Metodología de Encuestas
Vol 4, Núm 2, 2002, 161-173
ISSN: 1575-7803

ESTIMADORES DE CALIBRACIÓN: UNA NUEVA METODOLOGÍA PARA EL USO DE LA INFORMACIÓN AUXILIAR

María del Mar Rueda
Universidad de Granada

Sergio Martínez
Universidad de Almería

RESUMEN

El trabajo de Deville y Särdnal, de 1992, ha abierto una metodología nueva para el uso de la información auxiliar proporcionada por una o varias variables relacionadas con la variable objeto de estudio. En este trabajo se explican los fundamentos de este método, llamado de calibración, así como las distintas generalizaciones y aplicaciones de éste.

Palabras clave: calibración, estimador de regresión generalizado, estimador de Horvitz-Thompson, modelos de superpoblación.

Introducción

En la teoría de muestreo se ha estudiado en profundidad la estimación de parámetros lineales y cuadráticos, como son la media, el total y la varianza poblacionales. Se han desarrollado diseños muestrales cada vez más complejos (muestreos polietápicos, muestreos adaptativos, estratificación a posteriori…) con el fin de obtener estimadores más precisos de dichos parámetros. La mayoría de estos métodos sólo observan el comportamiento de la variable objeto de estudio para realizar la estimación. A menudo sucede que la característica que queremos estudiar, Y, se encuentra fuertemente relacionada con una característica auxiliar, X, y además los datos sobre X son conocidos o se pueden recoger fácilmente para todas las unidades de la población. En estas situaciones cabe plantearse cómo utilizar esta información auxiliar para mejorar las estimaciones, en el sentido de construir nuevos estimadores que, para el mismo tamaño muestral, tengan menor error de estimación (lo que implicaría mayor precisión en las estimaciones de los parámetros), o equivalentemente que tengan el mismo error que los ya conocidos pero con un menor tamaño muestral (lo que produciría una disminución en el coste de la realización de la encuesta).

Los primeros métodos utilizados para incorporar la información auxiliar en la fase de estimación, surgen en el contexto del modelo de población fija y son los llamados métodos indirectos de estimación, entre los que destacan los conocidos métodos de razón, de diferencia y de regresión. Estos estimadores se basan en modificar la forma del estimador directo (el estimador basado sólo en los datos que la muestra proporciona sobre la variable objeto de estudio) para construir un nuevo estimador que incorpora tanto la información que la muestra proporciona sobre la variable auxiliar como la información que se tiene de esta variable en toda la población.

Estos estimadores no siempre garantizan que se produzca una disminución del error de muestreo respecto a los estimadores que no usan información auxiliar. La precisión que ganamos con los métodos que emplean la información auxiliar estará en función del buen uso de las hipótesis que se supongan para emplear un procedimiento u otro, y el que dichas hipótesis se ajusten en mayor o menor medida al problema real.

Con la aparición de los modelos de superpoblación (ver p.e. Pérez 2002 y Sánchez-Crespo, 2002), la teoría de muestreo tuvo un gran empuje pues se le dotó de un instrumento muy valioso que permitió obtener resultados más concluyentes en la comparación de estrategias y, eventualmente, producir estrategias óptimas en varias situaciones.

En los modelos de superpoblación, cobra especial importancia el uso de variables auxiliares cuyos valores son conocidos para todos los individuos de la población. Concretamente, los modelos de superpoblación más usados, con gran diferencia, son los modelos de regresión dados por:

$$y_k = \boldsymbol{\mu}(X_k) + \boldsymbol{\varepsilon}_k$$

con

$$E_\xi\left[\boldsymbol{\varepsilon}^2\right] = 0, \qquad E_\xi\left[\boldsymbol{\varepsilon}_k^2\right] = \boldsymbol{\sigma}^2 V(X_i) \quad y \quad E_\xi\left[\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_k\right] = 0 \quad \forall k \neq j \tag{1}$$

donde ε_k es un error aleatorio.

