B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 2 2005

Barbara BUTRYN*
Marek FURA**

WYZNACZANIE PRAWDOPODOBIEŃSTWA PODJĘCIA DECYZJI Z UŻYCIEM MODELU *PROBITOWEGO* I *LOGITOWEGO*

Celem artykułu jest prezentacja modeli zmiennych dychotomicznych: *logitowego* i *probitowego* oraz zwrócenie uwagi na ich szerokie zastosowanie w różnych dziedzinach nauk. W artykule wykorzystano model regresji *probitowej* do wyznaczenia prawdopodobieństwa przyjęcia kandydata na Wydział Ekonomii, specjalność Handel i spółdzielczość, Uniwersytetu Rzeszowskiego.

Słowa kluczowe: *model logitowy, model probitowy, metoda największej wiarygodności*

Wiele zjawisk ekonomicznych i społecznych ma charakter jakościowy. Oznacza to, że zmienne opisujące dane zjawisko, zarówno zależne, jak i niezależne, przyjmują skończoną liczbę wartości. Z tego typu zjawiskami mamy z reguły do czynienia, gdy dane dotyczą pewnych jednostek ekonomicznych, np. gospodarstw domowych, gospodarstw rolniczych, pojedynczych konsumentów, indywidualnych przedsiębiorstw, przy czym każda z tych jednostek dokonuje wyboru spośród różnych możliwości. Przykładowo, dane gospodarstwo rolnicze może dokonać zakupu nowego ciągnika lub nie, osoba pozostająca bez pracy może ją znaleźć lub nie, pracownik może udać się do pracy samochodem, tramwajem czy pieszo. Wybór każdej z dostępnych możliwości jest zależny od różnorodnych czynników, pełniących rolę zmiennych objaśniających. Rozważając na przykład możliwość zakupu mieszkania, takimi czynnikami będą niewątpliwie dochód kupującego czy cena mieszkania.

Modele pozwalające określić prawdopodobieństwo podjęcia przez jednostkę ekonomiczną określonej decyzji to modele *probitowe* i *logitowe*.

W praktyce najczęściej podejmujemy decyzję o realizacji albo o odstąpieniu od realizacji jakiegoś przedsięwzięcia. Decyzje te oznaczmy odpowiednio przez 1 oraz 0. Niech

* Zakład Metod Ilościowych, Uniwersytet Rzeszowski, ul. Ćwiklińskiej 2, 35-959 Rzeszów, basiabutryn@o2.pl

** Wyższa Szkoła Inżynieryjno-Ekonomiczna, ul. Mickiewicza 10, 39-100 Ropczyce, marekfura@o2.pl

P_{1i} będzie prawdopodobieństwem, że *i*-ta jednostka ekonomiczna podejmie decyzję 1, P_{0i} prawdopodobieństwem, że *i*-ta jednostka ekonomiczna podejmie decyzję 0. Przykładowo niech badaną jednostką ekonomiczną będzie rodzina, mająca podjąć decyzję dotyczącą kupna samochodu. Oznaczmy przez z_i wektor zmiennych, opisujących preferencje *i*-tej rodziny względem samochodu, mających wpływ na decyzję dotyczącą zakupu. Niech

$$y_i = \beta^T z_i ,$$

gdzie β jest wektorem nieznanych parametrów. Przyjmijmy, że prawdopodobieństwo podjęcia decyzji jest uzależnione od y_i, tzn. $P_{1i} = P_1(y_i)$, $(P_{0i} = P_0(y_i))$.

Niech Φ będzie dystrybuantą standaryzowanego rozkładu normalnego. W modelu *probitowym* zakłada się, że P_{1i} jest wartością dystrybuanty Φ standaryzowanego rozkładu normalnego $N(0,1)$ dla y_i, tzn.

$$P_{1i} = \Phi(y_i) = \int_{-\infty}^{y_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds = \int_{-\infty}^{\beta^t z_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.$$

Informacje o *n* jednostkach ekonomicznych uzyskujemy na podstawie *n*-elementowej próby. Jej elementy porządkujemy w ten sposób, że przyjmujemy pierwszych m $(0 \le m \le n)$ jednostek, które podjęły decyzję 1, a pozostałych $n - m$, które podjęły decyzję 0. Funkcja wiarygodności dla tej próby zależy od parametru β i ma postać

$$L = \prod_{i=1}^{m} P_1(y_i) \prod_{i=m+1}^{n} P_0(y_i) = \prod_{i=1}^{m} \Phi(y_i) \prod_{i=m+1}^{n} [1 - \Phi(y_i)],$$

$$L = \prod_{i=1}^{m} \Phi(\beta^T z_i) \prod_{i=m+1}^{n} [1 - \Phi(\beta^T z_i)].$$

