Estudios de Economía Aplicada
Nº9, 1998 - Págs. 5-17

El estimador de regresión generalizado en el modelo de superpoblación: p-insesgadez asintótica y robustez

JOSÉ MIGUEL CASAS SANCHEZ
Universidad de Alcalá de Henares
MARTA GUIJARRO GARVI
Universidad de Cantabria

Esta nueva versión incluye todas las correcciones sugeridas por el evaluador, las cuales nos han parecido oportunas y por las que le quedamos muy agradecidos.

RESUMEN

Consideramos el modelo de superpoblación lineal múltiple y tomamos como estimador de la media poblacional el estimador de regresión generalizado. Estudiamos la p-insesgadez asintótica de este estimador bajo ciertas condiciones y vemos que, introduciendo condiciones algo más restrictivas que las anteriores, se llega a determinar métodos de selección de diseños de muestreo que, junto con el estimador de regresión generalizado, constituyen estrategias robustas frente a fallos de especificación de la matriz de variables explicativas.

Palabras clave: Modelo de Superpoblación, Estimador de Regresión Generalizado, P-insesgadez Asintótica, Estrategias Robustas.

Clasificación AMS: 62D05.

Artículo recibido en noviembre de 1997. Revisado en Febrero 1998.

1. Introducción

El planteamiento general de un modelo de muestreo consiste en considerar una población finita $U = \{1, \ldots, i, \ldots, N\}$ de tamaño N, en donde i representa la unidad i-ésima de la población, designando por s una muestra formada por un subconjunto de elementos de la población, por S el conjunto de todas las posibles muestras, s, y por r el conjunto de elementos de la población que no pertenecen a la muestra s.

Diremos que un diseño muestral es una función $p(\cdot)$ definida sobre S, tal que, $p(s) \geq 0, \quad \forall s \in S$ y $\sum_{s \in S} p(s) = 1$; su tamaño muestral lo representaremos por $\nu(s)$ y supondremos que es fijo, es decir, $\nu(s) = n$, cuando $p(s) \geq 0$.

Designaremos por $\pi_i = \sum_{s/i \in s} p(s)$ la probabilidad de que la unidad i-ésima de la población pertenezca a la muestra s. Asociada a ella definimos la variable I_i tal que:

$$I_i = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{si } i \notin s \end{cases}$$

Evidentemente,

$$\pi_i = p(I_i = 1) = 1 - p(I_i = 0)$$

Representaremos cada observación muestral por:

$$d = \{(i, y_i) : i \in s\}$$

donde s es una muestra fija e y_i es una valor desconocido correspondiente a una cierta característica que, junto con el vector $x_i = (x_{i1}, \ldots, x_{iq})'$ de dimensión $q \times 1$ de otra característica conocida, está asociado a la unidad i-ésima. Utilizaremos el vector $y = (y_i, \ldots, y_N)'$ para estimar la función paramétrica media poblacional:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

2. El modelo de superpoblación lineal múltiple

Supondremos ahora una población finita generada como muestra aleatoria de una superpoblación infinita, es decir, consideraremos que el vector de valores poblacionales $y = (y_1, \ldots, y_N)'$ es una realización del vector de variables aleatorias $Y = (Y_1, \ldots, Y_N)'$ cuya distribución conjunta la denotaremos por ξ.

Llamaremos modelo de superpoblación al conjunto específico de condiciones que definen una clase de distribuciones a la cual pertenece la distribución ξ. En nuestro caso admitiremos que el modelo de superpoblación lineal múltiple, ξ, viene dado por,

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E_\xi(Y) &= X\beta \\ E_\xi[(Y - X\beta)(Y - X\beta)'] &= \sigma^2 V \end{aligned}$$

donde la matriz $X = (x_1, \ldots, x_N)'$ es de rango completo q, β es un vector de parámetros desconocidos de dimensión $q \times 1$ y ε es un vector aleatorio de dimensión $N \times 1$ tal que,

$$E_\xi(\varepsilon) = 0$$

$$E_\xi(\varepsilon\varepsilon') = \sigma^2 V$$

siendo $E_\xi(\cdot)$ la esperanza respecto al modelo ξ, σ^2 una constante desconocida y V una matriz simétrica y definida positiva.

