

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka, Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska, Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński, Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz, Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel, Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień, Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska, Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan, Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska, Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka, Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański, Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk, Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk, Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura, Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk, Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski, Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka, Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk, Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarz , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Andrzej Bąk

Uniwersytet Ekonomiczny we Wrocławiu

MODELE KATEGORII NIEUPORZĄDKOWANYCH W BADANIACH PREFERENCJI

Streszczenie: Wśród mikroekonometrycznych modeli kategorii nieuporządkowanych wyróżnia się najczęściej wielomianowy model logitowy, warunkowy model logitowy i mieszany model logitowy. Podstawę rozróżnienia tych typów modeli stanowi głównie charakter zmiennych objaśniających uwzględnionych w modelu. Rozróżnienie to nie jest jednak jednoznacznie interpretowane. Celem artykułu jest wskazanie podstawowych różnic między typami modeli logitowych oraz prezentacja przykładów estymacji różnych typów tych modeli dla różnych typów danych z wykorzystaniem programu R.

Słowa kluczowe: preferencje, modele kategorii nieuporządkowanych, program R.

1. Wstęp

W badaniach preferencji zmierza się do identyfikacji czynników, którymi kierują się konsumenci, wybierając określone produkty lub usługi. Czynniki te są związane zarówno z cechami konsumentów, jak i z charakterystykami produktów lub usług. Empiryczne badania preferencji opierają się na teorii addytywnego jednoczesnego pomiaru łącznego [Coombs, Dawes, Tversky 1977, s. 50 i nast.] oraz teorii użyteczności losowej [Coombs, Dawes, Tversky 1977, s. 214 i nast.], które powstały na gruncie psychologii matematycznej i psychometrii. Teoria użyteczności losowej stanowi podstawę teoretyczną probabilistycznych modeli wyborów dyskretnych wykorzystywanych w badaniach preferencji, do których należą modele kategorii nieuporządkowanych [Greene 2008, s. 840-847].

Wśród mikroekonometrycznych modeli kategorii nieuporządkowanych wyróżnia się najczęściej wielomianowy model logitowy, warunkowy model logitowy [McFadden 1974] i mieszany model logitowy [Winkelmann, Boes 2006]. Podstawę rozróżnienia tych typów modeli stanowi głównie charakter zmiennych objaśniających. Rozróżnienie to nie jest jednak jednoznacznie interpretowane. Celem artykułu jest wskazanie podstawowych różnic między typami modeli logitowych oraz prezentacja przykładów estymacji różnych typów tych modeli dla różnych typów danych w badaniach preferencji z wykorzystaniem programu R. Przede wszystkim w artykule przedstawiono następujące zagadnienia: charakterystykę wybranych typów modeli

logitowych, interpretację zmiennych objaśniających w różnych typach tych modeli, organizację danych wykorzystywanych w estymacji parametrów różnych typów modeli, funkcje dostępne w programie R, które mogą znaleźć zastosowanie w estymacji modeli logitowych, funkcje napisane w języku programowania R, które zostały wykorzystane w obliczeniach.

2. Modele logitowe kategorii nieuporządkowanych¹

W badaniach preferencji wykorzystujących metody wyborów dyskretnych szczególnie ważną rolę odgrywają wielomianowe i warunkowe modele logitowe oraz ich połączenie w postaci tzw. mieszanych modeli logitowych [Cameron, Trivedi 2009, s. 500], nazywanych również hybrydowymi modelami logitowymi [Winkelmann, Boes 2006, s. 154]. Modele te mieszczą się w grupie wielomianowych modeli kategorii nieuporządkowanych.

Wielomianowy model logitowy jest uogólnieniem modelu logitowego dla danych binarnych (regresji logistycznej) i może być stosowany wówczas, kiedy zmienna objaśniana przyjmuje w sposób dyskretny wartości ze zbioru liczącego więcej niż dwie kategorie, których kolejność nie jest istotna. Model ten wywodzi się z teorii użyteczności losowej oraz tzw. aksjomatu wyboru Luce'a (modelu stałej użyteczności) [Coombs, Dawes, Tversky 1977, s. 217 i nast.]. Wielomianowy model logitowy można przedstawić w postaci [So, Kuhfeld 1995; Long 1997, s. 151 i nast.; Powers, Xie 2008, s. 243 i nast.; Cameron, Trivedi 2009, s. 500; Gruszczyński 2010, s. 161]]:

$$P_{ki} = \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}_i)}{\sum_{l=1}^n \exp(\mathbf{x}_k^T \boldsymbol{\beta}_l)}, \text{ przy czym } \boldsymbol{\beta}_n = \mathbf{0}, \quad (1)$$

gdzie: P_{ki} – prawdopodobieństwo wyboru i -tej kategorii przy k -tym stanie zmiennych objaśniających;

\mathbf{x}_k^T – wektor reprezentujący k -ty wiersz macierzy \mathbf{X} (zmiennych objaśniających);

$\boldsymbol{\beta}_i$ – wektor szacowanych parametrów związany z i -tą kategorią zmiennej objaśnianej.

