ugr | Universidad de Granada

Uso de Información Auxiliar en Encuestas con Diseños Muestrales Complejos

Uso de Información Auxiliar en Encuestas con Diseños Muestrales Complejos

TESIS DOCTORAL

JOSÉ MIGUEL CONTRERAS GARCÍA

Dirigida por ANTONIO ARCOS CEBRIÁN

UNIVERSIDAD DE GRANADA

USO DE INFORMACIÓN AUXILIAR EN ENCUESTAS CON DISEÑOS MUESTRALES COMPLEJOS

Tesis doctoral presentada por José Miguel Contreras García
dentro del Programa de Doctorado en Matemáticas y Estadística.
Dirigida por el Dr. Antonio Arcos Cebrián

Editor: Editorial de la Universidad de Granada
Autor: José Miguel Contreras García
D.L.: GR 1863-9083
ISBN: 978-84-9083-047-5

UNIVERSIDAD DE GRANADA

USO DE INFORMACIÓN AUXILIAR EN ENCUESTAS CON DISEÑOS MUESTRALES COMPLEJOS

Tesis doctoral presentada por José Miguel Contreras García
dentro del Programa de Doctorado en Matemáticas y Estadística.
Dirigida por el Dr. Antonio Arcos Cebrián

El doctorando El director

Granada, 13 de diciembre de 2013

Uso de información auxiliar en encuestas con diseños muestrales complejos

Autor: Dr. José Miguel Contreras García

Director: Dr. Antonio Arcos Cebrián

La siguiente página web contiene información actualizada acerca de esta tesis y temas relacionados:

`http://www.ugr.es/~jmcontreras`

Impresa en Granada

Edición, diciembre de 2013

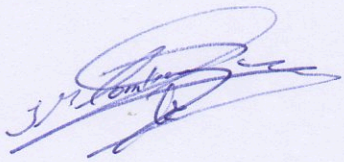
El doctorando José Miguel Contreras García y el director de la tesis Antonio Arcos Cebrián garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, Diciembre de 2013

Director/es de la Tesis

Fdo.: Antonio Arcos Cebrián

Doctorando



Fdo.: José Miguel Contreras García

Para Elena, la causa de toda mi suerte

Agradecimientos

En primer lugar quiero agradecer a mi director de tesis, el Dr. Antonio Arcos, por la oportunidad que me brindó en su día de trabajar con él, gracias a ello ahora tengo lo que tengo. Su sabiduría, su trabajo y sobre todo su paciencia, ha marcado este trabajo de investigación y al que lo firma.

Quiero agradecer al Departamento de Estadística e I.O. y especialmente a María del Mar Rueda, Andrés González, Ramón Gutiérrez Sánchez y Ramón Gutiérrez Jáimez, además de a Carmen Batanero, su apoyo y ayuda en estos años.

Agradecer a mis padres por permitirme estudiar en los malos momentos y por todo lo que me han inculcado. A mis tres hermanos que, aunque no estemos siempre juntos, sé que puedo contar con ellos. Y sobre todo, a Javi decirle una cosa, la siguiente es la tuya, ¡ánimo!

A mi otra familia, mis suegros y cuñados, gracias Manolo, Angustias, Francis y Celia. Y a mí sobry Fran, sigue tan bicho como siempre.

Y sobre todo agradecer a Elena todo lo que me ha proporcionado en estos 8 años, lo que he conseguido y conseguiré de aquí en adelante te lo debo a ti.

José Miguel Contreras García

Resumen

Gracias al gran número de investigaciones realizadas en los últimos años en muestreo en poblaciones finitas, podemos encontrar diferentes métodos de estimación que hacen uso de la información auxiliar para ganar precisión y eficacia en dichas estimaciones. Las encuestas no sólo recogen información relacionada con la variables objeto de estudio, sino que proporcionan gran información auxiliar (valores, totales, promedios, etc.) procedente de censos, proyecciones censales u otros estudios, que puede ser incluida en el proceso de estimación. Una de las metodologías que más auge ha experimentado en los últimos años, debido en gran medida al papel que las agencias de estadística, organismos oficiales y no oficiales le otorgan, ha sido la calibración (Deville y Särndal, 1992)[10] que proporciona una manera sistemática de incorporar la información auxiliar al proceso de estimación mejorando la eficiencia de los resultados.

Muchas encuestas, generalmente asumen que una muestra de unidades se observa por la selección en dos etapas a partir de una población finita, que se agrupan en grupos. Este diseño incluye muestras de dos poblaciones diferentes: la población de unidades primarias (familias, centros, hospitales, etc.) y la población de unidades secundarias (individuos). Los estimadores de calibración se pueden definir mediante el uso de la información combinada basada en los totales de las unidades primarias y secundarias.

El objetivo principal de este trabajo es aportar nuevos recursos metodológicos para la mejora de la eficiencia de las estimaciones, reduciendo la variabilidad de éstas, a partir de la combinación de estimadores que calibran a distintos niveles. Para tal propósito, partiremos de

la elección de las variables auxiliares pertinentemente elegidas según métodos que reducen los errores de estimación. Esta investigación se inscribe en un proyecto de investigación más general, actualmente en desarrollo en el Grupo de Investigación Diseño y Análisis Estadístico de Encuestas por Muestreo, en el cual se están abordando diferentes aplicaciones referentes al uso de la información auxiliar en el tratamiento de encuestas.

Los objetivos específicos de este trabajo son:

(1) la construcción de estimadores, adaptables a la información auxiliar a dos niveles (unidades muestrales primarias y secundarias), que sean más precisos que los estimadores simples para todas las variables de interés;

(2) la selección de métodos de elección de las variables auxiliares óptimas, que reduzcan el sesgo de no respuesta, del conjunto de posibles;

(3) la aplicación de los distintos métodos a datos reales con el fin de comprobar la eficiencias de tales técnicas (estudios de simulación).

En el primer capítulo se realiza una introducción al problema de investigación y a la notación que se va a utilizar en este trabajo. En términos generales, en este capítulo se realiza la descripción del estimador de Horvitz-Thompson (1952)[22], se especifica el uso de la información auxiliar mostrando algunos ejemplos de estimadores basados en su uso, y se describe la metodología de calibración, mostrando sus características principales, las distancias más utilizadas, ejemplos teóricos y prácticos y algunas extensiones de la calibración.

En el segundo capítulo se revisa algunas perspectivas del uso de la calibración en presencia de información compuesta, tales como información proveniente de diseños de muestreo en dos fases o en dos etapas. Se describen dos casos particulares que combinan la información disponible en ambas etapas: la integración de pesos (Estevao y Särndal,

2006)[16]) y el método de Lemaître y Dufour (1987)[32]. Para finalizar este capítulo, se describe el estimador propuesto en este trabajo y se realiza un estudio de simulación, con datos reales provenientes del estudio PISA 2006 y de la encuesta de presupuestos familiares, que evalúa el comportamiento empírico del estimador propuesto para dos tipos distintos de muestreos (muestreo aleatorio simple y de Midzuno). Los resultados se compararán con el estimador descrito por Estevao y Särndal (2006)[16].

En el tercer capítulo se describe el uso de la calibración para el ajuste del sesgo de no respuesta y se definen cuatro indicadores que nos permitirán elegir que variables auxiliares son más eficaces para construir el vector auxiliar. El capítulo finaliza con un estudio de simulación con datos reales (PISA 2006) en el que implementamos, y mostramos, como la elección apropiada del vector auxiliar permite reducir el sesgo de no repuesta.

Este trabajo finaliza con un apéndice donde se describen todas las poblaciones y programas implementados para los estudios de simulación, junto con las tablas no incluidas en el Capítulo 3.

Índice general

"...la calibración proporciona un enfoque sencillo y práctico para la incorporación de la información auxiliar en la estimación."

Rueda, Martínez, Martínez y Arcos

CAPÍTULO

1

Información auxiliar en encuestas por muestreo

En muestreo en poblaciones finitas podemos encontrar diferentes métodos de estimación que hacen uso de la información auxiliar para mejorar dichas estimaciones. Una de las metodologías que más auge han experimentado en los últimos años ha sido la calibración (Deville y Särndal, 1992)[10] que se apoya en la información auxiliar para mejorar la eficiencia de los resultados en comparación con otros estimadores basados únicamente en las variables de interés. En este primer capítulo se realiza una introducción al problema de investigación y a la notación que se va a utilizar en este trabajo. Partiremos de la descripción del estimador de Horvitz-Thompson, estimador que no utiliza información auxiliar, para seguidamente centrarnos en la información auxiliar mostrando algunos ejemplos de estimadores basados en su uso. Posteriormente se describe la metodología de calibración, indicando sus características principales, las distancias más utilizadas, ejemplos teóricos y prácticos y algunas extensiones de la calibración.

1.1 Notación

Para facilitar la lectura de este trabajo, primero vamos a introducir algo de terminología básica y notación, que ayudará en la lectura de las páginas siguientes. Sea $\mathcal{U}$ una población finita que contiene N elementos $\mathcal{U} = \{u_1, \ldots, u_k, \ldots, u_n\}$. Para simplificar, el elemento k-ésimo de la población estará representado por su etiqueta k. Por tanto, denotamos la población finita como

$$\mathcal{U} = \{1, \ldots, k, \ldots, N\}.$$

Sea y una variable de interés en la población e y_k su valor en el elemento k-ésimo de la población. Queremos hacer inferencia sobre cantidades desconocidas de esta variable, llamadas parámetros, tales como el total de la población

$$Y = \sum_{k=1}^{N} y_k,$$

o la media de la población

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{k=1}^{N} y_k.$$

Para el caso de una proporción, si A es una variable indicadora de la característica en estudio y a_k toma los valores $0, 1$, tendremos:

$$P = \frac{1}{N} \sum_{k=1}^{N} a_k.$$

Para estimar estos parámetros, los valores y_k se observan sólo para un subconjunto de la población $\mathcal{U}$, una muestra s, de tamaño $n < N$, seleccionada de la población con un procedimiento de muestreo.

Sea $\mathcal{S}$ es el conjunto de todas las muestras que se pueden obtener a partir de $\mathcal{U}$ con un procedimiento de muestreo y sea $p(s)$ la probabilidad de que la muestra s sea seleccionada.

La probabilidad $p(\ )$ verifica

$$p(s) \geq 0 \qquad \forall s \in \mathcal{S}, \qquad \text{y} \qquad \sum_{s \in \mathcal{S}} p(s) = 1,$$

y por tanto, esta función define una distribución de probabilidad sobre $\mathcal{S}$ llamada plan de muestreo o diseño muestral. Obviamente, puede haber varias maneras (esquemas de selección) para extraer una muestra con un diseño dado.

Una vez que el plan de muestreo se ha fijado, la probabilidad que tiene una determinada unidad k de ser incluida en la muestra depende de la aleatoriedad en la selección de s de $\mathcal{S}$. Sea δ_k la variable aleatoria definida como

$$\delta_k(s) = \begin{cases} 1 & \text{si} \quad k \in s \\ 0 & \text{en otro caso} \end{cases}, \tag{1.1}$$

entonces la probabilidad π_k de que el elemento k sea incluido en la muestra s puede ser calculada como la probabilidad de que el indicador de pertenencia a la muestra δ_k tome el valor 1, es decir,

$$\pi_k = P(k \in s) = P(\delta_k = 1) = \sum_{k \in s} p(s), \tag{1.2}$$

donde $k \in s$ indica que la suma se toma sobre todas las muestras que contienen a la unidad k. A π_k se le llama probabilidad de inclusión de primer orden.

Del mismo modo, la probabilidad de que dos elementos k y l sean incluidos en la muestra s, se denota por π_{kl} y se define como

$$\pi_{kl} = P(k, l \in s) = P(\delta_k \delta_l = 1) = \sum_{k,l \in s} p(s). \tag{1.3}$$

A π_{kl} se llama probabilidad de inclusión de segundo orden. Así $\pi_{kl} = \pi_{lk}$, para cada $k, l \in \mathcal{U}$ y $\pi_{kk} = \pi_k$ para cada $k \in \mathcal{U}$. Por tanto, cada diseño $p(s)$ induce unas probabilidades de inclusión de primer orden, π_k, y de segundo orden, π_{kl}.

Se define la esperanza de un estimador $\hat{\theta}(s)$ de un parámetro θ como

$$E[\hat{\theta}(s)] = \sum_{s \in \mathcal{S}} p(s)\hat{\theta}(s),$$

y la varianza como

$$V[\hat{\theta}(s)] = E\{[\hat{\theta}(s) - E(\hat{\theta}(s))]^2\} = \sum_{s \in \mathcal{S}} p(s)[\hat{\theta}(s) - E(\hat{\theta}(s))]^2.$$

Notaremos en general θ al parámetro de interés, $\hat{\theta}$ al estimador de θ y Θ al espacio paramétrico. Un estimador se llama insesgado si

$$E(\hat{\theta}) = \sum_{s \in \mathcal{S}} \hat{\theta}(s) p(s) = \theta, \qquad \forall \theta \in \Theta.$$

En otro caso se llama sesgado y se mide el sesgo con

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

El error cuadrático medio mide la dispersión de un estimador respecto al valor del parámetro y se define como

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \sum_{s \in \mathcal{S}} p(s)[\hat{\theta}(s) - \theta]^2, \tag{1.4}$$

aunque también se puede expresar como

$$ECM(\hat{\theta}) = V(\hat{\theta}) + Sesgo^2(\hat{\theta}), \tag{1.5}$$

y por tanto, si un estimador es insesgado su error cuadrático medio coincide con su varianza.

Para comparar distintos estimadores se suele usar el sesgo relativo (SR) y la eficiencia relativa (relativa a cierto estimador θ_0) (ER), que para un estimador $\hat{\theta}$ están definidos como

$$\text{SR} = E(\hat{\theta} - \theta)/\theta, \tag{1.6}$$

y

$$\text{ER} = ECM(\hat{\theta}_0)/ECM(\hat{\theta}). \tag{1.7}$$

1.2 El estimador de Horvitz-Thompson

Introducimos el estimador de Horvitz-Thompson (1952)[22] del total y de la media. Dado un diseño muestral, el estimador de Horvitz-Thompson del total de y es

$$\hat{Y}_{ht} = \sum_{k \in s} \frac{y_k}{\pi_k}. \tag{1.8}$$

Si llamamos $d_k = \pi_k^{-1}$ a los pesos del diseño de cada unidad k de la muestra, se tiene

$$\hat{Y}_{ht} = \sum_{k \in s} d_k y_k. \tag{1.9}$$

El estimador es insesgado ya que $E(\delta_k) = \pi_k$ y su varianza es:

$$V(\hat{Y}_{ht}) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N} \sum_{k \neq l}^{N} (\frac{\pi_{kl}}{\pi_k \pi_l} - 1) y_k y_l. \tag{1.10}$$

Cuando las probabilidades de inclusión de primer y segundo orden son todas conocidas y positivas para todas las unidades de la población, la varianza se puede estimar de forma insesgada con

$$\hat{V}(\hat{Y}_{ht}) = \sum_{k \in s} \frac{1-\pi_k}{\pi_k^2} y_k^2 + \sum_{k \in s} \sum_{\substack{l \in s \\ l \neq k}} (\frac{\pi_{kl}}{\pi_k \pi_l} - 1) \frac{y_k y_l}{\pi_{kl}}. \tag{1.11}$$

La estimación de la media es $\widehat{\bar{Y}}_{ht} = \frac{\hat{Y}_{ht}}{N} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$, su varianza y la estimación de su varianza se pueden deducir teniendo en cuenta que $V(\widehat{\bar{Y}}_{ht}) = \frac{V(\hat{Y}_{ht})^2}{N^2}$ y $\hat{V}(\widehat{\bar{Y}}_{ht}) = \frac{\hat{V}(\hat{Y}_{ht})^2}{N^2}$.

Por ejemplo, en muestreo aleatorio simple sin reemplazamiento, en el que $\pi_k = \frac{N}{n}$, $\forall k$, el estimador de Horvitz-Thompson es $\hat{Y} = \frac{N}{n} \sum_{k \in s} y_k = N\bar{y}$, donde $\bar{y} = \frac{1}{n} \sum_{k \in s} y_k$ es la media muestral de y. La varianza y su estimación son $V(\hat{Y}) = N^2 \frac{1-f}{n} S_y^2$ y $\hat{V}(\hat{Y}) = N^2 \frac{1-f}{n} s_y^2$, donde $f = \frac{n}{N}$ es la fracción de muestreo, $S_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} (y_k - \bar{Y})^2$ es la varianza poblacional de y y $s_y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$ es la varianza muestral.

1.3 El uso de información auxiliar

Cuando está disponible algún tipo de información auxiliar en la población, y de alguna manera está relacionada con la variable de estudio y, se puede utilizar para mejorar la precisión de las estimaciones. Generalmente, la información auxiliar adopta formas diferentes y, según su naturaleza, puede ser empleada en diferentes etapas de la encuesta. La calibración en gran medida generaliza el uso de esta

información auxiliar. Algunos ejemplos del uso de la información auxiliar son los siguientes.

Muestreo con probabilidades proporcionales al tamaño

Si los valores de una variable auxiliar x son conocidos para todas las unidades en la población, es decir, x_k es conocido para $k = 1, \ldots, N$ y son aproximadamente proporcionales a los valores de la variable y, entonces esta información puede ser utilizada en la fase de diseño mediante la selección de los elementos con probabilidades desiguales y sin reposición. Las probabilidades de inclusión de primer orden son $\pi_k = np_k$, con $p_k = \dfrac{x_k}{\sum_{k=1}^{N} x_k}$.

Muestreo estratificado

Es otra forma de utilizar la información auxiliar en la etapa del diseño. Existe una partición de la población en H estratos disjuntos, cada uno de tamaño N_h: $\mathcal{U}_1, \ldots, \mathcal{U}_h, \ldots, \mathcal{U}_H$, con $\bigcup_{h=1}^{H} \mathcal{U}_h = \mathcal{U}$.

La partición se realiza de acuerdo a los valores tomados por un conjunto de variables auxiliares sobre todas las unidades de la población con el fin de obtener subpoblaciones homogéneas. Posteriormente se extrae una muestra independiente s_h de tamaño n_h de cada estrato. El total de la población en cada estrato se estima con el estimador de Horvitz-Thompson, así como su varianza y su estimación en cada estrato h: $\hat{Y}_h = \displaystyle\sum_{k \in s_h} \frac{Y_{hk}}{\pi_{hk}},\ h = 1, \ldots, H.$

Como $Y = \sum_{h=1}^{H} Y_h$, entonces $\hat{Y}_{str} = \sum_{h=1}^{H} \hat{Y}_h$ y aplicando la propiedad de independencia se obtiene $V(\hat{Y}_{str}) = \sum_{h=1}^{H} V(\hat{Y}_h)$ y $\hat{V}(\hat{Y}_{str}) = \sum_{h=1}^{H} \hat{V}(\hat{Y}_h)$.

Estimador de razón

A veces la información auxiliar no es completa, o sólo se tiene acceso a la muestra observada, y sin embargo el total la población de x, X, es conocido. Es decir, los valores x_k se observan sólo en las unidades de la muestra.

En este caso, la información auxiliar no se puede emplear durante la fase de diseño, pero sí en la etapa de estimación.

Un ejemplo es el estimador de razón, $\hat{Y}_{ra} = \hat{Y}_{ht}\frac{X}{\hat{X}_{ht}}$, que es el estimador de Horvitz-Thompson de Y ajustado por el factor $\frac{X}{\hat{X}_{ht}}$, siendo $\hat{X}_{ht}$ el estimador de Horvitz-Thompson de x. El estimador de razón es asintóticamente insesgado de Y, por lo que necesita muestras grandes para dar buenos resultados.

Postestratificación

Este estimador está definido cuando la pertenencia de cada unidad a un estrato se tiene después que la muestra es observada y sin embargo el tamaño total del estrato, N_h para $h = 1, \ldots, H$, es conocido.

La información auxiliar toma la forma de sumas de las variables x_h, para $h = 1, \ldots, H$, donde x_k toma el valor uno si la unidad $k \in \mathcal{U}_h$ y cero en caso contrario. El estimador del total es $\hat{Y}_{ps} = \sum_{h=1}^{H} \hat{Y}_h \frac{N_h}{\hat{N}_h}$, donde $\hat{N}_h$ es el estimador del tamaño del estrato h, y es un estimador de razón para cada post-estrato.

Estimador de diferencia

De un modo más general, la información auxiliar tiene la forma de un vector de dimensión P, $\mathbf{x} = (x_1, \ldots, x_P)$, para el que son conocidos los totales de las variables auxiliares $\mathbf{x}$, $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$. El estimador de diferencia se define como

$$\hat{Y}_d = \hat{Y}_{ht} + (\mathbf{X} - \hat{\mathbf{X}}_{ht})\boldsymbol{\lambda}, \tag{1.12}$$

donde $\boldsymbol{\lambda}$ es un vector de constantes. Es decir, se ajusta el estimador de Horvitz-Thompson, $\hat{Y}_{ht}$, con la diferencia entre el valor real $\mathbf{X}$ y su estimación. Este estimador es insesgado y da buenos resultados cuando se elige adecuadamente el vector de constantes y hay una relación aproximadamente lineal entre y y $\mathbf{x}$.

Estimador de regresión generalizado

Suponiendo que el total de la población $\mathbf{X}$ es conocido, un estimador que utiliza esta información es el de regresión generalizada (estimador GREG). Este estimador se explica y se ilustra con varios ejemplos en Särndal, Swensson y Wretman (1992)[56], Capítulos 6 y 7. Se define como

$$\hat{Y}_{greg} = \hat{Y}_{ht} + (\mathbf{X} - \hat{\mathbf{X}}_{ht})\hat{\boldsymbol{B}}, \tag{1.13}$$

donde

$$\hat{\boldsymbol{B}} = (\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k c_k \mathbf{x}_k y_k), \tag{1.14}$$

es el vector de coeficientes de regresión obtenidos ajustando los datos en la muestra $\{(y_k, \mathbf{x}_k), k \in s\}$ y donde los factores c_k se especifican previamente (la elección más simple es $c_k = 1, \forall k$).

Aunque el estimador no es insesgado, lo es asintóticamente. Por cada especificación que se puede dar del vector auxiliar $\mathbf{x}_k$ y del factor c_k se tiene un estimador diferente, que queda completamente definido después de hacer efectivo el muestreo.

Lo usual es expresarlo como una combinación lineal de los valores observados de la variable de interés:

$$\hat{Y}_{greg} = \sum_s d_k g_k y_k, \tag{1.15}$$

donde

$$g_k = 1 + c_k (\sum_{\mathcal{U}} \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)' (\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \tag{1.16}$$

son valores entorno a la unidad, aunque pueden aparecer valores grandes y negativos.

El estimador de Horvitz-Thompson puede verse como un caso particular del de regresión generalizada cuando $\mathbf{x}_k = c_k = 1, \forall k$ y el diseño verifique $\sum_s d_k = N$.

Una propiedad importante es que es consistente con la información auxiliar, en el sentido que reproduce el total conocido de antemano, es decir:

$$\hat{X}_{greg} = \sum_s d_k g_k \mathbf{x}_k = \mathbf{X}. \tag{1.17}$$

La varianza del estimador $\hat{Y}_{greg}$ definido en (1.15), se puede aproximar a partir de los residuos $E_k = y_k - \mathbf{x}_k' \boldsymbol{B}$, donde

$$\boldsymbol{B} = (\sum_{\mathcal{U}} c_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_{\mathcal{U}} c_k \mathbf{x}_k y_k) \tag{1.18}$$

La varianza del estimador viene dada por:

$$V(\hat{Y}_{greg}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \left(\frac{d_k d_l}{d_{kl}} - 1 \right) E_k E_l, \tag{1.19}$$

y la estimación de la varianza del estimador se aproxima por:

$$\widehat{V}(\hat{Y}_{greg}) = \sum_{s} \sum_{s} \left(d_k d_l - d_{kl} \right) (g_k e_k)(g_l e_l); \tag{1.20}$$

donde $e_k = y_k - x_k' \widehat{\boldsymbol{B}}$, con $\widehat{\boldsymbol{B}}$ definido en (1.14).

Ejemplos del estimador de regresión generalizado

Ejemplo 1. Una clasificación

Para cierta población de individuos se supone conocidos el número de hombres, N_1, y mujeres, N_2. En este caso, el vector $\mathbf{x}_k$ sólo puede tomar dos valores: $\mathbf{x}_k = (1, 0)'$ para los hombres y $\mathbf{x}_k = (0, 1)'$ para las mujeres y por tanto, el total poblacional de $\mathbf{x}_k$, $(N_1, N_2)'$, es conocido.

Sea s_1 el subconjunto de la muestra s formado por hombres y s_2 el formado por mujeres. Tenemos una variante de los g-pesos, definidos en (1.16), donde $g_k = \frac{N_1}{\sum_{s_1} d_k}$ cuando el elemento k es un hombre y $g_k = \frac{N_2}{\sum_{s_2} d_k}$ si es una mujer. Como $d_k g_k$ satisface la propiedad (1.17), entonces el estimador GREG viene dado por $\hat{Y}_{greg} = N_1 \frac{\sum_{s_1} d_k y_k}{\sum_{s_1} d_k} + N_2 \frac{\sum_{s_2} d_k y_k}{\sum_{s_2} d_k}$. El estimador GREG resultante se denomina estimador post-estratificado (con dos post-estratos).

Ejemplo 2. Una clasificación doble

Sea una población de individuos distribuidos en función del sexo y de la región de procedencia como se indica en la Tabla 1.1. En este ejemplo la información auxiliar está compuesta por los seis totales $N_{11}, \ldots, N_{23}$ y por tanto, el vector $\mathbf{x}_k$ incluye información para las seis componentes, tomando el valor 1 si cumple la condición de pertenencia y cero en el resto de casos. Por ejemplo, para cada elemento de la población que cumple la condición ser hombre y vivir en la segunda región, el vector auxiliar tomará el valor $x_k = (0, 1, 0, 0, 0, 0)'$ y el vector de totales de la población viene dado por $(N_{11}, \ldots, N_{23})'$. Con este tipo de información auxiliar, el

Tabla 1.1: Distribución de la población por sexo y región

Sexo		Región 1	Región 2	Región 3	Total
Hombre	1	N_{11}	N_{12}	N_{13}	$N_{1\cdot}$
Mujer	2	N_{21}	N_{22}	N_{23}	$N_{2\cdot}$
Total		$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	N

estimador $\hat{Y}_{greg}$ es un estimador post-estratificado, aunque ahora con seis términos, que corresponden a seis post-estratos.

Hay situaciones en las que la clasificación cruzada de las variables es poco práctica o inconveniente. Por ejemplo, cuando algunos totales de las celdas son pequeños, éstos pueden provocar un estimador inestable. Una alternativa es utilizar sólo la información definida por los recuentos marginales. En nuestro ejemplo, el vector auxiliar tendrá cinco dimensiones, las dos primeras posiciones indicaran el sexo y las tres finales la región. Por tanto el individuo de nuestro ejemplo será representado en el vector auxiliar con el valor $\mathbf{x}_k = (\underbrace{1,0}_{\text{sexo}},\underbrace{0,1,0}_{\text{región}})'$ y el vector de totales de la población vendrá dado por $(N_{1\cdot}, N_{2\cdot}, N_{\cdot 1}, N_{\cdot 2}, N_{\cdot 3})'$.

Ejemplo 3. Estimación en dominios

Supongamos que para cierta población queremos estimar de forma separada hombres y mujeres, por lo que hemos de definir dos dominios de la población. Partiendo de un muestreo aleatorio simple, en el que conocemos el número de hombres y mujeres en la población, se decide utilizar el estimador GREG para estimar el total del dominio Y_d, $d = 1, 2$. El vector auxiliar es $\mathbf{x}_k = (1, 0)'$ si el elemento k es hombre y $\mathbf{x}_k = (0, 1)'$ si el elemento k es mujer, tomando $c_k = 1$ para todo k, y la varianza es

$$V(\hat{Y}_{dgreg}) = N^2 \left(\frac{(1-n)/N}{n} \frac{1}{N-1} \right) \sum_{\mathcal{U}} E_{dk}^2,$$

donde los residuos E_{dk} son

Tabla 1.2: Vectores auxiliares y totales poblacionales

Caso	Vector auxiliar $\mathbf{x}_k$	Total en la población $\sum_{\mathcal{U}} \mathbf{x}_k$
i	x_k	$\sum_{\mathcal{U}} x_k$
ii	$(1, x_k)'$	$(N, \sum_{\mathcal{U}} x_k)'$
iii	$(0, x_k, 0, 0, 0, 0)'$	$(\sum_{\mathcal{U}_{11}} x_k, \ldots, \sum_{\mathcal{U}_{23}} x_k)'$
iv	$(\underbrace{0,1,0,0,0,0}_{\text{recuento}}, \underbrace{0, x_k, 0, 0, 0, 0}_{\text{x-variable}})'$	$(N_{11}, \ldots, N_{23}, \sum_{\mathcal{U}_{11}} x_k, \ldots, \sum_{\mathcal{U}_{23}} x_k)'$
v	$(\underbrace{1,0}_{\text{sexo}}, \underbrace{0, x_k, 0}_{\text{región}})'$	$(N_{1\cdot}, N_{2\cdot}, \sum_{\mathcal{U}_{\cdot 1}} x_k, \sum_{\mathcal{U}_{\cdot 2}} x_k, \sum_{\mathcal{U}_{\cdot 3}} x_k)'$

$$E_{dk} = \begin{cases} y_k - \frac{Y_d}{N_d} & \text{para} \quad k \in \mathcal{U}_d \\ 0 & \text{para} \quad k \in \mathcal{U} - \mathcal{U}_d \end{cases}$$

Notar que el vector auxiliar coincide exactamente con el indicador del dominio.

Ejemplo 4. Combinación de una clasificación en una dirección con una variable cuantitativa

Supongamos un conjunto de datos en los que se especifica el sexo y la región de pertenencia, Tabla 1.1, así como el valor $\mathbf{x}_k$ de una variable auxiliar cuantitativa, por ejemplo ingresos. Unos ejemplos de vectores auxiliares que pueden utilizarse se muestran en la Tabla 1.2.

Algunos estimadores conocidos surgen de estos cinco casos. Consideremos dos de ellos para un diseño muestral realizado a partir de un muestreo aleatorio simple. Cuando $\mathbf{x}_k = x_k$ y $c_k = 1/x_k$, el estimador de razón se obtiene de la formulación del estimador GREG definida en (1.15), es decir

$$\hat{Y}_{greg} = \sum_{\mathcal{U}} x_k \frac{\overline{y}_s}{\overline{x}_s};$$

donde $\overline{y}_s = \dfrac{1}{n}\sum_s y_k$ y $\overline{x}_s = \dfrac{1}{n}\sum_s x_k$.

Cuando $\mathbf{x}_k = (1, x_k)'$ y $c_k = 1$ para todo k, se obtiene el estimador de regresión, es decir,

$$\hat{Y}_{greg} = N\{\overline{y}_s + (\overline{X} - \overline{x}_s)\widehat{B}\};$$

donde $\overline{X} = \dfrac{\sum_{\mathcal{U}} x_k}{N}$ y $\widehat{B} = \dfrac{Cov_{xys}}{S^2_{xs}}$, con $Cov_{xys} = \dfrac{1}{n-1}\sum_s (x_k - \overline{x}_s)(y_k - \overline{y}_s)$ y $S^2_{xs} = \frac{1}{n-1}\sum_s (x_k - \overline{x}_s)^2$.

1.4 Calibración

En este apartado nos centraremos en el método de calibración, una herramienta muy conocida y ampliamente empleada para el uso de la información auxiliar en la etapa de estimación.

1.4.1 Introducción a la estimación por calibración

Sea y la variable de interés en la estimación de la encuesta. Sin información auxiliar ninguna, el total de y, Y, se estima de forma insesgada con el estimador de Horvitz–Thompson (1.8) mediante $\hat{Y}_{ht} = \sum_{k\in s} d_k y_k$. Sea $\mathbf{x}$ un vector auxiliar asociado a y, del cual suponemos conocido el total en la población $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$. La estimación de calibración de Y consiste en la obtención de un nuevo vector de pesos w_k, para $k \in s$ que modifica lo menos posible los pesos de muestreo originales, d_k, que tienen la deseable propiedad de producir estimaciones insesgadas, respetando al mismo tiempo las ecuaciones de calibración:

$$\sum_{k\in s} w_k \mathbf{x}_k = \mathbf{X}. \tag{1.21}$$

Dada una distancia $G(w_k, d_k)$, el proceso de calibración consiste en encontrar la solución del problema de minimización

$$\min_{w_k} E\{\sum_{k\in s} G(w_k, d_k)\} \tag{1.22}$$

respetando al mismo tiempo la ecuación de calibración (1.21).

Dado que la ecuación depende de la distancia escogida $G(w_k, d_K)$, cada distancia diferente conduce a un sistema de ponderación específico y por tanto, a un

nuevo estimador. Para cada elemento k de cada muestra s, $G(w_k, d_k)$ deberá cumplir las siguientes características básicas de una distancia:

i) para cada $d_k > 0$ fijado, $G(w_k, d_k)$ es no negativa, diferenciable respecto de w_k, estrictamente convexa, definida en un intervalo que contenga a d_k y tal que $G(d_k, d_k) = 0$;

ii) $g(w_k, d_k) = \dfrac{\partial G(w_k, d_k)}{\partial w_k}$ es una función continua en el intervalo definida con $g(w_k, d_k)$, estrictamente creciente en w_k y $g(d_k, d_k) = 0$.

Llamando $\boldsymbol{\lambda}$ al vector de multiplicadores de Lagrange, el problema de minimización se resuelve con

$$g(w_k, d_k) - \mathbf{x}_k\boldsymbol{\lambda} = 0 \qquad \forall k \in s. \tag{1.23}$$

Si la solución existe, $i)$ e $ii)$ garantizan que es única y se puede escribir como

$$w_k = d_k F_k(q_k \mathbf{x}_k \boldsymbol{\lambda}), \tag{1.24}$$

donde $d_k F_k(\ )$ es la función recíproca de $g(\cdot, d_k)$, que es creciente; $F_k(\cdot)$ verifica $F_k(0) = 1$ y $F_k'(0) = q_k > 0$, donde $1/q_k$ son pesos positivos no relacionados con los pesos del diseño d_k. Habitualmente $1/q_k = 1$ y los pesos de calibración son

$$w_k = d_k F_k(\mathbf{x}_k \boldsymbol{\lambda}), \tag{1.25}$$

aunque también pueden usarse pesos diferentes.

Las ecuaciones necesarias para determinar $\boldsymbol{\lambda}$, obtenidas subtituyendo w_k en la ecuación de calibración, son:

$$\mathbf{X} = \sum_{k \in s} w_k \mathbf{x}_k = \sum_{k \in s} d_k F_k(\mathbf{x}_k' \boldsymbol{\lambda}) \mathbf{x}_k. \tag{1.26}$$

y el estimador de calibración resultante es:

$$\hat{Y}_{cal} = \sum_{k \in s} w_k y_k = \sum_{k \in s} d_k F_k(\mathbf{x}_k' \boldsymbol{\lambda}) y_k. \tag{1.27}$$

Si existe una fuerte relación lineal entre la variable y y la información auxiliar $\mathbf{x}$, el estimador de calibración proporciona estimaciones precisas del total Y. El

estimador de calibración puede ser escrito solamente si la solución de la ecuación existe.

Deville y Särndall (1992)[10] consideran diferentes distancias $G(w_k, d_k)$, que en algunos casos tienen siempre solución. Las distancias examinadas, Tabla 1.3, se pueden resumir según los diferentes valores de α (Tillé, 2005[68]), siendo:

$$G^{\alpha}(w_k, d_k) = \begin{cases} \dfrac{\frac{w_k^{\alpha}}{d_k^{\alpha-1}} + (\alpha-1)d_k - \alpha w_k}{\alpha(\alpha-1)} & \alpha \in \mathbb{R}\backslash\{0,1\} \\ w_k \log(\dfrac{w_k}{d_k}) + d_k - w_k & \alpha = 1 \\ d_k \log(\dfrac{d_k}{w_k}) + w_k - d_k & \alpha = 0 \end{cases}$$

Tabla 1.3: Distancias usuales utilizadas en calibración

Caso	α	$G_k^{\alpha}(w_k, d_k)$	$g^{\alpha}(w_k, d_k)$	$F_k^{\alpha}(u)$	Tipo
1	2	$\dfrac{(w_k - d_k)^2}{2d_k}$	$\dfrac{w_k}{d_k} - 1$	$1 + q_k u$	Chi-cuadrado
2	1	$w_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k$	$\log\left(\frac{w_k}{d_k}\right)$	$\exp(q_k u)$	Entropía
3	1/2	$\left(\sqrt{w_k} - \sqrt{d_k}\right)^2$	$2\left(1 - \sqrt{\frac{w_k}{d_k}}\right)$	$\left(1 - \frac{q_k u}{2}\right)^{-2}$	Distancia de Hellinger
4	0	$d_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k$	$1 - \left(\dfrac{w_k}{d_k}\right)^{-1}$	$(1 - q_k u)^{-1}$	Entropía Inversa
5	-1	$\dfrac{(w_k - d_k)^2}{2w_k}$	$\dfrac{1 - \left(\frac{w_k}{d_k}\right)^{-2}}{2}$	$(1 - 2q_k u)^{-1/2}$	Chi-Cuadrado inverso

Notar que derivando $G^{\alpha}(w_k, d_k)$ respecto w_k tenemos

$$g^{\alpha}(w_k, d_k) = \begin{cases} \dfrac{1}{\alpha - 1}\left(\dfrac{w_k^{\alpha-1}}{d_k^{\alpha-1}} - 1\right) & \alpha \in \mathbb{R}\backslash\{1\} \\ \log(\dfrac{w_k}{d_k}) & \alpha = 1 \end{cases}$$

donde la inversa de $g^{\alpha}(w_k, d_k)/q_k$ con respecto a w_k es

$$d_k F_k^{\alpha}(u) = \begin{cases} d_k \sqrt[\alpha-1]{1 + q_k u(\alpha - 1)} & \alpha \in \mathbb{R}\backslash\{1\} \\ d_k \exp(q_k u) & \alpha = 1 \end{cases}$$

Como indica Tillé, las distancias más usadas son los casos $\alpha = 2$ que hace referencia a la distancia Chi-cuadrado y $\alpha = 1$ que hace referencia a la entropía. Para cada una de las distancias, Deville y Särndall (1992)[10] deducen la función $F_k(q_k \mathbf{x}_k \boldsymbol{\lambda})$ y muestran que se puede escribir siempre como $(1 + \alpha q_k u)^{(1/\alpha)}$, donde $u = \mathbf{x}_k \boldsymbol{\lambda}$. Para los casos 1, 3, 4 y 5, α toma los valores 1, -1/2, -1 y -2. El caso 2 se obtiene cuando $\alpha \to 0$.

Los casos 1 y 2 siempre tienen solución, en los casos 3, 4 y 5 la solución no está garantizada. En el caso 1 los pesos pueden ser positivos o negativos, mientras que todas las otras distancias garantizan pesos positivos, aunque algunos pesos pueden ser muy grandes en comparación con los pesos del diseño.

Deville y Särndall (1992)[10] también consideran algunos casos más, que tienen la propiedad de que los pesos que proporcionan están incluidos en un intervalo que puede especificarse de antemano, de modo que los pesos extremos puede ser eliminados, conservando el estimador sus buenas propiedades desde el punto de vista de la estimación.

Para un tratamiento posterior definimos el $caso\ 1$ como linear o lineal, el $caso\ 2$ como raking, y el $caso\ 3$ como logit o logístico. Por su importancia definiremos de forma más explícita cada una de estas distancias.

Método lineal

Un caso importante, que denominamos método *lineal*, se logra usando como distancia una función Chi-cuadrado (caso $\alpha = 2$). Siguiendo con la terminología de Tillé (2005)[68] quedaría

$$w_k = d_k(1 + q_k \mathbf{x}_k \boldsymbol{\lambda}), \tag{1.28}$$

donde el vector de multiplicadores de Lagrange $\boldsymbol{\lambda}$ es

$$\boldsymbol{\lambda} = \mathbf{T}^{-1}(\mathbf{X} - \hat{\mathbf{X}})' \tag{1.29}$$

y

$$\mathbf{T} = \sum_{k \in s} d_k q_k \mathbf{x}_k' \mathbf{x}_k, \tag{1.30}$$

asumiendo que existe la inversa de $\mathbf{T}$.

Sustituyendo se obtiene

$$\hat{Y}_{cal} = \sum_{k \in s} w_k y_k = \hat{Y}_{ht} + (\mathbf{X} - \hat{X}_{ht})\hat{\mathbf{B}}, \tag{1.31}$$

donde $\hat{Y}_{ht} = \sum_{k \in s} d_k y_k$ es el estimador de Horvitz-Thompson de Y y $\hat{\mathbf{B}}$ es

$$\hat{\mathbf{B}} = \mathbf{T}^{-1} \sum_{k \in s} d_k q_k \mathbf{x}_k y_k. \tag{1.32}$$

Este estimador es el estimador de regresión generalizado y tiene la buena propiedad, como se comentó anteriormente, de que es una ponderación lineal de las observaciones y_k, por pesos w_k, que no dependen de la variable de interés y.

Método Raking

El método *raking*, que incluye el estimador de calibración sobre márgenes, se logra usando una pseudo-distancia del tipo *entropía* (caso $\alpha = 1$). Siguiendo con la terminología de Tillé (2005)[68] quedaría:

$$G^1(w_k, d_k) = w_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k,$$

de la que obtenemos una función lineal

$$F_k(u) = \exp(q_k u).$$

En este caso particular los pesos son siempre positivos. Éstos vienen dados por

$$w_k = d_k \exp\left(q_k \boldsymbol{\lambda} \mathbf{x}_k\right),$$

donde $\boldsymbol{\lambda}$ es calculado por la ecuación

$$\sum_{k \in s} d_k x_k \exp\left(q_k \boldsymbol{\lambda} \mathbf{x}_k\right) = \mathbf{X}.$$

Un caso particular es la denominada *calibración sobre márgenes*. En este caso, los $\mathbf{x}_k$ toman los valores 0 ó 1 según la unidad i esté o no en la subpoblación $\mathcal{U}_i \subseteq \mathcal{U}$. Si además, $q_k = 1$, con $k \in \mathcal{U}$, tenemos

$$w_k = d_k \prod_{i|k \in \mathcal{U}_i} \beta_i,$$

donde $\beta = \exp \boldsymbol{\lambda}$. Los elementos de β son calculados mediante la ecuación

$$\sum_{k \in s} d_k x_k \prod_{i|k \in \mathcal{U}_i} \beta_i = \mathbf{X}.$$

Método logit

Como destaca Tillé (2005)[68], a veces se requiere que los pesos w_k no sean demasiado variables. Este inconveniente se puede ajustar imponiendo que los pesos se encuentren entre dos valores L y H, con $L < 1 < H$, usando una función del tipo *logit*

$$G(w_k, d_k) = \begin{cases} a_k \log(\frac{a_k}{1-L}) + b_k \log(\frac{b_k}{H-1}) \frac{1}{A} & L < w_k < H \\ \infty & En\ otro\ caso \end{cases}$$

donde

$$a_k = \frac{w_k}{d_k} - L, \;\; b_k = H - \frac{w_k}{d_k}, \;\; A = \frac{H - L}{(1-L)(H-1)},$$

de la que obtenemos

$$F_k(u) = \frac{L(H-1) + H(1-L)\exp(Aq_k u)}{H - 1 + (1-L)\exp(Aq_k u)}.$$

que cumple las condiciones $F_k(-\infty) = L$, $F_k(\infty) = H$. De esta forma los pesos obtenidos siempre estarán en el intervalo $[Ld_k, Hd_k]$.

Otras funciones distancia alternativas se comparan en Deville, Särndal y Sautory (1993)[11], Singh y Mohl (1996)[61], Stukel, Hidiroglou y Särndal (1996)[62].

Algunas de estas funciones distancia garantizarán pesos dentro de los límites especificados, descartando pesos demasiados grandes, demasiados pequeños o negativos. Los cambios en la función distancia a menudo tienen poco efecto en la varianza del estimador de la calibración, incluso si el tamaño de la muestra es bastante pequeño. Las preguntas acerca de la existencia de una solución a la ecuación de calibración se analizan en Théberge (2000)[64]. Por ejemplo, si tomamos como función distancia $G(w_k, d_k) = \nu_k d_k (w_k/d_k - 1)^2/2$ y $F(u\nu_k^{-1}) = 1 + u\nu_k^{-1}$ se tiene como pesos $w_k = d_k \left(\frac{1+\mathbf{x}_k'\lambda_c}{\nu_k}\right)$ y como estimador

$$\widehat{Y}_{cal} = \widehat{Y}_{ht} + \sum_{k\in s} d_k(1 + \mathbf{x}_k'\boldsymbol{\lambda}_c/\nu_k)y_k.$$

El cálculo de los pesos calibrados plantea importantes cuestiones prácticas, tales como evitar pesos indeseados (o variables indebidas), o el planteamiento como requisito de que todos los pesos sean positivos (incluso mayores que la unidad) o evitar pesos demasiados grandes. Podemos encontrar casos, como algunos de los pesos calculados de acuerdo con el estimador GREG lineal, que pueden proporcionar pesos bastantes grandes o negativos. Cuando intervenimos en el cálculo de los pesos, con el fin de eliminar valores indeseables, se plantea la cuestión de hasta qué punto se pueden desviar de los pesos del diseño, d_k, sin comprometer la característica de estimación insesgada. Chambers (1996)[7] explora la idea de modificar el conjunto de restricciones para que las tolerancias sean respetadas por la diferencia entre el estimador para las variables auxiliares y los correspondientes totales conocidos de la población, minimizando una "función de pérdida de costo rígido".

Una causa de pesos extremos pueden ser los valores extremos en las variables. Autores como Duchesne (1999)[12] discuten como tratar la calibración en presencia de valores anómalos. Duchesne aplica la técnica de "calibración robusta" para introducir un cierto sesgo en las estimaciones, que se ve compensado por una reducción en la variabilidad. Cuando el conjunto de restricciones se extienden para hacer que los pesos se restrinjan a intervalos especificados, puede ocurrir que la solución al problema de optimización no esté garantizada. La existencia de una solución se considera en Théberge (2000)[64], que también propone métodos para tratar los valores atípicos.

