El estimador calibrado

Daniel Czarnievicz

2017

El estimador calibrado se puede utilizar cuando la información auxiliar se presenta en forma de conteos marginales en una tabla de frecuencias de dos variables. Los ponderadores del estimador calibrado reproducen los conteos marginales. La observación y_k tiene asociado el siguiente vector de información auxiliar: $\mathbf{x}_k = (x_{k1}; \ldots; x_{kj}; \ldots; x_{kJ})'$. Por lo tanto, $\mathbf{t}_x = \sum_U \mathbf{x}_k = (t_{x_1}; \ldots; t_{x_j}; \ldots; t_{x_J})'$, el cual puede ser visto como $(N\bar{x}_1; \ldots; N\bar{x}_j; \ldots; N\bar{x}_J)'$.

Se toma una muestra aleatoria de la población, donde $d_k = {}^1/_{\pi_k}$, por lo tanto, $\hat{t}_{y_\pi} = \sum_s d_k\, y_k$. El objetivo es modificar d_k en función de la información auxiliar de forma que los nuevos pesos, w_k, sean cercanos a los pesos originales d_k. Estos nuevos pesos se obtienen minimizando una función de distancia sujeto a $\sum_s w_k\, \mathbf{x}_k = \sum_U \mathbf{x}_k$. Para ello consideramos una función de distancia $G({}^{w_k}/_{d_k})$ con las siguientes propiedades:

- G es positiva y estrictamente convexa.
- $G(1) = G'(1) = 0$
- $G''(1) = 1$

G mide la distancia desde los pesos originales, d_k, a los nuevos pesos, w_k. La medida de distancia para toda la muestra s viene dada por: $\sum_s d_k\, G({}^{w_k}/_{d_k})$. El problema de optimización es por tanto:

$$\min_{w_k} \left\{ \mathscr{L} = \sum\nolimits_s d_k\, G({}^{w_k}/_{d_k}) - \boldsymbol{\lambda}' \left(\sum\nolimits_s w_k\, \mathbf{x}_k - \sum\nolimits_U \mathbf{x}_k \right) \right\}$$

$$\frac{\partial \mathscr{L}}{\partial w_k} = d_k\, \frac{\mathrm{d}\, G({}^{w_k}/_{d_k})}{\mathrm{d}({}^{w_k}/_{d_k})}\, \frac{1}{d_k} - \mathbf{x}_k'\, \boldsymbol{\lambda} = \mathbf{0} \Rightarrow \frac{\mathrm{d}\, G({}^{w_k}/_{d_k})}{\mathrm{d}({}^{w_k}/_{d_k})} = \mathbf{x}_k'\, \boldsymbol{\lambda} \Rightarrow$$

$$\Rightarrow g({}^{w_k}/_{d_k}) = \mathbf{x}_k'\, \boldsymbol{\lambda} \Rightarrow {}^{w_k}/_{d_k} = g^{-1}(\mathbf{x}_k'\, \boldsymbol{\lambda}) \Rightarrow \boxed{w_k = d_k\, F(\mathbf{x}_k'\, \boldsymbol{\lambda})} \quad (1)$$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\lambda}} = \sum\nolimits_s w_k\, \mathbf{x}_k - \sum\nolimits_U \mathbf{x}_k = \mathbf{0} \Rightarrow \boxed{\sum\nolimits_s w_k\, \mathbf{x}_k = \sum\nolimits_U \mathbf{x}_k} \quad (2)$$

Para hallar los nuevos pesos, primero debemos derivar $\boldsymbol{\lambda}$, resolviendo (generalmente) de forma númerica las **ecuaciones de calibración**:

$$\left.\begin{matrix} (1) \\ (2) \end{matrix}\right\} \Rightarrow \sum\nolimits_s d_k\, F(\mathbf{x}_k'\, \boldsymbol{\lambda})\, \mathbf{x}_k = \sum\nolimits_U \mathbf{x}_k$$

Una vez determinado $\boldsymbol{\lambda} = (\lambda_1; \ldots; \lambda_J)'$, podemos derivar los **pesos calibrados** a partir de la ecuación (1).

