UNIVERSIDAD DE GUANAJUATO

Estimador por calibración

Problemas en su implementación y propuestas de soluciones

TESIS

Que para obtener el título de

Licenciado en Matemáticas

PRESENTA:

Víctor Miguel García Sánchez

Director de tesis:

Dr. José Elías Rodríguez Muñoz

Guanajuato, Gto. mayo, 2019

Universidad de Guanajuato
Departamento de Matemáticas

Estimador por calibración

Problemas en su implementación y propuestas de soluciones

Víctor Miguel García Sánchez

19 de mayo de 2019

Dedicado a mi familia, amigos y maestros:
todos aquellos que me formaron.

Índice general

Agradecimientos

A Dios, por haberme permitido llegar hasta este punto. Por brindarme salud, amor, paciencia y oportunidades, pues por su infinita bondad he alcanzado cuanto tengo, sabiendo ver las puertas que abría frente a mi.

A mis padres, pues más allá de mi educación en casa y todas sus contribuciones para mi formación escolar, sembraron en mi algo que atesoraré toda la vida: Que en la vida no hay tareas demasiado pequeñas ni demasiado grandes para mis capacidades, y que gracias a personas como ellos, seré capaz de aprender a llevarlas a cabo para no esperar que alguien más las haga.

A mi hermana, por ser para mi siempre una de las personas que más admiro, por inspirarme a ser el mejor ejemplo a seguir que pueda. Por su tenacidad y perseverancia en la escuela y en la vida, que la han vuelto para mi un rival que me presiona a seguir mejorando.

A mi novia, por ponerme los pies sobre la tierra y recordarme constantemente que aún en la carrera de matemáticas, hay asuntos no matemáticos que no debo descuidar. Por su apoyo en las primeras revisiones de redacción y sobre todo, por no soltarme la mano cuando la situación se ponía difícil, en especial cuando era yo quien la volvía difícil.

A mis amigos, por cada vez que me contestaron una llamada, cada invitación a comer, cada noche de hospedaje, cada abrazo, cada lágrima a su lado, pues esos momentos nos volvieron esa familia que como foráneos hace tanta falta. Por cada problema que resolvimos juntos, cada desmañanada y cada discusión, todo lo que me hizo sentir que no estaba solo en la carrera y que había más personas en la misma situación que yo.

A mis maestros, por brindarme los conocimientos y la formalidad necesarios en un matemático. Por mostrarme lo que es ir a la escuela realmente, pues tras los amargos tragos de la educación básica, fueron ellos quienes con tanta pasión por enseñar lo que a ellos les apasiona, me mostraron como es el proceso de aprendizaje cuando de verdad te gusta cada aspecto de lo que aprendes.

A mi asesor y a mis tutores de la licenciatura y la maestría, por ser para mi guías en cada paso. Por exigirme trabajar a tal ritmo de trabajo y asegurarse que lo cumpliera. Por orillarme a hacerme tantas preguntas y apoyarme más allá de este trabajo final.

Estimador por calibración

Problemas en su implementación y propuestas de soluciones

Víctor Miguel García Sánchez

Resumen

La estimación de los totales de variables obtenidas en encuestas siempre ha sido de gran interés. En ocasiones, el no contar con información completa al final de la encuesta o desear predecir un total antes de contar con las respuestas de la población entera, nos lleva a tomar una muestra con información completa y, para evitar la sobrerrepresentación, calcular un total ponderado de la variable de interés para obtener una buena estimación.
Los ponderadores más utilizados, entre cuyas aplicaciones destaca la estimación de totales, son los inversos de las probabilidades de inclusión en la muestra, llamados también *factores de expansión*. Su uso se debe a que brindan una estimación insesgada, sin embargo, el error cuadrático medio (ECM) obtenido con tales estimaciones suele ser muy grande.
Una manera de reducir el ECM es la incorporación de información auxiliar para ajustar simultáneamente los ponderadores. Tal información está conformada por los totales de variables auxiliares. A los ponderadores obtenidos los denominamos *ponderadores calibrados*, que son lo más cercanos posibles a los iniciales bajo ciertas restricciones impuestas por dichos totales.
Éstas restricciones son que al calcular los totales de las variables auxiliares, usando los ponderadores calibrados, se obtengan los totales conocidos para dichas variables.

Índice de figuras

Índice de tablas

Introducción

El muestreo no es una mera sustitución de una cobertura parcial a la totalidad.

William Edwards Deming
Some Theory of Sampling(1950)

Pocas cosas son tan confusas para los investigadores aplicados como el rol de la ponderación de muestras[1]. La confusión se debe a la falta de claridad sobre cuál de los múltiples motivos potenciales de ponderación se aplica al proyecto de investigación en cuestión.

Consideremos la tasa de pobreza de Estados Unidos en 1967, que oficialmente se midió en 13 % según la Encuesta de Población Actual organizada por el U.S. Bureau of the Census (1968). Si se estima la tasa de pobreza de Estados Unidos en 1967 con la tasa de pobreza de la muestra completa del Panel Study of Income Dynamics (PSID) de 1968, sin ponderación alguna que ajuste la sobremuestra de bajos ingresos que caracteriza al PSID, el estimado sería de 26 %.

Sin embargo, uno puede lograr una estimación insesgada y consistente utilizando la tasa de pobreza ponderada de la muestra del PSID, ésto mediante el *estimador de Horvitz-Thompson*, que consiste en ponderar con los inversos de las probabilidades de selección, con la que resulta una tasa de pobreza del 12 %, una estimación más razonable que 26 %.

Este ejemplo, ilustrado en Haider y Wooldridge (2013), se muestra un caso simple de la estimación de una media poblacional, en la que nos basamos en una muestra que sistemáticamente no representa a la población objetivo, pero puede representarla por ponderación.

Utilizar como ponderadores a los inversos de las probabilidades de inclusión es lo más usual en la práctica, pues la estimación obtenida es insesgada. A pesar de tal ventaja, el Error Cuadrático Medio (ECM) y la varianza de las estimaciones para el anterior caso suele ser muy grande.

Cuando conocemos los totales de algunas variables auxiliares, sean cuantitativas o categóricas, podemos cambiar los ponderadores por otros, lo más cercanos posibles a los iniciales y sujetos a que al utilizarlos para calcular los totales ponderados de las variables auxiliares en la muestra, éstos coincidan con los totales conocidos.

[1] Angrist y Pischke (2009)

A los ponderadores obtenidos, los llamamos *ponderadores por calibración* o *ponderadores calibrados*, el estimador obtenido con ellos está justificado por una relación de regresión de la variable de estudio con el vector formado por las variables auxiliares, como se describe en Särndal y Deville (1992).
Aunque tal proceso suele reducir el ECM, también puede ocasionar problemas, en ocasiones asociados a la muestra[2]. En el presente trabajo se describen algunos de esos problemas junto con algunas propuestas para solucionarlos.

Justificación

Mi motivación en éste trabajo se dio en el Seminario internacional sobre edición de datos, imputación y no respuesta, celebrado en la ciudad de Guanajuato, Gto. el 26 y 27 de octubre de 2017. En una de sus participaciones en dicho evento, Joseph L. Schafer mencionó que en ocasiones la imputación en bases de datos, además de introducir errores y brindar intervalos de confianza inválidos, no satisface los objetivos con que se diseñaron los instrumentos de recolección de información.
Uno de los parámetros que más nos interesa obtener es el total de ciertas variables, sin importarnos realmente los valores de la misma para cada individuo de la población.
La simulación de bases de datos puede llevar a obtener algunas correlacionadas de manera poco natural, al grado de obtener multicolinealidad entre las variables de estudio y las variables auxiliares. Para evitar tal fenómeno se prefirió inicialmente el uso de bases de datos de encuestas reales, pues al contar con información real, no se tiene un modelo que pueda influir en el diseño de la estimación del total de interés.
La base de datos de la *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2016*[3], es de gran interés para la sociedad y organismos como el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). Información como el ingreso corriente total o la inversión total en programas sociales son cantidades de gran relevancia en los ámbitos político y económico de nuestro país. Por ello que algunos parámetros de interés son estimados al final del presente trabajo con la finalidad de poner en práctica el uso del estimador por calibración, además de hacer mención de otras aplicaciones no tan directas de la estimación por calibración.

Antecedentes

Para tener un contexto del desarrollo de la estimación por calibración, se expone un breve marco histórico de la estimación por calibración.
En el artículo de Stephan y Deming (1940) se realizó un ajuste mediante el método de mínimos cuadrados sobre la tabla de frecuencias del censo de población de 1940 para incorporar la información de totales conocidos.

[2] Huang y Fuller (1978) pp. 2
[3] INEGI (2017a)

En dicho artículo, los autores utilizan un enfoque de regresión por mínimos cuadrados. Esto dio origen a la estimación por calibración. Al ser el método de regresión el más usado para distintos fines, entre los que no podría faltar la estimación de totales, se mantuvo el nombre de *estimación por regresión* por varios años.

Es así como en Huang y Fuller (1978), la tesis doctoral de Huang, presentada en 1978, ya se hacía uso del método de calibración para la estimación de medias. En dicho documento, se refiere al estimador como *estimador por regresión* y a los ponderadores obtenidos en función de la información auxiliar los llama *ponderadores por regresión*.

Aunque en dicha tesis desarrolla un algoritmo de computadora diseñado para calcular los ponderadores, el enfoque que aún se daba en esa época para la calibración, como ya se mencionó, fue usando el método de mínimos cuadrados, es decir, se limitaban al uso de la distancia χ-cuadrada, como se puede verificar en Huang y Fuller (1978, sección IV.A).

Además en dicha tesis, Huang menciona que al contar con muestras muy pequeñas o muestras no aleatorias, algunos de los ponderadores obtenidos pueden llegar a ser negativos. Los ponderadores negativos, con justa razón, son inaceptables para algunos usuarios, debido a que pueden llevar a producir estimados negativos de parámetros poblacionales que sabemos que son positivos.

Autores como Särndal, Estevao, Deville y Wu popularizaron el tema; pues desde 1992 con el artículo de Särndal y Deville (1992), se ha vuelto un instrumento metodológico importante en la producción de estadísticas. Tanto es así, que se han desarrollado numerosos paquetes de software diseñado para calcular ponderadores calibrados basados en información auxiliar disponible en encuestas y registros de población. Por mencionar algunos nos encontramos con CALMAR, GES y CLAN97.

La popularidad del tema se adjudica a que en el artículo mencionado se prueba que el estimador por calibración es asintóticamente equivalente al estimador por regresión generalizada. Además de tras comenzar, de manera arbitraria, con la distancia χ-cuadrada, caracterizan la clase de distancias que cumplen el objetivo de mantener a los nuevos ponderadores cercanos a los iniciales.

Hipótesis

Lo que se espera, al proponer el método de estimación por calibración, es reducir el ECM de las estimaciones respecto al obtenido con las estimaciones de Horvitz-Thompson. Además de identificar, y de ser posible solucionar, distintos problemas que suelen presentarse en el proceso de estimación.

Objetivos

La finalidad de este trabajo es evaluar la efectividad del estimador por calibración para reducir el error cuadrático medio de las estimaciones de totales, al tratarse de estimadores insesgados, se buscará mejorar otras características

de las mismas. Determinar distintos tipos de problemas que se presentan al efectuar el proceso de calibración y proponer soluciones para resolverlos.

Objetivos generales

El objetivo general es evaluar de manera cuantitativa las ventajas de estimar totales poblacionales mediante el estimador por calibración y determinar los errores que se pueden presentar durante el proceso, así como identificar mecanismos de solución para los mismos.

Objetivos específicos

- Verificar las ventajas de la estimación por calibración por encima de la estimación de Horvitz-Thompson.
- Identificar distintas problemáticas que se suelen presentar durante las estimaciones por calibración.
- Exponer soluciones encontradas en la literatura para resolver las complicaciones que pueden presentarse.
- Evaluar las ventajas y desventajas de tales soluciones.
- Presentar propuestas propias de solución, incorporando lo mejor de las soluciones previas.

Delimitación

En este trabajo se simulan algunas bases de datos en las que se seleccionan muestras previamente elegidas para denotar algunos problemas que se presentan en el proceso de calibración. Tras presentar algunas propuestas de solución, se prueban en las bases de datos simuladas y posteriormente en la base de datos de la ENIGH 2016.
Por simplicidad, se trabajará unicamente con variables cuantitativas en todos los casos.

Estructura del documento

La presente tesis está formada por 4 capítulos, los cuales son: ¿Qué es la estimación por calibración?, Aspectos importantes de la calibración, Algunos problemas al calibrar y cómo evitarlos, Implementación y algunas aplicaciones del estimador por calibración. Cada uno de ellos está dividido en secciones y éstas últimas en subsecciones. Llevan un orden específico, siempre tomando en cuenta que cada tema va ligado con la teoría desarrollada en los apartados anteriores.
El primer capítulo de la tesis se divide en 5 secciones, comenzando por la definición del estimador por calibración, clave para comprender el resto del contenido, y tras mencionar las condiciones mínimas para efectuar la calibración,

obtener los ponderadores calibrados. Al final del capítulo se hace uso del enfoque de regresión lineal para la estimación de totales con la finalidad de verlo como un caso particular de la calibración.
El estudio de Aspectos importantes de la calibración corresponde al segundo capítulo. Se aborda el uso de otras distancias para determinar ponderadores calibrados y se estudian algunas medidas de tendencia y dispersión de las estimaciones. Además se estudian brevemente los aspectos computacionales involucrados en el cálculo de ponderadores, en los que se habla del uso de distintas distancias.
El tercer capítulo corresponde a la parte central de la tesis: Algunos problemas al calibrar y como evitarlos. Se comienza estudiando algunas bases de datos obtenidas mediante la simulación de variables aleatorias con determinación controlada, sección en la que se exponen a manera de ejemplos distintos problemas que surgen en la calibración. Posteriormente se explican los tipos de problemas y se revisan las propuestas de solución encontradas en la bibliografía.
En el capítulo final, con información de la base de datos de la ENIGH 2016, se implementa el método de calibración para realizar varias estimaciones del total de ingresos corrientes en hogares. Posteriormente se determinan sus medidas de tendencia y dispersión, con la finalidad de verificar en una base de datos real la hipótesis planteada.
Además del uso del método de calibración para la estimación de totales, en el cuarto capítulo se estudia el uso del mismo para la estimación de cuantiles, calibración en diseños de muestreo multifase, proporciones entre distintos totales, etc.
En el primer anexo se incluyen algunas demostraciones útiles para complementar el texto, entre las que se incluye que el estimador de Horvitz-Thompson es insesgado, para contrastar con la dificultad de obtener un resultado análogo para el estimador por calibración. Lo siguientes anexos exponen los códigos utilizados para los ejemplos, estimaciones y otros cálculos efectuados para la realización del trabajo.
Todo lo anterior hace de este texto una guía de consulta teórica y práctica para aprender el método de estimación por calibración y la identificación de motivos por los que se obtienen errores al usarlo, además de incluir posibles soluciones según sea el error presentado.

