Modele dla zmiennej binarnej w pakiecie STATA

materiały na ćwiczenia z ekonometrii
18.03.2005 r.
Piotr Wójcik, KTRG WNE UW

Dane

Dane wykorzystane w przykładzie pochodzą z pracy McCall, B.P., 1995, The Impact of Unemployment Insurance Benefit Levels on Recipiency, *Journal of Business and Economic Statistics*, vol. 13, pp. 189-198.

Rozpatrujemy próbę 4877 pracowników fizycznych, którzy stracili pracę w USA między rokiem 1982 i 1991. Nie wszyscy bezrobotni, którym przysługuje świadczenie z tytułu ubezpieczenia od utraty pracy ubiegają się o nie, być może z powodu związanych z tym kosztów psychologicznych. Procent uprawnionych do takiego świadczenia bezrobotnych, którzy się o nie ubiegają nazwijmy stopą objęcia świadczeniami – dla analizowanej próby wyniosła ona zaledwie 68%. Spróbujmy więc przeanalizować co powoduje, że ludzie bezrobotni decydują się zrezygnować z przysługującego im świadczenia.

Wysokość świadczenia zależy od stanu zamieszkania, roku utracenia pracy i poprzednich zarobków. Stopa zastąpienia (stosunek wysokości świadczenia do poprzednich dochodów) waha się od 33% do 54% ze średnią dla badanej próby na poziomie 44% i jest potencjalnie istotnym czynnikiem wpływającym na decyzję bezrobotnego o ubieganiu się o świadczenie ubezpieczeniowe. Wpływ na tę decyzję mogą również mieć charakterystyki osobiste (wykształcenie, wiek, płeć, rasa) jak i sytuacja rodzinna i względy budżetowe.

Ostatnim typem potencjalnie istotnych zmiennych jest powód zwolnienia: niesatysfakcjonujące wyniki w pracy, likwidacja stanowiska pracy, zakończenie pracy sezonowej.

Lista i opis zmiennych:

stateur: stanowa stopa bezrobocia (w %);
statemb: najniższy stanowy zasiłek dla bezrobotnych;
state: kod stanu;
age: wiek w latach;
tenure: długość okresu zatrudnienia w ostatnim miejscu pracy;
slack: zmienna 0-1, 1 jeśli zwolnienie z powodu niesatysfakcjonujących wyników w pracy;
abol: zmienna 0-1, 1 jeśli zwolnienie z powodu likwidacji stanowiska pracy;
seasonal: zmienna 0-1, 1 jeśli zwolnienie z powodu końca pracy sezonowej;
nwhite: zmienna 0-1, 1 jeśli pracownik rasy innej niż biała;
school12: zmienna 0-1, 1 jeśli pracownik ukończył więcej niż 12 lat szkoły;
male: zmienna 0-1, 1 jeśli mężczyzna;
smsa: zmienna 0-1, 1 jeśli mieszka w Standard Metropolitan Statistical Area, czyli w obszarze miejskim;
married: zmienna 0-1, 1 jeśli żonaty/zamężna;
dkids: zmienna 0-1, 1 jeśli posiada dzieci;
dykids: zmienna 0-1, 1 jeśli posiada małe dzieci (0-5 lat);
yrdispl: rok utraty pracy (1982=1,..., 1991=10);
rr: stopa zastąpienia (stosunek wysokości zasiłku do poprzednich dochodów);
rr2: rr do kwadratu;
head: zmienna 0-1, 1 jeśli pracownik jest głową rodziny;
y: zmienna 0-1, 1 jeśli ubiegał się (i otrzymał) zasiłek;

Estymacji dokonamy dla trzech różnych modeli: liniowego modelu prawdopodobieństwa, modelu logit i probit.

Estymacja liniowego modelu prawdopodobieństwa

Zaletą stosowania liniowego modelu prawdopodobieństwa do modelowania dychotomicznej zmiennej objaśnianej jest prostota jego estymacji. Ze względu na dwumianowy rozkład zmiennej objaśnianej nie jest jednak spełniony warunek homoskedastyczności reszt z modelu. Wadą podejścia tego typu jest również fakt, iż nie można zagwarantować, że wartości dopasowane znajdą się w przedziale [0,1].

Estymacji dokonujemy za pomocą polecenia `reg`:

```
reg y  rr rr2 age tenure slack abol seasonal head married dkids dykids smsa
nwhite yrdispl school12 male statemb stateur
```

```
      Source |       SS       df       MS              Number of obs =    4877
-------------+------------------------------           F( 18,  4858) =   19.00
       Model |  69.3608203    18  3.85337891           Prob > F      =  0.0000
    Residual |  985.092737  4858  .202777426           R-squared     =  0.0658
-------------+------------------------------           Adj R-squared =  0.0623
       Total |  1054.45356  4876  .216253806           Root MSE      =  .45031

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |   .6062438   .3842868     1.58   0.115    -.1471322     1.35962
         rr2 |  -1.010374   .4811835    -2.10   0.036    -1.953711   -.0670365
         age |    .004297   .0007798     5.51   0.000     .0027682    .0058258
      tenure |   .0054326   .0012124     4.48   0.000     .0030558    .0078095
       slack |   .1260833   .0142071     8.87   0.000      .098231    .1539356
        abol |  -.0082705   .0248301    -0.33   0.739    -.0569486    .0404077
    seasonal |   .0560999    .035809     1.57   0.117     -.014102    .1263017
        head |  -.0386771   .0165195    -2.34   0.019    -.0710629   -.0062914
     married |   .0500559   .0161317     3.10   0.002     .0184306    .0816813
       dkids |  -.0187618   .0167546    -1.12   0.263    -.0516085    .0140848
      dykids |   .0358303   .0195493     1.83   0.067    -.0024952    .0741557
        smsa |  -.0352754   .0140208    -2.52   0.012    -.0627624   -.0077883
      nwhite |   .0193939   .0186845     1.04   0.299    -.0172361     .056024
     yrdispl |  -.0130974   .0030688    -4.27   0.000    -.0191138   -.0070811
    school12 |  -.0096339   .0167537    -0.58   0.565    -.0424787     .023211
        male |  -.0388263   .0177931    -2.18   0.029    -.0737088   -.0039437
     statemb |   .0012594   .0002039     6.18   0.000     .0008597     .001659
     stateur |   .0181616   .0030859     5.89   0.000     .0121119    .0242113
       _cons |   .1279247   .0882052     1.45   0.147    -.0449975    .3008468
------------------------------------------------------------------------------
```

