PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Bąk
Uniwersytet Ekonomiczny we Wrocławiu

MODELE WYBORÓW DYSKRETNYCH I ICH ESTYMACJA W PROGRAMIE R

1. Wstęp

W badaniach preferencji znajdują zastosowanie dwie grupy metod dekompozycyjnych: metody *conjoint analysis* i metody wyborów dyskretnych. W tej drugiej grupie wykorzystuje się modele probabilistyczne opisujące prawdopodobieństwa wyboru profilów z oferowanego zbioru różnych wariantów produktów lub usług. Na wybór poszczególnych profilów wpływają zarówno ich atrybuty, jak i charakterystyki respondentów. Zmienne te mają najczęściej charakter dyskretny (są to kategorie i zmienne nominalne). Celem estymacji modeli wyborów dyskretnych jest oszacowanie prawdopodobieństw wyboru poszczególnych opcji (profilów) i parametrów wskazujących znaczenie atrybutów.

Program **R** nie oferuje pakietu bezpośrednio wspierającego badania preferencji za pomocą metod wyborów dyskretnych. Procedury obliczeniowe zawarte w różnych pakietach mogą być jednak wykorzystane tego typu badaniach. Celem artykułu jest prezentacja procedury estymacji modelu wyborów dyskretnych z wykorzystaniem funkcji dostępnych w wybranych pakietach programu **R** oraz funkcji napisanych w języku, umożliwiających realizację zadań, które nie są aktualnie oprogramowane. W tekście przedstawiono następujące zagadnienia:

– charakterystykę warunkowego modelu logitowego,
– pakiety i procedury obliczeniowe dostępne w programie **R**, które mogą znaleźć zastosowanie w badaniach preferencji z wykorzystaniem modeli wyborów dyskretnych,
– funkcje napisane w języku programowania **R** realizujące zadania obliczeniowe niewspomagane w aktualnie dostępnych pakietach,
– przykład zastosowania funkcji programu **R** w estymacji modelu wyborów dyskretnych.

2. Warunkowy model logitowy

W badaniach wyborów dyskretnych, w których zmienne objaśniające charakteryzują opcje wyboru (np. są to atrybuty opisujące produkty tworzące zbiór profilów), wykorzystywany jest warunkowy model logitowy (CLM – *Conditional Logit Model*) zaproponowany przez McFaddena [McFadden 1974; Agresti 2002, s. 298--300]. W modelu tym wartości zmiennych objaśniających (atrybutów) różnią się względem opcji wyboru, wartości parametrów są natomiast takie same dla wszystkich opcji wyboru (np. profilów produktów), co odróżnia ten model od wielomianowego modelu logitowego (zob. [Long 1997, s. 178]). W warunkowym modelu logitowym prawdopodobieństwo wyboru *i*-tego profilu ze zbioru opcji liczącego *n* elementów jest szacowane na podstawie zależności (por. [So, Kuhfeld 2005, s. 469; Long 1997, s. 151-183]):

$$P_{ki} = P\left(y_k = i\right) = \frac{\mathrm{e}^{v_i}}{\sum_{l=1}^{n} \mathrm{e}^{v_l}} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{ki})}{\sum_{l=1}^{n} \exp(\boldsymbol{\beta}^T \mathbf{x}_{kl})}, \tag{1}$$

gdzie: $\mathbf{x}_{ki}$ – *k*-ty wektor macierzy $\mathbf{X}$ opisujący *i*-tą opcję (np. *i*-ty profil wybrany przez *k*-tego respondenta); $\mathbf{x}_{kl}$ – *k*-ty wektor macierzy $\mathbf{X}$ opisujący *l*-ty profil; $\boldsymbol{\beta}$ – wektor parametrów (wartość β_j jest związana z *j*-tym atrybutem opisującym profile).

Prawdopodobieństwa wyboru między dwoma profilami *i* oraz *l* można przedstawić w postaci równania logitowego:

$$\log\left(\frac{P_{ki}}{P_{kl}}\right) = \boldsymbol{\beta}^T \left(\mathbf{x}_{ki} - \mathbf{x}_{kl}\right), \tag{2}$$

w którym lewa strona (logit) przedstawia logarytm ilorazu szans[1] zależny tylko od różnicy między wartościami atrybutów opisujących profile *i* oraz *l* postrzeganej przez *k*-tego respondenta. Na wybór między profilami *i* oraz *l* nie wpływają inne opcje oferowane w zbiorze profilów. Dlatego model ten może być stosowany tylko wtedy, gdy opcje wyboru nie są ze sobą powiązane (zob. m.in. [McFadden 1974; Agresti 2002, s. 299]).

