3 Modele wyborów dyskretnych

Przykłady, rozszerzenia i interpretacja

Dane do przykładu pochodzą z piątej fali badania *Health and Retirement Study* przeprowadzonego w Stanach Zjednoczonych Ameryki Północnej w 2002 roku. Analizowaną zmienną zależną będzie zakup dodatkowego ubezpieczenia medycznego `ins`. Zmienne objaśniające zawierają informacje dotyczące zdrowa, charakterystyk społeczno-ekonomicznych oraz informacji o żonie ubezpieczonego. Próba jest ograniczona do żonatych mężczyzn.

```
net from http://www.stata-press.com/data/mus
use mus14data.dta
* usuwanie zmiennych opisujacych interakcje
drop age2 agefem agechr agewhi

* zmienne glowne
global xlist age hstatusg hhincome educyear married hisp

* logarytm dochodu
generate linc=ln(hhinc)

* dodatkowe zmienne objasniajace
global extralist linc female white chronic adl sretire

summarize ins retire $xlist $extralist
```

Dokonujemy oszacowań wartości parametrów dla modelu logitowego, probitowego oraz liniowego modelu prawdopodobieństwa, oraz ich wariantów z odporną na zjawisko heteroscedastyczności macierzą wariancji-kowariancji, za każdym razem zapamiętujemy wynik uzyskanych ocen parametrów. Następnie tworzymy tabelę podsumowującą uzyskane wyniki

```
*Model logitowy logit ins retire $xlist
estimates store blogit

*Model probitowy
qui probit ins retire $xlist
estimates store bprobit

*Liniowy model prawdopodobieństwa qui reg ins retire $xlist
estimates store bols
```

```
*Logit z macierzą odporną qui logit ins retire $xlist, vce(robust)
estimates store blogitr

*Probit z macierzą odporną qui probit ins retire $xlist,
vce(robust) estimates store bprobitr

*Regresja z macierzą odporna qui reg ins retire $xlist,
vce(robust) estimates store bolsr

*Podsumowanie i porównanie modeli
estimates table blogit blogitr bprobit bprobitr bols bolsr, /*
*/ t stats(N ll) b(%7.3f) stfmt(%8.2f)
```

Następnie generujemy interakcje między zmiennymi objaśniającymi (kwadrat wieku, iloczyn wieku i płci, iloczyn wieku i indykatora choroby przewlekłej, iloczyn wieku i rasy białej). Dołączamy dodatkowe zmienne do specyfikacji modelu logitowego, szacujemy parametry modelu rozszerzonego i weryfikujemy istotność dołączonych zmiennych wykorzystując statystykę testu Walda.

```
* Test Walda (istotność interakcji)
generate age2 = age*age
generate agefem = age*female
generate agechr = age*chronic
generate agewhi = age*white

global intlist age2 agefem agechr agewhi

quietly logit ins retire $xlist $intlist
test $intlist
```

Uzyskana wartość statystyki testowej i jej *p-value* wskazują na brak podstaw do odrzucenia hipotezy o statystycznej nieistotności dodatkowych ocen parametrów.

Alternatywnym sposobem weryfikacji hipotez, preferowanym w przypadku wykorzystania metody największej wiarogodności do szacowania ocen parametrów modelu, jest przeprowadzenie testu ilorazu wiarogodności LR. W celu obliczenia wartości statystyki testowej należy oszacować parametry modelu bez ograniczeń, oraz parametry modelu z narzuconymi ograniczeniami. Następnie należy porównać dwa wektory oszacowań parametrów.

```
*Test LR (istotność interakcji)
quietly logit ins retire $xlist $intlist
estimates store B

quietly logit ins retire $xlist lrtest B
```

Wyniki testu są zbliżone do wyników testu Walda. Ale nie w każdym przypadku będzie zachodzić taka sytuacja.

Stuckel (1988) zaproponował aby badać poprawność specyfikacji modeli dla binarnej zmiennej zależnej, jego praca dotyczyła modeli logitowych, poprzez dodanie do modelu kwadratu wartości dopasowanej jako dodatkowego regresora. Test ten możemy przeprowadzić wykorzystując polecenie `linktest`

```
/* Test poprawności formy funkcyjnej */
quietly logit ins retire $xlist
linktest
```

Uzyskana wartość statystyki testowej i jej *p-value* wskazują, iż model nie jest dobrze dopasowany do danych.

