PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
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

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

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Bąk, Marcin Pełka, Aneta Rybicka
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE PAKIETU `dcMNM` PROGRAMU R W BADANIACH PREFERENCJI KONSUMENTÓW WÓDKI

Streszczenie: Głównym celem artykułu jest zastosowanie modeli logitowych w analizie preferencji konsumentów wódki. W tym celu wykorzystano pakiet `dcMNM` (*discrete choice MultiNominal Models*) programu R. Artykuł prezentuje podstawowe pojęcia z zakresu modeli logitowych oraz rezultaty szacowania preferencji. Rezultaty te pozwalają na wskazanie najlepszych oraz najgorszych marek wódki i tych atrybutów, które mają największe znaczenie dla konsumentów.

Słowa kluczowe: badania preferencji, wielomianowe modele logitowe, metody wyborów dyskretnych, program R.

1. Wstęp

Preferencje pozwalają na wyjaśnienie, jak i dlaczego konsumenci dokonują swoich wyborów. W badaniach preferencji znajdują zastosowanie metody wyborów dyskretnych. Pozwalają one na odzwierciedlenie wyborów konsumentów dokonywanych pomiędzy różnymi ofertami (profilami). Modele w ramach metod wyborów dyskretnych mogą być różnego typu. Są to: wielomianowy model logitowy (*MultiNominal Logit Model* – MNL), warunkowy model logitowy (*Conditional Logit Model* – CLM) i mieszany model logitowy (*Mixed Logit Model* – MLM). Podstawą rozróżnienia tych modeli jest charakter zmiennych objaśniających uwzględnionych w modelu.

Głównym celem artykułu jest wykorzystanie wielomianowych modeli logitowych i pakietu `dcMNM` opracowanego dla programu R [Bąk 2013a] w analizie preferencji konsumentów wódki. W artykule przedstawiono podstawowe pojęcia z zakresu modeli logitowych oraz wyniki szacowania preferencji. Rezultaty te pozwalają na wskazanie najlepszych i najgorszych marek wódki oraz tych atrybutów, które mają największe znaczenie dla konsumentów.

2. Wielomianowe modele kategorii nieuporządkowanych i ich estymacja

Realizacje zmiennych w modelach mikroekonometrycznych są najczęściej wynikami pomiarów na skalach słabych (niemetrycznych). Zgromadzone obserwacje są zwykle liczbowymi wartościami dyskretnymi.

W mikroekonometrii występują następujące rodzaje zmiennych objaśnianych [zob. Gatnar, Walesiak (red.) 2011, s. 113; Gruszczyński (red.) 2012; Bąk 2013b]:

a) zmienne dychotomiczne (dwukategorialne, np. binarne),

b) zmienne politomiczne (wielokategorialne), wśród których wyróżnia się zmienne o kategoriach uporządkowanych i nieuporządkowanych,

c) zmienne ograniczone, wśród których wyróżnia się zmienne cenzurowane i ucięte,

d) zmienne licznikowe.

W mikroekonometrii stosuje się najczęściej następujące rodzaje modeli [zob. Gatnar, Walesiak (red.) 2011, s. 113; Gruszczyński 2012 (red.); Bąk 2013b]:

a) modele dwumianowe:
- liniowe modele prawdopodobieństwa,
- modele logitowe i probitowe,
- modele komplementarne log-log,
- modele log-liniowe;

b) modele wielomianowe:
- modele kategorii nieuporządkowanych,
- modele kategorii uporządkowanych;

c) modele klas ukrytych;

d) modele przeżycia (trwania);

e) modele zmiennych ograniczonych.

Istotne miejsce wśród tych modeli zajmują: wielomianowy model logitowy, warunkowy model logitowy oraz mieszany model logitowy. Wielomianowy model logitowy jest uogólnieniem modelu logitowego dla danych binarnych. Model tego typu może być stosowany, gdy zmienna objaśniana przyjmuje w sposób dyskretny wartości ze zbioru liczącego więcej niż dwie kategorie [Gatnar, Walesiak (red.) 2011, s. 113]. Podstawami teoretycznymi tego modelu są teoria użyteczności losowej oraz aksjomat wyboru Luce'a [Coombs, Dawes i Tversky 1977, s. 217 i n.].