Oszacowane wartości prawdopodobieństw w modelu (1) sumują się do 1 w obrębie każdej konfiguracji zmiennych objaśniających. Rozkład prawdopodobieństw spełniających ten warunek można jednak uzyskać przy różnych wartościach parametrów $\boldsymbol{\beta}_i$, a zatem model pozostaje zdefiniowany niejednoznacznie. W celu rozwiązania tego problemu przyjmuje się pewne ograniczenia dotyczące wektora parametrów nazywane normalizacją, zakładając np. $\boldsymbol{\beta}_n = \mathbf{0}$. Oznacza to jednocześnie, że jedna

¹ Punkt opracowany na podstawie pracy [Bąk 2010].

z opcji wyboru (np. ostatnia) stanowi profil odniesienia, a pozostałe profile są różne od tej opcji [Agresti 2002, s. 268].

Macierz \mathbf{X} w modelu (1) zawiera charakterystyki respondentów, których preferencje dotyczące produktów lub usług są przedmiotem badań. Charakterystyki respondentów są stałe względem tych produktów lub usług.

Warunkowy model logitowy został zaproponowany przez McFaddena [1974] jako uogólnienie wielomianowego modelu logitowego. Podstawowym kryterium rozróżniania tych modeli jest charakter zmiennych objaśniających, tzn. macierzy \mathbf{X} w równaniu (1). Jeżeli zmienne objaśniające charakteryzują konsumentów, to na ogół wykorzystuje się wielomianowy model logitowy. Jeśli natomiast zmienne objaśniające opisują obiekty będące przedmiotem wyboru (produkty lub usługi), to z reguły stosuje się warunkowy model logitowy.

W warunkowym modelu logitowym prawdopodobieństwo wyboru i -tego profilu ze zbioru liczącego n elementów jest szacowane na podstawie zależności [So, Kuhfeld 1995, s. 7; Long 1997, s. 178 i nast.; Powers, Xie 2008, s. 256 i nast.; Cameron, Trivedi 2009, s. 500; Gruszczynski 2010, s. 172-173]:

$$P_{ki} = \frac{\exp(\mathbf{z}_{ki}^T \boldsymbol{\alpha})}{\sum_{l=1}^n \exp(\mathbf{z}_{kl}^T \boldsymbol{\alpha})}, \quad (2)$$

gdzie: \mathbf{z}_{ki}^T – k -ty wektor macierzy \mathbf{Z} (zmiennych objaśniających) opisujący i -tą opcję;

$\boldsymbol{\alpha}$ – wektor parametrów (wartość α_j jest związana z j -tą zmienną objaśniającą).

Macierz \mathbf{Z} w modelu (2) zawiera charakterystyki produktów lub usług, względem których badane są preferencje respondentów. Wartości zmiennych objaśniających opisujących produkty lub usługi są specyficzne w przekroju opcji wyboru oferowanych respondentom (np. w badaniu ankietowym).

Mieszany model logitowy (3) jest połączeniem modeli (1) i (2), a więc uwzględnia charakterystyki zarówno respondentów, jak i opcji wyboru (produktów lub usług) [So, Kuhfeld 1995; Powers, Xie 2008, s. 258 i nast.; Cameron, Trivedi 2009, s. 500]:

$$P_{ki} = \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}_i + \mathbf{z}_{ki}^T \boldsymbol{\alpha})}{\sum_{l=1}^n \exp(\mathbf{x}_k^T \boldsymbol{\beta}_l + \mathbf{z}_{kl}^T \boldsymbol{\alpha})}. \quad (3)$$

