PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 165
Społeczna rola statystyki 2011

Janusz L. Wywiał
Uniwersytet Ekonomiczny w Katowicach

SYMULACYJNA ANALIZA DOKŁADNOŚCI OCENY WARTOŚCI PRZECIĘTNEJ ZA POMOCĄ STRATEGII ZALEŻNYCH OD RÓŻNICY STATYSTYK POZYCYJNYCH ZMIENNEJ DODATKOWEJ*

Streszczenie: Praca dotyczy estymacji wartości średniej lub globalnej w populacji skończonej i ustalonej. Wnioskowanie jest wspomagane obserwacjami w całej populacji cechy dodatkowej. Brane są pod uwagę strategie estymacji zależne z jednej strony od planu losowania proporcjonalnego do dodatniej różnicy dwóch statystyk pozycyjnych, a z drugiej – od trzech estymatorów typu regresyjnego. Konstrukcja jednego z tych estymatorów została zainspirowana ideą regresji wyznaczanej metodą dwóch punktów, o których najpierw czerpałem wiadomości z prac Hellwiga [1; 2]. W artykule przeprowadzono analizę porównawczą dokładności tych estymatorów na podstawie badania symulacyjnego.

Słowa kluczowe: metoda reprezentacyjna, plan losowania próby, estymator regresyjny, statystyki pozycyjne, statystyka Horvitza–Thompsona.

1. Wstęp

Analizę porównawczą dokładności estymatorów typu regresyjnego przeprowadzono na podstawie komputerowego badania symulacyjnego. Rezultaty tych badań stanowią przyczynek do analizy praktycznej użyteczności strategii estymacji zależnych od tzw. cech dodatkowych obserwowanych w całej populacji. Okazuje się, że tego typu strategie są użyteczne wszędzie tam, gdzie jest możliwość obserwacji, jeszcze przed losowaniem próby, właśnie cech dodatkowych. W szczególności źródłem cech dodatkowych są Rejestr Rolny oraz system identyfikacji przedsiębiorstw REGON. Rejestr Rolny zawiera m.in. dane o powierzchni wszystkich polskich gospodarstw rolnych, co jest wykorzystywane do konstrukcji właśnie planów losowania prób lub estymatorów wartości globalnej np. plonów pszenicy. W REGON znajdujemy dane o liczbie zatrudnionych w przedsiębiorstwach, które także stanowią inspirację do tworzenia strategii losowania, które już z powodzeniem są wykorzystywane

* Artykuł jest rezultatem projektu nr N N111434137 finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego.

w praktyce badań reprezentacyjnych. Ponadto za obserwacje cechy dodatkowej są uważane wartości nominalne faktur, które podlegają kontroli zgodnej z odpowiednimi procedurami audytowymi.

2. Plan losowania próby

Populację skończoną i ustaloną o liczebności N oznaczmy przez $U = \{1, \ldots, N\}$, a przez $s \subseteq U$ próbę prostą losowaną bezzwrotnie. Obserwacje zmiennej badanej oznaczamy przez y_i, a cechy dodatkowej przez x_i, $i = 1, \ldots, N$. Niech $X_{(u)}$ i $X_{(r)}$ będą statystykami pozycyjnymi zmiennej dodatkowej z n-elementowej próby prostej. Autor [8] zaproponował plan losowania próby proporcjonalny do różnicy statystyk pozycyjnych zmiennej dodatkowej, który można natychmiastowo uogólnić na wersję warunkową następująco:

$$P_{r,u}(s) = \frac{f(x_{(u)}, x_{(r)}, c)}{z(r,u,c)}, \quad x_{(u)} < x_{(r)}, \tag{1}$$

gdzie:

$$f(x_{(u)}, x_{(r)}, c) = \begin{cases} x_i - x_j & \text{dla} \quad x_i - x_j \geq c \\ 0 & \text{dla} \quad x_i - x_j < c \end{cases}, \tag{2}$$

$$z(r,u,c) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} g(r,u,i,j) f(x_i, x_j, c), \tag{3}$$

