PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 84
Ekonometria 27 2009

**Andrzej Bąk**
Uniwersytet Ekonomiczny we Wrocławiu

# ZASTOSOWANIE MIKROEKONOMETRYCZNYCH MODELI ZMIENNYCH DYCHOTOMICZNYCH W BADANIACH PREFERENCJI I ICH ESTYMACJA W PROGRAMIE R

**Streszczenie:** Celem artykułu jest wskazanie (na przykładach empirycznych) możliwości wykorzystania modeli z dychotomiczną zmienną objaśnianą w analizie preferencji konsumentów. Omówiono w nim podstawowe założenia estymacji modeli dwumianowych z wykorzystaniem koncepcji uogólnionych modeli liniowych, a także model klas ukrytych szacowany na podstawie binarnych zmiennych obserwowanych.

**Słowa kluczowe:** badania preferencji, mikroekonometria, modele zmiennych dychotomicznych.

## 1. Preferencje

Pojęcie preferencji (relacji preferencji) odgrywa w ekonomii, zwłaszcza w teorii zachowań konsumentów i teoriach użyteczności, bardzo ważną rolę. Relacja preferencji jest podstawą badań indywidualnych decyzji (wyborów) konsumentów (popytu indywidualnego), co umożliwia z kolei analizę popytu rynkowego. Bezpośredni pomiar użyteczności (rozumianej jako poziom zadowolenia czy satysfakcji konsumenta) jest zadaniem trudnym. W celu jego rozwiązania stosuje się metody zmierzające do kwantyfikacji użyteczności w oparciu na pojęciu preferencji. Preferencje są relacją binarną odnoszoną do wektorów opisujących wielowymiarowe obiekty, tzw. profile dóbr lub usług, których zbiory tworzą koszyki towarów lub plany konsumpcji. Relacja preferencji umożliwia przypisanie konsumentowi indywidualnej skali preferencji, na której można wartościować profile i optymalizować wybory rynkowe (zob. [Bąk 2004]).

W pomiarze preferencji konsumentów wykorzystuje się obserwacje historyczne oraz dane o charakterze antycypacyjnym opisujące intencje konsumentów. W związku z tym rozróżnia się metody analizy preferencji ujawnionych (historycznych) i metody analizy preferencji wyrażonych (deklarowanych).

Pomiar preferencji wyrażonych jest bardzo często przeprowadzany na skalach niemetrycznych (respondenci przeprowadzają ranking oferowanych produktów lub

dokonują wyborów). Wybory konsumentów można odwzorowywać na skali binarnej. Dane zgromadzone w ten sposób analizuje się za pomocą modeli z dychotomiczną (binarną) zmienną objaśnianą.

Celem artykułu jest wskazanie, na przykładach empirycznych, możliwości wykorzystania modeli z dychotomiczną zmienną objaśnianą w analizie preferencji konsumentów. Omówiono w nim podstawowe założenia estymacji modeli dwumianowych z wykorzystaniem koncepcji uogólnionych modeli liniowych i model klas ukrytych szacowany na podstawie binarnych zmiennych obserwowanych.

## 2. Mikroekonometria

Teorie użyteczności mieszczą się w obrębie mikroekonomii, natomiast metody badania preferencji można zaklasyfikować jako narzędzia badawcze mikroekonometrii. W metodach tych wykorzystuje się dane o jednostkowych obiektach badania, szczególnie o konsumentach i produktach, które nazywa się w literaturze przedmiotu mikrodanymi w celu podkreślenia ich szczegółowości. Metody badawcze stosowane w mikroekonometrii, w tym metody stosowane do pomiaru preferencji, umożliwiają „wydobycie" ukrytych w mikrodanych informacji mogących służyć wspomaganiu rynkowych procesów decyzyjnych i wyjaśnieniu zasad postępowania konsumentów.

