

Grażyna Dehnel

Akademia Ekonomiczna w Poznaniu

EFEKTYWNE WYKORZYSTANIE ZMIENNYCH POMOCNICZYCH W BADANIU MAŁYCH PODMIOTÓW GOSPODARCZYCH

Badania statystyczne dotyczące działalności gospodarczej często sprawiają wiele kłopotów związanych z danymi. Problemy pojawiają się w odniesieniu do zmiennych. Rozkłady podmiotów według analizowanych cech charakteryzują: bardzo silna asymetria prawostronna, ogromne zróżnicowanie i silna koncentracja. Poza tym w przypadku wielu jednostek badana zmienna przybiera wartości zerowe bądź wartości skrajne. Problem zmiennych w statystyce gospodarczej dotyczy nie tylko Polski. Świadczą o tym publikacje, w których prezentowane są wyniki badań dotyczących szacunków parametrów dla podmiotów gospodarczych [1; 2; 3]. Rozwiązanie polegające na usunięciu wartości skrajnych nie zawsze dostarcza wiarygodnych szacunków. Dlatego też nieodzowne w takiej sytuacji wydaje się wprowadzenie do estymacji pewnych modyfikacji opartych np. na modelu Chambersa [2]. W modelu tym zaproponowano nowe podejście do estymacji pośredniej typu GREG – ze szczególnym uwzględnieniem wag przypisywanych poszczególnym obserwacjom.

Klasyczny estymator typu GREG (*Generalised REGression estimator*) jest ulepszoną postacią estymatora bezpośredniego wzbogaconego o dodatkowy element uwzględniający różnice między wartościami zmiennych pomocniczych dla populacji i dla próby w danym małym obszarze (domenie), [5, s. 17-18]. Standardowo zmienne pomocnicze są odpowiednio przekształcane i dostosowywane do wartości zmiennej badanej za pomocą modelu regresji. Estymator wartości globalnej typu GREG zmiennej badanej y w d -tej domenie można zapisać jako:

$$\hat{Y}_{d(\text{GREG})} = \hat{Y}_d + (\mathbf{X}_d - \hat{\mathbf{X}}_d)^T \hat{\boldsymbol{\beta}}_d, \quad (1)$$

gdzie $\hat{Y}_d$ oznacza bezpośredni estymator wartości globalnej zmiennej badanej y w d -tej domenie, $\hat{\mathbf{X}}_d$ – wektor bezpośrednich estymatorów wartości globalnej zmiennych pomocniczych x w d -tej domenie, $\mathbf{X}_d$ – wektor wartości globalnych

zmiennych pomocniczych x w d -tej domenie, $\hat{\beta}_d$ jest szacowany w oparciu o metodę najmniejszych kwadratów ważoną wagami określonymi przez schemat losowania:

$$\hat{\beta}_d = \left(\sum_{k \in s_d} w_k x_k x_k^T \right)^{-1} \sum_{k \in s_d} w_k x_k y_k, \quad (2)$$

gdzie s_d oznacza próbę w domenie d .

Jednym z założeń modelu regresji liniowej jest to, że składniki losowe mają jednakową wariancję. Jeżeli wariancja składnika losowego zwiększa się lub zmniejsza, gdy zmienia się któraś ze zmiennych objaśniających, to występuje zjawisko zwane heteroscedastycznością. W przypadku heteroscedastyczności oceny parametrów regresji są nieefektywne. Naruszenie założeń liniowego modelu regresji można czasem usunąć, dokonując pewnej transformacji.

W modelu Chambersa, który odnosi się do estymatorów typu GREG, zaproponowano transformację polegającą na włączeniu dodatkowej zmiennej pomocniczej z_k^Y . Estymator wartości globalnej zmiennej Y przybiera postać:

$$\hat{Y}_{\text{GREG}} = \sum_{k \in P} \hat{y}_k + \sum_{k \in s} w_k e_k, \quad (3)$$

gdzie: $\hat{y}_k = \mathbf{x}'_k \hat{\beta}$, e_k – reszty, w_k – wagi, s – próba, P – populacja generalna

$$\hat{\beta} = \left(\sum_{k \in s} w_k \mathbf{x}_k \mathbf{x}'_k / z_k^Y \right)^{-1} \left(\sum_{k \in s} w_k \mathbf{x}_k y_k / z_k^Y \right), \quad (4)$$

z_k^Y – zmienna pomocnicza określająca heteroscedastyczność w regresji y względem x , γ – parametr, określający poziom heteroscedastyczności.