Test Breuscha-Pagana na heteroskedastyczność:

```
hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: rr rr2 age tenure slack abol seasonal head married dkids
dykids smsa nwhite yrdispl school12 male statemb stateur

         chi2(18)     =    84.68
         Prob > chi2  =   0.0000
```

Czyli odrzucamy hipotezę zerową o braku heteroskedastyczności w modelu, co jest zgodne z oczekiwaniami. Heteroskedastyczność nie wpływa na obciążenie estymatora parametrów, powoduje jednak niewłaściwe oszacowanie błędów standardowych dla poszczególnych parametrów, co może wpływać na wnioski dotyczące (nie-)istotności zmiennych.

Rozwiązaniem tego problemu jest estymacja z wykorzystaniem błędów standardowych odpornych na heteroskedastyczność (macierz wariancji-kowariancji White'a). W naszym przypadku nie zmienia to jednak wniosków dotyczących (nie-)istotności poszczególnych zmiennych.

```
reg y  rr rr2 age tenure slack abol seasonal head married dkids dykids smsa
nwhite yrdispl school12 male statemb stateur, robust
```

```
Regression with robust standard errors                Number of obs =    4877
                                                      F( 18,  4858) =   20.77
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0658
                                                      Root MSE      =  .45031
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |   .6062438   .3957315     1.53   0.126     -.169569    1.382057
         rr2 |  -1.010374   .4951863    -2.04   0.041    -1.981163   -.0395847
         age |    .004297   .0007668     5.60   0.000     .0027936    .0058004
      tenure |   .0054326   .0011656     4.66   0.000     .0031476    .0077176
       slack |   .1260833   .0142299     8.86   0.000     .0981863    .1539802
        abol |  -.0082705   .0259736    -0.32   0.750    -.0591906    .0426496
    seasonal |   .0560999   .0373422     1.50   0.133    -.0171078    .1293075
        head |  -.0386771   .0166618    -2.32   0.020    -.0713418   -.0060125
     married |   .0500559   .0163331     3.06   0.002     .0180358    .0820761
       dkids |  -.0187618   .0168417    -1.11   0.265    -.0517793    .0142556
      dykids |   .0358303   .0196643     1.82   0.069    -.0027207    .0743813
        smsa |  -.0352754   .0139739    -2.52   0.012    -.0626705   -.0078803
      nwhite |   .0193939   .0184205     1.05   0.292    -.0167186    .0555064
     yrdispl |  -.0130974   .0030852    -4.25   0.000    -.0191459   -.0070489
    school12 |  -.0096339    .016992    -0.57   0.571    -.0429459    .0236782
        male |  -.0388263   .0178936    -2.17   0.030    -.0739059   -.0037466
     statemb |   .0012594   .0002038     6.18   0.000     .0008599    .0016588
     stateur |   .0181616   .0029548     6.15   0.000     .0123688    .0239544
       _cons |   .1279247   .0882096     1.45   0.147    -.0450061    .3008555
------------------------------------------------------------------------------
```

Testujemy łączną nieistotność zmiennych nieistotnych indywidualnie na poziomie 5% z wykorzystaniem standardowego testu F:

```
test rr abol seasonal dkids dykids nwhite school12
```

```
 ( 1)  rr = 0
 ( 2)  abol = 0
 ( 3)  seasonal = 0
 ( 4)  dkids = 0
 ( 5)  dykids = 0
 ( 6)  nwhite = 0
 ( 7)  school12 = 0

       F(  7,  4858) =    1.50
            Prob > F =    0.1621
```

P-value statystyki testowej F ponad 16% oznacza, że na poziomie 5% nie możemy odrzucić hipotezy zerowej o łącznej nieistotności siedmiu testowanych powyżej zmiennych.

