

Un enfoque multiparamétrico en encuestas multipropósito

Andrés Gutiérrez
Universidad Santo Tomás

Julio, 2009

- 1 Motivación
- 2 Introducción
- 3 Etapa de diseño
- 4 Etapa de estimación

Motivación

Si los estadísticos teóricos hacen caso omiso al reto de enfrentar las encuestas multi-propósito, entonces el vacío existente entre ellos y los estadísticos prácticos se hará cada vez más grande. El diseño y análisis de encuestas multivariantes debe ser una de las próximas áreas de mayor investigación.

T. M. F. Smith (1976)

Introducción

La mayoría de aplicaciones en encuestas por muestreo involucran múltiples variables de estudio. En este breve apartado, se presenta un marco de referencia para la estimación conjunta de los parámetros de interés, bajo algunos diseños de muestreo.

Con respecto al diseño de muestreo, en Holmberg (2002a, 2002b) se ha desarrollado la teoría pertinente para la sección de muestras probabilísticas en encuestas multipropósito, y con respecto a la estimación multiparamétrica, en Gutiérrez (2009) se propone un sistema general de estimación basado en resultados clásicos de la teoría de los modelos lineales y del álgebra lineal.

Introducción

- El propósito de un estudio por muestreo está enfocado en obtener información acerca de una población finita particular por medio de la estimación de parámetros poblacionales como medias, totales o proporciones o razones.
- Sin embargo, la mayoría de encuestas no involucra una sola característica sino varias características de interés.
- Los libros clásicos de muestreo parecen omitir el hecho de que raras veces se planea una encuesta con el fin de estimar un sólo parámetro

Introducción

- El propósito de un estudio por muestreo está enfocado en obtener información acerca de una población finita particular por medio de la estimación de parámetros poblacionales como medias, totales o proporciones o razones.
- Sin embargo, la mayoría de encuestas no involucra una sola característica sino varias características de interés.
- Los libros clásicos de muestreo parecen omitir el hecho de que raras veces se planea una encuesta con el fin de estimar un sólo parámetro

Introducción

- El propósito de un estudio por muestreo está enfocado en obtener información acerca de una población finita particular por medio de la estimación de parámetros poblacionales como medias, totales o proporciones o razones.
- Sin embargo, la mayoría de encuestas no involucra una sola característica sino varias características de interés.
- Los libros clásicos de muestreo parecen omitir el hecho de que raras veces se planea una encuesta con el fin de estimar un sólo parámetro

Introducción

¡¡Una encuesta típica en el sector económico involucra varias características de interés y varios parámetros objetivos... con múltiples parámetros de interés y múltiples requerimientos de precisión, el estadístico debería escoger un diseño de muestreo que tenga en cuenta las anteriores características y un sistema general de estimación que contemple las características multi-variantes del problema en cuestión.¡¡

Diseños óptimos = Diseños de Holmberg

- Suponga que las características de interés involucradas en la encuestas tienen todas la misma importancia (se pueden considerar variantes).
- Para cada una de las características de interés se debe proponer un diseño de muestreo, $p_q(\cdot)$ ($q = 1, \dots, Q$), que sea óptimo y tal que el tamaño esperado de muestra sea $E(n_S) = n_q$.
 - Cada uno de los Q diseños de muestreo pueden ser diferentes; aún más, los tamaños de muestra, en cada diseño propuesto, no necesariamente deben ser equivalentes.
- Cada uno de los diseños de muestreo $p_q(\cdot)$ induce un vector de probabilidades de inclusión de tamaño N para cada una de los elementos pertenecientes a la población finita.

Diseños óptimos = Diseños de Holmberg

- Suponga que las características de interés involucradas en la encuestas tienen todas la misma importancia (se pueden considerar variantes).
- Para cada una de las características de interés se debe proponer un diseño de muestreo, $p_q(\cdot)$ ($q = 1, \dots, Q$), que sea óptimo y tal que el tamaño esperado de muestra sea $E(n_S) = n_q$.
 - Cada uno de los Q diseños de muestreo pueden ser diferentes; aún más, los tamaños de muestra, en cada diseño propuesto, no necesariamente deben ser equivalentes.
- Cada uno de los diseños de muestreo $p_q(\cdot)$ induce un vector de probabilidades de inclusión de tamaño N para cada una de los elementos pertenecientes a la población finita.

