Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

1 (949)

ISSN 1898-6447
Zesz. Nauk. UEK, 2016; 1 (949): 5–21
DOI: 10.15678/ZNUEK.2016.0949.0101

Grażyna Dehnel
Elżbieta Gołata
Katedra Statystyki
Uniwersytet Ekonomiczny w Poznaniu

M-estymacja w badaniu małych przedsiębiorstw*

Streszczenie

W wielu badaniach z zakresu statystyki gospodarczej liczebność próby jest na tyle duża, że obserwacje odstające mają stosunkowo niewielki wpływ na wartości szacowanych parametrów. W badaniach prowadzonych na niskim poziomie agregacji w ramach statystyki krótkookresowej obecność obserwacji odstających może być jednak znacząca. Z tego powodu w przypadku populacji takich jak populacja przedsiębiorstw obok podejścia klasycznego w badaniach powinien być uwzględniany nurt metod odpornych na występowanie jednostek nietypowych. W literaturze przedmiotu zaproponowano wiele alternatywnych metod estymacji mniej wrażliwych na wartości odstające. W opracowaniu weryfikacji empirycznej poddano jedną z nich – *M*-estymację. Celem analizy była ocena jej użyteczności w odniesieniu do badania małych przedsiębiorstw.

Słowa kluczowe: regresja odporna, *M*-estymacja, statystyka przedsiębiorstw, obserwacje odstające.

1. Wprowadzenie

Założenia dotyczące rozkładów jednostek według badanych cech, które muszą być spełnione w przypadku stosowania metody najmniejszych kwadratów, zostały precyzyjnie określone. Ich niezachowanie prowadzi do obciążenia szaco-

* Artykuł powstał w ramach realizacji projektu sfinansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2015/17/B/HS4/00905.

wanego modelu. Problemy dotyczące niespełnienia wymaganych założeń można rozwiązać, dokonując transformacji zmiennych. Transformacja nie jest jednak wystarczającym rozwiązaniem w przypadku obecności obserwacji nietypowych. Nie zawsze bowiem prowadzi do wyeliminowania, czy chociaż złagodzenia ich oddziaływania obciążającego szacunek parametrów modelu. W takiej sytuacji regresja odporna, która może wykluczyć wpływ obserwacji odstających, stanowi propozycję godną rozważenia. W praktyce badań statystycznych dość często napotykamy populacje, które charakteryzują się obecnością obserwacji nietypowych. Zaproponowanych zostało wiele metod badawczych, które mają na celu złagodzenie wpływu jednostek odstających.

W literaturze przedmiotu metody te podzielono na trzy grupy [Cox i in. 1995]:

1) zmieniające wartości obserwacji odstających (*winsorization*, *trimming*),
2) redukujące wagi obserwacji odstających,
3) techniki estymacji (regresji) odpornej.

Metody zaliczane do pierwszej grupy należą do najprostszych, rzadziej dziś stosowanych. Metody wchodzące w skład drugiej grupy wykorzystywane są przede wszystkich w badaniach próbkowych. Pozwalają one na modyfikację wag wynikających z przyjętego schematu losowania próby. Metody zaliczane do trzeciej grupy, czyli techniki estymacji odpornej, w ostatnich latach coraz bardziej zyskują na znaczeniu. Wśród nich wyróżnić można *M*-estymację, *S*-estymację czy *MM*-estymację. Wiele z nich zostało opracowanych już w latach 70. i 80. XX w., z tego względu jednak, że wymagają one podejścia iteracyjnego, stosowanie ich było ograniczone, gdyż wiązało się m.in. z długim czasem obliczeń. Obecnie dostępne są pakiety statystyczne umożliwiające łatwą implementację metod. Wzrost zainteresowania technikami regresji odpornej wynika również z tego, że ich zastosowanie, w przeciwieństwie do innych metod, nie wymaga wcześniejszej detekcji obserwacji odstających.

Prowadząc badanie z wykorzystaniem technik regresji odpornej, stajemy przed koniecznością dokonania wyboru takiej metody, której zastosowanie pozwoli osiągnąć najlepszy kompromis pomiędzy obciążeniem szacunków a ich efektywnością. W wyborze kierować należy się własnościami poszczególnych metod, które zostały sformułowane m.in. w literaturze przedmiotu [Holland i Welsch 1977, Huber 1981, Hampel i in. 2011, Chen i Yin 2002], czy też na podstawie przeprowadzonych wcześniej badań empirycznych. Wybór metody wymusza konieczność podjęcia dalszych decyzji, ponieważ w ramach każdej z wyróżnionych metod stosowanych może być wiele podejść różniących się doborem parametrów lub wykorzystywanych funkcji.

W niniejszym artykule ograniczono się do analizy jednej z najczęściej stosowanych technik regresji odpornej, jaką jest *M*-estymacja. Celem badania było porównanie jakości szacunków otrzymanych w oparciu o pięć *M*-estymatorów,

w których do określenia wartości wag wykorzystano różne funkcje. Oceny estymacji dokonano na podstawie badania empirycznego, w którym wykorzystano dane dotyczące małych przedsiębiorstw działających w ramach sekcji PKD „Transport i gospodarka magazynowa".

2. M-estymacja

M-estymacja reprezentuje grupę regresyjnych odpornych estymatorów tzw. pierwszej generacji. Estymator *M* został wprowadzony przez P.J. Hubera w 1964 r. [Huber 1964] jako odporny na obserwacje nietypowe odpowiednik podejścia reprezentowanego przez metodę najmniejszych kwadratów. Minimalizuje on funkcję straty $\rho(\cdot)$:

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}(\theta)\right), \qquad r_i = y_i - X\theta, \tag{1}$$

gdzie:

ρ – funkcja celu,

s – parametr skali.

