EKONOMETRIA ECONOMETRICS 2(48) • 2015

ISSN 1507-3866
e-ISSN 2449-9994

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

REGRESJA LOGISTYCZNA DLA DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

LOGISTIC REGRESSION FOR INTERVAL-VALUED SYMBOLIC DATA

DOI: 10.15611/ekt.2015.2.04

Streszczenie: W praktyce badawczej często mamy do czynienia z sytuacją, gdy zmienna zależna ma postać zmiennej dwumianowej (binarnej, dychotomicznej). Ponieważ model regresji liniowej nie znajduje tutaj zastosowania, konieczne jest zastosowanie modeli nieliniowych. Modelem regresji stosowanym dla zmiennych dwumianowych jest model regresji logistycznej. Artykuł prezentuje adaptację modelu regresji logistycznej dla zmiennych symbolicznych interwałowych. W tym celu wskazano cztery różne rozwiązania, które zaproponowano w literaturze przedmiotu. W części empirycznej zaprezentowano wyniki badań z zastosowaniem sztucznych i rzeczywistych zbiorów danych. Otrzymane wyniki wskazują, że model regresji logistycznej, po odpowiedniej modyfikacji, może znaleźć zastosowanie dla zmiennych symbolicznych interwałowych. Najlepsze dopasowanie uzyskują modele budowane na podstawie środków bądź metody krańców o estymacji łącznej.

Słowa kluczowe: regresja logistyczna, zmienne symboliczne interwałowe, analiza danych symbolicznych.

Summary: When dealing with real data situation we often have a binary (biomial, dichotomous) dependent variable. As the linear probability model is not such a good solution in such a situation there is a need to use nonlinear models. A quite good solution for such a situation is the logistic regression model. The paper presents an adaptation of linear regression model when dealing with symbolic interval-valued variables. Four approaches poposed by de Souza et. al [2011] how to apply such variables are presented. In the empirical part results obtained with the application of artificial and real data sets are shown. The best results are obtained for midpoint and bounds (joint estimation) methods.

Keywords: logistic regression, interval-valued symbolic variables, symbolic data analysis.

1. Wstęp

W regresji logistycznej przedmiotem modelowania jest zmienna dwumianowa (binarna, dychotomiczna). Przykładami takich zmiennych mogą być na przykład (por. [Gruszczyński 2010, s. 17, 53-55; Gatnar, Walesiak 2011, s. 99]):

- y – stan aktywności zawodowej: 1 – pracuje, 0 – w pozostałych przypadkach,
- y – zmiana dotychczasowego operatora sieci komórkowej: 1 – zmiana nastąpiła, 0 – zmiana nie nastąpiła,
- y – polecenie produktu lub usługi innej osobie: 1 – produkt (usługa) został polecony, 0 – w pozostałych przypadkach.

Do typowych celów modelowania zmiennej dwumianowej zalicza się przede wszystkim prognozowanie wartości zmiennej y (w tym prognoza tego, że zmienna $y = 1$), czyli prognoza zmiany prawdopodobieństwa wywołanej zmianą wartości jednej ze zmiennych.

Drugim celem jest ustalanie zmiennych, które są istotne dla określenia prawdopodobieństwa dla zmiennej y. Innymi celami są także weryfikacja hipotezy na temat mechanizmu generującego wartości y oraz konstrukcja funkcji zmiennych objaśniających, która pozwoli rozróżnić dwie grupy zbiorowości – jednej odpowiadającej $y = 1$ oraz drugiej, która odpowiada $y = 0$ (zob. [Gruszczyński 2010, s. 54]).

Celem artykułu jest prezentacja adaptacji klasycznego modelu regresji logistycznej dla zmiennych symbolicznych interwałowych. Dodatkowo w artykule porównano dokładność oszacowań otrzymanych z zastosowaniem każdej z metod na przykładzie sztucznych i rzeczywistych zbiorów danych. W artykule przedstawiono zagadnienie danych symbolicznych oraz cztery różne rozwiązania, które zaproponowano w literaturze przedmiotu dla regresji logistycznej danych interwałowych (zob. [de Souza, Queiroz, Cysneiros 2011]): metodę środków, metodę krańców w dwóch różnych wariantach tej metody.