Mikroekonometria jest dynamicznie rozwijającą się gałęzią ekonometrii, o czym świadczą m.in. wyróżnienia nagrodą Nobla jej wybitnych reprezentantów[1] badających mikrodane za pomocą metod wyborów dyskretnych. Do głównych cech wyróżniających mikroekonometrię należą (por. [Winkelmann, Boes 2006; Gruszczyński 2002; Hozer 1993]):

- badanie zachowań ekonomicznych jednostek (konsumentów, gospodarstw domowych, firm),
- analiza danych (mikrodanych) na poziomie indywidualnym (jednostkowym),
- niski poziom agregacji mikrodanych (dane szczegółowe),
- możliwość zaobserwowania zjawisk lub zdarzeń niewidocznych w danych zagregowanych,
- nieliniowy rozkład obserwacji oraz wykorzystywanie nieliniowych modeli i metod estymacji parametrów,
- niejednorodność obserwacji (heterogeniczność badanych jednostek),
- duża liczba obserwacji (masowość mikrodanych),
- przekrojowy charakter mikrodanych.

W modelach wykorzystywanych w mikroekonometrii występują najczęściej następujące typy zmiennych objaśnianych:

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

b) zmienne politomiczne (wielokategorialne, np. wybór z wielu opcji):

[1] Nagrodę Banku Szwecji w dziedzinie ekonomii za rok 2000 otrzymali James J. Heckman i Daniel L. McFadden zajmujący się mikroekonometrią, analizą mikrodanych i metodami wyborów dyskretnych.

– zmienne o kategoriach uporządkowanych,
– zmienne o kategoriach nieuporządkowanych;

c) zmienne ograniczone:
– zmienne cenzurowane,
– zmienne ucięte;

d) zmienne licznikowe (wartości reprezentowane przez nieujemne liczby całkowite).

Realizacje zmiennych (mikrodane) są najczęściej wynikami pomiarów na skalach niemetrycznych (zgromadzone obserwacje są zwykle liczbowymi wartościami dyskretnymi lub symbolami).

Do najczęściej stosowanych w badaniach empirycznych modeli mikroekonometrycznych należą:

a) modele dwumianowe:
– modele liniowe prawdopodobieństwa,
– modele logitowe i probitowe,
– modele komplementarne log-log,
– modele log-liniowe (tablice kontyngencji);

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

c) modele klas ukrytych;

d) modele przeżycia (trwania);

e) modele zmiennych ograniczonych.

W artykule przedstawiono przykłady wykorzystania wybranych modeli dwumianowych (liniowego modelu prawdopodobieństwa, modelu logitowego i probitowego oraz modelu komplementarnego log-log) i modelu klas ukrytych dla danych binarnych w analizie danych o preferencjach wyrażonych, pochodzących z badań ankietowych.

## 3. Estymacja modeli dwumianowych w programie R

W celu estymacji modeli dwumianowych wykorzystuje się koncepcję uogólnionych modeli liniowych (GLM – *Generalized Linear Models*). Formułuje się model regresji, w którym wartość oczekiwana rozkładu zmiennej objaśnianej (preferencji konsumentów) $\mu$ zależy od liniowej kombinacji zmiennych objaśniających:

$$E(\mathbf{y}) = \mu = \mathbf{x}\boldsymbol{\beta}. \tag{1}$$

Rozkład wektora obserwacji $\mathbf{y}$ jest charakteryzowany przez wartość oczekiwaną $\mu$. Jeżeli obserwacje są niezależne i mają rozkład $N(\mu, \sigma^2)$, to parametry $\boldsymbol{\beta}$ są szacowane metodą najmniejszych kwadratów. Uzyskuje się w efekcie składnik systematyczny modelu GLM:

$$\eta = \mathbf{x}\boldsymbol{\beta} \tag{2}$$

nazywany predyktorem liniowym (*linear predictor*) $\eta$.

Rozkład wektora obserwacji **y** zależy m.in. od skali pomiaru i często nie jest to rozkład normalny, ale np. dwumianowy lub Poissona w przypadku skal niemetrycznych. Reprezentację różnych rozkładów wektora obserwacji **y** (w szczególności z rodziny rozkładów wykładniczych) w uogólnionym modelu liniowym umożliwia funkcja łącząca (*link function*) *g*, która wiąże wartość oczekiwaną $\mu$ z predyktorem liniowym $\eta$:

$$g(\mu) = \eta = \mathbf{x}\boldsymbol{\beta}. \tag{3}$$

Odwrotność funkcji łączącej:

$$g^{-1}(\eta) = \mu \tag{4}$$

jest nazywana funkcją średniej.

