PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 4 – 2012

KAMIL FIJOREK

PRZEDZIAŁ UFNOŚCI *PROFILE LIKELIHOOD* DLA PRAWDOPODOBIEŃSTWA SUKCESU W MODELU REGRESJI LOGISTYCZNEJ FIRTHA[1]

1. WSTĘP

Istotnym elementem teorii bankructwa jest nurt badań ukierunkowany na statystyczne modele przewidywania stopnia zagrożenia upadłością przedsiębiorstwa. Ich znaczenie i przydatność w działalności przedsiębiorstw oraz w prowadzonej przez władze publiczne polityce wzrasta wraz z narastającymi turbulencjami finansowymi oraz przemianami ekonomicznymi. W polskich warunkach modele te ciągle jeszcze nie znajdują właściwego miejsca w porównaniu do krajów o ugruntowanym modelu prywatnej kapitalistycznej gospodarki rynkowej.

Pomimo pojawienia się licznych nowych narzędzi statystycznej analizy klasyfikacji znaczenie modeli dyskryminacyjnych oraz modeli regresji logistycznej nie zmniejsza się, przede wszystkim dzięki ich użyteczności potwierdzonej w praktyce. Dla większości użytkowników metody te na tle metod nowszej generacji, jak chociażby sieci neuronowe, są mniej kosztowne, bardziej przejrzyste, ich wyniki są łatwiejsze do interpretacji i porównań, a co równie istotne rzadko odbywa się to kosztem zdolności predykcyjnych. Za modelem regresji logistycznej w porównaniu do modelu dyskryminacyjnego przemawia brak założeń czynionych w odniesieniu do probabilistycznej natury zmiennych objaśniających oraz bardziej naturalna interpretacja ocen parametrów modelu, wadą jest bardziej złożony proces wyznaczania ocen parametrów modelu.

W świetle poczynionych powyżej uwag w niniejszym artykule rozważane będzie jedynie wykorzystanie modelu regresji logistycznej jako narzędzia służącego do opisu związku pomiędzy wielowymiarowym stanem wskaźników kondycji ekonomiczno-finansowej przedsiębiorstwa, a stopniem zagrożenia przedsiębiorstwa upadłością.

Znacznym problemem stojącym przed badaczami zajmującymi się modelami przewidywania stopnia zagrożenia upadłością przedsiębiorstwa jest trudność w zdobywaniu danych o przedsiębiorstwach, zarówno tych które zbankrutowały jak i tych, które nie

[1] Artykuł powstał jako rezultat badań prowadzonych w projekcie „Instrument Szybkiego Reagowania" (UDA–PO KL.02.01.03–00–021/09–00, www.isr.parp.gov.pl), który jest realizowany w ramach Programu Operacyjnego Kapitał Ludzki współfinansowanego przez Unię Europejską z Europejskiego Funduszu Społecznego. Projekt ISR jest realizowany na zlecenie Ministerstwa Pracy i Polityki Społecznej przez Małopolską Szkołę Administracji Publicznej Uniwersytetu Ekonomicznego w Krakowie w ramach umowy partnerskiej z Polską Agencją Rozwoju Przedsiębiorczości (PARP) w Warszawie.

zbankrutowały. Modele budowane w polskich realiach bardzo rzadko przekraczają barierę 100 przedsiębiorstw. W obliczu małych prób należy dołożyć szczególnych starań, aby dane zostały wykorzystane maksymalnie efektywnie, wnioskowanie było obarczone jak najmniejszym błędem systematycznym, a niepewność ocen parametrów była mierzona rzetelnie. Na gruncie statystyki przekłada się to na postulat estymacji nieobciążonej oraz pośrednio na postulat konstruowania przedziałów ufności utrzymujących nominalne prawdopodobieństwo pokrycia.

Liczne badania pokazują, że w małych próbach estymatory parametrów modelu regresji logistycznej uzyskiwane metodą klasyczną i zarazem najbardziej popularną, tzn. metodą największej wiarygodności (MNW) charakteryzują się znacznym obciążeniem, ponadto klasyczne przedziały ufności oparte na teorii dużej próby rzadko osiągają nominalny poziom ufności (Firth, 1993; Heinze, 2006).

Wspomniane problemy niemal całkowicie eliminuje zastosowanie modelu regresji logistycznej Firtha, który można traktować jako stosunkowo niewielką modyfikację klasycznego modelu regresji logistycznej. Estymacja parametrów w tym modelu jest niemal nieobciążona, co szczególnie widać w bardzo małych próbach, natomiast przedziały ufności charakteryzują się lepszymi właściwościami probabilistycznymi (Firth, 1993; Heinze, 1999; Heinze, Schemper, 2002; Heinze, Ploner, 2004; Heinze, 2006).

Biorąc pod uwagę wymienione zalety modelu Firtha zasadne wydaje się stwierdzenie, że powinien on stać się jednym z podstawowych narzędzi w warsztacie badacza zajmującego się problematyką modelowania zjawiska upadłości przedsiębiorstw. Istotne znaczenie modelu Firtha dla praktyki uzasadnia dalsze badania jego teoretycznych właściwości.

Oszacowany model regresji logistycznej można traktować jako narzędzie, za pomocą którego dokonuje się przekształcenia informacji o kondycji przedsiębiorstwa opisywanej szeregiem wskaźników ekonomiczno-finansowych w liczbę z przedziału od 0 do 1, którą można interpretować jako swoistego rodzaju wskaźnik stopnia zagrożenia upadłością badanego przedsiębiorstwa. Zazwyczaj poprzestaje się na wskazaniu konkretnej wartości tego wskaźnika. Należy jednak mieć na uwadze, że wskaźnik jest jedynie oszacowaniem a tym samym jest z nim związany błąd szacunku, najczęściej tym większy im mniejszy zbiór danych służył estymacji modelu. W rezultacie w kontekście małych prób istotną kwestią staje się rzetelny sposób wyznaczania przedziału ufności dla miary stopnia zagrożenia upadłością.

