EKONOMETRIA ECONOMETRICS 34 • 2011

ISSN 1507-3866

Wojciech Gamrot
Uniwersytet Ekonomiczny w Katowicach

ESTYMACJA ŚREDNIEJ W POPULACJI SKOŃCZONEJ Z KOREKTĄ DWÓCH RODZAJÓW BŁĘDÓW NIELOSOWYCH NA PODSTAWIE DANYCH Z PODPRÓBY

Streszczenie: W niniejszym opracowaniu rozważono schemat losowania dwufazowego prób w sytuacji, gdy badanie statystyczne nie jest wolne od dwóch typów błędów nielosowych: błędu odpowiedzi i błędu braku odpowiedzi. Zaproponowano trzy estymatory wartości przeciętnej badanej cechy w populacji. Dwa z nich są prostymi liniowymi kombinacjami obserwacji badanej cechy, podczas gdy trzeci estymator jest typu regresyjnego. Własności estymatorów wobec występowania wymienionych błędów nielosowych zbadano z wykorzystaniem symulacji komputerowej.

Słowa kluczowe: błąd odpowiedzi, błąd braku odpowiedzi, losowanie dwufazowe, estymator regresyjny.

1. Wstęp

W praktyce badań próbkowych prowadzonych z wykorzystaniem metody reprezentacyjnej często zachodzi potrzeba uwzględnienia przy projektowaniu badania różnych typów błędów nielosowych, wynikających z wystąpienia takich zjawisk, jak nieaktualność operatu losowania, odmowa udziału w badaniu lub też fałszowanie danych. Do błędów tych zalicza się m.in. błędy pokrycia, błędy braku obserwacji cechy i błędy odpowiedzi [Kordos 1988, s. 35]. Opracowania teoretyczne poświęcone tym problemom w większości koncentrują się na jednym tylko typie błędu nielosowego. Dążenie do łącznego ujęcia różnych typów błędów nielosowych jest tendencją stosunkowo nową [Särndal, Lundstrom 2005, s. 179]. Niniejsze opracowanie sytuuje się w tym właśnie nurcie dociekań i poświęcone jest kwestii estymacji wartości przeciętnej w populacji skończonej i ustalonej w warunkach, gdy dla niektórych jednostek populacji wylosowanych do próby nie udaje się uzyskać odpowiedzi, a dla pozostałych uzyskane odpowiedzi mogą być fałszywe. Oznacza to równoczesne występowanie błędów odpowiedzi oraz błędów braku odpowiedzi. W rozdziale pierwszym pracy rozważana jest procedura losowania próby oparta na znanym schemacie losowania dwufazowego i wielokrotnie analizowana w literaturze przed-

miotu jako narzędzie badawcze umożliwiające redukcję obciążenia oszacowań parametrów populacji przy brakach odpowiedzi. Proponowana jest modyfikacja tej procedury umożliwiająca także pozyskanie wiedzy o rozbieżnościach między uzyskiwanymi w badaniu odpowiedziami i rzeczywistymi wartościami badanej cechy. W rozdziale drugim rozważane są trzy estymatory dwufazowe wartości przeciętnej w populacji, wykorzystujące w różnym stopniu i w różny sposób dostępne dane. Pierwszy z nich oparty jest jedynie na informacjach z podprób. Drugi wykorzystuje pełną informację z pierwszej fazy, jednak nie uwzględnia możliwości wystąpienia błędów odpowiedzi. Trzeci estymator to estymator regresyjny, w konstrukcji którego uwzględniono jawnie korektę niezgodności pomiędzy obserwacjami i rzeczywistymi wartościami badanej cechy opartą na modelu liniowym. Rozdział trzeci poświęcono porównaniu estymatorów i wykazaniu, iż estymator regresyjny może przynajmniej w niektórych sytuacjach okazać się dokładniejszy od pozostałych dwóch.

