Jacek Osiewalski, Jerzy Marzec

Uogólnienie dychotomicznego modelu probitowego z wykorzystaniem skośnego rozkładu Studenta*

1. Wprowadzenie

Najprostszym przypadkiem modelu dla jakościowej zmiennej endogenicznej jest model dwumianowy (dychotomiczny). Opisuje on zależność między prawdopodobieństwem wyboru jednej z dwóch możliwości (oznaczanych umownie jako 0 i 1) a egzogenicznymi zmiennymi objaśniającymi, które opisują cechy możliwych alternatyw lub indywidualne charakterystyki podmiotów podejmujących decyzję. Postać tego modelu jest następująca:

$$p_t \equiv \Pr(y_t = 1) = G(x_t \cdot \beta) = 1 - F(-x_t \cdot \beta) \quad \text{dla } t=1,\ldots,T, \tag{1}$$

gdzie β jest wektorem $k\times 1$ nieznanych parametrów ($\beta \in \mathrm{R}^k$), $x_t=(x_{t1}\ldots x_{tk})$ oznacza wektor ustalonych wartości k zmiennych egzogenicznych (lub ich znanych funkcji), zaś $G(\cdot)$ i $F(\cdot)$ są znanymi funkcjami wiążącymi p_t, czyli prawdopodobieństwo zaobserwowania sukcesu, z x_t i β. Funkcja $F(\cdot)$ ma wszystkie własności dystrybuanty rozkładu prawdopodobieństwa i określa klasę modelu. Równoważną specyfikację otrzymujemy przez wprowadzenie modelu regresji liniowej (ze względu na β) dla ukrytych (nieobserwowalnych) zmiennych ciągłych $z_1,\ldots,z_T$, których znaki determinują zaobserwowane wartości zmiennych y_t (0 lub 1):

$$\begin{aligned} z_t &= x_t \cdot \beta + \varepsilon_t, \\ y_t &= I_{[0,\infty)}(z_t) = \begin{cases} 1, & \text{gdy} \quad z_t \geq 0, \\ 0, & \text{gdy} \quad z_t < 0, \end{cases} \end{aligned} \tag{2}$$

czyli $I_A(.)$ jest funkcją charakterystyczną zbioru A. O składnikach losowych ε_t zakłada się zwykle, że są niezależne i posiadają ten sam rozkład o wartości oczekiwanej równej zero i jednostkowej wariancji. Jeśli rozkład jest symetryczny, to zapis (1) upraszcza się do bardziej znanego $p_t \equiv \Pr(y_t = 1) = F(x_t \cdot \beta)$. Szczegóły dotyczące niebayesowskiej estymacji modeli dla danych jakościowych oraz wiele ich zastosowań empirycznych z zakresu ekonomii prezentują m.in. Amemiya (1981, 1985), Maddala (1983) lub Greene (1993).

Najbardziej znanymi i powszechnie stosowanymi modelami dwumianowymi są modele probitowy i logitowy, które odpowiadają przyjęciu dla ε_t odpowiednio rozkładu normalnego lub

* Praca przygotowana w ramach badań statutowych Akademii Ekonomicznej w Krakowie w roku 2004.

logistycznego. Innymi, rzadziej stosowanymi, są np. modele krzywej Gompertza, Urbana i rozkładu Burra. Do estymacji tych modeli wykorzystywana jest zwykle metoda największej wiarygodności (MNW). W modelu probitowym i logitowym pokazano, że estymator MNW dla β jest jednoznacznie określony; wyprowadzono również postać asymptotycznej macierzy kowariancji estymatora MNW. Jednym z kierunków uogólnienia modelu probitowego jest przyjęcie dla ε_t rozkładu *t* Studenta o nieznanej liczbie stopni swobody $\nu>0$, co dopuszcza brak wariancji ($\nu \leq 2$) a nawet wartości oczekiwanej zmiennej ε_t ($\nu \leq 1$) Klasa rozkładów *t* Studenta zawiera rozkład normalny jako przypadek graniczny ($\nu = +\infty$), zaś – jak podają Albert i Chib (1993) – rozkład logistyczny może być przybliżany przez rozkład *t* Studenta o ok. 7 – 9 stopniach swobody. A zatem uogólnienie to pozwala testować (choćby w przybliżeniu) empiryczną adekwatność dwóch podstawowych modeli dwumianowych. Jednak zastosowanie MNW w tym przypadku jest niewskazane, ponieważ nie są znane własności estymatora MNW nawet w przypadku klasycznego modelu regresji liniowej ze składnikiem losowym o rozkładzie *t* Studenta z nieznanym parametrem ν.

Albert i Chib (1993) zaproponowali specyfikację i estymację bayesowskiego modelu dychotomicznego z rozkładem *t* Studenta. W celu numerycznej aproksymacji brzegowych rozkładów a posteriori interesujących wielkości wykorzystali algorytm Gibbsa, metodę Monte Carlo typu łańcuchów Markowa (ang. *Markov Chain Monte Carlo*, MCMC). Marzec (2003a,b,c) wykorzystał to podejście dla zbadania ryzyka pojedynczych umów kredytowych klientów detalicznych dużego banku komercyjnego; wyniki empiryczne wskazywały na zasadność zastosowania tego uogólnienia modelu probitowego, gdyż rozkład a posteriori parametru ν skupiony był w przedziale (1, 3) – świadcząc o relatywnie małej adekwatności modelu probitowego czy logistycznego.