CAPÍTULO 1

¿Qué es la estimación por calibración?

El método de **calibración** se refiere a lo siguiente:

1. El cálculo de ponderadores que agreguen información auxiliar y que estén restringidos por ciertas ecuaciones.

2. El uso de dichos ponderadores para calcular estimadores lineales de totales y otros parámetros finitos.

3. La obtención de **estimaciones por diseño casi insesgadas**, siempre que no haya respuestas incompletas y otros errores no muestrales.

En la literatura, lo usual es que el término se refiera solo al primer inciso. En contraste con Särndal (2007), quien lo usa solo para el primero y el tercero, en este trabajo nos referiremos a **calibración** indistintamente a cualquiera de las tres aplicaciones anteriores.
Una motivación a la definición formal del tema es que se trata de un método de re-ponderación que se usa cuando contamos con variables (cualitativas o cuantitativas), con las que se desea realizar, conjuntamente, un ajuste sobre los ponderadores.
Al utilizar como ponderadores a los inversos de las probabilidades de inclusión, la estimación de totales de variables de interés es insesgada. Tal estimador recibe el nombre de *estimador de Horvitz-Thompson*.
Por otro lado, la estimación por calibración ajusta los ponderadores por nuevos que son llamados *ponderadores por calibración*, *ponderadores calibrados* o ponderadores finales. Los ponderadores calibrados proporcionan estimados consistentes con el diseño y que generalmente tienen un ECM menor que el del estimador de Horvitz-Thompson.

1.1 Condiciones mínimas para la estimación basada en diseño en muestreo de encuestas

Consideremos muestreos de una sola fase y sin no respuesta. En la práctica, condiciones así no son tan sencillas, sin embargo, la mayoría de artículos teóricos lo asumen.
Tomemos un muestreo probabilístico S de la población finita $U = \{1, 2, \ldots, N\}$. El diseño de muestreo induce, para cada $k \in U$, una probabilidad de inclusión π_k, que conocemos, y que a su vez nos brinda un ponderador $d_k = \frac{1}{\pi_k} > 1$. Nuestro objetivo es estimar el total

$$t_y = \sum_{k \in U} y_k$$

a partir del uso de información auxiliar, donde la variable de estudio y puede ser continua o categórica.
Definamos la variable aleatoria $S_k = \mathbb{1}_S(k)$[1], con la que obtenemos el estimador insesgado básico de diseño de t_y:

$$\hat{t}_{y_{HT}} = \sum_{k \in U} d_k y_k S_k, \tag{1.1}$$

el cual es ineficiente[2] cuando contamos con variables auxiliares disponibles durante la fase de estimación, pues aún siendo insesgado, según los ponderadores de la muestra es usual obtener estimadores demasiado grandes o demasiado pequeños.
La notación general para los vectores formados por las variables auxiliares es $\mathbf{x}_k$. Distingamos dos casos respecto a $\mathbf{x}_k$:

I $\mathbf{x}_k$ es conocido para todo $k \in U$ (información auxiliar completa).

II $\mathbf{x}_k$ (observado) para todo $k \in S$ y $\sum_{k \in U} \mathbf{x}_k$ (importado) son conocidos. Llamamos importado al vector de totales porque a veces se toma de una fuente de información ajena a la encuesta.

El caso I proporciona una libertad considerable en la estructuración del vector auxiliar $\mathbf{x}_k$. Por ejemplo, si x_k es una variable continua con valor especificado para cada $k \in U$, entonces podemos considerar x_k^2 y otras funciones de x_k para su consideración, porque los totales como $\sum_{k \in U} x_k^2$ y $\sum_{k \in U} \log x_k$ son fácilmente calculables. Si la relación con la variable de estudio y no es lineal, puede ser una grave equivocación no tomar en cuenta totales conocidos, como el cuadrático o el logarítmico.
A diferencia del estimador por regresión, del que hablaremos en la subsección 1.5.1, el método por calibración no se refiere a algún modelo en específico.
El material disponible para estimar el total poblacional $t_y = \sum_{k \in U} y_k$ es:

1. Las variables observadas y_k para todo $k \in S$.

[1] Podemos notar que $\mathbb{E}(S_k) = \pi_k$
[2] Es decir, contando con la misma información se obtienen estimaciones menos precisas.

2. Las ponderaciones de diseño $d_k = \frac{1}{\pi_k}$ para todo $k \in S$.

3. Los vectores de valores conocidos $\mathbf{x}_k$ para todo $k \in U$ (o los valores observados $\mathbf{x}_k$ para todo $k \in S$ y el total importado $\sum_{k\in U} \mathbf{x}_k$).

Éstas condiciones fueron establecidas en Särndal y Deville (1992).

1.2 La distancia para calibrar

En éste método, se modifican los ponderadores iniciales d_k por nuevos ponderadores w_k, determinados para ser "cercanos" a los d_k. Para ello definamos una función distancia $G_k : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}_0^+$ continuamente diferenciable respecto a w, estrictamente convexa y con derivada $\frac{\delta G_k(w,d)}{\delta w} =: g_k(w, d)$, tal que $g_k(d, d) = 0$.
Usualmente elegimos G_k de manera que $g_k(w, d) = \frac{g\left(\frac{w}{d}\right)}{q_k}$, donde $g(\cdot)$ recibe un solo argumento, es continua y estrictamente creciente. Además g satisface $g(1) = 0$, $g'(1) = 1$ y los q_k son factores de escala positivos.
Para terminar este preámbulo, definamos $F(u) = g^{-1}(u)$, la función inversa de $g(\cdot)$.

1.3 Definición del modelo

Deseamos estimar el total desconocido

$$t_y = \sum_{k\in U} y_k. \tag{1.2}$$

Es así como a partir de conocer el estimador del estimador de Horvitz-Thompson 1.1, el estimador por calibración es

$$\hat{t}_{y_{cal}} = \sum_{k\in U} w_k y_k S_k, \tag{1.3}$$

donde los w_k son tales que minimizan $\sum_{k\in U} G_k(w_k, d_k)S_k$, sujetos a

$$\sum_{k\in U} w_k \mathbf{x}_k S_k = \sum_{k\in U} \mathbf{x}_k =: t_{\mathbf{x}}.$$

1.4 Determinación de los ponderadores calibrados w_k

Definamos el Lagrangiano

$$\mathcal{L}(d, w, \lambda) = \sum_{k\in U} G_k(w_k, d_k)S_k - \lambda^T \left(\sum_{k\in U} w_k \mathbf{x}_k S_k - t_{\mathbf{x}} \right).$$

Las derivadas parciales del Lagrangiano respecto a cada w_k son

$$\frac{\partial \mathcal{L}(d, w, \lambda)}{\partial w_k} = g_k(w_k, d_k)S_k - \lambda^T \mathbf{x_k} S_k.$$

Cuando g_k es de la forma descrita en la sección 1.2,

$$\frac{\partial \mathcal{L}(d,w,\lambda)}{\partial w_k} = \frac{g\left(\frac{w_k}{d_k}\right)}{q_k} S_k - \lambda^T \mathbf{x_k} S_k.$$

Para los $k \in S$ tenemos que, de existir tal λ, cuando igualamos a 0 para optimizar,

$$\begin{aligned} \frac{\partial \mathcal{L}(d,w,\lambda)}{\partial w_k} &= 0 \\ \Leftrightarrow \frac{g\left(\frac{w_k}{d_k}\right)}{q_k} - \lambda^T \mathbf{x_k} &= 0 \end{aligned}$$

y por ello

$$w_k = d_k F(q_k \lambda^T \mathbf{x_k}). \tag{1.4}$$

Por las condiciones pedidas, las w_k obtenidas minimizan $\sum_{k\in U} G_k(w_k, d_k) S_k$ y, como demuestran Särndal y Deville (1992), son únicas. También en palabras de Särndal y Deville, el estimador obtenido está justificado por una relación de regresión entre la variable de interés y y el vector de variables auxiliares $\mathbf{x}$.

1.5 Acerca de la existencia de λ

Para exponer la plausibilidad del método de calibración es necesario hablar sobre la existencia de λ, que en la sección 1.4 solo supusimos. En la siguiente subsección estudiaremos la aplicación de un modelo con el que estamos familiarizados, el de regresión lineal, para la estimación de totales y la manera en que podemos expresar tal estimación en términos de estimación por calibración.

1.5.1 Un caso especial: El estimador por regresión

Con la misma notación, tenemos el objetivo de estimar el total desconocido

$$t_y = \sum_{k\in U} y_k$$

a partir de conocer $(y_k, \mathbf{x}_k)$ para todos los $k \in S$ y contar con la información auxiliar completa.
El estimador de mínimos cuadrados de y bajo el modelo

$$y_k = B_{ols}^T \mathbf{x}_k + \sigma_k \epsilon_k, \tag{1.5}$$

usando para estimar solo los elementos de la muestra, está dado por

$$\hat{B}_{ols} = \left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(\sum_{k\in U} \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \right) \tag{1.6}$$

Así que podemos expresar la ecuación 1.2 mediante

$$\begin{aligned} t_y &= \sum_{k\in U} \left(\hat{B}_{ols}^T \mathbf{x}_k + \sigma_k \epsilon_k \right) \\ &= \hat{B}_{ols}^T \sum_{k\in U} \mathbf{x}_k + \sum_{k\in U} \sigma_k \epsilon_k, \end{aligned}$$

lo cual podemos reescribir como

$$= \hat{B}_{ols}^T t_{\mathbf{x}} + t_{\sigma\epsilon}.$$

A su vez, como

$$\begin{aligned} t_{\sigma\epsilon} &= \sum_{k\in U} \sigma_k \epsilon_k \\ &= \sum_{k\in U} (y_k - \hat{B}_{ols}^T \mathbf{x}_k) \end{aligned}$$

el cual es posible estimar mediante

$$\begin{aligned} \hat{t}_{\sigma\epsilon_{HT}} &= \sum_{k\in U} d_k (y_k - \hat{B}_{ols}^T \mathbf{x}_k) S_k \\ &= \sum_{k\in U} d_k y_k S_k - \sum_{k\in U} d_k \hat{B}_{ols}^T \mathbf{x}_k S_k \\ &= \sum_{k\in U} d_k y_k S_k - \hat{B}_{ols}^T \sum_{k\in U} d_k \mathbf{x}_k S_k \\ &= \hat{t}_{y_{HT}} - \hat{B}_{ols}^T \hat{t}_{\mathbf{x}_{HT}} \end{aligned}$$

Entonces el estimador para t_y, siguiendo el método de mínimos cuadrados es:

$$\begin{aligned} \hat{t}_{y_{ols}} &= \hat{B}_{ols}^T t_{\mathbf{x}} + \hat{t}_{y_{HT}} - \hat{B}_{ols}^T \hat{t}_{\mathbf{x}_{HT}} \\ &= \hat{t}_{y_{HT}} + \hat{B}_{ols}^T \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \end{aligned}$$

En Huang y Fuller (1978), tal expresión es llamada *estimador por regresión*. Veamos que es posible interpretar este estimador como uno obtenido mediante el método de calibración, pues al escribir lo anterior de la forma

$$\begin{aligned} \hat{t}_{y_{ols}} &= \hat{t}_{y_{HT}} + \hat{B}_{ols}^T \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \\ &= \hat{t}_{y_{HT}} + \left[\left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(\sum_{k\in U} \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \right) \right]^T \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \\ &= \hat{t}_{y_{HT}} + \left\{ \left[\left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(\sum_{k\in U} \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \right) \right]^T \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \right\}^T \end{aligned}$$

El paso anterior es posible al considerar al resultado escalar como una matriz de 1×1, que es simétrica. Al usar una propiedad de la transpuesta,

$$= \hat{t}_{y_{HT}} + \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right)^T \left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(\sum_{k\in U} \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \right)$$

Definamos ahora

$$\lambda = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \tag{1.7}$$

Con lo que

$$\begin{aligned} \hat{t}_{y_{ols}} &= \hat{t}_{y_{HT}} + \lambda^T \left(\sum_{k \in U} \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \right) \\ &= \hat{t}_{y_{HT}} + \sum_{k \in U} \lambda^T \frac{\mathbf{x}_k y_k d_k S_k}{\sigma_k^2} \end{aligned}$$

al reordenar los escalares,

$$\begin{aligned} &= \sum_{k \in U} d_k y_k S_k + \sum_{k \in U} d_k \frac{1}{\sigma_k^2} \lambda^T \mathbf{x}_k y_k S_k \\ &= \sum_{k \in U} d_k \left(1 + \frac{1}{\sigma_k^2} \lambda^T \mathbf{x}_k \right) y_k S_k \end{aligned}$$

Definamos $G_k(w_k, d_k) := \frac{\sigma_k^2}{2d_k} (w_k - d_k)^2$, la distancia χ-cuadrada, con lo que $g_k(w_k, d_k) = \frac{\sigma_k^2}{d_k} (w_k - d_k)$. Así que $q_k := \frac{1}{\sigma_k^2}$ y $g\left(\frac{w_k}{d_k}\right) := \frac{w_k}{d_k} - 1$, con inversa $F(u) = 1 + u$.
De la ecuación 1.4, podemos ver que bajo las condiciones anteriores, podemos tomar

$$w_k = d_k F\left(q_k \lambda^T \mathbf{x}_k \right)$$

y con ello

$$\hat{t}_{y_{ols}} = \sum_{k \in U} w_k y_k S_k$$

Con lo que obtuvimos $\hat{t}_{y_{ols}}$ en forma de estimador por calibración.
Como pudimos ver, tal elección de G_k satisface las condiciones pedidas en la sección 1.2 y los ponderadores w_k, inducidos por el uso de dicha distancia, brindan la estimación del total $\hat{t}_{y_{ols}}$. Sin embargo, también es necesario verificar que los ponderadores calibrados satisfacen las ecuaciones de calibración.