Bajo la perspectiva modelo-asistida (o inferencia asistida por el modelo) (ver, por ejemplo, Hedayat y Sinha, 1991 o Cassel, Särdnal y Wretman (1976) se proponen los estimadores de regresión generalizada (que pueden verse como una generalización del estimador de regresión, donde el estimador del coeficiente de regresión usual se substituye por un estimador ponderado por ciertos pesos que determina el modelo) que son óptimos si la población proviene de una superpoblación que sigue un modelo de regresión lineal.

Recientemente, han aparecido algunos métodos importantes para utilizar la información auxiliar, y que se pueden clasificar en dos tipos: estimadores de calibración (Deville y Särndal, 1992) y estimadores de verosimilitud empírica (Chen y Qin, 1993 y Chen y Setter, 1999). Todos estos métodos se han discutido sólo desde la perspectiva de los modelos de superpoblación, y concretamente usuando un modelo de regresión. Estos estimadores tienen muy buenas propiedades teóricas, sin embargo muy pocas son las encuestas por muestreo que hacen uso de estas metodologías.

En este trabajo se pretende presentar de una forma sencilla en qué consiste el método de calibración: cuándo puede aplicarse, cómo se obtienen los estimadores de los parámetros más importantes, así como la relación que tiene con otros estimadores más conocidos, con objeto de generalizar su conocimiento a un ambiente profesional más amplio que el de los teóricos del muestreo en poblaciones finitas.

Construcción de un estimador de calibración

De forma muy breve, el marco de trabajo usual en muestreo de poblaciones finitas es el descrito a continuación.

Una población consiste en N elementos distintos, identificados a través de sus etiquetas $i = 1,\ldots,N$. La característica de interés y_i asociada con el elemento i se conoce exactamente (sin error) observando el elemento i. Una muestra es un subconjunto, s, de U y sus valores asociados de Y, es decir $\{(k,y_k)\}$ seleccionados de acuerdo con un diseño de muestreo específico que asigna una probabilidad conocida $p(s)>0$ para todo $s \in S$, conjunto de las posibles muestras s, y $\sum_{s\in S} p(s)=1$.

Supongamos que queremos estimar el total poblacional de la variable Y, es decir, queremos estimar

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

y que $X = (X_1,\ldots,X_J)$, es un vector de variables auxiliares de forma que es perfectamente conocido para todos los elementos de la población U, entonces si consideramos el estimador de *Horvitz-Thompson* para T_Y, es decir

$$\hat{T}_{Y\pi} = \sum_{k\in s} d_k y_k$$

lo que pretendemos es modificar los pesos $d_k = \dfrac{1}{\pi_k}$, por otros pesos ω_k, de forma que el estimador basado en dichos pesos proporcione estimaciones perfectas para X, es decir

$$\sum_{s} \omega_k X_k = T_X = \left(T_{X_1}, ..., T_{X_J} \right) \tag{2}$$

y estén tan próximos como sea posible, respecto a una distancia dada, a los pesos originales d_k.

Usualmente la distancia elegida es la suma ponderada de cuadrados de las distancias, que viene dada por

$$\sum_{k \in s} \frac{(\omega_k - d_k)^2}{q_k d_k} \tag{3}$$

donde q_k son constantes positivas. Entonces, tenemos el siguiente problema:

$$\text{minimizar } \sum_{k \in s} \frac{(\omega_k - d_k)^2}{q_k d_k} \text{ sujeto a la condición } \sum_{s} \omega_k X_k = T_X$$

Utilizando el método de los multiplicadores de Lagrange, se obtienen los siguientes pesos calibrados:

$$\omega_k = d_k + d_k q_k \lambda X_k^{'} \tag{4}$$

supuesto que la inversa de

$$T_s = \sum_{k \in s} d_k q_k X_k X_k^{'}$$

existe.