Diseños óptimos = Diseños de Holmberg

- Suponga que las características de interés involucradas en la encuestas tienen todas la misma importancia (se pueden considerar variantes).
- Para cada una de las características de interés se debe proponer un diseño de muestreo, $p_q(\cdot)$ ($q = 1, \dots, Q$), que sea óptimo y tal que el tamaño esperado de muestra sea $E(n_S) = n_q$.
 - Cada uno de los Q diseños de muestreo pueden ser diferentes; aún más, los tamaños de muestra, en cada diseño propuesto, no necesariamente deben ser equivalentes.
- Cada uno de los diseños de muestreo $p_q(\cdot)$ induce un vector de probabilidades de inclusión de tamaño N para cada una de los elementos pertenecientes a la población finita.

Diseños óptimos = Diseños de Holmberg

- Suponga que las características de interés involucradas en la encuestas tienen todas la misma importancia (se pueden considerar variantes).
- Para cada una de las características de interés se debe proponer un diseño de muestreo, $p_q(\cdot)$ ($q = 1, \dots, Q$), que sea óptimo y tal que el tamaño esperado de muestra sea $E(n_S) = n_q$.
 - Cada uno de los Q diseños de muestreo pueden ser diferentes; aún más, los tamaños de muestra, en cada diseño propuesto, no necesariamente deben ser equivalentes.
- Cada uno de los diseños de muestreo $p_q(\cdot)$ induce un vector de probabilidades de inclusión de tamaño N para cada una de los elementos pertenecientes a la población finita.

- Estas probabilidades de inclusión deben tomar la siguiente forma

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \quad (1)$$

- Nótese que σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal).
- La optimalidad en el diseño de muestreo se obtiene si $\pi_{qk} \propto \sigma_{qk}$.
- Si el diseño de muestreo óptimo para la q -ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$.
- Con la escogencia de $\sigma_{qk}^2 = \sigma_q^2 x_{qk}^{\gamma_q}$, donde σ_q^2 es constante, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} . Es decir, $\pi_{qk} \propto x_{qk}^{\gamma_q/2}$.

- Estas probabilidades de inclusión deben tomar la siguiente forma

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \quad (1)$$

- Nótese que σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal).
- La optimalidad en el diseño de muestreo se obtiene si $\pi_{qk} \propto \sigma_{qk}$.
- Si el diseño de muestreo óptimo para la q -ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$.
- Con la escogencia de $\sigma_{qk}^2 = \sigma_q^2 x_{qk}^{\gamma_q}$, donde σ_q^2 es constante, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} . Es decir, $\pi_{qk} \propto x_{qk}^{\gamma_q/2}$.

- Estas probabilidades de inclusión deben tomar la siguiente forma

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \quad (1)$$

- Nótese que σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal).
- La optimalidad en el diseño de muestreo se obtiene si $\pi_{qk} \propto \sigma_{qk}$.
- Si el diseño de muestreo óptimo para la q -ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$.
- Con la escogencia de $\sigma_{qk}^2 = \sigma_q^2 x_{qk}^{\gamma_q}$, donde σ_q^2 es constante, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} . Es decir, $\pi_{qk} \propto x_{qk}^{\gamma_q/2}$.

- Estas probabilidades de inclusión deben tomar la siguiente forma

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \quad (1)$$

- Nótese que σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal).
- La optimalidad en el diseño de muestreo se obtiene si $\pi_{qk} \propto \sigma_{qk}$.
- Si el diseño de muestreo óptimo para la q -ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$.
- Con la escogencia de $\sigma_{qk}^2 = \sigma_q^2 x_{qk}^{\gamma_q}$, donde σ_q^2 es constante, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} . Es decir, $\pi_{qk} \propto x_{qk}^{\gamma_q/2}$.

