Jerzy Marzec, Uniwersytet Ekonomiczny w Krakowie.

JERZY MARZEC

BAYESOWSKA ANALIZA I TESTOWANIE MODELI DWUMIANOWYCH Z ROZKŁADEM *t* STUDENTA

1. WSTĘP

Początki analizy logitowej i probitowowej datuje się na lata trzydzieste poprzedniego stulecia; zob. [4]. Początkowo była ona stosowana w oparciu o dane otrzymywane z eksperymentów laboratoryjnych, więc pierwsze aplikacje pochodziły z obszaru nauk biologicznych, następnie psychologii i socjologii. Na gruncie nauk ekonomicznych ta klasa modeli nosi nazwę modeli jakościowych zmiennych endogenicznych lub modeli dyskretnego wyboru (ang. *qualitative dependent variables model, discrete choice model*)[1]. Od lat siedemdziesiątych i osiemdziesiątych dwudziestego wieku obserwuje się intensywne ich zastosowanie do opisu zjawisk ekonomicznych, przy czym Gourieroux [11] wyróżnia dwa główne kierunki rozwoju tych modeli: (1) konstrukcja modeli mających silne podstawy teoretyczne i opisujących zachowanie się jednostek (np. przedsiębiorstw, gospodarstw domowych) oraz (2) budowa modeli, w których zmienne endogeniczne mają charakter zmiennych jakościowych. Pionierskie badania Tobina i McFaddena wpisują się w ten pierwszy nurt; zob. [5], [21] i [32], natomiast niniejszy artykuł – w drugi.

Cechą charakterystyczną statystycznych modeli jakościowych zmiennych endogenicznych jest przyjmowanie założenia o rozkładzie logistycznym bądź normalnym, a zatem naturalnym kierunkiem uogólnienia jest zastosowanie rozkładu z szerszej klasy. Albert i Chib w pracy [1] zaproponowali, aby w tych modelach zastosować rozkład *t* Studenta o nieznanej liczbie stopni swobody $\nu > 0$. Wprowadzenie tego rozkładu jest w pełni uzasadnione, gdyż jak pokazali Mudholkar i George [22] rozkład logistyczny może być dobrze aproksymowany rozkładem *t* Studenta o 7-9 stopniach swobody. Umożliwia to statystyczną weryfikację obu najbardziej znanych modeli dla danych jakościowych: probitowego ($\nu \to +\infty$) i logistycznego, gdy $\nu \in [7;9]$. Testowanie ich zasadności w świetle posiadanych danych jest proste i sprowadza się do weryfikacji hipotezy dotyczącej parametru ν. Jednakże stosując jakikolwiek model z rozkładem *t* Studenta, w którym kluczowy parametr ν podlega estymacji, istotnym zagadnieniem staje się wybór odpowiedniej metody estymacji. W pracach [7], [29], [30] i [33] zwraca się uwagę, że już w przypadku modelu regresji liniowej z rozkładem *t* Studenta (o nieznanej liczbie stopni swobody) pojawiają się problemy z zastosowaniem metody największej wiarygodności – podstawowej metody estymacji danych jakościowych. Alternatywnym podejściem do estymacji i testowania wobec klasycznych metod jest wnioskowanie bayesowskie, które w ramach tej klasy modeli zostało po raz pierwszy zaproponowane przez Zellnera i Rossi; zob. [35] i [36]. Jest to małopróbkowe podejście, mające silne podstawy teoretyczne, które z powodzeniem było stosowane do estymacji i testowania modeli statystycznych z rozkładem *t* Studenta; zob. np. [1], [9].

Głównym celem niniejszej pracy jest prezentacja bayesowskiej konstrukcji i testowania modeli dwumianowych opartych na rozkładzie *t* Studenta, zilustrowana przykładem dotyczącym niespłacalności kredytów detalicznych. Niniejszy artykuł prezentuje nowe wyniki dalszych i pogłębionych badań, przedstawionych wcześniej w pracy [19], a także stanowi uzupełnienie badań prezentowanych w artykułach [27] i [28].

[1] Często z modelami zmiennych jakościowych łączone są modele zmiennych ograniczonych (ang. *limited–dependent variable models*), z których najbardziej znanym jest model tobitowy – model regresji cenzurowanej.

2. DEFINICJA MODELU

Model dla dychotomicznej zmiennej endogenicznej y_t z rozkładem *t* Studenta ma następującą postać

$$\begin{aligned} z_t &= x_t \cdot \beta + \varepsilon_t \\ y_t &= I_{[0,\infty)}(z_t), \end{aligned} \tag{1}$$

gdzie $I_\Omega(\omega)=1$, gdy $\omega \in \Omega$ i $I_\Omega(\omega)=0$, jeżeli $\omega \notin \Omega$. O składnikach losowych zakładamy, że posiadają identyczne, niezależne rozkłady z rodziny *t* Studenta o parametrze położenia równym zero, jednostkowej precyzji i nieznanej liczbie stopni swobody ν. Wektor x_t zawiera wartości zmiennych egzogenicznych lub ich znanych funkcji, zaś z_t jest nieobserwowalną zmienną reprezentującą użyteczność związaną z dokonanym wyborem przez jednostkę o numerze *t*. Zmienna y_t przyjmując dwie wartości, 0 albo 1, informuje o podjętej decyzji przez jednostkę.

Dla prawdopodobieństwa zaobserwowania y_t=1 (p_t) przyjmujemy zależność poprzez dystrybuantę zmiennej ε_t, $F_S(\cdot)$, w formie wielomianu drugiego stopnia względem oryginalnych zmiennych egzogenicznych w_{th}, czyli

$$\begin{aligned} &p_t \equiv \Pr(y_t = 1) = 1 - F_S(-x_t \cdot \beta), \quad \text{gdzie} \\ &x_t \cdot \beta = G(w_t, \beta) = \beta_1 + \sum_h \beta_h \cdot w_{th} + \sum_h \sum_{i \geq h} \beta_{hi} \cdot w_{th} \cdot w_{ti}. \end{aligned} \tag{2}$$

Osiewalski i Marzec [27] proponują nazywać powyższą specyfikację - z aproksymacją kwadratową - modelem dwumianowym II rzędu, w odróżnieniu od modelu dwumianowego I rzędu, czyli aproksymacji liniowej, gdy $\beta_{hi}=0$ dla każdego *h* oraz *i*. Takie rozszerzenie spotyka się w badaniach empirycznych, np. w pracach [4] i [14]. Z punktu widzenia estymacji, modele I i II rzędu różnią się jedynie liczbą parametrów, czyli wymiarem wektora β. Jednakże zasadnicza przewaga drugiego modelu nad pierwszym przejawia się w charakterze efektów krańcowych, co było głównym argumentem za wykorzystaniem modelu II rzędu; zob. [19] i następne prace autora.

Rozkład *t* Studenta można potraktować jako uogólnienie skończonej mieszanki rozkładów normalnych względem parametru skali (ang. *scale mixture of Normals*); zob. [8], [16]. Gęstość jednowymiarowej zmiennej losowej o rozkładzie *t* Studenta jest nieskończoną mieszanką gęstości rozkładów normalnych o identycznej wartości parametru położenia, gdy funkcja wagowa skonstruowana dla parametru skali (odwrotności wariancji) jest rozkładem gamma. Wynika to ze znanej tożsamości wiążącej te trzy rozkłady (zob. [25]):

$$\int_0^\infty f_N\left(\xi \middle| \mu, \tau^{-1}\right) \cdot f_G\left(\tau \middle| \nu_0/2, \nu_1/2\right) d\tau = f_S\left(\xi \middle| \mu, \nu_0, \nu_0/\nu_1\right), \tag{3}$$

gdzie $f_N(\cdot|a,b)$ jest funkcją gęstości zmiennej o rozkładzie normalnym, o wartości oczekiwanej *a* i wariancji *b*, zaś $f_G(\cdot|c,d)$ oznacza gęstość zmiennej o rozkładzie gamma o wartości oczekiwanej c/d i wariancji c/d^2, natomiast funkcja gęstości rozkładu *t* Studenta o parametrze niecentralności *e*, stopniach swobody *f*>0 i precyzji *g*>0 jest oznaczona przez $f_S(\cdot|e,f,g)$.