Niniejszy artykuł ma dwa główne cele. Pierwszym jest cel natury teoretycznej polegający na symulacyjnym zbadaniu właściwości przedziałów ufności wyznaczanych metodą *profile likelihood* (PL) dla prawdopodobieństwa (miara stopnia zagrożenia upadłością) w modelu regresji logistycznej Firtha oraz zaproponowanie efektywnej obliczeniowo metody wyznaczania tychże przedziałów. Cel ten znajduje swe źródło w badaniach Heinze'go (1999), który w wyniku przeprowadzenia symulacji pokazał, że przedziały ufności dla parametrów strukturalnych modelu regresji logistycznej Firtha konstruowane za pomocą metody PL charakteryzują się znacznie lepszymi własnościami w porównaniu do przedziałów asymptotycznych (przedziały ufności Walda).

Jednakże Heinze nie rozważa przedziałów ufności dla prawdopodobieństwa sukcesu. Również systematyczny przegląd literatury nie dostarcza informacji, aby tego rodzaju badania były wykonane. Biorąc pod uwagę pozytywne rezultaty uzyskane dla przedziałów ufności *profile likelihood* dla pojedynczych parametrów modelu można się również spodziewać dobrych rezultatów w przypadku przedziałów ufności dla ich funkcji. Drugim celem jest cel natury praktycznej. Cel ten będzie zrealizowany poprzez zbudowanie przykładowego modelu regresji logistycznej Firtha dla przedsiębiorstw handlowych a następnie na wyznaczeniu przedziałów ufności dla miary zagrożenia upadłością zarówno w ujęciu standardowym (przedziały ufności Walda) jak i w ujęciu proponowanym w części teoretycznej artykułu (przedziały ufności PL).

2. MODEL REGRESJI LOGISTYCZNEJ FIRTHA

Niech zmienna zależna $Y_i \in \{0, 1\}$ $(i = 1, \ldots, n)$ podlega rozkładowi Bernoulliego z prawdopodobieństwem sukcesu równym:

$$P(Y_i = 1) = F\left(x_i'\theta\right) = \frac{1}{1 + exp\left[-x_i'\theta\right]}, \tag{1}$$

gdzie F jest dystrybuantą rozkładu logistycznego, x_i to p-wymiarowy wektor zmiennych objaśniających, a $\theta \in R^p$ to zawierający wyraz wolny wektor parametrów strukturalnych (Long, 1997). W modelu regresji logistycznej Firtha funkcja wiarygodności, nazywana funkcją wiarygodności z karą (*penalized likelihood function*), jest postaci (Heinze, Schemper, 2002; Heinze, 2006):

$$L^*(\theta) = L(\theta)\,|I_\theta|^{1/2}, \tag{2}$$

gdzie

$$L(\theta) = \prod_{i=1}^{n} F(x_i'\theta)^{Y_i}\left[1 - F(x_i'\theta)\right]^{1-Y_i}, \tag{3}$$

$$I_\theta = -\sum\nolimits_{i=1}^{n} \frac{\partial^2 ln\, L_i(\theta)}{\partial\theta\partial\theta'} = X'WX, \tag{4}$$

X to macierz danych o wymiarach $n \times p$, a W jest macierzą diagonalną o wymiarach $n \times n$, której i-ty diagonalny element jest równy $F\left(x_i'\theta\right)(1 - F\left(x_i'\theta\right))$.

W celu oszacowania parametrów modelu oblicza się pochodne cząstkowe logarytmu funkcji wiarygodności $l(\theta)$ względem parametrów modelu:

$$U^*(\theta) = \sum\nolimits_{i=1}^{n}\left(Y_i - F\left(x_i'\theta\right) + h_i\left(\frac{1}{2} - F\left(x_i'\theta\right)\right)\right)x_i, \tag{5}$$

gdzie h_i to diagonalne elementy macierzy $H = W^{\frac{1}{2}}X(X'WX)^{-1}X'W^{\frac{1}{2}}$. Rozwiązanie układu równań $U^*(\theta) = 0$ jest równoważne ze znalezieniem wektora ocen parametrów

maksymalizujących funkcję wiarygodności $L^*(\theta)$. Wektor ocen jest uzyskiwany za pomocą następującej procedury iteracyjnej:

$$\theta^{(k+1)} = \theta^{(k)} + I^{-1}_{\theta^{(k)}} U^* \left(\theta^{(k)}\right), \tag{6}$$

gdzie indeks (k) oznacza k-tą iterację.

2.1. PRZEDZIAŁY UFNOŚCI DLA POJEDYNCZYCH PARAMETRÓW MODELU

Błąd średni szacunku j-tego ($j = 1, \ldots, p$) parametru można wyznaczyć obliczając pierwiastek kwadratowy z elementu leżącego w j-tym wierszu i j-tej kolumnie oceny odwrotnej macierzy informacyjnej $I^{-1}_{\hat{\theta}}$. Wówczas $(1-\alpha)\,\%$ przedział ufności Walda dla θ_j ma postać:

$$\left(\hat{\theta}_j - z_{1-\frac{\alpha}{2}} \sqrt{I^{-1}_{\hat{\theta}jj}};\ \hat{\theta}_j + z_{1-\frac{\alpha}{2}} \sqrt{I^{-1}_{\hat{\theta}jj}}\right), \tag{7}$$

gdzie $z_{1-\frac{\alpha}{2}}$ to kwantyl rzędu $1-\dfrac{\alpha}{2}$ standardowego rozkładu normalnego.

