6 Modele wyborów dyskretnych dla danych panelowych

Dane do notatek są danymi do podręcznika Cameron & Trivedi (2008), pochodzą z artykułu Deb i Triverdi (2002). Przedmiotem badania jest eksperyment związany z losowym przydzieleniem rodzinom różnych instrumentów w zakresie polityki zdrowotnej. Celem badania była weryfikacja w jaki sposób działa polityka współpłacenia przy różnej stopie współpłatności za usługi zdrowotne.

Dane mają charakter panelowy. Każda obserwacja opisuje jedną osobę w jednym roku. Każda osoba może być rejestrowana przez okres do pięciu lat. W rezultacie panel jest niezbilansowany. Identyfikatorem dla panelu jest zmienna `id`.

```
use mus18data.dta, clear
describe dmdu med mdu lcoins ndisease female age lfam child id year
```

```
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
dmdu            float   %9.0g                 any MD visit = 1 if mdu > 0
med             float   %9.0g                 medical exp excl outpatient men
mdu             float   %9.0g                 number face-to-fact md visits
lcoins          float   %9.0g                 log(coinsurance+1)
ndisease        float   %9.0g                 count of chronic diseases -- ba
female          float   %9.0g                 female
age             float   %9.0g                 age that year
lfam            float   %9.0g                 log of family size
child           float   %9.0g                 child
id              float   %9.0g                 person id, leading digit is sit
year            float   %9.0g                 study year
```

Podstawowe statystyki opisowe zmiennych

```
summarize dmdu med mdu lcoins ndisease female age lfam child id year
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        dmdu |     20186    .6875062    .4635214          0          1
         med |     20186    171.5892    698.2689          0   39182.02
         mdu |     20186    2.860696    4.504765          0         77
      lcoins |     20186    2.383588    2.041713          0   4.564348
    ndisease |     20186     11.2445    6.741647          0       58.6
-------------+--------------------------------------------------------
      female |     20186    .5169424    .4997252          0          1
```

```
         age |     20186    25.71844    16.76759          0   64.27515
        lfam |     20186    1.248404    .5390681          0   2.639057
       child |     20186    .4014168    .4901972          0          1
          id |     20186    357971.2    180885.6     125024     632167
-------------+---------------------------------------------------------
        year |     20186    2.420044    1.217237          1          5
```

Zmienna `dmdu` jest binarnym wskaźnikiem, którego wartość określa czy osoba w danym roku odwiedziła lekarza (wartość 1), czy nie (wartośc 0). Zmienna `med` pokazuje roczne wydatki związane z leczeniem. Zmienna `mdu` pokazuję liczbę wizyt u lekarza w danym roku.

Zmiennymi objaśniającymi są `lcoins` oznaczająca logarytm stopy współpłacenia plus 1. Dodano jedynkę, by w zbiorze danych nie pojawiły się brakujące wartości, z uwagi na fakt iż logarytm zera nie istnieje. Zmienna `ndisease` jest liczbą długotrwałych chorób jakie zdiagnozowano u pacjenta. Dodatkowo, zbiór zawiera cztery zmienne demograficzne: `female` zmienna zero-jedynkowa wskazująca, że osoba jest kobietą, `age` oznaczającą wiek pacjenta w danym roku, `lfam` logarytm liczby osób, które liczy rodzina, `child` zmienna zero-jedynkowa wskazująca, że osoba jest dzieckiem.

Zadeklarowane, że zbiór danych jest panelem i opis struktury panelowej zbioru.

```
xtset id year
xtdescribe
```

```
      id:  125024, 125025, ..., 632167                        n =       5908
    year:  1, 2, ..., 5                                       T =          5
           Delta(year) = 1 unit
           Span(year)  = 5 periods
           (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       2       3         3         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ------------------------+---------
     3710     62.80   62.80 |  111..
     1584     26.81   89.61 |  11111
      156      2.64   92.25 |  1....
      147      2.49   94.74 |  11...
       79      1.34   96.07 |  ..1..
       66      1.12   97.19 |  .11..
       33      0.56   97.75 |  ..111
       33      0.56   98.31 |  .1111
       29      0.49   98.80 |  ...11
```

```
       71      1.20  100.00 | (other patterns)
 --------------------------+---------
     5908    100.00         |  XXXXX
```

Eksperyment objął 5908 osób, jednak tylko 26,8 % osób jest obserwowanych przez 5 lat.