Zakładając, że parametr skali s jest znany, oszacowanie estymatora θ_M otrzymujemy przez rozwiązanie układu p równań normalnych ze względu na wektor θ, rozpisanych jako iloczyn zmiennych niezależnych i pochodnych cząstkowych funkcji ρ:

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \sum_{k=1}^{p} x_{ik}\theta_k}{s}\right) x_i = 0, \tag{2}$$

gdzie:

Ψ – pochodna funkcji ρ,

p – liczba zmiennych x.

Zakłada się tutaj, że s jest znane. W celu rozwiązania równania (2) proponowane jest zastosowanie metody iteracyjnej ważonych najmniejszych kwadratów (IRLS) z wagami określonymi wzorem [Trzpiot 2013]:

$$w_i = \frac{\Psi\left(\frac{y_i - \sum_{k=1}^{p} x_{ik}\theta_k}{s}\right)}{\left(\frac{y_i - \sum_{k=1}^{p} x_{ik}\theta_k}{s}\right)}. \tag{3}$$

Szacunku początkowej wartości $\hat{\theta}_0$ dokonujemy na podstawie KMNK. W każdej kolejnej iteracji $t+1$ wykorzystuje się wartości reszt oraz wag otrzymane w iteracji t aż do osiągnięcia zbieżności [Alma 2011].

Wartości wag zależne są od wyboru funkcji Ψ korespondującej z funkcją ρ. Doboru funkcji Ψ dokonujemy zatem m.in. w zależności od tego, jaką wagę chcemy przypisać obserwacjom odstającym. W literaturze przedmiotu proponowanych jest wiele podejść [Fair 1974, Holland i Welsch 1977, Hampel i in. 2011, Chen i Yin 2002, Banaś i Ligas 2014]. W celu przybliżenia ich własności, oceny przydatności oraz wpływu na wyniki szacunku w niniejszym artykule zastosowanych zostało pięć najczęściej wykorzystywanych w badaniach empirycznych funkcji: Andrewsa, Tukeya (*bisquare*), Cauchy'ego, Faira i Hampela (por. tabela 1):

Funkcja Andrewsa

$$W(x,c)=\begin{cases}\dfrac{\sin\left(\dfrac{x}{c}\right)}{\dfrac{x}{c}} & \text{jeżeli} \quad |x|\le \pi c,\\ 0 & \text{w przeciwnym przypadku,}\end{cases} \tag{4}$$

Funkcja Tukeya

$$W(x,c)=\begin{cases}\left(1-\left(\dfrac{x}{c}\right)^2\right)^2 & \text{jeżeli} \quad |x|\le c,\\ 0 & \text{w przeciwnym przypadku,}\end{cases} \tag{5}$$

Funkcja Cauchy'ego

$$W(x,c)=\frac{1}{1+\left(\dfrac{|x|}{c}\right)^2}, \tag{6}$$

Funkcja Faira

$$W(x,c)=\frac{1}{1+\dfrac{|x|}{c}}, \tag{7}$$

Funkcja Hampela

$$W(x,a,b,c)=\begin{cases}1 & |x|<a,\\ \dfrac{a}{|x|} & a<|x|\le b,\\ \dfrac{a}{|x|}\dfrac{c-|x|}{c-b} & b<|x|\le c,\\ 0 & \text{w przeciwnym przypadku.}\end{cases} \tag{8}$$

Tabela 1. Przyjęte wartości parametrów funkcji

Funkcje wag	Wartości parametrów
Andrewsa	$c = 1{,}339$
Tukeya	$c = 4{,}685$
Cauchy'ego	$c = 2{,}385$
Faira	$c = 1{,}4$
Hampela	$a = 2$, $b = 4$, $c = 8$

Źródło: opracowanie własne na podstawie [*User's Guide*… 2014].

W praktycznych zastosowaniach parametr skali s jest nieznany. Z uwagi na to, że wartość wariancji resztowej pozostaje pod silnym wpływem obserwacji odstających, do szacunku parametru skali wykorzystywane są różne metody. Wśród najczęściej stosowanych wyróżnić można łatwy i odporny estymator parametru skali, jakim jest mediana odchyleń bezwzględnych (MAD) (względem pewnego przyjętego centrum) (por. [Trzpiot 2013]). Można go zastosować dla reszt bliskich zero, dla pozostających w pewnym otoczeniu albo dla reszt z odpornego dopasowania. Iteracyjnie szacuje się:

$$\hat{s}^{(m+1)} = \frac{\underset{i=1}{\overset{n}{med}} \left| y_i - x_i^T \hat{\theta}^{(m)} \right|}{\beta_0}, \tag{9}$$

gdzie: $\beta_0 = \Phi^{-1}(0{,}75)$ jest stałą (por. [Hampel i in. 2011]).

M-estymator jest odporny jedynie na obserwacje odstające w kierunku y, nie jest natomiast odporny na punkty wysokiej dźwigni. Wpływa to na zakres jego zastosowań. Stosowany jest bowiem często, ale w sytuacjach, w których punkty wysokiej dźwigni nie są problemem. Jego punkt załamania nie jest wysoki i wynosi $1/n$.

