QÜESTIIÓ, Vol 18, 2 pp. 173-188, 1994

P–INSESGADEZ ASINTÓTICA Y ROBUSTEZ EN LA ESTIMACIÓN LINEAL CON MODELOS DE SUPERPOBLACIONES: UN CRITERIO DE SELECCIÓN

JOSÉ MIGUEL CASAS SÁNCHEZ* y MARTA GUIJARRO GARVI†

Se analiza la estimación lineal de la media poblacional desde el punto de vista de los modelos de superpoblaciones con parámetros desconocidos y correlación no nula. La posible existencia de errores de especificación en el modelo determina el estudio de propiedades tales como la insesgadez asintótica respecto al diseño de muestreo y la robustez débil, deseables para asegurar la robustez de los estimadores frente a este tipo de errores. Se establece, así, un criterio de selección entre estimadores lineales asintóticamente insesgados respecto al diseño y débilmente robustos.

Asymptotic design unbiasedness and robustness of linear estimation under superpopulation models: a selection procedure.

Key words: Modelo de superpoblación; Estimación lineal; P–insesgadez asintótica; Robustez débil.

Clasificación AMS: 62D05.

* José Miguel Casas Sánchez. Universidad de Alcalá de Henares.

† Marta Guijarro Garvi. Universidad de Cantabria.

-Article rebut el setembre de 1993.

-Acceptat el març de 1994.

1. INTRODUCCIÓN

En la estimación de la media poblacional bajo la suposición de un modelo de superpoblación que ligue la variable de interés a una o más variables auxiliares, resulta obligado considerar estimadores robustos frente a fallos en la formulación del modelo de trabajo. Si los parámetros del modelo son desconocidos (hecho, por otro lado, habitual), la estimación de dichos parámetros lleva consigo la dificultad para obtener estimadores de la media poblacional que sean insesgados según el diseño de muestreo. En este sentido, adoptaremos el punto de vista clásico, defendido por Hansen, Madow y Tepping (1983), exigiendo la insesgadez asintótica respecto del diseño de muestreo, o $p-$insesgadez asintótica, de los estimadores, como primera garantía de robustez; este hecho requiere la definición de un marco asintótico.

Sin embargo, aunque la $p-$insesgadez asintótica es necesaria, no es suficiente para que una estrategia de muestreo sea óptima en el sentido de minimizar el error cuadrático medio esperado, criterio que utilizaremos para medir la calidad de un estimador (Godambe, 1955; Särndal, 1980). Por ello, Tam (1988) considera otras propiedades de interés, tales como la insesgadez respecto al modelo y la robustez débil, o robustez frente a un error de especificación en la matriz de covarianzas del modelo.

En este trabajo, generalizamos la definición de estimador débilmente robusto debida a Tam (1988), con objeto de utilizarla en contextos con correlación no nula, considerando, únicamente, estimadores lineales.

El resultado que presentamos sugiere que, dados dos estimadores lineales asintóticamente $p-$insesgados y débilmente robustos, deberíamos elegir aquel con menor sesgo según el modelo porque, asintóticamente, es el de menor error cuadrático medio esperado. La extensión del resultado obtenido por Tam (1988) a un contexto con correlación conlleva, sin embargo, la necesidad de exigir una serie de hipótesis adicionales.

2. DESCRIPCIÓN DEL MODELO

Sea $\{i\}$ una sucesión de elementos, estando, cada uno de ellos, asociado a un número desconocido, y_i, y a un vector conocido, x_i, de dimensión $q \times 1$. Adoptando el marco de trabajo de Isaki y Fuller (1982), definimos una sucesión de poblaciones finitas, $\{U_t\}$, de tamaño N_t con $0 < N_1 < \ldots$, tal que U_1 está

formada por las N_1 primeras unidades de $\{i\}$, $U_2 \supset U_1$ contiene los N_2 primeros elementos de $\{i\}$, etc. Así, U_t se expande haciendo tender N_t a infinito, cuando t tiende a infinito.