W małych próbach przedziały ufności Walda rzadko osiągają nominalne prawdopodobieństwa pokrycia parametru θ_j. Twierdzi się, że w przypadku małych prób lepszymi właściwościami charakteryzują się przedziały ufności metody *profile likelihood* (Stryhn, Christensen, 2003). Metoda PL polega na odwróceniu testu ilorazu wiarygodności (*likelihood ratio test*) dla parametru będącego przedmiotem zainteresowania. Niech $LR = 2\left[l\left(\hat{\theta}\right) - l\left(\theta_{0j}, \hat{\theta}_{(-j)}\right)\right]$, gdzie $\hat{\theta}$ to wektor ocen parametrów. Dalej przyjmuje się, że $\hat{\theta}_{(-j)}$ są ocenami parametrów (poza j-tym) maksymalizującymi funkcję wiarygodności przy założeniu $\theta_j = \theta_{0j}$. Wówczas $(1-\alpha)\,\%$ przedziałem ufności dla parametru θ_j jest taki zbiór wartości θ_{0j}, dla których wartość statystyki *LR* jest nie większa niż kwantyl rzędu $1-\alpha$ z rozkładu χ^2 o jednym stopniu swobody.

Venzon i Moolgavkar (w skrócie ViM) zaproponowali iteracyjną metodę wyznaczania przedziałów ufności *profile likelihood* dla pojedynczego elementu wektora parametrów strukturalnych modelu, tj. θ_j (Venzon, Moolgavkar, 1988). Procedurę należy powtórzyć osobno dla każdego parametru oraz osobno dla górnej i dolnej granicy przedziału ufności. Procedurę inicjalizuje się poprzez przyjęcie, że θ jest równa wektorowi ocen maksymalizujących funkcję wiarygodności Firtha, tj. $\hat{\theta}$. Następnie niech $l_0 = l\left(\hat{\theta}\right) - \dfrac{1}{2}\chi^2_{1;\,1-\alpha}$ oraz $V = I^{-1}_{\theta}$, wówczas:

$$\lambda = \left(\frac{2\left(l_0 - l(\theta) - \frac{1}{2}\left[U^*(\theta)\right]' V \left[U^*(\theta)\right]\right)}{-\left[e_j\right]' V \left[e_j\right]} \right)^{\frac{1}{2}}, \tag{8}$$

gdzie e_j to wektor jednostkowy z jedynką na j-tej pozycji. Jeżeli celem jest znalezienie górnej granicy przedziału ufności to bieżącą wartość wektora θ należy zaktualizować w następujący sposób:

$$\theta = \theta + V\left(U^*(\theta) + \lambda e_j\right), \tag{9}$$

w przeciwnym razie:

$$\theta = \theta + V\left(U^*(\theta) - \lambda e_j\right). \tag{10}$$

Aktualizację wektora θ należy powtarzać do uzyskania zbieżności.

2.2. PRZEDZIAŁY UFNOŚCI DLA PRAWDOPODOBIEŃSTWA SUKCESU

$(1-\alpha)\,\%$ przedział ufności Walda dla prawdopodobieństwa sukcesu $F(x_i'\theta)$, dla z góry określonego wektora zmiennych objaśniających x_i, ma postać:

$$\left(F\left(x_i'\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{x_i'\left[I_{\hat{\theta}}^{-1}\right]x_i}\right); F\left(x_i'\hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{x_i'\left[I_{\hat{\theta}}^{-1}\right]x_i}\right)\right). \tag{11}$$

DiCiccio i Tibshirani (w skrócie DCiT) zaproponowali wyrafinowane uogólnienie algorytmu ViM, które umożliwia uzyskanie przedziału ufności *profile likelihood* dla dowolnej funkcji parametrów, a więc i w szczególności dla $F(x_i'\theta)$ (DiCiccio, Tibshirani, 1991). W rozważanym jednak przypadku można zauważyć, że nie ma konieczności sięgania po algorytm DCiT, gdyż istnieje przeformułowanie problemu umożliwiające zastosowanie algorytmu mniej złożonego, tj. algorytmu ViM.

W pierwszym kroku proponowanego postępowania budowany jest przedział ufności metodą PL dla liniowej kombinacji parametrów modelu $x_i'\theta$. W tym celu należy zmodyfikować oryginalną macierzy danych X poprzez odjęcie od każdego jej wiersza wektora x_i (kolumna macierzy X odpowiadająca wyrazowi wolnemu pozostaje niezmieniona). Następnie na tak przygotowanych danych należy oszacować model Firtha oraz wyznaczyć przedział ufności *profile likelihood* dla wyrazu wolnego za pomocą algorytmu ViM, przedział ten jest równoważny przedziałowi ufności dla kombinacji liniowej parametrów modelu $x_i'\theta$. W ostatnim kroku w celu uzyskania przedziału ufności dla prawdopodobieństwa sukcesu końce przedziału ufności dla kombinacji liniowej parametrów należy obłożyć funkcją $F(.)$.

3. BADANIA SYMULACYJNE

3.1. ZAŁOŻENIA BADAŃ SYMULACYJNYCH

W badaniach symulacyjnych zmienne niezależne generowano w następujący sposób:

– wariant A – dwie nieskorelowane zmienne ciągłe losowane z rozkładu normalnego N(0,1),

– wariant B – dwie zmienne ciągłe losowane z dwuwymiarowego rozkładu normalnego N(0,1) o współczynniku korelacji równym 0,8,

– wariant C – dwie nieskorelowane zmienne zero-jedynkowe (prawdopodobieństwo „jedynki" przyjęto na poziomie 0,5),

– wariant D – dwie nieskorelowane zmienne ciągłe losowane z rozkładu N(0,1) oraz dwie nieskorelowane zmienne zero-jedynkowe (prawdopodobieństwo „jedynki" przyjęto na poziomie 0,5),

– wariant E – dwie nieskorelowane zmienne ciągłe losowane z rozkładu t-Studenta o 5 stopniach swobody oraz dwie nieskorelowane zmienne zero-jedynkowe (prawdopodobieństwo „jedynki" przyjęto na poziomie 0,25).