Oszacowania parametrów warunkowego modelu logitowego (modelu wyborów dyskretnych) można uzyskać za pomocą metody częściowej wiarygodności stosowanej w estymacji modelu proporcjonalnego hazardu Coxa. Metoda estymacji modeli proporcjonalnego hazardu Coxa jest oprogramowana w wielu pakietach statystycznych i dlatego często wykorzystywana jest także w celu estymacji warunkowych modeli logitowych [So, Kuhfeld 2005].

[1] Iloraz szans oznacza tutaj stosunek prawdopodobieństwa wyboru profilu *i* do prawdopodobieństwa wyboru profilu *l*.

3. Pakiety i funkcje programu R

Metody wyborów dyskretnych znajdują zastosowanie na gruncie badań marketingowych, głównie w analizie preferencji konsumentów. W praktyce stosowana jest procedura badawcza przedstawiona w tab. 1.

Obecnie nie jest oferowany pakiet programu **R** wspomagający w sposób kompleksowy badania empiryczne (w tym badania preferencji) z wykorzystaniem metod wyborów dyskretnych. Niemniej jednak różne pakiety (opracowane w innych celach) zawierają funkcje, które mogą być wykorzystane w zadaniach obliczeniowych na wybranych etapach procedury badawczej prowadzonej metodami wyborów dyskretnych. Nie wszystkie etapy tej procedury wymagają stosowania metod obliczeniowych i korzystania z oprogramowania komputerowego.

Tabela 1. Procedura badawcza z zastosowaniem metod wyborów dyskretnych

Lp.	Etap procedury	Krok procedury
1	specyfikacja zadania badawczego	zmienna objaśniana (np. preferencje) zmienne objaśniające (atrybuty)
2	gromadzenie danych	metody gromadzenia danych (pomiar preferencji, dane symulacyjne) metody generowania profilów (układy czynnikowe, próba losowa) przygotowanie zbioru danych
3	prezentacja profilów	forma prezentacji (opis słowny, rysunek, model, produkt fizyczny) forma badań (wywiad bezpośredni, poczta, telefon, komputer, Internet)
4	kodowanie atrybutów za pomocą zmiennych sztucznych	kodowanie zero-jedynkowe kodowanie *quasi*-eksperymentalne kodowanie ortogonalne
5	estymacja modelu	probabilistyczny wielomianowy (warunkowy) model logitowy metoda częściowej wiarygodności (*partial likelihood method*)*
6	analiza i interpretacja wyników	analiza preferencji – szacowanie użyteczności całkowitych profilów analiza preferencji – szacowanie prawdopodobieństw wyboru profilów wykresy funkcji użyteczności i ilorazów hazardu

*Metoda estymacji modeli semiparametrycznych zaproponowana przez D.R. Coxa (zob. [Frątczak 1997, s. 91]).

Źródło: opracowano na podstawie: [Bąk 2004, s. 109].

W tabeli 2 zestawiono te etapy procedury (zob. tab. 1), w których wykorzystanie metod obliczeniowych i algorytmów komputerowych jest niezbędne, oraz wskazano na pakiety i funkcje programu **R**, które można w tych celach wykorzystać (zob. [Everitt, Hothorn 2006; Wheeler 2004; Therneau 1999; Linzer, Lewis 2007; 2008]).

Z zestawienia wynika, że przygotowanie zbioru danych oraz szacowanie użyteczności całkowitych profilów i szacowanie prawdopodobieństw wyboru profilów nie jest oprogramowane w obecnie dostępnych pakietach **R**.

