Estimación en dominios en Inferencia basada en Modelos

Muestreo y Planificación de Encuestas.

Primer Semestre 2023

Definición*

¿Qué es un dominio?

Un dominio es un subgrupo de la población de interés para el cual se requiere una estimación separada del total de la variable Y (o de su media, etc.)

Definimos a la variable D como la indicadora del dominio. Sea U_d la notación para indicar la subpoblación correspondiente al dominio d, se define a d_i como:

$$d_i = \begin{cases} 1 & \text{si } i \in U_d \\ 0 & \text{en otro caso} \end{cases}$$

$N_d = \sum_U d_i$ es el tamaño del dominio d, y $t_{dy} = \sum_U d_i y_i$

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

Inferencia en caso de D desconocida para la población.

La situación más frecuente es que conozcamos a D sólo en la muestra. Asumamos que la indicadora del dominio puede ser modelada como N realizaciones iid de una variable aleatoria Bernoulli y que, condicionales a D, los valores poblacionales de Y se encuentran incorrelacionados, con media y varianza constante. Entonces:

$$\begin{cases} E(y_i|d_i = 1) = \mu_d \\ Var(y_i|d_i = 1) = \sigma_d^2 \\ Cov(y_i, y_j|d_i, d_j; i \neq j) = 0 \\ E(d_i) = \theta_d \\ Var(d_i) = \theta_d(1 - \theta_d) \\ Cov(d_i, d_j|i \neq j) = 0 \end{cases}$$

Inferencia en caso de D desconocida para la población.

Asumimos que el diseño es No Informativo, por lo que la muestra y la variable D son independientes, y se puede estimar θ_d con la muestra.

Tenemos que:

$$E(d_i y_i) = E(y_i | d_i = 1) Pr(d_i = 1) = \theta_d \mu_d$$

y

$$Var(d_i y_i) = E(y_i^2 | d_i = 1) Pr(d_i = 1) - E^2(d_i y_i) = \sigma_d^2 \theta_d + \theta_d (1 - \theta_d) \mu_d^2$$

La covarianza es

$$Cov(d_i y_i, d_j y_j | i \neq j) = 0$$

Estimador de expansión para dominios con D desconocida

El predictor óptimo es

$$\hat{t}^{*}_{yd} = t_{y_d s} + E(t_{y_d r} | i \in s) = \sum_s d_i y_i + \sum_r \theta_d \mu_d$$

Si estimamos a θ_d con $\hat{\theta}_d = n_d/n$ y a μ_d con $\bar{y}_{sd} = \sum_s d_i y_i / n_d$, siendo n_d la cantidad de casos en la muestra que corresponden al dominio d, obtenemos el estimador BLUP:

$$\hat{t}^{BLUP}_{yd} = \hat{t}^{E}_{yd} = N \hat{\theta}_d \bar{y}_{sd}$$

que es un caso particular del estimador de expansión.

La varianza es:

$$Var(\hat{t}^{E}_{yd} - t_{yd}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \left(\sigma^2_d \theta_d + \theta_d (1 - \theta_d) \mu^2_d\right)$$

Inferencia en caso de D desconocida para la población.

El estimador de la varianza es:

$$\hat{V}(\hat{t}_{yd}^{E}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)s_{dy}^2$$

con

$$s_{dy}^2 = \frac{\sum_s (d_i y_i - \hat{\theta}_d \bar{y}_{sd})}{n-1}$$

De forma alternativa, tomando

$$\hat{\sigma}_d^2 = \frac{\sum_s (y_i - \bar{y}_{sd})}{n_d - 1}$$

se llega a que (ejercicio):

$$\hat{V}(\hat{t}_{yd}^{E}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{n}{n-1}\left\{\frac{n_d-1}{n_d}\hat{\theta}_d\hat{\sigma}_d^2 + \hat{\theta}_d(1-\hat{\theta}_d)\bar{y}_{sd}^2\right\}$$

Inferencia en caso de D conocida para la población.