Consideraremos un estimador $e = e(D)$ como una función sobre el espacio muestral $\{D : Y \in R^N, s \in S\}$ donde $D = \{(i, Y_i) : i \in s\}$ es la variable dato y s es una muestra extraída de la población según un diseño de muestreo, p, con probabilidades de inclusión de primer orden $\pi_i \quad (i = 1, \ldots, N)$. Dicho estimador nos proporciona la información sobre la media poblacional $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

Designaremos por $\Pi = diag(\pi_1, \ldots, \pi_N)$, matriz de dimensión $N \times N$.

3. El estimador de regresión generalizado

Consideremos una sucesión de poblaciones finitas U_t de tamaño N_t, con $0 < N_1 < N_2 \ldots$, tal que $U_t \quad (t = 1, 2, \ldots)$ está formada por las N_t primeras unidades de una sucesión dada $\{i\}$, así pues, $U_2 \supset U_1$ contiene las N_2 primeras unidades de la sucesión $\{i\}$, etc.

Supondremos que la correspondiente sucesión de medias poblacionales, $\bar{Y}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_i$, es convergente.

Sea s_t una sucesión de muestras obtenidas a partir de una secuencia de diseños de tamaños efectivos fijos $n_t \quad (n_t < N_t, \forall t)$, tal que,

$$s_t = \{i : I_{it} = 1, 1 \leq i \leq n_t\}$$

en donde I_{it} es una variable aleatoria que vale 1 si la unidad i-ésima está en la muestra t-ésima y 0 en caso contrario.

Denotaremos por π_{it} la probabilidad de que la i-ésima unidad esté en la t-ésima muestra.

De manera análoga a la indicada anteriormente, diremos que, asociado a la unidad i-ésima de la población U_t, tendremos un número desconocido, y_i, realización de una variable aleatoria Y_i, y un vector conocido $(x_{i1}, \ldots, x_{iq})'$, de dimensión $q \times 1$. Así pues, admitimos que, para cada t, $Y_t = (Y_1, \ldots, Y_{N_t})'$ está relacionado con $X_t = (x_1, \ldots, x_{N_t})'$ a través de un modelo de superpoblación lineal múltiple ξ.

3.1. P-insesgadez asintótica

Diremos que una sucesión de estimadores, e_t, construídos a partir de una sucesión de poblaciones U_t, es asintóticamente insesgada según el diseño de muestreo p, o *asintóticamente p-insesgada*, para la media poblacional $\bar{Y}_t$, si se verifica que:

$$\lim_{t \to \infty} [E_p(e_t) - \bar{Y}_t] = 0$$

Utilizaremos como estimador de la media poblacional, $\bar{Y}_t$, el estimador de regresión generalizado. [1]

$$e_{RG} = N_t^{-1} \sum_{i=1}^{N_t} \left[\frac{I_{it} Y_i}{\pi_{it}} - \sum_{j=1}^{q} \hat{\beta}_{jt} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right]$$

siendo $\hat{\beta}_{jt}$ la componente j-ésima del estimador del vector parámetrico $\hat{\beta}_{s_t}$. [2]

A continuación presentamos un teorema en el cual se dan condiciones suficientes para que el estimador de regresión generalizado, e_{RG}, sea asintóticamente p-insesgado para la media poblacional.

[1] Pertenece a la clase de estimadores de regresión generalizados propuesta por Cassel, Särndal y Wretman.

[2] En general el estimador de regresión generalizado no será p-insesgado pues en su expresión intervienen los estimadores $\hat{\beta}_{jt}$ que dependen de la muestra.