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}. \tag{4}$$

Ta definicja jest zgodna z ogólną ideą konstrukcji warunkowych planów losowania próby wprowadzoną przez Tillégo [5]. Niech $\delta(x) = 0$, gdy $x \leq 0$ oraz jeśli $x > 0$, to $\delta(x) = 1$. Zatem w szczególności $\delta(x)\delta(x-1) = \delta(x-1)$. Prawdopodobieństwa inkluzji rzędu pierwszego planu losowania są następujące. Dla $k < r$

$$\pi_k(r,u,c) = \frac{\delta(r-1)\delta(r-k)}{z(r,u,c)} \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-2}{r-2} \times \\ \times \binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f\left(x_j, x_i, c\right), \tag{5}$$

dla $r \leq k \leq N - n + u$

$$\pi_k(r,u,c) = \frac{\delta(k-r+1)\delta(N-n+u-k+1)}{z(r,u,c)}.$$

$$\cdot\Bigg(\delta(k-u)\delta(n-u)\sum_{i=r}^{k-u+r-1}\sum_{j=i+u-r}^{k-1}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f\left(x_j,x_i,c\right)+$$

$$+\delta(k-u+1)\binom{N-k}{n-u}\sum_{i=r}^{k-u+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f\left(x_k,x_j,c\right)+$$

$$+\delta(k-r)\delta(N-n+u-k)\sum_{i=r}^{k-1}\sum_{j=k+1}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j}{n-u}f\left(x_j,x_i,c\right)+$$

$$+\delta(N-n+r-k+1)\binom{k-1}{r-1}\sum_{j=k+u-r}^{N-n+u}\binom{j-k-1}{u-r-1}\binom{N-j}{n-u}f\left(x_j,x_k,c\right)+$$

$$+\delta(N-n+r-k)\sum_{i=k+1}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}f\left(x_j,x_i,c\right)\Bigg), \tag{6}$$

dla $k > N - n + u$

$$\pi_k(r,u,c)=\frac{\delta(k-N+n-u)}{z(r,u,c)}\times$$
$$\times\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f\left(x_j,x_i,c\right). \tag{7}$$

Dodajmy, że dla $c = 0$ rozważany plan $P_{r,u}(s|c)$ redukuje się do wersji bezwarunkowej $P_{r,u}(s|c)$. Dodajmy, że przestrzeń prób planu warunkowego zawiera się w przestrzeni prób planu bezwarunkowego.

Przy założeniu, że $x_i < x_j$, $i,j = 1, \ldots, N$, Wywiał (2010) określa następujący schemat losowania próby realizujący wyżej określony plan. Załóżmy, że $s = s_1 \cup \{i\} \cup s_2 \cup \{j\} s_3$, gdzie $s_1 = \{k\colon x_k < x_i\}$, $s_2 = \{k\colon x_i < x_k < x_j\}$ i $s_3 = \{k\colon x_k > x_j\}$. Ponadto, niech

$$U = U(1, i-1) \cup \{i\} \cup U(i+1, j-1\} \cup U(j+1, N),$$

gdzie $U(1,i-1)=\{1,\ldots,i-1\}$, $U(i+1,j-1)=\{i+1,\ldots,j-1\}$, $U(i-1,N)=\{j+1,\ldots,N\}$. Wówczas schemat losowania próby określa wyrażenie

$$P_1(s_1)p_{r,u}(i|c)P_2(s_2)q_{r,u}(j|c)\,P_3(s_3) = P_{r,u}(s|c), \tag{8}$$

gdzie

$$P_1(s_1)=\binom{i-1}{r-1}^{-1},\qquad P_2(s_2)=\binom{j-i-1}{u-r-1}^{-1},\qquad P_1(s_1)=\binom{N-j}{n-u}^{-1}, \tag{9}$$

$$p_{r,u}(i|c) = P\left(X_{(r)} = x_i \mid X_{(u)} = x_j, c\right) = \frac{P\left(X_{(r)} = x_i, X_{(u)} = x_j, c\right)}{q_{r,u}(j|c)}, \tag{10}$$

$$P\left(X_{(r)} = x_i \mid X_{(u)} = x_j, c\right) = \frac{g(r,u,i,j) f(x_i, x_j, c)}{z(r,u,c)}, \tag{11}$$

$$q_{r,u}(j|c) = P\left(X_{(u)} = x_j, c\right) = \frac{1}{z(r,u,c)} \sum_{i=r}^{N-n+r} g(r,u,i,j) f(x_i, x_j, c). \tag{12}$$