Estymujemy więc model bez tych zmiennych.

```
reg y  rr2 age tenure slack head married smsa yrdispl male statemb stateur,
robust
```

```
      Source |       SS       df       MS              Number of obs =    4877
-------------+------------------------------           F( 11,  4865) =   30.10
       Model |  67.1824206    11  6.10749278            Prob > F      =  0.0000
    Residual |  987.271137  4865   .20293343            R-squared     =  0.0637
-------------+------------------------------           Adj R-squared =  0.0616
       Total |  1054.45356  4876  .216253806            Root MSE      =  .45048
```

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rr2 |  -.2527459   .0851367    -2.97   0.003    -.4196522   -.0858395
         age |   .0039118   .0007254     5.39   0.000     .0024897     .005334
      tenure |   .0055123   .0012005     4.59   0.000     .0031588    .0078659
       slack |   .1247634   .0132395     9.42   0.000      .098808    .1507188
        head |  -.0376671   .0160508    -2.35   0.019    -.0691339   -.0062004
     married |   .0500746   .0140969     3.55   0.000     .0224383     .077711
        smsa |  -.0344193   .0139521    -2.47   0.014    -.0617717   -.0070668
     yrdispl |  -.0128674   .0030636    -4.20   0.000    -.0188734   -.0068614
        male |  -.0395063   .0175674    -2.25   0.025    -.0739463   -.0050663
     statemb |    .001258   .0002023     6.22   0.000     .0008615    .0016546
     stateur |   .0181462   .0030836     5.88   0.000     .0121008    .0241915
       _cons |   .2532724   .0504925     5.02   0.000     .1542842    .3522605
------------------------------------------------------------------------------
```

Test Breuscha-Pagana na heteroskedastyczność:

```
hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: rr2 age tenure slack head married smsa yrdispl male
statemb stateur

         chi2(11)     =    80.90
         Prob > chi2  =   0.0000
```

Heteroskedastyczność nadal występuje, ale wynika to z charakteru zmiennej objaśnianej i nie wpływa na oszacowania parametrów. Współczynniki oszacowane w liniowym modelu prawdopodobieństwa mają swoją interpretację. W powyższym przykładzie np. każdy kolejny rok stażu w ostatnim miejscu pracy (zmienna `tenure`) zwiększa prawdopodobieństwa ubiegania się o świadczenie z tytułu utraty pracy o 0,55 pp. Co jest zaskakujące, bycie głową rodziny (zmienna `head`) zmniejsza to prawdopodobieństwo o 3,77 pp, itd.

Modele dla zmiennych binarnych

Zmienną objaśnianą jest prawdopodobieństwo wystąpienia określonego zdarzenia uzależnione od wektora zmiennych egzogenicznych:

$$P(Y_j = 1 \mid X_{1j},...,X_{kj}) = F(x_i^{'}\beta)$$

przy czym w modelu **probit** wykorzystujemy dystrybuantę rozkładu normalnego:

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\Pi}} exp(-\frac{1}{2}t^2)dt\,,$$

natomiast w modelu **logit** wykorzystywana jest standardowa funkcja logistyczna:

$$F(w) = \frac{e^w}{1+e^w} = \frac{1}{1+e^{-w}}\,,$$

Estymacja modelu Logit

Estymacji modelu logit dokonujemy za pomocą polecenia `logit`:

```
logit y  rr rr2 age tenure slack abol seasonal head married dkids dykids smsa
nwhite yrdispl school12 male statemb stateur


Iteration 0:   log likelihood =  -3043.028
Iteration 1:   log likelihood = -2875.8198
Iteration 2:   log likelihood = -2873.2003
Iteration 3:   log likelihood = -2873.1965

Logit estimates                                   Number of obs   =       4877
                                                  LR chi2(19)     =     339.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -2873.1965                       Pseudo R2       =     0.0558

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |    3.06808   1.868225     1.64   0.101    -.5935732    6.729733
         rr2 |  -4.890618   2.333521    -2.10   0.036    -9.464236   -.3170007
         age |   .0676968   .0239095     2.83   0.005     .020835    .1145586
        age2 |  -.0059681   .0030383    -1.96   0.050    -.0119231    -.000013
      tenure |   .0312492   .0066443     4.70   0.000     .0182267    .0442717
       slack |    .624822   .0706385     8.85   0.000     .4863731    .7632709
        abol |  -.0361753   .1178082    -0.31   0.759    -.2670751    .1947245
    seasonal |    .270874   .1711711     1.58   0.114    -.0646152    .6063633
        head |  -.2106822    .081226    -2.59   0.009    -.3698822   -.0514821
     married |   .2422656   .0794099     3.05   0.002     .0866251    .3979061
       dkids |  -.1579269   .0862177    -1.83   0.067    -.3269105    .0110566
      dykids |   .2058941   .0974924     2.11   0.035     .0148126    .3969756
        smsa |  -.1703537   .0697808    -2.44   0.015    -.3071216   -.0335858
      nwhite |   .0740701   .0929562     0.80   0.426    -.1081208     .256261
     yrdispl |  -.0637001   .0149972    -4.25   0.000    -.0930941   -.0343062
    school12 |  -.0652576   .0824126    -0.79   0.428    -.2267834    .0962681
        male |   -.179829    .087535    -2.05   0.040    -.3513944   -.0082636
     statemb |    .006027    .001009     5.97   0.000     .0040494    .0080046
     stateur |   .0956198   .0159116     6.01   0.000     .0644336     .126806
       _cons |  -2.800499   .6041675    -4.64   0.000    -3.984645   -1.616352
------------------------------------------------------------------------------
```

Zapamiętujemy wyniki estymacji pod nazwą logit1:

```
est store logit1
```

Testowanie warunków ograniczających przy pomocy testu ilorazu wiarogodności (LR)

Nie możemy stosować testu F, bo estymacja nie jest dokonywana MNK, więc nie ma RSS. Ekwiwalentem jest w tym przypadku między innymi test ilorazu wiarogodności. Intuicyjnie rzecz biorąc sprawdzamy, czy narzucone ograniczenia w istotny sposób wpływają na logarytm wiarogodności.

Hipoteza zerowa: narzucone warunki ograniczające są prawdziwe. Jeśli H0 jest prawdziwa, statystyka testowa $LR=-2[\ln L_R-\ln R_U]$ ma w przybliżeniu rozkład chi-kwadrat z liczbą stopni swobody równą liczbie narzuconych ograniczeń.

Zanim przetestujemy łączną nieistotność zmiennych nieistotnych indywidualnie dokonujemy oszacowania modelu z warunkami ograniczającymi (czyli bez zmiennych nieistotnych):

```
logit y  rr2 age tenure slack head married smsa yrdispl male statemb stateur
```

```
Iteration 0:   log likelihood =  -3043.028
Iteration 1:   log likelihood = -2883.0745
Iteration 2:   log likelihood = -2880.5447
Iteration 3:   log likelihood = -2880.5412

Logit estimates                                   Number of obs   =       4877
                                                  LR chi2(11)     =     324.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -2880.5412                       Pseudo R2       =     0.0534

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rr2 |  -1.151087   .4204778    -2.74   0.006    -1.975208   -.3269655
         age |   .0195309   .0036723     5.32   0.000     .0123333    .0267286
      tenure |   .0308234   .0065936     4.67   0.000     .0179001    .0437467
       slack |   .6091688   .0659404     9.24   0.000      .479928    .7384096
        head |  -.1879037   .0781542    -2.40   0.016    -.3410832   -.0347242
     married |   .2403562   .0687984     3.49   0.000     .1055137    .3751986
        smsa |  -.1664986   .0693336    -2.40   0.016      -.30239   -.0306072
     yrdispl |  -.0615481   .0149413    -4.12   0.000    -.0908324   -.0322638
        male |  -.1876878   .0860323    -2.18   0.029     -.356308   -.0190676
     statemb |   .0061076   .0009994     6.11   0.000     .0041489    .0080663
     stateur |   .0953049   .0158821     6.00   0.000     .0641765    .1264332
       _cons |  -1.373795    .255853    -5.37   0.000    -1.875258   -.8723321
------------------------------------------------------------------------------
```

Zapamiętujemy wyniki estymacji pod nazwą logit2:

```
est store logit2
```

Korzystając z zapamiętanych wcześniej wyników dla modelu bez ograniczeń (logit1) i modelu z ograniczeniami (logit2) przeprowadzamy test LR wyświetlając jednocześnie statystyki opisowe.

```
lrtest logit1 logit2, stats

likelihood-ratio test                                 LR chi2(7)  =     10.86
(Assumption: logit2 nested in logit1)                 Prob > chi2 =    0.1449

-----------------------------------------------------------------------------
Model        |   nobs    ll(null)   ll(model)     df         AIC         BIC
-------------+---------------------------------------------------------------
      logit2 |   4877   -3043.028   -2880.541     12     5785.082     5862.99
      logit1 |   4877   -3043.028   -2875.112     19     5788.224    5911.577
-----------------------------------------------------------------------------
```

Nie możemy więc odrzucić H0, że narzucone ograniczenia są poprawne. Widać również, że kryteria informacyjne wskazują, że model „logit2" jest lepszy niż „logit1".

Testowanie liniowych warunków ograniczających przy pomocy statystyki Walda:

Alternatywną formą testowania hipotez dotyczących oszacowanych parametrów jest statystyka Walda, wywoływana komendą `test`:

```
test  rr abol seasonal dkids dykids nwhite school12

 ( 1)  rr = 0
 ( 2)  abol = 0
 ( 3)  seasonal = 0
 ( 4)  dkids = 0
 ( 5)  dykids = 0
 ( 6)  nwhite = 0
 ( 7)  school12 = 0

           chi2(  7) =   10.81
         Prob > chi2 =    0.1472
```

Testowanie nieliniowych warunków ograniczających przy pomocy statystyki Walda:

Dzięki statystyce Walda można również testować nieliniowe warunki ograniczające narzucane na parametry. Służy do tego polecenie `testnl`. Jego składnia różni się nieco od składni komendy `test` – zamiast podać nazwy zmiennych, przy których parametry testujemy, musimy użyć zapisu `_b[nazwa_zmiennej]`.

Przetestujmy hipotezę zerową mówiącą, że kwadrat oszacowania parametru przy zmiennej `age` jest równy 1 stosując dwa równoważne zapisy tej hipotezy:

$$H_0 : \beta_{age} * \beta_{age} - 1 = 0$$

```
testnl _b[age]*_b[age]-1=0

  (1)  _b[age]*_b[age]-1 = 0

             chi2(1) =    4.86e+07
         Prob > chi2 =      0.0000
```

$$H_0 : \beta_{age} - \frac{1}{\beta_{age}} = 0$$

```
testnl _b[age]-1/_b[age]=0

  (1)  _b[age]-1/_b[age] = 0

             chi2(1) =       28.24
         Prob > chi2 =      0.0000
```

Mimo że alternatywne sformułowania H0 są równoważne, oszacowana wartość statystyki Walda jest w obu przypadkach różna, co wskazuje, że statystyka Walda nie jest niezmiennicza względem zdefiniowania hipotezy zerowej.

Za pomocą polecenia `testnl` można oczywiście testować również liniowe warunki ograniczające oraz łącznie liniowe i nieliniowe ograniczenia narzucane na parametry, jednak w przypadku testowania wyłącznie hipotez liniowych polecenie `test` działa szybciej.

Interpretacja współczynników i ilorazy szans

Interpretacją współczynników z modelu logitowego jest procentowy wpływ jednostkowej zmiany wartości zmiennej objaśniającej na iloraz szans liczony jako stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki.

Aby zamiast oszacowań parametrów wygenerować ilorazy szans dla poszczególnych zmiennych należy użyć polecenia `logit` z opcją `or` (odds-ratio) lub użyć polecenia `logistic`.