Na gruncie bayesowskim model regresji z rozkładem *t* Studenta ma szczególną interpretację. Model ten jest równoważny regresji z rozkładem normalnym, w którym zakłada się heteroscedastyczność składników losowych, przy czym parametry odpowiedzialne za heterogeniczność mają rozkład gamma sparametryzowany poprzez liczbę stopni swobody; por [15] s. 125. Jest to odmienne podejście w kwestii modelowania heterogeniczności w stosunku do tego, w którym wariancja składnika losowego jest funkcją wybranych zmiennych egzogenicznych; zob. [13], a w przypadku modelu probitowego zob. [12].

Innym, ważnym argumentem za stosowaniem w modelach danych jakościowych rozkładu *t* Studenta jest jego związek z rozkładem logistycznym. Mudholkar i George [22] zauważyli, że

kształt funkcji gęstości rozkładu logistycznego jest zbliżony raczej do gęstości rozkładu *t* Studenta o 9 stopniach swobody niż do rozkładu normalnego, jak to się powszechnie głosi. Rozkłady *t* Studenta i logistyczny charakteryzują się grubszymi ogonami i są bardziej wyostrzone w otoczeniu wartości modalnej niż rozkład normalny. Eksces dla rozkładu logistycznego wynosi 6/5, zaś dla *t* Studenta $6/(\nu-4)$. Zatem dla $\nu = 9$ współczynnik spłaszczenia dla obu rozkładów przyjmuje identyczną wartość. Natomiast wartość $\nu = 7{,}3$ minimalizuje miarę podobieństwa między rozkładami, opartą na metryce odległości Kulbacka-Leiblera, zdefiniowaną jako $\int f_L(\xi)\cdot\left|\ln\left[f_L(\xi)/f_S(\xi;\nu)\right]\right| d\xi$, gdzie $f_L(\xi)$ i $f_S(\xi;\nu)$ jest odpowiednio gęstością standaryzowanej zmiennej losowej o rozkładzie logistycznym i gęstością zmiennej *t* Studenta o jednostkowej wariancji; zob. także [24]. W wielu innych badaniach potwierdzono spostrzeżenia Mudholkara i George'a, zob. [1], [2].

Rozkład *t* Studenta w modelach jakościowych zmiennych endogenicznych zaproponowali Albert i Chib [1]. Przyjęli oni, iż dodatkowym parametrem podlegającym estymacji jest liczba stopni swobody ($\nu>0$). Wówczas za cenę dodatkowego parametru uzyskali naturalne uogólnienie modelu probitowego i logitowego. W celu estymacji parametrów zastosowali wnioskowanie bayesowskie, używając metod Monte Carlo opartych na łańcuchach Markowa.

3. UWAGI NA TEMAT ESTYMACJI

W modelu z rozkładem *t* Studenta przyjmujemy, że zmienne z_t mają niezależne rozkłady *t* Studenta, o ν stopniach swobody, z parametrem położenia (niecentralności) $x_t\beta$ i jednostkowym parametrem skali. W efekcie dyskretny rozkład próbkowy ma postać

$$p(y|\beta,\nu)=\prod_{t:y_t=0}F_S(-x_t\beta|\beta,\nu)\cdot\prod_{t:y_t=1}\left(1-F_S(-x_t\beta|\beta,\nu)\right). \qquad (4)$$

Wydaje się, że zastosowanie metody największej wiarygodności dla modelu określonego poprzez funkcję wiarygodności daną wzorem (4) jest dopuszczalne. Niestety istnieją pewne przesłanki natury metodologicznej, które dyskryminują MNW już w takim prostym przypadku, jak model regresji liniowej z rozkładem *t* Studenta o nieznanej liczbie stopni swobody. Zauważmy, że zarówno wówczas, jak i w modelu (4), funkcja wiarygodności traktowana jako funkcja ν (przy ustalonym β) szybko zmierza do stałej wielkości odpowiadającej wartości wiarygodności przy rozkładzie normalnym ($\nu = +\infty$). W efekcie dla $\nu\to+\infty$ gradient logarytmu wiarygodności dąży do zera. Zastosowanie wtedy jakichkolwiek metod numerycznych nie gwarantuje uzyskania zbieżności tych procedur, a przecież ocena parametru ν jest decydująca z punktu widzenia wprowadzonego uogólnienia. Na te problemy zwracają uwagę Fernández i Steel [7], którzy rozważali model regresji liniowej dla zmiennej z_t o rozkładzie *t* Studenta (przy założeniu niezależności obserwacji). Zauważyli oni, iż w pewnych przypadkach funkcja wiarygodności może być nieograniczona, np. gdy za ν przyjmuje się dostatecznie małą liczbę dodatnią ν_0 taką, że $1+\nu_0^{-1} > T/s$, gdzie s to liczba obserwacji takich, że $z_t = \mu_t$, gdzie $\mu_t = E[z_t]$. Wówczas wartość funkcji wiarygodności zmierza do nieskończoności, gdy parametr precyzji $\tau \to +\infty$. Restrykcja $\nu > 1$ wydaje się rozwiązywać ten problem w większości praktycznych przypadków. W pracy [33] analizowano poprzez badania symulacyjne własności estymatora MNW dla parametrów β,ν,τ w modelu próby prostej i regresji nieliniowej z rozkładem *t* Studenta. Uzyskane wnioski są następujące: (i) wraz ze wzrostem wymiaru wektora β lub spadkiem liczby obserwacji obciążenie estymatora rośnie lub pojawiają się problemy ze zbieżnością algorytmu, (ii) jeżeli obserwacji jest przynamniej 400, to te problemy zanikają, ale nie dotyczy to kluczowego parametru ν, gdyż obciążenie estymatora wciąż się utrzymuje. Ruud [30] podaje, że funkcja wiarygodności dla modelu regresji liniowej może mieć wiele punktów maksimów lokalnych, zwłaszcza dla $\nu\in(0;1)$. W konsekwencji proponuje stosować MNW w dwóch etapach. W pierwszym wyznacza się oceny dla pozostałych parametrów przy ustalonych wartościach dla ν, np. ν =1,2,3…,30. W drugim etapie następuje pełna optymalizacja

wszystkich parametrów przy ograniczeniach na wartości ν. Obszerna literatura z zakresu badań własności metody MNW w przypadku rodziny rozkładów *t* Studenta zawarta jest w pracach [29] i [33].

Powyższe problemy ze stosowaniem MNW w modelach z rozkładem *t* Studenta skłoniły autora do wykorzystania wnioskowania bayesowskiego. W sposób przystępny, w formie podręcznika akademickiego, jest ono prezentowane w monografiach [16] i [17]. W języku polskim to podejście – aspekty teoretyczne i przykłady zastosowań – jest omówione w pracach [25] i [26].

4. ELEMENTY ESTYMACJI BAYESOWSKIEJ

Bayesowski model statystyczny, jako kombinacja rozkładu próbkowego $p(y|\beta,\nu)$, danego wzorem (4), i rozkładu a priori dla parametrów $p(\beta,\nu)$, jest określony przez łączną (uogólnioną) gęstość daną formułą:

$$p(y,\beta,\nu)=p(y|\beta,\nu)\cdot p(\beta,\nu). \tag{5}$$

Estymacja bayesowska sprowadza się do wyznaczenia wg wzory Bayesa warunkowej względem obserwacji *y* funkcji gęstości dla parametrów. W tym przypadku pełny rozkład a posteriori przyjmuje następującą formę:

$$p(\beta,\nu|\, y)=\frac{1}{c_0}\cdot p(\beta,\nu)\cdot p(y|\,\beta,\nu), \tag{6}$$

gdzie stała normująca c_0 jest brzegowym rozkładem obserwacji:

$$c_0 = p(y)=\int_{\Theta} p(y|\,\theta)\cdot p(\theta)d\theta, \tag{7}$$

gdy $\theta=(\beta'\ \nu)'$.

