Szacowanie modeli wielowartościowych w pakiecie STATA

Paweł Strawiński

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

25 kwietnia 2007

W badaniach ekonomicznych i społecznych często odpowiedzi na pytania są kodowane za pomocą skali Likerta zgodności z zaproponowanym stwierdzeniem. Wartości tej skali są przyjmowane arbitralne, zazwyczaj od 1 do 5, lub od 1 do 7. Z uwagi na arbitralne ustalenie wartości jakie przyjmuje zmienna nie posiada ona interpretacji ilościowej, ponadto z reguły wartości pośrednie nie posiadają interpretacji. Z tego powodu nie powinno się traktować takich zmiennych identycznie jak traktowane są zmienne kardynalne.

Gdy chcemy zmienną o wielu wartościach użyć jako zmienną zależną modelu to wygodnym narzędziem ekonometrycznym są uogólnione modele wyborów dyskretnych.

1 Dane

Dane do przykładu pochodzą z amerykańskiego rynku obligacji komercyjnych. Zbiór liczy 98 obserwacji i zawiera informacje firmach. Zmienną zależną jest rating obligacji `rating83` od AAA do C, który zakodowany jest jako wartość całkowita, im wyższa tym wyższy rating. Jego poziom jest tłumaczony za pomocą wskaźnika dochód do wartości `ia83` (ang. *income-to-asset ratio*), oraz zmianą poziomu tego wskaźnika między rokiem 1982 a 1983, zmienna `dia`.

Dane są dostępne w internecie. Wystarczy w Stacie wpisać

```
. use http://www.stata-press.com/data/imeus/panel84extract, clear
```

Zmienna oryginalna `rating` dla pewnych kategorii zawiera małą liczbę obserwacji.

```
. tab rating83

   rating83 |      Freq.     Percent        Cum.
------------+-----------------------------------
```

```
        AAA |         29       29.59       29.59
         AA |          2        2.04       31.63
          A |         13       13.27       44.90
        BAA |         28       28.57       73.47
         BA |         16       16.33       89.80
          B |          7        7.14       96.94
          C |          3        3.06      100.00
------------+-----------------------------------
      Total |         98      100.00
```

Gdybyśmy chcieli użyć zmiennej `rating83` jako objaśnianej to dla wyboru między ratingiem C a ratingiem B dysponowalibyśmy tylko 10 obserwacjami. W modelu są 2 zmienne, stała, należy oszacować wariancję, czyli zostaje tylko 6 stopnii swobody. Powodowało by to niską precyzję oszacowań.

Z tego powodu przed przystąpieniem do szacowania parametrów modelu należy ją przekształcić w taki sposób, aby w każda kategoria przekształconej zmiennej zawierała podobną ilość obserwacji. Taka operacja ułatwia, a czasami wręcz umożliwia oszacowanie parametrów modelu.

```
. tab rating83c
```

```
       Bond |
    rating, |
       1983 |      Freq.     Percent        Cum.
------------+-----------------------------------
     BA_B_C |         26       26.53       26.53
        BAA |         28       28.57       55.10
       AA_A |         15       15.31       70.41
        AAA |         29       29.59      100.00
------------+-----------------------------------
      Total |         98      100.00
```

Przekodowanie nastąpiło w sposób następujący

```
. tab rating83 rating83c
```

```
           |              Bond rating, 1983
  rating83 |    BA_B_C        BAA       AA_A        AAA |     Total
-----------+--------------------------------------------+----------
       AAA |         0          0          0         29 |        29
        AA |         0          0          2          0 |         2
         A |         0          0         13          0 |        13
       BAA |         0         28          0          0 |        28
        BA |        16          0          0          0 |        16
         B |         7          0          0          0 |         7
         C |         3          0          0          0 |         3
-----------+--------------------------------------------+----------
     Total |        26         28         15         29 |        98
```

Połączono rating AA z ratingiem A, oraz 3 najnizsze kategorie w jedną zmienną.