Przed przystąpieniem do dalszej analizy warto jest sprawdzić zróżnicowanie zmiennych niezależnych, które nie są stałe w czasie.

```
xtsum age lfam child
```

```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
age      overall |  25.71844   16.76759          0   64.27515 |     N =   20186
         between |             16.97265          0   63.27515 |     n =    5908
         within  |             1.086687   23.46844   27.96844 | T-bar = 3.41672
                 |                                           |
lfam     overall |  1.248404   .5390681          0   2.639057 |     N =   20186
         between |             .5372082          0   2.639057 |     n =    5908
         within  |             .0730824   .3242075    2.44291 | T-bar = 3.41672
                 |                                           |
child    overall |  .4014168   .4901972          0          1 |     N =   20186
         between |             .4820984          0          1 |     n =    5908
         within  |             .1096116  -.3985832   1.201417 | T-bar = 3.41672
```

Dla wszystkich analizowanych zmiennych większe jest zróżnicowanie międzygrupowe. Zatem oczekujemy, że estymator efektów stałych nie będzie efektywny, ponieważ jego wartość jest uzależniona od wariancji wewnątrzgrupowej.

Ujmując problem bardziej ogólnie nie ma podstaw by oczekiwać, że konieczne będzie wykorzystanie estymatora efektów stałych. Dane są danymi eksperymentalnymi, zatem nie powinno być problemu związanego potencjalną endogenicznością zmiennych. W rezultacie prawidłowe wyniki powinny dać estymatory efektów losowych albo uśrednionych efektów panelowych.

Analizę rozpoczynamy od opisu struktury panelowej zmiennej zależnej

```
xtsum dmdu
```

```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
dmdu     overall |  .6875062   .4635214          0          1 |     N =   20186
         between |             .3571059          0          1 |     n =    5908
         within  |             .3073307  -.1124938   1.487506 | T-bar = 3.41672
```

Widać, że zróżnicowanie międzygrupowe i wewnątrzgrupowe są na zbliżonym poziomie.

```
xttrans dmdu

    any MD |  any MD visit = 1 if
 visit = 1 |        mdu > 0
if mdu > 0 |         0          1 |     Total
-----------+----------------------+----------
         0 |     58.87      41.13 |    100.00
         1 |     19.73      80.27 |    100.00
-----------+----------------------+----------
     Total |     31.81      68.19 |    100.00
```

W wierszach tablicy przejść są wartości początkowe, z okresu $t-1$, a w kolumnach końcowe, z okresu t. Wartości w tablicy przejść pomiędzy latami wskazuje na trwałość zjawiska wizyt u lekarza.

```
corr dmdu l.dmdu l2.dmdu

(obs=8626)
             |                 L.      L2.
             |     dmdu     dmdu     dmdu
-------------+------------------------------
        dmdu |
         --. |   1.0000
         L1. |   0.3861   1.0000
         L2. |   0.3601   0.3807   1.0000
```

Warto zauważyć, że wartość współczynnika korelacji wizyt w kolejnych latach jest niemal stała w czasie.

6.1 Modelowanie

Pierwszym modelem będzie standardowy model logitowy. Panelowa struktura danych jest ignorowana. Zakładana jest niezależność obserwacji względem czasu i osób. W celu uwzględnienia powtarzających się obserwacji wykorzystywane jest odporne oszacowanie dla macierzy wariancji-kowariancji. Obserwacje są grupowana przez zmienną `id`.

```
logit dmdu lcoins ndisease female age lfam child, vce(cluster id) nolog

Logistic regression                             Number of obs   =      20186
                                                Wald chi2(6)    =     488.18
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -11973.392               Pseudo R2       =     0.0450

                                  (Std. Err. adjusted for 5908 clusters in id)
------------------------------------------------------------------------------
```

```
             |               Robust
        dmdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.1572107   .0109064   -14.41   0.000    -.1785869   -.1358345
    ndisease |    .050301   .0039657    12.68   0.000     .0425285    .0580735
      female |   .3091573   .0445772     6.94   0.000     .2217876     .396527
         age |   .0042689   .0022307     1.91   0.056    -.0001032     .008641
        lfam |  -.2047573   .0470287    -4.35   0.000    -.2969317   -.1125828
       child |   .0921709   .0728107     1.27   0.206    -.0505355    .2348773
       _cons |   .6039411   .1107712     5.45   0.000     .3868335    .8210486
------------------------------------------------------------------------------
```