3. Estymacja parametrów wielomianowych modeli logitowych w programie R

Szacowanie parametrów wielomianowych, warunkowych i mieszanych modeli logitowych w programie R można przeprowadzić z wykorzystaniem funkcji `optim()` z pakietu `stats`. Funkcję `optim` można wykorzystać do maksymalizacji funkcji największej wiarygodności, która umożliwia znalezienie najlepszego dopasowania modelu logitowego danych empirycznych (zob. [Jackman 2007]). Kryterium tego dopasowania jest wartość funkcji wiarygodności. Funkcja `optim()` korzysta z iteracyjnych algorytmów optymalizacji: sympleksu (Nelder-Meada), zmiennej metryki (*quasi*-Newtona, Broydena-Fletcher-Goldfarba-Shannona), gradientów sprzężonych (Fletcher-Reevesa), *quasi*-Newtona z ograniczeniami (algorytm L-BFGS-B), sieci neuronowych (SNN – *Simulated Neural Network*).

Wybrane argumenty funkcji `optim()` są następujące [R Development Core Team 2011]:

```
optim(par, fn, gr=NULL, method=c("Nelder-Mead",
"BFGS", "CG", "L-BFGS-B", "SANN"), control=list())
```

<code>par</code>	wartości początkowe szacowanych parametrów,
<code>fn</code>	funkcja, której wartość jest optymalizowana,
<code>gr</code>	gradient (wektor pochodnych cząstkowych),
<code>method</code>	algorytm optymalizacji,
<code>control</code>	parametry kontrolne.

Funkcja `optim()` domyślnie służy do minimalizacji funkcji największej wiarygodności, ale parametr kontrolny `control=list(fnscale=-1)` umożliwia zmianę kierunku optymalizacji w celu maksymalizacji wartości tej funkcji.

Funkcja największej wiarygodności (`fnw()`) przekazywana jako argument `fn` do funkcji `optim()` w celu oszacowania parametrów modeli logitowych ma postać:

```
# fnw(a,x,y) - funkcja największej wiarygodności dla modelu logitowego
# na podstawie: S. Jackman [2007]
# a - wartości startowe parametrów modelu
# x - wartości zmiennych objaśniających
# y - zmienna objaśniana o wartościach TRUE/FALSE (wybór/brak wyboru)
# wywołanie - jako parametr funkcji optim{stats}
fnw<-function(a,x,y) {
  mu<-x%*%a #składnik systematyczny modelu - użyteczność opcji
  eta<-exp(mu) #licznik
  suma<-tapply(eta,s,sum) #mianownik
  pr<-eta[y]/suma #prawdopodobieństwa wyboru opcji
  lnw<-sum(log(pr)) #suma logarytmów prawdopodobieństw
  return(lnw) #logarytm fnw }
```

4. Wykorzystanie funkcji `optim()` w estymacji modeli logitowych

W przykładach szacowania modeli logitowych (wielomianowego, warunkowego i mieszanego) wykorzystano dane (zbiór o nazwie `travel`) z pracy [So, Kuhfeld 1995] opisujące wybór środka podróży w zależności od czasu podróży (`travtime`) – jest to zmienna specyficzna dla opcji wyboru o poziomach `autotime`, `plantime` i `trantime`. Do wyboru jest jedna z trzech opcji: samochód (`auto`), samolot (`plane`) lub przewóz publiczny autobusem lub pociągiem (`transit`). Zawartość zbioru `travel` zapisana w pliku `travel.csv` jest następująca:

	<code>autotime</code>	<code>plantime</code>	<code>trantime</code>	<code>age</code>	<code>chosen</code>
1	10.0	4.5	10.5	32	2
2	5.5	4.0	7.5	13	1
3	4.5	6.0	5.5	41	3
4	3.5	2.0	5.0	41	3
5	1.5	4.5	4.0	47	1
6	10.5	3.0	10.5	24	2
7	7.0	3.0	9.0	27	1
8	9.0	3.5	9.0	21	2
9	4.0	5.0	5.5	23	1
10	22.0	4.5	22.5	30	2
11	7.5	5.5	10.0	58	2
12	11.5	3.5	11.5	36	3
13	3.5	4.5	4.5	43	1
14	12.0	3.0	11.0	33	2
15	18.0	5.5	20.0	30	2
16	23.0	5.5	21.5	28	2
17	4.0	3.0	4.5	44	2
18	5.0	2.5	7.0	37	3
19	3.5	2.0	7.0	45	1
20	12.5	3.5	15.5	35	2
21	1.5	4.0	2.0	22	1.

