Técnicas estadísticas para la estimación de líneas de pobreza

A. Arcos, J.F. Muñoz, S. Martínez(*) y M. Rueda

Departamento de Estadística e Investigación Operativa.
Facultad de Ciencias. 18071 Universidad de Granada. España

(*) Departamento de Estadística y Matemática Aplicada. Universidad de Almería.

e-mail: arcos@ugr.es, jfmunoz@ugr.es, spuertas@ual.es, mrueda@ugr.es

Resumen

En este trabajo tratamos algunos aspectos de la estimación de la función de distribución en una población finita y su aplicación en los estudios económicos de pobreza, proporcionando un software útil en la estimación de la proporción de bajos ingresos o en la estimación de líneas de pobreza.

Palabras clave: Función de distribución, muestreo en poblaciones finitas, proporción de ingresos bajos, líneas de pobreza.

1 Introduction

No existe un significado único del término pobreza, si bien un elemento común de las definiciones es la identificación de un nivel de vida que no puede ser alcanzado por ciertas personas. Diversos autores han desarrollado medidas que se pueden clasificar en: a) monetarias, que se basan en ingreso o gasto de las personas o de los hogares, y b) no monetarias, que consideran carencias en indicadores de bienestar (acceso a agua potable, electricidad, nivel de escolaridad, mortalidad infantil, inseguridad)

Además, también existen medidas que combinan distintos indicadores de bienestar en un índice único y otros métodos combinados que intentan integrar la información de carencias de bienes y servicios con la falta de ingresos.

En la medición de la pobreza monetaria se utiliza la determinación de las líneas de pobreza, si bien no todos los organismos que las computan ni los países que las requieren parten de una definición común (pueden referir si un ingreso cubre un gasto de referencia o si un gasto efectuado supera o no un valor de referencia, y el denominado valor de referencia puede ser muy variable). No obstante, se considera que la manera más sencilla y transparente de medir la pobreza resulta ser con el establecimiento de una línea de pobreza monetaria.

Por todo lo anterior, una medida que es habitualmente usada en la comparación de niveles de pobreza es la proporción por debajo de una cantidad α de un cuantil de orden β de una distribución de ingresos.

Por ejemplo, Eurostat define un salario bajo cuando es inferior al 60% ($\alpha = 0.6$) del salario mediano ($\beta = 0.5$) nacional mensual, para después comparar la proporción de empleados con salarios bajos en varias estados miembros.

Otro ejemplo, el IEA (Instituto de Estadística de Andalucía), siguiendo los trabajos del INE (Instituto Nacional de Estadística, España), define la línea

de pobreza relativa del 60% (LP60) de la mediana nacional calculada según el gasto medio.

Es por tanto de interés en los estudios cuyo fin es comparar niveles de pobreza disponer de una estimación de un cuantil (en los anteriores ejemplos la mediana poblacional) para posteriormente computar una fracción dada.

Por otra parte, en la mayor parte de las situaciones reales, los datos en los que se basan la estimaciones de los cuantiles de interés provienen de un muestreo en una población finita. Tiene entonces especial relevancia conservar las probabilidades de inclusión de cada unidad en la muestra hasta la fase final de estimación, y con ello, determinar de forma más acorde con la selección de la muestra los errores muestrales asociados a las estimaciones requeridas.

Otra forma de actuar, aunque usual, prescinde de los factores de corrección por finitud, y puede dar resultados trampa, especialmente cuando el interés radica en una población de tamaño moderado (que bien puede ser una subpoblación de interés de la población de partida)

También es bien cierto que cuando un organismo se encarga de un estudio de esta magnitud recoge, además de la variable de interés (por ejemplo, ingresos) otras variables sociodemográficas, económicas u de otro tipo (gastos, personas por hogar si se muestrean hogares, etc ...). Esto supone que aparte de la información de interés, se dispone de información auxiliar complementaria, por lo que puede ser de gran ayuda tratar de integrarla en la fase de estimación.