W tabeli 3 zestawiono funkcje programu **R** (pochodzące z różnych pakietów), które można wykorzystać w celu generowania danych symulacyjnych (danych o

Tabela 2. Etapy procedury wyborów dyskretnych w programie **R**

Etap procedury	Krok procedury	Wybrane pakiety i funkcje programu **R**
Gromadzenie danych	metody gromadzenia danych – dane symulacyjne (wybór profilów)	pakiet `stats` (funkcja `rmultinom`) pakiet `poLCA` (funkcja `rmulti`)
Gromadzenie danych	metody generowania profilów – układy czynnikowe	pakiet `AlgDesign` (funkcje: `gen.factorial`, `optFederov`)
Gromadzenie danych	przygotowanie zbioru danych	brak
Estymacja modelu	kodowanie atrybutów i szacowanie parametrów warunkowego model logitowego	pakiet `stats` (funkcje: `contrasts`, `lm`) pakiet `survival` (funkcje `coxph`, `Surv`, `strata`)
Analiza i interpretacja wyników	szacowanie użyteczności całkowitych profilów	brak
Analiza i interpretacja wyników	szacowanie prawdopodobieństw wyboru profilów	brak
Analiza i interpretacja wyników	wykresy funkcji użyteczności i ilorazów hazardu	pakiet `graphics` (funkcje: `plot`, `barplot`, `abline`)

Źródło: opracowanie własne.

preferencjach) i układów czynnikowych (zbiorów profilów), estymacji modelu (kodowania atrybutów niemetrycznych, szacowania parametrów), analizy i interpretacji wyników (ilustracji graficznej).

Oferowane aktualnie pakiety programu **R** nie zawierają funkcji wspierających takie elementy procedury wyborów dyskretnych, jak:

1. Przygotowanie zbioru danych o strukturze umożliwiającej oszacowanie warunkowego modelu logitowego z wykorzystaniem modelu proporcjonalnego hazardu Coxa.

2. Analiza i interpretacja wyników w zakresie szacowania użyteczności całkowitych profilów *i*. Ocena użyteczności całkowitych wynika z zależności: $y_k = \boldsymbol{\beta}^T \mathbf{x}_{ki}$.

3. Analiza i interpretacja wyników w zakresie szacowania prawdopodobieństw wyboru profilów. Prawdopodobieństwa wyboru profilów są szacowane na podstawie zależności (1).

W celu realizacji tych zadań opracowano funkcje w języku **R**, których kody źródłowe zamieszczono w postaci programów 1-5. Przeznaczenie i sposób użycia tych funkcji są następujące:

– `tabdat(W, M, m, n, p)` – funkcja zwracająca tablicę danych o strukturze umożliwiającej estymację warunkowego modelu logitowego na podstawie modelu Coxa; `W` – wektor z numerami profilów wybranych przez poszczególnych respondentów; `M` – macierz układu czynnikowego ze zmiennymi sztucznymi; `m` – liczba atrybutów; `n` – liczba respondentów; `p` – liczba profilów; wywołanie – `T <- tabdat(W, M, m, n, p)`,

Tabela 3. Opis wybranych funkcji programu R wykorzystywanych w metodach wyborów dyskretnych*

1. Generowanie danych symulacyjnych (wybór profilów)	
`rmultinom(n, size, prob)`	

`N`
`size`
`prob`
	liczba losowanych wektorów
	zakres, z którego losowane są elementy wektorów
	prawdopodobieństwa wylosowania elementów

`rmulti(p)`	
`P`	prawdopodobieństwa wylosowania elementów próby
2. Generowanie profilów – układy czynnikowe	
`Gen.factorial(levels, nVars=0, factors="none")`	

`levels`
`nVars`
`factors`
	liczby poziomów atrybutów
	liczba atrybutów
	sposób oznaczania poziomów atrybutów

`optFederov(frml, data)`	

`frml`
 	
`data`
	argument reprezentujący czynniki układu przechowywane w argumencie `data`; jeśli wykorzystuje się wszystkie czynniki z `data`, to należy użyć wyrażenia `~ .`
	zbiór profilów kandydujących w postaci macierzy (kolumny – atrybuty, wiersze – profile)

3. Estymacja modelu – kodowanie atrybutów	
`options(contrasts=c("contr.sum", "contr.poly"))`	
`contr.sum` `contr.poly`	kodowanie quasi-eksperymentalne poziomów nieuporządkowanych kodowanie wielomianowe poziomów uporządkowanych
`lm(formula, data)`	
`formula` `data`	argument reprezentujący model dane; jeśli jest pominięty, to danymi są wartości zmiennych argumentu `formula`
4. Estymacja modelu – szacowanie parametrów warunkowego modelu logitowego	
`coxph(formula, data, method=c("efron","breslow","exact"))`	