Losowanie próby s polega najpierw na wylosowaniu elementu populacji $\{j\}$, $j = 1, \ldots, N$ zgodnie z rozkładem prawdopodobieństwa $q_{r,u}(j|c)$. Potem spośród elementów zbioru $U - \{j\}$ jest losowany element populacji $\{i\}$ zgodnie z prawdopodobieństwem $p_{r,u}(i|c)$. W końcu próby proste s_1, s_2 oraz s_3 są losowane bezzwrotnie odpowiednio ze zbiorów $U(1, i-1)$, $U(i+1, j-1\}$ i $U(j+1, N)$, zgodnie z planami losowania odpowiednio $P_1(s_1)$, $P_2(s_2)$ i $P_3(s_3)$.

3. Strategie estymacji

Średnią z próby prostej losowanej bezzwrotnie będącą nieobciążonym estymatorem przeciętnej w populacji oznaczamy przez $\left(\bar{y}_S, P_0(s)\right)$, gdzie

$$\bar{y}_S = \frac{1}{n} \sum_{k \in S}^{n} y_k, \qquad V(\bar{y}_S, P_0(s)) = \frac{N-n}{Nn} v_{0,2}, \tag{13}$$

przy czym $v_{0,2}$ wyjaśniono niżej. Zwykły estymator regresyjny z próby prostej, dalej oznaczany strategią $\left(\bar{y}_{rS}, P_0(s)\right)$, jest jedynie granicznie nieobciążonym dla przeciętnej w populacji, przy czym

$$\bar{y}_{rS} = \bar{y}_S + a_S\left(\bar{x} - \bar{x}_S\right), \tag{14}$$

$$V\left(\bar{y}_{rS}, P_0(s)\right) \approx \frac{N-n}{Nn} v(1-\rho^2), \tag{15}$$

$$\begin{aligned} b &= MSE\left(\bar{y}_{rS}, P_0(s)\right) - V\left(\bar{y}_{rS}, P_0(s)\right) \approx \\ &\approx \frac{N-n}{(N-2)n} \sqrt{v_{0,2}\left(\beta_2 - 1\right)}\left(\theta_{21} - \rho\theta_3\right), \end{aligned} \tag{16}$$

$$\gamma = \frac{b^2 100\%}{V\left(\bar{y}_{rS}, P_0(s)\right)} \approx \frac{N}{N-2} \frac{\left(\beta_2 - 1\right)\left(\theta_{21} - \rho\theta_3\right)^2}{1-\rho^2} 100\%, \tag{17}$$

$$\rho = \frac{v_{1,1}}{\sqrt{v_{2,0} v_{0,2}}}, \qquad \beta_2 = \frac{v_{0,4}}{v_{2,0}^2}, \qquad \theta_{2,1} = \frac{v_{2,1}}{\sqrt{v_{0,2}\left(v_{4,0} - v_{2,0}^2\right)}}, \tag{18}$$

$$\theta_{3,0}=\frac{v_{3,0}}{\sqrt{v_{2,0}\left(v_{4,0}-v_{2,0}^2\right)}},\qquad v_{r,u}=\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k-\overline{x}\right)\left(y_k-\overline{y}\right). \tag{19}$$

Przez b oznaczono wyprowadzone przez Wywiała [11] obciążenie strategii regresyjnej $\left(\overline{y}_{rS},P_0(s)\right)$, natomiast przez ρ oraz β_2 oznaczono odpowiednio współczynnik korelacji i kurtozy rozkładu zmiennej dodatkowej, a przez $\theta_{3,0}$ unormowany na przedziale [0;1] współczynnik asymetrii wprowadzony przez Wywiała [7; 8].