W części empirycznej zaprezentowano wyniki badań z zastosowaniem sztucznych i rzeczywistych zbiorów danych. Artykuł stanowi pierwsze polskie opracowanie opisujące regresję logistyczną danych symbolicznych interwałowych, a dodatkowo porównuje różne podejścia estymacyjne i dokonuje ich ewaluacji.

2. Regresja logistyczna danych interwałowych

Obiekty symboliczne, w przeciwieństwie do obiektów w ujęciu klasycznym, mogą być opisywane przez następujące rodzaje zmiennych ([Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36]):

- zmienne nominalne, porządkowe, przedziałowe, ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli listy wartości z wagami.

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi znaleźć można m.in. w pracach: [Bock, Diday

(red.) (2000), s. 2-8; Dudek 2013, s. 42-43; 2004; Billard, Diday 2006, s. 7-66; Noirhomme-Fraiture, Brito 2011; Diday, Noirhomme-Fraiture 2008, s. 3-30].

W ogólnej postaci liniowy model regresji wielu zmiennych przedstawia się za pomocą następującego równania:

$$Y_t = b_0 X_{0t} + b_1 X_{1t} + \ldots + b_m X_{mt} + e_t = \sum_{j=0}^{m} b_j X_{jt} + e_t, \tag{1}$$

gdzie: Y – zmienna objaśniana (regresant), $X_0, X_1, \ldots, X_m$ – zmienne objaśniające (regresyjne), $b_0, b_1, \ldots, b_m$ – parametry strukturalne modelu, e – składnik losowy, $t = 1 \ldots, T$ – numer obserwacji, $j = 0, 1, \ldots, m$ – numer zmiennej objaśniającej.

W przypadku, gdy model przedstawiony równaniem 1 stosowany jest dla zmiennych dwumianowych, przedmiotem modelowania jest prawdopodobieństwo P_i, że zmienna objaśniana przyjmie wartość zero lub 1.

Niemniej jednak zastosowanie liniowego modelu regresji niesie za sobą ryzyko, że obliczone na jego podstawie prawdopodobieństwa będą większe od 1 lub mniejsze od zera (prezentuje to np. [Gatnar, Walesiak 2011, s. 100]). W związku z tym znacznie lepszym rozwiązaniem jest zastosowanie modelu logitowego.

W modelu logitowym zakłada się, że mamy do czynienia ze zmienną ukrytą y^*, która nie jest obserwowana bezpośrednio. Obserwujemy natomiast:

$$y_i = \begin{cases} 1, & \text{dla} \quad y^* > 0 \\ 0, & \text{dla} \quad y^* \leq 0 \end{cases}. \tag{2}$$

Zmienna ukryta y^* reprezentuje skłonność *i*-tego obiektu do przyjmowania wartości $y_i = 1$. Model logitowy ma zatem postać:

$$Y_t^* = b_0 X_{0t} + b_1 X_{1t} + \ldots + b_m X_{mt} + e_t = \sum_{j=0}^{m} b_j X_{jt} + e_t. \tag{3}$$

Prawdopodobieństwo, że zmienna niezależna y_i przyjmie wartość zero lub 1, jest zatem funkcją zmiennych objaśniających i parametrów:

$$P_i = F\left(x_i^{\mathrm{T}} b\right) = \frac{1}{1 + \exp\left(-x_i^{\mathrm{T}} b\right)} = \frac{\exp\left(x_i^{\mathrm{T}} b\right)}{1 + \exp\left(x_i^{\mathrm{T}} b\right)}, \tag{4}$$

gdzie: F – dystrybuanta rozkładu logistycznego.

Powstaje pytanie, w jaki sposób obliczyć prawdopodobieństwa z wykorzystaniem wzoru 4, jeżeli mamy do czynienia ze zmiennymi symbolicznymi interwało-

wymi. Zmienne te mają postać przedziału liczbowego: $[\underline{x}_i, \overline{x}_i]$, gdzie $\underline{x}_i$ to dolny kraniec przedziału *i*-tej zmiennej, a $\overline{x}_i$ to górny kraniec przedziału *i*-tej zmiennej.