3. Charakterystyka badania empirycznego

Miary wykorzystane do oceny szacunków otrzymanych w badaniu empirycznym

W ocenie szacunków wykorzystano R^2 – odporną wersję współczynnika determinacji (por. [Renaud i Victoria-Feser 2010]):

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} \tag{10}$$

oraz D – odporną miarę jakości dopasowania modelu[1] (por. [Chen 2003]):

$$D = 2(\hat{s})^2 \sum \rho\left(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}\right), \tag{11}$$

gdzie:

ρ – funkcja celu,
$\hat{\mu}$ – estymator parametru położenia,
$\hat{s}$ – estymator parametru skali.

Założenia badania

W badaniu empirycznym wykorzystano informacje pochodzące z badania przedsiębiorstw prowadzonego w ramach statystyki publicznej, oznaczonego symbolem *DG1*. Badanie to jest największym badaniem w Polsce zaliczanym do krótkookresowej statystyki gospodarczej. Objęte są nim przedsiębiorstwa, w których pracuje więcej niż 9 osób. Badanie dotyczy wszystkich średnich i dużych przedsiębiorstw oraz 10% małych. Prowadzone jest z częstotliwością miesięczną. Dostarcza ono informacji na temat takich zmiennych, jak: przychód, koszt, sprzedaż, transport, ceny, wynagrodzenia, obrót, pracujący, podatki i dotacje. W przeprowadzonym badaniu empirycznym ograniczono się do przedsiębiorstw małych i średnich (liczba pracujących zawiera się w przedziale od 10 do 250), które prowadziły działalność gospodarczą w grudniu 2011 r. Analizie poddano model, w którym jako zmienną zależną przyjęto przychód, zaś zmienną niezależną był koszt. Jako populację generalną przyjęto wszystkie małe i średnie przedsiębiorstwa biorące udział w badaniu *DG1*. Domeną studiów była jednostka powstała w wyniku uwzględnienia podziału na województwa i rodzaj prowadzonej działalności gospodarczej zgodnie z klasyfikacją NACE. W prezentacji wyników badania ograniczono się do 16 domen sekcji „Transport" w przekroju województw. Selekcji domen dokonano na podstawie wartości współczynnika determinacji charakteryzującego dobroć dopasowania modelu. Głównym celem wyboru domen do analizy było uwzględnienie tych jednostek, dla których wartości współczynnika determinacji charakteryzowały się dużą dyspersją. W przypadku sekcji PKD „Transport i gospodarka magazynowa" obszar zmienności zawierał się w granicach od 0,041 do 0,999 (por. tabela 3).

[1] W literaturze angielskojęzycznej określanej mianem: *the robust deviance*.

Pierwszym etapem analizy była ocena rozkładów przedsiębiorstw względem zmiennej „przychód" na podstawie informacji pochodzących z badania *DG1*. Wartości podstawowych charakterystyk, takich jak współczynnik zmienności (65%; 405%), skośność (0,27; 5,63), obszar zmienności (10 357 tys. PLN; 4 092 507 tys. PLN), wskazują na duże zróżnicowanie i bardzo silną asymetrię (por. tabela 2). Własności te przemawiają za użyciem metod uwzględniających obserwacje odstające. Ich obecność potwierdzona została na podstawie dwóch miar – RStudenta i statystyki D-Cooka [Rousseeuw i Leroy 1987]:

RStudent

$$r_i^* = \frac{e_i}{\sqrt{MSE_i} \cdot \sqrt{1-h_i}}, \qquad \left|r_i^*\right| \geq 3, \tag{12}$$

$$h_i = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2}, \tag{13}$$

gdzie:

MSE_i – średni kwadrat odchyleń dla reszty po wyeliminowaniu i-tej obserwacji,

h_i – odległość i-tej obserwacji od średniej wartości zmiennej X,

Statystyka D-Cooka

$$D_i = \frac{e_i^2}{MSE\left(k+1\right)} \cdot \frac{h_i}{\left(1-h_i\right)^2}, \qquad D_i > \frac{4}{n}. \tag{14}$$

Tabela 2. Charakterystyka statystyczna rozkładu zmiennej „przychód" (w tys. PLN) w małych i średnich przedsiębiorstwach stanowiących populację generalną w badaniu w przekroju województw, sekcja PKD „Transport i gospodarka magazynowa", 2011 r.

Województwo	Minimum	Średnia	Mediana	Maximum	*Vx* (%)	Skośność
Dolnośląskie	38	6 796	5 774	24 606	90	1,24
Kujawsko-pomorskie	622	6 935	4 563	25 200	100	1,45
Lubelskie	1 036	17 663	4 456	202 558	231	4,22
Lubuskie	553	56 246	5 938	751 871	303	3,99
Łódzkie	625	8 680	6 072	43 783	106	2,47
Małopolskie	37	21 612	5 649	417 352	327	5,62
Mazowieckie	491	146 475	11 776	4 092 998	405	5,63
Opolskie	1 125	9 173	6 073	25 034	89	0,91
Podkarpackie	758	5 298	5 278	11 115	72	0,27
Podlaskie	2 046	44 010	7 605	443 218	286	3,45
Pomorskie	26	10 459	4 621	64 424	138	2,38

cd. tabeli 2

Województwo	Minimum	Średnia	Mediana	Maximum	*Vx* (%)	Skośność
Śląskie	778	20 700	6 712	384 637	261	5,55
Świętokrzyskie	131	8 187	3 776	49 336	143	2,52
Warmińsko-mazurskie	707	5 795	5 396	13 965	65	0,76
Wielkopolskie	805	9 412	6 863	40 564	98	1,87
Zachodniopomorskie	316	12 876	6 237	76 897	138	2,21

