PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1189 — 2007
Zastosowania metod ilościowych

Wojciech Gamrot
Akademia Ekonomiczna w Katowicach

O WAŻENIU DANYCH PRZY BRAKACH ODPOWIEDZI Z WYKORZYSTANIEM METOD NIEPARAMETRYCZNYCH

1. Wstęp

Celem niniejszej pracy jest zaproponowanie estymatora wartości przeciętnej cechy w populacji skończonej i ustalonej, wyznaczanego na podstawie prób dwufazowych przy brakach odpowiedzi, z wykorzystaniem oszacowań prawdopodobieństw odpowiedzi uzyskanych na podstawie dostępnej informacji o wartościach cech pomocniczych. Zostaną też przedstawione wyniki eksperymentów symulacyjnych, przeprowadzonych w celu zbadania własności tego estymatora przy braku odpowiedzi o charakterze stochastycznym.

Niech celem estymacji jest wartość przeciętna $\overline{Y}$ pewnej cechy Y w N-elementowej ustalonej populacji U. Z populacji U losowana jest n-elementowa próba s, zgodnie z pewnym planem losowania $p(s)$, charakteryzującym się prawdopodobieństwami inkluzji pierwszego rzędu π_i dla $i \in U$ oraz drugiego rzędu π_{ij} dla $i \neq j \in U$. W ogólnym przypadku liczebność n próby s jest zmienną losową.

Przyjmijmy, podobnie jak Cassel i inni [1983], że w badaniu występuje brak odpowiedzi o charakterze stochastycznym, a więc, że każdej i-tej jednostce populacji można przypisać pewne indywidualne prawdopodobieństwo udzielenia odpowiedzi ρ_i, jeżeli jednostka ta zostanie wylosowana do próby. Udzielenie lub nieudzielenie odpowiedzi przez każdą jednostkę populacji jest zatem zdarzeniem losowym. Próbę s można w rezultacie podzielić na dwa podzbiory s_1 i s_2, o rozmiarach odpowiednio równych n_1 i n_2 takich, że jednostki należące do zbioru s_1 udzielają odpowiedzi, podczas gdy jednostki należące do zbioru s_2 nie udzielają. Tym samym zjawisko braku odpowiedzi może być traktowane jako dodatkowy etap wyboru próby, opisywany pewnym rozkładem prawdopodobieństwa $q(s_2 \mid s)$, który Särndal i inni [1992] nazywają rozkładem odpowiedzi. Na jego podstawie dla każdej i-tej jednostki populacji można określić indywidualne prawdopodobieństwa odpowiedzi:

$$\rho_{i|s} = P(i \in s_1 \mid s) = \sum_{s_1 \ni i} q(s_1 \mid s). \tag{1}$$

Ponadto dla każdej i-tej i j-tej jednostki populacji można określić prawdopodobieństwo równoczesnego udzielenia odpowiedzi:

$$\rho_{ij|s} = P(i, j \in s_1 \mid s) = \sum_{s_1 \ni i,j} q(s_1 \mid s). \tag{2}$$

Zauważmy przy tym, że $\rho_{ii|s} = P(i \in s_1 \mid s) = \rho_{i|s}$. W dalszych rozważaniach będziemy zakładać, że brak odpowiedzi nie zależy od próby s, czyli że:

$$\forall_{s \ni i} \rho_{i|s} = \rho_i. \tag{3}$$

oraz

$$\forall_{s \ni i,j} \rho_{i,j|s} = \rho_{ij}, \tag{4}$$

wskutek czego możemy zapisać: $P(i \in s_1 \mid i \in s) = \rho_i$ oraz $P(i, j \in s_1 \mid i, j \in s) = \rho_{ij}$.