Investigaciones como las de Huang y Fuller (1978)[23] o la de Park y Fuller (2005)[41] proponen métodos para evitar pesos indeseables. En el método de minimización de la distancia, la función distancia puede ser formulada de modo que los pesos negativos estén excluidos, mientras que aún satisfagan las ecuaciones de calibración dadas. Algunos tipos de software, como CALMAR (Deville, Särndal y Sautory 1993), o la versión más actualizada, CALMAR-2, descrita en Le Guennec y Sautory (2002)[31], permiten aplicar varias funciones distancia de este tipo. Otros organismos de estadística han desarrollado su propio software para el cálculo del pesos, entre ellos: GES (Statistics Canada), CLAN97 (Statistics Sweden), Bascula 4.0 (Central Bureau of Statistics, The Netherlands), descrito en Nieuwenbroek y Boonstra (2002)[39], g-CALIB-S (Statistics Belgium), descrito en Vanderhoeft, Waeytens y Museux (2001)[72] y en Vanderhoeft (2001)[71] o las últimas versiones de las bibliotecas del software estadístico R: sampling (Tillé y Matei, 2013[69]), survey (Lumley, 2013[33]), EVER (Zardetto, 2013)[77], TeachingSampling (Gutiérrez, 2013[19]) o laeken (Alfons, Holzer y Templ, 2013)[2]. Destacamos la biblioteca "survey" que permite calibrar con una distancia definida por el usuario.

1.4.2 Estimación de la varianza

Todos los estimadores de calibración son asintóticamente equivalentes al estimador de regresión generalizado, generado por la función lineal $F_k(u) = 1 + q_k u$. Así, la elección de la medida de la distancia se suele basar en el comportamiento de los pesos finales. La varianza asintótica es

$$AV(\hat{Y}_{cal}) = \sum_{k=1}^{N} \sum_{k \neq l}^{N} (\pi_{kl} - \pi_k \pi_l)(d_k e_k)(d_l e_l), \tag{1.33}$$

donde $e_k = y_k - \mathbf{x}_k \tilde{\boldsymbol{\beta}}$, con $\tilde{\boldsymbol{\beta}}$ verificando

$$(\sum_{k=1}^{N} q_k \mathbf{x}_k' \mathbf{x}_k) \tilde{\boldsymbol{\beta}} = \sum_{k=1}^{N} q_k \mathbf{x}_k' y_k. \tag{1.34}$$

La varianza es el límite de $\tilde{\boldsymbol{\beta}}$ y minimiza la expresión de mínimos cuadrados ponderados

$$\sum_{k=1}^{N} q_k (y_k - \mathbf{x}_k \tilde{\boldsymbol{\beta}})^2 = \sum_{k=1}^{N} q_k e_k^2. \tag{1.35}$$

Si se mira el término de la derecha como el total en la población de la variable $q_k e_k^2$, se puede estimar por calibración con $\sum_{k\in s} w_k q_k e_k^2$ minimizando con $\hat{\mathbf{B}}$ (descrito en 1.32) y cumpliendo las ecuaciones normales basadas en la muestra, $(\sum_{k\in s} w_k q_k \mathbf{x}_k \mathbf{x}_k')\hat{\mathbf{B}} = \sum_{k\in s} w_k q_k \mathbf{x}_k y_k$. Los residuos muestrales son $\hat{e}_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}$ y la estimación de la varianza es

$$\hat{V}(\hat{Y}_{cal}) = \sum_{k\in s} \sum_{l\neq k\in s} \frac{\Delta_{kl}}{\pi_{kl}} (w_k \hat{e}_k)(w_l \hat{e}_l). \tag{1.36}$$

con $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Siendo esta formulación una variante de la definida en 1.20.

1.4.3 Ejemplos de estimadores de calibración

El objetivo de esta sección es mostrar diferentes enfoques de la calibración para algunas especificaciones simples de la información auxiliar y mostrar la relación con algunos de los estimadores más utilizados. En el siguiente ejemplo asumimos un muestreo aleatorio simple con pesos del diseño $d_k = N/n$ para todo k, donde n es el tamaño de la muestra.

Ejemplo 1: Vector auxiliar simple

La formulación más simple del vector auxiliar es cuando se toma $\mathbf{x}_k = 1$ para todo k, por tanto este vector no reconoce las diferencias entre los elementos. Si tomamos también $c_k = 1$ para todo k, el estimador de calibración viene dado por

$$\hat{Y}_{cal} = \frac{N}{m} \sum_{s} y_k = \hat{Y}_{exp}, \tag{1.37}$$

con N el tamaño de la población y m el tamaño muestral.

El sufijo exp hace referencia al *estimador de expansión*. El estimador $\hat{Y}_{exp}$ es un estimador elemental. En ocasiones este estimador tiene un uso práctico, por ejemplo cuando no se dispone de información auxiliar útil.

Ejemplo 2: Una clasificación

En esta formulación, la población objetivo $\mathcal{U}$ se divide en grupos no-superpuestos y exhaustivos, $\mathcal{U}_p$, $p = 1, \ldots, P$, sobre la base de un criterio de clasificación especificado, por ejemplo la edad por grupos del mismo sexo. El vector auxiliar para el elemento k es el identificador del grupo $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk})'$, para $p = 1, \ldots, P$, donde

$$\gamma_{pk} = \begin{cases} 1 & \text{si} \quad k \in \mathcal{U}_p \\ 0 & \text{en otro caso} \end{cases} \tag{1.38}$$

Si $\sum_{\mathcal{U}} \mathbf{x}_k = (N_1, \ldots, N_p, \ldots, N_P)'$, donde N_p es el tamaño de $\mathcal{U}_p$, cumplimos el requisito de que el total de la población auxiliar equivalga a los tamaños conocidos de grupo P. Tomando $c_k = 1$ para todo k obtenemos de (1.16) que $q_k = \frac{N_p n}{N m_p}$ para $k \in s_p$, con s_p la muestra en el grupo p.

Con estas premisas el estimador de calibración se convierte en

$$\hat{Y}_{cal} = \sum_{p=1}^{P} N_p \bar{y}_k = \hat{Y}_{ps} \tag{1.39}$$

donde $\bar{y}_k = \frac{1}{m_p} \sum_{s_p} y_k$ y m_p es el número de encuestados en el grupo p. Este estimador, $\hat{Y}_{pst}$, es comúnmente denominado como *estimador de post-estratificación*.

Cuando el conocimiento del vector auxiliar $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk})'$ se limita a los elementos de la muestra s y tomamos $c_k = 1$ para todo k, obtenemos

$$\hat{Y}_{cal,s} = \sum_{p=1}^{P} \hat{N}_p \bar{y}_k = \hat{Y}_{pc} \tag{1.40}$$

donde $\hat{N}_p = \frac{N}{n} n_p$ y n_p es el número de elementos incluidos en la muestra del grupo p. Este estimador, denotado por $\hat{Y}_{pc}$, es conocido como *estimador de ponderación por clases*.

Ejemplo 3: Una sola variable cuantitativa

Supongamos que disponemos de una variable auxiliar cuantitativa, x_k, por ejemplo, el número de empleados de una empresa k en una encuesta de negocios, con $k = 1, \ldots, N$. Supongamos que la población total, $\sum_{\mathcal{U}} \mathbf{x}_k$, es conocida. Si ésta es la

única variable auxiliar, el vector auxiliar es unidimensional, y por tanto $\mathbf{x}_k = x_k$. Si tomamos $c_k = x_k^{-1}$, se obtiene el estimador

$$\hat{Y}_{cal} = \sum_{\mathcal{U}} x_k \frac{\bar{y}_s}{\bar{x}_s} = \hat{Y}_{ra} \tag{1.41}$$

con $\bar{y}_s = \frac{1}{m}\sum_s y_k$ y $\bar{x}_r = \frac{1}{m}\sum_s x_k$. Este estimador, $\hat{Y}_{ra}$, es conocido como *estimador de razón*.

Con la misma información podemos formular de forma alternativa el vector auxiliar como $\mathbf{x}_k = (1, x_k)'$. Esta opción es viable ya que se requiere de la información auxiliar $N = \sum_{\mathcal{U}} 1$ (tamaño de la población), que es conocido además de $\sum_{\mathcal{U}} x_k$. Cuando se toma $c_k = 1$ para todo k, tenemos

$$\hat{Y}_{cal} = N\{\bar{y}_s + (\bar{X} - \bar{x}_s)\hat{B}\} = \hat{Y}_{reg} \tag{1.42}$$

donde $\bar{X} = \frac{1}{N}\sum_{\mathcal{U}} x_k$ y $B = (\sum_s y_k x_k - \frac{1}{m}\sum_s y_k \sum_s x_k)/(\sum_s x_k^2 - \frac{1}{m}(\sum_s x_k)^2)$.

La notación $\hat{Y}_{reg}$ se utiliza para indicar el *estimador de regresión*.

Ejemplo 4: Clasificación en una dirección combinado con una variable cuantitativa

En este ejemplo, la información auxiliar toma una variable categórica con P-valores y una variable cuantitativa x, que puede ser el indicador del tamaño de un elemento. Supongamos que podemos colocar cada elemento k de la muestra, para el que conocemos su valor x_k, en un grupo adecuado y que para cada grupo $p = 1, \ldots, P$, también conocemos el tamaño N_p, y el total $\sum_{\mathcal{U}_P} y_k$. Hay más de una forma de usar esta información. Una opción es definir el vector auxiliar como

$$\mathbf{x}_k = (\gamma_{1k}x_k, \ldots, \gamma_{pk}x_k, \ldots, \gamma_{Pk}x_k)'$$

donde γ_{pk} se define como en (1.38). El total poblacional de $\mathbf{x}_k$ es por tanto un vector formado por las sumas conocidas de los P-grupos, $\sum_{\mathcal{U}_p} x_k$. Esta formulación de $\mathbf{x}_k$ ignora la información sobre los tamaños de los grupos, N_p, y esto puede equivaler a una pérdida no despreciable de información. Sin embargo, tomando $c_k = x_k^{-1}$ esta formulación conduce al estimador

$$\hat{Y}_{cal} = \sum_{p=1}^{p} (\sum_{\mathcal{U}_p} x_k) \frac{\bar{y}_{r_p}}{\bar{x}_{r_p}} = \hat{Y}_{sepra} \tag{1.43}$$

con $\bar{y}_{r_p} = \frac{1}{m_p} \sum_{r_p} y_k$ y $\bar{x}_{r_p} = \frac{1}{m_p} \sum_{r_p} \mathbf{x}_k$. Este estimador, $\hat{Y}_{sepra}$, es conocido como *estimador de razón separado*, es decir, se construye como suma de estimadores de razón para cada grupo.

Si la información auxiliar se obtiene a partir de los datos de la muestra s, el estimador toma la forma $\hat{Y}_{cal} = \frac{N}{n} \sum_{p=1}^{p} (\sum_{s_p} x_k) \frac{\bar{y}_{r_p}}{\bar{x}_{r_p}}$ que comúnmente se describe como *ponderación por grupos utilizando una variable tamaño*.

Para aprovechar las ventajas de la información completa, tamaño de N_p, así como los x-totales $\sum_{\mathcal{U}_p} x_k$, debemos formular el vector auxiliar como

$$\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk}, \gamma_{1k} x_k, \ldots, \gamma_{pk} x_k, \ldots, \gamma_{Pk} x_k)'.$$

Si tomamos $c_k = 1$ para todo k, se obtienen el estimador

$$\hat{Y}_{cal} = \sum_{p=1}^{p} N_p \{\bar{y}_{r_p} + (\bar{X}_p - \bar{x}_{r_p}) \hat{B}_p\} = \hat{Y}_{sepreg} \tag{1.44}$$

con $\bar{X}_p = \frac{1}{N_p} \sum_{\mathcal{U}_p} x_k$ y $B_p = \frac{Cov_{xyr_p}}{S^2_{xr_p}}$, donde $Cov_{xyr_p} = \frac{1}{m_p - 1} (\sum_{r_p} y_k x_k - \frac{1}{m_p} \sum_{r_p} y_k \sum_{r_p} x_k)$ y $S^2_{xr_p} = \frac{1}{m_p - 1} (\sum_{r_p} x_k^2 - \frac{1}{m_p} (\sum_{r_p} x_k)^2)$. El estimador $\hat{Y}_{sepreg}$ conoce como *estimador de regresión separado*.

Ejemplo 5: Clasificación doble

En este ejemplo presentamos el caso en el que la información auxiliar está compuesta de dos variables categóricas, aunque su razonamiento se puede extender a una clasificación de múltiples direcciones. Supongamos que tenemos P categorías como primer factor, por ejemplo, una clasificación geográfica, y H categorías en el segundo, por ejemplo, una clasificación socio-económica. Podemos representar la población $\mathcal{U}$ dividida en subconjuntos de $P \times H$ o en celdas, U_{ph}, con $p = 1, \ldots, P$, y $h = 1, \ldots, H$. Dependiendo de la información disponible sobre las celdas, son posibles varias formulaciones del $\mathbf{x}_k$.

Considerando la formulación del vector auxiliar

$$\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk}, \delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Pk})'$$

donde γ indica la primera clasificación con los P grupos y δ indica la segunda clasificación de los H grupos. γ_{pk} se define como en (1.38) y para, $h = 1, \ldots, H$, se define δ_{hk} como

$$\delta_{hk} = \begin{cases} 1 & \text{si} \quad k \in \text{grupo } h \\ 0 & \text{en otro caso} \end{cases} \tag{1.45}$$

Es fácil ver que esta formulación de $\mathbf{x}_k$ requiere del conocimiento de los $P + H$ totales marginales de los grupos, $N_{p.} = \sum_{h=1}^{H} N_{ph}$, con $p = 1, \ldots, P$, y $N_{.h} = \sum_{p=1}^{P} N_{ph}$, con $h = 1, \ldots, H$. Con esta formulación, podemos tratar tres situaciones:

i) Los $P \times H$ totales N_{ph}, con $p = 1, \ldots, P$ y $h = 1, \ldots, H$, son conocidos, pero se considera que el conjunto de $P + H$ totales marginales $N_{p.}$, con $p = 1, \ldots, P$, y $N_{.h}$, con $h = 1, \ldots, H$, contiene casi la misma información.

ii) Los $P \times H$ totales N_{ph}, con $p = 1, \ldots, P$ y $h = 1, \ldots, H$, son conocidos, pero algunos de ellos son muy pequeños o cero, una situación que surge frecuentemente en la práctica. Este problema podría causar una pérdida no despreciable de información auxiliar por lo que puede ser preferible simplemente utilizar los totales marginales.

iii) Los totales marginales $N_{p.}$ y $N_{.h}$, son conocidos, pero no los totales N_{ph} para cada celda. Un ejemplo de esta situación es cuando $N_{p.}$ y $N_{.h}$ se toman de dos registros diferentes.

Bajo esta formulación de información auxiliar $\mathbf{x}_k$, el enfoque del estimador de calibración $\hat{Y}_{cal}$ no tiene una forma simple, aunque computacionalmente es fácil de obtener utilizando el software existente.

1.4.4 Ejemplo de calibración lineal (GREG) en una encuesta preelectoral

En este ejemplo se analiza la estimación de voto al Partido Socialista Obrero Español (PSOE) y la participación en las elecciones generales de 2011 a partir de

los microdatos del barómetro preelectoral del CIS de octubre (2011)[5], última encuesta pre-electoral antes de las elecciones generales de 2011, utilizando la técnica de calibración bajo el modelo lineal (estimador de regresión generalizado). El objetivo pretendido es utilizar información auxiliar para encontrar alternativas para la estimación del total de las variables de interés, en nuestro caso, la intención de voto al PSOE y la participación en las elecciones generales del 2011.

Según la Ficha Técnica del estudio "Preelectoral elecciones Generales" (2011) del CIS[45], la selección de la muestra se realizó mediante un estudio bietápico. Primero se seleccionó los conglomerados o unidades primarias de muestreo (municipios) y posteriormente las unidades secundarias (individuos) de manera aleatoria. El tamaño muestral está compuesto por 6082 individuos, a los que se le realizó la encuesta completa, repartidos en 214 municipios. El informe metodológico de la encuesta pre-electoral y post-electoral elecciones Generales (2011)[25] del CIS, indica que las estimaciones se realizaron a partir de la variable de interés, sin tener en cuenta la información auxiliar, asignando a cada uno de los participantes de la encuesta un peso, según la comunidad autónoma de pertenencia. Estos pesos son calculados a partir de la metodología que se indica en la Figura 1.1

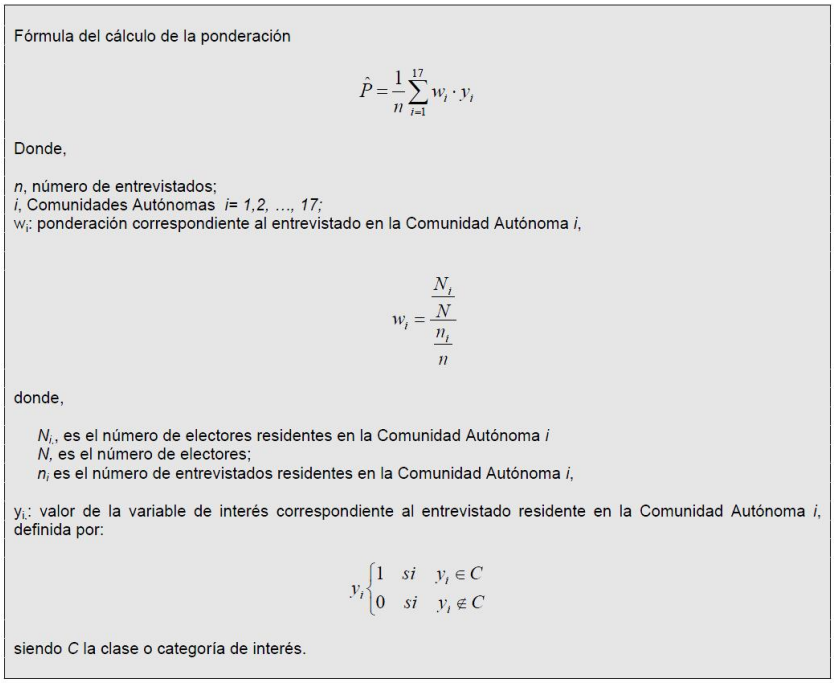
El CIS, en su estudio de octubre de 2011, previó una intención de voto al PSOE del 17,9 % y una participación en las elecciones del 71,8 %, respecto a un censo electoral de 35.776.115 españoles con derecho a voto.

Para este estudio, se han seleccionado dos variables de interés (variables principales o variables objeto de estudio) correspondientes al Barómetro Pre-electoral de Centro de Investigación Sociológicas de octubre de 2011[5]:

- y_1 Intención de voto al PSOE en las elecciones generales de 2011.
- y_2 Participación en las elecciones generales de 2011.

Como variables auxiliares se han tomado las variables:

- x_1 ¿Votó al PSOE en las elecciones generales de 2008?
- x_2 El encuestado está desempleado.
- x_3 El encuestado está jubilado.



Fórmula del cálculo de la ponderación

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{17} w_i \cdot y_i$$

Donde,

n, número de entrevistados;
i, Comunidades Autónomas *i= 1,2, …, 17;*
w_i: ponderación correspondiente al entrevistado en la Comunidad Autónoma *i*,

$$w_i = \frac{\frac{N_i}{N}}{\frac{n_i}{n}}$$

donde,

N_i, es el número de electores residentes en la Comunidad Autónoma *i*
N, es el número de electores;
n_i es el número de entrevistados residentes en la Comunidad Autónoma *i*,

y_i: valor de la variable de interés correspondiente al entrevistado residente en la Comunidad Autónoma *i*, definida por:

$$y_i \begin{cases} 1 & si \quad y_i \in C \\ 0 & si \quad y_i \notin C \end{cases}$$

siendo *C* la clase o categoría de interés.

Figura 1.1: Estimación encuesta CIS, octubre 2011

- x_4 El encuestado vive en un área metropolitana.

Los totales poblacionales de cada variable auxiliar se han tomado de la web del Instituto Nacional de Estadística (INE) [24] y de la web del Ministerio de Empleo y Seguridad Social[34].

Para este estudio de simulación se han seleccionado muestras de unidades primarias, en este caso conglomerados (municipios) de diferentes tamaños muestrales: m = 25, 50, 100, 150 y 200, seleccionando todos los encuestados (n, unidades secundarias) que componen el conglomerado seleccionado (es decir, los encuestados que viven en el determinado municipio). Para cada simulación, 1000 en total para cada tamaño de muestra, se ha calculado el estimador de calibración lineal con todas las variables auxiliares. Sólo se resume la media de las 1000 simulaciones para cada tamaño de muestra de municipios. Los resultados, que se exponen en la Tabla 1.4, dependiendo de la variable de interés elegida, muestran la insesgadez del estimador de calibración.

Tabla 1.4: Intención de voto al PSOE y participación en las elecciones generales de 2011

Resultados oficiales votos PSOE 2011		
6.973.880 (19,49 %)		
Votos al PSOE estimados por CIS		
17.9 %		
m	Promedio n 1000 simulaciones	Promedio estim. por cal. votos al PSOE en 2011 (1000 sim.)
25	712	6.271.250,57
50	1421	6.433.430,92
100	2845	6.644.325,60
150	4261	6.591.761,91
200	5684	6.684.868,86
Participación elecciones 2011		
25.241.538 (70.55 %)		
Participación estimada por CIS		
71.8 %		
m	Promedio n 1000 simulaciones	Promedio estim. por cal. participación
25	718,580	25.568.373,70
50	1.426,225	25.176.029,30
100	2.839,814	25.485.582,85
150	4.261,161	25.514.762,23
200	5.688,039	25.459.031,67

Para comprobar el comportamiento de las estimaciones según el vector auxiliar elegido se han realizado diferentes estimaciones para la muestra completa, 6082 individuos, eligiendo las posibles combinaciones de las cuatro variables auxiliares. Los resultados se muestran en la Tabla 1.5. De esta tabla surgen varias cuestiones, como qué variables auxiliares utilizar o si las variables auxiliares elegidas han de ser las mismas para todas las variables de interés, pues, como se observa, las estimaciones varían según qué información auxiliar se use. Y, ¿con qué criterio se eligen las variables? En el último capítulo se estudian distintos criterios para seleccionar las variables auxiliares disponibles, criterios objetivos basados en minimizar el sesgo de no respuesta, unos válidos para cualquier variables de interés y otros *ad hoc* para cada variable en estudio. Destacar que en esta encuesta, el porcentaje de no respuesta en la variable "participación" fue 4,1 %, mientras que en la variable

"voto PSOE" fue del 23,5 %.

Tabla 1.5: Estimación por calibración según v.a. $m = 214$, $n = 6082$

	Estimaciones (Porcentaje) votos PSOE	Estimaciones (Porcentaje) participación
CIS	17,9 %	71,8 %
Resultados reales votos	6.973.880 (19,49 %)	25.241.538 (70,55 %)
(x_1)	6.215.590 (17,37 %)	25.063.817 (70,06 %)
(x_2)	6.268.359 (17,52 %)	22.280.446 (62,27 %)
(x_3)	6.478.374 (18,10 %)	25.254.594 (70,59 %)
(x_4)	5.874.469 (16,42 %)	23.261.262 (65,01 %)
(x_1, x_2)	6.110.308 (17,08 %)	24.203.092 (67,65 %)
(x_1, x_3)	6.314.022 (17,64 %)	26.211.893 (73,26 %)
(x_1, x_4)	6.684.460 (18,68 %)	25.076.273 (70,09 %)
(x_2, x_3)	6.552.148 (18,31 %)	24.287.774 (67,88 %)
(x_2, x_4)	5.669.391 (15,84 %)	22.280.446 (62,27 %)
(x_3, x_4)	6.352.314 (17,76 %)	25.254.594 (70,59 %)
(x_1, x_2, x_3)	6.552.148 (18,31 %)	25.370.744 (70,91 %)
(x_1, x_2, x_4)	6.824.814 (19,07 %)	24.203.092 (67,65 %)
(x_1, x_3, x_4)	6.956.943 (19,44 %)	26.211.893 (65,01 %)
(x_2, x_3, x_4)	6.148.975 (17,18 %)	24.287.774 (67,89 %)
(x_1, x_2, x_3, x_4)	6.660.671 (18,62 %)	25.522.124 (71,34 %)

El uso de variables auxiliares adecuadas (en este caso solamente se ha utilizado cuatro variables sin establecer previamente la relación entre las variables) mejorará la precisión en la estimación cuanto más relacionadas estén las variables auxiliares con la variable de interés.

1.4.5 Extensiones de la calibración

Aunque en los capítulos siguientes veremos con más detalle la calibración con información auxiliar compuesta y la aplicación de la calibración para el tratamiento de la no respuesta, introducimos aquí brevemente extensiones de calibración para estimar otros parámetros más complejos y otros temas relacionados.

El método de calibración también se adapta a la estimación de otros parámetros que sean más complejos que el total poblacional, la media o una proporción. Un ejemplo es la estimación de cuantiles de la población o la estimación de funciones de totales. Más ejemplos de esta categoría, que no revisaremos aquí, son Théberge (1999)[63] para la estimación de parámetros bilineales, y Tracy, Singh y Arnab (2003)[70] para la calibración con respecto a los momentos de segundo orden.

Modelo-calibración

En el enfoque de calibración no existe un modelo explícito que relacione la variable de interés y las variables auxiliares. Wu y Sitter (2001)[76], Wu (2003)[75] y Montanari y Ranalli (2003[35], 2005[36]) definen un estimador de calibración que necesita la información auxiliar completa, (y permitir un uso más eficaz de los vectores $\mathbf{x}_k$, conocidos para todo $k \in \mathcal{U}$) lo que no es necesario en la calibración usual, donde con conocer el total, $\sum_{\mathcal{U}} \mathbf{x}_k$, es suficiente. Los pesos deben ser consistentes con el total de la población estimado a partir de las predicciones $\hat{y}_k$, obtenidas a través de la formulación de un modelo apropiado que relacione la variable de interés y las variables auxiliares.

Se considera un modelo del tipo $E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \theta)$ para $k \in \mathcal{U}$ con el que se estima el parámetro desconocido θ mediante $\hat{\theta}$, el cual proporciona los valores ajustados $\hat{y}_k = \hat{m}_k = \mu(\mathbf{x}_k, \hat{\theta})$, calculados con la ayuda de variables auxiliares $\mathbf{x}_k$, conocidas para todo $k \in \mathcal{U}$.

Suponiendo que el tamaño de la población N es conocido, los pesos del estimador de calibración $\hat{Y}_{mcal} = \sum_s w_k y_k$ se determinan minimizando la distancia chi-cuadrado, $\sum_s (w_k - d_k)^2/(2d_k q_k)$, para un q_k especificado, con pesos de diseño $d_k = 1/\pi_k$, sujeto a las ecuaciones de calibración:

$$\sum_s w_k = N;$$

$$\sum_s w_k \hat{y}_k = \sum_{\mathcal{U}} \hat{y}_k.$$

Tomando $q_k = 1$ para todo k, el estimador modelo-calibrado se puede escribir como

$$\hat{Y}_{mcal} = N\{\bar{y}_{s,d} + (\bar{\hat{y}}_{\mathcal{U}} - \bar{\hat{y}}_{s,d})\tilde{B}_{s,d}\}$$

donde

$$\bar{y}_{s,d} = \sum_s d_k y_k / \sum_s d_k,$$

$$\bar{\hat{y}}_{s,d} = \sum_s d_k \hat{y}_k / \sum_s d_k,$$

$$\tilde{B}_{s,d} = \frac{\sum_s d_k(\hat{y}_k - \bar{\hat{y}}_{s,d})y_k}{\left(\sum_s d_k(\hat{y}_k - \bar{\hat{y}}_{s,d})\right)^2},$$

donde el coeficiente de regresión $\tilde{B}_{s,d}$ representa la relación entre los valores observados y y los valores predichos $\hat{y}$.

Särndal (2007)[53] explica como un estimador GREG basado en un modelo no lineal es por lo general menos eficiente que $\hat{Y}_{mcal}$, aunque es posible modificarlo para que tenga en cuenta también la información dada por que el tamaño de la población sea conocido. Por otra parte, si se compara con el estimador de calibración usual (libre de modelo), el estimador modelo-calibrado $\hat{Y}_{mcal}$ puede ser más preciso, aunque ello implica una pérdida de consistencia en relación con el total de la población conocido, $\sum_{\mathcal{U}} x_k$. En este caso, los valores y en $\hat{Y}_{mcal}$ están ponderados linealmente, pero los pesos obtenidos también van a depender de los valores y.

Montanari y Ranalli (2005)[36] muestran un estudio con varias poblaciones creadas artificialmente, en el que comparan el estimador $\hat{Y}_{mcal}$ y el estimador GREG no lineal con el modelo auxiliar $y_k = \mu_k + \varepsilon_k$ ajustado mediante una regresión no paramétrica (suavizado polinomial) de las predicciones $\hat{y} = \hat{\mu}_k$, para $k \in \mathcal{U}$. Con este tipo de ajuste del modelo, la predicciones $\hat{y} = \hat{\mu}_k$ son altamente precisas.

En resumen, en el estimador modelo-calibrado se requiere información auxiliar completa y los pesos w_k dependen de los valores y, lo que implica una pérdida de la propiedad de usos múltiples (para cualquier y-variable).

La calibración para la estimación de cuantiles

La mediana y otros cuantiles de una población finita son importantes medidas descriptivas, sobre todo en las encuestas económicas. Para estimar los cuantiles, lo primero que debemos hacer es estimar la función de distribución de la población finita. Artículos recientes han conseguido un enfoque de la calibración para los mismos fines, incluyendo Kovaĉević (1997)[28], Wu y Sitter (2001)[76], Ren (2002)[49], Tillé (2002)[67], Harms (2003)[20], Harms y Duchesne (2006)[21] y Rueda et al. (2007)[51]. El carácter complejo de la función de distribución de la población finita hace que se den ciertos inconvenientes, que son resueltos por los distintos autores de diferentes maneras.

Sea $\Delta(\ )$ la función Heaviside, definida para cada real z tal que

$$\Delta(\ ) = \begin{cases} 1 \ \text{ si } \ z \geq 0 \\ 0 \ \text{ si } \ z < 0 \end{cases}$$

La función de distribución poblacional para la variable en estudio y es

$$F_y(t) = \frac{1}{N} \sum_{\mathcal{U}} \Delta(t - y_k).$$

El cuantil α de una población finita está definido como $Q_{y\alpha} = inf\{t/F_y(t) \geq \alpha\}$. Sean las variables auxiliares x_j, con valores x_{jk}, que tienen como función de distribución $F_{\mathbf{x}_j}(t) = \frac{1}{N} \sum_{\mathcal{U}} \Delta(t - \mathbf{x}_{jk})$ con el cuantil α conocido denotado por $Q_{x_j\alpha}, \ \ j = 1, 2, \cdots, J$.

El estimador de Horvitz Thompson de $F_y(t)$ basado en los pesos del diseño, $d_k = \frac{1}{\pi_k}$, es

$$\hat{F}_y(t) = \frac{1}{\sum_s d_k} \sum_s d_k \Delta(t - y_k).$$

Harms y Duchesne (2006)[21] proponen usar la información para aplicar la calibración dada por el tamaño de la población N y los cuantiles de las variables auxiliares $Q_{x_j\alpha}$, para $j = 1, 2, \ldots, J$. El estimador de calibración de $F_y(t)$ tiene la forma $\hat{F}_{ycal}(t) = \frac{1}{\sum_s w_k} \sum_s w_k \Delta(t - y_k)$, donde los pesos w_k se obtienen minimizando la distancia Chi-cuadrado $\frac{\sum_s (w_k - d_k)^2}{2 d_k q_k}$ para un especifico q_k, sujetos a la ecuación de calibración:

$$\begin{cases} \sum_s w_k = N \\ \hat{Q}_{x_j cal,\alpha} = Q_{x_j\alpha}; \ \ j = 1, \ldots, J \end{cases}$$

A partir de $\hat{F}_{ycal}(t)$ se obtiene el estimador para los cuantiles como $\hat{Q}_{y\alpha} = \text{ínf}\{t/\hat{F}_{ycal}(t) \geq \alpha\}$. Aquí no es necesaria la información auxiliar completa.

Otro método para la estimación de cuantiles puede verse en Rueda et al. (2007)[51]. Los autores presentan un estimador modelo-calibrado, en el que se calibra con respecto al total en la población de los valores predichos y y para lo que es necesaria la información auxiliar completa.

Usando los valores conocidos, $\mathbf{x}_k$, se calcula primero una predicción lineal $\hat{y}_k = \hat{\beta}\mathbf{x}_k$ para cada $k \in U$, con $\hat{\beta} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_s d_k q_k \mathbf{x}_k y_k\right)$ donde $d_k = 1/\pi_k$ y q_k especifica el factor de escala. Los pesos w_k se obtienen minimizando

la distancia chi-cuadrado sujeta a las ecuaciones de calibración indicadas desde el punto de vista de las predicciones, a fin de tener consistencia en los J puntos $t_j,\ j = 1, ..., J$, obtenidos arbitrariamente:

$$\frac{1}{N}\sum_s w_k \Delta(t_j - \hat{y}_k) = F_{\hat{y}}(t_j),\ \ j = 1, \ldots, J,$$

donde $F_{\hat{y}}(t_j)$ es la función de distribución de las predicciones y_k de la población finita, evaluada en t_j. Se sugiere que un número relativamente pequeño de los puntos seleccionados arbitrariamente t_j puede ser suficiente, por ejemplo menos de 10. Una vez que los pesos w_k se determinan, la estimación de los cuantiles se obtiene invirtiendo $\hat{F}_{ycal}(t) = \frac{1}{N}\sum_s w_k \Delta(t_j - \hat{y}_k)$.

Aunque ambos métodos mencionados ofrecen un estimador insesgado bajo el diseño, el de Harms y Duchesne (2006)[21] da unos pesos válidos para cualquier variable y mientras que el método de Rueda et al. (2007)[51] requiere de un nuevo conjunto de pesos para cada variable y.

Calibración de otros parámetros complejos

Krapavickaitè y Plikusas (2005)[29] y Plikusas (2006)[44] examinan la estimación por el método de calibración de ciertas funciones de totales de la población. Un ejemplo sencillo es la estimación de una razón de dos totales, $R = \sum_{\mathcal{U}} y_{1k} / \sum_{\mathcal{U}} y_{2k}$ donde y_{1k} e y_{2k} son los valores para el elemento k de las variables y_1 y y_2 respectivamente. Por ejemplo, la función de distribución $F_y(t) = \frac{1}{N}\sum_{\mathcal{U}} \Delta(t - y_k)$ es también del tipo razón cuando se toma $y_{2k} = 1$, y con $N = \sum_{\mathcal{U}} 1$ como el total del denominador.

Estos autores examinan el estimador de calibración $\hat{R}_{cal} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Los pesos w_k, comunes al numerador y al denominador, se determinan mediante la calibración de la información auxiliar de una variable auxiliar $\mathbf{x}_{1k}$ con cada y_{1k}, y otra $\mathbf{x}_{2k}$, con cada y_{2k}, y el cociente de los totales $R_0 = \sum_{\mathcal{U}} x_{1k} / \sum_{\mathcal{U}} x_{2k}$ es un valor conocido.

La ecuación de calibración propuesta es $\sum_{\mathcal{U}} w_k e_k = 0$, donde $e_k = \mathbf{x}_{1k} - R_0 \mathbf{x}_{2k}$ y como $\sum_U e_k = 0$, los pesos obtenidos minimizando la distancia chi-cuadrado, son

$$w_k = d_k \left\{ 1 - \left(\sum_s d_k e_k \right) \left(\sum_s d_k e_k^2 \right)^{-1} e_k \right\}.$$

Krapavickaitè y Plikusas (2005)[29] y Plikusas (2006)[44] muestran que este estimador es insesgado bajo el diseño y tiene menor varianza comparado con otros estimadores que usan la misma información auxiliar.

"Sin información auxiliar no hay enfoque de calibración..."

Särndal

CAPÍTULO

2

Información auxiliar en encuestas con diseños muestrales complejos

En los últimos años, la estimación por calibración se ha convertido en un importante campo de investigación en el muestreo de encuestas. En este capítulo se revisa algunas perspectivas del uso de la calibración en presencia de información compuesta, tales como: información proveniente de diseños de muestreo en dos fases o en dos etapas. Haremos hincapié en esta última opción, describiendo dos casos particulares que combinan la información disponible en ambas etapas: la integración de pesos (Estevao y Särndal, 2006)[16]) y, como caso particular, el estimador de Lemaître y Dufour (1987)[32].

Posteriormente se describe el estimador planteado en este trabajo, al que denominamos estimador de contracción, y que proponemos como alternativa a los descritos anteriormente. El capítulo se finaliza con un estudio de simulación, con datos reales provenientes del estudio PISA 2006 y de la encuesta de presupuestos familiares (EPA) realizada por el INE, que evalúa el comportamiento empírico de tres variantes del estimador de contracción y del estimador descrito por Estevao y Särndal para un muestreo aleatorio simple y un muestreo de Midzuno.

2.1 Estimación por calibración en presencia de información compuesta

2.1.1 Información compuesta para diseños de muestreo en dos fases

En un muestreo en dos fases se pretende obtener una muestra relativamente grande de elementos (primera fase) para obtener información auxiliar, para posteriormente seleccionar una submuestra (segunda fase) de la muestra de la primera fase, que permita construir un estimador más eficiente del parámetro de interés.

Sea una población $\mathcal{U} = \{1, \ldots, N\}$ de la que se extrae una muestra probabilística s_1 con probabilidades de inclusión π_{1k}, conocidas para $k \in \mathcal{U}$. Para esta primera fase se conocen los pesos del diseño d_{1k}, con $k \in \mathcal{U}$, y algunas variables auxiliares $\mathbf{x}_{1k}$, para $k \in s_1$. Posteriormente se extrae una submuestra aleatoria s_2 de s_1 con probabilidades de inclusión π_{2k} (conocidas), condicionada a s_1. En este caso, los pesos del diseño de la segunda fase (fase condicional) son $d_{2k} = 1/\pi_{2k}$, con $k \in s_1$. Denotando por $d_k = d_{1k}d_{2k}$ al peso del diseño general de la unidad k, el objetivo es utilizar la información auxiliar para encontrar estimaciones del total Y, alternativas a estimadores en dos fases del tipo

$$\hat{Y}_{ht} = \sum_{s_2} d_k y_k, \tag{2.1}$$

con y_k los valores de la variable de interés, observados para todos los $k \in s_2$.

La complejidad en este tipo de muestreos surge cuando se involucran dos vectores auxiliares, $\mathbf{x}_1$ y $\mathbf{x}_2$, para cada una de las muestras. Sea $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$ los respectivos valores para la unidad k y J_1 y J_2 sus respectivas dimensiones. La información auxiliar, en este caso, se puede presentar de tres formas:

i) Se conoce el total poblacional, $\sum_{\mathcal{U}} \mathbf{x}_{1k}$.

ii) Para cada $k \in s_1$, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$ son conocidos.

iii) Para cada $k \in s_2$, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$ son conocidos.

Los dos vectores $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$ se diferencian en que el total $\sum_{\mathcal{U}} \mathbf{x}_{2k}$ no es conocido.

Como es lógico, si se cumple la propiedad *ii*), se cumple la *iii*). Si $\mathbf{x}_{1k}$ es conocido para todos los $k \in \mathcal{U}$, entonces *i*), *ii*) y *iii*) se cumplen para todo k. Un caso particular ocurre cuando $\mathbf{x}_{1k} = 1$ para todo k, lo que implica que el único tipo de información disponible sea el tamaño de la población, N.

2.1.2 Estimación de calibración para el muestreo en dos fases

Los tipos de información auxiliar descritos en la sección anterior proporcionan una base para el cálculo de los pesos calibrados para el estimador de calibración directo en una sola fase

$$\hat{Y}_{cal} = \sum_{s_2} w_k y_k. \tag{2.2}$$

Alternativamente se puede calcular los pesos en dos fases donde en la primera fase se calculan pesos intermedios para su uso en la segunda fase. Cada paso requiere de unos pesos iniciales y un vector auxiliar. A partir de estos construimos un vector auxiliar $\mathbf{x}_k$ de dimensión $J_1 + J_2$ que combina la información de $\mathbf{x}_{1k}$ y de $\mathbf{x}_{2k}$, $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}$.

Algunos métodos alternativos son:

a) *Calibración en un solo paso*: A partir de $d_k = d_{1k} d_{2k}$, se calcula directamente los pesos finales w_k para $k \in s_2$, satisfaciendo $\sum_{s_2} w_k \mathbf{x}_k = \begin{pmatrix} \sum_{\mathcal{U}} \mathbf{x}_{1k} \\ \sum_{s_1} d_{1k} \mathbf{x}_{2k} \end{pmatrix}$.

b) *Calibración en dos fases, de arriba abajo*: En la primera fase, a partir de d_{1k} se calculan los pesos intermedios w_{1k}, para $k \in s_1$, satisfaciendo la ecuación $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_{\mathcal{U}} \mathbf{x}_{1k}$. En la segunda fase, a partir de $d_k = d_{1k} d_{2k}$ calculamos los pesos finales w_k, para $k \in s_2$, sujetos a la ecuación $\sum_{s_2} w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k$, usando los pesos w_{1k} de la primera fase. Los pesos finales satisfacen $\sum_{s_2} w_k x_k = \begin{pmatrix} \sum_{\mathcal{U}} \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k} \mathbf{x}_{2k} \end{pmatrix}$.

c) *Calibración en dos fases, de abajo arriba*: En la primera fase, a partir de $d_k = d_{1k} d_{2k}$, se calculan pesos intermedios w_{0k}, para $k \in s_2$, sujetos a la

ecuación $\sum_{s_2} w_{0k}\mathbf{x}_{2k} = \sum_{s_1} d_{1k}\mathbf{x}_{2k}$. En la segunda fase, a partir de los pesos w_{0k} obtenidos en la primera fase, se calculan los pesos finales w_k, para $k \in s_2$, de manera que satisfacen la condición $\sum_{s_2} w_k\mathbf{x}_{2k} = \begin{pmatrix} \sum_{\mathcal{U}} \mathbf{x}_{1k} \\ \sum_{s1} d_{1k}\mathbf{x}_{2k} \end{pmatrix}$. Una alternativa para la segunda fase es empezar por w_{0k} y calibrar para que se cumpla sólo la condición $\sum_{s_2} w_k\mathbf{x}_{1k} = \sum_{\mathcal{U}} \mathbf{x}_{1k}$, aunque estos pesos no satisfacen la condición $\sum_{s_2} w_k\mathbf{x}_{2k} = \sum_{s_1} d_{1k}\mathbf{x}_{2k}$.

Estevao y Särdal (2006)[16] indican que los métodos *a*), *b*) y *c*) son procedimientos correctos que, aunque cumplen los requisitos para la información auxiliar *i*), *ii*) y *iii*) descritos anteriormente, los pesos resultantes para cada uno de los procedimientos no son generalmente los mismos. De los tres métodos, el *b*) es considerado el más eficiente ya que implica más información que los casos *a*) y *c*), ya que a diferencia de éstos, requiere de los valores individuales $\mathbf{x}_{1k}$ para $k \in s_1$. Aunque el hecho de que los casos *a*) y *c*) hagan uso de menos información, no tiene que ser causa de una pérdida de eficiencia, ya que ésta dependerá de factores tales como el tamaño de la muestra en cada fase o la relación entre la variable de interés y_k y las variables auxiliares $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$.

En el caso *b*), los pesos intermedios, para $k \in s_1$, se calculan mediante la ecuación:

$$w_{1k} = d_{1k}(1 + \lambda'_{s_1}\mathbf{z}_{1k}) \tag{2.3}$$

con

$$\lambda'_{s_1} = (\sum_{\mathcal{U}} \mathbf{x}_{1k} - \sum_{s_1} d_{1k}\mathbf{x}_{1k})'(\sum_{s_1} d_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1} \tag{2.4}$$

para algún vector $\mathbf{z}_{1k}$. Estos pesos, w_{1k}, se utilizan para calcular los pesos calibrados finales, para $k \in s_2$, mediante:

$$w_k = d_k(1 + \lambda'_{s_2}\mathbf{z}_k), \tag{2.5}$$

con

$$\lambda'_{s_2} = (\sum_{s_1} w_{1k}\mathbf{x}_k - \sum_{s_2} d_k\mathbf{x}_k)'(\sum_{s_2} d_k\mathbf{z}_k\mathbf{x}'_k)^{-1}. \tag{2.6}$$

Otro aspecto importante es el de aproximar la varianza de $\hat{Y}_{cal}$, definida en (2.2); en este caso se utilizará el método de linealización automática. En primer lugar, se expresa w_k en función de w_{1k}. Simplificando y reordenando los términos, se llega a una expresión que incluye dos vectores, definidos como

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{\mathbf{B}}_1 \\ \hat{\mathbf{B}}_2 \end{pmatrix} = (\sum_{s_2} d_k \mathbf{z}_k \mathbf{x}_k')^{-1} (\sum_{s_2} d_k \mathbf{z}_k y_k), \tag{2.7}$$

y

$$\hat{\mathbf{B}}^* = (\sum_{s_1} d_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}')^{-1} (\sum_{s_1} d_{1k} \mathbf{z}_{1k} (x_k' \hat{\mathbf{B}})). \tag{2.8}$$

Tomamos los equivalentes poblacionales $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} = (\sum_{\mathcal{U}} \mathbf{z}_k \mathbf{x}_k')^{-1} (\sum_{\mathcal{U}} \mathbf{z}_k y_k)$ y $\mathbf{B}^* = (\sum_U d_k \mathbf{z}_{1k} \mathbf{x}_{1k}')^{-1} (\sum_{\mathcal{U}} \mathbf{z}_{1k} (x_k' \mathbf{B}))$ respectivamente, y simplificando, se define

$$\hat{Y}_{cal} = \sum_{s_2} w_k y_k = \hat{Y}_{cal.lin} + \mathbf{R}, \tag{2.9}$$

donde $\mathbf{R}$ es un término de orden inferior y $\hat{Y}_{cal.lin}$ es un estadístico lineal.

$$\hat{Y}_{cal.lin} = \sum_{s_2} d_k e_{2k} + \sum_{s_1} d_{1k} e_{1k}^* + \sum_{\mathcal{U}} \mathbf{x}_{1k}' \mathbf{B}^*, \tag{2.10}$$

con $e_{2k} = y_k - \mathbf{x}_k' \mathbf{B} = y_k - \mathbf{x}_{1k}' \mathbf{B}_1 - \mathbf{x}_{2k}' \mathbf{B}_2$ y $e_{1k}^* = \mathbf{x}_k' \mathbf{B} - \mathbf{x}_{1k}' \mathbf{B}^*$.