Do podstawowych zalet uogólnionych modeli liniowych zalicza się następujące cechy:

- w modelu mogą występować zmienne dyskretne i zmienne ciągłe,
- w modelu można uwzględnić wiele zmiennych objaśnianych,
- można stosować transformacje liniowe zmiennych o rozkładach nieliniowych,
- można uwzględnić rozkłady innych niż normalny, głównie rozkłady dyskretne,
- można szacować model w przypadku współliniowości zmiennych (gdy nie istnieje macierz odwrotna macierzy $\mathbf{X}'\mathbf{X}$, to wyznacza się tzw. uogólnioną macierz odwrotną).

**Tabela 1.** Wybrane funkcje łączące, ich odwrotności i transformacje zmiennych

| Rozkład (z rodziny rozkładów wykładniczych) | Funkcja łącząca $\eta = g(\mu)$ | Odwrotność funkcji łączącej $\mu = g^{-1}(\eta)$ | Transformacja zmiennej |
|---|---|---|---|
| Normalny | $\mu$ | $\eta$ | tożsamościowa (liniowy model prawdopodobieństwa) |
| Dwumianowy | $\log_e \frac{\mu}{1-\mu}$ | $\frac{1}{1+e^{-\eta}}$ | logitowa (model logitowy) |
| Dwumianowy | $\Phi^{-1}(\eta)$ | $\Phi(\mu)$ | probitowa (model probitowy) |
| Dwumianowy | $\log_e\left[-\log_e(1-\mu)\right]$ | $1-\exp\left[-\exp(\eta)\right]$ | log-log (komplementarny model log-log) |

Źródło: opracowano na podstawie pracy [Fox 2002].

W literaturze przedmiotu wskazuje się też na ograniczenia w zastosowaniach uogólnionych modeli liniowych, do których zalicza się: założenie o niezależności

obserwacji, nieadekwatność przyjętego rozkładu obserwacji, ograniczoną liczbę modeli nieliniowych, które mogą reprezentowane w modelu za pomocą funkcji łączącej (por. [Halekoh, Højsgaard 2008]).

W tabeli 1 zestawiono wybrane funkcje łączące i ich odwrotności umożliwiające transformacje zmiennych objaśnianych w modelach dwumianowych.

Estymację modeli dwumianowych przeprowadza się w programie R z wykorzystaniem koncepcji ogólnego modelu liniowego. W pakiecie `stats` oferowanym w wersji podstawowej programu R dostępna jest funkcja `glm(y~x,family)`, która umożliwia oszacowanie za pomocą funkcji łączących m.in. następujących modeli:

- liniowego modelu prawdopodobieństwa: `family=gaussian link="identity")`,
- modelu logitowego: `family=binomial(link="logit")`,
- modelu probitowego: `family=binomial(link="probit")`,
- komplementarnego modelu log-log: `family=binomial(link="cloglog")`.

Funkcje odwrotne do funkcji łączącej można uzyskać za pomocą klauzuli: `linkinv`.

## 4. Estymacja modeli klas ukrytych w programie R

Modele klas ukrytych uwzględniają heterogeniczność preferencji na poziome grupowym i znajdują zastosowanie w segmentacji konsumentów. Zakłada się, że w badanej próbie istnieje skończona liczba grup konsumentów o podobnych preferencjach. Między grupami natomiast występują istotne różnice. Grupy te nie są znane *a priori* (są „ukryte"), ponieważ nie jest znana ani przynależność poszczególnych konsumentów do określonych segmentów, ani liczba grup. Ogólną postać modelu ze zmiennymi ukrytymi (modelu klas ukrytych) reprezentuje zależność w postaci rozkładów warunkowych (por. [DeSarbo, Wedel 1994; Ramaswamy, Cohen 2000; Vriens 2001]):

$$f\left(\mathbf{y}\,|\,\mathbf{\Phi}\right)=\sum_{c=1}^{C}\pi_c f\left(\mathbf{y}\,|\,\mathbf{\theta}_c\right), \tag{5}$$