W drugiej fazie badania spośród jednostek zbioru s_2 losowana jest n'-elementowa podpróba s' zgodnie z pewnym schematem losowania $p'(s' \mid s, s_2)$ charakteryzującym się prawdopodobieństwami inkluzji pierwszego rzędu $\pi_{i|s,s_2}$ dla $i \in s_2$, drugiego rzędu $\pi_{ij|s,s_2}$ dla $i \neq j \in s_2$. Liczebność n' podpróby s' jest w ogólnym wypadku zmienną losową. Dla wszystkich jednostek w s' ponawia się następnie wysiłki w celu uzyskania danych. W dalszych rozważaniach przyjmiemy, że w drugiej fazie badania zaobserwowano wartości cechy dla wszystkich jednostek wylosowanych do podpróby. Tym samym indywidualne prawdopodobieństwa odpowiedzi ρ_i odnoszą się jedynie do pierwszej fazy badania. Zatem określono tu trzy źródła losowości próby, jakimi są rozkłady prawdopodobieństwa $p(s)$, $q(s_2 \mid s)$ i $p'(s' \mid s, s_2)$. Jeżeli nie określono inaczej, wszystkie wartości oczekiwane będą wyznaczane względem tych trzech rozkładów prawdopodobieństwa.

2. Estymatory

Jak podają Särndal i in. [1992], przy braku odpowiedzi nieobciążonym estymatorem wartości przeciętnej $\bar{Y}$ jest statystyka:

$$\bar{y}_\pi = \frac{1}{N}\left(\sum_{i \in s_1} \frac{y_i}{\pi_i} + \sum_{i \in s'} \frac{y_i}{\pi_i \pi_{i|s,s_2}}\right); \tag{5}$$

o wariancji:

$$V(\bar{y}_\pi) = \frac{1}{N^2} \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) + \frac{1}{N^2} E_{pq}\left(\sum_{i,j \in s'} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)\right), \tag{6}$$

gdzie symbol $E_{pq}(.)$ oznacza wartość oczekiwaną względem planu losowania w pierwszej fazie i rozkładu odpowiedzi.

W sytuacji gdy indywidualne prawdopodobieństwa odpowiedzi są znane, nieobciążonym estymatorem wartości przeciętnej w populacji jest również statystyka (por. [Bethlehem 1988]):

$$\bar{y}_{\rho} = \frac{1}{N} \sum_{i \in s_1} \frac{y_i}{\pi_i \rho_i}, \tag{7}$$

której wariancja wynosi:

$$V(\bar{y}_{\rho}) = \frac{1}{N^2} \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} \frac{\rho_{ij}}{\rho_i \rho_j} - 1 \right), \tag{8}$$

gdzie ρ_i oraz ρ_{ij} to indywidualne prawdopodobieństwa odpowiedzi określone wzorami (3) i (4). Każda uzyskana obserwacja mnożona jest zatem przez odpowiednią wagę równą odwrotności prawdopobieństwa odpowiedzi, stąd przedstawiona metoda estymacji określana jest zwykle mianem *metody ważenia*. Wykorzystanie estymatora (7) jest możliwe, gdy prawdopodobieństwa odpowiedzi są znane i przyjmują wartości niezerowe. W praktyce jednak pozostają one nieznane, co pociąga za sobą konieczność ich estymacji, a w rezultacie wprowadza obciążenie i uniemożliwia bezpośrednie zastosowanie wzoru (8). Nie ma też żadnej gwarancji, że będą one różne od zera. Dla wyeliminowania tej niedogodności rozważymy teraz możliwość łącznego zastosowania wymienionych wyżej podejść i zaproponujemy estymator wartości przeciętnej wyznaczany na podstawie próby dwufazowej, wykorzystujący ważenie danych.

Niech prawdopodobieństwa inkluzji w drugiej fazie badania nie zależą od s i s_2, a więc $\pi_{i|s,s_2} = \pi'_i$ dla $i \in U$ oraz $\pi_{ij|s,s_2} = \pi'_{ij}$ dla $i, j \in U$. Przede wszystkim warunek ten będzie spełniony, gdy w drugiej fazie badania podpróba s' o liczebności $n' = cn_2$ będzie losowana zgodnie ze schematem losowania prostego bez zwracania. Prawdopodobieństwo zaobserwowania wartości badanej cechy dla i-tej jednostki populacji pod warunkiem, że została ona wylosowana do próby s, można zatem zapisać w postaci:

$$\rho'_i = \rho_i + \pi'_i(1 - \rho_i). \tag{9}$$

gdzie ρ_i jest prawdopodobieństwem udzielenia odpowiedzi w pierwszej fazie badania, natomiast wyrażenie $\pi'_i(1-\rho_i)$ reprezentuje warunkowe prawdopodobieństwo zdarzenia polegającego tym, że i-ta jednostka w wypadku wylosowania do próby nie udzieli odpowiedzi i zostanie wylosowana do podpróby s'. Podobnie prawdopodobieństwo zaobserwowania wartości badanej cechy dla i-tej i j-tej jednostki populacji jednocześnie, pod warunkiem że jednostki te zostały wylosowane do próby s, wynosi:

$$\rho'_{ij} = \rho_{ij} + (\rho_i - \rho_{ij})\pi'_j + (\rho_j - \rho_{ij})\pi'_i + (1 - \rho_i - \rho_j + \rho_{ij})\pi'_{ij}. \qquad (10)$$

Wstawiając ρ'_i w miejsce ρ_i we wzorze (7) i biorąc pod uwagę jednostki zbioru zarówno s_1, jak również s', otrzymujemy statystykę:

$$\bar{y}_{\pi\rho} = \frac{1}{N} \sum_{i \in s_1 \cup s'} \frac{y_i}{\pi_i(\rho_i + \pi'_i(1 - \rho_i))}. \qquad (11)$$

która dla znanych ρ_i jest nieobciążonym estymatorem $\bar{Y}$. Jej wariancja wynosi:

$$V(\bar{y}_{\pi\rho}) = \frac{1}{N^2} \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} \frac{\rho'_{ij}}{\rho'_i \rho'_j} - 1 \right). \qquad (12)$$

Estymator (11), podobnie jak statystyka (7), został skonstruowany przy założeniu, że prawdopodobieństwa odpowiedzi ρ_i są znane. Gdy tak nie jest, można zastąpić je oszacowaniami $\hat{\rho}_i$, wyznaczanymi na podstawie dostępnej informacji o wartościach cech pomocniczych. W ten sposób uzyskujemy następujący estymator wartości przeciętnej w populacji:

$$\bar{y}_{\pi\rho^*} = \frac{1}{N} \sum_{i \in s_1 \cup s'} \frac{y_i}{\pi_i(\hat{\rho}_i + \pi'_i(1 - \hat{\rho}_i))}. \qquad (13)$$

Zauważmy, że oszacowania $\hat{\rho}_i$ są zmiennymi losowymi, wskutek czego estymator ten jest w ogólnym wypadku obciążony, a wzór (12) nie może być stosowany do wyznaczania jego wariancji. Warto jednak zwrócić uwagę, że mianowniki wyrażeń (11) i (13) są różne od zera, jeżeli tylko $\pi'_i \neq 0$ dla $i \in s_2$, niezależnie od wartości prawdopodobieństw odpowiedzi ρ_i i ich oszacowań $\hat{\rho}_i$, które w skrajnym wypadku mogą być nawet równe zeru. Jest to istotna zaleta zaproponowanych estymatorów w porównaniu ze statystyką (7).

3. Szacowanie prawdopodobieństw odpowiedzi

Indywidualne prawdopodobieństwa odpowiedzi można szacować na wiele sposobów. Do popularnych rozwiązań należy m.in. procedura, którą zaproponowali H.L. Oh i F. Scheuren [1983], opierająca się na założeniu, że prawdopodobieństwa odpowiedzi są stałe w pewnych podzbiorach populacji, a także modele parametryczne wykorzystujące regresję logistyczną (por. [Ekholm, Laaksonen 1991]). Słabością tych podejść jest konieczność specyfikacji modelu opisującego funkcjonalną zależność pomiędzy zmiennymi pomocniczymi i nieznanymi parametrami oraz prawdopodobieństwem ρ_i. Jakkolwiek w pewnych sytuacjach dostępne są wiarygodne informacje mogące stanowić podstawę ustalenia postaci takiego modelu, najczęściej istnieje znaczne ryzyko jego niewłaściwego zdefiniowania i nietrafnej oceny prawdopo-

dobieństw ρ_i. Przykłady błędów wynikających z błędnej specyfikacji modelu podaje między innymi Härdle [1994].

