Estadística Española • Volumen 58, número 190 / 2016, pp. 123-131

Estimación de regresión multivariante insesgada

Mariano Ruiz Espejo
Universidad Católica San Antonio de Murcia

Resumen

Proponemos un estimador de regresión multivariante insesgado para la media poblacional, que puede ser optimizado minimizando su varianza. En el caso bivariante obtenemos el estimador insesgado óptimo teórico, que puede ser aproximado por otro estimador insesgado práctico con varianza estimable insesgada.

Palabras clave: estimación insesgada de regresión multivariante, estimación insesgada de regresión bivariante, estimador insesgado óptimo teórico, estimación práctica.

Clasificación AMS: 62D05.

Unbiased multivariate regression estimation

Abstract

We propose an unbiased multivariate regression estimator for the population mean, which can be optimized minimizing its variance. In the bivariate case, we obtain the theoretic optimum unbiased estimator, which can be approximate for other practical unbiased estimation with unbiased estimable variance.

Keywords: multivariate regression unbiased estimation, bivariate regression unbiased estimation, theoretic optimum unbiased estimator, practical estimation.

AMS Classification: 62D05.

1. Introducción

Consideramos una población finita de tamaño N en cuyas unidades tenemos definidas la variable de interés y a observar en el mundo real, y las m variables auxiliares ya disponibles y almacenadas $x_1, x_2, \dots, x_m$, todas ellas (las $m+1$ variables) definidas y concretadas de modo fijo en cada unidad $k = 1, 2, \dots, N$ de la población finita. Nuestro objetivo es estimar sin sesgo la función paramétrica media poblacional definida por

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

Para ello, un estimador insesgado de $\bar{y}$ es la media muestral de la variable de interés definida por

$$\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k$$

Este estimador natural es insesgado junto con el diseño de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo n, en el cual la muestra no ordenada s, subconjunto de la población finita y de cardinal n, es una concreción del diseño muestral con probabilidades iguales de selección (Ruiz Espejo, 2013). Sin embargo este estimador media muestral no aprovecha la información de las m variables auxiliares que disponemos.

En este artículo proponemos un estimador insesgado general t_u que aprovecha toda la información auxiliar disponible, concretamente el estimador

$$t_u = \bar{y}_s + \sum_{i=1}^{m} k_i\left(\bar{x}_i - \bar{x}_{i,s}\right)$$

Donde los m valores k_i son constantes conocidas para todo $i = 1, 2, \dots, m$; $\bar{x}_i$ es la media poblacional de la variable auxiliar i-ésima; y $\bar{x}_{i,s}$ es la media muestral de la variable auxiliar i-ésima para la misma muestra aleatoria simple sin reemplazamiento s, de tamaño n, seleccionada. Así, tenemos

$$\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{i,k}$$

Y

$$\bar{x}_{i,s} = \frac{1}{n}\sum_{k\in s} x_{i,k}$$

Siendo $x_{i,k}$ el valor de la variable auxiliar i-ésima en la unidad k de la población finita, es decir, con uno de los valores posibles de $k = 1, 2, \dots, N$. Sabemos que la esperanza matemática de la media muestral coincide con la media poblacional de la misma variable. Por tanto, $E(\bar{y}_s) = \bar{y}$, y también para todo $i = 1, 2, \dots, m$ tenemos que $E\left(\bar{x}_{i,s}\right) = \bar{x}_i$, haciendo uso de las propiedades del diseño de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo n.

2. Insesgación del estimador general

Ya que $\bar{x}_{i,s}$ es una media muestral, es un estimador insesgado de la media poblacional $\bar{x}_i$, por lo que tomado la esperanza matemática de t_u tenemos

$$E(t_u) = E\left[\bar{y}_s + \sum_{i=1}^{m} k_i\left(\bar{x}_i - \bar{x}_{i,s}\right)\right] = E(\bar{y}_s) + \sum_{i=1}^{m} k_i\left[\bar{x}_i - E\left(\bar{x}_{i,s}\right)\right] = \bar{y}$$

Debido a las propiedades de la esperanza matemática, ya que para todos los valores posibles de $i = 1, 2, \dots, m$, tanto k_i como $\bar{x}_i$ son constantes. En resumen, el estimador general t_u es insesgado para estimar la media poblacional de interés, con muestreo irrestricto aleatorio.

