

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
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
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

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

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

Summaries

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

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

Agnieszka Sompolska-Rzechuła

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

e-mail: agnieszka.sompolska-rzechula@zut.edu.pl

OKREŚLENIE CZYNNIKÓW WPŁYWAJĄCYCH NA PRAWDOPODOBIENSTWO POPRAWY POZIOMU ROZWOJU SPOŁECZNEGO Z WYKORZYSTANIEM MODELU LOGITOWEGO

Streszczenie: Zróżnicowanie poziomu rozwoju społecznego w Polsce na poziomie lokalnym jest znacznie większe niż pomiędzy województwami. W związku z tym istnieje konieczność monitorowania i oceny poziomu rozwoju społecznego w powiatach oraz gminach Polski. Właściwy pomiar wymaga wykorzystania odpowiednich mierników. Jednym z najczęściej wykorzystywanych mierników do oceny poziomu rozwoju społecznego jest *Human Development Index* (HDI), a na poziomie lokalnym *Local Human Development Index* (LHDI). Celem pracy jest identyfikacja predyktorów poprawy wartości Lokalnego Wskaźnika Rozwoju Społecznego w powiatach Polski. Do realizacji celu wykorzystano regresję logitową, za pomocą której oszacowano prawdopodobieństwo poprawy LHDI. Jako predyktory wykorzystano wskaźniki: zdrowia, edukacji i zamożności i wskaźniki nakładów: zdrowotnych, edukacyjnych i wydatków na poziomie lokalnym oraz rodzaj powiatu. Jakość otrzymanego modelu oceniono za pomocą: R^2 zliczeniowego, testu Hosmera-Lemeshowa oraz krzywej ROC.

Słowa kluczowe: Lokalny Wskaźnik Rozwoju Społecznego, regresja logitowa, iloraz szans.

DOI: 10.15611/pn.2015.385.26

1. Wstęp

Podstawowym narzędziem w diagnozowaniu poziomu zaspokojenia potrzeb ludności są wskaźniki społeczne. Przez długi czas miarą rozwoju społecznego był produkt krajowy brutto (PKB). Jednak coraz częściej podważane jest stosowanie PKB jako głównego wskaźnika rozwoju społecznego. Prace nad jego modyfikacją, w celu dostosowania do pomiaru rozwoju społecznego, dotyczą między innymi eliminacji tych jego elementów, które nie przyczyniają się, bezpośrednio lub pośrednio, do zaspokojenia szeroko pojmowanych potrzeb konsumpcyjnych, przy jednoczesnym włączeniu do obliczeń szeregu dodatkowych elementów mających istotny wpływ na poziom i jakość życia. Efektem tych prac jest stworzenie Wskaź-

nika Rozwoju Społecznego (*Human Development Index*, HDI), który łączy PKB z miarami z zakresu zdrowia oraz poziomu edukacji. Zróżnicowanie poziomu rozwoju społecznego w Polsce na poziomie lokalnym jest znacznie większe niż pomiędzy województwami. Do pomiaru lokalnego poziomu rozwoju społecznego opracowano Lokalny Wskaźnik Rozwoju Społecznego (*Local Human Development Index*, LHDI), który opiera się na istniejącym już wskaźniku HDI, służącym do badania tego zjawiska na poziomie krajowym.

Celem pracy jest identyfikacja predyktorów poprawy wartości LHDI w powiatach Polski. Do realizacji celu wykorzystano regresję logitową, za pomocą której oszacowano prawdopodobieństwo poprawy LHDI. Jako predyktory wykorzystano wskaźniki: zdrowia, edukacji i zamożności oraz wskaźniki nakładów: zdrowotnych, edukacyjnych oraz wydatków na poziomie lokalnym. Uwzględniono także rodzaj powiatu: grodzki i ziemski.

2. Materiał badawczy i metoda

Poziom rozwoju społecznego Polski jest znacznie zróżnicowany, a stopień zróżnicowania tego poziomu jest znacznie większy na poziomie lokalnym niż pomiędzy województwami. Jak wskazują cele sformułowane w polityce społecznej Unii Europejskiej oraz w ramach Krajowej Strategii Rozwoju Regionalnego 2010-2020, istnieje konieczność badania zjawisk i procesów społeczno-ekonomicznych na poziomie regionalnym i lokalnym¹. Dzięki działaniom Ministerstwa Rozwoju Regionalnego Polska została objęta nowatorskim badaniem dotyczącym mierzenia rozwoju społecznego kraju na poziomie lokalnym. Innowacyjny charakter projektu polega na wykorzystaniu LHDI. Dla jak najlepszego uwzględnienia wielowymiarowego charakteru rozwoju metodologia opracowanego wskaźnika obejmowała trzy komponenty: zdrowie, edukację oraz zamożność².