2 Modele wielomianowe

Na początku potraktujmy rating jako zmienną bez ustalonej hierarchii[1]. Wobec tego aby zbudować model użyjemy modelu wielomianowego.

```
. mlogit rating83c ia83 dia

Iteration 0:   log likelihood = -133.04224
Iteration 1:   log likelihood = -119.74382
Iteration 2:   log likelihood = -118.09549
Iteration 3:   log likelihood = -118.00312
Iteration 4:   log likelihood = -118.00239

Multinomial logistic regression                 Number of obs   =         98
                                                LR chi2(6)      =      30.08
                                                Prob > chi2     =     0.0000
Log likelihood = -118.00239                     Pseudo R2       =     0.1130
------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
BA_B_C       |
        ia83 |  -.1303426   .0489834    -2.66   0.008    -.2263484   -.0343369
         dia |   .1542639   .0735878     2.10   0.036     .0100345    .2984934
       _cons |   .9820093   .4966503     1.98   0.048     .0085926    1.955426
-------------+----------------------------------------------------------------
BAA          |
        ia83 |  -.0454026   .0441004    -1.03   0.303    -.1318378    .0410326
         dia |   .1240656   .0709813     1.75   0.080    -.0150552    .2631865
       _cons |   .4024031   .5154444     0.78   0.435    -.6078494    1.412656
-------------+----------------------------------------------------------------
AA_A         |
        ia83 |   .1464149   .0593538     2.47   0.014     .0300836    .2627462
         dia |   .1748879   .0910389     1.92   0.055    -.0035451    .3533209
       _cons |  -2.909247   .9744167    -2.99   0.003    -4.819069   -.9994255
------------------------------------------------------------------------------
(rating83c==AAA is the base outcome)
```

Wyniki dla każdej kategorii są liczone w odniesieniu do poziomu bazowego. Stata jako poziom odniesienia ustaliła rating AAA, ponieważ dla tej kategorii dostępną jest największa liczba obserwacji. Za pomocą opcji `baseoutcome(#)`, gdzie # oznacza numer alternatywy, można kontrolować poziom odniesienia. W wierszu pierwszym są współczynniki porównujące rating BA_B_C z AAA, w drugim porównany jest rating BAA z AAA, a w trzecim AA_A. Zbliżone wyniki można uzyskać szacując osobno modele logitowe dla każdej kategorii. Wyniki nie będą takie same, bowiem w przypadku logitu funkcja wiarogodności zależy od 4 parametrów, a w przypadku wielomianowego logitu od 10.

Analogiczne wyniki można uzyskać szacując wielomianowy model probitowy

```
. mprobit rating83c ia83 dia
```

[1]Przyjmujemy takie założenie wyłącznie w celach szkoleniowych.

```
Iteration 0:   log likelihood = -117.80861
Iteration 1:   log likelihood = -117.62741
Iteration 2:   log likelihood =  -117.6271
Iteration 3:   log likelihood =  -117.6271

Multinomial probit regression                 Number of obs   =         98
                                              Wald chi2(6)    =      21.56
Log likelihood =  -117.6271                   Prob > chi2     =     0.0015

------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
BA_B_C       |
        ia83 |  -.0911987   .0331999    -2.75   0.006    -.1562694   -.026128
         dia |   .0995189    .050054     1.99   0.047     .0014149    .197623
       _cons |   .6635147   .3480602     1.91   0.057    -.0186707     1.3457
-------------+----------------------------------------------------------------
BAA          |
        ia83 |  -.0284042   .0307741    -0.92   0.356    -.0887203    .0319118
         dia |   .0834583   .0505054     1.65   0.098    -.0155304    .1824471
       _cons |   .2136007   .3606114     0.59   0.554    -.4931847    .9203861
-------------+----------------------------------------------------------------
AA_A         |
        ia83 |   .1113513   .0418002     2.66   0.008     .0294244    .1932782
         dia |   .1167499   .0636459     1.83   0.067    -.0079937    .2414935
       _cons |  -2.143089   .6606959    -3.24   0.001    -3.438029    .8481486
------------------------------------------------------------------------------
```

Porównując wartość logarytmu funkcji wiarogodności i wartość statystyki LR łatwo zauważyć, że wielomianowy model logitowy jest lepiej dopasowany do danych empirycznych.

