAM} = V_p(\hat{t}_y)$.

Los resultados presentados en esta sección establecen la necesidad de no ignorar el problema. La magnitud del sesgo introducido a las estimaciones a causa de la no respuesta guarda estrecha relación con la asociación existente (pero desconocida) entre la probabilidad de responder de los elementos en la muestra, y la variable de interés. Resulta necesario entonces aplicar métodos que permitan reducir el sesgo. A su vez, aún en el feliz caso en que el sesgo es aproximadamente cero, la no respuesta hará que incremente la varianza de las estimaciones respecto a estimaciones en presencia de respuesta completa.

5.2. Requisito indispensable: Información auxiliar

La clave para una calibración exitosa es el uso de información auxiliar poderosa; esto reduce tanto el sesgo como la varianza. Tanto en la etapa de construcción del diseño muestral como en la etapa de estimación existen variables que juegan un rol muy importante al respecto. En ambos casos se denominan variables auxiliares porque asisten y mejoran los procedimientos de inferencia. Las variables auxiliares pueden ser variables de registro (extraídas de otras encuestas o registros administrativos), respuestas dadas en olas anteriores, etc.

Cuando las variables auxiliares se utilizan para la construcción del diseño muestral se debe conocer su valor para todos los elementos de la población (por ejemplo para la construcción de estratos). Por contraste cuando éstas se utilizan en la etapa de la estimación no es necesario conocer los valores de la misma para toda la población: alcanza con conocer el total poblacional para cada variable auxiliar, mientras que el conocimiento de los valores individuales de la variable se limita a los elementos respondentes.

Para el cálculo de los ponderadores calibrados hay muy pocas restricciones acerca del vector de información auxiliar a ser utilizado. Aunque pueden ser creados vectores

que incluyan toda la información auxiliar disponible, esta puede no ser la solución preferida. En la construcción del vector de información auxiliar se siguen dos pasos importantes: preparar un inventario de todas la potenciales variables auxiliares y seleccionar aquellas que sean más apropiadas para incluir. Se deben ejercer juicios para seleccionar las variables auxiliares que finalmente serán retenidas para la estimación. Como guía para la constucción del vector auxiliar se deberían tomar en cuenta los siguientes principios:

Principio 1: El vector auxiliar debería permitir estimar la inversa de la probabilidad de respuesta, llamada la *influencia de respuesta* ($\phi_k = \frac{1}{\theta_k}$).

Principio 2: El vector auxiliar debería permitir estimar las principales variables del estudio.

Cuando se cumple el Principio 1, se reduce el sesgo de las estimaciones calibradas para todas las variables de estudio. Esta generalidad es importante ya que en las encuestas grandes se consideran muchas variables de estudio y es necesario la reducción efectiva del sesgo en todas las estimaciones.

Si se cumple el Principio 2, el sesgo se verá reducido en las estimaciones de las variables principales pero quizás no en las estimaciones de las demás.

La clave para obtener una estimación confiable en presencia de no respuesta es la utilización eficiente de la información auxiliar, y en el enfoque de calibración se distinguen tres niveles de información auxiliar:

- Info U: la información está disponible a nivel de la población U; así, $\mathbf{x}_k^\star$ es un vector de dimensión $J^\star \geqq 1$ tal que:

1. el vector de totales poblacionales $\sum_U \mathbf{x}_k^\star$ es conocido.
2. para todo $k \epsilon r$, el vector $\mathbf{x}_k^\star$ es conocido.

El vector auxiliar es en este caso $\mathbf{x} = \mathbf{x}_k^\star$, "vector estrella".

- Info S: la información está disponible a nivel de la muestra s, pero no a nivel poblacional; $\mathbf{x}_k^\circ$ es un vector de dimensión $J^\circ \geqq 1$ tal que:

1. para todo $k \epsilon s$, el vector $\mathbf{x}_k^\circ$ es conocido, sin embargo $\sum_U \mathbf{x}_k^\circ$ es desconocido.
2. para todo $k \epsilon r$, el vector $\mathbf{x}_k^\circ$ es conocido.

El vector auxiliar en este caso es $\mathbf{x} = \mathbf{x}_k^\circ$, "vector luna".

- Info US: se combinan ambos tipos de información para calcular los ponderadores. Una opción es formular el vector auxiliar como $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^\star \\ \mathbf{x}_k^\circ \end{pmatrix}$ de dimensión $J^\star + J^\circ$.

5.3. Estimador puntual bajo calibración

Es deseable que los estimadores afectados por la no respuesta sean útiles para estimar los totales de las variables de interés, que no estén sesgados y que tengan varianza reducida. El sistema de ponderadores que surge del enfoque de calibración verifica también, que al ser aplicado a las variables auxiliares reproduce el input de información auxiliar (aspecto clave en el enfoque de calibración).

Sea w_k el peso calibrado para $k \epsilon r$, luego, el estimador de $t_y = \sum_U y_k$ es:

$$\hat{t}_{y_W} = \sum_r w_k \, y_k \tag{5.13}$$

El caso más general es aquel en que la información disponible es la Info US : donde $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^\star \\ \mathbf{x}_k^\circ \end{pmatrix}$ de dimensión $J^\star + J^\circ$ y el input de información es $\mathbf{X} = \begin{pmatrix} \mathbf{X}^\star \\ \hat{\mathbf{X}}^\circ \end{pmatrix}$.

Se busca el conjunto de valores w_k para todo $k \epsilon r$ que satisfaga la *ecuación de calibración:*

$$\mathbf{X} = \sum_r w_k \, \mathbf{x}_k \tag{5.14}$$

esta ecuación implica que $\begin{pmatrix} \sum_r w_k \mathbf{x}_k^\star \\ \sum_r w_k \mathbf{x}_k^\circ \end{pmatrix} = \begin{pmatrix} \sum_U \mathbf{x}_k^\star \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$, donde $d_k = \frac{1}{\pi_k}$.

Se dice que estos pesos w_k están calibrados al input de información $\mathbf{X}$, ya que cuando se aplican al vector auxiliar reproducen exactamente la información dada en $\mathbf{X}$.

Como resultado de la selección de la muestra, a cada elemento k le corresponde el peso d_k. En presencia de no respuesta, $\sum_r d_k\, y_k$ subestima $\sum_U y_k$, en una magnitud $\sum_{s-r} d_k\, y_k$ (en el caso que la variable de interés tome solamente valores positivos). Es por esto que los d_k deben ser modificados. Se buscarán nuevos pesos que sean mayores que d_k al menos para la mayoría de los respondentes, de manera de compensar la pérdida de unidades. Los nuevos ponderadores $w_k = d_k\nu_k$ se obtienen "aumentando" los pesos originales mediante el factor ν_k, que reflejará las características individuales conocidas de los elementos $k\,\epsilon\, r$ (resumidas en el vector $\mathbf{x}_k$), y puede pensarse como una función del vector auxiliar $\nu_k = F(\lambda'\mathbf{x}_k)$, donde λ es un vector de la misma dimensión que $\mathbf{x}_k$ y se determinará para que se verifique la ecuación de calibración.

Una posible forma para ν_k es $\nu_k = 1+\lambda'\mathbf{x}_k$ (ν_k depende linealmente del valor conocido $\mathbf{x}_k$). Sustituyendo en la ecuación de calibración y despejando λ, se obtiene:

$$\mathbf{X}' = \sum_r d_k\,(1+\lambda_r'\mathbf{x}_k)\,\mathbf{x}_k' \quad \Rightarrow \quad \lambda_r' = (\mathbf{X} - \sum_r d_k\mathbf{x}_k)'(\sum_r d_k\mathbf{x}_k\mathbf{x}_k')^{-1} \qquad (5.15)$$

asumiendo que la inversa de $\sum_r d_k\mathbf{x}_k\mathbf{x}_k'$ existe.

Los nuevos ponderadores son $w_k = d_k + d_k\lambda_r'\mathbf{x}_k$ donde el término $d_k\lambda_r'\mathbf{x}_k$ que se agrega será positivo para la mayoría de los elementos (no necesariamente todos). En resumen, los ponderadores calibrados cuando ν_k depende linealmente del vector conocido $\mathbf{x}_k$ son:

$$w_k = d_k\nu_k; \qquad \nu_k = 1+\lambda'\mathbf{x}_k \qquad (5.16)$$

Aunque a $\hat{t}_{yW} = \sum_r w_k\, y_k$ se lo denomina *estimador calibrado*, éste es en realidad una familia entera de estimadores que corresponden a formulaciones diferentes de $\mathbf{x}_k$ y de $\nu_k = F(\lambda'\mathbf{x}_k)$ (Deville and Särndal (1992); Deville, Särndal and Sautory (1993)).

5.4. Conjuntos alternativos para los pesos calibrados

Los pesos calibrados no son únicos, ya que la única restricción sobre los ponderadores es que verifiquen la ecuación de calibración, y existen muchos conjuntos de ponderadores que la satisfacen dado un vector auxiliar $\mathbf{x}_k$ y un input de información $\mathbf{X}$.

5.4.1. Pesos iniciales alternativos

Es posible cumplir el requerimiento de que los ponderadores finales calibren a la información $\mathbf{X}$ aún cuando los ponderadores iniciales no sean los d_k. Sean $d_{\alpha k}$, para $k \epsilon r$, cualquier conjunto de ponderadores, se tiene que los pesos $w_{\alpha k}$ definidos por las siguientes condiciones verifican la ecuación de calibración:

$$w_{\alpha k} = d_{\alpha k}\nu_k; \qquad \nu_k = 1 + \lambda'_\alpha \mathbf{x}_k$$

$$\lambda'_\alpha = (\mathbf{X} - \sum_r d_{\alpha k}\mathbf{x}_k)'(\sum_r d_{\alpha k}\mathbf{x}_k\mathbf{x}'_k)^{-1} \tag{5.17}$$

Sustituyendo en la ecuación de calibración:

$$\begin{aligned}
\sum_r w_{\alpha k}\mathbf{x}'_k &= \sum_r d_{\alpha k}(1 + \lambda'_\alpha \mathbf{x}_k)\mathbf{x}'_k \\
&= \sum_r d_{\alpha k}\mathbf{x}'_k + \sum_r d_{\alpha k}\left[(\mathbf{X} - \sum_r d_{\alpha k}\mathbf{x}_k)'(\sum_r d_{\alpha k}\mathbf{x}_k\mathbf{x}'_k)^{-1}\right]\mathbf{x}_k\mathbf{x}'_k \\
&= \sum_r d_{\alpha k}\mathbf{x}'_k + (\mathbf{X} - \sum_r d_{\alpha k}\mathbf{x}_k)'\sum_r d_{\alpha k}\mathbf{x}_k\mathbf{x}'_k(\sum_r d_{\alpha k}\mathbf{x}_k\mathbf{x}'_k)^{-1} \\
&= \sum_r d_{\alpha k}\mathbf{x}'_k + \mathbf{X}' - \sum_r d_{\alpha k}\mathbf{x}'_k \\
&= \mathbf{X}
\end{aligned} \tag{5.18}$$

Por lo tanto, cualquiera sea el conjunto de ponderadores iniciales, los pesos calibrados estimarán sin error los totales poblacionales auxiliares conocidos.