Podstawową jednostką analizy rozwoju lokalnego, ze względu na bliskość usług publicznych zdrowotnych, rynku pracy oraz edukacji na poziomie średnim i wyższym, jest powiat. Konstrukcja LHDI jest następująca:

$$LHDI_i = \sqrt[3]{HI_i \cdot EI_i \cdot WI_i}, \quad (1)$$

gdzie: $LHDI_i$ – wartość wskaźnika rozwoju społecznego dla i -tego powiatu, HI_i – wskaźnik zdrowia w i -tym powiecie, EI_i – wskaźnik edukacji w i -tym powiecie, WI_i – wskaźnik zamożności w i -tym powiecie, n – liczba powiatów.

¹ Więcej informacji można znaleźć na <http://www.mir.gov.pl>.

² Szczegółowe informacje na temat HDI i LHDI oraz sposobów liczenia wskaźnika i jego składowych znajdują się w Krajowym Raporcie o Rozwoju Społecznym dostępnym na www.mir.gov.pl. W raporcie zawarte są także koncepcje rozwoju społecznego ze szczególnym uwzględnieniem podejścia reprezentowanego przez A. Sena, w którym rozwój społeczny rozumiany jest jako proces zwielokrotniania możliwości wyborów, przed jakimi staje każdy człowiek, osiągnięty przez umożliwienie skutecznego działania oraz efektywnego wypełniania obowiązków.

Rozwój społeczny na poziomie lokalnym obejmuje trzy wymiary: zdrowie, edukację i zamożność.

Wskaźniki zdrowia i edukacji wyrażone są jako średnia geometryczna wskaźników cząstkowych. Wskaźnik zdrowia obejmuje: wskaźnik przeciętnego trwania życia (liczba lat, jaką ma do przeżycia nowonarodzone dziecko, jeśli wzorzec umieralności według wieku pozostanie niezmienny przez cały okres życia) i zagregowany współczynnik zgonów na nowotwory i choroby serca (natężenie zgonów wywołanych dwoma głównymi przyczynami przedwczesnych zgonów w Polsce, w przeliczeniu na populację danego powiatu). Wskaźnik edukacji wyrażony jest za pomocą dwóch wskaźników cząstkowych: edukacji przedszkolnej (odsetek dzieci uczęszczających do przedszkoli w wieku 3–4 lata w ogólnej liczbie dzieci w wieku 3–4 lata) i wyników egzaminu gimnazjalnego (odchylenie od średniej z wyników egzaminu gimnazjalnego dla określonego powiatu). Edukacja przedszkolna jest postrzegana jako ważny czynnik późniejszego sukcesu uczniów w dorosłym życiu. Drugi ze wskaźników dotyczy wyniku egzaminu gimnazjalnego z części matematyczno-przyrodniczej, który w perspektywie długoterminowej jest kluczowym elementem budującym społeczeństwo oparte na wiedzy. Dane są dostępne za pośrednictwem Centralnej Komisji Egzaminacyjnej. Wskaźnik zamożności wyznaczany jest jako suma dochodu podatników ogółem przed opodatkowaniem³. Tak rozumiany dochód brutto jest najbliższy dochodowi rozporządzalnemu i jest szczególnie użyteczny w analizie rozkładu terytorialnego zamożności. Przeliczony na miesiąc daje przybliżony hipotetyczny miesięczny dochód na mieszkańca (średnią zamożność) w danej jednostce administracyjnej.

