estimación en dominios

Muestreo II

Licenciatura en Estadística

2023

- las encuestas no solo se utilizar para obtener estimaciones a nivel de toda la población U
- necesidad de obtener estimaciones "confiables" para distintos subconjuntos (dominios/áreas) U_d

ejemplos

- estimar el ingreso promedio de los hogares en Canelones
- estimar la proporción de niños menores de 6 años que se encuentran por debajo de la LP
- **estimar/predecir** el total de mujeres pobres entre 30 y 49 años que residen en Montevideo, tienen un hijo a cargo y tienen primaria incompleta.

Sea $U_1, ..., U_d, ..., U_D$ dominios disjuntos y N_d el tamaño del dominio U_d y se cumple que:

$$N = \sum_{d=1}^{D} N_d, \quad U = \bigcup_{d=1}^{D} U_d.$$

- el objetivo es estimar el total de la variable y en el dominio U_d

$$Y_d = \sum_{i \in U_d} y_i$$

los dominios pueden ser clasificados como:

- planeados = estrato (podemos definir en el diseño muestral tamaños de muestra específicos n_d)
- no planeado (la mayoría). El tamaño de muestra en el dominio n_d es aleatorio (agrega fuente EXTRA de variabilidad a las estimaciones)

un **dominio es pequeño** si el tamaño de muestra n_d no permite obtener estimaciones confiables utilizando estimadores basados en el diseño (asistidos o por no por modelos)

en la estimación del total de la variable y en un dominio U_d resulta útil definir una variable extendida y_d, la cual toma el valor $y_{di} = y_i$ si $i \in U_d$ y $y_{di} = 0$ en otro caso.
el total de la variable y en el dominio d se puede expresar como:

$$Y_d = \sum_{i \in U_d} y_i = \sum_{i \in U} y_{di},$$

y su respectivo estimador HT es:

$$\hat{Y}_d^{\mathrm{HT}} = \sum_{i \in s_d} w_i y_i = \sum_{i \in s} w_i y_{di},$$

- es insesgado y un estimador insesgado de la varianza se obtiene de remplazar la variable y por su correspondiente variable extendida y_d.

estimadores directos e indirectos

un estimador **directo** es aquel que utiliza únicamente datos de y de los individuos incluidos en la muestra s que pertencen a U_d

$$\hat{Y} = \sum i \in s_d w_i y_i$$

e.g. el estimador HT es directo.

un estimador **indirecto** es aquel que utiliza datos de la y de individuos que no pertenecen al dominio U_d.

- El objetivo es aumentar el tamaño de "muestra efectivo" del dominio para realizar estimaciones

Estimadores de regresión a nivel de dominios

Es necesario tener disponible información auxiliar especifica del dominio.

- Sea $\mathbf{x}_d$ el vector de información auxiliar extendido, el cual, $\mathbf{x}_{di} = \mathbf{x}_i \ \forall i \in U_d$ y $\mathbf{x}_{di}$ en otro caso.
- el total del vector de información auxiliar a nivel del dominio U_d es $\mathbf{X}_d = \sum_{i \in U_d} \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_{di}$.

El estimador GREG para el total de la variable de interés $Y_d = \sum_{i \in U_d} y_i$ queda definido como:

$$\hat{Y}_d^{\text{GREG}} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i (y_i - \hat{y}_i),$$

en donde existe distintas estrategias para definir el modelo de regresión lineal que asiste al estimador GREG

estimador GREG directo

Utilizando un modelo especifico del dominio $E(y_i) = \mathbf{x}_i^T \beta_d$, $V(y_i) = \sigma_i \ \forall i \in U_d$, en donde la estimación de los parámetros del modelo que asiste al estimador son calculados utilizando únicamente información de la variable y en el dominio.

$$\begin{aligned} \hat{Y}_d^{\text{GREG,D}} &= \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i (y_i - \hat{y}_i) \\ &= \mathbf{X}_d^T \hat{\mathbf{B}}_d + \sum_{i \in s_d} w_i (y_i - \mathbf{x}_i^T \hat{\mathbf{B}}_d) \end{aligned}$$

donde

$$\hat{\mathbf{B}}_d = (\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}_i^T / \sigma_i)^{-1} (\sum_{i \in s_d} w_i \mathbf{x}_i y_i / \sigma_i), \qquad (1)$$

estimador GREG directo

- El estimador $\hat{Y}_d^{\text{GREG,D}}$ es un estimador directo por definición ya que el mismo utiliza únicamente información de la variable y del dominio.
- como todo estimador GREG, estima sin error el total de las variables auxiliares utilizadas para su construcción

$$\mathbf{X}_d = \sum_{i \in s_d} w_i^* \mathbf{x}_i$$

,
y se computan ponderadores específicos $w_i^* \ \forall i \in s_d$.

- Otra estrategia es definir un único modelo a nivel de toda la población U y posteriormente utilizar las predicciones del mismo para el dominio de interés U_d
- la varianza del mismo dependerá de las sí el dominio presenta sus propias características.

$$\hat{Y}_d^{\text{GREG,U}} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i (y_i - \hat{y}_i)$$
$$= \mathbf{X}_d^T \hat{\mathbf{B}} + \sum_{i \in s_d} w_i (y_i - \mathbf{x}_i^T \hat{\mathbf{B}})$$

$\hat{Y}^{\text{GREG,U}}$ es **indirecto** debido a que valores de la variable y que no pertenecen al dominio, son utilizados para la estimación de los parámetros del modelo

Si la información utilizada para definir la ecuación de calibración es definida a nivel de U, el estimador puede expresarse como:

$$\hat{Y}_d^{\text{CAL,U}} = \sum_{i \in s_d} w_i^* y_i = \hat{Y}_d^{HT} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{R}}_U,$$

donde

$$\hat{\mathbf{R}}_U = (\sum\nolimits_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\sum\nolimits_{i \in s} w_i \mathbf{x} y_{di}),$$

y

$$w_i^* = w_i [1 + (\mathbf{X} - \hat{\mathbf{X}}^{HT})^T (\sum\nolimits_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} \mathbf{x}_i]$$

estimador calibrado en dominios

- $\hat{Y}_d^{CAL,U}$ por definición es directo, consistente en el diseño y aproximadamente insesgado en el diseño.
- $\hat{Y}_d^{\text{CAL,U}}$ es llamado estimador **uni-weight** debido a que con un único sistema de ponderadores se producen las estimaciones para todos los parámetros y dominios de interés de la encuesta.
- bajo el enfoque de regresión, el estimador $\hat{Y}_d^{\text{CAL,U}}$ puede verse como un estimador GREG a nivel de toda la población U en donde la variable dependiente es y_d.

- un estimador de la varianza de $\hat{Y}_d^{CAL,U}$ viene dada por:

$$\widehat{\text{var}}(\hat{Y}^{\text{CAL,U}}) = \sum_{i \in s} \sum_{j \in s} (\pi_{ij} - \pi_i \pi_j) \pi_{ij}^{-1} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}$$

donde los errores $e_i = y_i - \mathbf{x}_i^T \hat{\mathbf{R}}_U$ si $i \in U_d$ y $e_i = -\mathbf{x}_i^T \hat{\mathbf{R}}_U$ en otro caso.

$\hat{Y}_d^{CAL,P}$ tiene un rendimiento similar al estimador Horvitz-Thompson (el cual no utiliza información auxiliar) para la estimación en los dominios de interés producto de la cantidad de errores o residuos negativos.