Estimador asististido por modelos

Muestreo II

Licenciatura en Estadística

2023

set-up para la estimación

Objetivo: construir estimaciones para cantidades poblacionales

- utilizamos datos recolectados utilizando (generalmente) diseños muestrales complejos
- también tenemos disponible otras fuentes de datos $\mathbf{x}$ (e.g. censos, registros, etc.)
- combinando las dos fuentes anteriores de datos, podemos aumentar la **eficiencia** de nuestras estimaciones!

set-up para la estimación

Universo/Elegibles (e.g. hogares de Montevideo)

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
N-11	N-10	N-9	N-8	N-7	N-6	N-5	N-4	N-3	N-2	N-1	N

$$U = \{1, 2, ..., i, ...N\}$$

set-up para la estimación

Objetivo: estimar el total de la variable de interés (target) y

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}	y_{21}	y_{22}	y_{23}	y_{24}
y_{25}	y_{26}	y_{27}	y_{28}	y_{29}	y_{30}	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}	y_{36}
y_{37}	y_{38}	y_{39}	y_{40}	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}	y_{46}	y_{47}	y_{48}
y_{49}	y_{50}	y_{51}	y_{52}	y_{53}	y_{54}	y_{55}	y_{56}	y_{57}	y_{58}	y_{59}	y_{60}

y_{N-11}	y_{N-10}	y_{N-9}	y_{N-8}	y_{N-7}	y_{N-6}	y_{N-5}	y_{N-4}	y_{N-3}	y_{N-2}	y_{N-1}	y_N

$$Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$$

set-up para la estimación

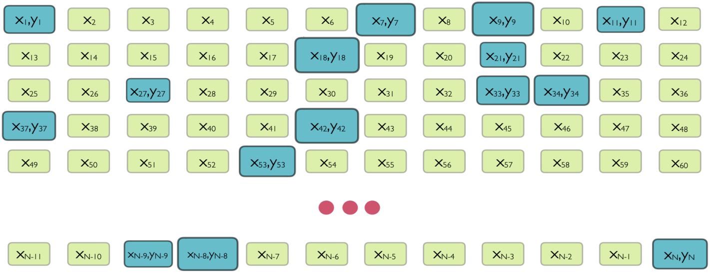
Definimos un diseño muestral en base a los datos disponibles $\mathbf{x}$ (e.g. estratos, MOS, etc.) de forma de asignarle a cada individuo una probabilidad de selección en la muestra (s)

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
π_{13}	π_{14}	π_{15}	π_{16}	π_{17}	π_{18}	π_{19}	π_{20}	π_{21}	π_{22}	π_{23}	π_{24}
π_{25}	π_{26}	π_{27}	π_{28}	π_{29}	π_{30}	π_{31}	π_{32}	π_{33}	π_{34}	π_{35}	π_{36}
π_{37}	π_{38}	π_{39}	π_{40}	π_{41}	π_{42}	π_{43}	π_{44}	π_{45}	π_{46}	π_{47}	π_{48}
π_{49}	π_{50}	π_{51}	π_{52}	π_{53}	π_{54}	π_{55}	π_{56}	π_{57}	π_{58}	π_{59}	π_{60}
					•••						
π_{N-11}	π_{N-10}	π_{N-9}	π_{N-8}	π_{N-7}	π_{N-6}	π_{N-5}	π_{N-4}	π_{N-3}	π_{N-2}	π_{N-1}	π_N

$$\pi_i = \text{Prob}[i \in s] > 0 \ \forall i \in U$$

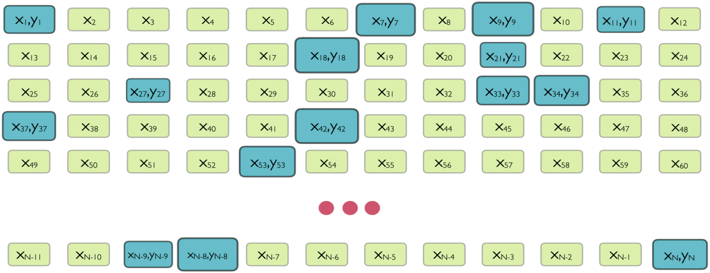
set-up para la estimación

- se selecciona la muestra
- se recolecta información de la variable de interés y e información adicional ($\mathbf{x}$)