5.4.2. Variables auxiliares alternativas: vector instrumento

También es posible cumplir el requerimiento de que los ponderadores finales calibren a la información $\mathbf{X}$ aún cuando los pesos calibrados estén definidos en base a un vector distinto al vector auxiliar $\mathbf{x}$. Se define al vector $\mathbf{z}$, de manera que tenga las mismas dimensiones que $\mathbf{x}$, $\mathbf{z}$ puede ser cualquier función específica de $\mathbf{x}_k$ o de otras variables que refieran a los elementos k. Dicho vector, $\mathbf{z}$, es conocido por el nombre de *vector instrumento para la calibración*.

$$w_k = d_k \nu_k; \qquad \nu_k = 1 + \lambda' \mathbf{z}_k$$

$$\lambda' = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1} \tag{5.19}$$

Los ponderadores calibrados definidos por (5.19) verifican la ecuación de calibración:

$$\begin{aligned}
\sum_r w_k \mathbf{x}_k' &= \sum_r d_k (1 + \lambda' \mathbf{z}_k) \mathbf{x}_k' \\
&= \sum_r d_k \mathbf{x}_k' + \sum_r d_k \left[(\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1} \right] \mathbf{z}_k \mathbf{x}_k' \\
&= \sum_r d_k \mathbf{x}_k' + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' \sum_r d_k \mathbf{z}_k \mathbf{x}_k' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1} \\
&= \sum_r d_k \mathbf{x}_k' + \mathbf{X}' - \sum_r d_k \mathbf{x}_k' \\
&= \mathbf{X}
\end{aligned} \tag{5.20}$$

Como demuestra la ecuación (5.20), el único requisito para que los ponderadores calculados en base al vector instrumento $\mathbf{z}$ verifiquen la ecuación de calibración es que la matriz $\sum_r d_k \mathbf{z}_k \mathbf{x}_k'$ sea invertible.

5.4.3. Ponderadores calibrados alternativos

Las propiedades demostradas en las subsecciones anteriores pueden verificarse de manera simultánea.

Los ponderadores $w_{\alpha k}$ satisfacen la ecuación de calibración para cualquier peso inicial

positivo $d_{\alpha k}$ y cualquier vector instrumento $\mathbf{z}_k$, siempre que la matriz $\sum_r d_{\alpha k}\,\mathbf{z}_k\mathbf{x}_k'$ sea invertible. Cuando $d_{\alpha k} = d_k$ y $\mathbf{z}_k = \mathbf{x}_k$ los ponderadores resultantes son conocidos como *pesos estándar*, definidos por (5.16).

Sean $d_{\alpha k}$, para $k\,\epsilon\,r$, ponderadores iniciales cualquiera, y $\mathbf{z}_k$ un vector instrumento, se tiene que los ponderadores calibrados $w_{\alpha k}$ están definidos por:

$$w_{\alpha k} = d_{\alpha k}\nu_k; \qquad \nu_k = 1 + \lambda_\alpha'\mathbf{z}_k$$

$$\lambda_\alpha' = (\mathbf{X} - \sum_r d_{\alpha k}\,\mathbf{x}_k)'\,(\sum_r d_{\alpha k}\,\mathbf{z}_k\mathbf{x}_k')^{-1} \tag{5.21}$$

En resumen para calcular los pesos calibrados se deberán especificar:

- los ponderadores iniciales $d_{\alpha k}$
- el vector auxiliar $\mathbf{x}_k$ y el input de información correspondiente $\mathbf{X}$
- los valores del vector instrumento $\mathbf{z}_k$ si es que estos difieren de $\mathbf{x}_k$

En resumen, existen muchos sistemas de ponderadores calibrados para un input de información $\mathbf{X}$ dado, debido a la libertad de elección de $d_{\alpha k}$ y $\mathbf{z}_k$[1], donde cada uno de estos sistemas se corresponde con un estimador calibrado $\hat{t}_{y_W} = \sum_r w_k y_k$. El hecho de que cada uno de estos sistemas de ponderadores w_k sean pesos calibrados implica que todos estimarán sin error totales conocidos de las variables auxiliares (ya sea a nivel muestral o poblacional), por lo que se espera que los estimadores resultantes no difieran sustancialmente.

5.5. Análisis del sesgo por no respuesta en el marco de la calibración

Limitar el sesgo de las estimaciones en presencia de no respuesta se tornará en la mayor precupación. La minimización de la varianza pasará a segundo plano ya que

[1] para las condiciones establecidas en (5.21)

de nada sirve que un estimador presente varianza chica cuando está fuertemente sesgado. Para lograr este fin es imprescindible la utilización de información auxiliar.

Särndal y Lundström (2005) realizaron un estudio de Simulación Monte Carlo a través del cual se obtiene evidencia empírica de la fuerte relación existente entre el sesgo del estimador de calibración que proviene de la no respuesta y la información auxiliar utilizada para calibrar.

Se desarrollará la manera de minimizar el sesgo del estimador calibrado con la información auxiliar US. Aunque no será posible obtener una expresión exacta del sesgo para demostrar de qué manera éste depende del vector auxiliar elegido, Lundström (1997) obtiene una aproximación cercana que denomina *sesgo aproximado (AB)*.

$$B_{pq}(\hat{t}_{y_W}) = E_{pq}(\hat{t}_{y_W}) - t_y \doteq AB(\hat{t}_{y_W}) \tag{5.22}$$

$$AB(\hat{t}_{y_W}) = -\sum_U (1-\theta_k) e_{\theta k} \tag{5.23}$$

donde θ_k es la probabilidad de respuesta del elemento k y:

$$e_{\theta k} = y_k - \mathbf{x}_k' \, \mathbf{B}_{U;\theta} \tag{5.24}$$

$$\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} \left(\sum_U \theta_k \mathbf{z}_k y_k \right) \tag{5.25}$$

De este planteo surge:

- Si θ_k es constante para todo $k \, \epsilon \, U$, $\hat{t}_{y_W}$ será aproximadamente insesgado.
- El sesgo de $\hat{t}_{y_W}$, o cualquier estimador alternativo, no se anula cuando existe no respuesta, excepto en circunstancias atípicas. Este sesgo es el causado por la no respuesta. El sesgo aproximado es independiente del diseño muestral utilizado para obtener la muestra s, aunque el estimador puntual $\hat{t}_{y_W} = \sum_r w_k y_k$ dependa del diseño a través de los ponderadores d_k. Para cualquier diseño, el sesgo aproximado siempre será el mismo mientras el vector $\mathbf{x}_k$ sea el mismo. Sin embargo, dependerá de la distribución de respuesta desconocida.
- No existe ninguna condición particular acerca de la composición del vector auxiliar $\mathbf{x}_k$: el sesgo aproximado será el mismo para cualquier vector de variables

auxiliares $\mathbf{x}_{0k}$; ya sea que contenga información solo hasta nivel de la muestra ($\mathbf{x}_k = \mathbf{z}_k = \mathbf{x}_k^\circ = \mathbf{x}_{0k}$) o hasta el nivel poblacional ($\mathbf{x}_k = \mathbf{z}_k = \mathbf{x}_k^\star = \mathbf{x}_{0k}$). Como herramienta para controlar el sesgo, el uso de información auxiliar a nivel de la muestra es tan efectivo como contar con información a nivel de la población. (Aún cuando la forma exacta del sesgo apenas difiere entre Info U e Info S, la aproximación de primer orden del sesgo resulta ser la misma). Para reducir el sesgo se necesita identificar la información auxiliar que cubra al menos el nivel de la muestra. La perspectiva será distinta al enfocar atención en reducir la varianza de las estimaciones, ya que habrá ventajas definidas de extender la información auxiliar a nivel de la población.

La efectiva eliminación del sesgo dependerá de propiedades del vector auxiliar $\mathbf{x}_k$. Si las probabilidades de respuesta son constantes en toda la población el sesgo es cero pero bajo condiciones menos severas de la distribución de respuesta el sesgo aproximado puede llegar a ser cero.

Cuanto mayor sea la *influencia de respuesta* del elemento k, ϕ_k, menor será su probabilidad de respuesta. En el contexto de este análisis, la diferencia entre "ponderador" e "influencia" radica en que los ponderadores son calculables, mientras que las influencias son cantidades desconocidas. A continuación se presentan dos proposiciones para calcular cantidades apropiadas de ϕ_k que puedan ser utilizadas en las estimaciones:

- El sesgo aproximado de $\hat{t}_{y_W}$ es cero si la influencia de respuesta tiene la forma:

$$\phi_k = 1 + \lambda' \mathbf{z}_k \qquad k \,\epsilon\, U \tag{5.26}$$

  para algún vector constante no aleatorio λ que no depende de k. Si $\mathbf{z}_k = \mathbf{x}_k$, la influencia de respuesta se transforma en $\phi_k = 1 + \lambda' \mathbf{x}_k$. Aún después de elegir el vector auxiliar $\mathbf{x}_k$, como último recurso para controlar el sesgo, se puede elegir apropiadamente el vector $\mathbf{z}_k$.

- El sesgo aproximado de $\hat{t}_{y_W}$ es cero si los valores de la variable en estudio tienen la forma:

$$y_k = \beta' \mathbf{x}_k \qquad k \,\epsilon\, U \tag{5.27}$$

para algún vector constante β que no dependa de k.

La primera proposición muestra lazos deseables entre la influencia de respuesta y el vector auxiliar (o el de instrumento) y la segunda muestra la relación deseable entre la variable de estudio y el vector auxiliar. La relación lineal perfecta expresada en la última proposición nunca se cumple en la práctica.

5.6. Varianza y su estimación

La estimación de la varianza del estimador es necesaria por dos motivos, primero para indicar la precisión de $\hat{t}_{y_W}$ y segundo para calcular un intervalo de confianza centrado en la estimación. Para este desarrollo se utilizará información auxiliar a nivel US, pesos iniciales $d_{\alpha k} = d_k$, y $\nu_k = 1 + \lambda' \mathbf{z}_k$.

Proposición 1

La varianza de $\hat{t}_{y_W}$ se estima mediante:

$$\hat{V}(\hat{t}_{y_W}) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{5.28}$$

La estimación del componente correspondiente a la varianza muestral es:

$$\hat{V}_{SAM} = \sum \sum_r (d_k d_l - d_{kl})(\nu_k \hat{e}_k^\star)(\nu_l \hat{e}_l^\star) - \sum_r d_k(d_k - 1)\nu_k(\nu_k - 1)(\hat{e}_k^\star)^2 \tag{5.29}$$

La estimación del componente de la varianza de no respuesta es:

$$\hat{V}_{NR} = \sum_r \nu_k(\nu_k - 1)(d_k \hat{e}_k)^2 \tag{5.30}$$

donde:

$$\hat{e}_k^\star = y_k - (\mathbf{x}_k^\star)' \mathbf{B}_{r;d\nu}^\star \tag{5.31}$$

$$\hat{e}_k = y_k - \mathbf{x}_k' \mathbf{B}_{r;d\nu} = y_k - (\mathbf{x}_k^\star)' \mathbf{B}_{r;d\nu}^\star - (\mathbf{x}_k^\circ)' \mathbf{B}_{r;d\nu}^\circ \tag{5.32}$$

$$\mathbf{B}_{r;d\nu} = \begin{pmatrix} \mathbf{B}_{r;d\nu}^\star \\ \mathbf{B}_{r;d\nu}^\circ \end{pmatrix} = \left(\sum_r d_k \nu_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} \left(\sum_r d_k \nu_k \mathbf{z}_k y_k \right) \tag{5.33}$$

El intervalo de confianza para $\hat{t}_{y_W}$ al $(1-\alpha)$ es $\hat{t}_{y_W} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{V}\left(\hat{t}_{y_W}\right)}$, donde $z_{\frac{\alpha}{2}}$ es el score normal estándar. Este nivel de confianza es aproximado, y no exacto, ya que $\hat{t}_{y_W}$ no tiene distribución normal exacta, además de contener indefectiblemente cierto sesgo a causa de la no respuesta. Para confiar en el intervalo se debe asegurar que el sesgo de $\hat{t}_{y_W}$ sea cercano a cero. Si esto no se cumple, el intervalo tenderá a estar centrado en un valor incorrecto. Es por esto que una reducción efectiva del sesgo es un pre requisito para lograr intervalos de confianza que tengan sentido.

El estimador de la varianza se basa en el paralelismo que existe con el estimador bajo el caso más favorable (pero imposible): distribución de respuesta conocida, o sea $q(r/s)$ conocida, con $P(k\,\epsilon\, r/s) = \theta_k$ y $P(k\,y\,l\,\epsilon\, r/s) = \theta_{kl} = \theta_k\theta_l$.