Ponieważ funkcja największej wiarygodności ma postać iloczynową, więc w celu znalezienia jej maksimum wygodnie jest ją zlogarytmować. Wiadomo, że maksimum funkcji wiarygodności oraz maksimum jej logarytmu znajdują się w tym samym punkcie. Wobec tego:

$$\ln L = \ln \prod_{i=1}^{m} \Phi(\beta^T z_i) + \ln \prod_{i=m+1}^{n} [1 - \Phi(\beta^T z_i)],$$

$$\ln L = \sum_{i=1}^{m} \ln \Phi(\beta^T z_i) + \sum_{i=m+1}^{n} \ln[1 - \Phi(\beta^T z_i)].$$

Różniczkując funkcję $\ln L$ względem β, dostajemy

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{m} \frac{1}{\Phi(\beta^T z_i)} \varphi(\beta^T z_i) z_i - \sum_{i=m+1}^{n} \frac{1}{1 - \Phi(\beta^T z_i)} \varphi(\beta^T z_i) z_i ,$$

gdzie φ oznacza gęstość standaryzowanego rozkładu normalnego. Przyrównując gradient $\frac{\partial \ln L}{\partial \beta}$ do zera, otrzymujemy układ równań, z którego za pomocą metod numerycznych wyliczamy wartości wektora parametrów β. Wartości P_{1i} odczytujemy z tablic rozkładu normalnego.

Niech Φ oznacza dystrybuantę rozkładu logistycznego. W modelu *logitowym* zakłada się, że P_{1i} jest wartością dystrybuanty $\overline{\Phi}$ rozkładu logistycznego dla y_i, tzn.

$$\mathrm{P}_{1i} = \overline{\Phi}(y_i) = \frac{1}{e^{-y_i}+1} = \frac{e^{y_i}}{e^{y_i}+1} = \frac{e^{\beta^T z_i}}{e^{\beta^T z_i}+1}.$$

Po przekształceniu otrzymujemy zależności:

$$e^{\beta^T z_i} = \frac{\mathrm{P}_{1i}}{1-\mathrm{P}_{1i}},$$
$$\beta^T z_i = \ln \frac{\mathrm{P}_{1i}}{1-\mathrm{P}_{1i}},$$
$$1 - P_{1i} = \frac{1}{e^{\beta^T z_i}+1}.$$

W celu określenia funkcji wiarygodności wprowadzamy zmienną

$$f_{1i} = \begin{cases} 1, & \text{gdy } i\text{-ta jednostka podejmuje decyzję } 1, \\ 0, & \text{gdy } i\text{-ta jednostka podejmuje decyzję } 0. \end{cases}$$

Funkcja wiarygodności n-elementowej próby wyraża się wzorem

$$L = \prod_{i=1}^{n} \mathrm{P}_{1i}^{f_{1i}} (1-\mathrm{P}_{1i})^{1-f_{1i}}.$$

Logarytmując otrzymujemy

$$\ln L = \sum_{i=1}^{n} \ln[\mathrm{P}_{1i}^{f_{1i}} (1-\mathrm{P}_{1i})^{1-f_{1i}}] = \sum_{i=1}^{n} [f_{1i} \ln \mathrm{P}_{1i} + (1-f_{1i}) \ln(1-\mathrm{P}_{1i})] =$$
$$= \sum_{i=1}^{n} \{f_{1i}[\ln \mathrm{P}_{1i} - \ln(1-\mathrm{P}_{1i})] + \ln(1-\mathrm{P}_{1i})\} = \sum_{i=1}^{n} \{f_{1i} \ln \frac{\mathrm{P}_{1i}}{1-\mathrm{P}_{1i}} + \ln(1-\mathrm{P}_{1i})\} =$$
$$= \sum_{i=1}^{n} \{f_{1i} \beta^T z_i - \ln(e^{\beta^T z_i}+1)\} = \sum_{i=1}^{n} f_{1i} \beta^T z_i - \sum_{i=1}^{n} \ln(e^{\beta^T z_i}+1).$$

Następnie maksymalizujemy logarytm funkcji wiarygodności, stosując jedną z numerycznych metod maksymalizacji. Procedura ta prowadzi do uzyskania ocen wektora parametrów β. Po ich uzyskaniu wyznaczamy wartość y_i, a następnie wartość dystrybuanty rozkładu logistycznego dla wyznaczonego y_i.

