Muestreo asistido por modelos

Daniel Czarnievicz

2017

El objetivo es utilizar información auxiliar para construir estimadores más precisos que el estimador $\hat{t}_\pi$. Asumimos que tenemos información complete para J variables auxiliares x, las cuales covarían con y. Para el k-ésimo elemento definimos el vector $\mathbf{x_k} = (x_{1;k}; \ldots; x_{j;k}; \ldots; x_{J;k})'$.

Se extrae una muestra s según el diseño $p(.)$ con $\pi_j > 0$ y $\pi_{kl} > 0 \ \forall k; l \in U$. A partir de esta muestra se desea estimar $t_y = \sum_U y_k$.

El estimador de diferencia

La idea principal de este estimador es utilizar la información auxiliar para formar un set de N valores proxy para la variable de análisis: $y_1^0; \ldots; y_k^0; \ldots; y_N^0$, de forma tal que sean buenas aproximaciones para los valores $y_1; \ldots; y_k; \ldots; y_N$. Una opción es construir estos valores proxy a partir de una combinación lineal de las J variables auxiliares:

$$y_k^0 = \sum_{j=1}^{J} A_j \, x_{jk} = \mathbf{A}' \, \mathbf{x_k}$$

donde A_j son coeficientes conocidos $\forall j$. De esta forma, y_k^0 puede calcularse para toda la población (dado que los valores de las variables auxiliares se asumen disponibles para toda la población). Dado esto, el total poblacional podría expresarse como:

$$t_y = \sum_U y_k = \sum_U y_k^0 + \sum_U \left(y_k - y_k^0\right) = \sum_U y_k^0 + \sum_U D_k$$

$\sum_U y_k^0$ es conocida para toda la población, pero $\sum_U D_k$ no lo es, y debe estimarse. Para esto se utiliza la estimación π, de forma de lograr un estimador insesgado, el cual se conoce como estimador de diferencias:

$$\star \ \hat{t}_{y_{diff}} = \sum_U y_k^0 + \sum_s D_k^{\checkmark}$$

$$\star \ E\left(\hat{t}_{y_{diff}}\right) = E\left(\sum_U y_k^0 + \sum_s D_k^{\checkmark}\right) = E\left(\sum_U y_k^0\right) + E\left(\sum_s D_k^{\checkmark}\right) =$$

$$= \sum_U y_k^0 + \sum_U D_k = t_y$$

$$\star \ V\left(\hat{t}_{y_{diff}}\right) = \sum\sum_U \Delta_{kl} \, D_k^{\checkmark} \, D_l^{\checkmark}$$

$$\star \ \hat{V}\left(\hat{t}_{y_{diff}}\right) = \sum\sum_s \Delta_{kl}^{\checkmark} \, D_k^{\checkmark} \, D_l^{\checkmark}$$

$$\star \ E\left[\hat{V}\left(\hat{t}_{y_{diff}}\right)\right] = \sum\sum_U \Delta_{kl} \, D_k^{\checkmark} \, D_l^{\checkmark}$$

Si el diseño es de tamaño fijo, entonces se cumple que:

$$\star \ V\left(\hat{t}_{y_{diff}}\right) = -\frac{1}{2} \sum\sum_U \Delta_{kl} \left(D_k^{\checkmark} - D_l^{\checkmark}\right)^2$$

$$\star \ V\left(\hat{t}_{y_{diff}}\right) = -\frac{1}{2} \sum\sum_s \Delta_{kl}^{\checkmark} \left(D_k^{\checkmark} - D_l^{\checkmark}\right)^2$$

Alternativamente, el estimador $\hat{t}_{y_{diff}}$ podría considerarse como una mejora sobre el estimador π. Si los valores de la variable proxy se generar como una combinación lineal, entonces:

$$\star \ \hat{t}_{y_{diff}} = \hat{t}_{y_\pi} + \sum_{j=1}^{J} A_j (t_{x_j} - \hat{t}_{x_{j\pi}})$$