Lema 1.1. *Los ponderadores* $w_k = d_k \left(1 + \frac{1}{\sigma_k^2} \lambda^T \mathbf{x}_k \right)$ *satisfacen las ecuaciones*

$$\sum_{k \in U} w_k \mathbf{x}_k S_k = t_{\mathbf{x}}$$

Demostración.

$$\begin{aligned} \sum_{k \in U} w_k \mathbf{x}_k S_k &= \sum_{k \in U} d_k \left(1 + \frac{1}{\sigma_k^2} \lambda^T \mathbf{x}_k \right) \mathbf{x}_k S_k \\ &= \sum_{k \in U} d_k \mathbf{x}_k S_k + \sum_{k \in U} d_k \frac{1}{\sigma_k^2} (\lambda^T \mathbf{x}_k) \mathbf{x}_k S_k \end{aligned}$$

Considerando la matriz $\lambda^T \mathbf{x}_k$, de 1×1, como un escalar, que al igual que d_k y S_k, puede conmutar con el resto de factores,

$$\sum_{k \in U} w_k \mathbf{x}_k S_k = \sum_{k \in U} d_k \mathbf{x}_k S_k + \sum_{k \in U} \frac{1}{\sigma_k^2} \mathbf{x}_k d_k S_k (\lambda^T \mathbf{x}_k)$$

Considerando ahora que $\lambda^T \mathbf{x}_k$ es simétrica,

$$\begin{aligned} &= \sum_{k \in U} d_k \mathbf{x}_k S_k + \sum_{k \in U} \frac{1}{\sigma_k^2} \mathbf{x}_k d_k S_k (\mathbf{x}_k^T \lambda) \\ &= \sum_{k \in U} d_k \mathbf{x}_k S_k + \sum_{k \in U} \frac{1}{\sigma_k^2} \mathbf{x}_k \mathbf{x}_k^T d_k S_k \lambda \\ &= \sum_{k \in U} d_k \mathbf{x}_k S_k + \sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \lambda \end{aligned}$$

Para terminar basta con sustituir λ y notar que la primera suma es $\hat{t}_{\mathbf{x}_{HT}}$, así que

$$\begin{aligned} \sum_{k \in U} w_k \mathbf{x}_k S_k &= \hat{t}_{\mathbf{x}_{HT}} + \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right) \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^T d_k S_k}{\sigma_k^2} \right)^{-1} \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \\ &= \hat{t}_{\mathbf{x}_{HT}} + \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}} \right) \\ &= t_{\mathbf{x}} \end{aligned}$$

Con lo anterior concluimos que los ponderadores w_k satisfacen las ecuaciones de calibración. ■

Aunque con lo anterior pudimos ver que es posible expresar el estimador por regresión en términos del estimador por calibración, no es parte del pensamiento de regresión lineal, cuya idea central se formuló con la ecuación 1.5. Sin embargo, resulta interesante interpretar al estimador por regresión como un caso particular del estimador por calibración.

1.5.2 Una forma explícita de λ

Al aplicar el método de multiplicadores de Lagrange se obtienen simultáneamente tanto los ponderadores por calibración w_k, para cada $k \in S$, como cada entrada del vector λ de multiplicadores.
En la sección 1.4 obtuvimos tales w_k, tras suponer la existencia de λ. Por otro lado en esta subsección, suponiendo que conocemos los ponderadores w_k, obtendremos λ.
Como veremos en la sección 2.4, algunos autores, entre los que destacan Deville *et al.* y Särndal, han utilizado herramientas computacionales para obtener los valores de los ponderadores w_k.
Ya que contamos con $\{w_k\}_{k \in S}$, procedamos a obtener una expresión para λ, partiendo de que al efectuar el proceso de multiplicadores de Lagrange,

$$\begin{aligned} g_k(w_k, d_k) S_k - \lambda^T \mathbf{x}_k S_k &= 0 \\ g_k(w_k, d_k) \quad S_k = \quad \lambda^T \mathbf{x}_k \quad S_k & \end{aligned}$$

Al multiplicar por $q_k d_k \mathbf{x}_k^T$,

$$q_k d_k g_k(w_k, d_k)\mathbf{x}_k^T S_k = q_k d_k \lambda^T \mathbf{x}_k \mathbf{x}_k^T S_k$$
$$d_k g\left(\frac{w_k}{d_k}\right)\mathbf{x}_k^T S_k = \lambda^T q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k$$

Al sumar sobre todas las $k \in U$

$$\sum_{k\in U} d_k g\left(\frac{w_k}{d_k}\right)\mathbf{x}_k^T S_k = \sum_{k\in U} \lambda^T q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k$$
$$= \lambda^T \sum_{k\in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k$$

Si $\sum_{k\in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k$ es invertible,

$$\lambda^T = \left[\sum_{k\in U} d_k g\left(\frac{w_k}{d_k}\right)\mathbf{x}_k^T S_k\right]\left[\sum_{k\in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k\right]^{-1}$$

Para finalizar,

$$\lambda = \left[\sum_{k\in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k\right]^{-1}\left[\sum_{k\in U} d_k g\left(\frac{w_k}{d_k}\right)\mathbf{x}_k S_k\right] \tag{1.8}$$

Podemos notar que al tomar como G_k a la distancia χ-cuadrada con factores de escala $q_k = \frac{1}{\sigma_k^2}$, se obtiene la ecuación 1.7.

1.6 ¿La calibración se hace para algún modelo específico?

No hay un modelo asistido explícito, a menos que uno insistiera en que ciertas variables para su inclusión en el vector $\mathbf{x}_k$ se sumen a un esfuerzo de modelado serio. En cambio, los ponderadores se justifican principalmente por su coherencia con los controles establecidos.

Esto plantea la pregunta: ¿es importante motivar tal "calibración sin modelo" con una declaración explícita de un modelo? Los estadísticos piensan en términos de modelos, y acompañan procedimientos estadístico con la declaración de un modelo. De hecho, también en la explicación de la calibración, para establecer la relación de y con $\mathbf{x}$, de manera casi inconsciente pensamos en un modelo, incluso si es tan simple como un modelo lineal estándar.

En ese sentido cobra importancia la correcta elección de variables auxiliares como parte de la definición "modelo". Es por eso que descartamos variables, como se explica al final de la sección 3.2, antes de iniciar el proceso de calibración y no al momento de tener problemas durante la implementación del mismo.

Pero, ¿un modelo establecido ayudará a los usuarios y profesionales a comprender mejor el enfoque de calibración? Para la mayoría de ellos, el enfoque está perfectamente claro y transparente de todos modos. Así que, ¿La búsqueda de "el verdadero modelo con la verdadera estructura de varianza" traerá una precisión significativamente mejor para la mayor parte de las muchas estimaciones producidas en una gran encuesta del gobierno? Para tales aplicaciones no se requiere otra justificación más que la coherencia con controles establecidos y totales plausibles.

CAPÍTULO 2

Aspectos importantes de la calibración

2.1 Otras distancias y ponderadores

Además de la distancia $G_k(w_k, d_k) = \frac{\sigma_k^2}{2d_k}(w_k - d_k)^2$, con la que obtenemos el estimador de mínimos cuadrados, aquí exploraremos otras funciones distancia que cumplen las condiciones pedidas en la sección 1.2.

Tabla 2.1: Algunas funciones distancia entre ponderadores.

Caso	$G_k(w_k, d_k)$	$F(q_k u) = F(q_k \lambda^T \mathbf{x}_k)$
Mínimos cuadrados generalizada	$\frac{(w_k - d_k)^2}{2d_k q_k}$	$1 + q_k \lambda^T \mathbf{x}_k$
Raking ratio	$\frac{w_k \log(w_k/d_k) - w_k + d_k}{q_k}$	$\exp(q_k \lambda^T \mathbf{x}_k)$
Hellinger	$\frac{2\left(\sqrt{w_k} - \sqrt{d_k}\right)}{q_k}$	$(1 - q_k \lambda^T \mathbf{x}_k/2)^{-2}$
Entropía mínima	$\frac{-d_k \log(w_k/d_k) + w_k - d_k}{q_k}$	$(1 - q_k \lambda^T \mathbf{x}_k)^{-1}$
Mínimos cuadrados modificada	$\frac{(w_k - d_k)^2}{2w_k q_k}$	$(1 - 2q_k \lambda^T \mathbf{x}_k)^{-1/2}$
Mínimos cuadrados restringidos	$\begin{cases} \frac{(w_k - d_k)^2}{2d_k q_k} & \text{, si } L < \frac{w_k}{d_k} < U \\ \infty & \text{, en otro caso.} \end{cases}$	$\begin{cases} L & \text{, si } q_k \lambda^T \mathbf{x}_k \leq L - 1 \\ 1 + q_k \lambda^T \mathbf{x}_k & \text{, si } L - 1 < q_k \lambda^T \mathbf{x}_k < U - 1 \\ U & \text{, si } q_k \lambda^T \mathbf{x}_k \geq U - 1. \end{cases}$ [a]
Logit	$\frac{d_k}{A q_k}\left[\left(\frac{w_k}{d_k} - L\right) log\left(\frac{w_k/d_k - L}{1-L}\right) + \left(U - \frac{w_k}{d_k}\right) log\left(\frac{U - w_k/d_k}{U-1}\right)\right]$	$\frac{L(U-1) + U(1-L)\exp(A q_k \lambda^T \mathbf{x}_k)}{(U-1) + (1-L)\exp(A q_k \lambda^T \mathbf{x}_k)}$. [b]

[a] Para restringir los ponderadores, seleccionemos constantes L, U tales que $L < 1 < U$.
[b] El factor de ajuste A está definido como $A = \frac{U-L}{(1-L)(U-1)}$.

La tabla 2.1, que se encuentra en Särndal y Deville (1992), muestra algunas funciones distancia conocidas muy usadas para calibrar.
La distancia de *mínimos cuadrados generalizada* es la única que puede dar ponderadores negativos, lo cual puede llegar a ser inaceptable para algunos usuarios. Como mencionan Huang y Fuller (1978), los ponderadores negativos pueden brindar estimaciones negativas de parámetros poblacionales positivos.
La distancia conocida como *raking ratio* nos conduce a obtener valores de w_k que son extremadamente grandes respecto a los ponderadores de muestreo d_k, lo cual también puede ser inaceptable por motivos similares.
La distancia de *mínimos cuadrados restringidos* y la distancia *logit* tienen la propiedad de conducir los ponderadores calibrados a intervalos controlados. Gracias a dichas distancias, los ponderadores extremos pueden ser eliminados manteniendo las propiedades favorables de los estimadores resultantes.
En la sección 2.4 se describirán a detalle las ventajas y desventajas de cada una de las distancias.

2.2 Algunas propiedades estadísticas del estimador

El sesgo, el error cuadrático medio y la varianza son algunas de las principales características de los estimadores. Las medidas de dispersión nos interesan debido a que el estimador por calibración suele reducir la varianza del estimador de Horvitz-Thompson que, aunque es insesgado, suele presentar estimaciones demasiado variadas.

2.2.1 Sesgo

El sesgo de $\hat{t}_{y_{cal}}$ está dado por la expresión

$$Sesgo(\hat{t}_{y_{cal}}, t_y) = \mathbb{E}(\hat{t}_{y_{cal}}) - t_y.$$

Debido a que el estimador de Horvitz-Thompson es insesgado, como se demuestra en el Teorema A.1, podemos escribir lo anterior como

$$\begin{aligned}
&= \mathbb{E}(\hat{t}_{y_{cal}}) - \mathbb{E}(\hat{t}_{y_{HT}}) \\
&= \mathbb{E}(\hat{t}_{y_{cal}} - \hat{t}_{y_{HT}}) \\
&= \mathbb{E}\left(\sum_{k\in U} w_k y_k S_k - \sum_{k\in U} d_k y_k S_k\right) \\
&= \mathbb{E}\left(\sum_{k\in U} (w_k - d_k) y_k S_k\right) \\
&= \mathbb{E}\left(\sum_{k\in U} (d_k F(q_k \lambda^T \mathbf{x}_k) - d_k) y_k S_k\right)
\end{aligned}$$

$$Sesgo(\hat{t}_{y_{cal}}, t_y) = \mathbb{E}\left(\sum_{k\in U} d_k[F(q_k\lambda^T\mathbf{x}_k) - 1]y_k S_k\right)$$
$$= \sum_{k\in U} d_k y_k \mathbb{E}\left([F(q_k\lambda^T\mathbf{x}_k) - 1]S_k\right)$$

Para cumplir con el objetivo de obtener estimadores casi insesgados, se requiere que $\mathbb{E}\left(\sum_{k\in U} d_k[F(q_k\lambda^T\mathbf{x}_k) - 1]y_k S_k\right) \approx 0$. Entonces la calibración deberá esforzarse por tener desviaciones $|F(q_k\lambda^T\mathbf{x}_k) - 1|$ pequeñas y que faciliten la cancelación de sumandos. De hecho podemos notar que cuando F es la identidad, $\hat{t}_{y_{cal}}$ coincide con $\hat{t}_{y_{HT}}$.
La expresión $\mathbb{E}\left([F(q_k\lambda^T\mathbf{x}_k) - 1]y_k S_k\right)$ depende de λ^T, la cual, como vimos en la ecuación 1.8, incorpora información de toda la muestra, así que una expresión explícita del sesgo se vuelve mas complicada que para el estimador de Horvitz-Thompson.
Sea $\mathcal{S}$ es la colección de todas las muestras posibles de la población, podemos expresar la esperanza anterior como

$$\mathbb{E}\left([F(q_k\lambda^T\mathbf{x}_k) - 1]S_k\right) = \sum_{S\in\mathcal{S}} S_k[F(q_k\lambda^T\mathbf{x}_k) - 1]\mathbb{P}(k \in S). \tag{2.1}$$

La ecuación 2.1 es mucho más compleja de obtener que la mostrada en el Teorema A.1 para el sesgo del estimador de Horvitz-Thompson. En la sección 2.3, abordaremos una manera natural de aproximar el sesgo.

2.2.2 Error cuadrático medio

El segundo momento (centrado en el origen) del error es denominado error cuadrático medio y está dado por la expresión

$$ECM(\hat{t}_{y_{cal}}) = \mathbb{E}\left[(\hat{t}_{y_{cal}} - t_y)^2\right]$$
$$= \mathbb{E}\left[\hat{t}^2_{y_{cal}} - 2t_y\hat{t}_{y_{cal}} + t_y^2\right]$$
$$= \mathbb{E}\left[\hat{t}^2_{y_{cal}}\right] - 2t_y\mathbb{E}\left[\hat{t}_{y_{cal}}\right] + t_y^2.$$

Análogamente a la expresión del sesgo, podemos expresar lo anterior como

$$ECM(\hat{t}_{y_{cal}}) = \mathbb{E}\left[\left(\sum_{k\in U} d_k F(q_k\lambda^T\mathbf{x}_k)y_k S_k\right)^2\right] - 2t_y\mathbb{E}\left[\sum_{k\in U} d_k F(q_k\lambda^T\mathbf{x}_k)y_k S_k\right] + t_y^2,$$

donde de nuevo se presenta la dificultad de la incorporación de información de toda la muestra.

