

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 385

**Taksonomia 25**

**Klasyfikacja i analiza danych –  
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

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 stronie internetowej Wydawnictwa  
[www.pracnaukowe.ue.wroc.pl](http://www.pracnaukowe.ue.wroc.pl)  
[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

Publikacja udostępniona na licencji Creative Commons  
Uznanie autorstwa-Użycie niekomercyjne-Bez utworów zależnych 3.0 Polska  
(CC BY-NC-ND 3.0 PL)



© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2015

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

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:  
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
tel./fax 71 36 80 602; e-mail:[econbook@ue.wroc.pl](mailto:econbook@ue.wroc.pl)  
[www.ksiegarnia.ue.wroc.pl](http://www.ksiegarnia.ue.wroc.pl)

Druk i oprawa: TOTEM

## Spis treści

Wstęp.....	9
<b>Tomasz Bartłomowicz:</b> Segmentacja konsumentów na podstawie preferencji wyrażonych uzyskanych metodą Maximum Difference Scaling .....	11
<b>Barbara Batóg, Jacek Batóg, Andrzej Niemiec, Wanda Skoczylas, Piotr Waśniewski:</b> Zastosowanie metod klasyfikacyjnych w identyfikacji kluczowych indyktorów osiągnięć w zarządzaniu wynikami przedsiębiorstw .....	20
<b>Iwona Bąk:</b> Wykorzystanie statystycznej analizy danych w badaniach turystyki transgranicznej na obszarach chronionych.....	28
<b>Beata Bieszk-Stolorz:</b> Ocena stopnia deprecjacji kapitału ludzkiego z wykorzystaniem nieliniowych modeli regresji.....	37
<b>Mariola Chrzanowska, Nina Drejerska:</b> Małe i średnie przedsiębiorstwa w strefie podmiejskiej Warszawy – określenie znaczenia lokalizacji z wykorzystaniem drzew klasyfikacyjnych.....	45
<b>Adam Depta:</b> Próba modelowania strukturalnego jakości życia osób jękaących się jako konstrukt ukrytego na podstawie kwestionariusza SF-36v2 .....	53
<b>Katarzyna Dębkowska:</b> Wielowymiarowa analiza kondycji finansowej przedsiębiorstw sektora e-usług .....	63
<b>Krzysztof Dmytrów, Mariusz Doszyń:</b> Taksonomiczna procedura wspomagania kompletacji produktów w magazynie .....	71
<b>Mariusz Doszyń, Sebastian Gnat:</b> Propozycja procedury taksonomiczno-ekonometrycznej w indywidualnej wycenie nieruchomości.....	81
<b>Marta Dziechciarz-Duda, Anna Król:</b> Zastosowanie analizy <i>unfolding</i> i regresji hedonicznej do oceny preferencji konsumentów .....	90
<b>Katarzyna Frodyma:</b> Współzależność między poziomem rozwoju gospodarczego a udziałem energii ze źródeł odnawialnych w końcowym zużyciu w krajach Unii Europejskiej.....	99
<b>Hanna Gruchociak:</b> Porównanie struktury lokalnych rynków pracy wyznaczonych przy wykorzystaniu różnych metod w Polsce w latach 2006 i 2011 .	111
<b>Alicja Grześkowiak, Agnieszka Stanimir:</b> Postrzeganie środowiska pracy przez starszą i młodszą generację pracowników .....	120
<b>Marta Hozer-Koćmiel, Christian Lis:</b> Klasyfikacja krajów nadbałtyckich ze względu na czas prac wykonywanych w gospodarstwie domowym .....	129
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel:</b> Zegar cyklu koniunkturalnego państw UE i USA w latach 1995-2013 w świetle badań synchronizacji.....	138
<b>Aleksandra Łuczak:</b> Wykorzystanie rozszerzonej interwałowej metody TOPSIS do porządkowania liniowego obiektów .....	147