LHDI ujmuje także różne nakłady polityki publicznej na rozwój społeczny w zakresie: zdrowia, edukacji i zamożności. Liczony jest także jako średnia geometryczna ze wskaźników: nakładów zdrowotnych i edukacyjnych oraz wydatków lokalnych. W określeniu miary nakładów na sferę zdrowia wykorzystywane są następujące wskaźniki: liczba lekarzy i lekarzy dentyistów według podstawowego miejsca pracy na 100 tys. osób i liczba pielęgniarek i położnych podstawowej opieki zdrowotnej według podstawowego miejsca pracy na 100 tys. osób. Wskaźniki te pokazują stopień zaangażowania zasobów ludzkich w świadczenie opieki zdrowotnej. Liczba lekarzy podstawowej opieki medycznej pokazuje ogólną dostępność opieki zdrowotnej, a liczba pielęgniarek i położnych różnicuje jakość opieki zapewnianej przez placówki medyczne w danym regionie. Dane pochodzą z informacji GUS. Wskaźnik nakładów edukacyjnych liczony jest jako średnia geometryczna z dwóch wskaźników cząstkowych: wydatki na edukację na ucznia ogółem i liczba uczniów przypadająca na jednego nauczyciela. Natomiast miarą nakładu dla ekonomicznego wymiaru standardu życia w ujęciu nakładów polityk publicznych jest suma wydatków z budżetów gmin i powiatów na terenie danego powiatu oraz pozycji budżetowych sfi-

³ Więcej informacji można znaleźć w Krajowym Raportcie o Rozwoju Społecznym [<http://www.mir.gov.pl>].

nansowanych ze środków unijnych w budżetach samorządu terytorialnego podzielona przez liczbę mieszkańców.

Do określenia predyktorów poprawy wartości LHDI można wykorzystać model logitowy⁴:

$$\text{logit } p = \ln\left(\frac{p}{1-p}\right) = \alpha_0 + \sum_{i=1}^k \alpha_i x_i, \quad (2)$$

gdzie: p – prawdopodobieństwo występowania określonego zdarzenia, $p/(1-p)$ – szansa wystąpienia określonego zdarzenia, $\alpha_0, \alpha_1, \dots, \alpha_k$ są współczynnikami regresji logistycznej, X_1, \dots, X_k – zmienne niezależne, które mogą mieć charakter zarówno jakościowy, jak i ilościowy.

Funkcja logistyczna umożliwia obliczenie wartości prawdopodobieństwa p_i ma postać:

$$p_i = \frac{\exp\left(\alpha_0 + \sum_{i=1}^k \alpha_i x_i\right)}{1 + \exp\left(\alpha_0 + \sum_{i=1}^k \alpha_i x_i\right)}. \quad (3)$$

Nieznane parametry modelu (3) szacuje się uogólnioną metodą największych kwadratów lub metodą największej wiarygodności. Do interpretacji wykorzystuje się iloraz szans rozumiany jako stosunek prawdopodobieństwa, że wystąpi dane zdarzenie, do prawdopodobieństwa, że ten przypadek nie pojawi się. Określa się zatem, o ile większa bądź mniejsza jest szansa wystąpienia zdarzenia. Po oszacowaniu parametrów modelu logitowego dokonuje się oceny jego poprawności, zliczając trafnie oraz błędnie zaklasyfikowane przypadki. W tabeli 1 umieszczono także liczebności trafnie i błędnie sklasyfikowanych powiatów.

Tabela 1. Macierz klasyfikacji powiatów

Wartości oczekiwane	Wartości obserwowane		Suma
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	$n_{11} = 93$	$n_{12} = 99$	$n_{1\bullet} = 159$
$\hat{y}_i = 0$	$n_{21} = 85$	$n_{22} = 135$	$n_{2\bullet} = 220$
Suma	$n_{\bullet 1} = 178$	$n_{\bullet 2} = 201$	$N = 379$

Źródło: opracowanie własne na podstawie [Dobosz 2004].

⁴ Szerzej o modelu logitowym w pracach [Gatnar, Walesiak 2004; Gruszczyński 2010; Maddala 2008].

Do oceny stopnia dopasowania modelu regresji logistycznej do danych empirycznych można wykorzystać miarę *zliczeniowy* R^2 , która przyjmuje wartości z przedziału $\langle 0, 1 \rangle$, zdefiniowaną następująco [Maddala 2008]:

$$R^2_{zlicz} = \frac{n_{11} + n_{22}}{n}. \quad (4)$$

Im wartość tej miary jest bliższa jednego, tym lepsze dopasowanie modelu logistycznego do danych empirycznych badanego zjawiska, R^2_{zlicz} oznacza procent trafnie zaklasyfikowanych przypadków. Model dobrze sprawdza się w prognozowaniu badanego zjawiska, gdy $R^2_{zlicz} > 50\%$. Oznacza to, że klasyfikacja na podstawie modelu jest lepsza od przypadkowej. Jakość zbudowanego modelu regresji logistycznej można również ocenić innymi miarami, np. za pomocą krzywej ROC, a w szczególności pola pod krzywą ROC, która powstaje przez połączenie punktów w kartezjańskim układzie o współrzędnych (1-swoistość, czułość).