Wyznaczenie brzegowych rozkładów a posteriori lub ich charakterystyk - jak wartości oczekiwane i odchylenia standardowe - dla poszczególnych składowych wektora parametrów θ lub ich nieliniowych funkcji np. Pr(y_t=1), wymaga dodatkowego całkowania. Niestandardowa postać rozkładu próbkowego powoduje, iż całkowanie analityczne jest niewykonalne. W praktyce stosuje się metody Monte Carlo. Albert i Chib w artykule [1] zaproponowali próbnik Gibbsa (zob. także [19]), zaś w pracach [27] i [28] wykorzystano algorytm Metropolisa i Hastingsa, gdy rozważano uogólnienie omawianego w artykule modelu, w postaci rozkładu *t* Studenta dopuszczającego asymetrię. Omówienie szczegółowych technik w ramach metod Monte Carlo można znaleźć m.in. w pracach: [3], [6], [10], [18] i [31].

4.1. UWAGI NA TEMAT ROZKŁADÓW A PRIORI

Pełną wiedzę o parametrach po zaobserwowaniu danych niesie rozkład a posteriori (6). Istnieje on, gdy całka – $p(y)$ – jest skończona. Jeżeli przyjmiemy niewłaściwy rozkład a priori dla ν o dziedzinie (0,+∞), to funkcja wiarygodności (4) dla $\nu\to+\infty$ jest zbieżna do dodatniej stałej:

$$\lim_{\nu\to+\infty}\prod_{t=1}^{T}\left(1-F_S\left(-x_t\beta|\beta,\nu\right)\right)^{y_t}\cdot F_S\left(-x_t\beta|\beta,\nu\right)^{1-y_t}=\prod_{t=1}^{T}\left(1-\Phi\left(-x_t\beta|\beta\right)\right)^{y_t}\cdot\Phi\left(-x_t\beta|\beta\right)^{1-y_t} \tag{8},$$

gdzie $\Phi(a)$ to wartość w punkcie *a* funkcji gęstości standaryzowanej zmiennej o rozkładzie normalnym. W konsekwencji całka z funkcji wiarygodności (po całej przestrzeni parametrów) jest nieskończona, tj. $\int_0^{+\infty} p(y|\,\nu)\,d\nu=+\infty$. Zatem konieczne jest przyjęcie właściwego rozkładu a priori dla parametru stopni swobody ν, gdyż niewłaściwy prowadziłby do braku rozkładu a posteriori.

Analogiczny warunek musi być spełniony dla parametru β, tzn. $\int_{R^k} p(\beta)\cdot p(y|\beta)\,d\beta < +\infty$. Ponadto w praktyce rozkład *t* Studenta o $\nu > 30$ może być przybliżany rozkładem normalnym. Przyjmując niewłaściwy rozkład o priori postaci $p(\nu) \propto c \cdot I_{(0;+\infty)}(\nu)$ otrzymujemy, iż iloraz szans a priori, że zaobserwujemy wartości mniejsze niż 30 albo większe niż 30, wynosi zero, czyli $\Pr(\nu \le 30)/\Pr(\nu > 30) = 0$. Oznacza to, że taki rozkład jest tylko na pozór nieinformacyjny. W rzeczywistości jest rozkładem silnie informacyjnym, gdyż z prawdopodobieństwem równym jeden dopuszcza a priori wyłącznie normalność.

W artykule [1] zastosowano niewłaściwy rozkład a priori dla β, którego także użyto pracach [19] i [28]. Z drugiej strony, brak jest formalnego dowodu, że wówczas rozkład a posteriori $p(\beta|y)$ istnieje, nawet w przypadku modelu probitowego. Jedynie w przypadku modelu logitowego Zellner i Rossi [36] pokazali, że gdy $p(\beta) \propto c$, to $p(\beta|y)$ istnieje przy pewnych założeniach dotyczących zmiennych objaśniających x_t, które mają charakter warunków wystarczających. Ponadto zauważyli oni, że skoro w modelu dwumianowym (w tym także wielomianowym) wartość funkcji wiarygodności jest ograniczona, $0 < L(\beta; y) < 1$, to $\int p(\beta)L(\beta; y)d\beta < \int p(\beta)d\beta$. Jeżeli $p(\beta)$ jest gęstością rozkładu właściwego, to $\int p(\beta)L(\beta; y)d\beta$ istnieje i jest skończona, więc $p(\beta|y)$ zawsze istnieje bez względu na typ rozkładu składnika losowego ε_t. Z drugiej strony przyjęcie właściwych rozkładów a priori jest konieczne w porównywaniu konkurencyjnych modeli, gdyż gwarantuje jednoznaczną identyfikowalność czynnika Bayesa.

Własne badania autora pokazały, że przyjęcie w modelu *t* Studenta $p(\beta) \propto c$ stwarza komplikacje natury numerycznej, które wskazują na problemy z istnieniem właściwego rozkładu a posteriori dla β. Ilustrują to wyniki prostego eksperymentu. Niech y_t będzie realizacją zmiennej losowej z modelu dwumianowego (1), gdzie ε_t ma rozkład *t* Studenta o pięciu stopniach swobody (jednostkowej precyzji i modalnej równej zero), zaś wartości wektora parametrów strukturalnych są ustalone $\beta = [-1\ -2\ 2\ 1]'$. Dla parametru ν przyjęto a priori rozkład wykładniczy o wartości oczekiwanej i odchyleniu standardowym równym 10. Ponadto $x_t = [1\ x_{t2}\ x_{t3}\ x_{t4}]$, gdzie x_{t2} i x_{t3} zostały wygenerowane ze standaryzowanego rozkładu normalnego, zaś x_{t4} ma rozkład zero-jedynkowy. Wygenerowano w ten sposób próbki o różnej liczebności. Wszystkie wyniki świadczyły o problemach z estymacją parametrów β. Przykładowe wyniki a posteriori dla T=100 przedstawia poniższa tabela (E(| *y*) i D(| *y*) oznaczają wartość oczekiwaną i odchylenie standardowe a posteriori):

Tabela 1

Jeżeli przyjmiemy informacyjny rozkład a priori dla β, np. normalny o zerowej wartości oczekiwanej i diagonalnej macierzy kowariancji, to rozkład a posteriori istnieje, więc nie obserwujemy jakichkolwiek problemów numerycznych. Natomiast w przypadku niewłaściwego rozkładu a priori wraz ze zwiększaniem liczby losowań w algorytmie Gibbsa bądź Metropolisa i Hastingsa nie obserwujemy stabilizacji wyników a posteriori, wręcz przeciwnie następuje gwałtowana ich rozbieżność. Dodatkowo uzyskano zaniżone wartości a posteriori dla parametru ν, a w efekcie precyzji rozkładu składnika losowego, który w tej sytuacji wyjaśnia zmienność zmiennej endogenicznej y_t. Zauważmy, że zachowane są znaki i proporcje między ocenami a prawdziwymi wartościami parametrów β. Sugeruje to, że rozkład a posteriori dla β nie istnieje, ponieważ $\int_{R^k} p(y|\beta)d\beta = +\infty$. Identyczne wnioski uzyskuje się na podstawie takich samych badań symulacyjnych, przeprowadzonych dla modelu wielomianowego kategorii uporządkowanych.