Para cada población, U_t, consideraremos que $y_t = (y_1, \ldots, y_{N_t})'$ es una realización del vector aleatorio $Y_t = (Y_1, \ldots, Y_{N_t})'$ relacionado con la matriz $X_t = (x_1, \ldots, x_{N_t})'$ a través del modelo de superpoblación ξ dado por

$$E_\xi(Y_t) = X_t\beta$$
$$E_\xi[(Y_t - X_t\beta)(Y_t - X_t\beta)'] = \sigma^2 V_t$$

donde $E_\xi(\cdot)$ denota la esperanza respecto al modelo ξ, β es un vector de dimensión $q \times 1$ desconocido, σ^2 constante conocida y $V_t = (v_{ik})$, matriz simétrica y definida positiva con la siguiente estructura:

$$v_{ik} = \begin{cases} v_i & i = k \\ \rho(v_i v_k)^{1/2} & i \neq k \end{cases}$$

con $v_i > 0$ conocido ($i = 1, \ldots, N_t$) y ρ conocido verificando la condición $-(N_t - 1)^{-1} \leq \rho < 1$. Supondremos que X_t es de rango completo q.

Sea $\{s_t\}$ una sucesión de muestras obtenidas a partir de $\{U_t\}$ mediante una secuencia de diseños de tamaño efectivo fijo $\{n_t\}$, de modo que s_1 está formada por n_1 elementos distintos de U_1, s_2 está formada por n_2 elementos distintos de U_2, etc.; $n_1 < n_2 < \ldots$ y $n_t < N_t$ $\forall t$. El hecho de que también n_t tiende a infinito está implícito en la condición C.3 de la siguiente sección. Destaquemos la no exigencia de que n_t crezca con la misma rapidez que N_t.

Llamaremos π_{it} a la probabilidad de que la i–ésima unidad esté incluida en la t–ésima muestra, e I_{it} a la variable aleatoria que vale 1 si la unidad i–ésima está en la muestra t–ésima y 0 en caso contrario.

Sin pérdida de generalidad listaremos primero las unidades de la muestra, realizando, así, las siguientes particiones:

$$Y_t = (Y_{s_t}', Y_{r_t}')'$$
$$X_t = (X_{s_t}', X_{r_t}')'$$
$$\Pi_t = \operatorname{diag}(\pi_{1t}, \ldots, \pi_{N_t t}) = \begin{pmatrix} \Pi_{s_t} & 0 \\ 0 & \Pi_{r_t} \end{pmatrix}$$
$$1_t = (1_{s_t}', 1_{r_t}')'$$

donde 1_t es el vector de dimensión $N_t \times 1$ y s_t y r_t indican el conjunto de unidades de la población que están y no están en la muestra, respectivamente.

Denotaremos por $\{e_t\}$, una sucesión de estimadores lineales homogéneos, es decir,

$$e_t = L_t' Y_{s_t} = N_t^{-1}(l_{1s_t}, \ldots, l_{n_t s_t}) Y_{s_t}$$

Definición 1

Diremos que una sucesión de estimadores lineales, $\{e_t\}$, construida a partir de la sucesión de poblaciones, es asintóticamente insesgada según un diseño, p, o asintóticamente $p-$insesgada para $\bar{Y}_t$ si

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

donde $E_p(\cdot)$ es la esperanza respecto al diseño de muestreo p e $\bar{Y}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_i$.

Por simplicidad, hablaremos de estimadores asintóticamente insesgados según el diseño o asintóticamente $p-$insesgados.

Definición 2

Un estimador, e_t, es insesgado respecto al modelo, o $\xi-$insesgado, si

$$E_\xi(e_t - \bar{Y}_t) = 0 \qquad \forall s.$$

Admitiremos que se cumplen las siguientes condiciones generales, habituales en la literatura del muestreo en superpoblaciones (Robinson y Särndal, 1983; Wright, 1983).

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

C.2 $\limsup_{t\to\infty} E_p(\hat{\beta}_{jt})^2 \qquad j = 1, \ldots, q$

C.3 $\liminf_{t\to\infty} \frac{N_t}{n_t} \min_{1\le i\le N_t} \pi_{it} > 0$

C.4 $\limsup_{t\to\infty} \frac{N_t^2}{n_t} \max_{i\neq k} |\pi_{ikt} - \pi_{it}\pi_{kt}| < \infty$ donde $\pi_{ikt} = p(I_{it} = I_{kt} = 1)$ son las probilidades de inclusión de segundo orden

C.5 $\limsup_{t\to\infty} \frac{1}{N_t} \sum_{i=1}^{N_t} v_i < \infty$

C.6 Dado un diseño de muestreo p, existe una constante k tal que para un t suficientemente grande,