Así se construye el estimador de regresión generalizado en dos fases cuya utilidad se limitará a ser utilizado como herramienta para construir el estimador de la varianza de $\hat{t}_{y_W}$. Sea:

$$\hat{t}_{y_{G2P}} = \sum_r w_{G2P,k} y_k \tag{5.34}$$

donde:

$$w_{G2P,k} = d_k \phi_r g_{k\phi} \tag{5.35}$$

$$g_{k\phi} = 1 + \lambda_r' \mathbf{z}_k \tag{5.36}$$

$$\lambda_r' = \left(\mathbf{X} - \sum_r d_k \phi_k \mathbf{x}_k\right)' \left(\sum_r d_k \phi_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1} \tag{5.37}$$

Las dos fases de selección le dan al elemento k el ponderador combinado de selección: $(1/\pi_k)\,(1/\theta_k) = d_k\phi_k$, siendo $g_{k\phi}$ un ajuste que proviene de la calibración. Linealizando se derivan los dos componentes de la varianza aproximada de $\hat{t}_{y_{G2P}}$:

$$\hat{V}\left(\hat{t}_{y_{G2P}}\right) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{5.38}$$

$$\hat{V}_{SAM} = \sum \sum_r (d_k d_l - d_{kl})(\phi_k \hat{e}_{k\phi}^\star)(\phi_l \hat{e}_{l\phi}^\star) - \sum_r d_k(d_k - 1)\phi_k(\phi_k - 1)(\hat{e}_{k\phi}^\star)^2 \tag{5.39}$$

$$\hat{V}_{NR} = \sum_r \phi_k(\phi_k - 1)(d_k \hat{e}_{k\phi})^2 \tag{5.40}$$

donde:

$$\hat{e}_{k\phi}^\star = y_k - (\mathbf{x}_k^\star)' \mathbf{B}_{r;d\phi}^\star \tag{5.41}$$

$$\hat{e}_{k\phi} = y_k - \mathbf{x}_k' \mathbf{B}_{r;d\phi} = y_k - (\mathbf{x}_k^{\star})' \mathbf{B}_{r;d\phi}^{\star} - (\mathbf{x}_k^{\circ})' \mathbf{B}_{r;d\phi}^{\circ} \tag{5.42}$$

$$\mathbf{B}_{r;d\phi} = \begin{pmatrix} \mathbf{B}_{r;d\phi}^{\star} \\ \mathbf{B}_{r;d\phi}^{\circ} \end{pmatrix} = \left(\sum_r d_k \phi_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} \left(\sum_r d_k \phi_k \mathbf{z}_k y_k \right) \tag{5.43}$$

La similitud entre $\hat{t}_{y_W}$ y $\hat{t}_{y_{G2P}}$ radica en que ambos están calibrados por el mismo input de información $\mathbf{X} = \begin{pmatrix} \mathbf{X}^{\star} \\ \hat{\mathbf{X}}^{\circ} \end{pmatrix}$. Esta información es el determinante principal de la varianza de los dos estimadores. La diferencia es que los ponderadores de $\hat{t}_{y_{G2P}}$ dependen explícitamente de las influencias, ϕ_k, y los de $\hat{t}_{y_W}$ no lo hacen. Como la influencia de respuesta no es conocida, ni $\hat{t}_{y_{G2P}}$ ni su varianza son calculables; en cambio $\hat{t}_{y_W}$ sí se puede calcular, pero para estimar su varianza aparece el obstáculo de la distribución de respuesta desconocida.

Proposición 2

Sea $\hat{\phi}_k$ un valor aproximado para ϕ_k.

$\hat{t}_{y_{G2P/\phi=\hat{\phi}}} = \hat{t}_{y_W}$ se cumple para $\phi_k = \hat{\phi}_k = \nu_k$ con $\nu_k = 1 + \lambda_r' \mathbf{z}_k$.

Sea $g_{k\hat{\phi}}$ el valor de $g_{k\phi}$ cuando se fija $\phi_k = \hat{\phi}_k = \nu_k$. Usando la ecuación de calibración $\sum_r d_k \nu_k \mathbf{x}_k = \mathbf{X}$ se obtiene:

$$g_{k\hat{\phi}} = 1 + \left(\mathbf{X} - \sum_r d_k \nu_k \mathbf{x}_k \right)' \left(\sum_r d_k \nu_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} \mathbf{z}_k = 1 \tag{5.44}$$

ya que $\left(\mathbf{X} - \sum_r d_k \nu_k \mathbf{x}_k \right) = 0$.

Por lo tanto

$$\hat{t}_{y_{G2P/\phi=\hat{\phi}}} = \sum_r d_k \nu_k g_{k\hat{\phi}} y_k = \sum_r d_k \nu_k y_k = \sum_r w_k y_k = \hat{t}_{y_W} \tag{5.45}$$

Sustituyendo $\hat{\phi}_k = \nu_k$ en el estimador de la varianza se obtiene:

$$\hat{V}\left(\hat{t}_{y_{G2P/\phi=\hat{\phi}}} \right) = \hat{V}\left(\hat{t}_{y_W} \right) \tag{5.46}$$

El resultado de esta técnica es lograr que el cálculo de $\hat{t}_{y_W}$ y $\hat{V}\left(\hat{t}_{y_W}\right)$ no requiera conocer el valor de las influencias desconocidas ϕ_k.

Generalización

Cuando se conoce la distribución de respuesta y sus influencias ϕ_k, el sesgo de $\hat{t}_{y_{G2P}}$ obtenido como promedio de todas las posibles muestras s y conjuntos de respuesta r, es cercano a cero. La igualdad de $\hat{t}_{y_{G2P}} = \hat{t}_{y_W}$ se obtuvo para una muestra y un conjunto de respuesta particular, no se puede decir que $\hat{t}_{y_W}$ es insesgado para todas las muestras y todas las distribuciones de respuesta. De hecho, se sabe que $B_{pq}(\hat{t}_{y_W}) = -\sum_U (1-\theta_k) e_{\theta k} \neq 0$.

Aún cuando las aproximaciones de $\hat{\phi}_k$ sean sustitutos mediocres de las influencias desconocidas, la estimación de la varianza de $\hat{t}_{y_W}$ obtenida por este procedimiento de sustitución será una indicación valiosa de la precisión de $\hat{t}_{y_W}$, y podrá utilizarse para construir intervalos de confianza. Si el sesgo de $\hat{t}_{y_W}$ es modesto, el intervalo de confianza será aproximadamente válido, con un nivel de confianza real que se aproxima al indicado $(1-\alpha)$.

5.7. Ejemplos de Estimadores Calibrados

El enfoque de calibración abarca a una familia de estimadores $\hat{t}_{y_W}$cuyos miembros corresponden a diferentes inputs de información. Se desarrollarán en esta sección algunas formulaciones del enfoque calibrado.

Se asume que se extrae una muestra s de tamaño n de una población U con N elementos bajo un diseño $p(s)$ que genera los ponderadores $d_k = \frac{1}{\pi_k}$. Se está frente a una situación de no respuesta, donde el conjunto de respondentes es de tamaño m menor que n.

5.7.1. El vector auxiliar más simple

Si se considera el vector auxiliar constante para todos los elementos: $\mathbf{x}_k = \mathbf{x}_k^\star = 1$, el estimador de calibración toma la forma:

$$\hat{t}_{y_W} = \sum_r w_k y_k = \sum_r d_k \frac{N}{\sum_r d_k} y_k = N\,\bar{y}_{r:d} \tag{5.47}$$

Asumir información auxiliar constante para todos los elementos es equivalente a no realizar ningún tratamiento para la no respuesta. El sesgo de este estimador puede ser grande debido a la gran debilidad de la información auxiliar, una aproximación de éste se estableció en (5.4).

5.7.2. Post estratificación

La información auxiliar utilizada para la calibración es la clasificación de cada uno de los elementos de la población en P grupos o categorías exclusivas y exhautivas, que pueden ser grupos de edad, nivel educativo, nivel socio económico, entre otras, o la interacción entre más de una de ellas. Para el caso de info U, $U = \bigcup_{p=1}^{P} U_p$ y $N = \sum_{p=1}^{P} N_p$ siendo N_p es el tamaño poblacional conocido del grupo U_p .

De esta manera el vector auxiliar para el elemento k está definido por:

$$\mathbf{x}_k^\star = \gamma_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk})' \tag{5.48}$$

$$\gamma_{pk} = \begin{cases} 1 & \text{si } k \in \text{ grupo } p \\ 0 & \text{o/c} \end{cases}$$

y el vector de totales poblacionales es

$$\mathbf{X} = \sum_U \mathbf{x}_k^\star = (N_1, ..., N_p, ..., N_P)' \tag{5.49}$$

Sea r_p el conjunto de respondentes del grupo p; el conjunto total de respondentes es $r = \bigcup_{p=1}^{P} r_p$, y sean $d_{\alpha k} = d_k$, $\mathbf{z}_k = \mathbf{x}_k$. Para cada grupo p, $\nu_k = \frac{N_p}{\sum_{r_p} d_k}$, esto es, dentro de cada grupo el vector auxiliar es constante para todos los respondentes. En otras

palabras, se está afirmando que la no respuesta ocurre de forma aleatoria dentro de cada grupo. El estimador calibrado se transforma en

$$\hat{t}_{y_W} = \sum_{p=1}^{P} N_p \, \bar{y}_{r_p:d} \tag{5.50}$$

siendo $\bar{y}_{r_p:d} = \sum_{r_p} d_k y_k / \sum_{r_p} d_k$ la media de y para los respondentes del grupo p con los pesos del diseño. En este estimador se requiere que $\sum_{r_p} d_k > 0$ para $p = 1, ..., P$.

5.7.3. Raking

De igual manera a la post estratificación, la información auxiliar que da lugar al raking está basada en dos o más variables auxiliares categóricas. Por simplicidad el desarrollo se realiza para el caso de dos variables auxiliares, pero puede extenderse a una clasificación múltiple fácilmente. Una de las variables, digamos la variable 1, presenta P categorías indexadas con $p = 1, ..., P$ y la otra, variable 2, H modalidades $h = 1, ..., H$. Todos los elementos respondentes k pueden ser asignados a una de las $P \, \mathrm{x} \, H$ celdas que resultan de la interacción entre las dos variables.

De esta manera el vector auxiliar para el elemento k está definido por:

$$\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk}, \delta_{1k}, ..., \delta_{hk}, ..., \delta_{(H-1)k})' \tag{5.51}$$

$$\gamma_{pk} = \begin{cases} 1 & \text{si } k \in \text{la modalidad } p \text{ de la variable 1} \\ 0 & \text{o/c} \end{cases}$$

$$\delta_{hk} = \begin{cases} 1 & \text{si } k \in \text{la modalidad } q \text{ de la variable 2} \\ 0 & \text{o/c} \end{cases}$$

Cada elemento k es clasificado en una única modalidad de cada una de las variables auxiliares, por lo que $\sum_{p=1}^{P} \gamma_{pk} = \sum_{h=1}^{H} \delta_{hk} = 1$, y la matriz a ser invertida para el cálculo de los ponderadores calibrados es singular. Para evitarlo, no resulta en pérdida de

información alguna desechar una modalidad cualquiera, así, el vector $\mathbf{x}_k$ definido por (5.51) tiene dimensión $P + (H - 1)$.

Considerando información a nivel poblacional, el input de información $\mathbf{X}$ está determinado por:

$$\mathbf{X} = \sum_U \mathbf{x}_k^\star = (N_{1\bullet}, ..., N_{P\bullet}, N_{\bullet 1}, ..., N_{\bullet(H-1)})' \tag{5.52}$$

Siendo $N_{p\bullet} = \sum_{h=1}^{H} N_{ph}$ para $p = 1, ..., P$ y $N_{\bullet h} = \sum_{p=1}^{P} N_{ph}$ para $h = 1, ..., H - 1$ los totales poblacionales marginales y N_{ph} el total poblacional correspondiente al cruce de modalidades p y h. Se obtiene de esta manera el procedimiento conocido como *calibración en totales marginales* o *raking*.