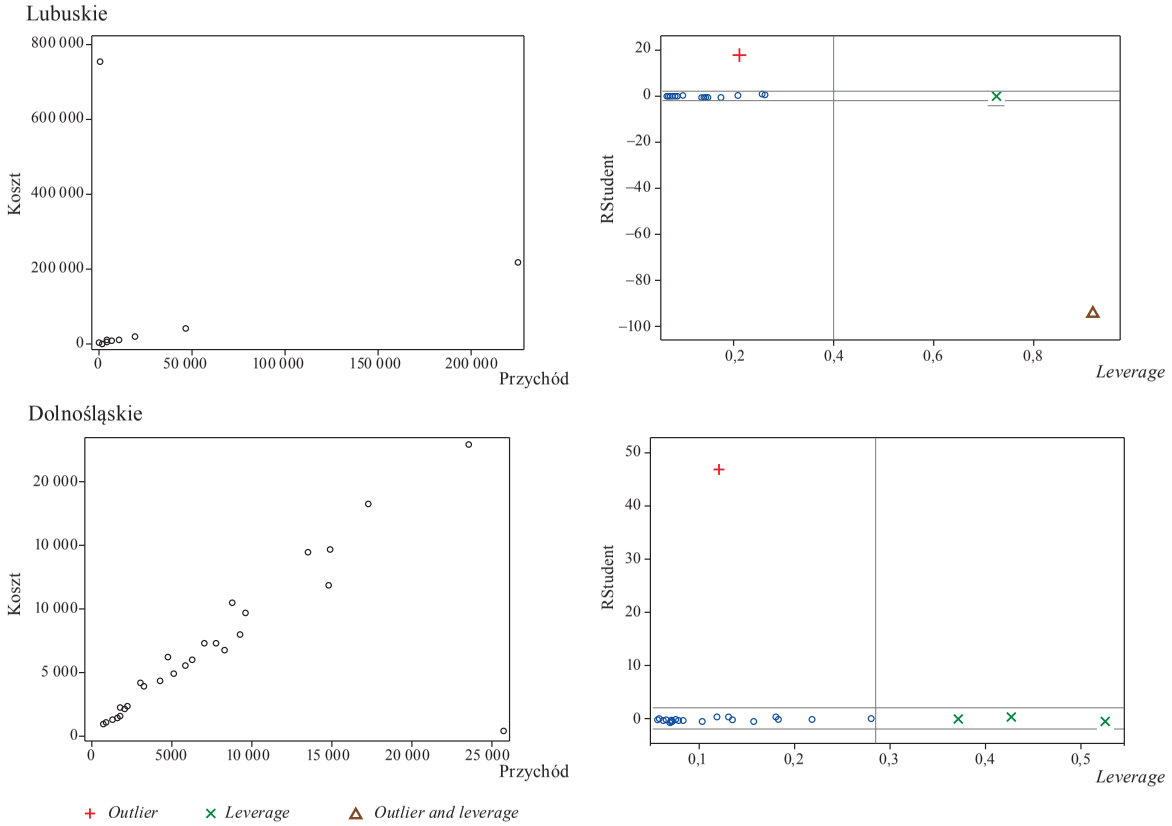
Źródło: opracowanie własne na podstawie wyników badania *DGI*, 2011.

Tabela 3. Charakterystyka statystyczna liczebności małych i średnich przedsiębiorstw stanowiących populację generalną w badaniu w przekroju województw, sekcja PKD „Transport i gospodarka magazynowa", 2011 r.

Województwo	Liczebność populacji	Liczba obserwacji odstających	Odsetek obserwacji odstających	R^2
Dolnośląskie	28	1	3,6	0,498
Kujawsko-pomorskie	24	2	8,3	0,873
Lubelskie	25	2	8,0	0,986
Lubuskie	20	2	10,0	0,041
Łódzkie	30	2	6,7	0,989
Małopolskie	34	4	11,8	0,998
Mazowieckie	73	2	2,7	0,935
Opolskie	14	2	14,3	0,921
Podkarpackie	16	2	12,5	0,977
Podlaskie	12	2	16,7	0,999
Pomorskie	33	3	9,1	0,972
Śląskie	64	4	6,3	0,992
Świętokrzyskie	23	3	13,0	0,996
Warmińsko-mazurskie	13	2	15,4	0,888
Wielkopolskie	49	6	12,2	0,982
Zachodniopomorskie	30	2	6,7	0,910

Źródło: opracowanie własne na podstawie wyników badania *DGI*, 2011.

Liczbę, udział procentowy obserwacji odstających i R^2 zaprezentowano w tabeli 3. Wartości współczynnika determinacji wskazują, że w przypadku liczniejszych sekcji nawet stosunkowo duża liczba obserwacji odstających niekoniecznie musi mieć wpływ na dopasowanie modelu. Z odwrotną sytuacją spotykamy się w sekcjach mniej licznie reprezentowanych, w których pojedyncze



Rys. 1. Obserwacje odstające w kierunku *x* (*leverage*) i *y* (*outlier*) w sekcji „Transport" w województwie dolnośląskim i lubuskim
Źródło: opracowanie własne na podstawie wyników badania *DG1*, 2011.

obserwacje odstające mogą mieć bardzo duży wpływ na jakość modelu. O mocy oddziaływania jednostek nietypowych decyduje bowiem poza ich liczbą także typ (obserwacje odstające w kierunku *x*, obserwacje odstające w kierunku *y*) oraz odległość od jednostek typowych.

W prezentacji graficznej pozwalającej na ocenę typu obserwacji odstającej ograniczono się do domen o najmniejszej wartości współczynnika determinacji, tj. województwa dolnośląskiego i lubuskiego (por. rys. 1).