En este trabajo nos centraremos en la estimación de una fracción de un cuantil en una población finita a partir de una muestra obtenida con un determinado diseño muestral, con y sin información auxiliar. En la segunda sección introducimos el problema, mostrando algunos estimadores que recientemente se han publicado. En la sección 3 presentamos a grandes rasgos un software implementado (usando R) ad hoc para la estimación de una proporción de ingresos bajos y en la estimación de una LP, con todos los estimadores presentados en la sección anterior. El mismo software comtempla la estimación jacknife de los errores muestrales asociados.

2 La estimación de la función de distribución: líneas de pobreza

Denotamos a la población $U = \{1, \ldots, i, \ldots, N\}$, donde N es el número de hogares en la poblacion. El gasto total en el hogar i lo denotamos y_i. La función de distribución de y la llamamos $F_y(t)$ y está definida como

$$F_y(t) = \frac{1}{N} \sum_{i \in U} \delta(t - y_i \leq t),$$

donde $\delta(\xi)$ toma el valor 1 si $\xi \geq 0$ y 0 en otro caso. El cuantil de orden β de y_k que denotamos $Q_y(\beta)$ está definido como:

$$Q_y(\beta) = \inf\{y : F_y(t) > \beta\} = F_y^{-1}(\beta)$$

Por ejemplo, $Q_y(0.5)$ es la mediana poblacional.

De la pobalción U se extrae una muestra s bajo un diseño muestral con probabilidades de inclusión π_i de primer orden y π_{ij} de segundo orden. Dado un estimador $\widehat{F}_y(t)$ de $F_y(t)$, el cuantil de interés $Q_y(\beta) = F_y^{-1}(\beta)$ se puede estimar usando $\widehat{Q}_y(\beta) = \widehat{F}_y^{-1}(\beta)$.

Así, puede emplearse el estimador de Horvitz-Thompson:

$$\widehat{F}_{HT}(t) = \frac{1}{N} \sum_{j \in s} \frac{\delta(t - y_j)}{\pi_j}$$

o el estimador de Hajeck, cuando el tamaño de la población ha de ser estimado también de la muestra:

$$\widehat{F}_H(t) = \sum_{j \in s} \frac{\delta(t - y_j)}{\pi_j} / \sum_{j \in s} \frac{1}{\pi_j}$$

Cuando además se dispone de información auxiliar (supondremos que de cierta variable x, relacionada con y, tenemos un censo), se pueden utilizar estimadores de razón:

$$\widehat{F}_r(t) = \frac{1}{N} \left[\sum_{j \in s} \frac{\delta(t - y_j)}{\pi_j} / \sum_{j \in s} \frac{\delta(t - \hat{R}x_j)}{\pi_j} \right] \sum_{j=1}^{N} \delta(t - \hat{R}x_j)$$

o de diferencia

$$\widehat{F}_d(t) = \frac{1}{N} \left[\sum_{j \in s} \frac{\delta(t - y_j)}{\pi_j} - \left\{ \sum_{j=1}^{N} \delta(t - \hat{R}x_j) - \sum_{j \in s} \frac{\delta(t - \hat{R}x_j)}{\pi_j} \right\} \right]$$

donde

$$\hat{R} = \sum_{j \in s} \frac{y_j}{\pi_j} / \sum_{j \in s} \frac{x_j}{\pi_j}$$

Otros estimadores que también utilizan información auxiliar más recientes pueden ser también usados, como el estimador de Chambers y Dunstan (1986)

$$\hat{F}_{CD} = \frac{1}{N} \left[\sum_{j \in s} \delta(t - y_j) + \frac{1}{n} \sum_{i \notin s} \sum_{j \in s} \delta\left(\frac{t - b_n x_i}{\nu(x_i)} - \hat{U}_j\right) \right]$$

con

$$\hat{U}_j = \frac{y_j - b_n x_j}{\nu(x_j)}, \quad \text{y} \qquad b_n = \sum_{j \in s} \frac{x_j y_j}{\nu(x_j)} / \sum_{j \in s} \frac{x_j^2}{\nu(x_j)},$$

el de Rao, Kovar y Mantel (1990)

$$\hat{F}_{RKM} = \frac{1}{N} \left[\sum_{j \in s} \frac{\delta(t - y_j)}{\pi_j} + \sum_{i=1}^{N} \hat{G}_i - \sum_{j \in s} \frac{\hat{G}_{jc}}{\pi_j} \right]$$

