A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 271, 2012

Grażyna Dehnel[*]

ESTYMACJA LOKALNA W SZACOWANIU INFORMACJI O DZIAŁALNOŚCI GOSPODARCZEJ MIKROPRZEDSIĘBIORSTW

Streszczenie. W wielu badaniach prowadzonych metodą reprezentacyjną często napotykamy na obserwacje, które znacząco różnią się, pod względem wartości badanych zmiennych, od pozostałych jednostek wylosowanych do próby. Dotyczy to zwłaszcza statystyki gospodarczej. Wpływ obserwacji odstającej na wartości estymatorów może być bardzo duży, zwłaszcza jeśli szacunek prowadzony jest na niskim poziomie agregacji. Należy jednak pamiętać, że jest ona jednym z elementów badanej zbiorowości i nie powinna być całkowicie pomijana w analizie. Stąd też konieczne jest prowadzenie badań dotyczących zastosowania nowych, nieklasycznych, technik estymacji, które są bardziej odporne na wartości odstające. W niniejszym artykule podjęto próbę zastosowania regresji lokalnej uwzględniającej lokalne zmiany w badaniu z zakresu statystyki gospodarczej.

Słowa kluczowe: estymacja jądrowa, statystka małych obszarów, statystyka gospodarcza.

I. WSTĘP

Wiele zmiennych opisujących podmioty gospodarcze charakteryzuje się silną prawostronną asymetrią, znacznym zróżnicowaniem i dużą koncentracją. Ponadto u wielu jednostek pojawiają się zerowe wartości zmiennych. Własności klasycznych estymatorów stosowanych w metodzie reprezentacyjnej takie jak nieobciążoność, czy duża efektywność w przypadku takich rozkładów zmiennych nie zostają zachowane. Ważny problem stanowi również obecność obserwacji odstających, których wpływ na wartości estymatorów może być bardzo duży, zwłaszcza jeśli szacunek prowadzony jest na niskim poziomie agregacji. Należy zatem poszukiwać metod estymacji, które w takich warunkach dostarczałyby wiarygodnych szacunków.

W niniejszym artykule podjęto próbę empirycznej weryfikacji możliwości wykorzystania lokalnego estymatora regresyjnego do szacowania informacji o działalności gospodarczej małych przedsiębiorstw na niskim poziomie agregacji tj. w przekroju województw i sekcji PKD. Celem badania było porównanie i ocena precyzji szacunku lokalnego estymatora regresyjnego ze znajdującym obecnie szerokie zastosowanie w praktyce badań statystycznych estymatorem GREG.

[*] Dr hab., Katedra Statystyki, Wydział Informatyki i Gospodarki Elektronicznej, Uniwersytet Ekonomiczny w Poznaniu.

II. ESTYMATOR GREG

Estymator GREG parametru Y można przedstawić jako sumę dwóch składników. Pierwszy z nich to model regresji, w którym uwzględnione są różnice między wartościami zmiennych pomocniczych dla populacji i dla próby w danym małym obszarze (domenie). Drugi składnik stanowi oszacowanie obciążenia.

$$\hat{Y}_{greg} = \sum_{i \in N} \hat{y}_i + \sum_{i \in s} w_i e_i \tag{1}$$

$$\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} \quad \text{gdzie: } \hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i \right) \tag{2}$$

gdzie:

y_i – wartość zmiennej badanej u *i*-tej jednostki,

$\mathbf{x}_i$ – wektor zmiennych pomocniczych u *i*-tej jednostki,

w_i – waga wynikające ze schematu losowania u *i*-tej jednostki,

$e_i = y_i - \hat{y}_i$ – składnik resztowy,

s – próba.

Wartość estymatora GREG zależy od dwóch rodzajów wag: wag wynikających ze schematu losowania oraz wag wyznaczanych na podstawie wartości zmiennej dodatkowej u jednostek, które zostały wylosowane do próby. Tę zależność można wyrazić wzorem (3):

$$\hat{Y}_{reg} = \sum_{i \in s} w_i g_i y_i \tag{3}$$

$$g_i = 1 + \left(X - \hat{X}_{HT} \right)' \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left(\mathbf{x}_i \right) \tag{4}$$

gdzie:

g_i – wagi zależne od wartości cechy *x* jednostek wylosowanych do próby

$\mathbf{X} = \left(X_1, ..., X_p \right)'$ – wektor wartości globalnych zmiennych *x* (5)

$\hat{\mathbf{X}}_{HT} = \sum_{i \in s} w_i \mathbf{x}_i$ – wektor bezpośrednich estymatorów Horvitza-Thompsona wartości globalnych zmiennych *x*

Zaletą estymatorów GREG jest to, że pozwalają na wykorzystanie różnorodnych zmiennych pomocniczych. Dotyczy to zarówno ilości zmiennych, ich rodzaju (ilościowe, jakościowe) oraz poziomu agregacji (dane jednostkowe i zagregowane).