<b>Aleksandra Łuczak, Feliks Wysocki:</b> Zintegrowane podejście do ustalania współczynników wagowych dla cech w zagadnieniach porządkowania linowego obiektów .....	156
<b>Małgorzata Markowska, Danuta Strahl:</b> Wykorzystanie klasyfikacji dynamicznej do identyfikacji wrażliwości na kryzys ekonomiczny unijnych regionów szczebla NUTS 2.....	166
<b>Aleksandra Matuszewska-Janica, Marta Hozer-Koćmiel:</b> Struktura zatrudnienia oraz wynagrodzenia kobiet i mężczyzn a przedmiotowa struktura gospodarcza w państwach UE.....	178
<b>Anna M. Olszewska:</b> Zastosowanie analizy korespondencji do badania związku pomiędzy zarządzaniem jakością a innowacyjnością przedsiębiorstw .....	187
<b>Małgorzata Podogrodzka:</b> Metoda aglomeracyjna w ocenie przestrzennego zróżnicowania starości demograficznej w Polsce .....	195
<b>Ewa Roszkowska, Tomasz Wachowicz:</b> Ocena ofert negocjacyjnych spoza dopuszczalnej przestrzeni negocjacyjnej.....	201
<b>Ewa Roszkowska, Tomasz Wachowicz:</b> Zastosowanie metody <i>unfolding</i> do wspomagania procesu negocjacji .....	210
<b>Małgorzata Rószkiewicz:</b> Próba diagnozy uwarunkowań poziomu wskaźnika braku odpowiedzi w środowisku polskich gospodarstw domowych.....	219
<b>Marcin Salamaga:</b> Próba identyfikacji muzycznych profili melomanów z wykorzystaniem drzew klasyfikacyjnych i regresyjnych .....	229
<b>Agnieszka Sompolska-Rzechuła:</b> Określenie czynników wpływających na prawdopodobieństwo poprawy poziomu rozwoju społecznego z wykorzystaniem modelu logitowego .....	239
<b>Iwona Staniec:</b> Wykorzystanie analizy czynnikowej w identyfikacji konstruktywów ukrytych determinujących ryzyko współpracy.....	248
<b>Agnieszka Stanimir:</b> Skłonność do zagranicznej mobilności młodszych i starszych osób .....	257
<b>Mirosława Sztemberg-Lewandowska:</b> Problemy decyzyjne w funkcjonalnej analizie głównych składowych.....	267
<b>Tomasz Szubert:</b> Demograficzno-społeczne determinanty określające subiektywny status jednostki w polskim społeczeństwie .....	276
<b>Piotr Tarka:</b> Własności 5- i 7-stopniowej skali Likerta w kontekście normalizacji zmiennych metodą Kaufmana i Rousseeuwa .....	286
<b>Joanna Trzęsiok:</b> Nielklasyczne metody regresji a problem odporności .....	296
<b>Katarzyna Wawrzyniak:</b> Ocena podobieństwa wyników uporządkowania województw uzyskanych różnymi metodami porządkowania .....	305
<b>Katarzyna Wójcik, Janusz Tuchowski:</b> Wykorzystanie metody opartej na wzorcach w automatycznej analizie opinii konsumenckich.....	314
<b>Anna Zamojska:</b> Zastosowanie analizy falkowej w ocenie efektywności funduszy inwestycyjnych .....	325

## Summaries

<b>Tomasz Bartłomowicz:</b> Segmentation of consumers based on revealed preferences obtained with the Maximum Difference Scaling method .....	19
<b>Barbara Batóg, Jacek Batóg, Andrzej Niemiec, Wanda Skoczylas, Piotr Waśniewski:</b> Application of classification methods to identify the key performance indicators of performance management .....	27
<b>Iwona Bąk:</b> The application of statistical data analysis in the studies of cross-border tourism in protected areas.....	36
<b>Beata Bieszk-Stolorz:</b> Evaluating human capital depreciation by means of non-linear regression models.....	44
<b>Mariola Chrzanowska, Nina Drejerska:</b> Small and medium enterprises in the Warsaw suburban zone – determination of a localization’s role using classification trees .....	52
<b>Adam Depta:</b> An attempt of structural modelling of the quality of life of stuttering people as a latent construct, based on SF-36v2 questionnaire ...	62
<b>Katarzyna Dębowska:</b> Multidimensional analysis of financial condition of e-business services .....	70
<b>Krzysztof Dmytrów, Mariusz Doszyń:</b> Taxonomic procedure of supporting order-picking of products in a warehouse .....	80
<b>Mariusz Doszyń, Sebastian Gnat:</b> Taxonomic and econometric methods in individual real estate evaluation.....	89
<b>Marta Dziechciarz-Duda, Anna Król:</b> The application of unfolding analysis and hedonic regression in the investigation of consumers’ preferences .....	98
<b>Katarzyna Frodyma:</b> Interdependence between the level of economic development and the share of renewable energy in gross final energy consumption in the European Union.....	110
<b>Hanna Gruchociak:</b> Comparison of local labour markets structure designated using different methods in Poland in 2006 and 2011 years.....	119
<b>Alicja Grześkowiak, Agnieszka Stanimir:</b> Perception of working environment by older and younger generation of workers.....	128
<b>Marta Hozer-Koćmiel, Christian Lis:</b> Classification of the Baltic Sea Region countries due to the time of household work.....	137
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel:</b> Business cycle clock for the EU and the USA in 1995-2013 in the light of synchronization research.....	146
<b>Aleksandra Łuczak:</b> The use of the extended interval TOPSIS methods for linear ordering of objects.....	155
<b>Aleksandra Łuczak, Feliks Wysocki:</b> Integrated approach for determining the weighting coefficients for features in issues of linear ordering of objects.....	165