Czułość opisuje zdolność wykrywania jednostek posiadających wyróżnioną cechę:

$$\text{czułość} = \frac{n_{11}}{n_{\bullet 1}}. \quad (5)$$

Swoistość opisuje zdolność wykrywania jednostek nieposiadających wyróżnionej cechy. Wyznaczana jest jako iloraz obserwacji nieposiadających wyróżnionej cechy i sklasyfikowanych poprawnie do liczby wszystkich obserwacji nieposiadających wyróżnionej cechy:

$$\text{swoistość} = \frac{n_{22}}{n_{\bullet 2}}. \quad (6)$$

Z krzywą ROC (*Receiver Operating Characteristic Curves*) [Zweig, Campbell 1993] związane jest pojęcie punktu odcięcia (*cut-off point*), który jest zdefiniowany jako pewna wartość zmiennej diagnostycznej dzieląca zbiorowość na dwie części: jedną, w której występowało pewne zdarzenie, oraz drugą, w której zdarzenie się nie pojawiło. Na podstawie modelu logitowego szacuje się prawdopodobieństwo występowania zjawiska i traktuje je jako wartości zmiennej diagnostycznej. Jeżeli badana jest n -elementowa próba, w której dla każdego obiektu możliwych jest K wartości zmiennej diagnostycznej $\hat{p}_1, \dots, \hat{p}_K$, to każda z tych wartości staje się punktem odcięcia \hat{p}_{cut} . Jeśli zmienna diagnostyczna $\hat{p}_i \geq \hat{p}_{cut}$, to obiekty takie zalicza się do grupy, w której występuje badane zjawisko ($\hat{y}_i = 1$), natomiast gdy $\hat{p}_i < \hat{p}_{cut}$, to uznaje się, że dla danych obiektów nie występuje badane zjawisko ($\hat{y}_i = 0$). Na tej podstawie tworzy się macierz klasyfikacji przypadków (tab. 1) i wyznacza wartości swoistości oraz czułości. Zatem powstaje n macierzy klasyfikacji przypadków oraz n punktów krzywej ROC [Sompolska-Rzechuła i in. 2014]. Powstała w ten sposób

krzywa ROC, a w szczególności pole pod nią, którego wartość odpowiada przedziałowi liczbowemu $\langle 0,1 \rangle$, obrazuje jakość klasyfikacyjną modelu. Dobry model to taki, który minimalizuje liczbę błędów, czyli wartości n_{12} oraz n_{21} . Zatem wysoką jakością klasyfikacyjną charakteryzuje się model, dla którego wyznaczane czułości i swoistości są możliwie największe, co oznacza, że dla takiego modelu pole pod krzywą ROC powinno być bliskie wartości 1. Gdy krzywa ROC pokrywa się z przekątną $x = y$, to decyzja o przyporządkowaniu przypadku do wybranej klasy podejmowana na podstawie modelu jest tak samo dobra jak losowy przydział badanych przypadków do tych grup. Jakość klasyfikacyjna modelu jest dobra, gdy pole pod krzywą ROC jest większe niż 0,5.

3. Wyniki badania

W ocenie poziomu rozwoju społecznego na poziomie lokalnym wykorzystano wskaźniki składające się na LHDİ w ujęciu zarówno nakładów, jak i efektów, przedstawione w punkcie 2 pracy. W budowie modelu logitowego uwzględniono następujące cechy: Y – poprawa poziomu rozwoju społecznego w roku 2010 w porównaniu z rokiem 2007 (1 – tak, 0 – nie), X_1 – wskaźnik zdrowia w i -tym powiecie, X_2 – wskaźnik edukacji w i -tym powiecie, X_3 – wskaźnik zamożności w i -tym powiecie, X_4 – wskaźnik nakładów zdrowotnych w i -tym powiecie, X_5 – wskaźnik nakładów edukacyjnych w i -tym powiecie, X_6 – wskaźnik wydatków lokalnych w i -tym powiecie, X_7 – typ powiatu (1 – grodzki, 0 – ziemski)⁵.

W celu znalezienia najlepszej kombinacji cech istotnie wpływających na poprawę poziomu rozwoju społecznego w powiatach Polski dokonano formalnej selekcji cech za pomocą regresji krokowej w tył i otrzymano następujący zestaw zmiennych: X_1, X_2, X_4 .