```
logit y  rr2 age tenure slack head married smsa yrdispl male statemb stateur, or
```

lub

```
logistic y  rr2 age tenure slack head married smsa yrdispl male statemb stateur


Logistic regression                               Number of obs   =       4877
                                                  LR chi2(11)     =     324.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -2880.5412                       Pseudo R2       =     0.0534

------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rr2 |   .3162928   .1329941    -2.74   0.006     .1387324    .7211086
         age |   1.019723   .0037448     5.32   0.000      1.01241    1.027089
      tenure |   1.031303      .0068     4.67   0.000     1.018061    1.044718
       slack |   1.838902   .1212579     9.24   0.000     1.615958    2.092605
        head |   .8286945    .064766    -2.40   0.016     .7109997    .9658717
     married |   1.271702   .0874911     3.49   0.000     1.111281     1.45528
        smsa |    .846624   .0586995    -2.40   0.016     .7390498    .9698565
     yrdispl |   .9403077   .0140494    -4.12   0.000     .9131708    .9682512
        male |   .8288735   .0713099    -2.18   0.029     .7002569    .9811131
     statemb |   1.006126   .0010055     6.11   0.000     1.004157    1.008099
     stateur |   1.099994   .0174702     6.00   0.000     1.066281    1.134774
------------------------------------------------------------------------------
```

Obliczanie efektów krańcowych (cząstkowych)

Ponieważ wartości parametrów oszacowane w modelach dla zmiennych binarnych są trudne w bezpośredniej interpretacji (interpretowalne są jedynie ich znaki) i jednocześnie nieporównywalne z innymi modelami (np. oszacowania liniowego modelu prawdopodobieństwa, modelu logit i probit dla tej samej zmiennej objaśnianej i analogicznego zbioru zmiennych objaśniających), oblicza się tzw. efekty krańcowe, które można bezpośrednio porównywać między modelami.

$$\frac{\partial E(y_i \mid x)}{\partial x} = f(x\beta)\beta$$

Służy do tego polecenie `mfx compute`:

```
mfx compute

Marginal effects after logit
      y  = Pr(y) (predict)
         =   .6970697
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     rr2 |  -.2430676      .08878   -2.74   0.006  -.417064 -.069071    .20344
     age |   .0041242      .00077    5.33   0.000   .002608  .005641     36.13
  tenure |   .0065088      .00139    4.69   0.000   .003788   .00923   5.66414
   slack*|   .1273793      .01353    9.42   0.000    .10087  .153889   .476112
    head*|  -.0391275      .01603   -2.44   0.015  -.070547 -.007708   .680541
 married*|   .0513508      .01485    3.46   0.001   .022251   .08045   .632766
    smsa*|   -.034793      .01433   -2.43   0.015  -.062873 -.006713   .652655
 yrdispl |  -.0129967      .00315   -4.12   0.000  -.019175 -.006818   5.20361
    male*|  -.0388319      .01742   -2.23   0.026  -.072971 -.004693   .764199
 statemb |   .0012897      .00021    6.12   0.000   .000877  .001702    180.66
 stateur |   .0201249      .00334    6.02   0.000   .013575  .026675   7.51103
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Efekty krańcowe są porównywalne dla różnych modeli i mogą być interpretowane jest jako wpływ zmiany zmiennej niezależnej o jednostkę na prawdopodobieństwo sukcesu (czyli przyjęcia przez zmienną objaśnianą wartości 1). Np. z powyższego zestawienia wynika, że bycie mężczyzną zmniejsza prawdopodobieństwo ubiegania się o świadczenie z tytułu utraty pracy o 3,88 pp. (fałszywie pojęta męska duma?), natomiast wzrost stanowej stopy bezrobocia o 1 pp. powoduje, że prawdopodobieństwo ubiegania się o świadczenie rośnie o 2 pp.

Efekty krańcowe zależą od wartości zmiennych objaśniających. Standardowo liczone są dla przeciętnych wartości zmiennych egzogenicznych. Można je również policzyć dla zadanych wartości wybranych (lub wszystkich) regresorów. Na przykład efekty krańcowe dla żonatego mężczyzny w wieku lat pięćdziesięciu policzymy stosując komendę:

```
   mfx compute, at(age=50, married=1, male=1)

Marginal effects after logistic
      y  = Pr(y) (predict)
         =  .75919788
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     rr2 |  -.2104376       .0781   -2.69   0.007  -.363513 -.057362    .20344
     age |   .0035706      .00058    6.15   0.000   .002432  .004709        50
  tenure |    .005635      .00127    4.42   0.000   .003137  .008133   5.66414
   slack*|   .1103696      .06594    1.67   0.094  -.018871   .23961   .476112
    head*|  -.0337421      .07815   -0.43   0.666  -.186922  .119437   .680541
 married*|   .0466213       .0688    0.68   0.498  -.088221  .181464         1
    smsa*|  -.0300339      .06933   -0.43   0.665  -.165925  .105858   .652655
 yrdispl |   -.011252      .00274   -4.11   0.000  -.016623 -.005881   5.20361
    male*|  -.0326293      .08603   -0.38   0.704   -.20125  .135991         1
 statemb |   .0011166      .00019    5.97   0.000    .00075  .001483    180.66
 stateur |   .0174233      .00293    5.94   0.000   .011675  .023172   7.51103
------------------------------------------------------------------------------
```

Miary dopasowania modelu

Standardowo liczone jest pseudo-R^2 – raportowane wraz z wynikami estymacji.