El término e_{1k}^* expresa los residuos de una regresión de $\mathbf{x}_k' \mathbf{B}$ sobre $\mathbf{x}_{1k}$. El término de orden inferior $\mathbf{R}$ viene definido como:

$$\mathbf{R} = -(\sum_{s_2} d_k \mathbf{x}_k - \sum_{s_1} d_{1k} \mathbf{x}_k)' (\hat{\mathbf{B}} - \mathbf{B}) - (\sum_{s_1} d_{1k} \mathbf{x}_{1k} - \sum_{\mathcal{U}} \mathbf{x}_{1k})' (\hat{\mathbf{B}}^* - \mathbf{B}^*). \tag{2.11}$$

Calculamos la varianza a partir de $\hat{Y}_{cal.lin}$, tomando $Var(\hat{Y}_{cal}) \approx Var(\hat{Y}_{cal.lin}) = V_1(E_c) + E_1(V_c)$, donde $E_c = \sum_{s_1} d_{1k} e_{1k}$, con $e_{1k} = e_{2k} + e_{1k}^* = y_k - \mathbf{x}_{1k}' \mathbf{B}^*$, es la esperanza condicional y V_c la varianza condicional de $\sum_{s_2} d_k e_{2k}$, dado s_1.

Para la estimación de la varianza, utilizamos $\hat{\mathbf{B}}$ y $\hat{\mathbf{B}}^*$ en lugar de $\mathbf{B}$ y $\mathbf{B}^*$. $V_1(E_c)$ se ajusta para $\mathbf{x}_{1k}$ pero no para $\mathbf{x}_{2k}$, por lo que su valor tiende a ser mayor que el

del segundo componente $E_1(V_c)$, ya que éste se ajusta para $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$ debido a que $\mathbf{x}_{2k}$ sólo proporciona información a nivel de la primera fase de la muestra y no ha de provocar un impacto en los residuos e_{1k}.

En los casos *a*) y *c*) la linealización automática de $\hat{Y}_{cal}$ produce una varianza de la forma $V_1(E_c) + E_1(V_c)$, con residuos e_{1k} para el primer componente y e_{2k} para el segundo. Aunque e_{1k} y e_{2k} no tienen la misma apariencia para cada método *a*), *b*) y *c*), existe un patrón común en todos ellos, la influencia de la información $\mathbf{x}_{1k}$ se extrae de e_{1k}, mientras que la influencia de ambas informaciones, x_{1k} y x_{2k} se extrae de e_{2k}. El resultado es que las varianzas de los estimadores $\hat{Y}_{cal}$ para los métodos *a*), *b*) y *c*), aunque no son iguales, no difieren demasiado. Pero teniendo en cuenta que dependiendo de la relación entre las variables y_k, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$, las diferencias a veces puede ser significativas.

Una cuestión que se plantea es cómo elegir los vectores $\mathbf{z}_{1k}$ y $\mathbf{z}_k$ de forma que minimicen la varianza. Estas opciones son tratadas en la literatura, aunque habitualmente la opción empleada es la más simple, tomar $\mathbf{z}_{1k} = \mathbf{x}_{1k}$ y $\mathbf{z}_k = \mathbf{x}_k$.

2.1.3 Información compuesta en diseños de muestreo en dos etapas

En un muestreo de conglomerados la población se encuentra dividida de manera natural en grupos que, se suponen, contienen toda la variabilidad de la población y por tanto representan fielmente la característica a estudiar para la realización del estudio. Dentro de los grupos (conglomerados) seleccionados se ubicarán las unidades elementales, por ejemplo las personas a encuestar, y podría aplicársele el instrumento de medición a todas las unidades (todos los miembros del conglomerado) o sólo a algunos de ellos seleccionados al azar, lo que simplificaría la recogida de información muestral. En forma resumida se define este método como muestreo de dos etapas o bietápico. La configuración tradicional de este método (grupos incluidos en la muestra en la primera etapa, de la que se extrae elementos de submuestras dentro de los grupos seleccionados en la segunda etapa) tiene en común con el muestreo en dos fases que la información auxiliar puede aparecer de diferentes maneras:

a) Información a nivel conglomerado.

***b*)** Información a nivel elemento, para todos los conglomerados.

***c*)** Información a nivel elemento, para los conglomerados seleccionados.

La calibración puede tratar eficazmente encuestas donde la información auxiliar existente puede encontrarse en diferentes niveles, y en el caso de muestreos en dos etapas donde la información puede existir tanto para la primera etapa de las unidades de muestreo (conglomerados) como para la segunda etapa (unidades de muestreo) no es una excepción.

En este tipo de encuestas la información auxiliar puede existir tanto "a nivel poblacional" (por ejemplo, conocer los totales de la población), como "a nivel de la muestra" (valores de la variable auxiliar para todos aquellos elementos incluidos en la muestra), también pueden estar disponible tanto para los conglomerados como para las unidades que forman parte de estos conglomerados, lo que crea una cierta complejidad en la información auxiliar. Un ejemplo de muestreo bietápico, siendo éste el muestreo más usado por las agencias de estadística y los organismos oficiales, es el muestreo de hogares. Gran número de encuestas se basan en la selección de hogares y dentro de cada hogar en la selección de uno, más de uno o de todos sus miembros. En los muestreos de hogares se recoge información a nivel persona, segunda etapa (ingresos, nivel de estudios, situación de desempleo, etc.) e información a nivel hogar, primera etapa (si el hogar dispone de calefacción, si es urbano o rural, ingresos del hogar, etc.) para posteriormente estimar parámetros de estas características a ambos niveles.

El interés para este tipo de diseños es abordar la incorporación de la información auxiliar compleja en el proceso de estimación, mediante la calibración, tratando de disminuir el error de estimación, de forma que las estimaciones que se obtengan a nivel unidad sean coherentes con las obtenidas a nivel conglomerado.

2.1.4 Estimación por calibración para muestreos en dos etapas

Supongamos que la extracción de una muestra de k unidades se realiza mediante la selección en dos etapas de elementos de una población finita $\mathcal{U} = \{1, \ldots, N\}$ agrupada en conglomerados. Este diseño incluye muestras de dos poblaciones diferentes:

i) Población de conglomerados (unidades primera etapa), $\mathcal{U}_i = \{1, \ldots, i, \ldots, N_I\}$.

ii) Población de unidades (unidades de segunda etapa), $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, que es la unión de todas las unidades N_I en todos los conglomerados $\mathcal{U}_i$, con $i \in \mathcal{U}$.

En primer lugar, se extrae de $\mathcal{U}_I$ una muestra de conglomerados s_I, con probabilidades de inclusión π_{Ii}, con $i \in U_i$, cuyos pesos de diseño para la primera etapa son $d_{Ii} = 1/\pi_{Ii}$, con $i \in \mathcal{U}_i$. En una segunda etapa, se extrae una muestra de las unidades dentro de cada uno de los grupos seleccionados. Para $i \in s_I$, enumeradas las unidades en $\mathcal{U}_i$, elegimos una muestra s_i de éstas, con probabilidades de inclusión $\pi_{k|i}$, con $k \in s_i$. La segunda etapa tiene como pesos de diseño $d_{k|i} = 1/\pi_{k|i}$ con $k \in s_i$. Se define el peso del diseño general como $d_k = d_{Ii} d_{I|i}$, para la unidad k, y como muestra de unidades secundarias $s = \bigcup_{i \in s_I}$.

Denotando la variable de interés, definida a nivel unidad, como y_u, cuyo valor para la unidad k es $y_{(u)k}$, observada para cada $k \in s$. El propósito es estimar el total de la variable de interés, $Y = \sum_{\mathcal{U}} y_{(u)k}$, con el uso de información auxiliar, para obtener mejores estimaciones que las producidas por el estimador simple en dos etapas $\hat{Y}_{ht} = \sum_s d_k y_{(u)k} = \sum_{s_1} d_{Ii} (\sum_{s_i} d_{k|i} y_{(u)k})$.

También se dispone de una variable de interés para los conglomerados, a la que denotaremos $y_{(c)}$, con valores $y_{(c)i}$ observados para cada conglomerado $i \in s_I$. En el caso del conglomerado, lo que se pretende es realizar una estimación del total $Y_I = \sum_{\mathcal{U}_I} y_{(c)i}$ con la ayuda de un uso eficiente de la información auxiliar.

Para esta formulación general de muestreo en dos etapas, podemos encontrar casos en los que los conglomerados pueden ser relativamente grandes, como cuando se muestrea distritos censales dentro de ciudades en la primera etapa y hogares en el segunda, o conglomerados relativamente pequeños, como cuando se seleccionan familias en la primera etapa y los miembros adultos del hogar en la segunda. Por ejemplo, cuando se toman hogares como conglomerados, la encuesta se pueden orientar al estudio de la variable $y_{(c)}$ = "ingreso familiar" con valores $y_{(c)i}$ para el hogar i, y al mismo tiempo, la variable $y_{(u)}$ = "situación laboral" con valor $y_{(u)k} = 1$ si persona k está desempleada y el valor $y_{(u)k} = 0$ si no lo está. En este caso la información auxiliar puede existir tanto para las unidades como para los conglomerados.

Denotamos por $\mathbf{x}_{(u)k}$ el valor del vector auxiliar asociado a la unidad k y por $\mathbf{x}_{(c)i}$ el valor del vector auxiliar asociado con el conglomerado i. En un muestreo en dos etapas podemos encontrar la información auxiliar de diferentes formas:

i) Se conoce el total poblacional de las variables auxiliares de los conglomerados, $\sum_{\mathcal{U}_I} \mathbf{x}_{(c)i}$.

ii) Para cada $i \in s_I$, los valores $\mathbf{x}_{(c)i}$ del vector auxiliar a nivel conglomerado son conocidos.

iii) Se conoce el total de las unidades poblacionales, $\sum_{\mathcal{U}} \mathbf{x}_{(u)k}$.

iv) Para cada $k \in s$, los valores del vector de unidades, $\mathbf{x}_{(u)k}$, son conocidos.

Si $\mathbf{x}_{(c)i}$ es conocido para todo $i \in \mathcal{U}$, entonces las categorías *i*) y *ii*) siempre se cumplen. Esto sucede, por ejemplo, en muestreo de áreas donde cada grupo es una entidad geográfica de las que tenemos mediciones auxiliares útiles, tales como la superficie y/o el número aproximado de habitantes. En cuanto a la información a nivel unidad, $\mathbf{x}_{(u)k}$, es menos probable que esté disponible para cada $k \in \mathcal{U}$, ya que es precisamente la ausencia de dicha información, la que obliga a realizar un muestreo en dos etapas en lugar de un muestreo directo de las unidades presentes. No ocurre lo mismo con las categorías *iii*) y *iv*), ya que éstas se cumplen si $\mathbf{x}_{(u)_k}$ es conocido para todas las unidades de la muestra y el total se puede importar desde una fuente precisa (censos o proyecciones censales).

El objetivo de esta sección es examinar los estimadores de calibración a partir de los diferentes tipos de información descritos en las cuatro categorías anteriores.

A nivel conglomerado:

$$\hat{Y}_{I.cal} = \sum_{s_I} w_{Ii} y_{(c)i}, \tag{2.12}$$

y a nivel unidad:

$$\hat{Y}_{cal} = \sum_{s} w_k y_{(u)k}, \tag{2.13}$$

con pesos a nivel conglomerado, w_{Ii}, que satisfacen la ecuación

$$\sum_{s_I} w_{Ii}\mathbf{x}_{(c)i} = \sum_{\mathcal{U}_I} \mathbf{x}_{(c)i} \tag{2.14}$$

y pesos a nivel unidad, w_k, que satisfacen la ecuación

$$\sum_{s} w_k\mathbf{x}_{(u)k} = \sum_{\mathcal{U}} \mathbf{x}_{(u)k}. \tag{2.15}$$

Para calibrar la información combinada de totales a nivel conglomerado y a nivel unidad, es necesario el vector "apilado" de totales

$$\mathbf{X} = \begin{pmatrix} \sum_{\mathcal{U}_I} \mathbf{x}_{(c)i} \\ \sum_{\mathcal{U}} \mathbf{x}_{(u)k} \end{pmatrix}. \tag{2.16}$$

2.1.5 Integración de pesos

El defecto de usar la calibración a nivel individuo es que los pesos se diferenciarán, por lo general, para cada unidad dentro de un mismo conglomerado. Por ello, para calibrar en la información auxiliar combinada, se requiere algún tipo de "ponderación integrada" en el que se imponga una relación conveniente entre el peso a nivel conglomerado, w_{Ii}, y los pesos de las unidades seleccionadas, w_k. Este puede causar inconsistencias cuando se estiman las variables de estudio a nivel conglomerado, puesto que no hay un peso común que represente al conglomerado. Este tema es tratado por diferentes autores, destacando el trabajo de Estevao y Särndal (2006)[16] y el de Lemaître y Dufour[32], quienes discuten diferentes vías para calcular los pesos finales que se usan en la estimación:

i) $\sum_{s_i} w_k = N_i w_{Ii}$ para cada $i \in s_I$, donde N_i es el tamaño del conglomerado conocido.

ii) $w_k = d_{k|i} w_{Ii}$ para la unidad k seleccionada en el conglomerado $i \in S_I$.

Para las dos opciones de integración de pesos se satisfacen las ecuaciones (2.14) y (2.15).

En el caso *i*) el número estimado de unidades en cualquier conglomerado es el mismo, se use el peso del conglomerado o el peso de la unidad para estimarlo. Para muestreo de conglomerados en una etapa (todas las unidades en el conglomerado

son observadas), el peso del conglomerado es la media de los pesos de las unidades que lo componen.

En el caso *ii*) se preserva los pesos condicionales en el sentido que $w_k/w_{Ii} = d_{k|i}$ imita la propiedad $d_k/d_{Ii} = d_{k|i}$.

Una variante de *ii*) es el muestreo en una etapa (Lemaître y Dufour (1987)[32], Andersson (1997)[1] y Nieuwenbroek (1993)[38]) en que todas las k unidades del conglomerado i son observadas de forma que $d_{k|i} = 1$ y por tanto *ii*) implica $w_k = w_{Ii}$. El peso de la unidad w_k será el mismo para todas las unidades dentro del conglomerado e igual al peso del conglomerado w_{Ii}. Es práctico tener todas las unidades dentro de un conglomerado con el mismo peso para estimar variables a nivel unidad y usar este peso para estimar variables a nivel conglomerado. En la aproximación de Lemaître y Dufour, aunque difiere de esta metodología, los pesos son obtenidos como un caso especial, como veremos en el apartado 2.1.6.

Existen diferentes alternativas para computar los pesos de los estimadores de calibración, $\widehat{Y}_{I.cal}$ y $\widehat{Y}_{cal}$ de forma que se satisfagan las ecuaciones de calibración (2.14) y (2.15).

a) Calibración sin integración

Partiendo de d_{Ii}, se calculan los pesos de los conglomerados w_{Ii}, para $i \in s_i$, con la ecuación de calibración (2.14). Independientemente, en una segunda calibración, partiendo de $d_k = d_{Ii}d_{k|i}$, se calculan los pesos de las unidades w_k, $k \in s$, con la ecuación de calibración (2.15).

b) Calibración en un paso con la opción de integración *i*)

En (2.14) se reemplaza los pesos w_{Ii} por $\sum_{s_i} w_k/N_i$, por lo que la ecuación dependerá de los pesos w_k. Se asigna el valor $\mathbf{x}_{(c)k} = \mathbf{x}_{(c)i}/N_i$ a todas las unidades seleccionadas en el conglomerado i. Se define el vector de variables auxiliares como $\mathbf{x}_{(cu)k} = \begin{pmatrix} \mathbf{x}_{(c)k} \\ \mathbf{x}_{(u)k} \end{pmatrix}$, cuyo total poblacional es conocido $\sum_{\mathcal{U}} \mathbf{x}_{(cu)k} = \mathbf{X}$.

Entonces, partiendo de $d_k = d_{Ii}d_{k|i}$, se calculan los pesos de las unidades w_k, $k \in s$, con la ecuación de calibración $\sum_s w_k \mathbf{x}_{(cu)k} = \mathbf{X}$. Posteriormente se calculan el peso del conglomerado $w_{Ii} = \sum_{s_i} w_k/N_i$.

c) Calibración en un paso con la opción de integración *ii*)

En (2.15) se reemplazan los pesos w_k por $d_{k|i}w_{Ii}$, haciendo que la ecuación sea función de los pesos w_{Ii}. Se define el vector de variables auxiliares como $\mathbf{x}_{(cu)k} = \begin{pmatrix} \mathbf{x}_{(c)i} \\ \widehat{\mathbf{x}}_{(u)i_{ht}} \end{pmatrix}$, donde $\widehat{\mathbf{x}}_{(u)i_{ht}} = \sum_{s_i} d_{k|i}\mathbf{x}_{(\mathbf{u})\mathbf{k}}$ es un estimador insesgado del total del conglomerado $\mathbf{x}_{(u)i} = \sum_{\mathcal{U}_i} \mathbf{x}_{(u)k}$. Entonces, partiendo de d_{Ii}, se calculan los pesos del conglomerado w_{Ii}, $i \in s_I$, con la ecuación de calibración $\sum_{s_I} w_{Ii}\mathbf{x}_{(cu)i} = \mathbf{X}$. Después se calcula el peso de la unidad $w_k = d_{k|i}w_{Ii}$.

d) Calibración en dos pasos con la opción de integración *i*)

Partiendo de d_{Ii}, se calcula en el primer paso el peso del conglomerado w_{Ii}, $i \in s_I$, con la ecuación de calibración (2.14). En el segundo paso, partiendo de $d_k = d_{Ii}d_{k|i}$, se calcula el peso de la unidad w_k, $k \in s$, con la ecuación de calibración (2.15) y tal que $\sum_{s_i} w_k = N_i w_{Ii}$ para todo $i \in s_i$, con w_{Ii} obtenido del paso uno.

Los cuatro casos anteriores dan diferentes sistemas de pesos calibrados. El caso *a*) da pesos no integrados, con w_{Ii} satisfaciendo (2.14) y w_k satisfaciendo (2.15). Los pesos obtenidos con *b*), *c*) y *d*) dan pesos integrados que satisfacen (2.14) y (2.15). *b*) y *d*) satisfacen (2.14) pero difieren en que *d*) usa valores verdaderos $\mathbf{x}_{(c)i}$ mientras que *b*) utiliza valores "imputados". En *c*) se usa la calibración en un paso bajo la condición (2.15) que es más restrictiva que la (2.14).

Los casos *b*) y *c*) se reducen a calcular los pesos de la forma simple, usada un muestreo en una etapa. Para ello se especifica un vector $\mathbf{z}_k$, de la misma dimensión que $\mathbf{x}_k$, y se calculan los pesos para $k \in s$

$$w_k = a_k(1 + \boldsymbol{\lambda}_s' \mathbf{z}_k), \tag{2.17}$$

con

$$\boldsymbol{\lambda}_s' = \left(\sum_{\mathcal{U}} \mathbf{x}_k - \sum_s a_k \mathbf{x}_k \right)' \left(\sum_s a_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1}. \tag{2.18}$$

Multiplicando los pesos originales d_k por el factor $(1 + \boldsymbol{\lambda}_s' \mathbf{z}_k)$ se obtienen los pesos calibrados. Se pueden elegir diferentes $\mathbf{z}_k$ que dan diferentes pesos. Una elección simple aunque no necesariamente óptima es tomar $\mathbf{z}_k = \mathbf{x}_k$, en este caso el estimador resultante es el estimador de regresión generalizado. Se dice que $\mathbf{z}_k$ es un vector válido si la matriz $J \times J$, $\sum_s d_k \mathbf{z}_k \mathbf{x}_k'$, es invertible para cada posible muestra s (la relación $\mathbf{z}_k$-w_k no es uno a uno, distintos $\mathbf{z}_k$ pueden dar el mismo w_k). Con los anteriores pesos se obtiene un estimador de calibración $\widehat{Y}_{cal}$, para el total Y que se puede escribir de la forma dada en (1.31).

En el muestreo en dos etapas la estimación puntual no es más compleja. Cualquier software que calcule el caso anterior puede ser adaptado para obtener estimadores de calibración en los casos *a*), *b*) y *c*). Pero para el caso de dos etapas el calculo de la varianza es más complejo ya que contamos con dos componentes, una por cada etapa de selección.

2.1.6 El método de Lemaître y Dufour

Dentro de la perspectiva de la integración de pesos, el método de Lemaître y Dufour (1987)[32] consiste en la ponderación de forma igualitaria de todas las unidades dentro del conglomerado seleccionado. Es decir, se considera un muestreo en una etapa de conglomerados, con la observación de todas las unidades que componen cada conglomerado seleccionado. El objetivo es producir el mismo peso para cada individuo del mismo conglomerado y usar ese peso para estimar características del conglomerado.

Lemaître y Dufour (1987)[32] proponen el reemplazo de los valores de la variable medida a nivel unidad (unidades secundarias) $\mathbf{x}_k$, donde cada unidad tiene un vector de fila con su propia información auxiliar, por una nueva matriz $\mathbf{z}_k$, basada en las variables-persona auxiliares, pero con los promedios para las características auxiliares correspondientes. Como las variables auxiliares son generalmente categóricas, y son así codificadas en variables indicadoras, la matriz $\mathbf{z}_k$ contendrá entonces las proporciones para el conglomerado. Así, el valor para cada miembro del conglomerado i, con un tamaño de N_i unidades, es $\mathbf{z}_{(c)i} = \sum_{k \in s_i} \mathbf{x}_{(c)i}/N_i$.

Cada miembro del conglomerado (unidad primaria) tendrá el mismo vector fila para la información auxiliar en la matriz. Así, si el peso de diseño de éste se asigna a todos los miembros del conglomerado, es decir, si lo adaptamos al muestreo de hogares, cada persona en un hogar tendrá el mismo peso final. Neethling y Galpin (2006)[37] muestran además que este peso, que está basado en variables auxiliares a nivel persona, también es un peso apropiado para ser usado en la estimación de variables a nivel conglomerado.

La cuestión que debe abordarse es si estas estimaciones resultantes, aplicadas a nivel unidad, son útiles para hacer estimaciones de los totales a nivel conglomerado, después de modificar la matriz $\mathbf{z}_k$ añadiendo columnas para cada categoría referente a una variable medida a nivel conglomerado. Por ejemplo, si seleccionamos dos hogares (conglomerados) con 4 y 3 personas respectivamente (unidades), de una matriz de variables auxiliares a la que denominamos $\mathbf{x}_k$, compuesta de dos variables (grupo de edad y sexo), con cuatro categorías de edades y dos de sexo, la matriz vendría dado por de la siguiente forma:

$$\mathbf{x}_k = \begin{pmatrix} & E_1 & E_2 & E_3 & E_4 & M & F \\ h_1p_1 & 1 & 0 & 0 & 0 & 0 & 1 \\ h_1p_2 & 0 & 1 & 0 & 0 & 1 & 0 \\ h_1p_3 & 0 & 1 & 0 & 0 & 1 & 0 \\ h_1p_4 & 0 & 0 & 0 & 1 & 1 & 0 \\ h_2p_1 & 0 & 0 & 1 & 0 & 1 & 0 \\ h_2p_2 & 0 & 0 & 1 & 0 & 0 & 1 \\ h_2p_3 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Supongamos que se añade la variable "Tipo de localidad", dividida en tres categorías: urbana, periurbana y rural. Definimos tres nuevas variables U, PU, R. Las

entradas de la nueva variables sería la inversa del tamaño del hogar para la variable a la que pertenece el hogar, y 0 para las otras dos variables de la zona. Cuando se dispone además de información auxiliar a nivel hogar, la matriz anterior puede ser ampliada añadiendo columnas indicadoras para cada variable-hogar considerada. Esta matriz incluye tanto variables auxiliares a nivel persona y como a nivel hogar.

$$\mathbf{z}_k = \begin{pmatrix} & E_1 & E_2 & E_3 & E_4 & M & F & U & PU & R \\ h_1p_1 & 1/4 & 2/4 & 0 & 1/4 & 3/4 & 1/4 & 1/4 & 0 & 0 \\ h_1p_2 & 1/4 & 2/4 & 0 & 1/4 & 3/4 & 1/4 & 1/4 & 0 & 0 \\ h_1p_3 & 1/4 & 2/4 & 0 & 1/4 & 3/4 & 1/4 & 1/4 & 0 & 0 \\ h_1p_4 & 1/4 & 2/4 & 0 & 1/4 & 3/4 & 1/4 & 1/4 & 0 & 0 \\ h_2p_1 & 0 & 1/3 & 2/3 & 0 & 1/3 & 2/3 & 0 & 0 & 1/3 \\ h_2p_2 & 0 & 1/3 & 2/3 & 0 & 1/3 & 2/3 & 0 & 0 & 1/3 \\ h_2p_3 & 0 & 1/3 & 2/3 & 0 & 1/3 & 2/3 & 0 & 0 & 1/3 \end{pmatrix}$$

Es decir, el peso $1/4$ para h_1p_1 quiere decir que la persona pertenece a un hogar donde $1/4$ de las personas son mujeres.

La cuestión ahora es cómo el método de Lemaître y Dufour (1987)[32] debe utilizarse cuando a nivel unidad y/o conglomerado existen distintas variables auxiliares. Lemaître y Dufour sólo consideran como variables auxiliares a nivel persona el sexo y grupo de edad. Chowdhury (1997)[9] incluye las variables sexo y grupo de edad, así como el estado del hogar (9 combinaciones posibles de los adultos y niños), por lo que la matriz de variables auxiliares se amplia para contener estas variables indicadoras de estos grupos. Nieuwenbroek (1993)[38] considera la aplicación del enfoque de Lemaître y Dufour por el modelado a nivel de hogares, utilizando variables auxiliares sólo persona.

Alternativamente, el caso *c*) (calibración en un paso con opción de integración *ii*)) parte de los pesos d_{Ii}, siendo $\mathbf{x}_{(cu)i} = \mathbf{x}_{(c)i}$ (la única información auxiliar para los conglomerados), donde $\mathbf{x}_{(c)i}$ es el vector de totales para el conglomerado i. Se obtiene primero w_{Ii}, y se toma $w_k = w_{Ii}$ para todas las unidades en el conglomerado i. Se puede comprobar que la elección de $\mathbf{z}_i = \mathbf{x}_{(c)i}/N_i$ proporciona exactamente los mismos pesos individuales que el método de Lemaître y Dufour. Esta perspectiva es más directa puesto que el vector auxiliar se define directamente al nivel conglomerado sin la construcción preliminar de igualdad de pesos para las unidades dentro de cada conglomerado.

2.2 Estimador de contracción

La contracción es una manera natural de mejorar las estimaciones disponibles, en términos del error cuadrático medio (1.7). Por ejemplo, los estimadores compuestos se utilizan en la estimación de áreas pequeñas para equilibrar el sesgo potencial del estimador sintético contra la inestabilidad del estimador directo (Rao, 2003[46]). El uso de estimadores de contracción en el contexto del análisis de regresión ha sido descrito por Copas (1983)[6]. Éste aplica contracción en un contexto donde el problema es predecir una respuesta binaria sobre la base de variables explicativas binarias. Del mismo modo, Schäfer y Strimmer (2005)[57] definen un estimador de contracción de la matriz de covarianza, y Rueda y Menéndez (2010)[52] utilizan la contracción de regiones en la estimación de áreas pequeñas.

La eficiencia de la utilización de estimadores ponderados depende de la relación entre las variables auxiliares y las variables de interés. Un estimador basado en la información auxiliar a nivel conglomerado será más eficiente cuando la información auxiliar está bien relacionada con la variable de interés. De manera similar, una información auxiliar basada en el nivel unidad, será más eficiente cuando la información esté relacionada con la variable de interés. Una forma natural de equilibrar la eficiencia de estos estimadores es tomar una media ponderada.

Sea $\widehat{Y}_{cal}$ el estimador de calibración, a nivel unidad, para el total de la población Y definido en (2.13), y sea $\widehat{Y}_{I,cal}$ el estimador de calibración, a nivel conglomerado, para el total de la población, definido en (2.12), donde los pesos unitarios w_k y los pesos a nivel conglomerado w_{Ii} satisfacen las ecuaciones (2.15) y (2.14) respectivamente.

Las ecuaciones de calibración imponen consistencia en el sistema de pesos, de manera que, cuando se aplican a las variables auxiliares son coherentes con los agregados conocidos. El estimador $\widehat{Y}_{cal}$ incorpora información auxiliar a nivel unidad, mientras que el estimador $\widehat{Y}_{I,cal}$ incorpora información auxiliar a nivel conglomerado, pero no en un patrón integrado. La calibración propone modificar los pesos iniciales d_k o d_{Ii} por nuevos pesos w_k o w_{Ii}, determinados tan cerca como los d_k o d_{Ii}, por lo que se obtiene estimaciones de diseños insesgados, ya que la información auxiliar utilizada para la calibración mejora la precisión de las estimaciones de la encuesta.

Proponemos un estimador basado en la información compuesta, de la siguiente manera. A partir de d_{Ii}, calculamos los pesos a nivel conglomerado w_{Ii} para $i \in s_I$, calibrando la información a nivel conglomerado. Después de calibrar, tomamos $w_k^I = w_{Ii}$, para todo k en el mismo $i \in s_I$ y se define $\widehat{Y}_{cal}^I = \sum_{k \in s} w_k^I y_{(u)k}$.

Siguiendo a Thompson (1968)[65], proponemos una contracción del estimador unitario $\widehat{Y}_{cal}$ hacia el estimador $\widehat{Y}_{cal}^I$. Obtenemos así $\widetilde{Y} = K\widehat{Y}_{cal} + (1-K)\widehat{Y}_{cal}^I$, donde K es una constante que satisface $0 < K < 1$.

$\widetilde{Y}$ es un estimador asintóticamente insesgado de Y porque $\widehat{Y}_{cal}$ y $\widehat{Y}_{cal}^I$ también son estimadores asintóticamente insesgados de Y.

Una elección óptima de K se puede calcular mediante la minimización de la varianza $\widetilde{Y}$, que viene dada por

$$V(\widetilde{Y}) = K^2 V(\widehat{Y}_{cal}) + (1-K)^2 V(\widehat{Y}_{cal}^I) + 2K(1-K)Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I). \quad (2.19)$$

Como esta ecuación es una ecuación cuadrática de K, su único extremo, al que denominaremos K_{opt}, se calcula fácilmente

$$K_{opt} = \frac{V(\widehat{Y}_{cal}^I) - Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)}{V(\widehat{Y}_{cal}) + V(\widehat{Y}_{cal}^I) - 2Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)}. \quad (2.20)$$

Obsérvese que el denominador en K_{opt} es la varianza de la diferencia $\widehat{Y}_{cal}^I - \widehat{Y}_{cal}$ y el numerador de K_{opt} es la covarianza de esta diferencia con $\widehat{Y}_{cal}^I$.

K_{opt} se puede utilizar para definir la expresión óptima

$$\widetilde{Y}_{opt} = K_{opt}\widehat{Y}_{cal} + (1 - K_{opt})\widehat{Y}_{cal}^I. \quad (2.21)$$

La varianza de este estimador está dada por:

$$V(\widetilde{Y}_{opt}) = V_{min}(\widetilde{Y}) = \frac{V(\widehat{Y}_{cal}^I)V(\widehat{Y}_{cal}) - Cov^2(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)}{V(\widehat{Y}_{cal}) + V(\widehat{Y}_{cal}^I) - 2Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)}. \quad (2.22)$$

Como el coeficiente óptimo, K_{opt}, depende de varianzas y covarianzas poblacionales, que son generalmente desconocidas en la práctica, $\widetilde{Y}_{opt}$ no se puede calcular. Por esta razón, estas varianzas y covarianzas deben ser estimadas.

Los valores poblacionales $V(\widehat{Y}_{cal}^I)$, $V(\widehat{Y}_{cal})$ y $Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)$ están definidos mediante dos procedimientos diferentes.

En el caso de $V(\widehat{Y}_{cal})$, la linealización automática identifica el estadístico linealizado y los residuos que determinan la varianza aproximada,

$$V(\widehat{Y}_{cal}) \approx V(\sum_{k\in s} d_k e_k) = \sum_{k,l\in U} F_{kl} e_k e_l, \tag{2.23}$$

con $F_{kl} = (d_k d_l/d_{kl}) - 1$ para $k \neq l$ y $F_{kk} = d_k - 1$ for $l = k$ y donde $e_k = y_{(u)k} - \mathbf{x}_{(u)k}^t \mathbf{B}$ y $\mathbf{B} = (\sum_{k\in U} \mathbf{x}_{(u)k}\mathbf{x}_{(u)k}^t)^{-1} \sum_{k\in U} \mathbf{x}_{(u)k} y_{(u)k}$. Por último, la varianza se puede estimar por

$$\hat{V}(\hat{Y}_{cal}) = \sum_{k\in s}\sum_{\ell\in s}(d_k d_l - d_{kl})\widehat{e}_k\widehat{e}_\ell, \tag{2.24}$$

donde $\widehat{e}_k = y_{(u)k} - \mathbf{x}_{(u)k}^t \widehat{\mathbf{B}}$ y $\widehat{\mathbf{B}} = (\sum_{k\in s} d_k \mathbf{x}_{(u)k}\mathbf{x}_{(u)k}^t)^{-1} \sum_{k\in s} d_k \mathbf{x}_{(u)k} y_{(u)k}$.

En el caso de $V(\widehat{Y}_{cal}^I)$, ya que $\widehat{Y}_{cal}^I$ es un estimador del total poblacional, la linealización automática da la estimación de la varianza como:

$$\hat{V}(\hat{Y}_{cal}^I) = \sum_{i\in s_I}\sum_{j\in s_I}(d_{Ii}d_{Ij} - d_{Iij})\widehat{e}_{Ii}\widehat{e}_{Ij}, \tag{2.25}$$

donde $\widehat{e}_{Ii} = y_{(c)i} - \mathbf{x}_{(c)i}^t \widehat{\mathbf{B}}_I$ y $\widehat{\mathbf{B}}_I = (\sum_{i\in s_I} d_{Ii}\mathbf{x}_{(c)i}\mathbf{x}_{(c)i}^t)^{-1} \sum_{i\in s_I} d_{Ii}\mathbf{x}_{(c)i} y_{(c)i}$.

En el caso de $Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I)$, tenemos

$$\begin{aligned} Cov(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I) \approx Cov(\sum_{k\in s} d_k e_k, \sum_{i\in s_I} d_{Ii} e_{Ii}) = Cov(\sum_{i\in s_I} d_{Ii} e_i^I, \sum_{i\in s_I} d_{Ii} e_{Ii}) = \\ = \sum_{i\in \mathcal{U}_I}\sum_{j\in \mathcal{U}_I} \frac{d_{Ii}d_{Ij} - d_{Iij}}{d_{Iij}} e_i^I e_{Ij}, \end{aligned} \tag{2.26}$$

siendo $d_k = d_{Ii}d_{k|i} = d_{Ii}$, ya que se trata de un muestreo de conglomerados de una sola etapa. Los residuos e_i^I se determinan mediante el cálculo del primer residuo e_k, basado en la regresión de $y_{(u)k}$ en $\mathbf{x}_{(u)k}^t$ a nivel de la unidad y posteriormente sumando los residuos e_k dentro de cada conglomerado para producir $e_i^I = \sum_{k\in \mathcal{U}_i^c} e_k$.

Finalmente, la covarianza puede ser estimada por

$$\widehat{Cov}(\widehat{Y}_{cal}, \widehat{Y}_{cal}^I) = \sum_{i\in s_I}\sum_{j\in s_I}(d_{Ii}d_{Ij} - d_{Iij})\widehat{e}_i^I\widehat{e}_{Ij}. \tag{2.27}$$

El estimador óptimo puede definirse mediante la siguiente ecuación:

$$\widehat{Y}_{opt_D} = \widehat{K}_{opt_D}\widehat{Y}_{cal} + (1 - \widehat{K}_{opt_D})\widehat{Y}^I_{cal}, \tag{2.28}$$

donde $\widehat{K}_{opt_D}$ denota que las estimaciones son sustituidos por las varianzas y covarianzas en la ecuación (2.20).

Como se ha descrito en el Capítulo 1, es posible obtener estimadores de calibración para varianzas y covarianzas de los estimadores de calibración. En nuestro caso, los valores poblacionales desconocidos $V(\widehat{Y}^I_{cal})$, $V(\widehat{Y}_{cal})$ y $Cov(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})$ pueden ser estimados mediante (Singh *et al.*, 1999[60]; Singh, 2001[58], 2010[59]):

$$\hat{V}_W(\hat{Y}_{cal}) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}(w_k\widehat{e}_k)(w_\ell\widehat{e}_\ell),$$

anteriormente definido en (1.36),

$$\hat{V}_W(\hat{Y}^I_{cal}) = \sum_{i\in s_I}\sum_{j\in s_I}\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}(w_{Ii}\widehat{e}_{Ii})(w_{Ij}\widehat{e}_{Ij}), \tag{2.29}$$

y

$$\widehat{Cov}_W(\widehat{Y}_{cal}, \widehat{Y}^I_{cal}) = \sum_{k\in s_I}\sum_{\ell\in s_I}\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}(w_{Ii}\widehat{e}^I_i)(w_{Ij}\widehat{e}_{Ij}). \tag{2.30}$$

Un segundo estimador óptimo puede ser definido como

$$\widehat{Y}_{opt_W} = \widehat{K}_{opt_W}\widehat{Y}_{cal} + (1 - \widehat{K}_{opt_W})\widehat{Y}^I_{cal}, \tag{2.31}$$

donde $\widehat{K}_{opt_W}$ denota que estas estimaciones de calibración son sustituidas por las varianzas y covarianzas en (2.20).

Consideremos ahora el problema de la estimación de las varianzas de los dos estimadores propuestos $\widehat{Y}_{opt_D}$ y $\widehat{Y}_{opt_W}$. Esto no es una cuestión sencilla, porque los factores $\widehat{K}_{opt_D}$ y $\widehat{K}_{opt_W}$ se derivan de los datos de la encuesta. Una manera de abordar este problema es considerar los métodos de remuestreo, como los grupos aleatorios y técnicas de muestras balanceadas, jackknife o bootstrap (Wolter, 2007[74] o Gershunskaya, Jiang y Lahiri, 2009[17]). Otra opción es utilizar los resultados asintóticos. Randles (1980)[48] estudia los efectos sobre la convergencia

de la sustitución de estimaciones de los parámetros en un estadístico. Este importante resultado (que ha sido utilizado por los autores como Chambers y Dustan, 1986[8], Rao *et al.*, 1990[47], Gijbels y Veraverbeke, 1988[18] y Rueda *et al.*, 2007[51]) supone que las observaciones de la muestra pueden ser tratadas como realizaciones independientes e idénticamente distribuidas. Recientemente Wand y Opsomer (2010)[73] extendieron este resultado a un diseño de muestreo general.

Asumimos una sucesión creciente de poblaciones finitas $\{U_N\}$ con $N \to \infty$. Usando la notación dada por Randles (1980)[48], el estimador propuesto se denota por $\widehat{Y}_{opt_D} = T_n(\widehat{K}_{opt_D}) = T_n(\widehat{\boldsymbol{\lambda}})$ y $\widetilde{Y}_{opt} = T_n(K_{opt}) = T_n(\boldsymbol{\lambda})$, donde el estimador consistente $\widehat{\boldsymbol{\lambda}}$ estima $\boldsymbol{\lambda}$. Si el diseño de muestreo verifica los supuestos propuestos por Wang y Opsomer (2010), el modelo propuesto por éstos se puede aplicar, debido a que el estimador $\widehat{\boldsymbol{\lambda}}$ y la función T verifican todos los supuestos necesarios. Por consiguiente, concluimos que la distribución asintótica de $T_n(\widehat{\boldsymbol{\lambda}})(= \widehat{Y}_{opt_D})$ es la misma que la de $T_n(\boldsymbol{\lambda})(= \widetilde{Y}_{opt})$.

Así, $\widehat{Y}_{opt_D}$ es un estimador asintóticamente insesgado para Y y la varianza asintótica de este estimador es:

$$VA(\widehat{Y}_{opt_D}) = V_{min}(\widetilde{Y}_{opt}) = \frac{V(\widehat{Y}^I_{cal})V(\widehat{Y}_{cal}) - Cov^2(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})}{V(\widehat{Y}_{cal}) + V(\widehat{Y}^I_{cal}) - 2Cov(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})}. \tag{2.32}$$

Por tanto, un estimador para la aproximación de la varianza vendrá dado por:

$$\widehat{VA}(\widehat{Y}_{opt_D}) = \frac{\widehat{V}(\widehat{Y}^I_{cal})\widehat{V}(\widehat{Y}_{cal}) - \widehat{Cov}^2(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})}{\widehat{V}(\widehat{Y}_{cal}) + \widehat{V}(\widehat{Y}^I_{cal}) - 2\widehat{Cov}(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})}. \tag{2.33}$$

El estimador de la varianza de $\widehat{Y}_{opt_W}$ se puede definir de la misma manera.

2.3 Estudio de simulación

Estimadores

En esta sección, evaluamos el comportamiento empírico de los estimadores propuestos a través de un estudio de simulación y comparamos los estimadores propuestos, $\widehat{Y}_{opt_S}$, $\widehat{Y}_{opt_D}$ y $\widehat{Y}_{opt_W}$, con el definido por Estevao y Särndal (2006)[16], denotado por $\widehat{Y}_{ES}$.

Se ha considerado tres estimadores para K_{opt}, para los cuales se han estimado $V(\widehat{Y}_{cal})$, $V(\widehat{Y}^I_{cal})$ y $Cov(\widehat{Y}_{cal}, \widehat{Y}^I_{cal})$. En primer lugar, se ha estimado las varianzas y covarianzas en función de los resultados de varias series de simulaciones (en concreto 1000). A partir de estos valores se ha calculado el estimador óptimo, denotado por $\widehat{Y}_{opt_S}$. También se ha realizado el cálculo de K_{opt_D} y K_{opt_W}, definidos en el apartado anterior, para calcular los estimadores $\widehat{Y}_{opt_D}$ y $\widehat{Y}_{opt_W}$. Hay que tener en cuenta que $\widehat{Y}_{opt_S}$ se calcula simplemente para utilizarlo de comparación, ya que este estimador no se puede conseguir en la práctica.

Poblaciones de estudio

La primera población utilizada en este estudio de simulación fue la del Programme for International Student Assessment (PISA)[42]. Este programa fue desarrollado para evaluar a los estudiantes de 15 años, de toda clase de escuelas, programas educativos y técnicos, de los países miembros del sistema de la OCDE, así como otros socios asociados. Los datos analizados corresponden al año 2006, para 57 países, y se centra en evaluación de las habilidades de lectura, matemáticas y ciencias.

La web de la OCDE[40] ofrece microdatos de varios años y en particular del estudio del 2006, en la que se incluyen información sobre los estudiantes, las familias y las escuelas, al tiempo que garantiza el anonimato de los participantes. El estudio de simulación lo realizamos con los datos de la OCDE para España.

El conjunto de microdatos del informe PISA-España contiene información sobre las pruebas que se llevaron a cabo en 686 escuelas, con la participación de 19.604 estudiantes. Se consideraron las unidades sin datos faltantes para las variables de estudio, obteniendo de este modo una población con tamaño $N = 18341$ estudiantes (unidades) agrupados en $N_I = 686$ escuelas (conglomerados).

Para la población en PISA, se optó por dos variables de interés, una variable cualitativa dicotómica “Estudiar una carrera de ciencias en el futuro” (Sci. future - After secondary school) y una variable cuantitativa “puntuación en matemáticas”. Como información auxiliar a nivel unidad elegimos el género (masculino y femenino), el nivel educativo de la madre, el nivel educativo del padre y el mayor nivel educativo de los padres con las categorías: sin estudios, educación primaria, educación secundaria y estudios superiores. Por último, como información auxiliar a nivel conglomerado, se consideró el tipo de escuela (pública o privada) y el tipo

localidad donde se encuentra la escuela (aldea o pueblo pequeño, pueblo, ciudad o gran ciudad).

La segunda población finita utilizada en la simulación se obtuvo de la Encuesta de Presupuestos Familiares (EPF) llevada a cabo en 2006. Se trata de una muestra aleatoria representativa de una encuesta que se realiza en los hogares españoles. La web del Instituto Nacional de Estadística español (INE) incluye estadísticas que se pueden utilizar para obtener los archivos de microdatos, cada uno de los cuales contiene datos individuales de una estadística determinada, filtradas adecuadamente para que la información se anónima y por lo tanto garantizar la confidencialidad.

Los microdatos EPF-2006 contiene 55.699 unidades agrupadas en 19435 conglomerados. Se consideraron sólo las unidades sin datos faltantes para las variables de estudio, y así obtuvo una población de tamaño $N = 9243$ individuos agrupados en $N_I = 5800$ hogares. Las principales características de la EPF-2006 se puede consultar en la página web del INE[24].

Variables

En este caso, la variable de interés es el ingreso de la población EPF. Se consideraron los géneros (masculino, femenino) y el nivel educativo (sin estudios, primaria, secundaria, superior) como información auxiliar para las unidades. Por último, se consideró el nivel educativo del sustentador principal (primaria, secundaria, universitaria) como información auxiliar para los conglomerados (hogares).

Tipos de muestreos

Se han realizado dos estudios de simulación. En el primer caso, para cada simulación, se ha seleccionado una muestra aleatoria simple de tamaño m de las unidades de primera etapa (escuelas y los hogares, respectivamente) de la población de conglomerados. Seleccionando todas las unidades del conglomerado, para una muestra de unidades de segunda etapa de tamaño n (individuos y estudiantes, respectivamente). En un segundo caso, se ha utilizado el método Midzuno para seleccionar una muestra de unidades (escuelas y hogares) con probabilidades desiguales (proporcional al tamaño del conglomerado).

Distancias para la calibración

Los pesos de los estimadores de calibración se puede calcular por medio de diferentes métodos de calibración, en función de la distancia seleccionada. Para fines comparativos, es interesante mostrar la precisión conseguida cuando se utilizan distintos métodos de calibración. En nuestro estudio de simulación, se han utilizado tres métodos de calibración: linear, raking y logit (Deville y Särndal, 1992[10]) descritos en el capítulo anterior.

Estimadores

Para cada muestra calculamos el estimador de Horvitz-Thompson $\widehat{Y}_{ht}$, el estimador de Estevao-Särndal $\widehat{Y}_{ES}$ y los estimadores propuestos, $\widehat{Y}_{opt_S}$, $\widehat{Y}_{opt_D}$, $\widehat{Y}_{opt_W}$. En el caso de una variable dicotómica, la proporción de la población P se estimó mediante $\widehat{P} = \widehat{Y}/N$ para cada estimador del total de la población $\widehat{Y}$.

Tamaño de la muestra

El proceso de simulación se repitió $B = 1000$ veces para distintos tamaños de muestra del conglomerado $m = 25, 50, 75, 150, 200$ y 250 para la población EPF. Con fracciones de muestreo, para los tamaños de muestra $n = 25$ a $n = 250$, $f = 25/5800 = 0{,}004$ a $f = 250/5800 = 0{,}04$. De forma similar, se tomaron los tamaños $m = 20, 25, 30, 40, 45$ y 50 para la población PISA, con fracciones de muestreo que varían de $f = 20/673 = 0{,}0297$ a $f = 50/673 = 0{,}0743$ en este caso.