Wyniki empiryczne badania

Celem badania było porównanie pięciu *M*-estymatorów wykorzystujących różne funkcje do określenia wartości wag pod względem dokładności dopasowania modelu, która była reprezentowana przez odporną wersję współczynnika determinacji R^2 oraz odporną miarę jakości dopasowania modelu *D*.

Różnice w wartościach wyżej wskazanych parametrów otrzymanych dla poszczególnych rodzajów *M*-estymatorów odzwierciedlają wrażliwość na obecność różnych typów obserwacji odstających oraz odległość jednostek nietypowych od pozostałych, standardowych jednostek.

Analiza otrzymanych wyników skłania do wniosku, że wykorzystanie *M*-estymacji poprawia jakość dopasowania modelu tylko wtedy, gdy obecne są obserwacje odstające w kierunku y. Jeśli bliżej przyjrzymy się sekcji „Transport", zauważymy, że zastosowanie *M*-estymacji ze względu na obecność punktów wysokiej dźwigni spowodowało spadek wartości współczynnika determinacji (w porównaniu z klasyczną metodą najmniejszych kwadratów) w prawie wszystkich województwach. Wyjątek stanowiło województwo dolnośląskie. Podobne wnioski można sformułować na podstawie wartości miary jakości dopasowania modelu *D* (por. tabela 4).

Najwyższe wartości współczynników determinacji i najniższe wartości charakterystyki *D* odnotowano dla funkcji Faira. W przypadku dwóch z zastosowanych funkcji wag – Andrewsa i Tukeya, jakość dopasowania modelu jest bardzo podobna. W pozostałych dwóch przypadkach, tzn. dla funkcji Cauchy'ego oraz Hampela, obserwujemy zdecydowanie niższe wartości współczynnika determinacji i wyższe odpornej miary *D*.

Zbadano także parametry modelu oraz wyznaczone dla nich przedziały ufności (por. tabela 5). Zarówno wartości estymatorów parametrów, jak i ich prezentacja graficzna wskazują na dużą zgodność oszacowań w przypadku wszystkich pięciu analizowanych funkcji wag (por. rys. 2 i 3). Dla dwóch województw – lubelskiego i mazursko-warmińskiego, oszacowania współczynnika kierunkowego są bardzo bliskie zera, co oznacza brak korelacji między zmiennymi. W pozostałych przypadkach wartość parametru kształtuje się na poziomie jedności.

Tabela 4. Wartości współczynnika determinacji R^2 oraz odpornej miary D dla M-estymatorów wykorzystujących pięć funkcji wag

Województwo	R^2						D				
	KMNK	Andrewsa	Tukeya	Cauchy'ego	Faira	Hampela	Andrewsa	Tukeya	Cauchy'ego	Faira	Hampela
Dolnośląskie	0,498	0,714	0,717	0,465	0,719	0,661	14318591	14410404	83994151	26216260	53591708
Kujawsko-pomorskie	0,873	0,639	0,641	0,396	0,796	0,602	11188519	11328444	69794042	16593167	39605375
Lubelskie	0,986	0,704	0,708	0,538	0,926	0,654	4443005	4596902	32815603	17380736	25044638
Lubuskie	0,041	0,016	0,016	0,002	0,000	0,005	439610000	444240000	3752000000	2487400000	1912400000
Łódzkie	0,989	0,801	0,802	0,446	0,941	0,659	20233998	20227889	165710000	9334610	96765572
Małopolskie	0,998	0,660	0,661	0,478	0,949	0,575	18282788	18511594	115340000	25866721	74159780
Mazowieckie	0,935	0,613	0,614	0,420	0,836	0,548	321670000	322680000	2308600000	3124500000	1339000000
Opolskie	0,921	0,716	0,717	0,480	0,836	0,650	9463874	9663267	49951895	10478068	51281648
Podkarpackie	0,977	0,790	0,792	0,616	0,941	0,682	852258	863363,8	5874447	893799	3971341
Podlaskie	0,999	0,861	0,863	0,495	0,995	0,683	9267824	9280576	125490000	4479108	72614839
Pomorskie	0,972	0,683	0,686	0,413	0,901	0,658	29649987	30157648	335030000	35278744	145490000
Śląskie	0,992	0,696	0,697	0,423	0,923	0,616	127210000	127980000	843260000	145710000	469890000
Świętokrzyskie	0,996	0,756	0,759	0,507	0,968	0,701	6520625	6548363	52576876	3369794	30685433
Warmińsko-mazurskie	0,888	0,783	0,784	0,153	0,807	0,391	16000423	16011595	354710000	8473638	149740000
Wielkopolskie	0,982	0,755	0,722	0,504	0,934	0,696	26695782	22685514	183960000	17597955	143140000
Zachodniopomorskie	0,910	0,629	0,629	0,442	0,818	0,620	32036788	31904599	174430000	67319033	127940000

