4 Logit wielomianowy, warunkowy i zagnieżdżony

Celem tej części notetek jest pokazanie różnic miedzy wartościami parametrów modeli szacowanych przy utrzymanym założeniu o niezależności niezwiązanych alternatyw a modelem który nie wymaga takiego założenia. Dane wykorzystane w notatkach są danymi do podręcznika Cameron & Trivedi. Ich źródłem jest artykułu Herriges i King (1999). Przedmiotem badania jest wybór przez wędkarza sposobu łowienia ryb. Ma on do dyspozycji cztery możliwości:

- z brzegu,
- z pomostu
- z prywatnej łódki
- z czarterowanej łodki

W zbiorze są trzy zmienne objaśniające. Z tym, że dochód `income` jest cechą charakteryzującą decydenta, a cena `price` i prawdopodobieństwo złowienia ryby `crate` charakteryzują dostępne wybory.

```
. describe

Contains data from mus15data.dta
  obs:         1,182
 vars:            16                          26 Nov 2008 17:16
 size:        75,648
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
mode            float  %9.0g       modetype   Fishing mode
price           float  %9.0g                  price for chosen alternative
crate           float  %9.0g                  catch rate for chosen alternative
dbeach          float  %9.0g                  1 if beach mode chosen
dpier           float  %9.0g                  1 if pier mode chosen
dprivate        float  %9.0g                  1 if private boat mode chosen
dcharter        float  %9.0g                  1 if charter boat mode chosen
pbeach          float  %9.0g                  price for beach mode
ppier           float  %9.0g                  price for pier mode
pprivate        float  %9.0g                  price for private boat mode
pcharter        float  %9.0g                  price for charter boat mode
qbeach          float  %9.0g                  catch rate for beach mode
qpier           float  %9.0g                  catch rate for pier mode
```

```
qprivate        float  %9.0g                 catch rate for private boat mode
qcharter        float  %9.0g                 catch rate for charter boat mode
income          float  %9.0g                 monthly income in thousands $
-------------------------------------------------------------------------------
Sorted by:
```

Zbiór liczy 1182 indywidualne obserwacje, po jednej dla każdego decydenta. Zmienne `mode`, `price` oraz `crate` opisują odpowiednio wybrany sposób połowu ryb, cenę i prawdopodobieństwo złowienia ryby przy wybranym sposobie ich połowu. Kolejne cztery zmienne są zmiennymi zero-jedynkowymi wskazującymi wybrany przez respondenta sposób połowu, przyjmują wartość 1 dla wybranej metody oraz 0 w przeciwnym przypadku. Kolejne osiem zmiennych charakteryzuje dostępne wybory zawierając informacje o cenach i szansie złowienia ryby w dostępnych najbliższych łowiskach. Zmienne z prefiksem `p` opisują ceny, a z prefiksem `q` jakość mierzoną prawdopodobieństwem złowienia ryby. Wartości tych zmiennych uzyskano z badania ankietowego, w którym pytano nie tylko o charakterystyki wybranego łowiska, ale również o charakterystyki innych łowisk. Ostatnia zmienna w zbiorze dochód `income` charakteryzuje respondenta.

```
. summarize, separator(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mode	1182	3.005076	.9936162	1	4
price	1182	52.08197	53.82997	1.29	666.11
crate	1182	.3893684	.5605964	.0002	2.3101
dbeach	1182	.1133672	.3171753	0	1
dpier	1182	.1505922	.3578023	0	1
dprivate	1182	.3536379	.4783008	0	1
dcharter	1182	.3824027	.4861799	0	1
pbeach	1182	103.422	103.641	1.29	843.186
ppier	1182	103.422	103.641	1.29	843.186
pprivate	1182	55.25657	62.71344	2.29	666.11
pcharter	1182	84.37924	63.54465	27.29	691.11
qbeach	1182	.2410113	.1907524	.0678	.5333
qpier	1182	.1622237	.1603898	.0014	.4522
qprivate	1182	.1712146	.2097885	.0002	.7369
qcharter	1182	.6293679	.7061142	.0021	2.3101
income	1182	4.099337	2.461964	.4166667	12.5

Zmienna określająca typ łowiska `mode` przyjmuje wartości od 1 do 4. Wykorzystując polecenie `tabulate` możemy poznać częstotliwość wyboru poszczególnych opcji.

