Estymatory kalibracyjne w NSP 2011

Marcin Szymkowiak

Urząd Statystyczny w Poznaniu
Uniwersytet Ekonomiczny w Poznaniu

18.10.2013

Plan prezentacji

1. Teoretyczne aspekty kalibracji
2. CALMAR
3. Kalibracja w NSP 2011

Kalibracja w ujęciu historycznym

Kalibracja w ujęciu historycznym

- Kalibracja, w swych różnych formach, stała się na przestrzeni ostatnich lat ważną metodą wykorzystywaną w estymacji różnych parametrów w badaniach statystycznych z brakami odpowiedzi.
- Kalibracja — jako nowy termin w metodzie reprezentacyjnej pojawił się w literaturze około 20 lat temu.
- Podstawy teoretyczne kalibracji zostały sformułowane w pionierskiej pracy Särndala i Deville'a (1992) z początku lat 90-tych XX wieku.

Formalne ujęcie kalibracji

Formalne ujęcie kalibracji

- Zakładamy, że populacja $U = \{1, 2, \ldots, N\}$ składa się z N elementów.
- Z populacji tej losujemy zgodnie z określonym schematem losowania próbę $s \subseteq U$, składającą się z n elementów.
- Niech π_i oznacza prawdopodobieństwo inkluzji pierwszego rzędu tj. $\pi_i = P(i \in s)$ a $d_i = 1/\pi_i$ wagę przypisaną i–tej jednostce w procesie losowania.
- Zakładamy, że głównym celem jest oszacowanie wartości globalnej zmiennej y:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

gdzie y_i oznacza wartość zmiennej y dla i–tej jednostki, $i = 1, \ldots, N$.

Formalne ujęcie kalibracji

Formalne ujęcie kalibracji

- Niech ponadto $x_1, \ldots, x_k$ oznaczają zmienne pomocnicze, a $\mathbf{X}_j$ oznacza wartość globalną zmiennej x_j, $j = 1, \ldots, k$, tj.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{2}$$

gdzie x_{ij} oznacza wartość j – tej zmiennej pomocniczej dla i – tej jednostki badania.

- Do oszacowania wartości globalnej zmiennej y wykorzystujemy estymator Horvitza-Thompsona:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} d_i y_i. \tag{3}$$

- W praktyce bardzo często zdarza się, że:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j \tag{4}$$

co oznacza, że pewna korekta wag (kalibracja) jest pożądana.

Formalne ujęcie kalibracji

Formalne ujęcie kalibracji

- Niech $\mathbf{d} = (d_1, \ldots, d_n)^T$ będzie wektorem wag wynikających ze schematu losowania próby, a $\mathbf{w} = (w_1, \ldots, w_n)^T$ poszukiwanym wektorem wag kalibracyjnych, gdzie n oznacza liczebność próby.
- Niech G będzie dowolną funkcją spełniającą następujące warunki:
  - $G(\cdot)$ jest dwukrotnie różniczkowalna,
  - $G(\cdot) \geq 0$,
  - $G(1) = 0$,
  - $G'(1) = 0$,
  - $G''(1) = 1$.
- Nowo wyznaczone wagi powinny nieznacznie się różnić od wag d_i oraz powinny spełniać warunek:

$$\sum_s w_i x_{ij} = \mathbf{X}_j. \tag{5}$$

Problem poszukiwania wag kalibracyjnych

Problem poszukiwania wag kalibracyjnych

(W1) Minimalizacja funkcji odległości:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) \longrightarrow \min, \tag{6}$$

(W2) Równania kalibracyjne:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{7}$$

(W3) Warunki ograniczające:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{gdzie: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, n. \tag{8}$$

Postać funkcji G

Postać funkcji G

- Istnieje pewna dowolność przy wyborze funkcji $G(\cdot)$.
- Najczęściej rozważa się w literaturze następujące jej postacie:

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{9}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{10}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{11}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{12}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt, \tag{13}$$

gdzie α jest dodatnim parametrem, pozwalającym sterować stopniem rozrzutu wag kalibracyjnych w stosunku do wag wynikających ze schematu losowania próby (domyślnie parametr przyjmuje wartość 1), a sinh jest funkcją sinusa hiperbolicznego zdefiniowanego jako $\sinh(x) = \frac{e^x - e^{-x}}{2}$.