Wszystkie trzy rozważane rozkłady prawdopodobieństwa (normalny, logistyczny, *t* Studenta) charakteryzują się symetrią, różnią się natomiast grubością ogonów (szybkością zbieżności dystrybuanty do wartości granicznych 0 i 1). Proponujemy więc w tej pracy dalsze uogólnienie modelu probitowego, które polega na przyjęciu dla ε_t klasy skośnych rozkładów *t* Studenta. Klasa ta ma dwa swobodne dodatnie parametry: stopnie swobody ν i współczynnik asymetrii γ, którego kwadrat jest równy ilorazowi mas prawdopodobieństwa na prawo i lewo od modalnej. Estymacja parametrów β, ν, γ i ich funkcji możliwa jest na gruncie bayesowskim przy wykorzystaniu metod Monte Carlo typu łańcuchów Markowa.

Asymetryczne rozkłady wielowymiarowe (w tym typu *t* Studenta) rozważali Fernández, Osiewalski i Steel (1995), natomiast szczegółową definicję i formalne własności skośnego rozkładu *t* w przypadku jednowymiarowym podali Fernández i Steel (1998). Autorzy ci zastosowali ten rozkład dla składnika losowego w modelu klasycznej regresji liniowej. Z kolei Osiewalski i Pipień

(1999, 2000) wykorzystali ten rozkład w modelach GARCH dla finansowych szeregów czasowych, wykazując jego użyteczność w badaniach empirycznych. W niniejszej pracy pokażemy, jak zastosować dystrybuantę skośnego rozkładu *t* w modelu dychotomicznym (część 2), jak przeprowadzić jego analizę bayesowską (część 3) oraz jak te propozycje metodologiczne wykorzystać w badaniu spłacalności kredytów (część 4). Część 5 zawiera uwagi końcowe.

2. Dystrybuanta skośnego rozkładu *t* w konstrukcji modelu dwumianowego

Przyjmijmy, że składnik losowy ε_t w równaniu (2) ma skośny rozkładu *t* Studenta o modalnej równej 0, jednostkowej precyzji, ν stopniach swobody (ν>0) i parametrze asymetrii γ > 0; jego funkcja gęstości ma postać:

$$p(\varepsilon_t|\theta) = f_{skS}(\varepsilon_t|\nu,\gamma) = \frac{2}{\gamma+\gamma^{-1}}\left\{f_\nu(\gamma\varepsilon_t)\cdot I_{(-\infty,0)}(\varepsilon_t) + f_\nu(\varepsilon_t\gamma^{-1})\cdot I_{[0,+\infty)}(\varepsilon_t)\right\}, \quad (3)$$

gdzie $\theta = (\beta' \nu\ \gamma)'$, zaś $f_\nu(\varepsilon_t)$ jest funkcją gęstości symetrycznego rozkładu *t* Studenta o zerowej modalnej, precyzji równej jeden i ν stopniach swobody. Stopień asymetrii określony jest przez kwadrat parametru γ:

$$\frac{\Pr(\varepsilon_t \geq 0|\gamma)}{\Pr(\varepsilon_t < 0|\gamma)} = \gamma^2. \quad (4)$$

Wielkość γ^2 informuje o ilorazie mas prawdopodobieństwa skupionych na prawo i na lewo od modalnej. Jeżeli parametr asymetrii γrówny jest jedności, to rozkład jest symetryczny.

Ze specyfikacji (2) wynika, że prawdopodobieństwo zaobserwowania y_t=1 wynosi

$$\Pr(y_t = 1|\theta) = \Pr(z_t \geq 0|\theta) = \Pr(\varepsilon_t \geq -x_t\beta|\theta) = 1 - \Pr(\varepsilon_t < -x_t\beta|\theta) = 1 - F_{skS}(-x_t\beta|\nu,\gamma), \quad (5)$$

gdzie dystrybuanta skośnego rozkładu *t* Studenta o modalnej 0, precyzji 1, ν stopniach swobody i parametrze asymetrii γ(obliczona w punkcie *a*) wyraża się formułą

$$F_{skS}(a|\nu,\gamma) = \frac{2}{\gamma+\gamma^{-1}}\left[\gamma^{-1}F_\nu(a\gamma)I_{(-\infty,0)}(a) + \left(\frac{\gamma^{-1}-\gamma}{2} + \gamma F_\nu(a\gamma^{-1})\right)I_{[0,+\infty)}(a)\right] \quad (6)$$

przy czym $F_\nu(a)$ jest dystrybuantą symetrycznego rozkładu *t* Studenta o modalnej 0, precyzji 1 i ν stopniach swobody. Rysunek 1 przedstawia przebieg dystrybuant skośnych rozkładów *t* Studenta w zależności od różnych wartości parametrów ν i γ (na tle dystrybuanty rozkładu normalnego, odpowiadającej $\nu = +\infty$ i γ=1). Wartość dystrybuanty w zerze jest funkcją parametru γ i wynosi 0.5 tylko dla γ=1.