Este procedimiento se aplica a las cantidades observadas m_{ph} (cantidad de respondentes correspondientes a la categoría p de la variable 1 y h de la variable 2) para, iterativamente, calcular estimaciones que satisfagan las restricciones sobre las marginales, utilizando una serie de constantes multiplicativas de las filas y de las columnas, de manera que en cada iteración se ajuste alguna de las marginales, hasta que se encuentra un valor para w_k que ajusta a todas las marginales en forma simultánea. Este ajuste iterativo proporcional es utilizado para ajustar las celdas al total de las marginales. La expresión analítica de los ponderadores w_k no es sencilla, pero su cálculo no presenta mayores dificultades.

Discusión

El raking permite obtener ponderadores calibrados en caso de no conocer los totales poblacionales correspondientes a cada celda, pero si conociendo los totales marginales. Esta situación es frecuente en la práctica, por ejemplo, cuando se consideran variables auxiliares de distintas fuentes. En este caso es posible identificar a qué categoría pertenece cada elemento en cada variable, pero no a qué celda en la doble clasificación.

Asimismo, en casos en que N_{ph} sea una cantidad conocida, el raking puede ser una opción preferible a la post estratificación incompleta cuando las celdas contengan can-

tidades muy pequeñas, o incluso celdas vacías. Esta dificultad se salvaría colapsando modalidades que contengan pocas observaciones para luego calcular el estimador de post estratificación.

5.8. Estimadores calibrados en encuestas de panel

La reiteración de entrevistas y la inevitable variación en el tiempo de la población, características por definición de las encuestas de panel, profundizan los mecanismos que llevan a la no respuesta de unidades. La no respuesta es el principal obstáculo para la obtención de estimadores que logren captar la dinámica de la población objetivo. La aplicación del enfoque de calibración en la instancia de estimación de medidas de interés se convierte prácticamente en una necesidad. Las formulaciones de estimadores calibrados desarrolladas en las secciones anteriores pueden aplicarse a estudios de paneles debiendo prestar particular atención a dos aspectos fundamentales: uno refiere a la definición del conjunto de respondentes en cada ola, y el otro a la selección de variables auxiliares.

El conjunto de respondentes r no es único, habrá uno para cada ola definidos como $r_1, r_2, ..., r_o$, siendo o la cantidad de olas del panel. Para los paneles en los que solo se admite no respuesta por desgaste se cumple que $s \supseteq r_1 \supseteq r_2 ... \supseteq r_o$, por lo que la estimación de totales y de cambios estará basada en conjuntos cada vez más pequeños de respondentes con el transcurso de las olas. Si se admite la no respuesta episódica, los totales de interés se estimarán en base al conjunto de respondentes de cada ola, pero para la estimación de cambios se debe utilizar la intersección de los conjuntos de respondentes en las olas consideradas.

La selección de variables auxiliares a ser utilizadas en el procedimiento de calibración requiere particular atención. Por un lado, se cuenta con un conjunto rico de variables auxiliares compuesto por la información coleccionada en olas previas. Esto no implica que todas puedan (y deban) ser incluidas en el vector auxiliar. La información auxiliar a ser utilizada en la calibración para la estimación de cambios debe cumplir los requisitos de ser estable en el tiempo, además de caracterizar a la población objetivo y no podrán utilizarse como variables auxiliares aquellas que sean objeto de medición de cambios.

Una forma de inclusión de respuestas en olas previas como información auxiliar para la calibración fue desarrollada por Little y David (1983) para el caso de no respuesta por desgaste:

Ola 1: En esta ola la única información disponible es el valor que toman las variables auxiliares utilizadas para estratificar. Esta información está disponible para todas las unidades de la muestra. De esta manera se obtienen los pesos correspondientes a la ola 1: w_1.

Ola 2: Las variables auxiliares para calcular los pesos de las unidades que se pierden en la segunda ola serán las del diseño y sus respuestas en la primera ola x_1. En base a esta información se calculan los nuevos pesos w_2, que tendrán un componente que proviene del ajuste por la pérdida de unidades en la primer ola, w_1, y otro calculado en esta instancia para ajustar por las unidades perdidas en la segunda ola, $w_{2,1}$, tomando como información auxiliar las respuestas en la ola 1.

Ola 3: Las variables auxiliares que servirán para calcular los pesos $w_{3,21}$ por la pérdida de unidades específica de esta ola serán las del diseño, x_1, y la información recopilada en la Ola 2, x_2.

La no respuesta en panel (manifiestada bajo los patrones de *no respuesta por desgaste* y *no respuesta episódica*) genera en cada ola, un conjunto de respondentes r_i, todos incluidos en la muestra s. Esto requerirá el cálculo de ponderadores calibrados particulares a cada individuo respondente en cada ola.

Estimación transversal

Para la estimación transversal se calcularán los ponderadores de las unidades respondentes en cada ola como se presentó en las secciones anteriores. El estimador calibrado del total correspondiente a la ola i, $\hat{t}_{y_{Wi}}$ se define por:

$$\hat{t}_{y_{Wi}} = \sum_{r_i} w_{k_i} y_{k_i} \tag{5.53}$$

siendo w_{k_i} los ponderadores calibrados del elemento k respondente en la ola i, y y_{k_i} el valor de la variable de interés para este individuo en dicha ola.

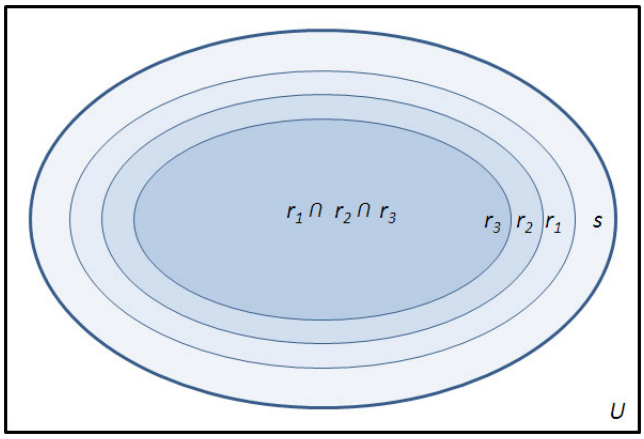
Estimación longitudinal

Adicionalmente, para las estimaciones longitudinales será necesario el cálculo de un nuevo conjunto de ponderadores aplicables únicamente a las unidades respondentes en todas las instancias de las que se quiere medir el cambio.

Cuando el patrón de respuesta admite únicamente no respuesta por desgaste, la medición de cambios longitudinales entre dos olas se hará sobre aquellos elementos respondentes en ambas olas: *respondentes simultáneos*.

Los respondentes de i-ésima ola también fueron respondentes en las olas anteriores $i-1$, $i-2$, ...,1, por lo tanto, los cambios solamente podrán ser medidos para las unidades respondentes en la ola más reciente. En la figura (5.1) puede verse la representación gráfica de un estudio de panel de tres olas con patrón de respuesta por desgaste. A modo de ejemplo, si el interés radica en la estimación de cambios de una variable entre las olas 1 y 3, se calibrarán los cambios individuales para cada respondente de la ola 3.

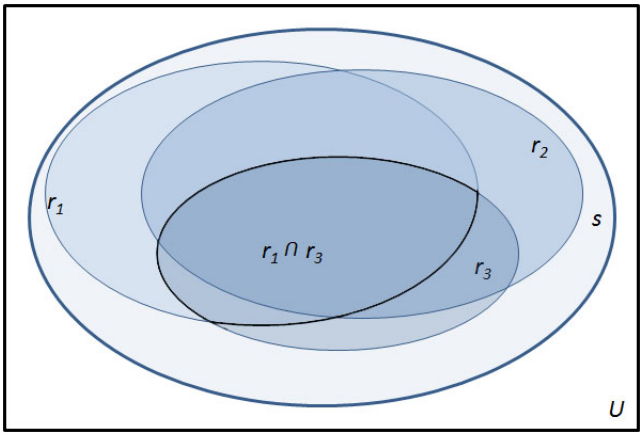
Figura 5.1: Conjunto de respondentes con patrón de respuesta: desgaste



Frente a un patrón de respuesta episódica, los cambios entre dos olas serán medidos en los respondentes simultáneos, pero en este caso, este conjunto no necesariamente

coincide con la ola más reciente de las sujetas a medición. Siguiendo el mismo ejemplo, la estimación de cambios entre la primera y la tercera ola se realizará en base al conjunto marcado en la figura (5.2).

Figura 5.2: Conjunto de respondentes con patrón de respuesta: episódica



Luego de definir el conjunto de respondentes simultáneos a las olas j y $j+h$, la estimación de las diferencias de la variable y entre dichas olas se obtiene mediante la siguiente fórmula:

$$\begin{aligned} \hat{A}_{W_{j,j+h}} &= \sum_{r_j \cap r_{j+h}} (y_{(j+h)\,k} - y_{j\,k})\, w_{(j,j+h)\,k;} \\ &= \sum_{r_j \cap r_{j+h}} a_{(j,j+h)\,k}\, w_{(j,j+h)\,k} \end{aligned} \tag{5.54}$$

Donde $w_{(j,j+h)\,k}$ son los ponderadores obtenidos mediante algún método de calibración para las unidades respondentes en las olas j y $j+h$, y $a_{(j,j+h)\,k} = y_{(j+h)\,k} - y_{j\,k}$ es el cambio en la variable de interés y entre estas olas para cada elemento respondente $k \in \{r_j \cap r_{j+h}\}$. La aplicación de la fórmula (5.54) en los patrones de respuesta episódicos y por desgaste difiere en la definición del conjunto de respondentes simultáneos $r_j \cap r_{j+h}$.

Capítulo 6

Aplicación: Las damas perdidas

6.1. Introducción

La aplicación del uso de ponderadores calibrados para encuestas de panel se realiza en el marco de la "Encuesta sobre Situaciones Familiares y Desempeños Sociales en Montevideo y Área Metropolitana", llevada a cabo por un equipo de investigadores de la Universidad de la República (Facultad de Ciencias Económicas y de Administración - Instituto de Economía - y de la Facultad de Ciencias Sociales - Departamento de Economía y Programa de Población).

La ventaja de un estudio con datos de panel, radica en que la información longitudinal resulta ser la más adecuada para captar la diversidad de procesos que caracterizan a la vida familiar. A partir de ella es posible estudiar las historias conyugales, permitiendo dar cuenta de los cambios que se procesan en la sociedad en materia de formación de familias y de los contextos en que los individuos procesan la reproducción cotidiana. La información recogida en esta encuesta de panel representa un importante avance respecto a trabajar con datos estrictamente transversales y permite al menos comenzar a plantearse preguntas y relaciones más complejas que aquellas que posibilita la información disponible en la actualidad.

Esta es una encuesta de panel que hasta el momento consiste en dos olas. La primera ola de entrevistas fue realizada entre marzo y octubre de 2001 a una muestra de 1806

mujeres, de 25 a 54 años de edad, residentes en el Área Metropolitana[1]. La encuesta es representativa de los hogares en dicha área que tienen al menos una mujer en ese tramo de edades. La segunda ola se realizó en el año 2008 logrando recontactar a 828 mujeres. El tratamiento de la *no respuesta* representa el objetivo de esta aplicación, e implica el cálculo de ponderadores que permitan compensar la pérdida de unidades entre 2001 y 2008.

La construcción de ponderadores adecuados permitirán concretar los objetivos centrales establecidos para esta encuesta, que pueden resumirse de la siguiente manera:

1. Reconstruir las trayectorias familiares de las mujeres entre 25 y 54 años, poniéndolas en relación con las particularidades de su entorno familiar y sus características socioeconómicas.
2. Reconstruir sus trayectorias laborales.
3. Evaluar los cambios de los comportamientos familiares entre las distintas cohortes encuestadas.
4. Caracterizar el desempeño educativo y laboral de las mujeres y sus hijos, en conexión con el tipo de hogar al que pertenecen y al tipo de trayectoria familiar.
5. Describir y evaluar las relaciones y las transferencias económicas entre hogares, particularmente aquellas resultantes de las rupturas conyugales.
6. Analizar las decisiones conyugales y los comportamientos familiares en función de un conjunto de dimensiones ideológicas escogidas.