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

C.7 $\limsup_{t\to\infty} N_t^{-1} \sum_{i=1}^{N_t} Y_i < \infty$

El estimador de regresión generalizado[1]:

$$e_{RG}(Q) = \frac{1}{N_t} 1'_{s_t} \Pi_{s_t}^{-1} Y_{s_t} + \frac{1}{N_t}(1'X_t - 1'_{s_t}\Pi_{s_t}^{-1} X_{s_t})\hat{\beta}(Q_{s_t})$$

con $\hat{\beta}(Q_{s_t}) = (X'_{s_t} Q_{s_t} X_{s_t})^{-1} X'_{s_t} Q_{s_t} Y_{s_t}$ y Q_{s_t}, matriz simétrica y definida positiva, es asintóticamente $p-$insesgado bajo las condiciones C.1-C.4 (Casas y Guijarro, 1993)[2].

Con objeto de simplificar notaciones, prescindiremos, en ocasiones, del subíndice t.

Definición 3

Un estimador lineal, e, de $\bar{Y}$ es robusto frente a un error en la especificación de la matriz de covarianzas, o débilmente robusto, si, para cada s con $p(s) > 0$, se cumple:

$$Q_s^{-1}\left(L - \frac{1}{N}\Pi_{0s}^{-1} 1_s\right) \in C(X_s)$$

para alguna matriz Q_s, simétrica y definida positiva. $C(X_s)$ denota el espacio generado por las columnas de X_s y Π_{0s}=diag $(\pi_{01}, \ldots, \pi_{0n}) = n\left(\sum_{i=1}^{N} v_i^{1/2}\right)^{-1}$ diag$(v_1^{1/2}, \ldots, v_n^{1/2})$.

Tam (1988) demuestra que una estrategia de muestreo, es decir, un par (e, p_0) donde e es un estimador lineal de $\bar{Y}$ y p_0 un diseño de muestreo con probabilidades de inclusión $\pi_{0i} = n\left(\sum_{i=1}^{N} v^{1/2}\right)^{-1} v_i^{1/2}$ $(\forall i)$, es óptima, en el sentido de minimizar el error cuadrático medio esperado, si el estimador es $\xi-$insesgado y

[1] Elemento de la *clase de estimadores QR* (Wright, 1983):

$$e_{QR} = \frac{1}{N} 1'_s R_s Y_s + \frac{1}{N}(1'X - 1'_s R_s X_s)\hat{\beta}(Q_s)$$

para $R_s = \Pi_s^{-1}$.

[2] De hecho Casas y Guijarro prueban la $p-$insesgadez asintótica del estimador de regresión generalizado bajo las condiciones C.1, C.2 y

$$\liminf_{t\to\infty} \frac{N_t}{n_t} \min_{1\le i\le N_t} \pi_{it} > 0$$

$$\liminf_{t\to\infty} \left| \frac{\pi_{ikt}}{\pi_{it}\pi_{kt}} - 1 \right|$$

menos restrictivas que las condiciones C.3 y C.4.

débilmente robusto. Este hecho justifica, por sí solo, la introducción del concepto de robustez débil.

3. ESTIMADORES LINEALES p–INSESGADOS Y DÉBILMENTE ROBUSTOS

En el siguiente resultado probamos que, dados dos estimadores lineales asintóticamente p–insesgados y débilmente robustos, aquel cuyo ξ–sesgo sea menor tiene, también, menor error cuadrático medio esperado.

Teorema

Sean $L'_m Y_s$ $(m = 1, 2)$ dos estimadores lineales de $\bar{Y}$, asintóticamente p–insesgados y débilmente robustos, verificando las hipótesis siguientes:

H.1 $\limsup_{t\to\infty} \frac{n_t}{N_t} \max_{1\le i\le N_t} E_p(l^2_{m_{is}} I_{it}) < \infty$

H.2 $\limsup_{t\to\infty} \max_{1\le i\le N_t} |E_p(l_{m_{is}} I_{it}) - 1| < \infty$

H.3 $\limsup_{t\to\infty} n_t \max_{i\neq k} |E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - 1| < \infty$

con $L'_m = N_t^{-1}(l_{m_{1s}}, \ldots, l_{m_{ns}})$.[3]

Si el ξ–sesgo de $L'_1 Y_s$ es mayor que el de $L'_2 Y_s$ (ambos en valor absoluto), entonces

$$\lim_{t\to\infty} n_t [E_p E_\xi (L'_1 Y_s - \bar{Y}_t)^2 - E_p E_\xi (L'_2 Y_s - \bar{Y}_t)^2] \ge 0$$

es decir, se verifica asintóticamente:

$$E_p E_\xi (L'_1 Y_s - \bar{Y}_t)^2 \ge E_p E_\xi (L'_2 Y_s - \bar{Y}_t)^2$$

para un diseño de muestreo dado p.