O problemach z istnieniem rozkładu a posteriori dla β w modelu regresji pisał Geweke [9]. Rozważał on liniowy model regresji przy założeniu, że składniki losowe mają niezależne rozkłady *t* Studenta o ustalonej liczbie stopni swobody. Pokazał, że jeżeli przyjmuje się nieinformacyjny

rozkład a priori dla β, to wartość oczekiwana a posteriori dla β istnieje i jest skończona, gdy $\nu > 2$ oraz odchylenie standardowe a posteriori rozkładu dla β istnieje i jest skończone, gdy $\nu > 4$.

Wobec powyższego, w przypadku rozkładu *t* Studenta stosujemy wyłącznie informacyjne rozkłady a priori, więc dla wektora parametrów strukturalnych β zakładamy rozkład normalny o wektorze wartości oczekiwanych β^* i macierzy precyzji H^*:

$$p(\beta) = f_N\left(\beta \middle| \beta^*, H^{*-1}\right). \tag{9}$$

W niniejszym artykule przyjęto, iż β^* jest wektorem zerowym, zaś H^* - macierzą jednostkową. Dla parametru ν przyjęty został rozkład z rodziny gamma, tj. wykładniczy o wartości oczekiwanej i odchyleniu standardowym *r*, o funkcji gęstości:

$$p(\nu) = f_{EXP}(\nu|r) = \frac{1}{r}\exp\left(-\frac{\nu}{r}\right)\cdot I_{(0,+\infty)}(\nu), \tag{10}$$

jak np. w pracy [1], [9], [27] i [28]. Wartość parametru *r* równa 10 implikuje rozkład a priori, który możemy uznać za mało informacyjny. Alternatywnie, można by zastosować nieinformacyjny rozkład a priori dla β, ale za cenę zgody, iż rozkład a priori dla ν jest ucięty na lewo od wartości 4.

4.2. SPECYFIKACJA I PORÓWNANIE MOCY WYJAŚNIAJĄCEJ KONKURENCYJNYCH MODELI

Proponując modele z rozkładem *t* Studenta oraz z aproksymacją II rzędu zamiast I rzędu należy postawić pytanie, czy takie uogólnienia w świetle posiadanych danych jest zasadne. Ponadto należy dokonać wyboru między modelem logitowym, probitowym a tym z rozkładem *t* Studenta. Odpowiedzi na te pytania uzyskamy dokonując testowania odpowiednich hipotez. Alternatywne modele są zdefiniowane poprzez różne rozkłady próbkowe $p(y|\theta, M_i)$ i rozkłady a priori $p(\theta|M_i)$, przy czym dla każdej specyfikacji wektor parametrów θ zwykle zawiera inne składowe. Zatem hipotezy dotyczyć będą postaci rozkładu dla składnika losowego i typu zależności dla nieobserwowalnej zmiennej z_t.

W ujęciu bayesowskim testowanie hipotez bądź równoważnie wybór najlepszego modelu statystycznego odbywa się w oparciu o prawdopodobieństwa a posteriori, które oblicza się wg wzoru Bayesa (zob. [15], [16], [17])

$$p(M_i|y) = \frac{p(y|M_i)\cdot p(M_i)}{\sum_{j=1}^{m} p(y|M_j)\cdot p(M_j)} \quad \text{dla } i\in\{1,\ldots,\text{m}\}, \tag{11}$$

gdzie $p(y|M_i)$ i $p(M_i)$ są odpowiednio brzegową gęstością wektora obserwacji i ustalanym przez badacza prawdopodobieństwem a priori modelu M_i, zaś *m* to liczba tych modeli. Przy wyborze najlepszego modelu wykorzystuje się także czynnik Bayesa $BF_{ij} = p(y|M_i)/p(y|M_j)$. Dla pary modeli wyraża on względną siłę wyjaśniającą przez dane zmienność zmiennej endogenicznej y_t. Kluczowym zagadnieniem jest wyznaczenie stałej $p(y|M_i)$. Różne propozycje w ramach metod Monte Carlo można znaleźć m.in. w pracach [15], [16], [17]. W niniejszym artykule brzegową gęstość wektora *y* liczono w oparciu o propozycję Newtona i Rafterego, czyli stosując aproksymację średnią geometryczną; zob. [23].

W pierwszej kolejności rozważamy cztery specyfikacje, z których najbardziej rozbudowany jest model II rzędu z rozkładem *t* Studenta (M_1), w którym zależność między nieobserwowalną zmienną z_t a zmiennymi objaśniającymi ma postać wielomianu drugiego stopnia. Najprostszą specyfikacją jest standardowy model probitowym (M_4), który otrzymujemy z M_1 zakładając $\nu\to+\infty$ oraz $\beta_{hi} = 0$ dla każdego *h* oraz *i* w równaniu (2). Modele M_2 i M_3 stanowią formy pośrednie. W

modelu M_3, czyli modelu probitowym z aproksymacją kwadratową, próbuje się uzyskać poprawę dopasowania jedynie poprzez zwiększenie liczby czynników wyjaśniających zaobserwowane wartości zmiennej y_t. Natomiast w M_2 – wyłącznie poprzez zastosowanie rozkładu z grubymi ogonami (ν swobodne, zaś $\beta_{hi}=0$ dla $\forall$ h oraz i). Wyboru najlepszego modelu dokonaliśmy w oparciu o wyniki jednoczesnego testowania czterech hipotez dotyczących wektora β i parametru ν, które mają postać restrykcji punktowych. Cztery rozważane modele stanowią parami wykluczające się specyfikacje, gdyż rozkłady a priori dla parametrów są rozkładami ciągłymi, więc prawdopodobieństwo spełnienia restrykcji $\nu\to+\infty$ lub $\beta_{hi}=0$ (dla $\forall$ h oraz i) wynosi zero. Wówczas istotną kwestią jest sposób ustalenia prawdopodobieństw $p(M_i)$. Najprościej przyjąć dla każdego identyczną wartość równą 1/m. Jednakże badane modele znacznie różnią się liczbą parametrów. W modelu M_1 łączna liczba parametrów wynosi 80, w modelu M_2 – 15, w M_3 – 79, zaś najprostszym M_4 – 14. Zatem uzasadnione jest faworyzowanie specyfikacji oszczędnie sparametryzowanych, czyli M_2 i M_4, więc przyjęliśmy, że $p(M_i)\propto 2^{-k_i'}$, gdzie k_i' jest liczbą parametrów w i-tym modelu; zob. [26].

W ramach modelu dwumianowego z rozkładem *t* Studenta o nieznanej liczbie stopni swobody jest możliwe testowanie modeli: probitowego i logitowego, Model logitowy otrzymujemy, gdy $\nu\in[7;9]$, zaś w przypadku probitu przyjmijmy, że $\nu>30$, choć formalnie odpowiada on przypadkowi $\nu\to+\infty$.[2] W obu przypadkach rozważamy wyłącznie specyfikację II rzędu dla zmiennej z_t, więc wektor β jest identyczny dla tych modeli. Wówczas hipotezy będące przedmiotem testowania dotyczą wyłącznie parametru ν i mają charakter nierówności. Aby w tym przypadku analizowany zbiór modeli był pełny, rozważyliśmy jeszcze trzeci model, w którym $0<\nu<7$ lub $9<\nu<30$. Taki sposób zdefiniowania modeli ułatwia ich testowanie, gdyż w praktyce nie ma potrzeby estymacji trzech wyżej wymienionych modeli. Testowanie odbywa się w oparciu o wyniki estymacji modelu M_1, *t* Studenta z $\nu\in(0,+\infty)$, i sprowadza się do porównania prawdopodobieństw a posteriori spełnienia nierównościowych restrykcji dla ν, definiujących trzy powyższe modele. Prawdopodobieństwa a priori poszczególnych specyfikacji ustaliliśmy w oparciu o rozkład a priori dla ν, tj. dla logitu jest ono równe prawdopodobieństwu a priori, że $\nu\in[7;9]$, dla probitu - $\Pr(\nu>30)$. W tym przypadku, z punktu widzenia metodologii warto podkreślić, że wnioskowanie na podstawie modelu M_1, *t* Studenta z $\nu\in(0;+\infty)$, odpowiada bayesowskiemu łączeniu wiedzy (ang. *Bayesian model averaging*, *Bayesian pooling*), gdy rozważamy modele logitowy i probitowy. To podejście pozwala zmniejszyć niepewność badacza w odniesieniu do rozkładu próbkowego i postaci analitycznej modelu. Wnioskowanie o badanym zjawisku czy prognozowanie może odbywać się w oparciu o łączną, uśrednioną informację płynącą z konkurencyjnych specyfikacji, np. poprzez konstruowanie prognoz kombinowanych zwłaszcza w sytuacji, gdyby dane nie dostarczają zdecydowanych dowodów wyłącznie na rzecz jednego z modeli. Ta przesłanka teoretyczna po raz kolejny potwierdza zasadność wprowadzenia modelu z rozkładem *t* Studenta. Kluczowym kwestią staje się wyznaczenie wag dla poszczególnych modeli, czyli $p(M_i|y)$.

