Metodología de Encuestas
Volumen 10, 2008, 77-91
ISSN: 1575-7803

Revisión de las principales técnicas de estimación de parámetros lineales en subpoblaciones

Muñoz Rosas, Juan Francisco.
Departamento de Métodos Cuantitativos para la Economía y la Empresa.
Universidad de Granada.
jfmunoz@ugr.es

Álvarez Verdejo, Encarnación.
Departamento de Comercialización e Investigación de Mercados.
Universidad de Granada.
encarniav@ugr.es

Sánchez Borrego, Ismael.
Departamento de Estadística e Investigación Operativa.
Universidad de Granada. ismasb@ugr.es
Dirección postal.: Departamento de Métodos Cuantitativos para la Economía y la Empresa. Facultad de Ciencias Económicas y Empresariales. Universidad de Granada. C.P. 18071 Granada. Tfno.: 958241955, Fax.: 958240620.

RESUMEN. Es conocido el interés de las distintas instituciones estadísticas por las técnicas apropiadas de muestreo que obtengan estimaciones fiables y eficientes para una determinada característica de interés. El objetivo principal de estos estudios suele ser la estimación de la media o total poblacional mediante el análisis, en general, de muestras grandes. Sin embargo, el objetivo de estos organismos no se centra exclusivamente en el análisis a nivel nacional, sino que también se realizan comparaciones entre grupos de poblaciones más pequeños llamados subpoblaciones. En este trabajo se recopilan las principales técnicas de estimación de la media poblacional en subpoblaciones.

PALABRAS CLAVE. Subpoblación, estimador sintético, estimador compuesto, estimadores basados en modelos.

ABSTRACT. It is known the interest of several statistical offices by the procedures related to survey sampling, which can provide efficient estimates to a given variable of interest. Assuming large samples, the main purpose derived from these studies generally is to estimate the population mean (or the total) at a national level. However, the statistical offices are also interested on the comparison between smaller groups of populations called subpopulations or domains. The main procedures for estimating the mean in the context of subpopulations are summarized in this paper.

KEYWORDS. Subpopulation,

Recibido: 20 de octubre 2007
Revisado: 8 de marzo 2008
Aceptado: 18 de abril 2008

Introducción

Uno de los objetivos de la teoría de muestreo es la obtención de estimaciones de medias o totales de una serie de características asociadas con los individuos de una determinada población en estudio.

Las metodologías relacionadas con este tópico han evolucionado enormemente en las últimas décadas hasta el punto de que hoy en día se disponen de numerosas técnicas para la estimación de parámetros y de múltiples y variados diseños muestrales que hacen que cualquier investigación por muestreo disponga de la suficiente capacidad para obtener resultados fiables y eficientes.

En otras palabras, la mayoría de las investigaciones llevadas a cabo en la teoría del muestreo en poblaciones finitas se han centrado en la búsqueda de nuevas técnicas de estimación y el estudio de éstos en diseños muestrales complejos, sin hacer especial hincapié en la estimación de parámetros no lineales, como los cuantiles, o el estudio de las técnicas de muestreo en otros aspectos de interés en las investigaciones estadísticas como es el caso de las subpoblaciones.

En este trabajo se realiza una revisión de las técnicas de muestreo utilizadas en el problema de la estimación de parámetros lineales en subpoblaciones.

En muestreo se habla de subpoblaciones cuando se desean investigar por separado las características de uno o varios grupos específicos de la población, es decir, se lleva a cabo el estudio según el plan previsto, pero además de los resultados obtenidos por la muestra para la población general, se desea obtener información sobre uno o varios de estos grupos concretos de la población, llamados subpoblaciones, que pueden estar definidos antes de que la investigación fuese llevada a cabo o después de que ésta empezase.

Un ejemplo usual de subpoblaciones aparece, por ejemplo, en los organismos nacionales de Estadística. Estas entidades llevan a cabo investigaciones basadas en grandes muestras, las cuales son muy fiables y presentan una precisión excelente a nivel nacional, pero sin embargo, la situación es muy distinta cuando los estudios van referidos a subpoblaciones. Las subpoblaciones surgen cuando estos organismos desean ofrecer resultados o conclusiones a nivel regional, provincial o incluso local, y la técnica a seguir es dividir la muestra inicial por regiones. Además de esta división, se hacen otras particiones como por ejemplo según el sexo, la edad, la raza, que los individuos estén desempleados o no, si las viviendas son de un miembro, de dos miembros, etc.

