PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Grażyna Dehnel
Uniwersytet Ekonomiczny w Poznaniu
e-mail: g.dehnel@ue.poznan.pl

MM-ESTYMACJA W BADANIU ŚREDNICH PRZEDSIĘBIORSTW W POLSCE[1]

MM-ESTIMATION IN THE MEDIUM-SIZED ENTERPRISES SURVEY IN POLAND

DOI: 10.15611/pn.2016.426.03

Streszczenie: Większość badań statystycznych dotyczących podmiotów gospodarczych jest prowadzona z wykorzystaniem metody reprezentacyjnej. Jak wiadomo, populacja przedsiębiorstw charakteryzuje się obecnością obserwacji odstających. Wykorzystanie klasycznych metod estymacji może prowadzić do wyników obarczonych dużym błędem. Poszukiwane są zatem metody, które pozwolą na podniesienie precyzji prowadzonych szacunków. W literaturze przedmiotu zaproponowano wiele technik estymacji mniej wrażliwych na wartości odstające. Celem badania będzie próba praktycznego zastosowania jednej z nich – *MM*-estymacji – do badania średnich przedsiębiorstw. Analizie poddane zostaną różne podejścia stosowane w ramach *MM*-estymacji. Ocena i wnioski sformułowane zostaną na podstawie przykładu empirycznego opartego na danych rzeczywistych pochodzących z badania DG1.

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

Summary: Most business surveys are conducted by using survey sampling. As we know, the population of enterprises is characterized by the presence of outliers. The use of classical methods of estimation may produce estimates that are very biased. One therefore looks for methods that will improve the precision of estimates. To deal with this problem, several alternative technique of estimation, less sensitive to outliers, have been proposed in the statistical literature. The aim of paper was to compare usefulness of *MM*-estimation – one of the robust regression methods- in the medium-sized businesses survey. In the study various approaches used in the *MM*-estimation will be analyzed. The assessments and conclusions we formulate on the basis of study relied on data from the DG1 survey.

Keywords: Robust regression, *MM*-estimation, business statistics, outliers.

[1] Projekt finansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2015/17/B/HS4/00905.

1. Wstęp

Wartości odstające są znanym problemem badawczym dotyczącym prawie wszystkich obszarów tematycznych objętych badaniami statystycznymi. Statystyka gospodarcza jest jednak tą dziedziną, w której ich obecność, ze względu na rodzaj badanych zmiennych, może powodować duże obciążenie szacunków. Sposób radzenia sobie z nietypowymi wartościami uzależniony jest m.in. od metody prowadzenia badania. W przypadku badań opartych na próbie proponowane jest zastosowanie tzw. estymacji odpornej, czyli takiej, która uwzględnia procedurę szacowania niewrażliwą na obecność odstających jednostek [Cox i in. 1995]. Estymacja odporna może być realizowana poprzez bezpośrednie korygowanie zidentyfikowanych obserwacji odstających lub pośrednio, przez stosowanie odpornych technik szacunku takich jak *M*-estymacja, *S*-estymacja czy *MM*-estymacja. W niniejszym artykule ograniczono się do analizy jednej z nich – *MM*-estymacji należącej do grupy najnowszych metod. Celem badania była ocena wpływu wyboru rodzaju estymatora wyróżnionego w ramach *MM*-estymacji na jakość szacunku parametrów. Oceny estymatorów dokonano na podstawie badania empirycznego, w którym wykorzystano dane dotyczące średnich przedsiębiorstw działających w ramach sekcji Budownictwo.

2. *MM*-estymacja

MM-estymacja została wprowadzona przez V. Yohai w 1987 r. [Yohai 1987]. Jej zaletą jest to, że łączy w sobie dużą efektywność i wysoki punkt załamania w prosty i intuicyjny sposób. Efekt ten może być osiągnięty poprzez rozszerzenie procesu estymacji do dwóch etapów, w ramach których wykorzystywane są różne rodzaje estymatorów znanych z regresji odpornej. Na pierwszym etapie zaleca się zastosowanie estymatora charakteryzującego się dużą odpornością, takiego jak np. *S*-estymator czy *LTS*-estymator. Drugi etap natomiast wymaga włączenia *M*-estymacji na podstawie szacunku parametru skali otrzymanego na pierwszym etapie [Copt, Hertier 2006]. Cała procedura estymacji przebiega zgodnie z następującym schematem [Alma 2011]:

1. Oszacowanie *S*-estymatora lub *LTS*-estymatora zgodnie z algorytmem, który przedstawili P.J. Rousseeuw i V. Yohai [1984]:

S-estymator: $$\hat{\theta}_S = \arg\min_{\theta} \hat{\sigma}\left(r(\theta)\right) \tag{1}$$

$\hat{\sigma}(r)$ jest *M*-estymatorem skali wyznaczonym jako rozwiązanie równania

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{Y_i - x_i'\theta}{\hat{\sigma}}\right) = K \qquad K = \text{const} = E_{\Theta}\left[\rho\right],$$

gdzie: n – liczebność próby, p – liczba parametrów, $\hat{\sigma}$ – szacunek parametru skali, Θ – rozkład normalny, jako funkcję wpływu $\rho(\cdot)$ przyjęto funkcję Tukeya.

$$\rho(x)=\begin{cases}3\left(\frac{x}{c}\right)^2-3\left(\frac{x}{c}\right)^4+\left(\frac{x}{c}\right)^6 & \text{dla} \quad |x|\le c\\ 1 & \text{dla} \quad |x|>c\end{cases}. \tag{2}$$

Za wartość c przyjęto 2,9366, co zapewniło 25% punkt załamania.

***LTS*-estymator:** $$\hat{\theta}_{LTS}=\arg\min_{\theta}\sum_{i=1}^{h}Q_{LTS}(\theta), \tag{3}$$

gdzie: $Q_{LTS}(\theta)=\sum_{i=1}^{h}r_{(i)}^2$, $h=\frac{3n+p+1}{4}$, p – liczba parametrów, $r_{(1)}^2\le r_{(2)}^2\le\ldots\le r_{(n)}^2$ – uporządkowane wartości kwadratów reszt.

2. Oszacowanie *MM*-estymatora

$$\hat{\theta}_{MM}=\arg\min_{\theta}\sum_{i=1}^{n}\rho\left(\frac{r_i(\theta)}{\hat{\sigma}^{S/LTS}}\right), \tag{4}$$

gdzie $\rho(\cdot)$ jest funkcją wpływu. Podobnie jak w przypadku *S*-estymatora za funkcję wpływu przyjęto funkcję Tukeya (2).

$\hat{\sigma}^{S/LTS}$ jest początkowym szacunkiem parametru skali, wartością startową określoną na pierwszym etapie szacunku *MM*-estymatora. Ostateczna wartość parametru skali ustalana jest na podstawie podejścia iteracyjnego:

$$\sigma_{m+1}^2=\frac{1}{(n-p)K}\sum_{i=1}^{n}\rho\left(\frac{r_i'(\theta)}{\sigma_m}\right)\sigma_m^2. \tag{5}$$

Ze względu na to, że *MM*-estymator jest szczególnym przypadkiem *M*-estymatora, w szacunku kowariancji *MM*-estymatora można skorzystać z metod stosowanych w przypadku *M*-estymacji. Wśród najczęściej stosowanych wskazuje się cztery estymatory [Huber 1973; Huber, Ronchetti 2009]:

C1: $$K^2\frac{\left[1/(n-p)\right]\sum\left(\Psi(r_i)\right)^2}{\left[(1/n)\sum\left(\Psi'(r_i)\right)\right]^2}\left(X^TX\right)^{-1}, \tag{6}$$

C2: $$K\frac{\left[1/(n-p)\right]\sum\left(\Psi(r_i)\right)^2}{\left[(1/n)\sum\left(\Psi'(r_i)\right)\right]}W^{-1}, \tag{7}$$

C3: $$K^{-1}\left[1/(n-p)\right]\sum\left(\Psi(r_i)\right)^2W^{-1}\left(X^TX\right)W^{-1}, \tag{8}$$

C4: $$K^2 \frac{\left[1/(n-p)\right]\sum\left(\Psi(r_i)\right)^2}{\left[(1/n)\sum\left(\Psi'(r_i)\right)\right]^2} W^{-1}, \tag{9}$$

gdzie: $\Psi = \rho'$, $W = \left(w_{jk}\right)$, $w_{jk} = \sum \Psi'(r_i) x_{ij} x_{ik}$ [Huber, Ronchetti 2009].