- Estas probabilidades de inclusión deben tomar la siguiente forma

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \quad (1)$$

- Nótese que σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal).
- La optimalidad en el diseño de muestreo se obtiene si $\pi_{qk} \propto \sigma_{qk}$.
- Si el diseño de muestreo óptimo para la q -ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$.
- Con la escogencia de $\sigma_{qk}^2 = \sigma_q^2 X_{qk}^{\gamma_q}$, donde σ_q^2 es constante, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} . Es decir, $\pi_{qk} \propto X_{qk}^{\gamma_q/2}$.

- Basado en el criterio de mínima pérdida de eficiencia relativa general (ANOREL), el tamaño de muestra óptimo para la encuesta multi-propósito estará dado por

$$n^* \geq \frac{(\sum_{k \in U} \sqrt{a_{qk}})^2}{(1+c)Q + \sum_{k \in U} a_{qk}}, \quad (2)$$

donde

$$a_{qk} = \sum_{q=1}^Q \frac{\sigma_{qk}^2}{\sum_{k \in U} \left(\frac{1}{\pi_{qk}} - 1 \right) \sigma_{qk}^2}, \quad (3)$$

y c es el máximo error permitido, bajo el criterio ANOREL, en una escala de cero hasta uno.

- Una vez que el tamaño de la muestra ha sido calculado, se debe crear un sólo vector de probabilidades de inclusión que sea óptimo para todas las características de interés. Este vector es inducido por el diseño de muestreo de Holmberg, el cual minimiza la pérdida de eficiencia relativa general, y está dado por la siguiente expresión

$$\pi_{(opt)k} = \frac{n^* \sqrt{a_{qk}}}{\sum_{k \in U} \sqrt{a_{qk}}} \quad (4)$$

- En la mayoría de los casos, el vector resultante $\pi_{(opt)} = (\pi_{(opt)1}, \dots, \pi_{(opt)N})'$ es un vector de probabilidades de inclusión desiguales. En esta situación, se debe usar un esquema de selección de muestras π_{PT} .

- Una vez que el tamaño de la muestra ha sido calculado, se debe crear un sólo vector de probabilidades de inclusión que sea óptimo para todas las características de interés. Este vector es inducido por el diseño de muestreo de Holmberg, el cual minimiza la pérdida de eficiencia relativa general, y está dado por la siguiente expresión

$$\pi_{(opt)k} = \frac{n^* \sqrt{a_{qk}}}{\sum_{k \in U} \sqrt{a_{qk}}} \quad (4)$$

- En la mayoría de los casos, el vector resultante $\pi_{(opt)} = (\pi_{(opt)1}, \dots, \pi_{(opt)N})'$ es un vector de probabilidades de inclusión desiguales. En esta situación, se debe usar un esquema de selección de muestras π PT.

Enfoque matricial

Suponga que la encuesta involucra el estudio de Q características de interés. Suponga que el k -ésimo elemento ($k \in U$) está asociado a un vector de Q características de interés, $\mathbf{y}_k = (y_{k1}, \dots, y_{kQ})$ cuyos valores son desconocidos para la población finita. De esta manera, la siguiente matriz será llamada la **matriz de interés**.

$$\mathbf{Y}_U = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \dots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \dots & y_{NQ} \end{pmatrix} = (\mathbf{y}^1 \quad \mathbf{y}^2 \quad \dots \quad \mathbf{y}^Q) \quad (5)$$

Los valores de cada característica de interés no son necesariamente continuos sino también discretos como indicadores de subgrupos poblacionales como dominios, estratos o post-estratos. El objetivo es estimar los Q componentes del vector de totales definido por la siguiente expresión

$$\mathbf{t} = (t_1, t_2, \dots, t_Q)' = \mathbf{Y}'_U \mathbf{1}_N, \quad (6)$$

donde $\mathbf{1}_N = (1, 1, \dots, 1)'_{N \times 1}$ y $t_q = \sum_{k \in U} y_{kq}$ es el total poblacional de la q -ésima característica de interés.