El estimador de regresión

Si los coeficientes A_j no son conocidos, entonces estos deben ser estimados. Sus estimaciones serán $\hat{B}_1; \ldots; \hat{B}_J$. El estimador de regresión será entonces:

$$\star\ \hat{t}_{y_r} = \hat{t}_{y_\pi} + \sum_{j=1}^{J} \hat{B}_j \left(t_{x_j} - \hat{t}_{x_{j\pi}}\right)$$

donde:

$$\hat{\mathbf{B}} = \left(\sum\nolimits_s \frac{\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)^{-1} \left(\sum\nolimits_s \frac{\mathbf{x_k}\,y_k}{\sigma_k^2\,\pi_k}\right) = \hat{\mathbf{T}}^{-1}\,\hat{\mathbf{t}}$$

Las estimaciones de los parámetros B_j se obtienen como el resultado de ajustar un modelo con las siguientes características:

- $y_1 : \ldots; y_N$ son realizaciones de N variables aleatorias iid $Y_1; \ldots; Y_N$
- $E_\xi(Y_k) = \sum_{j=1}^{J} \beta_j\, x_{jk}\ \ \forall k \in U$
- $V_\xi(Y_k) = \sigma_k^2\ \ \forall k \in U$

En un censo, el estimador MCP del vector β sería:

$$\mathbf{B} = \left(\sum\nolimits_U \frac{\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2}\right)^{-1} \left(\sum\nolimits_U \frac{\mathbf{x_k}\,y_k}{\sigma_k^2}\right) \Rightarrow \mathbf{B} = \left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y}\right) = \mathbf{T}^{-1}\mathbf{t}$$

donde:

$$\mathbf{\Sigma} = diag(\sigma_1^2; \ldots; \sigma_N^2) \Rightarrow \mathbf{\Sigma}^{-1} = diag(\sigma_1^{-2}; \ldots; \sigma_N^{-2})$$

Si el muestreo se hace con remplazo, el estimador de regresión puede construirse utilizando el estimador *pwr* en lugar del estimador π.

El modelo ajustado para la muestra s produce la estimación de $\mathbf{B}$, los valores ajustados $\hat{y}_k = \mathbf{x_k}'\hat{\mathbf{B}}$ para cada elemento de la población, y los residuos $e_{k_s} = y_k - \hat{y}_k$ para cada elemento en la muestra. El estimador de regresión puede expresarse también como:

$$\star\ \hat{t}_{y_r} = \sum\nolimits_U \hat{y}_k + \sum\nolimits_s e_{k_s}^{\checkmark}$$

El término de ajuste desaparece en muchas aplicaciones, aún cuando el ajuste del modelo no sea perfecto. Por ejemplo, si $\sigma_k^2 = \lambda'\,\mathbf{x_k} \Rightarrow \sum_s e_{k_s}^{\checkmark} = 0$.

$$\sum\nolimits_s e_{k_s}^{\checkmark} = \sum\nolimits_s y_k^{\checkmark} - \left(\sum\nolimits_s \frac{\mathbf{x_k}'}{\pi_k}\right)\hat{\mathbf{B}} = \sum\nolimits_s y_k^{\checkmark} - \left(\sum\nolimits_s \frac{\sigma_k^2\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)\hat{\mathbf{B}} = \sum\nolimits_s y_k^{\checkmark} - \left(\sum\nolimits_s \frac{\lambda'\,\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)\hat{\mathbf{B}} =$$

$$= \sum\nolimits_s y_k^{\checkmark} - \lambda'\left(\sum\nolimits_s \frac{\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)\left(\sum\nolimits_s \frac{\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)^{-1}\left(\sum\nolimits_s \frac{\mathbf{x_k}\,y_k}{\sigma_k^2\,\pi_k}\right) =$$

$$= \sum\nolimits_s y_k^{\checkmark} - \lambda'\left(\sum\nolimits_s \frac{\mathbf{x_k}\,y_k}{\sigma_k^2\,\pi_k}\right) = \sum\nolimits_s y_k^{\checkmark} - \sum\nolimits_s \frac{\lambda'\,\mathbf{x_k}\,y_k}{\sigma_k^2\,\pi_k} = \sum\nolimits_s y_k^{\checkmark} - \sum\nolimits_s \frac{\sigma_k^2\,y_k}{\sigma_k^2\,\pi_k} =$$

$$= \sum\nolimits_s y_k^{\checkmark} - \sum\nolimits_s \frac{y_k}{\pi_k} = \sum\nolimits_s y_k^{\checkmark} - \sum\nolimits_s y_k^{\checkmark} = 0$$