Wybór funkcji G

Wybór funkcji G

- W praktycznych zastosowaniach najczęściej wykorzystuje się funkcję G w postaci $G_1(x) = \frac{1}{2}(x-1)^2$. w tym przypadku mamy bowiem:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \quad (14)$$

Estymator kalibracyjny wartości globalnej

Estymator kalibracyjny wartości globalnej

Estymatorem kalibracyjnym wartości globalnej zmiennej Y jest:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i \,, \tag{15}$$

gdzie wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_v D(\mathbf{v}, \mathbf{d}) \,, \tag{16}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{17}$$

przy czym

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i} \,, \tag{18}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{19}$$

Twierdzenie o wagach kalibracyjnych

Twierdzenie o wagach kalibracyjnych

Rozwiązaniem zadania minimalizacji jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, którego składowe spełniają równanie

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{20}$$

przy czym:

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \tag{21}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{22}$$

CALMAR

CALMAR

- na potrzeby wyznaczenia wag kalibracyjnych w NSP 2011 wykorzystano makro CALMAR,
- makro jest napisane w języku 4GL w środowisku SAS i służy do wyznaczania wag kalibracyjnych,
- jest makrem napisanym na potrzeby prac francuskiego urzędu statystycznego (stąd dokumentacja techniczna jest dostępna tylko w języku francuskim),
- oferuje 4 sposoby wyznaczania wag kalibracyjnych w zależności od przyjętej postaci funkcji G.

CALMAR

CALMAR

CALMAR, jak stwierdzono powyżej, oferuje cztery sposoby wyznaczania wag kalibracyjnych, w zależności od postaci funkcji G:

- podejście liniowe

$$G(x) = \frac{1}{2}(x-1)^2, \tag{23}$$

- raking ratio

$$G(x) = x(\log x - 1) + 1, \tag{24}$$

- podejście logitowe

$$G(x) = \left[(x-L)\,log\frac{x-L}{1-L} + (U-x)\,log\frac{U-x}{U-1}\right]\frac{1}{A}, \tag{25}$$

gdzie:

$$A = \frac{U-L}{(1-L)(U-1)}, \tag{26}$$

- podejście liniowe z warunkami ograniczającymi

$$G(x) = \frac{1}{2}(x-1)^2, \quad L \leq \frac{w_i}{d_i} \leq U. \tag{27}$$

Kalibracja w NSP 2011 – ujęcie problemu

Kalibracja w NSP 2011 – ujęcie problemu

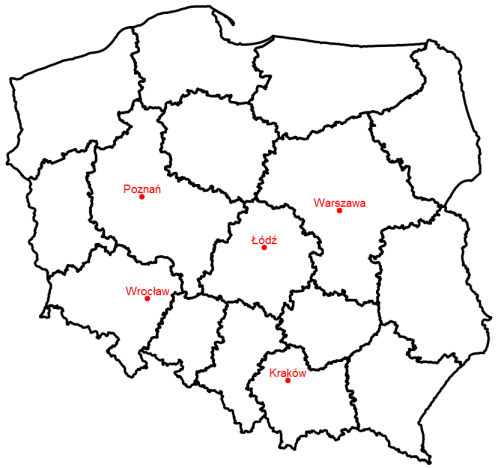
- Narodowy Spis Powszechny Ludności i Mieszkań 2011 – metoda mieszana
- Wykorzystanie danych pochodzących ze źródeł administracyjnych a także danych zbieranych od ludności w ramach przeprowadzonego na szeroką skalę badania reprezentacyjnego
- Uogólnianie wyników badania reprezentacyjnego – kalibracja wag
- Konieczność odpowiedniego zintegrowania i zachowania spójności pomiędzy wynikami badania reprezentacyjnego z danymi pochodzącymi z rejestrów

Wybór zmiennych pomocniczych

- płeć
- wiek
- miejsce zamieszkania - poziom powiatu z wyodrębnieniem części miejskiej i wiejskiej