```
. tabulate mode
```

```
    Fishing |
       mode |      Freq.     Percent        Cum.
------------+-----------------------------------
      beach |         134       11.34       11.34
       pier |         178       15.06       26.40
    private |         418       35.36       61.76
    charter |         452       38.24      100.00
------------+-----------------------------------
      Total |       1,182      100.00
```

Udziały poszczególnych opcji wynoszą około 1/3 dla połowów z brzegu (połów z plaży i pomostu łącznie), z własnej łódki i łódki wynajmowanej.

4.1 Analiza rozkładu zmiennych

4.1.1 Zmienne specyficzne dla decydenta

Aby poznać rozkład zmiennej specyficznej dla decydenta wystarczy zbudować tabelę jednokierunkową, Opcja `contents` definiuje zawartość kolumn tabeli. Będzie to liczba obserwacji `N`, wartość średnia `mean` oraz odchylenie standardowe `sd` dochodu decydenta.

```
* Cechy indywidualne (dochód)
. table mode, contents(N income mean income sd income)

----------------------------------------------------
Fishing   |
mode      |    N(income)  mean(income)    sd(income)
----------+-----------------------------------------
    beach |          134      4.051617       2.50542
     pier |          178      3.387172      2.340324
  private |          418      4.654107      2.777898
  charter |          452      3.880899      2.050029
----------------------------------------------------
```

Przeciętnie osoby łowiące z pomostu mają najniższe dochody, a wykorzystujące do połowów prywatne łodzie najwyższe

4.1.2 Zmienne specyficzne dla wyborów

Poznanie rozkładu zmiennych specyficznych dla wyborów wymaga wykorzystania tabel dwukierunkowych. W pierwszej tabeli zaprezentowano średnie wartości cen dla każdego typu łowiska dostępnego dla każdego decydenta.

```
. * Cechy dostepnych wyborow (lowiska)
. table mode, contents(mean pbeach mean ppier mean pprivate mean pcharter) /*
*/ format(%6.0f)
----------------------------------------------------------------------------
Fishing   |
mode      |   mean(pbeach)     mean(ppier)  mean(pprivate)  mean(pcharter)
----------+-----------------------------------------------------------------
    beach |             36             36             98             125
     pier |             31             31             82             110
  private |            138            138             42              71
  charter |            121            121             45              75
----------------------------------------------------------------------------
```

Przeciętnie decydenci wybierają najtańszy dostępny sposób łowienia lub drugi z najtańszych.

W podobny sposób można wygenerować średnie wartości dla prawdopodobieństw udanego połowu na łowisku każdego typu, które jest dostępne decydentowi.

```
. * Prawdopodobieństwo złowienia ryby
. table mode, contents(mean qbeach mean qpier mean qprivate mean qcharter) format(%6.2f)
----------------------------------------------------------------------------
Fishing   |
mode      |   mean(qbeach)     mean(qpier)  mean(qprivate)  mean(qcharter)
----------+-----------------------------------------------------------------
    beach |           0.28            0.22            0.16            0.52
     pier |           0.26            0.20            0.15            0.50
  private |           0.21            0.13            0.18            0.65
  charter |           0.25            0.16            0.18            0.69
----------------------------------------------------------------------------
```

4.2 Modelowanie

4.2.1 Wielomianowy model logit

Oszacowania modelu wielomianowego logitowego posłużą jako punkt odniesienia dla innych wyników. W przypadku wielomianowego modelu logitowego można wykorzystać wyłącznie dochód jako zmienną objaśniającą, gdyż jest on jedyną charakterystyką opisującą decydenta.

```
Multinomial logistic regression                 Number of obs   =       1182
                                                LR chi2(3)      =      41.14
                                                Prob > chi2     =     0.0000
Log likelihood = -1477.1506                     Pseudo R2       =     0.0137

------------------------------------------------------------------------------
        mode |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
beach        |  (base outcome)
-------------+----------------------------------------------------------------
pier         |
      income |  -.1434029   .0532884    -2.69   0.007    -.2478463   -.0389595
       _cons |   .8141503    .228632     3.56   0.000     .3660399    1.262261
-------------+----------------------------------------------------------------
private      |
      income |   .0919064   .0406637     2.26   0.024     .0122069    .1716058
       _cons |   .7389208   .1967309     3.76   0.000     .3533352    1.124506
-------------+----------------------------------------------------------------
charter      |
      income |  -.0316399   .0418463    -0.76   0.450    -.1136571    .0503774
       _cons |   1.341291   .1945167     6.90   0.000     .9600457    1.722537
------------------------------------------------------------------------------
```

Niska wartość statystyki Pseudo-R2 wskazuje, że model jest w niewielkim stopniu dopasowany do zbioru danych. Mimo wszystko, model jest statystycznie istotny, o czy świadczy wartość p statystyki F.

Niewątpliwym problemem jest fakt, iż w ostatnim równaniu parametr przy jedynym regresorze okazał się być nieistotny statystycznie. Zweryfikujmy łączną istotność statystyczną zmiennej dochód (`income`).