W artykule de Souzy i in. (por. [de Souza, Queiroz, Cysneiros 2011]) zaproponowano cztery modyfikacje pozwalające na szacowanie prawdopodobieństwa z wykorzystaniem wzoru 4, jeżeli mamy do czynienia ze zmiennymi symbolicznymi interwałowymi [de Souza, Queiroz, Cysneiros 2011, s. 275-278]):

1. **Metoda środków** (*centers*), która jest stosowana m.in. w odniesieniu do regresji liniowej czy w analizie głównych składowych dla danych symbolicznych interwałowych (por. np. [Billard, Diday 2006; Dudek 2013]).

W tym rozwiązaniu zamiast całego przedziału zmiennej symbolicznej we wzorze 4 wykorzystuje się jedynie środek jej przedziału $\frac{\underline{x}_i + \overline{x}_i}{2}$. Prawdopodobieństwo, że zmienna y_i przyjmie wartość zero lub 1, obliczane jest dla środków przedziałów wszystkich zmiennych.

2. **Metoda krańców** (*bounds*). W tym przypadku zamiast całego przedziału zmiennej symbolicznej wykorzystywane są jedynie krańce tej zmiennej $\underline{x}_i$ oraz $\overline{x}_i$.

Prawdopodobieństwo wyrażone wzorem 4 może być szacowane łącznie z wykorzystaniem obydwu krańców jednocześnie – estymacja łączna (*joint estimation*). W odniesieniu do estymacji łącznej (*joint estimation*) prawdopodobieństwo wyznacza się ze wzoru 4, wykorzystując zarówno krańce dolne, jak i krańce górne przedziałów wszystkich zmiennych jednocześnie (mamy tu do czynienia z 2*m* zmiennymi, gdzie: *m* – liczba zmiennych symbolicznych interwałowych).

Prawdopodobieństwo to może być również średnią obliczoną z dwóch modeli (por. [Alexandre, Campilho, Kamel 2001]) – jednego dla krańców dolnych i drugiego dla krańców górnych – estymacja rozdzielona (*separated estimation*). Dokonuje się więc oszacowania dwóch prawdopodobieństw – jednego dla krańców górnych oraz drugiego dla krańców dolnych zmiennych symbolicznych interwałowych.

3. **Metoda wierzchołków** (*vertices*), która jest stosowana m.in. w analizie dyskryminacyjnej czy analizie głównych składowych dla danych symbolicznych interwałowych (por. np. [Silva, Brito 2006]). W metodzie tej zamiast *m* zmiennych symbolicznych interwałowych $[\underline{x}_{i1}, \overline{x}_{i1}], \ldots, [\underline{x}_{it}, \overline{x}_{it}]$ stosowana jest macierz **M**, która jest kombinacją wszystkich wierzchołków we wszystkich zmiennych:

$$\mathbf{M} = \begin{bmatrix} \underline{x}_{i1} & \cdots & \underline{x}_{it} \\ \underline{x}_{i1} & \cdots & \overline{x}_{it} \\ \vdots & \ddots & \vdots \\ \overline{x}_{i1} & \cdots & \overline{x}_{it} \\ \overline{x}_{i1} & \cdots & \overline{x}_{it} \end{bmatrix}. \tag{5}$$

Na przykład jeżeli mamy jeden obiekt i dwie zmienne symboliczne interwałowe $[\underline{x}_{11}, \overline{x}_{11}], [\underline{x}_{21}, \overline{x}_{21}]$, to macierz **M** ma postać:

$$\mathbf{M} = \begin{bmatrix} \underline{x}_{11} & \underline{x}_{21} \\ \underline{x}_{11} & \overline{x}_{21} \\ \overline{x}_{11} & \underline{x}_{21} \\ \overline{x}_{11} & \overline{x}_{21} \end{bmatrix}. \tag{6}$$

W metodzie wierzchołków ostateczne prawdopodobieństwo to (por. [de Souza, Queiroz, Cysneiros 2011, s. 277]):

a) średnia z prawdopodobieństw obliczonych dla wszystkich kombinacji wierzchołków danego obiektu,

b) wartość maksymalna wśród prawdopodobieństw obliczonych dla wszystkich kombinacji wierzchołków danego obiektu,

c) wartość minimalna wśród prawdopodobieństw obliczonych dla wszystkich kombinacji wierzchołków danego obiektu.