Teorema 1

El estimador e_{RG} es asintóticamente p-insesgado si se verifican simultáneamente las siguientes condiciones:

C.1 $\lim_{t\to\infty} N_t^{-1} \sum_{i=1}^{N_t} x_{ij}^2 < \infty \quad j = 1, ..., q$

C.2 $\lim_{t\to\infty} E_p(\hat{\beta}_{jt})^2 < \infty \quad j = 1, ..., q$

C.3 $\lim_{t\to\infty} N_t \min_{1\leq i\leq N_t} \pi_{it} = \infty$

C.4 $\lim_{t\to\infty} \max_{i\neq k} \left| \frac{\pi_{ikt}}{\pi_{it}\pi_{kt}} - 1 \right| = 0$

con $\pi_{ikt} = p(I_{it} = I_{kt} = 1)$, *probabilidades de inclusión de segundo orden.*

Demostración

Definimos la siguiente variable aleatoria,

$$e_{RG}^* = N_t^{-1} \sum_{i=1}^{N_t} \left[\frac{I_{it}Y_i}{\pi_{it}} - \sum_{j=1}^{q} \beta_j \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right]$$

Sin más que tomar esperanzas se comprueba que $E_p(e_{RG}^*) = \bar{Y}_t$ para todo Y, es decir, $E_p(e_{RG}^*)$ es p-insesgado de $\bar{Y}_t$. Por tanto será suficiente demostrar que,

$$\lim_{t\to\infty} E_p(e_{RG} - e_{RG}^*) = 0$$

Pero,

$$\lim_{t\to\infty} E_p(e_{RG} - e_{RG}^*) = \sum_{j=1}^{q} \lim_{t\to\infty} E_p \left[\left(\beta_j - \hat{\beta}_{jt} \right) N_t^{-1} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right]$$

Basta comprobar que,

$$\lim_{t\to\infty} E_p \left| \left(\beta_j - \hat{\beta}_{jt} \right) N_t^{-1} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right| = 0 \qquad j = 1, ..., q$$

Aplicando la desigualdad de Schwartz,

$$E_p \left| \left(\beta_j - \hat{\beta}_{jt} \right) N_t^{-1} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right| \leq$$

$$\leq \quad \left[E_p(\beta_j - \hat{\beta}_{jt})^2 E_p \left(N_t^{-1} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right)^2 \right]^{1/2} \qquad j = 1, ..., q$$

Acotemos cada uno de los factores,

$$E_p\left[N_t^{-1}\sum_{i=1}^{N_t}\left(\frac{I_{it}}{\pi_{it}}-1\right)x_{ij}\right]^2 = \frac{1}{N_t^2}\sum_{i=1}^{N_t}x_{ij}^2\left(\frac{1}{\pi_{it}}-1\right)+ \\ + \frac{1}{N_t^2}\sum_{i=1}^{N_t}\sum_{i\neq k}x_{ij}x_{kj}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)$$

El primer sumando está acotado superiormente por,

$$N_t^{-1}\sum_{i=1}^{N_t}x_{ij}^2\frac{1}{N_t\min_{1\leq i\leq N_t}\pi_{it}}$$

que tiende a 0 cuando $t\to\infty$, aplicando C.1 y C.3.

El segundo sumando es menor o igual que,

$$\max_{i\neq k}\left|\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right|\left(\frac{1}{N_t}\sum_{i=1}^{N_t}|x_{ij}|\right)^2\leq\max_{i\neq k}\left|\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right|\frac{1}{N_t}\sum_{i=1}^{N_t}x_{ij}^2$$

que también converge a 0 cuando $t\to\infty$, por C.1 y C.4

Por último, por la condición C.2:

$$\lim_{t\to\infty}E_p(\beta_j-\hat{\beta}_{jt})^2<\infty \qquad j=1,...,q$$

c.q.d.