En el caso de las subpoblaciones, las muestras suelen contener muy pocas observaciones, el tamaño de la muestra es aleatorio y resulta imposible realizar estimaciones con una precisión aceptable. Por esta razón, el problema de la estimación en subpoblaciones dispone de técnicas distintas a las aplicadas para una muestra general. El objetivo de este trabajo es recopilar estas técnicas, mostrar varias estrategias que pueden ayudar a obtener mejores estimaciones y analizar otras propiedades relacionadas con el problema de las estimaciones en subpoblaciones.

Uno de los primeros estudios dedicados a la teoría de las subpoblaciones fue llevado a cabo por Hartley (1959), en el cual se estudiaron bajo varios diseños muestrales estimadores que no estaban basados en información auxiliar. Además de

estos estimadores, Hartley propuso un estimador de tipo razón, el cual usaba los totales poblacionales de la subpoblación en estudio. A partir de aquí, son varios los autores que han investigado sobre este problema y numerosas las definiciones y términos que han sido descritos exclusivamente para las subpoblaciones. Por ejemplo, Purcell y Kish (1979) introdujeron cuatro tipos de subpoblaciones dependiendo del tamaño de la subpoblación, y donde tal clasificación puede usarse para escoger el método de estimación apropiado que mejore la precisión de las estimaciones, mientras que Estevao, Hidiroglou y Särndal (1995) fueron los primeros en reconocer que los pesos de los estimadores basados en información auxiliar podían ser dependientes o no dependientes de la subpoblación. Otros trabajos interesantes sobre subpoblaciones se deben a Rao (1986), Chaudhuri (1994), Särndal, Swensson y Wretman (1992), Ghosh y Rao (1994), Estevao y Särndal (1999), Pfeffermann (2002), Dehnel, Golata y Klimanek (2004), Zhang y Chambers (2004), etc.

La estimación en subpoblaciones es un tema de gran interés para las economías de varios países de Europa y de la Unión Soviética. En la década de los 90, estos países se movilizaron para desarrollar determinadas decisiones centralizadas que dieron lugar al proyecto de investigación EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs, IST-2000-5.1.8, 2001-2003.), en el cual las investigaciones por muestreo no estaban centradas en producir estimaciones para poblaciones grandes, sino que al contrario, mejorar las estimaciones en pequeñas subpoblaciones. Además de estos estudios, se han organizado varias conferencias científicas sobre estimaciones en subpoblaciones o estimaciones en áreas pequeñas, nombre que reciben las subpoblaciones cuando los grupos se definen geográficamente. Estas conferencias se desarrollaron en Warsaw, Polonia (1992), en Riga, Letonia (1993), y más recientemente, en la Universidad de Jyväskylä, Finlandia, en Agosto de 2005. El objetivo de ésta última fue la de presentar los avances y principales conclusiones del proyecto EURAREA.

Notación y conceptos básicos.

En primer lugar, detallaremos el marco de trabajo que usualmente se sigue en muestreo de poblaciones finitas y en concreto el seguido en subpoblaciones.

Sea $U = \{1,\ldots,i,\ldots,N\}$ una población finita que contiene *N* elementos distintos identificados. Dentro de esta población interesa estudiar ciertas características de una variable de estudio que se denomina *y*. Asociado al elemento *i* de la población se conoce exactamente y sin error el valor de la característica de interés, esta cantidad se denotará como y_i. Observar el valor y_i en todas las unidades de la población va a resultar imposible o muy costoso, así que se utiliza una muestra para conocer los valores y_i de las unidades que pertenecen a la muestra. Una muestra es por tanto un subconjunto de *n* elementos $(n < N)$, *s*, de *U*

con sus valores asociados de *y*, es decir $\{(i, y_i)\}$, seleccionados de acuerdo con un diseño de muestreo específico que asigna una probabilidad conocida $p(s)$, tal que $p(s) > 0$ para todo $s \in S$, conjunto de las posibles muestras *s,* y $\sum_{s \in S} p(s) = 1$. Las probabilidades de inclusión de primer y segundo orden asociadas a este plan de muestreo se denotan como π_i y π_{ij}, respectivamente, y $w_i = \pi_i^{-1}$ son los pesos básicos del diseño. Además de la variable de interés se dispone de otras variables auxiliares $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_J\}$, donde $\boldsymbol{x}_j = \{x_{1j}, \ldots, x_{ij}, \ldots, x_{NJ}\}$, que también son conocidas para aquellos individuos seleccionados en la muestra. Se supone además, que se conocen los totales poblacionales de las variables auxiliares, es decir, las cantidades $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j, \ldots, \boldsymbol{X}_J\}$ son conocidas, donde $X_j = \sum_{i \in U} x_{ij}$.

