Poblaciones con estructura de regresión

Muestreo y Planificación de Encuestas.

Primer Semestre 2023

Predicción óptima bajo una relación proporcional*

Modelo de Población *Gamma*

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

El parámetro γ controla cuánto depende la varianza de la variable auxiliar Z y por lo general $0 \leq \gamma \leq 1$.

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

Predicción óptima bajo una relación proporcional

Cuando $\gamma = 0.5$ se tiene el Modelo de Razón Poblacional.

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

Por definición

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

En el caso del modelo de razón:

$$t_y^* = t_{ys} + \beta t_{zr}$$

Estimo β con $b = \hat{\beta}$ y obtengo a $\hat{t}_y^{EB}$

Predictor BLUP bajo el modelo poblacional gamma

$$\hat{t}_y = t_{ys} + t_{zr} \sum_s z_i^{1-2\gamma} y_i / \sum_s z_i^{2-2\gamma}$$

Obs: Se llega a la fórmula del estimador asumiendo que b es un estimador lineal, insesgado y de varianza mínima (se plantea el Lagrangiano y se obtiene como solución que $b = \sum_s z_i^{1-2\gamma} y_i / \sum_s z_i^{2-2\gamma}$)

Modelo de Razón

En el Modelo de Razón $\gamma = 0.5$, entonces:

$$b = \frac{\bar{y}_s}{\bar{z}_s} \quad \text{y}$$

$$\hat{t}_y^R = t_{ys} + b t_{zr} = \frac{\bar{y}_s}{\bar{z}_s} t_z$$

Si se quiere expresar como una combinación lineal $\hat{t}_y^R = \sum_s w_i y_i$ los pesos w_i son:

$$w_i = \frac{N \bar{z}_U}{n \bar{z}_s}$$

El Modelo de Razón es justificable cuando:

- Existe una relación proporcional entre y y z.
- Como mínimo se conocen los totales de z.
- Preferiblemente, se conoce $z \forall k \in U$.

Predicción óptima bajo una relación lineal

El modelo superpoblacional ahora es:

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

α y β se estiman por MCO con:

$$a_L = \hat{\alpha} = \bar{y}_s - b_L \bar{z}_s$$

y

$$b_L = \hat{\beta} = \frac{\sum_s (y_i - \bar{y}_s)(z_i - \bar{z}_s)}{\sum_s (z_i - \bar{z}_s)^2}$$

Predictor EB y EBLUP

El predictor EB es:

$$\hat{t}_y^{EB} = t_{ys} + \sum_r \alpha + \beta z_i$$

Estimador de regresión

El estimador EBLUP se obtiene estimando los parámetros α y β y sustituyendo en el estimador *EB*. En este caso particular, al resultado se lo denomina como *Estimador de Regresión*.

$$\hat{t}_y^{L} = t_{ys} + \sum_r a_L + b_L z_i = N(a_L + b_L \bar{z}_U)$$

Se puede demostrar que es el estimador BLUP con pesos (ejercicio)

$$w_i = \frac{N}{n}\left(1 + \frac{(\bar{z}_U - \bar{z}_s)(z_i - \bar{z}_s)}{(1 - n^{-1})s_z^2}\right)$$

Con cualquiera de los modelos de regresión utilizados, el supuesto de intercambiabilidad no se cumple (los momentos de primer y segundo orden dependen del valor de z_i). Entonces el diseño SI deja de ser el diseño más "acorde" a esta situación.

¿Cómo seleccionamos la muestra ahora?

Minimizando la varianza del predictor. En el caso del Modelo de Razón es:

$$Var(\hat{t}_y^R - t_y) = Var\left(b\sum\nolimits_r z_i - \sum\nolimits_r y_i\right) = \sigma^2 N\bar{z}_U\left(\frac{\sum_r z_i}{\sum_s z_i}\right)$$

$$\Rightarrow Var(\hat{t}_y^R - t_y) = \sigma^2\frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{\bar{z}_U\bar{z}_r}{\bar{z}_s}$$

Con este resultado, ¿qué diseño minimiza la varianza?

Extreme sampling, la muestra con los n valores más grandes de la variable Z.

La varianza del modelo se estima con:

$$\hat{\sigma}^2 = n^{-1} \sum_s \left\{ z_i \left(1 - \frac{z_i}{n\bar{z}_s} \right) \right\}^{-1} (y_i - bz_i)^2$$

Entonces el estimador de la varianza del predictor es:

$$\hat{V}(\hat{t}_y^R) = \hat{\sigma}^2 N \bar{z}_U \left(\frac{\sum_r z_i}{\sum_s z_i} \right)$$

Aplicando TCL se obtienen los intervalos de confianza para $\hat{t}_y^R$.

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

De la misma manera que para el Modelo de Razón, buscamos minimizar la varianza del predictor:

$$Var(\hat{t}_y^L - t_y) = Var\left(\sum_r a_L + b_L z_i - \sum_r y_i\right) = \frac{N^2}{n}\sigma^2\left[\left(1 - \frac{n}{N}\right) + \frac{(\bar{z}_U - \bar{z}_s)^2}{(1 - n^{-1})s_z^2}\right]$$

La demostración del resultado queda como ejercicio.

¿Cuando se minimiza la varianza?

Cuando la muestra se encuentra *balanceada*. Si se cumple que $\bar{z}_U = \bar{z}_s$, el estimador de regresión coincide con el de expansión (ejercicio).

El estimador de σ^2 es:

$$\hat{\sigma}^2 = \frac{\sum_s (y_i - a_l - b_l z_i)^2}{n-2}$$

Entonces, el estimador de la varianza del predictor es:

$$\hat{V}(\hat{t}_y^L) = \frac{N^2}{n}\hat{\sigma}^2 \left[\left(1 - \frac{n}{N}\right) + \frac{(\bar{z}_U - \bar{z}_s)^2}{(1-n^{-1})s_z^2} \right]$$

Cuando la muestra se encuentra balanceada, ¿la estimación de la varianza del predictor coincide con la del estimador de expansión?