```
. test income

 ( 1)  [beach]o.income = 0
 ( 2)  [pier]income = 0
 ( 3)  [private]income = 0
 ( 4)  [charter]income = 0
       Constraint 1 dropped

           chi2(  3) =   37.70
         Prob > chi2 =    0.0000
```

Wyniki testu wskazują, że zmienna jest łącznie istotna. Odrzucenie jednego ograniczenia jest spowodowane faktem, iż jedna z wartości zmiennej zależnej (połów z brzegu) została przyjęta jako kategoria odniesienia.

W celu dokonania interpretacji ilościowej parametrów warto jest przedstawić je w postaci ilorazów względnych szans (ryzyk). Parametry modelu są liniowe ze względu na szanse.

```
* Wyniki w postaci ilorazow wzglednych ryzyk
. mlogit mode income, rr baseoutcome(1) nolog

Multinomial logistic regression                 Number of obs   =       1182
                                                LR chi2(3)      =      41.14
                                                Prob > chi2     =     0.0000
```

```
Log likelihood = -1477.1506                     Pseudo R2       =     0.0137
------------------------------------------------------------------------------
        mode |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
beach        |  (base outcome)
-------------+----------------------------------------------------------------
pier         |
      income |   .8664049   .0461693    -2.69   0.007     .7804799    .9617896
       _cons |   2.257257    .516081     3.56   0.000     1.442013      3.5334
-------------+----------------------------------------------------------------
private      |
      income |   1.096262   .0445781     2.26   0.024     1.012282     1.18721
       _cons |   2.093675   .4118906     3.76   0.000     1.423808    3.078697
-------------+----------------------------------------------------------------
charter      |
      income |   .9688554    .040543    -0.76   0.450     .8925639    1.051668
       _cons |   3.823979   .7438278     6.90   0.000     2.611816    5.598715
------------------------------------------------------------------------------
```

Aby porównać oszacowania z wynikami innych alternatywnych modeli zapiszemy je pod nazwą MNL

```
* Zapisanie wynikow pod nazwa ML
. estimates store MNL
```

4.2.2 Warunkowy model logit

Przed przystąpieniem do szacowania parametrów modelu warunkowego musimy przekształcić zbiór danych do tzw. postaci długiej (*ang. long*). W forma szerokiej (*ang. wide*) jeden wiersz jest jedną obserwacją. W postaci długiej jeden wiersz zbioru danych opisuję jeden z dostępnych dla decydenta wyborów. Zatem, w przypadku danych dotyczących wyboru łowiska jedna obserwacja będzie zapisana w czterech wierszach.

Po pierwsze tworzymy identyfikator decydenta (`id`). Opcja `j()` tworzy zmienną `fishmode` o wartościach: beach, pier, private i charter. Jest to zmienna pomocnicza typu łańcuchowego zawierająca opis dostępnych wyborów. Następnie przekształcamy dane dotyczące wybranego sposobu łowienia (zmienne z prefiksem `d`), cen (zmienne z prefiksem `p`), oraz jakości (zmienne z prefiksem `q`). Opcja `i()` identyfikuje obserwacje, opcja `j` przy przekształcaniu na formę długą danych określa zmienną i opcjonalnie wartości, które mają być przekształcone. Opcja `string` informuje, że w opcji `j()` mogą pojawić się ciągi znaków zamiast liczb.

```
generate id=_n
reshape long d p q, i(id) j(fishmode beach pier private charter) string
```

```
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                     1182   ->    4728
Number of variables                  18   ->      10
j variable (4 values)                     ->   fishmode
xij variables:
              dbeach dpier ... dcharter   ->   d
              pbeach ppier ... pcharter   ->   p
              qbeach qpier ... qcharter   ->   q
-----------------------------------------------------------------------------
```

Zobaczmy jak teraz wygląda wiersz danych. Obserwacje są posortowane zgodnie z porządkiem leksykograficznym względem wartości zmiennej typ łowiska (`fishmode`).