La población *U* se puede dividir en *D* subgrupos mutuamente excluyentes y exhaustivos $U_1, \ldots, U_d, \ldots, U_D$, llamados subpoblaciones y donde el tamaño de U_d es N_d, es decir, se verifica:

$$U = \bigcup_{d=1}^{D} U_d \;\; ; \;\; N = \sum_{d=1}^{D} N_d .$$

El objetivo es estimar el total (o media) poblacional de la variable *y* para una subpoblación cualquiera, es decir, estimar $Y_d = \sum_{i \in U_d} y_i$. Para estimar este parámetro se utiliza la muestra *s*, la cual se puede dividir en *D* grupos $s_1, \ldots, s_d, \ldots, s_D$, donde $s_d = s \cap U_d$. El tamaño muestral de s_d es aleatorio, se denota como n_d y es un valor que puede ser muy pequeño o incluso igual a cero. Además, se verifica:

$$s = \bigcup_{d=1}^{D} s_d \;\; ; \;\; n = \sum_{d=1}^{D} n_d .$$

Para definir los estimadores y otras características en subpoblaciones suele resultar útil trabajar con variables indicadoras. Así, para la *d*-ésima subpoblación se tiene que $\delta_{di} = 1$ si $i \in U_d$ y $\delta_{di} = 0$ en otro caso. Del mismo modo, se pueden definir este tipo de variables indicadoras para las variables de interés y auxiliares. Por ejemplo, $y_{di} = y_i$ si $i \in U_d$ y $y_{di} = 0$ en caso contrario.

Los estimadores se pueden clasificar como estimadores de tipo Horvitz-Thompson o estimadores de tipo Háyek. La diferencia entre ellos está en que los estimadores de la familia Horvitz-Thompson usan el verdadero valor del tamaño de la subpoblación, N_d, mientras que los estimadores de tipo Háyek usan una

estimación de esta cantidad, dada por $\hat{N}_d = \sum_{i \in s_d} \pi_i^{-1}$. En el caso de estimar la media poblacional, si N_d es desconocido se puede usar tan solo estimadores de tipo Háyek, mientras que si es conocido se pueden computar ambos estimadores. En la práctica, cuando se estudian las subpoblaciones no suelen conocerse el tamaño de éstas.

Un primer estimador que puede usarse en subpoblaciones es el directo o de expansión. Los estimadores directos están destinados a usarse solamente en el caso de no disponer de información auxiliar, puesto que son mucho menos eficientes que el resto de estimadores. El estimador directo para la *d*-ésima subpoblación esta dado por:

$$\hat{Y}_d = \sum_{i \in s} w_i y_{di} = \sum_{i \in s_d} w_i y_i \ .$$

Notamos que este estimador es de tipo Horvitz-Thompson y que para el total poblacional también se pueden definir estimadores de tipo Háyek, que no serán vistos con el fin de no extender este trabajo.

Estimadores de regresión

El uso de estimadores de regresión generalizados (GREG) es la alternativa más conocida para mejorar las estimaciones de los estimadores directos. Bajo una muestra general, el estimador GREG del total de *y* se define como

$$\hat{Y}_{GR} = \hat{Y} + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{B}},$$

donde $\hat{Y} = \sum_{i \in s} w_i y_i$, $\hat{\boldsymbol{X}} = \sum_{i \in s} w_i \boldsymbol{x}_i$, $\hat{\boldsymbol{B}} = \left(\sum_{i \in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T / c_i \right)^{-1} \sum_{i \in s} w_i \boldsymbol{x}_i y_i / c_i$ es el parámetro de regresión poblacional estimado, y c_i son constantes positivas.

Bajo el contexto de subpoblaciones, se pueden plantear diferentes tipos de estimadores GREG. En primer lugar, se define el estimador dado por

$$\hat{Y}_{d,GR1} = \hat{Y}_d + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{B}}.$$

Según el valor que se le asigne a c_i se pueden obtener diferentes estimadores. El estimador de razón se puede obtener cuando bajo una sola variable auxiliar se considera $c_i = x_i$ en el término $\hat{\boldsymbol{B}}$. Haciendo este cambio, el estimador que resulta es:

$$\hat{Y}_{d,R1} = \hat{Y}_d \frac{X}{\hat{X}}.$$

El conocido estimador de regresión lineal también surge al considerar $\boldsymbol{x}_i = (1, x_i)^T$ y $c_i = 1$:

$$\hat{Y}_{d,reg1} = \hat{Y}_d + (X - \hat{X})^T \hat{B}_{reg},$$

donde $\hat{B}_{reg} = \sum_{i \in s} w_i (x_i - \hat{\bar{X}}_{HT})(y_i - \hat{\bar{Y}}_{HT}) \Big/ \sum_{i \in s} w_i (x_i - \hat{\bar{X}}_{HT})^2$, $\hat{\bar{X}}_{HT} = \hat{X}/N$ y $\hat{\bar{Y}}_{HT} = \hat{Y}/N$.