gdzie: $f\left(\mathbf{y}\,|\,\mathbf{\Phi}\right)$ – funkcja rozkładu obserwacji (np. preferencji konsumentów); $\sum_{c=1}^{C}\pi_c$ – rozkład prawdopodobieństw bezwarunkowych wyrażających przynależności do poszczególnych klas ukrytych (reprezentuje w modelu rozkład zmiennej ukrytej); $f\left(\mathbf{y}\,|\,\mathbf{\theta}_c\right)$ – funkcja opisująca prawdopodobieństwa warunkowe (reprezentuje w modelu rozkład zmiennych obserwowanych lub zmiennej objaśnianej); $\mathbf{\Phi}=\left(\pi,\mathbf{\theta}\right)$ – wszystkie nieznane parametry modelu; $\mathbf{\theta}_c$ – wektor nieznanych parametrów w *c*-tej klasie (np. $\mu_c$ i $\sigma_c$ dla rozkładu normalnego).

Główne cechy modeli klas ukrytych są następujące:
- umożliwiają identyfikację klas na podstawie zmiennych obserwowanych lub zmiennej objaśnianej (segmentacja konsumentów na podstawie preferencji),
- zawierają jedną kategorialną zmienną ukrytą (liczba kategorii jest równa liczbie klas),
- podstawą klasyfikacji obserwacji do klas są oszacowane na podstawie modelu prawdopodobieństwa przynależności,
- zmienne zaobserwowane mogą być mierzone na różnych skalach,
- do modelu można włączyć zmienne towarzyszące i zmienne objaśniające (segmentacja z wykorzystaniem informacji o respondentach – geograficznych, demograficznych, kulturowych, społeczno-ekonomicznych, psychologicznych).

Szacowanie modelu klas ukrytych z niemetrycznymi zmiennymi zaobserwowanymi (zmienne zaobserwowane mogą być dychotomiczne lub politomiczne) można przeprowadzić w programie R z wykorzystaniem pakietu `poLCA` i funkcji `poLCA(model, dane, nclass=2)` (zob. [Linzer, Lewis 2006]).

## 5. Przykłady

W przykładzie 1 wykorzystano część danych z badania ankietowego przeprowadzonego w 2007 r. i dotyczącego sposobu odżywania się studentów dwóch jeleniogórskich szkół wyższych: Wydziału Gospodarki Regionalnej i Turystyki Akademii Ekonomicznej oraz Kolegium Karkonoskiego[2]. W badaniu zgromadzono 200 poprawnie wypełnionych kwestionariuszy ankietowych. W przykładzie wykorzystano:
- jako zmienną objaśnianą pytanie nr 5 z kwestionariusza ankiety:
  **5. Czy zwracasz uwagę na zawartość kaloryczną spożywanych posiłków?**
  □ tak
  □ nie
- jako zmienne objaśniające wybrane charakterystyki respondentów:

| | |
|---|---|
| Płeć | □ kobieta<br>□ mężczyzna |
| Rok urodzenia | [.....] |
| Wzrost w [cm] | [.....] |
| Waga w [kg] | [.....] |
| Rok studiów | [.....] |
| Stan cywilny | □ panna/kawaler<br>□ zamężna/żonaty<br>□ inny<br>[.....] |

[2] Dane na potrzeby pracy magisterskiej zgromadziła Anna Zjawińska.

Oszacowano trzy modele dwumianowe (logitowy, probitowy i komplementarny log-log) z dwiema zmiennymi objaśniającymi: niemetryczną zmienną „płeć" i metryczną zmienną „waga". Wyniki estymacji modeli przedstawia tab. 2 (kryterium informacyjne AIC wskazuje jako najlepszy model logitowy).

**Tabela 2.** Wyniki estymacji modeli dwumianowych z dwiema zmiennymi objaśniającymi („płeć" i „waga")

```
> mlogit

Call:  glm(formula = p5 ~ waga + plec, family = binomial(link =
"logit"))

Coefficients:
(Intercept)         waga        plecm
   -4.33684      0.06492     -2.69857

Degrees of Freedom: 199 Total (i.e. Null);  197 Residual
Null Deviance:      246
Residual Deviance: 226.5        AIC: 232.5
```