Alternatywnym sposobem estymacji prawdopodobieństw ρ_i jest wykorzystanie metod regresji nieparametrycznej, a szczególnie tzw. estymatorów kernelowych (jądrowych – por. [Rosenblatt 1956]). Prawdopodobieństwa ρ_i są w tym wypadku również traktowane jako pewna nieznana funkcja zmiennych pomocniczych, lecz ich oceny są wyznaczane zgodnie z procedurą, która nie wymaga czynienia jawnych założeń o analitycznej postaci tej funkcji. Przyjmijmy, że dla wszystkich jednostek próby s obserwowana jest zmienna pomocnicza X przyjmująca wartości $x_1, ..., x_N$ i rozważmy statystykę:

$$\hat{\rho}_i = \frac{\sum_{j \in s_1} D(x_i - x_j)}{\sum_{j \in s} D(x_i - x_j)}, \tag{14}$$

gdzie wyrażenia $D(z)$ określane mianem jądra (*kernel*) są ciągłymi, ograniczonymi i parzystymi funkcjami zmiennej z, spełniającymi warunek:

$$\int_{-\infty}^{+\infty} D(z)dz = 1. \tag{15}$$

Statystyka (14) może być traktowana jako ocena prawdopodobieństwa ρ_i. W szczególnym przypadku, gdy:

$$D(z) = D(x_i - x_j) = \begin{cases} \frac{1}{2h} & \text{gdy } |x_i - x_j| \le h \\ 0 & \text{gdy } |x_i - x_j| > h \end{cases}, \tag{16}$$

gdzie h jest pewną stałą, to estymator (14) jest równy udziałowi respondentów wśród tych jednostek próby s, dla których wartości x_j cechy pomocniczej różnią się od x_i o nie więcej niż h. Dla odpowiednio dużych wartości h do otoczenia tego należą wszystkie wartości x_i odpowiadające jednostkom wylosowanych do próby i w rezultacie oceną prawdopodobieństwa ρ_i jest udział respondentów w próbie s, czyli stosunek n_1/n. Na wybór stałej h wpływ mają dwa kryteria. Z jednej strony musi to być liczba na tyle duża, aby w otoczeniu badanego punktu znalazła się liczba jednostek próby s, wystarczająca do obliczenia estymatora, przy czym zwykle postuluje się, aby była to liczba możliwie jak największa, co ma umożliwić uzyskanie ocen o możliwie niewielkiej zmienności. Z drugiej zaś strony nadmierna szerokość otoczenia może uniemożliwić wykrycie lokalnych anomalii bądź zaburzeń w rozkładzie prawdopodobieństwa ρ_i.

Ponieważ dla celów estymacji wartości przeciętnej cechy Y w populacji niezbędne jest wyznaczenie $\hat{\rho}_i$ jedynie dla jednostek zbioru s_1, w każdym ze wspomnianych otoczeń znajdzie się co najmniej jedna wartość x_i odpowiadająca jednostce

wylosowanej do próby. Gdyby jednak było konieczne wyznaczenie $\hat{\rho}_i$ dla wszystkich jednostek populacji, mogłoby dojść do sytuacji, w której dla pewnej wartości h żadna z wartości x_i nie należałaby do takiego otoczenia, co uniemożliwiłoby wyznaczenie wartości estymatora. Trudności tej można uniknąć, definiując funkcję $D(z)$ w taki sposób, aby przyjmowała niezerowe wartości dla dowolnego z. Inną przesłanką uzasadniającą konstrukcję funkcji jądra w taki sposób, jest to, iż w wypadku zastosowania formuły (16) wyniki estymacji zależą od doboru stałej h i jest to zależność o charakterze skokowym, a ponadto wszystkim jednostkom próby s, które należą do otoczenia, przypisywana jest jednakowa waga niezależnie od odległości od badanego punktu, co nie zawsze znajduje uzasadnienie.