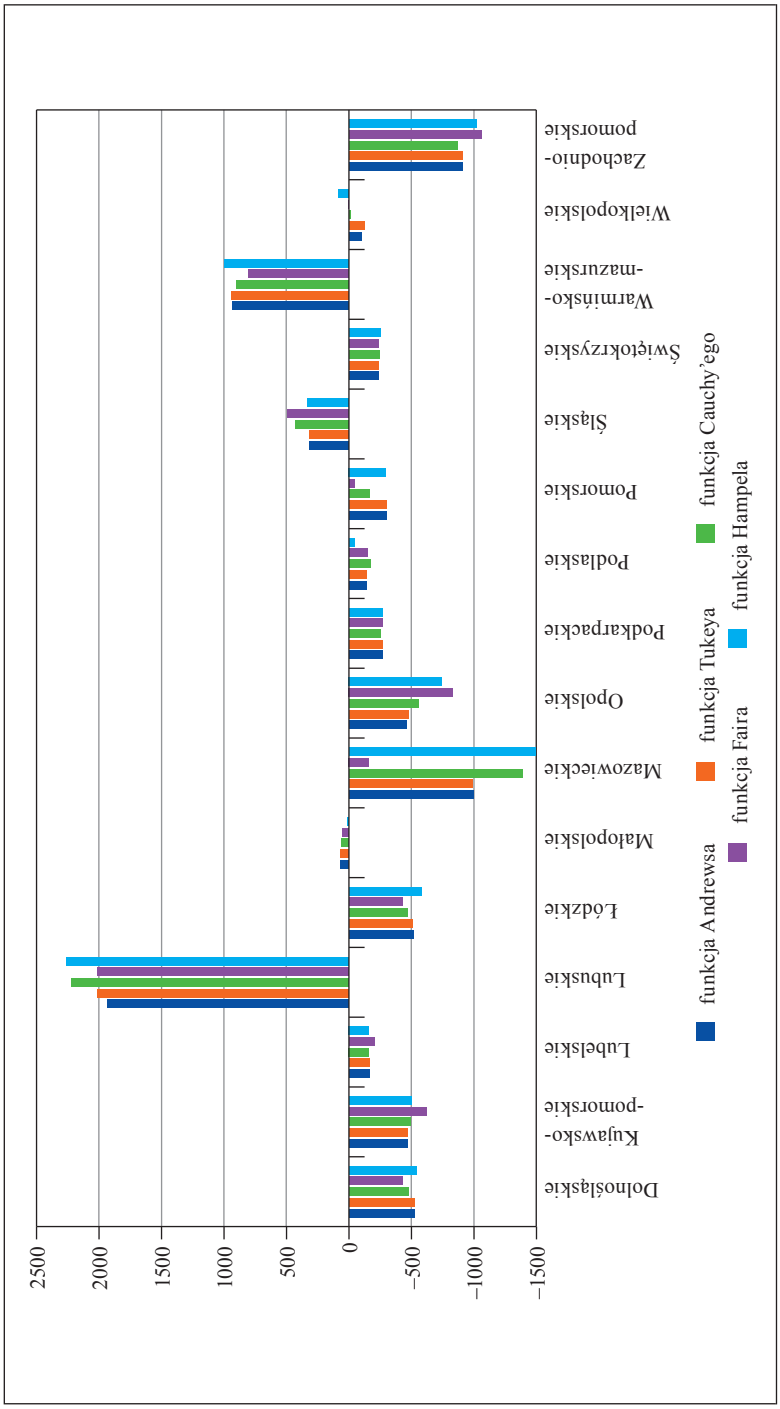
Źródło: opracowanie własne na podstawie wyników badania *DG1*, 2011.

Tabela 5. Szacunki parametrów równania oraz przedziały ufności (95%) dla *M*-estymatorów wykorzystujących pięć funkcji wag