Wybór zmiennych pomocniczych

- **Województwo: płeć x miejsce zamieszkania x pojedyncze roczniki wieku (0,1,...,83,84,85+)**
- **Powiaty: płeć x miejsce zamieszkania x grupy wieku (0–4,5–9,...,80–84,85+)**
- **Największe miasta: płeć x pojedyncze roczniki wieku (0,1,...,83,84,85+ lub 100+ dla Warszawy)**



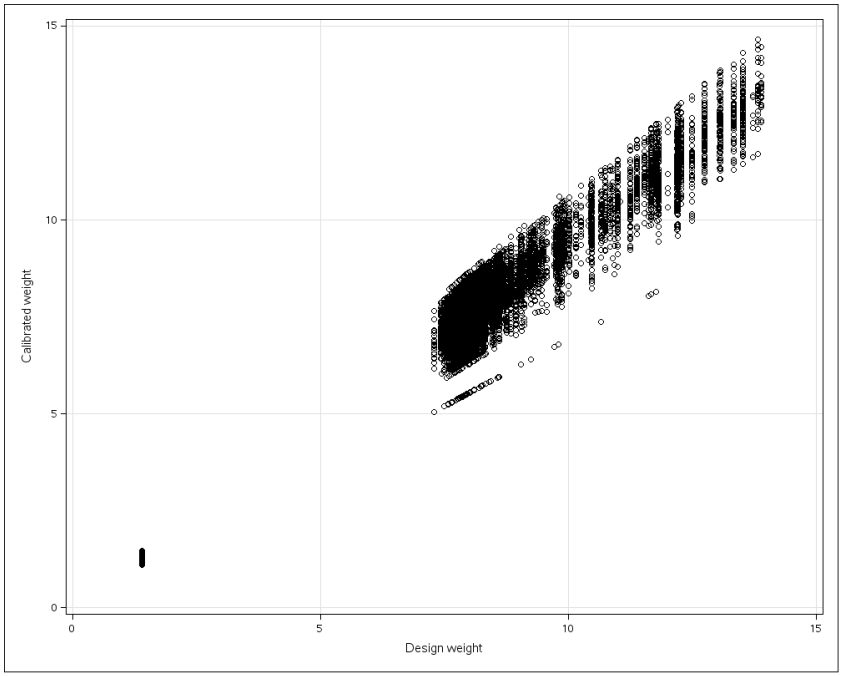
Wybór zmiennych pomocniczych

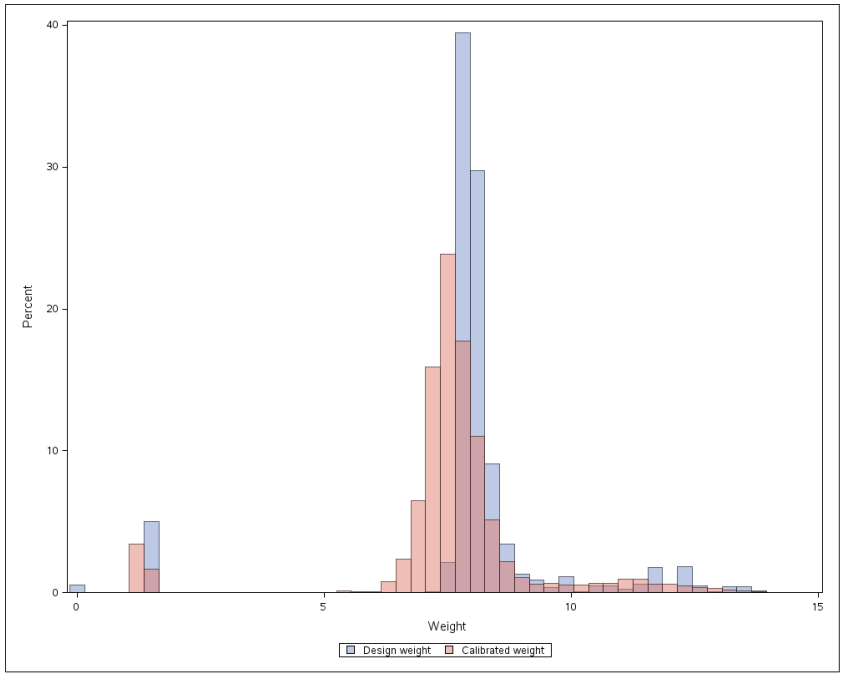
	Część miejska/ Część wiejska	Płeć	Grupy wieku	Pojedyncze roczniki wieku	Pojedyncze roczniki wieku
	1,2	1,2	0-4, 5-9,..., 80-84, 85+	0, 1, . . . ,83, 84; 85+	0, 1, . . . ,98 99, 100+
Polska	1	1	1	1	0
Województwa	1	1	1	1	0
Powiaty (bez 5 największych miast)	1	1	1	0	0
4 największe miasta	x	1	1	1	0
Warszawa	x	1	1	1	1
Dzielnice Warszawy	x	1	1	1	0
Delegatury 4 największych miast	x	1	1	1	0