<b>Małgorzata Markowska, Danuta Strahl:</b> The application of dynamic classification for the identification of vulnerability to economic crisis in the EU NUTS 2 regions .....	177
<b>Aleksandra Matuszewska-Janica, Marta Hozer-Koćmiel:</b> The structure of male and female employment and remuneration vs. the basic economy structure in the EU countries .....	186
<b>Anna M. Olszewska:</b> The application of the correspondence analysis for the study of the relations between quality management and innovation in the enterprises.....	194
<b>Małgorzata Podogrodzka:</b> Agglomeration method in the age and ageing in Poland by voivodships.....	200
<b>Ewa Roszkowska, Tomasz Wachowicz:</b> Scoring the negotiation offers from the outside of the feasible negotiation space .....	209
<b>Ewa Roszkowska, Tomasz Wachowicz:</b> Application of the unfolding analysis to negotiation support.....	218
<b>Małgorzata Rószkiewicz:</b> An attempt to diagnose the determinants of non-response rate in Polish households surveys .....	228
<b>Marcin Salamaga:</b> Attempt to identify music lovers profiles using classification and regression trees .....	238
<b>Agnieszka Sompolska-Rzechuła:</b> The definition of factors influencing the probability of improving the level of human development using the logit model.....	247
<b>Iwona Staniec:</b> The use of factor analysis to identify hidden constructs – determinants of the cooperation risk .....	256
<b>Agnieszka Stanimir:</b> Willingness to mobility abroad among younger and older persons .....	266
<b>Mirosława Sztemberg-Lewandowska:</b> Decision problems in functional principal components analysis.....	275
<b>Tomasz Szubert:</b> Socio-demographic factors determining subjective social status of an individual in Polish society .....	285
<b>Piotr Tarka:</b> Normalization methods of variables and measurement on 5 and 7 point Likert scale .....	295
<b>Joanna Trzęsiok:</b> Non-classical regression methods vs. robustness .....	304
<b>Katarzyna Wawrzyniak:</b> The evaluation of the similarity of the voivodships' orderings obtained by means of different methods.....	313
<b>Katarzyna Wójcik, Janusz Tuchowski:</b> Using pattern-based opinion mining.....	324
<b>Anna Zamojska:</b> Mutual funds performance measurement – wavelets analysis approach.....	333

**Joanna Trzęsiok**

Uniwersytet Ekonomiczny w Katowicach

e-mail: joanna.trzesiok@ue.katowice.pl

---

## NIEKLASYCZNE METODY REGRESJI A PROBLEM ODPORNOŚCI

---

**Streszczenie:** Artykuł poświęcony jest ważnemu zagadnieniu w statystyce, jakim jest problem odporności. Omówiono różne aspekty podejścia do tego tematu w analizie regresji. W ich kontekście badania poddane zostały wybrane nieparametryczne metody regresji, takie jak PPR, POLYMARS, MART i RANDOM FORESTS oraz parametryczna regresja grzbietowa. W przeprowadzonej analizie porównawczej testowano odporność tych metod na występowanie w zbiorze uczącym: wartości oddalonych, losowych zakłóceń wartości zmiennej zależnej oraz zmiennych nieistotnych. Porównań dokonano za pomocą procedur symulacyjnych na zbiorach danych wykorzystywanych standardowo do badania własności metod regresji. Pomimo dosyć powszechnych przekonań o odporności regresji nieparametrycznej, okazało się, iż uzyskane modele nie we wszystkich przypadkach są niewrażliwe na zakłócenia występujące w zbiorze uczącym.