El estimador calibrado obtenido viene así dado por

$$\hat{T}_{Yreg} = \sum_{k \in s} \omega_k y_k = \hat{T}_{Y\pi} + \left(\hat{T}_X - \hat{T}_{X\pi} \right) \cdot \hat{B}_s \tag{5}$$

siendo

$$\hat{B}_s = T_a^{-1} \cdot \sum_{k \in s} q_k X_k y_k \tag{6}$$

Dicho estimador es el estimador general de regresión (véase Cassel, Särndal y Wretman, 1976).

La forma de $\hat{T}_{Yreg}$ dependerá tanto del diseño muestral como de las constantes q_k elegidas. Por ejemplo, si trabajamos con una única variable auxiliar $X = X_1$, como q_k, elegimos $q_k = \frac{1}{x_k}$ y trabajamos con un muestreo aleatorio simple, entonces

$$\hat{T}_{Yreg} = \frac{\overline{x}}{\overline{y}} \cdot \overline{X}$$

que es el estimador de razón, donde $\bar{y}$ y $\bar{x}$ son las correspondientes medias muestrales de la variable Y y la variable X.

En general, el estimador $\hat{T}_{Yreg}$ no es insesgado, pero como los pesos ω_k son próximos a d_k, es asintóticamente insesgado. Para este estimador se han obtenido expresiones relativamente sencillas de la varianza así como diversos estimadores de ésta que pueden consultarse por ejemplo en Särndal y otros (1989).

Modificaciones en la construcción de los estimadores de calibración

En esta sección revisaremos otros estimadores calibrados que se obtienen, no modificando la distancia elegida, como en la sección anterior, sino modificando el proceso de construcción de un estimador de calibración. Dicho proceso se basa en dos puntos:

1. Minimizar una distancia.
2. Que los pesos equilibrados den estimaciones perfectas para las variables auxiliares.

Por tanto, las modificaciones que aquí revisaremos afectarán a una o a otra condición. Muchas de estas modificaciones harán uso de modelos de superpoblación para que den resultados deseables.

Estimadores de calibración para una familia de distancias

A partir de la distancia (3) se ha construido el estimador de calibración obteniendo el estimador general de regresión. Una primera vía de generalización de los estimadores de calibración se llevó a cabo mediante el uso de distintas medidas de distancia. En concreto, se consideró una familia de distancias de la forma:

$$\sum_{k\in s} G_k(\omega_k d_k)$$

sujeto a la condición ya conocida

$$\sum_{k\in s} \omega_k X_k = T_k$$

donde $g_k(\omega,d)$ verifica una serie de condiciones de regularidad (ver Deville y Särndal, 1992).

A partir de las condiciones impuestas, no podemos garantizar la existencia de solución, pero lo que sí se puede garantizar es que si existe es única, y además vendría dada por

$$\omega_k = d_k F_k\left(X_k^{'}\lambda\right)$$

donde λ es el vector de multiplicadores de Lagrange y $d_k F_k(\cdot)$ es la función inversa de

$$\frac{\partial G_k}{\partial \omega}(\cdot, d_k)$$

Para poder resolver la minimización necesitaremos en ocasiones aplicar métodos tales como *Newton-Rapshon*, etc. Ejemplos de este tipo de funciones se pueden ver en Deville y Särndal (1992) y Deville, Särndal y Sautory (1993), quienes consideran siete funciones diferentes de distancias y examinan las propiedades estadísticas de los correspondientes estimadores de calibración.

Es importante estudiar la relación existente entre esta familia de estimadores de calibración, con el estimador general de regresión. Imponiendo una serie de restricciones, Deville y Särndal (1992) demuestran que cualquier miembro de la clase anterior es asintóticamente equivalente al estimador general de regresión $\hat{T}_{Yreg}$ en el sentido de que

$$N^{-1}\left(\hat{T}_{Y\omega} - \hat{T}_{Yreg}\right) = O_p\left(n^{-1}\right)$$

Como consecuencia, los dos estimadores tienen la misma varianza asintótica. Este resultado es de una gran importancia ya que implica que cualquiera de los estimadores conocidos de la varianza del estimador de regresión pueden usarse para estimar la varianza de cualquier estimador $\hat{T}_{Y\omega}$ en esta clase, siempre y cuando estemos en las condiciones adecuadas.