W strukturze tych danych każdy wiersz przedstawia zbiór (sytuację wyboru), z którego respondent wybrał jedną z trzech opcji (profilów). Zmienną specyficzną dla respondentów jest wiek (`age`), liczba respondentów wynosi 21, natomiast wybrany środek podróży reprezentuje zmienna `chosen` (o wartościach 1 – `auto`, 2 – `plane`, 3 – `transit`).

W **warunkowym modelu logitowym** zmienne objaśniające charakteryzują przedmiot wyboru (profile tworzące zbiór, z którego respondent dokonuje wyboru), a więc są specyficzne dla opcji wyboru. Oszacowanie takiego modelu wymaga prze-

kształcenia danych z pliku `travel.csv` do postaci zapisanej w pliku `travel_2.csv` (10 pierwszych wierszy)²:

subject	option	choice	travtime	age
[1,]	1	1	0	10.0
[2,]	1	2	1	4.5
[3,]	1	3	0	10.5
[4,]	2	1	1	5.5
[5,]	2	2	0	4.0
[6,]	2	3	0	7.5
[7,]	3	1	0	4.5
[8,]	3	2	0	6.0
[9,]	3	3	1	5.5
[10,]	4	1	0	3.5

W strukturze tych danych każdy wiersz przedstawia jedną opcję wyboru (jeden profil), zmienna `subject` reprezentuje numer respondenta, zmienna `option` numer opcji wyboru (1 – auto, 2 – plane, 3 – transit), zmienna objaśniana `choice` wskazuje wybraną opcję (1 – wybrana opcja, 0 – niewybrana opcja), zmienna `travtime` jest specyficzna dla opcji wyboru i reprezentuje czas podróży, a zmienna `age` jest specyficzna dla respondentów i reprezentuje ich wiek (wartość tej zmiennej powtarza się trzykrotnie, ponieważ są trzy opcje wyboru). Liczba wierszy w tym zbiorze danych wynosi 63 (21 respondentów, 3 opcje wyboru).

Skrypt 1 wykorzystuje funkcję największej wiarygodności `fnw()` do estymacji warunkowego modelu logitowego za pomocą funkcji `optim()`.

Skrypt 1.

```
source("fnw.r")
dane<-read.csv2("travel_2.csv", header=TRUE)
head(dane, 6)
attach(dane)
s<-subject #identyfikator respondenta
p<-option #identyfikator opcji
w<-choice==1 #identyfikator wyboru TRUE/FALSE
y<-p[w] #numery wybranych opcji
X<-as.matrix(cbind(travtime)) #macierz danych
k<-dim(X)[2] #liczba zmiennych objaśniających
a<-rep(0, k) #wartości startowe parametrów modelu
nazwy<-colnames(X)
clm<-optim(par=a, fn=fnw, x=X, y=w, control=list(trace=TRUE, fnscale=-1),
method="BFGS", hessian=TRUE)
se<-sqrt(diag(solve(-clm$hessian))) #standardowe błędy parametrów
B<-clm$par #parametry
```

² Ze względu na ograniczoną objętość artykułu nie zamieszczono procedur i funkcji wykonujących przekształcenia danych.


```
wyniki<-cbind(B, se)
wyniki<-cbind(wyniki, wyniki[,1]/wyniki[,2], exp(wyniki[,1]))
dimnames(wyniki)<-list(nazwy, c("B", "se", "Z", "exp(B)"))
print(signif(wyniki, 4)).
```

W wyniku wykonania skryptu 1 otrzymuje się oszacowanie parametru (B) dla zmiennej specyficznej dla opcji wyboru (travtime):

```
          B          se          Z exp(B)
travtime -0.2655 0.1021 -2.599 0.7668.
```

W wielomianowym modelu logitowym zmienne objaśniające charakteryzują respondentów, a więc są specyficzne dla podmiotu dokonującego wyboru. Oszacowanie takiego modelu za pomocą funkcji `optim()` wymaga przekształcenia danych z pliku `travel_2.csv` do postaci zapisanej w pliku `travel_3.csv` (10 pierwszych wierszy):

	subject	option	choice	travtime	auto	plane	ageauto	ageplane
[1,]	1	1	0	10.0	1	0	32	0
[2,]	1	2	1	4.5	0	1	0	32
[3,]	1	3	0	10.5	0	0	0	0
[4,]	2	1	1	5.5	1	0	13	0
[5,]	2	2	0	4.0	0	1	0	13
[6,]	2	3	0	7.5	0	0	0	0
[7,]	3	1	0	4.5	1	0	41	0
[8,]	3	2	0	6.0	0	1	0	41
[9,]	3	3	1	5.5	0	0	0	0
[10,]	4	1	0	3.5	1	0	41	0.