W oparciu o rezultaty badań dotyczące statystyki gospodarczej, w których wykorzystano estymację GREG, można stwierdzić, że odpowiednie dobranie i wykorzystanie zmiennych pomocniczych może w znacznym stopniu wpłynąć na poprawę precyzji szacunku [Chambers, Falvey, Hedlin i Kokic, 2001]. Mamy tu bowiem do czynienia z zastosowaniem pewnego rodzaju kalibracji poprzez zmienne dodatkowe, prowadzącej do zmniejszenia obciążenia wynikającego z błędów nielosowych.

Estymator uogólniony regresyjny może być stosowany w przypadku różnych schematów losowania, gdyż uwzględnia prawdopodobieństwa wyboru jednostki do próby. W przypadku zmiennych binarnych preferowana jest wersja logitowa estymatora [Lehtonen, Veijanen, 1998].

Estymator GREG jest nieobciążony, jeśli spełniony jest warunek dobrego dopasowania modelu do danych. Jednakże, jeśli w próbie znajdzie się kilka jednostek, dla których reszty będą bardzo duże, wówczas zastosowanie tego estymatora może spowodować duże niedoszacowanie lub przeszacowanie wartości globalnej badanej zmiennej. Wyniki przeprowadzonych badań pokazują, jak ważną rolę odgrywa tu wybór „dobrego" modelu [Chambers, Falvey, Hedlin i Kokic, 2001]. W sytuacji złego dopasowania modelu, drugi składnik estymatora (por. wzór 1), będący oszacowaniem obciążenia (na podstawie reszt), przyjmuje wartość znacznie większą, niż składnik pierwszy, będący wartością teoretyczną, wyznaczaną na podstawie modelu.

Jednym z zagrożeń, jakie niesie za sobą stosowanie estymatora GREG jest to, że w skrajnych przypadkach może prowadzić do ujemnych szacunków [Chambers, Falvey, Hedlin i Kokic, 2001]. Sytuacja taka ma miejsce, gdy wagi g_i wyznaczane na podstawie cechy dodatkowej u jednostek wylosowanych do próby przybierają wartości ujemne. Kolejne zagrożenie dotyczy przyjętego w teorii założenia, że iloczyn wag w_i i g_i jest bliski wartości wagi w_i wynikającej ze schematu losowania (co oznacza, że waga g_i powinna być bliska jedności) [Deville, Särndal, 1992]. W praktyce nie zawsze jest ono spełnione. W badaniach empirycznych różnica między wartościami wag w_i i g_i bywa znaczna. W przypadku dużych prób prowadzi to do zniekształcenia ich struktur.

Przedstawione wyżej własności estymatora GREG mogą prowadzić do obciążonych szacunków charakteryzujących się niską precyzją. Ma to miejsce szczególnie w przypadku obecności w próbie obserwacji odstających, czy znacznego odsetka jednostek z zerowymi wartościami cech. Jeśli jednak zastąpimy pierwszy składnik estymatora GREG, który jest estymatorem zmiennej badanej opartym na modelu regresji ($\hat{y}_i$),szacunkiem wartości globalnej zmiennej badanej dokonanym na podstawie estymacji jądrowej ($\hat{y}_{loc,i}$). Dzięki takie-

mu zabiegowi estymator jest mniej wrażliwy na obserwacje odstające oraz na nieliniową zależność pomiędzy zmienną badaną i pomocniczą.

III. LOKALNY ESTYMATOR REGRESYJNY

Lokalny estymator regresyjny można przedstawić za pomocą wzoru [por. Breidt, Opsomer, 2000)]:

$$\hat{Y}_{loc} = \sum_{i \in U} \hat{y}_{loc,i} + \sum_{i \in s} w_i \left(y_i - \hat{y}_{loc,i} \right) \tag{6}$$

lub jako estymator oparty na modelu w którym szacunku dokonuje się dla każdej jednostki z populacji generalnej [Chambers, Dorfman, Wehrly, 1993; Dorfman, 2000]:

$$\hat{y}_{loc,i} = \mathbf{c}_j^{'} \left(\mathbf{D}_i' \mathbf{W}_i \mathbf{D}_i \right)^{-1} \mathbf{D}_i' \mathbf{W}_i \mathbf{y}_s \quad i=1, 2, \ldots, N \tag{7}$$

gdzie: U oznacza populację generalną, natomiast s próbę, $\mathbf{c}_j^{'}$ to wektor z wartością *1* na *j*-tej pozycji i zerami na pozostałych miejscach, $\mathbf{D}_i$, $i = 1, 2, \ldots, N$, jest macierzą budowaną dla każdej jednostki z populacji generalnej na podstawie wartości zmiennej pomocniczej, o wymiarach $n \times 2$, każda z $\left[1 \quad \left(x_j - x_i\right)\right]$ w *j*-tym wierszu, $j = 1, 2, \ldots, n$, $\mathbf{W}_i$, dla $i = 1, 2, \ldots, N$, jest diagonalną macierzą budowaną dla każdej jednostki z populacji generalnej, o wymiarach $n \times n$ i $w_i b_i^{-1} K\left[\left(x_j - x_i\right) b_i^{-1}\right]$ na miejscu (j, j), gdzie $K(\cdot)$ jest funkcją jądrową i b_i jest szerokością pasma dla *i*-tej jednostki.