`formula`
`data`
`method`
	argument reprezentujący model wyborów dyskretnych
	dane
	metoda estymacji parametrów

Surv(time, event)	
`time`	czas przeżycia – zmienna, której niższe wartości oznaczają wybór profilu
`event`	zmienna wskaźnikowa (status = 0 lub 1), której wartości 0 oznaczają profile ocenzurowane (obserwacje ucięte), a 1 oznacza zdarzenie (wybór profilu) w momencie `time`
strata(a)	
`a`	zmienna (lub zmienne) grupująca (np. zmienna reprezentująca numer respondenta)
5. Wykresy funkcji użyteczności i ilorazów hazardu	
`plot(x, type, ylab, xlab)`	

x
`type`
`ylab, xlab`
	współrzędne
	typ wykresu
	opis osi

`barplot (x, ylab, xlab)`	
x	współrzędne

* W opisie niektórych funkcji uwzględniono wybrane argumenty.

Źródło: opracowanie własne na podstawie dokumentacji programu **R**.

- `totalutils(B, M)` – funkcja obliczająca użyteczności całkowite profilów; `B` – wektor parametrów; `M` – układ czynnikowy ze zmiennymi sztucznymi; wywołanie – `U <- totalutils(B, M)`,
- `prob(B, M)` – funkcja obliczająca prawdopodobieństwa wyboru profilów na podstawie warunkowego modelu logitowego; `B` – wektor parametrów; `M` – układ czynnikowy ze zmiennymi sztucznymi; wywołanie – `P <- prob(B, M)`,
- `probs(P)` – funkcja malejąco sortująca prawdopodobieństwa wyboru profilów z przestawieniem numerów (etykiet) profilów; `P` – prawdopodobieństwa wyboru profilów; wywołanie – `Ps <- probs(P)`,
- `utilsgraph(B, M)` – funkcja rysująca wykres użyteczności całkowitych i prawdopodobieństw wyboru profilów; `B` – wektor parametrów; `M` – układ czynnikowy ze zmiennymi sztucznymi; wywołanie – `utilsgraph(B, M)`.

```
tabdat <- function(W, M, m, n, p)
{
T <- matrix(0, n*p, m+4) #macierz danych
nrprof <- c(1:p)       #numery profilów
s <- 1           #nr respondenta
k <- 1           #licznik od 1 do n*p
for(i in 1:n) {
   for(j in 1:p) {
   T[k, 1] <- s
   T[k, 2] <- ifelse(W[i]==nrprof[j], 1, 0)  #status (1-wbrany, 0-nie)
   T[k, 3] <- j          #nr profilu
   T[k, 4] <- (2 - T[k, 2])    #time (1-uwzględniony, 2-ucięty)
   for(l in 5:(m+4)) { T[k, l] <- M[j, l-4]     #zmienne sztuczne
   }
   k <- k + 1
   }
   s <- s + 1
}
Tnames <- c("resp", "status", "profil", "time", names(Z))
colnames(T) <- Tnames    #nazwy kolumn w macierzy danych
return(T)
}
```

Program 1. Funkcja `tabdat()`

Źródło: opracowanie własne.

```
totalutils <- function(B, M)
{ return(M %*% B) }
```

Program 2. Funkcja `totalutils()`

Źródło: opracowanie własne.

```
prob <- function(B, M)
{
expU <- exp(M %*% B); s <- sum(expU)
P <- expU / s
return(P)
}
```

Program 3. Funkcja `prob()`

Źródło: opracowanie własne.

```
probs <- function(P)
{
p <- length(P); Ps <- array(1:p, c(p, 2)); Ps[, 2] <- P
o <- order(Ps[, 2], decreasing = TRUE)
Ps <- cbind(Ps[o, 1], Ps[o, 2])
return(Ps)
}
```

Program 4. Funkcja `probs()`

Źródło: opracowanie własne.

```
utilsgraph <- function(B, M)
{
U <- totalutils(B, M); P <- prob(B, M)
o <- order(U); Us <- cbind(U[o], P[o])
plot(Us, type="b", ylab="prawdopodobieństwa", xlab="użyteczności")
}
```

Program 5. Funkcja `utilsgraph()`

Źródło: opracowanie własne.