5. PRZYKŁAD EMPIRYCZNY: BADANIA NIESPŁACALNOŚCI KREDYTÓW

Model dwumianowy oparty na rozkładzie *t* Studenta może mieć zastosowanie w bankowych systemach scoringowych, co ilustrujemy poniższym przykładem. Posiadając informacje o 39034 kredytach konsumpcyjnych i hipotecznych, udzielonych klientom detalicznym przez bank komercyjny w ciągu prawie dwóch lat, dokonaliśmy estymacji i testowania prezentowanych wcześniej modeli. Zmienna endogeniczna y_t informuje o kategorii ryzyka udzielonego kredytu,

[2] Korzystamy z faktu, że często przyjmuje się, iż rozkład normalny jest dobrze aproksymowany przez rozkład *t* Studenta o liczbie stopni swobody równej przynajmniej 30.

Wartość $y_t = 1$ oznacza, że kredyt został zakwalifikowany do kategorii zagrożonych (poniżej standardu, wątpliwych i straconych), a w przeciwnym przypadku kredyt należy do grupy należności normalnych ($y_t = 0$). Zbiór potencjalnych zmiennych egzogenicznych wyjaśniających ryzyko pojedynczej umowy kredytowej zawierał (jak we wcześniejszych pracach autora): płeć, wiek kredytobiorcy, wpływy na rachunki typu ROR, posiadanie rachunku ROR, informację o tym, czy kredytobiorca posiada karty płatnicze lub kredytowe wydane przez ten bank, sposób udzielenia kredytu (przez pośrednika kredytowego albo bezpośrednio przez bank), typ kredytu (kredyt konsumpcyjny albo hipoteczny), okres trwania umowy kredytowej, kwota przyznanego kredytu, waluta kredytu, podstawowe źródło dochodu uzyskiwanego przez kredytobiorcę (umowa o pracę, albo renta lub emerytura, albo własna działalność, umowa o dzieło lub umowa zlecenie, albo inne źródło). Szczegółowe informacje na ten temat znajdują się m.in. w pracy [20].

Tabela 1 przedstawia rezultaty porównywania mocy wyjaśniającej czterech modeli, w zależności od przyjętych prawdopodobieństw a priori dla każdej ze specyfikacji. Ich testowanie sprowadza się do weryfikacji hipotez dotyczących wektora β i parametru ν, więc dotyczą postaci równania dla zmiennej z_t (aproksymacja I czy II rzędu) oraz kwestii wyboru między rozkładem normalnym a *t* Studenta. Dane empiryczne zdecydowanie preferują model M_1, bez względu na założenie badacza dotyczące prawdopodobieństw a priori. Prawie cała masa prawdopodobieństwa jest przypisana tej specyfikacji, zaś prawdopodobieństwa a posteriori pozostałych modeli wynoszą praktycznie zero. Dane świadczą, iż następny w kolejności model M_3 jest pięć rzędów wielkości gorszy (mniej prawdopodobny) od M_1. Jeżeli preferujemy specyfikacje oszczędnie sparametryzowane, to prawdopodobieństwo a posteriori modelu M_3 jest kilkadziesiąt razy większe niż M_2, zaś standardowy model probitowy jest ostatni w tym rankingu.

Zatem, spośród dwóch możliwych rozszerzeń standardowego modelu probitowego (M_4), dane empiryczne bardziej wskazują na aproksymację II rzędu (model M_3) niż na przyjęcie rozkładu *t* Studenta o nieznanej liczbie stopni swobody. Aczkolwiek samo zastosowanie aproksymacji kwadratowej dla z_t, zamiast liniowej nie wystarcza, aby model M_3 był najlepszy. Dopiero wzbogacenie tego modelu o dodatkowy parametr ν rodzi najlepszy model M_1.

Tabela 2

Wyniki testowania modelu logitowego i probitowego potwierdzają wcześniej sformułowane wnioski. Z rozkładu a posteriori dla parametru ν wynika, że wartości ν większe niż 30 są bardzo mało prawdopodobne a posteriori, tzn. $\Pr(\nu>30|\, y) \approx 0$, gdy tymczasem a priori przyjęto, że $\Pr(\nu>30) = 0{,}05$ dla $r = 10$; zob. Rysunek 1. Wartość oczekiwana i odchylenie standardowe dla ν wynosi odpowiednio 6,76 i 1,55. Zatem model probitowy jest zdecydowanie odrzucany przez dane. Rozkład a posteriori dla ν jest tak zlokalizowany, że właściwie cała masa prawdopodobieństwa znajduje się w przedziale (4; 12). Model logitowy odpowiada hipotezie H_0: $\nu \in [7;9]$, której prawdopodobieństwo a posteriori wynosi 0,27, zaś dla alternatywnej H_1: $\nu \notin [7;9]$ wynosi 0,73, przy naturalnym założeniu, że prawdopodobieństwa a priori obu hipotez, wynikające z rozkładu a priori dla ν, wynoszą $\Pr(H_0)$=0,09 i $\Pr(H_1)$=0,91. Iloraz szans a priori wynosi 9/91 na niekorzyść modelu logitowego, zaś iloraz szans a posteriori $\Pr(H_0|y)/\Pr(H_1|y)$ wynosi 0,37. Zatem czynnik Bayesa ($BF_{H_0H_1}$) kształtuje się na poziomie 3,7 na korzyść modelu logitowego. Odnosząc się do interpretacji wartości czynnika Bayesa, przedstawionej w pracy [15], stwierdzamy, że dane solidnie świadczą przeciw hipotezie H_1. Posiadana próba opowiada się za hipotezą zerową, lecz gdy uwzględnimy wiedzę a priori, to model *t* Studenta z $\nu \in (0;7)\cup(9;\infty)$ jest prawie 3-krotnie bardziej prawdopodobny niż model logitowy. W tej sytuacji, gdy informacje płynące z próby są zdecydowanie odmienne od założeń a priori, nie ma potrzeby ograniczania się tylko do jednego z modeli. Podejście bayesowskie pozwala na wspólne wnioskowanie o parametrach w oparciu o obie specyfikacje. W tym przypadku wnioskowanie o niespłacalności kredytów na podstawie modelu ogólnego M_1 odpowiada bayesowskiemu łączeniu wiedzy, zawartej w modelu logitowym,

probitowym (ν>30) i w modelu *t* Studenta o liczbie stopni swobody $\nu \in (0;7) \cup (9;\infty)$, gdy waga przypisywana pierwszej specyfikacji wynosi 0,27, zaś pozostałym łącznie 0,73.