Podstawą lokalnego estymatora regresyjnego jest wartość $\hat{y}_{loc,i}$, która w wielu przypadkach jest zbliżona do wartości $\hat{y}_i$, wyznaczanej na podstawie klasycznej postaci estymatora GREG. Różnica między nimi polega na tym, że estymacja jądrowa w przypadku $\hat{y}_{loc,i}$ dzięki temu, że opiera się na wielu modelach budowanych na podstawie części próby, pozwala uwzględnić lokalne zmiany wartości zmiennej badanej, co w odniesieniu do pojedynczego, liniowego modelu regresji estymatora GREG jest niemożliwe.

Estymator (7) można przedstawić w postaci [Hedlin, 2004]:

$$\hat{y}_{loc,i} = \bar{y}_{loc,i} + \left(x_i - \bar{x}_{loc}\right) \frac{\sum_{j \in s} q_{ji} \left(x_j - \bar{x}_{loc}\right) y_j}{\sum_{j \in s} q_{ji} \left(x_j - \bar{x}_{loc}\right)^2} \tag{8}$$

$$q_{ji} = \max\left[0, \frac{3}{4}\left(1 - \left(\frac{\left(x_j - x_i\right)}{b_i^{(s)}}\right)^2\right)\right] \tag{9}$$

gdzie:
n – liczebność próby, $i = 1, 2, \ldots, n$
q_{ji} – diagonalne elementy macierzy $\boldsymbol{W_i}$,

$$\bar{y}_{loc} = \sum_{j \in s} q_{ji} y_j \left(\sum_{j \in s} q_{ji}\right)^{-1} \tag{10}$$

$$\bar{x}_{loc} = \sum_{j \in s} q_{ji} x_j \left(\sum_{j \in s} q_{ji}\right)^{-1} \tag{11}$$

Warto zauważyć, że $\bar{y}_{loc,i}$ jest szacunkiem wartości zmiennej badanej u i-tej jednostki dokonanym na podstawie estymacji lokalnej bez udziału zmiennych pomocniczych x.

Ważną rolę w regresji lokalnej odgrywa wybór postaci funkcji jądrowej oraz, niezwykle istotny, wybór odpowiedniej szerokości pasma. Zarówno postać funkcji jądrowej jak i szerokość pasma mają decydujący wpływ na efektywność uzyskanych wyników. W literaturze proponowane są różne podejścia [Chambers, Dorfman, Wehrly, 1993; Chambers, 1996; Kim, Breidt, Opsomer, 2001].

Funkcja jądrowa

W przeprowadzonym badaniu jako funkcję jądrową przyjęto jedną z najczęściej wykorzystywanych w estymacji jądrowej funkcję Epanechnikova [Hedlin, 2004]:

$$K\left(u_{ji}\right) = \max\left[0, \frac{3}{4}\left(1 - u_{ji}^2\right)\right] \tag{12}$$

gdzie:
$u_{ji} = \left(x_j - x_i\right) / b_i$ dla $i = 1, 2, \ldots, n$ oraz $j = 1, 2, \ldots, n$
b_i – szerokość pasma dla i-tej jednostki

$$K\left(u_{ji}\right)=\max\left[0,\frac{3}{4}\left(1-u_{ji}^2\right)\right]=\max\left[0,\frac{3}{4}\left(1-\left(\frac{\left(x_j-x_i\right)}{b_i}\right)^2\right)\right] \tag{13}$$

Funkcja jądrowa definiuje „okno" wokół każdej jednostki wylosowanej do próby. Jednostki znajdujące się poza min nie biorą udziału w szacunku wartości $\hat{y}_{loc,i}$. Warto zauważyć, że:

$$K\left(u_{ji}\right)=0 \quad \text{jeśli } u_{ji}=\left(x_j-x_i\right)\Big/b_i^{(s)}\geq 1 \tag{14}$$

Jeśli nie uwzględni się wartości wag zależnych od schematu losowania $w_i=\pi_i^{-1}$, to można przyjąć, że szacunek $\hat{y}_{loc,i}$ jest oparty na lokalnej standardowej regresji liniowej.

Funkcja jądrowa jest funkcją wag. W odniesieniu do każdej *i*-tej jednostki wyznaczane są niezależnie wartości wag dla wszystkich jednostek wylosowanych do próby. Tak więc, proces nadawania wag powtarzany jest tyle razy, ile wynosi liczebność próby. W pierwszej kolejności wokół wszystkich jednostek należących do próby określane są tak zwane „okna". Jednostkom znajdującym się poza „oknem" wyznaczonym dla *i*-tej jednostki przypisywana jest wartość wagi równa zero. Pozostałym jednostkom, należącym do „okna" określonego dla *i* -tej jednostki, nadawane są dodatnie wartości wag. Ich wielkość zależy od tego na ile poziom zmiennej pomocniczej różni się od poziomu tej zmiennej zanotowanego w przypadku jednostki *i*-tej (dla której zdefiniowano „okno"). Największe wartości wag nadawane są obserwacjom, u których wartość zmiennej pomocniczej *x* jest bliska wartości x_i.