Następna z rozważanych strategii jest oznaczana parą $\left(t_{HTS},P_{r,u}(s\,|\,c)\right)$, gdzie znany estymator Horvitza–Thompsona [3] ma postać

$$\overline{y}_{HTS}=\frac{1}{n}\sum_{k\in S}\frac{y_k}{\pi_k(r,u,c)} \tag{20}$$

Autor [9] analizuje własności strategii estymacji postaci $\left(\overline{y}_{r,uS},P_{r,u}(s\,|\,c)\right)$, gdzie

$$\overline{y}_{r,uHTS}=\overline{y}_{HTS}+b_{r,uS}\left(\overline{x}-\overline{x}_{HTS}\right),\quad b_{r,uS}=\frac{Y_u-Y_r}{X_{(u)}-X_{(r)}},\quad \overline{x}_{HTS}=\frac{1}{n}\sum_{k\in S}\frac{x_k}{\pi_k(r,u,c)} \tag{21}$$

Ta statystyka jest otrzymana z równania linii prostej przechodzącej przez punkty $(X_{(r)},Y_r)$ i $(X_{(r)},Y_r)$. Taki sposób wyznaczania regresji jest odporny na występowanie wartości oddalonych. Inspiracją konstrukcji statystyki $\overline{y}_{r,uHTS}$ jest rozważana przez Walda [6] lub Hellwiga [2] funkcja liniowa regresji przeprowadzana przez dwa punkty, którymi są średnie z odpowiednio uciętej próby z dwuwymiarowego rozkładu zmiennej, przy czym Wald rozważał problem obserwacji zmiennej objaśnianej i zmiennej objaśniającej, które są zanieczyszczone losowymi błędami pomiaru. Z kolei Hellwig zajmujący się m.in. zagadnieniami bliskimi ekonometrii rozważał zmienną objaśnianą jako losową, a objaśniającą jako z góry ustaloną. To podejście pozwoliło Hellwigowi [1] wykazać m.in. nieobciążoność i zgodność oceny współczynnika kierunkowego liniowej regresji wyznaczanej naszkicowaną metodą dwóch punktów.

Ostatnią z branych pod uwagę strategii oznaczono przez $\left(\overline{y}_{r,uHTS},P_{r,u}(s\,|\,c)\right)$, gdzie

$$\overline{y}_{r,uHTS}=\overline{y}_{HTS}+b_{HTS}\left(\overline{x}-\overline{x}_{HTS}\right),\qquad b_{HTS}=\frac{v_{xyHTS}}{v_{xHTS}},\qquad v_{xHTS}=v_{xxHTS}, \tag{22}$$

$$v_{xyHTS}=\frac{1}{N_{HTS}}\sum_{k\in S}\frac{\left(x_k-\hat{x}_{HTS}\right)\left(x_k-\hat{x}_{HTS}\right)}{\pi_k(r,u,c)},\qquad N_{HTS}=\sum_{k\in S}\frac{1}{\pi_k(r,u,c)}, \tag{23}$$

$$\hat{x}_{HTS}=\frac{N}{N_{HTS}}\overline{x}_{HTS},\qquad \hat{y}_{HTS}=\frac{N}{N_{HTS}}\hat{x}_{HTS}. \tag{24}$$

Z ogólnych rezultatów zamieszczonych np. w pracy [4] wynika, że w szczególności strategia $\left(\overline{y}_{r,uHTS}, P_{r,u}(s \mid c)\right)$ daje zgodne oceny średniej w populacji zmiennej badanej.