Medidas de sesgo y eficiencia

El rendimiento de cada estimador del total se mide y se compara en términos del sesgo relativo (*SR*) y la eficacia relativa (*ER*). Los valores simulados de *SR* y *ER*, para un estimador del total determinado T, se calculan como se indica en 1.6 y 1.7 (apartado 1.2).

Resultados y comentarios

Las Tablas 2.1, 2.2 y 2.3 y las Figuras 2.1, 2.2 y 2.3 muestran los resultados obtenidos en el muestreo aleatorio simple. En la Tabla y Figura 2.1 la variable principal

es una variable dicotómica, Sci. future (estudiar ciencias en el futuro), las variables auxiliares a nivel conglomerado (centro) son el tipo de escuela y el tipo de comunidad, las variables auxiliares a nivel unidades (estudiantes) son el sexo, el nivel educativo de la madre, nivel educativo del padre y el nivel educativo más alto del hogar del alumno (lo que no implica que sea el del padre o de la madre).

En la Tabla y Figura 2.2, la variable principal es puntuación en matemáticas (una variable continua), y la información auxiliar es la misma que en la Tabla 2.1. En la Tabla y Figura 2.3, la variable principal es la renta de la población EPF, la variable auxiliar a nivel hogar es el nivel educativo del sustentador principal, y las variables auxiliares a nivel persona son el género y el nivel educativo.

Las Tablas 2.1, 2.2 y 2.3 muestran que los tres métodos de calibración (linear, raking y logit) dan resultados similares en términos de *SR* y *ER*. Por tanto, los tres métodos conducen a las mismas conclusiones. Se puede observar que el estimador de Estevao-Särndal siempre es más eficiente que el estimador Horvitz-Thompson, y que como muestran las Figuras 2.1, 2.2 y 2.3 los tres estimadores óptimos dan mejores resultados respecto a la *ER* que el estimador de Estevao-Särndal. Ocurre lo mismo para el estudio PISA respecto al *SR*, como se observa en las Figuras 2.1 y 2.2.

No existe una relación clara entre $\widehat{Y}_{opt_D}$ y $\widehat{Y}_{opt_W}$ en términos de *SR* y *ER*, aunque el estimador de $\widehat{Y}_{opt_W}$ parece ser un poco más eficiente que $\widehat{Y}_{opt_D}$.

Para tamaños de muestra moderados, el estimador propuesto es claramente el estimador más eficiente, y la ganancia en eficiencia es mayor para tamaños de muestra pequeños. El propuesto y el estimador de Estevao-Särndal dan valores similares de *ER* a medida que aumenta el tamaño de la muestra. En las Figuras 2.1, 2.2 y 2.3 podemos ver como se estabilizan en torno a cero el sesgo relativo y como se equipara la eficiencia relativa de los cuatro estimadores al aumentar el número de conglomerados seleccionados.

Hay que tener en cuenta que $\widehat{Y}_{opt}$ tiene la desventaja añadida de tener que estimar K_{opt}, aunque el esfuerzo que se requiere no es tan grande para tamaños de muestra moderados, logrando así un estimador más eficiente.

Las Tablas 2.4, 2.5 y 2.6 y las Figuras 2.4, 2.5 y 2.6 muestran los resultados obtenidos en el muestreo con probabilidad proporcional al tamaño. Las variables auxiliares y objetivo de la Tabla 2.4 son las mismas que en la Tabla 2.1, las de la

Tabla 2.5 son las mismas que en la Tabla 2.2, y las de la Tabla 2.6 son las mismas que en la Tabla 2.3, más la edad (como variable continua) como variable auxiliar a nivel persona.

Las Tablas 2.4, 2.5 y 2.6 y las Figuras 2.4, 2.5 y 2.6 sólo incluyen el estimador de diseño óptimo $\widehat{Y}_{opt_D}$, ya que $\widehat{Y}_{opt_W}$ da resultados similares a $\widehat{Y}_{opt_D}$.

En las Tablas 2.4 y 2.5 (población PISA) se observa que el estimador de Estevao-Särndal no siempre es más eficiente que el estimador de Horvitz-Thompson (para $m = 50$, y utilizando el método lineal, el estimador de Estevao-Särndal es más eficiente que el de Horvitz-Thompson), mientras que la Tabla 2.6 (población EPF) muestra que el estimador de Estevao-Särndal es más eficiente que Horvitz-Thompson (excepto para $m = 25$ y con los métodos de raking y logit).

El estimador propuesto es claramente el estimador más eficiente, Figuras 2.4, 2.5 y 2.6, y la ganancia en eficiencia con respecto al estimador de Estevao-Särndal es mayor para tamaños de muestra pequeños (Tablas 2.4, 2.5 y 2.6). Este aumento de la eficiencia es moderada (*ER* $\simeq 2\,\%$) cuando se estima un total poblacional (una variable cuantitativa, Tabla 2.5) y aumenta (*ER* varía de 16 % a 18 %) al estimar una proporción poblacional (una variable cualitativa, Tabla 2.4), mientras que *ER* tiene de rango entre 32 % a 47 %, cuando se estima el total de los ingresos en la población EPF (Tabla 2.6).

De las Tablas 2.1, 2.2, 2.3, 2.4, 2.5 y 2.6, y el conjunto de Figuras tanto en el caso del muestreo aleatorio simple o bajo el esquema de muestreo Midzuno (probabilidades desiguales), observamos que el estimador óptimo propuesto produce valores de *ER* mayores que los del estimador de Estevao-Särndal o el estimador de Horvitz-Thompson. Los tres métodos de calibración (linear, raking y logit) producen las mismas conclusiones. Por otra parte, en el supuesto de probabilidades desiguales, el estimador de Estevao-Särndal no siempre es más preciso que Horvitz-Thompson. Destacamos que la ganancia en eficiencia con respecto al estimador de Estevao-Särndal aumenta cuando el tamaño de la muestra disminuye.

Tabla 2.1: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Sci. future. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo aleatorio simple.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{P}_{\pi}$	20	544.31	-.227	100.00	-.227	100.00	-.227	100.00
$\widehat{P}_{ES}$			-.409	115.62	-.117	104.42	-.051	103.47
$\widehat{P}_{opt_S}$			-.190	162.13	-.194	162.13	-.194	162.13
$\widehat{P}_{opt_W}$			-.250	148.24	-.245	148.10	-.246	148.15
$\widehat{P}_{opt_D}$			-.154	148.02	-.149	147.97	-.150	147.97
$\widehat{P}_{\pi}$	25	680.80	.242	100.00	.242	100.00	.242	100.00
$\widehat{P}_{ES}$			.027	127.58	.081	114.61	.125	115.10
$\widehat{P}_{opt_S}$			.245	179.28	.238	180.12	.237	180.12
$\widehat{P}_{opt_W}$			.250	151.13	.251	150.90	.251	150.94
$\widehat{P}_{opt_D}$			.317	151.79	.316	151.68	.316	151.68
$\widehat{P}_{\pi}$	30	817.03	.079	100.00	.079	100.00	.079	100.00
$\widehat{P}_{ES}$			-.108	143.35	-.105	125.96	-.059	127.81
$\widehat{P}_{opt_S}$			.077	178.73	.078	179.02	.078	179.02
$\widehat{P}_{opt_W}$			.072	160.41	.073	160.21	.073	160.23
$\widehat{P}_{opt_D}$			.132	160.00	.132	159.95	.132	159.95
$\widehat{P}_{\pi}$	35	953.69	-.388	100.00	-.388	100.00	-.388	100.00
$\widehat{P}_{ES}$			-.533	144.55	-.498	135.23	-.502	135.85
$\widehat{P}_{opt_S}$			-.424	168.52	-.423	168.55	-.423	168.55
$\widehat{P}_{opt_W}$			-.407	153.07	-.404	153.02	-.405	153.05
$\widehat{P}_{opt_D}$			-.361	153.54	-.361	153.52	-.361	153.52
$\widehat{P}_{\pi}$	40	1089.15	-.063	100.00	-.063	100.00	-.063	100.00
$\widehat{P}_{ES}$			.006	152.95	.101	145.99	.122	144.18
$\widehat{P}_{opt_S}$			.015	179.28	.014	178.92	.013	178.89
$\widehat{P}_{opt_W}$			.009	160.90	.009	160.95	.009	160.95
$\widehat{P}_{opt_D}$			.058	160.05	.058	160.05	.058	160.05
$\widehat{P}_{\pi}$	45	1227.41	-.035	100.00	-.035	100.00	-.035	100.00
$\widehat{P}_{ES}$			.034	153.70	.021	144.45	.089	146.84
$\widehat{P}_{opt_S}$			-.076	174.40	-.075	174.28	-.075	174.28
$\widehat{P}_{opt_W}$			-.044	160.31	-.003	160.26	-.044	160.28
$\widehat{P}_{opt_D}$			-.003	160.62	-.361	160.62	-.003	160.62
$\widehat{P}_{\pi}$	50	1363.28	.072	100.00	.072	100.00	.072	100.00
$\widehat{P}_{ES}$			-.172	150.44	-.140	143.80	-.134	148.72
$\widehat{P}_{opt_S}$			-.068	166.25	-.070	166.56	-.070	166.56
$\widehat{P}_{opt_W}$			-.033	154.42	-.033	154.32	-.033	154.34
$\widehat{P}_{opt_D}$			.003	154.15	.003	154.15	.003	154.15

Tabla 2.2: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Valor plausible en matemáticas. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo aleatorio simple.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{Y}_{\pi}$	20	543.57	-.289	100.00	-.289	100.00	-.289	100.00
$\widehat{Y}_{ES}$			-.028	87.95	-.037	78.37	-.018	77.45
$\widehat{Y}_{opt_S}$			-.012	127.55	-.017	128.20	-.017	128.20
$\widehat{Y}_{opt_W}$			.000	125.15	.001	125.15	.001	125.15
$\widehat{Y}_{opt_D}$			.004	125.94	.005	125.78	.005	125.78
$\widehat{Y}_{\pi}$	25	681.67	.117	100.00	.117	100.0	.117	100.00
$\widehat{Y}_{ES}$			.054	88.18	.039	82.71	.028	84.60
$\widehat{Y}_{opt_S}$			.047	104.60	.062	127.06	.062	127.06
$\widehat{Y}_{opt_W}$			.092	117.78	.092	117.64	.092	117.64
$\widehat{Y}_{opt_D}$			.096	117.50	.096	117.50	.096	117.50
$\widehat{Y}_{\pi}$	30	818.74	.042	100.00	.042	100.00	.042	100.00
$\widehat{Y}_{ES}$			-.108	95.51	-.095	91.32	-.094	92.50
$\widehat{Y}_{opt_S}$			-.101	112.23	-.103	116.95	-.103	116.95
$\widehat{Y}_{opt_W}$			-.076	115.74	-.076	115.74	-.076	115.74
$\widehat{Y}_{opt_D}$			-.073	115.74	-.073	115.74	-.073	115.74
$\widehat{Y}_{\pi}$	35	953.81	.022	100.00	.022	100.00	.022	100.00
$\widehat{Y}_{ES}$			.018	107.41	.019	103.19	.011	105.59
$\widehat{Y}_{opt_S}$			.021	122.39	.024	123.30	.024	123.30
$\widehat{Y}_{opt_W}$			.038	121.21	.038	121.21	.038	121.21
$\widehat{Y}_{opt_D}$			.042	121.21	.042	121.21	.042	121.21
$\widehat{Y}_{\pi}$	40	1091.05	.086	100.00	.086	100.00	.086	100.00
$\widehat{Y}_{ES}$			-.016	114.81	-.003	113.63	-.014	115.20
$\widehat{Y}_{opt_S}$			-.002	129.19	-.003	133.51	-.003	133.51
$\widehat{Y}_{opt_W}$			.015	129.19	.015	129.19	.015	129.19
$\widehat{Y}_{opt_D}$			.019	128.70	.019	128.70	.019	128.70
$\widehat{Y}_{\pi}$	45	1227.10	.053	100.00	.053	100.00	.053	100.00
$\widehat{Y}_{ES}$			.018	105.70	.006	100.20	.009	105.04
$\widehat{Y}_{opt_S}$			-.005	121.35	-.004	123.00	-.004	123.00
$\widehat{Y}_{opt_W}$			.006	122.10	.006	122.24	.006	122.24
$\widehat{Y}_{opt_D}$			.010	121.95	.010	121.95	.010	121.95
$\widehat{Y}_{\pi}$	50	1362.21	-.031	100.00	-.031	100.00	-.031	100.00
$\widehat{Y}_{ES}$			-.029	116.27	-.024	110.13	-.036	118.34
$\widehat{Y}_{opt_S}$			-.005	128.04	-.004	128.53	-.004	128.53
$\widehat{Y}_{opt_W}$			.000	126.10	.000	126.10	.000	126.10
$\widehat{Y}_{opt_D}$			.003	126.10	.003	126.10	.003	126.10

Tabla 2.3: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Renta. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel hogar: Nivel de estudios del sustentador principal. Variables auxiliares a nivel persona: género y nivel educativo. Población EPF. Muestreo aleatorio simple.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{Y}_{\pi}$	25	40.14	1.92	100.00	1.92	100.00	1.92	100.00
$\widehat{Y}_{ES}$			.605	155.86	.500	152.00	.503	152.53
$\widehat{Y}_{opt_S}$			.471	197.47	.460	198.49	.461	198.41
$\widehat{Y}_{opt_W}$			.106	193.54	.106	193.61	.105	193.61
$\widehat{Y}_{opt_D}$			.676	189.93	.676	189.93	.676	189.93
$\widehat{Y}_{\pi}$	50	79.93	.070	100.00	.070	100.00	.070	100.00
$\widehat{Y}_{ES}$			.223	203.21	.129	204.46	.135	204.62
$\widehat{Y}_{opt_S}$			-.013	231.43	-.010	231.27	-.011	231.32
$\widehat{Y}_{opt_W}$			-.394	226.35	-.395	226.30	-.395	226.30
$\widehat{Y}_{opt_D}$			.002	226.30	.002	226.30	.002	226.30
$\widehat{Y}_{\pi}$	75	119.46	-.086	100.00	-.086	100.00	-.086	100.00
$\widehat{Y}_{ES}$			.146	208.94	.034	210.13	.049	210.17
$\widehat{Y}_{opt_S}$			.007	225.73	.006	225.53	.006	225.58
$\widehat{Y}_{opt_W}$			-.309	223.26	-.309	223.21	-.309	223.21
$\widehat{Y}_{opt_D}$			-.007	224.92	-.007	224.92	-.007	224.92
$\widehat{Y}_{\pi}$	100	159.54	.418	100.00	.418	100.00	.418	100.00
$\widehat{Y}_{ES}$			.276	209.95	.227	205.80	.238	206.19
$\widehat{Y}_{opt_S}$			.174	223.16	.174	223.02	.174	223.02
$\widehat{Y}_{opt_W}$			-.029	225.12	-.029	225.12	-.029	225.12
$\widehat{Y}_{opt_D}$			.174	222.47	.174	222.47	.174	222.47
$\widehat{Y}_{\pi}$	150	239.36	.084	100.00	.084	100.00	.084	100.00
$\widehat{Y}_{ES}$			.101	211.82	.051	212.86	.064	212.77
$\widehat{Y}_{opt_S}$			.039	216.97	.031	217.06	.031	217.06
$\widehat{Y}_{opt_W}$			-.108	217.96	-.108	217.96	-.108	217.96
$\widehat{Y}_{opt_D}$			.032	215.38	.032	215.38	.032	215.38
$\widehat{Y}_{\pi}$	200	318.87	.117	100.00	.117	100.00	.117	100.00
$\widehat{Y}_{ES}$			.097	217.72	.060	218.67	.061	220.60
$\widehat{Y}_{opt_S}$			.041	225.63	.042	225.53	.042	225.53
$\widehat{Y}_{opt_W}$			-.055	222.67	-.055	222.67	-.055	222.67
$\widehat{Y}_{opt_D}$			.040	223.11	.040	223.11	.040	223.11
$\widehat{Y}_{\pi}$	250	398.39	-.102	100.00	-.102	100.00	-.102	100.00
$\widehat{Y}_{ES}$			-.012	220.51	-.048	220.65	-.042	220.26
$\widehat{Y}_{opt_S}$			-.054	222.57	-.056	222.62	-.056	222.62
$\widehat{Y}_{opt_W}$			-.127	220.70	-.127	220.75	-.127	220.75
$\widehat{Y}_{opt_D}$			-.058	221.39	-.058	221.39	-.058	221.39

Tabla 2.4: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Sci. future. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo de Midzuno.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{P}_{\pi}$	20	564.438	-.173	100	-.173	100	-.173	100
$\widehat{P}_{ES}$			-.109	75.89	.113	69.99	.136	68.21
$\widehat{P}_{opt_D}$			-.172	116.05	-.173	116.07	-.172	116.08
$\widehat{P}_{\pi}$	25	707.461	.142	100	.142	100	.142	100
$\widehat{P}_{ES}$			-.111	80.03	-.011	70.55	.041	70.32
$\widehat{P}_{opt_D}$			.054	114.25	.054	114.27	.054	114.27
$\widehat{P}_{\pi}$	30	848.861	-.487	100	-.487	100	-.487	100
$\widehat{P}_{ES}$			-.352	89.23	-.304	83.99	-.298	83.86
$\widehat{P}_{opt_D}$			-.350	117.29	-.350	117.38	-.350	117.37
$\widehat{P}_{\pi}$	35	991.364	-.318	100	-.318	100	-.318	100
$\widehat{P}_{ES}$			-.377	86.01	-.361	79.11	-.425	79.65
$\widehat{P}_{opt_D}$			-.359	115.09	-.358	115.09	-.359	115.09
$\widehat{P}_{\pi}$	40	1135.8	.030	100	.030	100	.030	100
$\widehat{P}_{ES}$			.180	95.44	.238	88.43	.231	89.55
$\widehat{P}_{opt_D}$			.049	116.34	.049	116.35	.049	116.35
$\widehat{P}_{\pi}$	45	1275.921	-.005	100	-.005	100	-.005	100
$\widehat{P}_{ES}$			-.075	92.51	-.016	87.70	-.005	88.56
$\widehat{P}_{opt_D}$			-.052	114.93	-.052	114.92	-.052	114.92
$\widehat{P}_{\pi}$	50	1416.123	-.321	100	-.321	100	-.321	100
$\widehat{P}_{ES}$			-.150	100.94	-.133	94.59	-.092	99.12
$\widehat{P}_{opt_D}$			-.217	118.46	-.218	118.45	-.218	118.45

Tabla 2.5: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Valor plausible en matemáticas. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo de Midzuno.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{Y}_{\pi}$	20	563.957	.021	100	.021	100	.021	100
$\widehat{Y}_{ES}$			.051	69.76	.028	67.14	.018	65.79
$\widehat{Y}_{opt_D}$			.014	102.29	.014	102.28	.014	102.28
$\widehat{Y}_{\pi}$	25	706.662	.002	100	.002	100	.002	100
$\widehat{Y}_{ES}$			-.065	76.43	-.079	70.31	-.085	69.20
$\widehat{Y}_{opt_D}$			.001	102.17	.001	102.15	.001	102.15
$\widehat{Y}_{\pi}$	30	848.241	-.101	100	-.101	100	-.101	100
$\widehat{Y}_{ES}$			-.114	76.70	-.103	70.25	-.116	72.25
$\widehat{Y}_{opt_D}$			-.101	101.61	-.101	101.63	-.101	101.62
$\widehat{Y}_{\pi}$	35	991.274	-.019	100	-.019	100	-.019	100
$\widehat{Y}_{ES}$			-.046	86.77	-.058	76.45	-.069	78.05
$\widehat{Y}_{opt_D}$			-.022	102.08	-.022	102.07	-.022	102.07
$\widehat{Y}_{\pi}$	40	1136.184	.004	100	.004	100	.004	100
$\widehat{Y}_{ES}$			.033	86.07	.037	78.72	.028	82.00
$\widehat{Y}_{opt_D}$			.004	102.68	.004	102.68	.004	102.68
$\widehat{Y}_{\pi}$	45	1275.299	.056	100	.056	100	.056	100
$\widehat{Y}_{ES}$			.034	84.84	.045	80.34	.040	82.14
$\widehat{Y}_{opt_D}$			.053	101.51	.053	101.50	.053	101.50
$\widehat{Y}_{\pi}$	50	1415.947	-.016	100	-.016	100	-.016	100
$\widehat{Y}_{ES}$			-.007	91.21	.008	88.68	.001	89.92
$\widehat{Y}_{opt_D}$			-.015	102.86	-.015	102.86	-.015	102.86

Tabla 2.6: SR % y ER % de los estimadores comparados. m: Número de conglomerados. $\overline{n}$: tamaño promedio de los conglomerados sobre B = 1000 repeticiones. Métodos de calibración: lineal, raking y logit. Variable principal: Renta. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel hogar: Nivel de estudios del sustentador principal. Variables auxiliares a nivel persona: género y nivel educativo. Población EPF. Muestreo de Midzuno.

			LINEAR		RAKING		LOGIT	
	m	$\overline{n}$	SR %	ER %	SR %	ER %	SR %	ER %
$\widehat{Y}_{\pi}$	25	48.058	.385	100	.385	100	.385	100
$\widehat{Y}_{ES}$			.148	100.86	-.009	96.09	-.026	97.10
$\widehat{Y}_{opt_D}$			.090	133.52	.035	132.71	.038	132.84
$\widehat{Y}_{\pi}$	50	96.576	-.134	100	-.134	100	-.134	100
$\widehat{Y}_{ES}$			.322	126.02	.163	125.20	.183	125.39
$\widehat{Y}_{opt_D}$			.265	139.46	.241	139.19	.243	139.27
$\widehat{Y}_{\pi}$	75	145.027	.251	100	.251	100	.251	100
$\widehat{Y}_{ES}$			.352	136.58	.250	133.92	.258	134.86
$\widehat{Y}_{opt_D}$			.318	146.04	.307	145.92	.308	145.99
$\widehat{Y}_{\pi}$	100	193.04	.072	100	.072	100	.072	100
$\widehat{Y}_{ES}$			.101	146.46	.010	145.36	.021	145.65
$\widehat{Y}_{opt_D}$			.098	147.43	.085	147.64	.086	147.62
$\widehat{Y}_{\pi}$	150	289.01	.102	100	.102	100	.102	100
$\widehat{Y}_{ES}$			.301	139.51	.231	140.03	.240	139.97
$\widehat{Y}_{opt_D}$			.175	146.31	.166	146.41	.167	146.40
$\widehat{Y}_{\pi}$	200	385.727	-.124	100	-.124	100	-.124	100
$\widehat{Y}_{ES}$			-.005	138.40	-.056	137.80	-.050	137.94
$\widehat{Y}_{opt_D}$			-.061	139.78	-.067	139.70	-.066	139.71
$\widehat{Y}_{\pi}$	250	481.883	-.085	100	-.085	100	-.085	100
$\widehat{Y}_{ES}$			.022	132.57	-.022	133.13	-.017	133.06
$\widehat{Y}_{opt_D}$			.089	136.55	.086	136.56	.086	136.56

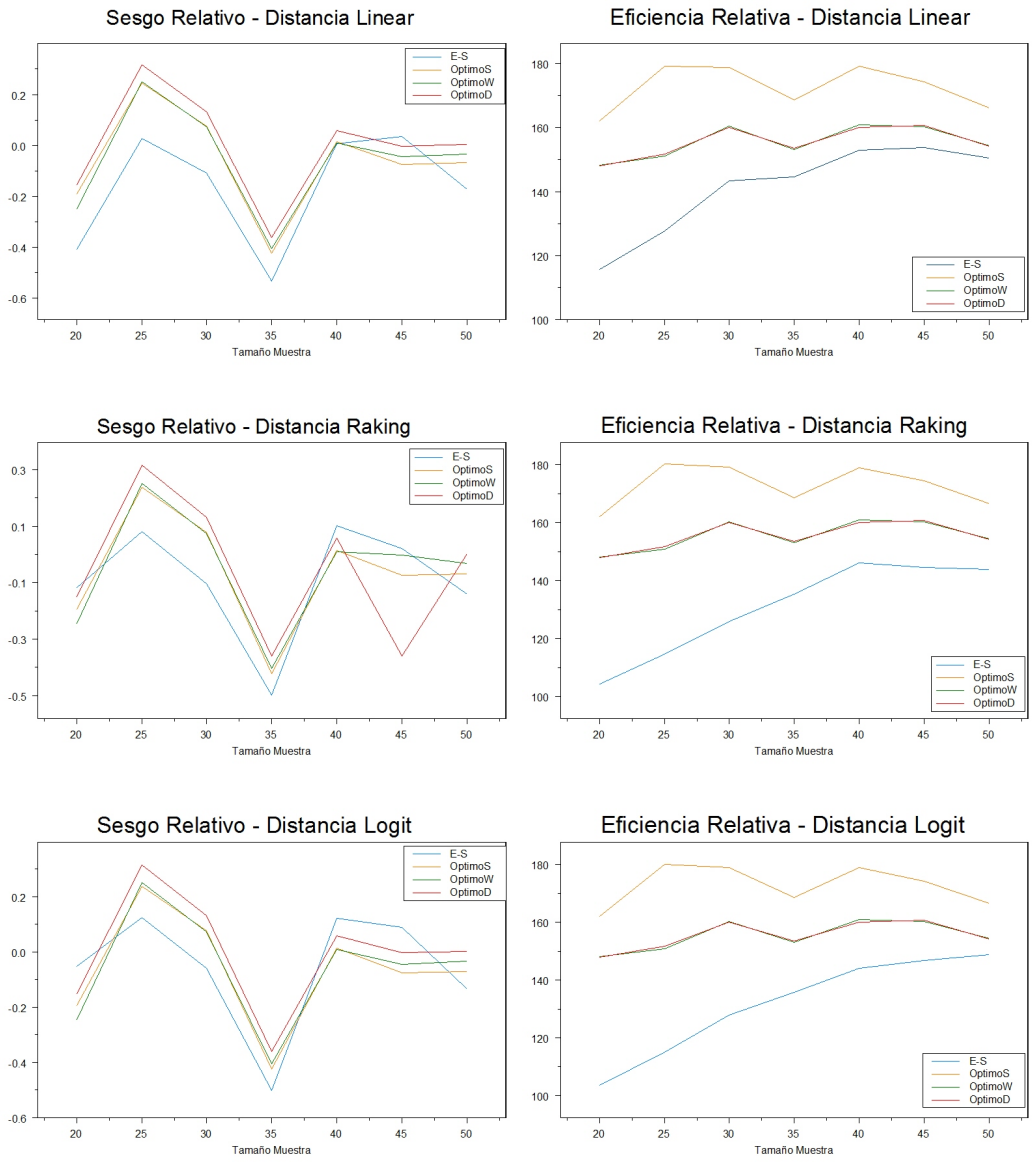


Figura 2.1: Comparativa del sesgo relativo y la eficiencia relativa para los cuatro estimadores por distancia y tamaño de muestra. Variable principal: Sci. future. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel. Población PISA-ESPAÑA. Muestreo aleatorio simple

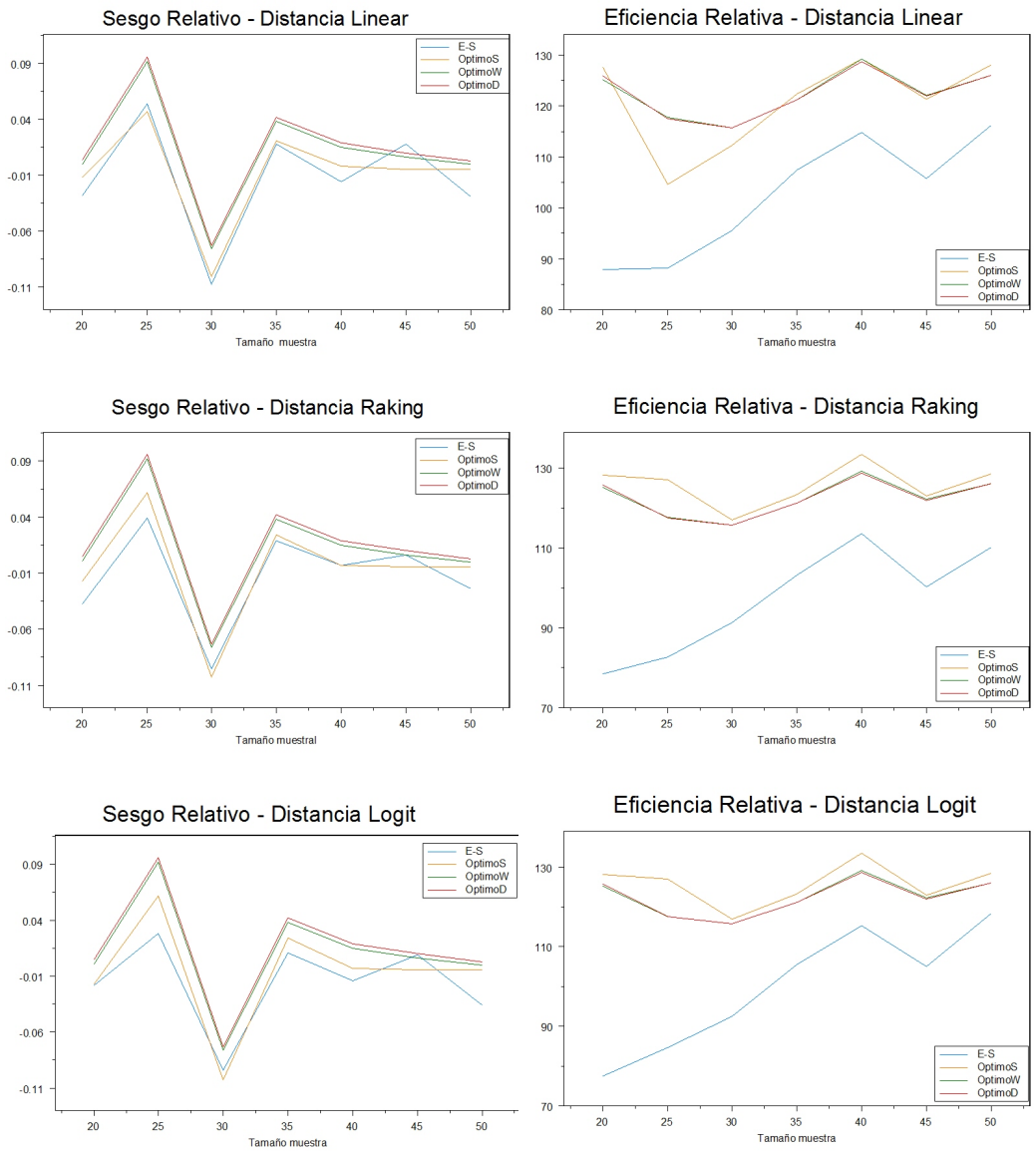


Figura 2.2: Comparativa del sesgo relativo y la eficiencia relativa para los cuatro estimadores por distancia y tamaño de muestra. Variable principal: puntuación en matemáticas. Variables auxiliares a nivel centro: tipo de escuela y el tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo aleatorio simple

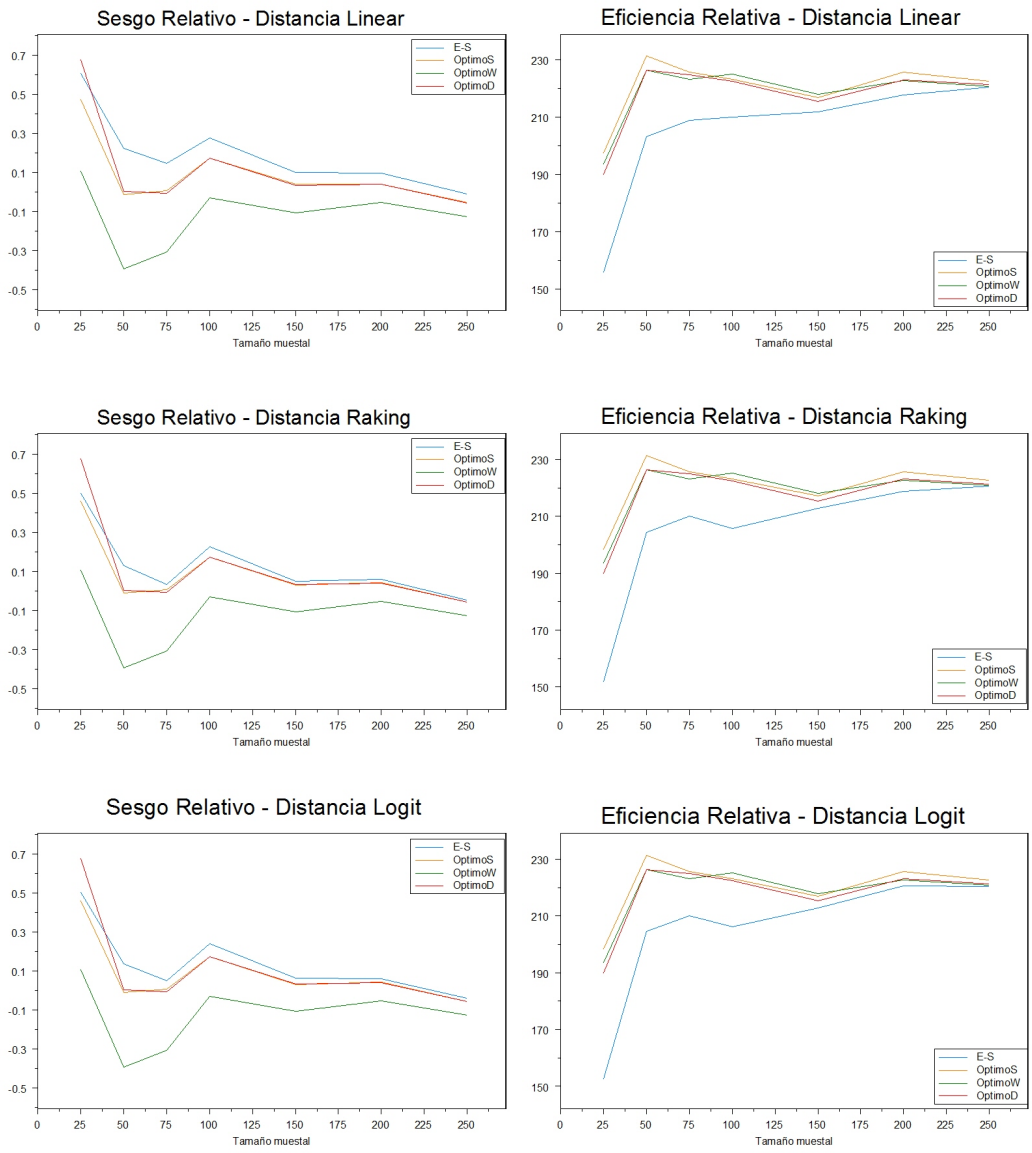


Figura 2.3: Comparativa del sesgo relativo y la eficiencia relativa para los cuatro estimadores por distancia y tamaño de muestra. Variable principal: Renta. Variables auxiliares a nivel hogar: Nivel de estudios del sustentador principal. Variables auxiliares a nivel persona: género y nivel educativo. Población EPF. Muestreo aleatorio simple

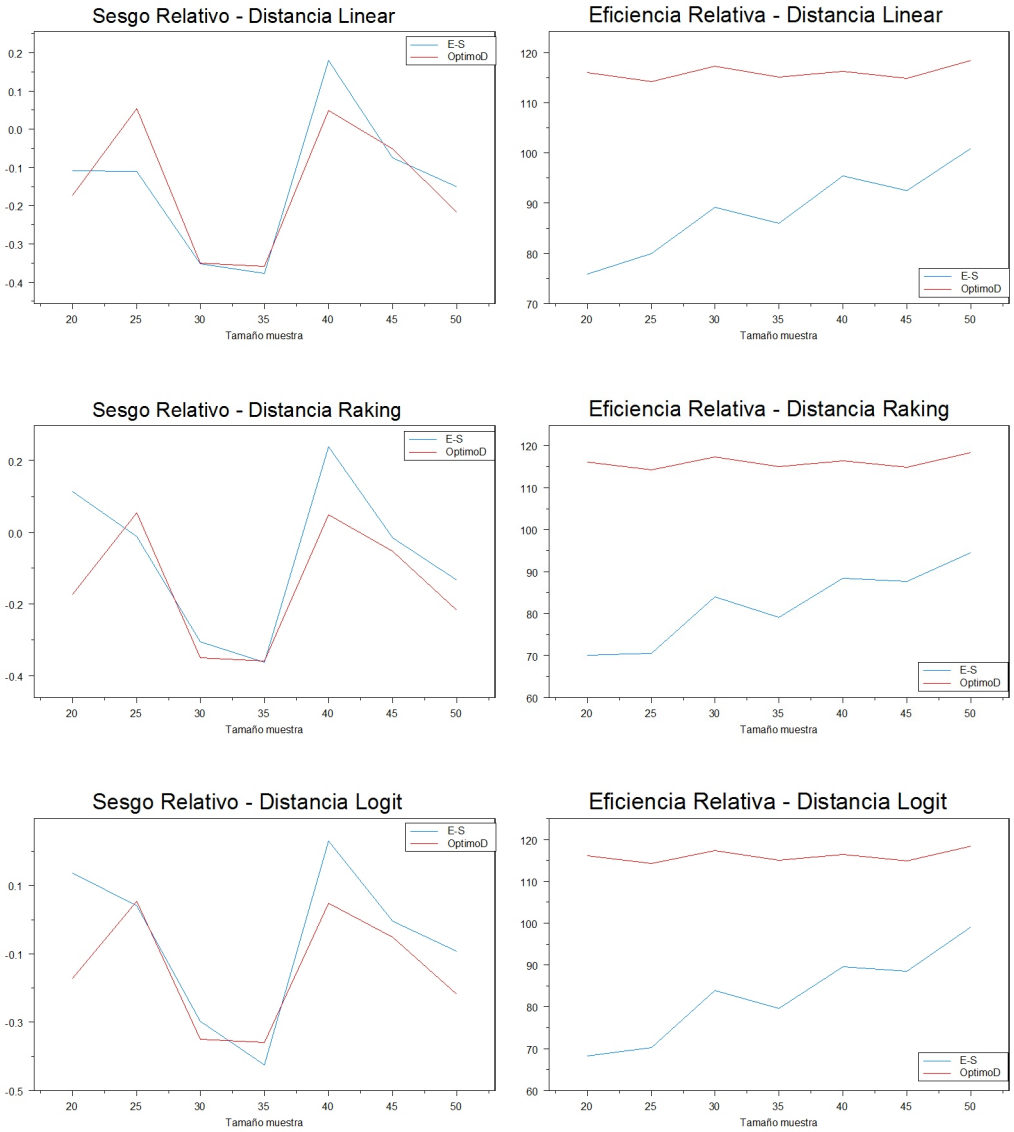


Figura 2.4: Comparativa del sesgo relativo y la eficiencia relativa para los estimadores $\widehat{Y}_{ES}$ y $\widehat{Y}_{opt_D}$ por distancia y tamaño de muestra. Variable principal: Sci. future. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo de Midzuno.

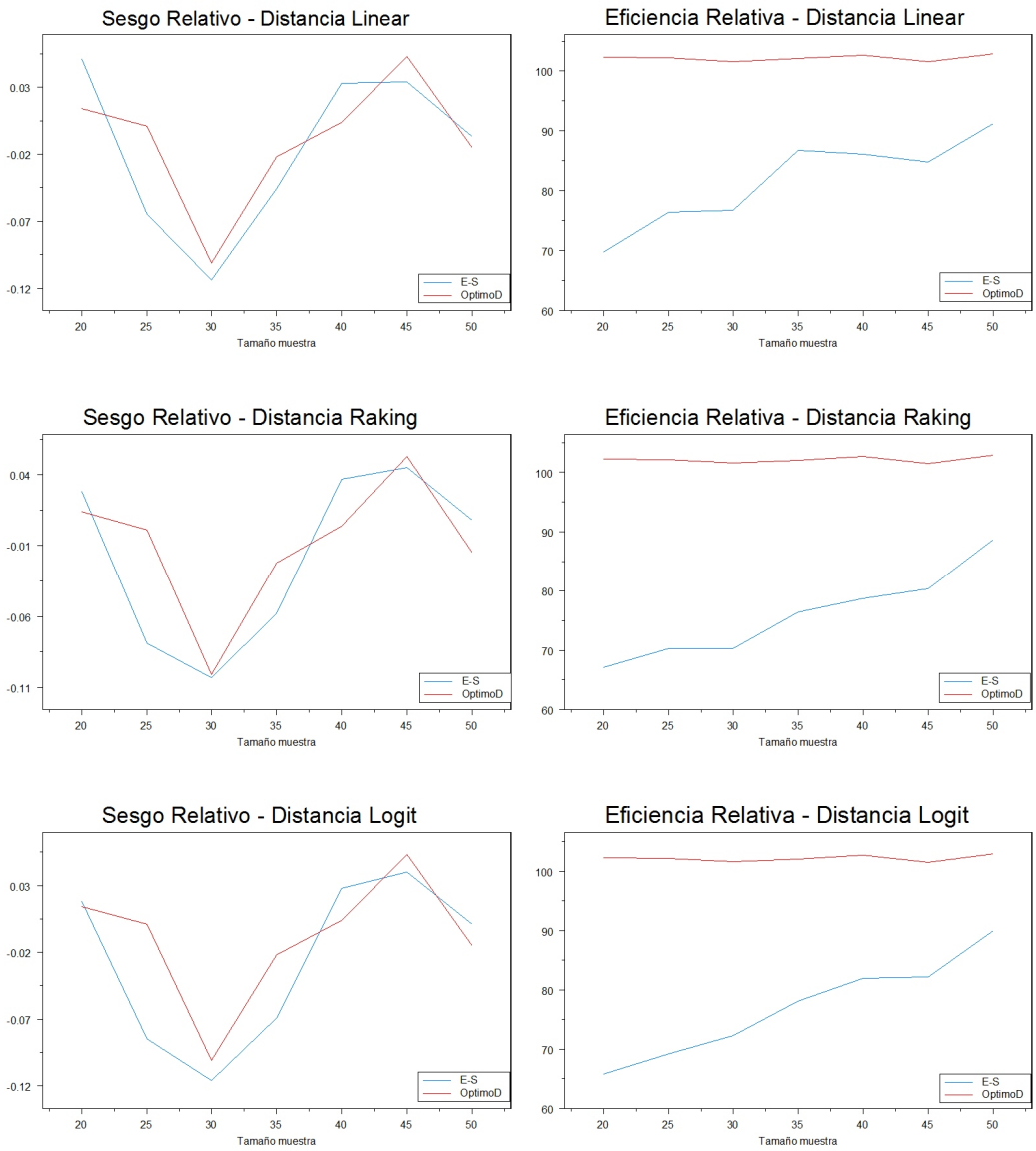


Figura 2.5: Comparativa del sesgo relativo y la eficiencia relativa para los estimadores $\widehat{Y}_{ES}$ y $\widehat{Y}_{opt_D}$ por distancia y tamaño de muestra. Variable principal: puntuación en matemáticas. Variables auxiliares a nivel centro: Tipo de escuela y tipo de comunidad. Variables auxiliares a nivel estudiante: sexo, nivel educativo madre, nivel de estudios del padre y de mayor nivel educativo. Población PISA-ESPAÑA. Muestreo de Midzuno.

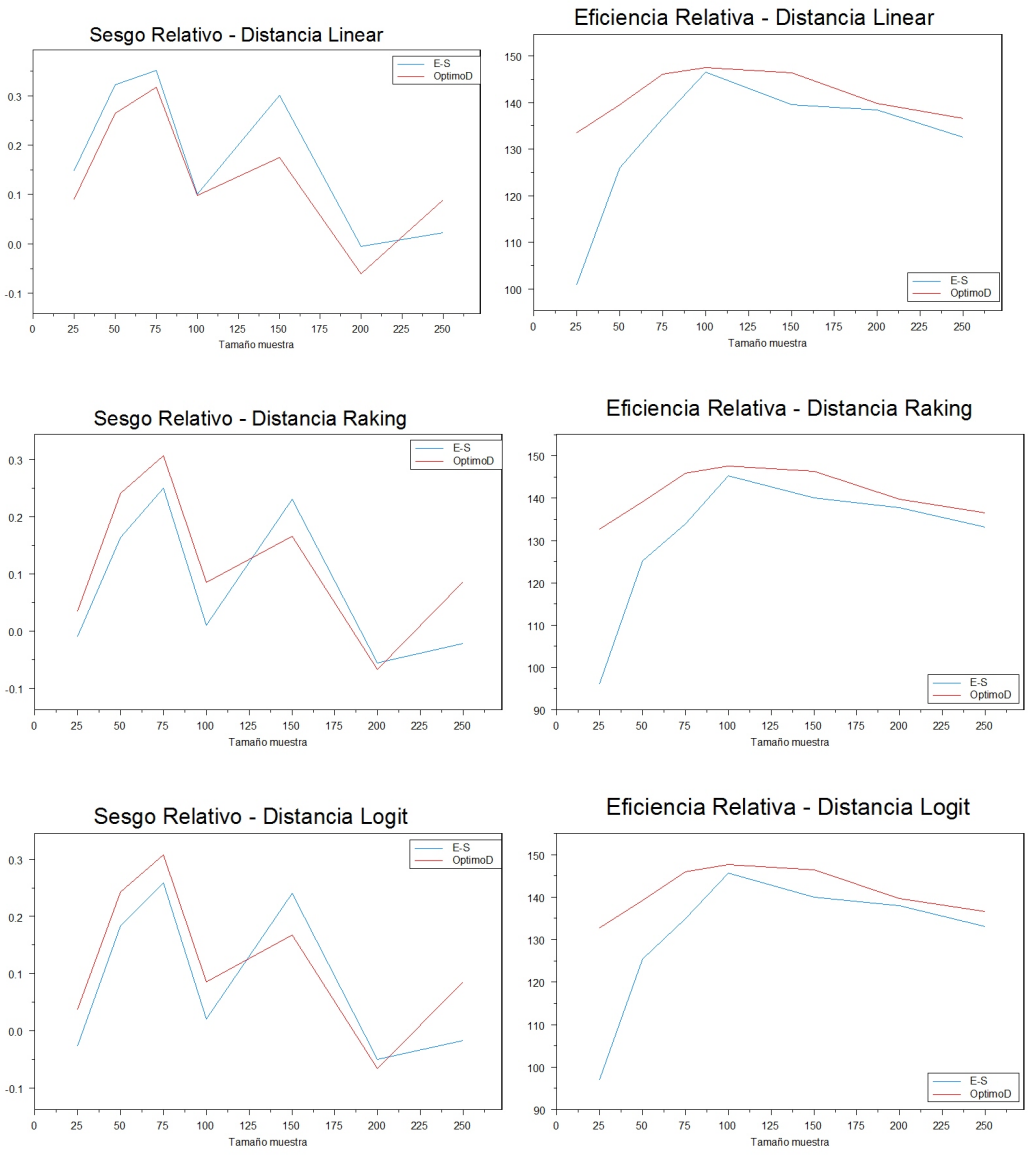


Figura 2.6: Comparativa del sesgo relativo y la eficiencia relativa para los estimadores $\widehat{Y}_{ES}$ y $\widehat{Y}_{opt_D}$ por distancia y tamaño de muestra. Variable principal: Renta. Variables auxiliares a nivel hogar: Nivel de estudios del sustentador principal. Variables auxiliares a nivel persona: género y nivel educativo. Población EPF. Muestreo de Midzuno.