donde

$$\hat{G}_i = \sum_{j\in s} \frac{1}{\pi_j} \delta(\frac{t-\hat{R}x_i}{\sqrt{x_i}} - \frac{y_j - \hat{R}x_j}{\sqrt{x_j}}) / \sum_{j\in s} \frac{1}{\pi_j}$$

y

$$\hat{G}_{jc} = \sum_{k\in s} \frac{\pi_j}{\pi_{jk}} \delta(\frac{t-\hat{R}x_j}{\sqrt{x_j}} - \frac{y_k - \hat{R}x_k}{\sqrt{x_k}}) / \sum_{k\in s} \frac{\pi_j}{\pi_{jk}}$$

Un estimador aún más reciente es el siguiente:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YH}(t) + \left(F_g(\mathbf{t}) - \widehat{F}_{GH}(\mathbf{t})\right)' \cdot \widehat{D} \tag{1}$$

donde

$$\widehat{D} = T^{-1} \cdot \sum_{k\in s} d_k q_k \Delta(\mathbf{t} - g_k)\Delta(t - y_k)$$

obtenido por el método de calibración definiendo la pseudo-variable $g_k = \widehat{\beta}' \mathbf{x}_k$ for $k = 1, 2, ...N$, donde

$$\widehat{\beta} = \left(\sum_{k\in s} d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \cdot \sum_{k\in s} d_k q_k \mathbf{x}_k y_k \tag{2}$$

es un estimador ponderado del coeficiente de regresión múltiple β entre y y $\mathbf{x}$, y los pesos se obtienen minimizando la distancia chi-cuadrado

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

sujeta a la ecuaciones de calibración

$$\frac{1}{N} \sum_{k\in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P \tag{4}$$

con q_k constantes positivas independientes de los pesos originales d_k. El término $F_g(t_j)$ denota a la función de distribución de la pseudo-variable g evaluada en los puntos t_j, $j = 1, 2, \ldots, P$ arbitrariamente elegidos y verificando $t_1 < t_2 < \ldots < t_P$.

De esta forma se obiene un estimador de calibración que es una verdadera función de distribución.

Bajo la metodología de la verosimilitud empírica, son dos los caminos que existen para hacer las estimaciones: bajo el diseño muestral y bajo la teoría de los modelos de superpoblación introducida en Wu y Sitter (2001). El estimador basado en el diseño muestral viene dado por

$$\widehat{F}_{ve}(t) = \sum_{j\in s} \widehat{p}_j \delta(t - y_j),$$

donde $\widehat{p}_j = \dfrac{1}{\pi_j}\dfrac{1}{1+\lambda' u_j}/\displaystyle\sum_{j\in s}\dfrac{1}{\pi_j}$, λ se obtiene como solución de la expresión $\displaystyle\sum_{j\in s}\dfrac{1}{\pi_j}\dfrac{u_j}{1+\lambda' u_j}/\sum_{j\in s}\dfrac{1}{\pi_j} = 0$, y $u_j = \delta(t_0 - \widehat{R}x_j) - \dfrac{1}{N}\displaystyle\sum_{j=1}^{N}\delta(t_0 - \widehat{R}x_j)$. El valor t_0 ha de ser constante para que $\widehat{F}_{ve}(t)$ sea una auténtica función de distribución. El estimador obtenido con este método es más eficiente si se considera $t_0 = t$.

Bajo la teoría de los modelos de superpoblación, el estimador varia dependiendo del modelo considerado. Por ejemplo, si se asume el altamente usado modelo de regresión de la forma $y_i = \mu(x_i, \theta) + \nu_i \epsilon_i, \quad i = 1, \ldots, N$, donde ν_i es una función conocida de x_i, y los errores ϵ_i son variables aleatorias independientes e idénticamente distribuidas com media 0 y varianza σ^2, el estimador de verosimilitud empírica viene dado por

$$\widehat{F}_{Mve}(t) = \sum_{j\in s}\widehat{p}_j\delta(t - y_j),$$

donde $\widehat{p}_j$ y λ tienen la misma expresión que en el caso anterior. La diferencia esta en que u_j viene dado por