**Słowa kluczowe:** analiza regresji, odporność, regresja nieparametryczna.

DOI: 10.15611/pn.2015.385.32

### 1. Wstęp

W modelowaniu zjawisk ekonomicznych ważnym zagadnieniem jest problem odporności metod na zakłócenia danych wynikające np. z błędów pomiaru, braku losowości próby czy występowania obserwacji oddalonych. Efektem zastosowania metod, które nie są odporne na zakłócenia wartości cech, może być zbudowanie modelu, który nie odzwierciedla głównych mechanizmów regulujących zachowanie badanego zjawiska. W związku z tym predykcja, wnioskowanie i podejmowanie decyzji na podstawie takiego modelu może być obarczone dużymi błędami. Szczególnego znaczenia nabiera to w przypadku modeli nieklasycznych, które charakteryzują się dużą elastycznością i zdolnością do adaptacyjnego, dokładnego dopasowania się do danych (uwzględniając również zmienność wynikającą z zakłóceń).

W artykule przedstawiono wybrane podejścia do problemu odporności w analizie regresji. W ich kontekście badaniu poddane zostały wybrane nieparametryczne metody regresji, takie jak metoda rzutowania PPR, metoda krzywych sklepanych POLYMARS, metody MART i RANDOM FORESTS wykorzystujące drzewa regresyjne, jak również nieklasyczna metoda parametryczna – regresja grzbietowa.

Celem pracy było zbadanie, które z metod regresji prowadzą do uzyskania modeli odpornych, czyli charakteryzujących się wysokimi wartościami miar dokładności predykcji pomimo zakłóceń występujących w zbiorach danych.

## 2. Różne aspekty problemu odporności

W najbardziej ogólnym rozumieniu zastosowanie odpornej metody regresji oznacza, że mamy do czynienia z modelem, który wskazuje tendencję reprezentowaną przez większość obserwacji. Model taki jest niewrażliwy na działanie czynników niezwiązanych z badanym zjawiskiem, które mogą zakłócić wyniki analizy. Jednak odporność regresji można rozpatrywać w kilku aspektach.

Metoda regresji może być odporna na:

- występowanie w zbiorze uczącym *wartości oddalonych* (nietypowych), które mogą zakłócić i istotnie zmienić równanie funkcji regresji,
- *losowe zakłócenia wartości zmiennej zależnej* (np. losowe błędy pomiaru o rozkładzie normalnym);
- występowanie w zbiorze uczącym *zmiennych nieistotnych*, które nie mają wpływu na postać modelu i wartości zmiennej zależnej;
- *dobór próby* do zbioru uczącego, na którym budowany jest model;
- *braki wartości* niektórych zmiennych w zbiorze uczącym;
- *niespełnienie założeń* nakładanych na tę metodę.

Najczęściej, mówiąc o odporności regresji, mamy na myśli niewrażliwość modelu na jakość danych, czyli przede wszystkim na obecność w zbiorze uczącym obserwacji oddalonych (nietypowych), które mogą wynikać z zakłóceń wartości, zarówno zmiennej zależnej, jak i zmiennych objaśniających, błędami pomiaru. Identyfikacja obserwacji oddalonych, jak również sposoby radzenia sobie z nimi, są ważnymi zagadnieniami związanymi z pojęciem odporności w statystyce [Trzpiot (red.) 2013].

W najbardziej ogólnym przypadku można rozpatrywać odporność metody regresji w kontekście niespełnienia części założeń wymaganych dla prawidłowego działania danej metody. Testujemy wtedy możliwość zastosowania tej metody i uzyskania poprawnych wyników, pomimo że nie wszystkie nałożone na nią warunki będą spełnione. Ten przypadek nie będzie jednak rozpatrywany w tym artykule, ponieważ nie dotyczy analizowanych tutaj nieparametrycznych metod regresji, które charakteryzują się właśnie brakiem lub niewielką liczbą założeń dotyczących zmiennych, reszt czy postaci modelu.



Niniejsza praca jest pierwszą próbą zmierzenia się autorki z problemem odporności regresji. Skupiono się w niej na przeprowadzeniu analizy odporności wybranych metod regresji w trzech pierwszych wymienionych aspektach i nie podjęto szerszego omówienia odporności regresji na dobór próby oraz braki wartości niektórych zmiennych. Temat ten będzie kontynuowany w dalszych badaniach.