Estimadores de calibración basados en una forma funcional

Comenzaremos por la modificación consistente en eliminar el proceso de minimización de la distancia entre d_k y ω_k e imponer que los pesos calibrados ω_k tengan una determinada forma funcional, en concreto:

$$\omega_k = d_k + p_k \lambda_s' Z_k$$

donde p_k son constantes positivas y Z_k es un vector, de forma que cumple las dos condiciones siguientes:

(a) $dim(Z_k) = J = dim(X_k)$

(b) La matriz $\sum_{k \in s} p_k Z_k X_k'$ es no singular.

De esta manera, obtenemos el estimador calibrado

$$\hat{T}_{YCALF} = \hat{T}_{Y\pi} + \left(T_X - \hat{T}_{X\pi}\right)' \hat{Q}$$

donde

$$\hat{Q} = \left(\sum_{k \in s} p_k Z_k X_k'\right)^{-1} \cdot \sum_{k \in s} p_k Z_k y_k$$

En primer lugar, podemos observar que el estimador general de regresión pertenece a esta nueva familia de estimadores calibrados, pues:

$$\hat{T}_{YCALF} = \hat{T}_{Yreg}$$

si $Z_k = X_k$ y $p_k = p_k q_k$, es decir, eligiendo el vector Z_k igual al vector de variables auxiliares, obtenemos $\hat{T}_{Yreg}$.

Esta nueva familia de estimadores calibrados también, en general, son estimadores sesgados. Un estudio comparativo entre las varianzas de estos nuevos estimadores y el estimador general de regresión fue dado por Estevao y Särndal (2000), quienes establecen que bajo ciertas condiciones el estimador $\hat{T}_{Yreg}$ tiene varianza asintótica mínima de entre todos los estimadores de la forma $\hat{T}_{YCALIF}$.

Estimadores de calibración cosméticos

Otra posible modificación de los estimadores de calibración consiste en construir un estimador de calibración que sea también un estimador cosmético. Los estimadores cosméticos se pueden interpretar tanto como un estimador predictivo o bien como un estimador basado en un diseño y, por tanto, pueden ser obtenidos por medio de estos dos procedimientos, o bien indirectamente por medio de la calibración, como veremos a continuación.

Para introducir este tipo de estimadores calibrados, asumiremos el siguiente modelo de superpoblación

$$y_k = X_k^{'}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k \quad \text{con} \quad E_\xi[\boldsymbol{\varepsilon}_k] = 0, \quad E_\xi[\boldsymbol{\varepsilon}_k^2] = \boldsymbol{\sigma}^2 a_k^2 \quad \text{y} \quad E_\xi[\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_k] = 0 \quad \forall k \neq j$$

donde $\boldsymbol{\varepsilon}_k$ es un error aleatorio, $\boldsymbol{\sigma}^2$ es un escalar desconocido y a_k^2 son constantes conocidas.

Una vez introducido el modelo de superpoblación, los estimadores calibrados cosméticos surgen al intentar encontrar un estimador de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{COSCAL}$, de manera que la forma estándar y predictirva (Royall, 1970) del estimador general de regresión sean numéricamente iguales, esto es

$$\hat{T}_{YCOSCAL} = \hat{T}_{Y\pi} + \left(T_X - \hat{T}_{X\pi}\right)\boldsymbol{\beta}_{COSCAL} = \tag{8}$$

$$= \hat{T}_{Ys} + \left(T_X - \hat{T}_{Xs}\right)\boldsymbol{\beta}_{COSCAL} \tag{9}$$

siendo $\hat{T}_{Ys}$ y $\hat{T}_{Xs}$ los totales muestrales de Y y de X respectivamente.