Wśród miar dopasowania dla modeli dwumianowych w literaturze przedmiotu zaproponowano (zob. np. [Gatnar, Walesiak 2011, s. 102-103; Gruszczyński i in. 2010, s. 71-72; Smith, McKenna 2013, s. 17-26; Hosmer, Lemeshow, Sturdivant 2013; Menard 2002]):

1. R^2 współczynnika korelacji między wartościami teoretycznymi i empirycznymi zmiennej objaśnianej.

2. Miara R^2 Efrona:

$$R^2 = 1 - \left[\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 / \left(n_1 - \frac{n_1^2}{n} \right) \right], \tag{7}$$

gdzie: y_i – wartości empiryczne zmiennej objaśnianej, $\hat{y}_i$ – wartości teoretyczne zmiennej objaśnianej, n_1 – liczba jedynek dla zmiennej y, n – liczba obserwacji.

1. Miara R^2 Nagelkerke:

$$R^2 = \frac{1 - \exp\left(\left(D - D_{null} \right) / n \right)}{1 - \exp\left(- D_{null} / n \right)}, \tag{8}$$

gdzie: $D = \ln L_{UR}$ – maksimum funkcji wiarygodności, przy maksymalizacji względem wszystkich parametrów (dla pełnego modelu), $D_{null} = \ln L_R$, L_R – maksimum funkcji wiarygodności przy maksymalizacji pod warunkiem $\forall_{j=1}^{m} b_j = 0$ (dla modelu tylko z wyrazem wolnym).

2. Miara R^2 McFaddena:

$$R^2 = 1 - \frac{D}{D_{null}}. \tag{9}$$

Miary dopasowania R^2 dla modeli dwumianowych należą do przedziału $[0;1]$ i im są większe, tym lepsze dopasowanie modelu.

Prognozę dla prawdopodobieństwa P_i można wyznaczyć na podstawie wektora zmiennych objaśniających. Dla próby zbilansowanej $\hat{y}_i = 0,$ jeżeli $\hat{P}_i \le 0,5$ oraz $\hat{y}_i = 1$ dla $\hat{P}_i > 0,5$. W próbie niezbilansowanej $\hat{y}_i = 0$, jeżeli $\hat{P}_i \le \alpha$ oraz $\hat{y}_i = 1$ dla $\hat{P}_i > \alpha$ (α – odsetek jedynek w próbie).

3. Wyniki badań empirycznych

Celem badania jest porównanie czterech proponowanych w literaturze rozwiązań pod względem jakości dopasowania modeli do danych (w sensie współczynnika R^2). Dotychczasowe badania z zastosowaniem sztucznych zbiorów danych (zob. [de Souza, Queiroz, Cysneiros 2011, s. 278-280]) wskazują, że zwykle to metoda krańców o estymacji rozdzielonej otrzymuje wyniki najlepsze dla różnych modeli, a najgorsze metoda środków.

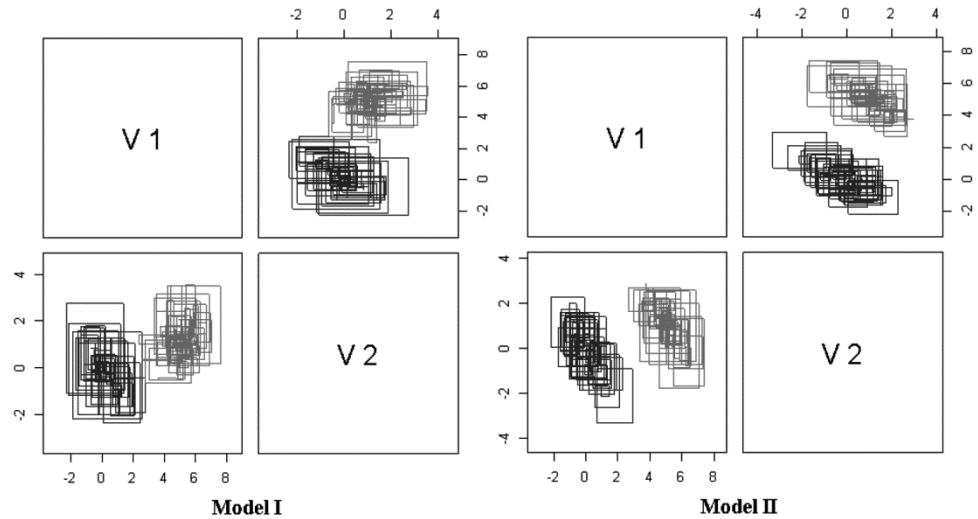
Na potrzeby badań empirycznych przygotowano w programie R z wykorzystaniem pakietu `clusterSim` dwa sztuczne zbiory danych (rys. 1):