Cuando la muestra de tamaño n es seleccionada, entonces y_{kq} es observado ($k \in s$) y es posible definir la siguiente matriz

$$\mathbf{Y}_s = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} \end{pmatrix}. \quad (7)$$

De esta manera, la matriz de probabilidades de inclusión está definida por la siguiente expresión

$$\mathbf{\Pi} = \text{diag}(\pi_1, \pi_2, \dots, \pi_n), \quad (8)$$

En este orden de ideas, el estimador de Horvitz-Thompson del vector de totales \mathbf{t} se define como

$$\hat{\mathbf{t}}_{\pi} = (\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \dots, \hat{t}_{Q,\pi})' = \mathbf{Y}'_s \mathbf{\Pi}^{-1} \mathbf{1}_n, \quad (9)$$

con $\mathbf{1}_N = (1, 1, \dots, 1)'_{n \times 1}$ y $\hat{t}_{q,\pi} = \sum_{k \in S} y_{kq} / \pi_k$ es el estimador de Horvitz-Thompson de t_q . Es fácil probar que $\hat{\mathbf{t}}_{\pi}$ corresponde a un estimador insesgado para \mathbf{t} .

La matriz de varianzas está dada por

$$\mathbf{V}(\widehat{\mathbf{t}}_{\pi}) = E(\widehat{\mathbf{t}}_{\pi} - \mathbf{t})(\widehat{\mathbf{t}}_{\pi} - \mathbf{t})'. \quad (10)$$

Nótese que, si $N \geq q$, entonces $\mathbf{V}(\widehat{\mathbf{t}}_{\pi})$ será una matriz simétrica y definida positiva cuyo elemento qq' es

$$\sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_{kq}}{\pi_k} \frac{y_{lq'}}{\pi_l}, \quad (11)$$

con $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Si $s \neq U$ es imposible calcular el valor de la anterior expresión. Sin embargo, si $n \geq q$, la varianza puede ser estimada mediante una matriz simétrica y definida positiva $\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_{\pi})$ cuyo elemento qq' es

$$\sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{kq}}{\pi_k} \frac{y_{lq'}}{\pi_l}. \quad (12)$$

Si los requerimientos de la encuesta están relacionados con la estimación del tamaño absoluto de un dominio o del total de alguna o varias características de interés en tal dominio se supone que la población está particionada en D dominios y se define la **matriz indicadora de dominios** como

$$\mathbf{Z} = \begin{pmatrix} z_{11} & \dots & z_{1d} & \dots & z_{1D} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{k1} & \dots & z_{kd} & \dots & z_{kD} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & \dots & z_{nd} & \dots & z_{nD} \end{pmatrix} \quad (13)$$

donde el elemento

$$z_{kd} = \begin{cases} 1 & \text{si } k \in U_d, \text{ y} \\ 0 & \text{en otro caso} \end{cases} \quad (14)$$

El vector de tamaños absolutos del dominio d esta dado por

$$\mathbf{N}_d = (N_1, N_2, \dots, N_D)' \quad (15)$$

donde

$$N_d = \sum_{k \in U} z_{kd}. \quad (16)$$

y \mathbf{N}_d es estimado insesgadamente por el estimador de Horvitz-Thompson de la siguiente manera

$$\hat{\mathbf{N}}_d = (\hat{N}_1, \hat{N}_2, \dots, \hat{N}_D)' = \mathbf{Z}'\boldsymbol{\pi}^{-1}\mathbf{1}_n, \quad (17)$$

su matriz de varianzas es estimada insesgadamente por $\hat{\mathbf{V}}(\hat{\mathbf{N}}_d)$.

el total de la q -ésima variable sobre todos los D dominios de interés está dado por

$$\mathbf{t}_{dq} = (t_{1q}, t_{2q}, \dots, t_{Dq})' \quad (18)$$

y una forma de estimarlo está dada por la siguiente expresión

$$\widehat{\mathbf{t}}_{dq\pi} = (\widehat{t}_{1q\pi}, \widehat{t}_{2q\pi}, \dots, \widehat{t}_{Dq\pi})' = (\mathbf{y}^q \mathbf{1}_D \odot \mathbf{Z})' \mathbf{\Pi}^{-1} \mathbf{1}_n \quad (19)$$

En donde, \mathbf{y}^q denota la q -ésima columna de la matriz \mathbf{Y}_s , $\mathbf{1}_D = (1, \dots, 1)'_{D \times 1}$ y \odot denota el producto matricial de Hadamard.