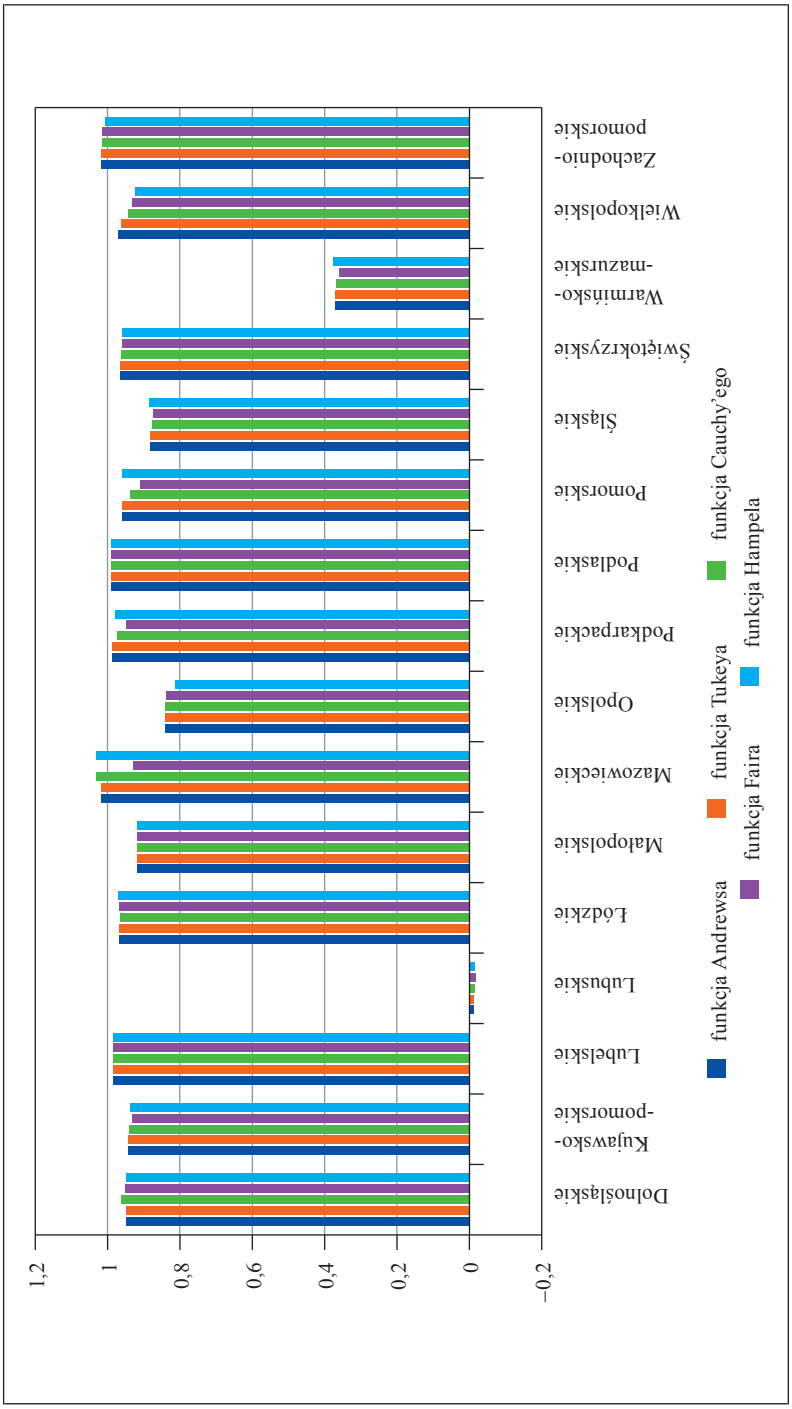
Województwo	Funkcja Andrewsa			Funkcja Tukeya			Funkcja Cauchy'ego			Funkcja Faira			Funkcja Hampela		
	koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności	
Szacunek współczynnika kierunkowego															
Dolnośląskie	1,000	0,949	1,052	1,000	0,949	1,052	1,008	0,962	1,054	1,004	0,953	1,055	1,002	0,950	1,055
Kujawsko-pomorskie	0,981	0,943	1,020	0,982	0,943	1,020	0,980	0,942	1,019	0,982	0,931	1,033	0,979	0,937	1,020
Lubelskie	0,990	0,986	0,994	0,990	0,986	0,994	0,990	0,986	0,994	0,990	0,985	0,995	0,990	0,985	0,995
Lubuskie	–0,004	–0,013	0,006	–0,004	–0,013	0,005	–0,004	–0,014	0,006	–0,001	–0,019	0,017	–0,004	–0,015	0,006
Łódzkie	1,005	0,968	1,041	1,004	0,968	1,041	1,000	0,965	1,034	0,998	0,967	1,029	1,007	0,971	1,043
Małopolskie	0,921	0,919	0,924	0,921	0,919	0,924	0,921	0,918	0,924	0,921	0,917	0,925	0,921	0,918	0,925
Mazowieckie	1,019	1,017	1,020	1,019	1,018	1,020	1,034	1,033	1,035	0,931	0,929	0,932	1,034	1,033	1,035
Opolskie	0,881	0,842	0,920	0,881	0,841	0,921	0,886	0,840	0,932	0,906	0,838	0,974	0,875	0,815	0,935
Podkarpackie	1,014	0,988	1,040	1,014	0,988	1,040	1,002	0,973	1,030	0,985	0,949	1,020	1,008	0,980	1,036
Podlaskie	0,995	0,989	1,000	0,995	0,989	1,000	0,995	0,990	1,000	0,995	0,990	1,000	0,994	0,989	0,999
Pomorskie	0,980	0,961	0,999	0,980	0,960	0,999	0,961	0,939	0,983	0,937	0,911	0,963	0,982	0,961	1,004
Śląskie	0,889	0,884	0,895	0,889	0,884	0,895	0,883	0,877	0,889	0,882	0,874	0,889	0,892	0,885	0,898
Świętokrzyskie	0,984	0,965	1,003	0,984	0,965	1,004	0,982	0,962	1,001	0,979	0,960	0,998	0,980	0,960	1,001
Warmińsko-mazurskie	0,478	0,372	0,584	0,478	0,372	0,584	0,478	0,368	0,588	0,479	0,360	0,598	0,479	0,378	0,579
Wielkopolskie	0,990	0,970	1,010	0,987	0,963	1,011	0,967	0,942	0,992	0,961	0,934	0,988	0,952	0,925	0,980
Zachodniopomorskie	1,037	1,019	1,055	1,037	1,018	1,056	1,034	1,015	1,053	1,041	1,015	1,067	1,032	1,007	1,056

Województwo	Funkcja Andrewsa			Funkcja Tukeya			Funkcja Cauchy'ego			Funkcja Faira			Funkcja Hampela		
	koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności		koszt	przedział ufności	
Szacunek wyrazu wolnego															
Dolnośląskie	–88	–528	351	–88	–528	351	–82	–473	308	8	–429	444	–94	–542	354
Kujawsko-pomorskie	–122	–472	228	–124	–469	222	–146	–492	200	–161	–623	300	–126	–502	249
Lubelskie	–8	–164	147	–5	–166	156	4	–162	170	9	–203	222	45	–160	250
Lubuskie	3590	1927	5252	3622	2010	5234	3977	2220	5735	5167	2009	8325	4097	2263	5930
Łódzkie	–65	–513	384	–65	–509	379	–45	–469	379	–50	–427	327	–141	–585	303
Małopolskie	270	65	474	268	66	471	273	57	488	336	51	620	262	10	515
Mazowieckie	–575	–995	–156	–580	–991	–170	–937	–1391	–482	443	–162	1047	–943	–1492	–395
Opolskie	–28	–463	407	–29	–481	423	–38	–556	480	–67	–833	698	–69	–746	607
Podkarpackie	–102	–265	61	–102	–265	61	–76	–257	105	–46	–272	181	–92	–267	84
Podlaskie	487	–144	1117	488	–143	1118	451	–172	1074	443	–153	1038	562	–47	1170
Pomorskie	52	–299	403	55	–304	414	237	–166	640	430	–49	908	96	–297	490
Śląskie	631	313	949	633	317	948	750	429	1071	907	491	1322	689	328	1050
Świętokrzyskie	37	–233	306	36	–238	310	27	–245	300	31	–234	297	39	–249	328
Warmińsko-mazurskie	1944	935	2953	1944	936	2952	1942	897	2986	1931	802	3061	1950	996	2904
Wielkopolskie	149	–100	398	175	–126	475	297	–10	604	347	6	689	423	81	765
Zachodniopomorskie	–526	–909	–144	–521	–911	–130	–472	–870	–73	–516	–1063	30	–505	–1019	9