Wraz ze wzrostem wartości parametru γ zwiększa się rozproszenie punktów na diagramie rozrzutu. Silniejsza jest tendencja do „odskakiwania” – rozproszenia punktów od linii wyznaczonej przez funkcję. Z praktyki badań statystycznych wiadomo, że w przypadku wielu populacji wzrost wartości zmiennej objaśniającej powoduje zwiększenie wariancji składnika losowego, co świadczy o tym, że parametr γ przybiera wartość większą od zera. Najczęściej jest to wartość z przedziału $1 \leq \gamma \leq 2$ [6, s. 255].

Wzór (3) można przedstawić również w postaci:

$$\hat{Y}_{\text{GREG}} = \sum_{k \in s} w_k g_k y_k, \quad (5)$$

gdzie g_k jest wagą k -tej jednostki zdefiniowaną jako:

$$g_k = 1 + (X - \hat{X}_{HT}) \left(\sum_{k \in s} w_k \mathbf{x}_k \mathbf{x}_k' / z_k^\gamma \right)^{-1} (\mathbf{x}_k / z_k^\gamma), \quad (6)$$

gdzie $\hat{X}_{HT}$ oznacza estymator bezpośredni Horvitz-Thompsona wartości globalnej wyznaczony dla zmiennej pomocniczej x , X – wartość globalną zmiennej pomocniczej x .

Wagi wyznaczone dla estymatora typu GREG pozostają w silnym związku z miarą odległości DFBETA, która pozwala na zidentyfikowanie wartości skrajnych. Wartości DFBETA dla k -tej jednostki można wyznaczyć na podstawie wzoru:

$$\text{DFBETA}_k = \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k' / z_k^\gamma \right)^{-1} \left(\frac{\mathbf{x}_k}{z_k^\gamma} \right) \left(\frac{e_k}{1 - h_k} \right), \quad (7)$$

gdzie: $h_k = (\mathbf{x}_k / z_k^\gamma) \left(\sum_{k \in n} \mathbf{x}_k \mathbf{x}_k' / z_k^\gamma \right)^{-1} (\mathbf{x}_k / z_k^\gamma)$.

Przeprowadzono badanie, w którym poddano analizie następujące estymatory:

1. $\hat{Y}_{\text{exp}} = \sum_{i \in s} w_k y_k$ – estymator bezpośredni.
2. $\hat{Y}_{\text{GREG}}^0 - \gamma = 0 \Rightarrow z_k^0$ – estymator regresyjny oparty na liniowym modelu regresji zakładającym homoscedastyczność.
3. $\hat{Y}_{\text{GREG}}^1 - \gamma = 1 \Rightarrow z_k^1$
4. $\hat{Y}_{\text{GREG}}^{1,5} - \gamma = 1,5 \Rightarrow z_k^{1,5}$
5. $\hat{Y}_{\text{GREG}}^2 - \gamma = 2 \Rightarrow z_k^2$

W punktach 3-5 przedstawiono estymatory regresyjne oparte na liniowym modelu regresji zakładającym heteroscedastyczność.

W ocenie jakości estymacji pośredniej wykorzystano względny błąd szacunku

$$\text{REE}(\hat{Y}) = \frac{\sqrt{\text{MSE}(\hat{Y})}}{\hat{Y}}, \quad (8)$$

gdzie:

$$\text{MSE}(\hat{Y}) = \frac{\sum_{i=1}^{300} (Y_i - \hat{Y}_i)^2}{300} \quad i = 1 \dots 300 - \text{liczba przeprowadzonych symulacji}. \quad (9)$$

Wyznaczono również wartość zaproponowanego przez L. Kisha współczynnika deff jako względnej miary precyzji [4]. Jest on zdefiniowany następującym wzorem:

$$\text{deff}(\hat{Y}_{\text{GREG}}) = \frac{\text{MSE}(\hat{Y})}{\text{Var}(\hat{Y}_{\text{exp}})} \quad (10)$$

Współczynnik ten pozwala na porównanie precyzji zastosowanej metody estymacji z efektywnością estymatora bezpośredniego. Wartość współczynnika deff mniejsza od jedności oznacza, że zastosowana estymacja typu GREG jest bardziej precyzyjna od estymacji bezpośredniej.