2. Błędy nielosowe i losowanie dwufazowe

Rozważmy skończoną, N-elementową populację $\{u_1, ..., u_N\}$ reprezentowaną przez zbiór indeksów $\{1 ... N\}$ oraz pewną charakterystykę Y jednostek populacji przyjmującą ustalone wartości $y_1, ..., y_N$. Przedmiotem estymacji jest parametr:

$$\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i. \tag{1}$$

Z populacji U losowana jest n-elementowa próba s, według arbitralnie przyjętego planu losowania $p(s)$, charakteryzującego się prawdopodobieństwami inkluzji pierwszego rzędu $\pi_i = P(i \in s)$ dla $i \in U$ oraz drugiego rzędu $\pi_{ij} = P(i,j \in s)$ dla $i \neq j \in U$. W ogólności liczebność próby nie musi być stała. Założymy teraz podobnie jak Cassel i in. [1983], że w badaniu statystycznym może wystąpić brak odpowiedzi o charakterze stochastycznym, a więc, że udzielenie lub nieudzielenie odpowiedzi jest zdarzeniem losowym. Podzbiór próby s zawierający jednostki, które udzieliły odpowiedzi, oznaczmy symbolem s_1, natomiast podzbiór próby s zawierający jednostki, które nie udzieliły odpowiedzi, symbolem s_2. Liczebności tych zbiorów oznaczmy jako n_1 i n_2. Zjawisko braku odpowiedzi może być traktowane jako kolejny etap wyboru próby, opisywany warunkowym rozkładem prawdopodobieństwa $q(s_1 \mid s) = q(s_1, s_2 \mid s)$, zwanym rozkładem odpowiedzi [Särndal i in. 1992, s. 576]. Rozkład ten można scharakteryzować, wyznaczając indywidualne prawdopodobieństwa odpowiedzi pierwszego rzędu: $\rho_{i|s} = \sum_{s_1 \ni i} q(s_1 \mid s)$, oraz rzędu drugiego: $\rho_{ij|s} = \sum_{s_1 \ni i,j} q(s_1 \mid s)$. Oznaczmy uzyskaną w badaniu odpowiedź dotyczącą wartości y_i symbolem x_i. W ogólnym wypadku relacja $x_i = y_i$ nie musi zachodzić. Może się tak zdarzyć m.in. wtedy, gdy odpowiedzi udzielane przez badane jednostki są

niedokładne (np. ze względu na brak pełnej, dokładnej informacji) lub gdy są przedmiotem celowego zawyżania lub zaniżania. Zjawisko to można próbować opisać za pomocą modelu:

$$x_i = f(y_i, \varepsilon, \boldsymbol{\alpha}) , \tag{2}$$

gdzie $\varepsilon \sim N(0,1)$, natomiast wektor parametrów $\boldsymbol{\alpha}$ opisuje mechanizm generowania odpowiedzi. Wartości x_i będą dalej traktowane jako realizacje pewnej, niekoniecznie ustalonej cechy X. W drugiej fazie badania spośród jednostek zbioru s_1 losowana jest n_{u1}-elementowa podpróba $s_{u1,}$ natomiast spośród jednostek zbioru s_2 losowana jest n_{u2}-elementowa podpróba s_{u2}, zgodnie ze schematami losowania $p_1(s_{u1} \mid s, s_1)$ oraz $p_2(s_{u2} \mid s, s_2)$ charakteryzującymi się prawdopodobieństwami inkluzji pierwszego rzędu $\pi_{i|s,s_1}$ dla $i \in s_1$, $\pi_{i|s,s_2}$ dla $i \in s_2$ oraz drugiego rzędu $\pi_{ij|s,s_1}$ dla $i \neq j \in s_1$ oraz $\pi_{ij|s,s_2}$ dla $i \neq j \in s_2$. Rozmiary n_{u1} i n_{u2} obu podprób nie muszą być stałe, i często takie nie będą. Dla każdej jednostki z obu podprób podejmuje się ponowną, bezpośrednią próbę pozyskania danych. Będziemy zakładać, że dla wszystkich jednostek kończy się ona pełnym sukcesem, a więc że wszystkie jednostki w podpróbach uczestniczą w badaniu statystycznym i równocześnie udzielają pełnej, prawdziwej odpowiedzi. Oznacza to, że obserwowane w podpróbie wartości wolne są od obu rozważanych typów błędów nielosowych. W wyniku działania przedstawionej procedury uzyskuje się dokładne obserwacje cechy Y w zbiorach s_{u1} i s_{u2} oraz obarczone błędem odpowiedzi obserwacje cechy X w zbiorze $s_1 - s_{u1}$. Zatem mają na nie wpływ trzy plany losowania wykorzystywane w obu fazach badania, rozkład odpowiedzi oraz rozkład prawdopodobieństwa zmiennej ε. Można zatem w tej sytuacji mówić o pięciu różnych źródłach losowości danych w próbie.