Este supuesto respecto de la estructura de σ_k^2 se cumplirá si:

- $\sigma_k^2 = \sigma^2 \ \forall k \in U$ y $x_{1k} = 1 \ \forall k \in U$
- $\exists\, x_j \ / \ \sigma_k^2 \propto x_{jk} \ \forall k \in U$
- $\sigma_k^2 \propto \sum_{j=1}^{J} a_j \, x_{jk} \ \forall k \in U$

Sean $\mathbf{t}_x = \left(t_{x_1}; \ldots; t_{x_J}\right)'$ el vector J-dimensional de los totales de las J variables auxiliares, y $\hat{\mathbf{t}}_{x_\pi} = \left(\hat{t}_{x_1\pi}; \ldots; \hat{t}_{x_J\pi}\right)'$ el vector J-dimensional de sus estimadores π, entonces el estimador de regresión puede escribirse como:

$$\star \ \hat{t}_{y_r} = \hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{B}} = \sum\nolimits_s y_k^{\checkmark} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\left(\sum\nolimits_s \frac{\mathbf{x_k}\, y_k^{\checkmark}}{\sigma_k^2}\right) \Rightarrow$$

$$\Rightarrow \hat{t}_{y_r} = \sum\nolimits_s y_k^{\checkmark} \underbrace{\left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x_k}}{\sigma_k^2}\right]}_{g_{k_s}} \Rightarrow \boxed{\hat{t}_{y_r} = \sum\nolimits_s y_k^{\checkmark}\ g_{k_s}}$$

La última forma de expresar el estimador de regresión utiliza los valores ajustados $y_k^0 = \mathbf{x_k}'\mathbf{B}$ y los residuos del modelo ajustado: $E_k = y_k - y_k^0$. Dado que $y_k = y_k^0 + E_k$, el estimador de regresión toma la forma:

$$\star \ \hat{t}_{y_r} = \sum\nolimits_s g_{k_s}\left(y_k^{0^{\checkmark}} + E_k^{\checkmark}\right)$$

$$\sum\nolimits_s g_{k_s}\, \mathbf{x_k}^{\checkmark\prime} = \sum\nolimits_s \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x_k}}{\sigma_k^2}\right]\frac{\mathbf{x_k}'}{\pi_k} =$$

$$= \sum\nolimits_s \mathbf{x_k}^{\checkmark\prime} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\underbrace{\left(\sum\nolimits_s \frac{\mathbf{x_k}\,\mathbf{x_k}'}{\sigma_k^2\,\pi_k}\right)}_{\hat{\mathbf{T}}} = \sum\nolimits_s \mathbf{x_k}^{\checkmark\prime} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)' =$$

$$= \hat{\mathbf{t}}'_{x_\pi} + \mathbf{t}'_x - \hat{\mathbf{t}}'_{x_\pi} = \mathbf{t}'_x = \sum\nolimits_U \mathbf{x_k}'$$

Post multiplicando por $\mathbf{B}$ obtenemos:

$$\sum\nolimits_s g_{k_s} \underbrace{\mathbf{x_k}^{\checkmark\prime}\mathbf{B}}_{y_k^{0^{\checkmark}}} = \sum\nolimits_U \underbrace{\mathbf{x_k}'\mathbf{B}}_{y_k^0} \Rightarrow \sum\nolimits_s g_{k_s}\, y_k^{0^{\checkmark}} = \sum\nolimits_U y_k^0$$