Szerokość pasma

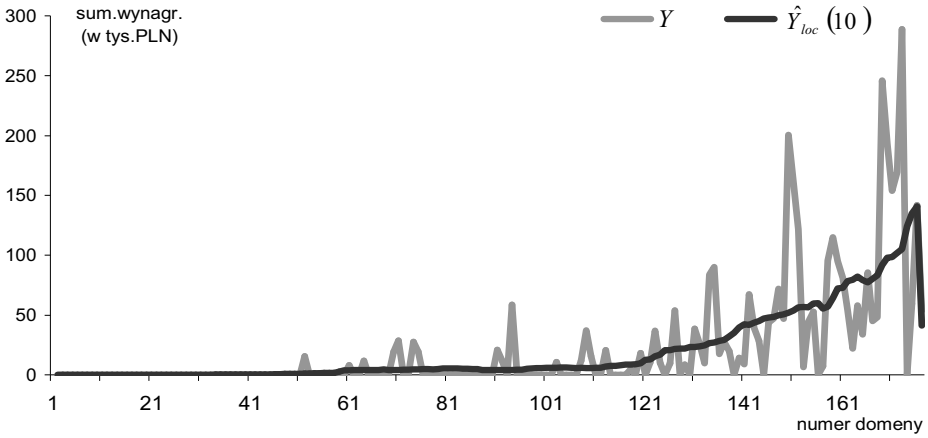
Istnieje wiele metod wyznaczania szerokości pasma. Można je podzielić na dwa rodzaje. Jeden rodzaj stanowią metody, w których określana jest tylko jedna, stała dla całej próby szerokość pasma. Drugi rodzaj reprezentują metody, w których zalecany jest dobór wielu szerokości pasma, a ich wartość jest związana z poszczególnymi obserwacjami z próby.

W przeprowadzonym badaniu wykorzystano cztery różne metody określania szerokości pasma:

1. $b_i=\frac{1}{4}\left(x_{\max}-x_{\min}\right)$ => $t_{loc}(\max\min)$
2. $b_i=x_{i+10}-x_{i-10}$ => $t_{loc}(10)$
3. $b_i=x_{i+20}-x_{i-20}$ => $t_{loc}(20)$
4. $b_i=x_{i+40}-x_{i-40}$ => $t_{loc}(40)$

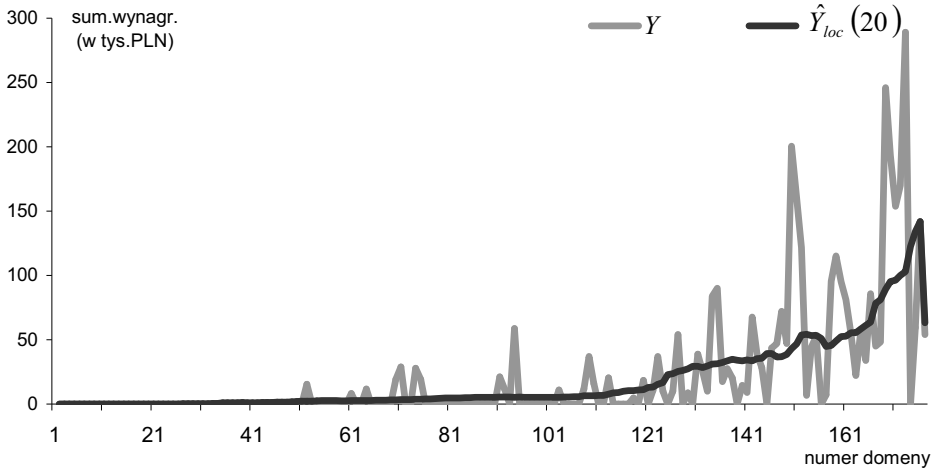
Zgodnie z pierwszą metodą szerokość pasma jest stała i wyznaczana jako ¼ z obszaru zmienności cechy pomocniczej.

Pozostałe trzy metody, określane są mianem metod „najbliższego sąsiada", traktują szerokość pasma jako wielkość zmienną. Parametr b_i stanowi różnicę pomiędzy wartościami zmiennej pomocniczej u dwóch jednostek wybieranych z wszystkich, posortowanych uprzednio według wzrastającej wartości zmiennej *x,* i oddalonych od jednostki x_i dla której jest ta szerokość określana odpowiednio na 10, 20 i 40 jednostek. Jeśli numer jednostki należącej do próby (oznaczony przez i, gdzie i przyjmuje wartości $i = 1....n$) jest tak mały, że nie można wyznaczyć jednostki o numerze $i-10$, $i-20$ lub $i-40$, a tym samym wartość zmiennej pomocniczej x_{i-10}, x_{i-20} lub x_{i-40} nie istnieje, to w zastępstwie za nią przyjmuje się minimalny poziom cechy x. Podobnie postępujemy w przypadku x_{i+10}, x_{i+20} i x_{i+40} biorąc wartość maksymalną [Hedlin, 2004] gdyż dla wąskiej szerokości pasma otrzymany na jego podstawie szacunek oparty jest na wielu modelach lokalnych. Jednak wraz ze wzrostem szerokości pasma w coraz większym stopniu przypomina klasyczny estymator typu GREG. Na wykresach przedstawiono wartości rzeczywiste zmiennej badanej oraz szacunek na podstawie czterech lokalnych estymatorów regresyjnych (por. rys. 1, 2, 3, 4). Dla każdego z estymatorów podano zakres w jakim zmieniała się szerokość pasma. Wraz z jej skróceniem w coraz większym stopniu uwzględniane były lokalne zmiany wartości zmiennej badanej.