$K = 1 + \frac{p}{n} \frac{\mathrm{Var}(\Psi')}{(E\Psi')^2}$ – współczynnik korygujący obciążenie estymatora.

W praktyce zarówno $E\Psi'$, jak i $\mathrm{Var}(\Psi')$ są nieznane. Ich wartości mogą być szacowane na podstawie: $E(\Psi') \cong m = \frac{1}{n}\sum \Psi'(r_i)$ oraz $\mathrm{Var}(\Psi') \cong \frac{1}{n}\sum\left[\Psi'(r_i) - m\right]^2$.

3. Charakterystyka badania

Badanie empiryczne oparto na danych pochodzących z badania statystycznego DG1. Podlegają mu przedsiębiorstwa, w których liczba pracujących jest nie mniejsza niż 10 osób. Badaniem objęta jest 10-procentowa próba małych jednostek oraz wszystkie średnie i duże podmioty gospodarcze. Prowadzone jest ono z częstotliwością miesięczną. Dostarcza informacji m.in. na temat takich zmiennych, jak przychód, koszt czy wynagrodzenia. W przeprowadzonym badaniu empirycznym ograniczono się do przedsiębiorstw średnich (liczba pracujących zawiera się w przedziale od 50 do 249 osób), które prowadziły działalność gospodarczą w grudniu 2011 r. Analizie poddano model, w którym za zmienną zależną przyjęto *przychód*, zaś zmiennymi niezależnymi były *koszt*, *dochód* oraz *liczba pracujących*. Źródłem informacji o zmiennych niezależnych był rejestr administracyjny tworzony na podstawie zeznań podatkowych. Szacunku dokonano w przekroju regionalnym, z uwzględnieniem rodzaju prowadzonej działalności gospodarczej. Przekrój regionalny obejmował jednostki na poziomie województw, a rodzajowi prowadzonej działalności odpowiadały sekcje PKD zgodne z klasyfikacją NACE. Ze względu na to, że otrzymane wyniki estymacji są bardzo obszerne, ich prezentacja przedstawiona w dalszej części artykułu zostanie ograniczona do szacunków dla sekcji Budownictwo. W wyborze sekcji kierowano się tym, by obserwacje odstające, obecne w badanych domenach, reprezentowały nie tylko jednostki nietypowe, lecz także wpływowe.

4. Ocena szacunków otrzymanych w badaniu empirycznym

Oceny szacunków dokonano na podstawie względnych standardowych błędów szacunku oraz odpornej wersji współczynnika determinacji [Renaud, Victoria-Feser 2010]:

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

gdzie: ρ – funkcja celu; $\hat{\mu}$, $\hat{s}$ – estymatory parametru położenia oraz skali.

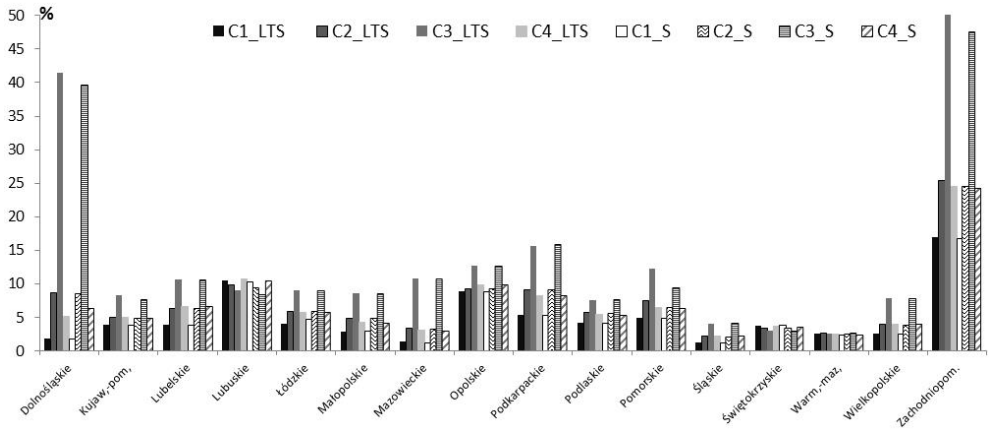
Ponadto w celu porównania podstawowych własności estymatorów parametrów równania regresji, takich jak efektywność i obciążenie, zastosowano jedną ze znanych metod przybliżonych, opartą na podpróbkach i zasadzie *bootstrap*. Na podstawie 1000 podprób wyznaczono miary [Choudhry, Rao 1993]:

- względnej efektywności $$CV\left(\hat{Y}_d\right) = \frac{\sqrt{\mathrm{Var}\left(\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)} = \frac{\sqrt{\frac{1}{999}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d} - \hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)}, \tag{11}$$
- względnego obciążenia $$\mathrm{ARB}\left(\hat{Y}_d\right) = \frac{1}{1000}\left|\sum_{b=1}^{1000}\frac{\hat{Y}_{b,d} - Y_d}{Y_d}\right|, \tag{12}$$
- względnego MSE $$\mathrm{RMSE}\left(\hat{Y}_d\right) = \frac{\sqrt{\frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d} - Y_d\right)^2}}{Y_d}. \tag{13}$$

5. Wyniki empiryczne badania

Celem badania było porównanie jakości szacunków otrzymanych w oparciu o osiem *MM*-estymatorów różniących się między sobą sposobem wyznaczania. O ich rodzaju decydowały dwa elementy: typ estymatora przyjętego jako tzw. punkt startowy w *MM*-estymacji oraz rodzaj zastosowanego estymatora kowariancji. W odniesieniu do pierwszego elementu przebadano dwa podejścia: *S*-estymację oraz *LTS*-estymację. Dodatkowo, w ramach każdego z tych podejść, uwzględnione zostały cztery różne estymatory kowariancji opisane wzorami (6)–(9).

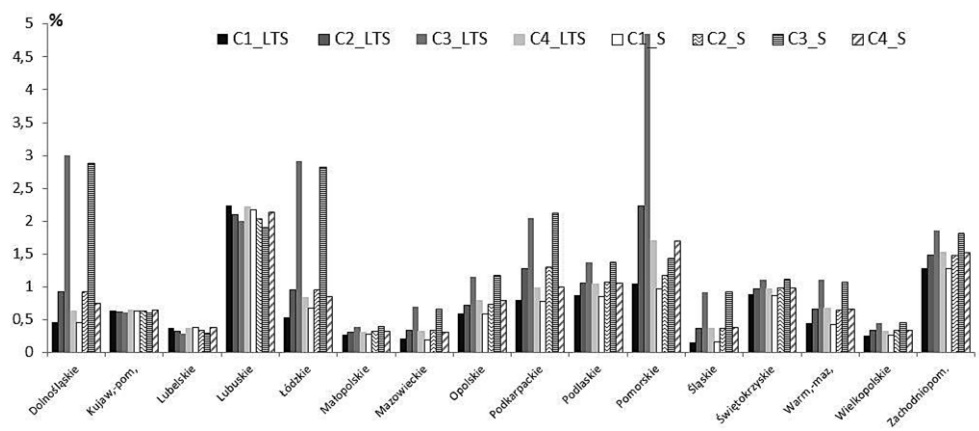
Badaniu, w pierwszej kolejności, poddano szacunki parametrów modeli oraz wyznaczone dla nich przedziały ufności. W przypadku szacunków parametrów otrzymane wartości dla wszystkich ośmiu analizowanych estymatorów kształtują się na tym samym poziomie. Z kolei w przypadku przedziałów ufności widoczne jest, w ramach każdego województwa, nieznaczne zróżnicowanie. Na rysunkach 1–3 zaprezentowano wykorzystane do budowy przedziałów ufności standardowe błędy



Rys. 1. Wartości względnych standardowych błędów szacunku dla zmiennej *koszt*

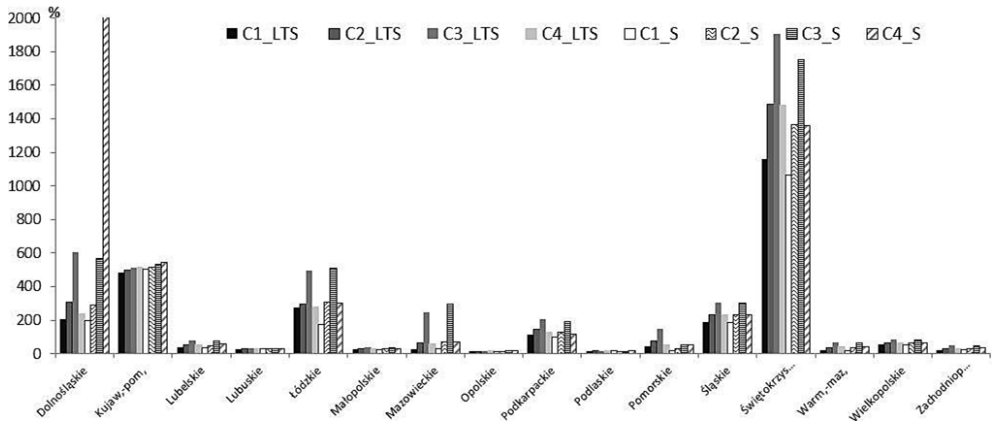
Źródło: opracowanie własne na podstawie badania DG1.