### 3. Metody regresji wykorzystane w badaniu

Tak jak już wspomniano we wstępie, problem odporności nabiera szczególnego znaczenia w przypadku nieklasycznych metod regresji, które pozwalają na budowę modeli, które w sposób elastyczny i adaptacyjny dopasowują się do danych ze zbioru uczącego. Modele budowane na zbiorach danych, w których wartości zakłócone są np. błędami pomiaru, wartościami oddalonymi czy nieistotnymi, mogą mieć niewielkie zdolności predykcyjne, a więc również małą wartość poznawczą dla badacza.

Wiele z metod nieparametrycznych ma wbudowany mechanizm regularyzacji, który pozwala ograniczyć problem nadmiernego dopasowania modelu do danych ze zbioru uczącego i polega na przyjęciu kompromisu pomiędzy właśnie dopasowaniem tego modelu a jego złożonością [Trzęsiok 2011]. Mechanizm regularyzacji prowadzi do zwiększenia zdolności predykcyjnych modelu, przez co metody nieparametryczne uchodzą za bardziej odporne. Zachodzi jednak pytanie, w jakim stopniu mechanizm ten jest skuteczny, a omawiane metody są rzeczywiście odporne na wartości oddalone, zakłócenia zmiennej zależnej czy występowanie zmiennych nieistotnych w zbiorze uczącym.

Analizie poddane zostały cztery wybrane metody nieparametryczne, często wykorzystywane w badaniach i charakteryzujące się dobrymi własnościami predykcyjnymi [Meyer, Leisch, Hornik 2003]:

- metoda rzutowania PPR [Koooperberg, Bose, Stone 1997],
- wielowymiarowa metoda krzywych sklepanych POLYMARS [Friedman, Stuetzle 1981],
- addytywna metoda drzew regresyjnych MART [Friedman 1999a; Friedman 1999b],
- metoda zagregowanych drzew Breimana – RANDOM FORESTS [Breiman 2001].

W badaniu wykorzystano również metodę parametryczną – regresję grzbietową (*ridge regression*) [Hoerl, Kennard 1970], którą można nazwać „próbą poprawienia” liniowego modelu regresji wielorakiej. Regresja grzbietowa nie wymaga spełnienia tak wielu założeń jak metoda najmniejszych kwadratów, przez co większy jest zakres jej zastosowań.

## 4. Analiza porównawcza

Ze względu na odmienne mechanizmy działania nieparametrycznych metod regresji niemożliwe jest analityczne porównanie generowanych przez nie modeli. Z tego względu badania porównawcze przeprowadzono za pomocą procedur symulacyjnych na zbiorach danych standardowo wykorzystywanych do testowania własności różnych metod regresji. Posłużono się zbiorami danych opisanymi w literaturze, które również przez innych autorów zostały zastosowane do testowania odporności metod regresji. Wszystkie analizy i obliczenia przeprowadzono z wykorzystaniem programu statystycznego **R** z dołączonymi bibliotekami tego programu.

W każdym przypadku badano zdolności predykcyjne modelu za pomocą błędu średniokwadratowego  $MSE$ , obliczonego na zbiorze testowym (jeśli był dostępny) lub metodą sprawdzania krzyżowego (ozn.  $MSE_{cv}$ ).

### 4.1. Odporność na występowanie obserwacji oddalonych

Testowanie odporności wybranych metod regresji na występowanie w zbiorze uczącym wartości oddalonych przeprowadzono na trzech zbiorach danych:

- *hbk*, który przedstawiono w pracy [Rousseeuw, Leroy 2009]. Jest to zbiór danych generowanych komputerowo, zawierający 75 obserwacji, z czego 14 ma wartości odstające o charakterze wpływowym<sup>1</sup>;
- *crime*, zaproponowany w pracy [Agresti, Finlay 2009]. Jest to zbiór danych rzeczywistych przedstawiających przestępczość w poszczególnych stanach USA (51 obserwacji). Zawiera trzy obserwacje odstające – wpływowe o dużych resztach;
- *Friedman 2* – zbiór danych generowanych komputerowo, stworzony przez Friedmana i opisany w pracy [Friedman 1991]. W tym przypadku wykorzystano zbiór 500 obserwacji, w którym 5% losowo wybranych wartości zmiennej  $Y$  zostało zakłóconych tak, by wykraczały poza typowy obszar zmienności znany chociażby z wykresów pudełkowych  $\langle Q_1 - 3Q, Q_3 + 3Q \rangle$  (gdzie  $Q_1$  i  $Q_3$  to pierwszy i trzeci kwartył, zaś  $Q$  – odchylenie ćwiartkowe).