3. Estymatory

W wypadku, gdy w obu fazach badania wykorzystywany jest schemat losowania prostego bez zwracania, natomiast występuje jedynie błąd braku odpowiedzi, często rozważa się wykorzystanie jako estymatora wartości przeciętnej $\overline{Y}$ statystyki w postaci [Hansen, Hurwitz 1946]:

$$\overline{y}_{HH} = \frac{n_1}{n}\overline{y}_{s_1} + \frac{n_2}{n}\overline{y}_{s_{u2}}, \tag{3}$$

gdzie:

$$\overline{y}_{s_1} = \frac{1}{n_1}\sum_{i \in s_1} y_i, \tag{4}$$

$$\overline{y}_{s_{u2}} = \frac{1}{n_{u2}}\sum_{i \in s_{u2}} y_i. \tag{5}$$

W rozważanej tutaj sytuacji wykorzystanie tak skonstruowanego estymatora nie jest możliwe, ponieważ wartości cechy Y są bezpośrednio obserwowane jedynie dla tej części jednostek ze zbioru s_1, które równocześnie należą do s_{u1}. Można jednak rozważać trzy alternatywne estymatory w postaci:

$$\overline{y}_{MIN} = \frac{1}{N}\left(\sum_{i \in s_{u1}} \frac{y_i}{\pi_i \pi_{i|s,s_1}} + \sum_{i \in s_{u2}} \frac{y_i}{\pi_i \pi_{i|s,s_2}} \right), \tag{6}$$

$$\overline{y}_{MAX} = \frac{1}{N}\left(\sum_{i \in (s_1 - s_{u1})} \frac{x_i}{\pi_i} + \sum_{i \in s_{u1}} \frac{y_i}{\pi_i} + \sum_{i \in s_{u2}} \frac{y_i}{\pi_i \pi_{i|s,s_2}} \right), \tag{7}$$

$$\overline{y}_{REG} = \frac{1}{N}\left(\sum_{i \in s_{u1}} \frac{y_i}{\pi_i \pi_{i|s,s_1}} + \left(\sum_{i \in s_1} \frac{x_i}{\pi_i} - \sum_{i \in s_{u1}} \frac{x_i}{\pi_i \pi_{i|s,s_1}} \right) \hat{B}_{xy} + \sum_{i \in s_{u2}} \frac{y_i}{\pi_i \pi_{i|s,s_2}} \right), \tag{8}$$

gdzie

$$\hat{B}_{xy} = \frac{\sum_{i \in s_{u1}} \frac{x_i y_i}{\pi_i \pi_{i|s,s_1}} - \sum_{i \in s_{u1}} \frac{x_i}{\pi_i \pi_{i|s,s_1}} \sum_{i \in s_{u1}} \frac{y_i}{\pi_i \pi_{i|s,s_1}}}{\sum_{i \in s_{u1}} \frac{x_i^2}{\pi_i \pi_{i|s,s_1}} - \left(\sum_{i \in s_{u1}} \frac{x_i}{\pi_i \pi_{i|s,s_1}} \right)^2}. \tag{9}$$

Pierwszy z wymienionych estymatorów wykorzystuje jedynie bezpośrednie obserwacje cechy Y. Jest więc wolny od błędu odpowiedzi oraz nieobciążony, ale dzieje się tak za cenę ograniczenia liczebności próby do $n_{u1} + n_{u2}$, co zapewne uczyni go relatywnie nieprecyzyjnym. Drugi z wymienionych estymatorów wykorzystuje wszystkie dostępne dane, ale nie uwzględnia w żaden sposób możliwości wystąpienia błędu odpowiedzi. Trzeci estymator skonstruowano przy założeniu, że pomiędzy x_i oraz y_i występuje zależność i że ma ona przynajmniej w przybliżeniu charakter liniowy. Pozwala to podjąć próbę korekty ewentualnych błędów odpowiedzi.