2.2.3 Varianza

La varianza del estimador por calibración está dada por

$$Var(\hat{t}_{y_{cal}}) = Var\left(\sum_{i\in U} w_i y_i S_i\right).$$

Al expresar la varianza de la suma como

$$\begin{aligned} Var(\hat{t}_{y_{cal}}) &= \sum_{i,j\in U} Cov(w_i y_i S_i, w_j y_j S_j) \\ &= \sum_{i,j\in U} Cov(d_i F(q_i \lambda^T \mathbf{x}_i) y_i S_i, d_j F(q_j \lambda^T \mathbf{x}_j) y_j S_j) \\ &= \sum_{i,j\in U} d_i d_j y_i y_j Cov(F(q_i \lambda^T \mathbf{x}_i) S_i, F(q_j \lambda^T \mathbf{x}_j) S_j), \end{aligned}$$

como se mencionó anteriormente, es deseable que $F(\cdot) = 1$, caso en el que la expresión anterior se logra simplificar a

$$\begin{aligned} &= \sum_{i,j\in U} d_i d_j y_i y_j Cov(S_i, S_j) \\ &= \sum_{i\in U} d_i d_j y_i y_j [\mathbb{E}(S_i S_j) - \mathbb{E}(S_i)\mathbb{E}(S_j)] \\ &= \sum_{i,j\in U} d_i d_j y_i y_j (\pi_{ij} - \pi_i \pi_j) \\ &= \sum_{i,j\in U} d_i d_j y_i y_j \left(\pi_{ij} - \frac{1}{d_i}\frac{1}{d_j}\right) \\ &= \sum_{i,j\in U} d_i d_j y_i y_j \pi_{ij} - \sum_{i,j\in U} y_i y_j, \end{aligned}$$

donde se ven involucradas las probabilidades de inclusión de segundo orden. Sin embargo, por simplicidad haremos uso de la relación

$$ECM(\hat{t}_{y_{cal}}) = Var(\hat{t}_{y_{cal}}) + Sesgo(\hat{t}_{y_{cal}}, t_y)^2.$$

De aquí que

$$Var(\hat{t}_{y_{cal}}) = ECM(\hat{t}_{y_{cal}}) - Sesgo(\hat{t}_{y_{cal}}, t_y)^2, \tag{2.2}$$

entonces usaremos las expresiones obtenidas en estas últimas secciones para la obtención de la varianza.

2.3 Estimando medidas de tendencia y dispersión de las estimaciones

Como vimos en la ecuación 2.1, para determinar las características del estimador requerimos de los valores $\mathbb{P}(k \in S)$ para cada k en cada muestra $S \in \mathcal{S}$. Una manera de evitar esta complicación es mediante el uso de la ley fuerte de los grandes números que, como nos explican Vapnik *et al.* (1971), al realizar varias estimaciones por calibración $\hat{t}_{y_{cal,1}}, \ldots, \hat{t}_{y_{cal,m}}$, es posible estimar $\mathbb{P}(\hat{t}_{y_{cal}})$ mediante la expresión empírica,

$$\frac{\sum_{i=1}^{m} \mathbb{1}_{(\hat{t}_{y_{cal,i}} = \hat{t}_y)}}{m} \quad \underset{m\to\infty}{\longrightarrow} \quad \mathbb{P}(\hat{t}_{y_{cal}} = \hat{t}_y) \qquad c.s. \tag{2.3}$$

La ecuación 2.3 nos permite estimar $\mathbb{E}(\hat{t}_{y_{cal}})$ mediante la esperanza empírica,

$$\sum_{i=1}^{m} \frac{\hat{t}_{y_{cal,i}}}{m} \quad \underset{m\to\infty}{\longrightarrow} \quad \mathbb{E}(\hat{t}_{y_{cal}}) \qquad c.s.,$$

con lo que a su vez es posible estimar el sesgo mediante

$$\sum_{i=1}^{m} \frac{\hat{t}_{y_{cal,i}} - t_y}{m} \quad \underset{m\to\infty}{\longrightarrow} \quad Sesgo(\hat{t}_{y_{cal}}, t_y) \qquad c.s. \tag{2.4}$$

Análogamente a la expresión anterior, podemos estimar empíricamente el $ECM(\hat{t}_{y_{cal}})$ con la expresión

$$\sum_{i=1}^{m} \frac{(\hat{t}_{y_{cal,i}} - t_y)^2}{m} \quad \underset{m\to\infty}{\longrightarrow} \quad ECM(\hat{t}_{y_{cal}}) \qquad c.s. \tag{2.5}$$

Para la varianza, al sustituir las ecuaciones 2.4 y 2.5 en la identidad 2.2 se tiene

$$\sum_{i=1}^{m} \frac{(\hat{t}_{y_{cal,i}} - t_y)^2}{m} - \left(\sum_{i=1}^{m} \frac{\hat{t}_{y_{cal,i}} - t_y}{m}\right)^2 \quad \underset{m\to\infty}{\longrightarrow} \quad Var(\hat{t}_{y_{cal}}) \qquad c.s. \tag{2.6}$$

2.4 Aspectos computacionales

Para calcular el estimador de la ecuación 1.3, requerimos de la obtención de los w_k y del vector de multiplicadores de Lagrange. En el artículo de Deville *et al.* (1993) explican el proceso llevado a cabo por el macro para SAS llamado *CALMAR*, el cual es ampliamente utilizado en la literatura. Aunque hacemos uso de la librería icarus para *R*, una versión para dicho lenguaje de CALMAR, el proceso que se sigue es descrito a continuación.
Se encuentra primero tal λ mediante el método de Newton al definir $\phi_S(\lambda) = t_{\mathbf{x}} - \hat{t}_{x_{HT}}$ y $\phi'_S(\lambda) = \frac{\partial \phi_S(\lambda)}{\partial \lambda}$.
Al iniciar con $\lambda_0 = 0$, se obtienen valores sucesivos λ_ν, con $\nu = 1, 2, \ldots$ a partir de

$$\lambda_{\nu+1} = \lambda_\nu + [\phi'_S(\lambda_\nu)]^{-1} \left[t_{\mathbf{x}} - \hat{t}_{x_{HT}} - \phi_S(\lambda_\nu)\right] \tag{2.7}$$

En la función de calibración de tal librería, se cuenta con la opción de elegir la función distancia entre los siguientes métodos:

1. El método *linear*, que usa la distancia de mínimos cuadrados.
2. El método *raking* o multiplicativo.
3. El método *logit*.
4. El método *truncated*, que usa la distancia de mínimos cuadrados restringidos.

Mencionaremos algunas ventajas y desventajas de cada método, descritas también por Estevao y Särndal (2006) y Deville *et al.* (1993):

- El método *linear* es computacionalmente el más rápido, pues requiere de solo de una iteración para la obtención de los multiplicadores λ y posteriormente de los ponderadores w_k. Como se observa al ejecutar el código del anexo B, en ocasiones aún tras 2500 iteraciones, no llega a converger.

  La ecuación 1.8 nos permite notar que esto se debe a que en tales casos no es posible invertir la matriz

  $$\sum_{k \in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k,$$

  debido a los vectores de información auxiliar $\mathbf{x}_k$ seleccionados en la muestra. Esto repercute a su vez en la existencia de $[\phi'_S(\lambda_\nu)]^{-1}$ en la ecuación 2.6. Otra desventaja de este método es que ocasionalmente proporciona ponderadores negativos o muy grandes, como se mencionó en el capítulo 2.1. Los consideraremos muy grandes cuando sean más de 4 veces mayores que los ponderadores iniciales.

- El método *raking* garantiza la obtención de ponderadores positivos, sin embargo no están acotados superiormente.

- Los métodos *logit* y *truncated* fueron creados para imponer restricciones adicionales a los ponderadores. Algunos usuarios solicitan valores estrictamente positivos, pues consideran que toda observación debe contribuir de manera positiva al total. Por otro lado, también es usual recurrir a la calibración para justificar la reducción de un exceso de representatividad. La desventaja que suele presentar esta opción es que limitar demasiado los ponderadores, puede llevar a no tener solución.

  Las cotas L y U no pueden ser tomadas arbitrariamente. Además de la restricción pedida $L < 1 < U$, en el anexo D se muestra un ejemplo en el que se requieren más restricciones. En la práctica, los valores extremos de L y U son determinados mediante simulaciones sucesivas en las que L crece hasta 1 (y U decrece hasta 1) hasta que el paquete es incapaz de resolver las ecuaciones de calibración.

  Con el fin de mantener la dispersión de los ponderadores similar a la inicial, se sugiere elegir L y U de manera que sus respectivas proporciones con la media geométrica de los ponderadores, sea la misma que la que poseen el menor y el mayor de los ponderadores iniciales a dicha cantidad.

CAPÍTULO 3

Algunos problemas al calibrar y cómo evitarlos

El objetivo del presente capítulo es mostrar algunos de los problemas que se presentan en la calibración, para luego explicar algunas propuestas de solución a los mismos encontradas en la literatura.
Los principales errores surgidos durante la calibración son principalmente de 3 tipos. Al igual que Deville *et al.* (1993) mostraremos, a manera de introducción, algunos ejemplos sencillos de los mismos.
Para contar con una base de datos de tamaño similar a la que se estudiará en la sección 4.1, simularemos unas bases de datos con $(266)^2$ observaciones de las variables x_1, x_2, y y ponderadores d.
Para los Ejemplos 1 y 2, donde $x_1 \sim U(9990, 10000)$, x_2 está en 2 subpoblaciones del mismo tamaño: $U_{x_{2,1}} \sim U(0, 10)$ y $U_{x_{2,2}} \sim U(9990, 10000)$. La variable de estudio y sigue el modelo lineal $y = 100 + 2x_1 + 2x_2 + e$, con $e \sim \mathcal{N}\left(0, \frac{1-R^2}{R^2} Var(\hat{y})\right)$[1] y un coeficiente de determinación $R^2 = 0.95$. Los ponderadores d son inversos a probabilidades de inclusión en una muestra con probabilidad proporcional al tamaño (PPT) del 3 %.

Ejemplo 1. Un primer caso de error se presenta cuando se selecciona una muestra en la que los vectores de variables auxiliares $\mathbf{x}_k = (1, x_{1_k}, x_{2_k})$ son de la forma $(1, x_{1_k}, 0)$. Dicho problema se debe a que las ecuaciones $\sum_{k \in U} w_k \mathbf{x}_k S_k = t_{\mathbf{x}}$ no tienen solución al no estar definidas para el total de x_2. Más aún, la matriz

$$\sum_{k \in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k,$$

que aparece en la ecuación 1.8 no es invertible, así que no existen el

[1] Ver teorema A.2

vector de multiplicadores λ y los ponderadores calibrados w_k.

Ejemplo 2. Cuando una de las variables auxiliares está separada en subpoblaciones muy alejadas, el total de dicha variable se ve muy influido por la subpoblación en la que el valor de tal variable es mayor.
Para ilustrar tal problema, realicemos lo siguiente:
Al notar que en una muestra con PPT del 3 % la cantidad de elementos de $U_{x_{2,2}}$ en la muestra son 704 ó 705, tomemos una muestra con PPT de 705 elementos de $U_{x_{2,2}}$ y otra de 1418 de $U_{x_{2,1}}$, obteniendo una muestra "artificial" del 3 % de la población, para luego calibrar los ponderadores. En el anexo E se realiza tal simulación, en la que además de obtener obtener una estimación con un error relativo de solo 4.14 %, aumentó la dispersión de los ponderadores como podemos ver en la figura 3.1 y se obtuvieron ponderadores menores a 1.

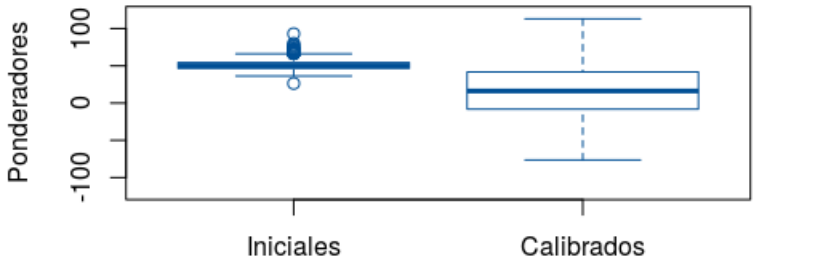

Figura 3.1: Efecto al calibrar en los ponderadores del Ejemplo 2.

Si se interpretaran como inversos de probabilidades de inclusión, éstas implicarían la existencia de probabilidades de inclusión mayores a 1. Además de los ponderadores menores a 1, la tabla 3.1 nos muestra que se obtienen ponderadores negativos y un ponderador demasiado grande.

Tabla 3.1: Ponderadores que al calibrar se vuelven menores a 1.

	Mínimo	Mediana	Media	Máximo
Iniciales	26.37	50.28	50.90	93.07
Calibrados	−76.33	16.23	33.33	35 383.27

La sobrerrepresentación de $U_{x_{2,2}}$ en la muestra, ocasiona que el total de x_2 se dispare. Para compensarlo, durante la calibración tales elementos son ajustados. El ajuste sobre los ponderadores d_k más pequeños es lo que brinda ponderadores calibrados menores a 1. El efecto de tal ajuste en x_2 es compensado en x_1, lo cual nos da el ponderador demasiado grande.

Para los Ejemplos 3 y 4, consideremos una base de datos con $x_1 \sim U(0, 10)$ y x_2 particionada en 2 subpoblaciones: $U_1 \sim U(0, 10)$, que representa el 99 % de la población, y $U_2 \sim U(9990, 10000)$. La variable de estudio y sigue el modelo lineal $y = 3000 + 2x_1 + 2x_2 + e$, donde el error e conserva las mismas características que en la base de datos anterior. Los ponderadores d son inversos

a probabilidades de inclusión en una muestra con PPT del 3 %.

Ejemplo 3. Con las condiciones anteriores, tomemos la muestra del 3 % de la manera "artificial" propuesta en el Ejemplo 2, aunque con $u_1 = 1$. Asistidos por el código del anexo F, notemos en la figura 3.2 que aún tras haber iniciado con ponderadores relativamente cercanos, obtenemos un ponderador muy grande, asociado al único elemento de la muestra perteneciente a U_1.

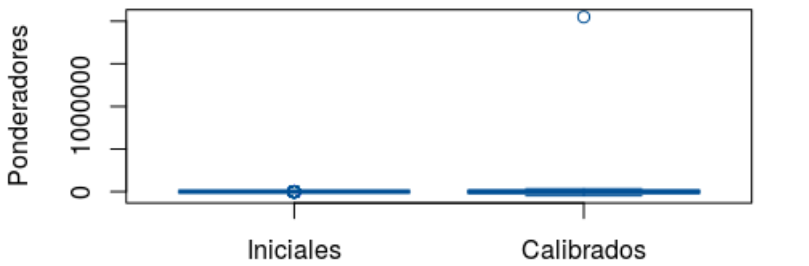

Figura 3.2: Cambio en los ponderadores del Ejemplo 3.