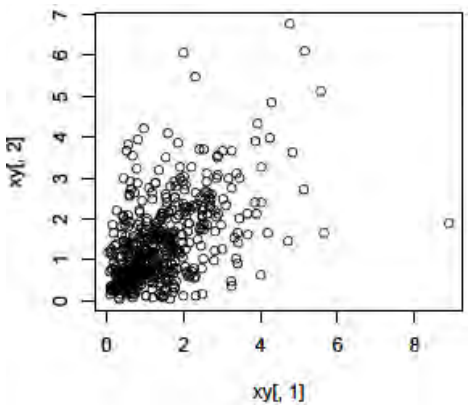
4. Porównanie dokładności estymacji

Dokładność strategii estymacji analizowano na podstawie przeprowadzonego komputerowego badania symulacyjnego. W referacie autora wygłoszonym na VI konferencji „Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych 2009" pokazano, również na podstawie analiz symulacyjnych, że strategie kwantylowo-regresyjne dają dokładniejsze oceny niż zwykła strategia regresyjna, gdy w populacji występują jednocześnie wartości oddalone (inaczej nazywane nietypowymi lub odstającymi) zmiennej badanej i dodatkowej. Ponadto wysunięto przypuszczenie, że również strategia $\left(\overline{y}_{r,uS}, P_{r,u}(s \mid c)\right)$ może być użyteczna w praktyce, gdy łączny rozkład zmiennej badanej i dodatkowej w populacji jest rozkładem silnie prawostronnie asymetrycznym. Ponadto zauważono, że dokładność strategii regresyjno-kwantylowych jest największa, gdy $r = 1$ oraz $u = n$. W związku z tym dalej rozważamy tylko przypadek planu losowania $\left(\overline{y}_{1,n,S}, P_{r,u}(s \mid c)\right)$. Założenia komputerowo-symulacyjnego eksperymentu były następujące: generowano pseudowartości dwuwymiarowego rozkładu wykładniczego (X, Y): $X = Z + U$, $Y = Z + V$, gdzie U, V oraz Z są niezależne, przy czym $Z \sim \exp(\alpha^{-1})$, $V \sim \exp(\beta^{-1})$, $U \sim \exp(\beta^{-1})$, $\alpha^2 + \beta^2 = 1$, $\rho(X, Y) = \alpha^2$. Stąd wynika, że ten dwuwymiarowy rozkład ma jeden parametr $\rho = \rho(X, Y) = \alpha^2$. Za pomocą przygotowanego programu komputerowego wygenerowano po 300 obserwacji tej dwuwymiarowej zmiennej dla parametrów $\rho = 0{,}5$, $\rho = 0{,}8$ i $\rho = 0{,}95$. Ich wykresy rozrzutu prezentują rys. 1-3. Wiadomo, że faktyczne współczynniki korelacji w wygenerowanych zbiorach obserwacji dwuwymiarowej zmiennej różnią się od założonych dla rozkładów teoretycznych. Te rozbieżności wynikają z wierszy 1 i 2 tab. 1. Następnie przeprowadzono symulacyjne wyznaczanie wartości błędu średniokwadratowego rozważanych strategii estymacji przeciętnej w populacji. Względne współczynniki efektywności wyznaczono na podstawie wzoru:

$$e(t_S, P(s)) = \frac{MSE(t_S, P(s))}{V(\overline{y}_S, P_0(s))} 100\%. \tag{25}$$

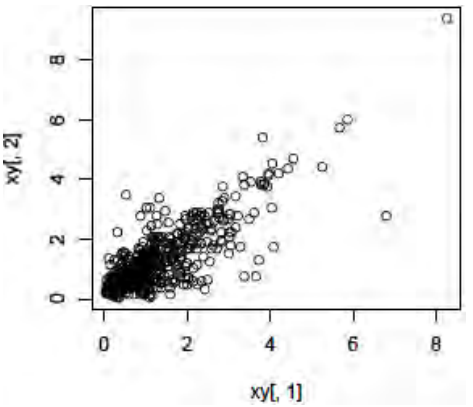
W szczególności dalej są używane następujące skróty oznaczeń współczynników względnej efektywności $e_1 = e\left(\overline{y}_{rS}, P_0(s)\right)$, $e_2 = e\left(\overline{y}_{HTS}, P_{r,u}(s \mid c)\right)$, $e_3 = e\left(\overline{y}_{r,uHTS}, P_{r,u}(s \mid c)\right)$ i $e_4 = e\left(\overline{y}_{r,uS}, P_{r,u}(s \mid c)\right)$.