Este teorema mejora los resultados obtenidos por otros autores, pues para que el estimador de regresión generalizado sea asintóticamente p-insesgado, no es necesario que se cumpla la condición adicional $\lim_{t\to\infty}\frac{1}{N_t}\sum_{i=1}^{N_t}Y_i^2<\infty$, como se ha probado con el teorema anterior.[3]

3.2. Robustez

Para determinar una elección óptima del diseño de muestreo asociado al estimador de regresión generalizado, utilizaremos el criterio de minimizar el error cuadrático medio esperado,

[3]Los resultados obtenidos son independientes de que el modelo tenga o no errores de especificación.

$$E_p E_\xi (e_{RG} - Y_t)^2$$

Pero la utilización de este criterio exige ciertas especificaciones sobre la relación existente entre la variable aleatoria Y_i y el vector asociado x_i. Así pues, supondremos que:

$$E_\xi(Y_i) = \sum_{j=1}^{q} \beta_j x_{ij} + \sum_{j=1}^{m} \gamma_j z_{ij} \quad i = 1, \ldots, N_t \text{ [4]}$$

donde $Z_t = (z_{ij})$ es la matriz de regresores adicionales de dimensión $N \times m$ y γ es un vector de parámetros desconocidos de dimensión $m \times 1$.

Admitimos además que:

$$Var_\xi(Y_i) = \sigma^2 v_i \quad i = 1, \ldots, N_t$$

$$Cov_\xi(Y_i, Y_k) = \sigma^2 \rho (v_i, v_k)^{\frac{1}{2}} \quad i \neq k$$

donde $v_i > 0$ conocido, σ^2 constante conocida, y ρ valor verificando la condición:

$$-(N_t - 1)^{-1} \leq \rho < 1$$

Como pretendemos probar que el estimador es robusto, necesitaremos dar un método que nos permita obtener estrategias robustas frente a errores de especificación en la matriz de regresores, partiendo de un diseño de muestreo óptimo. [5] Para ello presentamos el siguiente teorema:

Teorema 2

Si se verifican simultáneamente las condiciones:

C.1 $\lim_{t \to \infty} N_t^{-1} \sum_{i=1}^{N_t} x_{ij}^2 < \infty \quad j = 1, \ldots, q$

C.5 Existe una constante K tal que para un t suficientemente grande,

$$n_t \sum_{j=1}^{q} E_\xi(\beta_j - \hat{\beta}_{jt})^2 < K < \infty$$

[4]Evidentemente, si no hubiera error de especificación no aparecería el segundo sumando $\sum_{j=1}^{m} \gamma_j z_{ij}$

[5]Llamaremos clase de diseños óptimos, y la denotaremos por P_0, a la formada por diseños, p_0, de tamaño efectivo fijo n que cumplan la condición de que sus probabilidades de inclusión de primer orden responden a la expresión: $\pi_{0i} = n v_i^{1/2} v^{-1}$.

C.6 $\lim_{t\to\infty} \frac{1}{N_t}\sum_{i=1}^{N_t} v_i < \infty$

C.7 $\lim_{t\to\infty} \frac{N_t}{n_t} \min_{1\leq i\leq N_t} \pi_{it} > 0$

C.8 $\lim_{t\to\infty} \frac{N_t^2}{n_t} \max_{i\neq k} \left|\pi_{ikt} - \pi_{it}\pi_{kt}\right| < \infty$ [6]

resulta que:

$$n_t E_p E_\xi (e_{RG} - \bar{Y}_t)^2 = A_t + B_t + C_t$$

con,

$$A_t = \frac{n_t}{N_t^2}\sigma^2 \left[\sum_{i=1}^{N_t} v_i \left(\frac{1}{\pi_{it}} - 1\right) + \rho \sum_{i=1}^{N_t}\sum_{i\neq k} (v_i v_k)^{1/2} \left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}} - 1\right)\right]$$

$$B_t = \frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\left(\frac{1}{\pi_{it}} - 1\right) b_{it}^2 + \frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\sum_{i\neq k}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}} - 1\right) b_{it} b_{kt}$$

$$b_{it} = \sum_{j=1}^{m} \gamma_j z_{ij} \qquad i = 1, ..., N_t$$

y

$$\lim_{t\to\infty} C_t = 0$$

siendo el estimador de regresión generalizado, e_{RG}, robusto frente a errores en la especificación de la matriz de diseño. [7]