Oszacowany model logistyczny przybiera zatem następującą postać:

$$\text{logit} \hat{p} = 1,354 - 0,023x_1 - 0,028x_2 + 0,038x_4.$$

(0,449)
(0,008)
(0,010)
(0,5944)

Ilorazy szans dla zmiennych istotnie wpływających na poprawę poziomu rozwoju społecznego wynoszą odpowiednio: 0,977, 0,972 i 1,038.

Interpretując ilorazy szans przy i -tej zmiennej (zakładając, że pozostałe zmienne uwzględnione w modelu pozostaną bez zmian), uzyskuje się następujące informacje:

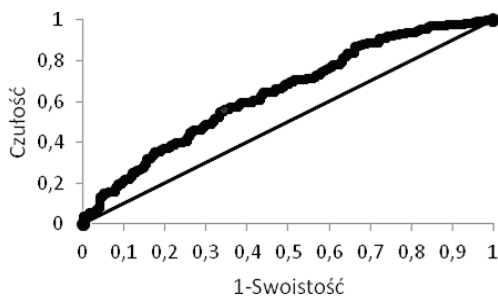
- zwiększenie wartości wskaźnika zdrowia o jeden powoduje spadek szansy na poprawę poziomu rozwoju społecznego o 2,3%,
- jeżeli wartość wskaźnika edukacji ulegnie zwiększeniu o jeden, to szansa na poprawę poziomu rozwoju społecznego spadnie o 2,8%,

⁵ W badaniu wykorzystano dane zawarte w Krajowym Raporcie o Rozwoju Społecznym [<http://www.mir.gov.pl>].

- zwiększenie wskaźnika nakładów zdrowotnych o jeden spowoduje wzrost szansy na poprawę poziomu rozwoju społecznego o 3,8%.

O ile interpretacja ilorazu szans przy zmiennej X_4 wydaje się zrozumiała, o tyle wyjaśnienia wymagają interpretacje ilorazów szans dla pozostałych dwóch zmiennych. Jednym z czynników uwzględnionych w wyznaczeniu wskaźnika zdrowia jest zagregowany współczynnik zgonów na nowotwory i choroby serca. Według informacji GUS głównymi przyczynami zgonów w Polsce są choroby układu krążenia i choroby nowotworowe, stanowiące ponad 70% wszystkich zgonów. Ponadto niekorzystnym zjawiskiem jest bardzo szybki wzrost liczby zgonów powodowanych chorobami nowotworowymi przy jednoczesnym wzroście liczby nowych zachorowań⁶. Zatem właśnie ten czynnik wpływa na zmniejszenie szansy poprawy poziomu rozwoju społecznego. W przypadku wskaźnika edukacji jednym z komponentów jest wynik egzaminu gimnazjalnego z części matematyczno-przyrodniczej liczony jako odchylenie od średniej z wyników egzaminu gimnazjalnego dla określonego powiatu. Ujemny wpływ na zwiększenie szansy poprawy poziomu rozwoju społecznego rozpatrywanego czynnika można tłumaczyć dużą wartością odchylenia standardowego wyników egzaminu w powiatach.

W kolejnym kroku analizy dokonano oceny poprawności oszacowanego modelu, zliczając trafność klasyfikacji powiatów i wyznaczając wartości czułości i swoistości modelu na podstawie wartości zawartych w tab. 1, i otrzymano, że wynoszą one odpowiednio: 52,25% i 67,16%. Model charakteryzuje się wyższą swoistością. Ma zatem większą zdolność wykrywania jednostek nieposiadających wyróżnionej cechy, czyli tych powiatów, w których nie zaobserwowano poprawy poziomu rozwoju społecznego. Wartość R^2_{zlicz} wynosi 60,20%, co oznacza, że klasyfikacja na podstawie modelu jest lepsza od przypadkowej. Wyniki testu Hosmera–Lemeshowa ($\chi^2 = 5,01$, $p = 0,756$) wskazują na brak istotnych różnic między liczebnościami empirycznymi a teoretycznymi, co jest wynikiem pożądanym



Rys. 1. Krzywa ROC

Źródło: opracowanie własne.

⁶ Szerzej na stronie <http://stat.gov.pl/> w publikacji *Zachorowalność i umieralność na nowotwory a sytuacja demograficzna Polski*.

i potwierdza dobrą jakość modelu [Hosmer, Lemenshow 2000]. O dość dobrej jakości modelu świadczy także wielkość pola pod krzywą ROC, które wynosi 64,5% (rys. 1).