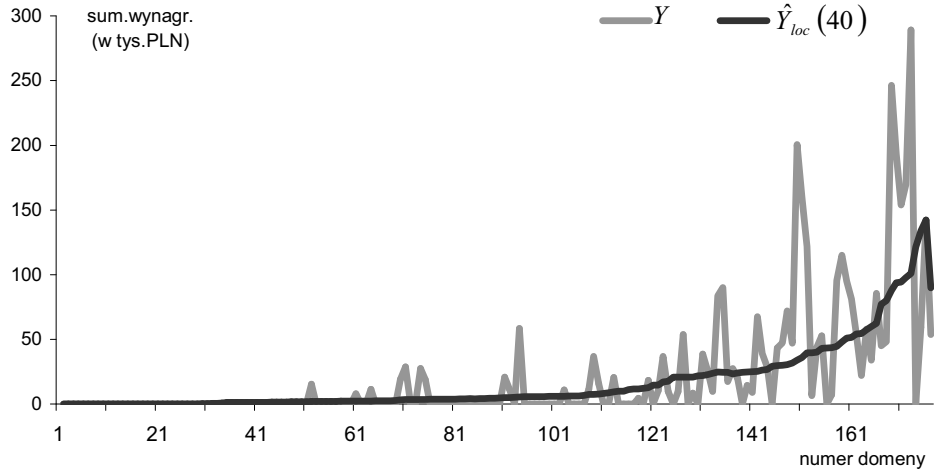
Rys. 1. Wartości rzeczywiste i szacunek sumy wynagrodzeń brutto na podstawie lokalnego estymatora $t_{loc}(10)$ w sekcji budownictwo (b_i <60 000 – 700 000>)

Źródło: Opracowanie własne na podstawie badania SP3 oraz rejestru podatkowego.



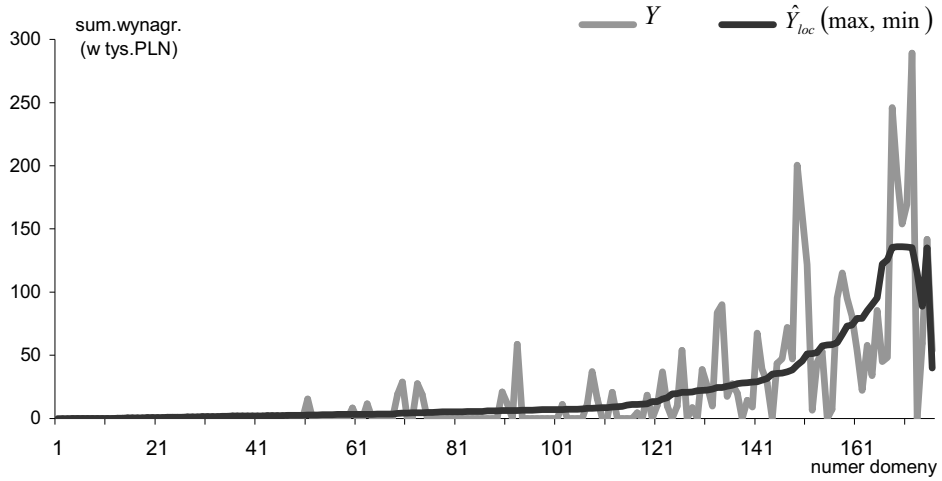
Rys. 2. Wartości rzeczywiste i szacunek sumy wynagrodzeń brutto na podstawie lokalnego estymatora $t_{loc}(20)$ w sekcji budownictwo (b_i <113 000 – 980 000>)

Źródło: Opracowanie własne na podstawie badania SP3 oraz rejestru podatkowego.



Rys. 3. Wartości rzeczywiste i szacunek sumy wynagrodzeń brutto na podstawie lokalnego estymatora $t_{loc}(40)$ w sekcji budownictwo (b_i <440 000–1 000 000>)

Źródło: Opracowanie własne na podstawie badania SP3 oraz rejestru podatkowego.



Rys. 4. Wartości rzeczywiste i szacunek sumy wynagrodzeń brutto na podstawie lokalnego estymatora $t_{loc}(\max \min)$ w sekcji budownictwo ($b_i = \frac{1}{4}(1015272{,}47 - 6097) = 252000$)

Źródło: Opracowanie własne na podstawie badania SP3 oraz rejestru podatkowego.

IV. BADANIE SYMULACYJNE

W celu porównania i oceny precyzji szacunku lokalnego estymatora regresyjnego i estymatora GREG przeprowadzono badania symulacyjne.

