Inferencia basada en modelos: poblaciones homogéneas

Muestreo y Planificación de Encuestas.

Primer Semestre 2023

Poblaciones Homogéneas*

El modelo más sencillo que se puede considerar es aquel en donde no se incluye información auxiliar. En este caso $Y|Z$ no depende de Z, de forma que el modelo para las y_i es igual $\forall i \in U$.

* La presentación se encuentra basada en el libro *An Introduction to Model-Based Survey Sampling* de Chambers y Clark, 2012

Intercambiabilidad

Las variables aleatorias cuya realización son los valores y_i de Y se dicen que son *intercambiables* de orden k si la distribución conjunta de $y_i; k \in A$ es la misma para cualquier permutación A, de las $k = 1, ..., K$ etiquetas de la población.

En una población *intercambiable* todos los momentos de orden k son los mismos.

Si $k \geq 2$ todas las esperanzas y varianzas son iguales para todo $i \in U$, y los pares $i \neq j$ tienen igual covarianza la población se denomina como **homogénea de segundo orden** o simplemente homogénea.

Un modelo para una población homogénea

Asumimos que los y_i son independientes.

$$\begin{cases} E(y_i) = \mu \\ Var(y_i) = \sigma^2 \\ y_i \text{ y } y_j \text{ son independientes cuando } i \neq j \end{cases}$$

Si asumimos que la población tiene una media constante, ¿cuál es el diseño más apropiado para estimar μ?

Un modelo para una población homogénea

Asumimos que los y_i son independientes.

$$\begin{cases} E(y_i) = \mu \\ Var(y_i) = \sigma^2 \\ y_i \text{ y } y_j \text{ son independientes cuando } i \neq j \end{cases}$$

Si asumimos que la población tiene una media constante, ¿cuál es el diseño más apropiado para estimar μ?

Uno que asigne probabilidades constantes a todas las muestras posibles de tamaño n.

Predictor EB

Por definición

$$t_y^* = t_{ys} + E\left[t_{yr} | y_i, i \in s\right]$$

$$t_y^* = t_{ys} + (N-n)\mu$$

Estimo μ con $\hat{\mu} = \bar{y}_s$ y obtengo a $\hat{t}_y^{EB}$:

Estimador de Expansión

$$\hat{t}_y^E = \hat{t}_y^{EB} = t_{ys} + (N-n)\bar{y}_s = \frac{N}{n} t_{ys}$$

Obs: Es igual al estimador π con un diseño SI, pero en este caso la muestra no fue obtenida con un diseño SI necesariamente. Es el mejor estimador cuando la población se comporta de acuerdo al modelo de poblaciones homogéneas.

Estimador BLUP (*Best Linear Unbiased Predictor*)

¿Cómo sabemos si el estimador de expansión es el que "mejor" estima a t_y?

Para que un predictor sea el mejor predictor lineal insesgado se deben cumplir tres cosas:

- Tiene que ser lineal. Puede ser escrito como $\hat{t}_y^{EB} = \sum_s w_i y_i$, donde los w_i son pesos que deben ser determinados. No hay restricciones para los valores de w_i salvo que no pueden depender de los valores de Y.
- Tiene que ser insesgado para estimar t_y, $E(\hat{t}_y^{EB} - t_y) = 0$.
- Para cualquier muestra s su error tiene mínima varianza:

$$Var(\hat{t}_y^{EB} - t_y) \leq Var(\hat{t}_y - t_y)$$

Estimador BLUP (*Best Linear Unbiased Predictor*)

Notemos que cualquier predictor lineal de t_y se puede escribir como:

$$\hat{t}_y = \sum_s w_i y_i = \sum_s y_i + \sum_s (w_i - 1) y_i = t_{ys} + \underbrace{\sum_s u_i y_i}_{\hat{t}_{yr}}$$

con $u_i = w_i - 1$

Consecuentemente, el error de muestreo puede ser expresando como:

$$\hat{t}_y - t_y = \sum_s u_i y_i - \sum_r y_i$$

u_i es el "peso" de la predicción. Para determinar el estimador BLUP debemos determinar los u_i de forma que el predictor sea insesgado y de varianza mínima.

Estimador BLUP (*Best Linear Unbiased Predictor*)

Partiendo de que

$$E(\sum_s u_i y_i - \sum_r y_i) = 0$$

se llega a que para que sea insesgado se debe cumplir que:

$$\sum_s u_i - (N - n) = 0$$

Estimador BLUP (*Best Linear Unbiased Predictor*)

La varianza de la predicción es:

$$Var(\hat{t}_y - t_y) = Var(\hat{t}_{yr} - t_{yr}) = Var(\hat{t}_{yr}) + Var(t_{yr}) - 2Cov(\hat{t}_{yr}, t_{yr})$$

donde

$$Var(\hat{t}_{yr}) = \sigma^2 \sum_r u_i^2$$

$$Var(t_{yr}) = (N-n)\sigma^2$$

y

$$Cov(\hat{t}_{yr}, t_{yr}) = 0$$

Estimador BLUP (*Best Linear Unbiased Predictor*)

Hay un sólo término de $Var(\hat{t}_y - t_y)$ que depende de u_i, o sea que minimizar $Var(\hat{t}_y - t_y)$ es equivalente a minimizar $Var(\hat{t}_{yr})$ sujeto a la restricción $\sum_s u_i - (N - n) = 0$. El problema a resolver es:

$$\min \sum_s u_i^2$$

sujeto a

$$\sum_s u_i - (N - n) = 0$$

entonces

$$\mathcal{L} = \sum_s u_i^2 - 2\lambda\left(\sum_s u_i - (N - n)\right)$$

$$\frac{\partial \mathcal{L}}{\partial u_i} = \sum_s 2u_i - 2\lambda = 0$$

Estimador BLUP (*Best Linear Unbiased Predictor*)

La solución al problema anterior es

$$u_i = \frac{N-n}{n} \Rightarrow w_i = \frac{N}{n}$$

Por lo tanto el estimador $\hat{t}_y^{EB}$ es el estimador de expansión

$$\hat{t}_y^{E} = \frac{N}{n} t_{ys}$$

$$Var(\hat{t}_y^E - t_y) = Var(\hat{t}_{yr}) + Var(t_{yr})$$

Sustituyendo los valores de $Var(\hat{t}_{yr})$ y $Var(t_{yr})$ se llega a que:

$$Var(\hat{t}_y^E - t_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

σ^2 se estima con

$$s_y^2 = \frac{1}{n-1} \sum_s (y_i - \bar{y}_s)^2$$

El estimador de la varianza es

$$\hat{V}(\hat{t}_y^E) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}$$

Estimación de la Varianza e Intervalos de Confianza

Si la muestra es grande, aplico TCL. Se tiene entonces que

$$z = \frac{\hat{t}_y^E - t_y}{\sqrt{\hat{V}(\hat{t}_y^E)}} \sim N(0,1)$$

y los intervalos de confianza son:

$$\hat{t}_y^E \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{t}_y^E)}$$

Generalización del modelo homogéneo.

El modelo se generaliza asumiendo que la covarianza entre y_i y y_j (con $i \neq j$) es una constante.

$$\begin{cases} E(y_i) = \mu \\ \\ Var(y_i) = \sigma^2 \\ \\ Cov(y_i, y_j) = \rho\sigma^2 \text{ para } i \neq j \end{cases}$$

Se puede demostrar que el estimador de expansión sigue siendo el BLUP (ejercicio).