Za pomocą analizy *probitowej* chcemy wyznaczyć prawdopodobieństwo przyjęcia kandydata na studia wyższe, na kierunek Ekonomia, specjalność Handel i spółdzielczość Uniwersytetu Rzeszowskiego. Posłużą nam do tego dane pochodzące z przeprowadzonej rekrutacji w czerwcu 2004 r.

Przyjęcie kandydata na studia odbywało się na podstawie konkursu świadectw. Zaliczane były oceny ze świadectwa dojrzałości z przedmiotów: matematyka, geografia (w przypadku jej braku – historia), język obcy. O przyjęcie na studia ubiegało się 826 kandydatów, z czego zostało przyjętych 112 osób.

Zmienna zależna w modelu (decyzja) jest dychotomiczna, czyli przyjmuje dwie wartości: 1 – gdy kandydat został przyjęty na studia i 0 – w przeciwnym razie. Zmienne niezależne w modelu to: ocena z języka obcego, ocena z matematyki i ocena z geografii (historii) na świadectwie dojrzałości. Są to zmienne jakościowe mogące przyjmować wartości: 6, 5, 4, 3, 2. Ponieważ o przyjęciu na studia decyduje suma punktów uzyskana z trzech przedmiotów, zmienne objaśniające zostały więc zastąpione jedną zmienną niezależną: *suma punktów*.

Prawdopodobieństwo przyjęcia *i*-kandydata na studia zależy od wartości:

$$y_i = [\beta_0 \;\; \beta_1]^T \cdot [1 \;\; z_i], \quad i = 1, \ldots, 826,$$

gdzie:

$\beta = [\beta_0 \;\; \beta_1]$ – wektor nieznanych parametrów,

z_i – wartość zmiennej niezależnej dla *i*-tego kandydata

i wynosi

$$P_{1i} = \Phi(y_i) = \int_{-\infty}^{y_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds = \int_{-\infty}^{\beta^t z_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.$$

W modelu *probitowym* wartości ocen parametru β, uzyskane metodą największej wiarygodności polegającą na wyznaczeniu z próby takich ocen parametrów modelu, które maksymalizują wiarygodność próby statystycznej przedstawiono w tabeli 1.

Wartość statystyki dobroci dopasowania χ^2 analizowanego modelu wskazuje na istotność różnicy między aktualnym modelem, a modelem tylko z wyrazem wolnym. Możemy stwierdzić, że zmienna *suma punktów* istotnie wpływa na decyzję o przyjęciu. Na podstawie testu *t*-Studenta stwierdzamy, że parametr dla zmiennej *suma punktów* i wyraz wolny są statystycznie istotne.

Tabela 1

Wyniki estymacji

n = 826	Model: regresja probit; liczba 0:714 1: 112 (Butryn) Zmienna zależna: decyzja χ^2 (1) = 583,56 p = 0,0000	
	Stała	Suma punktów
Ocena	–31,6440	2,173110
Błąd standardowy	3,2039	0,221648
t (824)	–9,8767	9,804322
poziom p	0,0000	0,000000

Źródło: opracowanie własne za pomocą pakietu *Statistica*.

Prawdopodobieństwo *sukcesu* dla *i*-tego kandydata w modelu *probitowym* ma postać

$$P_{1i} = \Phi(-31{,}6440 + 2{,}17311 \cdot z_i),$$

gdzie Φ – dystrybuanta standaryzowanego rozkładu normalnego.

Obliczmy za pomocą oszacowanego modelu prawdopodobieństwo przyjęcia na studia kandydata, który uzyskał w konkursie świadectw sumę punków 14:

$$P_{1i} = \Phi(-31{,}644 + 2{,}17311 \cdot 14) = \Phi(-1{,}22046) = 0{,}111145.$$

Wnioskujemy więc, iż rozważany kandydat ma bardzo małe szanse przyjęcia na wybrany kierunek studiów. Znajomość wyznaczonego prawdopodobieństwa byłaby niezmiernie ważną informacją dla owego kandydata.

Calculation of decision making probability using probit and logit models

The aim of this article is presentation of *logit* and *probit* models and their wide application in many different science. *Logit* and *probit* regression are used for analyzing the relationship between one or more independent variables with categorical dependent variable. There are a lot of advantages of *logit* (*probit*) models over linear multiple regression. These methods imply that the dependent variable is actually the result of a transformation of an underlying variable, which is not restricted in range. For example, the *probit* model assumes that the actual underlying depedent variable is measured in terms of values for normal curve; if one transforms those values for probabilities then the predictions for the dependent variable will always fall between 0 ond 1. Thus, we are actually predicting probabilities from the independent variables The *probit* model was used to calculate the probability of admissions in Rzeszów Uniwersity, speciality Handel i spółdzielczość.

Keywords: *logit model, probit model, maximum likelihood*