Wielomianowy model logitowy można przedstawić jako [Agresti 2002, s. 267-268; Gatnar, Walesiak (red.) 2011, s. 113-114; Bąk 2004, s. 118-120]:

$$P_{ki} = \frac{\exp\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}_i\right)}{\sum_{l=1}^{n}\exp\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}_l\right)}, \tag{1}$$

gdzie: P_{ki} – prawdopodobieństwo wyboru *i*-tej kategorii przy *k*-tym stanie zmiennych objaśniających; $\mathbf{x}_k^{\mathrm{T}}$ – wektor reprezentujący *k*-ty wiersz macierzy $\mathbf{X}$ (zmiennych objaśniających), $\boldsymbol{\beta}_i$ – wektor parametrów związany z *i*-tą kategorią zmiennej objaśnianej, przy czym $\boldsymbol{\beta}_n = \mathbf{0}$.

Warunkowy model logitowy został zaproponowany przez McFaddena [1974] jako uogólnienie wielomianowego modelu logitowego. Podstawową różnicą jest charakter zmiennych objaśniających, tzn. macierzy $\mathbf{X}$, w równaniu (1). Jeżeli zmienne objaśniające charakteryzują konsumentów, to na ogół wykorzystuje się wielomianowy model logitowy. Jeżeli natomiast zmienne objaśniające opisują obiekty będące przedmiotem wyboru, to z reguły stosuje się warunkowy model logitowy [Gatnar, Walesiak (red.) 2011, s. 114].

Warunkowy model logitowy można przedstawić jako [Gatnar, Walesiak (red.) 2011, s. 114; So, Kuhfeld 1995; Bąk 2004, s. 120-122]:

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

gdzie: $\mathbf{z}_{ki}^{\mathrm{T}}$ – *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 $\mathbf{Z}$ we wzorze (2) zawiera charakterystyki produktów lub usług, względem których badane są preferencje respondentów.

Mieszany model logitowy stanowi połączenie cech jednostek (osób) i cech opcji (alternatyw) w jednym modelu. Mieszany model logitowy można przedstawić jako [So, Kuhfeld 1995]:

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

gdzie: oznaczenia jak we wzorach (1) i (2).

Do estymacji parametrów warunkowego modelu logitowego wykorzystuje się funkcję największej wiarygodności w postaci:

$$\log L\left(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}\right) = \sum_{k=1}^{p}\sum_{i=1}^{n} d_{ki}\log P_{ki}, \tag{4}$$

gdzie: $d_{ki} = 1$, jeśli wybrano opcję *i* ($d_{ki} = 0$ w przeciwnym przypadku); *k* – numer konsumenta (osoby); *i* – numer opcji wyboru; $\mathbf{y}$ – zmienna zależna (objaśniana); $\mathbf{X}$ – zmienne objaśniające (np. charakterystyki konsumentów).

W programie R do optymalizacji funkcji największej wiarygodności (4) wykorzystuje się funkcję optim z pakietu stats. Wymaga ona podania wartości początkowych parametrów do optymalizacji oraz funkcji, która będzie poddawana optymalizacji, a także zmiennych zależnych oraz objaśniających. Na potrzeby pakietu dcMNM [Bąk 2013a] programu R opracowano funkcję największej wiarygodności na podstawie pracy [Jackman 2007].

3. Zastosowanie modeli wielomianowych w badaniu preferencji konsumentów wódki

W 2012 r. przygotowano i przeprowadzono badanie ankietowe dotyczące preferencji konsumentów wódki (w badaniu wzięli udział mieszkańcy Jeleniej Góry i okolic). Głównym celem badania była identyfikacja i analiza preferencji respondentów w odniesieniu do przedmiotu badania.

W badaniu ankietowym uwzględniono pięć atrybutów opisujących wódkę:

a) kraj – z pięcioma poziomami: Polska, Rosja, Szwecja, Finlandia, Niemcy,

b) objętość – z trzema poziomami: 0,5 l, 0,7-0,75 l, 1 litr i więcej,

c) smak – z czterema poziomami: czysta, owocowa, ziołowa, wytrawna,

d) zawartość alkoholu – z trzema poziomami: 40%, 45%, 50% i więcej,

e) cena – z czterema poziomami: poniżej 20 zł, 20-40 zł, 40-100 zł, powyżej 100 zł.

Pełny układ czynnikowy liczy 720 profilów ($5 \times 3 \times 4 \times 3 \times 4$). Za pomocą funkcji z pakietu AlgDesign programu R wygenerowano cząstkowy układ czynnikowy z podziałem na bloki liczące 360 profilów. Liczba bloków wynosiła 4 (przykładowy zbiór z bloku 1 prezentuje tab. 1). Liczba zbiorów w każdym bloku – 15. Liczba opcji w każdym zbiorze – 6 (5 profilów plus opcja rezygnacji z wyboru). Łącznie w bloku 1 zebrano 157 ankiet, w bloku 2 – 123 ankiety, w bloku 3 – 140 ankiet, a w bloku 4 – 134 ankiety. Łącznie zgromadzono 49 860 obserwacji. W trakcie badania ankietowego (próba miała charakter przypadkowy) zebrano 544 kwestionariusze poprawnie wypełnionych ankiet.