1. Zbiór 100 obiektów symbolicznych, podzielony na trzy klasy o wydłużonym kształcie, które są opisywane przez dwie zmienne symboliczne interwałowe. Obserwacje są losowane niezależnie z rozkładu normalnego o średnich (0, 0), (1,5, 7), (3, 14) oraz macierzy kowariancji $\sum(\sigma_{jj} = 1, \sigma_{jl} = -0,9)$.

2. Zbiór 100 obiektów symbolicznych, podzielony na dwie klasy o wydłużonym kształcie, które są opisywane przez dwie zmienne symboliczne interwałowe. Obserwacje są losowane z rozkładu normalnego o średnich (0, 0), (1, 5) i macierzach kowariancji $\sum_1 = \begin{bmatrix} 1 & -0,9 \\ -0,9 & 1 \end{bmatrix}, \sum_2 = \begin{bmatrix} 1 & 0,5 \\ 0,5 & 1 \end{bmatrix}$.

W badaniach empirycznych wykorzystano także zbiór danych opisujący oleje (zbiór danych przygotowali M. Ichino i H. Yaguchi). Zbiór opisuje 8 różnych tłuszczów roślinnych i zwierzęcych, które są opisywane przez cztery zmienne symboliczne interwałowe (zob. [Ichino, Yaguchi 1994]) oraz zbiór `cars` (pochodzący z programu SODAS 2.50[1]). Zbiór `cars` zawiera 33 modeli samochodów różnych marek, które są opisywane przez 11 zmiennych (w tym 8 interwałowych). Do analiz wykorzystano jedynie zmienne interwałowe, a zbiór danych podzielono na dwie grupy samochodów: użytkowe (10 obiektów) oraz pozostałe (23 obiekty).

[1] Program jest dostępny pod adresem www.info.fundp.ac.be/asso/.



Rys. 1. Zbiory danych wygenerowane na potrzeby badań empirycznych

Źródło: opracowanie własne z wykorzystaniem programu R.

Tabela 1. Wyniki badań empirycznych

Metoda szacowania		Środków	Krańców (estymacja łączna)	Krańców (estymacja rozdzielona)[a]	Wierzchołków (wyniki uśrednione)[b]
Zbiór danych I	dokładność prognozy	1	1	1	1
	R^2 Efrona	1	1	1	1
	R^2 Nagelkerke	1	1	1	1
	R^2 McFaddena	1	1	1	1
Zbiór danych II	dokładność prognozy	1	1	1	1
	R^2 Efrona	1	1	1	1
	R^2 Nagelkerke	1	1	1	1
	R^2 McFaddena	1	1	1	1
Zbiór Ichnino i Yaguchiego	dokładność prognozy	1	1	1	1
	R^2 Efrona	1	1	1	1
	R^2 Nagelkerke	1	1	1	1
	R^2 McFaddena	1	1	1	1
Zbiór `cars`	dokładność prognozy	1	1	0,94	0,95
	R^2 Efrona	1	1	0,87	0,89
	R^2 Nagelkerke	1	0,99	0,91	0,95
	R^2 McFaddena	1	0,99	0,86	0,89

[a] Wyniki uśredniono na podstawie wyników otrzymanych dla krańca górnego i dolnego; [b] w tabeli zaprezentowano wyniki dla rozwiązania, które polega na uśrednianiu wyników; pozostałe rozwiązania (wartość minimalna i maksymalna) uzyskały nieco gorsze wyniki.