W strukturze tych danych każdy wiersz przedstawia jedną opcję wyboru. Trzy wiersze tworzą zbiór, z którego respondent wybrał jedną opcję (profil). Opcje wyboru są w tym zbiorze reprezentowane przez zmienne zero-jedynkowe (`auto`, `plane`), które zostały pomnożone przez zmienną specyficzną dla respondentów (`age`). W wyniku takiej interakcji powstały zmienne `ageauto` i `ageplane`, które reprezentują zmienną specyficzną dla respondentów `age` i są jednocześnie specyficzne dla opcji wyboru. Profilem odniesienia jest opcja trzecia (`transit`), pominięta w tym zbiorze.

Skrypt 2 wykorzystuje funkcję największej wiarygodności `fnw()` do estymacji wielomianowego modelu logitowego za pomocą funkcji `optim()`.

Skrypt 2.

```
source("fnw.r")
dane<-read.csv2("travel_3.csv", header=TRUE)
head(dane, 6)
attach(dane)
s<-subject      #identyfikator respondenta
p<-option       #identyfikator opcji
```

```
w<-choice==1 #identyfikator wyboru TRUE/FALSE
y<-p[w] #numery wybranych opcji
X<-as.matrix(cbind(auto,plane,ageauto,ageplane)) #macierz danych
k<-dim(X)[2] #liczba zmiennych objaśniających
a<-rep(0,k) #wartości startowe parametrów modelu
nazwy<-colnames(X)
clm<-optim(par=a,fn=fnw,x=X,y=w,control=list(trace=TRUE,fnscale=-1),
method="BFGS",hessian=TRUE)
se<-sqrt(diag(solve(-clm$hessian))) #standardowe błędy parametrów
B<-clm$par #parametry
wyniki<-cbind(B,se)
wyniki<-cbind(wyniki,wyniki[,1]/wyniki[,2],exp(wyniki[,1]))
dimnames(wyniki)<-list(nazwy,c("B","se","Z","exp(B)"))
print(signif(wyniki,4)).
```

W wyniku wykonania skryptu 2 otrzymuje się oszacowania parametrów (B) dla zmiennej specyficznej dla respondentów (age):

	B	se	Z	exp(B)
auto	3.04900	2.42700	1.256	21.0800
plane	2.72400	2.29400	1.187	15.2400
ageauto	-0.07106	0.06518	-1.090	0.9314
ageplane	-0.05007	0.05961	-0.840	0.9512.

W tym modelu są dwa wyrazy wolne (auto i plane) i dwa parametry dla zmiennej age (ageauto i ageplane), które reprezentują odpowiednio efekt wpływu na prawdopodobieństwo wyboru samochodu w odniesieniu do przewozu publicznego i samolotu w odniesieniu do przewozu publicznego.

W **mieszanym (hybrydowym) modelu logitowym** występują zmienne objaśniające charakteryzujące zarówno respondentów, jak i opcje wyboru. Struktura zbioru danych jest taka sama jak w przypadku wielomianowego modelu logitowego (travel_3.csv).

Skrypt 3 wykorzystuje funkcję największej wiarygodności fnw() do estymacji mieszanego modelu logitowego za pomocą funkcji optim().

Skrypt 3.

```
source("fnw.r")
dane<-read.csv2("travel_3.csv", header=TRUE)
head(dane,6)
attach(dane)
s<-subject #identyfikator respondenta
p<-option #identyfikator opcji
w<-choice==1 #identyfikator wyboru TRUE/FALSE
y<-p[w] #numery wybranych opcji
X<-as.matrix(cbind(auto,plane,ageauto,ageplane,traveltime)) #macierz
danych
k<-dim(X)[2] #liczba zmiennych objaśniających
a<-rep(0,k) #wartości startowe parametrów modelu
nazwy<-colnames(X)
```

```

clm<-optim(par=a, fn=fnw, x=X, y=w, control=list(trace=TRUE, fnscale=-1),
method="BFGS", hessian=TRUE)
se<-sqrt(diag(solve(-clm$hessian))) #standardowe błędy parametrów
B<-clm$par #parametry
wyniki<-cbind(B, se) #parametry i błędy
wyniki<-cbind(wyniki, wyniki[,1]/wyniki[,2], exp(wyniki[,1]))
dimnames(wyniki)<-list(nazwy, c("B", "se", "Z", "exp(B)"))
print(signif(wyniki, 4)).