"De entre las variables disponibles es conveniente utilizar aquellas que mejoren la estimación."

Särndal y Lundström

CAPÍTULO

3

Información auxiliar en encuestas con falta de respuesta

Si una encuesta tiene elementos faltantes en las respuestas de determinados ítem se considera que tiene "falta de respuesta". Esta falta de respuesta afecta principalmente al tratamiento de la estimación. Uno de los campos de investigación relacionado con el análisis de la falta de respuesta se centra en el análisis del sesgo que produce y en su reducción. En este capítulo vamos a considerar técnicas, basadas en el uso eficaz de la información auxiliar, que reduzcan al mismo tiempo el error de muestreo y el sesgo debido a la falta de respuesta.

En primer lugar se describe el uso de la calibración para el ajuste del sesgo de no respuesta y se definen cuatro indicadores que nos permitirán elegir qué variables auxiliares son más eficaces para construir el vector auxiliar. Posteriormente finalizamos con un estudio de simulación con datos reales (PISA 2006) en el que mostramos como la elección apropiada del vector auxiliar permite reducir el sesgo de no repuesta.

3.1 Introducción

El ajuste de falta de respuesta es un término general que hace referencia a los diversos modelos estadísticos utilizados para hacer frente a la falta de respuesta una vez que ha ocurrido, es decir, después de la aceptación del hecho de que algunos datos deseados faltarán. Se entiende como "ajuste" la realización de cambios en el procedimiento de estimación original o "ideal", es decir, en aquel destinado al uso en el caso ideal de que se tenga el 100 % de respuesta. Los principales métodos para el ajuste de no respuesta son la imputación y la ponderación.

La imputación implica reemplazar los valores perdidos por valores próximos. El estadístico puede optar por utilizar la imputación sólo para el ítem con falta de respuesta y luego tratar la no respuesta por reponderación, enfoque ITIMP, o imputar los valores para el ítem con falta de respuesta, así como para las unidades con falta de respuesta, enfoque UNIMP.

La ponderación altera los pesos de los encuestados, en comparación con los pesos que se habrían utilizado en el caso de tener el 100 % de respuesta. Dado que la falta de respuesta supone una perdida de observaciones, la ponderación implicará un aumento del peso de todos, o casi todos, los elementos que responden.

Investigaciones como las de Eltinge y Yansaneh (1997)[13], Kalton y Flores-Cervantes (2003)[26], y Thomsen, Kleven, Wang y Zhang (2006)[66] hacen hincapié en la ponderación de encuestas con falta de respuesta y en especial en la selección de las mejores variables auxiliares para tratarla. Rizzo, Kalton y Brick (1996)[50] sugieren que la elección de las variables auxiliares es más importante, incluso, que la elección de la metodología de ponderación.

En este capítulo nos centramos en este método. Más concretamente, el capítulo está dedicado a llevar a la práctica, con datos procedentes del estudio PISA descritos en el capítulo 2, los métodos de selección de variables auxiliares propuestos por Särndal y Lundstöm (2010)[55] para el ajuste de no respuesta.

Primeramente se describe el estimador de calibración a usar en presencia de no respuesta. Se analizan las propiedades de los nuevos pesos de calibración que constituyen los factores de ajuste de no respuesta: promediar la inversa de la tasa de respuesta y su varianza va a permitir relacionar el sesgo de no respuesta con la

información auxiliar. De esta relación se pueden deducir indicadores que permitan seleccionar las variables auxiliares que más ayuden a reducir el sesgo de no respuesta. Estos criterios de selección son objetivos, basados exclusivamente en reducir el sesgo de no respuesta. En la última parte del capítulo, las variables auxiliares que los criterios sugieren son usadas para obtener nuevas estimaciones de algunas variables del informe PISA.

Para concluir esta introducción, planteamos la notación que será la que se mantendrá en todo el texto. Sea una muestra probabilística s extraída de una población $\mathcal{U} = \{1, \ldots, k, , \ldots, N\}$ en la que se conocen las probabilidades de inclusión de primer orden $\pi_k = Pr(k \in s) > 0$, para cualquier elemento k y los pesos del diseño $d_k = 1/\pi_k$. Además, ahora suponemos que en nuestra encuesta hay falta de respuesta, por lo que tenemos un subconjunto r de s ($r \subset s \subset \mathcal{U}$) generado de forma desconocida y no vacío. Entonces se define la tasa de respuesta (para un diseño ponderado) como:

$$P = \frac{\sum_r d_k}{\sum_s d_k} \tag{3.1}$$

Por tanto, si $k \in r$, y_k está disponible y por contra si $k \in \mathcal{U} - r$, y_k no lo estará.

3.2 La calibración para el ajuste de sesgo de no respuesta

La información auxiliar

La información auxiliar aparecerá de dos formas, a las que le corresponderá dos tipos de vectores, $\mathbf{x}_k^*$ y $\mathbf{x}_k^o$. La *información auxiliar a nivel poblacional* se denota por $\mathbf{x}_k^*$, vector cuyos valores son conocidos para cada $k \in \mathcal{U}$. Y por tanto $\sum_{\mathcal{U}} \mathbf{x}_k^*$ es el total poblacional conocido. Como ya se trató en los capítulos anteriores, el total poblacional $\sum_{\mathcal{U}} \mathbf{x}_k^*$ se puede importar de fuentes externas, con lo que sólo debemos conocer los elementos del vector $\mathbf{x}_k^*$ para cada $k \in s$. La *información auxiliar a nivel muestral* se denota por $\mathbf{x}_k^o$, vector cuyos valores son conocidos para cada $k \in s$. El total $\sum_{\mathcal{U}} \mathbf{x}_k^o$ es desconocido pero puede ser estimado sin sesgo por

$\sum_s d_k \mathbf{x}_k^o$. Denotamos por $\mathbf{x}_k$ al vector auxiliar formado por la combinación de los vectores $\mathbf{x}_k^*$ y $\mathbf{x}_k^o$, y a $\mathbf{X}$ al formado por los totales.

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix} \tag{3.2}$$

$$\mathbf{X} = \begin{pmatrix} \sum_{\mathcal{U}} \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^o \end{pmatrix} \tag{3.3}$$

Resumiendo, para cada $k \in \mathcal{U}$ conocemos π_k; y_k es conocido para $k \in r$; el componente $\mathbf{x}_k^*$ de $\mathbf{x}_k$ aporta la información auxiliar a nivel poblacional y el componente $\mathbf{x}_k^o$ de $\mathbf{x}_k$ aporta la información auxiliar a nivel muestral.

De los posibles vectores auxiliares que se pueden formar con la ayuda de las variables auxiliares procedentes de registros administrativos, datos de encuestas o de otras fuentes, se tratará de identificar aquellos con más posibilidades de reducir el sesgo de no-respuesta.

El estimador de calibración

El estimador de calibración del total $Y = \sum_{\mathcal{U}} y_k$ con falta de respuesta, se calcula para y_k, con $k \in r$, mediante

$$\hat{Y}_{cal} = \sum_r w_k y_k \tag{3.4}$$

con pesos $w_k = d_k\{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k\}$ que calibran en ambos tipos de información, es decir, $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, por tanto $\sum_r w_k \mathbf{x}_k^* = \sum_{\mathcal{U}} \mathbf{x}_k^*$ y $\sum_r w_k \mathbf{x}_k^o = \sum_s d_k \mathbf{x}_k^o$. Asumiendo, por razones computacionales, la no singularidad de $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$, es posible definir el estimador de calibración como $\hat{Y}_{cal} = \sum_r w_k y_k$, con $w_k = d_k \nu_k$ donde $\nu_k = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$. Estos pesos satisfacen la ecuación $\sum_r d_k \nu_k \mathbf{x}_k = \mathbf{X}$, donde $\mathbf{X}$ es la matriz acumulada descrita en (3.3).

Otro estimador de calibración (Särndal y Lundström, 2005[54]) relacionado con el anterior, basado en el mismo vector de dos niveles $\mathbf{x}_k$, pero que sólo calibra a nivel de la muestra es:

$$\tilde{Y}_{cal} = \sum_r m_k y_k, \tag{3.5}$$

donde

$$m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k, \tag{3.6}$$

cumpliendo $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$.

El vector auxiliar $\mathbf{x}_k$ descrito en (3.2) cumple dos propósitos: conseguir una varianza y un sesgo de no respuesta bajos. Por lo general $\hat{Y}_{cal}$ es preferible a $\tilde{Y}_{cal}$ ya que éste tiene en cuenta el total poblacional (conocido) $\sum_{\mathcal{U}} \mathbf{x}_k^*$. Pero desde la perspectiva del sesgo, prácticamente no hay diferencias entre $\hat{Y}_{cal}$ y $\tilde{Y}_{cal}$ por lo que es preferible utilizar el segundo de ellos. De hecho, la diferencia entre el sesgo de $N^{-1}\hat{Y}_{cal}$ y el de $N^{-1}\tilde{Y}_{cal}$ es del orden n^{-1}, lo que en la práctica tiene pocas consecuencias, incluso para tamaños de muestra pequeños.

Una expresión alternativa para (3.5) es

$$\tilde{Y}_{cal} = (\sum_s d_k \mathbf{x}_k)'\mathbf{B}_{\mathbf{x}|r;d}, \tag{3.7}$$

donde

$$\mathbf{B}_{\mathbf{x}|r;d} = \mathbf{B}_{\mathbf{x}} = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_r d_k \mathbf{x}_k y_k, \tag{3.8}$$

es el vector de coeficientes de regresión derivado del ajuste de mínimos cuadrados basado en los datos $(y_k, \mathbf{x}_k)$ para $k \in r$.

Los factores de ajuste

El factor de ajuste de no respuesta $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ expande el peso del diseño d_k. Podemos tomar m_k como el valor de una variable, definida para un caso particular (r, s) y cuya elección de $\mathbf{x}_k$ es independiente de todas las variables y de interés, y se puede calcular para todo $k \in s$ (pero se utiliza en $\tilde{Y}_{cal}$ sólo para $k \in r$). Usando (3.6), se obtiene:

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k,$$

$$\sum_r d_k m_k = \sum_s d_k,$$

$$\sum_r d_k m_k^2 = \sum_s d_k m_k.$$

Se definen los promedios ponderados, $\bar{m}_{r;d}$ y $\bar{m}_{s;d}$ como:

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P}, \tag{3.9}$$

$$\bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k}, \tag{3.10}$$

donde P es la tasa de respuesta (3.1). Por tanto, el factor de ajuste medio en $\tilde{Y}_{cal} = \sum_r m_k y_k$ es $1/P$, independiente de la elección del vector auxiliar. Entonces, que sea elegido o no un vector auxiliar para reducir el sesgo dependerá de momentos de orden superior de m_k. La varianza ponderada de m_k se define como:

$$S_m^2 = s_{m|r;d}^2 = \frac{\sum_r d_k (m_k - \bar{m}_{r;d})^2}{\sum_r d_k} = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}). \tag{3.11}$$

y el coeficiente de variación de m_k será:

$$cv_m = \frac{s_m}{\bar{m}_{r;d}} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1}. \tag{3.12}$$

La varianza ponderada de la variable de estudio y viene dada por:

$$S_y^2 = s_{y|r;d}^2 = \frac{\sum_r d_k (y_k - \bar{y}_{r;d})^2}{\sum_r d_k}, \tag{3.13}$$

y la covarianza se define como:

$$cov(y, m) = cov(y, m)_{r;d} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})(y_k - \bar{y}_{r;d}), \tag{3.14}$$

y el coeficiente de correlación es:

$$R_{y,m} = \frac{cov(y, m)}{S_y \cdot S_m}, \tag{3.15}$$

satisfaciendo $-1 \leq R_{y,m} \leq 1$.

Desde el punto de vista del sesgo de no respuesta, los estimadores $\widetilde{Y}_{ful}$ (con respuesta completa, no computable), $\widetilde{Y}_{cal}$ (estimador de calibración basado en un

vector $\mathbf{x}_k$) y $\widetilde{Y}_{exp}$ (basado en el vector auxiliar $x_k = 1$ con no respuesta, computable) ocupan la posición que muestran la Figura 3.1.

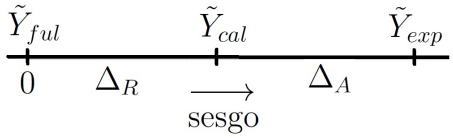


Figura 3.1: Comparativa de estimadores según sesgo

El objetivo entonces será encontrar vectores x_k que hagan Δ_A grande. Exactamente, Δ_A se define como

$$\Delta_A = \frac{\tilde{Y}_{exp} - \tilde{Y}_{cal}}{\hat{N}}, \tag{3.16}$$

y se demuestra (véase Särndal y Lundstöm (2010)[55]) que la desviación Δ_A se puede escribir como:

$$\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}, \tag{3.17}$$

lo cual permite definir los indicadores que vemos a continuación.

Definición de indicadores

Esta desviación nos va a permitir controlar cual de las variables x son las más efectivas. Podemos factorizar Δ_A / S_y como

$$\Delta_A / S_y = -R_{y,m} \times cv_m. \tag{3.18}$$

Dos factores multiplicativos simples determinan Δ_A / S_y: el coeficiente de variación cv_m, independiente de y_k, calculado solamente sobre $\mathbf{x}_k$, y el coeficiente de correlación (positivo o negativo) $R_{y,m}$. Otra tipo de factorización de Δ_A / S_y es

$$\Delta_A / S_y = F \times R_{y,\mathbf{x}} \times cv_m. \tag{3.19}$$

donde $R_{y,\mathbf{x}} = \sqrt{R^2_{y,\mathbf{x}}}$ es el coeficiente de correlación múltiple entre y y $\mathbf{x}$, $R^2_{y,\mathbf{x}}$ es la proporción de la y-varianza S^2_y explicada por el predictor $\mathbf{x}$, tal que $|R_{y,m}| \leq R_{y,\mathbf{x}}$ para cualquier vector x y variable y y $F = \frac{-R_{y,m}}{R_{y,\mathbf{x}}}$ con $-1 \leq F \leq 1$.

Como cv_m y $R_{y,\mathbf{x}}$ son términos no negativos, mientras que $R_{y,m}$ y F pueden tener cualquier signo (o ser cero), se toma:

$$|\Delta_A|/S_y = |R_{y,m}| \times cv_m = |F| \times R_{y,\mathbf{x}} \times cv_m. \tag{3.20}$$

A partir de (3.18) se deduce que $0 \leq |\Delta_A|/S_y \leq cv_m$ para cualquier variable y. En cambio la desigualdad $|\Delta_A|/S_y \leq R_{y,\mathbf{x}} \times cv_m$ si depende de la variable y. Notar que todas las cantidades S_y, cv_m, $R_{y,\mathbf{x}}$, $R_{y,m}$ y F se pueden calcular fácilmente en una encuesta. cv_m y $R_{y,\mathbf{x}}$ aumentan al introducir nuevas variables en el vector aunque $R_{y,m}$ puede no aumentar. Sin embargo, aunque se puede calcular la desviación Δ_A para cualquier vector x y cualquier conjunto (s, r), ésta no proporciona información de la proporción de sesgo. Principalmente Δ_A proporciona herramientas para el cálculo de indicadores que proporcionen una comparación de las variables auxiliares. Las definiciones de cada indicador se dan a continuación.

A partir de (3.18), se define

$$H_0 = \Delta_A/S_y = -R_{y,m} \times cv_m. \tag{3.21}$$

La relación entre el promedio de H_0 y la desviación media $\tilde{Y}_{cal} - Y$ (que mide el sesgo de $\tilde{Y}_{cal}$) es lineal y casi perfecta cuando se produce cambios en las variables auxiliares, independientemente de la distribución de respuesta que genera r en s. Ya que H_0 puede tener cualquier signo, es práctico trabajar con su valor absoluto, lo que proporciona el indicador H_1:

$$H_1 = |\Delta_A|/S_y = |R_{y,m}| \times cv_m. \tag{3.22}$$

A partir de (3.19) y (3.20), se definen H_2 y H_3 como:

$$H_2 = R_{y,\mathbf{x}} \times cv_m; \tag{3.23}$$

$$H_3 = cv_m. \tag{3.24}$$

De estas alternativas, H_1 está motivada por su relación directa con Δ_A, que se desea que sea la mayor posible, para una determinada variable y. Una razón de peso para considerar H_3 es su independencia de todas las variables y de la encuesta. El

indicador de H_2 es una alternativa *ad-hoc*, aunque H_2 depende del coeficiente de correlación múltiple $R_{y,\mathbf{x}}$, y por tanto es menos apropiado que H_1 debido a que la razón $F = -R_{y,m}/R_{y,\mathbf{x}}$ puede variar considerablemente de un vector x a otro. En contra de lo que sucede para H_1 tanto H_2 y H_3 aumentan cuando se añaden más variables x al vector $\mathbf{x}$.

Construcción del vector auxiliar

La información auxiliar se transmite por un vector auxiliar $\mathbf{x}_k$ cuyo valor es conocido por todos los elementos de la respuesta, $k \in r$, pero existe información para un conjunto mayor que r. Esta información contribuirá tanto a la protección contra la falta de respuesta como a la reducción de la varianza.

Las fuentes de información oficiales disponibles proporcionan una rica fuente de información auxiliar, en particular para las encuestas sobre individuos y hogares. Estos registros contienen muchas variables x potenciales entre las que elegir para formar diferentes vectores $\mathbf{x}$. Los indicadores (3.21) a (3.24) proporcionan herramientas computacionales para la obtención de un orden de preferencias de los vectores $\mathbf{x}$, con el objetivo de reducir tanto como sea posible el sesgo que queda en el estimador de calibración. Podemos encontrar dos posibles escenarios:

- Caso 1: Debido a que el sesgo que produce el estimador de calibración depende de la variable y (algunas variables y son más propensas que otras al sesgo), se selecciona una variable y específica de la encuesta y buscamos identificar un vector $\mathbf{x}$ que reduzca el sesgo de esta variable tanto como sea posible. Para este propósito se utiliza el indicador dependiente de la variable y, $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times cv_m$, y se elige el vector $\mathbf{x}$ a fin de que H_1 sea lo mayor posible. Una alternativa *ad hoc* es utilizar el indicador $H_2 = R_{y,\mathbf{x}} \times cv_m$ y tratar de que sea lo mayor posible.

- Caso 2: El objetivo es identificar un vector $\mathbf{x}$ eficiente para todas o la mayoría de las variables y de la encuesta. Este caso sugiere tomar el indicador $H_3 = cv_m$ para elegir aquél vector $\mathbf{x}$ que maximize H_3. Notar que la varianza de los factores de ajuste $S_m^2 = H_3^2/P^2$ también se puede utilizar en el mismo sentido, es decir, que la variable m_k se puede considerar como un predictor

de la inversa de la probabilidad de respuesta desconocida y que la elección del vector $\mathbf{x}$ a partir de S_m^2 provoca una reducción del sesgo en el estimador de calibración, con independencia de la variable y.

Para cada escenario se pueden distinguir dos procedimientos:

- *Se toman todos los vectores propuestos como variables auxiliares*: Se selecciona una lista de vectores $\mathbf{x}$ candidatos en base a un indicador. Se calcula el indicador elegido para cada vector $\mathbf{x}$, y se conforma el vector con aquellos vectores $\mathbf{x}$ cuyo valor del indicador sea más alto. El vector $\mathbf{x}$ resultante puede no ser el mismo para H_1 (que refiere a una variable y específica) como para H_3 (que busca un compromiso para todos las variables y en la encuesta).

- *Se eligen las variables auxiliares en un procedimiento paso a paso*: Supongamos que existe una colección disponible de variables x. A partir de ellas construimos el vector $\mathbf{x}$ paso a paso seleccionando de entre las variables x disponibles una variable cada vez. Este proceso se realizará a partir de los valores del indicador elegido, que indica la inclusión (o exclusión) de una variable x dada en un paso determinado. Los indicadores H_1, H_2 y H_3, en general, no proporcionan la misma selección de variables.

  Por ejemplo, consideremos dos $\mathbf{x}$-vectores, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$, de manera que $\mathbf{x}_{2k}$ esta compuesto de $\mathbf{x}_{1k}$ y un vector adicional $\mathbf{x}_{+k}$: $\mathbf{x}_{2k} = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$, de manera que la transición de $\mathbf{x}_{1k}$ a $\mathbf{x}_{2k}$ incrementa el valor de H_2 y H_3. En cada paso del procedimiento de selección (hacia adelante) seleccionamos la variable que tenga el mayor aumento de H_2 o H_3, aunque la incorporación no garantiza un aumento del valor del indicador H_1 (el más apropiado en este caso).

3.3 Aplicación a la selección de la información auxiliar en la encuesta PISA

En esta sección, evaluamos el comportamiento empírico de los indicadores propuestos para la elección de aquellas variables auxiliares que reducen el sesgo de la

variable de interés a través de un estudio de simulación. Para ello partimos de los supuestos descritos en la sección 3.2.3 (Caso 1, Procedimiento 2) es decir partimos de la elección de una variable y específica de la encuesta para la que queremos identificar un vector $\mathbf{x}$ que reduzca el sesgo de no respuesta. Para ello vamos a tomar una colección disponible de variables x y a partir de ellas construir un vector $\mathbf{x}$ seleccionando de entre las variables x disponibles una variable paso a paso. Este proceso se realizará partiendo de los valores del indicador elegido (el valor máximo de cada indicador).

Población de estudio

La población utilizada en este estudio de simulación es la del Programme for International Student Assessment (PISA) descrita en el capítulo 2. El conjunto de microdatos del informe PISA-España contiene información sobre las pruebas que se llevaron a cabo en 686 escuelas, con la participación de 19.604 estudiantes. Se consideraron las unidades sin datos faltantes para las variables auxiliares (variables x) y con datos faltantes para las variables de interés, obteniendo de este modo una población de tamaño $N = 17{,}463$ estudiantes agrupados en $N_I = 686$ centros. Hay que tener en cuenta que la población difiere de la vista en el capítulo 2 ya que en esta se ha tenido en cuenta tres nuevas variables.

Variables de estudio

Para este estudio se han seleccionado tres variables de interés, las dos primeras pertenecientes al primer estudio: la variable cualitativa dicotómica "Estudiar una carrera de ciencias en el futuro" , la variable cuantitativa "puntuación en matemáticas", y como tercera una dicotómica "recibe clases particulares de ciencias".

Variables auxiliares

Como información auxiliar a nivel unidad elegimos el género (masculino y femenino), el nivel educativo de la madre, el nivel educativo del padre y el mayor nivel educativo de los padres con las categorías: sin estudios, educación primaria, educación secundaria y estudios superiores, y tres variables dicotómicas relacionadas

con si el padre, la madre o alguien en el hogar tiene un empleo relacionado con las ciencias (Si, No). Todas las variables son del tipo $\mathbf{x}_k^o$.

Las variables auxiliares se han enumerado como: sexo (4), nivel educativo padre (5), nivel de estudios del madre (6), mayor nivel educativo (7), el trabajo del padre está relacionado con la ciencia (8), el trabajo de la madre está relacionado con la ciencia (9) y el trabajo de algún miembro del hogar está relacionado con la ciencia (10).

Elección del vector auxiliar

Para la elección del vector auxiliar, dada una variable de interés, se han realizado 1000 repeticiones con muestras de centros de diferentes tamaños, calculando en cada repetición todos los indicadores H_0, H_1, H_2 y H_3. En cada repetición, cada indicador sugiere que sea seleccionada una variable. Finalmente, las variables son ordenadas de acuerdo a la frecuencia con la que han sido seleccionadas.

Se han realizado dos tipos de simulaciones:

- En la primera, la variable de interés es "Puntuación en matemáticas". La falta de respuesta se ha forzado de manera aleatoria para los valores 20 %, 30 %, 40 % y 50 %, multiplicando la variable de interés original, con respuesta completa, por un vector de unos, creado de manera aleatoria, con una falta de respuesta forzada del 20 %, 30 %, 40 % y 50 %. Los tamaños de muestra de centros son: 10, 20, 30, 40, 50 y 100.

- En la segunda, las variables de interés son "Carrera ciencias" y "Clases particulares". La falta de respuesta real es del 11,5 % y 23.92 %, respectivamente. Los tamaños de muestra de centros son: 10, 20, 30, 40, 50 y 100.

Se ha llevado a cabo la selección "hacia adelante" de la siguiente manera: Primeramente se ha seleccionado el vector auxiliar trivial $\mathbf{x}_k = 1$, y se ha calculado el valor del indicador para cada una de las P variables auxiliares; seleccionando la variable que produce el mayor valor para cada indicador, este proceso se ha realizado una vez para cada simulación, eligiendo de entre todas las variables aquella que con mayor frecuencia era seleccionada en las 1000 simulaciones. En el segundo paso, el valor del indicador se calcula con las $P - 1$ variables restantes formando

matrices cuyas columnas contienen la variable seleccionada en el paso 1 y una de las variables restantes. Al igual que anteriormente se selecciona la variable que da el mayor valor para cada indicador y así sucesivamente en los siguientes pasos. Hay que destacar que el orden de selección es diferente para cada indicador.

Estimadores y medidas de referencia

Una vez seleccionadas las variables, se trata de comprobar con las simulaciones, cómo se reduce el sesgo de no respuesta. Se calcula con ese fin la estimación por calibración $\tilde{Y}_{cal}$ con la variable elegida (comenzando con la primera variable auxiliar elegida, siguiendo con la primera y la segunda variables, y así sucesivamente). Además, se calcula la desviación relativa (DR), en el caso de no respuesta generada, definida como:

$$DR = \frac{(\tilde{Y}_{cal} - \tilde{Y}_{ful})}{\tilde{Y}_{ful}} \times 10^2. \tag{3.25}$$

Tablas de resultados

Las Tablas 3.1 y 3.2 muestran el orden en el que las variables auxiliares: sexo(4), nivel educativo padre (5), nivel de estudios del madre (6), mayor nivel educativo (7), el trabajo del padre está relacionado con la ciencia (8), el trabajo de la madre está relacionado con la ciencia (9) y el trabajo de algún miembro del hogar está relacionado con la ciencia (10), han sido seleccionadas por cada indicador con mayor frecuencia, de entre 1000 repeticiones. Las tablas también muestran el número de conglomerados (centros) seleccionados y el promedio de individuos (alumnos) en las 1000 simulaciones. La Tabla 3.1 para la variable principal puntuación en matemáticas, con unas tasas de no respuesta generadas del 11.5 % y del 23.92 % y la Tabla 3.2 para las variables principales: Carrera ciencias (con una tasa de no respuesta real del 11.5 %) y Clases particulares (con una tasa de no respuesta real del 23.92 %)

La Tabla 3.3 muestran la estimación por calibración, la desviación relativa (DR) y los indicadores H_0 y H_1 H_2 y H_3 para la variable de interés puntuación en matemáticas, con las variables que propone la Tabla 3.1 con el indicador H_3 y para las tasas de respuesta antes citadas del 11.5 % y del 23.92 %.

El estudio anterior sobre la frecuencia más alta en las repeticiones sobre el orden de selección de la variables se ha llevado a cabo para todas las variables principales y para tasas de no respuesta de entre el 20 % hasta el 50 %, y se incluyen algunos resultados en el anexo como Tablas A.1 y A.2 y las Tablas desde la A.15 a la A.18.

Una vez sugerido ese orden por simulación, los resultados que muestran la relación de la estimación por calibración con el valor del indicador correspondiente y con la desviación relativa, se han incluido en el anexo en las Tablas desde la A3 a la Tabla A10.

Las Tablas de A.11 a A.14 muestran la estimación por calibración, la desviación relativa (DR) y los indicadores H_0 y H_1 H_2 y H_3 para la variable de interés puntuación en matemáticas y las tasas de respuesta del 11.5 % y del 23.92 %, con las variables propuestas por la tabla 3.1.

Comentarios sobre la selección de variables

Todas las tablas muestran, incluso para tamaños de conglomerados pequeños, una cierta estabilización a la hora de elección de aquellas variables auxiliares, sea cual sea el indicador usado, que han de reducir el sesgo de no respuesta. En el indicador H_3, independiente de la variable de interés, destacan aquellas variables auxiliares ligadas al sexo y a la relación de los trabajos de los padres con las ciencias. Para los indicadores dependientes de la variable de interés, aparece además en ocasiones el nivel de estudios del padre.

Para comprobar el comportamiento del estimador de calibración con falta de respuesta para las variables seleccionadas se han calculado diferentes ejemplos para cada una de los tamaños de conglomerados, faltas de respuestas y variables de interés. Para cada ejemplo, como se muestra en la Tabla 3.3, se ha calculado el indicador correspondiente para cada variable paso a paso, teniendo en cuenta que la primera fila hace referencia a la primera variable auxiliar elegida, la segunda fila a la primera más la segunda y así sucesivamente. El estimador de calibración $\tilde{Y}_{cal}$ y la desviación relativa (DR) definida como $\frac{(\tilde{Y}_{cal}-\tilde{Y}_{ful})}{\tilde{Y}_{ful}} \times 10^2$ se muestran también en la Tabla 3.3. Hay que tener en cuenta que en un entorno de encuesta real, DR es desconocida. Otros ejemplos con distintas faltas de respuesta se muestran el Anexo **??**.

En el ejemplo concreto de la Tabla 3.3, con falta de respuesta del 11.5 % y 23.93 % y un tamaño de 100 centros (2500 alumnos), se muestran los diferentes indicadores, de H_0 a H_3. En estos indicadores se observan similares tendencias, por ejemplo para H_1 se muestra una reducción mayor de DR cuando se van acumulando las variables auxiliares hasta que aparece la variable número 5 (Nivel educativo del padre) en el que se observa una desaceleración de la disminución de DR, cuando la tasa de no respuesta es la menor. Cuando la tasa de respuesta es mayor la variable número 5 ocupa el primer lugar. Para H_2 ocurre algo diferente a H_1 si la tasa de respuesta es más alta y algo más similar si la tasa de respuesta es la más baja. Para para H_3 se muestra una reducción mayor de DR cuando se van acumulando las variables auxiliares hasta que aparece la variable número 5 (Nivel educativo del padre) en el que se observa una desaceleración de la disminución de DR, independientemente de la tasa de respuesta. Además las Tablas A.15 a A.18 sugieren la inclusión de la variable (5), nivel educativo del padre, según el estudio más completo de todas las simulaciones realizadas que se encuentra en el Anexo A.

Estimaciones

Atendiendo al criterio H_3 con un tamaño de muestra mayor de la Tabla 3.1, junto con las variaciones de la DR de la Tabla 3.3, se han elegido como variables auxiliares: sexo (4), nivel educativo padre (5), si el trabajo del padre está relacionado con la ciencia (8), si el trabajo de la madre está relacionado con la ciencia (9) y si el trabajo de algún miembro del hogar está relacionado con la ciencia (10). Con éstas variables se han conseguido las estimaciones por calibración de la Tabla 3.4 con la muestra completa del informe PISA. Para los totales poblacionales, debido a la complejidad de encontrar algunos de ellos, se han estimado con el estimador de Horvitz-Thompson con la muestra completa.

Esta tabla también incluye los resultados propuestos por la OCDE en el Informe Español PISA (2007)[30], que indican que el 28 % de los alumnos de 15 años en España tienen la intención de estudiar una carrera científica, que el 13,1 % recibe clases particulares de ciencias y el rendimiento de los alumnos en matemáticas en España y en diez comunidades autónomas.

Comentarios sobre las estimaciones

En resumen, este capítulo muestra la posibilidad de tratar situaciones en las que podemos considerar muchos vectores auxiliares alternativos (vectores x) dentro de una encuesta para su uso en estimaciones a partir del estimador de calibración. La idea principal ha sido utilizar los indicadores descritos por Särndal y Lundström (2010)[55] para realizar una elección apropiada del vector $\mathbf{x}$ con el fin de reducir el sesgo lo máximo posible.

El indicador H_1 es el idóneo si el estudio es para una variable de interés y en particular, mientras que el indicador de H_3 es el que mejor se adapta a la elección de un conjunto de variables auxiliares para cualquier variable de interés, ya que sólo depende de las varaibles auxiliares $\mathbf{x}_k$ pero no de y.

Sobre los resultados obtenidos con las estimaciones destacar que no disponemos de resultados regionales para la variables "estudiar ciencias en el futuro" y "recibir clases particulares". Las estimaciones por calibración aumentan en casi todos los pasos con lo que no influyen en el orden que ocupan las regiones si se ordenan por puntuación en matemáticas.

Con este ejemplo se pone de manifiesto la importancia de considerar la información auxiliar y como hacerlo mediante calibración. En una situación más ideal se podría disponer de una información auxiliar mejor, más relacionada con las variables de interés, bien sea al mismo nivel que para la variable de interés (aspectos relacionados con estudiantes, con su familia, etc.) o a nivel centro (dotaciones presupuestarías, ratio profesor alumno, etc.) que desde un punto de vista más experto se crea que puede influir o estar relacionada con el rendimiento.

En este sentido tendría interés probar con las variables que el último informe PISA[43] sugiere que más influyen en el rendimiento, tales como el nivel socioeconómico del alumno o del profesor, el ratio alumno por aula o la autonomía del centro.

Tabla 3.1: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. Promedio de estudiantes (n) en (m) centros. Variables principales: Puntuación matemáticas (falta de respuesta simulada: 11,5 % y 23,92 %).

		PUNTUACIÓN MATEMÁTICAS 11,5 %							PUNTUACIÓN MATEMÁTICAS 23,92 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
							H_0								
254,6	10	6	8	10	9	5	7	4	9	8	10	4	5	6	7
508,8	20	4	8	10	9	5	6	7	4	9	8	10	5	6	7
761,1	30	4	8	10	9	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	8	10	9	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	8	10	9	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	8	10	9	5	6	7	4	9	8	10	5	6	7
							H_1								
254,6	10	4	9	8	10	5	6	7	5	9	10	8	6	7	4
508,8	20	4	8	10	9	5	6	7	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	5	9	10	8	6	7	4
1019,1	40	4	9	8	10	5	6	7	5	9	10	8	6	7	4
1273,7	50	4	9	8	10	5	6	7	5	9	10	8	6	7	4
2545,58	100	4	9	8	10	5	6	7	5	9	10	8	6	7	4
							H_2								
254,6	10	4	5	7	6	9	10	8	4	5	7	6	9	8	10
508,8	20	4	5	7	6	9	10	8	4	5	7	6	9	10	8
761,1	30	4	5	7	6	9	10	8	4	5	7	6	8	10	9
1019,1	40	4	5	9	10	8	7	6	4	5	7	6	9	10	8
1273,7	50	4	5	9	10	8	7	6	5	4	7	6	8	10	9
2545,5	100	4	5	9	10	8	7	6	5	4	7	6	9	10	8
							H_3								
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	9	8	10	5	6	7

Tabla 3.2: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. Promedio de estudiantes (n) en (m) centros. Variables principales: Carrera ciencias (falta de respuesta real: 11,5 %) y Clases particulares (falta de respuesta real: 23.92 %).

		Carrera ciencias 11,5 %							Clases particulares 23,92 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
							H_0								
254,6	10	5	10	8	9	4	7	6	4	9	8	10	5	6	7
508,8	20	5	10	8	9	4	7	6	4	8	9	10	5	6	7
761,1	30	10	9	8	4	5	7	6	5	9	10	8	6	7	4
1019,1	40	5	10	8	9	4	7	6	5	9	10	8	6	7	4
1273,7	50	10	9	8	4	5	7	6	4	8	9	10	5	6	7
2545,5	100	10	9	8	4	5	7	6	4	8	9	10	5	6	7
							H_1								
254,6	10	5	10	8	9	4	6	7	8	9	10	5	6	7	4
508,8	20	5	10	8	9	4	7	6	4	8	9	10	5	6	7
761,1	30	5	10	8	9	4	7	6	5	9	10	8	6	7	4
1019,1	40	5	10	8	9	4	7	6	8	9	10	5	6	7	4
1273,7	50	5	10	8	9	4	7	6	4	8	9	10	5	6	7
2545,5	100	10	9	8	4	5	7	6	4	8	9	10	5	6	7
							H_2								
254,6	10	4	5	6	7	9	10	8	4	6	7	8	5	10	9
508,8	20	6	5	9	10	8	7	4	4	9	6	7	5	8	10
761,1	30	6	5	7	9	10	8	4	4	6	8	9	10	7	5
1019,1	40	6	5	9	10	8	7	4	4	6	7	8	9	10	5
1273,7	50	6	5	9	10	8	7	4	8	4	6	9	10	7	5
2545,5	100	6	5	9	10	8	7	4	4	6	7	8	9	10	5
							H_3								
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	9	8	10	5	6	7

Tabla 3.3: Estimador de calibración $\tilde{Y}_{cal}$, Desviación relativa (DR) e Indicadores H_0, H_1, H_2 y H_3. Falta de respuesta 11,5 % y 23,92 %. Variable de interés: Puntuación matemáticas. m=100

PUNTUACIÓN MATEMÁTICAS 11,5 %				PUNTUACIÓN MATEMÁTICAS 23,92 %			
v.a.	$H_i \times 10^3$	$\tilde{Y}_{cal} \times 10^{-3}$	DR	v.a.	$H_i \times 10^3$	$\tilde{Y}_{cal} \times 10^{-3}$	DR
			H_0				
4	177.866	8431.980	0.175	4	176.202	8438.939	0.174
8	192.734	8441.939	0.174	9	191.197	8444.336	0.174
10	242.765	8444.344	0.174	8	231.513	8453.042	0.173
9	585.536	8444.340	0.174	10	577.828	8456.761	0.172
5	585.534	8969.759	0.122	5	580.297	8978.400	0.121
6	587.102	9046.449	0.115	6	586.077	9039.126	0.115
7	593.600	9069.240	0.112	7	589.660	9061.712	0.113
			H_1				
4	177.866	8431.980	0.175	5	176.202	8739.512	0.145
9	192.734	8436.172	0.174	9	247.919	8746.225	0.144
8	242.765	8443.691	0.174	10	278.277	8757.569	0.143
10	585.536	8444.340	0.174	8	374.700	8762.723	0.142
5	585.960	8969.759	0.122	6	378.122	8907.962	0.128
6	590.865	9046.449	0.115	7	385.653	8953.688	0.124
7	593.600	9069.240	0.112	4	390.110	9061.712	0.113
			H_2				
4	6.067	8431.980	0.175	5	8.158	8739.512	0.145
5	6.853	8959.013	0.123	4	8.236	8970.380	0.122
9	8.426	8962.070	0.123	7	8.506	9051.186	0.114
10	8.319	8969.148	0.122	6	9.762	9052.728	0.114
8	9.185	8969.759	0.122	9	9.926	9058.470	0.113
7	10.351	9063.841	0.113	10	12.013	9060.850	0.113
6	8.019	9069.240	0.112	8	13.511	9061.712	0.113
			H_3				
4	179.008	8431.980	0.175	4	181.175	8438.939	0.174
9	186.783	8436.172	0.174	9	189.546	8444.336	0.174
8	207.701	8443.691	0.174	8	211.827	8453.042	0.173
10	330.707	8444.340	0.174	10	332.558	8456.761	0.172
5	330.852	8969.759	0.122	5	332.252	8978.400	0.121
6	332.542	9046.449	0.115	6	333.327	9039.126	0.115
7	333.473	9069.240	0.112	7	334.353	9061.712	0.113

Tabla 3.4: Población PISA-ESPAÑA. Estimaciones por calibración a partir de las variables auxiliares proporcionadas por H_3. Variables auxiliares: sexo (4), nivel educativo padre (5), el padre tiene un trabajo relacionado con la ciencia (8), la madre tiene un trabajo relacionado con la ciencia (9), algún miembro del hogar tiene un trabajo relacionado con la ciencia (10).

			PUNTUACIÓN MATEMÁTICAS		CARRERA CIENCIAS	CLASES PARTICULARES
m	n		PISA	$\tilde{Y}_{cal}$	$\tilde{Y}_{cal}$	$\tilde{Y}_{cal}$
686	17,463	España	480	497	28,50 %	17,40 %
52	1,325	Andalucía	463	474	28,30 %	16,47 %
54	1,346	Cataluña	488	493	28,68 %	21,19 %
55	1,420	Galicia	494	497	28,58 %	16,37 %
53	1,392	Asturias	497	500	27,64 %	20,62 %
140	3,473	País Vasco	501	503	28,71 %	17,24 %
56	1,335	Cantabria	502	510	28,23 %	15,17 %
56	1,367	Aragón	513	521	30,91 %	17,65 %
58	1,405	Navarra	515	519	28,44 %	20,82 %
53	1,339	Castilla y León	515	517	26,28 %	16,24 %
45	1,198	La Rioja	526	531	30,38 %	13,32 %

(*) Los datos del informe PISA dan para la variable Carrera ciencias una estimación del 28 % y del 13,1 % para la variable Clases particulares.

Bibliografía

[1] ANDERSSON, C. (1997). *Continuous labour force surveys: performance analysis of a single weight procedure*, Internal report, Statistical Methodology Unit, Statistics Sweden. 45

[2] ALFONS, A. HOLZER, J. y TEMPL. M. (2013). *laeken: Estimation of Indicators on Social Exclusion and Poverty.* R package version 2.0.1. URL `CRAN.R-project.org/package=laeken`. 19

[3] ARCOS, A. y CONTRERAS, J.M. (2014). Selección de variables para el ajuste de no respuesta. XIX SIMPOSIO INTERNACIONAL DE MÉTODOS MATEMÁTICOS APLICADOS A LAS CIENCIAS. San José (Costa Rica).

[4] ARCOS, A., CONTRERAS, J.M. y RUEDA, M.M (2013). A Novel Calibration Estimator In Social Surveys. SOCIOLOGICAL METHODS AND RESEARCH. published online. DOI: 10.1177/0049124113507906.

[5] BARÓMETRO PREELECTORAL DE CENTRO DE INVESTIGACIÓN SOCIOLÓGICAS (CIS) DE OCTUBRE (2011). `www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=11904`. 25, 126

[6] CORPAS, J.B. (1983). Regression, Prediction and Shrinkage. JOURNAL OF THE ROYAL STATISTICAL SOCIETY. SERIES B (METHODOLOGICAL), **45(3)**, 311-354. 50

[7] CHAMBERS, R.L. (1996). Robust case weighting for multipurpose establishment surveys. JOURNAL OF OFFICIAL STATISTICS, **12**, 332. 18

[8] CHAMBERS, R.L. y DUNSTAN, R. (1986). Estimating distribution function from survey data. BIOMETRIKA, **73**, 597-604. 54

[9] CHOWDHURY, S. (1997). Integrated weighting for estimation in household surveys. STATIST, **3**, 335-356. 49

[10] DEVILLE, J.C. y SÄRNDAL, C.E. (1992). Calibration estimators in survey sampling. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **87**, 376–382. vii, 1, 14, 15, 57

[11] DEVILLE, J.C., SÄRNDAL, C.E. y SANTOURY, O. (1993). Generalised ranking procedure in survey sampling. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **88**, 1013-1020. 17

[12] DUCHESNE, P. (1999). Robust calibration estimators. SURVEY METHODOLOGY, **25**, 43–56. 18

[13] ELTINGE, J. y YANSANEH, I. (1997). Diagnostics for the formation of nonresponse adjustment cells with an application to income nonresponse in the US Consumer Expenditure Survey. SURVEY METHODOLOGY, **23**, 33-40. 74

[14] ESTEVÁO, V.M. y SÄRNDAL, C.E. (2000). A functional form approach to calibration. JOURNAL OF OFFICIAL STATISTICS, **16**, 379-399.

[15] ESTEVÁO, V.M. y SÄRNDAL, C.E. (2002). The ten cases of auxiliary information for calibration in twophase sampling. JOURNAL OF OFFICIAL STATISTICS, **18**, 233-255.

[16] ESTEVÁO, V.M. y SÄRNDAL, C.E. (2006). Survey estimates by calibration on complex auxiliary information. INTERNATIONAL STATISTICAL REVIEW, **74**, 127-147. ix, 35, 38, 44, 54

[17] GERSHUNSKAYA, J., JIANG, J. y LAHIRI, P. (2009). Resampling methods in surveys in Handbook of Statistics. SAMPLE SURVEYS: INFERENCE AND ANALYSIS, Vol. 29B. D. Pfeermann y C.R. Rao (editors), The Netherlands: North-Holland. 53

[18] GIJBELS, I. y VERAVERBEKE, N. (1988). Weak asymptotic representations for quantiles of the product-limit estimator. JOURNAL OF STATISTICAL PLANNING AND INFERENCE, **18**, 151-160. 54

[19] GUTIÉRREZ, H.A. (2013). TEACHINGSAMPLING. R package version 2.0.1, URL: `cran.r-project.org/web/packages/TeachingSampling/TeachingSampling.pdf`. 19

[20] HARMS, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. CHINTEX WORKING PAPER, **13**, Federal Statistical Office, Germany. 30

[21] HARMS, T. y DUCHESNE, P. (2006). On calibration estimation for quantiles. SURVEY METHODOLOGY, **32**, 37-52. 30, 31, 32

[22] HORVITZ,D. G. y THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **47**, 663–685. viii, 4

[23] HUANG, E.T. y FULLER, W.A. (1978). Nonnegative regression estimation for survey data. PROCEEDINGS OF THE SOCIAL STATISTICS SECTION, AMERICAN STATISTICAL ASSOCIATION, Alexandria, 300–303. 19

[24] INE `www.ine.es/en/prensa/epf-prensa-en.htm` 26, 56, 124, 127

[25] INFORME METODOLÓGICO DE LA ENCUESTA PRE-ELECTORAL Y POST-ELECTORAL ELECCIONES GENERALES (2011). Centro de Investigación Sociológicas (CIS). `www.cis.es/cis/export/sites/default/-Archivos/Marginales/2900_2919/2915/Informemetodologico7711.pdf` 25

[26] KALTON, G. y FLORES-CERVANTES, I. (2003). Weighting methods. JOURNAL OF OFFICIAL STATISTICS, **19**, 81-98. 74

[27] KOTT, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. SURVEY METHODOLOGY, **32**, 133-142.