- **Legenda:** 1–kalibracja możliwa, 0–kalibracja niemożliwa, x–przekrój nieadekwatny

Ocena wag kalibracyjnych

Korelogram i histogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – powiat poznański



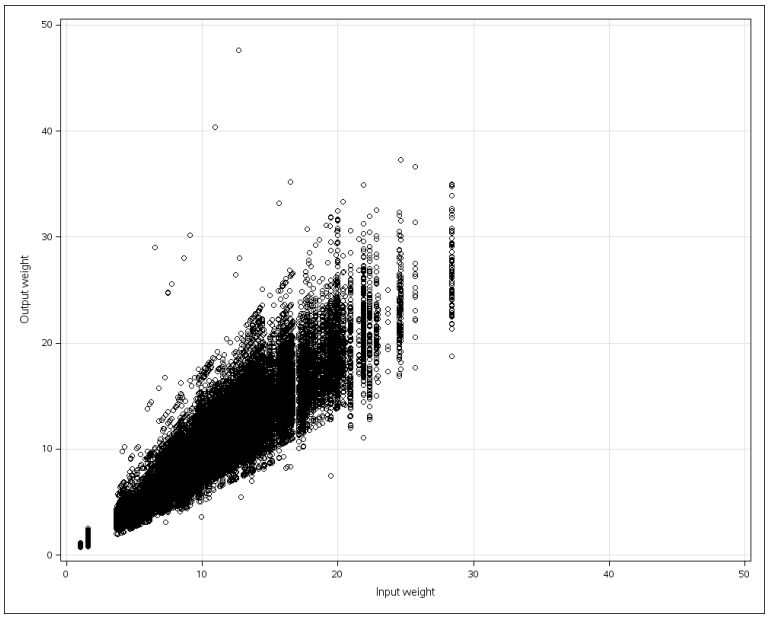


Ocena wag kalibracyjnych

Statystyki opisowe wag w powiecie poznańskim					
Waga	Minimum	Maksimum	Suma	Mediana	Odch. std
d_i	1.3919308	13.8937500	350920.53	7.9896301	1.8675295
w_i	1.0884322	14.4946168	331525.00	7.5480397	1.8096110

Ocena wag kalibracyjnych

Korelogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – Warszawa



Podsumowanie

Podsumowanie

- Kalibracja umożliwiła dopasowanie struktur z badania reprezentacyjnego do znanych wartości globalnych z rejestrów administracyjnych
- Analiza wag kalibracyjnych we wszystkich badanych przekrojach umożliwiła kompleksową ich ocenę
- Wyznaczone wagi kalibracyjne mogą stanowić postawę uogólniania wyników z wykorzystaniem danych pochodzących z badania reprezentacyjnego

Literatura

Literatura


Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.


Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.


Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.

Dziękuję za uwagę!

Podziękowania: Serdeczne podziękowania składam wszystkim moim kolegom z Ośrodka Statystyki Małych Obszarów (Jan, Łukasz, Tomasz, Tomasz i Maciej), bez których proces kalibracji nie osiągnąłby zbieżności!