PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Tomasz Klimanek, Jan Paradysz, Marcin Szymkowiak

Uniwersytet Ekonomiczny w Poznaniu

ESTYMATORY KALIBRACYJNE KWANTYLI ROZKŁADU DOCHODÓW I WYDATKÓW GOSPODARSTW DOMOWYCH W BBGD

1. Wstęp

W ostatnich latach dużą rolę w badaniach próbkowych przypisuje się metodom szacowania parametrów w populacji generalnej z uwzględnieniem wszelkich dostępnych źródeł informacji (spisy, rejestry administracyjne itd.). W ramach tego nurtu badań – określanego mianem statystyki małych obszarów (SMO) – szczególne znaczenie przypisuje się podejściu kalibracyjnemu, które odgrywa istotną rolę w badaniach z brakami odpowiedzi.

Zgodnie z definicją zaproponowaną przez jednego z jej twórców [Särndal 2007] kalibracja jest metodą estymacji parametrów w odniesieniu do skończonych populacji, która składa się z:

- obliczenia wag z uwzględnieniem informacji dodatkowych, tak aby spełnione było odpowiednie równanie – tzw. równanie kalibracyjne,
- wykorzystania tych wag do estymacji wartości globalnej bądź innych parametrów, przy czym wartość zmiennej mnożona jest przez wagę, a sumowanie odbywa się po zbiorze wszystkich respondentów,
- uzyskania w ten sposób nieobciążonych oszacowań parametrów, w sytuacji gdyby w badaniu nie wystąpiły braki odpowiedzi oraz inne błędy nielosowe.

Podejście to zyskało akceptację – przede wszystkim w badaniach prowadzonych przez urzędy statystyczne państw skandynawskich – a także w Kanadzie, Belgii i Holandii. Na potrzeby realizacji badań stworzono odpowiednie programy (GES w Kanadzie, CLAN97 w Szwecji, g-CALIB-S w Belgii oraz Bascula w Holandii), które w procesie estymacji różnych parametrów wykorzystują wiele dostępnych w tych państwach zbiorów zmiennych pomocniczych z rejestrów administracyjnych i spisów.

W polskiej praktyce badań statystycznych kalibracja nie była do tej pory wykorzystywana. Jak jednak pokazują pierwsze prace z tego zakresu [Paradysz, Szym-

kowiak 2007; 2008], jej zastosowanie może prowadzić do znacznej redukcji obciążenia i zmniejszenia wariancji estymatorów, które są często konsekwencją występowania w badaniu braków odpowiedzi.

W podejściu kalibracyjnym ze względu na fakt, że w wielu badaniach, oprócz braków odpowiedzi, rozkłady badanych cech charakteryzują się silnymi asymetriami – dużą wagę zaczęto przywiązywać do problemu estymacji kwantyli, które w takich przypadkach są często miarami położenia bardziej pożądanymi aniżeli średnia.

W związku z tym celem artykułu jest zastosowanie estymatorów kalibracyjnych kwantyli na potrzeby oszacowania wydatków gospodarstw domowych na wybrane towary i usługi konsumpcyjne w przekroju powiatów województwa wielkopolskiego w 2002 r. na podstawie danych pochodzących z Badania Budżetów Gospodarstw Domowych (BBGD).

Potrzeba zastosowania estymatorów kalibracyjnych wynika z dwóch zasadniczych powodów. Po pierwsze, w badaniu tym jednym z istotnych problemów jest duża frakcja braków odpowiedzi oraz skrajnie asymetryczne rozkłady wielu kategorii wydatków i dochodów. Po drugie zaś, mimo iż w wielu publikacjach podawane są informacje o przeciętnych dochodach i wydatkach w gospodarstwach domowych jako podstawowych miarach określających zamożność, poziom życia oraz strukturę konsumpcji, informacje te nie są jednak wystarczające, aby określić stopień zróżnicowania gospodarstw pod względem uzyskiwanych dochodów oraz ponoszonych wydatkach. Stąd też prezentuje się wartości różnych wskaźników mówiących o nierównościach dochodowych i strukturze konsumpcji opartych na decylach (wskaźnik zróżnicowania decylowego, wskaźnik wahania decylowego, wskaźnik dyspersji decylowej itd.).