Rysunek 1

Wzór (5) określa rozkład pojedynczej obserwacji (przy danych parametrach) jako rozkład dwupunktowy o funkcji prawdopodobieństwa:

$$p(y_t|\theta) = F_{skS}(-x_t\beta|\nu,\gamma)I_{\{0\}}(y_t) + [1 - F_{skS}(-x_t\beta|\nu,\gamma)]I_{\{1\}}(y_t).$$

W przypadku *T* niezależnych obserwacji ich łączne prawdopodobieństwo można zapisać jako:

$$p(y|\theta) = p(y_1,\ldots,y_T|\theta) = \prod_{t=1}^{T} p(y_t|\theta) = \left[\prod_{t:y_t=0} F_{skS}(-x_t\beta|\nu,\gamma)\right]\cdot\left[\prod_{t:y_t=1}(1 - F_{skS}(-x_t\beta|\nu,\gamma))\right].$$

Przy ustalonych obserwacjach, powyższa formuła określa funkcję wiarygodności dla modelu dychotomicznego proponowanego w tej pracy. Funkcja ta ma ważną własność – traktowana jako funkcja argumentu ν (przy pozostałych ustalonych) bardzo szybko zmierza do dodatniej stałej równej wartości wiarygodności przy (skośnym) rozkładzie normalnym ($\nu = +\infty$). Ta praktyczna stałość wiarygodności dla dużych ν może być poważną przeszkodą w klasycznej estymacji parametrów modelu dwumianowego. Oczywiście, tę samą własność ma już funkcja wiarygodności w modelach z symetrycznym rozkładem Studenta. Autorzy nie znają żadnej pracy określającej własności estymatora MNW w takich przypadkach. Własne, wstępne badania symulacyjne ukazują jego znaczne, systematyczne obciążenie.

3. Elementy analizy bayesowskiej

Podstawowym elementem analizy bayesowskiej jest statystyczny model bayesowski, czyli łączny rozkład obserwacji i parametrów, określony przez dyskretny warunkowy rozkład wektora obserwacji *y*, $p(y|\theta)$, i ciągły brzegowy rozkład wektora parametrów (tzw. rozkład a priori). Często wykorzystuje się uogólniony model bayesowski, w którym rozkład a priori jest miarą σ–skończoną, ale nie jest miarą probabilistyczną. Podobnie w tej pracy – zakładamy, że brzegowy rozkład a priori wektora β jest niewłaściwy jednostajny na całej przestrzeni R^k. Dla ν i γ przyjmujemy rozkłady właściwe: dla ν wykładniczy o wartości oczekiwanej *r* (*r*=10), zaś dla γ standardowy rozkład logarytmiczno-normalny.

Z uwagi na to, że zbiory dopuszczalnych wartości parametrów ν i γ są równe R_+, warto dokonać reparametryzacji $\theta_{k+1} = \ln(\nu/r)$, $\theta_{k+2} = \ln(\gamma)$ i redefiniować wektor wszystkich parametrów jako $\theta = [\beta' \ \theta_{k+1} \ \theta_{k+2}]'$. Wówczas przestrzeń parametrów jest całym zbiorem R^{k+2}, co bardzo upraszcza stronę numeryczną analizy bayesowskiej. Dla tak określonego θ mamy następującą strukturę a priori:

$$p(\theta) = p(\beta)\cdot p(\theta_{k+1})\cdot p(\theta_{k+2}), \quad \text{gdzie}$$
$$p(\beta) = c \quad \text{dla } \beta \in R^k, \quad p(\theta_{k+1}) = \exp(\theta_{k+1})\exp(-\exp(\theta_{k+1})), \quad p(\theta_{k+2}) = f_N(\theta_{k+2}\,|\,0,1). \quad (7)$$

W szczególności wykładniczy rozkład a priori dla ν (o wartości oczekiwanej r) prowadzi do rozkładu wartości ekstremalnych (rozkładu Gumbela) dla $\theta_{k+1} = \ln(\nu / r)$, zaś informacja o parametrze skośności jest reprezentowana przez standaryzowany rozkład normalny dla $\ln(\gamma)$. Określona powyżej struktura a priori reprezentuje bardzo słabą wstępną wiedzę obserwatora o parametrach. Prowadzi ona, wraz z modelem próbkowym zdefiniowanym w poprzedniej części, do pełnego modelu bayesowskiego określonego przez uogólnioną gęstość postaci

$$p(y,\theta) = p(y|\theta)p(\theta) = p(\theta)\left[\prod_{t:y_t=0} F_{skS}\left(-x_t\beta|\nu,\gamma\right)\right]\cdot\left[\prod_{t:y_t=1}\left(1-F_{skS}\left(-x_t\beta|\nu,\gamma\right)\right)\right]. \qquad (8)$$