Aby uzyskać inne miary można ściągnąć pakiet FITSTAT (należy będąc w pakiecie STATA nacisnąć „clrl+3", następnie „s" jak search i zaznaczywszy „Search net resources" wpisać w okienku wyszukiwania „fitstat"; po znalezieniu zainstalować – wykonując tę operację na WNE należy utworzyć na dysku L:\ katalog STATA8 – inaczej nie zadziała), a następnie po estymacji modelu użyć komendy `fitstat`:

```
fitstat

Measures of Fit for logit of y

Log-Lik Intercept Only:    -3043.028     Log-Lik Full Model:        -2875.112
D(4858):                    5750.224     LR(18):                      335.832
                                         Prob > LR:                     0.000
McFadden's R2:                 0.055     McFadden's Adj R2:             0.049
Maximum Likelihood R2:         0.067     Cragg & Uhler's R2:            0.093
McKelvey and Zavoina's R2:     0.099     Efron's R2:                    0.069
Variance of y*:                3.653     Variance of error:             3.290
Count R2:                      0.698     Adj Count R2:                  0.045
AIC:                           1.187     AIC*n:                      5788.224
BIC:                     -35505.300      BIC':                       -182.971
```

Tablica trafności dopasowań

Tablica trafności dopasowań może być innym sposobem mierzenia jakości dopasowania modelu do danych. Dla ostatnio przeprowadzonej estymacji uzyskamy ją stosując komendę `lstat`.

```
lstat

Logistic model for y

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      3183          1317  |       4500
     -     |       152           225  |        377
-----------+--------------------------+-----------
   Total   |      3335          1542  |       4877

Classified + if predicted Pr(D) >= .5
True D defined as y != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   95.44%
Specificity                     Pr( -|~D)   14.59%
Positive predictive value       Pr( D| +)   70.73%
Negative predictive value       Pr(~D| -)   59.68%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   85.41%
False - rate for true D         Pr( -| D)    4.56%
False + rate for classified +   Pr(~D| +)   29.27%
False - rate for classified -   Pr( D| -)   40.32%
--------------------------------------------------
Correctly classified                        69.88%
--------------------------------------------------
```

Standardowo przyjmowane jest, że model przewiduje $y_i=1$ jeżeli oszacowane prawdopodobieństwo jest większe niż 0,5. Jeżeli chcemy zastosować inny punkt odcięcia, np. 0,25, należy użyć następującej komendy:

```
lstat, cutoff(0.25)
```

Wartości dopasowane

Wygenerujmy wartości dopasowane z oszacowanego modelu logitowego oraz oszacowane prawdopodobieństwo sukcesu (ubiegania się o przysługujące świadczenie):

```
predict xb, xb
predict p, p
```

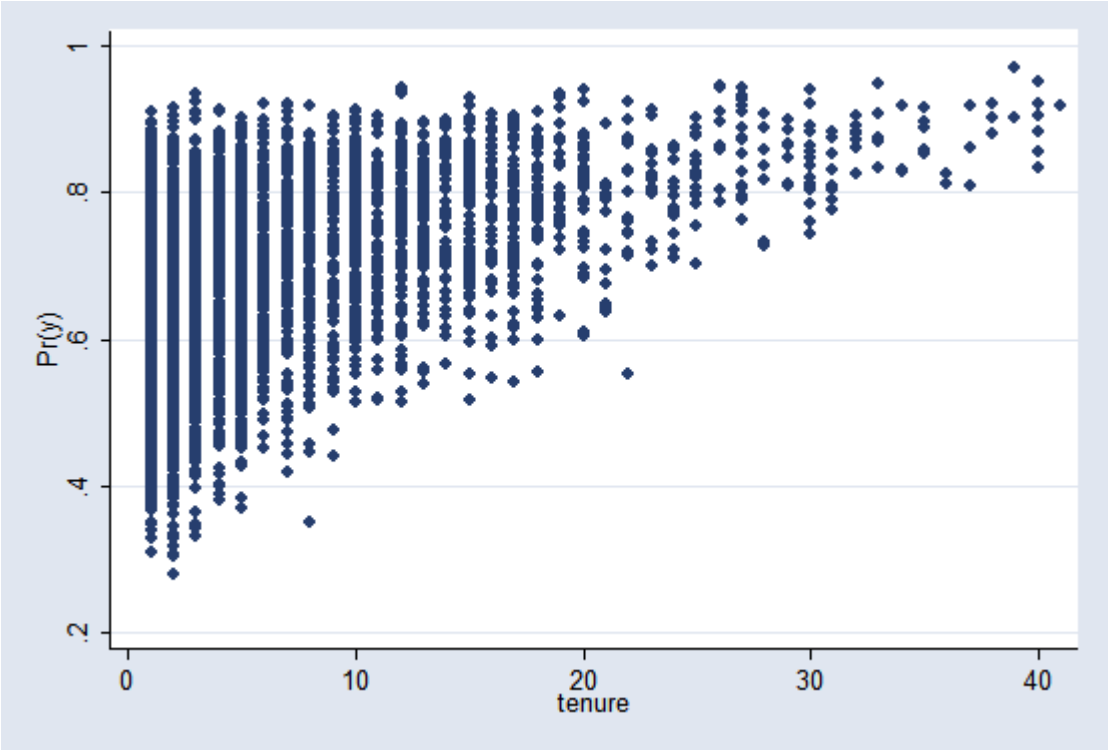
Między wartościami dopasowanymi z modelu a oszacowanym prawdopodobieństwem zachodzi zależność zgodna z przyjętą funkcją "gęstości":