W badaniu wykorzystano dwa źródła informacji:

1) wyniki **badania SP3** przeprowadzonego w 2001 roku. Jest to badanie reprezentacyjne obejmujące mikroprzedsiębiorstwa. Stanowiło ono źródło informacji o zmiennej badanej – ***y.***

Próba wylosowana do badania SP3 w 2001 roku liczyła ponad 114 tysięcy jednostek (4%). Jednak ostatecznie informacje pozyskano jedynie od 44 807 podmiotów gospodarczych.

2) **zbiory danych z systemu podatkowego Ministerstwa Finansów** (rejestr podatkowy) – stanowiło 907580 zeznań podatkowych od osób fizycznych i prawnych.

Rejestr podatkowy wykorzystano jako źródło cech dodatkowych – ***x,*** których zadaniem jest poszerzenie informacji uzyskanych z badania SP3.

Estymacji dokonano dla zmiennej **suma wynagrodzeń brutto** (*y*).

Jako zmienną pomocniczą (*x)* wykorzystano zmienną **koszty.** Przy doborze zmiennej dodatkowej kierowano się przede wszystkim stopniem skorelowania informacji z badania SP3 oraz rejestru podatkowego.

Estymacji dokonano w przekroju: województwo i rodzaj prowadzonej działalności gospodarczej (sekcja PKD). Wyróżniono 160 domen (16 województw

x 10 sekcji PKD). Prezentację wyników w artykule zawężono do województwa zachodniopomorskiego w przekroju sekcji (por. tab.1):

Tabela 1. Wielkość próby w przekroju sekcji w województwie zachodniopomorskim

Sekcja		N	n	n/N (%)
Przetwórstwo przemysłowe	D	3430	338	9,85
Budownictwo	F	2844	176	6,19
Handel i naprawy	G	16856	611	3,62
Hotele i restauracje	H	2467	63	2,55
Transport, łączność	I	2681	170	6,34
Pośrednictwo finansowe	J	2559	119	4,65
Obsługa nieruchom. i firm, nauka	K	10967	200	1,82
Ochrona zdrowia i opieka społ.	N	5221	144	2,76
Pozostała działalność usługowa	O	1716	73	4,25
Suma		49793	1920	3,86

Źródło: Wyniki badania SP3.

Do wyznaczenia ocen precyzji badanych estymatorów zastosowano metodę bootstrapową. Wykonano 500 repetycji losowania podprób, na podstawie których wyznaczono wartość wariancji z ocen szacowanego parametru

$$Var\left(\hat{Y}\right)=\frac{1}{500-1}\sum_{b=1}^{500}\left(\hat{Y}_b-\hat{Y}\right)^2 \tag{15}$$

dla każdej iteracji dokonano modyfikacji oryginalnych wag wynikających ze schematu losowania $w_{i(b)}=w_i\frac{n}{n-1}m_{i(b)}$

gdzie:

$\hat{Y}_b$ – ocena szacowanego parametru na podstawie podróby b,

$\hat{Y}$ – ocena szacowanego parametru na podstawie całej próby,

b – numer repetycji $\left(b=1,2,....,500\right)$,

$m_{i(b)}$ – ile razy jednostka i-ta została wybrana do podróby b,

w_i – oryginalna waga jednostki i (wynikająca ze schematu losowania),

$w_{i(b)}$ – waga dla jednostki i w podpróbce b.

Oceny precyzji danego estymatora dokonano na podstawie dwóch parametrów. Jednym z nich był współczynnik zmienności estymatora

$$CV = \frac{\sqrt{Var}}{\hat{Y}} \tag{16}$$

Drugi parametr charakteryzował stopień redukcji zmienności lokalnego estymatora regresyjnego w przypadku zastosowania jednej z czterech metod określania szerokości pasma w porównaniu z estymacją bezpośrednią:

$$\mathrm{Re}\,dCV = \frac{CV(\hat{t}) - CV(\hat{t}_{DIR})}{CV(\hat{t}_{DIR})} \tag{17}$$