Tak jak w modelu logitowym wartości współczynników nie mają interpretacji ekonomicznej. Aby im ją nadać należy przedstawić wyniki w postaci ilorazów szans (ryzyk) (ang. *relative risk ratio*)

```
.estimates restore mlogit

.mlogit, rrr

Multinomial logistic regression               Number of obs   =         98
                                              LR chi2(6)      =      30.08
                                              Prob > chi2     =     0.0000
Log likelihood = -118.00239                   Pseudo R2       =     0.1130
------------------------------------------------------------------------------
   rating83c |       RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
BA_B_C       |
        ia83 |   .8777946   .0429974    -2.66   0.008     .7974402    .9662459
         dia |   1.166799   .0858622     2.10   0.036     1.010085    1.347827
-------------+----------------------------------------------------------------
BAA          |
        ia83 |   .9556127   .0421429    -1.03   0.303     .8764831    1.041886
         dia |    1.13209   .0803573     1.75   0.080     .9850576    1.301069
```

```
-------------+----------------------------------------------------------------
AA_A         |
        ia83 |  1.157676   .0687125     2.47  0.014   1.030541    1.300497
         dia |  1.191113   .1084376     1.92  0.055   .9964612    1.423788
------------------------------------------------------------------------------
(rating83c==AAA is the base outcome)
```

Im wyższa była wartość wskaźnika dochód do wartości, tym większa szansa na wyższy rating, natomiast wpływ zmiany tego wskaźnika jest ujemny.

Porównanie modeli Mając oszacowane i zapamiętane parametry dla dla obu modeli możemy je porównać za pomocą wartości kryteriów informacyjnych

```
. estimates restore mlogit (results mlogit are already active)

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
      mlogit |     98   -133.0422   -118.0024      9     254.0048    277.2695
-----------------------------------------------------------------------------

. estimates restore mprobit (results mprobit are active now)

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
     mprobit |     98           .   -117.6271      9     253.2542    276.5189
-----------------------------------------------------------------------------
```

Wyższa wartość logarytmu funkcji wiarogodności oraz niższa wartość kryteriów informacyjnych wskazują na lepsze dopasowanie do danych empirycznych modelu probitowego.

Testowanie istotności Na wydruku ze Staty widzimy, że zmienne są łącznie istotne (statystyka LR). Natomiast warto jest przetestować istotność poszczególnych zmiennych. Można zrobić to dwoma metodami. Testem Walda

```
. test ia83

 ( 1)  [BA_B_C]ia83 = 0
 ( 2)  [BAA]ia83 = 0
 ( 3)  [AA_A]ia83 = 0

           chi2(  3) =   17.07
         Prob > chi2 =    0.0007

. test dia
```

```
 ( 1)  [BA_B_C]dia = 0
 ( 2)  [BAA]dia = 0
 ( 3)  [AA_A]dia = 0

           chi2(  3) =    5.82
         Prob > chi2 =    0.1208
```

Na podstawie tego testu odrzucamy hipotezę o łącznej nieistotności zmiennej `ia83`, natomiast zmienna `dia` jest statystycznie nieistotna.

Alternatywnym testem jest test ilorazu wiarogodności. W uzyskaniu jego wyników, jak również innych testów, bardzo przydatny jest dodatkowy pakiet `mlogtest`.

Aby go zainstalować należy napisać

```
. net search mlogtest
```

następnie wybrać pakiet spost9 i go zainstalować.

W kolejnym kroku należy aktywować oszacowania wielomianowego modelu logitowego

```
. estimates restore mlogit

. mlogtest, l w

**** Likelihood-ratio tests for independent variables

 Ho: All coefficients associated with given variable(s) are 0.

    rati~83c |       chi2   df   P>chi2
-------------+------------------------
        ia83 |     23.935    3    0.000
         dia |      6.741    3    0.081
---------------------------------------

**** Wald tests for independent variables

 Ho: All coefficients associated with given variable(s) are 0.

    rati~83c |       chi2   df   P>chi2
-------------+------------------------
        ia83 |     17.068    3    0.001
         dia |      5.819    3    0.121
---------------------------------------
```

Wartości statystyk nieznacznie różnią się w obu testach, jednak dają one takie same konkluzję.

Testowanie niezależności niezwiązanych alternatyw W pakiecie Stata za zaimplementowane dwa testy sprawdzające założenie o niezależności niezwiązanych alternatyw. W obu hipotezą zerową jest niezależność niezwiązanych alternatyw. Oba mają podobną konstrukcję i porównują oszacowania przy pełnym zestawie alternatyw i pominięciu jednej z nich.