Tabela1. Przykładowy zbiór profilów wyboru z bloku pierwszego

Kraj produkcji	Objętość	Smak	Zawartość alkoholu	Cena	Wybieram opcję
Niemcy	1 l i więcej	ziołowa	40%	poniżej 20 zł	1
Niemcy	0,5 l	czysta	45%	poniżej 20 zł	2
Szwecja	0,7-0,75 l	czysta	45%	poniżej 20 zł	3
Finlandia	1 l i więcej	czysta	45%	poniżej 20 zł	4
Szwecja	1 l i więcej	owocowa	45%	poniżej 20 zł	5
Żaden z powyższych					6

Źródło: opracowanie własne.

W prezentowanej analizie uwzględniono pierwszy blok danych – tj. 90 profilów (15 zbiorów po 6 opcji), które oceniło 157 respondentów, co daje łącznie 14 130 obserwacji. Do estymacji warunkowego modelu logitowego zastosowano skrypt korzystający z pakietu `dcMNM`:

```
library(dcMNM)
x<-read.csv2("wodka_X1.csv",header=TRUE)
y<-read.csv2("wodka_Y1.csv",header=TRUE)
head(CLMdata(x,y))
AS<-CLMattrsel(x,y)
print(AS$vif)
print(head(AS$W))
clm<-CLMmodel(x,y)
print(clm)
CLMgraph(x,y)
Pclm<-CLMprob(x,y)
print(Pclm)
```

Polecenie `head(CLMdata(x,y))` wyświetla fragment zbioru danych o strukturze wymaganej do estymacji warunkowego modelu logitowego.

```
choice kraj1 kraj2 kraj3 kraj4 kraj5 objetosc1 objetosc2 objetosc3 smak1
     0     0     0     0     0     1         0         0         1     0
     0     0     0     0     0     1         1         0         0     1
     0     0     0     1     0     0         0         1         0     1
     0     0     0     0     1     0         0         0         1     1
     0     0     0     1     0     0         0         0         1     0
     1     0     0     0     0     0         0         0         0     0
smak2 smak3 smak4 alkohol1 alkohol2 alkohol3 cena1 cena2 cena3 cena4
    0     1     0         1         0         0     1     0     0     0
    0     0     0         0         1         0     1     0     0     0
    0     0     0         0         1         0     1     0     0     0
    0     0     0         0         1         0     1     0     0     0
    1     0     0         0         1         0     1     0     0     0
    0     0     0         0         0         0     0     0     0     0
```

Atrybuty opisujące opcje wyboru są reprezentowane za pomocą zmiennych sztucznych otrzymanych w wyniku kodowania zero-jedynkowego (są to zmienne objaśniające w modelu). Zmienną objaśnianą, reprezentującą wybory respondentów, jest `choice`. Polecenie `AS<-CLMattrsel(x,y)` umożliwia usunięcie współliniowości występującej między zmiennymi sztucznymi na podstawie czynnika wzrostu wariancji *VIF* (*Variance Inflation Factor*) [zob. np. Maddala 2006; Greene 2008]. W rezultacie w dalszej analizie zostały uwzględnione zmienne sztuczne, dla których czynnik wzrostu wariancji *VIF* nie przekroczył wartości 10[1]:

[1] Przyjmuje się, że wartość czynnika wzrostu wariancji *VIF* większa lub równa 10 wskazuje na istotną współliniowość zmiennych [zob. np. Dobosz 2004, s. 207; Orme, Combs-Orme 2009, s. 27].

```
print(AS$vif)
   kraj1    kraj2    kraj3    kraj4    kraj5    objetosc1 objetosc2
2.766260 1.600265  3.610946  2.817511  4.276752  1.959654  1.984138
 smak1      smak2     smak3  alkohol1  alkohol2     cena1     cena2
2.060787 1.788195  1.900510  1.623302  2.703496  1.565596  3.643379
    cena3
 1.919827
print(head(AS$W))
  choice kraj1 kraj2 kraj3 kraj4 kraj5 objetosc1 objetosc2 smak1 smak2
1      0     0     0     0     0     1         0         0     0     0
2      0     0     0     0     0     1         1         0     1     0
3      0     0     0     1     0     0         0         1     1     0
4      0     0     0     0     1     0         0         0     1     0
5      0     0     0     1     0     0         0         0     0     1
6      1     0     0     0     0     0         0         0     0     0
  smak3 alkohol1 alkohol2 cena1 cena2 cena3
1     1        1        0     1     0     0
2     0        0        1     1     0     0
3     0        0        1     1     0     0
4     0        0        1     1     0     0
5     0        0        1     1     0     0
6     0        0        0     0     0     0
```