Warto zwrócić uwagę, że wymuszenie odpornych błędów standardowych spowodowało wykorzystanie metody pseudo-największej wiarogodności. Zatem uzyskane rozwiązanie jest jedynie przybliżone. Uzyskane znaki dla ocen parametrów są zgodne z oczekiwaniami.

W modelu uśrednionych efektów panelowych można założyć różną postać struktury korelacyjnej wewnątrz panelu. W tym celu należy posłużyć się opcją `corr`

Struktura niezależna wartość opcji `independent` jest zdefiniowana jako

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ 0 & \text{jeżeli } t \neq s \end{cases}$$

Struktura wymienna wartość opcji `exchangeable` jest zdefiniowana jako

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ \rho & \text{jeżeli } t \neq s \end{cases}$$

odpowiada ona modelowi o stałej wartości korelacji między obserwacjami.

Struktura autoregresyjna rzędu p wartość opcji `ar p` jest zdefiniowana jako

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ \rho^{|t-s|} & \text{jeżeli } t \neq s \end{cases}$$

Struktura stacjonarna rzędu g wartość opcji `stationary g` jest zdefiniowana jako

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ \rho & \text{jeżeli } \mid t - s \mid = 1 \\ 0 & \text{w pp.} \end{cases}$$

Struktura niestacjonarna rzędu g wartość opcji `nonstationary g` jest zdefiniowana jako

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ \rho_{ts} & \text{jeżeli } 0 < \mid t - s \mid < g, \rho_{ts} = \rho_{st} \\ 0 & \text{w pp.} \end{cases}$$

Struktura nieustrukturyzowana wartość opcji `unstructured` wymaga by elementy na diagonali macierzy korelacji były równe 1.

$$\mathbb{V}_{t,s} = \begin{cases} 1 & \text{jeżeli } t = s \\ \rho_{ts} & \text{w pp., } \rho_{ts} = \rho_{st} \end{cases}$$

Struktura nazwana `exachangeable` zakłada, że wartość korelacji jest identyczna niezależnie od tego jak daleko względem czasu są obserwacje od siebie. Ze względu na wartości w tablicy przejść model ten wydaje się być odpowiedni.

```
xtlogit dmdu lcoins ndisease female age lfam child, pa corr(exch) vce(robust) nolog
```

```
GEE population-averaged model                   Number of obs      =     20186
Group variable:                           id    Number of groups   =      5908
Link:                                  logit    Obs per group: min =         1
Family:                             binomial                   avg =       3.4
Correlation:                   exchangeable                   max =         5
                                               Wald chi2(6)       =    521.45
Scale parameter:                           1   Prob > chi2        =    0.0000

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
        dmdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.1603179   .0107779   -14.87   0.000    -.1814422   -.1391935
    ndisease |   .0515445   .0038528    13.38   0.000     .0439931    .0590958
      female |   .2977003   .0438316     6.79   0.000      .211792    .3836086
         age |   .0045675   .0021001     2.17   0.030     .0004514    .0086836
        lfam |  -.2044045   .0455004    -4.49   0.000    -.2935837   -.1152254
       child |   .1184697   .0674367     1.76   0.079    -.0137039    .2506432
       _cons |   .5776986    .106591     5.42   0.000      .368784    .7866132
------------------------------------------------------------------------------
```

Jak widać wartości oszacowań parametrów są zbliżone do oszacowań uzyskanych w standardowym, nieprawidłowym, modelu regresji logistycznej.