estimador HT

los estimadores simples utilizan únicamente información de la variable de interés y de las unidades incluidas en la muestra s



$$\hat{Y}^{\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i \times y_i$$

propiedades

buenas propiedades teóricas

- insesgado $\rightarrow \mathrm{E}(\hat{Y}^{\mathrm{HT}}) = Y$
- varianza "simple"

$$\mathrm{var}(\hat{Y}^{\mathrm{HT}}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j)(y_i \pi_i^{-1})(y_j \pi_j^{-1})$$

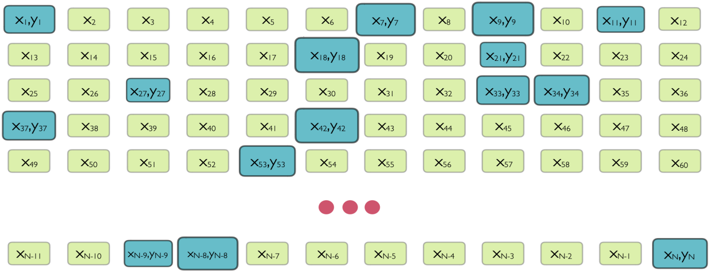
donde $\pi_{ij} = \mathrm{Prob}(i \in s \& j \in s)$

- un estimador insesgado de la varianza (si el diseño lo permite)

$$\widehat{\mathrm{var}}(\hat{Y}^{\mathrm{HT}}) = \sum_{i \in s} \sum_{j \in s} \pi_{ij}^{-1} (\pi_{ij} - \pi_i \pi_j)(y_i \pi_i^{-1})(y_j \pi_j^{-1})$$

estimador de diferencia

supongamos que tenemos una forma o "método" $m(.)$ para poder predecir la variable y, el cual, no depende de la muestra



con las "salidas" del método $m(.)$ se define el estimador de diferencia:

$$\hat{Y}^{\text{DIFF}} = \sum_{i \in U} m(\mathbf{x}_i) + \sum_{i \in s} w_i (y_i - m(\mathbf{x}_i)) = \sum_{i \in U} m(\mathbf{x}_i) + \text{HT}(y - m)$$

propiedades (1)

el estimador es exactamente insesgado independientemente de la calidad del método $m(.)$

$$\mathrm{E}[\hat{Y}^{\mathrm{DIFF}}] = \sum_{i \in U} m(\mathbf{x}_i) + \mathrm{E}[\mathrm{HT}(y - m)] = \sum_{i \in U} m(\mathbf{x}_i) + Y - \sum_{i \in U} m(\mathbf{x}_i) = Y$$

propiedades (2)

Teniendo en cuenta que $\sum_{i \in U} m(\mathbf{x}_i)$ no depende de la muestra(i.e. NO es aleatorio) y la varianza en el diseño es:

$$\text{var}(\hat{Y}^{\text{DIFF}}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{(y_i - m(\mathbf{x}_i))}{\pi_i} \frac{(y_j - m(\mathbf{x}_j))}{\pi_j}$$

- la varianza del estimador será mas pequeña que la del estimador HT si los "residuos" $(y_i - m(\mathbf{x}_i))$ tiene una menor variación que los datos puros y_i
- si el método elegido $m(.)$ tiene un pobre poder predicitivo la varianza del estimador de diferencia será similar a la varianza del estimador HT

propiedades (3)

Un estimador insesgado de la varianza es:

$$\widehat{\text{var}}(\hat{Y}^{\text{DIFF}}) = \sum_{i\in s}\sum_{j\in s} \pi_{ij}^{-1}(\pi_{ij} - \pi_i\pi_j)\frac{(y_i - m(\mathbf{x}_i))}{\pi_i}\frac{(y_j - m(\mathbf{x}_j))}{\pi_j}$$

conclusiones

- si bien el estimador $\hat{Y}^{\text{DIFF}}$ resulta atractivo, el problema se encuentra que en la práctica es difícil tener un método $m(.)$ que sea independiente de la muestra y que proporcione buenas "predicciones" de la variable y
- una alternativa (razonable) es estimar el método $m(.)$ utilizando los datos de la muestra