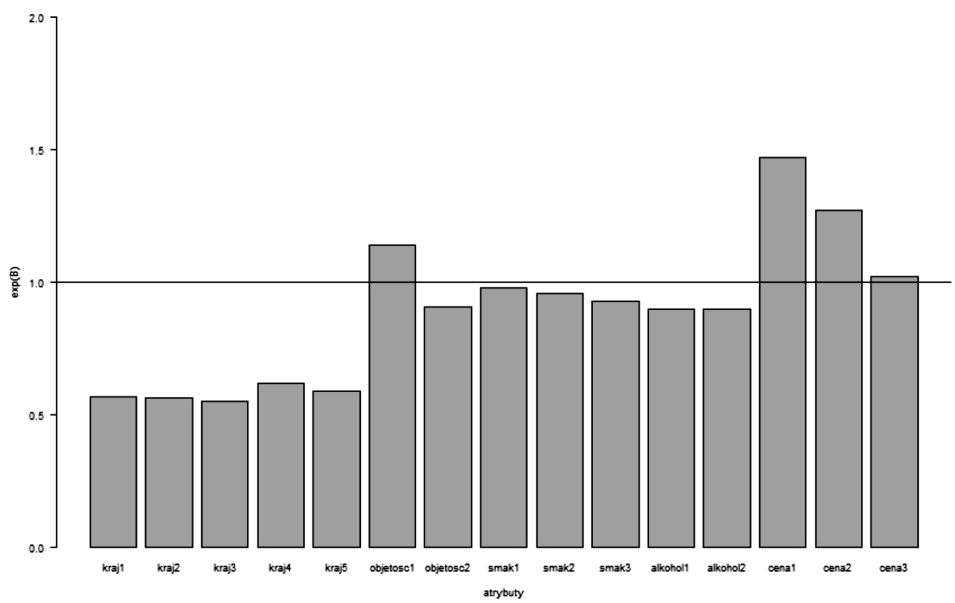
Polecenie `clm<-CLMmodel(x,y)` szacuje parametry modelu (`coef`) oraz oblicza błędy standardowe (`se`), wartości statystyki (`t`), wartości *p-value* (`Pr(>|t|)`) i ilorazy hazardu (`exp(coef)`):

```
print(clm)
$estimate
                 coef         se          t    Pr(>|t|) exp(coef)
kraj1       -0.5654864  0.1422622 -3.9749589  0.0000704 0.5680838
kraj2       -0.5680860  0.1690173 -3.3611116  0.0007763 0.5666089
kraj3       -0.5921168  0.1399808 -4.2299858  0.0000234 0.5531551
kraj4       -0.4759962  0.1270159 -3.7475324  0.0001786 0.6212658
kraj5       -0.5286539  0.1353328 -3.9063250  0.0000937 0.5893978
objetosc1    0.1319658  0.0605365  2.1799377  0.0292621 1.1410693
objetosc2   -0.0967864  0.0667818 -1.4492931  0.1472557 0.9077499
smak1       -0.0189597  0.0754349 -0.2513386  0.8015523 0.9812189
smak2       -0.0419719  0.0915944 -0.4582365  0.6467825 0.9588967
smak3       -0.0737781  0.0801963 -0.9199689  0.3575890 0.9288778
alkohol1    -0.1033741  0.0978124 -1.0568609  0.2905751 0.9017896
alkohol2    -0.1040915  0.0847939 -1.2275824  0.2196038 0.9011428
cena1        0.3853024  0.1651003  2.3337474  0.0196090 1.4700588
cena2        0.2416567  0.1154108  2.0938829  0.0362704 1.2733570
cena3        0.0238967  0.1305776  0.1830077  0.8547920 1.0241845

$logLik
[1] -4166.118

$McFaddenR2
[1] 0.0007013

$LR
df Chisq Pr(>Chisq)
[1,] 10 5.847191 0.827934
```

Dopasowanie modelu do danych empirycznych jest oceniane na podstawie wartości współczynnika determinacji pseudo R^2 McFaddena (`McFaddenR2`) [McFadden 1974].