En segundo lugar, se pueden construir estimadores GREG con información auxiliar dependiente del dominio:

$$\hat{Y}_{d,GR2} = \hat{Y}_d + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_d)^T \hat{\boldsymbol{B}}_d,$$

donde $\hat{\boldsymbol{B}}_d = \left(\sum_{i \in s_d} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T / c_i \right)^{-1} \sum_{i \in s_d} w_i \boldsymbol{x}_i y_i / c_i$. Asignando a c_i las expresiones correspondientes y operando adecuadamente, se puede obtener el estimador de tipo razón:

$$\hat{Y}_{d,R2} = \hat{Y}_d \frac{X_d}{\hat{X}_d},$$

y el estimador de regresión lineal:

$$\hat{Y}_{d,reg2} = \hat{Y}_d + (X_d - \hat{X}_d)^T \hat{B}_{d,reg},$$

donde $\hat{B}_{d,reg}$ se define igual que $\hat{B}_{reg}$ pero para la subpoblación *d*.

Más recientemente, en Hidiroglou y Patak (2004), se han definido nuevos estimadores de tipo regresión con buenas propiedades de eficiencia. Estos estimadores se diferencian de los anteriores en el modo de definir el término $\hat{\boldsymbol{B}}$ y en cómo combinarlo con la información auxiliar. Así, se puede definir el parámetro de regresión estimado $\hat{\boldsymbol{B}}_{ds} = \left(\sum_{i \in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T / c_i \right)^{-1} \sum_{i \in s_d} w_i \boldsymbol{x}_i y_i / c_i$, para formar el estimador:

$$\hat{Y}_{d,GR3} = \hat{Y}_d + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{B}}_{ds},$$

donde asignando los valores correspondientes a c_i, se pueden obtener de nuevo estimadores de tipo razón y regresión lineal. El estimador de tipo razón que surge en este caso fue propuesto por Hidiroglou (1991), y discutido con mayor detalle en Estevao et al. (1995).

El estimador GREG planteado en último lugar está dado por

$$\hat{Y}_{d,GR4} = \hat{Y}_d + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_d)^T \hat{\boldsymbol{B}},$$

donde la información auxiliar depende de la subpoblación mientras que el parámetro de regresión está al nivel de la muestra general.

Otras propiedades, como varianzas y sus estimaciones, de estos estimadores GREG pueden consultarse en Hidiroglou y Patak (2004) y Rao (2003).

Otros estimadores indirectos

En la sección anterior se han introducido los estimadores de tipo regresión para la estimación de parámetros lineales en una determinada subpoblación. Aunque para un diseño general estos estimadores presentan buenas propiedades de eficiencia con aceptables errores de muestreo, la situación es bastante diferente en el contexto de las subpoblaciones, es decir, estos estimadores pueden producir grandes errores en muestras demasiado pequeñas provenientes de la subpoblación de interés. Esta situación hace necesaria la búsqueda de nuevos estimadores que aprovechen al máximo la información disponible en un diseño muestral basado en el estudio de una subpoblación, y que por tanto, se disminuya el error de muestreo que produciría un estimador de regresión.

Existen varios métodos indirectos para la estimación en subpoblaciones, de entre los que destacamos los estimadores sintéticos y compuestos.

Estimadores sintéticos

En el contexto de subpoblaciones, los estimadores se llaman sintéticos cuando éstos se basan en un estimador directo definido para una población mayor que cubre varias subpoblaciones y bajo el supuesto de que todas las subpoblaciones poseen las mismas características que la población mayor.

El caso más simple de definición de un estimador sintético se presenta cuando no se dispone de información auxiliar. En este caso, el estimador sintético para el total poblacional en una subpoblación d está dado por $\hat{Y}_{d,s} = \hat{Y} = \sum_{i \in s} w_i y_i$, mientras que el estimador para la media es

$$\hat{\bar{Y}}_{d,s} = \hat{\bar{Y}}_{HT} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i},$$

es decir, el estimador sintético para la media, por ejemplo, es el propio estimador directo definido para toda la media poblacional. Si se cumple el supuesto de que las subpoblaciones son idénticas entre sí, entonces la media de la subpoblación será aproximadamente igual a la media general, obteniendo de este modo un estimador sintético más eficiente que los estimadores no sintéticos, puesto que su error cuadrático medio tenderá a ser pequeño.