Test Hausmana opiera się o statystykę Walda.

```
. mlogtest, h sm
**** Hausman tests of IIA assumption

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

 Omitted |      chi2   df   P>chi2   evidence
---------+----------------------------------
  BA_B_C |    -1.834    6    1.000   for Ho
     BAA |    -5.012    6    1.000   for Ho
    AA_A |    -1.956    6    1.000   for Ho
-------------------------------------------
```

Test Small'a i Hsiao bazuje na statystyce ilorazu wiarogodności.

```
 **** Small-Hsiao tests of IIA assumption

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

 Omitted |  lnL(full)  lnL(omit)    chi2   df   P>chi2   evidence
---------+--------------------------------------------------------
  BA_B_C |    -35.112    -33.256   3.713    3    0.294   for Ho
     BAA |    -29.480    -26.636   5.688    3    0.128   for Ho
    AA_A |    -36.001    -34.593   2.816    3    0.421   for Ho
-------------------------------------------------------------------
```

Oba testy wskazują, że założenie o niezależności alternatyw jest spełnione. Jest to rzadko spotykany przypadek. Szczególnie w dużych próbach testy mogą dawać przeczące sobie wyniki.

Warto również sprawdzić czy pewnych kategorii zmiennej zależnej nie da się połączyć w jedną.

```
. mlogtest, c lrc
 **** Wald tests for combining outcome categories

 Ho: All coefficients except intercepts associated with given pair
     of outcomes are 0 (i.e., categories can be collapsed).

Categories tested |      chi2   df   P>chi2
------------------+-----------------------
  BA_B_C-     BAA |     3.287    2    0.193
  BA_B_C-    AA_A |    17.198    2    0.000
  BA_B_C-     AAA |     8.067    2    0.018
     BAA-    AA_A |    10.214    2    0.006
     BAA-     AAA |     3.212    2    0.201
    AA_A-     AAA |    10.273    2    0.006
----------------------------------------------
```

```
**** LR tests for combining outcome categories

 Ho: All coefficients except intercepts associated with given pair
     of outcomes are 0 (i.e., categories can be collapsed).

Categories tested |      chi2   df   P>chi2
------------------+-----------------------
  BA_B_C-     BAA |     3.534    2    0.171
  BA_B_C-    AA_A |    24.973    2    0.000
  BA_B_C-     AAA |     9.476    2    0.009
     BAA-    AA_A |    13.573    2    0.001
     BAA-     AAA |     3.392    2    0.183
    AA_A-     AAA |    14.280    2    0.001
-------------------------------------------
```

Oba testy dają podobne wyniki wskazując, że kategorię BAA można połączyć z kategorią BA_B_C, oraz można ją połączyć z kategorią AAA.

Wszystkie testy można wywołać poleceniem

```
. mlogtest, all
```

3 Modele uporządkowane

Czasami zmienna zależna o charakterze nominalnym posiada naturalną hierarchię. Wobec tego modelując zjawisko można i wskazane jest taką informację wykorzystać.

Przystępując do analizy danych o ratingu w poprzednim punkcie pominęliśmy fakt uszeregowania zmiennej zależnej.

```
. ologit rating83c ia83 dia

Iteration 0:   log likelihood = -133.04224
Iteration 1:   log likelihood = -127.30126
Iteration 2:   log likelihood = -127.27148
Iteration 3:   log likelihood = -127.27146

Ordered logistic regression                     Number of obs   =         98
                                                LR chi2(2)      =      11.54
                                                Prob > chi2     =     0.0031
Log likelihood = -127.27146                     Pseudo R2       =     0.0434

------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ia83 |   .0939166   .0296196     3.17   0.002     .0358633    .1519699
         dia |  -.0866925   .0449789    -1.93   0.054    -.1748496    .0014646
-------------+----------------------------------------------------------------
       /cut1 |  -.1853053   .3571432                     -.8852931    .5146825
       /cut2 |   1.185726   .3882098                      .4248489    1.946603
       /cut3 |   1.908412   .4164895                      1.092108    2.724717
```

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

. estimates store ologit
```

Podobne wyniki uzyskamy szacując uszeregowany model probitowy

```
. oprobit rating83c ia83 dia

Iteration 0:   log likelihood = -133.04224
Iteration 1:   log likelihood = -127.87966
Iteration 2:   log likelihood = -127.87756

Ordered probit regression                       Number of obs   =         98
                                                LR chi2(2)      =      10.33
                                                Prob > chi2     =     0.0057
Log likelihood = -127.87756                     Pseudo R2       =     0.0388

------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ia83 |   .0512509   .0167257     3.06   0.002     .0184691    .0840327
         dia |  -.0496082   .0254406    -1.95   0.051    -.0994709    .0002545
-------------+----------------------------------------------------------------
       /cut1 |  -.1877442   .2048421                     -.5892274     .213739
       /cut2 |   .6344613   .2156936                      .2117095    1.057213
       /cut3 |   1.064523   .2231879                      .6270823    1.501963
------------------------------------------------------------------------------

. estimates store oprobit
```

Tak jak w przypadku modelu wielomianowego model logitowy wydaje się być lepiej dopasowany do danych ze względu na wyższą wartość logarytmu funkcji wiarogodności oraz statystyki LR.

Modele uporządkowane mogą być traktowane w pewnym przybliżeniu jako model wielomianowy z narzuconymi ograniczeniami. Wobec tego można porównać oba modele.