3. Varianza del estimador general

Haciendo uso de las propiedades de la varianza de una variable aleatoria, tenemos que

$$V(t_u) = V\left[\bar{y}_s + \sum_{i=1}^{m} k_i(\bar{x}_i - \bar{x}_{i,s})\right]$$

$$= V(\bar{y}_s) + \sum_{i=1}^{m} k_i^2 V(\bar{x}_{i,s}) - 2\sum_{i=1}^{m} k_i Cov(\bar{y}_s, \bar{x}_{i,s})$$

$$+ \sum_{i=1}^{m}\sum_{j\neq i}^{m} k_i k_j Cov(\bar{x}_{i,s}, \bar{x}_{j,s})$$

Aquí, en el último miembro, todo son constantes conocidas antes de proceder al muestreo y a la fase de estimación, salvo las funciones paramétricas $V(\bar{y}_s)$ y $Cov(\bar{y}_s, \bar{x}_{i,s})$, con $i = 1, 2, \dots, m$. Por esto, la varianza del estimador general t_u puede ser estimada sin sesgo del modo

$$\hat{V}(t_u) = \hat{V}(\bar{y}_s) + \sum_{i=1}^{m} k_i^2 V(\bar{x}_{i,s}) - 2\sum_{i=1}^{m} k_i \widehat{Cov}(\bar{y}_s, \bar{x}_{i,s})$$

$$+ \sum_{i=1}^{m}\sum_{j\neq i}^{m} k_i k_j Cov(\bar{x}_{i,s}, \bar{x}_{j,s})$$

Donde $\hat{V}(\bar{y}_s)$ y $\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s})$ son los estimadores insesgados respectivos uno a uno de las funciones paramétricas $V(\bar{y}_s)$ y $Cov(\bar{y}_s, \bar{x}_{i,s})$, de modo similar a como expliqué en el artículo reciente de Ruiz Espejo *et al.* (2013). A continuación vamos a obtener dichos estimadores insesgados en el muestreo irrestricto aleatorio de tamaño muestral efectivo n.

$$\hat{V}(\bar{y}_s) = \frac{N-n}{Nn}\widehat{(S_y^2)} = \frac{N-n}{Nn} s_y^2 = \frac{N-n}{Nn(n-1)} \sum_{k\in s} (y_k - \bar{y}_s)^2$$

Y

$$\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s}) = \frac{N-n}{Nn}\widehat{(S_{y,x_i})} = \frac{N-n}{Nn} s'_{y,x_i} = \frac{N-n}{(N-1)n^2} \sum_{k\in s} y_k(x_{i,k} - \bar{x}_i)$$

4. Estimador insesgado óptimo teórico

Hasta aquí hemos supuesto que los valores constantes k_i estaban fijados de antemano y eran conocidos para concretar el estimador insesgado t_u. Sin embargo, es posible estudiar qué valores concretos de k_i minimizan la varianza del estimador general insesgado multivariante t_u. Para ello, derivamos parcialmente la expresión de la varianza $V(t_u)$ con respecto a k_i, e igualándolas a cero obtenemos un sistema de m ecuaciones lineales con m incógnitas (que son las constantes óptimas $k_i = k_{i,ópt}$). En efecto, el sistema de ecuaciones lineales es el siguiente

$$\begin{cases} \dfrac{\partial V(t_u)}{\partial k_i} = 0 \\ i = 1, 2, \dots, m \end{cases}$$

Que resulta ser entonces

$$\begin{cases} k_i V(\bar{x}_{i,s}) + \displaystyle\sum_{j \neq i}^{m} k_j Cov(\bar{x}_{i,s}, \bar{x}_{j,s}) = Cov(\bar{y}_s, \bar{x}_{i,s}) \\ i = 1, 2, \dots, m \end{cases}$$

También se puede comprobar que

$$\frac{\partial^2 V(t_u)}{\partial k_i^2} = 2V(\bar{x}_{i,s})$$

Que es una constante positiva, salvo que la variable auxiliar i-ésima sea constante en todas las unidades de la población finita, en cuyo caso el término correspondiente a dicha variable auxiliar se anula en la fórmula del estimador t_u, por lo que su expresión se reduciría a una estimación basada en $m - 1$ variables auxiliares al eliminar aquélla en la que la variable auxiliar no aportara una información con alguna variabilidad.
Para $i \neq j$, tenemos que

$$\frac{\partial^2 V(t_u)}{\partial k_i \partial k_j} = 2Cov(\bar{x}_{i,s}, \bar{x}_{j,s})$$

Finalmente, las derivadas parciales de orden tres se anulan en todos los casos, por lo cual concluimos que se obtiene un mínimo global de la función real m-dimensional para ciertos valores $k_i = k_{i,ópt}$ que son óptimos y calculables teóricamente en cada caso concreto.