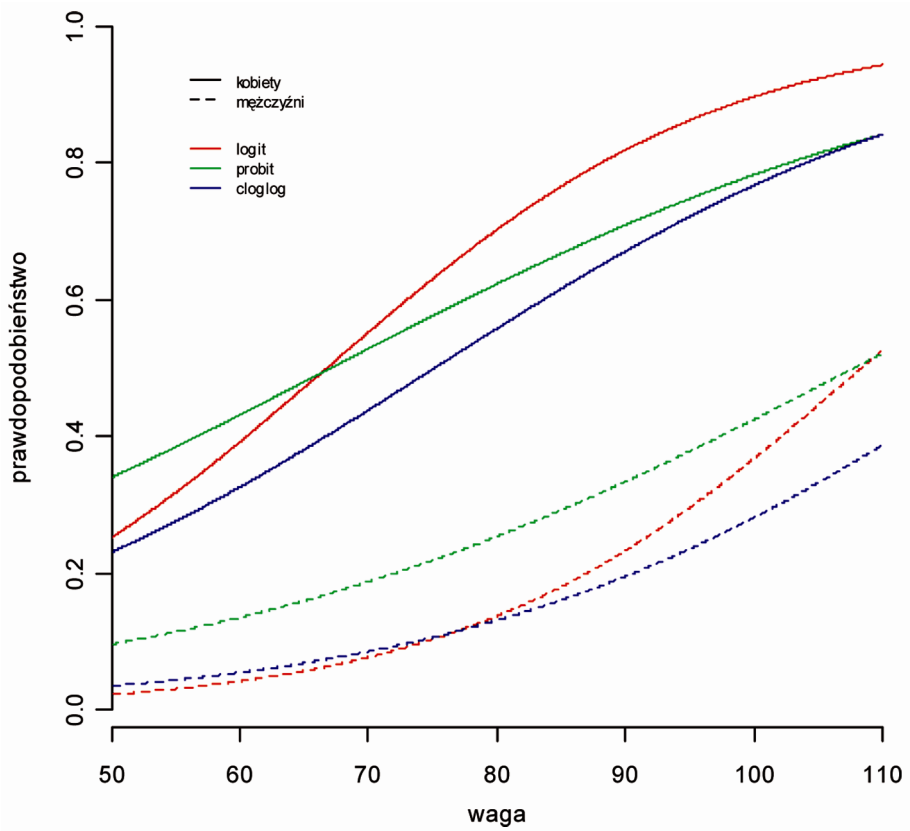
```
> mprobit

Call:  glm(formula = p5 ~ waga + plec, family = binomial(link =
"probit"))

Coefficients:
(Intercept)         waga        plecm
   -2.61152      0.03890     -1.58115

Degrees of Freedom: 199 Total (i.e. Null);  197 Residual
Null Deviance:      246
Residual Deviance: 226.6        AIC: 232.6
```

```
> mcloglog

Call:  glm(formula = p5 ~ waga + plec, family = binomial(link =
"cloglog"))

Coefficients:
(Intercept)         waga        plecm
   -3.60630      0.04796     -2.12786

Degrees of Freedom: 199 Total (i.e. Null);  197 Residual
Null Deviance:      246
Residual Deviance: 226.9        AIC: 232.9
```

Źródło: opracowanie własne z wykorzystaniem programu R.

Na rysunku 1 przedstawiono zależność prawdopodobieństwa odpowiedzi „tak" na pytanie 5 od cech „płeć" i „waga" w przekroju trzech modeli. Prawdopodobieństwo odpowiedzi „tak" kobiet jest wyraźnie większe niż mężczyzn i bardziej zależne od wzrostu wagi. Potwierdza to powszechne przekonanie, że kobiety przywiązują

większą wagę do sposobu odżywiania się, a przez to bardziej dbają o zdrowie. Wyniki estymacji wszystkich modeli zilustrowane na rys. 1 pozwalają też sformułować wniosek, że waga jest bardziej istotnym elementem samooceny wizerunku u kobiet niż u mężczyzn. Kobiety z nadwagą zwracają większą uwagę na kaloryczność posiłków niż mężczyźni. Niemniej jednak waga jest czynnikiem, który wpływa na ocenę kaloryczności spożywanych posiłków w przypadku obu płci w badanej próbie.