Por lo tanto,

$$\hat{t}_{y_r} = \sum\nolimits_s g_{k_s}\left(y_k^{0^{\checkmark}} + E_k^{\checkmark}\right) = \sum\nolimits_s g_{k_s}\, y_k^{0^{\checkmark}} + \sum\nolimits_s g_{k_s}\, E_k^{\checkmark} \Rightarrow$$

$$\Rightarrow \boxed{\hat{t}_{y_r} = \sum\nolimits_U y_k^0 + \sum\nolimits_s g_{k_s}\, E_k^{\checkmark}}$$

La varianza del estimador de regresión

El estimador de regresión no es insesgado, pero es aproximadamente insesgado para muestras grandes. Él mismo puede aproximarse por un desarrollo de Taylor de primer orden:

$$\hat{t}_{y_0} = \hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\mathbf{B} = \sum\nolimits_U y_k^0 + \sum\nolimits_s E_k^{\checkmark}$$

Demostración:

$$\hat{t}_{y_r} = \hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} = f(\hat{t}_{y_\pi}; \hat{t}_{x_\pi}; \hat{\mathbf{T}}; \hat{\mathbf{t}})$$

- $\star\ \dfrac{\partial f}{\partial \hat{t}_{y_\pi}} = 1$
- $\star\ \dfrac{\partial f}{\partial \hat{t}_{x_\pi}} = -\hat{B}_j\ \ \forall j = 1; \ldots; J$
- $\star\ \dfrac{\partial f}{\partial \hat{t}_{jj'_\pi}} = \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\left(-\hat{\mathbf{T}}^{-1}\,\mathbf{\Lambda}_{jj'}\,\hat{\mathbf{T}}^{-1}\right)\hat{\mathbf{t}}\ \ \forall j \leq j' = 1; \ldots; J$
- $\star\ \dfrac{\partial f}{\partial \hat{t}_{j0_\pi}} = \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\hat{\mathbf{T}}^{-1}\lambda_j\ \ \forall j = 1; \ldots; J$ donde λ_j es un vector J-dimensional con valor 1 en la j-ésima componente, y valor 0 en las demás. $\Lambda_{jj'}$ es una matriz de tamaño $J \times J$ con valor 1 en los elementos (jj') y $(j'j)$ y 0 en los demás componentes. Evaluamos en $\hat{\theta} = \theta$ para obtener:

$$\star\ \hat{t}_{y_r} \doteq \hat{t}_{y_{r_0}} = t_y + \mathbf{1}\left(\hat{t}_{y_\pi} - t_y\right) - \sum_{j=1}^{J} B_j\left(\hat{t}_{x_j\pi} - t_{x_j}\right) =$$

$$= \hat{t}_{y_\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}\right)'\mathbf{B} = \sum\nolimits_U y_k^0 + \sum\nolimits_s E_k^{\checkmark}$$

$$\star\ E\left(\hat{t}_{y_r}\right) \doteq E\left(\hat{t}_{y_{r_0}}\right) = E\left(\sum\nolimits_U y_k^0 + \sum\nolimits_s E_k^{\checkmark}\right) = \sum\nolimits_U y_k^0 + E\left(\sum\nolimits_s E_k^{\checkmark}\right) =$$

$$= \sum\nolimits_U y_k^0 + \sum\nolimits_U E_k = t_y$$

$$\star\ AV\left(\hat{t}_{y_r}\right) \doteq V\left(\hat{t}_{y_{r_0}}\right) = V\left(\sum\nolimits_s E_k^{\checkmark}\right) = \sum\sum\nolimits_U \Delta_{kl}\, E_k^{\checkmark}\, E_l^{\checkmark}$$

$$\star\ \hat{V}\left(\hat{t}_{y_r}\right) = \sum\sum\nolimits_s \Delta_{kl}^{\checkmark}\, E_k^{\checkmark}\, E_l^{\checkmark}$$

$$\star\ IC_{t_y}^{(1-\alpha)100\,\%} = \left[\hat{t}_{y_r} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{t}_{y_r}\right)}\right]$$