W celu interpretacji wyników estymacji parametrów wykorzystano ilorazy hazardu przedstawione graficznie za pomocą polecenia `CLMgraph(x,y)` na rys. 1. Wartości ilorazów hazardu informują o stymulującym (wartości większe od 1) lub destymulującym (wartości mniejsze od 1) wpływie atrybutu na prawdopodobieństwo wyboru określonej opcji.

Analiza wyników zilustrowanych na rys. 1 pozwala stwierdzić, że stymulująco na prawdopodobieństwo wyboru wódki wpływają następujące poziomy atrybutów: objętość 0,5 l, cena do 20 zł, cena 20-40 zł oraz cena 40-100 zł. Najbardziej destymulująco na wybór wpływa kraj pochodzenia (wszystkie poziomy atrybutu są tu traktowane tak samo), a nieco mniej destymulująco – smak wódki, zwartość alkoholu powyżej 40%.



Rys. 1. Wykres ilorazów hazardu dla parametrów różnych od zera

Źródło: opracowanie własne z wykorzystaniem programu R.

4. Podsumowanie

Mikroekonometryczne wielomianowe modele logitowe znajdują zastosowanie w analizie preferencji wyrażonych opartych na metodzie wyborów dyskretnych. W metodzie tej różne typy modeli wielomianowych mogą być szacowane z wyko-

rzystaniem tej samej procedury wykorzystującej iteracyjne algorytmy optymalizacyjne. Jednakże estymacja różnych typów modeli wielomianowych wymaga różnej struktury danych empirycznych. Pakiet dcMNM umożliwia reorganizację danych empirycznych do postaci struktur wymaganych dla różnych typów modeli oraz estymację parametrów tych modeli i szacowanie prawdopodobieństw wyboru poszczególnych profilów.

Kierunkiem dalszych badań powinny być oszacowania modelu wielomianowego i mieszanego i interpretacja wyników oraz integracja pakietu dcMNM z rozwiązaniami typowymi dla stron WWW w celu gromadzenia danych empirycznych.

Literatura

Agresti A. (2002), *Categorical Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.

Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław.

Bąk A. (2013a), *Discrete choice multinomial models – package* dcMNM, [URL:] http://keii.ue.wroc.pl/dcMNM.

Bąk A. (2013b), *Mikroekonometryczne metody badania preferencji konsumentów z wykorzystaniem program R*, Wydawnictwo C.H. Beck, Warszawa.

Cameron A.C., Trivedi P.K. (2005), *Microeconometrics. Methods and applications*, Cambridge University Press, New York.

Coombs C.H., Dawes R.M., Tversky A. (1977), *Wprowadzenie do psychologii matematycznej*, PWN, Warszawa.

Dobosz M. (2004), *Wspomagana komputerowo statystyczna analiza wyników badań*, wydanie drugie uaktualnione, Akademicka Oficyna Wydawnicza EXIT, Warszawa.

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

Greene W.H. (2008), *Econometric Analysis*, Sixth Edition, Pearson Prentice Hall, Upper Saddle River.

Gruszczyński M. (red.) (2012), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, wydanie drugie rozszerzone, Wolters Kluwer, Warszawa.

Jackman S. (2007), *Models for unordered outcomes*, Political Science 150C/350C [URL:] http://jackman.stan-ford.edu/classes (2.05.2009).

Maddala G.S. (2006), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.

McFadden D. (1974), *Conditional logit analysis of qualitative choice behavior*, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, Academic Press, New York – San Francisco – London.

Orme J.G., Combs-Orme T. (2009), *Multiple Regression with Discrete Dependent Variables*, Oxford University Press, Oxford – New York.

So Y., Kuhfeld W.F. (1995), *Multinomial Logit Models*, [URL:] http://support.sas.com/techsup/technote/ mr2010g.pdf (12.03.2012).

APPLICATION OF THE MMLM PACKAGE OF R SOFTWARE FOR VODKA CONSUMERS PREFERENCE ANALYSIS

Summary: The main aim of the paper is to apply logit models in preference analysis of vodka consumers. In order to obtain such a goal `dcMNM` (*discrete choice MultiNominal Models*) package of R software was applied. The paper presents basic terms of logit models and the results of preference analysis. The estimates allowed to identify the best and the worst vodka brands as well as the attributes that have the biggest influence on preferences.

Keywords: preference analysis, multinominal logit model, discrete choice models, R program.