Cuando se dispone de los totales de las variables auxiliares para la subpoblación de estudio, se puede definir un estimador sintético de tipo regresión para el total como:

$$\hat{Y}_{d,GRs} = \boldsymbol{X}_d^T \hat{\boldsymbol{B}}.$$

Otro caso particular surge en la presencia de una única variable auxiliar y al considerar $c_i = x_i$. En este caso se obtiene el siguiente estimador sintético de tipo razón:

$$\hat{Y}_{d,Rs} = \hat{Y}\frac{X_d}{\hat{X}}.$$

Estimadores compuestos

Los estimadores compuestos surgen al intentar solapar las mejores propiedades de un estimador sintético con las de un estimador directo. Esto es, si denominamos $\hat{Y}_{d1}$ el estimador directo para el total poblacional y $\hat{Y}_{d2}$ un estimador sintético, el estimador compuesto del total para la *d*-ésima subpoblación se define como

$$\hat{Y}_{d,c} = \theta_d \hat{Y}_{d1} + (1-\theta_d)\hat{Y}_{d2},$$

para un peso θ_d $(0 \le \theta_d \le 1)$ escogido convenientemente.

El método más común para obtener el peso apropiado θ_d es minimizar el error cuadrático medio del estimador compuesto $\hat{Y}_{d,c}$, el cual viene dado por la expresión

$$ECM(\hat{Y}_{d,c}) = \theta_d^2 ECM(\hat{Y}_{d1}) + (1-\theta_d)^2 ECM(\hat{Y}_{d2}) + 2\theta_d(1-\theta_d)E(\hat{Y}_{d1}-Y_d)(\hat{Y}_{d2}-Y_d).$$

Minimizando $ECM(\hat{Y}_{d,c})$ respecto θ_d se obtienen los pesos óptimos

$$\theta_d^* = \frac{ECM(\hat{Y}_{d2}) - E(\hat{Y}_{d1}-Y_d)(\hat{Y}_{d2}-Y_d)}{ECM(\hat{Y}_{d1}) + ECM(\hat{Y}_{d2}) - 2E(\hat{Y}_{d1}-Y_d)(\hat{Y}_{d2}-Y_d)}.$$

Por otro lado, asumiendo que los términos de covarianza $E(\hat{Y}_{d1}-Y_d)(\hat{Y}_{d2}-Y_d)$ son pequeños respecto $ECM(\hat{Y}_{d2})$, se puede considerar que

$$\theta_d^* \approx \frac{ECM(\hat{Y}_{d2})}{ECM(\hat{Y}_{d1}) + ECM(\hat{Y}_{d2})},$$

concluyendo que $\theta_d^* \in [0,1]$. El último paso seria estimar θ_d^*, puesto que depende de cantidades desconocidas. En Rao (2003) se propone el siguiente estimador

$$\hat{\theta}_d^* = \frac{(\hat{Y}_{d2}-\hat{Y}_d)^2 - \upsilon(\hat{Y}_d)}{(\hat{Y}_{d2}-\hat{Y}_{d1})^2},$$

donde $\upsilon(\cdot)$ es un estimador de la varianza de $\hat{Y}_d$.

Existen otras formas de seleccionar los pesos θ_d. Por ejemplo, Purcell y Kish (1979) proponen usar un peso común, $\theta_d = \theta$, y entonces minimizar el total del error cuadrático medio de todas las subpoblaciones, $\sum_d ECM(\hat{Y}_{d,c})$, con respecto a θ. Bajo este nuevo esquema y operando adecuadamente, puede obtenerse el peso estimado

$$\hat{\theta}^* = 1 - \frac{\sum_d \upsilon(\hat{Y}_d)}{\sum_d (\hat{Y}_{d2} - \hat{Y}_d)^2}.$$

Estimadores basados en modelos

En las últimas décadas, se ha venido desarrollando en teoría de muestreo nuevas técnicas de estimación basadas en modelos de superpoblación. Estas técnicas también pueden ser aplicadas en el contexto de la estimación de parámetros lineales en subpoblaciones.

Debido a la variedad de modelos, en este trabajo se tratará exclusivamente uno en concreto, pero sin embargo, la preocupación sobre la eficiencia de estos estimadores no debe centrarse en la elección adecuada del modelo, sino por el contrario, disponer de las variables auxiliares adecuadas. En la teoría de los modelos, las variables auxiliares juegan un papel aún más importante que en las técnicas basadas en el diseño, de tal modo que el éxito de los métodos basados en modelos dependen fundamentalmente de la disponibilidad de buenas variables auxiliares que funcionen como buenos predictores en los modelos considerados. Véase Rao (2003) para mayor detalle de estimadores basados en modelos definidos en subpoblaciones.