2. Estymatory kalibracyjne kwantyli

Niech dana będzie *N*-elementowa populacja $U = \{1,2,\ldots,N\}$. Z populacji tej zgodnie z określonym schematem losowania pobieramy *n*-elementową próbę $s \subseteq U$. Niech π_i oznacza prawdopodobieństwo inkluzji *i*-tej jednostki do próby tzn. $\pi_i = P(i \in s)$ dla $i = 1,2,\ldots,N,$ a $d_i = \frac{1}{\pi_i}$ będzie wagą odpowiadającą jednostce *i*. Załóżmy, że celem naszego badania jest oszacowanie kwantyla rzędu $\alpha - Q_{y,\alpha}$ zmiennej y. W szczególnym przypadku, gdy $\alpha = 0,5$, dokonujemy oszacowania mediany zmiennej y.

Załóżmy w dalszym ciągu, że r oznacza *m*-elementowy zbiór respondentów, dla których znana jest wartość zmiennej y. W konsekwencji zbiór $s \setminus r$ oznacza zbiór nierespondentów, tj. jednostek w próbie, które z różnych powodów nie udzieliły odpowiedzi i nie jest znana dla nich wartość zmiennej y.

Załóżmy ponadto, że dysponujemy wektorem zmiennych pomocniczych $Q_{\mathbf{x},\alpha} = \left(Q_{\mathbf{x}_1,\alpha},\ldots,Q_{\mathbf{x}_J,\alpha}\right)^T$, którego poszczególne składowe są kwantylami rzędu α na poziomie populacji. Informacje na temat poszczególnych J kwantyli można pozyskać ze spisów bądź z odpowiednich rejestrów administracyjnych. W przypadku braku takich danych zastępujemy wektor $Q_{\mathbf{x},\alpha}$ wektorem oszacowanych kwantyli $\hat{Q}_{\mathbf{x},\alpha} = \left(\hat{Q}_{\mathbf{x}_1,\alpha},\ldots,\hat{Q}_{\mathbf{x}_J,\alpha}\right)^T$ na podstawie próby s. Zakładamy przy tym, że w przeciwieństwie do zmiennej y w odniesieniu do zmiennych $x_1,\ldots,x_J$ dysponujemy informacjami dla wszystkich jednostek badania.

Zanim zdefiniowany zostanie estymator kalibracyjny kwantyla rzędu α wprowadzimy definicje dystrybuanty zmiennej y, kwantyla rzędu α oraz dystrybuanty interpolacyjnej [Harms, Duchesne 2006].

Definicja 1. Dystrybuantą zmiennej y jest funkcja o postaci:

$$F_y(t) = \frac{\sum_{i=1}^{N} H(t - y_i)}{N}, \tag{1}$$

gdzie:

$$H(t - y_i) = \begin{cases} 1 & \text{dla } t \geq y_i, \\ 0 & \text{dla } t < y_i, \end{cases} \tag{2}$$

dla $i = 1,\ldots,N$ i $t \in \mathbb{R}$.

Definicja 2. $Q_{y,\alpha}$ nazywamy kwantylem rzędu $\alpha \in (0,1)$ zmiennej y, jeżeli:

$$Q_{y,\alpha} = \inf\left\{t \middle| F_y(t) \geq \alpha\right\}, \tag{3}$$

gdzie $t \in \mathbb{R}$.

Definicja 3. Dystrybuantą interpolacyjną zmiennej y nazywamy funkcję $\hat{F}_{y,cal}$ zdefiniowaną w następujący sposób:

$$\hat{F}_{y,cal}(t) = \frac{\sum_{i=1}^{m} w_i H_{y,r}(t, y_i)}{\sum_{i=1}^{m} w_i}, \tag{4}$$

gdzie funkcja $H_{y,r}(t, y_i)$ jest zmodyfikowaną postacią funkcji (2), tj.

