

Jan Paradysz, Marcin Szymkowiak
Akademia Ekonomiczna w Poznaniu

IMPUTACJA I KALIBRACJA JAKO REMEDIUM NA BRAKI ODPOWIEDZI W BADANIU BUDŻETÓW GOSPODARSTW DOMOWYCH

1. Wstęp

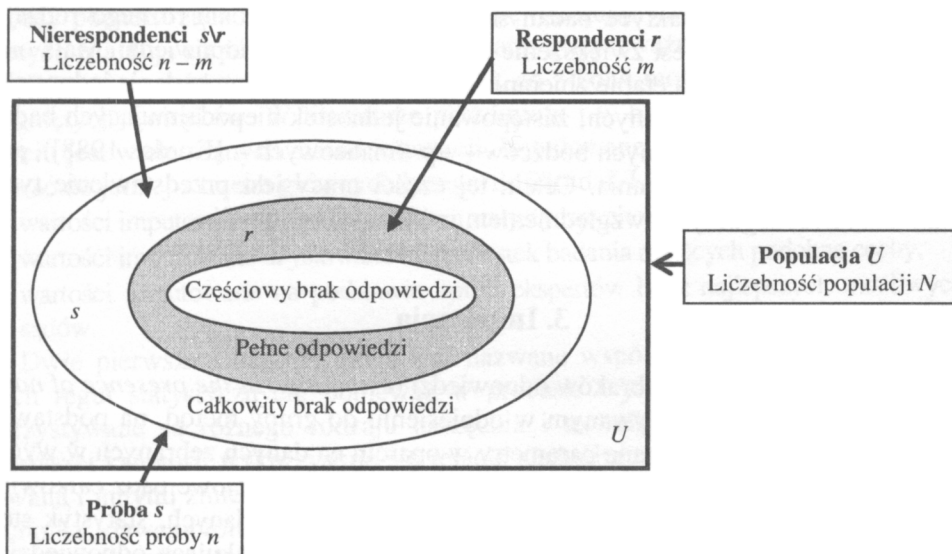
W literaturze przedmiotu przedstawia się dwie podstawowe metody stosowane w przypadku wystąpienia braków odpowiedzi w badaniach reprezentacyjnych: imputację i kalibrację. Pierwsza polega na zastąpieniu brakujących danych konkretnymi wartościami celem uzyskania kompletnego zbioru danych. Druga jest procedurą polegającą na skorygowaniu wyjściowych wag do redukcji obciążenia wynikającego z istnienia braków odpowiedzi. W rezultacie uzyskuje się równowagę rozumianą w ten sposób, że po zastosowaniu kalibracji próba jest „wyglądem” zbliżona do całej populacji. Celem artykułu jest przedstawienie idei imputacji i kalibracji zgodnie z modelami zaproponowanymi przez S. Lundströma i C.-E. Särndala [2005] na potrzeby badania budżetów gospodarstw domowych w Polsce.

2. Braki odpowiedzi jako główne źródło błędów nielosowych w badaniach statystycznych

Braki odpowiedzi występują w badaniach zarówno pełnych, jak i częściowych. Może się wydawać, że w spisach ludności czy w sprawozdawczości statystycznej przypadki nieudzielania odpowiedzi są rzadsze aniżeli w badaniach dobrowolnych, takich jak np., badanie budżetów gospodarstw domowych. Nie jest to jednak takie oczywiste. W badaniach towarzyszących spisowi, jak np. badanie diety kobiet w NSP 1970 i w NSP 1988, braki odpowiedzi sięgały 30%. Z kolei w badaniach

budżetów gospodarstw domowych wskaźnik braku odpowiedzi wahał się w ostatnich latach od 30% do 50%.

Wśród braków odpowiedzi można wyróżnić całkowity oraz częściowy brak odpowiedzi. Z pierwszą sytuacją mamy do czynienia, gdy znane są tylko dane identyfikacyjne, a nie uzyskaliśmy żadnych informacji od badanej jednostki, np. na skutek odmowy bądź nieobecności w czasie przeprowadzenia badania. Z drugą sytuacją spotykamy się, gdy jednostki badania nie udzieliły odpowiedzi na niektóre pytania. Najczęściej niechęć podania informacji podyktowana jest drażliwością pytania lub obawą o ich wykorzystanie przeciwko respondentowi (pytania o plany lub zachowania prokreacyjne albo o wysokość dochodów).



Rys. 1. Zbiór respondentów i nierespondentów w badaniach statystycznych
Źródło: na podstawie [Särndal, Lundström 2005].