Funkcję jądra, przyjmującą niezerowe wartości dla dowolnej wartości z, rozważa m.in. A. Giommi [1987]. Funkcja ta przyjmuje postać:

$$D(z) = \frac{1}{h\sqrt{2\pi}} e^{\frac{-z^2}{2h^2}}. \tag{17}$$

Stała h spełnia tu podobną funkcję jak w przypadku formuły (16), określając z jednej strony czułość metody na lokalne „anomalie", z drugiej zaś zmienność uzyskiwanych ocen $\hat{\rho}_i$. Podobnie jak poprzednio, w miarę wzrostu h wartość wyrażenia (14) dąży do n_1/n. Zauważmy, że ten sposób szacowania prawdopodobieństw odpowiedzi, mimo że pozwala uniknąć konieczności jawnego specyfikowania modelu, nie jest wolny od pewnej arbitralności.

Do oceny prawdopodobieństw odpowiedzi ρ_i za pomocą opisanych metod można także wykorzystać więcej niż jedną zmienną pomocniczą. W takim wypadku zamiast zmiennej z we wzorach (14)-(17) można wykorzystać dowolną miarę odległości dwóch punktów reprezentowanych przez wektory $\boldsymbol{x}_i$ i $\boldsymbol{x}_j$ wartości zmiennych pomocniczych. Przede wszystkim może to być odległość euklidesowa, wyznaczana zgodnie z wzorem:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'}. \tag{18}$$

Estymator (14) prawdopodobieństwa odpowiedzi ρ_i w punkcie $\mathbf{x}_i$, można zapisać w postaci:

$$\hat{\rho}_i = \frac{\sum_{j \in s_1} D(d(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{j \in s} D(d(\mathbf{x}_i, \mathbf{x}_j))}. \tag{19}$$

Dodajmy, że wykorzystanie w konstrukcji estymatora euklidesowej miary odległości pociąga za sobą konieczność odpowiedniej standaryzacji obserwacji cech pomocniczych.

Zastępując we wzorze (13) nieznane prawdopodobieństwa odpowiedzi ρ_i oszacowaniami $\hat{\rho}_i$, uzyskujemy ostatecznie estymator wartości przeciętnej.

4. Wyniki symulacji

Własności zaproponowanego estymatora są trudne do określenia metodami analitycznymi, dlatego też dla ich zbadania przeprowadzono eksperyment symulacyjny. Celem eksperymentu było wyznaczenie i porównanie średniego błędu kwadratowego oraz obciążenia estymatorów $\bar{y}_{\pi p^*}$ i $\bar{y}_{\pi}$. Przyjęto, że próba losowana jest zgodnie ze schematem losowania prostego bez zwracania charakteryzującym się prawdopodobieństwami inkluzji pierwszego i drugiego rzędu, odpowiednio równymi $\pi_i = n/N$ dla $i \in U$ oraz $\pi_{ij} = n(n-1)/N(N-1)$ dla $i \neq j \in U$. W symulacji wykorzystano dane uzyskane podczas spisu rolnego przeprowadzonego w roku 1996 w wybranych gminach powiatu Dąbrowa Tarnowska. Dane te, obejmujące 2422 jednostki populacji (gospodarstwa rolne), reprezentowały badaną populację. Przyjęto, że zmienną badaną Y jest sprzedaż ogółem w danym gospodarstwie w roku 1995. Zmienną pomocniczą X była powierzchnia gospodarstwa w hektarach przeliczeniowych. Dla każdej jednostki populacji wygenerowano prawdopodobieństwa odpowiedzi na podstawie modelu logistycznego, określonego wzorem:

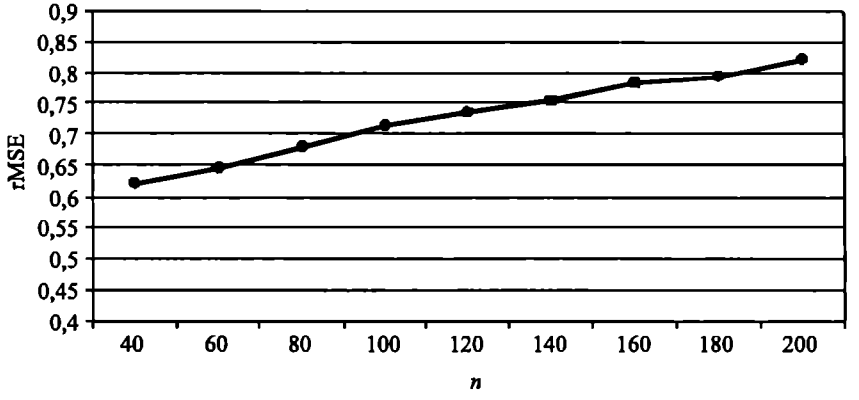
$$\rho_i = \frac{1}{1 + e^{\beta \mathbf{x}_i}}, \tag{20}$$

gdzie $\mathbf{x}_i = [1, x_i]'$ to wektor wartości zmiennych pomocniczych odpowiadający i-tej jednostce populacji (pierwszy element reprezentuje wyraz wolny), natomiast $\beta = [-4, 0.003]$ stanowi wektor parametrów arbitralnie dobranych w taki sposób, aby prawdopodobieństwo udzielenia odpowiedzi malało w miarę wzrostu wartości badanej cechy. Wartość przeciętna indywidualnych prawdopodobieństw odpowiedzi wyniosła 0,89. Założono przy tym, że odpowiedzi poszczególnych jednostek są niezależne, czyli $\rho_{ij} = \rho_i \rho_j$ dla $i \neq j \in U$.

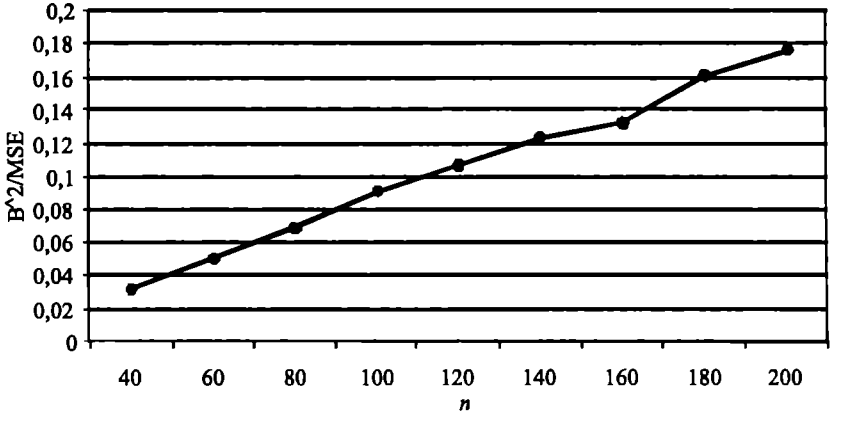
Eksperyment symulacyjny polegał na wielokrotnym losowaniu próby s z populacji zgodnie ze schematem losowania prostego bez zwracania. Następnie dla każdej jednostki populacji wylosowanej do podpróby przeprowadzano niezależne doświadczenie losowe, w którym prawdopodobieństwo sukcesu było równe odpowiadającemu tej jednostce indywidualnemu prawdopodobieństwu odpowiedzi. W wypadku sukcesu przyjmowano, że badana jednostka udzieliła odpowiedzi, natomiast w przeciwnym wypadku przyjmowano, że odpowiedzi nie udzieliła, wskutek czego wyodrębniano podzbiory s_1 i s_2. Następnie ze zbioru nierespondentów losowano zgodnie ze schematem losowania prostego bez zwracania podpróbę s', której liczebność stanowiła 30% liczebności n_2 zbioru s_2. Na podstawie tak uzyskanych prób i podprób każdorazowo wyznaczano wartości estymatorów. Na podstawie rozkładu empirycznego wartości każdego z tych estymatorów wyznaczano jego przybliżone obciążenie i średni błąd kwadratowy.