La tabla 3.2 nos permite notar que, para compensar la presencia de tal ponderador en el total t_{x_1}, se obtuvieron varios ponderadores negativos, los cuales causan que el estimador del total $t_y = 2.29 \cdot 10^8$ por calibración sea $\hat{t}_{y,cal} = -0.9 \cdot 10^8$. Con tal muestra fija se pasó de tener un error relativo cercano al 0 %, con el estimador de Horvitz-Thompson, a tener un error relativo de casi 140 %.

Tabla 3.2: Ponderadores que al ser calibrados dan uno demasiado grande y varios negativos.

	Mínimo	Mediana	Media	Máximo
Iniciales	24.57	35.90	36.97	86.50
Calibrados	−48 158.0	−895.6	33.3	2 049 737.8

Tal error se puede adjudicar a la muestra, pues en otra realización de la simulación obtenemos los ponderadores que se muestran en la tabla 3.3.

Tabla 3.3: Ponderadores que al ser calibrados se vuelven demasiado grandes o negativos.

	Mínimo	Mediana	Media	Máximo
Iniciales	24.57	35.83	36.85	86.50
Calibrados	−26 165.5	−589.0	33.3	1 410 780.6

Tal muestra pasó de tener un error relativo cercano al 0 %, con el estimador de Horvitz-Thompson, a tener un error relativo de 183.86 % con el estimador por calibración, al haber estimado con $\hat{t}_{y,cal} = 6.52 \cdot 10^8$ lo cual es más aceptable por ser positivo.

A medida que la presencia de u_1 en la muestra aumenta, los ponderadores

empiezan a normalizarse, aunque siguen presentando errores. Sin embargo, notemos que $|U_2| = 798$, por lo que al haber tomado una muestra de 2122 en el caso anterior, el efecto de sobrerrepresentación se ve acrecentado por el reemplazo en el muestreo.

Ejemplo 4. Bajo las mismas condiciones que se tienen en el Ejemplo 3, aunque tomando $u_2 = 16$, la cantidad de elementos de U_2 usual en una muestra con PPT del 3 %, podemos notar en la figura 4 que además de aumentar la dispersión, siguen apareciendo valores extremos.

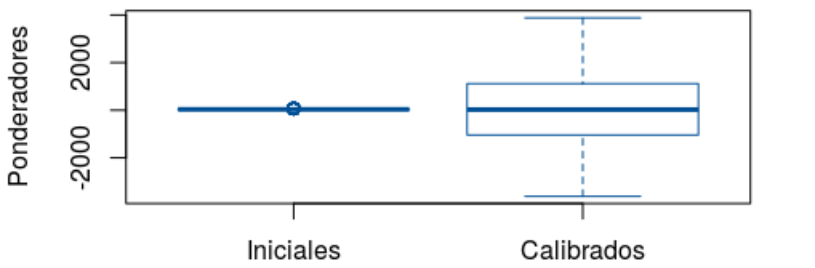

Figura 3.3: Cambio en los ponderadores de la primera parte del Ejemplo 4.

En contraste con el Ejemplo 2, los ponderadores calibrados pertenecientes a U_2 toman valores grandes de signos opuestos, aunque de menor tamaño por ser los elementos de U_2 más representativos en el total t_{x_2}. En este caso el error relativo de $\hat{t}_{y,cal}$ disminuye al 18.57 %.

Tabla 3.4: Efecto del aumento de la representación de una subpoblación.

	Mínimo	Mediana	Media	Máximo
Iniciales	22.75	36.67	38.17	92.36
Calibrados	−3619.97	25.70	33.33	3880.13

Al tomar ahora $u_2 = 2$, sobrerrepresentando la subpoblación U_1, podemos ver en la tabla 3.5 que se logra disminuir ligeramente la dispersión, además el error relativo de $\hat{t}_{y,cal}$ se reduce al 10.52 %. En otras realizaciones se reduce inclusive a 3.58 %

Tabla 3.5: Efecto del aumento de la representación de una subpoblación.

	Mínimo	Mediana	Media	Máximo
Iniciales	23.92	36.73	37.92	75.16
Calibrados	−3559.02	30.68	33.33	3371.67

Comenzamos mostrando que la alta representación de 0's en alguna de las variables auxiliares puede llegar a impedir la inversión de una matriz requerida en la calibración.

Posteriormente vimos que con la aparición de ponderadores negativos se puede llegar a obtener estimaciones negativas que, para variables positivas, es un error

bastante grave.
También notamos en algunos ejemplos que la aparición de ponderadores demasiado alejados del resto puede causar estimaciones alejadas del total real.

3.1 Limitantes numéricas

Como se mostró en la sección 1.5.1, es posible pensar en el estimador por regresión lineal como un caso particular de la estimación por calibración. Algunas limitantes numéricas están asociados a la imposibilidad de obtener una inversa para la matriz

$$\sum_{k \in U} q_k d_k \mathbf{x}_k \mathbf{x}_k^T S_k, \tag{3.1}$$

requerida para la obtención de λ como se vió en la ecuación 1.8. Aunque el uso de técnicas como la regresión Ridge permiten la inversión de la matriz, también ocasionan la pérdida del sentido del modelo. Lo anterior se refiere a que modificar la matriz 3.1 significa la alteración de respuestas en las variables auxiliares.
Aún si entre las variables auxiliares no se cuenta con variables categóricas, que se ven más afectadas por el efecto de la modificación antes mencionada, modificar la base de datos representa un error mayor que la modificación del modelo durante la etapa de estimación.
En Théberge (1999) se trabaja la calibración como un problema algebraico y, además de proponer funciones distancia inspiradas en el álgebra matricial, se usa como una técnica para obtener otros parámetros además de medias poblacionales. Cuando las ecuaciones de calibración no tienen solución, la población se descompone en dominios pequeños en los que si la tienen. Posteriormente se usa el producto de Kronecker para combinar los estimadores obtenidos en cada dominio.

3.2 Ponderadores negativos o demasiado grandes

Como mencionan Huang y Fuller (1978), el uso de ponderadores negativos puede llevar a obtener estimaciones negativas de parámetros que sabemos que son positivos. Es así como partiendo del método *linear*, con el objetivo de calcular ponderadores calibrados que no sean negativos, se añade como restricción a las soluciones de las ecuaciones de calibración, que éstas deben satisfacer $L < w_k < U$, con

$$L > 0.$$

Tal método, llamado *truncated* en la computación, corresponde a la distancia de mínimos cuadrados restringidos expuesta en la tabla 2.1. Como ya se mencionó en la sección 2.1, consiste en imponer cotas a las soluciones de las ecuaciones de calibración. Además de brindar una cota inferior, que trata con problemas como los presentados en los Ejemplos 3 y 4, el presente método acota superiormente las soluciones, lo cual permite tratar con casos como el del Ejemplo 2.

Otra manera de obtener ponderadores que no sean negativos, es partiendo de la distancia *raking* que, como se explica en la sección 2.1, proporciona solo ponderadores positivos, aunque no están acotados. En el método *logit* los ponderadores obtenidos mediante la distancia *raking* son ajustados para pertenecer a un intervalo $[L, U]$. La expresión de dicho ajuste está dada explícitamente en la distancia *logit* de la tabla 2.1.

3.3 Descarte de restricciones

Bankier *et al.* (1992) proponen una técnica a la se refieren como *descarte de restricciones*. En dicha técnica se determinan variables que muestran ser "problemáticas" durante la fase de estimación para luego descartarlas, ajustar el modelo y repetir la estimación usando la misma distancia. A continuación se describe dicha técnica:

1. Definir el **tamaño** de una variable como la cantidad de observaciones en las que la variable no es 0.

2. Ordenar de manera descendente las variables según su tamaño y considerar, por ahora, solo aquellas con tamaño positivo. Descartar además las variables de tamaño menor a 60[2], entre otros motivos, para ahorrar recursos computacionales.

3. Verificamos si la matriz 3.1 es invertible, de no serlo, procedemos de la siguiente manera.

4. Comenzando con la matriz en la que se consideran solo las dos variables de mayor tamaño. Calculamos su número de condición[3], y si es mayor a 1000, descartamos la variable de menor tamaño considerada, en otro caso conservamos ambas.

5. Agregamos la variable que sigue en el orden del tamaño, si el número de condición de la nueva matriz es mayor de 1000, se descarta, en otro caso, la conservamos.

6. Repetimos el proceso hasta que todas las variables sean probadas.

¿Cuál es nuestro interés en mantener un número de condición acotado? Coenders y Saez (2000) mencionan que si deseamos evitar la multicolinealidad, lo aceptable es contar con números de condición menores a 900. Por otro lado, en Pizer (1975) se presentó un criterio más relajado, pues usar al 1000 para tal diagnóstico permite mantener un alto número de variables mientras se reduce significativamente el número de condición final.

[2] El 60 está relacionado con el tamaño de población de la encuesta sobre la que trabajaron Bankier *et al.*

[3] Definido como el valor absoluto de la razón entre el mayor y el menor de los eigenvalores de la matriz.

7. Si después del proceso anterior, el número de condición de la matriz sigue siendo mayor que 10000, lo cual rara vez ocurre con datos de un censo, descartamos algunas variables adicionales. Para este segundo descarte, ahora seguimos un orden descendiente respecto al incremento causado por cada variable en el paso 5.

8. Calculamos los ponderadores calibrados considerando solo las variables restantes.

9. Repetimos los pasos 4 al 8 sobre el conjunto de variables de tamaño 0. Adicionalmente, en caso de no encontrarse tales calibradores en cierto rango, se descartan más restricciones.

10. Los ponderadores calibrados finales son tomados al promediar los ponderadores que obtuvimos en los pasos 8 y 9.

El descarte de restricciones fue formulado por Bankier *et al.* como una medida a llevarse a cabo en la fase de estimación. Aunque tal técnica pretende mantener todas las variables auxiliares, en principio modifica el modelo, lo cual es inaceptable.
Los ponderadores finales no tienen porque satisfacer las ecuaciones de calibración del conjunto total de variables auxiliares que no fueron descartadas, y esto es otra notable desventaja. Sin embargo, es rescatable que podemos aplicar el descarte de restricciones desde la elección de variables auxiliares.
Otra característica relevante es que trata de manera separada las variables de tamaño 0 del resto, para luego promediar los ponderadores obtenidos en cada grupo.
Las anteriores propuestas de solución buscan resolver las limitantes numéricas o concentrar los ponderadores calibrados en cierto rango para evitar obtener estimaciones erradas. Cabe destacar que antes de la fase de estimación se requiere de una buena elección de variables auxiliares y modelo. En el siguiente capítulo veremos la implementación de estas propuestas de solución para evaluar su efectividad.

Capítulo 4

Implementación y otros usos del estimador por calibración

El objetivo del presente capítulo es ilustrar el uso del estimador por calibración con datos reales. En la sección 4.1 usaremos una base de datos para evaluar la efectividad de las propuestas presentadas en el capítulo 3. En las secciones siguientes describiremos algunos usos del estimador por calibración para la obtención de parámetros poblacionales distintos a totales o medias, además los implementaremos en la misma base de datos.

4.1 Implementacion del metodo de calibracion

Estudiaremos la población artificial creada con los datos del *Concentrado de Hogares* de la ENIGH 2016. Como se explica en INEGI (2017b), el diseño muestral de la ENIGH es estratificado, polietápico y con PPT. Los factores de expansión son los inversos a tales probabilidades[1].
Tomaremos como total a estimar el *Ingreso Corriente* total de nuestra población artificial. Para elegir las variables auxiliares descartemos primero aquellas que involucran ingresos, pues es natural pensar que si en la encuesta no fue contestada la pregunta de *Ingreso corriente*, tampoco lo fueron *Ingresos por sueldos*, *otros ingresos*, entre otras. Tampoco usaremos las variables categóricas, aunque añadirlas no requiere mayor modificación al código.
Al seguir un método de *stepwise selection* para elegir a las variables para un modelo lineal, conservamos 50 variables que contribuyen significativamente al modelo. Por simplicidad, empecemos tomando solo las variables *tot_integ*, *mayores*, *percep_ing*, *estim_alqu* y *pago_tarje* para formar al vector de variables auxiliares, además de un componente 1 que servirá como intercepto. Algunas

[1] Además, fueron reajustados por proyección en cada dominio o según la tasa de no respuesta a nivel estrato.

características de tales variables se describen en la tabla 4.1.

Tabla 4.1: Características de algunas variables de la ENIGH 2016.

Variable	Descripción	Coeficiente de correlación con ing_cor	Porcentaje de respuesta positiva
tot_integ	Número de integrantes del hogar	0.044	100
mayores	Integrantes del hogar de al menos 12 años	0.065	100
percep_ing	Número de personas que perciben ingreso corriente monetario	0.042	99.856
estim_alqu	El valor estimado del alquiler que habrían de pagar por un alojamiento de las mismas características.	0.171	86.371
pago_tarje	Pago por tarjeta de crédito.	0.187	9.978

Al generar estimaciones por calibración incorporando tal información auxiliar, podemos notar que como la variable *pago_tarje* está altamente representada por 0's, lo que ocasiona que en algunas realizaciones se presentan errores como en el Ejemplo 1.
Tras descartar variables mediante el proceso descrito en la sección 3.3 e implementado en el anexo D, proponemos un segundo modelo en el que solo consideramos las variables *tot_integ*, *mayores*, *percep_ing* y el componente 1.
Aunque tales limitantes numéricas desaparecen, la tabla 4.2 nos permite notar que al igual que en el Ejemplo 2, en ocasiones se obtienen ponderadores negativos, sin embargo las estimaciones siguen siendo aceptables, pues con la misma muestra el estimador de Horvitz-Thompson presentó un error del 1.11 %, mientras que el estimador por calibración tuvo un error del 4.17 %.

Tabla 4.2: Efecto de la calibración en ponderadores de los datos de la ENIGH tras el descarte de variables.

	Mínimo	Mediana	Media	Máximo
Iniciales	113.0	670.0	976.2	5386.0
Calibrados	−43.61	785.39	1004.44	4425.28

Para tratar con la aparición de ponderadores negativos, haremos uso del método "*raking*" en la misma muestra. La tabla 4.3 nos muestra que a pesar de que la distancia *raking* brinda ponderadores que no están acotados superiormente, es posible obtener ponderadores dentro de un rango aceptable, lo cual podemos notar también en otras realizaciones. En ésta realización se obtuvo un error relativo incluso menor al del método *linear*, pues fue del 3.55 %.

Tabla 4.3: Efecto de la calibración en ponderadores de los datos de la ENIGH con el método raking.

	Mínimo	Mediana	Media	Máximo
Iniciales	113.0	670.0	976.2	5386.0
Calibrados	90.64	782.96	1004.44	4239.82

En la sección 2.3 mencionamos que para obtener mayor precisión en la estimación de las propiedades estadísticas del estimador son necesarias "varias estimaciones", es natural preguntarse cuántas son requeridas.
En el anexo B se realizaron 100000 estimaciones con el método *linear* usando muestras del 0.001 % de la población, agrupadas en 100 diagramas de cajas. Con la finalidad de determinar si dichas cajas se encuentran en rangos cercanos se graficaron juntas en la imagen 4.3.