Istnieje wiele powodów, dla których w badaniach występują braki odpowiedzi. Do najczęstszych należą niemożność wzięcia udziału w badaniu ze względu na wiek, chorobę, nieobecność w domu, zmiana miejsca zamieszkania. Czynniki te mają charakter obiektywny. Istnieją również powody mające charakter subiektywny, a więc dana jednostka mogłaby wziąć udział w badaniu, ale ze względu na brak czasu czy niechęć do badania odmawia udzielenia odpowiedzi. Bez względu na przyczyny, jakie towarzyszą brakom odpowiedzi, ich występowanie jest źródłem wielu zaburzeń. Jest tak dlatego, że osoby, które odmawiają wzięcia udziału w badaniu bądź nie udzielają na niektóre pytania odpowiedzi, na ogół różnią się od tych, które biorą w nim udział i dostarczają niezbędnych danych. Wskutek tego, po pierwsze, zmniejsza się efektywny rozmiar badanej próby bądź populacji, co ma niekorzystny wpływ na wariancję estymatorów, powodując ich zwiększenie. Po drugie, uzyskane wyniki obciążone są zbyt dużymi błędami. Wyznaczone oceny

parametrów (średnia, wariancja, współczynnik korelacji itd.) znacznie odbiegają od ich prawdziwych wartości, a skonstruowane na podstawie próby przedziały ufności różnych parametrów koncentrują się wokół „złych” wartości. Po trzecie, w przypadku braków odpowiedzi rozkłady wielu cech będą zniekształcone i niemożliwe będzie zastosowanie wielu klasycznych metod statystycznych, gdyż nie będą spełnione założenia, które im towarzyszą. Po czwarte, zbyt niski wskaźnik udzielonych odpowiedzi nie wpływa korzystnie na pozytywne postrzeganie badania przez jego użytkowników i w skrajnych przypadkach może się ono okazać dla nich całkowicie bezużyteczne. W badaniach już na etapie ich projektowania należy zatem uwzględniać występowanie braków odpowiedzi i podejmować stosowne kroki służące ich redukcji. W praktyce badań statystycznych stosuje się różnego rodzaju metody, których celem jest zwiększenie frakcji udzielonych odpowiedzi. Mają one zarówno zastosowanie na etapie zbierania danych (np. powtórne badanie jednostek, od których nie uzyskano danych; zastępowanie jednostek niepodających badania innymi; stosowanie różnych bodźców – np. finansowych – [Kordos 1988]), jak i na etapie ich opracowywania. Celem tej części pracy jest przedstawienie tych ostatnich, ze szczególnym uwzględnieniem imputacji i kalibracji.

3. Imputacja

Estymacja w przypadku braków odpowiedzi (*estimation in the presence of non-response*) jest terminem używanym w odniesieniu do grupy metod, na podstawie których szacowane są nieznane parametry w oparciu na danych zebranych w wyniku przeprowadzonego badania, w którym występują ich częściowe bądź całkowite braki. Gdy zawiodą metody stosowane na etapie zbierania danych, statystyk stoi przed trudnym do rozwiązania dylematem: czy pominąć brakujące odpowiedzi i proces estymacji ograniczyć tylko do obiektów, dla których dysponujemy danymi, czy też uwzględnić je w badaniu i starać się, korzystając z dostępnych narzędzi, wyszacować brakujące dane bądź wprowadzić korektę na uzyskane wyniki? Okazuje się, że tylko w niektórych przypadkach można zastosować pierwsze podejście (np. gdy frakcja braków odpowiedzi jest niewielka). Ze względu jednak na problemy, o których mowa była wcześniej (różnice między respondentami i nierepondentami), bardziej stosowne wydaje się uwzględnianie faktu, że dla niektórych obiektów brak jest całkowicie bądź częściowo danych i dokonywanie prób ich wyszacowania bądź skorygowania uzyskanych wyników na podstawie odpowiednio dobranych wag.

Imputacja to proces szacowania brakujących lub niepoprawnych danych, oparty na wykrytych relacjach w zbiorze wartości tych samych lub innych zmiennych (lub obserwacji), dla których danych nie brakuje. Wynikiem imputacji jest przypisanie jakiejś wartości dla każdej jednostki w miejsce brakujących lub nieważnych danych. Oznacza to, że brakujące dane uzupełniane są ich „substytutami” i są one z

samej definicji „wartościami sztucznymi”. Należy jednak podkreślić, że aby imputacja odegrała swoją rolę w badaniu, muszą być spełnione trzy ważne założenia:

- imputacja nie powinna prowadzić do obciążeń bądź zmian rozkładów cech w zbiorze danych oraz do wzrostu wariancji stosowanych estymatorów,
- proces imputacji w większym stopniu powinien być uzależniony od danych pochodzących z próby aniżeli odwoływać się do założeń co do natury brakujących danych,
- oszacowania ważnych statystyk z próby nie powinny „zbyt mocno” opierać się na imputowanych danych.