Para conseguir este objetivo, es suficiente utilizar las técnicas de calibración ya revisadas, pero minimizando una nueva distancia, y obtenemos

$$\boldsymbol{\omega}_k = d_k + (d_k - 1)q_k X_k^{'}\left(\sum_{k\in s}(d_k - 1)q_k X_k X_k^{'}\right)^{-1}\left(T_X - \hat{T}_{X\pi}\right)$$

y el estimador de calibración, definido como $\sum_{k\in s}\boldsymbol{\omega}_k y_k$, se reduce a la fórmula dada para $\hat{T}_{YCOSCAL}$ en (8).

Brewer (1999) obtiene una expresión para la varianza exacta de este estimador bajo el modelo de superpoblación (7), así como de la varianza asintótica. Esta expresión ya fue presentada por Godambe (1955) como la mínima varianza asintótica posible para cualquier estimador insesgado, con respecto al diseño de T_Y. Por tanto, será preferible, si el modelo de trabajo es útil, estimar la varianza asintótica en lugar de la varianza con respecto al diseño.

Estimador de calibración modelo-asistido

Los estimadores de calibración, que ahora consideraremos en este apartado, surgen al sustituir, en el proceso de calibración, la restricción

$$\sum_{k\in s}\omega_k y_k = T_X$$

por otra restricción más adecuada.

Esta idea surge al plantearse cómo de efectivo es el uso que estamos haciendo de la información adicional a través de la condición

$$\sum_{k\in s}\omega_k y_k = T_X$$

Usando la información adicional de esta manera, estamos implícitamente asumiendo un modelo lineal de regresión, entre la variable de estudio *Y* y las variables incluidas en el vector *X*, en la población de estudio; es decir, si asumimos una relación lineal

$$y_k = X_k^{'}\boldsymbol{\beta} \quad \text{para todo } k \in U$$

los estimadores de calibración hasta ahora estudiados coinciden y reproducen exactamente el total poblacional de la variable de estudio *Y*.

Por tanto, los estimadores calibrados construidos hasta ahora están motivados por un modelo lineal. Si otro tipo de curva es la que relaciona la variable de estudio *Y* con las variables auxiliares que forman el vector *X*, los estimadores de calibración vistos hasta el momento pueden resultar ineficaces.

De este modo, cuando trabajamos con otro tipo de relación entre *X* e *Y*, se hace necesario modificar la condición $\sum_{k\in s}\omega_k y_k = T_X$ y adaptarla a cada tipo de situación, para así poder usar la información auxiliar de la mejor manera posible.

A este tipo de estimadores calibrados, que se adaptan al modelo de trabajo, se les denominan estimadores calibrados modelo-asistidos.

A partir de ahora asumiremos que la relación entre las variables *Y* y el vector *X* puede ser descrita por el siguiente modelo de superpoblación

$$E_\xi(y_k \mid X_k) = \boldsymbol{\mu}(X_k, \boldsymbol{\theta}) \qquad V_\xi(y_k \mid X_k) = \upsilon_k^2 \boldsymbol{\sigma}^2 \tag{10}$$

con $k = 1,2,\ldots,N$ y donde $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)'$ y σ^2 son parámetros poblacionales desconocidos, $\boldsymbol{\mu}(X,\boldsymbol{\theta})$ es una función conocida de X y $\boldsymbol{\theta}$, υ_k es una función conocida de X_k o bien de $\boldsymbol{\mu}_k = \boldsymbol{\mu}(X_k,\boldsymbol{\theta})$ y E_ξ y V_ξ denotan la esperanza y la varianza con respecto al modelo de superpoblación.

Este modelo es bastante general e incluye dos casos muy importantes:

(a) El modelo de regresión lineal o no lineal (1)

(b) El modelo lineal generalizado

$$g(\boldsymbol{\mu}_k)X_k^{'}\boldsymbol{\theta} \qquad V_\xi(y_k \mid X_k) = \upsilon(\boldsymbol{\mu}_k) \qquad k = 1,2,\ldots N$$

donde $\boldsymbol{\mu}_k = E_\xi(y_k \mid X_k)$, $g(\cdot)$ es una función de enlace y $\upsilon(\cdot)$ es la función varianza.