W przeprowadzonej analizie wykorzystano wyniki badania reprezentacyjnego SP3 dla 2001 r. Badaniem tym objęta jest 4-procentowa próba podmiotów gospodarczych o liczbie pracujących do 9 osób [7]. Zakres ograniczono do województwa wielkopolskiego.

Wyniki badania potraktowano jako populację generalną. Wylosowano 10-procentową próbę. W estymacji wykorzystano informacje o czterech zmiennych:

- 1) przeciętnych wynagrodzeniach,
- 2) przeciętnej liczbie zatrudnionych,
- 3) przeciętnych przychodach,
- 4) przeciętnych kosztach.

Zmienną badaną y były wynagrodzenia, zaś zmienną pomocniczą x zamiennie przychód (p) lub przeciętna liczba zatrudnionych (plz). Za zmienną z_k^y przyjęto, w zależności od podejścia, przychód (p) lub koszt (k) lub przeciętną liczbę zatrudnionych (plz).

Podobnie jak to czyni GUS w badaniu SP3, wydzielono dwie warstwy:

- 1) podmioty zatrudniające 0-5 osób,
- 2) podmioty zatrudniające ponad 6 osób.

Domeną był rodzaj prowadzonej działalności gospodarczej (według obowiązującego podziału na sekcje zgodnie z Polską Klasyfikacją Działalności – PKD). Wyniki badania zostaną przedstawione dla 4 sekcji: przemysłu, budownictwa, handlu i hoteli. Liczebności poszczególnych domen, w podziale na populację generalną i próbę, z uwzględnieniem warstw przedstawiono w tab. 1.

Tabela 1. Liczba jednostek w populacji generalnej (PG) i w próbie, biorących udział w badaniu w podziale na warstwy i domeny

Domena	Warstwa 1		Warstwa 2	
	PG	Próba	PG	Próba
Przemysł	3 834	208	391	98
Budownictwo	4 043	193	339	85
Handel	12 271	967	544	136
Hotele	737	42	49	13
Suma	20 885	1 410	1 323	332

Źródło: obliczenia własne na podstawie wyników badania SP3.

Tabela 2. Precyzja estymacji przeciętnego wynagrodzenia na 1 zakład na podstawie badania SP3 z 2001 roku w województwie wielkopolskim