Demostración

El hecho de que los estimadores considerados sean débilmente robustos nos permite la siguientes descomposición:

[3] Deberíamos escribir $L'_{m_t} Y_{s_t} = N_t^{-1}(l_{m_{1s_t}}, \ldots, l_{m_{n_t s_t}}) Y_{s_t}$, estimador asintóticamente p–insesgado de $\bar{Y}_t$; mantendremos, sin embargo, la notación inicial a fin de simplificar las expresiones.

$$L'_m Y_s = e_{RG} + \left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) \hat{\beta}$$

para alguna matriz Q, simétrica y definida positiva. [4]

Así, podemos escribir

$$n_t E_p E_\xi (L'_m Y_s - \bar{Y}_t)^2 = n_t E_p E_\xi \left[e^*_{RG} + \left(L' X_s - \frac{1'_t X_t}{N_t} \right) \beta - \bar{Y}_t \right]^2 +$$

$$n_t E_p E_\xi \left[e_{RG} - e^*_{RG} - \left(L'_m X_s - \frac{1' X}{N_t} \right) (\beta - \hat{\beta}) \right]^2 +$$

$$+2 n_t E_p E_\xi \left[e^*_{RG} + \left(L'_m X_s - \frac{1' X}{N_t} \right) \beta - \bar{Y}_t \right] \cdot$$

$$\cdot \left[e_{RG} - e^*_{RG} - \left(L'_m X_s - \frac{1' X}{N_t} \right) (\beta - \hat{\beta}) \right]$$

donde e^*_{RG} es la variable aleatoria que resulta de sustituir $\hat{\beta}$ en e_{RG} por β, parámetro desconocido.

Acotemos cada uno de los sumandos:

$$e^*_{RG} + \left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) \beta - \bar{Y}_t = e^*_{RG} - \bar{Y}_t + \frac{1}{N_t} \sum_{j=1}^{q} \beta_j \sum_{i=1}^{N_t} (l_{m_{is}} I_{it} - 1) x_{ij}$$

con β_j componente j–ésima del vector β.

Elevando al cuadrado y tomando esperanzas tendremos que

$$n_t E_p E_\xi \left[e^*_{RG} + \left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) \beta - \bar{Y}_t \right]^2$$

puede expresarse como

$$n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 + \frac{n_t}{N_t^2} E_p \left(\sum_{j=1}^{q} \beta_j c_{m_{jt}} \right)^2$$

con

$$c_{m_{jt}} = \sum_{i=1}^{N_t} (l_{is} I_{it} - 1) x_{ij}$$

[4] Para dar mayor fluidez a las notaciones escribiremos e_{RG} y $\hat{\beta}$ en vez de $e_{RG}(Q)$ y $\hat{\beta}(Q_s)$.

Bajo las condiciones C.3-C.5:

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

cuando $t \to \infty$ (Casas y Guijarro, 1993).

Además,

$$\frac{n_t}{N_t^2} E_p \left(\sum_{j=1}^{q} \beta_j c_{m_{jt}} \right)^2 \leq \frac{n_t}{N_t^2} \sum_{j=1}^{q} \beta_j^2 \sum_{j=1}^{q} E_p(c^2_{m_{jt}})$$

Pero,

$$\frac{n_t}{N_t^2} E_p(c^2_{m_{jt}}) = \frac{n_t}{N_t^2} E_p \left[\sum_{i=1}^{N_t} (l_{m_{is}} I_{it} - 1) x_{ij} \right]^2 =$$

$$= \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} E_p(l_{m_{is}} I_{it} - 1)^2 x_{ij}^2 + \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} E_p(l_{m_{is}} I_{it} - 1)(l_{m_{ks}} I_{kt} - 1) x_{ij} x_{kj}$$

Desarrollando cada sumando:

$$\frac{n_t}{N_t^2} \sum_{i=1}^{N_t} E_p(l_{m_{is}} I_{it} - 1)^2 x_{ij}^2 \leq \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} E_p(l^2_{m_{is}} I_{it}) x_{ij}^2 +$$

$$+2 \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} [1 - E_p(l_{m_{is}} I_{it})] x_{ij}^2 \leq$$

$$\leq \frac{n_t}{N_t} \max_{1 \leq i \leq N_t} E_p(l^2_{m_{is}} I_{it}) \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}^2 \right) +$$

$$+2 \frac{n_t}{N_t} \max_{1 \leq i \leq N_t} |E_p(l_{m_{is}} I_{it}) - 1| \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}^2 \right) < \infty$$

cuando $t \to \infty$ por las condiciones H.1, H.2 y C.1.