W szczególnym wypadku, gdy w obu fazach badania i dla obu podprób stosowany jest schemat losowania prostego bez zwracania, powyższe wzory przyjmują postać:

$$\overline{y}_{MIN} = \frac{n_1}{n}\overline{y}_{s_{u1}} + \frac{n_2}{n}\overline{y}_{s_{u2}}, \tag{10}$$

$$\overline{y}_{MAX} = \frac{n_1 - n_{u1}}{n}\overline{x}_{s_{u1}} + \frac{n_{u1}}{n}\overline{y}_{s_{u1}} + \frac{n_2}{n}\overline{y}_{s_{u2}}, \tag{11}$$

$$\overline{y}_{REG} = \frac{n_1}{n}\left(\overline{y}_{s_{u1}} + (\overline{x}_{s_1} - \overline{x}_{s_{u1}})\hat{B}_{xy} \right) + \frac{n_2}{n}\overline{y}_{s_{u2}}, \tag{12}$$

gdzie

$$\bar{y}_{s_{u1}} = \frac{1}{n_{u1}} \sum_{i \in s_{u1}} y_i, \tag{13}$$

$$\bar{x}_{s_{u1}} = \frac{1}{n_{u1}} \sum_{i \in s_{u1}} x_i, \tag{14}$$

$$\bar{x}_{s_1} = \frac{1}{n_1} \sum_{i \in s_1} x_i \tag{15}$$

oraz

$$\hat{B}_{xy} = \frac{C_{s_{u1}}(y)}{S^2_{s_{u1}}(x)}, \tag{16}$$

$$C_{s_{u1}}(y) = \frac{1}{n_{u1}} \sum_{i \in s_{u1}} (x_i - \bar{x}_{s_{u1}})(y_i - \bar{y}_{s_{u1}}), \tag{17}$$

$$S^2_{s_{u1}}(x) = \frac{1}{n_{u1}} \sum_{i \in s_{u1}} (x_i - \bar{x}_{s_{u1}})^2. \tag{18}$$

Dla porównania własności rozważanych estymatorów przeprowadzone zostało badanie symulacyjne.

4. Wyniki symulacji

Symulacje polegały na wielokrotnym losowaniu prób, symulowaniu błędów nielosowych i każdorazowym dolosowywaniu odpowiednich podprób. Wartości wszystkich trzech estymatorów wyznaczano na podstawie tych samych prób i podprób. Na podstawie zarejestrowanych rozkładów empirycznych oceniano względne obciążenie estymatorów (RB – stosunek obciążenia do szacowanego parametru) oraz ich względny średni błąd szacunku (RRMSE – pierwiastek ze średniego błędu kwadratowego podzielony przez szacowany parametr).

Badaną populację skończoną reprezentował zbiór danych uzyskany w wyniku spisu rolnego przeprowadzonego w roku 1996 w gminach Bolesław, Gręboszów i Radgoszcz powiatu Dąbrowa Tarnowska. Jako badaną ustaloną zmienną Y wykorzystano sprzedaż ogółem gospodarstw rolnych. Przyjęto, że wszystkie losowania wykonywane są według schematu losowania prostego bez zwracania, przy czym liczebności podprób są odpowiednio proporcjonalne do liczebności zbiorów respondentów i nierespondentów, czyli $n_{u1} = c_1 n_1$ oraz $n_{u2} = c_2 n_2$. W eksperymentach przyjęto $c_1 = c_2 = 0{,}1$, co oznacza, że łączna liczebność obu podprób jest stała (z dokładnością do zaokrąglenia). Założono, podobnie jak w pracy Ekholma i Laaksonena [1991], że zdarzenia polegające na udzieleniu lub nieudzieleniu odpowiedzi

przez badane jednostki są niezależne oraz że indywidualne prawdopodobieństwo odpowiedzi zależy od wartości badanej zmiennej Y zgodnie z formułą:

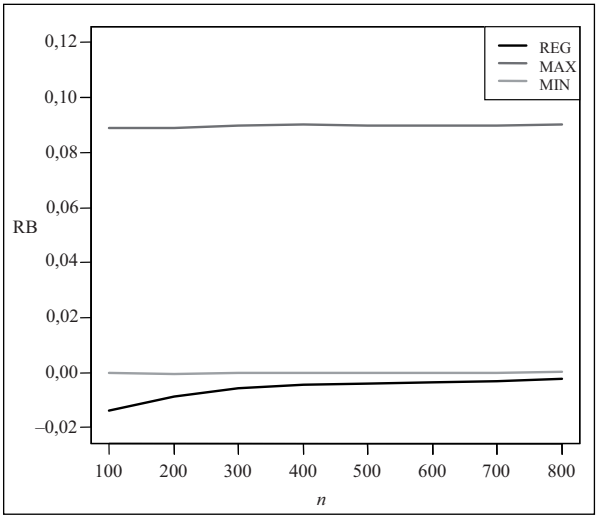
$$\rho_i = \frac{1}{1+\exp(-\beta_0 + \beta_0 y_i)}, \tag{19}$$