Wnioskowanie bayesowskie wykorzystuje faktoryzację tego modelu na rozkład a posteriori, tj. rozkład ciągły o funkcji gęstości:

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)} \propto p(\theta)\left[\prod_{t:y_t=0} F_{skS}\left(-x_t\beta|\nu,\gamma\right)\right]\cdot\left[\prod_{t:y_t=1}\left(1-F_{skS}\left(-x_t\beta|\nu,\gamma\right)\right)\right],$$

oraz brzegowy rozkład obserwacji (dyskretny):

$$p(y) = \int_{\Theta} p(y|\theta)\,p(\theta)\,d\theta\ .$$

Aby faktoryzacja ta była możliwa (i tym samym istniał rozkład a posteriori), powyższa całka musi być skończona. Dlatego przyjęto właściwy rozkład a priori parametru stopni swobody. Zbieżność funkcji wiarygodności przy $\nu \to +\infty$ do dodatniej stałej oznacza bowiem, że całka tej funkcji (po całej przestrzeni parametrów) jest nieskończona, więc niewłaściwy jednostajny rozkład a priori dla ν prowadziłby do braku rozkładu a posteriori.

Rozkład a posteriori parametrów proponowanego (jak i każdego innego) modelu dwumianowego jest niestandardowym rozkładem wielowymiarowym. Ponadto, przedmiotem wnioskowania są nie tylko oryginalne parametry, ale przede wszystkim ich skomplikowane funkcje nieliniowe takie jak:

- prawdopodobieństwo $p_* \equiv \Pr(y_* = 1) = 1 - F(-x_* \cdot \beta)$ określonego zachowania (wyboru) w przypadku nie obserwowanego obiektu o ustalonych charakterystykach,
- efekt krańcowy wzrostu *h*-tej charakterystyki o małą jednostkę; w przypadku ciągłej zmiennej objaśniającej (nie powiązanej z pozostałymi) jest to $\partial p_*/\partial x_{*h} = \beta_h f(-x_*\beta)$, gdzie f jest gęstością odpowiadającą dystrybuancie F, definiującej model dychotomiczny – u nas jest to gęstość (3); wielkość $\beta_h f(-x_*\beta)$ szacujemy również dla dyskretnych zmiennych objaśniających, choć traci ona wtedy swą pierwotną interpretację.

Uzyskanie brzegowej funkcji gęstości a posteriori dla wielkości będącej przedmiotem analizy jest złożonym problemem całkowania w przestrzeni (k+2)-wymiarowej. Proponujemy w tym celu

zastosowanie metod Monte Carlo typu łańcuchów Markowa (MCMC), a w szczególności losowania Metropolisa i Hastingsa.

Metody MCMC polegają na tym, że ciąg kolejnych losowań w przestrzeni parametrów $(\theta^{(1)},...,\theta^{(n)},...)$ tworzy łańcuch Markowa (o nieprzeliczalnej liczbie stanów) z rozkładem stacjonarnym równym rozkładowi a posteriori $p(\theta|\, y)$. W efekcie, po osiągnięciu zbieżności łańcucha do rozkładu stacjonarnego, generujemy realizacje (otrzymujemy próbę) z rozkładu a posteriori; zob. np. O'Hagan (1994), Gamerman (1997). Algorytm Metropolisa i Hastingsa buduje łańcuch Markowa poprzez zadanie $\theta^{(0)}$ – arbitralnego punktu startowego oraz gęstości $q(\theta^*;\theta^{(m-1)})$ rozkładu losowań kandydackich θ^* (m=1,2,...); dla danego $\theta^{(m-1)}$ przyjmujemy

$\theta^{(m)} = \theta^*$ z prawdopodobieństwem $P(\theta^*,\theta^{(m-1)})$,

$\theta^{(m)} = \theta^{(m-1)}$ z prawdopodobieństwem $1 - P(\theta^*,\theta^{(m-1)})$,

przy czym prawdopodobieństwo akceptacji wylosowanego wstępnie θ^* dane jest wzorem

$$P(\theta^*,\theta^{(m-1)}) = \min\left\{\frac{f(\theta^*)q(\theta^{(m-1)};\theta^*)}{f(\theta^{(m-1)})q(\theta^*;\theta^{(m-1)})},1\right\},$$

gdzie $f(\theta)$ to jądro gęstości rozkładu a posteriori. Dogodny mechanizm losowań wstępnych wykorzystuje $q(\theta^*;\theta^{(m-1)}) = f_S(\theta^*\,|\,3,\theta^{(m-1)},3C^{-1})$, wielowymiarowy rozkład t Studenta o 3 stopniach swobody, modalnej równej poprzedniemu stanowi łańcucha oraz macierzy precyzji takiej, że C jest macierzą kowariancji (równą wstępnej ocenie macierzy kowariancji rozkładu a posteriori). W tym przypadku gęstość rozkładu losowań kandydackich $q(\theta^*;\theta^{(m-1)})$ jest symetryczna względem obu argumentów, więc prawdopodobieństwo akceptacji zależy tylko od ilorazu gęstości a posteriori:

$$P(\theta^*,\theta^{(m-1)}) = \min\left\{\frac{f(\theta^*)}{f(\theta^{(m-1)})},1\right\}.$$