Para diseños estratificados se la población finita U se divide en H grupos. Se selecciona una muestra aleatoria en todos los H estratos existentes. La matriz \mathbf{Y}_s se particiona en H bloques de la siguiente manera

$$\mathbf{Y}_s = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_h \\ \vdots \\ \mathbf{Y}_H \end{pmatrix}, \quad (20)$$

Note que $\mathbf{Y}_s \in \mathfrak{R}^{Hn \times Q}$ y $\mathbf{Y}_h \in \mathfrak{R}^{n_h \times Q}$. Definido $\mathbf{n} = (n_1, \dots, n_H)'$, entonces $n = \mathbf{n}'\mathbf{1}_H = n_1 + \dots + n_H$.

El objetivo es la estimación de los Q componentes del vector de totales en el h -ésimo estrato dado por

$$\mathbf{t}_h = (t_{1h}, t_{2h}, \dots, t_{Qh})' = \mathbf{Y}'_h \mathbf{1}_{N_h}, \quad (21)$$

donde N_h es el tamaño del h -ésimo estrato. El total poblacional puede ser escrito como

$$\mathbf{t} = (t_1, t_2, \dots, t_Q)' = \sum_{h=1}^H \mathbf{t}_h, \quad (22)$$

donde \mathbf{t}_h es estimado insesgadamente por la siguiente expresión

$$\hat{\mathbf{t}}_{h\pi} = (\hat{t}_{1h\pi}, \hat{t}_{2h\pi}, \dots, \hat{t}_{Qh\pi})' = \mathbf{Y}'_h \Pi_h \mathbf{1}_{n_h}, \quad (23)$$

con n_h el tamaño de la muestra en el h -ésimo estrato. El total poblacional está dado por

$$\hat{\mathbf{t}}_{\pi} = (\hat{t}_{1\pi}, \hat{t}_{2\pi}, \dots, \hat{t}_{Q\pi})' = \sum_{h=1}^H \hat{\mathbf{t}}_h, \quad (24)$$

Si el k -ésimo elemento está asociado con un vector de P características de información auxiliar, contenidas en un vector $\mathbf{x}_k = (x_{k1}, \dots, x_{kP})$ cuyos valores se suponen conocidos para la población finita, entonces la siguiente matriz

$$\mathbf{X}_U = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kP} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} = (\mathbf{x}^1 \quad \mathbf{x}^2 \quad \dots \quad \mathbf{x}^P) \quad (25)$$

que será llamada la **matriz de información auxiliar**.

Es posible asumir que existe una relación lineal entre cada uno de los componentes de las características de interés y las características de información auxiliar mediante un modelo de superpoblación ξ_q , $q = 1, \dots, Q$, tal que

$$\mathbf{Y}^q = \mathbf{X} \beta^q + \boldsymbol{\varepsilon}^q.$$

$(N \times 1)$ $(N \times P)$ $(P \times 1)$ $(N \times 1)$

El modelo ξ_q tiene las siguientes propiedades:

$$\begin{aligned} E_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \mathbf{0} \\ V_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \boldsymbol{\Sigma}_q. \end{aligned} \tag{26}$$

$\boldsymbol{\Sigma}_q$ establece la estructura de varianza del vector $\boldsymbol{\varepsilon}^q$. Nótese que las anteriores relaciones pueden reescribirse mediante un modelo conjunto ξ tal que

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$(N \times Q)$ $(N \times P)$ $(P \times Q)$ $(N \times Q)$