Dla każdego zbioru wykonano analizę w dwóch wariantach. Zbudowano model na zbiorze z wartościami oddalonymi, a następnie obserwacje te usunięto i zbudowano nowy model. W każdym przypadku (dla każdego zbioru, dla każdej metody regresji) obliczono, metodą sprawdzania krzyżowego, błąd średniokwadratowy. Wyniki przedstawiono w tab. 1.

<sup>1</sup> Szczegółowo typy obserwacji oddalonych zostały przedstawione w pracy [Trzęsiok 2014].

**Tabela 1.** Wartości błędów średniokwadratowych  $MSE_{CV}$ , obliczone dla różnych modeli regresji zbudowanych na zbiorach danych z obserwacjami oddalonymi oraz po ich usunięciu

Metody	Zbiory danych					
	<i>Crime</i>	<i>crime</i> bez obs. oddalonych	<i>hbk</i>	<i>hbk</i> bez obs. oddalonych	<i>Friedman 2</i> z obs. oddalonymi	<i>Friedman 2</i>
1	2	3	4	5	6	7
PPR	78 236	31 311	2,72	0,29	24 717	18 483
POLYMARS	109 334	29 628	1,74	0,33	16 180	15 319
MART	95 359	22 888	1,00	0,26	23 037	16 820
R.FORESTS	61 893	21 669	0,81	0,22	27 645	17 957
RIDGE	54 115	32 690	4,74	0,31	37 299	35 652

Źródło: opracowanie własne.

Analizując wyniki przedstawione w tab. 1 dla poszczególnych metod, należy porównywać parami wartości  $MSE_{CV}$  otrzymane dla modeli zbudowanych:

- na zbiorze, w którym występowały wartości oddalone,
- oraz na zbiorze bez tych wartości nietypowych.

Nie jest tutaj ważne, dla jakiego modelu otrzymujemy najmniejsze wartości  $MSE_{CV}$ , tylko jak te wielkości (w odpowiednich parach) zmieniają się po usunięciu obserwacji nietypowych. Porównując liczby w kolumnach 2 i 3, 4 i 5 oraz 6 i 7 tab. 1, można zauważyć, że w każdym przypadku nastąpił stosunkowo duży spadek wartości błędu średniokwadratowego, co oznacza, że żadna z badanych metod nie jest odporna na występowanie w zbiorze uczącym wartości oddalonych.

## 4.2. Odporność na zaburzenia wartości zmiennej zależnej szumem

W tym przypadku badanie odporności metod regresji na losowe zakłócenia wartości zmiennej  $Y$  przeprowadzono na zbiorach danych:

- *Friedman 1*, który został utworzony jako realizacje dziesięciu niezależnych zmiennych objaśniających, o rozkładzie jednostajnym na przedziale  $[0,1]$ , oraz zmiennej  $Y$ , która zależy tylko od pięciu z nich i wyznaczona jest według wzoru

$$y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0,5)^2 + 10x_4 + 5x_5 + \varepsilon; \quad (1)$$

- *Friedman 2*, którego obserwacje to realizacje czterech zmiennych niezależnych na przedziałach:

$$0 \leq x_1 \leq 100, \quad 4\pi \leq x_2 \leq 560\pi, \quad 0 \leq x_3 \leq 1, \quad 1 \leq x_4 \leq 11,$$

zaś zmienna  $Y$  zadana jest wzorem

$$y = \sqrt{x_1^2 + \left(x_2 x_3 - \frac{1}{x_2 x_4}\right)^2} + \varepsilon; \quad (2)$$

- *Friedman 3*, który również jest realizacjami tych samych zmiennych objaśniających jak dla zbioru *Friedman 2* oraz zmiennej  $Y$  wyznaczonej wzorem

$$y = \arctg \frac{\left( x_2 x_3 - \frac{1}{x_2 x_4} \right)}{x_1} + \varepsilon . \quad (3)$$

W każdym przypadku  $\varepsilon$  jest zmienną zakłócającą (nazywaną również szumem gaussowskim) o rozkładzie normalnym  $N(0, \sigma)$ .