```
browse in 1/4
```

id	fishmode	mode	price	crate	d	p	q	income	_est_MNL
1	beach	charter	182.93	.5391	0	157.93	.0678	7.083332	1
1	charter	charter	182.93	.5391	1	182.93	.5391	7.083332	1
1	pier	charter	182.93	.5391	0	157.93	.0503	7.083332	1
1	private	charter	182.93	.5391	0	157.93	.2601	7.083332	1

Wszystkie zmienne o wartościach specyficznych dla dostępnych wyborów przyjmują taką samą wartość dla czterech wartości zmiennej zależnej. Nie stanowi to problemu w przypadku zmiennej dochód (`income`), gdyż dochód respondenta jest niezależny od możliwych wyborów. Problem jest związany ze zmiennymi `mode`, `price`, `crate`. Po prostu usuńmy je ze zbioru.

```
drop mode price crate
```

Polecenie `asclogit` pozwala na oszacowanie parametrów warunkowego modelu logitowego ze zmiennymi objaśniającymi o wartościach specyficznych dla dostępnych wyborów (ang. *alternatives*). Zmienne niezależne wymienione po poleceniu `asclogit` są traktowane jako zmienne o wartościach specyficznych dla dostępnych wyborów. Opcja `alternatives` definiuje dostępne wybory, a opcja `casevars` wskazuje na zmienne o wartościach specyficznych dla dokonanych wyborów.

Zaletą tego wariantu modelu warunkowego jest możliwość uwzględnienia różnych zbiorów możliwych wyborów dla każdego decydenta oraz modelowania wskazania więcej niż jednego wyboru przez decydenta.

Oszacujemy parametry modelu warunkowego wyjaśniającego dokonany wybór sposobu połowu ryb wykorzystując informacje specyficzne dla łowiska cenę (`p`) oraz jakość (`q`), i specyficzny dla decydenta dochód (`income`), oraz

stałą. Tak jak w przypadku modelu wielomianowego deklarujemy łowienie z brzegu jako kategorię odniesienia, wykorzystując opcję `basealternative`.

```
/* McFadden's choice model */
* Conditional logit with alternative-specific and case-specific regressors
asclogit d p q, case(id) alternatives(fishmode) casevars(income) /*
*/ basealternative(beach) nolog

Alternative-specific conditional logit          Number of obs      =       4728
Case variable: id                               Number of cases    =       1182

Alternative variable: fishmode                  Alts per case: min =          4
                                                               avg =        4.0
                                                               max =          4

                                                  Wald chi2(5)     =     252.98
Log likelihood = -1215.1376                       Prob > chi2      =     0.0000