6.2. Diseño Muestral

6.2.1. Diseño Muestral[2] (2001)

De acuerdo a los objetivos generales de la investigación, el universo de interés está constituido por las mujeres que pertenecen al tramo de edad entre 25 y 54 años y

[1] El área metropolitana o Gran Montevideo comprende, además de la totalidad del departamento de Montevideo, a las localidades urbanas de Canelones y San José en un radio de 30 Km a partir del Km 0, según la definición usada por el INE en la ECH.

[2] Extraído del Documento de Trabajo de la "Encuesta sobre Situaciones Familiares y Desempeños Sociales en Montevideo y Área Metropolitana", 2001.

residen en hogares pertenecientes a zonas urbanas del área metropolitana. El tipo de diseño elegido para la selección de las unidades a ser encuestadas es estratificado en cuatro etapas. Las unidades primarias de selección son los segmentos censales, las unidades secundarias, las zonas censales urbanas y las unidades de tercera etapa están constituidas por hogares con por lo menos una mujer de entre 25 y 54 años de edad. En la última etapa de muestreo se selecciona una de dichas mujeres.

El criterio seguido para la estratificación es prácticamente el mismo que en la actualidad utiliza la ECH. Los segmentos censales de localidades rurales y urbanas para Montevideo se clasifican según el ingreso medio per cápita de los hogares. Se distinguen cuatro estratos: de ingreso bajo, ingreso medio bajo, ingreso medio alto e ingreso alto. La periferia de Montevideo[3] fue dividida en tres estratos por razones de administración de la encuesta. Éstos corresponden a la periferia de San José, periferia de Canelones y la Ciudad de la Costa. Los recursos disponibles para la instrumentación de la encuesta permiten un relevamiento de alrededor de 1800 casos. Trabajando con un tamaño total de muestra en dicho entorno la asignación de casos por estrato se realizó de manera proporcional al número de hogares con mujeres entre 25 y 54 años según el censo de 1996. La primera etapa de muestreo se desarrolla por estrato. Se seleccionan segmentos censales con probabilidades de inclusión proporcionales al número de hogares objetivo en el segmento. El número de segmentos a ser seleccionados por estrato resulta de asignar el tamaño de muestra de manera proporcional al número de hogares objetivo por estrato. En la segunda etapa de muestreo se seleccionan dos zonas, para cada uno de los segmentos seleccionados en la primera etapa; la selección se realiza con probabilidades de inclusión proporcional al número de viviendas particulares que releva el Censo de 1996 para cada zona[4]. La tercera etapa de muestreo consiste en localizar cuatro hogares de la población objetivo por zona seleccionada. Esta etapa de selección y la siguiente es realizada por las propias encuestadoras. Para cada zona seleccionada se parte de un punto elegido previamente al azar en la zona y la encuestadora la recorre hasta lograr contactar cuatro hogares con mujeres entre 25 y 54 años. Por último, se elige dentro del hogar a la persona que será efectivamente encuestada[5]. El número de unidades de primera, segunda y

[3]La ECH considera a la periferia de Montevideo como un único estrato en virtud de algunas dificultades para aplicar el mismo criterio usado para la capital.

[4]La elección del numero de viviendas por zona como aproximación al número de hogares objetivo se debe a que es la única información del censo de 1996 que se dispone desagregada a nivel de zonas.

[5]Para asegurar una selección aleatoria de los casos en hogares con más de una mujer comprendida

tercera etapa en la población (según el censo de 1996) y en la muestra se resumen en el cuadro 6.1.

Cuadro 6.1: Totales poblacionales y muestrales según Segmento, Zona y Hogar por Estrato

	Población			Muestra			
Estrato	Seg.	Zonas	Hog.	Seg.	Zonas	Hog.	Expansor
1: Montevideo Bajo	143	1540	36876	30	60	240	153,65
2: Montevideo Medio Bajo	283	2587	65580	52	104	416	157,64
3: Montevideo Medio Alto	335	2536	80000	63	123	492	162,60
4: Montevideo Alto	212	1507	53624	43	86	344	155,88
5: Periferia Canelones	174	2807	34129	27	54	216	158,00
6: Periferia San José	28	519	3212	3	6	24	133,83
7: Periferia C. de la Costa	72	1378	12664	10	20	80	158,30
Total	1247	12874	286085	228	453	1812	157,88

El cálculo de los expansores para los datos muestrales presenta la dificultad de que no es posible determinar con exactitud las probabilidades de inclusión de última etapa ya que el número de hogares objetivo por zona seleccionada es desconocido. Las probabilidades de inclusión de los elementos por estrato pueden aproximarse por el cociente entre el número hogares objetivo y el número de hogares efectivamente encuestados en dicho estrato[6]. El uso de los expansores que aparecen en el cuadro se justifica porque el diseño utilizado conduce a similares probabilidades de inclusión

en el tramo de edad de 25 a 54 años se optó por encuestar a la mujer cuya primer letra de su nombre de pila estuviera más próxima al comienzo del abecedario. La selección de un procedimiento como este permite evitar que se seleccione sistemáticamente a la persona que está disponible en el hogar, lo cual podría introducir sesgos no deseados.

[6]Este cociente no es constante entre estratos debido a los ajustes necesarios para lograr un número entero de segmentos a seleccionar por estrato.

para cada elemento de un estrato. Lo anterior se explica de la siguiente forma: en cada zona se seleccionan cuatro hogares de manera aleatoria y sin considerar el número de hogares potencialmente encuestables; por su parte, dicha zona es seleccionada con una probabilidad aproximadamente proporcional a la cantidad de hogares objetivo de la zona, así, una alta probabilidad de inclusión para la zona determina bajas probabilidades de selección para los hogares objetivo incluidos en ella. Lo contrario ocurre para zonas con un número reducido de hogares objetivo. Este comportamiento inverso entre probabilidades de inclusión de las zonas y probabilidades de inclusión para los hogares objetivo dentro de cada zona determina probabilidades de inclusión aproximadamente iguales para cada elemento de la población e iguales a la constante de proporcionalidad. Luego, es sencillo de ver que si los expansores son aproximadamente constantes por estrato y el tamaño de muestra en el estrato es fijo, se tiene que dicha constante debe ser la inversa de la fracción de muestreo por estrato, es decir, el número de hogares objetivo del estrato sobre el número de casos en la muestra. Por último, vale la pena aclarar que la expansión de los valores muestrales reproducirán los totales poblacionales que se verificarían si la estructura de hogares objetivo por estrato no hubiera experimentado cambios desde el censo de 1996. Para obtener una estimación más ajustada se debería contar con algún dato externo que permitiera ajustar tal evolución.

6.2.2. La segunda ola del panel: entrevistas 2008

Dado que el período transcurrido entre las dos olas de este panel es extenso, la estrategia de entrevistas en la segunda ola considera la potencial dificultad de recontacto de las unidades participantes en la primera ola. Asimismo, el transcurso del tiempo se refleja en un envejecimiento del panel, que se debe a que las unidades entrevistadas ya no representarán a las mujeres de Gran Montevideo de edades entre 25 y 54 años. Por estos motivos la estrategia de entrevistas llevada a cabo en la segunda ola de este panel consiste en invertir esfuerzos en intentar recontactar la mayor cantidad de mujeres encuestadas en 2001, y suplementar la muestra con nuevas unidades que servirán de instrumento para "rejuvenecer" el panel (mediante la inclusión de un grupo de mujeres dentro de la franja etárea de 25 a 31 años), y ampliar el panel (mediante la inclusión de otro grupo de mujeres que contemplen todas las edades: de 25 a 61 años).

El diseño muestral para 2008 incluye tres muestras "enlazadas":

1. Muestra original utilizada en 2001: 1.806 mujeres, 997 de las cuales fueron recontactadas en 2008.
2. Una muestra de 1.000 hogares situados a "la derecha" del hogar mencionado en el punto 1. En los hogares de esta muestra se encuesta una mujer entre 25 y 61 años en cada hogar (si es que hay mujeres de ese rango etario en ese hogar).
3. Una muestra de 1.000 hogares situados a "la izquierda" del hogar mencionado en el punto 1. En los hogares de esta muestra se encuesta una mujer entre 25 y 31 años en cada hogar (si es que hay mujeres de ese rango etario en ese hogar).

Atendiendo al porcentaje de hogares con mujeres en los rangos etarios correspondientes a los hogares mencionados en los puntos 2 y 3, se esperaba entrevistar a unas 630 mujeres para el primer caso (Derechas), y 160 para el segundo (Izquierdas).

Los tamaños de muestra efectivos en la segunda ola para cada una de las tres muestras "enlazadas" fue:

Cuadro 6.2: Cantidad de entrevistas efectivas en 2008

Centro: Recontacto muestra original, edades 32 a 61	828
Derechas: Suplemento edades 25 a 61	308
Izquierdas: Suplemento edades 25 a 31	93
Total de Entrevistas (todas las edades)	1229

El diseño original de la investigación realizada en el 2001 es estratificado multietápico, y es un diseño aproximadamente autoponderado. La selección de hogares del 2001 que participan en la investigación en 2008 no se realiza de manera aleatoria, ya que se consideran en esta última instancia únicamente los hogares que pudieron ser contactados (1000 hogares). Las diferencias en las características de los hogares se verán reflejadas en la calibración de los correspondientes factores de expansión muestrales. La calibración de los factores de expansión permite que la muestra 2008 estime sin error algunas características conocidas para la población objetivo.

6.3. Obtención de ponderadores

Los ponderadores a ser utilizados en la estimación longitudinal de cambios entre olas y en la estimación transversal de totales no podrán ser los mismos, ya que estas estimaciones estarán basadas en distintas unidades. Los cambios longitudinales solamente podrán ser estimados utilizando las mediciones efectivas sobre las 828 mujeres que fueron encuestadas en las dos olas del estudio, mientras que para las estimaciones transversales se cuenta con información sobre el total de mujeres entrevistadas (1229 mujeres). Por este motivo se obtendrán dos conjuntos de ponderadores, utilizando el raking como técnica de calibración.

6.3.1. Ponderadores para Estimaciones Longitudinales

Entre la instancia inicial en 2001 y la segunda ola en 2008, la imposibilidad de contacto de algunas mujeres (fallecimiento, movilidad, etc.) y la negativa de la mujer seleccionada a seguir participando genera la pérdida de 978 mujeres, provocando el desgaste del panel. Las 1806 mujeres entrevistadas en la primera instancia representan bien a la población de mujeres del 2001 con las características ya mencionadas. Si fuera posible medir los cambios entre olas en ciertas variables en cada una de estas mujeres, los mismos también serían representativos de la población objetivo 2008. La no respuesta imposibilita dicha medición para todas las unidades de la muestra, y debe compensarse. Este objetivo se puede lograr siguiendo distintas estrategias: calculando ponderadores calibrados para las 828 mujeres respondentes para que representen a las 1806 iniciales (que a su vez, representan la población objetivo 2001), o calibrar los ponderadores de las 828 respondentes para que representen a la población objetivo 2008 de forma directa (las mujeres de 32 a 61 años de Gran Montevideo). Se opta por seguir la segunda estrategia.

Como ya se dijo, el procedimiento de calibración utilizado para el cálculo de estos ponderadores es el raking. Del pool de variables disponibles se deben seleccionar aquellas que serán utilizadas como información auxiliar para el cálculo de los nuevos ponderadores. Algunas de las características deseables para las variables auxiliares son su estabilidad en el tiempo y que permitan una buena caracterización de la población objeto de estudio. Para el caso particular de este estudio de panel, se descarta la

utilización de variables provenientes del cuestionario de 2001 como variables auxiliares, ya que las mismas serán objeto de la medición de cambios longitudinales. Las variables finalmente elegidas para la calibración son la Edad y Nivel Educativo, al entenderse que verifican las condiciones mencionadas. Los totales poblacionales correpondientes se estiman a partir de la Encuesta Nacional de Hogares Ampliada del año 2006 (ENHA 2006).