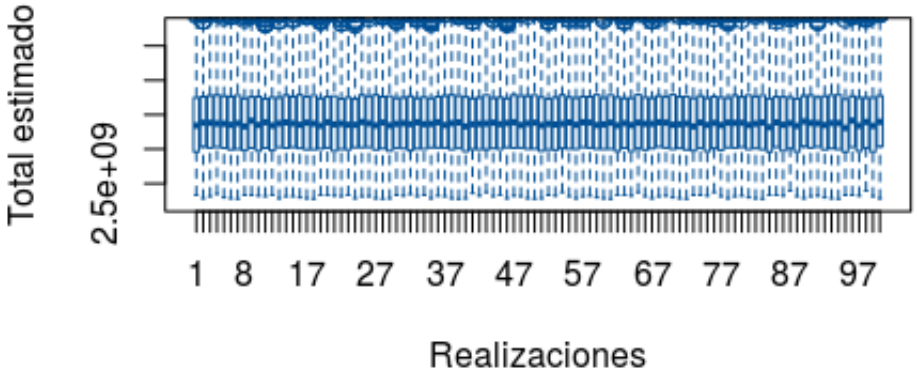

Figura 4.1: Diagrama de cajas para la cantidad de estimaciones.

A pesar de contar con estimaciones que sobreestiman o subestiman el total (cercano a 3 mil millones), la mayoría permanecen en nuestro rango de interés, así que consideraremos 1000 estimaciones como suficientes para nuestras estimaciones.
Usando el código del Anexo C, se realizaron 1000 simulaciones para las que se calcularon los estimadores de Horvitz-Thompson y por calibración según distintas distancias, obteniendo los errores relativos que se muestran en la tabla 4.4.

Tabla 4.4: Comparación de características de las estimaciones usando distintas distancias en el ejemplo.

		Calibración con distancia			
	Horvitz-Thompson	Linear	Raking	Logit	Truncated
Sesgos	−0.0015	0.4086	0.3302	0.1538	0.0424
ECM	0.7563	0.1689	0.1091	0.0237	0.0017
V	0.7563	0.0019	0.0001	0.0001	0.0000

Los estimadores de Horvitz-Thompson son insesgados, ésto justifica que al

contar con errores cuadráticos medios menores en los estimadores por calibración que en los estimadores de Horvitz-Thompson, también tengan una varianza menor. Naturalmente tendremos intervalos de confianza de menor tamaño.
Además de los totales poblacionales, la calibración se puede adaptar para estimar parámetros más complejos que la estimación de medias y totales. A continuación hablaremos sobre el uso de la calibración para la estimación de cuantiles y la estimación en diseños de muestreo de 2 fases.

4.2 Estimación de cuantiles

La mediana y otros cuantiles de la población finita son medidas descriptivas importantes. Para estimar los cuantiles, primero se debe estimar la función de distribución del parámetro de interés, que en la población finita es una suma. Consideremos la función Heaviside, definida como

$$\begin{array}{rccc} \Delta(z): & \mathbb{R} & \to & \mathbb{R} \\ & z & \mapsto & \begin{cases} 0, & \text{si } z < 0 \\ 1, & \text{si } z \geq 0. \end{cases} \end{array}$$

La función de distribución empírica de la variable de estudio y es

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{4.1}$$

El cuantil α está definido como

$$Q_{y,\alpha} = \text{ínf} \left\{ t \,|F_y(t) \geq \alpha \right\}$$

En la práctica, la información completa es necesaria, porque es poco factible que los cuantiles de varias variables sean importados de fuentes externas.
La variable auxiliar x_j, de la que se observaron los valores x_{jk}, tiene función de distribución conocida $F_{x_j}(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - x_{jk})$ y denotamos a su cuantil α por $Q_{x_j,\alpha}$.
Un estimador natural de $F_y(t)$ basado en los ponderadores de diseño $d_k = \frac{1}{\pi_k}$ es

$$\hat{F}_{y_{HT}}(t) = \frac{1}{\sum_{k \in U} d_k S_k} \sum_{k \in U} d_k \Delta(t - y_k) S_k,$$

así que un estimador de $F_y(t)$ por calibración tiene la forma

$$\hat{F}_{y_{cal}}(t) = \frac{1}{\sum_{k \in U} w_k S_k} \sum_{k \in U} w_k \Delta(t - y_k) S_k, \tag{4.2}$$

donde los ponderadores w_k están calibrados adecuadamente a una información auxiliar especificada. A partir de la ecuación 4.2 obtenemos la estimación del cuantil α como

$$\hat{Q}_{y,\alpha_{cal}} = \text{ínf} \left\{ t \,\middle|\, \hat{F}_{y_{cal}}(t) \geq \alpha \right\}$$

y definimos de forma análoga $\hat{Q}_{x_j,\alpha_{cal}}$.

Un caso simple es el abordado por Harms y Duchesne (2006) y Wu (2003), quienes sin especificar modelo alguno, además de contar con los cuantiles conocidos $Q_{x_j\alpha}$ para $j = 1, 2, \ldots, J$, consideran la variable auxiliar 1 entre la información disponible para la calibración al total poblacional, y determinan los w_k que minimizan la distancia χ-cuadrada, sujeto a las ecuaciones de calibración

$$\begin{cases} \sum_{k\in U} w_k S_k &= N \\ \hat{Q}_{x_j,\alpha_{cal}} &= Q_{x_j,\alpha_{cal}} \qquad j = 1, 2, \ldots, J \end{cases}$$

para estimados $\hat{Q}_{x_j,\alpha_{cal}}$ adecuadamente definidos.
Al aplicar lo anterior para estimar los cuantiles $Q_{ing\_cor,0.25}$, $Q_{ing\_cor,0.5}$ y $Q_{ing\_cor,0.75}$ usando el código del anexo G, la figura 4.2 nos permite ver que las funciones de distribución empíricas inducidas por cada estimador parecen ser cercanas.

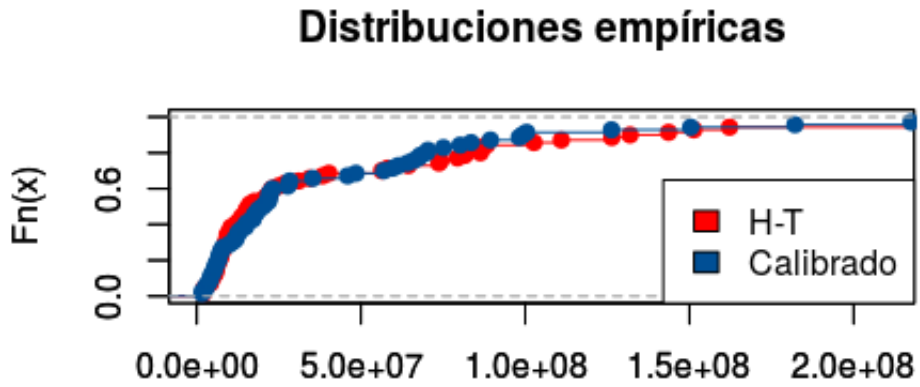

Figura 4.2: Distribuciones empíricas inducidas por los estimadores.

Mientras que la tabla 4.5 nos permite ver a detalle tales estimaciones.

Tabla 4.5: Cuantiles empíricos estimados inducidos por los estimadores por Horvitz-Thompson y por Calibración.

	Horvitz-Thompson	Calibración
0.25	7 849 837	7 591 824
0.5	15 948 311	20 191 357
0.75	74 309 298	66 050 342

4.3 Información compuesta para diseños de muestreo de 2 fases

El muestreo de doble fase tradicional se refiere a diseños que involucran dos muestreos probabilísticos, S_1 y S_2, de la misma población $U = \{1, \ldots, N\}$, donde $S_2 \subset S_1 \subset U$.
La información auxiliar se registra para U y S_1, se registran los valores de la variable de estudio y_k solo para $k \in S_2$ con el objetivo de estimar $t_y = \sum_{k\in U} y_k$. Los ponderadores de diseño son $d_{1k} = \frac{1}{\pi_{1k}}$ y $d_{2k} = \frac{1}{\pi_{2k}}$[2], además denotemos por

[2] $\pi_{2k} = \pi_{k|S_1}$

$S_{ik} = \mathbb{1}_{S_i}(k)$. Los ponderadores combinados de diseño son $d_k = d_{1k}d_{2k}$, con los cuales el estimador básico insesgado queda de la forma

$$\hat{t}_y = \sum_{k \in U} d_k y_k S_{1k} S_{2k}.$$

El cual puede mejorarse mediante el uso de información auxiliar en dos niveles:

1. Nivel poblacional: El valor del vector $\mathbf{x}_{1k}$ es conocido para cada $k \in U$, por lo tanto, se conoce para cada $k \in S_1$ y para cada $k \in S_2$; $\sum_{k \in U} \mathbf{x}_{1k} S_{1k}$ es el vector de totales de la población conocida.

2. Nivel de primera muestra: El valor del vector $\mathbf{x}_{2k}$ es conocido para cada $k \in S_1$, y por lo tanto conocido para cada $k \in S_2$; el total desconocido $\sum_{k \in U} \mathbf{x}_{2k} S_{2k}$ se estima mediante el estimador insesgado $\sum_{k \in U} d_{1k} \mathbf{x}_{2k} S_{1k}$.

Dupont (1995) propone dos alternativas para obtener estimadores de calibración

$$\hat{t}_{y_{cal}} = \sum_{k \in U} w_k y_k S_{2k}.$$

La primera consiste en una calibración de dos pasos: Determinar ponderadores calibrados iniciales w_{1k} que satisfagan la ecuación

$$\sum_{k \in U} w_{1k} \mathbf{x}_{1k} S_{1k} = \sum_{k \in U} \mathbf{x}_{1k}.$$

Posteriormente, partiendo de los ponderadores w_{1k}, obtener ponderadores calibrados finales que satisfagan

$$\sum_{k \in U} w_k \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix} S_{2k} = \sum_{k \in U} w_{1k} \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix} S_{1k} = \begin{pmatrix} \sum_{k \in U} \mathbf{x}_{1k} \\ \sum_{k \in U} w_{1k} \mathbf{x}_{2k} S_{1k} \end{pmatrix}.$$

La segunda propuesta busca determinar directamente los ponderadores w_k tales que

$$\sum_{k \in U} w_k \mathbf{x}_k S_{2k} = \begin{pmatrix} \sum_{k \in U} \mathbf{x}_{1k} \\ \sum_{k \in U} d_{1k} \mathbf{x}_{2k} S_{1k} \end{pmatrix}.$$

No es de sorprender que cada alternativa brinda ponderadores diferentes y su eficiencia depende de manera sutil del patrón de correlación entre y_k, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$. Para verificar la utilidad de la calibración de dos pasos, tomemos de nuestra población artificial una primera muestra S_1 con PPT del 1 % en la que nuestro vector de información auxiliar $\mathbf{x}_{1k}$ está conformado por las variables *tot_integ* y un componente 1. En la siguiente fase de muestreo, tomemos el vector de información auxiliar $\mathbf{x}_{2k}$ como el formado por las variables *mayores* y *percep_ing*. Mediante el código del anexo G podemos notar que en algunas realizaciones, algunos ponderadores calibrados iniciales w_{1k}, lo cual no es aceptado por el software para el segundo paso, sin embargo en la mayoría de los casos, como el que se ilustra en la tabla 4.6, cumple su cometido.

Tabla 4.6: Efecto de la calibración de dos pasos en ponderadores de los datos de la ENIGH.

	Mínimo	Mediana	Media	Máximo
Iniciales	212.0	1544.0	1827.0	4974.0
Primer paso	101.7	793.3	901.2	2452.6
Segundo paso	62.03	799.00	901.20	2347.78

La calibración de dos pasos brindó estimadores muy cercanos al total real, pues el error relativo pasó del 109.25 % en el estimador de Horvitz-Thompson a tener un error relativo del 2.34 % en el primer paso y un 1.70 % tras el segundo paso, aunque ésta reducción del error no es persistente.

4.4 Breve mención de otras aplicaciones

En Plikusas (2006) se aborda el uso de la calibración para la estimación de la razón entre los totales de dos variables y_1 y y_2,

$$R = \frac{\sum_{k\in U} y_{1k}}{\sum_{k\in U} y_{2k}},$$

mediante el estimador

$$\hat{R}_{cal} = \frac{\sum_{k\in U} w_k y_{1k} S_k}{\sum_{k\in U} w_k y_{2k} S_k},$$

donde al conocer la razón R_0 entre los totales de las variables auxiliares x_1 y x_2, auxiliares de y_1 y y_2 respectivamente, los ponderadores w_k son aquellos que satisfacen la ecuación de calibración

$$\sum_{k\in U} w_k \left(x_{1k} - R_0 x_{2k}\right) = 0.$$

Además de éstas aplicaciones, en la literatura reciente se ha usado la calibración para la estimación de parámetros bilineales, ajuste de ponderadores para no respuesta, calibración con respecto a momentos empíricos de segundo orden y algunas otras que son exploradas en S. y Ted (2010) y Plikusas (2006).