przy czym przyjęto $\beta_0 = 1$ oraz $\beta_1 = 10^{-4,}$ wskutek czego prawdopodobieństwa odpowiedzi maleją w miarę wzrostu wartości badanej cechy przy średnim prawdopodobieństwie odpowiedzi wynoszącym 0.64. Ponadto przyjęto, że:

$$x_i = y_i(1 + \alpha_0 + \alpha_1 \varepsilon_i) \tag{20}$$

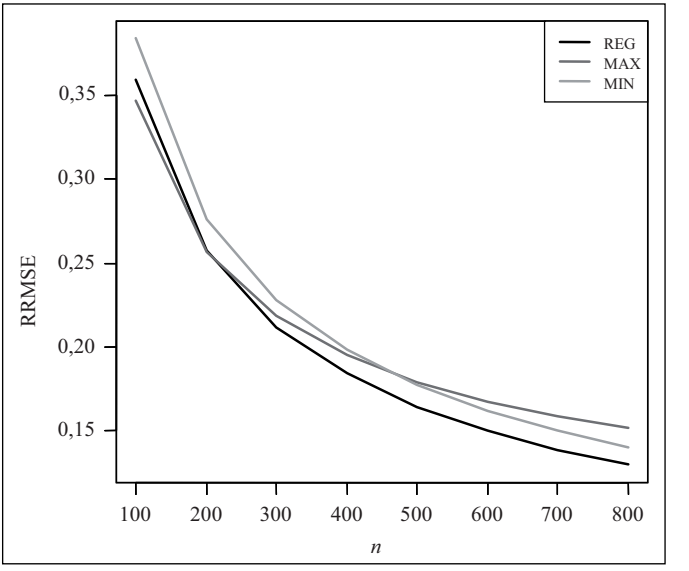
przy arbitralnie dobranych wartościach parametrów: $\alpha_0 = 0{,}2$ oraz $\alpha_1 = 1$. Dodatnia wartość pierwszego z tych parametrów reprezentuje sytuację, w której odpowiedzi udzielane w pierwszej fazie badania obciążone są znacznym błędem systematycznym polegającym na zawyżaniu badanej wartości, a dodatkowo także zmiennym błędem niesystematycznym. Skala obu błędów zależy od wartości badanej cechy, co często ma miejsce w tego rodzaju badaniach.

Symulacje wykonano $3 \cdot 10^5$-krotnie dla rozmiaru próby początkowej n = 100, 200, ... 800 z wykorzystaniem skryptów w środowisku *R*. Zarejestrowane względne obciążenie i względny średni błąd szacunku estymatorów przedstawiono na rys. 1 i 2. Dodatkowo na rys. 3 oraz na rys. 4 zaprezentowano względne wskaźniki efektywności estymatorów obliczane jako stosunki odpowiednich względnych średnich błędów szacunku.



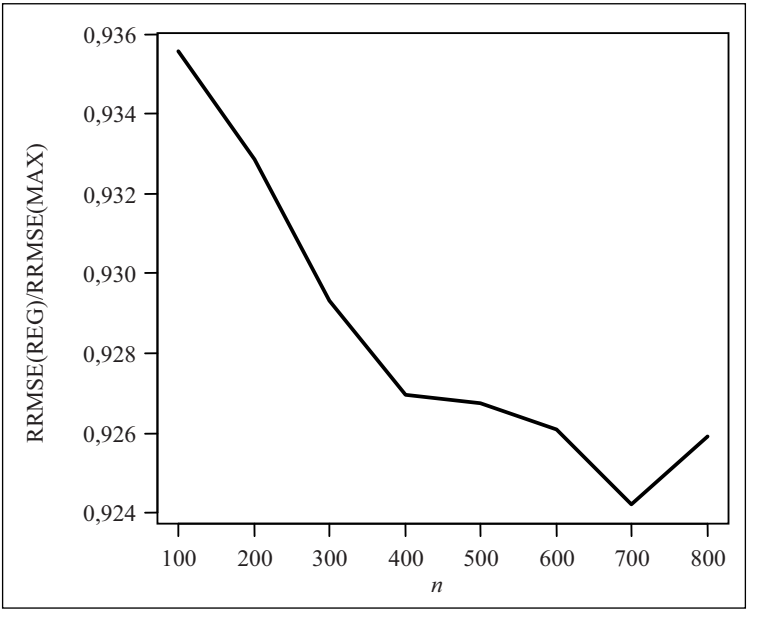
Rys. 1. Względne obciążenie estymatorów

Źródło: opracowanie własne.