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
fishmode     |
           p |  -.0251166   .0017317   -14.50   0.000    -.0285106   -.0217225
           q |    .357782   .1097733     3.26   0.001     .1426302    .5729337
-------------+----------------------------------------------------------------
beach        |  (base alternative)
-------------+----------------------------------------------------------------
charter      |
      income |  -.0332917   .0503409    -0.66   0.508     -.131958    .0653745
       _cons |   1.694366   .2240506     7.56   0.000     1.255235    2.133497
-------------+----------------------------------------------------------------
pier         |
      income |  -.1275771   .0506395    -2.52   0.012    -.2268288   -.0283255
       _cons |   .7779593   .2204939     3.53   0.000     .3457992    1.210119
-------------+----------------------------------------------------------------
private      |
      income |   .0894398   .0500671     1.79   0.074    -.0086898    .1875694
       _cons |   .5272788   .2227927     2.37   0.018     .0906132    .9639444
------------------------------------------------------------------------------
```

W górnej części wydruku znajduje się informacja, że zbiór liczy 4728 obserwacji dotyczących 1182 wyborów. Parametry obok zmiennych modelu są łącznie istotne w sensie statystycznym.

Pierwsza tabela wartości współczynników prezentuje oszacowania dla zmiennych o wartościach specyficznych dla wyborów. Kolejne wiersze prezentują wybór bazowy, oraz oszacowania wartości współczynników modelu dla charakterystyk decydenta, które są niezależne od możliwych opcji wyboru.

Wszystkie parametry są łącznie istotne w sensie statystycznym. Dopasowanie modelu do danych jest dużo wyższe niż w przypadku wielomianowego modelu logitowego.

Współczynniki warunkowego modelu logitowego ze zmiennymi objaśniającymi specyficznymi dla dostępnych wyborów posiadają relatywnie prostą interpretację. Niech β_s będzie wartością współczynnika dla zmiennej objaśniającej o wartościach specyficznych dla jednego z możliwych wyborów s, x_{sik}. Niech p_{ij} oznacza prawdopodobieństwo wyboru możliwości j przez jednostkę i. Wówczas zmiana wartości x_{sik}, oznaczającej wartość zmiennej x_s dla jednostki i oraz wyboru k wynosi:

$$\frac{\partial p_{ij}}{\partial x_{sik}} = \begin{cases} p_{ij}(1-p_{ij})\beta_s & j = k \\ -p_{ij}p_{ik}\beta_s & j \neq k \end{cases}$$

Zatem, dla wartości β_s większej od zera efekt zmiany wartości zmiennej objaśniającej jest pozytywny. Oznacza to, że w przypadku wzrostu wartości tej zmiennej ta kategoria, która została wybrana będzie wybierana częściej.

Gdyby oszacowania wartości współczynników specyficznych dla wyborów nie różniły się od wartości zero w sensie statystycznym, to model zostałby zredukowany do wielomianowego modelu logitowego.

Wartości oszacowania dla parametrów specyficznych dla dostępnych wyborów mają interpretację analogiczną do interpretacji dla parametrów modelu logitowego.

Parametry modelu warunkowego można uzyskać również wykorzystując polecenie `clogit`. To polecenie jest podobne w konstrukcji do polecenia dla modelu panelowego logitowego (`xtlogit`). Oba wymagają grupowania danych.

Polecenie `clogit` nie pozwala na wprowadzenie do modelu zmiennych specyficznych dla możliwych wyborów. W zamian tego typu zmienne są dołączane poprzez interakcje ze stałymi specyficznymi dla możliwych wyborów.

Przed oszacowaniem warunkowego modelu logitowego należy utworzyć zmienne o wartościach specyficznych dla dostępnych wyborów. Postępowanie jest dwukrokowe. W pierwszym tworzone są zmienne wskazujące czy dana możliwość został wybrana. W drugim kroku otworzone zmienne wskazujące możliwość mnożone są przez wartość zmiennej określającej dochód (`income`).

```
. clogit d dbeach dprivate dcharter ybeach yprivate ycharter income, group(id)
note: income omitted because of no within-group variance.

Iteration 0:   log likelihood = -1492.5126
Iteration 1:   log likelihood = -1477.3262
Iteration 2:   log likelihood = -1477.1506
Iteration 3:   log likelihood = -1477.1506
```

```
Conditional (fixed-effects) logistic regression   Number of obs   =       4728
                                                  LR chi2(6)      =     322.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -1477.1506                       Pseudo R2       =     0.0985

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dbeach |     .81415   .2286319     3.56   0.000     .3660396     1.26226
    dprivate |    .738921   .1967309     3.76   0.000     .3533355    1.124507
    dcharter |   1.341292   .1945167     6.90   0.000     .9600459    1.722537
      ybeach |  -.1434028   .0532884    -2.69   0.007    -.2478462   -.0389595
    yprivate |   .0919063   .0406637     2.26   0.024     .0122068    .1716057
    ycharter |    -.03164   .0418463    -0.76   0.450    -.1136572    .0503773
      income |          0  (omitted)
------------------------------------------------------------------------------
```

Jak widać, jeżeli w modelu zostanie umieszczona zmienna, która przyjmuje identyczne wartości dla wszystkich dostępnych wyborów zostanie ona usunięta z szacowanego modelu. Oszacowane wartości współczynników przy zmiennych `d` obrazują relatywne prawdopodobieństwo wyboru wskazanego typu połowu w stosunku do połowu z pomostu (ang. *pier*). Dodatnia wartość współczynnika przy zmiennej `yprivate` wskazuje, że wraz ze wzrostem dochodu decydenta wzrasta prawdopodobieństwo, iż w celu połowu ryb wykorzysta on prywatną łódź.

Aby porównać oszacowania z wynikami innych modeli zapiszemy je pod nazwą CL