Standardowe modele probit i logit zakładają stałą wariancję składnika losowego. Założenie o stałości wariancji modelu probitowego może być testowane poprzez oszacowanie modelu heteroscedastycznego probitu, który nie zakłada stałej wariancji. Model logitowy z heteroscedastycznością nie jest standardową komendą, niemniej jego zaprogramowanie nie jest trudne.

```
hetprob ins retire $xlist, het(chronic) //Heteroscedastic Probit
```

Test Hosmera-Lemeshowa sprawdza dopasowanie modelu do danych empirycznych poprzez porównanie częstotliwości próbkowych zmiennej zależnej z jej wartościami dopasowanymi z modelu. Konstrukcja testu jest zbliżona do testu χ^2 Pearsona. Pierwszym krokiem jest podział próby na G podpróbek według kwantyli rozkładu. Następnie obliczana jest wartość statystyki

$$HL = \sum_{g=1}^{G} \frac{(\hat{p_g} - \bar{y}_g)^2}{\bar{y}_g(1 - \bar{y}_g)}$$

Statystyka testowa ma rozkład $\chi^2(G-2)$, gdzie g jest liczbą podgrup.

```
/* Miary dopasowania */

quietly logit ins retire $xlist
estat gof, group(4)
estat gof, group(10)
```

Tablica klasyfikacji jest standardowym narzędziem diagnostycznym porównującym wartości rzeczywiste z próby i dopasowane wynikające z wartości oszacowań parametrów modelu

```
estat classification
```

Do obliczenia dopasowanych prawdopodobieństw wykorzystujemy polecenie `predict`. Różnice w wartościach dopasowanych modelu probitowego i logitowego są na ogół nieznaczne.

```
/* Predykcja */
* calculate predicted probabilities
quietly logit ins hhincome
predict plogit, pr

quietly probit ins hhincome
predict pprobit, pr

quietly reg ins hhincome
predict pols, xb
```

Aby się o tym przekonać porównamy metodą graficzną wartości dopasowane z obu modeli

```
sort hhincome
graph twoway (scatter ins hhincome,msize(vsmall) jitter(3)) /*
*/ (line plogit hhincome, clstyle(p1) ) /*
*/ (line pprobit hhincome, clstyle(p2) ) /*
*/ (line pols hhincome, clstyle(p3) ) /*
*/ scale(1.2) plotregion(style(none)) /*
*/ title("Predicted probabilities across models") /*
*/ xtitle("HHINCOME (hhincome)", size(medlarge)) xscale(titlegap(*5)) /*
*/ ytitle("Predicted probability", size(medlarge)) yscale(titlegap(*5)) /*
*/ legend(pos(1)) ring(0) col(1)) legend(size(small))  /*
*/ legend(label(1 "Actual Data (jittered)") label(2 "Logit") /*
*/ label(3 "Probit") label(4 "OLS"))
```

Jak widać wartości dopasowane uzyskane z modelu logitowego i probitowego kształtują się w podobny sposób, wyraźnie różniąc się od wartości dopasowanych pochodzących z modelu liniowego prawdopodobieństwa.

```
/* efekty krańcowe */
quietly logit ins retire $xlist
mfx
mfx, at(1 75 1 35 12 1 1)
/* przecietna wielkość pochodnej */
net serach margeff
margeff
/* zmiana efektów krańcowych wywołana jednostkową zmianą regresora */
prchange

/* model ze zmiennymi endogenicznymi */
global xlist2 female age age2 educyear married hisp white chronic /*
*/ adl hstatusg
probit ins linc $xlist2, vce(robust)

global ivlist retire sretire
ivprobit ins $xlist2 (linc=$ivlist),vce(robust)
```

W zastosowania w równaniu modelu obok zmiennych objaśniających często występują interakcje. Ich zadaniem jest wskazanie jaki wpływ na zmianę wartości jednej ze zmiennych objaśniających ma inna zmienna objaśniająca. Pomimo tego, że interakcje są często wykorzystywane przez badaczy, równie często wartości parametrów przy takich zmiennych są w nieprawidłowy sposób interpretowane.