Program komputerowy najpierw wyliczał prawdopodobieństwa inkluzji rzędu pierwszego, potem tysiąckrotnie losował próbę n-elementową spośród populacji 300 obserwacji dwuwymiarowej zmiennej. Następnie dla każdej z tak dobranych prób wyznaczano wartości estymatorów. W końcu na podstawie tak wyliczonych tysiąca wartości estymatora wyznaczano błąd średniokwadratowy. Dodać należy, że rów-



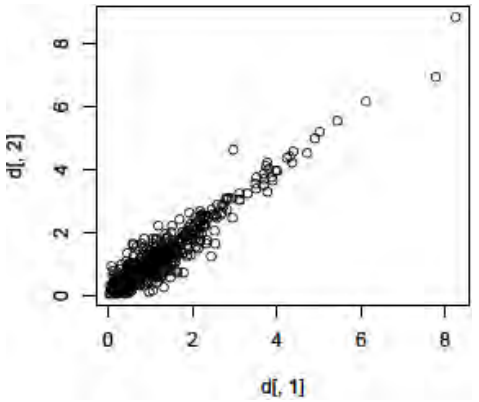
Rys. 1. Wykres rozrzutu dla $\rho = 0{,}5$

Źródło: opracowanie własne.



Rys. 2. Wykres rozrzutu dla $\rho = 0{,}8$

Źródło: opracowanie własne.



Rys. 3. Wykres rozrzutu dla $\rho = 0{,}95$

Źródło: opracowanie własne

nież tą drogą symulacyjną wyznaczano błąd średniokwadratowy zwykłego estymatora regresyjnego z próby prostej, dlatego by uwzględnić obciążenie tego estymatora, którego poziom (por. [11]) może być znaczny z powodu silnej prawostronnej asymetrii rozkładu cech badanej i dodatkowej, jakkolwiek dla dużych wartości N oraz n te obciążenia w przybliżeniu nie są stabilne, ponieważ stosunkowo mała zmiana np. wartości kurtozy rozkładu zmiennej dodatkowej powoduje dużą wartość obciążenia względnego estymatora regresyjnego, co wynika z tab. 1.

Z rysunków 1- 3 oraz z tabeli 1 wynika, że wraz ze wzrostem wartości współczynnika korelacji ρ rozproszenie łącznego rozkładu wartości dwuwymiarowej cechy (X, Y) zmniejsza się, przy czym wskaźniki rozproszenia i skośności zmiennych składowych są w przybliżeniu podobne.

Tabela 1. Parametry symulowanych rozkładów empirycznych

Parametr	Rysunek		
	1	2	3
ρ	0,500	0,800	0,950
r	0,562	0,826	0,959
$\overline{x}$	1,382	1,290	1,246
$\overline{y}$	1,389	1,275	1,244
v_{20}	0,935	1,040	1,047
v_{02}	0,990	1,105	1,064
β_2	5,689	11,071	5,315
θ_{21}	0,378	0,677	0,701
θ_{30}	0,668	0,667	0,706
γ %	0,0	50,6	3,3

Źródło: obliczenia własne.

Tabela 2. Współczynniki względnej efektywności strategii (w %), $c = 0$

n	$\rho = 0{,}5$				$\rho = 0{,}8$				$\rho = 0{,}95$			
	e_1	e_2	e_3	e_4	e_1	e_2	e_3	e_4	e_1	e_2	e_3	e_4
3	325	174	93	276	148	119	39	204	47	84	11	52
9	82	118	87	1316	41	126	37	120	11	96	11	46
15	79	117	77	373	39	110	38	168	9	242	10	305
30	73	117	77	234	36	103	34	396	8	516	9	796

Źródło: obliczenia własne.