Domena	Warianty	x	z	REE					deff			
				$\hat{Y}_{exp}$	$\hat{Y}_{GREG}^0$	$\hat{Y}_{GREG}^1$	$\hat{Y}_{GREG}^{1,5}$	$\hat{Y}_{GREG}^2$	$\hat{Y}_{GREG}^0$	$\hat{Y}_{GREG}^1$	$\hat{Y}_{GREG}^{1,5}$	$\hat{Y}_{GREG}^2$
BUDOWNICTWO	1	p	p	<u>0,049</u>	<u>0,066</u>	<u>0,119</u>	<u>0,145</u>	<u>0,129</u>	<u>1,80</u>	<u>5,83</u>	<u>8,62</u>	<u>6,85</u>
		p	k	0,055	0,054	0,114	0,122	0,068	0,96	4,40	4,99	1,53
		plz	p	<u>0,054</u>	<u>0,024</u>	<u>0,024</u>	<u>0,025</u>	<u>0,025</u>	<u>0,20</u>	<u>0,20</u>	<u>0,21</u>	<u>0,21</u>
	2	p	p	0,051	0,048	0,052	0,056	0,063	0,90	1,04	1,22	1,51
		p	k	0,055	0,052	0,055	0,058	0,063	0,89	0,98	1,10	1,30
		plz	k	0,061	0,050	0,050	0,050	0,050	0,67	0,67	0,67	0,67
	1,2	p	p	<u>0,035</u>	<u>0,039</u>	<u>0,060</u>	<u>0,070</u>	<u>0,067</u>	<u>1,25</u>	<u>2,87</u>	<u>4,02</u>	<u>3,61</u>
		p	k	0,042	0,040	0,061	0,065	0,049	0,89	2,09	2,36	1,34
		plz	p	0,042	0,031	0,031	0,031	0,031	0,53	0,53	0,53	0,53
PRZEMYSŁ	1	p	p	<u>0,051</u>	<u>0,051</u>	<u>0,061</u>	<u>0,077</u>	<u>0,101</u>	<u>1,00</u>	<u>1,41</u>	<u>2,25</u>	<u>3,87</u>
		p	k	0,048	0,046	0,046	0,046	0,046	0,92	0,90	0,90	0,92
		plz	p	<u>0,057</u>	<u>0,021</u>	<u>0,027</u>	<u>0,029</u>	<u>0,030</u>	<u>0,14</u>	<u>0,21</u>	<u>0,25</u>	<u>0,27</u>
	2	p	p	0,020	0,039	0,051	0,060	0,073	<u>3,95</u>	<u>6,81</u>	<u>9,34</u>	<u>13,70</u>
		p	k	0,020	0,040	0,051	0,057	0,067	<u>4,17</u>	<u>6,59</u>	<u>8,38</u>	<u>11,34</u>
		plz	k	<u>0,022</u>	<u>0,014</u>	<u>0,016</u>	<u>0,017</u>	<u>0,018</u>	<u>0,43</u>	<u>0,53</u>	<u>0,60</u>	<u>0,67</u>
	1,2	p	p	0,019	0,031	0,040	0,047	0,057	2,61	4,21	5,80	8,64
		p	k	0,018	0,033	0,040	0,044	0,051	3,18	4,66	5,76	7,60
		plz	p	<u>0,023</u>	<u>0,012</u>	<u>0,013</u>	<u>0,014</u>	<u>0,015</u>	0,28	0,35	0,40	0,44
HANDL	1	p	p	0,103	0,026	0,065	0,132	0,114	0,07	0,40	1,66	1,24
		p	k	0,103	<u>0,023</u>	<u>0,072</u>	0,143	0,041	0,05	0,49	1,95	0,16
		plz	p	<u>0,095</u>	<u>0,086</u>	<u>0,088</u>	<u>0,089</u>	<u>0,089</u>	0,82	0,87	0,88	0,88
	2	p	p	0,065	<u>0,059</u>	0,076	0,109	0,169	0,83	1,39	2,86	6,80
		p	k	0,062	0,063	0,079	0,105	0,154	1,02	1,60	2,85	6,05
		plz	k	<u>0,068</u>	<u>0,074</u>	<u>0,071</u>	<u>0,071</u>	<u>0,071</u>	<u>1,20</u>	<u>1,11</u>	<u>1,10</u>	<u>1,10</u>
	1,2	p	p	0,062	0,032	0,049	0,085	0,098	0,28	0,65	1,89	2,53
		p	k	0,062	0,032	0,051	0,087	0,075	0,26	0,68	1,95	1,46
		plz	p	<u>0,058</u>	<u>0,056</u>	<u>0,056</u>	<u>0,056</u>	<u>0,056</u>	0,96	0,95	0,95	0,95
HOTEL	1	p	p	<u>0,130</u>	<u>0,078</u>	<u>0,079</u>	<u>0,089</u>	<u>0,108</u>	0,37	0,38	0,47	0,70
		p	k	<u>0,123</u>	<u>0,073</u>	<u>0,075</u>	<u>0,085</u>	<u>0,104</u>	0,35	0,37	0,48	0,72
		plz	p	0,123	0,020	0,019	0,019	0,020	0,03	0,02	0,03	0,03
	2	p	p	0,246	0,321	0,324	0,359	0,421	1,71	1,74	2,14	2,94
		p	k	0,242	0,331	0,334	0,373	0,432	1,87	1,91	2,38	3,19
		plz	k	<u>0,240</u>	<u>0,198</u>	<u>0,204</u>	<u>0,205</u>	<u>0,205</u>	<u>0,68</u>	<u>0,72</u>	<u>0,73</u>	<u>0,73</u>
	1,2	p	p	0,161	0,196	0,197	0,217	0,254	1,483	1,5	1,824	2,4973
		p	k	0,155	0,200	0,202	0,225	0,261	1,66	1,69	2,10	2,83
		plz	p	<u>0,155</u>	<u>0,118</u>	<u>0,121</u>	<u>0,122</u>	<u>0,122</u>	<u>0,58</u>	<u>0,61</u>	<u>0,62</u>	<u>0,62</u>

Oznaczenia: w – wynagrodzenia, p – przychód, k – koszt, plz – przeciętna liczba zatrudnionych.