Raking

Para el cálculo de los ponderadores se utilizó el programa R Development Core Team (2008) con el paquete *survey* (T. Lumley (2009)) y la función *rake*. Los insumos necesarios para calcular estos nuevos ponderadores son el diseño que genera los expansores originales y los totales marginales sobre los cuales se calibrará. En los cuadros siguientes se presenta dicha información.

Cuadro 6.3: Totales Poblacionales, Muestrales y Expansores originales por Estrato

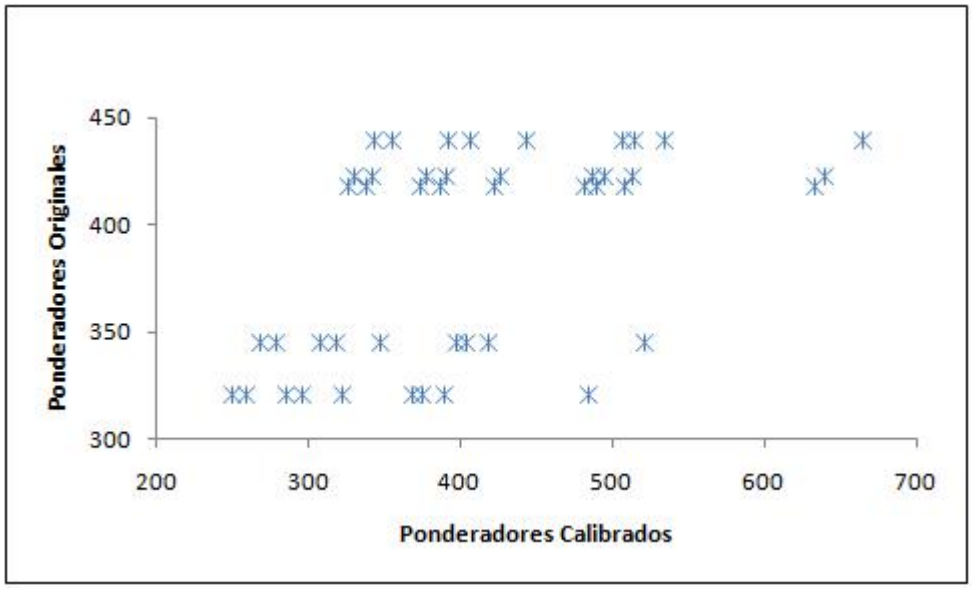
Estrato	Tot. Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	46924	111	533,738
MVD Medio Bajo	68596	199	344,703
MVD Medio Alto	76267	238	320,449
MVD Alto	57736	138	418,377
Periferia	62418	142	439,563
Total	311941	828	

Cuadro 6.4: Totales Poblacionales de las Variables Auxiliares

Nivel Educativo	Frecuencia
Primaria	83112
Secundaria	147627
Terciaria	81202
Total	311941

Edad	Frecuencia
32 a 40 años	94225
41 a 50 años	112040
51 a 61 años	105676
Total	311941

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.



En el gráfico puede notarse que los ponderadores dentro de cada estrato (representados por los conjuntos de puntos alineados de forma paralela al eje de las abscisas)

dejan de ser iguales. Dentro de cada estrato, hay nueve pesos diferentes, determinados por la interacción entre tramo de edad y nivel educativo, utilizados en el raking. En el siguiente cuadro se presentan los totales muestrales que dan origen a 45 nuevos ponderadores.

Cuadro 6.5: Totales muestrales según Edad por Estrato y Nivel Educativo

Estrato y Nivel Educativo/Edad	32 a 40	41 a 50	51 a 61
Mvd Bajo - Primaria	9	17	20
Mvd Bajo - Secundaria	22	18	14
Mvd Bajo - Terciaria	4	2	5
Mvd Medio Bajo - Primaria	8	13	22
Mvd Medio Bajo - Secundaria	33	46	38
Mvd Medio Bajo - Terciaria	13	11	15
Mvd Medio Alto - Primaria	2	5	16
Mvd Medio Alto - Secundaria	27	49	45
Mvd Medio Alto - Terciaria	26	45	23
Mvd Alto - Primaria	1	1	4
Mvd Alto - Secundaria	9	17	19
Mvd Alto - Terciaria	17	36	34
Periferia - Primaria	10	20	21
Periferia - Secundaria	22	26	20
Periferia - Terciaria	11	10	2

Del cuadro surge la justificación de la utilización del raking frente a la post estratificación, ya que existen celdas de la clasificación estrato, nivel educativo y edad

con muy pocas observaciones. Frente a la alternativa de colapsar categorías para luego aplicar la post estratificación para la obtención de ponderadores, se opta por la aplicación directa del raking.

Estimador de cambios longitudinales

Las diferencias para la variable y entre 2001 y 2008 se estiman por $\hat{A}_{W_{2001,2008}}$:

$$\begin{aligned} \hat{A}_{W_{2001,2008}} &= \sum_{k=1}^{828} (y_{2008\,k} - y_{2001\,k})\, w_{Ck}; \\ &= \sum_{k=1}^{828} a_{(2001,2008)\,k}\, w_{Ck} \end{aligned} \tag{6.1}$$

donde $a_{(2001,2008)\,k} = (y_{2008\,k} - y_{2001\,k})$ representa la diferencia de los valores de la variable de interés y medidos en 2008 y 2001 para las 828 mujeres de la muestra original entrevistadas en 2001 y en 2008, y w_{Ck} el ponderador que surge de la aplicación del método de calibración raking para este grupo de mujeres.

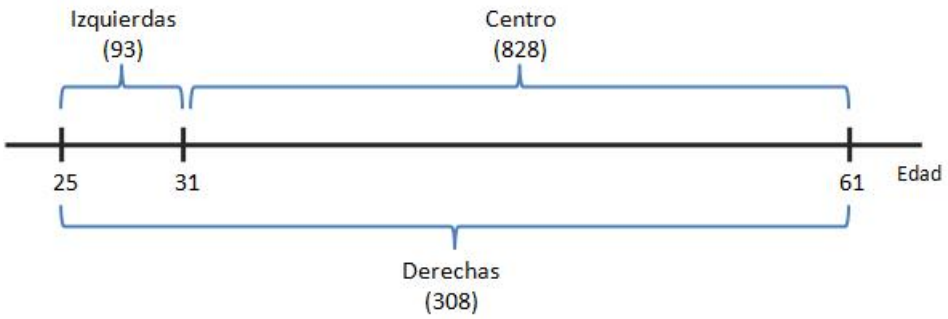
6.3.2. Ponderadores para Estimaciones Transversales

Las estimaciones transversales para el 2008 se calcularán en base a las respuestas dadas por las 1229 mujeres respondentes: 828 provenientes del panel original, 93 mujeres de edades entre 25 y 31 años (las "rejuvenecedoras" del panel) y 308 de 25 a 61 años de edad (las "ampliadoras" del panel). El objetivo de lograr que las 1229 mujeres representen bien a las mujeres de Gran Montevideo con edades entre 25 y 61 años en 2008 requiere el cálculo por separado de ponderadores en cada uno de estos tres grupos, ya que las mujeres que integran cada grupo provienen de muestras distintas. La libertad de elección de ponderadores iniciales $d_{\alpha k}$ mencionada en la sección (5.5.1) no contempla el caso en que las unidades provengan de muestras distintas. Es por esto que en cada uno de estos grupos se calcularán los nuevos ponderadores con el método raking para luego combinar los resultados. Las variables auxiliares a ser utilizadas en el procedimiento de calibración seleccionado para cada

uno de estos grupos serán edad y nivel educativo, de igual manera que en la sección anterior.

En la siguiente figura se representa el rango de edad de las 1229 mujeres relevadas en 2008 según la muestra de la que provienen: centro, derechas e izquierdas.

Figura 6.1: Representación gráfica de las edades de las 1229 mujeres relevadas en 2008 según muestra de origen



Raking Centro

Los nuevos ponderadores para estas 828 mujeres centro son los mismos que fueron calculados en la parte anterior.

Raking Izquierdas

En este caso la única variable auxiliar para calibrar es nivel educativo dado que todas estas mujeres tienen entre 25 a 31 años, que corresponde a una única categoría de la variable edad. Los ponderadores originales y los totales poblacionales se presentan en el cuadro 6.7 y el cuadro 6.8.

Cuadro 6.7: Totales Poblacionales, Muestrales y Expansores Originales por Estrato

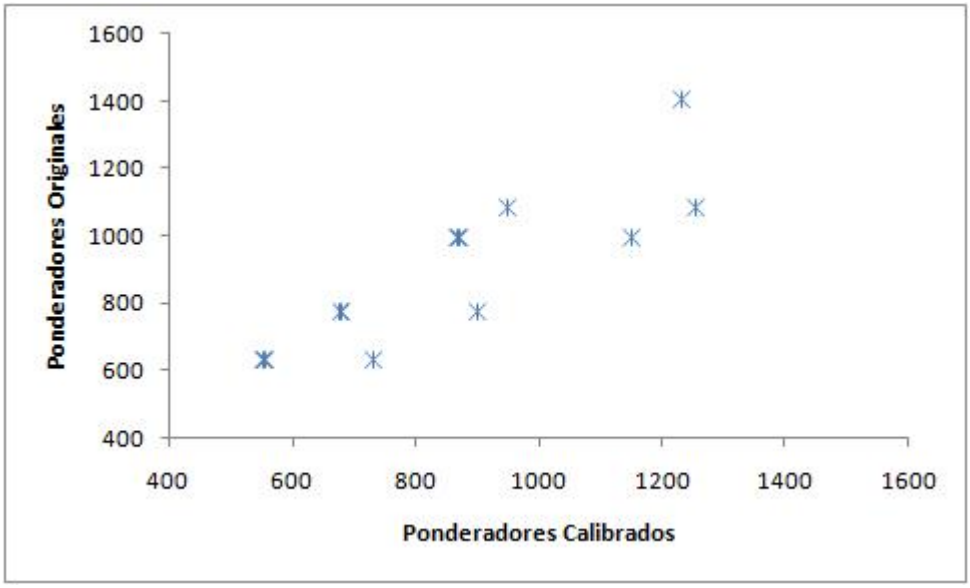
Estrato	Tot.Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	13931	14	995,071
MVD Medio Bajo	18622	24	775,917
MVD Medio Alto	21677	20	1083,850
MVD Alto	14050	10	1405,000
Periferia	15804	25	632,160
Total	84084	93	

Cuadro 6.8: Totales Poblacionales de la Variable Auxiliar

Nivel Educativo	Frecuencia
Primaria	12804
Secundaria	42852
Terciaria	28428
Total	84084

En este caso el método raking genera los mismos resultados que el método de post estratificación, ya que calibrar en las marginales de una única variable auxiliar es equivalente a calibrar en las celdas.

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.



En este caso hay tres ponderadores por estrato, las tres categorías de nivel educativo, excepto para los estratos Montevideo Medio Alto y Alto; esto último se debe a que en esta muestra no hay mujeres con solamente primaria completa en estos estratos, como se muestra en el siguiente cuadro.

Cuadro 6.9: Totales Muestrales por Nivel Educativo según Estrato

Estrato/Nivel Educativo	Primaria	Secundaria	Terciaria	Total
MVD Bajo	4	9	1	14
MVD Medio Bajo	4	14	6	24
MVD Medio Alto	0	8	12	20
MVD Alto	0	2	8	10
Periferia	12	9	4	25
Total	20	42	31	93

Raking Derechas

Los insumos necesarios para calcular estos nuevos ponderadores son el diseño que genera los expansores originales y los totales marginales sobre los cuales se calibrará, de igual manera a las partes anteriores, pero con la excepción de que la variable auxiliar edad tiene ahora cuatro categorías.