El estimador calibrado viene dado por:

$$\star\ \hat{t}_{y_{cal}} = \sum\nolimits_s w_k\, y_k = \sum\nolimits_s d_k\, F(\mathbf{x}_k'\, \boldsymbol{\lambda})\, y_k$$

Medidas de distancia

Sean $x = {}^{w_k}/_{d_k}$, G una función de distancia, y F la inversa de su derivada (de argumento u).

1. Método lineal:
   - $G(x) = \frac{1}{2}(x-1)^2 \; x \in \mathbb{R}$
   - $F(u) = 1 + u \; u \in \mathbb{R}$
2. Método multiplicativo:
   - $G(x) = x \log x - x + 1 \; x > 0$
   - $F(u) = e^u \; u > 0$
3. Método logit $(L; U)$:
   - Sean L y U dos números reales tales que $L < 1 < U$
   - Sea $A = \dfrac{U-L}{(1-L)(U-1)}$
   - $G(x) = \begin{cases} \dfrac{1}{A}\left[(x-L)\log\left(\dfrac{x-L}{1-L}\right) + (U-x)\log\left(\dfrac{U-x}{U-1}\right)\right] & \text{si } L < x < U \\ +\infty \;\; \text{en otro caso} \end{cases}$
   - $F(u) = \dfrac{L(U-1) + U(1-L)e^{Au}}{U-1+(1-L)e^{Au}} \; u \in (L; U)$
4. Método lineal truncado $(L; U)$:
   - Sean L y U dos números reales tales que $L < 1 < U$
   - $G(x) = \begin{cases} \dfrac{1}{2}(x-1)^2 & \text{si } L < x < U \\ +\infty \;\; \text{en otro caso} \end{cases}$
   - $F(u) = \begin{cases} L & \text{si } u < L-1 \\ 1+u & \text{si } u \in [L-1, U-1] \\ U & \text{si } u > U-1 \end{cases}$

Estimación de la varianza

En el método lineal[1] $w_k = d_k(1 + \mathbf{x}_k' \boldsymbol{\lambda})$, donde $\boldsymbol{\lambda}$ es la solución a:

$$\left(\sum\nolimits_s d_k \, \mathbf{x}_k \, \mathbf{x}_k'\right) \boldsymbol{\lambda} = \mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi}$$

El estimador generalizado de regresión es entonces:

$$\hat{t}_{y_{greg}} = \sum\nolimits_s w_k \, y_k = \hat{t}_{y_\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x_\pi})' \hat{\mathbf{B}}$$

donde $\hat{\mathbf{B}}$ es la solución a las ecuaciones normales:

$$\left(\sum\nolimits_s d_k \, \mathbf{x}_k \, \mathbf{x}_k'\right) \hat{\mathbf{B}} = \sum\nolimits_s sss d_k \, \mathbf{x}_k \, \mathbf{y}_k$$

[1] Las propiedades para muestras grandes coinciden en todos los métodos.

Su varianza aproximada viene dada por:

$$\star\ AV(\hat{t}_{y_{greg}}) = \sum\sum_U \Delta_{kl}(d_k\, E_k)(d_l\, E_l) \quad \text{donde } E_k = y_k - \mathbf{x}_k'\mathbf{B}$$

Un estimador para su varianza aproximada viene dada por:

$$\star\ \hat{V}(\hat{t}_{y_{greg}}) = \sum\sum_s \Delta_{kl}^{\checkmark}(w_k\, e_k)(w_l\, e_l) \quad \text{donde } e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$$

Post-estratificación completa

Sea una tabla de datos con r filas y c columnas ($r \times c$ celdas). U_{ij} contiene N_{ij} elementos, $i = 1; \ldots; r$, $j = 1; \ldots; c$, de forma tal que $N = \sum_i \sum_j N_{ij}$. N_{ij} es conocido $\forall i \forall j$, y se utilizan para la calibración.

$\mathbf{x}_k$ está compuesto por $rc - 1$ componentes iguales a cero, y 1 componente igual a uno, indicando la celda a la que pertenece el elemento k. De esta forma, $\mathbf{t}_x = \sum_U \mathbf{x}_k = (N_{11}; \ldots; N_{ij}; \ldots; N_{rc})'$. Para todos los elementos en la ij-ésima celda, $\mathbf{x}_k'\,\boldsymbol{\lambda}$ es constante e igual a λ_{ij}.