Źródło: obliczenia własne na podstawie wyników badania SP3.

Aby umożliwić ocenę precyzji uzyskanych wyników, przeprowadzono 300 symulacji. W ramach każdej z nich dla wyróżnionych 4 sekcji PKD wyznaczono

wartości 5 wyżej opisanych estymatorów. W tab. 2 zamieszczono wartości względnego błędu szacunku (REE) i współczynnika deff dla wszystkich analizowanych estymatorów. Pogrubioną czcionką zaznaczono przypadki, w których zastosowanie estymacji typu GREG poprawiło precyzję. Natomiast podkreślone wartości odnoszą się do sytuacji, w której estymacja bezpośrednia charakteryzowała się większą precyzją niż estymacja pośrednia.

Otrzymane wyniki wskazują na duże zróżnicowanie precyzji estymacji. Dla przykładu w budownictwie w warstwie (1), czyli w podmiotach zatrudniających do 5 osób, zastosowanie zmiennych pomocniczych: x – przeciętna liczba zatrudnionych, z – przychód dało wyraźną poprawę precyzji. Zysk na precyzji po zastosowaniu estymacji typu GREG był pięciokrotny. Z kolei w przemyśle w warstwie (2), którą stanowiły podmioty zatrudniające więcej niż 5 osób, dla zmiennych pomocniczych: x – przychód, z – przychód, estymacja bezpośrednia okazała się najlepsza (por. tab. 2). Analizując wyniki, można zauważyć, że zysk na precyzji zależy od siły związku między zmienną badaną i zmienną pomocniczą (por. tab. 3).

Tabela 3. Współczynniki korelacji liniowej między wartościami zmiennych szacowanych i pomocniczych na podstawie badania SP3 z 2001 roku

Warstwa	Zmienna	Budownictwo			Warstwa	Zmienna	Przemysł		
		<i>plz</i>	<i>p</i>	<i>k</i>			<i>plz</i>	<i>p</i>	<i>k</i>
1	<i>w</i>	0,893	0,261	0,302	1	<i>w</i>	0,9	<u>0,281</u>	<u>0,255</u>
	<i>p</i>	0,237		0,928		<i>p</i>	0,171		0,92
	<i>k</i>	0,256				<i>k</i>	0,145		
2	<i>w</i>	0,56	0,465	0,492	2	<i>w</i>	0,86	<u>0,43</u>	<u>0,292</u>
	<i>p</i>	0,01		0,987		<i>p</i>	0,45		
	<i>k</i>	0,064				<i>k</i>	0,216	0,957	
1,2	<i>w</i>	0,881	0,435	0,571	1,2	<i>w</i>	0,981	0,645	0,575
	<i>p</i>	0,401				<i>p</i>	0,635		0,951
	<i>k</i>	0,516	0,928			<i>k</i>	0,547		
Warstwa	Zmienna	Handel			Warstwa	Zmienna	Hotele		
		<i>plz</i>	<i>p</i>	<i>k</i>			<i>plz</i>	<i>p</i>	<i>k</i>
1	<i>w</i>	0,478	0,968	0,968	1	<i>w</i>	0,991	0,837	0,831
	<i>p</i>	0,319				<i>p</i>	0,878	0,999	
	<i>k</i>	0,313		0,999		<i>k</i>	0,872		
2	<i>w</i>	-0,335	0,474	0,499	2	<i>w</i>	0,603	0,963	0,956
	<i>p</i>	-0,394				<i>p</i>	0,364		
	<i>k</i>	-0,409	0,997			<i>k</i>	0,343		0,999
1,2	<i>w</i>	0,542	0,779	0,777	1,2	<i>w</i>	0,836	0,931	0,919
	<i>p</i>	0,185				<i>p</i>	0,59		
	<i>k</i>	0,181		0,999		<i>k</i>	0,563		0,999