Se define el estimador de calibración modelo-asistido de T_Y como

$$\hat{T}_{YMC} = \sum_{k\in s}\omega_k y_k$$

donde los pesos calibrados ω_k minimizan una medida de distancia con respecto a d_k, sujeto a la nueva condición

$$\sum_{k\in s}\omega_k\,\mu\left(X_k\hat{\theta}\right)=\sum_{k=1}^{N}\mu\left(X_k\hat{\theta}\right)$$

Minimizando la medida de distancia, obtenemos el siguiente estimador

$$\hat{T}^{*}_{YMC}=\hat{T}_{Y\pi}+\left(\sum_{k=1}^{N}\hat{\mu}_k-\sum_{k\in s}d_k\hat{\mu}_k\right)\ddot{B}^{*}_{N}$$

donde

$$\ddot{B}^{*}_{N}=\frac{\sum_{k\in s}d_kq_k\hat{\mu}_ky_k}{\sum_{k\in s}d_kq_k\hat{\mu}_k^2}$$

Si se impone además la restricción $N^{-1}\sum_{k\in s}\omega_k = 1$, entonces se obtiene un estimador alternativo $\hat{T}_{YMC}$.

Las propiedades más importantes de estos dos nuevos estimadores fueron obtenidas por Wu y Setter (2001) y vienen dadas por:

(1) Bajo ciertas condiciones, tanto $\hat{T}_{YMC}$ como $\hat{T}^{*}_{YMC}$ son iguales a $\hat{T}_{Y\pi}+O_p\left(n^{-1/2}\right)$ y son, de este modo, asintóticamente insesgados para T_Y con respecto al diseño, sin tener en cuenta si el modelo es correcto o no. Estos estimadores son también aproximadamente insesgados bajo el modelo.

(2) Si $q_k=\dfrac{1}{\upsilon_k^2}$ entonces $\hat{T}_{YMC}$ y $\hat{T}^{*}_{YMC}$ son modelo-asistidos, en el sentido de que pueden manejar tanto modelos lineales como no lineales; esto es, ambos son consistentes respecto al diseño, sin tener en cuenta el modelo; y particularmente eficientes si el modelo es correcto. También, en el caso de un modelo sin error (esto es, $y_k = \mu_k$), tenemos que

$$\hat{T}_{YMC}=\hat{T}^{*}_{YMC}=\hat{T}_{Y}$$

Además, tanto $\hat{T}_{YMC}$ como $\hat{T}^{*}_{YMC}$ se reducen al estimador convencional (Deville y Särndal, 1992) si trabajamos con un modelo lineal.

Algunos ejemplos

Para ilustrar el comportamiento de los distintos estimadores de calibración expuestos, hemos realizado un pequeño estudio de simulación sobre dos poblaciones representativas.

La primera población considerada es natural, constituida por 430 granjas de ganado, usada por Chambers, Dorfman y Wherly en un informe técnico de la *Australian National University*. En él, la variable objeto de estudio, Y, son los ingresos obtenidos del ganado, y la variable auxiliar, X, es el número de cabezas de ganado.

En esta población se seleccionaron 100 muestras de tamaño 50%, por medio de los siguientes esquemas de muestreo:

- Un muestreo aleatorio simple.
- Un esquema de Midzumo usando la variable número de cabezas de ganado.

Para cada muestra obtenida se calcularon los estimadores $\hat{T}_{Yreg}$, $\hat{T}_{YCALF}$ y $\hat{T}_{COSCAL}$.

Tanto para $\hat{T}_{Yreg}$ como para $\hat{T}_{COSCAL}$,se tomó q_k = 1, para todo $k \in s$.

Para realizar las estimaciones con $\hat{T}_{YCALF}$, se tomó como $z_k = \left(x_{k1}^2\right)$, y como p_k = 430/50, pata todo $k \in s$.