Tabela 2 prezentuje wartości współczynników względnej efektywności rozważanych strategii estymacji. Z jej analizy wynika, że dokładność strategii estymacji $\left(\overline{y}_{rS}, P_0(s)\right)$ i $\left(\overline{y}_{r,uHTS}, P_{r,u}(s \mid c)\right)$ jest podobna oraz że są one znacznie dokładniejsze od pozostałych dwóch strategii.

Tabela 3. Współczynniki względnej efektywności strategii (w %), $c = \sqrt{v_x}$

n	$\rho = 0{,}5$			$\rho = 0{,}8$			$\rho = 0{,}95$		
	e_2	e_3	e_4	e_1	e_2	e_3	e_1	e_2	e_3
3	351	104	363	228	46	207	1237	9	202
9	359	93	327	124	43	106	104	11	17
15	154	84	130	102	36	70	274	11	15
30	109	76	111	99	34	78	101	9	14

Źródło: obliczenia własne.

Tabela 4. Współczynniki względnej efektywności strategii (w %), $c = 2\sqrt{v_x}$

n	$\rho = 0{,}5$			$\rho = 0{,}8$			$\rho = 0{,}95$		
	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3
3	328	96	172	465	59	418	161	11	16
9	149	92	107	198	56	105	424	28	66
15	123	87	102	129	45	73	180	15	16
30	104	81	95	98	31	56	104	9	10

Źródło: obliczenia własne.

Tabela 5. Współczynniki względnej efektywności strategii (w %), $c = 3\sqrt{v_x}$

n	$\rho = 0{,}5$			$\rho = 0{,}8$			$\rho = 0{,}95$		
	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3
3	667	108	274	183	56	60	223	12	18
9	208	108	120	128	42	42	96	10	9
15	173	97	114	94	36	32	97	9	9
30	145	78	97	89	34	31	165	8	8

Źródło: obliczenia własne.

Jednoczesna analiza zawartości tab. 2-5 prowadzi do wniosku, że dokładność strategii $\left(\overline{y}_{rS}, P_0(s)\right)$ i $\left(\overline{y}_{r,uHTS}, P_{r,u}(s \mid c)\right)$ rośnie wraz ze wzrostem współczynnika korelacji między zmiennymi badaną i dodatkową oraz wraz ze wzrostem liczebności próby, czego należało oczekiwać.

Ciekawą rzeczą jest tendencja wzrostu dokładności strategii $\left(\overline{y}_{r,uS}, P_{r,u}(s \mid c)\right)$ wraz ze wzrostem wartości parametru c warunkowego planu losowania próby, który kolejno przybiera wartości *0,* $\sqrt{v_x}$ *,* $2\sqrt{v_x}$ *i* $3\sqrt{v_x}$. Wzrost tych wartości skutkuje stopniowym zawężeniem przestrzeni możliwych do wylosowania prób do takich, w których obserwowane w nich różnice kwantyli $X_{(u)}$-$X_{(r)}$ przekraczają wartość c. Z konstrukcji estymatora $\overline{y}_{r,uS}$, a dokładniej występującego w nim kwantylowego estymatora $b_{r,uS} = \dfrac{Y_u - Y_r}{X_{(u)} - X_{(r)}}$ współczynnika kierunkowego regresji wynika, iż należy oczekiwać dużego zróżnicowania wartości statystyki $b_{r,uS}$, gdy wartość warunku c jest mała i odwrotnie. To pociąga duże zróżnicowanie (dużą wariancję) estymatora $\overline{y}_{r,uS}$ dla małych wartości warunku c. Zatem należy oczekiwać, że zwiększanie wartości warunku c prowadzi do poprawy dokładności strategii $\left(\overline{y}_{r,uS}, P_{r,u}(s \mid c)\right)$. Wartość warunku c jednak nie może być zwiększana bez pojawienia się negatywnych konsekwencji. Przekroczenie przez wartość parametru c pewnej granicznej liczby c_0 powoduje to, że przestrzeń prób będzie tak zawężona, iż we wchodzących

w jej skład wszystkich próbach zabraknie pewnych elementów populacji, co spowoduje, że przypisane im prawdopodobieństwa inkluzji będą równe zero. To z kolei może doprowadzić do obciążenia estymacji przeciętnej w populacji prowadzonej za pomocą strategii zależnych od planu losowania $P_{r,u}(s|c)$. W końcu to obciążenie może doprowadzić do istotnego wzrostu wariancji.