Oznaczenia: *w* – wynagrodzenia, *p* – przychód, *k* – koszt, *plz* – przeciętna liczba zatrudnionych

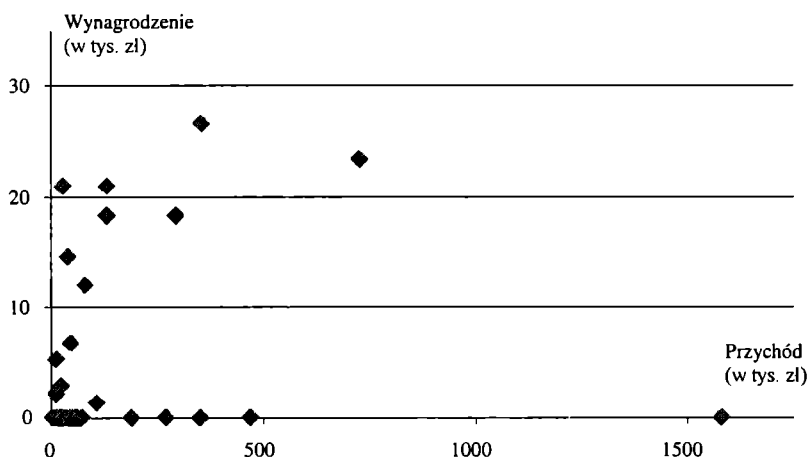
Źródło: obliczenia własne na podstawie wyników badania SP3.

Im silniejsza jest zależność, tym zysk na precyzji jest większy. Kolejną prawidłowością jest to, że precyzja szacunku estymacji typu GREG zależy od liczebności próby. Jest ona z reguły niższa w przypadku mniej licznej warstwy (2), co najlepiej widać w przemyśle ($\text{deff}(\hat{Y}_{reg}^2) = 13,7$). Oba przedstawione wnioski zdają się potwierdzeniem teorii znanej z metody reprezentacyjnej, zgodnie z którą w przypadku zastosowania estymacji typu GREG zysk na precyzji jest większy, jeśli większa jest liczebność próby oraz jeśli korelacja między zmienną badaną a zmienną pomocniczą jest silna.

Nie można więc jednoznacznie wskazać, który z zaprezentowanych estymatorów charakteryzuje się największą precyzją. W każdym przypadku wymagany jest dobór odpowiedniego modelu. Dla domen, dla których żaden estymator nie zapewnia odpowiedniej precyzji, proponuje się np. poststratyfikację. Polega ona na wydzieleniu w ramach warstwy 3 kolejnych warstw i zastosowaniu „dobrego” modelu regresyjnego tylko dla tej środkowej, na którą obserwacje odstające nie wpływają. Warstwa ta powinna być tak duża, jak to jest tylko możliwe (maksymalizując korzyść wynikającą z estymacji regresyjnej). W pozostałych warstwach należałoby zastosować estymację bezpośrednią, na którą obserwacje odstające nie mają wpływu. Warstwy te powinny być na tyle duże, aby wartość wariancji nie była zbyt duża.

Aby bliżej przeanalizować rozpatrywane estymatory należy zapoznać się ze strukturą domeny. Jako przykład wybrano sekcję przemysł dla zmiennych pomocniczych: x – przychód, z – przychód.