Una vez realizadas todas las estimaciones, se usó la varianza en el muestreo de los estimadores para aproximar su varianza teórica. Los resultados de estas varianzas son:

(1). En el muestreo aleatorio simple:

$$\upsilon_{\mathrm{Yreg}} = 231272053$$
$$\upsilon_{\mathrm{YCALF}} = 158926882$$
$$\upsilon_{\mathrm{YMC}} = 231272053$$

(2). En el muestreo de Midzumo:

$$\upsilon_{\mathrm{Yreg}} = 196353502$$
$$\upsilon_{\mathrm{YCALF}} = 148087682$$
$$\upsilon_{\mathrm{YMC}} = 197006992$$

Como se puede observar, el mejor comportamiento en ambos tipos de muestreo lo tiene el estimador de calibración basado en una forma funcional, $\hat{T}_{YCALF}$, con la adecuada selección de Z_k. También se puede observar cómo el método de selección muestral influye considerablemente en la precisión de los estimadores.

La segunda población es una población artificial, usada para ilustrar la diferencia en el comportamiento de los estimadores en un modelo claramente no lineal. Para ello se ha generado una población, de tamaño 1000, bajo el siguiente modelo:

$$Y = \sqrt{X_1 + 2 \cdot X_2} + \varepsilon$$

donde

$$X_1 \sim \mathit{Gamma}\ (50,1)$$
$$X_2 \sim \mathit{Exp}\ (1)$$
$$\varepsilon \sim \mathit{Normal}\ (0;0,5)$$

Se han seleccionado también 100 muestras de tamaño 100, mediante muestreo aleatorio simple sin reemplazamiento.

De esta manera se realizó la estimación del total de la variable Y, usando como variables auxiliares a X_1 y X_2, para las 100 muestras, por medio de $\hat{T}_{Yreg}$, $\hat{T}_{YCALF}$ y $\hat{T}^*_{YMC}$.

Tanto para $\hat{T}_{Yreg}$ como para $\hat{T}^*_{YMC}$, se tomó q_k = 1, para todo $k \in s$.

Para realizar las estimaciones, con $\hat{T}_{YCALF}$, se tomó como $z_k = \left(x_{k1}^2, x_{k2}^2\right)$, y como p_k = 1000/100, para todo $k \in s$.

Para realizar las estimaciones con $\hat{T}^{*}_{YMC}$, se tomó como:

$$\mu_k\left(x_{k1}, x_{k2}, \theta_1, \theta_2\right) = \sqrt{\left(b_1 \cdot x_{1k} + b_2 \cdot x_{2k}\right)}$$

Como resultado, las diferentes varianzas implicadas son:

$$\upsilon_{\text{Yreg}} = 5699{,}969$$
$$\upsilon_{\text{YCALF}} = 5660{,}784$$
$$\upsilon_{\text{YMC}} = 2651{,}443$$

Como puede observarse, en esta población el estimador de calibración modelo asistido es mucho más preciso que los otros dos, reduciéndose su varianza en más de la mitad de la correspondiente a los otros estimadores. Así pues, si se tiene algún conocimiento del modelo que liga las variables, es fundamental incorporar esta información en la construcción de un estimador de calibración modelo asistido.

Otros trabajos sobre calibración

En los últimos años, la idea de la calibración ha tenido una gran aceptación entre los teóricos de muestreo en poblaciones finitas y ha surgido así una gran cantidad de trabajos que utilizan el modelo de calibración bajo distintos tipos de diseños muestrales, así como para la estimación de otros parámetros distintos al total poblacional.

Algunos de estos trabajos a destacar son los siguientes:

Dupont (1995) considera estimadores de calibración de un total cuando la información auxiliar es obtenida en diferentes fases mediante un procedimiento de muestreo bifásico.

Mukhopadhayay (2000ª) obtiene estimadores de calibración para la función de distribución poblacional y también formula estimadores de calibración para la varianza poblacional (2000b).