Symulacje przeprowadzono dla rozmiaru próby początkowej n = 40, 80, ..., 200. Dla każdej z wartości n wylosowano 10000 prób z populacji. Względny wskaźnik

efektywności estymatora $\bar{y}_{\pi p^*}$ (wyznaczany jako stosunek średniego błędu kwadratowego tego estymatora do średniego błędu kwadratowego estymatora $\bar{y}_{\pi}$) w funkcji rozmiaru próby początkowej przedstawiono na rys. 1. Każdemu punktowi na wykresie odpowiada 20000 prób wylosowanych z populacji.



Rys. 1. Względny wskaźnik efektywności estymatora $\bar{y}_{\pi p^*}$ jako funkcja rozmiaru n próby początkowej



Rys. 2. Udział obciążenia estymatora $\bar{y}_{\pi p^*}$ w jego średnim błędzie kwadratowym jako funkcja rozmiaru n próby początkowej

Dla wszystkich badanych wartości n względny wskaźnik efektywności przyjmuje wartości mniejsze od jedności, co oznacza, że estymator $\bar{y}_{\pi p^*}$ jest dokładniejszy od $\bar{y}_{\pi}$. Przewaga ta w skrajnym wypadku przekracza nawet 35%, jednak maleje w miarę wzrostu rozmiaru próby początkowej n. Zależność udziału obciążenia estymatora $\bar{y}_{\pi p^*}$ od n przedstawiono na rys. 2. Udział ten wzrasta wraz z n (głównie wskutek spadku MSE) i dla $n = 200$ przekracza 17%.

5. Podsumowanie

W niniejszym artykule zaproponowano estymator wartości przeciętnej dla schematu losowania dwufazowego przy brakach odpowiedzi. W konstrukcji tego estymatora uwzględniono oszacowania prawdopodobieństw odpowiedzi uzyskane z wykorzystaniem estymatora jądrowego. Przedstawione wyniki symulacji wskazują, iż zaproponowany estymator jest dokładniejszy (w sensie średniego błędu kwadratowego) od klasycznego estymatora dla próby dwufazowej rozważanego przez Särndala i innych [1992]. Dalszą poprawę jego dokładności można też zapewne uzyskać, dobierając inny kształt lub rozpiętość funkcji jądra – zagadnienie optymalnego ich doboru nie było tu bowiem rozważane. Trzeba jednak dodać, że udział obciążenia zaproponowanego estymatora w jego średnim błędzie kwadratowym rośnie wraz z rozmiarem próby, co czyni jego zastosowanie celowym głównie w przypadku prób o niewielkiej liczebności.

Literatura

Bethlehem J.G. , *Reduction of Nonresponse Bias Through Regression Estimation* „Journal of Official Statistics" 1988, vol. 4, no. 3, 251-160.

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*, red. W.G. Madow, I. Olkin, 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, 325-338.

Giommi A., *Nonparametric Methods for Estimating Individual Response Probabilities*, „Survey Methodology" 1987, vol 13, no. 2, 127-134.

Härdle W., *Applied Nonparametric Regression*, Cambridge University Press, 1992.

Oh H.L., Sheuren F., *Weighting Adjustment for Unit Nonresponse*, [w:] *Incomplete Data in Sample Surveys*, red. W.G. Madow, I. Olkin, Academic Press, New York 1983.

Rosenblatt M., *Remarks on Some Nonparametric Estimates for the Density Function*, *Annals of Mathematical Statistics*, 1956, no. 27, 832-837.

Särndal C.E., Swensson B., Wretman J.H., *Model Assisted Survey Sampling* Springer-Verlag, New York 1992.

ON WEIGHTING ADJUSTMENTS FOR NONRESPONSE USING NONPARAMETRIC METHODS

Summary

The occurrence of nonresponse in a sample survey usually introduces the bias in population parameter estimates and reduces their precision. Several estimation methods have been developed to compensate for these effects. One of the procedures broadly discussed in the literature is the two-phase sampling for nonresponse which is based on subsampling initial nonrespondents and recontacting them. Another popular method known as weighting adjustment relies on constructing the estimator using weights dependent on individual response probabilities which are known in advance or estimated. In this paper an attempt is made to construct another mean value estimator by combining these two approaches, and assuming that individual response probabilities are estimated using nonparametric methods. The results of simulation experiments assessing its properties are presented.