$$H_{y,r}(t, y_i) = \begin{cases} 1 & \text{dla } y_i \leq L_{y,r}(t), \\ \beta_{y,r}(t) & \text{dla } y_i = U_{y,r}(t), \\ 0 & \text{dla } y_i > U_{y,r}(t), \end{cases} \tag{5}$$

gdzie:

$$L_{y,r}(t) = \max\left\{\left\{y_i, i \in r \middle| y_i \le t\right\} \cup \left\{-\infty\right\}\right\}, \tag{6}$$

$$U_{y,r}(t) = \min\left\{\left\{y_i, i \in r \middle| y_i > t\right\} \cup \left\{\infty\right\}\right\}, \tag{7}$$

$$\beta_{y,r}(t) = \frac{t - L_{y,r}(t)}{U_{y,r}(t) - L_{y,r}(t)}. \tag{8}$$

Do wyznaczania estymatorów kalibracyjnych kwantyli wykorzystywana jest dystrybuanta interpolacyjna bazująca na zmodyfikowanej funkcji H. Dla funkcji tej w punktach $t \in \{y_i,\ i \in r\}$ spełniony jest warunek $H_{y,r}(t, y_i) = H(t - y_i)$. Występujące w definicji (3) funkcje L i U mają następującą interpretację: L oznacza dolnego, a U − górnego sąsiada punktu t w zbiorze respondentów r dla każdego y_i, $i = 1,\ldots,m$.

W podobny sposób definiujemy dystrybuantę kwantyla rzędu α i dystrybuantę interpolacyjną dla każdej zmiennej pomocniczej x_i, $i = 1,\ldots,J$. W tym celu w definicjach (1)-(3) należy dokonać zamiany oznaczenia zmiennej y na odpowiednią zmienną pomocniczą x_i oraz zbioru r na zbiór s. Wynika to z założenia poczynionego poprzednio odnośnie do tego, że dla zmiennych pomocniczych dysponujemy informacjami na poziomie całej próby.

Definicja 4. Estymatorem kalibracyjnym kwantyla rzędu α jest $\hat{Q}_{y,cal,\alpha}$, który jest rozwiązaniem równania:

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{9}$$

gdzie $\hat{F}_{y,cal}$ jest dystrybuantą interpolacyjną zmiennej y, a wektor wag $\mathbf{w} = (w_1,\ldots,w_m)^T$ jest rozwiązaniem zadania optymalizacyjnego:

$$\mathbf{w} = \arg\min_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{10}$$

przy warunku:

$$\hat{Q}_{\mathbf{x},\alpha,cal} = \left(\hat{Q}_{\mathbf{x}_1,cal,\alpha},\ldots,\hat{Q}_{\mathbf{x}_J,cal,\alpha}\right)^T = Q_{\mathbf{x},\alpha} = \left(Q_{\mathbf{x}_1,\alpha},\ldots,Q_{\mathbf{x}_J,\alpha}\right)^T, \tag{11}$$

oraz

$$\sum_{i=1}^{m} w_i = N, \tag{12}$$

gdzie $D(\mathbf{v},\mathbf{d})$ jest funkcją odległości o postaci:

$$D(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i}. \tag{13}$$

Występujące w powyższej definicji estymatora kalibracyjnego kwantyla rzędu α równania (11) i (12) są tzw. równaniami kalibracyjnymi. Pierwsze z nich mówi

o tym, że wektor wag kalibracyjnych $\mathbf{w}$ powinien zostać wyznaczony w taki sposób, aby kwantyle poszczególnych zmiennych pomocniczych w populacji równały się odpowiednim kwantylom wyznaczonym na podstawie próby. Drugie z kolei równanie orzeka, że suma wag powinna być równa liczebności populacji generalnej. Zwróćmy uwagę, że gdy nie są znane kwantyle rzędu α w odniesieniu do zmiennych pomocniczych – w powyższej definicji należy w miejsce wektora $Q_{\mathbf{x},\alpha} = \left(Q_{\mathbf{x}_1,\alpha},\ldots,Q_{\mathbf{x}_J,\alpha}\right)^T$ wprowadzić wektor $\hat{Q}_{\mathbf{x},\alpha} = \left(\hat{Q}_{\mathbf{x}_1,\alpha},\ldots,\hat{Q}_{\mathbf{x}_J,\alpha}\right)^T$. Oszacowania poszczególnych kwantyli każdej zmiennej pomocniczej można dokonać, wykorzystując w tym celu estymator Horvitza-Thompsona o postaci:

$$\hat{Q}_{x_i,HT,\alpha} = \hat{F}_{x_i}^{-1}\left(\alpha\right), \tag{14}$$

gdzie dla $i = 1,\ldots,J$:

$$\hat{F}_{x_i}\left(t\right) = \frac{\sum_{i=1}^{n} d_i H_{x_i,s}\left(t,x_i\right)}{\sum_{i=1}^{n} d_i}. \tag{15}$$

Twierdzenie 1 rozstrzyga postać wag estymatora kalibracyjnego określonego w definicji (4).