Conclusiones y comentarios finales

El capítulo 1 introduce el estimador por calibración y los aspectos que nos lo brindan. A partir del estimador de Horvitz-Thompson, se ajustan simultaneamente los ponderadores sujetos a restricciones impuestas por las variables auxiliares. El uso de ponderadores que incorporan información de toda la muestra está justificado por la relación entre la variable de estudio y las variables auxiliares.
En el segundo capítulo, hemos explorado las características que nos hacen elegirlo por sobre el estimador de Horvitz-Thompson. Como tal estimador es insesgado, reducir el ECM de las estimaciones ocasiona la reducción de la varianza y el tamaño de los intervalos de confianza.
Poner la estimación por calibración en lugar de la de Horvitz-Thompson no es garantía para contar con una estimación sin errores. Es así como en el tercer capítulo, hemos estudiado algunos problemas que suelen presentarse durante la estimación por calibración, además de algunas propuestas de solución a los mismos junto a sus ventajas y desventajas.
Para lograr una buena estimación por calibración, se requiere iniciar con una buena elección de las variables auxiliares. Un método de elección de las mismas se expuso en la sección 3.3.
Tener una partición en la población inducida por una variable que toma valores repartidos entre varios grupos alejados, puede ocasionar problemas en la fase de estimación. Lo anterior se debe usualmente a haber obtenido una muestra que sobrerrepresenta a alguna de las subpoblaciones, lo que puede traer consigo ponderadores problemáticos y que brindan estimadores sesgados a ciertos valores.
Consideramos como problemáticos a los ponderadores menores que 1 y a los ponderadores negativos. Se propuso la elección de alguna otra distancia que evitara tal problema. En ocasiones las ecuaciones de calibración con la nueva distancia no tienen solución, preferiremos entonces quedarnos con el estimador de Horvitz-Thompson.
Aunque en ocasiones podemos considerar aceptables los ponderadores menores que 1, aludiendo a que se asocian a elementos de la población demasiado raros, no es el caso de los ponderadores negativos.
También adjudicado a la partición de la población por alguna variable, aunque ahora para compensar la aparición de ponderadores menores a 1 (o negativos) que ajustan el total de una primera variable, suelen aparecer ponderadores demasiado grandes y alejados del resto, como una medida para satisfacer la ecuación de calibración de otra de las variables.
Otra causa de tal problema es la infrarrepresentación en la muestra de elementos

de una subpoblación con respuestas demasiado grandes a alguna variable.
En el capítulo final se implementaron las propuestas de solución en datos reales. Se obtuvieron resultados satisfactorios que verificaron lo mencionado en capítulos anteriores. Además, se expusieron algunas aplicaciones del método de calibración. Más allá de la estimación de totales, al tratarse de sumas ponderadas, es posible estimar cuantiles, generar estimaciones para muestras de 2 fases, imputar respuestas, etc.
A pesar de que el estimador por calibración no es una propuesta tan reciente, tampoco goza de la popularidad del estimador de Horvitz-Thompson, aunque está presente en el estimador por regresión. Un buen estimador por calibración requiere de un buen diseño de muestreo y una correcta elección de variables, las cuales también requieren de un análisis exploratorio previo. Para usos requeridos en problemas de actualidad en los que más que un insesgamiento, requerimos la reducción del error cuadrático medio en términos de precisión y se sugiere considerar el uso del método de calibración.

Anexos

APÉNDICE A

Algunos resultados mencionados

Teorema A.1. *El estimador de Horvitz-Thompson es insesgado*

Demostración. Con la misma notación, veamos que

$$\begin{aligned} Sesgo(\hat{t}_{y_{HT}}, t_y) &= \mathbb{E}(\hat{t}_{y_{HT}}) - t_y \\ &= \mathbb{E}\left(\sum_{k\in U} d_k y_k S_k\right) - t_y \\ &= \sum_{k\in U} d_k y_k \mathbb{E}\left(S_k\right) - t_y \end{aligned}$$

Para determinar $\mathbb{E}\left(S_k\right)$, de nuevo consideremos la colección de todas las posibles muestras $\mathcal{S}$ para ver que

$$\mathbb{E}\left(S_k\right) = \sum_{S\in\mathcal{S}} S_k \mathbb{P}(k \in S)$$

Como S_k es una indicadora, la cantidad anterior es la suma de las probabilidades de pertenencia del elemento k a cada muestra S, es decir,

$$\begin{aligned} &= \pi_k \\ &= \frac{1}{d_k} \end{aligned}$$

luego,

$$\begin{aligned} Sesgo(\hat{t}_{y_{HT}}, t_y) &= \sum_{k\in U} d_k y_k \frac{1}{d_k} - t_y \\ &= \sum_{k\in U} y_k - t_y \\ &= t_y - t_y \\ &= 0 \end{aligned}$$

■

Teorema A.2. *Consideremos el modelo* $y = \hat{B}^T\mathbf{x}$. *Sea* $\hat{y} = \hat{B}^T\mathbf{x} + e$ *el estimador de* y *según el modelo.*
La varianza del error e *del modelo está dada por*

$$Var(e) = \frac{1 - R^2}{R^2} Var(\hat{y}),$$

donde R^2 *es el coeficiente de determinación entre* y *y su estimador* $\hat{y}$.

Demostración. El coeficiente de determinación está dado por

$$\begin{aligned} R^2 &= \frac{Cov(y, \hat{y})^2}{Var(y)Var(\hat{y})} \\ &= \frac{\left[Cov(\hat{B}^T\mathbf{x}, \hat{B}^T\mathbf{x} + e)\right]^2}{Var(y)Var(\hat{y})} \\ &= \frac{\left[Cov(\hat{B}^T\mathbf{x}, \hat{B}^T\mathbf{x}) + Cov(\hat{B}^T\mathbf{x}, e)\right]^2}{Var(y)Var(\hat{y})} \\ &= \frac{\left[Var(y) + Cov(y, e)\right]^2}{Var(y)Var(\hat{y})}. \end{aligned}$$

Supongamos que e y y son independientes, así que

$$\begin{aligned} &= \frac{Var(y)^2}{Var(y)Var(\hat{y})} \\ &= \frac{Var(y)}{Var(\hat{y})} \\ &= \frac{Var(y)}{Var(y + e)} \\ &= \frac{Var(y)}{Var(y) + Var(e)} \\ R^2\left[Var(y) + Var(e)\right] &= Var(y) \\ R^2 Var(e) &= (1 - R^2)Var(y) \\ Var(e) &= \frac{1 - R^2}{R^2} Var(y) \end{aligned}$$

Aunque supusimos la independencia de e y y, generan un espacio de probabilidad donde son dependientes, así que el resultado es válido. ■

APÉNDICE B

Determinando la cantidad necesaria de simulaciones

```
############################Librerías
   ##############################
        #Librerías necesarias para trabajar
  library(icarus)        #Para la calibración
  library(parallel)      #Para paralelizar las
     operaciones
  library(sampling)      #Para calcular las PPT
  library(pps)           #Para tomar muestras con PPT
############################Funciones
   ##############################
        #Declaración de funciones necesarias para
           trabajar
#Carga la base de datos y declara variables globales
   y para el cluster
BaseDatos<-function()
{
  #Leemos toda la base de datos
  MyData <- read.csv(file="/.../concentradohogar.csv"
     , header=TRUE, sep=",")
  #Ya que se cargaron los 70311 observaciones de 127
     variables, descartamos las variables categóricas
  Data=MyData[,c(14:127)]
  Data[,colnames(MyData[c(0),][9])] <- MyData[,c(9)]
  #Quitamos las variables relacionadas con ingresos
  Data=Data[,-c(12:42,44,97:100)]
  #Columna de 1's para el intercepto
  Data$intercepto=rep(1,nrow(Data))
```

```
  #Reordenamos las variables por simplicidad
  Data=cbind(Data[,-c(11,79)],Data[,c(11,79)])
  #Data es la matriz de datos para realizar
     inferencia y Factores los factores de expansión
  return(Data)
}
#Encuentra el apuntador al primer elemento de la
   segunda subpoblación
separacion<-function(Data)
{
  V=mean(Data$y[1:200])+sqrt(var(Data$y[1:200]))
  Division=min(which(Data$y>2*V))
  return(Division)
}
#En la siguiente función Data es nuestra matriz de
   datos, ycolname el nombre de la variable de interés
    vectaux las columnas que forman los vectores de
   información auxiliar, samplesize el tamaño de
   muestra y mtd el método de calibración
estim_cal<-function(Data, ycolname, vectaux,
   samplesize, mtd, control)
{
  #Diseño de muestreo: muestreo con PPT
  if(is.na(control))
    s<-ppss(Data$factor, nrow(Data)*samplesize)
  else
  {
    div<-separacion(Data)
    if(control==0)#Para permitir muestras con
       sobrerrepresentación extrema
      s2<-c()
    else
      s2<-ppss(Data$factor[div:N],control)

    if(control==nrow(Data)*samplesize)
      s1<-c()
    else
      s1<-ppss(Data$factor[1:(div-1)],nrow(Data)*
         samplesize-control)

    s=c(s1, s2)
  }
  S=Data[s,]
  y= unlist(S[ycolname])        #y es nuestro pará
     metro de interés
  #Determinamos ahora los ponderadores por calibració
     n
```

```
  dimaux=ncol(Data) #Tamaño original del vector de
     información auxiliar
  margins <- matrix(NA, nrow=dimaux, ncol=3)#vector
     completo de información auxiliar
  for(n in 1:dimaux)#inserta todos los vectores
     auxiliares excepto el de los factores de expansió
     n
    margins[n,]<-c(colnames(S[n]), 0, Totales[n])
  margins<- margins[vectaux,] #considera solo las
     columnas de la información auxiliar
  #que están en el arreglo vectaux
  #Realiza la calibración de ponderadores con la
     muestra S, la información auxiliar
  #de margins los ponderadores iniciales de factor,
     por el método ingresado mtd
  w <- calibration(data=S, marginMatrix=margins,
     colWeights="factor",method=mtd
                   ,description=FALSE)

  boxplot(S$factor, w, border = rgb
     (0,0.3094340,0.5849057), ylab=c("Ponderadores"),
     names=c("Iniciales","Calibrados"))
  print(summary(S$factor))
  print(summary(w))

  Y=Totales[which(names(Data)==ycolname)]
  Y_HT=weightedTotal(y,S$factor)
  Y_cal=weightedTotal(y,w)
  Tot<-matrix(c(Y,Y_HT,Y_cal), nrow = 1)
  colnames(Tot)<-c("Real","H-T","Cal")

  return(Y_cal)
}


#En la siguiente linea llamamos a la función de
   calibración e imprimimos los resultados de los
   totales, el método de calibración mtd puede ser
   linear, logit, truncated o raking
estim_ingcor<-function(x){
  return(try(estim_cal(Data, "ing_cor", InfAux,
     0.00215, "linear", NA),TRUE))
}
#La siguiente función proporciona nestimaciones*
   ncajas agrupadas en ncajas conjuntos
estimators<-function(ncajas, nestimaciones, Data,
   ycolname, InfAux){
```

```
  for(j in 1:ncajas){
      Aux<-parSapply(clust, 1:nestimaciones, function
         (x) estim_ingcor(x))
    if(j==1){
      Cajas<-matrix(data=Aux, nrow = ncajas)
    }else{
      Cajas<-cbind(Cajas, Aux)}
    rm(Aux)
    print(c("Cargando␣", j*100/ncajas, "%"))
  }
  Cajas<-as.numeric(Cajas)
  Cajas<-matrix(data=Cajas, ncol = ncajas, nrow =
     nestimaciones)
  #Hasta aquí puede tener NA's debido a errores de
     convergencia
  VectorNA=which(is.na(Cajas))
  while (length(VectorNA)>0){
    Cajas[which(is.na(Cajas))]=parSapply(clust, 1:
       length(VectorNA), function(x) estim_ingcor(x))#
       Llena los na's
    VectorNA=which(is.na(Cajas))
  }
  #Se asegura de darle estructura de matriz
  Cajas<-as.numeric(Cajas)
  Cajas<-matrix(data=Cajas, ncol = ncajas, nrow =
     nestimaciones)
  return(Cajas)
}
#Declaración de las funciones para paralelizar
paraleliza<-function()
{
  no_cores <- detectCores()
  clust <- makeCluster(no_cores)
  clusterExport(clust, "estim_cal")
  clusterExport(clust, "multiestim_cal")
  clusterExport(clust, "Data")
  clusterExport(clust, "weightedTotal")
  clusterExport(clust, "Totales")
  clusterExport(clust, "InfAux")
  clusterExport(clust, "calibration")
  clusterExport(clust, "estim_ingcor")
  clusterExport(clust, "multiestim_ingcor")
  return(clust)
}
############################Programa
   ##############################
  Data=BaseDatos()
```

```
#Consideramos los totales ponderados para obtener
   los totales poblacionales, pues tales factores se
    obtuvieron para el muestreo con probabilidad
   proporcional al tamaño en la población entera.
Totales=colSums(Data$factor*Data)
#Cajas y bigotes de 100 000 simulaciones
clust=paraleliza()
Cajas=estimators(100,1000, Data, "ing_cor", InfAux,
    "linear")
stopCluster(clust)
Cajas
boxplot(Cajas, border=rgb(0,0.3094340,0.5849057),
   ylim=c(2.2e+09,4.8e+09), ylab="Total␣estimado",
   xlab="Realizaciones")
abline(h=Totales["ing_cor"], col="red")
```

APÉNDICE C

Estimando las propiedades estadísticas del estimador bajo distintas distancias

```
############################Librerías
   ##############################
        #Librerías necesarias para trabajar
  library(icarus)       #Para la calibración
  library(parallel)     #Para paralelizar las
     operaciones
############################Funciones
   ##############################
multiestim_cal<-function(Data, ycolname, vectaux,
   samplesize, CotaU){
  #Muestra simple del samplesize\%
  s<-ppss(Data$factor,nrow(Data)*samplesize)
  #De la matriz data, consideramos solo los datos de
     la muestra elegida
  S=Data[s,]
  #d son los factores de expansión d_k
  d=S$factor
  #y es nuestro parámetro de interés
  y= unlist(S[ycolname])
  #Estimador de HT del total de y
  Y_HT=weightedTotal(y  ,d)
  #Determinamos las cotas para logit y truncated
  L=min(Data[,"factor"])/exp(mean(log(Data[,"factor"
     ])))
  U=max(Data[,"factor"])/exp(mean(log(Data[,"factor"
     ])))
  #Determinamos ahora los ponderadores calibrados
```

```
  #Primero definimos la matriz de ecuaciones de
     calibración para todos los datos
  #Tamaño original del vector de información auxiliar
  dimaux=ncol(Data)-1
  #Matriz de tamaño px3, p-tamaño del vector auxiliar
     , 3: nombre de la variable auxiliar[,1], cantidad
      de categorías, que fijamos en 1[,2], total por
     cada categoría[,3]
  margins <- matrix(NA, nrow=dimaux, ncol=3)
  for(n in 1:dimaux)
    #inserta todos los vectores auxiliares excepto el
       de los factores de expansión
    margins[n,]<-c(colnames(S[n]), 0, Totales[n])
  #Considera solo las columnas de información
     auxiliar que están en el arreglo vectaux
  margins<- margins[vectaux,]
  #Realiza la calibración de ponderadores con los de
     la muestra S, la información auxiliar de margins,
      los ponderadores iniciales de ''factor'', por
     los distintos métodos
  wlin <- try(calibration(data=S, marginMatrix=
     margins, colWeights="factor",method="linear" ,
     description=FALSE),TRUE)
  wran <- try(calibration(data=S, marginMatrix=
     margins, colWeights="factor",method="raking" ,
     description=FALSE), TRUE)
  wlog <- try(calibration(data=S, marginMatrix=
     margins, colWeights="factor",method="logit" ,
     bounds=c(L,U*CotaU),description=FALSE),TRUE)#0 a
     55500
  wtru <- try(calibration(data=S, marginMatrix=
     margins, colWeights="factor",method="truncated" ,
     bounds=c(L,U*CotaU),description=FALSE),TRUE)
  Y_cal_lin=weightedTotal(y,wlin)
  Y_cal_ran=weightedTotal(y,wran)
  Y_cal_log=weightedTotal(y,wlog)
  Y_cal_tru=weightedTotal(y,wtru)
  return(c(Y_HT, Y_cal_lin, Y_cal_ran, Y_cal_log, Y_
     cal_tru))
}
multiestim_ingcor<-function(x){
 return(try(multiestim_cal(Data, "ing_cor", InfAux,
    0.001, 2),TRUE))
}
comparativa<-function(nestimaciones){
  Simulacion=parSapply(clust, 1:nestimaciones,
     function(x) multiestim_ingcor(x))#aqui es un
```

```
    arreglo multidimensional
  VectorNA=unique((which(is.na(Simulacion), arr.ind=
    TRUE))[,2])#debe determinar solo el renglon
  while (length(VectorNA)>0){
    Aux=parSapply(clust, 1:length(VectorNA), function
      (x) multiestim_ingcor(x))
    for(i in 1:length(VectorNA))
      Simulacion[,VectorNA[i]]=Aux[,i]
    rm(Aux)
    VectorNA=unique((which(is.na(Simulacion), arr.ind
      =TRUE))[,2])
  }
  rownames(Simulacion)<-c("Y_HT", "Y_cal_lin", "Y_cal
    _ran", "Y_cal_log", "Y_cal_tru")
  Y=Totales["ing_cor"]
  Sesgos<-mean(Simulacion[,1]-Y)/Y
  for(i in 2:5)
    Sesgos<-c(Sesgos, mean(Simulacion[,i]-Y)/Y)
  ECM<-mean((Simulacion[1,]-Y)^2)/Y^2
  for(i in 2:5)
    ECM<-c(ECM, mean((Simulacion[,i]-Y)^2)/Y^2)
  V<-ECM-Sesgos^2
  Descripcion=rbind(Sesgos, ECM, V)
  colnames(Descripcion)<-c("HT", "lin", "ran", "log",
    "tru")
  print(Descripcion)
  return(Simulacion)
}
##########################Programa
  ############################
  clust=paraleliza()
  Com=comparativa(1000)
  stopCluster(clust)
```