Demostración

Trivialmente,

$$\begin{aligned} n_t E_p E_\xi (e_{RG} - \bar{Y}_t)^2 &= n_t E_p E_\xi (e_{RG} - e_{RG}^*)^2 + n_t E_p E_\xi (e_{RG}^* - \bar{Y}_t)^2 + \\ &+ 2 n_t E_p E_\xi (e_{RG} - e_{RG}^*)(e_{RG}^* - \bar{Y}_t) \end{aligned}$$

Acotemos cada uno de los sumandos:

$$n_t E_p E_\xi (e_{RG}^* - \bar{Y}_t)^2 = n_t E_p E_\xi \left[\frac{1}{N_t}\sum_{i=1}^{N_t}\left(\frac{I_{it}}{\pi_{it}} - 1\right)\left(Y_i - \sum_{j=1}^{q}\beta_j x_{ij}\right)\right]^2 =$$

$$= \frac{n_t}{N_t^2} E_p E_\xi \left[\sum_{i=1}^{N_t}\left(\frac{I_{it}}{\pi_{it}} - 1\right)^2 \left(Y_i - \sum_{j=1}^{q}\beta_j x_{ij}\right)^2\right] +$$

[6] Las condiciones C.7 y C.8 son más restrictivas que C.3 y C.4.

[7] Este teorema generaliza el dado por Robinson y Särndal para $\rho = 0$.

$$\frac{n_t}{N_t^2}E_pE_\xi\left[\sum_{i=1}^{N_t}\sum_{i\neq k}\left(\frac{I_{it}}{\pi_{it}}-1\right)\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\left(Y_i-\sum_{j=1}^{q}\beta_jx_{ij}\right)\left(Y_k-\sum_{j=1}^{q}\beta_jx_{kj}\right)\right]=$$

$$=\frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\left(\frac{1}{\pi_{it}}-1\right)Var_\xi(Y_i)+\frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\sum_{i\neq k}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)Cov_\xi(Y_i,Y_k)=$$

$$=\frac{n_t}{N_t^2}\sigma^2\sum_{i=1}^{N_t}\left(\frac{1}{\pi_{it}}-1\right)v_i+\frac{n_t}{N_t^2}\sigma^2\rho\sum_{i=1}^{N_t}\sum_{i\neq k}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)(v_iv_k)^{1/2}\leq$$

$$\leq\frac{n_t}{N_t}\frac{1}{\min_{1\leq i\leq N_t}\pi_{it}}\frac{\sigma^2}{N_t}\sum_{i=1}^{N_t}v_i+\frac{n_t}{N_t}\sigma^2|\rho|\max_{i\neq k}\left|1-\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}\right|\sum_{i=1}^{N_t}v_i\leq$$

$$\leq\frac{n_t}{N_t}\frac{1}{\min_{1\leq i\leq N_t}\pi_{it}}\frac{\sigma^2}{N_t}\sum_{i=1}^{N_t}v_i+$$

$$+\frac{N_t^2}{n_t}\max_{i\neq k}|\pi_{ikt}-\pi_{it}\pi_{kt}|\frac{n_t^2}{N_t^2(\min_{1\leq i\leq N_t}\pi_{it})^2}\frac{1}{N_t}\sigma^2|\rho|\sum_{i=1}^{N_t}v_i$$

Puede comprobarse que la expresión anterior es menor que ∞, aplicando las condiciones C.6-C.8.

Además,

$$e_{RG}-e^*_{RG}=\frac{1}{N_t}\textstyle\sum_{j=1}^{q}(\beta_j-\hat{\beta}_{jt})a_{jt}$$

con,

$$a_{jt}=\sum_{j=1}^{N_t}\left(\frac{I_{it}}{\pi_{it}}-1\right)x_{ij}\qquad j=1,\ldots,q$$

Por tanto,

$$n_t(e_{RG}-e^*_{RG})^2\leq\frac{n_t}{N_t^2}\sum_{j=1}^{q}(\beta_j-\hat{\beta}_{jt})^2\sum_{j=1}^{q}a_{jt}^2$$

Tomando esperanzas respecto al modelo ξ:

$$n_tE_\xi(e_{RG}-e^*_{RG})^2\leq\frac{1}{N_t^2}\sum_{j=1}^{q}a_{jt}^2n_t\sum_{j=1}^{q}E_\xi(\beta_j-\hat{\beta}_{jt})^2$$