Este enfoque sugiere que \mathbf{Y} , \mathbf{X} y ε son matrices aleatorias definidas en el modelo de superpoblación ξ , para el cual \mathbf{Y}_U y \mathbf{X}_U se suponen meras realizaciones de las anteriores matrices aleatorias. Más precisamente, el modelo ξ tiene las siguientes características:

$$\begin{aligned}
 E_{\xi}(\varepsilon) &= \mathbf{0}_{(N \times Q)} \\
 V_{\xi}(\vec{\varepsilon}) &= \sum_{(NQ \times NQ)} = \text{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_Q)
 \end{aligned} \tag{27}$$

Note que el subíndice ξ se refiere a la esperanza bajo la estructura particular que ese modelo de superpoblación induce. En situaciones prácticas, es común asumir $\Sigma_q = \sigma_q^2 \text{diag}(c_{1q}, \dots, c_{Nq})$, donde $c_{kq} = f_q(x_{k1}, \dots, x_{kP})$ y f_q es una función de valor real.

Sea $D(\mathbf{X})$ una medida de dispersión invariante ante traslaciones tal que $D(\mathbf{X} + \mathbf{K}) = D(\mathbf{X})$, con \mathbf{K} una matriz de constantes. Entonces al estimación de β corresponderá a aquel vector que minimize la anterior medida de dispersión. Particularmente, $D(\cdot)$ podría estar dada por la varianza total multivariante definida como

$$\text{trace}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta). \quad (28)$$

Entonces, (28) es minimizada por

$$\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_Q), \quad (29)$$

donde

$$\mathbf{B}_q = (\mathbf{X}'_U \Sigma_q^{-1} \mathbf{X}_U)^{-1} (\mathbf{X}'_U \Sigma_q^{-1} \mathbf{Y}_U). \quad (30)$$

Para poder calcular esta estimación, se deben conocer todos los valores poblacionales.

Como sólo se selecciona una muestra no es posible calcular \mathbf{B} . Por lo tanto, debe ser estimado. Puede ser demostrado que la siguiente expresión corresponde a un estimador asintóticamente insesgado para \mathbf{B}

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1, \widehat{\mathbf{B}}_2, \dots, \widehat{\mathbf{B}}_Q), \quad (31)$$

donde

$$\widehat{\mathbf{B}}_q = (\mathbf{X}'_s \mathbf{A}_q^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}'_s \mathbf{A}_q^{-1} \mathbf{Y}_s), \quad (32)$$

$q = 1, \dots, Q$, \mathbf{X}_s similarmente definido, y

$$\mathbf{A}_q = \boldsymbol{\Pi}^{1/2} \boldsymbol{\Sigma}_q \boldsymbol{\Pi}^{1/2}. \quad (33)$$

El **estimador múltiple de regresión general** para el vector de totales poblacionales se define como

$$\hat{\mathbf{t}}_{Mgreg} = \hat{\mathbf{t}}_{y\pi} + \hat{\mathbf{B}}'(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}), \quad (34)$$

con, $\hat{\mathbf{t}}_{y\pi}$, $\hat{\mathbf{t}}_{x\pi}$ los estimadores de Horvitz-Thompson de \mathbf{t}_y y \mathbf{t}_x , respectivamente. Nótese que $\hat{\mathbf{B}}_q$ también puede ser escrito como

$$\hat{\mathbf{B}}_q = (\mathbf{X}'_s \mathbf{D}_\lambda \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{D}_\lambda \mathbf{Y}_s \quad (35)$$

$$= \left(\sum_{k \in S} \mathbf{x}_k \lambda_k^q \mathbf{x}'_k \right)^{-1} \left(\sum_{k \in S} \mathbf{x}_k \lambda_k^q \mathbf{y}'_k \right) \quad (36)$$

donde $\mathbf{D}_\lambda = \text{diag}(\lambda_1^q, \dots, \lambda_n^q)$ y λ_k^q son funciones de valor real de las probabilidades de inclusión y de la información auxiliar. Las propiedades del estimador múltiple de regresión general (esperanza y varianza) también se definen desde una perspectiva de inferencia basada en el diseño de muestreo.