W praktyce początkowe S stanów łańcucha Markowa służy uzyskaniu zbieżności (cykle spalone), a następne M stanów – generowaniu próby z rozkładu stacjonarnego i aproksymacji jego charakterystyk zgodnie z ogólnym wzorem:

$$E\left[g(\theta)|y\right] \approx \frac{1}{M}\sum_{q=S+1}^{S+M} g\left(\theta^{(q)}\right).$$

Wyniki badań empirycznych prezentowanych w następnej części uzyskano na podstawie S=10000 cykli spalonych i M=500000 realizacji tworzących próbę z rozkładu a posteriori. Kilka krótkich łańcuchów wstępnych pozwoliło wcześniej wykalibrować macierz C.

4. Badanie spłacalności kredytów

Zaprezentowany powyżej bayesowski modelu dwumianowy ze skośnym rozkładem *t* Studenta zastosowano do badania spłacalności kredytów detalicznych w oparciu o dane, które wykorzystał wcześniej Marzec (2003a,b,c). Przyjęto, iż zmienna objaśniana y_t przyjmuje dwie wartości, tzn. y_t=1, gdy kredytobiorca na dzień 30.09.2001 miał zaległości w spłacie rat kapitałowo-odsetkowych (opóźnienie w spłacie ostatniej raty wynosiło więcej niż miesiąc), natomiast y_t=0 w przeciwnym przypadku. Jako potencjalne zmienne wyjaśniające ryzyko pojedynczej umowy kredytowej wprowadziliśmy (jak we wcześniejszych pracach):

- płeć (zmienna przyjmuje wartość 1, jeżeli klientem jest mężczyzna, 0 w przypadku kobiety),
- wiek kredytobiorcy (w setkach lat),
- wpływy, tzn. wielkość kwartalnych wpływów w latach 2000-2001 (w setkach tys. zł) na rachunki typu ROR kredytobiorcy w badanym banku,
- posiadanie ROR w analizowanym banku (1 – posiada, 0 – nie posiada),
- informację o tym, czy kredytobiorca posiada karty płatnicze lub kredytowe wydane przez ten bank (1 – posiada choć jedną kartę płatniczą, 0 – nie posiada),
- sposób udzielenia kredytu (1 – poprzez pośrednika kredytowego, 0 – bezpośrednio przez rozważany bank),
- typ kredytu (1 – kredyt konsumpcyjny, 0 – kredyt hipoteczny),
- okres trwania umowy kredytowej (w dziesiątkach lat),
- podstawowe źródło dochodu uzyskiwanego przez kredytobiorcę (zmienne *zrdoch*), tj. umowa o pracę, albo renta lub emerytura, albo własna działalność, umowa o dzieło lub umowa zlecenie, albo inne źródło (np. stypendium).

Ostatnia zmienna może przyjmować cztery różne wartości. Chcąc ją uwzględnić w równaniu regresji z wyrazem wolnym, wprowadziliśmy trzy zmienne zerojedynkowe, a za punkt odniesienia przyjęliśmy umowę o pracę (*zrdoch1* = *zrdoch2* = *zrdoch3* = 0); w pozostałych przypadkach:

- *zrdoch1* = 1, gdy źródłem dochodu kredytobiorcy jest renta lub emerytura,
- *zrdoch2* = 1, gdy źródłem dochodu kredytobiorcy jest własna działalność, umowa o dzieło lub umowa zlecenie,
- *zrdoch3* = 1 w przypadku innego źródła dochodu, np. stypendium.

Poniżej przedstawiamy wyniki bayesowskiej estymacji modelu dwumianowego I – ze skośnym rozkładem *t* Studenta dla zmiennej ukrytej – na tle modelu II (z symetrycznym rozkładem *t* Studenta) oraz modelu III (probitowego), który jest najczęściej wykorzystywany w praktyce. Tabela 1 zawiera wartości oczekiwane i odchylenia standardowe a posteriori dla parametrów tych modeli. Wszystkie wyniki uzyskano za pomocą algorytmu Metropolisa i Hastingsa, przy czym

wyniki bayesowskie dla modelu III są prawie identyczne z wynikami uzyskanymi za pomocą MNW. Spodziewaliśmy się takiego rezultatu, ponieważ (zgodnie z teorią) w modelu probitowym oceny MNW w przypadku dużej liczby obserwacji można interpretować jako wartości oczekiwane rozkładu a posteriori (przy dość dowolnym ciągłym rozkładzie a priori). Alternatywnym do algorytmu Metropolisa i Hastingsa podejściem numerycznym w estymacji modelu II jest wykorzystanie próbnika Gibbsa, co zaproponowali Albert i Chib (1993). Łańcuch Markowa generowany przez algorytm Gibbsa wydaje się jednak wolno zbieżny do rozkładu a posteriori w modelu II, prawdopodobnie na skutek bardzo silnej autokorelacji. Wyniki uzyskane przez losowanie Gibbsa w modelu II prezentuje (dla tego samego zbioru danych) Marzec (2003c).