1. Przykład

W przykładzie przedstawiono sposób użycia funkcji programu **R** zamieszczonych w tab. 3 oraz funkcji opracowanych w postaci programów 1-5 z wykorzystaniem danych hipotetycznych wygenerowanych losowo.

A. Generowanie układu czynnikowego (3 atrybuty po 2 poziomy):

```
Z <- gen.factorial(c(2, 2, 2), factors="all") #układ
czynnikowy pełny.
```

W wyniku otrzymano zbiór 8 profilów:

```
  X1 X2 X3
1  1  1  1
2  2  1  1
3  1  2  1
4  2  2  1
5  1  1  2
6  2  1  2
7  1  2  2
8  2  2  2
```

B. Kodowanie atrybutów za pomocą zmiennych sztucznych:

```
profil <- c(1:8)
ml = lm(profil ~ factor(X1) + factor(X2) + factor(X3),
data = Z)
M <- model.matrix(ml)# macierz układu czynnikowego ze
zmiennymi sztucznymi
M <- M[, 1:m+1]; colnames(M) <- names(Z).
```

W wyniku otrzymano zbiór 8 profilów (kodowanie zero-jedynkowe):

```
  X1 X2 X3
  1  0  0  0
  2  1  0  0
  3  0  1  0
  4  1  1  0
  5  0  0  1
  6  1  0  1
  7  0  1  1
  8  1  1  1
```

C. Generowanie wyboru profilów:

```
p1 <- c(0.1,0.2,0.1,0.1,0.2,0.1,0.1,0.1)
p2 <- matrix(p1,nrow=n,ncol=8,byrow=TRUE)
W <- rmulti(p2).
```

W wyniku otrzymano wektor numerów profilów (10 pierwszych elementów):

```
[1] 7 5 5 8 5 2 1 6 5 2 ...
```

D. Przygotowanie tablicy danych o strukturze wymaganej przez funkcję `coxph()`:

```
T <- tabdat(W, M, m, n, p).
```

W wyniku otrzymano macierz (10 pierwszych wierszy):

```
      resp    status profil time X1 X2 X3
 [1,]  1       0       1      2    0  0  0
 [2,]  1       0       2      2    1  0  0
 [3,]  1       0       3      2    0  1  0
 [4,]  1       0       4      2    1  1  0
 [5,]  1       0       5      2    0  0  1
 [6,]  1       0       6      2    1  0  1
 [7,]  1       1       7      1    0  1  1
 [8,]  1       0       8      2    1  1  1
 [9,]  2       0       1      2    0  0  0
[10,]  2       0       2      2    1  0  0
```

E. Szacowanie warunkowego modelu logitowego:

```
coxph(formula=Surv(time,  status)~X1+X2+X3+strata(resp),
data=T, method = "exact").
```

W wyniku otrzymano oszacowania parametrów:

```
n = 800
    coef   exp(coef)  se(coef)  z     p
X1  0.282  1.326      0.202     1.40  0.1600
X2 -0.575  0.563      0.208    -2.76  0.0057
X3 -0.405  0.667      0.204    -1.99  0.0470
   exp(coef)  exp(-coef)  lower .95  upper .95
X1 1.326      0.754       0.892      1.969
X2 0.563      1.778       0.374      0.846
X3 0.667      1.500       0.447      0.995
Rsquare= 0.017 (max possible= 0.405 )
Likelihood ratio test= 13.9 on 3 df, p=0.00299
Wald test = 13.5 on 3 df, p=0.00364
Score (logrank) test = 13.8 on 3 df, p=0.00319
```

F. Użyteczności całkowite profilów:

```
U <- totalutils(B, M).
```

W wyniku otrzymano wektor użyteczności:

```
    [,1]
1  0.0000000
2  0.2818512
3 -0.5753641
4 -0.2935130
5 -0.4054651
6 -0.1236140
7 -0.9808293
8 -0.6989781
```

G. Prawdopodobieństwa wyboru profilów (`P`) i prawdopodobieństwa uporządkowane (`Ps`):

```
P <- prob(B, M); Ps <- probs(P).
```