Por otro lado, en el contexto de las subpoblaciones, se pueden diferenciar entre dos modelos distintos, dependiendo del nivel al que está orientado el modelo. De este modo, se habla de modelos a nivel subpoblacional y modelos a nivel individual. Los primeros, relacionan las medias de la variable de interés en las subpoblaciones con las medias de las variables auxiliares, mientras que los segundos, relacionan los valores individuales de la variable de estudio con los valores individuales de las variables auxiliares.

En ambos casos, se van a definir los estimadores bajo las *D* subpoblaciones que consta el estudio, aunque en la práctica es más común usar un subconjunto *m* $(m \subset D)$ de subpoblaciones, puesto que resulta imposible o es muy costoso disponer de toda la información de las *D* subpoblaciones. Notamos que estos estimadores se pueden obtener de forma similar en el caso de usar *m* subpoblaciones.

Modelos a nivel subpoblacional

Se asume que $\theta_d = g(\overline{Y}_d)$ para alguna función $g(\cdot)$ específica que está relacionada con los datos auxiliares $\boldsymbol{z}_d = (z_{1d},\ldots,z_{pd})^T$ a través del modelo lineal

$$\theta_d = \boldsymbol{z}_d^T\boldsymbol{\beta} + b_d\upsilon_d, \qquad d = 1,\ldots,D,$$

donde los valores b_d son constantes positivas, $\boldsymbol{\beta} = (\beta_1,\ldots\beta_p)^T$ es un vector de coeficientes de regresión y las cantidades υ_d son efectos aleatorios independientes e idénticamente distribuidos con esperanza bajo el modelo igual a 0 y varianza constante σ^2.

Para obtener las estimaciones de las medias de las subpoblaciones, se asume que los estimadores directos están disponibles y que los parámetros θ_d verifican

$$\hat{\theta}_d = g(\overline{Y}_d) = \theta_d + e_d, \qquad d = 1,\ldots,D,$$

donde los errores muestrales e_d son independientes con $E(e_d|\theta_d) = 0$, y $V(e_d|\theta_d) = \psi_d$,

donde las varianzas muestrales, ψ_d son conocidas. La relación anterior puede expresarse también como

$$\hat{\theta}_d = \boldsymbol{z}_d^T\boldsymbol{\beta} + b_d\upsilon_d + e_d, \qquad d = 1,\ldots,D.$$

Para estimar los parámetros desconocidos $\boldsymbol{\beta}$ se suele usar el criterio de minimizar el error cuadrático medio en la clase de estimadores lineales insesgados. Bajo esta metodología se obtiene que el estimador insesgado lineal de mínimo error cuadrático medio de θ_d está dado por

$$\tilde{\theta}_d = \gamma_d\hat{\theta}_d + (1-\gamma_d)\boldsymbol{z}_d^T\tilde{\boldsymbol{\beta}},$$

donde $\gamma_d = \dfrac{\sigma^2 b_d^2}{\psi_d + \sigma^2 b_d^2}$, y $\tilde{\boldsymbol{\beta}} = \dfrac{\sum_{d=1}^{D}\boldsymbol{z}_d\hat{\theta}_d/(\psi_d + \sigma^2 b_d^2)}{\sum_{d=1}^{D}\boldsymbol{z}_d\boldsymbol{z}_d^T/(\psi_d + \sigma^2 b_d^2)}$.

Modelos a nivel individual

En este tipo de modelos se asume que los datos auxiliares a nivel individual $\boldsymbol{x}_{di} = (x_{di1},\ldots,x_{dip})^T$ están disponibles para cualquier individuo *i* en cada

subpoblación *d*, y que las medias poblacionales $\boldsymbol{X}_d$ son conocidas. El modelo usual que se asume esta dado por

$$y_{di} = \boldsymbol{x}_{di}\boldsymbol{\beta} + \upsilon_d + e_{di}, \quad i = 1,\ldots,n_d, \quad d = 1,\ldots,D,$$

donde υ_d son variables aleatorias independientes e idénticamente distribuidas con $E(\upsilon_d) = 0$ y $V(\upsilon_d) = \sigma_\upsilon^2$, $e_{di} = k_{di}\widetilde{e}_{di}$ y $\widetilde{e}_{di}$ son variables aleatorias idénticamente distribuidas e independientes de los valores υ_d, satisfaciendo $E(\widetilde{e}_{di}) = 0$ y $V(\widetilde{e}_{di}) = \sigma_e^2$.