```
* Zapisanie wynikow pod nazwa CL
estimates store CL
```

4.2.3 Zagnieżdżony model logit

W literaturze opisywane są dwa warianty zagnieżdźonego modelu logitowego (ang. *nested logit*). Preferowany jest ten, który zostanie zaprezentowany w tych notatkach. Jest on zgodny z warunkiem maksymalizacji addytywnej funkcji losowej użyteczności (ARUM). Zagnieżdżony model logitowy, w odróżnieniu od modelu wielomianowego logitowego, czy warunkowego modelu logitowego nie wymaga spełnienia założenia o niezależności niezwiązanych wyborów (ang. *Independence of Irrelevant alternatives*). W zamian wymaga, by miedzy wyborami występowała zagnieżdżona struktura. Pozwala ona na występowanie niezerowej korelacji w ramach gniazda (grupy wyborów), ale gniazda (grupy wyborów) muszą być nieskorelowane.

W przypadku danych dotyczących sposobów rekreacyjnego połowu ryb można wskazać na fundamentalną różnicę między połowem z brzegu oraz łódki. Natomiast czy w ramach pierwszej możliwości zostanie wybrany brzeg czy pomost, a w ramach drugiej łódź wynajmowana czy prywatna może mieć dużo mniejsze znaczenie. Taka sytuacja skłania do utworzenia drzewa wyborów, w którym na pierwszym poziomie znajduje się decyzja czy połów odbywa się z brzegu czy wykorzystywana jest w tym celu łódka, natomiast na drugim poziomie wyboru znajduje się konkretny sposób połowu. Warto podkreślić, iż hierarchiczność wyboru nie jest warunkiem koniecznym wykorzystania modelu zagnieżdżonego. Jego zaletą jest fakt, iż pozwala na korelację pomiędzy dostępnymi wyborami w ramach gniazda, czyli na drugim poziomie.

Przed przystąpieniem do szacowania parametrów modelu zagnieżdżonego należy zadeklarować strukturę drzewa wyborów. W tym celu wykorzystywane jest polecenie `nlogitgen`. Po poleceniu wskazujemy nową nazwę zmiennej wskazującej na wybór i za pomocą znaku = przypisujemy jej dostępną w zbiorze danych zmienną określającą wybory. W nawiasach definiujemy rozdzielone przecinkiem gniazda. W ramach gniazda wybory oddzielane są znakiem "||". Elementy przez znakiem dwukropka ":" to etykiety opisujące gniazda. Ich definiowanie nie jest konieczne, jednak poprawia czytelność wyników. Cała składnia na następującą postać: (gniazdo 1: wybor1 | wybor2, gniazdo2: wybor3 | wybor 4).

Dla zbioru danych dotyczących wyboru łowiska przez decydenta drzewo wyboru konstruowane jest w następujący sposób:

```
* Definiowanie drzewa wyboru
. nlogitgen type = fishmode(shore:pier | beach, boat: private |
charter) new variable type is generated with 2 groups label list
lb_type lb_type:
           1 shore
           2 boat
```

Użytecznym poleceniem pomocniczym jest komenda `nlogittree`. Wyświetla ona na ekranie w trybie pesudograficznym strukturę zdefiniowanego drzewa wyborów. W opcji `choice` można zadeklarować zmienną zależną modelu. Wówczas oprogramowanie wyświetli częstotliwości wyboru każdej z dostępnych opcji.

```
* Sprawdzenie drzewa
nlogittree fishmode type, choice(d)

tree structure specified for the nested logit model

 type   N       fishmode  N    k
```

```
-----------------------------------
 shore 2364 --- beach    1182  134
             +- pier     1182  178
 boat  2364 --- charter  1182  452
             +- private  1182  418
-----------------------------------
                  total  4728 1182

k = number of times alternative is chosen
N = number of observations at each level
```

Po poleceniu `nlogit` wskazujemy zmienną zależną, a następnie wymieniamy zmienne o wartościach specyficznych dla dostępnych wyborów. Następnie po znaku "||" zapisujemy zmienne objaśniające wybory na pierwszym poziomie drzewa (pomiędzy gniazdami), a po kolejnym znaku "||" zmienne objaśniające wybory na drugim poziomie drzewa (wewnątrz gniazd). Opcja `case` definiuje identyfikator obserwacji.

Wyniki oszacowań ukazane poniżej uzyskano w Stata 11.1. W przypadku nowszych wersji pakietu statystycznego wartości parametrów mogą się nieznacznie różnić. By w nowszej wersji pakietu odtworzyć poniższe wyniki wystarczy przed uruchomieniem szacowania modelu wpisać polecenie `version 11.1`. Wówczas nowsza wersja pakietu Stata będzie zachowywać się tak jakby była wersją 11.1.