W toku symulacji generowano N = 10000 zestawów danych (macierzy danych X) liczących $n = \{50, 100, 150\}$ przypadków (wierszy). Wektor parametrów modelu dla wariantów A, B, C jest postaci $\theta = (0; 0,5; 0,5)$ oraz $\theta = (0; 1; 1)$ natomiast dla wariantów D, E jest postaci $\theta = (0; 0,5; 0,5; 0,5; 0,5)$ oraz $\theta = (0; 1; 1; 1; 1)$. Wygenerowane zestawy danych oraz zdefiniowane parametry modelu stanowiły podstawę do wylosowania zmiennej zależnej $y_i (i = 1, \ldots, n)$. Dla i-tego przypadku y_i stanowiło losową realizację z rozkładu Bernoulliego, w którym prawdopodobieństwo sukcesu było równe $F\left(x_i'\theta\right)$.

Symulację przeprowadzono w środowisku obliczeń statystycznych R (R Development Core Team, 2011). Oceny parametrów modelu regresji logistycznej Firtha oraz odpowiednie przedziały ufności uzyskiwano za pomocą zmodyfikowanej przez autora biblioteki *logistf* (Heinze, Ploner, 2004; Fijorek, Sokołowski, 2012). Jakość estymacji przedziałowej określano osobno dla przedziałów Walda oraz przedziałów metody *profile likelihood*. Empiryczna wartość prawdopodobieństwa pokrycia była określana poprzez wylosowanie po jednej obserwacji z N kolejnych macierzy danych, wyznaczenie dla nich przedziałów ufności obu typów oraz sprawdzenie czy prawdziwe wartości prawdopodobieństw sukcesu należą do przedziałów ufności. Współczynnik ufności został przyjęty na poziomie 95%.

3.2. WYNIKI BADAŃ SYMULACYJNYCH

Wyniki przeprowadzonych badań symulacyjnych zestawiono w tabeli 1. Pierwszym wnioskiem, który można wyciągnąć z analizy wyników symulacji jest naturalna i w pełni oczekiwana obserwacja, że im większa liczba przypadków tym prawdopodobieństwo pokrycia jest bliższe poziomowi nominalnemu dla obu typów przedziałów ufności. Drugim spostrzeżeniem jest to, że przedziały ufności metody PL osiągają prawdopodobieństwo pokrycia znacznie bliższe poziomowi nominalnemu w porównaniu do przedziałów Walda niemal we wszystkich rozważonych scenariuszach symulacyjnych. Średnie odchylenie od poziomu nominalnego, który wynosił 95%, dla przedziałów Walda wyniosło niemal 1%, natomiast dla przedziałów PL około 0,3%, czyli w przybliżeniu trzykrotnie mniej. Odchylenia te są jednak znacznie mniejsze w porównaniu do odchyleń zaobserwowanych przez Heinze'go (1999) w przypadku przedziałów ufności obu typów dla pojedynczych parametrów modelu.

Z praktycznego punktu widzenia można powiedzieć, że wyniki symulacji nie dyskwalifikują przedziałów Walda. Jednak w przypadku, gdy rozważane są bardzo małe próby a probabilistyczne właściwości stosowanych metod statystycznych mają spełniać

najwyższe standardy zaleca się stosowanie przedziałów ufności metody PL, pomimo ich znacznej złożoności obliczeniowej.

Tabela 1.

Prawdopodobieństwo pokrycia przez przedział ufności prawdopodobieństwa sukcesu

Wariant	**n**	**Wald**	**PL**	**Wald**	**PL**
		$\theta = (0;\ 0,5;\ 0,5)$ $\theta = (0; 0,5; 0,5; 0,5; 0,5)$		$\theta = (0;\ 1;\ 1)$ $\theta = (0; 1; 1; 1; 1)$	
A	50	0,9684	0,9565	0,9643	0,9533
	100	0,9603	0,9533	0,9560	0,9508
	150	0,9572	0,9522	0,9546	0,9503
B	50	0,9688	0,9559	0,9508	0,9523
	100	0,9583	0,9524	0,9534	0,9557
	150	0,9529	0,9496	0,9515	0,9493
C	50	0,9696	0,9544	0,9709	0,9560
	100	0,9595	0,9507	0,9587	0,9504
	150	0,9568	0,9510	0,9577	0,9549
D	50	0,9720	0,9536	0,9573	0,9526
	100	0,9631	0,9537	0,9617	0,9590
	150	0,9545	0,9492	0,9541	0,9500
E	50	0,9752	0,9561	0,9480	0,9581
	100	0,9624	0,9530	0,9455	0,9493
	150	0,9550	0,9492	0,9536	0,9551

Źródło: Opracowanie własne.

4. BUDOWA MODELU PRZEWIDYWANIA STOPNIA ZAGROŻENIA UPADŁOŚCIĄ PRZEDSIĘBIORSTWA HANDLOWEGO

W dalszej części artykułu zaprezentowano proces budowania przykładowego modelu zagrożenia upadłością przedsiębiorstwa handlowego jako etap pośredni w celu zademonstrowania praktycznego znaczenia rezultatów uzyskanych w części teoretycznej artykułu. Z tego względu niektóre aspekty procesu modelowania zostały pominięte lub opisane w sposób skrótowy. Bardziej dogłębne ujęcie tematyki zawiera artykuł Fijorek, Grotowski (2012).