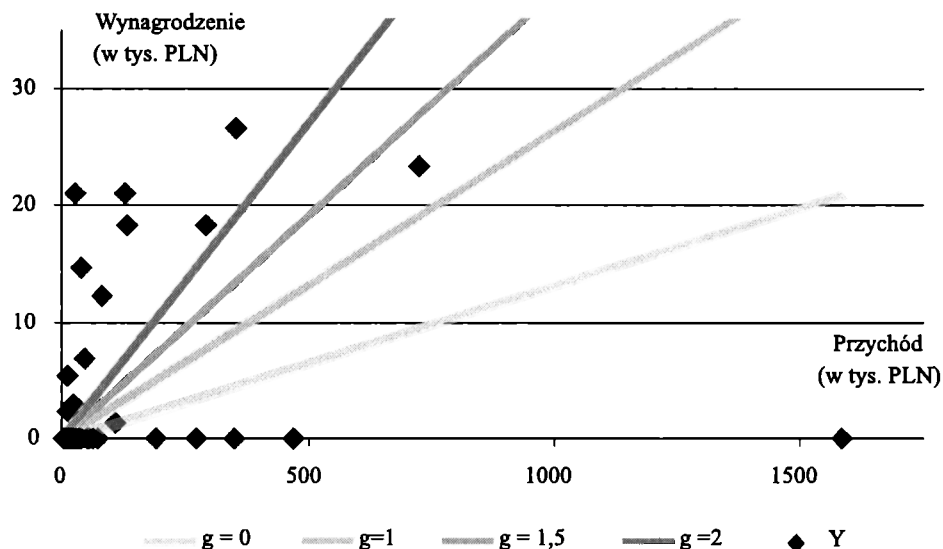
Na diagramie rozrzutu (por. rys. 1) widoczne są obserwacje odstające zarówno dla zmiennej y (duże wynagrodzenie) jak i dla zmiennej x (duży przychód).



Rys. 1. Diagram rozrzutu przedstawiający zależność pomiędzy przeciętnym wynagrodzeniem na 1 firmę a przeciętnym przychodem na 1 firmę na podstawie badania SP3

Źródło: obliczenia własne na podstawie wyników badania SP3.

Po naniesieniu na diagram rozrzutu linii regresji wyznaczonych na podstawie analizowanych estymatorów można dostrzec pewną prawidłowość (por. rys. 2).



Rys. 2. Linie regresji wyznaczone na podstawie badanych estymatorów $\hat{Y}_{\text{GREG}}^{\gamma}$

Źródło: obliczenia własne na podstawie wyników badania SP3.

Regresja liniowa $\hat{Y}_{\text{GREG}}^0$, dla której $\gamma = 0$ w największym stopniu uwzględnia wartości odstające w przestrzeni x . Wraz ze wzrostem parametru γ w coraz większym stopniu są brane pod uwagę wartości odstające dla zmiennej y . Takie położenie linii regresji można było zaobserwować w prawie wszystkich rozpatrywanych przypadkach.

Wyjątkiem jest sekcja handel w warstwie (1), czyli wśród podmiotów, w których liczba zatrudnionych nie przekracza 5 osób. Kolejność linii regresji była tu odwrotna do tej, którą zaprezentowano na rys. 2. W wyniku szczegółowej analizy danych ustalono, że w sekcji handel na 961 jednostek wylosowanych do próby wartość wynagrodzeń (zmiennej badanej) była równa zero w przypadku 765 jednostek. Zera pojawiały się również wówczas, gdy zmienna pomocnicza (przychód) przybierała dużą wartość.

Przyczyny tak zróżnicowanych oszacowań należy dopatrywać się w wartościach wag nadawanych poszczególnym jednostkom. Rozkład wartości wag dla 3 sekcji przemysłu, budownictwa i handlu ukazano w tab. 4, 5, 6.

W przypadku modeli, dla których $\gamma = 0$ oraz $\gamma = 1$, wartości odstające dla zmiennej x otrzymały jedne z najwyższych wag spośród wszystkich modeli. W modelach, w których $\gamma = 1,5$ oraz $\gamma = 2$, wartościom odstającym (x) przyporządkowano najmniejsze wartości wag. Zgodnie z wyżej przedstawionymi wnios-

Tabela 4. Rozkład wartości wag dla 4 estymatorów typu GREG w sekcji przemysł, dla zmiennej x – przychód oraz zmiennej z – koszt

Waga g_k^γ	Min	Max	Waga dla obserwacji odstającej (x)	Mediana
g_k^0	0,8613	1,0000	0,9999	0,9994
g_k^1	0,7189	0,9831	0,8239	0,9706
$g_k^{1,5}$	-1,5694	0,9940	-1,5694	0,9261
g_k^2	-12,1605	0,9995	-12,1605	0,9407

Źródło: obliczenia własne na podstawie wyników badania SP3.