Teniendo en cuenta la condición C.5, a partir de un t suficientemente grande, $n_t\sum_{j=1}^{q}E_\xi(\beta_j-\hat{\beta}_{jt})^2$ está acotado, con lo cual, tomando esperanzas respecto al diseño, y para dicho t, podremos escribir:

$$n_tE_pE_\xi(e_{RG}-e^*_{RG})^2\leq\frac{K}{N_t^2}\sum_{j=1}^{q}\left[\sum_{i=1}^{N_t}x_{ij}^2\left(\frac{1}{\pi_{it}}-1\right)+\sum_{i=1}^{N_t}\sum_{i\neq k}x_{ij}x_{kj}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)\right]$$

expresión que tiende a 0 cuando $t \to \infty$, por C.1, C.3 y C.4 como se demostró en el teorema anterior.

Se prueba de modo inmediato que:

$$2n_t E_p E_\xi |(e_{RG} - e^*_{RG})(e^*_{RG} - \bar{Y}_t)| \leq 2[n_t E_p E_\xi (e_{RG} - e^*_{RG})^2 n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2]^{1/2}$$

converge a 0 cuando $t \to \infty$.

Hemos demostrado, así, que:

$$n_t E_p E_\xi (e_{RG} - \bar{Y}_t)^2 = n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 + C_t$$

donde,

$$\lim_{t \to \infty} n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 < \infty$$

y

$$\lim_{t \to \infty} C_t = 0$$

Bastará comprobar entonces que:

$$n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 = A_t + B_t$$

Pero,

$$\begin{aligned} e^*_{RG} - \bar{Y}_t &= \frac{1}{N_t} \sum_{i=1}^{N_t} \left[Y_i \left(\frac{I_{it}}{\pi_{it}} - 1 \right) - \sum_{j=1}^{q} \beta_{jt} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) x_{ij} \right] = \\ &= \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \mu_{it}) \left(\frac{I_{it}}{\pi_{it}} - 1 \right) + \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) b_{it} = \\ &= a_t + b_t \end{aligned}$$

donde,

$$\mu_{it} = \sum_{j=1}^{q} \beta_j x_{ij} + \sum_{j=1}^{m} \gamma_t z_{ij}$$

$$a_t = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \mu_{it}) \left(\frac{I_{it}}{\pi_{it}} - 1 \right)$$

$$b_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\frac{I_{it}}{\pi_{it}} - 1 \right) b_{it}$$

Elevando al cuadrado y tomando esperanzas:

$$E_pE_\xi(c^*_{RG} - \bar{Y}_t)^2 = E_pE_\xi a_t^2 + E_pE_\xi b_t^2 + 2E_pE_\xi a_t b_t$$

Desarrollando cada sumando,

$$E_pE_\xi a_t^2 =$$

$$E_p\left[\frac{1}{N_t^2}\sigma^2\sum_{i=1}^{N_t} v_i\left(\frac{I_{it}}{\pi_{it}}-1\right)^2 + \frac{1}{N_t^2}\sigma^2\rho\sum_{i=1}^{N_t}\sum_{i\neq k}(v_iv_k)^{1/2}\left(\frac{I_{it}}{\pi_{it}}-1\right)\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\right] =$$

$$= \frac{1}{N_t}\sigma^2\left[\sum_{i=1}^{N_t} v_i\left(\frac{1}{\pi_{it}}-1\right) + \rho\sum_{i=1}^{N_t}\sum_{i\neq k}(v_iv_k)^{1/2}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)\right]$$

Por tanto,

$$n_tE_pE_\xi a_t^2 = A_t$$

Además:

$$n_tE_pE_\xi b_t^2 = \frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\left(\frac{1}{\pi_{it}}-1\right)b_{it}^2 + \frac{n_t}{N_t^2}\sum_{i=1}^{N_t}\sum_{i\neq k}\left(\frac{\pi_{ikt}}{\pi_{it}\pi_{kt}}-1\right)b_{it}b_{kt} = B_t$$

Por último:

$$E_pE_\xi a_tb_t = E_p\left\{\frac{1}{N_t^2}\sum_{i=1}^{N_t}\left(\frac{I_{it}}{\pi_{it}}-1\right)b_{it}E_\xi\left[\frac{1}{N_t}\sum_{i=1}^{N_t}(Y_i-\mu_{it})\left(\frac{I_{it}}{\pi_{it}}-1\right)\right]\right\} = 0$$

c.q.d.