Con respecto al segundo sumando:

$$\frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} E_p(l_{m_{is}} I_{it} - 1)(l_{m_{ks}} I_{kt} - 1) x_{ij} x_{kj} =$$

$$= \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} [E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - E_p(l_{m_{is}} I_{it}) - E_p(l_{m_{ks}} I_{kt}) + 1] x_{ij} x_{kj} =$$

$$= \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} [E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - 1] x_{ij} x_{kj} +$$

$$+ \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} [1 - E_p(l_{m_{is}} I_{it})] x_{ij} x_{kj} +$$

$$+ \frac{n_t}{N_t^2} \sum_{i=1}^{N_t} \sum_{i \neq k} [1 - E_p(l_{m_{ks}} I_{kt})] x_{ij} x_{kj} \leq$$

$$\leq n_t \max_{i \neq k} |E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - 1| \frac{1}{N_t^2} \left(\sum_{i=1}^{N_t} |x_{ij}| \right)^2 +$$

$$+ 2n_t \max_{1 \leq i \leq N_t} |1 - E_p(l_{m_{is}} I_{it})| \frac{1}{N_t^2} \left(\sum_{i=1}^{N_t} |x_{ij}| \right)^2 \leq$$

$$\leq n_t \max_{i \neq k} |E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - 1| \frac{1}{N_t} \left(\sum_{i=1}^{N_t} x_{ij}^2 \right) +$$

$$+ 2n_t \max_{1 \leq i \leq N_t} |1 - E_p(l_{m_{is}} I_{it})| \frac{1}{N_t} \left(\sum_{i=1}^{N_t} x_{ij}^2 \right) < \infty$$

cuando t tiende a infinito, aplicando H.2 y H.3.

Acotemos ahora la expresión,

$$n_t E_p E_\xi \left[\epsilon_{RG} - \epsilon^*_{RG} - \left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) (\beta - \hat{\beta}) \right]^2 =$$

$$= n_t E_p E_\xi (\epsilon_{RG} - \epsilon^*_{RG})^2 +$$

$$n_t E_p E_\xi \left[\left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) (\beta - \hat{\beta}) \right]^2 +$$

$$+ 2n_t E_p E_\xi (\epsilon_{RG} - \epsilon^*_{RG}) \left[\left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) (\beta - \hat{\beta}) \right]$$

Teniendo en cuenta que

$$\left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) (\beta - \hat{\beta}) = \frac{1}{N_t} \sum_{j=1}^{q} (\beta_j - \hat{\beta}_j) \sum_{i=1}^{N_t} (l_{m_{is}} I_{it} - 1) x_{ij}$$

Elevando al cuadrado y tomando esperanzas respecto al modelo resultará:

$$\frac{n_t}{N_t^2} E_\xi \left[\sum_{j=1}^{q} (\beta_j - \hat{\beta}_j) c_{m_{jt}} \right]^2 \leq \frac{n_t}{N_t^2} \sum_{j=1}^{q} E_\xi (\beta_j - \hat{\beta}_j)^2 \sum_{j=1}^{q} c_{m_{jt}}^2$$

Por la condición C.6, dado un diseño p, existe una constante K tal que, a partir de un t suficientemente grande:

$$\begin{aligned}
\frac{K}{N_t^2} \sum_{j=1}^{q} E_p(c_{m_{jt}}^2) &= \frac{K}{N_t^2} \sum_{j=1}^{q} E_p \left[\sum_{i=1}^{N_t} (l_{m_{is}} I_{it} - 1)^2 x_{ij}^2 \right] + \\
&+ \frac{K}{N_t^2} \sum_{j=1}^{q} E_p \left[\sum_{i=1}^{N_t} \sum_{i \neq k} (l_{m_{is}} I_{it} - 1)(l_{m_{ks}} I_{kt} - 1) x_{ij} x_{kj} \right] \leq \\
&\leq \sum_{j=1}^{q} \left[\frac{K}{n_t} \frac{n_t}{N_t} \max_{1 \leq i \leq N_t} E_p(l_{m_{is}}^2 I_{it}) \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}^2 \right) \right] + \\
&+ \sum_{j=1}^{q} \left[2 \frac{K}{N_t} \max_{1 \leq i \leq N_t} |E_p(l_{m_{is}} I_{it}) - 1| \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}^2 \right) \right] + \\
&+ k \sum_{j=1}^{q} \left[\max_{i \neq k} |E_p(l_{m_{is}} l_{m_{ks}} I_{it} I_{kt}) - 1| \left(\frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}^2 \right) \right] + \\
&+ k 2 \frac{1}{N_t} \sum_{j=1}^{q} \left(\max_{1 \leq i \leq N_t} |E_p(l_{m_{is}} I_{it}) - 1| \sum_{i=1}^{N_t} x_{ij}^2 \right)
\end{aligned}$$