**Rys. 1.** Prawdopodobieństwa odpowiedzi „tak" w zależności od płci i wagi (modele: logitowy, probitowy i komplementarny log-log)

Źródło: opracowanie własne z wykorzystaniem programu R.

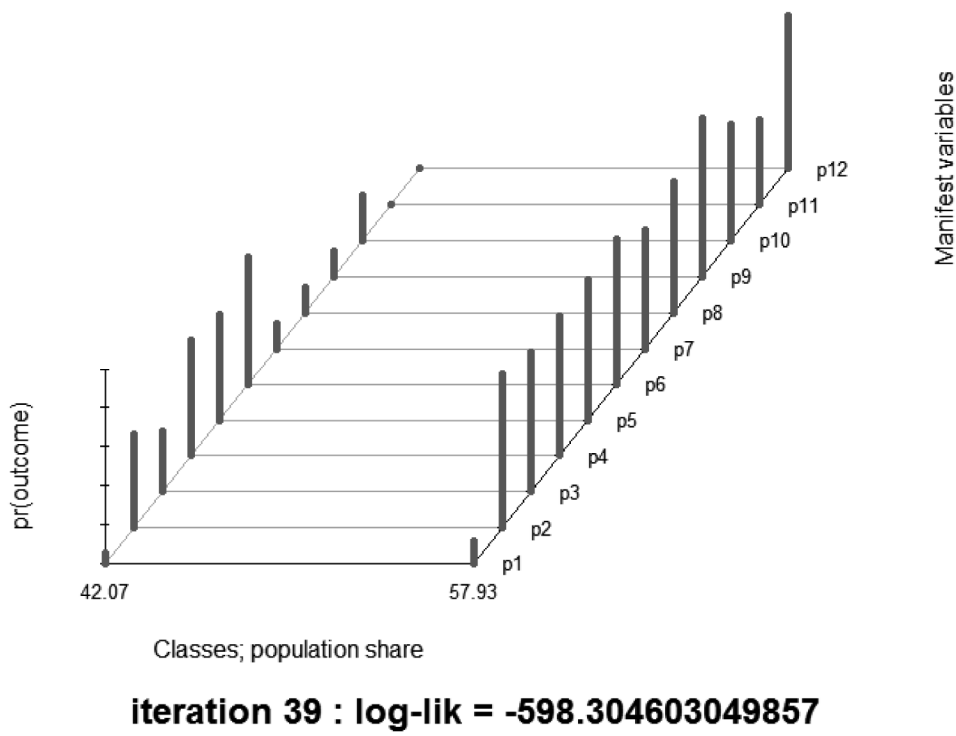
W przykładzie 2 wykorzystano część danych z badania ankietowego przeprowadzonego w 2008 r. i dotyczącego motywacji w zakresie dokonywanych zakupów roślin, jakimi kierują się osoby urządzające ogrody[3]. W badaniu zgromadzono 88 poprawnie wypełnionych kwestionariuszy ankietowych (dane zbierane były

[3] Dane na potrzeby pracy magisterskiej zgromadziła Marta Waliszkiewicz.

wyłącznie za pośrednictwem kwestionariusza zamieszczonego na stronie internetowej). W przykładzie wykorzystano pytania z części II kwestionariusza ankiety, na które odpowiedzi były mierzone na skali dychotomicznej:

CZĘŚĆ II. PREFERENCJE ZAKUPU (*Proszę wstawić znak X w odpowiednie pole*)

| Lp. | Czy kupuje Pani/Pan rośliny ogrodowe i inne produkty? | TAK | NIE |
|---|---|---|---|
| 1 | trawa z rolki | | |
| 2 | nasiona trawy | | |
| 3 | skalniaki (np. rozchodnik, macierzanka) | | |
| 4 | kwiaty doniczkowe (np. orchidee) | | |
| 5 | kwiaty balkonowe (np. surfinie, pelargonie) | | |
| 6 | kwiaty gruntowe (np. astry, piwonie) | | |
| 7 | krzewy liściaste owocowe (np. porzeczki) | | |
| 8 | krzewy liściaste ozdobne (np. forsycja, barwinek) | | |
| 9 | krzewy iglaste (np. jałowiec, cis, cyprys) | | |
| 10 | drzewa liściaste owocowe (np. jabłoń, grusza) | | |
| 11 | drzewa liściaste ozdobne (np. wierzba, klon) | | |
| 12 | drzewa iglaste (np. świerk, jodła) | | |




**Rys. 2.** Wyniki segmentacji respondentów

Źródło: opracowanie własne z wykorzystaniem programu R (pakietu `poLCA`).