Rysunek 1

W odniesieniu do rezultatów estymacji parametrów poszczególnych modeli można sformułować kilka wniosków. W przypadku badań o tak dużej liczbie obserwacji rola rozkładów a priori jest niewielka. Oceny i błędy średnie szacunku MNW dla standardowego modelu probitowego oraz z aproksymacją II rzędu są prawie identyczne z wynikami a posteriori modeli M_4 i M_3. W modelu M_1 wartość oczekiwana a posteriori dla parametru ν jest równa 6,76, a odchylenie standardowe ±1,55, zaś w M_2 wartości tych charakterystyk wynoszą odpowiednio 4,81 i ±0,55. Zatem przyjęcie dla zmiennej z_t w modelu M_1 aproksymacji kwadratowej, zamiast liniowej, jak w M_2, powoduje, że rozkład a posteriori składników losowych ε_t charakteryzuje się mniejszym rozproszeniem mierzonym wariancją $\nu/(\nu-2)$. Odchylenie standardowe w M_1 wynosi 1,42 (±0,14), zaś w M_2 kształtuje się na poziomie 1,71 (±0,14). Oznacza to, że rola zmiennych objaśniających w wyjaśnianiu zmienności y_t, w stosunku do składnika losowego, jest większa w modelu M_1 niż w M_2.

Informacyjną rolę danych ilustruje zestawienie rozkładów a priori i a posteriori w przypadku podstawowej charakterystyki badanego zjawiska tj. prawdopodobieństwa niespłacenia kredytu p_t. Rysunek 2 ilustruje pełną wiedzę o p_t przedstawiając te rozkłady dla wybranych czterech kredytobiorców.

Rysunek 2

Dane empiryczne zdecydowanie zmodyfikowały wstępne założenia o p_t. Rozkłady a priori są u-kształtne, o medianie równej 0,5 i odchyleniu ćwiartkowym równym 0,41, więc są płaskie na przedziale (0,1; 0,9). Natomiast rozkłady a posteriori są skupione w wąskim przedziale i charakteryzują się większą precyzją. Rozproszenie rozkładu a posteriori jest mniejsze dla starszej pani i typowego klienta, któremu udzielono kredyt bezpośrednio w banku. W przypadku pozostałych dwóch kredytobiorców występuje większe ryzyko kredytowe mierzone nie tylko wartością p_t, ale także większym rozproszeniem rozkładu a posteriori.

Wszystkie modele zgodnie prognozują, że największe ryzyko kredytowe związane jest z klientem będącym młodym mężczyzną, który utrzymuje się z prowadzenia własnej działalności i nie korzysta z jakichkolwiek innych usług badanego banku oprócz kredytu, który został mu udzielony poprzez pośrednika. Model M_1 wskazuje, że prawdopodobieństwo niedotrzymania umowy kredytowej (p_t) przez tego młodego biznesmena jest bardzo wysokie i wynosi 0,55 (±0,04). Jednocześnie we wszystkich modelach precyzja wnioskowania o p_t dla tego kredytobiorcy jest najmniejsza w porównaniu do pozostałych klientów. Najmniejsze ryzyko kredytowe, spośród czterech rozważanych kredytobiorców, związane jest ze starszą panią utrzymującą się z emerytury, której udzielono kredyt hipoteczny. Wszystkie modele zgodnie wskazują, że w tym przypadku prawdopodobieństwo niespłacenia kredytu wynosi praktycznie zero, zaś niepewność wnioskowania o tej wielkości jest niewielka. Największe różnice w oszacowaniu p_t dotyczą typowego klienta, tzn. o cechach najczęstszych w próbie (dotyczy zmiennych jakościowych) i przeciętnych (dla zmiennych ciągłych) w badanej zbiorowości, który uzyskał kredyt poprzez pośrednika. Dla tego klienta, w modelu M_1 prawdopodobieństwo „złego kredytu" wynosi 0,28 (±0,04) i jest mniejsze niż w modelach probitowych I i II rzędu, w których ta wielkość jest na poziomie 0,31-0,32. W przypadku najbardziej licznej grupy kredytobiorców, reprezentowanej przez typowego klienta, który uzyskał kredyt konsumpcyjny bezpośrednio w banku, model M_1 szacuje p_t na poziomie 0,04 (±0,004), co z praktycznego punktu widzenia oznacza minimalne ryzyko bądź jego brak. Pozostałe modele prognozują to prawdopodobieństwo na podobnym poziomie.

Podsumowując, uzyskane wyniki pozwalają stwierdzić, iż ryzyko kredytowe w przypadku typowego kredytobiorcy jest na niskim, akceptowalnym poziomie. Ponadto prawdopodobieństwo

niespłacenia kredytu jest bardzo duże, gdy mamy do czynienia z klientami w młodym wieku, którzy nie korzystali wcześniej z produktów badanego banku, a kredyt otrzymali poprzez pośrednika, kupując np. sprzęt AGD w systemie ratalnym. Długoletni klient banku, reprezentowany przez starszą panią, jawi się jako bardzo wiarygodny kredytobiorca. Odnosząc się do wyników, uzyskanych na podstawie rozważanych modeli, można stwierdzić, że są widoczne różnice w wielkości oszacowanego prawdopodobieństwa „złego kredytu" między preferowanym przez dane modelem M_1, a pozostałymi, zwłaszcza tym często stosowanym, najprostszym modelem probitowym.

6. PODSUMOWANIE

Głównym celem niniejszego artykułu była prezentacja analizy bayesowskiej i testowania modeli dwumianowych opartych na rozkładzie *t* Studenta o nieznanej liczbie stopni swobody. Budując model bayesowski należy stosować wyłącznie właściwe rozkłady a priori, gdyż w przeciwnym przypadku - jak pokazały badania symulacyjne – występują problemy numeryczne, wskazujące na nieistnienie właściwych rozkładów a posteriori.

Zaproponowane w artykule uogólnienie polegało na wprowadzeniu rozkładu dopuszczającego grube ogony oraz zastosowaniu aproksymacji kwadratowej dla zależności między nieobserwowaną zmienną reprezentującą użyteczność decyzji kredytobiorcy (spłaty w terminie rat kapitałowo-odsetkowych bądź nie) a czynnikami egzogenicznymi, wpływającymi na ryzyko kredytowe. W świetle przeprowadzonych testów okazało się ono zasadne. Dane empiryczne zdecydowanie wskazują na aproksymację II rzędu (zależność kwadratowa) niż na formułę liniową. Aczkolwiek samo zastosowanie aproksymacji kwadratowej dla z_t, zamiast liniowej nie wystarcza, aby uzyskany w ten sposób model był najlepszy. Dopiero jego wzbogacenie o dodatkowy parametr ν, czyli dopuszczenie rozkładu o grubych ogonach, zrodziło najlepszy model. Wprowadzenie dodatkowego parametru – liczby stopni swobody ν – pozwoliło także na równoczesne testowanie dwóch najczęściej stosowanych modeli dwumianowych: probitowego i logitowego. W tym przypadku dane solidnie świadczą na rzecz modelu logitowego i zdecydowanie odrzucają model probitowy. Model *t* Studenta o nieznanej, podlegającej estymacji, liczbie stopni swobody jest interesującą alternatywą dla obu wspomnianych modeli. Wnioskowanie na jego podstawie odpowiada bayesowskiemu łączeniu wiedzy, gdy rozważamy modele logitowy i probitowy.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Albert J., Chib S., (1993), *Bayesian Analysis of Binary and Polychotomous Response Data*, Journal of the American Statistical Association, vol. 88, s. 669-679.

[2] Arabmazar A., Schmidt P., (1981), *Further Evidence on the Robustness of the Tobit Estimator to Heteroscedasticity*, Journal of Econometrics, 17, s. 253-258.

[3] Chib S., Greenberg E., (1995), *Understanding the Metropolis–Hastings Algorithm*, The American Statistician, 49, s. 327-335.