```
* Model
nlogit d p q || type:, base(shore) || fishmode:income, case(id)

RUM-consistent nested logit regression         Number of obs      =      4,728
Case variable: id                              Number of cases    =       1182

Alternative variable: fishmode                 Alts per case: min =          4
                                                              avg =        4.0
                                                              max =          4

                                                  Wald chi2(5)    =     212.37
Log likelihood = -1192.4236                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
fishmode     |
           p |  -.0267625   .0018937   -14.13   0.000    -.0304741    -.023051
           q |    1.34009   .3080485     4.35   0.000     .7363261    1.943854
------------------------------------------------------------------------------
fishmode equations
------------------------------------------------------------------------------
beach        |
```

```
      income |          0  (base)
       _cons |          0  (base)
-------------+------------------------------------------------------------------
charter      |
      income |  -8.402089   78.35381    -0.11   0.915    -161.9727    145.1686
       _cons |   69.96961   558.8719     0.13   0.900    -1025.399    1165.338
-------------+------------------------------------------------------------------
pier         |
      income |   -9.45814   80.30031    -0.12   0.906    -166.8438    147.9276
       _cons |   58.94376   500.7214     0.12   0.906    -922.4522     1040.34
-------------+------------------------------------------------------------------
private      |
      income |  -1.634909    8.58798    -0.19   0.849    -18.46704    15.19722
       _cons |   37.52499   230.8861     0.16   0.871    -415.0034    490.0533
------------------------------------------------------------------------------
dissimilarity parameters
------------------------------------------------------------------------------
type         |
  /shore_tau |   83.46907   718.5086                     -1324.782     1491.72
   /boat_tau |   52.56016    542.889                     -1011.483    1116.603
------------------------------------------------------------------------------
LR test for IIA (tau=1): chi2(2) = 45.43                 Prob > chi2 = 0.0000
```

Model zagnieżdżony redukuje się do modelu warunkowego jeżeli oba parametry niepodobieństwa τ są równe 1. Pod tabelą zaprezentowany jest wynik testu ilorazu wiarogodności weryfikującego taką hipotezę. Wskazuje on, iż celowe jest wykorzystanie modelu zagnieżdżonego.

Jeżeli model ma być zgodny z warunkiem maksymalizacji addytywnej funkcji losowej użyteczności (ARUM) to wartości parametrów τ powinny być między 0 a 1 i sumować się do jedności. Zatem uzyskane oszacowania parametrów wskazują, iż nie jest on zgodny z addytywną funkcją losowej użyteczności.

Oszacowana wartość parametru przy zmiennej `p` jest zbliżona do oszacowania uzyskanego z modelu warunkowego logitowego, natomiast w przypadku parametru przy zmiennej `q` różnica między modelami jest znaczna.

W celu porównania wyników zapamiętujemy je pod nazwą NL:

```
estimates store NL
```

Mając oszacowane przy modele, możemy porównać wartości ocen ich parametrów i dopasowanie do danych.

```
. estimates table MNL CL NL, keep(p q) stats(N ll aic bic) equation(1)

----------------------------------------------------
    Variable |    MNL           CL           NL
```