En el caso bidimensional es obvio, salvo casos triviales, que los valores críticos son los óptimos que minimizan la varianza del estimador t_u, ya que los menores principales de la matriz de covarianzas son positivos. Excluimos el caso trivial en que exista un coeficiente de correlación 1 ó -1 entre las medias muestrales de las dos variables auxiliares.

Veamos a continuación la solución óptima teórica en el caso de disponer de dos variables auxiliares con un coeficiente de correlación absoluto menor que 1.

5. Estimador insesgado bivariante óptimo

En el caso en que el número de variables auxiliares sea $m = 2$, tenemos que la solución concreta del sistema de ecuaciones lineales viene dada por estas fórmulas.

$$k_{1,\acute{o}pt} = \frac{V(\bar{x}_{2,s})Cov(\bar{y}_s, \bar{x}_{1,s}) - Cov(\bar{y}_s, \bar{x}_{2,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}$$

$$k_{2,\acute{o}pt} = \frac{V(\bar{x}_{1,s})Cov(\bar{y}_s, \bar{x}_{2,s}) - Cov(\bar{y}_s, \bar{x}_{1,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}$$

Que son constantes óptimas desconocidas, pues son funciones paramétricas que dependen de todos los valores de la variable de interés en las unidades de la población finita. Con estas constantes, si las conociéramos antes de realizar el muestreo y de observar en la muestra seleccionada la variable de interés, el estimador insesgado de regresión bivariante sería

$$t_u = \bar{y}_s + \sum_{i=1}^{2} k_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s})$$

Y alcanzaría su varianza el valor mínimo global con $(k_{1,\acute{o}pt}, k_{2,\acute{o}pt})$ entre todos los posibles valores del plano real para (k_1, k_2). Pero la realidad es que no conocemos estas constantes óptimas teóricas en un estudio concreto, por lo que cabe estimarlas sin sesgo sustituyendo, en el numerador de la expresión de cada una de dichas constantes óptimas, las funciones paramétricas $Cov(\bar{y}_s, \bar{x}_{i,s})$ por sus estimadores insesgados (al variar $i = 1, 2$) que obtenemos a continuación.

$$\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s}) = \frac{N-n}{Nn} s'_{y,x_i} = \frac{N-n}{(N-1)n^2} \sum_{k \in s} y_k(x_{i,k} - \bar{x}_i)$$

De ese modo, ya que los demás términos de $k_{i,\acute{o}pt}$ son constantes conocidas de antemano, obtenemos los valores óptimos estimados sin sesgo siguientes

$$\hat{k}_{1,\acute{o}pt} = \frac{V(\bar{x}_{2,s})\widehat{Cov}(\bar{y}_s, \bar{x}_{1,s}) - \widehat{Cov}(\bar{y}_s, \bar{x}_{2,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}$$

$$\hat{k}_{2,\acute{o}pt} = \frac{V(\bar{x}_{1,s})\widehat{Cov}(\bar{y}_s, \bar{x}_{2,s}) - \widehat{Cov}(\bar{y}_s, \bar{x}_{1,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}$$

Por todo ello, parece indicado partir del estimador

$$t' = \bar{y}_s + \sum_{i=1}^{2} \hat{k}_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s})$$

Este estimador es similar al que hemos estudiado como bivariante insesgado t_u al sustituir los valores k_i por los valores que estiman sus valores óptimos, es decir, por $\hat{k}_{i,\acute{o}pt}$. Pero como estos últimos estimadores no son constantes sino variables aleatorias, tienen un efecto en t' que lo hacen sesgado para estimar la media poblacional $\bar{y}$.

6. Estimación insesgada de la varianza óptima

El estimador bivariante óptimo teórico es

$$t_u = \bar{y}_s + \sum_{i=1}^{2} k_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s})$$

Tendría una varianza

$$V_{\acute{o}pt}(t_u) = V(\bar{y}_s) + \sum_{i=1}^{2} k_{i,\acute{o}pt}^2 V(\bar{x}_{i,s})$$

$$-2\sum_{i=1}^{2} k_{i,\acute{o}pt} Cov(\bar{y}_s, \bar{x}_{i,s}) + 2\, k_{1,\acute{o}pt} k_{2,\acute{o}pt} Cov(\bar{x}_{1,s}, \bar{x}_{2,s})$$