Los siguientes escenarios resultan ser casos especiales del estimador múltiple de regresión general; en la mayoría de los casos su particularidad está inducida por la escogencia de los valores de λ_k .

- Si $P = 1$, $\mathbf{x}_k = x_k$, y $\lambda_k^q = (\pi_k x_k)^{-1}$, entonces se tiene el estimador de razón para cada característica de interés.
- Si $P = 2$, $\mathbf{x}_k = (1, x_k)'$, y $\lambda_k^q = (\pi_k)^{-1}$, entonces se tiene el estimador de regresión clásico.
- Si $P = M$ (number of post-strata), $\mathbf{x}_k = \delta_k = (0, \dots, 0, 1, 0, \dots)$ y $\lambda_k^q = (\pi_k)^{-1}$, donde δ_k representa M variables indicadoras (cada indicadora representa la membresía del elemento poblacional al post-estrato en cuestión), entonces tenemos el estimador de post-estratificación.

Los siguientes escenarios resultan ser casos especiales del estimador múltiple de regresión general; en la mayoría de los casos su particularidad está inducida por la escogencia de los valores de λ_k .

- Si $P = 1$, $\mathbf{x}_k = x_k$, y $\lambda_k^q = (\pi_k x_k)^{-1}$, entonces se tiene el estimador de razón para cada característica de interés.
- Si $P = 2$, $\mathbf{x}_k = (1, x_k)'$, y $\lambda_k^q = (\pi_k)^{-1}$, entonces se tiene el estimador de regresión clásico.
- Si $P = M$ (number of post-strata), $\mathbf{x}_k = \delta_k = (0, \dots, 0, 1, 0, \dots)$ y $\lambda_k^q = (\pi_k)^{-1}$, donde δ_k representa M variables indicadoras (cada indicadora representa la membresía del elemento poblacional al post-estrato en cuestión), entonces tenemos el estimador de post-estratificación.

Los siguientes escenarios resultan ser casos especiales del estimador múltiple de regresión general; en la mayoría de los casos su particularidad está inducida por la escogencia de los valores de λ_k .

- Si $P = 1$, $\mathbf{x}_k = x_k$, y $\lambda_k^q = (\pi_k x_k)^{-1}$, entonces se tiene el estimador de razón para cada característica de interés.
- Si $P = 2$, $\mathbf{x}_k = (1, x_k)'$, y $\lambda_k^q = (\pi_k)^{-1}$, entonces se tiene el estimador de regresión clásico.
- Si $P = M$ (number of post-strata), $\mathbf{x}_k = \delta_k = (0, \dots, 0, 1, 0, \dots)$ y $\lambda_k^q = (\pi_k)^{-1}$, donde δ_k representa M variables indicadoras (cada indicadora representa la membresía del elemento poblacional al post-estrato en cuestión), entonces tenemos el estimador de post-estratificación.

Nótese que el estimador múltiple de regresión general puede también escribirse de la siguiente manera

$$\hat{\mathbf{t}}_{Mgreg} = (\mathbf{W}' \odot \mathbf{Y}'_s) \mathbf{1}_n, \quad (37)$$

donde

$$\mathbf{W} = \begin{pmatrix} w_1^1 & w_1^2 & \dots & w_1^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_k^1 & w_k^2 & \dots & w_k^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_n^1 & w_n^2 & \dots & w_n^Q \end{pmatrix} = (\mathbf{w}^1 \quad \mathbf{w}^2 \quad \dots \quad \mathbf{w}^Q). \quad (38)$$

Se tiene que $\mathbf{w}^q = (w_1^q, \dots, w_k^q, \dots, w_n^q)'$ es un vector de pesos o ponderaciones tales que

$$w_k^q = \frac{1}{\pi_k} \left(1 + \lambda_k^q \mathbf{x}'_k \left(\sum_{k \in s} \mathbf{x}_k \lambda_k^q \mathbf{x}'_k \right)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}) \right). \quad (39)$$