4.1. ZMIENNE OPISUJĄCE KONDYCJĘ EKONOMICZNO-FINANSOWĄ PRZEDSIĘBIORSTWA

Do budowy modelu predykcji zagrożenia upadłością wykorzystano wskaźniki, które opisują w sposób syntetyczny, ale równocześnie wielopłaszczyznowy, stan i wyniki ekonomiczno-finansowe przedsiębiorstwa. Wskaźniki te opisują zachowanie się przedsiębiorstwa w obszarach produktywności, płynności, finansowania, rentowności, zadłużenia oraz wydajności. Ich doboru dokonano na podstawie analiz, studiów literaturowych oraz wiedzy merytorycznej. Zmienne objaśniające w przypadku przedsiębiorstwa, które upadło opisują kondycję ekonomiczno-finansową przedsiębiorstwa na jeden rok przed ogłoszeniem upadłości.

4.2. ZDEFINIOWANIE ZBIORU UCZĄCEGO DLA MODELU PRZEWIDYWANIA STOPNIA ZAGROŻENIA UPADŁOŚCIĄ PRZEDSIĘBIORSTWA HANDLOWEGO

W standardowym przypadku, tworzenie zbioru danych na potrzeby budowy modeli predykcji sprowadza się do pobrania z populacji próby losowej jednostek statystycznych. Następnie dla każdej jednostki określana jest jej przynależność do jednej z uprzednio zdefiniowanych klas. W przypadku predykcji upadłości takie podejście jest rzadko spotykane. Częściej stosowanym postępowaniem jest zgromadzenie zbioru danych o przedsiębiorstwach upadłych, a następnie dobranie do nich przedsiębiorstw, które nie upadły. W niewielkich zbiorach danych stosuje się zazwyczaj dobieranie przedsiębiorstw upadłych do nieupadłych oparte na wiedzy eksperckiej oraz wnikliwej analizie każdej pojedynczej obserwacji.

Podstawową statystyczną metodą „parowania" obiektów posiadających wyróżnioną cechę do obiektów nie posiadających takiej cechy jest technika *case-control*. Polega ona na określeniu kilku kluczowych charakterystyk jednostek statystycznych oraz dopasowaniu do każdej jednostki posiadającej wyróżnioną cechę jednostki bez takiej cechy, która jest do niej najbardziej podobna ze względu na zmienne służące do parowania. W artykule przyjęto, że każdemu przedsiębiorstwu upadłemu będą towarzyszyć przedsiębiorstwa nieupadłe podobne pod względem sumy bilansowej oraz przychodów netto ze sprzedaży. Parowane przedsiębiorstwa będą mieć zgodne dwie pierwsze cyfry symbolu rodzaju działalności według PKD (poziom działów) oraz formę prawno-organizacyjną. Ponadto dane ekonomiczno-finansowe przedsiębiorstwa upadłego oraz dopasowanych do niego przedsiębiorstw nieupadłych będą pochodzić z tego samego roku. W praktycznych zastosowaniach najczęściej wykorzystuje się parowanie „1 do 1", lecz z teoretycznego punktu widzenia uzasadnione jest parowanie nawet „1 do 5" (Hosmer, Lemeshow, 1989). W artykule przyjęto, że do każdego przedsiębiorstwa upadłego zostanie dobranych pięć przedsiębiorstw nieupadłych.

W rezultacie wykonanych prac zgromadzony materiał liczbowy wstępnie obejmował informacje dotyczące około 15 tys. przedsiębiorstw nieupadłych oraz około 2 tys. przedsiębiorstw upadłych. Odpowiednie dane, które posłużyły do stworzenia zbioru uczącego, zgromadzono z publicznie dostępnych baz danych o przedsiębiorstwach. W wyniku zawężenia zbioru danych tylko do przedsiębiorstw handlowych,

w następstwie eliminacji przedsiębiorstw z niekompletnymi danymi oraz po uwzględnieniu kryteriów dobierania przedsiębiorstw upadłych do nieupadłych końcowy zbiór uczący przedsiębiorstw handlowych liczył 84 przedsiębiorstwa upadłe oraz 405 przedsiębiorstw nieupadłych (nie do każdego przedsiębiorstwa upadłego możliwe było dobranie dokładnie 5 przedsiębiorstw nieupadłych). W rezultacie jest to jeden z większych, choć obiektywnie rzecz ujmując nadal niewielki, przedstawionych w polskiej literaturze przedmiotu zbiór danych o przedsiębiorstwach handlowych rozważany w kontekście modelowania stopnia zagrożenia upadłością.

4.3. ESTYMACJA MODELU REGRESJI LOGISTYCZNEJ FIRTHA

W pierwszym etapie prac dokonano analizy jednowymiarowych rozkładów zmiennych objaśniających oraz analizy korelacji zachodzących pomiędzy nimi. W wyniku tej procedury zidentyfikowano obserwacje odstające oraz grupy zmiennych silnie skorelowanych. W drugim etapie oszacowano parametry modelu regresji logistycznej Firtha opisującego zależność pomiędzy kondycją ekonomiczno-finansową przedsiębiorstwa handlowego a stopniem zagrożenia upadłością. W celu określenia optymalnego zbioru zmiennych objaśniających tworzących model regresji logistycznej wykorzystano metodę najlepszego podzbioru (Fijorek, Fijorek, 2011). W wyniku zastosowania opisanej procedury otrzymano następujące równanie służące do wyznaczania wskaźnika zagrożenia upadłością (WZU) przedsiębiorstwa handlowego:

$$\mathrm{WZU} = \frac{1}{1 + exp\left[-(-0,27 - 0,29\,Z_1 - 0,39\,Z_2 + 0,27\,Z_3 - 0,75\,Z_4)\right]}. \tag{12}$$

Wskaźnik WZU przyjmuje wartości z przedziału (0,1), przy czym wyższe jego wartości wskazują na wyższe zagrożenie upadłością. Ujemna (dodatnia) wartość współczynnika odpowiadającego określonej zmiennej objaśniającej oznacza, że jej wzrost przekłada się na spadek (wzrost) stopnia zagrożenia upadłością.