Por lo que esta varianza óptima teórica $V_{\acute{o}pt}(t_u)$ puede ser estimada sin sesgo a partir de las estimaciones insesgadas siguientes.

$$\hat{V}(\bar{y}_s) = \frac{N-n}{Nn} s_y^2 = \frac{N-n}{Nn(n-1)} \sum_{k\in s} (y_k - \bar{y}_s)^2$$

También

$$\left\{\left[\widehat{Cov(\bar{y}_s, \bar{x}_{\imath,s})}\right]^2\right\} = \left(\frac{N-n}{Nn}\right)^2 \widehat{(S^2_{y,x_\imath})}$$

$$= \left(\frac{N-n}{Nn}\right)^2 \left[(s'_{y,x_i})^2 - \hat{V}(s'_{y,x_i})\right]$$

Donde

$$s'_{y,x_i} = \frac{N}{(N-1)n} \sum_{k\in s} y_k (x_{i,k} - \bar{x}_i)$$

Y

$$\hat{V}(s'_{y,x_i}) = \frac{N^2}{(N-1)^2} \hat{V}\left[\frac{1}{n}\sum_{k\in s} y_k (x_{i,k} - \bar{x}_i)\right]$$

$$= \frac{N^2}{(N-1)^2}\frac{N-n}{Nn}\left[\widehat{S^2_{y(x_i-\bar{x}_i)}}\right] = \frac{N(N-n)}{(N-1)^2 n} s^2_{y(x_i-\bar{x}_i)}$$

$$= \frac{N(N-n)}{(N-1)^2 n(n-1)} \sum_{k\in s}\left[y_k\left(x_{i,k}-\bar{x}_i\right) - a_{1;y(x_i-\bar{x}_i)}\right]^2$$

Siendo

$$a_{1;y(x_i-\bar{x}_i)} = \frac{1}{n}\sum_{k\in s} y_k\left(x_{i,k}-\bar{x}_i\right)$$

Y también

$$\left[Cov\left(\bar{y}_s, \bar{x}_{1,s}\right)\widehat{Cov}\left(\bar{y}_s, \bar{x}_{2,s}\right)\right] = \left(\frac{N-n}{Nn}\right)^2 \left(S_{y,x_1}\widehat{S_{y,x_2}}\right)$$

$$= \left(\frac{N-n}{Nn}\right)^2 \left[s'_{y,x_1}s'_{y,x_2} - \widehat{Cov}\left(s'_{y,x_1}, s'_{y,x_2}\right)\right]$$

Donde

$$\widehat{Cov}\left(s'_{y,x_1}, s'_{y,x_2}\right) = \frac{N^2}{(N-1)^2}\widehat{Cov}\left[a_{1;y(x_1-\bar{x}_1)}, a_{1;y(x_2-\bar{x}_2)}\right]$$

$$= \frac{N^2}{(N-1)^2}\frac{N-n}{Nn}\left[\widehat{S_{y(x_1-\bar{x}_1),y(x_2-\bar{x}_2)}}\right]$$

$$= \frac{N(N-n)}{(N-1)^2 n} s_{y(x_1-\bar{x}_1),y(x_2-\bar{x}_2)}$$

$$= \frac{N(N-n)}{(N-1)^2 n(n-1)}$$

$$\times \sum_{k\in s}\left[y_k\left(x_{1,k}-\bar{x}_1\right) - a_{1;y(x_1-\bar{x}_1)}\right]\left[y_k\left(x_{2,k}-\bar{x}_2\right) - a_{1;y(x_2-\bar{x}_2)}\right]$$

El resto de la demostración es un ejercicio algebraico relativamente asequible.

7. Estimador de regresión multivariante corregido insesgado

El estimador que hemos estudiado en la sección anterior no es posible llevarlo a la práctica pues aunque tiene muy buenas propiedades teóricas depende de funciones paramétricas que son desconocidas y que deben ser estimadas sin sesgo. Así si sustituimos los valores óptimos $k_{i,\acute{o}pt}$ por sus estimadores insesgados $\hat{k}_{i,\acute{o}pt}$, el estimador resultante t' es sesgado, concretamente

$$t' = \bar{y}_s + \sum_{i=1}^{m} \hat{k}_{i,ópt}(\bar{x}_i - \bar{x}_{i,s})$$

Sin embargo, se puede corregir para que sea insesgado, del modo siguiente

$$t'_u = \bar{y}_s + \sum_{i=1}^{m} \hat{k}_{i,ópt}(\bar{x}_i - \bar{x}_{i,s}) - \sum_{i=1}^{m} \widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s})$$