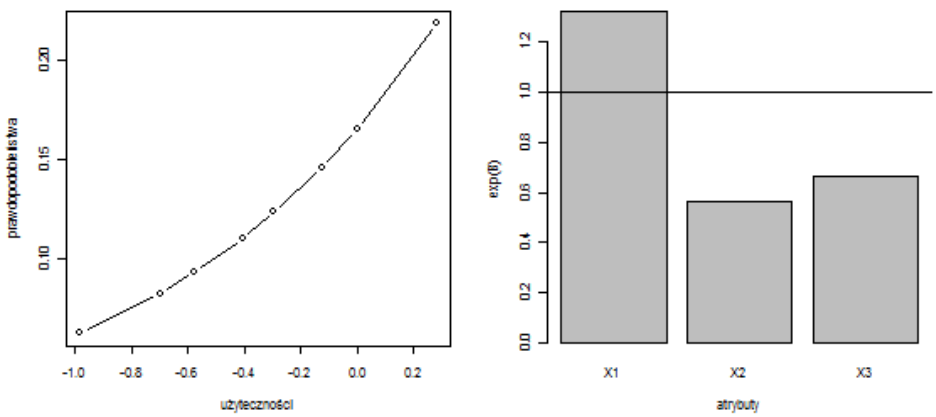
W wyniku otrzymano wektory prawdopodobieństw wyboru profilów:

```
  [,1]                      [,1]   [,2]
1 0.16512                [1,]  2    0.21888
2 0.21888                [2,]  1    0.16512
3 0.09288                [3,]  6    0.14592
4 0.12312                [4,]  4    0.12312
5 0.11008                [5,]  5    0.11008
6 0.14592                [6,]  3    0.09288
7 0.06192                [7,]  8    0.08208
8 0.08208                [8,]  7    0.06192
```

H. Wykresy funkcji użyteczności i ilorazów hazardu

```
utilsgraph(B, M)
barplot(exp(B), ylab="exp(B)", xlab="atrybuty"); abline(h=1, v=0, col=1).
```

W wyniku otrzymano wykresy:



Literatura

Agresti A. (2002), *Categorical data analysis*, second edition, Wiley, New York.

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

Everitt B.S., Hothorn T. (2006), *An handbook of statistical analysis using R. Chapman i Hall,* Boca Raton, London, New York.

Frątczak E. (1997), *Analiza historii zdarzeń – elementy teorii, wybrane przykłady zastosowań z wykorzystaniem pakietu TDA*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.

Linzer D.A., Lewis J. (2007), *poLCA: polytomous variable latent class analysis*, R package version 1.1, http://userwww.service.emory.edu/~dlinzer/poLCA (25.09.2008).

Linzer D.A., Lewis J. (2008), *poLCA: polytomous variable latent class analysis*, "Journal of Statistical Software" (w redakcji) (25.09.2008).

Long J.S. (1997), *Regression models for categorical and limited dependent variables*, SAGE Publications, Thousand Oaks-London-New Delhi.

McFadden D. (1974), *Conditional logit analysis of qualitative choice behavior*, [w:] *Frontiers in Econometrics*, red. P. Zarembka, Academic Press, New York-San Francisco-London, s. 105-142.

So Y., Kuhfeld W.F. (2005), *Multinomial logit models*, http://support.sas.com/ resources/papers/ tnote/tnote_marketresearch.html, SAS Institute, Cary (25.09.2008).

Therneau T.M. (1999), *A package for survival analysis in S*, Mayo Foundation.

Wheeler R.E. (2004), *AlgDesign. The R project for statistical computing,* http://www.r-project.org/ (25.09.2008).

DISCRETE CHOICE MODELS AND THEIR ESTIMATION IN R COMPUTER PROGRAMME

Summary

The aim of the paper is to present discrete choice models estimation procedure using **R** computer programme function and own function wrote in **R** computer language.

The paper presents:

– basic description of conditional logit model,
– packages and procedures available in **R** computer programme that are useful in discrete choice methods,
– own functions wrote in **R** computer language that are useful in some computing tasks not supported yet in standard **R** packages,
– the example of application **R** functions in estimation of discrete choice model.