Twierdzenie 1. Rozwiązaniem zadania minimalizacji funkcji odległości (10) przy warunku (11) i (12) jest wektor wag kalibracyjnych o postaci $\mathbf{w} = \left(w_1,\ldots,w_m\right)^T$, którego składowe spełniają warunek:

$$w_i = d_i\left(1 + \mathbf{a}_i^T \lambda_r\right), \tag{16}$$

przy czym

$$\lambda_r = \left(\sum_{i-1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T\right)^{-1}\left(\mathbf{T}_{\mathbf{a}} - \sum_{i=1}^{m} d_i \mathbf{a}_i\right), \tag{17}$$

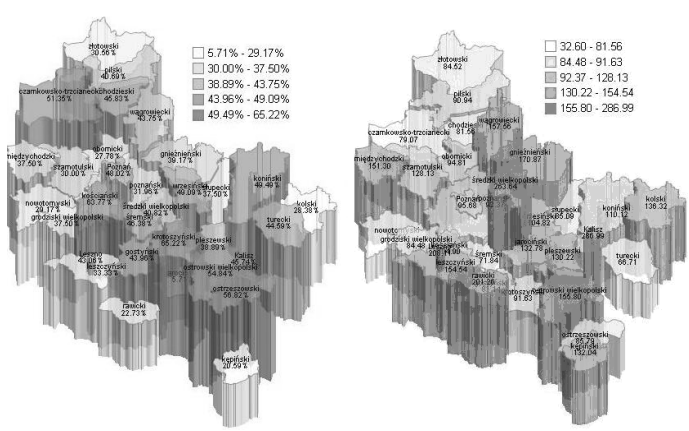
gdzie $\mathbf{T}_{\mathbf{a}} = \left(N,\alpha,\ldots,\alpha\right)^T$, a składowe wektora $\mathbf{a}_i = \left(1,a_{1i},\ldots,a_{Ji}\right)^T$, dla $j = 1,\ldots,J$, wyrażają się następującym wzorem:

$$a_{ji} = \begin{cases} \dfrac{1}{N} & \text{dla } \mathrm{x}_{ji} \le L_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ \dfrac{1}{N}\beta_{x_j,s}\left(Q_{x_j,\alpha}\right) & \text{dla } \mathrm{x}_{ji} = U_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ 0 & \text{dla } \mathrm{x}_{ji} > U_{x_j,s}\left(Q_{x_j,\alpha}\right). \end{cases} \tag{18}$$

Analogicznie do uwagi poczynionej poprzednio, gdy nie znamy kwantyla $Q_{x_j,\alpha}$ zmiennej x_j na poziomie populacji, zastępujemy go we wzorze (18) przez $\hat{Q}_{x_j,\alpha}$ oszacowanym na podstawie próby s.