[28] KOVAĈEVIĆ, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. PROCEEDINGS OF THE SURVEY METHODS SECTION, STATISTICAL SOCIETY OF CANADA, 139-144. 30

[29] KRAPAVICKAITĖ, D. y PLIKUSAS, A. (2005). Estimation of a ratio in the finite population. INFORMATICA, **16**, 347-364. 32, 33

[30] INFORME ESPAÑOL PISA (2007) Ministerio de Educación y Ciencia Secretaría General de Educación `www.mec.es/multimedia/00005713.pdf` 87

[31] LE GUENNEC, J. y SAUTORY, O. (2002). Calmar 2: une nouvelle version de la macro calmar de redressement d'échantillon par calage. Insee-Méthodes: ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE 2002. INSEE. 19

[32] LEMAÎTRE, G. y DUFOUR, J. (1987). An integrated method for weighting persons and families. SURVEY METHODOLOGY, **13**, 199-207. ix, 35, 44, 45, 47, 48, 49

[33] LUMLEY, T. (2013). SURVEY. R package version 3.29-5, URL: `cran.r-project.org/web/packages/survey/index.html`. 19

[34] MINISTERIO DE EMPLEO Y SEGURIDAD SOCIAL (2013). URL: `www.empleo.gob.es`. 26, 127

[35] MONTANARI, G.E. y RANALLI, M.G. (2003). On calibration methods for design based finite population inferences. BULLETIN OF THE INTERNATIONAL STATISTICAL INSTITUTE, 54 th session, volume LX, contributed papers, book 2, 81-82. 29

[36] MONTANARI, G.E. y RANALLI, M.G. (2005). Nonparametric model calibration estimation in survey sampling. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **100**, 1429-1442. 29, 30

[37] NEETHLING, A. y GALPIN, J. (2006). Weighting of Household Survey Data: A Comparison of Various Calibration, Intergrated and Cosmetic Estimators. SOUTH AFRICAN STATISTICS JOUNAL, **40(2)** , 123-150. 48

[38] NIEUWENBROEK, N.J. (1993). AN INTEGRATED METHOD FOR WEIGHTING CHARACTERISTICS OF PERSONS AND HOUSEHOLDS USING THE LINEAR REGRESSION ESTIMATOR. Internal report, Central Bureau of Statistics, The Netherlands. 45, 49

[39] NIEUWENBROEK, N.J. y BOONSTRA, H.J. (2002). BASCULA 4.0 FOR WEIGHTING SAMPLE SURVEY DATA WITH ESTIMATION OF VARIANCES. The Survey Statistician, Software Reviews. 19

[40] OCDE (2013). Organización para la Cooperación y el Desarrollo Económico. URL `www.oecd.org/`. 55, 122

[41] PARK, M. y FULLER, W.A. (2005). Towards nonnegative regression weights for survey samples. SURVEY METHODOLOGY, **31**, 85-93. 19

[42] PISA `pisa2006.acer.edu.au` 55

[43] PISA 2012. Programa para la Evaluación Internacional de los Alumnos. Informe español. Volumen I: Resultados y contexto. `www.mecd.gob.es/dctm/inee/internacional/pisa2012` 88

[44] PLIKUSAS, A. (2006). Nonlinear calibration. PROCEEDINGS WORKSHOP ON SURVEY SAMPLING, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia. 32, 33

[45] PREELECTORAL ELECCIONES GENERALES (2011). Centro de Investigación Sociológicas (CIS). URL `datos.cis.es/pdf/Es2915mar_A.pdf`. 25, 126

[46] RAO, J.N.K. (2003). SMALL AREA ESTIMATION. John Wiley and Sons. 50

[47] RAO, J.N.K., KOVAR, J.G. y MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. BIOMETRIKA, **77**, 365-375. 54

[48] RANDLES, R.H. (1980). On the asymptotic normality of statistics with estimated parameters. ANN. STATIST, **10**, 462-474. 53, 54

[49] REN, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE, INSEE Méthodes, tome 1, 263-289. 30

[50] RIZZO, L., KALTON, G. y BRICK, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. SURVEY METHODOLOGY, **22**, 43-53. 74

[51] RUEDA, M., MARTÍNEZ, S., MARTÍNEZ, H. y ARCOS, A. (2007). Estimation of the distribution function with calibration methods. JOURNAL OF STATISTICAL PLANNING AND INFERENCE, **137**, 435-448. 30, 31, 32, 54

[52] RUEDA, C. y MENÉNDEZ, J.A. (2010). The Selection of the Shrinkage Region in Small Area Estimation. ADVANCES IN SOFT COMPUTING, **77**, 553-560. 50

[53] SÄRNDAL, C. (2007). The calibration approach in survey theory and practice. SURVEY METHODOLOGY, **33(2)**, 99–119. 30

[54] SÄRNDAL, C.E. y LUNDSTRÖM, S. (2005). ESTIMATION IN SURVEYS WITH NONRESPONSE. New York: John Wiley and Sons, Inc. 76

[55] SÄRNDAL, C.E. y LUNDSTRÖM, S. (2010). Design for estimation: identifying auxiliary vectors to reduce nonresponse bias. SURVEY METHODOLOGY, **36(2)**, 131–144. 74, 79, 88

[56] SÄRNDAL, C.E., SWENSSON, B. y WRETMAN, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag. 7

[57] SCHÄFER, J. y STRIMMER K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY, **4(1)**, 32. 50

[58] SINGH, S. (2001). Generalized Calibration Approach for Estimating Variance in Survey Sampling. ANNALS OF THE INSTITUTE OF STATISTICAL MATHEMATICS, **53**, 404-417. 53

[59] SINGH, S. (2010). On the calibration of design weights using a displacement function. METRIKA, **75(1)**, 85-107. 53

[60] SINGH, S. HORN, S., CHOWDHURY, S. y YU, F. (1999). Calibration of the estimators of variance. AUSTRALIAN AND NEW ZEALAND JOURNAL OF STATISTICS, **41**, 199-212. 53

[61] SINGH, A.C. y MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. SURVEY METHODOLOGY, **22**, 107-115. 17

[62] STUKEL, D.M., HIDIROGLOU, M.A. y SÄRNDAL, C.E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. SURVEY METHODOLOGY, **22**, 117-125. 17

[63] THÉBERGE, A. (1999). Extensions of calibration estimation in survey sampling. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 94, 635-644. 28

[64] THÉBERGE, A. (2000). Calibration and restricted weights. SURVEY METHODOLOGY, **26**, 99-107. 18

[65] THOMPSON, J.R. (1968). Some shrinkage technique for estimating the mean. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **63**, 113-123. 51

[66] THOMSEN, I., KLEVEN, O., WANG, J.H. y ZHANG, L.C. (2006). COPING WITH DECEASING RESPONSE RATES IN STATISTICS NORWAY. RECOMMENDED PRACTICE FOR REDUCING THE EFFECT OF NONRESPONSE. Oslo: Statistics Norway. 74

[67] TILLÉ, Y. (2002). Unbiased estimation by calibration on distribution in simple sampling designs without replacement. SURVEY METHODOLOGY, **28**, 77-85. 30

[68] TILLÉ, Y. (2005). TEORÍA DE MUESTREO. Gruope de Statistique, Université de Neuchâtel, Suisse. 14, 15, 16, 17

[69] TILLÉ, Y. y MATEI, A. (2013). SAMPLING: SURVEY SAMPLING. R package version 2.3, URL: `CRAN.R-project.org/package=sampling`. 19, 133

[70] TRACY, D.S., SINGH, S. y ARNAB, R. (2003). Note on calibration in stratified and double sampling. SURVEY METHODOLOGY, **29**, 99–104. 28

[71] VANDERHOEFT, C. (2001). GENERALISED CALIBRATION AT STATISTICS BELGIUM. SPSS MODULE GCALIBS AND CURRENT PRACTICES. Statistics BelgiumWorking Paper n. 3. 19

[72] VANDERHOEFT, C., WAEYTENS, E. y MUSEUX, J.M. (2001). GENERALISED CALIBRATION WITH SPSS 9.0 FOR WINDOWS. In Enquêtes, Modèles et Applications (Eds. J.J. Droesbeke and L. Lebart), Paris: Dunod. 19

[73] WAND, Y. y OPSOMER, J. (2011). On the asymptotic normality and variance estimation of non-differentiable survey estimators. BIOMETRIKA, **98(1)**, 91-106. 54

[74] WOLTER, K. M. (2007). INTRODUCTION TO VARIANCE ESTIMATION, 2nd Edition. Springer. 53

[75] WU, C. (2003). Optimal calibration estimators in survey sampling. BIOMETRIKA, **90**, 937-951. 29

[76] WU, C. y SITTER, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, **96**, 185-193. 29, 30

[77] ZARDETTO, D. (2013). EVER. R package version 1.2, URL: `cran.r-project.org/web/packages/EVER/index.html`. 19

CAPÍTULO

A

Anexo. Tablas Capítulo 3

A. ANEXO. TABLAS CAPÍTULO 3

Tabla A.1: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Puntuación matemáticas (falta de respuesta simulada: 20 % y 30 %).

		PUNTUACIÓN MATEMÁTICAS 20 %							PUNTUACIÓN MATEMÁTICAS 30 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
					H_0							H_0			
254,6	10	4	9	8	10	5	6	7	4	9	10	8	5	6	7
508,8	20	6	8	10	9	5	7	4	4	9	8	10	5	6	7
761,1	30	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	8	9	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	8	9	10	5	6	7	4	9	8	10	5	6	7
					H_1							H_1			
254,6	10	4	9	8	10	5	6	7	4	9	10	8	5	6	7
508,8	20	6	8	10	9	5	7	4	4	9	8	10	5	6	7
761,1	30	6	8	10	9	5	7	4	4	9	8	10	5	6	7
1019,1	40	6	8	10	9	5	7	4	4	9	8	10	5	6	7
1273,7	50	6	8	10	9	5	7	4	4	9	8	10	5	6	7
2545,5	100	4	8	9	10	5	6	7	4	8	9	10	5	6	7
					H_2							H_2			
254,6	10	4	5	7	6	9	10	8	4	9	8	10	5	6	7
508,8	20	4	5	9	10	8	7	6	4	5	7	6	9	10	8
761,1	30	4	5	9	10	8	7	6	4	5	7	6	9	10	8
1019,1	40	4	5	9	10	8	7	6	4	5	7	6	9	10	8
1273,7	50	4	5	9	10	8	7	6	4	9	8	10	5	7	6
2545,5	100	5	9	4	10	8	7	6	4	8	9	10	5	7	6
					H_3							H_3			
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	6	8	10	9	5	7	4	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	6	8	10	9	5	7	4	4	9	8	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	9	8	10	5	6	7

Tabla A.2: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Puntuación matemáticas (falta de respuesta simulada: 40 % y 50 %).

		Puntuación matemáticas 40 %							Puntuación matemáticas 50 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
		H_0							H_0						
254.600	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508.800	20	4	8	9	10	5	6	7	4	9	8	10	5	6	7
761.100	30	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1019.100	40	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1273.700	50	4	8	9	10	5	6	7	4	9	8	10	5	6	7
2545.500	100	4	8	9	10	5	6	7	4	9	8	10	5	6	7
		H_1							H_1						
254.600	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508.800	20	4	9	10	8	5	6	7	6	8	10	9	5	7	4
761.100	30	4	8	9	10	5	6	7	5	9	10	8	6	7	4
1019.100	40	4	8	9	10	5	6	7	5	9	10	8	6	7	4
1273.700	50	4	8	9	10	5	6	7	5	9	8	10	4	6	7
2545.500	100	4	8	9	10	5	6	7	5	9	8	10	4	6	7
		H_2							H_2						
254.600	10	4	5	7	6	9	10	8	4	5	9	7	6	8	10
508.800	20	4	5	7	6	9	10	8	4	5	7	6	8	10	9
761.100	30	4	5	7	6	9	10	8	4	5	7	6	8	10	9
1019.100	40	4	5	7	6	9	10	8	4	5	7	6	8	10	9
1273.700	50	4	5	7	6	9	10	8	4	5	7	6	8	10	9
2545.500	100	4	5	7	6	9	10	8	4	5	9	10	8	7	6
		H_3							H_3						
254.600	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508.800	20	4	8	9	10	5	6	7	4	9	8	10	5	6	7
761.100	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019.100	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273.700	50	4	9	8	10	5	6	7	4	9	8	10	5	6	7
2545.500	100	4	8	9	10	5	6	7	4	9	8	10	5	6	7

Tabla A.3: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 20 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	206.206	8419.774	0.176	4	206.206	8419.774	0.176
9	210.816	8443.220	0.174	9	210.816	8443.220	0.174
8	276.026	8440.040	0.174	8	276.026	8440.040	0.174
10	542.463	8440.440	0.174	10	542.463	8440.440	0.174
5	542.706	8879.061	0.132	5	542.706	8879.061	0.132
6	540.774	8986.413	0.121	6	540.774	8986.413	0.121
7	555.016	8994.001	0.120	7	555.016	8994.001	0.120
	m=20				m=20		
6	216.410	8692.607	0.109	6	216.410	8692.607	0.109
8	280.252	8695.158	0.108	8	280.252	8695.158	0.108
10	298.625	8697.738	0.108	10	298.625	8697.738	0.108
9	347.469	8703.882	0.108	9	347.469	8703.882	0.108
5	351.386	8780.497	0.100	5	351.386	8780.497	0.100
7	353.030	8809.316	0.097	7	353.030	8809.316	0.097
4	354.657	8909.457	0.086	4	354.657	8909.457	0.086
	m=30				m=30		
4	238.699	8468.948	0.156	6	238.699	8631.885	0.140
8	249.546	8473.399	0.155	8	344.484	8635.986	0.139
9	302.726	8485.749	0.154	10	378.196	8639.756	0.139
10	563.439	8485.374	0.154	9	460.367	8641.825	0.139
5	563.192	8881.107	0.115	5	461.731	8766.553	0.126
6	571.328	8961.828	0.107	7	464.214	8817.724	0.121
7	574.261	8978.294	0.105	4	466.916	8978.294	0.105
	m=40				m=40		
4	203.594	8384.385	0.152	6	203.594	8615.645	0.128
8	212.878	8398.147	0.150	8	290.365	8616.007	0.128
9	257.524	8404.889	0.150	10	317.926	8618.481	0.128
10	543.891	8403.986	0.150	9	404.725	8617.819	0.128
5	543.303	8843.997	0.105	5	404.294	8751.188	0.114
6	547.691	8912.596	0.098	7	405.904	8793.537	0.110
7	556.648	8926.863	0.097	4	406.140	8926.863	0.097
	m=50				m=50		
4	203.130	8432.017	0.152	6	203.130	8689.620	0.126
8	210.772	8446.876	0.151	8	282.431	8691.824	0.126
9	254.798	8449.440	0.151	10	306.454	8695.907	0.126
10	557.067	8451.319	0.150	9	394.874	8698.570	0.125
5	558.299	8912.106	0.104	5	396.621	8833.360	0.112
6	559.982	8979.220	0.097	7	399.299	8869.982	0.108
7	569.729	8990.870	0.096	4	400.745	8990.870	0.096
	m=100				m=100		
4	180.596	8428.725	0.175	4	180.596	8428.725	0.175
8	195.171	8443.550	0.174	8	195.171	8443.550	0.174
9	233.914	8451.824	0.173	9	233.914	8451.824	0.173
10	567.114	8454.065	0.173	10	567.114	8454.065	0.173
5	568.584	8962.212	0.123	5	568.584	8962.212	0.123
6	574.009	9021.297	0.117	6	574.009	9021.297	0.117
7	583.730	9043.525	0.115	7	583.730	9043.525	0.115

Tabla A.4: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 20 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-18.146	8419.774	0.176	4	193.668	8419.774	0.176
5	-19.533	8872.585	0.132	9	196.775	8443.220	0.174
7	-17.143	8972.485	0.122	8	226.903	8440.040	0.174
6	-15.243	8993.350	0.120	10	332.729	8440.440	0.174
9	-12.284	8990.308	0.121	5	331.412	8879.061	0.132
10	-9.314	8993.013	0.120	6	330.214	8986.413	0.121
8	-3.754	8994.001	0.120	7	330.400	8994.001	0.120
	m=20				m=20		
4	-10.529	8339.786	0.145	6	187.372	8692.607	0.109
5	-7.721	8766.461	0.101	8	212.591	8695.158	0.108
9	-4.563	8766.451	0.101	10	221.526	8697.738	0.108
10	-5.066	8769.710	0.101	9	245.196	8703.882	0.108
8	-2.722	8773.222	0.100	5	246.239	8780.497	0.100
7	-2.785	8891.219	0.088	7	247.411	8809.316	0.097
6	2.405	8909.457	0.086	4	244.607	8909.457	0.086
	m=30				m=30		
4	2.449	8468.948	0.156	4	204.770	8468.948	0.156
5	3.233	8876.855	0.115	9	210.757	8485.090	0.154
9	5.432	8876.717	0.115	8	231.948	8485.749	0.154
10	4.955	8880.016	0.115	10	327.273	8485.374	0.154
8	8.172	8881.107	0.115	5	327.292	8881.107	0.115
7	7.837	8970.756	0.106	6	327.492	8961.828	0.107
6	14.572	8978.294	0.105	7	328.641	8978.294	0.105
	m=40				m=40		
4	3.626	8384.385	0.152	4	194.968	8384.385	0.152
5	4.533	8844.299	0.105	9	200.653	8395.638	0.150
9	7.952	8843.487	0.105	8	222.764	8404.889	0.150
10	7.661	8843.257	0.105	10	328.841	8403.986	0.150
8	11.112	8843.997	0.105	5	328.769	8843.997	0.105
7	10.264	8921.526	0.097	6	330.358	8912.596	0.098
6	17.749	8926.863	0.097	7	329.246	8926.863	0.097
	m=50				m=50		
4	3.664	8432.017	0.152	6	191.795	8689.620	0.126
5	4.900	8908.603	0.104	8	222.569	8691.824	0.126
9	5.948	8908.320	0.104	10	233.935	8695.907	0.126
10	5.142	8909.122	0.104	9	265.163	8698.570	0.125
8	8.112	8912.106	0.104	5	265.810	8833.360	0.112
7	7.952	8984.117	0.097	7	267.868	8869.982	0.108
6	14.866	8990.870	0.096	4	268.785	8990.870	0.096
	m=100				m=100		
5	5.603	8729.213	0.146	4	182.941	8428.725	0.175
9	5.963	8729.077	0.146	9	190.951	8442.854	0.174
4	7.801	8956.883	0.123	8	212.257	8451.824	0.173
10	7.346	8960.379	0.123	10	332.859	8454.065	0.173
8	8.004	8962.212	0.123	5	332.775	8962.212	0.123
7	6.768	9041.085	0.115	6	334.523	9021.297	0.117
6	6.715	9043.525	0.115	7	334.285	9043.525	0.115

Tabla A.5: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 30 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	189.543	8458.822	0.173	4	189.543	8458.822	0.173
9	188.720	8458.761	0.173	9	188.720	8458.761	0.173
10	256.235	8491.394	0.169	10	256.235	8491.394	0.169
8	513.935	8501.039	0.168	8	513.935	8501.039	0.168
5	519.600	8939.798	0.126	5	519.600	8939.798	0.126
6	538.767	9054.749	0.114	6	538.767	9054.749	0.114
7	538.731	9053.347	0.114	7	538.731	9053.347	0.114
	m=20				m=20		
4	183.338	8493.363	0.129	4	183.338	8493.363	0.129
9	183.453	8493.308	0.129	9	183.453	8493.308	0.129
8	244.010	8526.236	0.126	8	244.010	8526.236	0.126
10	526.764	8526.088	0.126	10	526.764	8526.088	0.126
5	526.671	8977.916	0.079	5	526.671	8977.916	0.079
6	547.277	9074.682	0.069	6	547.277	9074.682	0.069
7	547.243	9074.865	0.069	7	547.243	9074.865	0.069
	m=30				m=30		
4	202.674	8633.028	0.139	4	202.674	8633.028	0.139
9	216.482	8631.853	0.140	9	216.482	8631.853	0.140
8	250.025	8647.719	0.138	8	250.025	8647.719	0.138
10	487.313	8647.560	0.138	10	487.313	8647.560	0.138
5	487.211	9017.768	0.101	5	487.211	9017.768	0.101
6	497.381	9070.101	0.096	6	497.381	9070.101	0.096
7	496.627	9091.643	0.094	7	496.627	9091.643	0.094
	m=40				m=40		
4	188.309	8511.181	0.139	4	188.309	8511.181	0.139
9	201.022	8513.411	0.139	9	201.022	8513.411	0.139
8	240.336	8529.594	0.137	8	240.336	8529.594	0.137
10	496.653	8531.371	0.137	10	496.653	8531.371	0.137
5	497.790	8931.906	0.096	5	497.790	8931.906	0.096
6	508.146	8993.340	0.090	6	508.146	8993.340	0.090
7	509.573	9013.206	0.088	7	509.573	9013.206	0.088
	m=50				m=50		
4	183.126	8544.351	0.141	4	183.126	8544.351	0.141
9	194.738	8547.052	0.141	9	194.738	8547.052	0.141
8	235.534	8571.341	0.138	8	235.534	8571.341	0.138
10	504.693	8572.886	0.138	10	504.693	8572.886	0.138
5	505.685	8992.132	0.096	5	505.685	8992.132	0.096
6	521.279	9055.677	0.090	6	521.279	9055.677	0.090
7	523.013	9073.764	0.088	7	523.013	9073.764	0.088
	m=100				m=100		
4	169.933	8435.863	0.174	4	169.933	8435.863	0.174
9	188.465	8439.758	0.174	8	188.465	8446.430	0.173
8	230.229	8447.890	0.173	9	230.229	8447.890	0.173
10	578.198	8448.395	0.173	10	578.198	8448.395	0.173
5	578.524	8987.293	0.120	5	578.524	8987.293	0.120
6	583.775	9051.974	0.114	6	579.467	9051.974	0.114
7	586.290	9080.674	0.111	7	586.290	9080.674	0.111

Tabla A.6: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 30 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-5.477	8458.822	0.173	4	190.608	8458.822	0.173
9	-5.340	8458.761	0.173	9	189.852	8458.761	0.173
8	-4.785	8500.752	0.169	8	221.068	8500.752	0.169
10	16.625	8501.039	0.168	10	332.343	8501.039	0.168
5	19.242	8939.798	0.126	5	331.276	8939.798	0.126
6	11.969	9054.749	0.114	6	329.179	9054.749	0.114
7	11.937	9053.347	0.114	7	329.135	9053.347	0.114
	m=20				m=20		
4	-5.787	8493.363	0.129	4	192.151	8493.363	0.129
5	-6.165	8973.537	0.080	9	192.247	8493.308	0.129
7	-9.147	9047.465	0.072	8	225.858	8526.236	0.126
6	-9.736	9060.142	0.071	10	330.894	8526.088	0.126
9	-6.871	9071.747	0.070	5	330.855	8977.916	0.079
10	-5.908	9074.692	0.069	6	329.743	9074.682	0.069
8	8.471	9074.865	0.069	7	329.716	9074.865	0.069
	m=30				m=30		
4	18.032	8633.028	0.139	4	213.293	8633.028	0.139
5	17.753	9016.773	0.101	9	219.139	8631.853	0.140
7	17.716	9086.240	0.094	8	238.862	8647.719	0.138
6	19.127	9084.450	0.094	10	333.762	8647.560	0.138
9	19.080	9091.646	0.094	5	333.708	9017.768	0.101
10	20.462	9091.652	0.094	6	333.375	9070.101	0.096
8	28.706	9091.643	0.094	7	332.943	9091.643	0.094
	m=40				m=40		
4	20.481	8511.181	0.139	4	200.102	8511.181	0.139
5	20.517	8929.883	0.096	9	205.838	8513.411	0.139
7	18.192	9007.834	0.089	8	228.518	8529.594	0.137
6	19.233	9007.922	0.089	10	333.021	8531.371	0.137
9	19.238	9012.330	0.088	5	332.781	8931.906	0.096
10	20.992	9013.054	0.088	6	331.475	8993.340	0.090
8	27.180	9013.206	0.088	7	331.983	9013.206	0.088
	m=50				m=50		
4	15.856	8544.351	0.141	4	196.708	8544.351	0.141
9	15.924	8547.052	0.141	9	202.267	8547.052	0.141
8	15.775	8571.341	0.138	8	224.506	8571.341	0.138
10	24.235	8572.886	0.138	10	331.616	8572.886	0.138
5	23.647	8992.132	0.096	5	331.630	8992.132	0.096
7	16.965	9070.337	0.088	6	331.016	9055.677	0.090
6	16.539	9073.764	0.088	7	331.525	9073.764	0.088
	m=100				m=100		
4	7.434	8435.863	0.174	4	177.676	8435.863	0.174
8	7.674	8446.430	0.173	9	185.448	8439.758	0.174
9	9.017	8447.890	0.173	8	207.544	8447.890	0.173
10	11.741	8448.395	0.173	10	331.200	8448.395	0.173
5	11.732	8987.293	0.120	5	331.299	8987.293	0.120
7	11.506	9078.813	0.111	6	332.607	9051.974	0.114
6	10.605	9080.674	0.111	7	333.393	9080.674	0.111

Tabla A.7: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 40 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	256.681	8321.860	0.189	4	256.681	8321.860	0.189
9	268.043	8317.444	0.189	9	268.043	8317.444	0.186
8	342.754	8296.331	0.186	8	342.754	8296.331	0.186
10	621.773	8298.712	0.187	10	621.773	8298.712	0.188
5	623.251	8748.227	0.144	5	623.251	8748.227	0.144
6	610.146	8868.591	0.133	6	610.146	8868.591	0.133
7	607.405	8886.896	0.131	7	607.405	8886.896	0.131
	m=20				m=20		
4	224.586	8268.120	0.155	4	224.586	8268.120	0.152
8	217.632	8249.195	0.154	9	217.632	8267.715	0.152
9	328.630	8249.237	0.154	10	328.630	8251.780	0.150
10	652.815	8249.525	0.154	8	652.815	8249.525	0.150
5	652.997	8761.100	0.102	5	651.386	8761.100	0.102
6	653.024	8936.259	0.084	6	641.288	8936.259	0.084
7	641.031	8925.285	0.085	7	641.031	8925.285	0.085
	m=30				m=30		
4	239.555	8444.247	0.161	4	239.216	8444.247	0.158
8	246.658	8430.148	0.160	8	267.027	8444.247	0.158
9	323.345	8429.183	0.160	9	599.754	8430.148	0.156
10	599.754	8427.599	0.160	10	598.742	8429.183	0.156
5	598.742	8860.320	0.117	5	598.126	8427.599	0.156
6	598.126	8980.376	0.105	6	589.120	8948.487	0.108
7	589.120	8991.495	0.104	7	589.120	8992.027	0.104
	m=40				m=40		
4	196.130	8362.451	0.156	4	196.130	8362.451	0.156
8	203.756	8348.218	0.155	8	203.756	8348.218	0.155
9	277.269	8347.792	0.155	9	277.269	8347.792	0.155
10	596.307	8345.143	0.154	10	596.307	8345.143	0.154
5	594.587	8836.742	0.106	5	594.587	8836.742	0.106
6	594.310	8950.016	0.094	6	594.310	8950.016	0.094
7	585.074	8961.766	0.093	7	585.074	8961.766	0.093
	m=50				m=50		
4	192.141	8426.447	0.155	4	192.141	8426.447	0.155
8	200.272	8414.839	0.154	8	200.272	8414.839	0.154
9	267.387	8414.020	0.154	9	267.387	8414.020	0.154
10	574.453	8412.969	0.153	10	574.453	8412.969	0.154
5	573.773	8888.383	0.106	5	573.773	8888.383	0.106
6	573.245	8992.295	0.096	6	573.245	8992.295	0.096
7	565.747	9004.883	0.095	7	565.747	9004.883	0.095
	m=100				m=100		
4	174.973	8395.457	0.178	4	174.973	8395.457	0.178
8	188.115	8396.612	0.178	8	188.115	8396.612	0.178
9	250.615	8398.234	0.178	9	250.615	8398.234	0.178
10	597.425	8398.620	0.178	10	597.425	8398.620	0.178
5	597.676	8930.983	0.126	5	597.676	8930.983	0.126
6	598.733	9026.922	0.117	6	598.733	9026.922	0.117
7	599.485	9047.095	0.115	7	599.485	9047.095	0.115

Tabla A.8: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 40 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-19.346	8321.860	0.186	4	223.719	8321.860	0.189
5	-19.500	8742.558	0.145	9	228.580	8317.444	0.189
7	-8.681	8867.364	0.133	8	253.512	8296.331	0.189
6	-15.745	8876.553	0.132	10	343.104	8298.712	0.188
9	-11.827	8897.366	0.130	5	343.420	8748.227	0.144
10	-18.364	8887.396	0.131	6	334.817	8868.591	0.133
8	-10.791	8886.896	0.131	7	327.132	8886.896	0.131
	m=20				m=20		
4	-11.125	8268.120	0.152	4	224.648	8268.120	0.154
5	-10.007	8752.958	0.102	8	218.123	8249.195	0.154
7	-4.186	8875.936	0.090	9	247.182	8249.237	0.154
6	-8.896	8907.648	0.087	10	332.005	8249.525	0.154
9	-5.466	8930.306	0.084	5	332.039	8761.100	0.102
10	-6.211	8928.071	0.085	6	331.980	8936.259	0.084
8	4.500	8925.285	0.085	7	321.230	8925.285	0.085
	m=30				m=30		
4	2.365	8444.247	0.158	4	222.467	8444.247	0.159
5	2.351	8853.176	0.118	9	227.023	8441.452	0.159
7	11.246	8956.905	0.107	8	248.399	8429.183	0.159
6	10.102	8972.262	0.106	10	334.311	8427.599	0.159
9	11.298	8992.271	0.104	5	333.112	8860.320	0.117
10	12.132	8991.490	0.104	6	327.850	8980.376	0.105
8	18.813	8991.495	0.104	7	325.470	8991.495	0.104
	m=40				m=40		
4	0.644	8362.451	0.154	4	209.597	8362.451	0.156
5	0.186	8828.567	0.107	9	214.203	8360.436	0.156
7	11.406	8932.859	0.096	8	235.245	8347.792	0.155
6	11.576	8944.707	0.095	10	334.182	8345.143	0.154
9	13.326	8962.082	0.093	5	333.541	8836.742	0.106
10	11.528	8960.893	0.093	6	328.603	8950.016	0.094
8	11.793	8961.766	0.093	7	327.215	8961.766	0.093
	m=50				m=50		
4	1.120	8426.447	0.153	4	207.120	8421.447	0.154
5	0.786	8875.858	0.108	9	211.812	8423.286	0.154
7	9.526	8974.429	0.098	8	231.372	8424.020	0.153
6	9.709	8984.821	0.097	10	333.892	8424.969	0.153
9	11.147	9002.038	0.095	5	333.661	8888.383	0.106
10	9.181	9003.663	0.095	6	330.307	8992.295	0.096
8	10.549	9004.883	0.095	7	328.406	9004.883	0.095
	m=100				m=100		
4	2.482	8395.457	0.178	4	185.019	8395.457	0.178
5	2.256	8916.941	0.127	8	192.571	8396.612	0.178
7	3.870	9021.988	0.117	9	211.564	8398.234	0.178
6	6.722	9028.842	0.116	10	332.527	8398.620	0.178
9	7.521	9037.128	0.116	5	332.604	8930.983	0.126
10	6.694	9046.600	0.115	6	332.945	9026.922	0.117
8	0.717	9047.095	0.115	7	333.333	9047.095	0.115

Tabla A.9: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 50 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	247.456	8308.170	0.187	4	247.456	8308.170	0.187
9	265.643	8285.006	0.190	9	265.643	8285.006	0.190
8	310.678	8254.954	0.193	8	310.678	8254.954	0.193
10	529.810	8252.752	0.193	10	529.810	8252.752	0.193
5	528.467	8612.118	0.158	5	528.467	8612.118	0.158
6	510.142	8685.972	0.150	6	510.142	8685.972	0.150
7	496.017	8715.798	0.147	7	496.017	8715.798	0.147
	m=20				m=20		
4	201.929	8308.539	0.148	6	201.929	8588.440	0.119
9	241.917	8308.609	0.148	8	284.081	8583.688	0.120
8	312.125	8302.369	0.149	10	338.085	8592.917	0.119
10	539.405	8305.426	0.148	9	368.300	8584.568	0.120
5	541.278	8676.212	0.110	5	363.182	8633.861	0.115
6	537.454	8790.749	0.099	7	368.839	8721.963	0.106
7	537.496	8855.987	0.092	4	365.927	8855.987	0.092
	m=30				m=30		
4	219.927	8379.024	0.165	5	219.927	8487.519	0.154
9	244.839	8384.668	0.164	9	305.242	8489.873	0.154
8	294.387	8379.277	0.165	10	354.168	8503.514	0.152
10	557.659	8378.563	0.165	8	484.277	8496.925	0.153
5	557.217	8803.207	0.123	6	480.192	8706.782	0.132
6	553.874	8883.124	0.115	7	488.649	8785.697	0.124
7	557.374	8923.306	0.111	4	490.109	8923.306	0.111
	m=40				m=40		
4	197.895	8349.664	0.155	5	197.895	8501.890	0.140
9	216.326	8352.848	0.155	9	286.767	8503.581	0.140
8	258.843	8354.303	0.155	10	325.620	8512.801	0.139
10	542.028	8350.235	0.155	8	438.535	8514.404	0.138
5	539.464	8799.447	0.110	6	439.546	8693.521	0.120
6	540.381	8866.892	0.103	7	445.358	8755.153	0.114
7	542.388	8896.128	0.100	4	446.425	8896.128	0.100
	m=50				m=50		
4	197.616	8399.792	0.156	5	197.616	8597.829	0.136
9	214.180	8401.249	0.155	9	214.180	8602.097	0.135
8	265.189	8405.032	0.155	8	265.189	8624.089	0.133
10	569.189	8404.271	0.155	10	428.830	8623.104	0.133
5	568.701	8878.237	0.107	4	428.199	8878.237	0.107
6	571.127	8957.765	0.099	6	442.304	8957.765	0.099
7	572.062	8983.590	0.097	7	445.041	8983.590	0.097
	m=100				m=100		
4	200.623	8395.959	0.178	5	200.623	8701.381	0.148
9	221.476	8398.192	0.178	9	221.476	8709.983	0.148
8	264.124	8416.769	0.176	8	264.124	8713.578	0.147
10	602.183	8416.051	0.176	10	408.358	8713.874	0.147
5	601.716	8935.497	0.125	4	408.550	8935.497	0.125
6	613.806	9001.027	0.119	6	410.890	9001.027	0.119
7	615.259	9033.068	0.116	7	416.488	9033.068	0.116

Tabla A.10: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 50 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-30.618	8308.170	0.187	4	251.414	8308.170	0.187
5	-25.672	8614.197	0.157	9	255.143	8285.006	0.190
9	-11.858	8616.787	0.157	8	280.598	8254.954	0.193
7	-8.964	8724.654	0.147	10	373.122	8252.752	0.193
6	-6.321	8726.552	0.146	5	366.649	8612.118	0.158
8	-10.759	8719.341	0.147	6	354.625	8685.972	0.150
10	0.116	8715.798	0.147	7	334.228	8715.798	0.147
	m=20				m=20		
4	-19.573	8308.539	0.148	4	229.287	8308.539	0.148
5	-17.104	8680.216	0.110	9	212.174	8308.609	0.148
7	-13.400	8870.050	0.090	8	244.551	8302.369	0.149
6	-10.007	8860.833	0.091	10	341.579	8305.426	0.148
8	-12.097	8860.105	0.091	5	333.016	8676.212	0.110
10	-20.648	8861.845	0.091	6	330.546	8790.749	0.099
9	-15.568	8855.987	0.092	7	330.570	8855.987	0.092
	m=30				m=30		
4	-3.758	8379.024	0.165	4	221.657	8379.024	0.165
5	-3.058	8802.831	0.123	9	228.028	8384.668	0.164
7	-1.329	8918.087	0.111	8	250.387	8379.277	0.165
6	5.459	8919.178	0.111	10	337.044	8378.563	0.165
8	5.580	8920.715	0.111	5	326.792	8803.207	0.123
10	4.921	8926.418	0.110	6	325.041	8883.124	0.115
9	19.744	8923.306	0.111	7	326.202	8923.306	0.111
	m=40				m=40		
4	0.179	8349.664	0.155	4	197.899	8349.664	0.155
5	1.947	8801.151	0.109	9	205.033	8352.848	0.155
7	3.876	8895.433	0.100	8	228.654	8354.303	0.155
6	8.489	8896.340	0.100	10	329.661	8350.235	0.155
8	8.739	8898.058	0.100	5	328.043	8799.447	0.110
10	11.806	8900.472	0.099	6	328.445	8866.892	0.103
9	20.543	8896.128	0.100	7	328.531	8896.128	0.100
	m=50				m=50		
4	1.084	8399.792	0.156	4	189.717	8399.792	0.156
5	1.083	8872.023	0.108	9	196.236	8401.249	0.155
7	3.233	8972.854	0.098	8	220.106	8405.032	0.155
6	7.656	8976.396	0.098	10	325.557	8404.271	0.155
8	8.317	8981.059	0.097	5	325.418	8878.237	0.107
10	9.498	8983.588	0.097	6	326.409	8957.765	0.099
9	20.185	8983.590	0.097	7	326.636	8983.590	0.097
	m=100				m=100		
4	3.160	8395.959	0.178	4	177.669	8395.959	0.178
5	3.440	8932.775	0.126	9	186.151	8398.192	0.178
9	6.260	8936.507	0.125	8	209.493	8416.769	0.176
10	6.687	8937.014	0.125	10	331.828	8416.051	0.176
8	6.758	8935.497	0.125	5	331.476	8935.497	0.125
7	9.187	9031.883	0.116	6	331.066	9001.027	0.119
6	9.499	9033.068	0.116	7	331.567	9033.068	0.116

Tabla A.11: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 11,5 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
6	161.363	8806.386	0.139	4	161.363	8453.594	0.173
8	239.288	8807.478	0.139	9	165.402	8462.873	0.172
10	249.328	8816.254	0.138	8	232.670	8468.460	0.172
9	295.799	8819.412	0.137	10	507.984	8468.622	0.172
5	297.709	8896.239	0.130	5	508.082	8923.778	0.127
7	303.018	8912.837	0.128	6	511.462	9034.988	0.116
4	303.678	9041.664	0.116	7	517.074	9041.664	0.116
	m=20				m=20		
4	160.493	8441.227	0.134	4	160.493	8441.227	0.134
8	165.060	8450.038	0.134	8	165.060	8450.038	0.134
10	248.835	8451.968	0.133	10	248.835	8451.968	0.133
9	532.971	8455.302	0.133	9	532.971	8455.302	0.133
5	535.105	8899.328	0.087	5	535.105	8899.328	0.087
6	536.340	9030.245	0.074	6	536.340	9030.245	0.074
7	541.978	9037.383	0.073	7	541.978	9037.383	0.073
	m=30				m=30		
4	208.056	8499.274	0.153	4	208.056	8499.274	0.153
8	219.067	8505.819	0.152	9	219.067	8505.037	0.152
10	280.254	8506.832	0.152	8	280.254	8509.628	0.152
9	540.265	8509.660	0.152	10	540.265	8509.660	0.152
5	542.091	8912.349	0.112	5	540.286	8912.349	0.112
6	542.745	9007.112	0.102	6	543.250	9007.112	0.102
7	546.971	9024.166	0.100	7	546.971	9024.166	0.100
	m=40				m=40		
4	184.002	8415.270	0.148	4	184.002	8415.270	0.148
8	196.438	8419.535	0.148	9	196.438	8417.022	0.148
10	252.941	8420.157	0.148	8	252.941	8420.391	0.148
9	544.299	8420.078	0.148	10	544.299	8420.078	0.148
5	544.249	8871.986	0.102	5	544.098	8871.986	0.102
6	544.650	8959.624	0.093	6	546.270	8959.624	0.093
7	547.399	8978.912	0.091	7	547.399	8978.912	0.091
	m=50				m=50		
4	181.913	8460.809	0.149	4	181.913	8460.809	0.149
8	193.689	8472.063	0.148	9	193.689	8462.925	0.149
10	245.485	8473.530	0.148	8	245.485	8472.565	0.148
9	552.950	8473.859	0.148	10	552.950	8473.859	0.148
5	553.163	8948.028	0.100	5	553.789	8948.028	0.100
6	554.114	9027.909	0.092	6	560.040	9027.909	0.092
7	561.412	9046.069	0.091	7	561.412	9046.069	0.091
	m=100				m=100		
4	177.866	8431.980	0.175	4	177.866	8431.980	0.175
8	192.734	8441.939	0.174	9	192.734	8436.172	0.174
10	242.765	8444.344	0.174	8	242.765	8443.691	0.174
9	585.536	8444.340	0.174	10	585.536	8444.340	0.174
5	585.534	8969.759	0.122	5	585.960	8969.759	0.122
6	587.102	9046.449	0.115	6	590.865	9046.449	0.115
7	593.600	9069.240	0.112	7	593.600	9069.240	0.112

Tabla A.12: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 11,5 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-12.257	8453.594	0.173	4	190.960	8453.594	0.173
5	-12.325	8918.108	0.128	9	193.617	8462.873	0.172
7	-11.292	9016.569	0.118	8	219.194	8468.460	0.172
6	-12.169	9036.189	0.116	10	326.870	8468.622	0.172
9	-7.365	9039.253	0.116	5	327.016	8923.778	0.127
10	-0.259	9041.756	0.116	6	328.533	9034.988	0.116
8	20.584	9041.664	0.116	7	331.462	9041.664	0.116
	m=20				m=20		
4	-9.167	8441.227	0.134	4	191.384	8441.227	0.134
5	-9.208	8896.204	0.088	9	194.392	8448.045	0.134
7	-8.335	9005.768	0.077	8	222.861	8454.909	0.133
6	-8.555	9028.116	0.074	10	327.519	8455.302	0.133
9	-5.191	9035.543	0.074	5	327.018	8899.328	0.087
10	2.369	9037.135	0.073	6	328.592	9030.245	0.074
8	10.288	9037.383	0.073	7	330.668	9037.383	0.073
	m=30				m=30		
4	7.905	8499.274	0.153	4	206.376	8499.274	0.153
5	8.014	8909.360	0.112	9	212.140	8505.037	0.152
7	10.145	9006.623	0.102	8	231.005	8509.628	0.152
6	10.948	9016.094	0.101	10	326.818	8509.660	0.152
9	12.376	9022.274	0.101	5	326.687	8912.349	0.112
10	17.906	9024.221	0.100	6	327.595	9007.112	0.102
8	24.240	9024.166	0.100	7	329.256	9024.166	0.100
	m=40				m=40		
4	6.843	8415.270	0.148	4	195.309	8415.270	0.148
5	8.129	8866.025	0.103	9	200.736	8417.022	0.148
9	13.800	8868.891	0.103	8	218.796	8420.391	0.148
10	13.432	8870.735	0.102	10	328.150	8420.078	0.148
8	17.158	8871.986	0.102	5	328.145	8871.986	0.102
7	18.186	8971.120	0.092	6	328.831	8959.624	0.093
6	23.406	8978.912	0.091	7	329.270	8978.912	0.091
	m=50				m=50		
4	9.000	8460.809	0.149	4	190.546	8460.809	0.149
5	10.394	8941.039	0.101	9	195.465	8462.925	0.149
9	12.285	8943.998	0.101	8	214.038	8472.565	0.148
10	12.006	8946.368	0.101	10	326.363	8473.859	0.148
8	14.363	8948.028	0.100	5	326.507	8948.028	0.100
7	14.914	9038.366	0.091	6	328.076	9027.909	0.092
6	18.375	9046.069	0.091	7	328.495	9046.069	0.091
	m=100				m=100		
4	6.067	8431.980	0.175	4	179.008	8431.980	0.175
5	6.853	8959.013	0.123	9	186.783	8436.172	0.174
9	8.426	8962.070	0.123	8	207.701	8443.691	0.174
10	8.319	8969.148	0.122	10	330.707	8444.340	0.174
8	9.185	8969.759	0.122	5	330.852	8969.759	0.122
7	10.351	9063.841	0.113	6	332.542	9046.449	0.115
6	8.019	9069.240	0.112	7	333.473	9069.240	0.112

Tabla A.13: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_0 y H_1 ($\times 10^3$). Falta de respuesta 23,92 %. Variable de interés: Puntuación matemáticas

v.a.	H_0	$\tilde{Y}_{cal}$	DR	v.a.	H_1	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	153.596	8589.568	0.160	5	153.596	8636.576	0.155
9	150.358	8599.619	0.159	9	274.465	8635.813	0.155
8	216.741	8605.040	0.158	10	277.461	8639.685	0.155
10	455.334	8601.617	0.159	8	431.552	8641.012	0.155
5	453.268	8996.849	0.120	6	432.353	8896.266	0.130
6	456.540	9106.814	0.109	7	434.691	8901.230	0.129
7	462.608	9101.449	0.110	4	434.230	9101.449	0.110
	m=20				m=20		
4	185.474	8485.126	0.130	4	185.474	8485.126	0.130
9	187.163	8495.493	0.129	9	187.163	8495.493	0.129
8	273.805	8495.116	0.129	8	273.805	8495.116	0.129
10	503.903	8496.291	0.129	10	503.903	8496.291	0.129
5	504.663	8851.871	0.092	5	504.663	8851.871	0.092
6	504.419	8985.762	0.079	6	504.419	8985.762	0.079
7	511.127	8988.372	0.078	7	511.127	8988.372	0.078
	m=30				m=30		
4	217.721	8560.109	0.147	5	217.721	8535.969	0.149
9	224.309	8571.194	0.146	9	353.952	8539.244	0.149
8	276.104	8572.289	0.146	10	381.531	8542.231	0.149
10	516.223	8571.233	0.146	8	530.922	8549.195	0.148
5	515.519	8931.239	0.110	6	535.567	8773.173	0.126
6	516.249	9008.894	0.102	7	537.559	8814.523	0.121
7	523.642	9018.770	0.101	4	539.743	9018.770	0.101
	m=40				m=40		
4	208.850	8408.199	0.149	5	208.850	8559.226	0.134
9	216.976	8411.225	0.149	9	302.621	8562.675	0.134
8	266.961	8413.712	0.149	10	326.579	8568.760	0.133
10	558.843	8414.425	0.149	8	451.727	8576.259	0.132
5	559.315	8855.407	0.104	6	456.691	8765.335	0.113
6	560.961	8930.926	0.096	7	460.718	8801.532	0.109
7	562.964	8943.203	0.095	4	463.001	8943.203	0.095
	m=50				m=50		
4	200.678	8489.334	0.147	5	200.678	8618.898	0.133
9	206.548	8491.038	0.146	9	299.157	8623.523	0.133
8	249.161	8493.653	0.146	10	322.197	8637.554	0.132
10	545.590	8496.560	0.146	8	446.134	8646.885	0.131
5	547.513	8944.606	0.101	6	452.308	8834.214	0.112
6	549.244	9009.015	0.094	7	461.591	8869.038	0.108
7	550.371	9017.887	0.093	4	464.651	9017.887	0.093
	m=100				m=100		
4	176.202	8438.939	0.174	5	176.202	8739.512	0.145
9	191.197	8444.336	0.174	9	247.919	8746.225	0.144
8	231.513	8453.042	0.173	10	278.277	8757.569	0.143
10	577.828	8456.761	0.172	8	374.700	8762.723	0.142
5	580.297	8978.400	0.121	6	378.122	8907.962	0.128
6	586.077	9039.126	0.115	7	385.653	8953.688	0.124
7	589.660	9061.712	0.113	4	390.110	9061.712	0.113

Tabla A.14: Estimador de calibración $\tilde{Y}_{cal}$ ($\times 10^{-3}$), Desviación relativa (DR) e Indicador H_2 y H_3 ($\times 10^3$). Falta de respuesta 23,92 %. Variable de interés: Puntuación matemáticas

v.a.	H_2	$\tilde{Y}_{cal}$	DR	v.a.	H_3	$\tilde{Y}_{cal}$	DR
	m=10				m=10		
4	-16.573	8589.568	0.160	4	213.332	8589.568	0.160
5	-17.079	9003.237	0.119	9	209.787	8599.619	0.159
7	-17.182	9087.972	0.111	8	233.470	8605.040	0.158
6	-16.531	9108.280	0.109	10	337.111	8601.617	0.159
9	-9.316	9106.732	0.109	5	333.903	8996.849	0.120
8	1.208	9106.210	0.109	6	334.074	9106.814	0.109
10	5.566	9101.449	0.110	7	336.941	9101.449	0.110
	m=20				m=20		
4	-16.250	8485.126	0.130	4	209.877	8485.126	0.130
5	-15.138	8851.383	0.092	9	211.313	8495.493	0.129
7	-14.948	8967.951	0.080	8	239.387	8495.116	0.129
6	-15.344	8987.386	0.078	10	335.574	8496.291	0.129
9	-10.605	8989.160	0.078	5	331.974	8851.871	0.092
10	0.837	8989.432	0.078	6	331.875	8985.762	0.079
8	0.745	8988.372	0.078	7	334.178	8988.372	0.078
	m=30				m=30		
4	11.554	8560.109	0.147	4	214.504	8560.109	0.147
5	11.166	8930.039	0.110	9	219.056	8571.194	0.146
7	11.277	9010.803	0.102	8	240.992	8572.289	0.146
6	12.387	9017.958	0.101	10	332.867	8571.233	0.146
8	13.853	9020.888	0.101	5	330.362	8931.239	0.110
10	19.901	9018.407	0.101	6	330.515	9008.894	0.102
9	29.245	9018.770	0.101	7	332.205	9018.770	0.101
	m=40				m=40		
4	11.244	8408.199	0.149	4	198.255	8408.199	0.149
5	11.201	8849.451	0.105	9	203.734	8411.225	0.149
7	13.046	8931.052	0.096	8	227.376	8413.712	0.149
6	13.764	8935.865	0.096	10	328.843	8414.425	0.149
9	14.877	8940.508	0.095	5	328.345	8855.407	0.104
10	19.080	8942.634	0.095	6	328.712	8930.926	0.096
8	24.368	8943.203	0.095	7	329.132	8943.203	0.095
	m=50				m=50		
5	9.343	8618.898	0.133	4	196.458	8489.334	0.147
4	9.126	8934.607	0.102	9	200.586	8491.038	0.146
7	9.343	9001.714	0.095	8	222.254	8493.653	0.146
6	12.465	9007.041	0.094	10	329.322	8496.560	0.146
8	13.070	9014.655	0.094	5	329.260	8944.606	0.101
10	14.446	9016.144	0.094	6	329.919	9009.015	0.094
9	10.281	9017.887	0.093	7	330.287	9017.887	0.093
	m=100				m=100		
5	8.158	8739.512	0.145	4	181.175	8438.939	0.174
4	8.236	8970.380	0.122	9	189.546	8444.336	0.174
7	8.506	9051.186	0.114	8	211.827	8453.042	0.173
6	9.762	9052.728	0.114	10	332.558	8456.761	0.172
9	9.926	9058.470	0.113	5	332.252	8978.400	0.121
10	12.013	9060.850	0.113	6	333.327	9039.126	0.115
8	13.511	9061.712	0.113	7	334.353	9061.712	0.113

Tabla A.15: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Estudiar carrera de ciencias (falta de respuesta simulada: 20 % y 30 %).