Théberge (1999) considera estimadores de calibración para estimar cualquier parámetro lineal dando una solución al caso de que no exista solución al problema de calibración mediante inversas de Moore-Penrose.

Lundström y Särndal (1999) aplican la técnica de calibración para el tratamiento de la no respuesta, reduciendo tanto el error muestral como el sesgo de no respuesta.

Welsh y Ronchetti (1998) proponen estimadores calibrados bajo la suposición de que la muestra contiene *outliers* (valores raros) representativos.

Singh y otros (1999) investigan nuevas técnicas de calibración para estimadores de la varianza de medias muestrales, estimadores de razón y regresión bajo diferentes esquemas de muestreo. Singh (2001) también propone estimadores de la varianza poblacional.

Estos trabajos pueden darnos una idea de lo rápido que se ha extendido esta nueva forma de utilizar la información auxiliar, aplicándose a muy variados problemas de estimación. Si bien en su comienzo los estimadores calibrados tenían el inconveniente de que su utilización sólo era sencilla en un caso muy concreto de modelos de superpoblación y para una cierta métrica, los últimos trabajos han permitido extender la técnica para modelos de superpoblación muy generales, parámetros más amplios y una amplia gama de métricas, resolviendo los problemas de no existencia de solución de la ecuación de calibración. Además, permiten englobar como casos particulares a los estimadores más usuales.

En nuestra opinión, la propiedad más importante a destacar en estos estimadores es que pueden abordarse desde los distintos enfoques hoy seguidos en el muestreo: el enfoque del modelo de población fija, el enfoque predictivo y el enfoque modelo asistido, siendo pues una metodología que puede ser aceptada por cualquiera de los estudiosos en muestreo, independientemente de su forma de hacer inferencia en este área.

Referencias

Brewer, K.R.W. (1999) Cosmetic calibration with unequal probability sampling. *Survey Methogology*, 25, 205-212.

Cassel, C.M., Särndal, C.E. y Wretman, J.H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.

Chen, J. y Qin, J. (1993) Empirical likelihood estimation from finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, J. y Sitter, R.R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.

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

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.

Dupont, F. (1995) Alternative adjustements where there are several levels of auxiliary information. *Survey Methodology*, 21 (2), 125-135.

Estevao, V.M. y Särndal, C.E. (2000) A functional form approach to calibration. *Journal of Official Statistics*, 16 (4), 379-399.

Godambe, V.P. (1955) A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Ser. B, 17, 269-278.

Hedayat, A.S. y Sinha, B.K. (1991) *Desing and inference in finite population sampling*. New York: John Wiley & Sons.

Lundström, S. y Särndal, C.E. (1999) Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15 (2), 305-327.

Mukhopadhyay, P. (2002a) Calibration estimators of a finite population variance. *Parisankhyan Samikkha*.

Mukhopadhyay, P. (2002b) On estimating a finite population distribution function. *Parisankhyan Samikkha*.

Pérez, R.A. (2002) ¿Qué es un modelo de superpoblación? *Metodología de Encuestas*, 4 (1), 79-86.

Royal, R.M. (1970) On finite population sampling theory under certain linear regresión models. *Biométrica*, 57, 377-387.

Sánchez-Crespo, G. (2002) Introducción a los modelos de superpoblación en las técnicas de muestreo con probabilidades desiguales. *Metodología de Encuestas*, 4 (1), 87-104.

Singh, S. (2001) generalized calibration approach for estimating variance in survey sampling. *Am. Inst. Statist. Math.* 53, 404-417.

Singh, S.; Horn, S; Showdhury, S. y Yu, F. (1999) Calibration of the estimators variance. *Australian Statistical Publishing Association*.

Théberge, a. (1999) Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Tillé, Y. (1995) Auxiliary information and conditional inference, en *Bulletin of the International Statistical Institute*, *Proceedings fo the 50th session*, 1, 303-319.

Welsh, A.H. y Ronchetti, E. (1998) Bias calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, Ser. B, 60 (2) 413-428.

Wu, C. y Sitter, R. (2001) Approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.