En forma matricial, el modelo anterior se puede expresar como

$$\boldsymbol{y}_d = \boldsymbol{X}_d\boldsymbol{\beta} + \upsilon_d \boldsymbol{1}_{n_d} + \boldsymbol{e}_d, \qquad d = 1,\ldots,D.$$

Tomando medias, se tiene que $\mu_{\boldsymbol{d}} = \theta_{\boldsymbol{d}} = \overline{\boldsymbol{X}}_d^T\boldsymbol{\beta} + \upsilon_d$, de donde se obtiene que el estimador lineal insesgado de mínimo error cuadrático medio para μ_d está dado por

$$\widetilde{\mu}_{\boldsymbol{d}} = \overline{\boldsymbol{X}}_d^T\widetilde{\boldsymbol{\beta}} + \gamma_d(\overline{y}_{da} - \boldsymbol{x}_{da}^T\widetilde{\boldsymbol{\beta}}),$$

donde $\overline{y}_{da}$ y $\overline{\boldsymbol{x}}_{da}$ están dadas por $\overline{y}_{da} = \sum_i a_{di}y_{di}/a_{d\cdot}$, $\overline{\boldsymbol{x}}_{da} = \sum_i a_{di}\boldsymbol{x}_{di}/a_{d\cdot}$, siendo $a_{di} = k_{di}^{-2}$ y $a_{d\cdot} = \sum_i a_{di}$. Por otro lado, $\gamma_d = \sigma_\upsilon^2/(\sigma_\upsilon^2 + \sigma_\upsilon^2/a_{d\cdot})$ y $\widetilde{\boldsymbol{\beta}}$ es el estimador lineal insesgado de menor error cuadrático medio de $\boldsymbol{\beta}$ dado por

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{\boldsymbol{d}} A_d\right)^{-1}\left(\sum_{\boldsymbol{d}} B_d\right),$$

donde

$$A_d = \frac{1}{\sigma_e^2}\left(\sum_i a_{di}\boldsymbol{x}_{di}\boldsymbol{x}_{di}^T - \gamma_d a_{d\cdot}\overline{\boldsymbol{x}}_{da}\overline{\boldsymbol{x}}_{da}^T\right),$$

y

$$B_d = \frac{1}{\sigma_e^2}\left(\sum_i a_{di}\boldsymbol{x}_{di}y_{di}^T - \gamma_d a_{d\cdot}\overline{\boldsymbol{x}}_{da}\overline{y}_{da}\right).$$

Software para la estimación de parámetros lineales en subpoblaciones

Como complemento a la revisión de los distintos estimadores de parámetros lineales en subpoblaciones descrita en este artículo, en esta sección se comentan algunos aspectos sobre el software disponible para realizar estimaciones en el contexto de subpoblaciones.

Existen un gran número de programas o paquetes estadísticos que abordan el campo del muestreo en poblaciones finitas, aunque no profundizan, en general, en la estimación de parámetros en subpoblaciones. El programa estadístico *R*, por ejemplo, dispone del paquete o librería *survey* para el estudio y análisis de distintos aspectos del muestreo en poblaciones finitas, entre los que se encuentran las subpoblaciones. Las funciones disponibles en la librería *survey* permiten la obtención de estimadores de parámetros en subpoblaciones de tipo razón y regresión, es decir, de momento no se han incorporado estimadores más complejos y recientes en esta librería. La librería *survey* puede descargarse, para su uso, en la dirección web *http://cran.r-project.org/*, en la cual también se dispone de una serie de manuales explicativos de cada librería.

Para un software más completo y actualizado recomendamos al lector el software desarrollado por los distintos miembros del proyecto de investigación EURAREA, cuyo objetivo fue, tal como se indicó en la Sección 1, el estudio de estimaciones en subpoblaciones o área pequeñas.

Los miembros del proyecto EURAREA escogieron el lenguaje de programación *SAS* para la elaboración del software de estimación en subpoblaciones, puesto que la mayoría de los institutos de estadística europeos (incluyendo el Instituto Nacional de Estadística) usan generalmente dicho programa. *SAS* es un lenguaje de programación que dispone de una batería de procedimientos estadísticos incorporados que pueden ser utilizados directamente y de forma sencilla en la fase de estimación

Tanto los distintos informes, en los que se integran los resultados producidos por el proyecto EURAREA, como la teoría y el software asociados con la estimación en subpoblaciones, están disponibles en la web oficial del proyecto:

http://www.statistics.gov.uk/eurarea

Debido a que los institutos de estadística de algunos países europeos, como por ejemplo Polonia, no utilizan el programa SAS, y con el fin de que los distintos métodos puedan ser aplicados en tales países, en la comentada web del proyecto EURAREA también se dispone de una descripción de los algoritmos de todo el software, de modo que éste pueda ser descargado y programado en cualquier otro lenguaje de una manera relativamente sencilla. En resumen, los métodos de estimación en subpoblaciones abordados en el proyecto EURAREA pueden ser usados por cualquier usuario mediante el programa estadístico *SAS*. Si el usuario no dispone de tal programa, la aplicación de éstas técnicas de estimación también es posible con unos conocimientos básicos de cualquier lenguaje de programación.