		Estudiar ciencias 20 %							Estudiar ciencias 30 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
		H_0							H_0						
254,6	10	5	8	9	10	4	7	6	4	9	8	10	5	7	6
508,8	20	6	10	9	8	4	7	5	5	10	8	9	6	4	7
761,1	30	6	10	9	8	4	5	7	5	10	8	9	4	7	6
1019,1	40	5	10	8	9	4	7	6	5	9	8	10	4	7	6
1273,7	50	5	10	8	9	4	7	6	5	9	8	10	4	7	6
2545,5	100	5	10	8	9	4	7	6	5	9	8	10	4	7	6
		H_1							H_1						
254,6	10	5	8	9	10	4	7	6	5	8	9	10	4	6	7
508,8	20	5	10	8	9	4	6	7	5	7	8	9	10	6	4
761,1	30	5	10	8	9	4	6	7	5	10	8	9	4	6	7
1019,1	40	5	10	8	9	4	7	6	5	10	8	9	4	6	7
1273,7	50	5	10	8	9	4	6	7	5	10	8	9	4	6	7
2545,5	100	5	7	10	9	8	6	4	5	10	8	9	4	6	7
		H_2							H_2						
254,6	10	5	6	9	10	8	7	4	5	9	6	7	10	8	4
508,8	20	8	9	10	5	6	7	4	6	5	9	10	8	7	4
761,1	30	8	9	10	5	6	7	4	6	5	9	10	8	7	4
1019,1	40	8	9	10	5	6	7	4	6	5	9	10	8	7	4
1273,7	50	7	6	5	9	10	8	4	6	5	9	10	8	7	4
2545,5	100	7	6	5	8	10	9	4	6	5	9	10	8	7	4
		H_3							H_3						
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	8	9	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	8	9	10	5	6	7	4	9	8	10	6	5	7

Tabla A.16: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Estudiar carrera de ciencias (falta de respuesta simulada: 40 % y 50 %).

		Estudiar ciencias 40 %							Estudiar ciencias 50 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
		H_0							H_0						
254,6	10	5	8	9	10	4	7	6	5	8	10	9	4	7	6
508,8	20	5	8	9	10	4	7	6	5	8	10	9	4	7	6
761,1	30	5	8	9	10	4	7	6	5	8	10	9	4	7	6
1019,1	40	5	10	8	9	4	7	6	5	8	10	9	4	7	6
1273,7	50	5	10	8	9	4	7	6	5	10	8	9	4	7	6
2545,5	100	5	10	8	9	4	7	6	5	10	8	9	4	7	6
		H_1							H_1						
254,6	10	5	8	9	10	4	6	7	6	9	8	10	4	5	7
508,8	20	7	8	9	10	5	6	4	6	8	9	10	4	5	7
761,1	30	5	8	9	10	4	7	6	6	10	8	9	4	5	7
1019,1	40	7	10	5	8	9	6	4	6	10	8	9	4	5	7
1273,7	50	6	10	8	9	4	5	7	5	10	8	9	4	6	7
2545,5	100	5	7	10	8	9	6	4	5	10	8	9	4	7	6
		H_2							H_2						
254,6	10	6	5	7	9	10	8	4	9	8	10	5	6	7	4
508,8	20	6	5	7	9	10	8	4	9	8	10	5	6	7	4
761,1	30	6	5	9	10	8	7	4	9	8	10	5	6	7	4
1019,1	40	6	5	7	9	10	8	4	9	8	10	5	6	7	4
1273,7	50	6	5	9	10	8	7	4	9	8	10	5	6	7	4
2545,5	100	6	5	9	10	8	7	4	9	8	10	5	6	7	4
		H_3							H_3						
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	9	8	10	5	6	7

Tabla A.17: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Clases particulares (falta de respuesta simulada: 20 % y 30 %).

		Clases particulares 20 %							Clases particulares 30 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
		H_0							H_0						
254,6	10	4	9	8	10	5	6	7	5	9	10	8	6	4	7
508,8	20	4	8	9	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	8	9	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	8	9	10	5	6	7	4	9	8	10	5	6	7
		H_1							H_1						
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	8	9	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	8	9	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	8	9	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	8	9	10	5	6	7	4	9	8	10	5	6	7
		H_2							H_2						
254,6	10	4	6	7	5	8	10	9	4	7	6	5	9	10	8
508,8	20	4	6	7	5	8	10	9	4	7	6	5	9	8	10
761,1	30	6	7	8	5	10	9	4	4	7	6	5	9	10	8
1019,1	40	4	7	10	6	5	8	9	4	5	7	6	9	10	8
1273,7	50	6	7	10	5	9	8	4	4	5	7	6	9	10	8
2545,5	100	8	9	10	5	7	6	4	4	5	7	6	9	10	8
		H_3							H_3						
254,6	10	4	9	8	10	5	6	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	9	8	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	8	10	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	9	8	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	9	8	10	5	6	7

Tabla A.18: Frecuencia en el orden de selección de las variables auxiliares. Indicadores H_0, H_1, H_2 y H_3. promedio de estudiantes (n) en (m) centros. Variables principales: Clases particulares (falta de respuesta simulada: 40 % y 50 %).

		Clases particulares 40 %							Clases particulares 50 %						
n	m	1.º	2.º	3.º	4.º	5.º	6.º	7.º	1.º	2.º	3.º	4.º	5.º	6.º	7.º
		H_0							H_0						
254,6	10	4	9	10	8	5	6	7	9	8	10	5	6	4	7
508,8	20	4	9	10	8	5	6	7	4	9	8	10	6	5	7
761,1	30	4	9	8	10	5	6	7	4	9	10	8	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	10	8	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	9	10	8	6	5	7
2545,5	100	4	9	8	10	5	6	7	4	9	10	8	6	5	7
		H_1							H_1						
254,6	10	9	8	10	5	6	4	7	9	8	10	5	6	4	7
508,8	20	4	9	10	8	5	6	7	4	9	8	10	6	5	7
761,1	30	4	9	8	10	5	6	7	4	8	9	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	9	10	8	6	5	7
1273,7	50	4	9	8	10	5	6	7	4	9	10	8	6	5	7
2545,5	100	4	9	8	10	5	6	7	4	9	10	8	6	5	7
		H_2							H_2						
254,6	10	4	7	6	5	9	8	10	4	7	6	5	8	10	9
508,8	20	4	7	5	6	8	10	9	4	7	6	5	8	10	9
761,1	30	4	7	5	6	9	10	8	8	9	10	4	5	7	6
1019,1	40	4	7	5	6	9	8	10	8	9	10	4	5	7	6
1273,7	50	9	8	10	5	6	7	4	8	9	10	4	5	7	6
2545,5	100	9	8	10	6	5	7	4	8	9	10	4	5	7	6
		H_3							H_3						
254,6	10	9	8	10	5	6	4	7	4	9	8	10	5	6	7
508,8	20	4	9	8	10	5	6	7	4	8	9	10	5	6	7
761,1	30	4	9	8	10	5	6	7	4	8	9	10	5	6	7
1019,1	40	4	9	8	10	5	6	7	4	8	9	10	5	6	7
1273,7	50	4	9	8	10	5	6	7	4	8	9	10	5	6	7
2545,5	100	4	9	8	10	5	6	7	4	8	9	10	5	6	7

CAPÍTULO

B

Anexo. Poblaciones

En esta sección, analizamos las poblaciones utilizadas en este trabajo.

B.1 PISA España

La población "PISA-España" son datos tomados del programa PISA (Programme for International Student Assessment) para España referente al año 2006. El estudio PISA es un programa desarrollado para evaluar a los estudiantes de 15 años, sus familias y centros donde realizan sus estudios, de países pertenecientes a la OCDE[40] (Organización para la Cooperación y el Desarrollo). El estudio del año 2006, que contó con la participación de 57 países, se ha centrado en las habilidades de lectura, matemáticas y ciencias.
Los datos han sido tomados de la web de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). Estos microdatos, que incluyen información sobre los estudiantes, las familias y las escuelas, han sido tratados para garantizar el anonimato de los participantes. Para nuestro estudio se utilizaron solamente los datos de España. Los microdatos del informe PISA-España contienen información sobre las pruebas realizadas en 686 escuelas, con la participación de 19604 estudiantes.
Un ejemplo, para una muestra concreta de los datos del estudio de simulación del capítulo 2 (que coincidiría con lo que llamaremos en el código R fichero XYP), estaría formado por las siguientes variables:

- v1 Centro (XYP[,1]; 1ª columna)
- v2 Nº de estudiantes del centro (XYP[,2]; 2ª columna)
- v3 Sexo del estudiante (XYP[,3]; 3ª columna)
- v4 Estudios del padre del estudiante (XYP[,4]; 4ª columna)
- v7 Tipo de centro (XYP[,7]; 7ª columna)
- v8 Tipo de localidad (XYP[,8]; 8ª columna)
- v9 Interés por estudiar una carrera científica (XYP[,9]; 9ª columna)

- v10 Total de alumnos de cada centro con interés por estudiar una carrera científica (XYP[,10]; 10ª columna)
- v11 Indicador de centro (XYP[,11]; 11ª columna)
- v12 Estudiate (XYP[,12]; 12ª columna)

y por datos como los siguientes:

Centro	Número estud.	Sexo	Est. padre	Tipo centro	Tipo local	Est. ciencias	Total centro	Indic. centro	Alumno
v1	v2	v3	v4	v7	v8	v9	v10	v11	v12
319	32	1	1	2	3	0	18	1	8752
319	32	1	2	2	3	1	18	0	8753
319	32	1	2	2	3	1	18	0	8754
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
319	32	1	2	2	3	0	18	0	8783
320	22	2	1	1	1	0	11	1	8784
320	22	2	1	1	1	0	11	0	8785
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
320	22	1	3	1	1	1	11	0	8805

Un ejemplo para una muestra concreta (que coincidiría con lo que llamaremos en el código R fichero DXYP) que hemos utilizado para el método de Integración de Pesos y de Contracción, estaría formado por el siguiente conjunto de datos, que han sido calculados mediante la función $disjunctive$ del paquete $sampling$. Esta función transforma una variable categórica en una matriz de indicadores. Los valores de la variable categórica son números enteros (positivos o negativos).

Ce.	N. est.	Sexo		Est. padre			Tipo centro		Tipo localidad			Est. cien.	Tot. ce.	In. ce.	Al.
v1	v2	v3	v4	v5	v6	v7	v14	v15	v16	v17	v18	v19	v20	v21	v22
319	32	1	0	1	0	0	0	1	0	0	1	0	18	1	8752
319	32	1	0	0	1	0	0	1	0	0	1	1	18	0	8753
319	32	1	0	0	1	0	0	1	0	0	1	1	18	0	8754
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
319	32	0	1	0	1	0	0	1	0	0	1	1	18	0	8783
319	22	0	1	1	0	0	1	0	1	0	0	0	11	1	8784
319	22	0	1	1	0	0	1	0	1	0	0	0	11	0	8785
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
320	22	1	0	0	0	1	1	0	1	0	0	1	11	0	8805

Un ejemplo para una muestra concreta (que coincidiría con lo que llamaremos en el código R fichero XYPIW) que hemos utilizado para el método de Integración de Pesos y de Contracción, estaría formado por el siguiente conjunto de datos, que han sido calculados mediante la media de los valores de cada centro de las variables a nivel persona y la media respecto al total de las variables a nivel centro.

Ce.	N. est.	Sexo		Estudios padre			Tipo centro		Tipo localidad			Est. cien.	Tot. ce.	In. ce.	Al.
v1	v2	v3	v4	v5	v6	v7	v14	v15	v16	v17	v18	v19	v20	v21	v22
319	32	0.531	0.469	0.469	0.313	0.219	0	0.031	0	0	0.031	0	18	1	8752
319	32	0.531	0.469	0.469	0.313	0.219	0	0.031	0	0	0.031	1	18	0	8753
319	32	0.531	0.469	0.469	0.313	0.219	0	0.031	0	0	0.031	1	18	0	8754
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
319	32	0.531	0.469	0.469	0.313	0.219	0	0.031	0	0	0.031	0	18	0	8783
320	22	0.455	0.545	0.5	0.182	0.318	0.045	0	0.045	0	0	0	11	1	8784
320	22	0.454	0.545	0.5	0.182	0.318	0.045	0	0.045	0	0	0	11	0	8785
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
320	22	0.454	0.545	0.5	0.182	0.318	0.045	0	0.045	0	0	1	11	0	8805

B.2 Encuesta de Presupuestos Familiares

La segunda población finita utilizada es la Encuesta de Presupuestos Familiares (EPF) llevada a cabo en 2006. Se trata de una muestra aleatoria representativa de una encuesta que se realiza en los hogares españoles. La web del Instituto Nacional de Estadística español (INE) incluye estadísticas que se pueden utilizar para obtener los archivos de microdatos, cada uno de los cuales contiene datos individuales de una estadística determinada, filtradas adecuadamente para que la información se anónima y por lo tanto garantizar la confidencialidad. Esta encuesta, que ofrece información sobre el gasto en consumo de los hogares y se utiliza para calcular el Índice de Precios de Consumo (IPC), tiene periodicidad anual e incluye cerca de 24.000 viviendas en su muestra.

Los microdatos EPF-2006 contiene 55.699 unidades agrupadas en 19435 conglomerados. Se consideraron sólo las unidades sin datos faltantes para las variables de estudio, y así se obtuvo una población de tamaño $N = 9243$ individuos agrupados en $N_I = 5800$ hogares. Las principales características de la EPF-2006 se puede consultar en la página web del INE[24]. Cada registro contiene las variables que se solicitan a nivel individual para todos los miembros, es decir, hay un registro

para cada miembro del hogar. Aquellas variables que se investigan exclusivamente para el sustentador principal, figuran en el fichero de hogar. Por lo que respecta a los ingresos individuales, en los ficheros correspondientes a 2006 no se ha realizado imputación alguna, de forma que en los ficheros hay en algunos casos falta de respuesta.
Un ejemplo, para una muestra concreta de los datos del estudio de simulación del capítulo 2 (que coincidiría con lo que llamaremos en el código R fichero XYP), estaría formado por las siguientes variables:

- v1 Hogar (XYP[,1]; 1ª columna)
- v2 Nº de residentes en el hogar (XYP[,2]; 2ª columna)
- v3 Sexo del residente (XYP[,3]; 3ª columna)
- v4 Nivel educativo del residente (XYP[,4]; 4ª columna)
- v7 Nivel educativo del sustentador principal(XYP[,7]; 7ª columna)
- v8 Nivel de ingresos (XYP[,8]; 8ª columna)
- v10 Total de ingresos por hogar (XYP[,9]; 9ª columna)
- v11 Indicador de hogar (XYP[,10]; 10ª columna)
- v12 Indicador unidad (XYP[,11]; 11ª columna)

y por datos como los siguientes:

Hogar	Número resi.	Sexo	Nivel estudios	Nivel est. s.p.	Nivel ingresos	Total ingresos	Indic. hogar	Indic. unidad
v1	v2	v3	v4	v7	v8	v9	v10	v11
1451	3	1	1	1	30000	48000	1	8752
1451	3	2	2	1	18000	48000	0	8753
1451	3	1	2	1	0	48000	0	8754
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1452	2	1	2	2	15000	30000	1	8755
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1453	2	2	2	2	17000	31000	1	8757

Los ficheros DXYP y XYPIW, que hemos utilizado para el método de Integración de Pesos y de Contracción, estarían formados de igual forma que los descritos en la población PISA- España.

B.3 Encuesta Pre-electoral Octubre 2011 CIS

La tercera población utilizada, Capítulo 1, está formado por datos de la encuesta pre-electoral del CIS[45] (Centro de Investigaciones Sociológicas) de octubre de 2011, anterior a las elecciones generales al Congreso de los Diputados de noviembre de 2011. Esta encuesta se concibe para recabar información útil para analizar las claves del comportamiento de los electores en las elecciones generales y realizar estimaciones de las mismas.
La población investigada son los ciudadanos con derecho a voto que residen en España, constituido por ciudadanos españoles mayores de 18 años, residentes en España o en el extranjero. En las elecciones generales del 20 de noviembre de 2011 el censo electoral estaba compuesto de un total de 35.776.615 electores, siendo los españoles residentes en España su componente más numeroso (96 %).
El CIS realizó un muestreo polietápico, estratificado por conglomerados, con selección de las unidades primarias de muestreo (municipios) y de las unidades secundarias (secciones) de forma aleatoria proporcional. Los estratos se formaron mediante el cruce de las 17 CC.AA. con el tamaño de hábitat, dividido en 7 categorías: menor o igual a 2.000 habitantes; de 2001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000, y más de 1.000.000 de habitantes. Los cuestionarios se han aplicado mediante entrevista personal en los domicilios. El tamaño muestral está compuesto por 6082 individuos, a los que se le realizó la encuesta completa, repartidos en 214 municipios.
Para este estudio, se han seleccionado dos variables de interés (variables principales o variables objeto de estudio) correspondientes al Barómetro Pre-electoral de Centro de Investigación Sociológicas de octubre de 2011[5]:

- y_1 Intención de voto al PSOE en las elecciones generales de 2011.
- y_2 Participación en las elecciones generales de 2011.

Como variables auxiliares se ha tomado las variables correspondientes al Barómetro pre-electoral de Centro de Investigación Sociológicas de octubre de 2011:

- x_1 ¿Votó al PSOE en las elecciones generales de 2008?
- x_2 El encuestado está desempleado.
- x_3 El encuestado está jubilado.
- x_4 El encuestado vive en un área metropolitana.

Los totales poblacionales de cada variable auxiliar se han tomado de la web del Instituto Nacional de Estadística (INE) [24] y de la web del Ministerio de Empleo y Seguridad Social[34].
Un ejemplo del conjunto de datos es el siguiente

Munic.	Nº enc.	Votó PSOE 2008	Desemple.	Jubilado	Tipo local.	Votará PSOE 2011	Votará 2011	Enc.
99	87	1	1	0	3	0	1	2012
99	87	0	0	1	3	0	0	2013
99	87	1	0	1	3	1	0	2014
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	87	0	0	0	3	1	1	2099

CAPÍTULO

C

Anexo. Funciones en R

C.1 Funciones básicas para muestras y calibración

Muestreo aleatorio simple de centros

Resumimos en la siguientes líneas los programas realizados en $\mathcal{R}$. Para estos ejemplos utilizaremos la nomenclatura utilizada para los ejemplos con la Población PISA-España

En primer lugar vamos a guardar en memoria la población de trabajo XYP, descrita en el Anexo B.1, para ello leeremos los datos con la función *read.table* y le asignaremos un nombre, en este caso XYP.

Nota: *Para los ejemplos partiremos del conjunto de datos XYP (ver anexo B), que representan a una muestra de alumnos españoles que participaron en las pruebas de evaluación del informe PISA. Los datos DXYP son variantes de XYP en los que se han transformado algunas variables categóricas en una matriz de indicadores y en el caso XIPIW las valores de las variables han sido reemplazados por medias ponderadas según la unidad primaria a la que pertenezcan. Todos estos datos y las variables que los componen se especifica en el anexo.*

```
>read.table("XYP.txt")->XYP
```

La primera columna del conjunto de datos XYP indica el centro al que pertenece cada estudiante, que almacenamos en la variable *centro* con la orden:

```
>XYP[,1]->centro
```

El número de centros se puede calcular con la orden (lo habitual en las aplicaciones prácticas es conocer el número de centros):

```
>nlevels(factor(centro))->M
```

Para extraer una muestra aleatoria simple de centros, de tamaño m=20, de la población de centros, y las probabilidades de inclusión de primer y segundo orden, escribimos las órdenes:

```
>m<-20
>s1<-sample(M,m)
>pi1.mas<-rep(m/M,m)
>kk<-matrix(m*(m-1)/(M*(M-1)),m,m)
>diag(kk)<-m/M
>pi2.mas<-kk
```

Podemos visualizar los centros que componen la muestra y las probabilidades anteriores escribiendo:

```
>s1
 [1] 455 485 190 503 57 122 220 237 424 232 79 342 ...
>pi1.mas
 [1] 0.02971768 0.02971768 0.02971768 0.02971768
 [5] 0.02971768 0.02971768 0.02971768 0.02971768
 [9] 0.02971768 0.02971768 0.02971768 0.02971768 ...

>pi2.mas
             [,1]         [,2]         [,3]         [,4] ...
 [1,] 0.029717682 0.000840232 0.000840232 0.000840232 ...
 [2,] 0.000840232 0.029717682 0.000840232 0.000840232 ...
 [3,] 0.000840232 0.000840232 0.029717682 0.000840232 ...
 [4,] 0.000840232 0.000840232 0.000840232 0.029717682 ...
 ...
```

Los pesos de diseño serían:

```
>di1.mas<-1/pi1.mas; di1.mas
 [1] 33.65 33.65 33.65 33.65 33.65 33.65 33.65 33.65
 [9] 33.65 33.65 33.65 33.65 33.65 33.65 33.65 33.65 ...
```

Obsérvese que hasta ahora no se ha observado ninguna variable en la muestra. El ejemplo, al igual que el siguiente, permite seleccionar las unidades, centros educativos, que componen una muestra aleatoria simple de una población de centros educativos. Con posterioridad veremos cómo seleccionar todas las unidades (estudiantes) que componen un centro.

Muestreo aleatorio simple de estudiantes

Otro ejemplo de muestreo aleatorio simple, pero en este caso de estudiantes, sería el que se muestra a continuación. Calculamos el número de unidades secundarias, estudiantes, disponibles en nuestro fichero de datos (este dato es habitual que sea conocido en las aplicaciones prácticas), *N*, y mediante la función *sample*, una muestra de estudiantes de tamaño m=20, y las respectivas probabilidades de inclusión de primer y segundo orden:

```
>length(centro)->N
>sample(N,m)->s2
>pi1.mas2<-rep(m/N,m)
>kk2<-matrix(m*(m-1)/(N*(N-1)),m,m)
>diag(kk2)<-m/N
>pi2.mas2<-kk2
```

Podemos visualizar el número de estudiantes disponibles en el fichero de datos, N, los $m = 20$ estudiantes que componen la muestra y las probabilidades de inclusión anteriores y los pesos del diseño, escribiendo:

```
>N
 [1] 18341
>s2
 [1] 1786  7391  7263  10596 17806 10492 15654
 [8] 2962  7866  7496  2087  6764  9260  13444...
>pi1.mas2
 [1] 0.001090453 0.001090453 0.001090453 0.001090453
 [5] 0.001090453 0.001090453 0.001090453 0.001090453
 [9] 0.001090453 0.001090453 0.001090453 0.001090453 ...
>pi2.mas2
      [,1]         [,2]         [,3]         [,4]         ...
 [1,] 1.090453e-03 1.129695e-06 1.129695e-06 1.129695e-06 ...
 [2,] 1.129695e-06 1.090453e-03 1.129695e-06 1.129695e-06 ...
 [3,] 1.129695e-06 1.129695e-06 1.090453e-03 1.129695e-06 ...
 [4,] 1.129695e-06 1.129695e-06 1.129695e-06 1.090453e-03 ...
 ...
>di1.mas2<-1/pi1.mas2; di1.mas2
 [1] 917.05 917.05 917.05 917.05 917.05 917.05 917.05
 [8] 917.05 917.05 917.05 917.05 917.05 917.05 917.05 ...
```

Muestreo de Midzuno

Para extraer una muestra, con probabilidades desiguales, por el método de Midzuno, vamos a utilizar las funciones definidas por Tillé y Matei (2009)[69] en el paquete de R *sampling*. Para calcular las probabilidades de inclusion por Midzuno debemos tomar una variable auxiliar tamaño, en este caso tomaremos la variable que hace referencia al conjunto de alumnos de cada centros. Calcularemos un vector, llamado *tam_centro*, seleccionando aquellos valores de la variables que poseen información a nivel centro. Calcularemos las probabilidades de inclusión de primer orden mediante la función *inclusionprobabilities*, para muestras de tamaño 20, y la de segundo orden mediante la función *UPmidzunopi2*, ambas funciones del paquete *sampling*, las cuales nos proporcionaran probabilidades de inclusión para todos los elementos de la población. Mediante la función *UPmidzuno*, seleccionamos los elementos que componen la muestra. En realidad esta función crea un vector de ceros y unos que indican los elementos elegidos, que posteriormente utilizaremos para calcular la muestra.

Nota: *XYP[,2] hace referencia al tamaño de cada centro (número de alumnos), que guardaremos en memoria como "estudiantes.centro", y XYP[,11] es el indicador de las variables centro, que guardaremos en memoria como "indicador.centro".*

```
>XYP[,2]->estudiantes.centro
>XYP[,11]->indicador.centro

>library(sampling)
>tam_centro<-subset(estudiantes.centro,indicador.centro==1)
>pi1.Mz<-inclusionprobabilities(tam_centro,m)
>pi2.Mz<-UPmidzunopi2(pi1.Mz)
>sm=UPmidzuno(pi1.Mz)
>(1:length(pi1.Mz))[sm==1]->s3; s3
 [1]  9  14  34  63 125 154 185 187 244 256 360 364 ...
>pi1.Mz[s3]->pi1.mz; pi1.mz
 [1] 0.03271359 0.03489450 0.03489450 0.03162314
 [5] 0.02508042 0.03053269 0.03053269 0.03271359 ...
>pi2.Mz[s3,s3]->pi2.mz; pi2.mz
              [,1]         [,2]         [,3]         [,4] ...
 [1,] 0.0327135925 0.0010854378 0.0010854378 0.0009830434 ...
 [2,] 0.0010854378 0.0348944987 0.0011537008 0.0010513063 ...
 [3,] 0.0010854378 0.0011537008 0.0348944987 0.0010513063 ...
```

```
 [4,] 0.0009830434 0.0010513063 0.0010513063 0.0316231394 ...
 ...
```

Muestreo aleatorio simple de conglomerados

La única precaución a la hora de calcular las probabilidades de inclusión en este caso, es que éstas son en número tantas como indique el tamaño aleatorio de la muestra. Así, si los centros que componen la muestra y sus probabilidades de inclusión son las anteriores:

```
>s1
 [1] 455 485 190 503 57 122 220 237 424 232 79 342 ...
>pi1.mas
 [1] 0.02971768 0.02971768 0.02971768 0.02971768
 [5] 0.02971768 0.02971768 0.02971768 0.02971768 ...
>pi2.mas
              [,1]         [,2]         [,3]         [,4] ...
 [1,] 0.029717682 0.000840232 0.000840232 0.000840232 ...
 [2,] 0.000840232 0.029717682 0.000840232 0.000840232 ...
 [3,] 0.000840232 0.000840232 0.029717682 0.000840232 ...
 [4,] 0.000840232 0.000840232 0.000840232 0.029717682 ...
 ...
```

se tendrá que observar cuántos estudiantes hay en los centros seleccionados en la muestra de centros:

```
>subset(estudiantes.centro,(centro %in% s1)&
+ (indicador.centro==1))->tam.mas
>rep(pi1.mas,tam.mas)->pi1.mas.p
```

Muestreo de conglomerados con probabilidades proporcionales al tamaño

Igual que en caso anterior, las probabilidades de inclusión son en número tantas como indique el tamaño aleatorio de la muestra. Así, si los centros que componen la muestra y sus probabilidades de inclusión son las anteriores:

```
>s3
 [1] 9  14 34 63 125 154 185 187 244 256 360 364 ...
>pi1.mz
 [1] 0.03271359 0.03489450 0.03489450 0.03162314
```

```
 [5] 0.02508042 0.03053269 0.03053269 0.03271359 ...
>pi2.Mz[s3,s3]->pi2.mz; pi2.mz
              [,1]         [,2]         [,3]         [,4] ...
 [1,] 0.0327135925 0.0010854378 0.0010854378 0.0009830434 ...
 [2,] 0.0010854378 0.0348944987 0.0011537008 0.0010513063 ...
 [3,] 0.0010854378 0.0011537008 0.0348944987 0.0010513063 ...
 [4,] 0.0009830434 0.0010513063 0.0010513063 0.0316231394 ...
 ...
```

se tendrá que observar cuántos estudiantes hay en los centros seleccionados en la muestra de centros:

```
>subset(estudiantes.centro,(centro %in% s3)&
+ (indicador.centro==1))->tam.mz
>rep(pi1.mz,tam.mz)->pi1.mz.p
```

Estimación sin información auxiliar

En este ejemplo se pretende estimar mediante el estimador de Horvitz-Thompson, a partir de una muestra de 20 centros elegidos mediante m.a.s. y Midzuno y otra muestra de 20 alumnos, el número total de alumnos que tienen intención de estudiar una carrera universitaria de ciencias. En primer lugar definimos la variable principal (alumnos que estudiarán una carrera de ciencias), en este caso esta variable ocupa la columna 9 del conjunto de datos XYP (ver anexo). Para realizar la estimación necesitamos conocer los valores de XYP para la muestra de centros (o para la muestra de alumnos), datos que llamaremos xys, esto se consigue mediante la función *subset*, teniendo en cuenta que la columna 1 indica el código de centro y la 12 el código del alumno. Calculamos, también, el estimador de la varianza del estimador de Horvitz-Thompson a partir de la función *varest* del paquete *sampling*. Como en el caso de los centros tenemos que tomar los resultados de todos los alumnos encuestados, por lo que asignamos a todos los estudiantes el mismo peso de diseño que coincide con el de su centro. Calculamos por tanto el número de alumnos de cada centro mediante *subset*, indicando que el centro pertenezca a la muestra y que tomando solo la fila con el indicador del centro (columna 11 con valor 1).

Nota: *XYP[,12] hace referencia a la variable que identifica al alumno dentro de la población, la guardaremos en memoria como "estudiante". los vectores:*

xys.mas[,9], xys.mas2[,9] y xys.mz[,9] proporcionan los valores de la variable de interés para cada una de las muestras.

```
>XYP[,12]->estudiante
>subset(XYP, centro %in% s1)->xys.mas
>subset(XYP, estudiante %in% s2)->xys.mas2
>subset(XYP, centro %in% s3)->xys.mz
>HT.mas<-sum(xys.mas[,9]*(1/pi1.mas.p))
>HT.mas
 [1] 8849.95
>varHT.mas<-varest(xys.mas[,9],pik=pi1.mas.p)
>varHT.mas
 [1] 160419.24
>HT.mas2<-sum(xys.mas2[,9]*(1/pi1.mas2))
>HT.mas2
 [1] 10087.55
>varHT.mas2<-varest(xysp.mas2[,9],pik=pi1.mas2)
>varHT.mas2
 [1] 4377173.77
>HT.mz<-sum(xys.mz[,9]*(1/pi1.mz.p))
>HT.mz
 [1] 7323.686
>varHT.mz<-varest(xys.mz[,9],pik=pi1.mz.p)
>varHT.mz
 [1] 136902.5
```

Estimación con información auxiliar por calibración

Nota: *totalXp hace referencia a los totales de las variables con información a nivel persona, dxys a los elementos de la muestra, xsp a las variables de la muestra con información a nivel persona.*

Con una variable auxiliar

Calculamos ahora los diferentes estimadores de la varianza mediante la función *varest* del paquete *sampling*, para las diferentes muestras y distancias.

Nota: *Para este ejemplo, en el que sólo realizaremos los cálculos para la distancia "linear", necesitamos cargar el fichero DXYP y totalXp que hace referencia a los totales de las variables con información a nivel persona y que calculamos a partir*

del vector de totales totalXY. En este ejemplo sólo vamos a utilizar los dos primeros valores de totalXp y las dos primeras columnas de cada muestra (xsp.mas, xsp.mas2 y xsp.mz) ya que hacen referencia a la variable auxiliar S="sexo".

```
>read.table("DXYP.txt")->DXYP
>colSums(DXYP)->totalXY
>c(totalXY[3:7])->totalXp
># Calculamos las diferentes muestras
>subset(DXYP, centro %in% s1)->dxys.mas
>subset(DXYP, estudiante %in% s2)->dxys.mas2
>subset(DXYP, centro %in% s3)->dxys.mz
>data.frame(dxys.mas[,3:7])->xsp.mas
>data.frame(dxys.mas2[,3:7])->xsp.mas2
>data.frame(dxys.mz[,3:7])->xsp.mz
># Calibración est. total y varianza,
># distancia Linear, m.a.s centros
>gl.S.mas=calib(xsp.mas[,1:2],d=1/pi1.mas.p,
+ totalXp[1:2],method="linear"); gl.S.mas
 [1] 0.8285508 1.0422658 0.8285508 1.0422658
 [5] 0.8285508 0.8285508 0.8285508 1.0422658 ...
>Calp.gl.S.mas<-sum((1/pi1.mas.p)*gl.S.mas*xys.mas[,9])
>Calp.gl.S.mas
 [1] 8102.125
>VCalp.gl.S.mas<-varest(xys.mas[,9], xsp.mas[,1:2],
+ pi1.mas.p, gl.S.mas/pi1.mas.p)
>VCalp.gl.S.mas
 [1] 159628.24
># Calibración est. total y varianza,
># distancia Linear, m.a.s alumnos
>gl.S.mas2=calib(xsp.mas2[,1:2],d=1/pi1.mas2,
+ totalXp[1:2],method="linear")
>gl.S.mas2
 [1] 1.005834 0.994166 1.005834 0.994166
 [5] 1.005834 1.005834 0.994166 1.005834 ...
>Calp.gl.S.mas2<-sum((1/pi1.mas2)*gl.S.mas2*xys.mas2[,9])
>Calp.gl.S.mas2
 [1] 10071.5
>VCalp.gl.S.mas2<-varest(xys.mas2[,9], xsp.mas2[,1:2],
+ pi1.mas2, gl.S.mas2/pi1.mas2)
>VCalp.gl.S.mas2
 [1] 3979249
```

```
>#  Calibración est. total y varianza,
>#  distancia Linear, Midzuno centros
>gl.S.mz=calib(xsp.mz[,1:2],d=1/pi1.mz.p,totalXp[1:2],
+ method="linear")
>gl.S.mz
 [1] 0.998945 1.001070 0.998945 0.998945
 [5] 0.998945 1.001070 0.998945 1.001070 ...
>Calp.gl.S.mz<-sum((1/pi1.mz)*gl.S.mz*xys.mz[,9])
>Calp.gl.S.mz
 [1] 7323.46
>VCalp.gl.S.mz<-varest(xys.mz[,9], xsp.mz[,1:2],
+ pi1.mz.p, gl.S.mz/pi1.mz.p)
>VCalp.gl.S.mz
 [1] 136602.9
```

Más de una variable auxiliar

Realizamos el mismo proceso para comprobar el comportamiento de los estimadores cuando aumentamos el número de variables auxiliares. En este ejemplo vamos a calcular los estimadores del total y de la varianza del total, para cada tipo de muestra, tomando como variables auxiliares "sexo" y "nivel de estudios del padre".

Nota: *Para este ejemplo, utilizando la distancia linear, partiremos de los mismos ficheros, y valores, iniciales del anterior (totalXp y xsp.mas), ya que hacen referencia a las variables con información a nivel persona S="sexo" y E="nivel de estudios del padre".*

```
>#  Calibración est. total y varianza,
>#  distancia Linear, m.a.s centros
>gl.SE.mas=calib(xsp.mas,d=1/pi1.mas.p,totalXp,method="linear")
>gl.SE.mas
 [1] 0.9365234 1.1406250 0.9365234 0.9169922
 [5] 0.9365234 0.7128906 0.7128906 1.1406250 ...
>Calp.gl.SE.mas<-sum((1/pi1.mas.p)*gl.SE.mas*xys.mas[,9])
>Calp.gl.SE.mas
 [1] 7947.775
>VCalp.gl.SE.mas<-varest(xys.mas[,9], xsp.mas, pi1.mas.p,
+ gl.SE.mas/pi1.mas.p)
>VCalp.gl.SE.mas
 [1] 154645.1
```

```
>#  Calibración est. total y varianza,
>#  distancia Linear, m.a.s alumnos
>gl.SE.mas2=calib(xsp.mas2,d=1/pi1.mas2,totalXp,method="linear")
>gl.SE.mas2
 [1] 1.2720048 1.0180715 0.5370022 0.8789205
 [5] 0.6761532 0.5370022 0.8789205 0.6761532 ...
>Calp.gl.SE.mas2<-sum((1/pi1.mas2)*gl.SE.mas2*xys.mas2[,9])
>Calp.gl.SE.mas2
 [1] 9342.823
>VCalp.gl.SE.mas2<-varest(xys.mas2[,9], xsp.mas2, pi1.mas2,
+ gl.SE.mas2/pi1.mas2)
>VCalp.gl.SE.mas2
 [1] 3694156
>#  Calibración est. total y varianza,
>#  distancia Linear, Midzuno centros
>gl.SE.mz=calib(xsp.mz,d=1/pi1.mz.p,totalXp,method="linear")
>gl.SE.mz
 [1] 0.9726562 1.0878906 0.9726562 0.9648438
 [5] 0.9726562 1.0878906 1.1015625 0.9511719 ...
>Calp.gl.SE.mz<-sum((1/pi1.mz)*gl.SE.mz*xys.mz[,9])
>Calp.gl.SE.mz
 [1] 7426.54
>VCalp.gl.S.mz<-varest(xys.mz[,9], xsp.mz, pi1.mz.p,
+ gl.S.mz/pi1.mz.p)
>VCalp.gl.S.mz
 [1] 135737.7
```

Información auxiliar compleja. Integración de pesos

Nota: *Para utilizar el estimador propuesto por Estevao y Särndal seleccionamos las variables auxiliares: S="sexo", E="nivel de estudios del padre", ambas a nivel alumno y la variable C="tipo de centro"que indica si este es público o privado, a nivel centro. Para este ejemplo, a partir de la distancia linear, utilizamos la base de datos XYPIW (ver anexo) y los totales a nivel alumno (totalXp), ya definido anteriormente, y a nivel centro (totalXh), que calculamos a partir de los totales de la variable auxiliar a nivel centro, tanto para la variable "tipo de centro" como "tipo de localidad" que utilizaremos posteriormente.*

```
>read.table("XYPIW")->XYPIW
```

```
>#  Calculamos totales a nivel centro
>#  (Tipo de centro y tipo de localidad)
>colSums(XYPIW)->totalXYPIW
>c(totalXYPIW[14:18])->totalXh
>#  Agrupamos los dos totales (alumno y centro)
>c(totalXp,totalXh)->totalXiwp
>#  Calculamos las muestras
>subset(XYPIW, centro %in% s1)->xysiw.mas
>subset(XYPIW, centro %in% s3)->xysiw.mz
>data.frame(xysiw.mas[,3:7],xysiw.mas[,14:18])->xsiwp.mas
>data.frame(xysiw.mz[,3:7],xysiw.mz[,14:18])->xsiwp.mz
>#  I. P. total y varianza, distancia Linear, m.a.s centros
>gliw.SEC.mas=calib(xsiwp.mas[,1:7],d=1/pi1.mas.p,
+ totalXiwp[1:7],method="linear")
>gliw.SEC.mas
 [1] 0.8855058 0.8855058 0.8855058 0.8855058 0.8855058
 [6] 0.8855058 0.8855058 0.8855058 0.8855058 0.8855058 ...
>Cal.gliw.SEC.mas<-sum((1/pi1.mas.p)*gliw.SEC.mas*xys.mas[,9])
>Cal.gliw.SEC.mas
 [1] 7688.254
>VCal.gliw.SEC.mas<-varest(xys.mas[,9], xysiw.mas[,1:7],
+ pi1.mas.p, gliw.SEC.mas/pi1.mas.p)
>VCal.gliw.SEC.mas
 [1] 153292.4
>#  I. P. total y varianza, distancia Linear, Midzuno de centros
>gliw.SEC.mz=calib(xsiwp.mz[,1:7],d=1/pi1.mz.p,
+ totalXiwp[1:7],method="linear")
>gliw.SEC.mz
 [1] 1.1939640  1.1939640  1.1939640  1.1939640
 [5] 1.1939640  1.1939640  1.1939640  1.1939640 ...
>Cal.gliw.SEC.mz<-sum((1/pi1.mz)*gliw.SEC.mz*xys.mz[,9])
>Cal.gliw.SEC.mz
 [1] 7375.279
>VCal.gliw.SEC.mz<-varest(xys.mz[,9], xysiw.mz[,1:7],
+ pi1.mz.p, gliw.SEC.mz/pi1.mz.p)
>VCal.gliw.SEC.mz
 [1] 135376.8
```

Para comprobar el comportamiento de este estimador realizamos otro ejemplo incrementando el número de variables auxiliares a nivel centro.

Nota: *Para este ejemplo seleccionamos las variables auxiliares: S="sexo", E="nivel*

de estudios del padre", ambas a nivel alumno y las variable C="tipo de centro" que indica si este es público o privado y L="tipo de localidad", a nivel centro. Para este ejemplo, a partir de la distancia linear, utilizamos los mismos ficheros y variables que en el ejemplo anterior adaptados a las cuatro variables.

```
>#  I. P. total y varianza, distancia Linear, m.a.s centros
>gliw.SECL.mas=calib(xsiwp.mas,d=1/pi1.mas.p,totalXiwp,method="linear")
>gliw.SECL.mas
 [1] 1.11224855  1.11224855  1.11224855  1.11224855
 [5] 1.11224855  1.11224855  1.11224855  1.11224855 ...
>Cal.gliw.SECL.mas<-sum((1/pi1.mas.p)*gliw.SECL.mas*xys.mas[,9])
>Cal.gliw.SECL.mas
 [1] 8103.85
>VCal.gliw.SEC.mas<-varest(xys.mas[,9], xysiw.mas, pi1.mas.p,
+ gliw.SEC.mas/pi1.mas.p)
>VCal.gliw.SEC.mas
 [1] 152941.7
>#  I. P. total y varianza, distancia Linear, Midzuno de centros
>gliw.SECL.mz=calib(xsiwp.mz,d=1/pi1.mz.p,totalXiwp,method="linear")
>gliw.SECL.mz
 [1] 1.0744735  1.0744735  1.0744735  1.0744735
 [5] 1.0744735  1.0744735  1.0744735  1.0744735 ...
>Cal.gliw.SECL.mz<-sum((1/pi1.mz)*gliw.SECL.mz*xys.mz[,9])
>Cal.gliw.SECL.mz
 [1] 7300.264
>VCal.gliw.SECL.mz<-varest(xys.mz[,9], xysiw.mz, pi1.mz.p,
+ gliw.SECL.mz/pi1.mz.p)
>VCal.gliw.SECL.mz
 [1] 134122.9
```

Método de contracción

Para el ejemplo en cuestión, vamos a calcular los estimadores a nivel centro, con información a nivel centro, para posteriormente realizar la combinación con los estimadores con información a nivel alumno. Calcularemos en primer lugar los elementos que componen la muestra de centros, particularizando a las variables con información a nivel centro. Calcularemos los totales de las variables a nivel centro, totalXh, y utilizaremos la variables, calculadas anteriormente, tam.mas y tam.mz que indican el número de alumnos por centro.