Tabela 1

Wyniki uzyskane w modelu I świadczą o dużej asymetrii rozkładu ε_t; $\gamma^2 \approx 0.117 \approx 1/9$; niespełna 12% masy prawdopodobieństwa rozkładu próbkowego zmiennej ε_t znajduje się na prawo od zera. Małe odchylenie standardowe a posteriori dla γ wskazuje, że korzystając z prostego bayesowskiego testu na redukcję modelu (testu Lindleya) musimy dojść do wniosku o pełnej zasadności uogólnienia modelu II (symetrycznego *t* Studenta) do modelu I poprzez wprowadzenie asymetrii. Wartość oczekiwana a posteriori dla stopni swobody ν wynosi około 0.5 w modelu I i 1.4 w modelu II, przy czym małe odchylenia standardowe wskazują na precyzyjny szacunek parametru ν. Dystrybuanty odpowiadające modelom I i II przedstawiono na omawianym już wcześniej Rysunku 1; przyjęte wartości swobodnych parametrów to ich wartości oczekiwane a posteriori. Uzyskane wyniki oznaczają, iż zakładanie normalności składnika losowego i stosowanie modelu probitowego jest nieuzasadnione w przypadku naszych danych. Również przyjęcie symetrycznego rozkładu *t* Studenta nie jest w pełni adekwatne, gdyż ważne są nie tylko ogony rozkładu zmiennej ukrytej, ale również możliwość jego asymetrii.

Wartości parametrów β_i nie są bezpośrednio interpretowalne, zaś informacje o sile i kierunku wpływu zmiennych egzogenicznych na p_t (prawdopodobieństwo niespłacalności) uzyskujemy na podstawie efektów krańcowych, których wartości przedstawia Tabela 2.

Tabela 2

Spośród rozważanych jakościowych zmiennych egzogenicznych największy wpływ na prawdopodobieństwo niepłacenia kredytu ma sposób udzielenia tego kredytu; udzielenie go przez pośrednika, zamiast bezpośrednio przez bank, powoduje najbardziej znaczący wzrost p_t. Posiadanie ROR oraz uzyskanie kredytu konsumpcyjnego (a nie hipotecznego) również zwiększa p_t, ale

znacznie słabiej. Posiadanie kart płatniczych lub kredytowych, jako przejaw aktywnego korzystanie z usług badanego banku, zmniejsza ryzyko złego kredytu. Wzrost wpływów kwartalnych kredytobiorcy o 1 tys. zł zmniejsza wartość p_t o 0.033 (±0.003). Prawdopodobieństwo niepłacenia kredytu przez osobę starszą o rok jest mniejsze o prawie 0.001. Wraz ze wzrostem o 1 rok okresu na jaki został udzielony kredyt, p_t maleje o ok. 0.003 – 0.004. Spośród czterech źródeł dochodu największe ryzyko kredytowe związane jest z kredytobiorcą prowadzącym własną działalność gospodarczą. Kredytobiorcami o niższym ryzyku są osoby pobierające emeryturę lub rentę, a także studenci spłacający kredyty studenckie, przy czym udział (ilościowy i wartościowy) tej ostatniej grupy kredytów jest znikomy.

Głównym sposobem wykorzystania modeli jest prognozowanie prawdopodobieństwa nieterminowej spłaty rat kapitałowo-odsetkowych bądź całkowitego zaniechania ich spłat. W tym celu rozważamy cztery hipotetyczne sylwetki kredytobiorców, których charakterystykę zawiera Tabela 3. Zauważmy, że rozkłady a posteriori uzyskane dla p_t w modelach I i II są dość zbliżone (z wyjątkiem „starszej pani") i różnią się znacznie od wyników z modelu III (probitowego). Jeśli redukcje modelu I do modeli II i (zwłaszcza) III nie są zasadne (jak w przypadku naszych danych), to stosowanie w ocenie ryzyka kredytowego tylko modelu probitowego może prowadzić do błędnych wniosków. Przydatność modelu dwumianowego zależy nie tylko od zachowania stosowanej dystrybuanty w ogonach rozkładu, ale również od jej kształtu w centralnej części tego rozkładu.

Tabela 3

5. Uwagi końcowe

Proponowane uogólnienie modelu probitowego uzależnia kształt dystrybuanty $F(\cdot)$ od dwóch dodatkowych swobodnych parametrów, odpowiedzialnych za szybkość jej zbieżności do wartości granicznych 0 i 1 (czyli za grubość ogonów) oraz za jej wartość w punkcie 0. Uogólnienie to, a w szczególności parametryzacja $F(0)$ poprzez wprowadzenie swobodnego parametru asymetrii, wydaje się istotne z punktu widzenia wnioskowania statystycznego, co pokazał przykład empiryczny w poprzedniej części pracy. Problemem otwartym pozostaje ocena poprawy dopasowania modelu do danych binarnych i jego własności prognostycznych. Stanowi to przedmiot dalszych badań autorów.