$$u_j = G[t_0 - \mu(x_j, \theta)/\nu_j] - \frac{1}{N}\sum_{j=1}^{N} G[t_0 - \mu(x_j, \theta)/\nu_j],$$

donde $G(\cdot)$ es la función de distribución de los términos ϵ_j. En las situaciones donde sea razonable asumir que los términos ϵ_j están normalmente distribuidos, se obtiene que

$$u_j = \Phi[t_0 - \mu(x_j, \theta)/(\nu_j\sigma)] - \frac{1}{N}\sum_{j=1}^{N}\Phi[t_0 - \mu(x_j, \theta)/(\nu_j\sigma)],$$

donde $\Phi(\cdot)$ es la función de distribución de la distribución normal standard. Los parámetros del modelo, θ y σ^2, se pueden sustituir por $\widehat{\theta}$ y $\widehat{\sigma}^2$, las respectivas estimaciones basadas en la muestra.

Cada uno de los estimadores anteriores posee ventajas e inconvenientes frente a los demás, que son medidas desde el punto de vista de sus cualidades estadísticas como sesgo o error cuadrático medio, que además pueden ser considerados con respecto al diseño muestral o con respecto a cierto modelo de superpoblación.

No todos los estimadores anteriores poseen propiedades deseables de un estimador de la función de distribución como ser monótonos crecientes o tomar valores comprendidos entre 0 y 1. Tampoco todos poseen propiedades como pueda ser la unicidad en su definición o ser sencillos de computar.

Por ello en la práctica lo más conveniente consiste en calcularlos todos, estimar su error muestral y elegir entre estos estimadores alternativos el más preciso. Para ello es fundamental la estimación de los errores muestrales, si bien en algunos casos estas estimaciones son muy laboriosas de computar. En este trabajo damos una estimación común a todos mediante la técnica jacknife:

3 Software

No hay en la actualidad un programa que integre todos estos estimadores. Algunos paquetes muy especializados en muestreo de poblaciones finitas como por ejemplo SUDAAN se limita exclusivamente al estimador de Hajeck.

El software elegido para implementar los estimadores anteriores es el entorno estadístico R gratuito y disponible en la siguiente dirección de Internet "http://www.r-project.org/".

R, incorpora varios libros adicionales que puede invocar en un momento dado, y además R brinda la posibilidad al usuario de crear e incorporar sus propios libros, cargarlos y descargarlos a voluntad y acceder a las funciones que contienen.

El software que presentamos permite: dado un censo de x $(x_i, i = 1, \ldots, N)$ y la muestra $(y_j, \pi_j), j = 1, \ldots, n$ calcula, para cada estimador: $\widehat{F}_{HT}$, $\widehat{F}_H$, $\widehat{F}_r$, $\widehat{F}_d$, $\widehat{F}_{CD}$, $\widehat{F}_{RKM}$, $\widehat{F}_{cal}$, $\widehat{F}_{ve}$ y $\widehat{F}_{Mve}$ la estimación del cuantil de interés, la LP estimada y la estimación jacknife del error de muestreo, adaptable a cualquier tipo de muestreo. No obstante, invitamos a las personas interesadas en este tema o relacionados nos soliciten vía correo eléctrónico la información al respecto que requieran.

References

[1] Chambers, R.L and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 597–604.

[2] Deville, J.C. and Särndal, C.E.(1992). Calibration Estimators in Survey Sampling.(1992) *Journal of the American Statistical Association*, **87** 376-382.

[3] Eurostat (2000) Low-wage employees in EU countries. In Statistics in Focus: Populations and Social Conditions. Luxembourg: Office for Oficial Publicationsof the EC.

[4] Fernández, F. R. and Mayor, J. A.(1994). *Muestreo en Poblaciones Finitas: Curso Básico.* P.P.U. Barcelona,

[5] Rao, J.N.K., Kovar, J.G. and Mantel, H.J.(1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika **77**, 365-375.

[6] Wu, C., Sitter, R.R, (2001), Approach to Using Complete Auxiliary Information from Survey Data, Journal of the American Statistical Association, 96, 185-193.