Powyższe zbiory Friedman w pracy [Friedman 1991] skonstruował tak, że badacz do wartości zmiennej zależnej może dodać szum gaussowski, regulując jego poziom przez odpowiednie dobieranie parametru  $\sigma$ .

Do przeprowadzonej analizy przygotowano zbiory uczące w *pięciu wariantach* (każdy o liczebności 500 obserwacji) o wzrastającym poziomie zakłóceń wartości zmiennej zależnej – od 0% do 40% zmienności mierzonej wariancją. Na tak stworzonych zbiorach zbudowano modele regresji, a następnie wykorzystując zbiory testowe (zawierające 1000 obserwacji, bez zakłóceń wartości zmiennej zależnej), obliczono wartości błędu średniokwadratowego. Otrzymane wyniki przedstawiono w tab. 2, 3 i 4.

**Tabela 2.** Wartości błędów średniokwadratowych  $MSE$  obliczone dla różnych modeli regresji zbudowanych na zbiorze *Friedman 1*, w którym wartości zmiennej zależnej zostały zaburzone szumem na różnym poziomie

Metody	Poziom szumu				
	0%	10%	20%	30%	40%
PPR	0,17	2,65	6,07	6,30	6,81
POLYMARS	2,50	2,96	3,18	3,16	4,73
MART	1,36	1,51	2,01	1,98	2,60
R.FORESTS	3,66	3,78	4,07	3,69	4,70
RIDGE	6,19	6,42	6,23	6,43	6,57

Źródło: opracowanie własne.

**Tabela 3.** Wartości błędów średniokwadratowych  $MSE$  obliczone dla różnych modeli regresji zbudowanych na zbiorze *Friedman 2*, w którym wartości zmiennej zależnej zostały zaburzone szumem na różnym poziomie

Metody	Poziom szumu				
	0%	10%	20%	30%	40%
PPR	4,5	5 151,2	6 871,6	5 645,1	6 768,5
POLYMARS	30,1	211,9	179,6	278,8	605,2
MART	621,5	2 157,3	3 702,0	4 061,1	5 245,8
R.FORESTS	589,9	3 364,8	5 376,5	6 718,5	10 535,8
RIDGE	19 154,2	19 972,9	19 471,4	19 853,4	20 800,6

Źródło: opracowanie własne.

**Tabela 4.** Wartości błędów średniokwadratowych *MSE* obliczone dla różnych modeli regresji zbudowanych na zbiorze *Friedman 3*, w którym wartości zmiennej zależnej zostały zaburzone szumem na różnym poziomie

Metody	Poziom szumu				
	0%	10%	20%	30%	40%
PPR	0,003	0,006	0,016	0,010	0,022
POLYMARS	0,004	0,005	0,006	0,014	0,009
MART	0,006	0,006	0,009	0,009	0,013
R.FORESTS	0,007	0,008	0,010	0,011	0,018
RIDGE	0,040	0,041	0,041	0,041	0,042

Źródło: opracowanie własne.

Analizując kolejno poszczególne wiersze tab. 2-4, można zauważyć, że wraz ze wzrostem zakłóceń zmiennej *Y* następuje na ogół stosunkowo niewielki wzrost wartości błędów średniokwadratowych. Jedynie w przypadku zbioru *Friedman 2* (tab. 3) obserwujemy dosyć duży „przeskok” wartości od braku zakłóceń *Y* do szumu na poziomie 10%. Największy wzrost wartości *MSE* widzimy dla metody PPR, której niestety nie można uznać za odporną w tym przypadku. Pozostałe badane metody dla analizowanych zbiorów danych wykazują się odpornością na losowe zaburzenia szumem gaussowskim wartości zmiennej zależnej.

### 4.3. Odporność na występowanie zmiennych nieistotnych

Badając odporność metod regresji na występowanie w zbiorze uczącym zmiennych nieistotnych, wykorzystano ponownie zbiór *Friedman 1*, lecz tym razem w dwóch wariantach:

- ze wszystkimi 10 zmiennymi objaśniającymi,
- tylko z 5 pierwszymi zmiennymi objaśniającymi, które mają istotny wpływ na zmienną zależną (zob. wzór (1)).