Cuando D es conocida para toda la población (D_U) se tiene que

$$\begin{cases} E(y_i) = \mu_d \\ Var(y_i) = \sigma_d^2 \\ Cov(y_i, y_j | i \neq j) = 0 \end{cases}$$

En este caso se tiene un modelo de población homogénea para cada dominio, por lo que también lleva a un caso particular del estimador de expansión:

Estimador de expansión para dominios con D_U conocida

$$\hat{t}_{yd}^E = N_d \bar{y}_{sd}$$

y

$$\hat{V}(\hat{t}_{yd}^E) = N_d^2 \left(1 - \frac{n_d}{N_d} \frac{\hat{\sigma}_d^2}{n_d}\right)$$

Predicción para Áreas Pequeñas

Definición

Un área pequeña es el caso de un dominio con muy pocos casos en la muestra. Al tener un tamaño de muestra pequeño, las estimaciones son muy inestables.

Se tiene una muestra s de U que se encuentra particionada en $d = 1, \ldots, D$ áreas. El objetivo es predecir el valor de la media de una variable Y en cada una de ellas.

En el caso que la muestra contenga casos del área d, podemos estimar la media de dos formas:

- En el caso que no conozca N_d, utilizamos el estimador de Hájek

$$\hat{\bar{y}}_d^{Ha} = \left(\sum_{sd} w_i\right)^{-1} \sum_{sd} w_i y_i$$

- Si se conoce N_d, utilizamos el estimador HT

$$\hat{\bar{y}}_d^{HT} = \frac{\sum_{sd} w_i y_i}{N_d}$$

Predicción para Áreas Pequeñas

Los pesos w_i son pesos muestrales. Ambos estimadores son los basados en el diseño, y en contexto de inferencia basada en modelos, se les llama "directos".

La idea central de la estimación en áreas pequeñas es mitigar el problema de los pocos casos en la muestra "fortaleciendo" (*strengthen*) los datos de la muestra con la ayuda de un conjunto de variables auxiliares **Z**.

Los valores de **Z** pueden ser por ejemplo:

- Valores de la variable *Y* en el pasado.
- Valores de la variable *Y* en unidades vecinas.
- Cualquier variable que se encuentre relacionada con la variable de interés.

El punto clave de este enfoque es el supuesto de que alguno de los parámteros del modelo que relaciona a *Y* con **Z** es común a todas las áreas, entonces para la inferencia en un área particular se puede "pedir prestada fuerza" (*borrow strength*) de otras áreas usando datos de toda la muestra para la estimación de los parámetros comunes.

Existen dos formas de pedir prestada "fuerza" al resto de la muestra:

- Métodos sintéticos: es la situación en donde el modelo asumido para un área particular funciona en todas. Es decir, la variabilidad entre áreas se encuentra explicada por el conjunto de variables **Z** del modelo.
- *Small Area Models*: para cada área funciona un modelo diferente. Se puede tener un conjunto de parámetros común, pero el conjunto de variables **Z** no captura completamente la variabilidad entre áreas. Se utiliza en este caso Modelos Mixtos.

Supongamos que el modelo superpoblacional es un modelo lineal general, tal que:

$$Y_U = \mathbf{Z}_U \beta + \mathbf{e}_U$$

Dada una estimación $\hat{\beta}$ de β, se puede predecir el promedio $\bar{y}_d$, asumiendo que el modelo funciona para todas las áreas. De esta forma se obtiene el predictor *sintético* de $\bar{y}_d$.

$$\hat{\bar{y}}_d^{SYN} = N_d^{-1} \left(\sum_{sd} y_i + \sum_{rd} z_i' \hat{\beta} \right) = \hat{\beta} \bar{z}_d + N_d^{-1} n_d (\bar{y}_{sd} - \hat{\beta} \bar{z}_{sd})$$

Necesito conocer a N_d y a $\bar{z}_d$. ¿Qué sucede si $n_d = 0$? O sea, si no cae ninguna observación en el área d.

$$\hat{\bar{y}}_d^{SYN} = \hat{\beta} \bar{z}_d$$

Puedo predecir un valor para el área d aunque no tenga datos en la muestra del área de interés.