Cuadro 6.10: Totales Poblacionales, Muestrales y Expansores Originales por Estrato

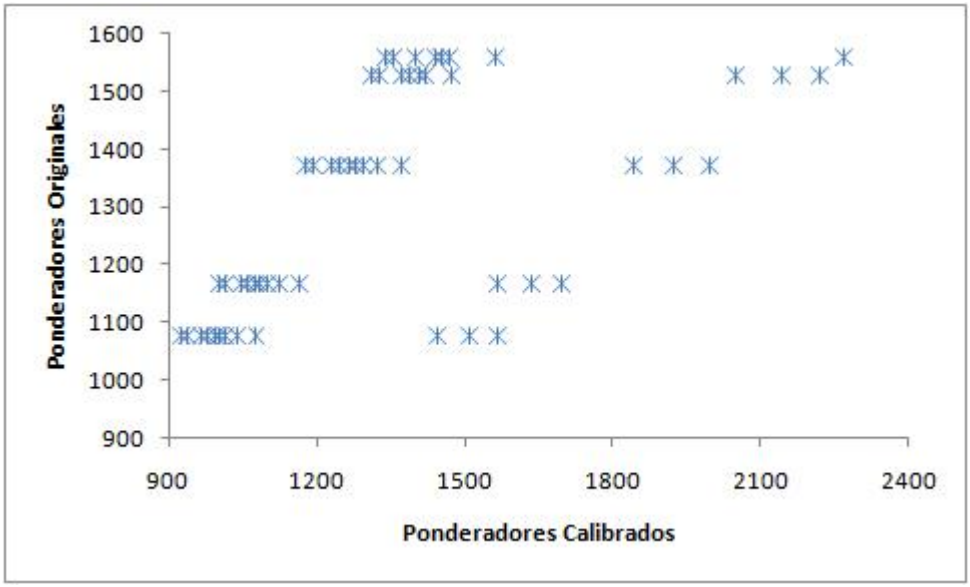
Estrato	Tot. Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	60855	39	1560,385
MVD Medio Bajo	87218	81	1076,765
MVD Medio Alto	97944	84	1166,000
MVD Alto	71786	47	1527,362
Periferia	78222	57	1372,316
Total	396025	308	

Cuadro 6.11: Totales Poblacionales de las Variables Auxiliares

Nivel Educativo	Frecuencia
Primaria	95916
Secundaria	190479
Terciaria	109630
Total	396025

Edad	Frecuencia
25 a 31 años	84084
32 a 40 años	94225
41 a 50 años	112040
51 a 61 años	105676
Total	396025

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.



La interacción entre las variables nivel educativo (3 categorías) y edad (4 categorías) para cada uno de los cinco estratos debería generar 60 pesos diferentes. En los hechos, se pueden distinguir 54 ponderadores diferentes debido a que no todas las celdas correspondientes a la interacción de las variables auxiliares contienen observaciones, como se puede ver en el cuadro 6.12.

Cuadro 6.12: Totales Muestrales según Edad por Estrato y Nivel Educativo

Estrato y Nivel Educativo/Edad	25 a 31	32 a 40	41 a 50	51 a 61
Mvd Bajo - Primaria	4	6	4	3
Mvd Bajo - Secundaria	4	7	10	0
Mvd Bajo - Terciaria	0	1	0	0
Mvd Medio Bajo - Primaria	5	5	4	6
Mvd Medio Bajo - Secundaria	15	14	9	7
Mvd Medio Bajo - Terciaria	4	5	4	3
Mvd Medio Alto - Primaria	1	3	1	4
Mvd Medio Alto - Secundaria	8	7	21	7
Mvd Medio Alto - Terciaria	7	11	8	6
Mvd Alto - Primaria	0	2	0	2
Mvd Alto - Secundaria	4	5	6	2
Mvd Alto - Terciaria	6	1	8	11
Periferia - Primaria	2	4	10	2
Periferia - Secundaria	9	10	9	5
Periferia - Terciaria	1	2	2	1

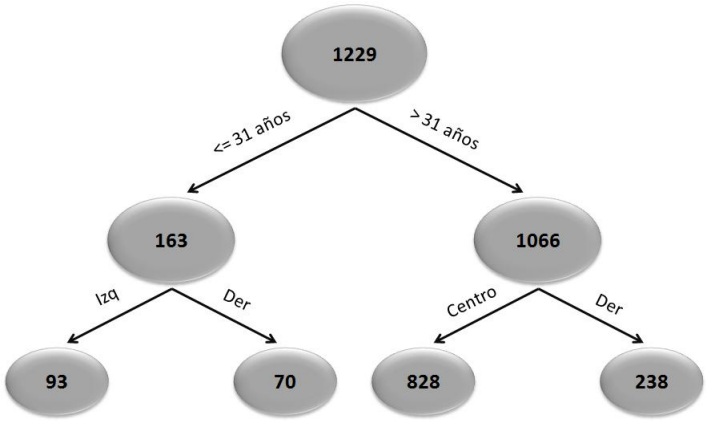
Ponderadores Combinados

Para cada una de las 1229 mujeres se calculó su respectivo ponderador, en relación a la submuestra a la que pertenece. Las 828 mujeres del centro representan al total de mujeres de Gran Montevideo con edades entre 32 y 61 años (311941 mujeres); las 93 mujeres de la izquierda representan a las 84084 mujeres dentro de la franja etaria de

25 a 31 años; y las 308 de la derecha a aquellas de edades entre 25 y 61 años, cuyo total asciende a 396025 mujeres. Si para realizar cálculos de totales se utilizaran los ponderadores obtenidos en cada uno de los grupos de manera directa, dichos totales se estarían sobreestimando. De hecho, se estaría estimando el total correspondiente a una población compuesta por el doble de las mujeres existentes en Gran Montevideo de edades entre 25 y 61 años. Es por este motivo que los ponderadores calculados en las partes anteriores deben utilizarse de forma combinada, para lograr la estimación sobre el total efectivo de mujeres de dichas características: 396025 mujeres.

Las 1229 mujeres entrevistadas en 2008 se clasifican en cuatro subpoblaciones en relación a su edad (mayor o menor a 31 años) y muestra de procedencia (centro, izquierdas, derechas), como muestra la Figura 6.2.

Figura 6.2: Clasificación de las mujeres entrevistadas en 2008



El primer nivel de división (corte por edad) permite identificar dos grandes grupos: 163 mujeres de edades entre 25 y 31 años que deberán representar a 84084 mujeres, y 1066 mayores a 31 años, que deberán representar a 311941 mujeres existentes en dicha franja etaria. El segundo corte queda determinado por la muestra de la que provienen cada una de las mujeres: las 163 mujeres menores a 32 años pueden pertenecer a la muestra de izquierdas o derechas, y las 1066 mayores de 31 pueden ser parte del panel original o de la muestra de derechas.

La combinación de resultados sigue la misma lógica que la figura: en primer lugar, se particiona la muestra total de acuerdo a la edad (mayores o menores de 31 años)

formando dos grandes grupos. Dentro de cada grupo los nuevos ponderadores combinados tendrán en cuenta la muestra de la que provienen a través de la relación que existe entre la cantidad de mujeres en esta última partición y la cantidad de mujeres en el grupo etario que corresponda. Por ejemplo, los pesos obtenidos a través del raking para las 828 mujeres provenientes del panel original se multiplicarán por 828/1066, la proporción de mujeres del panel que corresponde a la franja etaria establecida. Considerando que las varianzas de los estimadores son del orden de $1/n_i$ siendo n_i el tamaño de cada una de las cuatro particiones, se combinan los ponderadores resultantes del raking de manera que las mujeres provenientes de subgrupos más grandes adquieran mayor importancia en el análisis.

Los ponderadores combinados a ser utilizados en las estimaciones transversales resultan ser:

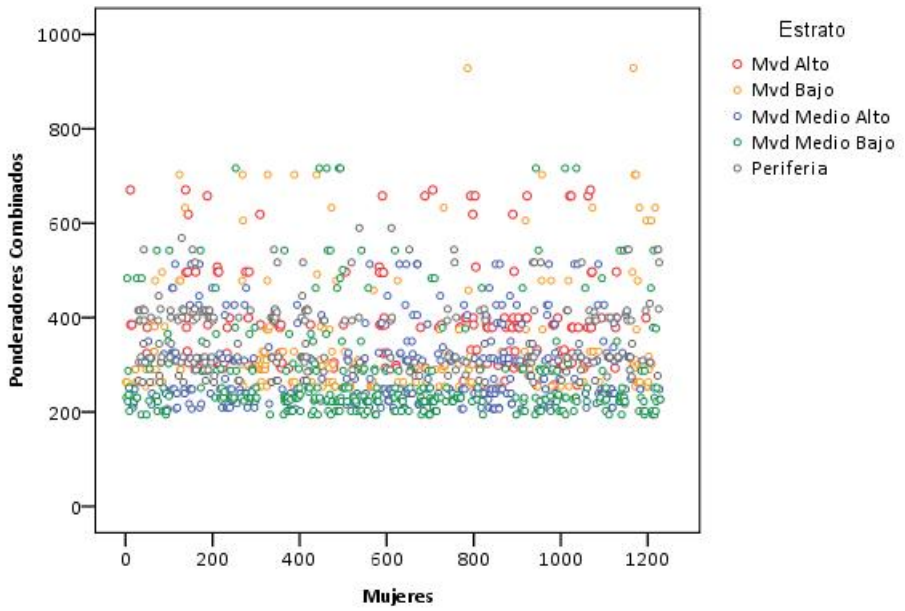
$$w_{comb_k} = \begin{cases} \frac{93}{163} w_{Ik} & k \in \text{muestra de izquierdas} \\ \frac{70}{163} w_{Dk} & k \in \text{muestra de derechas con edades entre 25 y 31 años} \\ \frac{238}{1066} w_{Dk} & k \in \text{muestra de derechas con edades entre 32 y 61 años} \\ \frac{828}{1066} w_{Ck} & k \in \text{panel original} \end{cases}$$

Estos ponderadores combinados estiman sin error la cantidad total de mujeres pertenecientes a la población objetivo estimada por la ENHA 2006, que se considera equivalente a la población 2008.

$$\begin{aligned} \hat{N} &= \sum_{k=1}^{1229} w_{comb_k} = \\ &= \sum_{k=1}^{93} \frac{93}{163} w_{Ik} + \sum_{k=1}^{70} \frac{70}{163} w_{Dk} + \sum_{k=1}^{238} \frac{238}{1066} w_{Dk} + \sum_{k=1}^{828} \frac{828}{1066} w_{Ck} \\ &= 47974,3 + 36109,7 + 69645,4 + 242295,6 \\ &= 396025 = N \end{aligned}$$

La siguiente figura muestra los ponderadores combinados para las 1229 mujeres según el estrato al que pertenecen.