Analiza tab. 2-5 prowadzi do wniosku, że strategia $\left(\overline{y}_{HTS}, P_{r,u}(s \mid c)\right)$ wykorzystująca estymator Horvitza–Thompsona jest ewidentnie najmniej dokładną spośród analizowanych.

Konkludując przeprowadzoną analizę, dochodzimy do wniosku, że należy oczekiwać, iż odpowiednie zmniejszanie wartości parametru c planu $P_{r,u}(s|c)$ prowadzi do wzrostu dokładności estymacji strategii estymacji zależnych od tego planu. Z kolei najdokładniejszymi w każdych warunkach określonych przez założenia symulacji okazały się strategia $\left(\overline{y}_{r,uHTS}, P_{r,u}(s \mid c)\right)$ oraz zwykły estymator regresyjny z próby prostej.

Literatura

[1] Hellwig Z., *Regresja liniowa i jej zastosowanie w ekonomii,* PWN, Warszawa 1963.

[2] Hellwig Z., *Wyznaczanie parametrów regresji liniowej metodą dwóch punktów*, „Zastosowania Matematyki" 1956, nr 3, s. 60-81.

[3] Horvitz D.G., Thompson D.J., *A generalization of sampling without replacement from finite universe*, „Journal of the American Statistical Association" 1952, vol. 47, s. 663-685.

[4] Särndal C.E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, Springer Verlag, New York 1992.

[5] Tillé Y., *Sampling Algorithms*, Springer, New York 2006.

[6] Wald A., *The fitting the straight lines if both variables are subject to errors*, „Annals of Mathematical Statistics" 1940, vol. 11, s. 284.

[7] Wywiał J.L., *Badanie zmienności parametrów rozkładów warunkowych*, „Przegląd Statystyczny" 1983, vol. 30, s. 213-222.

[8] Wywiał J.L., *O pewnych unormowanych współczynnikach asymetrii i spłaszczenia*, „Przegląd Statystyczny" 1981, vol. 28, s. 263-269.

[9] Wywiał J.L., *Performing quantiles in regression sampling strategy*, „Model Assisted Statistics and Applications" 2009, vol. 4, no. 2, s. 131-142.

[10] Wywiał J.L., *Sampling designs proportionate to non-negative functions of two quantiles of auxiliary variable*, praca recenzowana w „Journal of Statistical Research", 2010.

[11] Wywiał J.L., *Statystyczna metoda reprezentacyjna w badaniach ekonomicznych*, Prace Naukowe Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Wydawnictwo Akademii Ekonomicznej, Katowice 1992.

ACCURACY OF POPULATION MEAN ESTIMATION ON THE BASIS OF STRATEGIES DEPENDENT ON DIFFERENCE OF AUXILIARY VARIABLE ORDER STATISTICS. SIMULATION ANALYSIS

Summary: Estimation of the population average in a finite population by means of sampling strategy dependent on the sample quantile of an auxiliary variable is considered. The sampling design is proportionate to the difference of two quantiles of an auxiliary variable. The derived inclusion probabilities are applied to estimation the population mean using the well known Horvitz-Thompson estimator which is called Horvitz-Thompson-regression estimator. The next estimator was the regression estimator whose parameters are assessed by the Horvitz-Thompson estimator. Particularly, the quantile-regression estimator is defined as the function of the slope coefficient dependent on the quantiles (order statistics) of the auxiliary variable. The construction of the slope coefficient was inspired by the estimator of this coefficient considered by Hellwig. The simulation analysis of accuracy estimators leads to conclusion that in general Horvitz-Thompson-regression and the ordinary regression one are the best among the considered ones.