szacunku w ujęciu względnym, dla zmiennych uwzględnionych w modelu tzn. *kosztu*, *dochodu* oraz *liczby pracujących*. Prezentacja graficzna wskazuje na dość umiarkowaną dyspersję oszacowań w odniesieniu do pojedynczego województwa (por. rys. 1–3). Zmienność błędów szacunku wynika przede wszystkim z dużych ich wartości zanotowanych w przypadku estymatora kowariancji C3 w porównaniu do pozostałych estymatorów. Dotyczy to zarówno podejścia, w którym wykorzystuje się zarówno *LTS*-estymację, jak i *S*-estymację. W obu podejściach poziom względnych standardowych błędów szacunku dla estymatora C3 jest zbliżony.



Rys. 2. Wartości względnych standardowych błędów szacunku dla zmiennej *dochód*

Źródło: opracowanie własne na podstawie badania DG1.

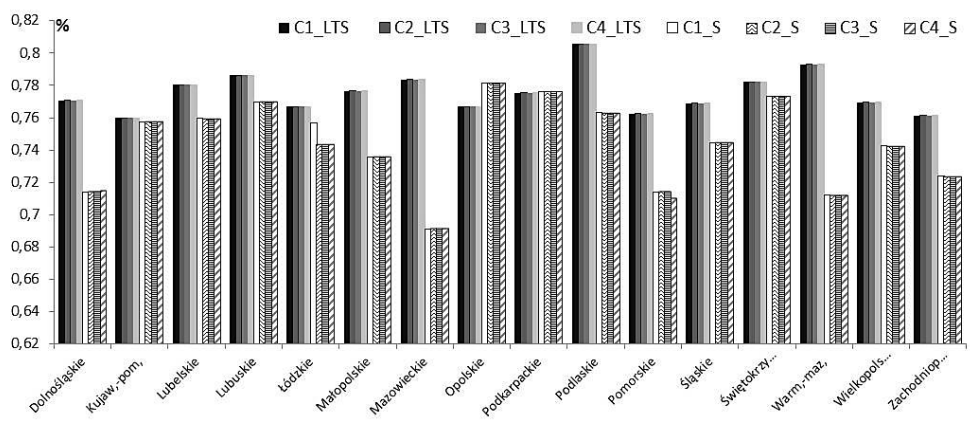


Rys. 3. Wartości względnych standardowych błędów szacunku dla zmiennej *liczba pracujących*

Źródło: opracowanie własne na podstawie badania DG1.

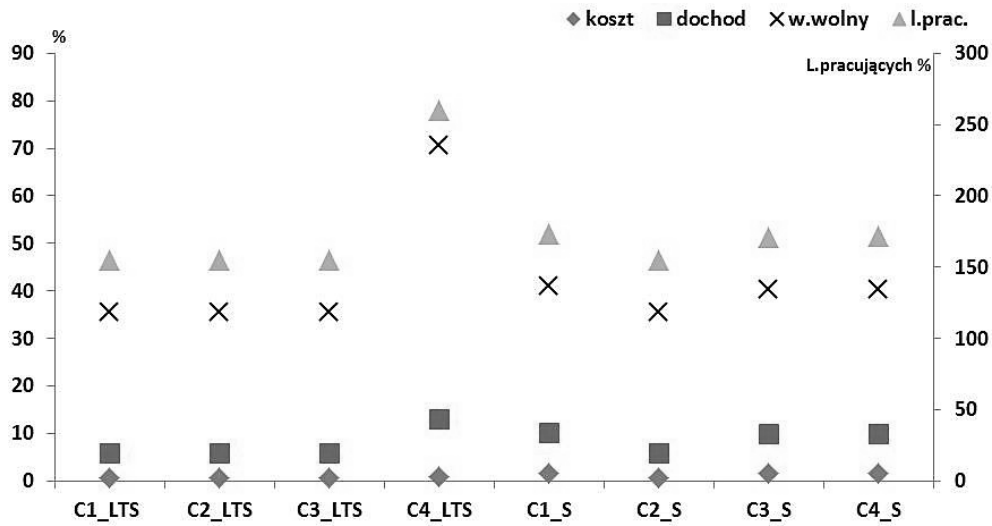
Widoczna jest również zależność między rodzajem zmiennej uwzględnionej w modelu a wartościami błędów. W przypadku *liczby pracujących* poziom błędów szacunku jest znacznie wyższy niż dla pozostałych dwóch zmiennych. W kilku województwach przyjmują one nawet nieakceptowalnie duże wartości.