estimación asistida por modelos

- el estimador $\hat{Y}^{\text{DIFF}}$ necesita un método $m(.)$ que sea independiente de la muestra
- en la práctica tenemos que utilizar los datos de la muestra para construir el método para predecir
- las estimaciones asistidas por modelos abordan este problema introduciendo un "working model"

$$y_i = m(\mathbf{x_i}) + \epsilon_i$$

donde ϵ_i tiene media cero.
- asumimos que $y_i \ \forall i \in U$ son realizaciones de un modelo **superpoblacional**

estimación asistida por modelos

- el modelo superpoblacional elegido no tiene porque ser verdadero para la población U
- es de ayuda que el modelo tenga cierto poder de ajuste respecto a la variable de interés y

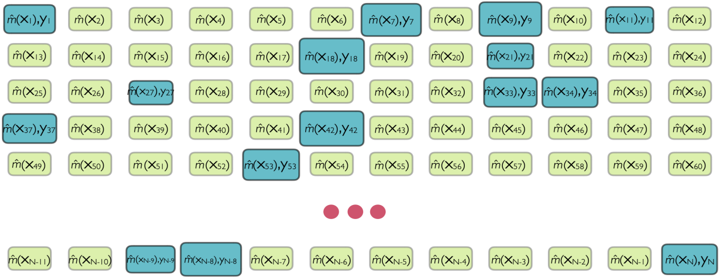
receta

una receta general para la estimación e inferencia utilizando información auxiliar es:

- si $(y_i, \mathbf{x}_i)$ $\forall i \in U$ fueran observadas (conocidas) para toda la población podríamos computar $m(.)$ utilizando métodos estadísticos (e.g. regresión lineal) que son independientes de la muestra
- debido a que en la práctica, únicamente tenemos datos de y para la muestra, el modelo $m(.)$ es desconocido y lo tenemos que estimar $\hat{m}(.)$. Notemos que $\hat{m}(.)$ depende de la muestra seleccionada

la idea

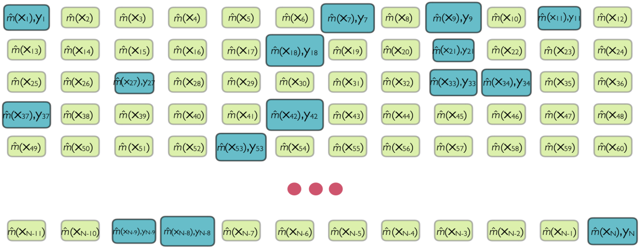
utilizando los datos de la muestra estimamos un modelo $\hat{m}(\mathbf{x}_i)$ para predecir la variable de interés y



$$\hat{y}_i = \hat{m}(\mathbf{x}_i)$$

la idea

construimos un estimador que sea robusto si el modelo/método elegido no es verdadero



Reemplazamos el método $m(\mathbf{x}_i)$ en $\hat{Y}^{\mathrm{DIFF}}$ por su correspondiente estimación $\hat{m}(\mathbf{x}_i)$:

$$\hat{Y}^{\mathrm{PRED}} = \sum_{i \in U} \hat{m}(\mathbf{x}_i) + \sum_{i \in s} w_i (y_i - \hat{m}(\mathbf{x}_i))$$

propiedades

- es asintoticamente insesgado $\rightarrow \mathrm{E}(\hat{Y}^{\mathrm{PRED}}) = Y$
- no tiene una forma exacta de la varianza
- el estimador de la varianza se basa en la formula del estimador de diferencia

$$\widehat{\mathrm{var}}(\hat{Y}^{\mathrm{PRED}}) = \sum_{i \in s} \sum_{j \in s} \pi_{ij}^{-1} (\pi_{ij} - \pi_i \pi_j) \frac{(y_i - \hat{m}(\mathbf{x}_i))}{\pi_i} \frac{(y_j - \hat{m}(\mathbf{x}_j))}{\pi_j}$$

modelos que asisten al estimador

existe una amplia gama de modelos que se han utilizado para asistir al estimador. Algunos...