Skośne rozkłady *t* mogą stanowić podstawę budowy nie tylko modeli dwumianowych, ale także modeli wielomianowych dla kategorii uporządkowanych czy modeli dla zmiennych uciętych

lub cenzurowanych. To proste dwuparametrowe uogólnienie daje możliwość badania empirycznej adekwatności modeli opartych na założeniach symetrii i normalności rozkładu zmiennej ukrytej, w praktyce najczęściej przyjmowanych.

Akademia Ekonomiczna w Krakowie

LITERATURA

[1] Albert J. Chib S., 1993, *Bayesian analysis of binary and polychotomous response data*, JASA (Journal of the American Statistical Association) vol. 88, 669-679.

[2] Amemiya T., 1981, *Qualitative response models: A survey*, Journal of Economic Literature vol.19, 1483-1536.

[3] Amemiya T., 1985, *Advanced Econometrics*, Harvard University Press, Cambridge (Massachusetts).

[4] Fernández C., Osiewalski J., Steel M., 1995, *Modeling and inference with* υ*-spherical distributions*, JASA (Journal of the American Statistical Association) vol. 90, 1331-1340.

[5] Fernández C., Steel M., 1998, *On Bayesian modeling of fat tails and skewness*, JASA (Journal of the American Statistical Association) vol. 93, 359-371.

[6] Gamerman D., 1997, *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*, Chapman and Hall, London.

[7] Greene W.H., 1993, *Econometric Analysis*, Macmillan, New York.

[8] Maddala G.S., 1983, *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.

[9] Marzec J., 2003a, *Badanie niewypłacalności kredytobiorcy na podstawie modeli logitowych i probitowych*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie nr 628, 103-117.

[10] Marzec J., 2003b, *Badanie niespłacalności kredytów za pomocą bayesowskich modeli dychotomicznych - założenia i wyniki*, Metody ilościowe w naukach ekonomicznych (red. A. Welfe), Wydawnictwo SGH w Warszawie, 73-86.

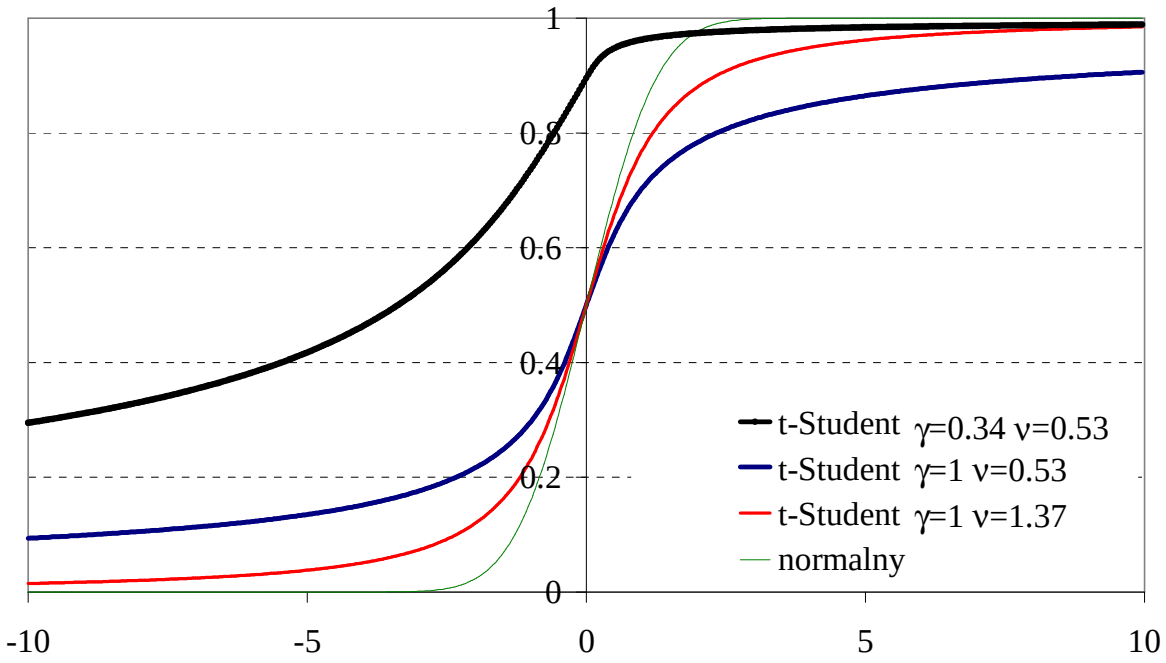
[11] Marzec J., 2003c, *Bayesowska analiza modeli dyskretnego wyboru (dwumianowych)*, Przegląd Statystyczny t. 50, 129-146.

[12] O'Hagan A., 1994, *Bayesian Inference*, Edward Arnold, London.

[13] Osiewalski J., Pipień M., 1999, *Bayesian forecasting of foreign exchange rates using GARCH models with skewed t conditional distributions*, MACROMODELS'98 - Conference Proceedings (red. W. Welfe), Vol. 2, Absolwent, Łódź, 195-218.