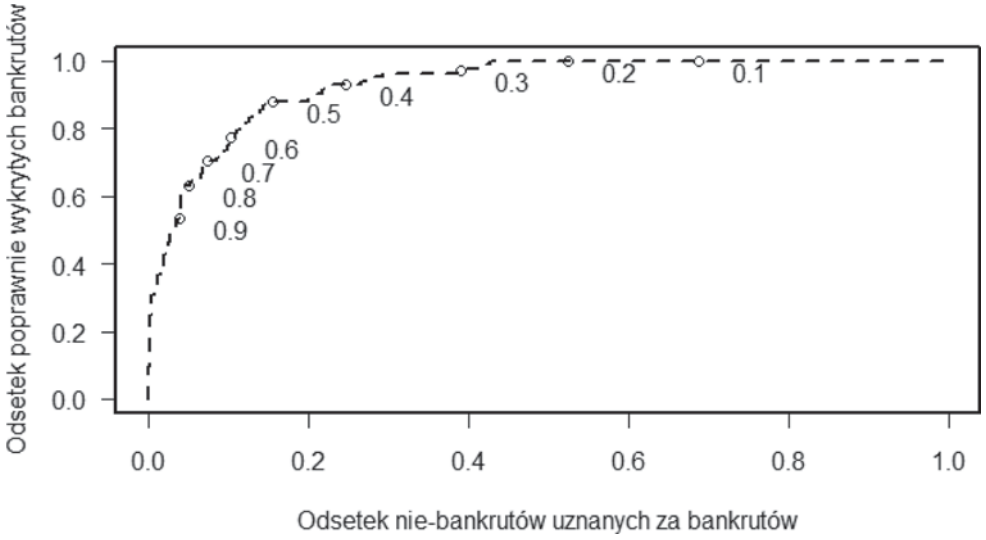
4.4. ZDEFINIOWANIE KLAS ZAGROŻENIA UPADŁOŚCIĄ ORAZ OKREŚLENIE SPRAWNOŚCI WYKRYWANIA STANU ZAGROŻENIA UPADŁOŚCIĄ

Optymalny punkt odcięcia dla wskaźnika zagrożenia upadłością wyznaczono za pomocą krzywej ROC (*Receiver Operating Characteristic*). Krzywa ROC to dwuwymiarowy wykres, który prezentuje *czułość* (odsetek bankrutów uznanych za bankrutów) oraz *1 – specyficzność* (odsetek nie-bankrutów uznanych za bankrutów), obliczone dla różnych wartości punktu odcięcia. W rezultacie przyjęto następującą regułę definiującą przynależność przedsiębiorstwa handlowego do klasy przedsiębiorstw zagrożonych upadłością:

$$WZU > 0,5 \rightarrow \textit{klasa wysokiego zagrożenia upadłością}.$$

Zaprezentowana na rysunku 1 krzywa ROC ukazuje zachowanie się reguły decyzyjnej w przypadku przyjęcia innych wartości punktu odcięcia. Generalną regułą jest to, że

im niższy punkt odcięcia tym więcej wykrywanych jest bankrutów, lecz odbywa się to kosztem uznawania coraz większej liczby nie-bankrutów za bankrutów.



Rysunek 1. Krzywa ROC – przedsiębiorstwa handlowe

Zdolności predykcyjne modelu regresji logistycznej Firtha zostały zmierzone za pomocą *czułości* oraz *specyficzności* (Fijorek, Fijorek, Wiśniowska, Polak, 2011). Pierwsza z tych miar dla oszacowanego modelu wynosi 88,1%, podczas gdy druga 84,7%. Biorąc pod uwagę wartości miar należy stwierdzić, że model charakteryzuje się wysokimi zdolnościami przewidywania stanu zagrożenia upadłością przedsiębiorstwa handlowego.

4.5. OKREŚLENIE NIEPEWNOŚCI ZWIĄZANEJ Z SZACOWANIEM STOPNIA ZAGROŻENIA UPADŁOŚCIĄ

Rolę każdego wskaźnika opisującego kondycję ekonomiczno-finansową przedsiębiorstwa w kształtowaniu stopnia zagrożenia upadłością zbadano w ramach analizy scenariuszowej. Założenia analizy przedstawiono w tabeli 2 natomiast jej wyniki zilustrowano na rysunkach 2-5. W celu zapewnienia porównywalności wyników w ramach danego scenariusza manipulowano jedynie wartością pojedynczego wskaźnika ekonomiczno-finansowego, podczas gdy wartości pozostałych wskaźników pozostawały na ustalonym poziomie.