3. Estymatory kalibracyjne kwantyli w BBGD

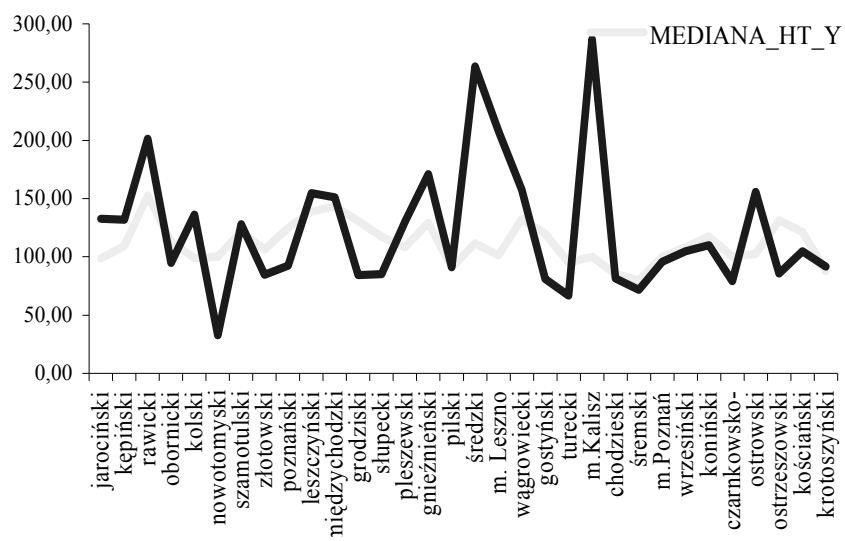
Na rysunkach 1 i 2 zestawione zostały wyniki estymacji wydatków gospodarstw domowych na energię elektryczną. Rysunek 1 zawiera ponadto informacje na temat frakcji braków odpowiedzi w przekroju powiatów województwa wielkopolskiego w 2002 r. Frakcja braków odpowiedzi dla wybranej grupy wydatków – na energię elektryczną – wahała się od 6 do 65%. Z analizy – ze względu na to, że trzy powiaty nie były reprezentowane w próbie przez żadne gospodarstwo domowe – wyłączono powiat kaliski, wolsztyński i miasto Konin. Jako zmienną pomocniczą, dla której dostępne były informacje na poziomie całej próby, wykorzystano dochody gospodarstw domowych ogółem. Ponadto dane o liczbie wszystkich gospodarstw domowych w poszczególnych powiatach województwa wielkopolskiego (niezbędne do utworzenia pierwszej składowej wektora $\mathbf{T_a}$) zostały zaczerpnięte ze spisu z roku 2002. W analizie wydatków gospodarstw domowych na energię elektryczną przyjęto, że $\alpha = 0{,}5$. Oznacza to, że szacowaniu poddana została mediana tej grupy wydatków. Ze względu na nieznajomość mediany dochodów gospodarstw domowych w przekroju powiatów województwa wielkopolskiego (poziom populacji) dokonano jej oszacowania z wykorzystaniem estymatora Horvitza-Thompsona (14) w na podstawie wyników BBGD.



Rys. 1. Frakcja braków odpowiedzi i oszacowania mediany wydatków gospodarstw domowych na energię elektryczną (w zł) z wykorzystaniem estymatora kalibracyjnego w przekroju powiatów województwa wielkopolskiego w 2002 r.

Źródło: opracowanie własne.

Jeśli za podstawę wyciąganych wniosków przyjmie się medianę oszacowaną w oparciu na estymatorze kalibracyjnym, to największą wartość przyjmie ona w powiatach chodzieskim i średzkim. Oznacza to, że 50% gospodarstw domowych w tych powiatach wydaje na energię elektryczną ponad 250 zł. Najmniejszą wartość mediana przyjmuje z kolei w powiecie nowotomyskim, z czego wynika, że w 50% gospodarstw domowych wydatki na energię elektryczną nie przekraczają 35 zł. Jak pokazuje rys. 2, estymator kalibracyjny mediany pogłębia różnice między powiatami, choć w niektórych powiatach daje zbliżone oceny jak estymator Horvitza-Thompsona. Co więcej, estymator ten w swych ocenach znacznie odbiega od ocen estymatora bezpośredniego dla powiatów, w których frakcja braków odpowiedzi jest stosunkowo duża[1]. Dla powiatów, w których frakcja braków odpowiedzi mieściła się w przedziale 30-40%, oceny wydatków na energię elektryczną uzyskane na podstawie estymatora kalibracyjnego i Horvitza-Thompsona były zbliżone (z wyjątkiem powiatu nowotomyskiego).



Rys. 2. Oszacowania mediany wydatków na energię elektryczną (w zł) z wykorzystaniem estymatora kalibracyjnego i Horvitza-Thompsona w przekroju powiatów województwa wielkopolskiego w 2002 r.

Źródło: opracowanie własne.

Wydaje się ponadto, że oceny mediany dokonane z wykorzystaniem estymatora kalibracyjnego (z wyjątkiem powiatu nowotomyskiego) dają bardziej przekonujące oceny wydatków na energię elektryczną. Wynika to zapewne z tego, że przy ocenie

[1] Oceny mediany wydatków na energię elektryczną na rys. 2 z uwzględnieniem obydwu estymatorów zostały przedstawione w przekroju powiatów województwa wielkopolskiego według rosnących wartości wskaźnika braków odpowiedzi.

wyników estymacji należy mieć na uwadze fakt, że istnieją różnice między postawami konsumpcyjnymi respondentów i nierespondentów. Jest to również konsekwencją tego, że estymatory znane ze statystyki małych obszarów (w tym również estymatory bezpośrednie) mają tendencję do niwelowania różnic między obszarami, w przeciwieństwie do estymatorów kalibracyjnych, które te różnice uwypuklają.