Tabela 5. Rozkład wartości wag dla czterech estymatorów typu GREG w sekcji budownictwo, dla zmiennej x – przychód oraz zmiennej z – przychód

Waga g_k^γ	Min	Max	Waga dla obserwacji odstającej (x)	Mediana
g_k^0	1,0016	19,1115	19,1115	1,3373
g_k^1	5,0168	5,0168	5,0168	5,0168
$g_k^{1,5}$	2,5555	35,0902	2,5555	10,7400
g_k^2	1,3753	181,2438	1,3753	15,7135

Źródło: obliczenia własne na podstawie wyników badania SP3.

Tabela 6. Rozkład wartości wag dla 4 estymatorów typu GREG w sekcji handel, dla zmiennej x – przychód oraz zmiennej z – przychód

Waga g_k^γ	Min	Max	Waga dla obserwacji odstającej (x)	Mediana
g_k^0	1,0001	49,3728	47,6579	1,1364
g_k^1	8,0300	8,0300	8,0300	8,0300
$g_k^{1,5}$	2,9976	126,4889	2,9976	28,5191
g_k^2	1,2712	1071,4305	1,2712	52,4775

Źródło: obliczenia własne na podstawie wyników badania SP3.

kami modele wraz ze wzrostem wartości γ przenoszą ciężar estymacji z obserwacji skrajnych (x) na obserwację skrajną (y). Parametr γ wpływa również w znacznym stopniu na zakres zmienności wag. Im większa jego wartość, tym większa zmienność wag.

Wnioski

1. Stosowanie estymatorów typu GREG prowadzi do bardzo zróżnicowanych wyników.
2. Dobór odpowiedniego modelu wpływa na poprawę precyzji estymacji.
3. Zastosowanie dodatkowej zmiennej (z) pozwala na przeniesienie problemu estymacji dla wartości skrajnych zmiennej pomocniczej na wartości skrajne zmiennej badanej.
4. Wagi wyznaczone dla estymatora typu GREG pozostają w silnym związku z miarą odległości DFBETA, która pozwala na zidentyfikowanie wartości skrajnych i wyeliminowanie ich lub zastąpienie wartością mediany.
5. W przypadku domen, dla których trudny jest dobór odpowiedniego modelu, proponuje się poststratyfikację.
6. Wykorzystanie zmiennej dodatkowej (z) stwarza możliwość dokonania szacunku podstawowych wskaźników podmiotów gospodarczych na niskim poziomie agregacji (tj. w przekroju, dla którego zastosowanie metod klasycznych jest niemożliwe).

Literatura

- [1] Chambers R.L., *Robust Case-weighting for Multipurpose Establishment Surveys*, „Journal of Official Statistics” 1996, Vol. 12, No. 1, 3-32.
- [2] Chambers R.L., Falvey H., Hedlin D., Kokic P., *Does the Model Matter for GREG Estimation? A Business Survey Example*, „Journal of Official Statistics” 2001, Vol. 17, No. 4, 527-544.
- [3] Karlberg F., *Survey Estimation for Highly Skewed Populations in the Presence of Zeroes*, „Journal of Official Statistics” 2000, Vol. 16, No. 1, 229-241.
- [4] Kish L., *Discussion*, [in:] *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Särndal, M.P. Singh, (ed.), John Wiley & Sons, New York 1987.
- [5] Rao J.N.K., *Small Area Estimation*, John Wiley & Sons, Hoboken, New Jersey, 2003.
- [6] Särndal C.E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.
- [7] Zagożdżińska I., Kotowski J., Borowska J., Starzyńska M., *Działalność gospodarcza przedsiębiorstw o liczbie pracujących do 9 osób w 2001 r.*, Główny Urząd Statystyczny, Warszawa 2003.

EFFECTIVE USE OF AUXILIARY VARIABLE IN SMALL BUSINESS CHARACTERISTICS ESTIMATION

Summary

According to the Central Statistical Office, there were about 1.6 million small enterprises (called *small business*) in Poland in 2001 – over 95% of all enterprises. Although the small economic entities constitute the basis of local economy, there is no information about this group of enterprises available

at regional and local level. One of the major problems involved in estimating information about economic activity across domains is the distribution of small companies variables. The study variable may be highly skewed, there may be a large proportion of zero responses and there may be several auxiliary variables that can be used to improve estimation but these may include some extreme values. This article presents Chamber's Model which examines the behaviour of GREG estimators taking into account data classification. It shows how an efficient Chambers Model was found for a business survey.