$$p = P(Y_j = 1 \mid X_{1j},...,X_{kj}) = F(x_i'\beta)$$

czyli dla modelu logitowego `p=1/[1+exp(-xb)]`, natomiast dla modelu probitowego `p` jest wartością dystrybuanty standardowego rozkładu normalnego w punkcie `xb` (`p=Φ(xb)`).

Narysujmy wykres zależności przewidywanego prawdopodobieństwa sukcesu od długości zatrudnienia w ostatnim miejscu pracy.

```
scatter p tenure
```



Z powyższego wykresu widać, że wraz ze wzrostem stażu w ostatnim miejscu zatrudnienia prawdopodobieństwo ubiegania się o świadczenie przysługujące z tytułu ubezpieczenia od utraty pracy wyraźnie wzrasta. Dla osób, które zanim straciły pracę przepracowały w danej firmie ponad 35 lat prawdopodobieństwo ubiegania się o świadczenie z tytułu utraty pracy wynosi ponad 80%.

Estymacja modelu Probit

Estymacji modelu probit dokonujemy za pomocą polecenia `probit`:

```
probit y rr rr2 age tenure slack abol seasonal head married dkids dykids smsa
nwhite yrdispl school 12 male statemb stateur


Iteration 0:   log likelihood =  -3043.028
Iteration 1:   log likelihood = -2877.2378
Iteration 2:   log likelihood = -2876.2341
Iteration 3:   log likelihood = -2876.2339

Probit estimates                                  Number of obs   =       4877
                                                  LR chi2(18)     =     333.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -2876.2339                       Pseudo R2       =     0.0548

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |    1.81413   1.126008     1.61   0.107    -.3928044    4.021065
         rr2 |  -2.965477   1.409031    -2.10   0.035    -5.727126    -.203827
         age |   .0128577    .002348     5.48   0.000     .0082556    .0174597
      tenure |   .0171165   .0038063     4.50   0.000     .0096562    .0245768
       slack |    .369949   .0423106     8.74   0.000     .2870218    .4528763
        abol |   -.027034   .0717981    -0.38   0.707    -.1677557    .1136877
    seasonal |   .1576748   .1038971     1.52   0.129    -.0459599    .3613094
        head |   -.111508   .0487228    -2.29   0.022    -.2070029   -.0160131
     married |   .1492617   .0477113     3.13   0.002     .0557493     .242774
       dkids |  -.0669911   .0498136    -1.34   0.179    -.1646241    .0306418
      dykids |   .1065534   .0579546     1.84   0.066    -.0070355    .2201423
        smsa |  -.1000473    .041819    -2.39   0.017    -.1820112   -.0180835
      nwhite |   .0589318   .0558523     1.06   0.291    -.0505367    .1684003
     yrdispl |  -.0379514   .0090638    -4.19   0.000    -.0557162   -.0201866
    school12 |  -.0300342   .0493823    -0.61   0.543    -.1268218    .0667534
        male |  -.1135565   .0526911    -2.16   0.031    -.2168291   -.0102839
     statemb |   .0036919   .0006066     6.09   0.000     .0025031    .0048807
     stateur |   .0567939   .0094476     6.01   0.000      .038277    .0753108
       _cons |  -1.181325   .2624662    -4.50   0.000    -1.695749   -.6669005
------------------------------------------------------------------------------
```

Zapamiętujemy wyniki estymacji pod nazwą probit1:

```
est store probit1
```

Łączną nieistotność parametrów w modelach logit i probit możemy również testować za pomocą **statystyki Walda**, wywoływanego komendą `test`. Testujemy ponownie zmienne indywidualnie nieistotne na poziomie 5% (są to te same zmienne, które były indywidualnie nieistotne w obu poprzednich regresjach):

```
test  rr abol seasonal dkids dykids nwhite school12

 ( 1)  rr = 0
 ( 2)  abol = 0
 ( 3)  seasonal = 0
 ( 4)  dkids = 0
 ( 5)  dykids = 0
 ( 6)  nwhite = 0
 ( 7)  school12 = 0

           chi2(  7) =   10.98
         Prob > chi2 =    0.1395
```