4. Zakończenie

W pracy przedstawiono wykorzystanie modelu logitowego do identyfikacji predyktorów poprawy wartości LHDI w powiatach Polski. LHDI opiera się na wskaźnikach dotyczących trzech aspektów życia: zdrowia, edukacji i zamożności. W badaniu zastosowano podejście ujmujące zarówno nakłady, jak i efekty dotyczące wymienionych aspektów życia. Prawdopodobieństwo poprawy poziomu rozwoju społecznego oszacowano, wykorzystując model logitowy. Za zmienną objaśnianą przyjęto zmienną dychotomiczną określoną jako wystąpienie poprawy poziomu rozwoju społecznego, wyrażonej wartością LHDI, w roku 2010 w porównaniu z rokiem 2007. Za pomocą regresji krokowej w tył wyodrębniono zbiór zmiennych istotnie wpływających na poprawę poziomu rozwoju społecznego i otrzymano wskaźniki: zdrowia, edukacji i nakładów zdrowotnych. Dwie pierwsze zmienne mają ujemny wpływ na prawdopodobieństwo poprawy poziomu rozwoju społecznego. Taka sytuacja wynika z uwzględnienia we wskaźniku zdrowia zagregowanego współczynnika zgonów na nowotwory i choroby serca, a w przypadku wskaźnika edukacji wyrażenia wyniku egzaminu gimnazjalnego z części matematyczno-przyrodniczej jako odchylenia standardowego wyników egzaminu gimnazjalnego dla określonego powiatu. Natomiast wskaźnik nakładów na zdrowie ma dodatni wpływ na prawdopodobieństwo poprawy poziomu rozwoju społecznego. Zbudowany model charakteryzuje się dobrą jakością, o czym świadczą wartości miar oceniających jakość modelu: zliczeniowy współczynnik R^2 , wyniki testu Hosmera-Lemenshowa oraz wielkość pola pod krzywą ROC.

Zaproponowana metoda okazała się przydatna do identyfikacji czynników wpływających na prawdopodobieństwo poprawy wartości LHDI, a tym samym poziomu rozwoju społecznego na poziomie lokalnym.

Literatura

- Dobosz M., 2004, *Wspomagana komputerowo statystyczna analiza wyników badań*, Akademicka Oficyna Wydawnicza EXIT, Warszawa, s. 261-263.
- Gatnar E., Walesiak M. (red.), 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, s. 159-163.
- Gruszczyński M., 2010, *Modele zmiennych jakościowych dwumianowych*, [w:] *Mikroekonometria. Modele i metody analizy danych indywidualnych*, red. Gruszczyński M., Wolters Kluwer Polska, Warszawa.
- Hosmer D., Lemenshow S., 2000, *Applied Logistic Regression*, John Wiley & Sons, New Jersey.
- <http://www.mir.gov.pl> (19.12.2014).

<http://stat.gov.pl/> (19.12.2014).

Maddala G.S., 2008, *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.

Sompolska-Rzechuła A., Machowska-Szewczyk M., Chudecka-Głaz A., Cymbaluk-Płoska A., Menkiszak J., 2014, *The Use of Logistic Regression in the Ovarian Cancer Diagnostic*, *Ekonometria*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław (w druku).

Zweig M.H., Campbell G., 1993, Receiver Operang Characterisc (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine, 39, 561-577.

THE DEFINITION OF FACTORS INFLUENCING THE PROBABILITY OF IMPROVING THE LEVEL OF HUMAN DEVELOPMENT USING THE LOGIT MODEL

Summary: The diversity of social development level in Poland on the local level is much higher than between regions. Therefore, there is a need to monitor and assess the level of social development in poviats and municipalities in Poland. Correct measurement requires the use of suitable measurers. One of the most common measures used to assess the level of social development is Human Development Index and Local Human Development Index. The aims of this paper include the identification of predictors of the Local Human Development Index value improvement in poviats in Poland. In this study, a logistic regression model is used to estimate the probability of improvement of LHDI. The following factors were used: indicators for health, education and wealth as well as health expenditures index, educational expenditures index and local expense index and the type of poviat. The quality of the built up logistic regression model was evaluated using measures e.g.: count-R2, the Hosmer-Lemenshow test and ROC Curve.

Keywords: Local Human Development Index, logit model, odds ratios.