Conclusiones y futuras líneas de investigación

En este artículo se hace una breve revisión de los estimadores diseñados para el estudio de subpoblaciones o áreas pequeñas. El análisis en estas regiones se realiza frecuentemente por la mayoría de los organismos estadísticos, agencias estadísticas, investigaciones sociales, etc., cuando, por ejemplo, comparan grupos de poblaciones (comunidades autónomas, provincias, ciudades, etc.) entre ellas.

En el estudio de subpoblaciones, existen diferentes técnicas de estimación para la media poblacional. Los estimadores más conocidos desde la perspectiva de estimación basada en el diseño son los estimadores directos, de regresión, sintéticos y compuestos, mientras que por otro lado, también se encuentran los estimadores basados en modelos. Según las conclusiones del proyecto EURAREA, los métodos de estimación basados en modelos obtienen mejores resultados que los métodos basados en el diseño para subpoblaciones muy pequeñas, mientras que obtienen un nivel similar de precisión para subpoblaciones de tamaño medio. De entre los estimadores basados en el diseño, los sintéticos y los compuestos presentan importantes ganancias en eficiencia respecto al estimador directo y el GREG.

El inconveniente surge en que todos los estudios desarrollados en este tópico se han centrado en la media o el total de la característica de interés, quedando ignorada la estimación de otros parámetros no lineales como varianzas, funciones de distribución y cuantiles.

El problema de la estimación de la función de distribución es un tema actual y muy importante del muestreo en poblaciones finitas, por tratarse de una función que permite determinar las características más importantes de la población en estudio, proporcionando información relevante acerca del comportamiento global de la población. El problema de la estimación de cuantiles y de otros parámetros de tipo no funcional también queda resuelto con el conocimiento de la función de distribución, puesto que éstos pueden obtenerse mediante inversión directa de la función de distribución. De este modo, estos parámetros son muy importantes en algunas investigaciones o estudios, y por tanto, se debe de disponer de buenos métodos y técnicas para obtener las mejores estimaciones posibles. Bajo la metodología de subpoblaciones resulta conveniente definir estimadores para la función de distribución basados en un uso efectivo de la información auxiliar y con mejores estimaciones.

Por otro lado, los cuantiles son muy utilizados en los organismos estadísticos por la información que recogen y por las numerosas aplicaciones que tienen en la práctica. Por ejemplo, los cuantiles son necesarios para obtener medidas como las líneas de pobreza, proporción de bajos ingresos, etc.

En resumen, las estimaciones de parámetros no lineales carecen de estudios en el contexto de subpoblaciones, pero sin embargo, estos parámetros son muy útiles y necesarios en las principales instituciones estadísticas. En este trabajo se han recopilado las principales técnicas de muestreo desarrolladas en subpoblaciones para el problema de la estimación de medias y totales. La extensión de estos métodos a la estimación de otros parámetros no lineales es, por tanto, un campo nuevo para la investigación todavía sin explorar.

Referencias

Chaudhuri, A. (1994). Small domain statistics: a review. *Statistica Neerlandica*, 48, 215-236.

Dehnel, G.; Golata, E. y Klimanek, T. (2004). Consideration on optimal sampling design for small area estimation. *Statistics in transition*, 6, 725-754.

Estevao, V.M.; Hidiroglou, M.A. y Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11 (2), 181-204.

Estevao, V.M. y Särndal, C.E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 213-231.

Ghosh, M. y Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 55-93.

Hartley, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.

Hidiroglou, M.A. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada report, September, 1991.

Hidiroglou, M.A. y Patak (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67-78.

Pfeffermann, D. (2002). Small area estimation. New development and directions. *International Statistical Review*, 70, 125-143.

Purcell, N.J. y Kish, L. (1979). Estimations for small domains. *Biometrics*, 35, 365-384.

Rao. J.N.K. (1986). Synthetic Estimators, SPREE and best model based Predictors. *Proceedings of the Conference on Survey Research Methods in Agriculture*, U.S. Department of Agriculture, Washington, DC, pp.1-16.

Rao. J.N.K. (2003). *Small area estimation*. Wiley, New York.

Särndal, C.E.; Swensson, B. y Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics.

Zhang, L. y Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 479-496.