```
. lrtest ologit mlogit, force

Likelihood-ratio test                                 LR chi2(4)=     18.54
(Assumption: ologit nested in mlogit)                 Prob > chi2 =    0.0010

. lrtest oprobit mprobit, force

Likelihood-ratio test                                 LR chi2(4)=     20.50
(Assumption: oprobit nested in mprobit)               Prob > chi2 =    0.0004
```

Oba testy wskazują, że model uporządkowany nie jest zagnieżdżony w modelu wielomianowym, wobec tego wnosi dodatkową informację.

Uogólniony uporządkowany model logitowy Uporządkowany model probitowy i uporządkowany model logitowy muszą spełniać założenie o niezależności niezwiązanych alternatyw. Model uogólniony pozwala na zgodne oszacowanie parametrów w sytuacji, gdy to założenie nie jest spełnione. Model ten pozwala, aby macierz wariancji-kowariancji składnika losowego miała postać procesu autoregresyjnego rzędu (1). Przy czym niezerowa korelacja występuje jedynie między składnikami losowymi z sąsiadujących równań.

```
. gologit rating83c ia83 dia
Iteration 0:  Log Likelihood = -133.04224
(unproductive step attempted)
Iteration 1:  Log Likelihood = -125.80487
(unproductive step attempted)
Iteration 2:  Log Likelihood = -124.22611
(unproductive step attempted)
Iteration 3:  Log Likelihood = -121.02729
(unproductive step attempted)
Iteration 4:  Log Likelihood = -116.91873
(unproductive step attempted)
Iteration 5:  Log Likelihood = -114.70916
(unproductive step attempted)
Iteration 6:  Log Likelihood =  -113.4419
(unproductive step attempted)
Iteration 7:  Log Likelihood = -111.05564
(unproductive step attempted)
Iteration 8:  Log Likelihood = -110.13773
(unproductive step attempted)
Iteration 9:  Log Likelihood = -109.72069
Iteration 10:  Log Likelihood = -106.39091
Iteration 11:  Log Likelihood = -104.56663
Iteration 12:  Log Likelihood = -104.53306
Iteration 13:  Log Likelihood = -104.53292
Iteration 14:  Log Likelihood = -104.53292

Generalized Ordered Logit Estimates                   Number of obs    =      98
                                                      Model chi2(6)    =   57.02
                                                      Prob > chi2      =  0.0000
Log Likelihood =   -104.5329203                       Pseudo R2        =  0.2143

------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mleq1        |
        ia83 |   .2253086   .0576035     3.91   0.000     .1124078    .3382094
         dia |  -.0813404    .077436    -1.05   0.294    -.2331121    .0704313
       _cons |  -1.022971   .5193287    -1.97   0.049    -2.040836   -.0051051
-------------+----------------------------------------------------------------
mleq2        |
        ia83 |   .2570442   .0581859     4.42   0.000     .1430019    .3710865
         dia |  -.0718418   .0684521    -1.05   0.294    -.2060056    .0623219
       _cons |  -3.038211   .6498426    -4.68   0.000    -4.311879   -1.764543
-------------+----------------------------------------------------------------
```

```
mleq3        |
        ia83 |  -.0448163    .041431    -1.08   0.279    -.1260197     .036387
         dia |  -.1532988   .0707874    -2.17   0.030    -.2920395    -.014558
       _cons |   .0401501   .5286311     0.08   0.939    -.9959479    1.076248
------------------------------------------------------------------------------
```

Wyniki wyglądają analogicznie do wyników modelu wielomianowego, bowiem każdej pary alternatyw szacowane jest osobne równanie.

Literatura

[1] Kit Baum (2006) *An Introduction to Econometrics Using Stata*, Stata Press.

[2] J. Scott Long, Jeremy Freese (2003) *Regression Models for Categorical Dependent Variables Using Stata. Revised Edition*, Stata Press.