W praktyce badań statystycznych bardzo trudno jest przestrzegać tych założeń. Należy ponadto zachować szczególną uwagę, operując na zbiorze danych, wśród których znajdują się również dane imputowane. Nierozważne użycie imputacji może poważnie zniekształcić uzyskane wyniki, co z kolei może być źródłem źle wyciągniętych wniosków. W literaturze przedmiotu i w badaniach statystycznych w użyciu jest wiele różnych metod imputacji. Imputowane wartości można zaklasyfikować do jednej z trzech głównych kategorii [Särndal, Lundström 2005]:

- wartości imputowane z wykorzystaniem statystycznych reguł predykcyjnych,
- wartości imputowane uzyskiwane od jednostek badania mających podobne cechy,
- wartości imputowane na podstawie opinii ekspertów bądź najlepszych możliwych sądów.

Dwie pierwsze kategorie mogą być nazwane wspólnym terminem imputacyjnych reguł statystycznych, ponieważ w procesie wyznaczania substytutów wykorzystywane są różnego rodzaju narzędzia i techniki statystyczne. W ramach pierwszej kategorii wykorzystuje się relacje zachodzące pomiędzy zmienną imputowaną i innymi zmiennymi. Natomiast w drugim przypadku wykorzystywana jest technika „dawca-biorca”, w której obiekt, dla którego imputujemy wartości jakiejś zmiennej, „pożycza” wartości od innych, bardzo podobnych obiektów. Trzecia kategoria obejmuje metody bazujące na wiedzy i doświadczeniu specjalistów z zakresu danego badania.

Dokonując innego rozróżnienia, możemy traktować imputowane wartości jako w pełni losowe (gdy procedura imputacyjna przypisuje różne na ogół wartości zmiennym imputowanym dla różnych obiektów – np. imputacja typu *hot-deck*) oraz mające charakter deterministyczny (kiedy różnym obiektom dla imputowanych zmiennych przypisywane są te same wartości – np. imputacja z wykorzystaniem średniej).

4. Kalibracja

Kalibracja w swych różnych formach stała się w ostatnich latach istotnym narzędziem w badaniach statystycznych. Pojęcie kalibracji, które zostało zasygnalizowane wcześniej, opiera się na odpowiednim doborze wag tak, aby zostały zre-

kompensowane straty informacji związane z występującymi brakami odpowiedzi. Wagi te obliczane są na podstawie informacji, jaką niesie wektor zmiennych pomocniczych.

Istnieje kilka powodów, dla których kalibracja jest wykorzystywana, poza oczywistym argumentem poprawy efektywności szacowanych parametrów poprzez włączenie dodatkowych informacji. Do innych ważnych powodów, dla których w badaniach statystycznych odgrywać zaczęła ona istotną rolę, należy zaliczyć [Gambio 2004]:

- „równowagę” – rozumianą w ten sposób, że po zastosowaniu kalibracji próba jest zbliżona wyglądem do całej populacji,
- „kotwiczenie szacunków” – co należy rozumieć w ten sposób, że zmienna, która jest skorelowana ze zmienną, w stosunku do której zastosowana została kalibracja, nie będzie „dryfowała” i będzie dostarczała stabilnych oszacowań w czasie,
- „wygodę” – zastosowanie kalibracji umożliwia wykorzystanie klasycznych metod statystycznych (np. regresji), które nie dałyby poprawnych wyników, gdyby nie uwzględniono braków odpowiedzi,
- „logiczne szacunki” – po zastosowaniu kalibracji wyniki są bardziej logiczne – np. poszczególne wartości sumują się do wartości globalnych.

5. Estymator kalibracyjny wartości globalnej

Założmy, że celem naszego badania jest oszacowanie wartości globalnej pewnej zmiennej $Y = \sum_U y_k$. Klasyczny estymator Horvitz-Thompsona wartości globalnej

Y wyraża się wzorem $\hat{Y} = \sum_s d_k y_k$, gdzie $d_k = \frac{1}{\pi_k}$, a π_k oznacza prawdopodobieństwo wylosowania (inkluzji) k – tej jednostki do próby. W przypadku występowania w badaniu braków odpowiedzi wagi d_k dla $k \in r$ są z reguły zaniżone,

w związku z czym ważona suma $\sum_r d_k y_k$ jest niedoszacowana w stosunku do wartości globalnej $Y = \sum_U y_k$. Wagi d_k muszą być zatem zwiększone. Poszukujemy zatem nowych wag, większych od d_k dla wszystkich jednostek w próbie, dla których mamy informację o zmiennej Y . Oznaczmy przez w_k poszukiwane wagi.