```
-------------+----------------------------------------
           p |                            -.02676254
           q |                               1.34009
-------------+----------------------------------------
           N |       1182         4728         4728
          ll | -1477.1506   -1477.1506   -1192.4236
         aic |  2966.3011    2966.3011    2404.8471
         bic |  2996.7509    3005.0687    2469.4597
-------------------------------------------------------
```

Na podstawie wartości wszystkich kryteriów można przyjąć iż najlepiej dopasowanym do danych spośród rozpatrywanych modeli jest zagnieżdżony model logitowy.

4.2.4 Wielomianowy model probit

- W przypadku modelu o zmiennych z wartościami specyficznymi dla wyborów szacowanie wartości parametrów trwa bardzo długo
- Przy szacowaniu wykorzystywana jest pseudo-funkcja wiarogodności, ponieważ problem nie ma rozwiązania analitycznego. Rozwiązanie jest znajdowane w sposób przybliżony z wykorzystaniem symulacji

4.2.5 Model logit o losowych parametrach

Model o losowych parametrach (ang. *random parameter model*) nazywany również modelem mieszanym (ang. *mixed model*) pozwala na opuszczenie założenia o niezależności niezwiązanych alternatyw (ang. *independence of irrelevant alternatives*). W zamian zakładane jest, że parametry modeli posiadają znany rozkład, najczęściej normalny lub log-normalny.

Rozwój modelowania w kierunku modeli o losowych parametrach nastąpił w latach 1990-tych. Punktem przełomowym był rozwój metod symulacyjnych, na przykład symulowanej funkcji największej wiarogodności. Pozwoliło to na numeryczne szukanie rozwiązań problemów obliczeniowych, które nie posiadają rozwiązania analitycznego. Dzięki temu szacowanie parametrów modeli takich jak wielominowy model probitowy czy model mieszany (model o losowych parametrach) jest relatywnie szybkie.

Symulowana funkcja wiarogodności W przypadku modeli, które nie posiadają rozwiązania analitycznego nie jest możliwe znalezienie optymalnej wartości funkcji wiarogodności. W takich przypadkach, jednym z możliwych rozwiązań tego problemu jest posłużenie się procedurą maksymalizacji symulowanej funkcji wiarogodności (ang. *maximum simulated likelihood*). Procedura polega na zastąpieniu oryginalnej funkcji wiarogodności $F(X, \beta)$ jest

oszacowaniem uzyskanym z wykorzystaniem metod symulacyjnych $\tilde{F}(X, \beta)$. Do tego estymatora można zastosować standardową teorię asymptotyczną, a więc w próbie o dużej liczebności jest on zgodny i nieobciążony, oraz zbieżny w tempie nie wolniejszym niż $\sqrt{N}$, gdzie N jest liczbą obserwacji na podstawie której szacowane są parametry modelu.

Zaletą modeli o losowych parametrach jest fakt, iż pozwalają one na modelowanie heterogeniczności zachowań decydentów poprzez uchylenie dość restrykcyjnego założenia o homogeniczności preferencji. W tym modelu liczbowe wartości parametrów są wielkościami specyficznymi dla każdego decydenta. Oczywiście w tak ogólnej postaci parametry modelu byłyby niemal niemożliwe do oszacowania. W trakcie modelowania przyjmuje się pewien rozkład dla nieznanych parametrów, zwyczajowym wyborem jest rozkład normalny. Niektóre pakiety statystyczne pozwalają na wykorzystanie innych rozkładów.

Prawdopodobieństwo wyboru możliwości i jest dane przez:

$$P_{ni}(y = i|X = x) = \int \frac{exp(x_{ni}\beta_n)}{\sum_{j=1}^{J} exp(x_{ni}\beta_n)} f(\beta|\theta)d\beta \tag{1}$$

gdzie θ jest zbiorem parametrów od których uzależniona jest heterogeniczność preferencji w populacji, $f(\beta|\theta)$ jest funkcją gęstości prawdopodobieństwa parametru β. Sumowanie (całkowanie) odbywa się względem wszystkich możliwych wartości β. W tym modelu wektor parametrów β jest specyficzny dla decydenta, zatem potencjalnie każdy decydent może być charakteryzowany przez inną funkcję opisującą jego preferencje.

W przypadku mieszanych modeli logitowych (modeli logitowych o losowych parametrach) występuje w literaturze pewna niezręczność terminologiczna. Zarówno β jak i θ są określane jako parametry modelu. Wektor β to zestaw parametrów funkcji logistycznej, są one opisane przez funkcję gęstości $f(\beta)$. Drugi zestaw parametrów θ opisuje tę funkcję gęstości. Należy pamiętać, że prawdopodobieństwo wyboru możliwości i w mieszanym modelu logitowym nie zależy od wartości parametru β. Zgodnie ze wzorem (1) całkowanie odbywa się względem wszystkich wartości parametrów β, wobec tego poszczególne realizacje wartości tego parametru nie mają znaczenia dla opisania prawdopodobieństwa wyboru możliwości i. Te prawdopodobieństwo zależy wyłącznie od wartości parametru θ.

Parametry modelu (1) można oszacować wykorzystując algorytm symulowanej funkcji wiarogodności. Ta metoda jest wykorzystywana przez polecenie `mixlogit` w pakiecie Stata.

Literatura

[1] Cameron, A.C. i Trivedi, P.K.. (2009): *Microeconometrics Using Stata*, Stata Press.

[2] Cameron, A.C. i Windmeijer, F.A.G. (1993): *R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization*, Dept. of Economics Working Paper 93-24, University of California at Davis.

[3] Veall, Michael R. i Zimmermann, Klaus F. (1996) *Pseudo-R2 Measures for Some Common Limited Dependent Variable Models*. Collaborative Research Center 386, Discussion Paper 18.

[4] Williams Richard (2011) *Comparing Logit and Probit Coefficients Between Models and Across Groups*.