[4] Cramer J.S, (2003), *Logit Models From Economics and Other Fields*, Cambridge University Press, Cambridge.

[5] Domencich T.A., McFadden D.L., (1975), *Urban Travel Demand: A Behavioral Analysis*, North-Holland Publishing Co., Amsterdam.

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

[7] Fernández C., Steel M., (1999), *Multivariate Student t Regression Models: Pitfalls and Inference*, Biometrika, vol. 86, 153-167.

[8] Fernández C., Steel M.F.J, (2000), *Bayesian Regression Analysis with Scale Mixture of Normals*, Econometric Theory, 16, s. 80-101.

[9] Geweke J., (1993), *Bayesian Treatment of the Independent Student–t Linear Model*, Journal of Applied Econometrics, vol. 8, s. 19-40.

[10] Geweke J., (1996), *Monte Carlo Simulation and Numerical Integration*, [w:] Handbook of Computational Economics, red. H. Amman, D. Kendrick, J. Rust, North-Holland, Amsterdam.

[11] Gourieroux C., (2000), *Econometrics of Qualitative Dependent Variables*, Cambridge University Press, Cambridge.

[12] Greene W.H., (2003), *Econometric Analysis* (5th editon), Prentice Hall, New York.

[13] Harvey A., (1976), *Estimating Regression Models with Multiplicative Heteroscedasticity*, Econometrica, 44, s. 461-465.

[14] Hosmer D., Lemeshow S., (2000), *Applied Logistic Regression*, Wiley, New York.

[15] Kass R.E., Raftery A., (1995), *Bayes Factor*, Journal of the American Statistical Association, vol. 90, nr 90, s. 773-795.

[16] Koop G., (2003), *Bayesian Econometrics*, Wiley, Chichester.

[17] Lancaster T., (2004), *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing, Oxford.

[18] Magiera R., (2005), *Modele i metody statystyki matematycznej. Część I: Rozkłady i symulacje stochastyczne*, Oficyna wydawnicza GiS, Wrocław.

[19] Marzec J., (2003), *Bayesowska analiza modeli dyskretnego wyboru (dwumianowych)*, Przegląd Statystyczny, tom 50 nr 4, s. 129-145.

[20] Marzec J., (2006), *Bayesowski model wielomianowy z rozkładem t Studenta dla kategorii uporządkowanych*, Metody ilościowe w naukach ekonomicznych (red. A. Welfe), Wydawnictwo SGH w Warszawie, s. 123-144.

[21] McFadden D.L., (1984), *Econometrics Analysis of Qualitative Response Models*, [w:] Handbooks of Econometrics vol II, red.Z. Griliches M. Intriligator, Elsevier Science Publishers.

[22] Mudholkar G ., George E., (1978), *A Remark on the shape of the logistic distribution*, Biometrika, nr 65, s. 667-668.

[23] Newton M.A., A.E. Raftery (1994), *Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion)*, Journal of the Royal Statistical Society B vol.56, s. 3-48.

[24] O'Brien S.M., Dunson D.B., (2004), *Bayesian Multivariate Logistic Regression*, Biometrics, 60, s. 739-746.

[25] Osiewalski J., (1991), *Bayesowska estymacja i predykcja dla jednorównaniowych modeli ekonometrycznych*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria specjalna: Monografie nr 100, Kraków.

[26] Osiewalski J., (2001), *Ekonometria bayesowska w zastosowaniach*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

[27] Osiewalski J., Marzec J., (2004), *Model dwumianowy II rzędu i skośny rozkład Studenta w analizie ryzyka kredytowego*, Folia Oeconomica Cracoviensia, vol. 45. s. 63-84.

[28] Osiewalski J., Marzec J., (2004), *Uogólnienie dychotomicznego modelu probitowego z wykorzystaniem skośnego rozkładu Studenta*, Przegląd Statystyczny, t. 51, s. 13-24.

[29] Pawitan Y., (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Clarendon Press, Oxford.

[30] Ruud P.A., (2000), *An Introduction to Classical Econometric Theory*, Oxford University Press, Oxford.

[31] Tierney L., (1994), *Markov Chains for Exploring Posterior Distributions (with discussion)*, Annals of Statistics, 22, s. 1701-1762.

[32] Tobin J., (1958), *Estimation of Relationships for Limited Dependent Variables*, Econometrica, vol. 26, nr 1, s. 24-36.

[33] Vasconcellos K., Gomes da Silva S., (2005), *Corrected Estimations for Student t Regression Models with Unknown Degrees of Freedom*, Journal of Statistical Computation and Simulation, vol. 75, nr 6, s. 409-423.

[34] Zellner A., (1971), *An Introduction to Bayesian Inference in Econometrics*, J.Wiley, New York.
[35] Zellner A., (1983), *Bayesian Analysis of Simple Multinomial Logit Model*, Economics Letters, 11, s. 133-136.
[36] Zellner A., Rossi P. (1984), *Bayesian Analysis of Dichotomous Quantal Response Models*, Journal of Econometrics, 25, s. 365-393.

Tabela 1

Porównanie wyników bayesowskich przy różnych rozkładach a priori dla danych symulowanych

Typ rozkładu a priori:			Właściwy rozkład $p(\beta) \sim N(0, 9\cdot I)$		Niewłaściwy rozkład $p(\beta) \propto c$	
Parametr	Zmienna	Prawdziwa wartość	$E(\beta \mid y)$	$D(\beta \mid y)$	Numeryczna $E(\beta \mid y)$	Numeryczne $D(\beta \mid y)$
β_1	$x_{t1} \equiv 1$	–1	–1,51	0,60	-46	42
β_2	x_{t2}	–2	–2,68	0,91	-96	40
β_3	x_{t3}	2	2,71	0,79	103	42
β_4	x_{t4}	1	0,86	0,66	29	25
ν	–	5	4,53	3,15	0,35	0,09

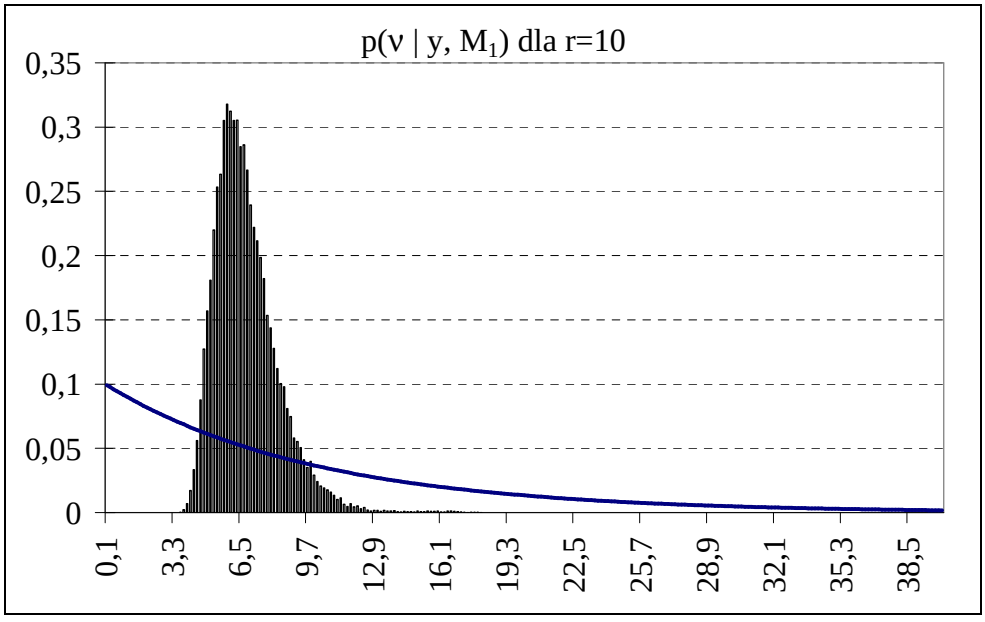
Źródło: obliczenia własne.