Źródło: opracowanie własne na podstawie wyników badania *DG1*, 2011.



Rys. 2. Ocena współczynnika kierunkowego modelu regresji na podstawie wybranych funkcji
Źródło: opracowanie własne na podstawie wyników badania *DGI*, 2011.



Rys. 3. Ocena wyrazu wolnego modelu regresji na podstawie wybranych funkcji
Źródło: opracowanie własne na podstawie wyników badania *DGI*, 2011.

4. Wnioski

Zastosowanie każdej z pięciu badanych funkcji z punktu widzenia dopasowania modelu przyniosło zbliżone rezultaty. W wielu praktycznych sytuacjach wykorzystania *M*-estymacji wybór funkcji Ψ nie jest kluczowy dla uzyskania dobrego odpornego oszacowania. Największe rozbieżności w ocenie szacowanych parametrów dotyczyły funkcji Cauchy'ego i Hampela. Dla tych funkcji jakość dopasowania modelu była też najsłabsza.

Zastosowanie *M*-estymatora w przypadku obecności wartości odstających może wpłynąć na poprawę jakości dopasowania modelu w porównaniu z klasycznymi metodami szacunków, zależy to jednak w dużym stopniu od rodzaju obserwacji nietypowej (odległości). *M*-estymator nie jest odporny na punkty wysokiej dźwigni, a więc powinien być stosowany w sytuacjach, w których punkty wysokiej dźwigni nie występują.

Literatura

Alma Ö.G. [2011], *Comparison of Robust Regression Methods in Linear Regression*, „International Journal of Contemporary Mathematical Sciences", vol. 6, nr 9, http://dx.doi.org/10.12988/ijcms.

Banaś M., Ligas M. [2014], *Empirical Tests of Performance of Some M-estimators*, „Geodesy and Cartography", vol. 63, nr 2, http://dx.doi.org/10.2478/geocart-2014-0015.

Chen C. [2003], *Robust Tools in SAS* [w:] *Developments in Robust Statistics. International Conference on Robust Statistics*, red. R. Dutter i in., Springer Science and Business Media, Berlin–Heidelberg, http://dx.doi.org/10.1007/2F978-3-642-57338-5.

Chen C., Yin G. [2002], *Computing the Efficiency and Tuning Constants for M-Estimation*, Proceedings of the 2002 Joint Statistical Meetings, American Statistical Association, Alexandria.

Cox B.G., Binder A., Chinnappa N.B., Christianson A., Colledge M.J., Kott P.S. [1995], *Business Survey Methods*, John Wiley and Sons, Hoboken, NJ, http://dx.doi.org/10.1002/9781118150504.fmatter.

Fair R.C. [1974], *On the Robust Estimation of Econometric Models*, „Annals of Economic and Social Measurement", vol. 3.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. [2011], *Robust Statistics: The Approach Based on Influence Functions*, John Wiley and Sons, Hoboken, NJ, http://dx.doi.org/10.1002/9781118186435.fmatter.

Holland P., Welsch R. [1977], *Robust Regression Using Iteratively Reweighted Least-Squares*, „Communications in Statistics – Theory and Methods", vol. 6, http://dx.doi.org/10.1080/03610927708827533.

Huber P.J. [1964], *Robust Estimation of a Location Parameter*, „Annals of Mathematical Statistics", vol. 35.

Huber P.J. [1981], *Robust Statistics*, John Wiley and Sons, New York.

Renaud O., Victoria-Feser M. [2010], *A Robust Coefficient of Determination for Regression*, „Journal of Statistical Planning and Inference", vol. 140, nr 7, http://dx.doi.org/10.1016/j.jspi.2010.01.008.
Rousseeuw P.J., Leroy A.M. [1987], *Robust Regression and Outlier Detection*, Wiley-Interscience, New York.
Trzpiot G. [2013], *Wybrane statystyki odporne*, „Studia Ekonomiczne", nr 152.
User's Guide. The Robustreg Procedure [2014], SAS Institute, Cary, NC.

M-estimation in a Small Business Survey

(Abstract)

In many business surveys, sample sizes are large enough to compensate for the presence of outliers, which have a relatively small impact on estimates. However, at low levels of aggregation, the impact of outliers might be significant. Therefore, in the case of a population such as the population of enterprises, the classical approach should be accompanied by methods that resist the occurrence of outliers. To deal with this problem, several alternative technique of estimation, less sensitive to outliers, have been proposed in the statistics literature. In this paper we look at one of them – *M*-estimation, and compare its usefulness in the small businesses survey.

Keywords: robust regression, *M*-estimation, business statistics, outliers.