Aquí $\widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s})$ es un estimador insesgado de la covarianza $Cov(\hat{k}_{i,ópt}, \bar{x}_{i,s})$, que más adelante pasaremos a concretar cómo obtenerlo para que sea útil en la práctica. Para demostrar que t'_u es insesgado nos basamos en que $\widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s})$ es un estimador insesgado de la esperanza matemática de $\hat{k}_{i,ópt}(\bar{x}_i - \bar{x}_{i,s})$. En concreto se puede ver que

$$\{E[\hat{k}_{\iota,ópt}\widehat{(\bar{x}_\iota - \bar{x}_{\iota,s})}]\} = [E(\hat{k}_{\iota,ópt})\widehat{E(\bar{x}_\iota - \bar{x}_{\iota,s})}] + \widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s})$$

$$= [E(\widehat{\hat{k}_{\iota,ópt}}) \times 0] + \widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s}) = \widehat{Cov}(\hat{k}_{i,ópt}, \bar{x}_{i,s})$$

Para calcular este último estimador, es un ejercicio asequible pero cuidadoso en el caso bivariante a partir de los estimadores insesgados necesarios siguientes.

$$\widehat{Cov}[\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s}), \bar{x}_{i,s}] = \widehat{Cov}\left(\frac{N-n}{Nn}\hat{S}_{y,x_i}, \bar{x}_{i,s}\right)$$

$$= \widehat{Cov}\left[\frac{N-n}{(N-1)n} a_{1;y(x_i-\bar{x}_i)}, \bar{x}_{i,s}\right] = \frac{N-n}{(N-1)n}\frac{N-n}{Nn}\left[\widehat{S^2_{y(x_\iota-\bar{x}_\iota)^2}}\right]$$

$$= \frac{(N-n)^2}{N(N-1)n^2} s^2_{y(x_i-\bar{x}_i)^2}$$

$$= \frac{(N-n)^2}{(N-1)^2 n^2 (n-1)} \sum_{k\in s} \left[y_k(x_{i,k} - \bar{x}_i)^2 - a_{1;y(x_i-\bar{x}_i)^2}\right]^2$$

Y de modo similar, en el caso bivariante,

$$\widehat{Cov}[\widehat{Cov}(\bar{y}_s, \bar{x}_{2,s}), \bar{x}_{1,s}] = \frac{(N-n)^2}{N(N-1)n^2}\hat{S}_{y(x_2-\bar{x}_2),x_1}$$

$$= \frac{(N-n)^2}{N(N-1)n^2} s'_{y(x_2-\bar{x}_2),x_1}$$

$$= \frac{(N-n)^2}{N(N-1)n^3} \sum_{k\in s} y_k(x_{2,k} - \bar{x}_2)(x_{1,k} - \bar{x}_1)$$

Etc.

De todo ello, y con razonamientos similares, es posible también estimar sin sesgo la varianza $V(t'_u)$, pero no lo detallamos en este artículo por su complejidad y laboriosidad de las fórmulas que resuelven este problema adicional.

8. Conclusiones

Hemos propuesto un estimador insesgado basado en m variables aleatorias auxiliares para estimar la media poblacional de interés en el muestreo aleatorio simple sin reemplazamiento de tamaño n, a partir de una población finita de tamaño N. El estimador propuesto aproxima al estimador de regresión multivariante óptimo teórico, ya que este no puede ser conocido pues requeriría tener el censo de la variable de interés, algo que haría innecesario estimar por muestreo la media poblacional ya que sería deducible del censo. Además indicamos que este estimador insesgado propuesto en el artículo admite un estimador insesgado de su varianza al menos para el caso de información auxiliar bivariante.

Finalmente indicamos que el método de análisis estadístico con el que hemos desarrollado esta teoría y práctica es también aplicable a otras clases de estimadores de la media poblacional en el muestreo aleatorio simple sin reemplazamiento. Su estudio será explicado en futuras aportaciones al muestreo y estimación insesgada tanto en la media poblacional como en la varianza del "estimador insesgado de la media poblacional".

Referencias

RUIZ ESPEJO, MARIANO (2013). «Exactitud de la Inferencia en Poblaciones Finitas». Madrid: Bubok.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2013; 2016). «Optimal unbiased estimation of some population central moments». *Metron* 71, 39-62; 74, 139.