Tabela 2

Brzegowe gęstości wektora obserwacji i prawdopodobieństwa a posteriori badanych modeli.

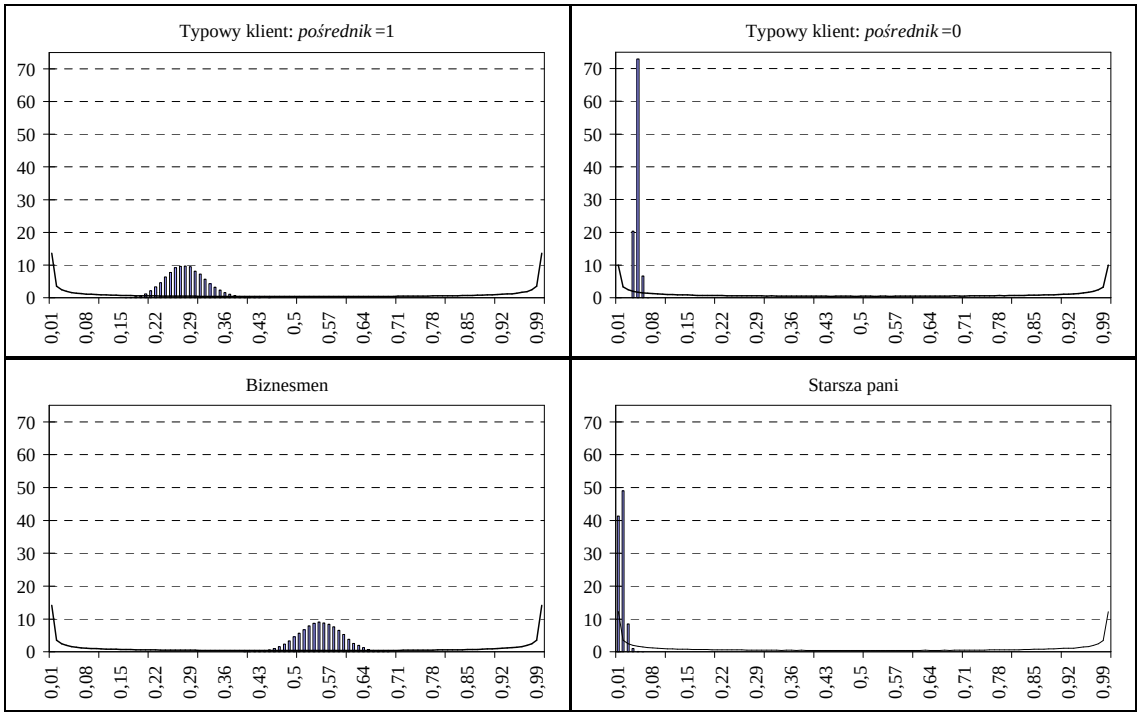
Model	M_1	M_2	M_3	M_4
Liczba parametrów (k')	80	15	79	14
$\ln p(y \mid M_i)$	-13599,5	-13808,7	-13610,8	-13845,8
BF	1	$1{,}45\times10^{-91}$	$1{,}33\times10^{-5}$	$1{,}18\times10^{-107}$
Log_{10} BF	–	–90,8	–4,9	–106,9
$p(M_i)$	0,25	0,25	0,25	0,25
$p(M_i \mid y)$	≈1	$1{,}45\times10^{-91}$	$1{,}33\times10^{-5}$	$1{,}18\times10^{-107}$
$p(M_i) \propto 2^{-k_i'}$	9×10^{-21}	0,3333	$1{,}8\times10^{-20}$	0,6666
$p(M_i \mid y)$	≈1	$5{,}4\times10^{-72}$	$2{,}7\times10^{-5}$	$8{,}7\times10^{-88}$

Źródło: opracowanie własne.



Rysunek 1. Histogram brzegowego rozkładu a posteriori (słupki) i rozkład a priori (linia) dla parametru ν w modelu M_1 (r=10)

Źródło: opracowanie własne.



Rysunek 2. Brzegowe rozkłady a priori (linia ciągła) i a posteriori (słupki) Pr(y_t=1) dla wybranych kredytobiorców w modelu M_1

Źródło: opracowanie własne.

BAYESOWSKA ANALIZA I TESTOWANIE MODELI DWUMIANOWYCH Z ROZKŁADEM t STUDENTA

Streszczenie

Głównym celem niniejszego artykułu była prezentacja bayesowskiej konstrukcji i testowania modeli dwumianowych opartych na rozkładzie *t* Studenta, zilustrowana przykładem dotyczącym analizy niespłacalności kredytów detalicznych. Omówiono podstawowe problemy związane z estymacją za pomocą metody największej wiarygodności. Przedstawiono konstrukcję rozkładów a priori, zwracając uwagę, że wyłącznie właściwe rozkłady gwarantują istnienie rozkładów a posteriori. W celu bayesowskiej estymacji wykorzystano jedną z metod Monte Carlo łańcuchów Markowa, tj. algorytm Metropolisa i Hastingsa. Testowanie modeli bayesowskich przeprowadzono w oparciu o czynniki Bayesa i prawdopodobieństwa a posteriori.

Zaproponowane uogólnienie polegało na przyjęciu rozkładu dopuszczającego grube ogony oraz zastosowaniu aproksymacji kwadratowej dla zależności między nieobserwowalną zmienną reprezentującą użyteczność decyzji kredytobiorcy (spłaty w terminie rat kapitałowo-odsetkowych bądź nie) a czynnikami egzogenicznymi, wyjaśniającymi ryzyko kredytowe. W świetle wyników przeprowadzonych testów okazało się ono zasadne. Wprowadzenie dodatkowego parametru – liczby stopni swobody ν – pozwoliło także na równoczesne testowanie dwóch najczęściej stosowanych modeli dwumianowych: probitowego i logitowego. Dane solidnie świadczą na rzecz modelu logitowego i zdecydowanie odrzucają model probitowy. Model *t* Studenta o nieznanej, podlegającej estymacji, liczbie stopni swobody jest interesującą alternatywą dla obu wspomnianych modeli, gdyż wnioskowanie na jego podstawie odpowiada bayesowskiemu łączeniu wiedzy, gdy rozważamy te dwa, najczęściej stosowane w literaturze modele.

Jerzy Marzec, Uniwersytet Ekonomiczny w Krakowie.

BAYESIAN ANALYSIS AND TESTING FOR STUDENT T-DICHOTOMOUS QUANTAL RESPONSE MODELS

Summary

This paper is concerned with statistical inference in Student *t*-Dichotomous Quantal Response Models. We analyze discrete choice models from Bayesian point of view. Bayesian methods for modeling binary data are applied because of a nonstandard property of maximum likelihood function in the case of the Student *t* model. Details of the construction of the prior are presented. We observed than only a proper prior density guarantees existence of the posterior density of parameters. Bayesian inference in this model is feasible using a Markov chain Monte Carlo posterior simulator i.e. Metropolis– Hastings algorithm.

We took advantage of the fact that, approximately, one can view the logistic distribution as a member of the *t* family. Thus, there is possibility testing probit and logit models within the confines of *t*-model. Then we illustrate probit and logit models extensions for retail loan data. Thus, the first generalization relies on it, that the errors are Student-t distributed with unknown degrees of freedom. Secondly, we assume that a latent variable (represent a utility associated with loans repayment) is a second-order polynomial of the explanatory variables.

An important concern of this paper is the question of comparing the fit of alternative models. We show that the posterior model probabilities and the Bayes factors framework are quite useful for this purpose. The data give very strong evidence that Student-t model with second-order approximation compared to the other models. The Student-t model with unknown parameter, degrees of freedom, is equivalent the Bayesian model averaging which involves keeping all models (i.e. probit and logit models), but presenting results averaged over these models.