[14] Osiewalski J., Pipień M., 2000, *GARCH-In-Mean through skewed t conditional distributions: Bayesian inference for exchange rates*, MACROMODELS'99 – Conference Proceedings (red. W. Welfe, P. Wdowiński), Absolwent, Łódź, 354–369.

Rysunek 1. Dystrybuanty wybranych rozkładów typu *t* Studenta.



Źródło: obliczenia własne.

Tabela 1. Wartości oczekiwane i odchylenia standardowe a posteriori parametrów.

	Model I		Model II		Model III	
	$\varepsilon_t \sim$ St(0, 1, ν, γ)		$\varepsilon_t \sim$ St(0, 1, ν, γ=1)		$\varepsilon_t \sim$St(0,1,ν=∞, γ=1)	
Zmienna (parametr)	E(y)	D(y)	E(y)	D(y)	E(y)	D(y)
Stała	-0.272	0.550	-3.363	0.860	-1.006	0.097
Płeć	0.152	0.067	0.003	0.027	0.041	0.018
Wiek	-1.992	0.358	-1.200	0.129	-0.865	0.085
Wpływy	-70.261	7.422	-94.050	11.314	-1.757	0.186
ROR	1.409	0.189	1.446	0.189	-0.290	0.038
Karty	-0.272	0.110	-0.391	0.147	-0.175	0.034
Pośrednik	3.197	0.370	2.196	0.134	1.290	0.032
Typ Kredytu	1.084	0.513	1.773	0.826	0.022	0.080
Okres kredytu	-0.733	0.256	-0.719	0.096	-0.185	0.054
Zrdoch1	-0.752	0.203	-0.135	0.040	-0.092	0.029
Zrdoch2	0.734	0.192	0.010	0.082	0.319	0.040
Zrdoch3	-1.334	0.263	-0.470	0.156	-0.187	0.077
ν	0.530	0.051	1.372	0.088	-	-
γ	0.342	0.021	-	-	-	-

Źródło: obliczenia własne.

Tabela 2. Wartości oczekiwane i odchylenia standardowe a posteriori uśrednionych efektów krańcowych $T^{-1}\sum_t \beta_h f(-x_t\beta)$.

	Model I		Model II		Model III	
	$\varepsilon_t \sim$ St(0, 1, ν, γ)		$\varepsilon_t \sim$ St(0, 1, ν, γ= 1)		$\varepsilon_t \sim$St(0,1,$\nu=\infty$, γ=1)	
Zmienna (x_{th})	E($|y$)	D($|y$)	E($|y$)	D($|y$)	E($|y$)	D($|y$)
Płeć	0.007	0.003	0.0004	0.0035	0.008	0.003
Wiek	-0.094	0.020	-0.155	0.016	-0.171	0.017
Wpływy	-3.297	0.314	-12.140	1.240	-0.347	0.037
ROR	0.066	0.008	0.187	0.022	-0.057	0.008
Karty	-0.013	0.005	-0.050	0.019	-0.035	0.007
Pośrednik	0.149	0.004	0.284	0.013	0.255	0.006
Typ Kredytu	0.050	0.022	0.229	0.105	0.004	0.016
Okres kredytu	-0.035	0.014	-0.093	0.012	-0.037	0.011
Zrdoch1	-0.035	0.006	-0.017	0.005	-0.018	0.006
Zrdoch2	0.034	0.008	0.001	0.011	0.063	0.008
Zrdoch3	-0.062	0.009	-0.061	0.020	-0.037	0.015

Źródło: obliczenia własne.

Tabela 3 .Wartości oczekiwane i odchylenia standardowe a posteriori prawdopodobieństwa niespłacenia kredytu $p_t = \Pr(y_t=1)=1-F(-x_t\beta)$.

	najczęstszy klient		„młody	„starsza
Zmienna	pośrednik=1	pośrednik=0	biznesmen”	pani”
1 (wyraz wolny)	1	1	1	1
Płeć	1	1	1	0
Wiek (w latach)	40.2	40.2	21	60
Wpływy (w tys. zł)	10.2	10.2	0	1
ROR	1	1	0	1
Karty płatnicze	0	0	0	1
Pośrednik	1	0	1	0
Typ kredytu: konsumpcyjny	1	1	1	0
Okres kredytu (w latach)	2.61	2.61	2.61	2.61
Zrdoch1	0	0	0	1
Zrdoch2	0	0	1	0
Zrdoch3	0	0	0	0
Model I				
E($p_t \mid y$)	0.023	0.015	0.547	0.027
D($p_t \mid y$)	(0.002)	(0.001)	(0.013)	(0.004)
Model II				
E($p_t \mid y$)	0.020	0.015	0.559	0.049
D($p_t \mid y$)	(0.001)	(0.001)	(0.028)	(0.011)
Model III				
E($p_t \mid y$)	0.303	0.036	0.668	0.017
D($p_t \mid y$)	(0.014)	(0.002)	(0.015)	(0.003)

Źródło: obliczenia własne.