Krzywe umieszczone „centralnie" na wykresach 2-5 oznaczają poziom wskaźnika zagrożenia upadłością. Na podstawie analizy ich przebiegu można powiedzieć, że rola wskaźnika Z_2 oraz Z_3 w kształtowaniu WZU jest znacznie większa niż rola pozostałych dwóch wskaźników. Krzywe umieszczone powyżej i poniżej krzywej centralnej stanowią odpowiednio górne i dolne granice 95% przedziału ufności dla WZU. Li-

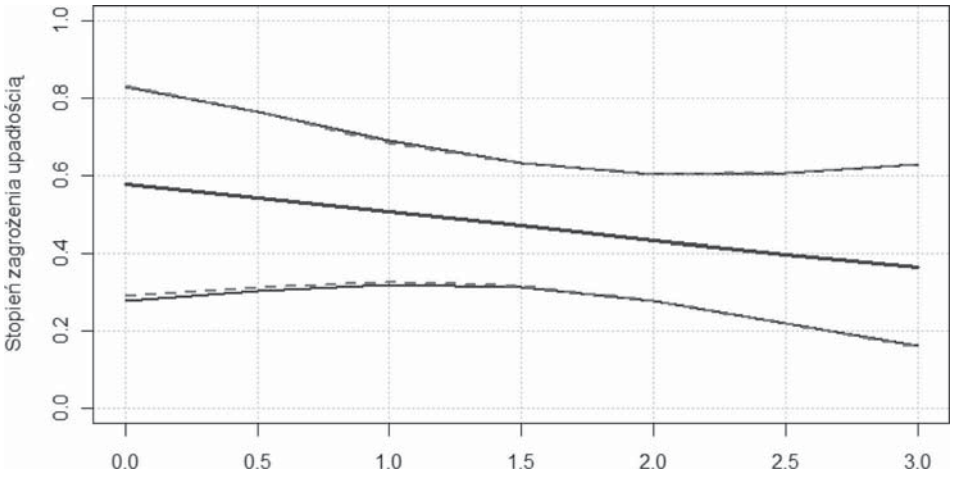
Tabela 2.
Założenia analizy scenariuszowej dotyczącej wpływu wskaźników ekonomiczno-finansowych na stopnień zagrożenia upadłością

Numer scenariusza	**Numer wykresu**	**Wartość wskaźnika finansowego**			
		Z_1	**Z_2**	**Z_3**	**Z_4**
1	2	od 0 do 3	2	5	0
2	3	1	od 0 do 8	5	0
3	4	1	2	od 1 do 10	0
4	5	1	2	5	od -1 do 1

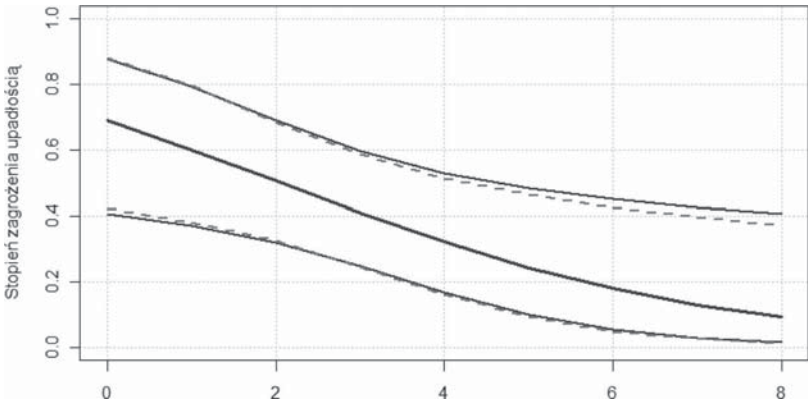
Źródło: Opracowanie własne.

nią ciągłą oznaczono przedziały ufności Walda natomiast linią przerywaną przedziały uzyskane metodą *profile likelihood.* Skonstruowane przedziały ufności ukazują bardzo dużą niepewność związaną z szacowanym wskaźnikiem zagrożenia upadłością. Należy przypuszczać, że w mniejszych próbach, tak często spotykanych w polskich modelach zagrożenia upadłością, niepewność szacunków WZU będzie na jeszcze wyższym poziomie.

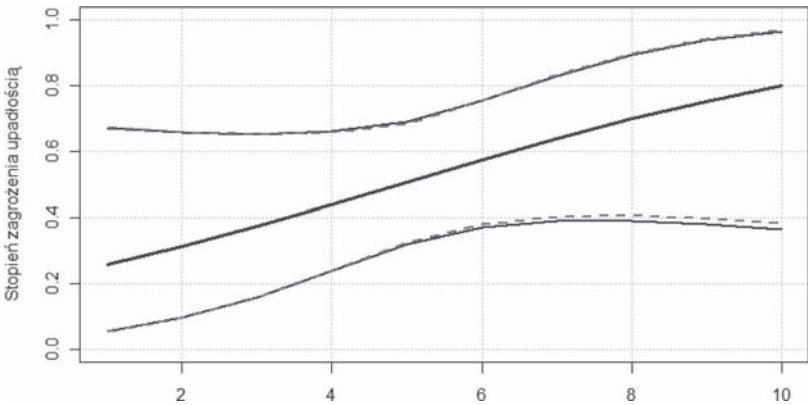
Kolejnym wnioskiem płynącym z analizy wykresów jest praktyczny brak różnic pomiędzy przedziałami ufności Walda oraz przedziałami uzyskiwanymi metodą *profile likelihood*. Tym samym w celu uzyskania przedziału ufności dla wskaźnika zagrożenia upadłością w rozważanym przykładzie należy zalecać stosowanie przedziałów Walda jako metody znacznie mniej złożonej obliczeniowo.