En el supuesto de que no exista error en la especificación de la matriz de diseño, el término B_t se anula y el A_t se minimiza mediante un diseño de muestreo con probabilidades de inclusión $\pi_{0i} = nv_iv^{-1}$.[8]

En cambio, si el modelo considerado es falso, B_t no se anula, y no podríamos obtener una expresión simple para las probabilidades de inclusión asintóticamente óptimas.

Sin embargo, la contribución del valor de B_t a la varianza asintótica esperada puede reducirse mediante una adecuada elección del diseño de muestreo $p_o \in P_0$.En efecto:

[8] Guijarro, M (1991). El Modelo de Superpoblación: Estimaciones y estrategias óptimas.

Expresando B_t como,

$$B_t = n_t \frac{N_t^{-2}}{2} \sum_{i=1}^{N_t} \sum_{i \neq k} \left(\frac{b_{it}}{\pi_{it}} - \frac{b_{kt}}{\pi_{kt}} \right)^2 (\pi_{it}\pi_{kt} - \pi_{ikt})$$

Observamos que el término $\left(\frac{b_{it}}{\pi_{it}} - \frac{b_{kt}}{\pi_{kt}}\right)^2$ tenderá a ser mayor cuanto más grande sea la diferencia entre los elementos i-ésimo y k-ésimo, esto es, cuanto más "difieran" x_i' y x_k'. Por tanto, para minimizar B_t, elegiremos $p_0 \in P_0$ tal que $\pi_{it}\pi_{kt} - \pi_{ikt}$ sea negativa o nula para aquellos individuos tales que $|x_i'\beta - x_k'\beta|$ sea mayor.

Esta condición puede conseguirse, por ejemplo, estratificando la población en estratos *homogéneos* respecto de x y eligiendo de cada estrato una muestra independiente con probabilidades de inclusión verificando la condición $\pi_{0i} = nv_iv^{-1}$.

Por tanto, una aleatorización apropiada -mediante un proceso de estratificación- nos lleva a que el estimador e_{RG} es robusto frente a errores en la especificación de la matriz de diseño.

Referencias

AZORÍN, F. Y SÁNCHEZ-CRESPO, J. L. (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza.

CASSEL, C., SÄRNDAL, C. Y WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**,615-620.

CASSEL, C., SÄRNDAL, C. Y WRETMAN, J. H. (1977). *Foundations of Inference in Survey Sampling*, New York: John Wiley.

FULLER, W. A. E ISAKI, C. T. (1982). Survey design under the regression superpopulation model. *Journal of American Statistical Association* **77**, 89-96.

GODAMBE, V. P. (1982). Estimation in survey sampling: robustness and optimality. *Journal of American Statistical Association* **77** 393-406.

GUIJARRO, M. (1991). *El modelo de superpoblación: estimaciones y estrategias óptimas.* Tesis Doctoral. Universidad de Alcalá de Henares.

HERSON, J. Y ROYALL, R. M. (1973). Robust estimation in finite populations. *Journal of American Statistical Association* **68**, 880-893.

KALTON, G. (1983). Models in the practice of survey sampling. *International Statistical Review* **51**, 175-188.

ROBINSON, P. M. Y SÄRNDAL, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankyā*. Ser. B, **45**, 240-248.

ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377-387.

TAM, S. M. (1984). Optimal estimation in survey sampling under a regression superpopulation model. *Biometrika* **71**, 645-647.

TAM, S. M. (1988b). Some results on robust estimation in finite population sampling. *Journal of American Statistical Association* **83**, 242-248.