Kolejnym krokiem w badaniu była analiza odpornej wersji współczynnika determinacji (por. rys. 4). Wskazuje ona na to, że rodzaj estymacji, który zostanie użyty na pierwszym etapie szacunku *MM*-estymatora, ma wpływ na dopasowanie modelu. Zmiana wartości współczynnika nie jest jednak znaczna, nie przekracza



Rys. 4. Wartości odpornego współczynnika determinacji

Źródło: opracowanie własne na podstawie badania DG1.

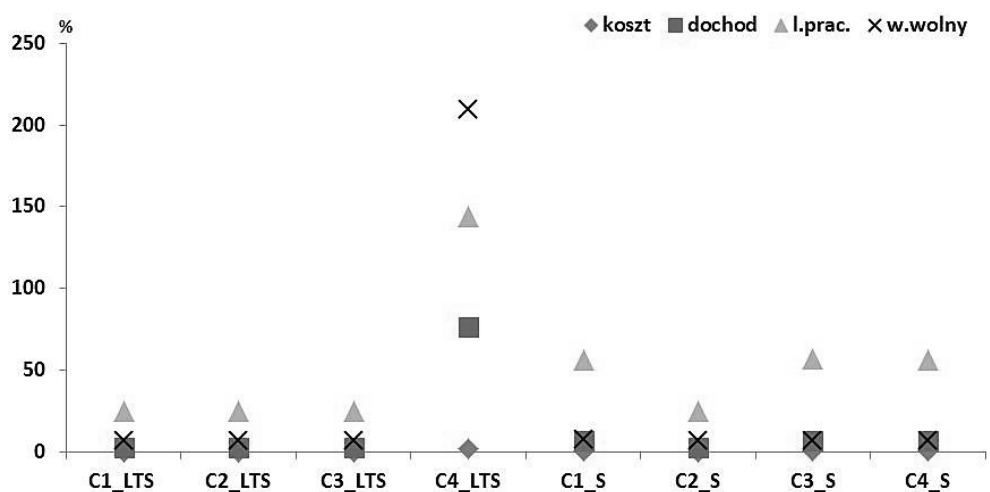


Rys. 5. Względna efektywność – CV estymatorów parametrów równania regresji

Źródło: opracowanie własne na podstawie badania DG1.

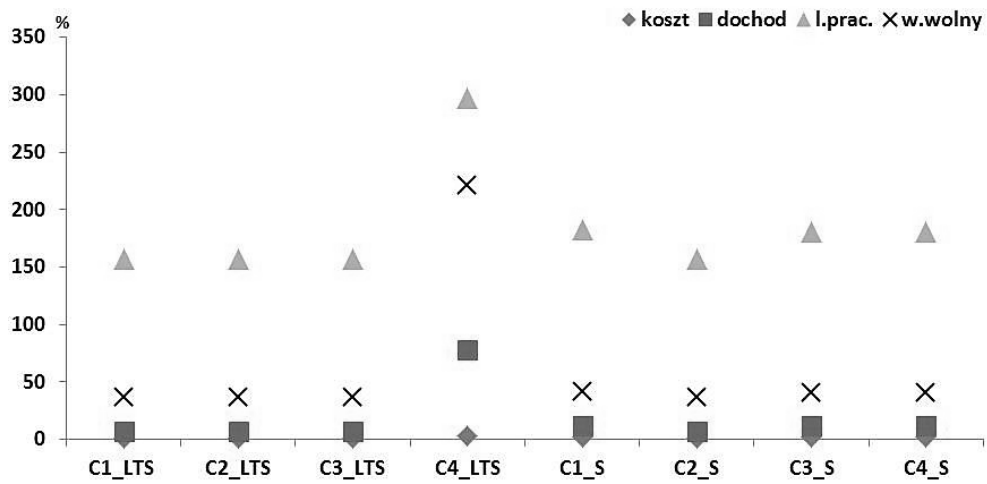
bowiem 9 punktów procentowych. W przeprowadzonym badaniu wykorzystanie *LTS*-estymatora poprawiło jakość modelu w 14 województwach.