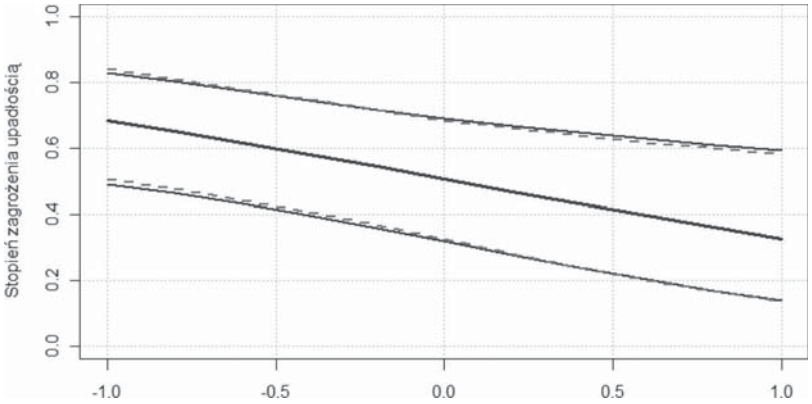
Rysunek 2. Wskaźnik stopnia zagrożenia upadłością w funkcji wartości wskaźnika Z_1



Rysunek 3. Wskaźnik stopnia zagrożenia upadłością w funkcji wartości wskaźnika Z_2



Rysunek 4. Wskaźnik stopnia zagrożenia upadłością w funkcji wartości wskaźnika Z_3



Rysunek 5. Wskaźnik stopnia zagrożenia upadłością w funkcji wartości wskaźnika Z_4

5. PODSUMOWANIE

W pierwszej części artykułu za pomocą symulacji zbadano właściwości przedziałów ufności Walda oraz przedziałów ufności wyznaczanych metodą *profile likelihood* budowanych dla wskaźnika zagrożenia upadłością w modelu regresji logistycznej Firtha. W wyniku symulacji stwierdzono, że różnice pomiędzy obydwoma typami przedziałów ufności chociaż są zauważalne to jednak nie są duże. Wniosek ten dodatkowo potwierdziły badania na rzeczywistym zbiorze danych omówionym w drugiej części artykułu.

Analiza rzeczywistego zbioru danych ponadto dostarczyła istotnego z punktu widzenia praktyki wniosku, tzn. skonstruowane przedziały ufności ukazały niepokojąco dużą niepewność związaną z szacowanym wskaźnikiem zagrożenia upadłością. Należy przypuszczać, że w mniejszych próbach, tak często spotykanych w polskich modelach zagrożenia upadłością, niepewność szacunków miar zagrożenia znajduje się na jeszcze wyższym poziomie.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] DiCiccio T., Tibshirani R., (1991), *Technical Report No. 9107: On the Implementation of Profile Likelihood*, Department of Statistics, University of Toronto.
[2] Fijorek K., Fijorek D., (2011), *Dobór zmiennych objaśniających metodą najlepszego podzbioru do modelu regresji logistycznej Firtha*, Metody Informatyki Stosowanej, 2, 15-23.
[3] Fijorek K., Fijorek D., Wiśniowska B., Polak S., (2011), *BDTcomparator: A Program for Comparing Binary Classifiers*, Bioinformatics, 27 (24), 3439-3440.
[4] Fijorek K., Grotowski M., (2012), *Bankruptcy Prediction: Some Results From a Large Sample of Polish Companies*, International Business Research, 5 (9).
[5] Fijorek K., Sokołowski A., (2012), *Separation-Resistant and Bias-Reduced Logistic Regression: STATISTICA macro*, Journal of Statistical Software, 47, 1-12.
[6] Firth D., (1993), *Bias Reduction of Maximum Likelihood Estimates*, Biometrika, 80, 27-38.
[7] Heinze G., (1999), *Technical Report 10: The Application of Firth's Procedure to Cox and Logistic Regression*, Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna.
[8] Heinze G., (2006), *A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data*, Statistics in Medicine, 25, 4216-4226.
[9] Heinze G., Ploner M., (2004), *Technical Report 2/2004: A SAS Macro, S-PLUS Library and R Package to Perform Logistic Regression without Convergence Problems*, Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna.
[10] Heinze G., Schemper M., (2002), *A Solution to the Problem of Separation in Logistic Regression*, Statistics in Medicine, 21, 2409-2419.
[11] Hosmer D.W., Lemeshow S., (1989), *Applied Logistic Regression*, Wiley, New York.
[12] Long J.S., (1997), *Regression Models for Categorical and Limited Dependent Variables*, SAGE.
[13] R Development Core Team, (2011), R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

[14] Stryhn H., Christensen J., (2003), *Confidence Intervals by the Profile Likelihood Method, with Applications in Veterinary Epidemiology*, ISVEE X, Chile.
[15] Venzon D.J., Moolgavkar S.H., (1988), *A Method for Computing Profile-Likelihood Based Confidence Intervals*, Applied Statistics, 37, 87-94.

PRZEDZIAŁ UFNOŚCI *PROFILE LIKELIHOOD* DLA PRAWDOPODOBIEŃSTWA SUKCESU W MODELU REGRESJI LOGISTYCZNEJ FIRTHA

Streszczenie

W pierwszej części artykułu za pomocą symulacji zbadano właściwości przedziałów ufności Walda oraz przedziałów ufności wyznaczanych metodą *profile likelihood* (zaproponowano również efektywny algorytm wyznaczania tychże przedziałów) budowanych dla prawdopodobieństwa sukcesu w modelu regresji logistycznej Firtha. W drugiej części artykułu zaprezentowano przykładowy model zagrożenia upadłością przedsiębiorstwa handlowego jako etap pośredni w celu zademonstrowania praktycznego znaczenia rezultatów uzyskanych w części teoretycznej artykułu.

Słowa kluczowe: regresja logistyczna, przedziały ufności, metoda *profile likelihood*

PROFILE LIKELIHOOD CONFIDENCE INTERVAL FOR THE PROBABILITY OF A SUCCESS IN THE FIRTH'S LOGISTIC REGRESSION

Abstract

In the first part of the paper the results of the simulation study, comparing the coverage properties of Wald's and the profile likelihood confidence intervals for the probability of a success in the Firth's logistic regression, are described. The efficient algorithm for computing profile likelihood confidence intervals is proposed. In the second part of the paper the theoretical results are applied to the bankruptcy model.

Key words: logistic regression, confidence intervals, profile likelihood