Figura 6.3: Ponderadores Combinados por Estrato



Estimación de totales transversales

La estimación transversal de los totales de interés debe realizarse utilizando los ponderadores combinados de acuerdo a la siguiente fórmula:

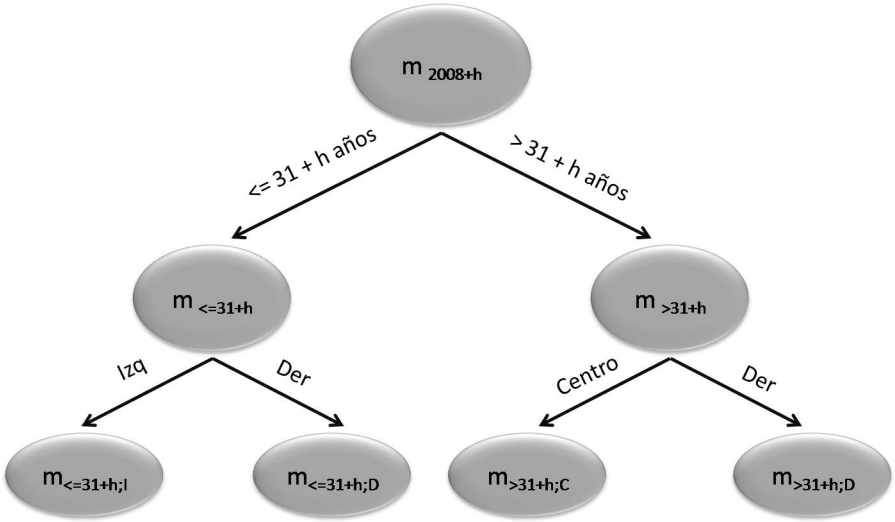
$$\begin{aligned}\hat{t}_{yWcomb} &= \sum_{k=1}^{1229} w_{comb_k} y_k \\ &= \sum_{k=1}^{93} \frac{93}{163} w_{Ik} y_k + \sum_{k=1}^{70} \frac{70}{163} w_{Dk} y_k + \sum_{k=1}^{238} \frac{238}{1066} w_{Dk} y_k + \sum_{k=1}^{828} \frac{828}{1066} w_{Ck} y_k \end{aligned} \quad (6.2)$$

6.4. Sugerencias para una potencial tercera ola

En una potencial tercera ola a realizarse h años después se debería intentar recontactar a las 1229 mujeres efectivamente entrevistadas en 2008. Las estimaciones

longitudinales deberán realizarse en base a todas las mujeres entrevistadas en 2008 y 2008+h. Si en esta tercera instancia no se suplementa la muestra con nuevas unidades, los ponderadores a ser utilizados en estimaciones longitudinales y transversales serán los mismos. El razonamiento seguido para el cálculo de ponderadores transversales desarrollado en la sección anterior se extiende para obtener los ponderadores combinados correspondientes a la ola 2008+h.

Las mujeres encuestadas en 2008 respondentes en la tercera ola ($m_{2008+h} \leq 1229$) pueden particionarse según edad y muestra de procedencia, segun muestra la figura:



donde $m_{\leq 31+h}$ y $m_{>31+h}$ representan la cantidad de mujeres respondentes, en la tercera ola, menores y mayores a $31+h$ años de edad respectivamente; y $m_{\leq 31+h;I}$, $m_{\leq 31+h;D}$, $m_{>31+h;D}$, $m_{>31+h;C}$ son la cantidad de mujeres según partición de edad (menores o mayores de $31+h$ años) y muestra de origen en 2008 (centro, derechas e izquierdas).

De esta manera, los ponderadores combinados a ser utilizados tanto en estimaciones transversales como longitudinales[7] son:

[7] para comparar 2008 y 2008+h. Los cambios entre 2001 y 2008+h solo serán medibles en las mujeres del centro que respondan en esta última instancia.

$$w_{comb_k} = \begin{cases} \frac{m_{\leq 31+h;I}}{m_{\leq 31+h}} w_{Ik} & k \in \text{muestra izq 2008} \\ \frac{m_{\leq 31+h;D}}{m_{\leq 31+h}} w_{Dk} & k \in \text{muestra der 2008, de edades entre } 25+h \text{y } 31+h \\ \frac{m_{>31+h;D}}{m_{>31+h}} w_{Dk} & k \in \text{muestra der 2008, de edades entre } 32+h \text{y } 61+h \\ \frac{m_{>31+h;C}}{m_{>31+h}} w_{Ck} & k \in \text{panel original 2001} \end{cases}$$

Los ponderadores w_{Ck}, w_{Dk} y w_{Ik} son aquellos que surgen de la aplicación de algún método de calibración en base a totales conocidos de variables auxiliares para centros, derechas e izquierdas en el año $2008 + h$.

Capítulo 7

Conclusiones

La medición de cambios entre distintas instancias en el tiempo es el principal objetivo de las encuestas de panel. Se presentaron estimadores del cambio de las variables de interés que, bajo el supuesto de respuesta perfecta, son insesgados para estimar el cambio total en la población objetivo. Para su desarrollo, se partió de supuestos muy restrictivos rara vez presentes en la práctica: la existencia de respuesta perfecta y población fija en el tiempo, reflejados en ponderadores constantes en el tiempo para cada elemento e iguales al inverso de su probabilidad de inclusión en la muestra.

La no respuesta es un fenómeno presente en la mayoría de las encuestas por muestreo y es necesario su tratamiento para evitar sesgos en las estimaciones. En las encuestas de panel la inclusión del factor tiempo provoca un agravamiento del problema de no respuesta, reflejado en reducciones considerables en el tamaño de muestra período a período. Si la no respuesta se presentara de manera completamente aleatoria, el único incoveniente al que se enfrenta el investigador resulta en la reducción del tamaño de muestra y su respectivo aumento en la varianza de las estimaciones, pero las unidades que no contestan suelen diferir de aquellas que sí lo hacen, y el sesgo introducido en las estimaciones por esta causa constituye el obstáculo más importante por corregir.

Aún cuando se invierta esfuerzo en intentar alcanzar la mayor tasa de respuesta posible, la no respuesta existe y es deber de los investigadores realizar algún tratamiento en la etapa de análisis de datos para controlar el sesgo introducido a las estimaciones por su causa. Como método de tratamiento se descartan la sustitución y submuestreo de no respondentes y los métodos basados en la quasi-ramdomnization por entender

que no son las opciones preferibles en encuestas de panel. La imputación como forma de tratar la no respuesta de unidades en la ola también se descarta por generar una fabricación masiva de datos, que pueden distorisionar las asociaciones entre las variables que representan el principal objetivo del panel. Es preferible entonces utilizar una estrategia global como es la calibración cuando es la unidad la que no provee de respuesta. De todas maneras la imputación es el tratamiento elegido como manera de compensar la no respuesta en los ítems, previo a la calibración por la no respuesta de unidades.

La idea central de los estimadores calibrados es sencilla, en base a información auxiliar se modifican los ponderadores originales de la muestra minimizando alguna función de distancia entre dichos ponderadores y los ponderadores finales (o calibrados) y de manera que estos últimos estimen sin error totales poblacionales conocidos de las variables auxiliares que asisten al procedimiento de calibración. El estimador calibrado corresponde en realidad a una familia entera de estimadores que dependen de formulaciones diferentes del vector auxiliar y de la función de distancia.

La efectividad del estimador de calibración para controlar el sesgo ocacionado por la no respuesta dependerá de propiedades del vector auxiliar. Se obtiene una expresión del sesgo aproximado que demuestra que éste será menor cuanto más estrecha sea la relación entre la información auxiliar y la probabilidad de respuesta o la variable de interés. También se propone una formulación para la estimación de la varianza del estimador calibrado que podrá ser utilizada en la construcción de intervalos de confianza. Si el sesgo es modesto el intervalo de confianza será válido y la probabilidad de cobertura será cercana al nivel de confianza especificado.

El estimador de cambios entre olas propuesto para el caso de respuesta perfecta se extiende utilizando los ponderadores calibrados calculados para las unidades respondentes. De esta manera, los cambios entre las olas de interés medido en los respondentes simultáneos ponderado por los pesos calibrados representarán los cambios de la población objetivo.

Bibliografía

[1] Bailar, B.A. (1975). "The Effects of Rotation Group Bias on Estimates from Panel Surveys". Journal of the American Statistical Association, 70(349): 23–30.

[2] Chhikara, R.S.; and Deng, L.Y. (1992). "Estimation Using Multiyear Rotation Design Sampling in Agricultural Surveys". Journal of the American Statistical Association, 87(420): pp 924–932.

[3] Copeland, K.R. (2004). "Nonresponse Adjustment in the Current Employment Statistics Survey". U.S. Bureau of Labour Statistics, Washington DC.

[4] De Leeuw, E.D.; Hox, J.; and Huisman, M. (2003). "Prevention and Treatment of Item Nonresponse". Journal of Official Statistics, 19 (2): 153–176.

[5] De Leeuw, E.D. (2006). "Introduction to Survey Nonresponse". Summer Institute in Survey Research Techniques. Survey Methodology 988.223.

[6] Dennis, J.M. and Li, R. (2003). "Effects of Panel Attrition on Survey Estimates". For Presentation at the 2003 Annual Meeting of the American Association for Public Opinion Research in Nashville, Tennesse.

[7] Deville, J.C. and Särndal, C.E. (1992). "Calibration Estimators in Survey Sampling". Journal of the American Statistical Association, 87(418).

[8] Deville, J.C.; Särndal, C.E.; and Sautory, O. (1993). "Generalized Raking Procedures in Survey Sampling". Journal of the American Statistical Association, 88(423).

[9] Deville, J.C. and Särndal, C.E. (1994).“Variance estimation for the regression imputed Horvitz – Thompson estimator”. Journal of Official Statistics. 10: 381–394.

[10] Ernst, L.R (1986). “Weighting Issues for Longitudinal Household and Family Estimates”. Bureau of Census, SRD Research Report Number: CENSUS/SRD/RR-86/23.

[11] Fuller, W.A. and Breidt, F.J. (1999). “Estimation for Supplemented Panels”. The Indian Journal of Statistics Special Issue on Sample Surveys, 61: 58–70.

[12] Gross Sobol, M. (1959). “Panel Mortality and Panel Bias”. Journal of the American Statistical Association, 54(285): 52 - 68.

[13] Groves, R.M. and Couper M.P (1998). Nonresponse in Household Interview Surveys. New York: John Wiley & Sons.

[14] Kalton, G. and Brick, J.M . “Weighting Schemes for Household Panel Surveys: The Survey of Income and Program Participation”. U.S. Department of Commerce Bureau of Census.

[15] Kish, Leslie (1995). Diseño Estadístico para la Investigación. España: Siglo XXI de España Editores SA.

[16] Little, R.J.A and David, H.M. (1983). “Weighting Adjustment for Nonresponse in Panel Surveys”. Working Paper: U.S. Bureau of Census, Washington DC.

[17] Little, R.J.A and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley & Sons.

[18] Lumley, T. (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19

[19] Lumley, T. (2009) "survey: analysis of complex survey samples". R package version 3.11-2.

[20] Lynn, P.; Buck, N.; Burton, J.,; Jäckle, A.; and Laurie, H. (2005). “A Review of Methodological Research Pertinent to Longitudinal Survey Design and Data Collection”. Working Papers of the Institute for Social and Economic Research, paper 2005 – 29. Colchester: University of Essex.

[21] Rubin, Donald B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley & Sons.

[22] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[23] Särndal, C.E.; Swensson, B.; and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

[24] Särndal, C.E. and Lundström, S. (1999). "Calibration as a Standard Method for Treatment of Nonresponse". Journal of Official Statistics. 15(2): 305–327.

[25] Särndal, C.E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. New York: John Wiley and Sons, Ltd.

[26] Särndal,C.E. and Lundström, S. (2008). "Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator". Journal of Official Statistics, 24 (2):167–191.

[27] Sastry, N; Ghosh-Dastidar, B; Adams, J; and Pebley, A (2000). "The Design of a Multilevel Longitudinal Survey of Children, Families, and Communities: The Los Angeles Family and Neighborhood Survey". Labour and Population Program, Working Paper Series 00-18.

[28] Sharot, T. (1991). "Attrition and Rotation in Panel Surveys". The Statistician, 40(3): 325–331.

[29] Simard, M. (2002). "Generation and Gender Survey Sample Design Guidelines". Statistics Canada.

[30] Taylor, B.; Brook, L.; and Lynn, P. (1997). "Incentives Information and Number of Contacts: Testing the Effects of these Factors of Response to a Panel Survey". Survey Methods Centre Newsletter, 17(3): 712.

[31] Wadsworth, R.N. (1952). "The experience of a User of a Consumer Panel". Applied Statistics, 1(3): 169–178.

[32] Williams, W.H. and Mallows, C.L. (1970). "Sistematic Biases in Panel Surveys Due to Differential Nonresponse". Journal of the American Statistical Association, 65(331).

[33] Woodruff, R.S. (1963). "The Use of Rotating Samples on the Census Bureau´s Monthly Surveys". Journal of the American Statistical Association, 58(302): 454–467.