```

W wyniku wykonania skryptu 3 otrzymuje się oszacowania parametrów (B) dla zmiennych specyficznych dla respondentów i opcji wyboru:

	B	se	Z	exp(B)
auto	2.50100	2.39600	1.0440	12.20000
plane	-2.78000	3.53000	-0.7873	0.06206
ageauto	-0.07828	0.06333	-1.2360	0.92470
ageplane	0.01696	0.07442	0.2279	1.01700
travtime	-0.60850	0.27130	-2.2430	0.54420.

Oszacowane parametry mieszanego modelu można interpretować w kategoriach prawdopodobieństwa [Gruszczyński 2002]. Jeżeli czas podróży wzrasta (ujemna wartość parametru przy zmiennej *travtime*), to zmniejsza się prawdopodobieństwo wyboru takiej opcji. Wraz z rosnącym wiekiem respondentów będzie malało prawdopodobieństwo wyboru samochodu (ujemna wartość parametru przy zmiennej *ageauto*) w odniesieniu do prawdopodobieństwa wyboru przewozu publicznego, natomiast będzie wzrastać prawdopodobieństwo wyboru samolotu (dodatnia wartość parametru przy zmiennej *ageplane*) w odniesieniu do prawdopodobieństwa wyboru przewozu publicznego.

5. Podsumowanie

Metoda największej wiarygodności umożliwia estymację różnych typów modeli kategorii nieuporządkowanych. Oferowana w programie R funkcja `optim()` z pakietu `stats` wymaga opracowania funkcji największej wiarygodności dla modeli logitowych, którą można wykorzystać w takiej samej postaci do estymacji warunkowego, wielomianowego i mieszanego modelu logitowego.

W procedurach estymacji modeli kategorii nieuporządkowanych ważną rolę odgrywa struktura danych empirycznych. Szacowanie różnych typów tych modeli wymaga przekształcania danych do odpowiednich formatów. W tym celu warto kontynuować prace zmierzające do opracowania uniwersalnych procedur transformacji danych do pożądanej postaci w zależności od szacowanego modelu.

Literatura

- Agresti A., *Categorical Data Analysis*, Second Edition, Wiley, New York 2002.
- Bąk A., *Analiza danych o preferencjach z wykorzystaniem mikroekonometrycznych modeli kategorii nieuporządkowanych i programu R*, [w:] K. Jajuga, M. Walesiak, *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.
- Cameron A.C., Trivedi P.K., *Microeconometrics. Methods and Applications*, Cambridge University Press, New York 2009.
- Coombs C.H., Dawes R.M., Tversky A., *Wprowadzenie do psychologii matematycznej*, PWN, Warszawa 1977.
- Greene W.H., *Econometric Analysis*, 6th ed., Prentice Hall, Upper Saddle River, 2008.
- Gruszczyński M. (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Wolters Kluwer, Warszawa 2010.
- Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa 2002.
- Jackman S., *Models for Unordered Outcomes*, Political Science 150C/350C, 2007, <http://jackman.stanford.edu/classes/350C/07/unordered.pdf> (14.10.2011).
- Long J.S., *Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, Thousand Oaks-London-New Delhi 1997.
- McFadden D., *Conditional Logit Analysis of Qualitative Choice Behavior*, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, Academic Press, New York-San Francisco-London 1974.
- Powers D.A., Xie Y., *Statistical Methods for Categorical Data Analysis*, 2nd ed., Emerald, Bingley 2008.
- R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2011, <http://cran.r-project.org/>.
- So Y., Kuhfeld W.F., *Multinomial Logit Models*, 1995, <http://www.sascommunity.org/sugi/SUGI95/>, 14.10.2011.
- Winkelmann R., Boes S., *Analysis of Microdata*, Springer-Verlag, Berlin, Heidelberg 2006.

MODELS FOR UNORDERED CATEGORIES IN PREFERENCE ANALYSIS

Summary: Among microeconomic models for unordered categories multinomial logit model, conditional logit model and mixed logit model are most frequently mentioned. The character of the independent variables included in the model is mainly the basis for distinguishing among these types of models. This distinction is not clearly interpreted. The main aim of this article is to identify the fundamental differences among the types of logit models and to present the examples of estimation of various types of this models for various types of data using R program.

Keywords: preferences, models for unordered categories, R program.