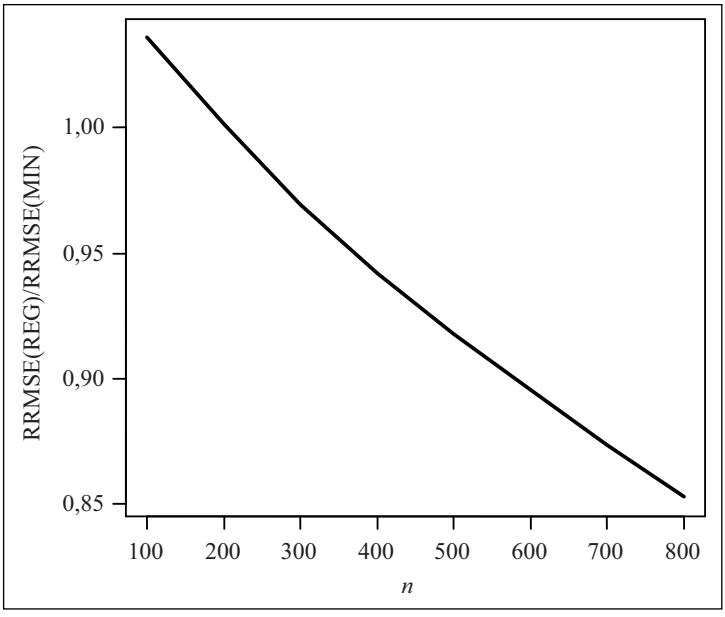
Rys. 2. Względny średni błąd szacunku estymatorów

Źródło: opracowanie własne.



Rys. 3. Względny wskaźnik efektywności estymatorów REG i MAX

Źródło: opracowanie własne.



Rys. 4. Względny wskaźnik efektywności estymatorów REG i MIN

Źródło: opracowanie własne.

Uzyskane wyniki wskazują, że estymator $\overline{y}_{MIN}$ jest nieobciążony. Pozostałe dwa estymatory charakteryzują się wyraźnym obciążeniem. Dla estymatora $\overline{y}_{MAX}$ jest ono dodatnie i pozostaje stałe, gdy zmienia się rozmiar próby początkowej *n*. Taki rezultat sugeruje, że estymator ten nie jest nawet asymptotycznie nieobciążony ani też zgodny (w sensie ciągu populacji skończonych rozważanego m. in. w pracy [Getka-Wilczyńska 2000]). Stawia to pod znakiem zapytania sens stosowania tego estymatora w praktyce. Dla estymatora $\overline{y}_{REG}$ zaobserwowane obciążenie jest natomiast ujemne i szybko maleje do zera wraz ze wzrostem rozmiaru próby *n*. Pozwala to domniemywać, że przy odpowiednio dużym rozmiarze próby obciążenie będzie pomijalnie małe.

Pomimo zaobserwowanego obciążenia, największą dokładnością w sensie względnego średniego błędu kwadratowego (RRMSE) charakteryzował się na ogół estymator regresyjny $\overline{y}_{REG}$, a jedynie dla bardzo małych prób ustępował on nieco estymatorowi $\overline{y}_{MAX}$. W świetle obserwacji poczynionych na temat obciążenia tego ostatniego można jednak spodziewać się, że jego średni błąd kwadratowy w miarę wzrostu rozmiaru próby będzie stabilizować się na poziomie wyraźnie różnym od zera, podczas gdy dla estymatorów $\overline{y}_{REG}$ oraz $\overline{y}_{MIN}$ można mieć nadzieję, iż będzie on malał do zera. Przewaga dokładności estymatora $\overline{y}_{REG}$ nad pozostałymi wyraźnie rośnie w miarę wzrostu rozmiaru próby początkowej, pomijając drobne fluktuacje

względnego wskaźnika efektywności widoczne przy $n = 800$ na rys. 3 i będące zapewne wynikiem niedokładności eksperymentu symulacyjnego.