Ostatni etap analizy dotyczył porównania własności estymatorów parametrów równania regresji. Na rysunkach 5–7 zaprezentowano wartości trzech mierników



Rys. 6. Względne obciążenie – ARB estymatorów parametrów równania regresji

Źródło: opracowanie własne na podstawie badania DG1.



Rys. 7. Względny średni błąd kwadratowy – RMSE estymatorów parametrów równania regresji

Źródło: opracowanie własne na podstawie badania DG1.

oceniających efektywność, obciążenie oraz MSE. W ramach każdej z analizowanych zmiennych relacje pomiędzy wartościami charakterystyk *MM*-estymatorów są bardzo zbliżone, tak w przypadku efektywności, jak i obciążenia, co ma bezpośredni wpływ na wartości MSE. Wyjątek od tej reguły stanowi estymator C4_LTS charakteryzujący się (w przypadku wyrazu wolnego i dwóch badanych zmiennych) najniższą efektywnością oraz najwyższym, znacznie odbiegającym od pozostałych wartości poziomem obciążenia.

6. Zakończenie

Analiza otrzymanych wyników skłania do wniosku, że zastosowanie *MM*-estymacji w populacji charakteryzującej się obecnością jednostek odstających, niezależnie od rodzaju obserwacji nietypowych, pozwala na budowę dobrze dopasowanego modelu. Na poziom dopasowania korzystnie wpływa użycie *LTS*-estymacji na pierwszym etapie procedury szacunku *MM*-estymatora.

Wykorzystanie każdego z wyróżnionych w ramach *MM*-estymacji estymatorów przyniosło dość podobne rezultaty w ocenie ich jakości. Wśród tych, których ocena wypadła najgorzej znalazły się dwa: C3 ze względu na stosunkowo duże średnie błędy szacunku oraz C4_LTS z uwagi na znaczne obciążenie estymatora parametru funkcji regresji. Oceny pozostałych sześciu estymatorów są bardzo zbliżone.

Literatura

Alma Ö.G., 2011, *Comparison of robust regression methods in linear regression*, International Journal of Contemporary Mathematical Sciences, vol. 6, no. 9, s. 409–421.

Choudhry G.H., Rao J.N.K., 1993, *Evaluation of small area estimators and empirical study*, [w:] Small Area Statistics and Survey Designs, GUS, Warszawa.

Copt S., Hertier S., 2006, *Robust MM-estimation and Inference in Mixed Linear Models*, http://www.unige.ch/ses/metri/cahiers/2006_01.pdf (1.09.2015).

Cox B.G., Binder A., Chinnappa N.B., Christianson A., Colledge M.J., Kott P.S., 1995, *Business Survey Methods*, John Wiley & Sons, New York.

Huber P.J., 1973, *Robust regression: Asymptotics, conjectures and Monte Carlo*, Annals of Statistics, vol. 1, no. 5, s. 799–821.

Huber P.J., Ronchetti E.M., 2009, *Robust Statistics*, John Wiley & Sons, Hoboken, NJ.

Renaud O., Victoria-Feser M., 2010, *A robust coefficient of determination for regression*, Journal of Statistical Planning and Inference, vol. 140, no. 7, s. 1852–1862.

Rousseeuw P.J., Yohai V., 1984, *Robust regression by means of S-estimators*, [w:] W.H.J. Franke, D. Martin (red.), *Robust and Nonlinear Time Series Analysis*, Springer-Verlag, New York, s. 256–272.

Yohai V., 1987, *High breakdown-point and high efficiency robust estimates for regression*, The Annals of Statistics, vol. 15, no. 2, s. 642–656.