W przypadku modelu liniowego interpretacja oszacowanej wartości współczynnika dla interakcji jest niemal bezpośrednia. Przyjmijmy następujące oznaczenia. Niech y będzie ciągłą zmienna zależną, której wartość zależy od dwóch zmiennych objaśniających x_1 oraz x_2, ich interakcji, oraz macierzy dodatkowych zmiennych $\mathbb{X}$. Wartości β oznaczają nieznane wartości parametrów modelu, które wymagają oszacowania. Jeżeli zmienne x_1 i x_2 są ciągłe, efekt ich interakcji jest obliczany jako pierwsza pochodna wartości oczekiwanej y względem obu zmiennych

$$\frac{\partial^2 E(y \mid x_1, x_2, \mathbb{X})}{\partial x_1 \partial x_2} = \beta_{12}$$

Jeżeli x_1 i x_2 są dyskretne, to efekt interakcji jest dyskretną różnicą

$$\frac{\partial^2 E(y \mid x_1, x_2, \mathbb{X})}{\Delta x_1 \Delta x_2} = \beta_{12}$$

Ale tak jest w przypadku modelu liniowego. W przypadku modelu nieliniowego, na przykład modelu probit, postać analityczna wyrażenia opisującego warunkową wartość oczekiwaną zmiennej zależnej jest bardziej skomplikowana. Przyjmijmy tym razem, że y jest wskazującą zmienną zero-jedynkową. Wówczas jej warunkowa wartość oczekiwana wynosi

$$E(y \mid x_1, x_2, \mathbb{X}) = \Phi(x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_{12} + \mathbb{X}\beta) = \Phi(\cdot)$$

gdzie Φ jest dystrybuantą standardowego rozkładu normalnego. Gdy x_1 oraz x_2 są ciągłe efekt interakcji jest pochodną krzyżową wartości oczekiwanej y.

$$\frac{\partial^2 \Phi(\cdot)}{\partial x_1 \partial x_2} = \beta_{12}\Phi\prime(\cdot) + (\beta_1 + x_2\beta_{12})(\beta_2 + x_1\beta_{12})\Phi\prime\prime(\cdot)$$

Jednak większość pakietów statystycznych, a w konsekwencji badaczy, oblicza wartość efektu interakcji jako

$$\frac{\partial^2 \Phi(\cdot)}{\partial x_1 \partial x_2} = \beta_{12}\Phi\prime(\cdot)$$

Ponadto, warto wskazać za Ai i Norton (2003):

- Efekt interakcji może być niezerowy nawet jeżeli parametr $\beta_{12} = 0$
- Statystyczna istotność efektu interakcji nie może być weryfikowana testem t dla parametru β_{12}
- Wartość efektu jest uzależniona od wartości pozostałych zmiennych niezależnych. Odróżnia to modele nieliniowe od modeli liniowych.
- Efekt interakcji może mieć różny znak dla różnych wartości zmiennych objaśniających.

Standardowe polecenie pakietu Stata licząc pochodną zwróci wartość parametru β_{12}. Buis (2010) pokazał sposób w jaki można zinterpretować efekty interakcji w modelach nieliniowych bez odwoływania się do dodatkowych pakietów.

Interpretację wyników wyjaśnimy na podstawie przykładu zaproponowanego przez Buisa (2010). W przykładzie wykorzystano dane z amerykańskiego badania *National Longitudinal Survey (NLSW) for employed women.* Analizujemy w jaki sposób wpływa ukończenie koledżu (collgrad) na szansę uzyskania "dobrej" pracy (high occ), tzn. wymagającej wysokich kwalifikacji, przez kobietę rasy czarnej i białej.

```
logit high_occ black##collgrad baseline, or noconstant nolog

Logistic regression                                     Number of obs   =       2211
                                                        Wald chi2(4)    =     504.62
Log likelihood = -1199.4399                             Prob > chi2     =     0.0000

------------------------------------------------------------------------------
      high_occ | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
       1.black |   .4194072   .0655069    -5.56   0.000     .3088072    .5696188
    1.collgrad |   2.465411    .293568     7.58   0.000     1.952238    3.113478
               |
black#collgrad |
          1 1  |   1.479715   .4132536     1.40   0.161     .8559637    2.558003
               |
      baseline |   .3220524   .0215596   -16.93   0.000     .2824512    .3672059
------------------------------------------------------------------------------
```