Expresión que tiende a 0 cuando $t \to \infty$ por las condiciones H.1 y H.3.

Además, las condiciones C.1, C.3, C.4 y C.6 permiten asegurar que

$$\lim_{t \to \infty} n_t E_p E_\xi (e_{RG} - e_{RG}^*)^2 = 0$$

Por último, por la desigualdad de Schwartz se demuestra que

$$2 n_t E_p E_\xi (e_{RG} - e_{RG}^*) \left[\left(L_m' X_s - \frac{1_t' X_t}{N_t} \right) (\beta - \hat{\beta}) \right]$$

y

$$\begin{aligned}
&2 n_t E_p E_\xi \left[e_{RG}^* + \left(L_m' X_s - \frac{1_t' X_t}{N_t} \right) \beta - \bar{Y}_t \right] \left[e_{RG} - e_{RG}^* - \right. \\
&- \left. \left(L_m' X_s - \frac{1_t' X_t}{N_t} \right) (\beta - \hat{\beta}) \right]
\end{aligned}$$

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

De todo lo visto se deduce que

$$n_t E_p E_\xi (L'_m Y_s - \bar{Y}_t)^2 = n_t E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 + n_t E_p \left[\left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) \beta \right]^2 + O_t$$

con $\lim_{t \to \infty} O_t = 0$, es decir, asintóticamente,

$$E_p E_\xi (L'_m Y_s - \bar{Y}_t)^2 = E_p E_\xi (e^*_{RG} - \bar{Y}_t)^2 + E_p \left[\left(L'_m X_s - \frac{1'_t X_t}{N_t} \right) \beta \right]^2$$

Por tanto, si

$$\left| \left(L'_1 X_s - \frac{1'_t X_t}{N_t} \right) \beta \right| \geq \left| \left(L'_2 X_s - \frac{1'_t X_t}{N_t} \right) \beta \right|$$

esto es, si el ξ–sesgo de $L'_1 Y_s$ es mayor que el de $L'_2 Y_s$ (ambos en valor absoluto), resulta

$$E_p \left[\left(L'_1 X_s - \frac{1'_t X_t}{N_t} \right) \beta \right]^2 \geq E_p \left[\left(L'_2 X_s - \frac{1'_t X_t}{N_t} \right) \beta \right]^2$$

ya que el valor medio esperado respecto al diseño del cuadrado del ξ–sesgo es una función creciente. Se concluye, así, la prueba del teorema.

4. APLICACIÓN A UN MODELO ESPECÍFICO

Sea el modelo

$$\begin{aligned} &E_\xi(Y_i) = \beta_0 + \beta_1 x_i \\ &E_\xi[(Y_i - \beta_0 - \beta_1 x_i)^2] = \sigma^2 x_i \\ &E_\xi[(Y_i - \beta_0 - \beta_1 x_i)(Y_k - \beta_0 - \beta_1 x_k)] = \sigma^2 \rho (x_i x_k)^{1/2} \end{aligned}$$

con $\beta_0 \neq 0$, $\beta_1 \neq 0$ y σ constantes desconocidas y $x_i > 0$ para $i = 1, \ldots, N$.

Consideremos las estrategias de muestreo $(L'_1 Y_s, p_0)$ y $(L'_2 Y_s, p_0)$, donde

$$L'_1 Y_s = \bar{x} \frac{\sum_{i \in s} Y_i / \pi_i}{\sum_{i \in s} x_i / \pi_i}$$

es el estimador de razón generalizado (Brewer, 1963).

$$L'_2 Y_s = \frac{1}{N} \sum_{i \in s} \frac{Y_i}{\pi_i}$$

es el estimador de Horvitz-Thompson y, por último, p_0 es un diseño de muestreo con probabilidades de inclusión

$$\pi_{i0} = \frac{n x_i^{1/2}}{\sum_{i=1}^{N} x_i^{1/2}}$$

Se demuestra de modo sencillo que $L_1' Y_s$ es asintóticamente insesgado según el diseño de muestreo p_0 (Särndal, 1980). El estimador de Horvitz-Thompson, como ya es sabido, es insesgado para cualquier diseño de muestreo.