Nie możemy odrzucić hipotezy zerowej o łącznej nieistotności tych zmiennych, więc redukujemy model:

```
probit y  rr2 age tenure slack head married smsa yrdispl male statemb stateur

Iteration 0:   log likelihood =  -3043.028
Iteration 1:   log likelihood = -2882.6368
Iteration 2:   log likelihood = -2881.7372
Iteration 3:   log likelihood = -2881.7371

Probit estimates                                  Number of obs   =       4877
                                                  LR chi2(11)     =     322.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -2881.7371                       Pseudo R2       =     0.0530

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rr2 |  -.7020753   .2523342    -2.78   0.005    -1.196641   -.2075092
         age |   .0117275   .0021911     5.35   0.000     .0074331    .0160219
      tenure |   .0173067   .0037642     4.60   0.000      .009929    .0246844
       slack |   .3661648   .0395464     9.26   0.000     .2886553    .4436743
        head |  -.1101841   .0473191    -2.33   0.020    -.2029279   -.0174403
     married |   .1441728   .0415553     3.47   0.001     .0627259    .2256198
        smsa |  -.0974368    .041559    -2.34   0.019     -.178891   -.0159826
     yrdispl |  -.0371914   .0090398    -4.11   0.000    -.0549092   -.0194736
        male |  -.1139709   .0519568    -2.19   0.028    -.2158042   -.0121375
     statemb |    .003684   .0006012     6.13   0.000     .0025057    .0048624
     stateur |   .0566916   .0094341     6.01   0.000     .0382012     .075182
       _cons |  -.8093586   .1530845    -5.29   0.000    -1.109399   -.5093185
------------------------------------------------------------------------------
```

Zapamiętujemy wyniki estymacji pod nazwą probit2:

```
est store probit2
```

Prawidłowość narzuconych warunków ograniczających możemy potwierdzić również testem LR:

```
lrtest probit1 probit2, stats

likelihood-ratio test                                 LR chi2(7)  =     11.01
(Assumption: probit2 nested in probit1)               Prob > chi2 =    0.1383

-----------------------------------------------------------------------------
Model        |   nobs    ll(null)   ll(model)     df         AIC         BIC
-------------+---------------------------------------------------------------
     probit2 |   4877   -3043.028   -2881.737     12     5787.474    5865.382
     probit1 |   4877   -3043.028   -2876.234     19     5790.468    5913.821
-----------------------------------------------------------------------------
```

Miary dopasowania modelu

```
fitstat

Measures of Fit for probit of y

Log-Lik Intercept Only:    -3043.028     Log-Lik Full Model:        -2881.737
D(4865):                    5763.474     LR(11):                      322.582
                                         Prob > LR:                     0.000
McFadden's R2:                 0.053     McFadden's Adj R2:             0.049
Maximum Likelihood R2:         0.064     Cragg & Uhler's R2:            0.090
McKelvey and Zavoina's R2:     0.110     Efron's R2:                    0.067
Variance of y*:                1.124     Variance of error:             1.000
Count R2:                      0.699     Adj Count R2:                  0.048
AIC:                           1.187     AIC*n:                      5787.474
BIC:                       -35551.495    BIC':                        -229.167
```

Obliczanie efektów krańcowych

W przypadku modelu probit oszacowane parametry nie mają żadnej bezpośredniej interpretacji. Jedynie ich znak może być odczytywany jako pozytywny lub negatywny wpływ analizowanej zmiennej na prawdopodobieństwo sukcesu. Interpretować (i porównywać z innymi modelami) można dopiero efekty krańcowe.

```
mfx compute

Marginal effects after probit
      y  = Pr(y) (predict)
         =  .69402302
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     rr2 |   -.246271      .08851   -2.78   0.005  -.419752  -.07279    .20344
     age |   .0041137      .00077    5.36   0.000   .002608  .005619     36.13
  tenure |   .0060708      .00132    4.60   0.000   .003485  .008657   5.66414
   slack*|   .1273458      .01356    9.39   0.000   .100772   .15392   .476112
    head*|  -.0382403      .01624   -2.36   0.019  -.070064 -.006417   .680541
 married*|   .0510228      .01482    3.44   0.001   .021973  .080073   .632766
    smsa*|   -.033908      .01434   -2.36   0.018  -.062018 -.005798   .652655
 yrdispl |  -.0130458      .00317   -4.12   0.000  -.019259 -.006832   5.20361
    male*|  -.0393392      .01763   -2.23   0.026  -.073892 -.004786   .764199
 statemb |   .0012923      .00021    6.13   0.000   .000879  .001705    180.66
 stateur |    .019886       .0033    6.02   0.000   .013409  .026363   7.51103
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Tablica trafności dopasowań

```
lstat

Probit model for y

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      3197          1330  |       4527
     -     |       138           212  |        350
-----------+--------------------------+-----------
   Total   |      3335          1542  |       4877

Classified + if predicted Pr(D) >= .5
True D defined as y != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   95.86%
Specificity                     Pr( -|~D)   13.75%
Positive predictive value       Pr( D| +)   70.62%
Negative predictive value       Pr(~D| -)   60.57%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   86.25%
False - rate for true D         Pr( -| D)    4.14%
False + rate for classified +   Pr(~D| +)   29.38%
False - rate for classified -   Pr( D| -)   39.43%
--------------------------------------------------
Correctly classified                        69.90%
--------------------------------------------------
```

Literatura:

McCall, B.P., 1995,The Impact of Unemployment Insurance Benefit Levels on Recipiency, *Journal of Business and Economic Statistics*, vol. 13, pp. 189-198.
Verbeek, M., 2004, A guide to Modern Econometrics, John Wiley & Sons Ltd., wydanie 2.