Źródło: opracowanie własne z zastosowaniem autorskich procedur programu R.

Wyniki otrzymane z zastosowaniem każdej z proponowanych metod dla poszczególnych zbiorów danych zawarto w tab. 1.

Z danych zawartych w tab. 1 wynika, że w odniesieniu do zbiorów danych o typowych (wydłużonych) kształtach wszystkie metody zaproponowane w pracy de Souzy, Queiroza i Cysneirosa [2011] uzyskują stuprocentową dokładność prognozy oraz wszystkie mierniki R^2 są równe jedności.

Jeśli mamy do czynienia z nieco bardziej skomplikowanym zbiorem danych – które tworzą skupienia o klasach trudno separowalnych czy nierozłącznych i które dodatkowo mają nietypowe kształty skupień – (jak np. zbiór `cars`), to najlepsze wyniki uzyskuje metoda środków, następnie metoda krańców o estymacji łącznej. Najsłabiej wypadają tu metoda wierzchołków oraz metoda krańców o estymacji rozdzielonej.

4. Zakończenie

Regresja logistyczna może znaleźć zastosowanie do analizowania zjawisk opisywanych przez zmienne symboliczne interwałowe oraz zmienne metryczne, które opisują obiekty symboliczne.

Przeprowadzone badania empiryczne wskazują, że w odniesieniu do zbiorów danych o klasycznym wydłużonym kształcie wszystkie rozwiązania zaproponowane w literaturze przedmiotu osiągają takie same wyniki, jeżeli chodzi o dokładność prognozy oraz dopasowanie modelu do danych (w sensie miary R^2). Gdy mamy do czynienia ze zbiorami danych o nieco bardziej skomplikowanej strukturze danych (tj. zbiorów danych tworzących skupienia trudno separowalne lub nierozłączne o kształtach niesferycznych), wtedy najlepsze wyniki uzyskała metoda środków oraz metoda krańców o estymacji łącznej. Najgorsze wyniki uzyskały metoda wierzchołków, która uśrednia wyniki, oraz metoda krańców o estymacji rozdzielonej.

Celem dalszych badań będzie analiza porównawcza proponowanych w literaturze przedmiotu rozwiązań w zakresie regresji interwałowych z zastosowaniem sztucznych i rzeczywistych zbiorów danych różnego typu (w tym zbiorów danych zawierających obserwacje odstające i zmienne zakłócające).

Literatura

Alexandre L.A., Campilho A.C., Kamel M., 2001, *On combining classifiers using product and sum rules*, Pattern Recogintion Letters, vol. 22, issue 12, s. 1283-1289.

Bock H.-H., Diday E. (red.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg.

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

de Souza R.M.C.R., Queiroz D.C.F, Cysneiros F.J.A., 2011, *Logistic regression-based pattern classifiers for symbolic interval data*. Pattern Analysis and Applications, vol. 14, issue 3, s. 273-282.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2004, *Tworzenie obiektów symbolicznych z baz danych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021, s. 107-114.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Gatnar E., Walesiak M. (red.), 2011, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Gruszczyński M. (red.), 2010, *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Wolters Kulwer Polska, Warszawa.

Hosmer D.W., Lemeshow S., Sturdivant R.X., 2013, *Applied logistic regression*, John Wiley & Sons, Chichester.

Ichino M., Yaguchi H., 1994, *Generalized Minkowski metrics for mixed feature-type data analysis*, IEEE Transactions on Systems, Man and Cybernetics, vol. 24, no. 4, s. 698-708.

Menard S., 2002, *Applied logistic regression*, second edition, Sage Publishing, Thousand Oaks, California.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: Symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, s. 157-170.

Silva A.P.D., Brito P., 2006, *Linear discriminant analysis for interval data*, Computational Statistics, vol. 21, issue 2, s. 289-308.

Smith T.J., McKenna C.M, 2013, *A comparison of logistic regression pseudo R^2 indices*, Multiple Linear Regression Viewpoints, vol. 39(2), s. 17-26.

Walesiak M., Dudek A., 2014, *The clusterSim package*, www.r-project.org.