Nota: *Calculamos los estimadores a nivel centro, para m.a.s y muestro por Midzuno de centros, para la distancia linear. Posteriormente reponderamos los pesos según el número de alumnos de cada centro para calcular el estimador, no calculado por calibración, a nivel alumno. Las variables dxyph.mas[,19] y dxyph.mz[,19] hacen referencia a las variables de estudio para cada muestra.*

```
>subset(DXYP, centro %in% s1 & indicador.centro==1)->dxyph.mas
>data.frame(dxyph.mas[,14:18])->xsh.mas
>subset(DXYP, centro %in% s3 & indicador.centro==1)->dxyph.mz
>data.frame(dxyph.mz[,14:18])->xsh.mz
>colSums(XYPIW)->totalXYPIW
>c(totalXYPIW[14:18])->totalXh
>glh.C.mas=calib(xsh.mas,d=1/pi1.mas,totalXh,method="linear")
>glh.C.mas
 [1] 1.0116019 1.0116019 2.5070918 0.7081213
 [5] 0.7657556 0.7081213 0.7081213 0.7081213 ...
>Cal.glh.C.mas<-sum(gl.h.mas*(1/pi1.mas)*dxyph.mas[,19])
>Cal.glh.C.mas
 [1] 299.2929
>glh.C.mz=calib(xsh.mz,d=1/pi1.mz,totalXh,method="linear")
>glh.C.mz
 [1] 0.8966428 0.8966428 0.2533553 1.6795068
 [7] 0.9716596 0.2533553 0.8966428 0.8966428 ...
>Cal.glh.C.mz<-sum(gl.h.mz*(1/pi1.mz)*dxyph.mz[,19])
>Cal.glh.C.mz
 [1] 251.4239
```

A partir de los estimadores a nivel centro, con información a nivel centro, reponderamos los pesos para calcular estimaciones a nivel alumno a partir del número de alumnos de cada centro (*tam.mas* y *tam.mz*). Posteriormente calculamos los estimadores de la varianza a nivel centro, para m.a.s y muestro por Midzuno de centros.

```
>wigl.h.mas<-gl.h.mas*(1/pi1.mas)
```

```
>wigl.h.mas.p<-rep(wigl.h.mas,tam.mas)
>noCal.p.gl.mas<-sum(wigl.h.mas.p*xys.mas[,9])
>noCal.p.gl.mas
 [1] 8792.576
>Vnocal.gl.h.mas<-varest(dxyph.mas[,19], xsh.mas, pi1.mas,
+ gl.h.mas/pi1.mas)
>Vnocal.gl.h.mas
 [1] 5453.269
># Est. a nivel alumno a partir de las estimaciones a nivel centro,
># Midzuno centros
>wigl.h.mz<-gl.h.mz*(1/pi1.mz)
>wigl.h.mz.p<-rep(wigl.h.mz,tam.mz)
>noCal.p.gl.mz<-sum(wigl.h.mz.p*xys.mz[,9]); noCal.p.gl.mz
 [1] 7558.498
>Vnocal.gl.h.mz<-varest(dxyph.mz[,19], xsh.mz, pi1.mz,
+ gl.h.mz/pi1.mz)
>Vnocal.gl.h.mz
 [1] 4065.532
```

Para calcular los estimadores óptimos, a partir de las variables auxiliares: sexo y nivel estudios del padre, a nivel persona, necesitamos los valores g (gl.SE.mas y gl.SE.mz), estimadores del total (Calp.gl.SE.mas y Calp.gl.SE.mz) y de la varianza (VCalp.gl.SE.mas y VCalp.gl.SE.mz), de igual manera que se realizó anteriormente en el ejemplo de la sección 3.4.

```
>Calp.gl.SE.mas
 [1] 7947.775
>VCalp.gl.SE.mas
 [1] 154645.1
> Calp.gl.SE.mz
 [1] 7426.54
>VCalp.gl.S.mz
 [1] 135737.7
```

Calculamos los alphas óptimos a partir de los estimadores de las varianzas a nivel centro y alumno.

```
># Alphas optimos
>alpha.l.mas<-(Vnocal.gl.h.mas)/(VCalp.gl.SE.mas+Vnocal.gl.h.mas)
>alpha.l.mas
```

```
 [1] 0.03406235
>alpha.l.mz<-(Vnocal.gl.h.mz)/(VCalp.gl.SE.mz+Vnocal.gl.h.mz)
>alpha.l.mz
 [1] 0.03015906
```

Calculamos los estimadores óptimos del total y de la varianza de los estimadores, a partir de los estimadores del total, a nivel centro y alumnos, y de los alphas óptimos calculados anteriormente para la distancia linear.

```
>opt.p.gl.mas<-(1-alpha.l.mas)*Calp.gl.SE.mas +
+ (alpha.l.mas)*noCal.p.gl.mas
>opt.p.gl.mas
 [1] 7976.551
>Vopt.p.gl.mas<-(1-alpha.l.mas)^2*Vp.gl.SE.mas +
+ (alpha.l.mas)^2*V.nocal.gl.h.mas
>Vopt.p.gl.mas
 [1] 144294.1
>opt.p.gl.mz<-(1-alpha.l.mz)*Calp.gl.SE.mz +
+ (alpha.l.mz)*noCal.p.gl.mz
>opt.p.gl.mz
 [1] 7398.143
>Vopt.p.gl.mz<-(1-alpha.l.mz)^2*Vp.gl.SE.mz +
(alpha.l.mz)^2*V.nocal.gl.h.mz
>Vopt.p.gl.mz
 [1] 122974.3
```

C.2 Funciones capítulo 2

Función calibecov (creada). Calcula la estimación de la covarianza utilizando el método de los residuos.

```
#################################################
# Función calibecov, estimación de la covarianza" #
#################################################
calibecov<-function(Ys1,Ys2,Xs1,Xs2,total1,total2,pikl1,
pikl2,d1,d2,g1,g2,q1=rep(1,length(d1)),q2=rep(1,length(d2)),
with = FALSE, EPS = 1e-06)
{
    stopifnot((ns <- length(g1)) >= 1)
    stopifnot((ns <- length(g2)) >= 1)
    piks1 = as.vector(diag(pikl1))
    piks2 = as.vector(diag(pikl2))
    if (is.data.frame(Xs1))
        Xs1 = as.matrix(Xs1)
    if (is.data.frame(Xs2))
        Xs2 = as.matrix(Xs2)
    if (!is.vector(Ys1))
        Ys1 = as.vector(Ys1)
    if (!is.vector(Ys2))
        Ys2 = as.vector(Ys2)
    if (is.matrix(Xs1))
        n1 = nrow(Xs1)
    else n1 = length(Xs1)
    if (is.matrix(Xs2))
        n2 = nrow(Xs2)
    else n2 = length(Xs2)
    #if (ns!=length(Ys)|ns!=length(Ys)|ns!=length(piks)|
     ns != n | ns !=length(d))
    #   stop("The parameters have different sizes.\n")
    w1 = g1 * d1
    wtilde1 = w1 * q1
    B1 = t(Xs1 * wtilde1)
    beta1 = ginv(B1 %*% Xs1) %*% B1 %*% Ys1
    e1 = Ys1 - Xs1 %*% beta1
    w2 = g2 * d2
    wtilde2 = w2 * q2
    B2 = t(Xs2 * wtilde2)
```

```
    beta2 = ginv(B2 %*% Xs2) %*% B2 %*% Ys2
    e2 = Ys2 - Xs2 %*% beta2
    ei1<- rep(0,m);    wi1<- rep(0,m);    di1<- rep(0,m)
    piksi1<- rep(0,m)
    l<-1
    for (i in 1:m) {
                suma1<-0; suma2<-0; suma3<-0; suma4<-0
                for (j in 1:hT[i]) {
                                    suma1=suma1+e1[l]
                                    suma2=suma2+w1[l]
                                    suma3=suma3+d1[l]
                                    suma4=suma4+piks[l]
                                    l<-l+1
                                   }
                ei1[i]<- suma1/hT[i]
                wi1[i]<- suma2/hT[i]
                di1[i]<- suma3/hT[i]
                piksi1[i]<- suma4/hT[i]
                }
    ss <- 0
    for (k in 1:ns) {
        ss2 <- 0
        for (l in 1:ns) if (!with)
            ss2 <- ss2+(1-piksi1[k]*
            piks2[l]/pikl2[k,l])*wi1[k]*ei1[k]*w2[l]*e2[l]
            else ss2<-ss2+(1-piksi1[k]*piks2[l]/pikl2[k,l])*
            di1[k]*ei1[k]*d2[l]*e2[l]
            ss<-ss+ss2
    }
    list( ecovar = as.numeric(ss))
}


################################################
# Programa "Comparativa de estimadores para EPF" #
################################################

# Selecionamos carpeta de trabajo
setwd("H:/EPF")

# Inicializo la simulación
```

```
a<-0
while (a<1){

# Número de simulaciones
nsim<-1000

# Cargo en  memoria los datos
# Datos normales
read.table("datos/XYP.txt")->XYP
# Datos dicotómicos
read.table("datos/DXYP.txt")->DXYP
# Datos integrados
read.table("datos/XYPIW.txt")->XYPIW

# Guardo en memória el número de Conglomarados
nlevels(factor(XYP[,1]))->M

# Guardo en memória el total de la variable de interés
totalp<-sum(XYP[,9])

# Inicializo salidas
arcsal<-"estimadores"
arcsal2<-"muestra"
salres<- "resumen.txt"

# Inicilizo totales
colSums(DXYP)->totalXY
c(totalXY[3:8])->totalXp
colSums(XYPIW)->totalXYPIW
c(totalXYPIW[9:11])->totalXh
c(totalXp,totalXh)->totalXiwp

# Tamaños muestras conglomerados
tam<-c(25,50,75,100,150,200,250)

# Simulación por tamaño
for (m in tam) {

# Salidas
sal <- paste(arcsal,m,".txt",sep="")
sal2<- paste(arcsal2,m,".txt",sep="")
```

```
sim<-1
while (sim<=nsim){

# Inicializo g-pesos
HTp<-0; Cal.p.gr<-0;noCal.p.gr<-0;Cal.iw.p.gr<-0
Calpgr<-0; Caliwpgr<-0

# Selecionamos la muestra por m.a.s.
sample(M,m)->s

#Seleccionamos los datos para la muestra
# Para unidades
subset(XYP, XYP[,1] %in% s)->xys
subset(DXYP, DXYP[,1] %in% s)->dxys
subset(XYPIW, XYPIW[,1] %in% s)->xysiw
# Para conglomerados
subset(DXYP, DXYP[,1] %in% s  & DXYP[,15]==1)->dxysh
subset(XYPIW, XYPIW[,1] %in% s  & XYPIW[,15]==1)->xysiwh

# Convertimos los dataframes en matrices
as.matrix(xys)->xys
as.matrix(dxys)->dxys
as.matrix(xysiw)->xysiw
as.matrix(dxysh)->dxysh
as.matrix(xysiwh)->xysiwh

# Seleccionamos las variables a nivel persona y hogar
data.frame(dxys[,3:8])->xsp
data.frame(dxysh[,9:11])->xsh
data.frame(xysiw[,3:11])->xsiwp
data.frame(xysiwh[,3:11])->xsiwph

# Número de unidades
nrow(dxys)->n
# Número de conglomerados seleccionados
nrow(dxysh)->kk #tiene que ser m

# Pesos-unidad iguales a pesos-hogar
Hpiks<-rep(m/M,kk)
piks<-rep(m/M,n)
```

```
# Variables de interés individuo-conglomerado
xys[,9]->yp
dxysh[,13]->yh

# Estimadores

# Horvitz-Thopmson individuo
HT.p<-sum(yp*(1/piks))

# Horvitz-Thopmson  hogar
HT.h<-sum(yh*(1/Hpiks)) #HT nivel persona

#Estimadores a nivel unidad con informacion a nivel unidad
# calibrado linear individuo
gl=calib1(xsp,d=1/piks,totalXp,method="linear")
Cal.p.gl<-sum((1/piks)*gl*yp)

# calibrado raking individuo
gr=calib1(xsp,d=1/piks,totalXp,method="raking")
Cal.p.gr<-sum((1/piks)*gr*yp)

# calibrado logit individuo
gt=calib1(xsp,d=1/piks,totalXp,method="logit")
Cal.p.gt<-sum((1/piks)*gt*yp)

# Número de unidades en el conglomerado
hT<-dxysh[,2]

#Calculo los pesos medios por conglomerado
(1/piks)*gl->wgl
(1/piks)*gr->wgr
(1/piks)*gt->wgt

hwgl<- rep(0,m); hwgr<- rep(0,m); hwgt<- rep(0,m)
k<-1
for (i in 1:m) {
   suma1<-0;suma2<-0;suma3<-0
        for (j in 1:hT[i]) {
            suma1=suma1+wgl[k]
            suma2=suma2+wgr[k]
```

```
            suma3=suma3+wgt[k]
            k<-k+1
            }
   hwgl[i]<- suma1/hT[i]
   hwgr[i]<- suma2/hT[i]
   hwgt[i]<- suma3/hT[i]
   }

# Estimadores (no calibrados a nivel) conglomerado
noCal.h.gl<-sum(hwgl*yh) # linear
noCal.h.gr<-sum(hwgr*yh) # raking
noCal.h.gt<-sum(hwgt*yh) # logit

#Est. cali. a nivel conglomerado
#con informacion a nivel conglomerado
glh=calib1(xsh,d=1/Hpiks,totalXh,method="linear")
Cal.h.gl<-sum(glh*(1/Hpiks)*yh)

grh=calib1(xsh,d=1/Hpiks,totalXh,method="raking")
Cal.h.gr<-sum(grh*(1/Hpiks)*yh)

gth=calib1(xsh,d=1/Hpiks,totalXh,method="logit")
Cal.h.gt<-sum(gth*(1/Hpiks)*yh)

# Calculo pesos que ponderan el estimador a nivel hogar
# Linear
wil<- glh*(1/Hpiks)
wkil <- rep(wil, hT); wkil<- as.vector(wkil)
noCal.p.gl<-sum(wkil*yp)

# raking
wir<- grh*(1/Hpiks)
wkir <- rep(wir, hT); wkir<- as.vector(wkir)
noCal.p.gr<-sum(wkir*yp)

# logit
wit<- gth*(1/Hpiks)
wkit <- rep(wit, hT); wkit<- as.vector(wkit)
noCal.p.gt<-sum(wkit*yp)
```

```
# Estimador Estevao-Sardall
xsiwp<-as.matrix(xsiwp);  dk<-as.vector(1/piks)

gliw<-rep(0,n); gliw<-as.vector(gliw);
gliw=calib1(xsiwp,d=dk,totalXiwp,method="linear")

griw<-rep(0,n); griw<-as.vector(griw);
griw=calib1(xsiwp,d=dk,totalXiwp,method="raking")

gtiw<-rep(0,n); gtiw<-as.vector(gtiw);
gtiw=calib1(xsiwp,d=dk,totalXiwp,method="logit")

# Elimino g nulos
if (!is.null(gliw) & !is.null(gl) & !is.null(glh) &
!is.null(griw) & !is.null(gr) & !is.null(grh) &
!is.null(gtiw) & !is.null(gt) & !is.null(gth))
{
# Individuos
Cal.iw.p.gl<-sum((1/piks)*gliw*yp)
Cal.iw.p.gr<-sum((1/piks)*griw*yp)
Cal.iw.p.gt<-sum((1/piks)*gtiw*yp)

# Conglomerado

Cal.iw.h.gl<-sum((1/piks)*gliw*dxys[,13]*dxys[,15])
Cal.iw.h.gr<-sum((1/piks)*griw*dxys[,13]*dxys[,15])
Cal.iw.h.gt<-sum((1/piks)*gtiw*dxys[,13]*dxys[,15])

# Sacamos estimadores en tex
write.table(data.frame(m,n,totalp,round(HT.p,2),round(Cal.p.gl,2),
round(noCal.p.gl,2),round(Cal.iw.p.gl,2),round(Cal.p.gr,2),
round(noCal.p.gr,2),round(Cal.iw.p.gr,2),round(Cal.p.gt,2),
round(noCal.p.gt,2),round(Cal.iw.p.gt,2),round(HT.h,2),
round(Cal.h.gl,2),round(noCal.h.gl,2),round(Cal.iw.h.gl,2),
round(Cal.h.gr,2),round(noCal.h.gr,2),round(Cal.iw.h.gr,2),
round(Cal.h.gt,2),round(noCal.h.gt,2),round(Cal.iw.h.gt,2)),
sal, append =TRUE,row.names=FALSE,col.names=FALSE, quote = FALSE)
write(s,sal2,append=TRUE, ncol=m)
sim <- sim + 1
}
```

```
}#del while
sim <- sim
}#del for en m

for (m in tam)
{
ent <- paste(arcsal,m,".txt",sep="")
read.table(ent)->resu

# Estimador óptimo simulaciones individuos
# Linear
var(resu[,5])*(length(resu[,5])-1)/(length(resu[,5]))->vp1
var(resu[,6])*(length(resu[,6])-1)/(length(resu[,6]))->vp2
cov(resu[,5],resu[,6])*(length(resu[,5])-1)/(length(resu[,5]))->covp12
alpha.p.l<-(vp2-covp12)/(vp1+vp2-2*covp12)
opt.p.gl<- rep(0,nsim)
for (j in 1:nsim) {
opt.p.gl[j]<-alpha.p.l*resu[j,5]+(1-alpha.p.l)*resu[j,6]
}
# Raking
var(resu[,8])*(length(resu[,8])-1)/(length(resu[,8]))->vp3
var(resu[,9])*(length(resu[,9])-1)/(length(resu[,9]))->vp4
cov(resu[,8],resu[,9])*(length(resu[,8])-1)/(length(resu[,8]))->covp34
alpha.p.r<-(vp4 - covp34)/(vp3 + vp4 - 2*covp34)
opt.p.gr<- rep(0,nsim)
for (j in 1:nsim) {
opt.p.gr[j]<-alpha.p.r*resu[j,8]+(1-alpha.p.r)*resu[j,9]
}
# Logit
var(resu[,11])*(length(resu[,11])-1)/(length(resu[,11]))->vp5
var(resu[,12])*(length(resu[,12])-1)/(length(resu[,12]))->vp6
cov(resu[,11],resu[,12])*(length(resu[,11])-1)/(length(resu[,11]))->covp56
alpha.p.t<-(vp6 - covp56)/(vp5 + vp6 - 2*covp56)
opt.p.gt<- rep(0,nsim)
for (j in 1:nsim) {
opt.p.gt[j]<-alpha.p.t*resu[j,11]+(1-alpha.p.t)*resu[j,12]
}
# Estimador óptimo simulaciones Conglomerado
# Linear
var(resu[,15])*(length(resu[,15])-1)/(length(resu[,15]))->vh1
var(resu[,16])*(length(resu[,16])-1)/(length(resu[,16]))->vh2
```

```
cov(resu[,15],resu[,16])*(length(resu[,15])-1)/(length(resu[,15]))->covh12
alpha.h.l<-(vh2 - covh12)/(vh1 + vh2 - 2*covh12)
opt.h.gl<- rep(0,nsim)
for (j in 1:nsim) opt.h.gl[j]<-alpha.h.l*resu[j,15]+(1-alpha.h.l)*resu[j,16]

# Raking
var(resu[,18])*(length(resu[,18])-1)/(length(resu[,18]))->vh3
var(resu[,19])*(length(resu[,19])-1)/(length(resu[,19]))->vh4
cov(resu[,18],resu[,19])*(length(resu[,18])-1)/(length(resu[,18]))->covh34
alpha.h.r<-(vh4 - covh34)/(vh3 + vh4 - 2*covh34)
opt.h.gr<- rep(0,nsim)
for (j in 1:nsim) opt.h.gr[j <-alpha.h.r*resu[j,18]+(1-alpha.h.r)*resu[j,19]

# Logit
var(resu[,21])*(length(resu[,21])-1)/(length(resu[,21]))->vh5
var(resu[,22])*(length(resu[,22])-1)/(length(resu[,22]))->vh6
cov(resu[,21],resu[,22])*(length(resu[,21])-1)/(length(resu[,21]))->covh56
alpha.h.t<-(vh6 - covh56)/(vh5 + vh6 - 2*covh56)
opt.h.gt<- rep(0,nsim)
for (j in 1:nsim) opt.h.gt[j]<-alpha.h.t*resu[j,21]+(1-alpha.h.t)*resu[j,22]

file.remove(ent)
write.table(data.frame(resu,round(opt.p.gl,2),round(opt.p.gr,2),
round(opt.p.gt,2),round(opt.h.gl,2),round(opt.h.gr,2),
round(opt.h.gt,2),round(alpha.p.l,4),round(alpha.p.r,4),
round(alpha.p.t,4),round(alpha.h.l,4),round(alpha.h.r,4),
round(alpha.h.t,4)),ent,append =TRUE,row.names=FALSE,
col.names=FALSE, quote = FALSE)
}

for (m in tam){
   ent <- paste(arcsal,m,".txt",sep="")
 read.table(ent)->resu
 ncol(resu)-9->nesti
 rep(0,nesti)->rb
 rep(0,nesti)->mse
 for (j in 1:nesti) rb[j]<- sum(resu[,j+3]-resu[,3])/resu[1,3]
 for (j in 1:nesti) mse[j]<- sum((resu[,j+3]-resu[,3])^2)
 re <- mse/mse[1]
 mean(resu[,2])->np
 write.table(data.frame(m, np, round(rb,4), round(mse,0),
```

```
round(re,4)), salres,append=TRUE,row.names=FALSE,col.names=FALSE,
quote = FALSE)
}

read.table("resumen.txt")->bb
if(bb[4,5]<bb[21,5] && bb[7,5]<bb[22,5] && bb[10,5]<bb[23,5]) a<-0
if(bb[4,5]>=bb[21,5] && bb[7,5]>=bb[22,5] && bb[10,5]>=bb[23,5]) a<-1

}
```

C.3 Funciones capítulo 3

```
#####################################################
# mk "estimador de calibración para no respuesta"  #
#####################################################
calmk<-function (Xr, Xs, dr, ds, q = rep(1, length(dr)))
{
    EPS = .Machine$double.eps
    am=(as.vector(t(ds) %*% Xs))
    bm=ginv(t(Xr * dr * q) %*% Xr, tol = EPS)
    m = q * as.vector(Xr %*% (bm%*%am))
    return(m)
}

################################################
# "Selección de valores sin datos faltantes"  #
################################################
subset(XYP ,!is.na(XYP[,1])==TRUE &  !is.na(XYP[,2])==TRUE &
 !is.na(XYP[,3])==TRUE )->k
write.table(k, file = "XYP.SinNA.txt", sep = " ",
 row.names = FALSE, col.names = FALSE)

############################################
# "Creación de una variable dicotómica     #
# con un porcentaje de falta de respuesta" #
############################################
# Por ejemplo un 40% de f.r. con un tamaño de 20000 unidades
na.p<-0.40
a.p<-runif(1)
b.p<-1-na.p-a.p
sample(c(0,1,NA), 20000, replace=T, prob=c(a.p,b.p,na.p))->fr40

#################################################
# Programa "Cálculo de indicadores para una v.a."#
#################################################
calculos.1 <- function (xys, xys.sinNA, yp.sinNA, piks, pikr)
{
y.k=yp.sinNA
y.media.r.d=sum((1/pikr)*y.k)/sum((1/pikr))
P=sum((1/pikr))/sum((1/piks))
S.2.y=sum((1/pikr)*(y.k-y.media.r.d)^2)/sum((1/pikr))
S.y=sqrt(S.2.y)
```

```
m.k.r= as.vector(t(sum((1/piks)*xys[,i])) /
 (sum((1/pikr)*xys.sinNA[,i]%*% t(xys.sinNA[,i]))))*xys.sinNA[,i]
m.k.s= as.vector(t(sum((1/piks)*xys[,i])) /
 (sum((1/pikr)*xys.sinNA[,i]%*% t(xys.sinNA[,i]))))*xys[,i]
m.media.r.d=sum((1/pikr)*m.k.r)/sum(1/pikr)
m.media.s.d=sum((1/piks)*m.k.s)/sum(1/piks)
S.2.m=sum((1/pikr)*(m.k.r-m.media.r.d)^2)/sum((1/pikr))
S.m=sqrt(S.2.m)
cv.m= S.m / m.media.r.d
cov.y.m=sum((1/pikr)*(m.k.r-m.media.r.d)*(y.k-y.media.r.d))/
sum((1/pikr))
R.y.m = (cov.y.m)/(S.y*S.m)
y.k=yp.sinNA
x.media.r.d=sum((1/pikr)*xys.sinNA[,i])/sum((1/pikr))
x.media.s.d=sum((1/piks)*xys[,i])/sum((1/piks))
delta.A=t(x.media.r.d-x.media.s.d)*(sum((1/pikr)*xys.sinNA[,i]*y.k ))/
 (sum((1/pikr)*xys.sinNA[,i] %*%t(xys.sinNA[,i])))
H0= delta.A/S.y
H1=abs(delta.A)/S.y
H2= R.y.m*cv.m
H3=cv.m
v=c(i,H0,H1,H2,H3)
write.table(t(v), file = "v.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
}

################################################
# Programa "Cálculo de indicadores para dos v.a."#
################################################
calculos.2 <- function (xys, xys.sinNA, yp.sinNA, piks, pikr)
{
y.k=yp.sinNA
y.media.r.d=sum((1/pikr)*y.k)/sum((1/pikr))
P=sum((1/pikr))/sum((1/piks))
S.2.y=sum((1/pikr)*(y.k-y.media.r.d)^2)/sum((1/pikr))
S.y=sqrt(S.2.y)
as.matrix(data.frame(xys[,i],xys[,j])) -> xys.ij
as.matrix(data.frame(xys.sinNA[,i],
 xys.sinNA[,j]))->xys.sinNA.ij
m.k.r= as.vector(t(sum((1/piks)*xys.ij )) /
 (sum((1/pikr)* xys.sinNA.ij  %*%  t(xys.sinNA.ij))))*
```

```
xys.sinNA.ij
m.k.s= as.vector(t(sum((1/piks)*xys.ij)) /
 (sum((1/pikr)*xys.sinNA.ij  %*%  t(xys.sinNA.ij))))*
xys.ij
m.media.r.d=sum((1/pikr)*m.k.r)/sum(1/pikr)
m.media.s.d=sum((1/piks)*m.k.s)/sum(1/piks)
S.2.m=sum((1/pikr)*(m.k.r-m.media.r.d)^2)/sum((1/pikr))
S.m=sqrt(S.2.m)
cv.m= S.m / m.media.r.d
cov.y.m=sum((1/pikr)*(m.k.r-m.media.r.d)*
 (y.k-y.media.r.d))/sum((1/pikr))
R.y.m = (cov.y.m)/(S.y*S.m)
y.k=yp.sinNA
x.media.r.d=sum((1/pikr)*xys.sinNA.ij)/sum((1/pikr))
x.media.s.d=sum((1/piks)*xys.ij)/sum((1/piks))
delta.A=t(x.media.r.d-x.media.s.d)*(sum((1/pikr)*
 xys.sinNA.ij*y.k ))/
 (sum((1/pikr)* xys.sinNA.ij %*%t(xys.sinNA.ij)))
H0= delta.A/S.y
H1=abs(delta.A)/S.y
H2= R.y.m*cv.m
H3=cv.m
v=c(i,j,H0,H1,H2,H3)
write.table(t(v), file = "v.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
}
```

Para el resto de variables se calculan de igual manera. Por ejemplo para 7 variables:

```
#####################################################
# Programa "Cálculo de indicadores para siete v.a."#
#####################################################
calculos.7 <- function (xys, xys.sinNA, yp.sinNA, piks, pikr)
{
y.k=yp.sinNA
y.media.r.d=sum((1/pikr)*y.k)/sum((1/pikr))
P=sum((1/pikr))/sum((1/piks))
S.2.y=sum((1/pikr)*(y.k-y.media.r.d)^2)/sum((1/pikr))
S.y=sqrt(S.2.y)
as.matrix(data.frame(xys[,i],xys[,j],xys[,k],xys[,l],xys[,m],
 xys[,n],xys[,o])) -> xys.ijklmno
```

```
as.matrix(data.frame(xys.sinNA[,i],xys.sinNA[,j],xys.sinNA[,k],
xys.sinNA[,l], xys.sinNA[,m], xys.sinNA[,n], xys.sinNA[,o]))->
 xys.sinNA.ijklmno
m.k.r= as.vector(t(sum((1/piks)*xys.ijklmno ))/
 (sum((1/pikr)* xys.sinNA.ijklmno  %*%  t(xys.sinNA.ijklmno))))*
xys.sinNA.ijklmno
m.k.s= as.vector(t(sum((1/piks)*xys.ijklmno))/(sum((1/pikr)*
xys.sinNA.ijklmno
 %*%  t(xys.sinNA.ijklmno)))) * xys.ijklmno
m.media.r.d=sum((1/pikr)*m.k.r)/sum(1/pikr)
m.media.s.d=sum((1/piks)*m.k.s)/sum(1/piks)
S.2.m=sum((1/pikr)*(m.k.r-m.media.r.d)^2)/sum((1/pikr))
S.m=sqrt(S.2.m)
cv.m= S.m / m.media.r.d
cov.y.m=sum((1/pikr)*(m.k.r-m.media.r.d)*(y.k-y.media.r.d))/
 sum((1/pikr))
R.y.m = (cov.y.m)/(S.y*S.m)
y.k=yp.sinNA
x.media.r.d=sum((1/pikr)*xys.sinNA.ijklmno)/sum((1/pikr))
x.media.s.d=sum((1/piks)*xys.ijklmno)/sum((1/piks))
delta.A=t(x.media.r.d-x.media.s.d)*(sum((1/pikr) *
 xys.sinNA.ijklmno *y.k ))/(sum((1/pikr)* xys.sinNA.ijklmno
 %*%t(xys.sinNA.ijklmno)))
H0= delta.A/S.y
H1=abs(delta.A)/S.y
H2= R.y.m*cv.m
H3=cv.m
v=c(i,j,k,l,m,n,o, H0,H1,H2,H3)
write.table(t(v), file = "v.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
}

#######################################################
# Programa "Ordenación de variables según indicador" #
#######################################################
ordena.1 <- function(ms){
#sort ordena de mayor a menor
ms[sort.list(ms[,2]),]->ms2
ms[sort.list(ms[,3]),]->ms3
ms[sort.list(ms[,4]),]->ms4
ms[sort.list(ms[,5]),]->ms5
```

```
ord<-data.frame(t(ms2[,1]),t(ms2[,2]),t(ms3[,1]),t(ms3[,3]),
 t(ms4[,1]),t(ms4[,4]), t(ms5[,1]),t(ms5[,5]))
write.table(ord, file = "ord.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
}

##########################################################
# Programa "Ordenación de dos variables según indicador" #
##########################################################
ordena.2 <- function(ms){
#sort ordena de mayor a menor
ms[sort.list(ms[,3]),]->ms2
ms[sort.list(ms[,4]),]->ms3
ms[sort.list(ms[,5]),]->ms4
ms[sort.list(ms[,6]),]->ms5
ordms2<-data.frame(t(ms2[,2]),t(ms2[,3]))
write.table(ordms2, file = "ordms2.txt", sep = " ",
 append =TRUE, row.names = FALSE, col.names = FALSE)
}
```

⋮

```
##########################################################
# Programa "Ordenación de 7 variables según indicador" #
##########################################################
ordena.7 <- function(ms){
#sort ordena de mayor a menor
ms[sort.list(ms[,8]),]->ms2
ms[sort.list(ms[,9]),]->ms3
ms[sort.list(ms[,10]),]->ms4
ms[sort.list(ms[,11]),]->ms5
ordms7<-data.frame(t(ms2[,7]),t(ms2[,8]))
write.table(ordms7, file = "ordms7.txt", sep = " ",
append =TRUE, row.names = FALSE, col.names = FALSE)
}
```

Nota: *Simulación para calcular las estimaciones por calibración según ordenación para cada indicador H_0, H_1, H_2 y H_3. Datosh0, ..., Datosh3 hacen referencia a los ficheros donde se han descrito la ordenación de las variables auxiliares para cada indicador.*

```
############################################################
# Simulación Estimaciones por calibración según ordenación #
############################################################
library(MASS)
library(lpSolve)
library(sampling)

#Leemos los el orden de los datos según ordenación
read.table("Datosh0.txt")->datosh0
read.table("Datosh1.txt")->datosh1
read.table("Datosh2.txt")->datosh2
read.table("Datosh3.txt")->datosh3

# Leo datos sin fata de respuesta para las
# variables auxiliares
read.table("XYP.txt")->XYP

# Número de centros
nlevels(factor(XYP[,1]))->M

# Primer indicador H0
bb<-1
zz=nrow(datosh0)
while(bb<zz+1)
{
aa<-datosh0[bb,1]
m<-datosh0[bb,2]
#Muestra
sample(M,m)->s2
subset(XYP, XYP[,1] %in% s2)->xys; as.matrix(xys)->xys
n=nrow(XYP)
centro=XYP[,1]
Pob=429385
# m= 17445; M= 429385 (Alumnos matriculados en españa en 2006)
# p=0.04063
# probabilidades de inclusión
piks<-rep(n/Pob,nrow(xys))
yp<-xys[,16]
# Estimador de H-T
HT.p<-sum(yp*(1/piks))
# Leo orden de variables
```

```
j<-datosh0[bb,3]
k<-datosh0[bb,4]
l<-datosh0[bb,5]
o<-datosh0[bb,6]
t<-datosh0[bb,7]
h<-datosh0[bb,8]
w<-datosh0[bb,9]

# Selecciono las muestras según v.a.
as.matrix(data.frame(xys[,j]))->xys.j
as.matrix(data.frame(xys[,j],xys[,k]))->xys.jk
as.matrix(data.frame(xys[,j],xys[,k],xys[,l]))->xys.jkl
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o]))->xys.jklo
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],xys[,t]))->
 xys.jklot
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],xys[,t],
 xys[,h]))->xys.jkloth
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],xys[,t],
 xys[,h],xys[,w]))->xys.jklothw
# Totales
total<-c(637923, sum(xys[,5]*(1/piks)), sum(xys[,6]*(1/piks)),
 sum(xys[,7]*(1/piks)), sum(xys[,8]*(1/piks)), sum(xys[,9]*(1/piks)),
 sum(xys[,10]*(1/piks)))

# Selecciono totales por muestra
total.j<-c(total[j-3])
total.jk<-c(total[j-3],total[k-3])
total.jkl<-c(total[j-3],total[k-3],total[l-3])
total.jklo<-c(total[j-3],total[k-3],total[l-3],total[o-3])
total.jklot<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3])
total.jkloth<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3], total[h-3])
total.jklothw<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3], total[h-3], total[w-3])

#Estimadores de calibración por variables auxilares
g1=calib(xys.j,d=1/piks,total.j,method="linear")
Ycalm1=sum((1/piks)*g1*yp)
g2=calib(xys.jk,d=1/piks,total.jk,method="linear")
Ycalm2=sum((1/piks)*g2*yp)
```

```
g3=calib(xys.jkl,d=1/piks,total.jkl,method="linear")
Ycalm3=sum((1/piks)*g3*yp)
g4=calib(xys.jklo,d=1/piks,total.jklo,method="linear")
Ycalm4=sum((1/piks)*g4*yp)
g5=calib(xys.jklot,d=1/piks,total.jklot,method="linear")
Ycalm5=sum((1/piks)*g5*yp)
g6=calib(xys.jkloth,d=1/piks,total.jkloth,method="linear")
Ycalm6=sum((1/piks)*g6*yp)
g7=calib(xys.jklothw,d=1/piks,total.jklothw,method="linear")
Ycalm7=sum((1/piks)*g7*yp)
Ycalmh0<-c(round(HT.p,2), round(Ycalm1,2), round(Ycalm2,2),
  round(Ycalm3,2), round(Ycalm4,2), round(Ycalm5,2), round(Ycalm6,2),
  round(Ycalm7,2))
write.table(t(Ycalmh0), file = "RDFh0.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
bb=bb+1
}

# Segundo indicador H1
bb<-1
zz=nrow(datosh1)
while(bb<zz+1){
aa<-datosh1[bb,1]
m<-datosh1[bb,2]
sample(M,m)->s2
subset(XYP, XYP[,1] %in% s2)->xys; as.matrix(xys)->xys
n=nrow(XYP)
centro=XYP[,1]
Pob=429385
piks<-rep(n/Pob,nrow(xys))
yp<-xys[,16]
HT.p<-sum(yp*(1/piks))
j<-datosh1[bb,3]
k<-datosh1[bb,4]
l<-datosh1[bb,5]
o<-datosh1[bb,6]
t<-datosh1[bb,7]
h<-datosh1[bb,8]
w<-datosh1[bb,9]
as.matrix(data.frame(xys[,j])) -> xys.j
as.matrix(data.frame(xys[,j],xys[,k]))->xys.jk
```

```
as.matrix(data.frame(xys[,j],xys[,k],xys[,l]))->xys.jkl
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o]))->xys.jklo
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],
 xys[,t]))->xys.jklot
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],
 xys[,t],xys[,h]))->xys.jkloth
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],
 xys[,t],xys[,h],xys[,w]))->xys.jklothw
total<-c(637923, sum(xys[,5]*(1/piks)), sum(xys[,6]*(1/piks)),
sum(xys[,7]*(1/piks)), sum(xys[,8]*(1/piks)), sum(xys[,9]*(1/piks)),
 sum(xys[,10]*(1/piks)))
total.j<-c(total[j-3])
total.jk<-c(total[j-3], total[k-3])
total.jkl<-c(total[j-3], total[k-3], total[l-3])
total.jklo<-c(total[j-3], total[k-3], total[l-3], total[o-3])
total.jklot<-c(total[j-3], total[k-3], total[l-3], total[o-3],
 total[t-3])
total.jkloth<-c(total[j-3], total[k-3], total[l-3], total[o-3],
 total[t-3], total[h-3])
total.jklothw<-c(total[j-3], total[k-3], total[l-3], total[o-3],
 total[t-3], total[h-3], total[w-3])
g1=calib(xys.j,d=1/piks,total.j,method="linear")
Ycalm1=sum((1/piks)*g1*yp)
g2=calib(xys.jk,d=1/piks,total.jk,method="linear")
Ycalm2=sum((1/piks)*g2*yp)
g3=calib(xys.jkl,d=1/piks,total.jkl,method="linear")
Ycalm3=sum((1/piks)*g3*yp)
g4=calib(xys.jklo,d=1/piks,total.jklo,method="linear")
Ycalm4=sum((1/piks)*g4*yp)
g5=calib(xys.jklot,d=1/piks,total.jklot,method="linear")
Ycalm5=sum((1/piks)*g5*yp)
g6=calib(xys.jkloth,d=1/piks,total.jkloth,method="linear")
Ycalm6=sum((1/piks)*g6*yp)
g7=calib(xys.jklothw,d=1/piks,total.jklothw,method="linear")
Ycalm7=sum((1/piks)*g7*yp)
Ycalmh1<-c(round(HT.p,2), round(Ycalm1,2), round(Ycalm2,2),
 round(Ycalm3,2), round(Ycalm4,2), round(Ycalm5,2), round(Ycalm6,2),
 round(Ycalm7,2))
#Ycalmh1=Ycalmh1/1000
write.table(t(Ycalmh1), file = "RDFh1.txt", sep = " ", append =TRUE,
 row.names = FALSE, col.names = FALSE)
```

```
bb=bb+1
}

# Tercer indicador H2
bb<-1
zz=nrow(datosh2)
while(bb<zz+1){
aa<-datosh2[bb,1]
m<-datosh2[bb,2]
sample(M,m)->s2
subset(XYP, XYP[,1] %in% s2)->xys; as.matrix(xys)->xys
n=nrow(XYP)
centro=XYP[,1]
Pob=429385
# Los pesos del diseño: m= 17445; M= 429385 (Alumnos matriculados
# en españa en 2006)   p=0.04063
piks<-rep(n/Pob,nrow(xys))
yp<-xys[,16]
HT.p<-sum(yp*(1/piks))
j<-datosh2[bb,3]
k<-datosh2[bb,4]
l<-datosh2[bb,5]
o<-datosh2[bb,6]
t<-datosh2[bb,7]
h<-datosh2[bb,8]
w<-datosh2[bb,9]
as.matrix(data.frame(xys[,j])) -> xys.j
as.matrix(data.frame(xys[,j],xys[,k])) -> xys.jk
as.matrix(data.frame(xys[,j],xys[,k],xys[,l])) -> xys.jkl
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o])) -> xys.jklo
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],
 xys[,t])) -> xys.jklot
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],xys[,t],
 xys[,h])) -> xys.jkloth
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],xys[,o],xys[,t],
 xys[,h],xys[,w])) -> xys.jklothw
total<-c(637923, sum(xys[,5]*(1/piks)), sum(xys[,6]*(1/piks)),
 sum(xys[,7]*(1/piks)), sum(xys[,8]*(1/piks)), sum(xys[,9]*(1/piks)),
  sum(xys[,10]*(1/piks)))
total.j<-c(total[j-3])
total.jk<-c(total[j-3],total[k-3])
```

```
total.jkl<-c(total[j-3],total[k-3],total[l-3])
total.jklo<-c(total[j-3],total[k-3],total[l-3],total[o-3])
total.jklot<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3])
total.jkloth<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3], total[h-3])
total.jklothw<-c(total[j-3],total[k-3],total[l-3],total[o-3],
 total[t-3], total[h-3], total[w-3])
g1=calib(xys.j,d=1/piks,total.j,method="linear")
Ycalm1=sum((1/piks)*g1*yp)
g2=calib(xys.jk,d=1/piks,total.jk,method="linear")
Ycalm2=sum((1/piks)*g2*yp)
g3=calib(xys.jkl,d=1/piks,total.jkl,method="linear")
Ycalm3=sum((1/piks)*g3*yp)
g4=calib(xys.jklo,d=1/piks,total.jklo,method="linear")
Ycalm4=sum((1/piks)*g4*yp)
g5=calib(xys.jklot,d=1/piks,total.jklot,method="linear")
Ycalm5=sum((1/piks)*g5*yp)
g6=calib(xys.jkloth,d=1/piks,total.jkloth,method="linear")
Ycalm6=sum((1/piks)*g6*yp)
g7=calib(xys.jklothw,d=1/piks,total.jklothw,method="linear")
Ycalm7=sum((1/piks)*g7*yp)
Ycalmh2<-c(round(HT.p,2), round(Ycalm1,2), round(Ycalm2,2),
 round(Ycalm3,2), round(Ycalm4,2), round(Ycalm5,2),
 round(Ycalm6,2), round(Ycalm7,2))
write.table(t(Ycalmh2), file = "RDFh2.txt", sep = " ",
 append =TRUE, row.names = FALSE, col.names = FALSE)
bb=bb+1
}

# Cuarto indicador H3
bb<-1
zz=nrow(datosh3)
while(bb<zz+1){
aa<-datosh3[bb,1]
m<-datosh3[bb,2]
sample(M,m)->s2
subset(XYP, XYP[,1] %in% s2)->xys; as.matrix(xys)->xys
n=nrow(XYP)
centro=XYP[,1]
Pob=429385
```

```
# Los pesos del diseño: m= 17445; M= 429385
# (Alumnos matriculados en españa en 2006)   p=0.04063
piks<-rep(n/Pob,nrow(xys))
yp<-xys[,16]
HT.p<-sum(yp*(1/piks))
j<-datosh3[bb,3]
k<-datosh3[bb,4]
l<-datosh3[bb,5]
o<-datosh3[bb,6]
t<-datosh3[bb,7]
h<-datosh3[bb,8]
w<-datosh3[bb,9]
as.matrix(data.frame(xys[,j])) -> xys.j
as.matrix(data.frame(xys[,j],xys[,k])) -> xys.jk
as.matrix(data.frame(xys[,j],xys[,k],xys[,l])) -> xys.jkl
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],
 xys[,o])) -> xys.jklo
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],
 xys[,o],xys[,t])) -> xys.jklot
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],
 xys[,o],xys[,t],xys[,h])) -> xys.jkloth
as.matrix(data.frame(xys[,j],xys[,k],xys[,l],
 xys[,o],xys[,t],xys[,h],xys[,w])) -> xys.jklothw
total<-c(637923, sum(xys[,5]*(1/piks)), sum(xys[,6]*(1/piks)),
 sum(xys[,7]*(1/piks)), sum(xys[,8]*(1/piks)),
 sum(xys[,9]*(1/piks)), sum(xys[,10]*(1/piks)))
total.j<-c(total[j-3])
total.jk<-c(total[j-3], total[k-3])
total.jkl<-c(total[j-3], total[k-3], total[l-3])
total.jklo<-c(total[j-3], total[k-3], total[l-3],
 total[o-3])
total.jklot<-c(total[j-3], total[k-3], total[l-3],
 total[o-3], total[t-3])
total.jkloth<-c(total[j-3], total[k-3], total[l-3], total[o-3],
 total[t-3], total[h-3])
total.jklothw<-c(total[j-3], total[k-3], total[l-3], total[o-3],
 total[t-3], total[h-3], total[w-3])
g1=calib(xys.j,d=1/piks,total.j,method="linear")
Ycalm1=sum((1/piks)*g1*yp)
g2=calib(xys.jk,d=1/piks,total.jk,method="linear")
Ycalm2=sum((1/piks)*g2*yp)
```

```
g3=calib(xys.jkl,d=1/piks,total.jkl,method="linear")
Ycalm3=sum((1/piks)*g3*yp)
g4=calib(xys.jklo,d=1/piks,total.jklo,method="linear")
Ycalm4=sum((1/piks)*g4*yp)
g5=calib(xys.jklot,d=1/piks,total.jklot,method="linear")
Ycalm5=sum((1/piks)*g5*yp)
g6=calib(xys.jkloth,d=1/piks,total.jkloth,method="linear")
Ycalm6=sum((1/piks)*g6*yp)
g7=calib(xys.jklothw,d=1/piks,total.jklothw,method="linear")
Ycalm7=sum((1/piks)*g7*yp)
Ycalmh3<-c(round(HT.p,2), round(Ycalm1,2), round(Ycalm2,2),
round(Ycalm3,2), round(Ycalm4,2), round(Ycalm5,2),
 round(Ycalm6,2), round(Ycalm7,2))
#Ycalmh3=Ycalmh3/1000
write.table(t(Ycalmh3), file = "RDFh3.txt", sep = " ",
 append =TRUE, row.names = FALSE, col.names = FALSE)
bb=bb+1
}
```

Índice de figuras

ÍNDICE DE FIGURAS

Índice de tablas

ÍNDICE DE TABLAS