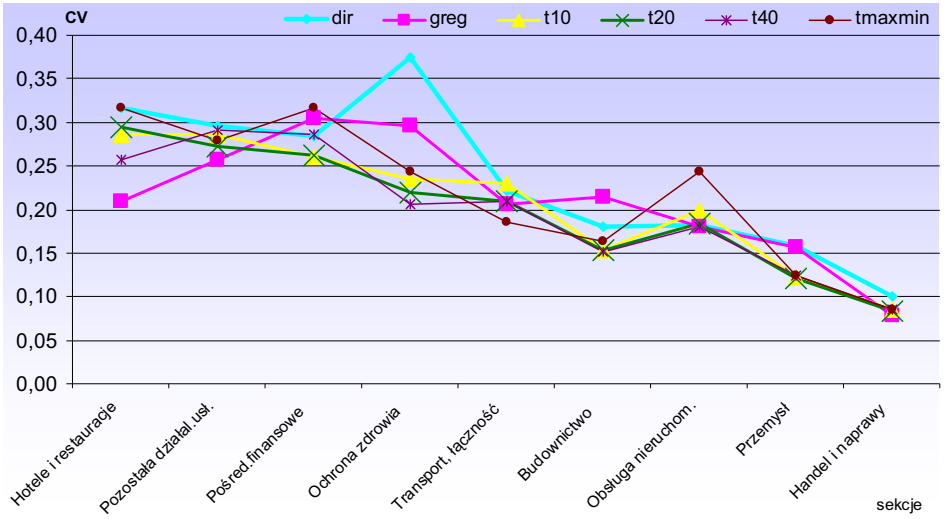
W tabeli 2 zamieszczono wartości współczynnika zmienności estymatora *(CV)* oraz jego podstawowe charakterystyki takie jak: wartość minimalna i maksymalna, średnia oraz mediana. W drugiej kolumnie znajdują się wielkości dotyczące klasycznej estymacji bezpośredniej. Kolumna trzecia odnosi się do estymacji typu GREG. Cztery kolejne kolumny charakteryzują zaś lokalny estymator regresyjny, przy czym każda z nich prezentuje inną metodę określania szerokości pasma. Dane zawarte w tabeli świadczą jednoznacznie o tym, że z największą zmiennością (największe wartości parametrów) mamy do czynienia w przypadku estymacji bezpośredniej. Zróżnicowanie precyzji estymatora lokalnego w porównaniu z estymatorem GREG jest nieduże. Niższe wartości współczynnika, a co za tym idzie i charakterystyk otrzymano dla lokalnego estymatora regresyjnego $t_{loc}(\max\min)$ (średnia 0,217; mediana 0,244). Jeszcze mniejsza zmienność charakteryzuje estymatory: $t_{loc}(10)$ (średnia 0,206; mediana 0,229) oraz klasyczny estymator GREG (średnia 0,211; mediana 0,209). Najniższy poziom zmienności dotyczy jednak lokalnych estymatorów regresyjnych, w których do określenia szerokości pasma zastosowano metodę „najbliższego sąsiada" $t_{loc}(40)$ (średnia 0,199; mediana 0,207), $t_{loc}(20)$ (średnia 0,200; mediana 0,209).

Szczegółowe wartości współczynników zmienności estymatorów w przekroju sekcji w województwie zachodniopomorskim przedstawiono na wykresie (por. rys. 5). Sekcje uporządkowano według rosnącej liczebności próby. W większości wyróżnionych domen stosunkowo wysoki poziom *CV* zanotowano dla estymacji bezpośredniej, zaś niski dla lokalnej estymacji regresyjnej $\hat{Y}_{loc}(20)$ oraz $\hat{Y}_{loc}(40)$. Domeny nielicznie reprezentowane w próbie charakteryzuje duża dyspersja ocen estymatorów. Ponadto można zauważyć, że w przypadku większości sekcji wraz ze wzrostem liczebności próby maleje zarówno zróżnicowanie wartości współczynników zmienności wyznaczonych dla różnych estymatorów, jak i poziom zmienności ocen estymatorów.

Tabela 2. Współczynniki zmienności estymatorów i jego charakterystyki (CV) w przekroju sekcji w województwie zachodniopomorskim

Sekcja\Estymator	**DIR**	**GREG**	$t_{loc}(10)$	$t_{loc}(20)$	$t_{loc}(40)$	$t_{loc}(\max\min)$
Przetwórstwo przemysłowe	0,16	0,16	0,12	0,12	0,12	0,12
Budownictwo	0,18	0,21	0,15	0,15	0,15	0,16
Handel i naprawy	0,10	0,08	0,09	0,08	0,09	0,08
Hotele i restauracje	0,32	0,21	0,29	0,29	0,26	0,32
Transport, łączność	0,22	0,21	0,23	0,21	0,21	0,19
Pośrednictwo finansowe	0,28	0,30	0,26	0,26	0,29	0,32
Obsługa nieruchom. i firm, nauka	0,18	0,18	0,20	0,18	0,18	0,24
Ochrona zdrowia i opieka społ.	0,37	0,30	0,23	0,22	0,21	0,24
Pozostała działalność usługowa	0,30	0,26	0,29	0,27	0,29	0,28
min	**0,100**	**0,078**	**0,085**	**0,083**	**0,085**	**0,085**
max	**0,374**	**0,304**	**0,285**	**0,294**	**0,292**	**0,316**
średnia	**0,235**	**0,211**	**0,206**	**0,200**	**0,199**	**0,217**
mediana	**0,221**	**0,209**	**0,229**	**0,209**	**0,207**	**0,244**

Źródło: Obliczenia własne.



Rys. 5. Wartości współczynników zmienności estymatorów w województwie zachodniopomorskim przekroju sekcji

Źródło: Obliczenia własne.