Las condiciones C.3 y C.4, junto con la aplicación del teorema de Slutsky (Särndal, 1980), mediante el cual sustituimos en el límite una función de las medias muestrales por la misma función de sus valores esperados, nos permiten demostrar que $L_1' Y_s$ y $L_2' Y_s$ cumplen las condiciones H.1-H.3.

El estimador $L_1' Y_s$ es débilmente robusto, sin más que considerar la matriz Q_s=diag $(x_1^{1.5}, \ldots, x_n^{1.5})$ (Tam, 1988). Además, como

$$L_2' Y_s = \frac{1}{N} 1_s' \Pi_s^{-1} Y_s$$

basta tomar el estimador de Horvitz-Thompson con el diseño de muestreo óptimo p_0 para que, trivialmente, se cumpla la definición de estimador débilmente robusto.

Estamos, pues, en condiciones de aplicar el resultado que nos sugiere la elección del estimador con el menor ξ–sesgo. Sencillas operaciones nos conducen a

$$E_\xi(L_1' Y_s - \bar{Y}) = \left[\bar{x} \left(\sum_s x_i^{-1/2} \right) \left(\sum_s x_i^{1/2} \right)^{-1} - 1 \right] \beta_0$$

y

$$\begin{aligned} E_\xi(L_2' Y_s - \bar{Y}) &= \left[\frac{\sum_{i=1}^{N} x_i^{1/2}}{Nn} \left(\sum_s x_i^{-1/2} \right) - 1 \right] \beta_0 + \\ &+ \left[\frac{\sum_{i=1}^{N} x_i^{1/2}}{Nn} \left(\sum_s x_i^{1/2} \right) - \bar{x} \right] \beta_1 \end{aligned}$$

con $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

El hecho de que el ξ–sesgo del estimador de Horvitz-Thompson, $L_2' Y_s$, dependa, no sólo de β_0, sino también de $\beta_1 \neq 0$, nos lleva a descartarlo, prefiriendo, en este caso, el estimador de razón generalizado, $L_1' Y_s$.

5. BIBLIOGRAFÍA

[1] **Azorín, F.** y **Sánchez-Crespo, J.L.** (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza.

[2] **Brewer, K.R.W.** (1963). "Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process". *Australian Journal of Statistics,* **5**, 93–105.

[3] **Brewer, K.R.W.** (1979). "A class of robust sampling designs for large-scale surveys". *Journal of American Statistical Association,* **74**, 911–915.

[4] **Cassel, C., Särndal, C.** y **Wretman, J.H.** (1977). *Foundations of Inference in Survey Sampling.* New York: John Wiley.

[5] **Casas, J.M.** y **Guijarro, M.** (1993). "El estimador de regresión generalizado en el modelo de superpoblación: $p-$insesgadez asintótica y robustez". *Estadística Española,* **35**, 425–437.

[6] **Fuller, W.A.** e **Isaki, C.T.** (1982). "Survey design under the regression superpopulation model". *Journal of American Statistical Association,* **77**, 89–96.

[7] **Godambe, V.P.** (1955). "A unified theory of sampling from finite populations". *Journal of the Royal Statistical Society,* Ser. B, **17**, 369–378.

[8] **Godambe, V.P.** (1982). "Estimation in survey sampling: robustness and optimality". *Journal of American Statistical Association,* **77**, 393–406.

[9] **Godambe, V.P.** y **Thompson, M. E.** (1977). "Robust near optimal estimation in survey practice". *Bulletin of the International Statistical Institute,* **47**, 129–146.

[10] **Hansen, M.H., Madow, W.G.** y **Tepping, B. J.** (1983). "An evaluation of model-dependent and probability-sampling inferences in sample surveys". *Journal of American Statistical Association,* **78**, 776–807.

[11] **Herson, J.** y **Royall, R.M.** (1973). "Robust estimation in finite populations". *Journal of American Statistical Association,* **68**, 880–893.