APÉNDICE D

Elección de las variables auxiliares

```
#Función para la elección de variables que no causan
   multicolinealidad
#Para la variable ing_cor, descartamos las variables
   que involucran ingresos, al llamar la función con
   rest_adicionales=c(24:55,57,110:113, 115)-13
#Para descartar además las variables que presentan
   singularidades al aplicar lm y aquellas que son
   sumas o sumandos de otras variables consideradas, a
   ñadimos
#rest_adicionales=c(rest_adicionales
   ,2,3,5:8,61,63,66,71,75,79,86,87,91,95,96,113,114)
constrains<-function(Data, rest_adicionales)
{
  cant_variables<-ncol(Data)-1
  #Obtiene el orden de las variables según su tamaño
  constrain<-matrix(0,1,cant_variables)
  for(i in 1:cant_variables)
    constrain[i]<-sum(Data[,i]>0)
  colnames(constrain)<-colnames(Data[1:cant_variables
     ])
  orden=order(constrain[1,], decreasing = TRUE)
# orden[c(1,3)]=orden[c(3,1)] #con ésta línea
   elegimos cuál de las variables de mayor tamaño será
    la inicial para el proceso
  ini_const<-length(which(constrain==max(constrain)))
     #cantidad de variables con el tamaño mayor
  variables=orden[1]
```

```
  for(i in 2:cant_variables)#Se agregan variables sin
      exceder las condiciones descritas
  {
    if(!orden[i]%in%rest_adicionales)#Para aquellas
       variables que no involucran ingresos,
    {
      Mat<-matrix(0,ini_const,ini_const)#Inicializa
         con la de 2x2
      for(j in 1:nrow(Data))#Determina la matriz a
         invertir incorporando información de toda la
         base de datos
      {
        aux<-as.matrix(Data[j,c(variables, orden[i])
           ])
        Mat<-Mat+Data$factor[j]*t(aux)%*%aux
      }
      eig<-eigen(Mat)$values
      condition<-abs(max(eig)/min(eig))#Calcula el nú
         mero de condición de la matriz obtenida
      if(condition<1000)#Si no se excede el límite,
         se agrega la nueva variable
      {
        variables=c(variables, orden[i])
        ini_const=ini_const+1
      }
      rm(Mat)
    }
  }
  return(variables)
}
```

APÉNDICE E

Simulaciones para ilustrar algunos errores al calibrar

```
#Librerías necesarias para el diseño de los
   ponderadores iniciales
library(sampling)
library(ineq)
tamx<-round(sqrt(70311)+1)
N=tamx*tamx
n=round(0.03*N)

#Ejemplos 1 y 2, para ponderadores menores a 1
x1=seq(9990, 10000, length.out = tamx)
x2=c(seq(0,10,length.out = tamx/2), seq(9990,10000,
   length.out = tamx/2))
#Genera una tabla con todas las combinaciones
   posibles de las variables x1 y x2
Data<-expand.grid(x1,x2)

#Coeficientes para generar la variable y,
   correlacionada con x1 y x2
beta0=100
beta1=2
beta2=2
R2=0.95
ve=((1-R2)/R2)*((beta1^2)*var(x1)+(beta2^2)*var(x2))
DE<-sqrt(ve)
y=beta0 + beta1*Data[,1] + beta2*Data[,2]+ rnorm(N,
   mean=0, sd=DE)
```

```
#Genera los ponderadores
pi=inclusionprobabilities(y, n)
boxplot(1/pi, border = rgb(0,0.3094340,0.5849057),
   ylab="Ponderadores")

#Agrega a la tabla Data una columna de 1's para el
   intercepto y las variables creadas
Data<-cbind(rep(1,N), Data, y, 1/pi)
colnames(Data)<-c("1","x1","x2","y", "factor")
#Finalmente le da estructura Dataframe, para volverla
    compatible con el código del anexo D
Data=as.data.frame(Data)

#########################################################
################ Para calibrar ##########################
#########################################################
Totales=colSums(Data)

#Con una muestra aleatoria, siguiendo el código del
   anexo D
source("/.../Paralelizado.R")
cal=estim_cal(Data, "y", c(1,2,3), n/N, "linear",
   control=NA)#control=NA para el ejemplo 1
```

APÉNDICE F

Otras simulaciones para ilustrar algunos errores al calibrar

```
#Librerías necesarias para el diseño de los
   ponderadores iniciales
library(sampling)
library(ineq)

tamx<-round(sqrt(70311)+1)
N=tamx*tamx
n=round(0.03*N)

#Ejemplos 3 y 4, ponderadores muy grandes y negativos
x1=seq(0, 10, length.out = tamx)
x2=c(seq(0,10,length.out = (tamx*0.99-1)), seq
   (9990,10000,length.out = tamx*0.01))
#Genera una tabla con todas las combinaciones
   posibles de las variables x1 y x2
Data<-expand.grid(x1,x2)

#Coeficientes para generar la variable y,
   correlacionada con x1 y x2
beta0=3000
beta1=2
beta2=2
R2=0.95
ve=((1-R2)/R2)*((beta1^2)*var(x1)+(beta2^2)*var(x2))
DE<-sqrt(ve)
y=beta0 + beta1*Data[,1] + beta2*Data[,2]+ rnorm(N,
   mean=0, sd=DE)
```

```
#Genera los ponderadores
pi=inclusionprobabilities(1:N, n)
boxplot(1/pi, border = rgb(0,0.3094340,0.5849057),
   ylab="Ponderadores")

#Agrega a la tabla Data una columna de 1's para el
   intercepto y las variables creadas
Data<-cbind(1,Data, y, 1/pi)
colnames(Data)<-c("1","x1","x2","y", "factor")
Data=as.data.frame(Data)#Finalmente le da estructura
   Dataframe, para volverla compatible con el código
   del anexo D

#########################################################
################ Para calibrar ##########################
#########################################################
Totales=colSums(Data)
#con una muestra aleatoria
source("../Paralelizado.R")
cal=estim_cal(Data, "y", c(1,2,3), n/N, "linear",
   200)#(2123-200))
```

APÉNDICE G

Otros usos del estimador por calibración

```
        #Librerías necesarias para trabajar
  library(icarus)        #Para la calibración
  library(parallel)      #Para paralelizar las
     operaciones
  library(sampling)      #Para calcular las PPT
  library(pps)           #Para tomar muestras con PPT

  Cuantile_comparsion<-function(Data, ycolname,
     vectaux, samplesize, mtd, alpha)
{
  s<-ppss(Data$factor, nrow(Data)*samplesize)
  S=Data[s,]
  y= unlist(S[ycolname])         #y es nuestro pará
     metro de interés
  #Determinamos ahora los ponderadores por calibració
     n
  dimaux=ncol(Data) #Tamaño original del vector de
     información auxiliar
  margins <- matrix(NA, nrow=dimaux, ncol=3)#vector
     completo de información auxiliar
  for(n in 1:dimaux)#inserta todos los vectores
     auxiliares excepto el de los factores de exp.
    margins[n,]<-c(colnames(S[n]), 0, Totales[n])
  margins<- margins[vectaux,] #considera solo las
     columnas de la información auxiliar
  #que están en el arreglo vectaux
  #Realiza la calibración de ponderadores con la
     muestra S, la información auxiliar
```

```
  #de margins los ponderadores iniciales de factor,
     por el método ingresado mtd
  w <- calibration(data=S, marginMatrix=margins,
     colWeights="factor",method=mtd
                      ,description=FALSE)
  P=ecdf(y*S$factor)#Función de distribución empírica
      inducida por el estimador de H-T
  plot(P, xlim=c(0,2.1e+08), ylim = c(0,1), col="red"
     , main = "")
  Q=ecdf(y*w)#Función de distribución empírica
     inducida por el estimador por calibración
  par(new=TRUE)
  plot(Q, xlim=c(0,2.1e+08), ylim = c(0,1), col=rgb
     (0,0.3094340,0.5849057) ,main = "Distribuciones␣
     empíricas")#
  legend("bottomright",c("H-T","Calibrado"),fill=c("
     red",rgb(0,0.3094340,0.5849057)))
  return(c(  quantile(P,alpha, type = 1),quantile(Q,
     alpha, type = 1)))
}

#2 FASES
fases<-function(Data, ycolname, vectaux1, vectaux2,
   size1, size2, mtd)
{
  if(size2>size1)
  {
    warning("S2␣debe␣estar␣contenida␣en␣S1,␣así␣que␣
       debe␣ser␣menor")
    return(0)
  }
  vectaux=c(vectaux1,vectaux2)#Mezcla la información
     auxiliar completa
  s1<-ppss(Data$factor, nrow(Data)*size1)
  S1=Data[s1,]#La primera fase de muestreo

  dimaux=ncol(Data) #Tamaño original del vector de
     información auxiliar
  margins <- matrix(NA, nrow=dimaux, ncol=3)#vector
     completo de información auxiliar
  for(n in 1:dimaux)#inserta todos los vectores
     auxiliares excepto el de los factores de exp.
    margins[n,]<-c(colnames(Data[n]), 0, Totales[n])
  margins<- margins[vectaux,] #considera solo las
     columnas de la información auxiliar
  #que están en el arreglo vectaux
```

```
  #Realiza la calibración de ponderadores con la
     muestra S, la información auxiliar
  #de margins los ponderadores iniciales de factor,
     por el método ingresado mtd
  w1<- calibration(data=S1, marginMatrix=margins[-
     vectaux2,], colWeights="factor",method=mtd
                    ,description=FALSE)
 #Los ponderadores w1 fueron calibrados solo respecto
     a vectaux1, la información obtenida en la primera
     fase de muestreo.

  s2<-ppss(S1$factor, nrow(Data)*size2)#Toma
     submuestra ppt de S1
  S2=Data[s1[s2],]#La segunda fase de muestreo
  S2$w1=w1[s2]
  y=unlist(S2[ycolname])

  #Agrega las marginales de las variables x2
  for (i in length(vectaux1):length(vectaux2))
    margins[i,3]=weightedTotal(S2[,vectaux[i]],S2$w1)

  w<- calibration(data=S2, marginMatrix=margins,
     colWeights="w1",method=mtd
                   ,description=FALSE)
  #En w se incorpora la información de la segunda
     fase
  print(summary(S2$factor))
  print(summary(S2$w1))
  print(summary(w))
  Y_HT=weightedTotal(y,S2$factor)
  Y_cal=weightedTotal(y,S2$w1)
  Y_cal2=weightedTotal(y,w)
  print(c(Y_HT,Y_cal, Y_cal2))
  return(Y_cal2)
}

######Programa
Data=BaseDatos()
Totales=colSums(Data)

InfAux<-c(1,4,9,78)
Cuantile_comparsion(Data, "ing_cor", InfAux, 0.001, "
   linear", c(0.025,0.5,0.975))

fases(Data, "ing_cor", InfAux[-c(2,3)], InfAux[c(2,3)
   ], 0.002, 0.001, "linear")
```

Bibliografía

Angrist, J. D. y Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion Princeton University Press. *Statistical Papers*, **52**(2):503–504

Bankier, M. D., Rathwell, S., y Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. págs. 764–769

Coenders, G. y Saez, M. (2000). Collinearity, heteroscedasticity and outlier diagnostics in regression. Do they always offer what they claim? *Metodoloski Zvezki*, **16**:79–94

Deville, J.-C., Särndal, C.-E., y Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88**:1013–1020

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, **21**:125–135

Estevao, V. M. y Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, **74**(2):127–147

Haider, G. S. S. J. y Wooldridge, J. (2013). What are we weighting for? *NBER Working paper series*, (C1):1–29

Harms, T. y Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, **32**(1):37–52

Huang, E. T. y Fuller, W. A. (1978). *Nonnegative regression estimation for sample survey data*

INEGI (2017a). Encuesta Nacional de Ingresos y Gastos de los Hogares 2016: Descripción de la base de datos. *Nueva Serie*, **1**:154–170

INEGI (2017b). Encuesta Nacional de Ingresos y Gastos de los Hogares 2016: Diseño muestral. *Nueva Serie*, **1**:154–170

of the Census, U. B. (1968). Current Population Reports: Consumer Income. *U.S. Government Printing Office*, **1**:55–60

Pizer, S. M. (1975). *Numerical Computing and Mathematical Analysis*

Plikusas, A. (2006). Nonlinear calibration. Proceedings, Workshop on Survey Sampling. *Central Statistical Bureau of Latvia*, **1**(1)

S., K. P. y Ted, C. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, **105**(491):1265–1275

Stephan, F. F. y Deming, W. E. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, **11**(4):427–444

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**(2):99–119

Särndal, C.-E. y Deville, J.-C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**:376–382

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, **94**:635–644

Vapnik, V. N., Chervonenkis, A. Y., y Seckler, B. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**:264–280

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, **90**(4):937–951