W panelowym modelu logitowym o efektach losowych zakładane jest, że efekt indywidualny ma rozkład normalny $\mathcal{N}(0, \sigma_\alpha^2)$. Z uwagi na brak analitycznych rozwiązań dla modelu uzyskiwane jest rozwiązanie przybliżone

```
* Logit random-effects estimator
xtlogit dmdu lcoins ndisease female age lfam child, re nolog
```

```
Random-effects logistic regression              Number of obs      =     20186
Group variable: id                              Number of groups   =      5908
```

```
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       3.4
                                                               max =         5

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(6)       =    549.76
Log likelihood  = -10878.687                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        dmdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.2403864   .0162836   -14.76   0.000    -.2723017    -.208471
    ndisease |    .078151   .0055456    14.09   0.000     .0672819    .0890201
      female |   .4631005   .0663209     6.98   0.000     .3331138    .5930871
         age |   .0073441   .0031508     2.33   0.020     .0011687    .0135194
        lfam |  -.3021841   .0644721    -4.69   0.000    -.4285471    -.175821
       child |   .1935357   .1002267     1.93   0.053     -.002905    .3899763
       _cons |   .8629898   .1568968     5.50   0.000     .5554778    1.170502
-------------+----------------------------------------------------------------
    /lnsig2u |   1.225652   .0490898                      1.129438    1.321866
-------------+----------------------------------------------------------------
     sigma_u |    1.84564    .045301                      1.758953    1.936599
         rho |   .5087003   .0122687                      .4846525     .532708
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  2189.41 Prob >= chibar2 = 0.000
```

Oszacowana wartość parametru `rho` wskazuje, że około 50% wariancji jest generowane przez zróżnicowanie wewnątrzgrupowe.

W modelu efektów stałych efekty indywidualne mogą być skorelowane ze zmiennymi objaśniającymi. Model efektów stałych jest trudny do oszacowania ze względu na problemy numeryczne. W estymacji wykorzystywana jest metoda warunkowej największej wiarogodności.

```
* Logit fixed-effects estimator
xtlogit dmdu lcoins ndisease female age lfam child, fe nolog

Conditional fixed-effects logistic regression   Number of obs      =      9025
Group variable: id                              Number of groups   =      2449

                                                Obs per group: min =         2
                                                               avg =       3.7
                                                               max =         5

                                                LR chi2(3)         =     10.74
Log likelihood  = -3395.5996                    Prob > chi2        =    0.0132

------------------------------------------------------------------------------
```

```
        dmdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |          0  (omitted)
    ndisease |          0  (omitted)
      female |          0  (omitted)
         age |  -.0341815   .0183827    -1.86   0.063     -.070211     .001848
        lfam |    .478755   .2597327     1.84   0.065    -.0303116    .9878217
       child |    .270458   .1684974     1.61   0.108    -.0597907    .6007068
------------------------------------------------------------------------------
```

Zgodnie z oczekiwaniami uzyskanie oszacowań dla parametrów stałych w czasie nie jest możliwe. Dodatkowo, warto zwrócić uwagę, że oszacowania uzyskano na podstawie mniejszej liczby obserwacji. Dzieje się tak, gdyż odrzucono obserwacje dla osób, dla których zmienna zależna nie zmieniała wartości.

Możemy podsumować modele w jednej tabeli.

```
* Panel logit estimator comparison
global xlist lcoins ndisease female age lfam child

quietly logit dmdu \$xlist, vce(cluster id) estimates store POOLED

quietly xtlogit dmdu \$xlist, pa corr(exch) vce(robust) estimates store PA

quietly xtlogit dmdu \$xlist, re     // SEs are not cluster-robust estimates store RE

quietly xtlogit dmdu \$xlist, fe     // SEs are not cluster-robust estimates store FE

estimates table POOLED PA RE FE, equations(1) se b(\%8.4f) stats(N ll) stfmt(\%8.0f)
```

Model efektów losowych pozwala na wprowadzenie do modelu stałej o rozkładzie normalnym. Polecenie `xmmelogit` pozwala na uzyskanie oszacowania dla modelu w którym parametry nachylenia również są zmienną losową o rozkładzie normalnym.

```
* Logit mixed-effects estimator (same as xtlogit, re)
* xtmelogit dmdu lcoins ndisease female age lfam child || id:
```

Ten model jest jednak częściej wykorzystywany w statystyce niż w ekonometrii, do analizy skupionych danych (ang. clustered data).

Literatura

[1] Colin Cameron and Pravin K. Trivedi (2008) "Microeconometrics using Stata"