4. Podsumowanie

Przedstawiona metodologia wyznaczania kwantyli w badaniach reprezentacyjnych z wykorzystaniem wszelkich dostępnych źródeł danych – charakterystyczna dla statystyki małych obszarów – może stanowić godną rozważenia alternatywę w stosunku do estymatorów znanych z klasycznej metody reprezentacyjnej – zwłaszcza w odniesieniu do badań z brakami odpowiedzi, w których rozkłady cech charakteryzują się dodatkowo silnymi asymetriami. Jak jednak pokazują doświadczenia państw stosujących w praktyce podejście kalibracyjne oraz wyniki badań symulacyjnych, estymatory kalibracyjne w przypadku istnienia odpowiednich źródeł zmiennych pomocniczych (spisy, rejestry administracyjne) charakteryzują się mniejszym obciążeniem i wariancją w porównaniu z estymatorami klasy SMO.

W polskiej praktyce badań statystycznych metodologia ta nie była dotąd wykorzystywana przez GUS. Warto jednak wspomnieć, że w ramach przygotowań do spisu 2011 własności tych estymatorów będą testowane na podstawie danych z NSP 2002 przez pracowników AE w Poznaniu wchodzących w skład powołanej przez GUS podgrupy roboczej ds. metod statystyczno-matematycznych na rzecz spisów. Ponieważ Spis Powszechny 2011 ma mieć nową formułę bazującą na integracji różnego rodzaju rejestrów administracyjnych, metodologia ta zyskałaby na znaczeniu ze względu na bardzo duże możliwości wykorzystania różnego rodzaju zmiennych pomocniczych. W przypadku szacowania parametrów opisujących wydatki gospodarstw domowych na różne artykuły i dobra konsumpcyjne z danych BBGD szczególnie ważnym źródłem – w kontekście wykorzystania zmiennych pomocniczych – mógłby być na przykład rejestr podatników, płatników i innych podmiotów oraz dokumentów „POLTAX" czy baza danych o podatnikach podatku dochodowego od osób fizycznych (PIT).

Pełne wykorzystanie informacji pomocniczych zawartych w tych źródłach mogłoby dodatkowo poprawić proces estymacji oraz uwiarygodnić wyniki badań z BBGD, w których oprócz braków odpowiedzi istotnym problemem jest wiarygodność podawanych przez gospodarstw domowe danych.

Literatura

Deville J.-C., Särndal C.-E. (1992), *Calibration estimators in survey sampling*, "Journal of the American Statistical Association", vol. 87, s. 376-382.

Harms T., Duchesne P. (2006), *On Calibration Estimation for Quantiles*, "Survey Methodology", vol. 32, June 2006, s. 37-52.

Paradysz J., Szymkowiak M. (2007), *Imputacja i kalibracja jako remedium na braki odpowiedzi w badaniu budżetów gospodarstw domowych,* [w:] *Taksonomia 14*, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław.

Paradysz J., Szymkowiak M. (2008), *Taksonometryczne podstawy kalibracji w statystyce małych obszarów,* [w:] Taksonomia 15, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7(1207), UE, Wrocław.

Särndal C.-E. (2007), *The calibration approach in survey theory and practice*, "Survey Methodology", vol. 33, no 2, s. 99-119.

Särndal C.-E., Lundström S. (2005), *Estimation in surveys with nonresponse*, John Wiley & Sons, Ltd.

THE CALIBRATION ESTIMATORS OF QUANTILES OF HOUSEHOLDS' INCOMES AND EXPENDITURES DISTRIBUTION IN HOUSEHOLD BUDGET SURVEY

Summary

The main objective of the paper is to present the calibration estimators used in Household Budget Survey (HBS) which are essential as far as methodology in surveys with nonresponse is concerned. HBS is the survey in which besides the nonresponse one has to deal with the problem of highly skewed distributions. The article presents selected calibration estimators of quantiles according to the idea of calibration proposed by J.-C. Deville, C.-E. Särndal and S. Lundström and developed by T. Harms and P. Duchesne. The theoretical foundations are accompanied by the empirical attempt to apply the discussed estimators in HBS. Moreover, the comparison between calibration estimators of quantiles and classical ones, well-known in literature, is made.