Pytanie te zostały uwzględnione w modelu klas ukrytych jako zmienne obserwowane i stanowiły one podstawę segmentacji respondentów. Na podstawie analizy wyników przeprowadzonych obliczeń przyjęto ostatecznie podział badanej próby respondentów na dwie klasy. Ilustrację przyjętego modelu przedstawia rys. 2.

Wysokość słupków oznacza prawdopodobieństwo odpowiedzi „tak" na każde z 12 pytań. Uwzględniając cechy demograficzne, można powiedzieć, że klasa pierwsza obejmuje respondentów nieposiadających działek lub ogrodów (mieszkania w starej części miasta lub na dużych osiedlach mieszkaniowych bez ogródków działkowych), natomiast w klasie drugiej znajdują się w większości właściciele domów jednorodzinnych z ogrodami przydomowymi. Są też rośliny (np. kwiaty doniczkowe, balkonowe i gruntowe), których prawdopodobieństwa zakupu są zbliżone w obu klasach.

## 6. Podsumowanie

Modele zmiennych dychotomicznych umożliwiają analizę mikrodanych ankietowych (m.in. reprezentujących preferencje konsumentów). Wyboru modelu można dokonać na podstawie kryterium informacyjnego (najczęściej wykorzystuje się AIC). W przytoczonych przykładach modele logitowe charakteryzowały się najniższymi wartościami kryterium AIC.

Na podstawie mikrodanych (w zamieszczonym przykładzie – danych binarnych) o preferencjach można przeprowadzać segmentację konsumentów. Interpretację wyników segmentacji ułatwia analiza cech demograficznych respondentów.

Program R zawiera pakiety i funkcje, które można wykorzystać w celu zarówno analizy mikrodanych o preferencjach, jak i wizualizacji wyników.

W dalszych badaniach należy uwzględnić politomiczne zmienne objaśniane i zmienne towarzyszące w modelach klas ukrytych.

## Literatura

Bąk A., *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1013, Seria Monografie i Opracowania nr 157, AE, Wrocław 2004.

DeSarbo W.S., Wedel M., *A review of recent developments in latent class regression models*, [w:] R.P. Bagozzi (red.), *Advanced Methods of Marketing Research*, Blackwell, Cambridge 1994, s. 352-388.

Fox J., *An R and S-PLUS Companion to Applied Regression*, SAGE Publications, Thousand Oaks 2002.

Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa 2002.

Halekoh U., Højsgaard S., *Generalized Linear Models (GLM)*, 2008, http://gbi.agr sci.dk/ statistics/courses/Rcourse-DJF2008/, 25.01.2010.

Hozer J., *Mikroekonometria. Analizy, diagnozy, prognozy*, PWE, Warszawa 1993.

Linzer D.A., Lewis J., *poLCA: Polytomous Variable Latent Class Analysis*, http://dlinzer.bol.ucla.edu/polka, 2006.

Ramaswamy V., Cohen S.H., *Latent Class Models for Conjoint Analysis*, [w:] A. Gustafsson, A. Herrmann, F. Huber (red.), *Conjoint Measurement: Methods and Applications,* Springer, Berlin 2000, s. 361-392.

Vriens M., *Market Segmentation. Analytical Developments and Application Guidelines*. Millward Brown IntelliQuest, 2001.

Winkelmann R., Boes S., *Analysis of Microdata*, Springer-Verlag, Berlin, Heidelberg 2006.

## AN APPLICATION OF MICROECONOMETRIC DICHOTOMOUS VARIABLES MODELS IN THE RESEARCH OF PREFERENCES AND THEIR ESTIMATION USING R PROGRAM

**Summary:** The main aim of the paper is presentation, with empirical examples, possibilities to use models with dichotomous dependent variable in the consumer preferences analysis. There are presented basic approaches of the dichotomous models estimation using generalized linear models and latent class model estimated on the base of the binary manifest variables.