W standardowym podejściu zostałyby oszacowane parametry modelu, a następnie obliczone wartości efektów krańcowych dla zmiennych objaśniających tłumaczące prawdopodobieństwo uzyskania dobrej pracy. W tym przykładzie wykorzystano polecenie `logit` z opcją `or` zatem oszacowane wartości współczynników są szansami. W tym przypadku można je interpretować jako oczekiwana liczba osób zatrudnionych w dobrym zawodzie przypadająca na jedną osobę zatrudnioną w złym zawodzie.

Szansa dla zmiennej `baseline`, czyli kategorii odniesienia została uzyskana dzięki zastosowaniu sztuczki. W modelu pominięto stałą dzięki wykorzystaniu opcji `noconst` a w zamian do zbioru regresorów dołączono zmienną `baseline` równą 1 dla każdej obserwacji. Szansę dla zmiennej `baseline` intrepretować należy jako szansę, że biała kobieta która ukończyła koledż posiada pracę wymagającą wysokich kwalifikacji. Wynosi ona 0,32 oznaczając, że należy spodziewać się że na każdą pracującą białą kobietę, która ukończyła koledź i pracuje w zawodzie wymagającym niskich kwalifikacji należy spodziewać się 0,32 kobiety pracującej w zawodzie wymagającym wysokich kwalifikacji.

Iloraz szans dla zmiennej `collgrad` wynosi 2,47 oznaczając, że szansa posiadania zajęcia wymagającego wysokich kwalifikacji jest 2,47 razy wyższa dla kobiet które skończyły koledź. W modelu jest również zawarta zmienna obrazująca interakcję między zmienną `collgrad` a wskaźnikiem rasy czarnej `black`, zatem oszacowanie efektu ukończenia koledżu odnosi się do kobiet rasy białej.

Efekt ukończenia koledżu dla czarnych kobiet wynosi 1,48 efektu dla białych kobiet. Zatem parametr przy interakcji pokazuje w jaki sposób efekt ukończenia koledżu jest zróżnicowany między białymi a czarnymi kobietami.

Wyniki pokazują, iż to ostatnie oszacowanie nie jest statystycznie istotne.

Polecenie `margins` pokazuje szansę otrzymania dobrej pracy dla każdej kombinacji zmiennych `black` i `colgrad`. Szansa uzyskania dobrej pracy przez białą kobietę bez ukończonego koledżu wynosi 0,32, a dla kobiety z ukończonym koledźem 0,)79. Zatem efekt krańcowy ukończenia koledżu dla białej kobiety wynosi 0,47. Analogiczny efekt dla kobiety czarnej wynosi jedynie 0,36.

Zatem efekt krańcowy zmiennej (collgrad, oznaczające ukończenie koledżu, jest wyższy dla białych kobiet niż czarnych kobiet, podczas gdy efekt mnożnikowy (efekt interakcji) zmiennej (collgrad jest wyższy dla czarnych kobiet niż dla białych kobiet.

```
. margins, over(black collgrad) expression(exp(xb())) post

Predictive margins                                 Number of obs=       2211
Model VCE    : OIM

Expression   : exp(xb()) over         : black collgrad

------------------------------------------------------------------------------
               |            Delta-method
               |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
black#collgrad |
          0 0  |   .3220524   .0215596    14.94   0.000     .2797964    .3643084
          0 1  |   .7939914    .078188    10.15   0.000     .6407457    .9472371
          1 0  |   .1350711   .0190606     7.09   0.000      .097713    .1724292
          1 1  |   .4927536   .1032487     4.77   0.000       .29039    .6951173
------------------------------------------------------------------------------
```

Literatura

[1] Ai, C i Norton E. (2003) *Interaction terms in logit and probit models*, Economic Letters, vol. 80, pp. 123-129.

[2] Buis, M. (2010) *Stata tip 87: Interpretation of interactions in nonlinear models*, Stata Journal, vol 10, Number 2, pp. 305-310.

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

[4] 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.

[5] 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.

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