A estos pesos, como se estudió en capítulos anteriores, se le conocen con el nombre de **ponderaciones de calibración** y ellos reproducen con exactitud el vector de totales \mathbf{t}_x cuando son aplicados a la información auxiliar disponible. Entonces, \mathbf{W} es llamada **matriz de calibración**. No es difícil mostrar que la siguiente relación

$$\sum_{k \in s} w_k^q \mathbf{x}_k = \mathbf{X}'_s \mathbf{w}^q = \mathbf{t}_x, \quad (40)$$

se satisface para cada $q = 1, \dots, Q$. Es interesante observar que \mathbf{t}_x resulta calibrado bajo diferentes escogencias de los pesos w^q . Por otra parte, note que

se supone la existencia de una matriz de información conjunta cuya estructura algebraica está definida por la siguiente expresión

$$\mathbf{V} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1Q} & x_{11} & x_{12} & \cdots & x_{1P} \\ y_{21} & y_{22} & \cdots & y_{2Q} & x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} & x_{n1} & x_{n2} & \cdots & x_{nP} \end{pmatrix}. \quad (42)$$

El estimador del vector de totales de las características de interés y de las características de información auxiliar está dado por $\hat{\mathbf{t}}_{\mathbf{v}\pi}$, el cual está definido como

$$\hat{\mathbf{t}}_{\mathbf{v}\pi} = \mathbf{V}'\boldsymbol{\Pi}^{-1}\mathbf{1}_n. \quad (43)$$

Suponga que $\widehat{\mathbf{t}}_{\mathbf{v}\pi}$ sigue una distribución normal multivariante con media

$$E(\widehat{\mathbf{t}}_{\mathbf{v}\pi}) = (\mathbf{t}'_{\mathbf{y}\pi}, \mathbf{t}'_{\mathbf{x}\pi})' = \mathbf{t}_{\mathbf{v}},$$

y matriz de varianzas definida como

$$V(\widehat{\mathbf{t}}_{\mathbf{v}\pi}) = \begin{pmatrix} V(\widehat{\mathbf{t}}_{\mathbf{y}\pi}) & C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}) \\ C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}) & V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}) \end{pmatrix},$$

Por lo tanto, la distribución condicional de $\widehat{\mathbf{t}}_{\mathbf{y}\pi}$ dado $\widehat{\mathbf{t}}_{\mathbf{x}\pi}$ sigue también una distribución normal multivariante con media condicional dada por

$$E(\widehat{\mathbf{t}}_{\mathbf{y}\pi} | \widehat{\mathbf{t}}_{\mathbf{x}\pi}) = \mathbf{t}_{\mathbf{y}\pi} + C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})(V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}))^{-1}(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}), \quad (44)$$

Y varianza condicional dada por

$$V(\widehat{\mathbf{t}}_{\mathbf{y}\pi} | \widehat{\mathbf{t}}_{\mathbf{x}\pi}) = V(\widehat{\mathbf{t}}_{\mathbf{y}\pi}) - C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})(V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}))^{-1}C(\widehat{\mathbf{t}}_{\mathbf{x}\pi}, \widehat{\mathbf{t}}_{\mathbf{y}\pi}). \quad (45)$$

Observe que el anterior estimador luce como el estimador múltiple de regresión general. Sin embargo, su pendiente, $\hat{\mathbf{B}}$, sería diferente: mientras la pendiente del estimador de regresión general está inducida por el método de mínimos cuadrados, la pendiente de éste último corresponde, según los resultados de la inferencia estadística multivariante, a un conjunto de regresiones múltiples de \mathbf{X} sobre \mathbf{Y} .

Este estimador del vector de totales de la característica de interés debería ser llamado **estimador óptimo de regresión general** y ha sido estudiado por diversos autores en el contexto de la inferencia basada en modelos poblacionales para la estimación del total de una sola característica de interés.

Software

TeachingSampling: R Packackage

disponible en el CRAN desde principios de
2009

Agradecimientos

Muchas gracias !!!

<http://predictive.wordpress.com>