[12] **Robinson, P.M.** y **Särndal, C.E.** (1983). "Asymptotic properties of the generalized regression estimator in probability sampling". *Sankyā.* Ser. B, **45**, 240–248.

[13] **Särndal, C.E.** (1980a). "On $\pi-$inverse weighting versus best linear unbiased weighting in probability sampling". *Biometrika,* **67**, 639–650.

[14] **Tam, S.M.** (1988b). "Some results on robust estimation in finite population sampling". *Journal of American Statistical Association,* **83**, 242–248.

[15] **Wright, R.L.** (1983). "Finite population sampling with multivariate auxiliar information". *Journal of American Statistical Association,* **78**, 879–884.

ENGLISH SUMMARY:

ASYMPTOTIC DESIGN UNBIASEDNESS AND ROBUSTNESS OF LINEAR ESTIMATION UNDER SUPERPOPULATION MODELS: A SELECTION PROCEDURE

José Miguel Casas Sánchez and Marta Guijarro Garvi

1. INTRODUCTION

Estimation of means under superpopulation models leads to the necessity of looking for robust estimators when the model is misspecified.

This work assumes a superpopulation model with correlated residuals. In such context, a rule for choosing among weakly robust and asymptotically design unbiased linear estimators, is suggested.

2. MODEL DESCRIPTION

Following the asymptotic framework of Isaki and Fuller, a generalized definition of weakly robust estimator is presented.

This definition assumes $y_t = (y_1, \ldots, y_{N_t})'$ to be the realized outcome of a random vector $Y_t(Y_1, \ldots, Y_{N_t})'$ related to the matrix $X_t = (x_1, \ldots, x_{N_t})'$ through the superpopulation model ξ

$$E_\xi(Y_t) = X_t\beta$$
$$E_\xi[(Y_t - X_t\beta)(Y_t - X_t\beta)'] = \sigma^2 V_t$$

where $V_t = (v_{ik})$ is a positive definite matrix with non cero correlation coefficient.

Weakly robustness, asymptotic design unbiasedness, model unbiasedness and C.1-C.7 conditions, play a central role in this paper.

3. ASYMPTOTICALLY DESIGN UNBIASED AND WEAKLY ROBUST LINEAR ESTIMATORS

Under H.1-H.3 and given two veackly robust and asymptotically design unbiased linear estimators of $\bar{Y}, L'_m Y_s (m = 1, 2)$:

$$E_p E_\xi (L'_1 Y_s - \bar{Y}_t)^2 \geq E_p E_\xi (L'_2 Y_s - \bar{Y}_t)^2$$

asymptotically and for all sampling designs p, if the model bias of $L'_1 Y_s$ is bigger than that of $L'_2 Y_s$ (in absolute terms).

4. APPLICATION

Under the model

$$E_\xi(Y_i) = \beta_0 + \beta_1 x_i$$

$$E_\xi[(Y_i - \beta_0 - \beta_1 x_i)^2] = \sigma^2 x_i$$

$$E_\xi[(Y_i - \beta_0 - \beta_1 x_i)(Y_k - \beta_0 - \beta_1 x_k)] = \sigma^2 \rho (x_i x_k)^{1/2}$$

it can be easily proved that

$$E_\xi(L'_1 Y_s - \bar{Y}) = \left[\bar{x} \left(\sum_s x_i^{-1/2} \right) \left(\sum_s x_i^{1/2} \right)^{-1} - 1 \right] \beta_0$$

and

$$E_\xi(L_2'Y_s - \bar{Y}) = \left[\frac{\sum_{i=1}^N x_i^{1/2}}{Nn}\left(\sum_s x_i^{-1/2}\right) - 1\right]\beta_0 + \left[\frac{\sum_{i=1}^N x_i^{1/2}}{Nn}\left(\sum_s x_i^{1/2}\right) - \bar{x}\right]\beta_1$$

with $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and

$$\pi_{i0} = \frac{n x_i^{1/2}}{\sum_{i=1}^N x_i^{1/2}}$$

where

$$L_1'Y_s = \bar{x}\frac{\sum_{i\in s} Y_i/\pi_i}{\sum_{i\in s} x_i/\pi_i}$$

is the generalized ratio estimator and

$$L_2'Y_s = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i}$$

the Horvitz-Thompson estimator.

Since $E_\xi(L_2'Y_s - \bar{Y})$ depends not only on B_0, but in $B_1 \neq 0$, application of the theorem shows that $L_1'Y_s$ is better than $L_2'Y_s$.