**Tabela 5.** Wartości błędów średniokwadratowych *MSE* obliczone dla różnych modeli regresji zbudowanych na zbiorze *Friedman 1* ze wszystkimi oraz tylko istotnymi zmiennymi objaśniającymi

Metody	Zbiory danych	
	<i>Friedman 1</i> z 5 zmiennymi	<i>Friedman 1</i> z 10 zmiennymi
PPR	1,30	1,78
POLYMARS	1,15	1,19
MART	2,21	2,36
R.FORESTS	3,90	4,51
RIDGE	7,15	7,28

Źródło: opracowanie własne.

W obu zbiorach uczących było po 500 obserwacji zakłóconych szumem na poziomie 10%. Dla zbudowanych modeli regresji obliczono ponownie wartości błędu średniokwadratowego na zbiorach testowych, złożonych z 1000 elementów, bez zakłóceń zmiennej  $Y$ . Otrzymane wyniki prezentuje tab. 5.

Porównując parami wartości  $MSE$  w kolejnych wierszach tab. 5, można powiedzieć, iż nie następuje znaczny ich wzrost. Oznacza to, że badane metody charakteryzują się odpornością w tym ostatnim aspekcie, którym jest występowanie w zbiorze uczącym zmiennych nieistotnych.

## 5. Zakończenie

W artykule przedstawiono wybrane podejścia do problemu odporności regresji, przy czym w analizie skupiono się na badaniu wrażliwości modeli regresji na występowanie w zbiorze uczącym: wartości oddalonych, losowych zakłóceń wartości zmiennej zależnej oraz zmiennych nieistotnych.

Wyniki analiz, przeprowadzonych metodami symulacyjnymi na zbiorach danych wykorzystywanych do badania własności nieklasycznej regresji, pokazują, że:

- żadna z przedstawionych metod nie jest odporna na występowanie obserwacji oddalonych, a usunięcie tych obserwacji ze zbioru uczącego może znacznie poprawić dokładność predykcji uzyskanego modelu;
- badane metody, poza metodą rzutowania PPR, są odporne na zakłócenia wartości zmiennej zależnej szumem o rozkładzie normalnym;
- wszystkie omawiane metody są odporne na występowanie w zbiorze uczącym zmiennych nieistotnych.

## Literatura

- Agresti A., Finlay B., 2009, *Statistical Methods for the Social Sciences*, 4th ed., Pearson.
- Breiman L., 2001, *Random forests*, Machine Learning, no. 45, s. 5-32.
- Friedman J., 1991, *Multivariate adaptive regression splines*, The Annals of Statistics, vol. 1, no. 19, s. 1-67.
- Friedman J., 1999a, *Greedy Function Approximation: a Gradient Boosting Machine*, Technical Report, Stanford University, Dept. of Statistics.
- Friedman J., 1999b, *Stochastic Gradient Boosting*, Technical Report, Stanford University, Dept. of Statistics.
- Friedman J., Stuetzle W., 1981, *Projection pursuit regression*, Journal of the American Statistical Association, no. 76, s. 817-823.
- Hoerl A.E., Kennard R.W., 1970, *Ridge regression: applications to nonorthogonal problems*, Technometrics, no. 12, s. 69-82.
- Kooperberg C., Bose S., Stone C., 1997, *Polychotomous regression*, Journal of the American Statistical Association, no. 92, s. 117-127.
- Meyer D., Leisch F., Hornik K., 2003, *The support vector machine under test*, Neurocomputing, vol. 1-2, no. 55, s. 169-186.

- Rousseeuw P., Leroy A., 2003, *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc.
- Trzęsiok J., 2011, *Przegląd metod regularyzacji w zagadnieniach regresji nieparametrycznej*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Jajuga K., Walesiak M. (red.), *Taksonomia 18. Klasyfikacja i analiza danych*, s. 330-339.
- Trzęsiok M., 2014, *Wybrane metody identyfikacji obserwacji oddalonych*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 327, Jajuga K., Walesiak M. (red.), *Taksonomia 22. Klasyfikacja i analiza danych – teoria i zastosowania*, s. 157-166.
- Trzpiot G. (red.), 2013, *Wybrane elementy statystyki odpornej*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

## NON-CLASSICAL REGRESSION METHODS VS. ROBUSTNESS

**Summary:** The paper presents an important problem of robustness in regression. Various aspects of the approach to this problem are discussed, but the paper focuses on the sensitivity of the model to outliers, noise of the values of dependent variable and to the presence of insignificant variables. The study assesses the robustness of the following methods: PPR, POLYMARS, RANDOM FORESTS, MART and ridge regression.

**Keywords:** regression analysis, robustness, nonparametric regression.