Oznaczając przez v_k czynnik, przez który przemnożymy wyjściowe wagi d_k , otrzymujemy nowe wagi postaci $w_k = d_k v_k$.

Niech $\hat{Y}_w$ oznacza estymator kalibracyjny (*calibration estimator*) wartości globalnej Y . Skonstruowany estymator kalibracyjny powinien odznaczać się kilkoma zasadniczymi własnościami: małym obciążeniem, małą wariancją, systemem wag,

który „naśladuje” informacje zawarte w wektorze zmiennych pomocniczych, systemem wag, który będzie użyteczny do szacowania wartości globalnej każdej zmiennej Y w badaniu.

Estymator kalibracyjny wartości globalnej dany jest następującym wzorem:

$$\hat{Y}_w = \sum_r w_k y_k. \quad (1)$$

To, czy i w jaki sposób wagi kalibracyjne mogą poprawić oszacowanie wartości globalnej Y oraz skutecznie chronić przed brakami odpowiedzi, zależy od rodzaju i siły informacji, jaką niesie wektor zmiennych pomocniczych. Niech $\mathbf{x}$ oznacza wektor zmiennych pomocniczych, a $\mathbf{x}_k$ niech oznacza wartość wektora zmiennej pomocniczej dla $k \in r$. Załóżmy ponadto, że znany jest wektor wartości globalnych $\sum_U \mathbf{x}_k$. Punktem wyjścia do skonstruowania estymatora kalibracyjnego (1) jest rozwiązanie tzw. równania kalibracyjnego:

$$\sum_r w_k \mathbf{x}_k = \sum_U \mathbf{x}_k. \quad (2)$$

Wagi spełniające to równanie nazywane są wagami kalibracyjnymi i naśladują informacje zawarte w wektorze zmiennych pomocniczych. Ponieważ $w_k = d_k v_k$ to równanie kalibracyjne (2), przyjmuje następującą postać:

$$\sum_r d_k v_k \mathbf{x}_k = \sum_U \mathbf{x}_k. \quad (3)$$

Następnym krokiem jest określenie v_k , przez który przemnożone będą wyjściowe wagi d_k celem uzyskania wag kalibracyjnych w_k . Jedną z postaci spotykanych w literaturze jest $v_k = 1 + \lambda^T \mathbf{x}_k$, gdzie λ jest wektorem, który musi zostać określony. Po wstawieniu v_k do (3) otrzymujemy następujące równanie:

$$\sum_r d_k (1 + \lambda^T \mathbf{x}_k) \mathbf{x}_k = \sum_U \mathbf{x}_k, \quad (4)$$

którego rozwiązaniem jest wektor λ^T o postaci:

$$\lambda_r^T = \left(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k \right)^T \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1}, \quad (5)$$

przy założeniu, że macierz odwrotna $\left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1}$ istnieje.

Ostatecznie estymatorem kalibracyjnym wartości globalnej $Y = \sum_U y_k$ jest:

$$\hat{Y}_w = \sum_r d_k (1 + \lambda_r^T \mathbf{x}_k) y_k, \quad (6)$$

gdzie wektor λ_r^T wyraża się wzorem (5).

6. Wyniki badań empirycznych

W celu zilustrowania prezentowanej metodologii wyznaczania estymatora kalibracyjnego wartości globalnej $Y = \sum_U y_k$ posłużono się danymi indywidualnymi, nieiden-

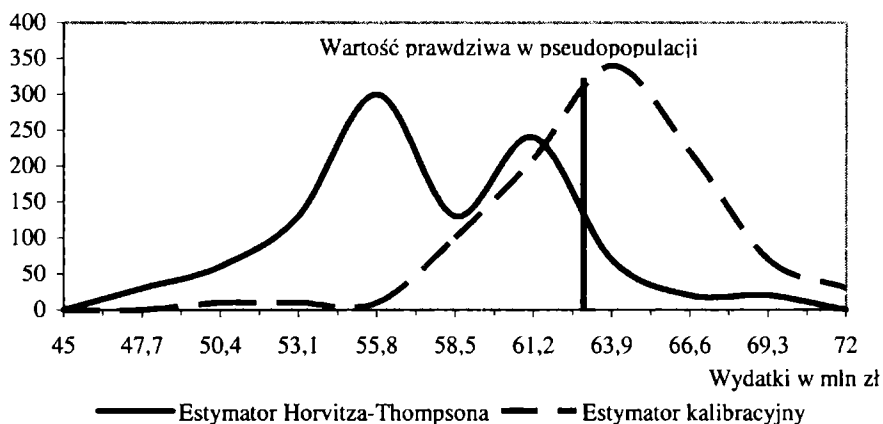
tyfikowalnymi z badania budżetów gospodarstw domowych w Polsce z 2002 r. i przeprowadzono badania symulacyjne. Za zmienną Y przyjęto wydatki gospodarstw domowych ogółem, a za zmienną pomocniczą X – dochody gospodarstw domowych ogółem. Populację określono w ten sposób, że stanowiły ją wszystkie gospodarstwa domowe, które zostały wylosowane na potrzeby badania budżetów gospodarstw domowych w Polsce (łącznie 32 215 gospodarstw domowych). Procedura badawcza przebiegała według następującego schematu:

1. Z tak określonej pseudopopulacji wylosowano zgodnie ze schematem losowania prostego bez zwracania 100 prób 1000-elementowych.

2. W każdej próbie w sposób losowy dla 100 gospodarstw zastąpiono brakiem danych dane o wydatkach gospodarstwa domowego.

3. Na podstawie tak skonstruowanej próby złożonej z 1000 gospodarstw domowych, dla których mieliśmy kompletną informację o ich dochodach, a dla 100 gospodarstw nie mieliśmy danych o wydatkach (tę informację zastąpiono brakiem danych), dokonano estymacji wartości globalnej zmiennej Y w oparciu na klasycznym estymatorze Horvitz-Thompsona oraz estymatorze kalibracyjnym danym wzorem (6).

4. Obliczono obciążenia obydwu estymatorów, porównując wartości, jakie dają w każdej próbie, z „wartością prawdziwą” zmiennej Y obliczoną dla wszystkich gospodarstw domowych wchodzących w skład tak zdefiniowanej populacji.



Rys. 2. Porównanie rozkładów estymatorów Horvitz-Thompsona i kalibracyjnego
Źródło: opracowanie własne.

W wyniku przeprowadzonego eksperymentu okazało się, że szacunki uzyskiwane za pomocą estymatora kalibracyjnego o postaci (6) w przypadku występowania braków odpowiedzi dawały o wiele lepsze rezultaty w porównaniu z zastosowaniem klasycznego estymatora Horvitz-Thompsona, który dawał zazwyczaj заниżone oszacowania wartości globalnej wydatków. W 85% prób estymator kalibracyjny charakteryzował się mniejszym obciążeniem w porównaniu z estymatorem Horvitz-Thompsona (por. rys. 2).

7. Podsumowanie

Jak się wydaje, zaprezentowana metodologia wyznaczania estymatorów kalibracyjnych może być zastosowana w każdym badaniu, w którym występują braki odpowiedzi. Podejście to nie wymaga specjalnych założeń w odniesieniu do stosowanych modeli. Co więcej, może być z powodzeniem wykorzystywane przez agencje statystyczne, w których wiele badań jest obarczonych brakami odpowiedzi, a które mają charakter badań powtarzalnych. Gdy dysponujemy dostatecznie „silnym” wektorem zmiennych pomocniczych, wyniki uzyskiwane na podstawie estymatorów kalibracyjnych dają lepsze oszacowania w porównaniu z estymatorami znanymi z klasycznej metody reprezentacyjnej.

Literatura

- Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.
- Deville J-C., Särndal C-E. (1992), *Calibration Estimators in Survey Sampling*, „Journal of the American Statistical Association” vol. 87, s. 376-382.
- Estevao V.M., Särndal C-E. (2006), *Survey Estimates by Calibration on Complex Auxiliary Information*, „International Statistical Review” vol. 74, s. 127-147.
- Gambio J., (2004), *Discussion of „Issues in Weighting Household and Business Surveys”* – materiały niepublikowane.
- Särndal C-E., Swensson B., Wretman J.H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Särndal C-E., Lundström S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd., New York.

IMPUTATION AND CALIBRATION AS A REMEDY FOR NONRESPONSE IN HOUSEHOLD BUDGET SURVEYS

Summary

The main goal of the paper was to present theoretical and empirical aspects of imputation and calibration as standard methods of dealing with the lack of response in statistical surveys. The authors focused on Polish household budget surveys where the lack of participation of some of which is a significant problem. Because the lack of response occurs on a large scale, it influences the quality of the results. The authors showed how calibration estimators which were proposed by Sixten Lundström and Carl-Erik Särndal, can be used in order to improve estimations in Polish household budget surveys.