Oceny precyzji estymacji dokonano również na podstawie parametru *RedCV* charakteryzującego stopień redukcji współczynników zmienności estymatora GREG oraz wyróżnionych w badaniu rodzajów lokalnego estymatora regresyjnego w porównaniu z estymacją bezpośrednią (por. tab. 3). Na podstawie danych zawartych w tabeli można stwierdzić, że największa redukcja nastąpiła w wyniku zastosowania lokalnego estymatora regresyjnego $t_{loc}(20)$ *(średnia –0,139, mediana –0,080)* oraz $t_{loc}(40)$ *(średnia –0,137, mediana –0,150).* Ze zmniejszeniem zmienności mamy do czynienia także w przypadku pozostałych estymatorów. W najmniejszym stopniu zmienność estymatora spadła w wyniku zastosowania lokalnego estymatora regresyjnego $t_{loc}(\max \min)$ *(średnia –0,065, mediana –0,095)*.

Tabela 3. Redukcja wartości współczynników zmienności w porównaniu z estymacją bezpośrednią (*RedCV*)

Estymator	**GREG**	$t_{loc}(10)$	$t_{loc}(20)$	$t_{loc}(40)$	$t_{loc}(\max \min)$
min	–0,337	–0,374	–0,412	–0,448	–0,348
max	0,180	0,093	0,004	0,003	0,336
średnia	–0,083	–0,110	–0,139	–0,137	–0,065
mediana	–0,069	–0,098	–0,080	–0,150	–0,095

Źródło: Obliczenia własne.

V. WNIOSKI

Przeprowadzone badanie dotyczące lokalnej estymacji regresyjnej pozwala na sformułowanie następujących wniosków:

- Oceny parametrów otrzymane na podstawie lokalnych estymatorów regresyjnych wraz ze wzrostem szerokości pasma, coraz bardziej stają się podobne do ocen otrzymanych na podstawie modelu wyznaczonego dla estymatora typu GREG.
- Najbardziej precyzyjne, biorąc pod uwagę wartości współczynnika zmienności estymatora, okazały się lokalne estymatory regresyjne ze zmienną szerokością pasma: $\hat{Y}_{loc}(10)$, $\hat{Y}_{loc}(20)$ i $\hat{Y}_{loc}(40)$.
- Lokalne estymatory regresyjne, w których szerokość pasma jest zmienna ($\hat{Y}_{loc}(10), \hat{Y}_{loc}(20), \hat{Y}_{loc}(40)$) charakteryzują się mniejszą dyspersją w porównaniu do estymatora bezpośredniego.
- W przypadku wąskich pasm szacunki oparte są na wielu modelach lokalnych, co znacznie wydłuża proces przetwarzania danych.

- Wraz ze zmniejszaniem się szerokości pasma, w coraz większym stopniu uwzględniane są lokalne zmiany wartości zmiennej badanej. Poszerzanie pasma wpływa na zwiększenie efektu wygładzenia.
- Wagi wyznaczone w oparciu o funkcję jądrową nie zależą od wartości zmiennej badanej, tylko od zmiennych pomocniczych. Oznacza to, że mogą być wykorzystane w przypadku wielu zmiennych badanych, jeśli skład zmiennych pomocniczych jest stały.

BIBLIOGRAFIA

Breidt, F.J., Opsomer, J.D. (2000). *Local Polynomial Regression Estimation in Survey Sampling.* The Annals of Statistics, 28, 1026–1053.

Chambers, R. (1996), *Robust case-weighting for multipurpose establishment surveys*, Journal of Official Statistics, 12, s. 3–32.

Chambers, R., Dorfman, A.H., Wehrly, T.E. (1993). *Bias Robust Estimation in Finite Populations Using Nonparametric Calibration.* Journal of the American Statistical Association, 88, s. 268–277.

Chambers R.L, Falvey H., Hedlin D., Kokic P. (2001), *Does the Model Matter for GREG Estimation? A Business Survey Example*, [w:] Journal of Official Statistics, Vol.17, No.4, 527–544.

Deville, J.C., Särndal, C.E. (1992), *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, 87, 376–382.

Dorfman, A.H. (2000), *Non-Parametric Regression for Estimating Totals in Finite Populations.* Proceedings of the Survey Research Methods. American Statistical Association, s. 47–54.

Hedlin D. (2004), *Business Survey Estimation*, R&D, Sweden.

Kim, J.Y., Breidt, F.J. and Opsomer, J.D. (2001), *Local polynomial regression estimation in two-stage sampling.* Proceedings of the Section on Survey Research Methods, American Statistical Association, s. 55–61.

Lehtonen R., Veijanen A., 1998, *On multinomial logistic generalized regression estimators*, Maszynopis Department of statistics, University of Jyväskylä, No. 22, Jyväskylä.

Rousseeuw, P.J., and Leroy, P.M., *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.

Särndal, C.E., Swensson, B. and Wretman, J.H., *Model Assisted Survey Sampling*, Springer–Verlag, 1992.

Grażyna Dehnel

LOCAL ESTIMATION IN SMALL BUSINESS RESEARCH

Abstract

There are many surveys of populations that contain a number of extreme values. This is particularly true in surveys of business enterprises. Outliers observations can have an important effect on work with estimation especially on low level of the aggregation. Although the values are extreme, they need not necessarily be false; extremely large observations are a natural component in survey populations. So we shall explore some alternative technique estimation less sensitive to outliers. In this paper we examine local regression which has ability to accommodate local departures from the underlying linear model in business statistics.