- **Regresión lineal** (Cassel, Sarndal, y Wretman 1976)
- Regresión logística (Lehtonen y Veijanen 1998)
- Redes nueronales (Montanari y Ranalli 2005)
- Arboles de regresión (McConville y Toth 2020)

modelos que asisten al estimador

Cómo elegimos el modelo que asiste al estimador?

Recordemos que el estimador asistido por modelos es asintoticamente insesgado

- para una amplia gama de de métodos/modelos/algoritmos
- no importa que el modelo elegido sea correcto
- PERO su **precisión** depende del poder predictivo del modelo!

$$\widehat{\text{var}}(\hat{Y}^{\text{PRED}}) = \sum_{i \in s} \sum_{j \in s} \pi_{ij}^{-1}(\pi_{ij} - \pi_i \pi_j) \frac{(y_i - \hat{m}(\mathbf{x}_i))}{\pi_i} \frac{(y_j - \hat{m}(\mathbf{x}_j))}{\pi_j}$$

estimador de regresión

$$\hat{Y}^{\text{PRED}} = \sum_{i \in U} \hat{m}(\mathbf{x}_i) + \sum_{i \in s} w_i (y_i - \hat{m}(\mathbf{x}_i))$$

El método que asiste al estimador es un modelo de regresión lineal:

$$\begin{aligned} y_i &= m(\mathbf{x}) + \epsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_1 x_{iJ} + \epsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \end{aligned}$$

donde $\epsilon_i \sim (0, \sigma_i^2)$

estimador de regresión

el estimador de regresión queda definido como:

$$\begin{aligned}\hat{Y}^{\text{GREG}} &= \sum_{i\in U}\hat{y}_i + \sum_{i\in s} w_i(y_i - \hat{y}_i)\\ &= \sum_{i\in U}\mathbf{x}_i^T\hat{\mathbf{B}} + \sum_{i\in s} w_i(y_i - \mathbf{x}_i^T\hat{\mathbf{B}})\end{aligned}$$

donde

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

es el estimador de β a nivel muestral utilizando el método de mínimos cuadrados ponderados.

casos especiales

- **estimador post-estratificado**: una sola variable categórica de entrada x (e.g. tramo de edad, departamento, sexo)
- **estimador de razón:** una única variable numérica de entrada x en donde el modelo no tiene incerpeto (la recta de regresión pasa por el origen)

ventajas del estimador de regresión

Flexibilidad en el requisito de información auxiliar.

únicamente es necesario conocer:

- $\mathbf{x}_i$ $\forall i \in s$ (i.e. podemos relevar los datos en el formulario de la encuesta).
- conocer los totales de las covariables $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$. No es neceario tener la información a nivel micro!

el estimador de regresión se puede expresar como:

$$\hat{Y}^{\text{GREG}} = \mathbf{X}^T \hat{\mathbf{B}} + \sum_{i \in s} w_i (y_i - \mathbf{x}_i^T \hat{\mathbf{B}})$$

propiedades del estimador de regresión

bajo un modelo de regresión lineal, el estimador puede ser expresado como una suma ponderada:

$$\hat{Y}^{\mathrm{GREG}} = \sum_{i \in s} w_i^* y_i,$$

en donde los ponderadores $w_i^* = g_i w_i$ con un factor de ajuste:

$$g_i = 1 + (\mathbf{X} + \hat{\mathbf{X}}^{\mathrm{HT}})^T (\textstyle\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T / \sigma_i)^{-1} \mathbf{x}_i / \sigma_i,$$

- los ponderadores w_i^* **no dependen** de la variable y
- los podemos utilizar para distintas variables de interés!

propiedades del estimador de regresión

el estimador se encuentra **calibrado** a la información auxiliar utilizada:

$$\sum_{i \in s} w_i^* \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$$

propiedad atractiva para la producción de estadísticas oficiales debido a que brinda consistencia entre datos de distintas fuentes.

problemas

Existen algunos problemas:

- puede existir mucha variabilidad en los ponderadores w_i^* lo que repercute en un aumento innecesario de los SEs
- pueden incluso existir ponderadores w_i^* negativos!
- si el modelo no es correcto puede ocurrir que no se obtenga una mejora alguna en la precisión en comparación al estimador HT