De las ecuaciones de calibración obtenemos que: $F(\mathbf{x}_k'\,\boldsymbol{\lambda}) = F(\lambda_{ij}) = \dfrac{N_{ij}}{\hat{N}_{ij}}$. Por lo tanto, los pesos calibrados están dados por $w_k = \dfrac{d_k\, N_{ij}}{\hat{N}_{ij}} \ \ \forall k \in (ij)$, independientemente de F. El estimador calibrado es entonces:

$$\star\ \hat{t}_{y_{cal}} = \hat{t}_{y_{pos}} = \sum_i \sum_j N_{ij}\, \tilde{y}_{s_{ij}} \quad \text{donde} \quad \tilde{y}_{s_{ij}} = \sum_{s_{ij}} \frac{y_k^{\checkmark}}{\hat{N}_{ij}}$$

Post-estratificación incompleta

En este caso el conteo por celda no es conocido o no puede utilizarse. $\mathbf{x}_k$ se define de forma tal de resumir los conteos marginales:

$$\mathbf{x}_k = (\delta_{1\cdot k}; \ldots; \delta_{r\cdot k};\ \delta_{\cdot 1k}; \ldots; \delta_{\cdot ck})'$$

donde:

$$\delta_{i\cdot k} = \begin{cases} 1 & \text{si } k \in \text{ a la fila } i \\ 0 & \text{si } k \notin \text{ a la fila } i \end{cases} \qquad \delta_{\cdot jk} = \begin{cases} 1 & \text{si } k \in \text{ a la columna } j \\ 0 & \text{si } k \notin \text{ a la columna } j \end{cases}$$

Por lo tanto, $\mathbf{x}_k$ tiene dos entradas iguales a uno, y $r + c - 2$ entradas iguales a cero. Entonces, $\sum_U \mathbf{x}_k = (N_{1\cdot}; \ldots,\ N_{r\cdot};\ N_{\cdot 1};\ \ldots;\ N_{\cdot c})'$.

Se define $\boldsymbol{\lambda} = (u_1; \ldots;\ u_r;\ v_1; \ldots;\ v_c)' \Rightarrow \mathbf{x}_k'\,\boldsymbol{\lambda} = u_i + v_j$ cuando $k \in (ij)$. Con $N_{ij} = \sum_{s_{ij}} d_k$ las ecuaciones de calibración son:

$$\begin{cases} \displaystyle\sum_{j=1}^{c} \hat{N}_{ij}\, F(u_i + v_j) = N_{i\cdot} \quad i = 1; \ldots; r \\[2ex] \displaystyle\sum_{i=1}^{r} \hat{N}_{ij}\, F(u_i + v_j) = N_{\cdot j} \quad j = 1; \ldots; c \end{cases}$$

Para resolver el sistema se debe fijar un componente igual a cero, por ejemplo $v_c = 0$. El sistema es invariante a qué componente sea fijada. Una vez resuelto el sistema se obtienen los **conteos calibrados** para las celdas, y los pesos calibrados:

$$\star\ \hat{N}_{ij}^{w} = \hat{N}_{ij} F(u_i + v_j) \qquad \star\ w_k = d_k\, \frac{\hat{N}_{ij}^{w}}{\hat{N}_{ij}}$$

El estimador calibrado queda definido como:

$$\star\ \hat{t}_{y_{cal}} = \sum\nolimits_s w_k\, y_k = \sum\nolimits_i \sum\nolimits_j \hat{N}^w_{ij}\, \tilde{y}_{s_{ij}} = \hat{t}_{y_{marg}}$$

Para el método multiplicativo los factores por celda vienen dados por:

$$F(u_i + v_j) = e^{u_i + v_j} = \alpha_i\, \beta_j > 0 \ \text{ donde } \ \alpha_i = e^{u_i} \ \text{ y } \ \beta_j = e^{v_j}$$

La varianza del estimador $\hat{t}_{y_{marg}}$ es apenas mayor a la de $\hat{t}_{y_{pos}}$ si las dos variables que conforman los márgenes explican la variable Y mediante efectos aditivos, sin interacción.