5. Podsumowanie

Przeprowadzone eksperymenty symulacyjne opierają się na licznych założeniach dotyczących rozkładu badanej charakterystyki populacji i mechanizmów generujących błędy nielosowe. Przede wszystkim należy tu podkreślić kwestię znajomości rodzaju zależności pomiędzy obserwacjami badanej cechy w pierwszej fazie i ich rzeczywistymi wartościami. Konstrukcja estymatora $\bar{y}_{REG}$ opiera się w istocie na założeniu, że zależność ta ma charakter liniowy, a warunkowa wariancja obserwacji cechy nie zależy od ich rzeczywistych wartości (por. m.in. [Särndal i in. 1992, s. 219-242]). Wykorzystywany w eksperymencie symulacyjnym model (20) jest zgodny z pierwszym z tych założeń, choć nie jest zgodny z drugim. Kwestia własności tego estymatora w sytuacji, gdy pierwsze założenie również nie będzie spełnione, pozostaje otwarta. Drugim kluczowym założeniem, na którym oparto symulacje, jest założenie o kompletności i prawdziwości odpowiedzi uzyskiwanych w podpróbach w drugiej fazie badania. Spełnienie tego założenia może wymagać przeprowadzenia badania w drugiej fazie za pomocą innych i droższych metod akwizycji danych niż w fazie pierwszej (np. wizyta ankietera zamiast ankiety telefonicznej lub pocztowej). Jeśli nie będzie ono spełnione, to można oczekiwać, że własności oszacowań ulegną zniekształceniu. Także zmiana każdego z pozostałych warunków eksperymentu może wymusić rewizję ocen własności estymatorów. Dlatego też należy zachować dużą ostrożność, próbując uogólniać wyniki przeprowadzonych eksperymentów na inne sytuacje. Jedynym sposobem obiektywnego sformułowania ogólnych wniosków dotyczących porównania dokładności rozważanych estymatorów pozostają rozważania prowadzone na drodze analitycznej.

Równocześnie należy stwierdzić, że postawiony we wstępie cel pracy został osiągnięty. Uzyskane wyniki wskazują, że istnieją takie warunki, w których estymator $\bar{y}_{REG}$ jest wyraźnie dokładniejszy od dwóch pozostałych. W połączeniu z rozważanym schematem losowania dwufazowego może się on okazać atrakcyjnym narzędziem umożliwiającym równoczesną redukcję obu rozważanych rodzajów błędów nielosowych. Uzasadnia to badanie własności tego estymatora na drodze analitycznej. Próby takie zostaną podjęte w odrębnym artykule.

Literatura

Cassel C.M., Särndal C.E., Wretman J.H., *Some Uses of Statistical Models in Connection with the Nonresponse Problem*, [w:] *Incomplete Data in Sample Surveys*, W.G. Madow I. Olkin (red.), Academic Press, New York 1983.

Ekholm A., Laaksonen S. , *Weighting via response modelling in the Finnish household budget survey*, „Journal of Official Statistics" 1991, vol. 7, no 3.

Getka-Wilczyńska E. (2000), *Estymacja zjawisk rzadkich w populacji skończonej,* Praca doktorska, Szkoła Główna Handlowa, Warszawa.
Hansen M.H., Hurwitz W.N., *The problem of nonresponse in sample surveys*, „Journal of the American Statistical Society" 1946, no 41.
Kordos J., *Jakość danych statystycznych*, PWE, Warszawa 1988.
Särndal C.E., Lundström S., *Estimation in Surveys with Nonresponse*, Wiley, New York 2005.
Särndal C.E., Swensson B., Wretman J.H., *Model Assisted Survey Sampling*, Springer-Verlag, New York 1992.

ESTIMATION OF FINITE POPULATION MEAN WITH A SUBSAMPLE-BASED CORRECTION FOR TWO TYPES OF NON-SAMPLING ERRORS

Summary: In this study the double sampling scheme is considered. It was adopted to the situation where a survey had two types of non-sampling errors, namely response errors and nonresponse errors. Three estimators of a finite population mean were proposed for such a situation. Two of them are linear combinations of sample values while the third is a regression-type one. Properties of estimators were assessed in a simulation study.

Keywords: response error, nonresponse error, two-phase sampling, regression estimator.