EKONOMETRIA
22

PRACE NAUKOWE nr 27
Uniwersytetu Ekonomicznego
we Wrocławiu

EKONOMETRIA
22

Zastosowania metod ilościowych

Redaktor naukowy
Józef Dziechciarz

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2008

PL ISSN 1899-3192
PL ISSN 1507-3866

Druk i oprawa: Zakład Graficzny UE we Wrocławiu. Zam. 56/2009

Spis treści

Summaries

Wstęp

W kolejnym numerze zeszytu Prac Naukowych „Ekonometria" zamieszczono dziewięć artykułów. Danuta Strahl prezentuje rozważania na temat wykorzystania klasyfikacji pozycyjnej w analizach dynamicznych. Prace Alicji Grześkowiak, Marcina Pełki oraz Piotra Michalskiego mają charakter prac metodyczno-teoretycznych z zakresu statystyki i ekonometrii. Aneta Rybicka zajęła się badaniami preferencji konsumentów za pomocą metod dekompozycyjnych. Kolejne prace, Agnieszki Przybylskiej-Mazur, Grzegorza Michalskiego oraz Marcina Wojciela, dotyczą problemów związanych z analizą rynku ubezpieczeń oraz analizą danych finansowych.

Wrocław, grudzień 2008 *Józef Dziechciarz*

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 2008
Ekonometria 22

Danuta Strahl

KLASYFIKACJA POZYCYJNA W ANALIZACH DYNAMICZNYCH

1. Wstęp

Wśród wielu zadań i celów analiz ekonomicznych znajduje się ocena zmian tendencji rozwoju i wzajemnych relacji badanych obiektów. W badaniach regionalnych zachodzi z kolei potrzeba łączenia kryteriów oceny uwzględniających wymiar dynamiczny oraz przestrzenny. Wielowymiarowa analiza danych zawiera zdecydowanie więcej propozycji dla badań opartych wyłącznie na danych przekrojowych lub czasowych, a znacznie mniej opartych na danych przekrojowo-czasowych czy też przestrzenno-czasowych (por. [1; 6; 7; 8; 12]). Do cennych narzędzi opisu i identyfikacji szczególnych właściwości zjawisk ekonomicznych należą metody klasyfikacji. Metody te są licznie prezentowane w literaturze przedmiotu, ale na ogół ograniczają się do badań statycznych. Jest też kilka propozycji pokazujących możliwości wykorzystania metod klasyfikacji w badaniach dynamicznych (por. [5; 9; 11]). Zasadniczym celem tego artykułu jest opracowanie procedury klasyfikacji obiektów z wykorzystaniem statystyk pozycyjnych w badaniach dynamicznych.

2. Podstawy formalne

Dany jest zbiór obiektów $P = \{P_1, P_2, \ldots, P_n, \ldots, P_N\}$ opisany zbiorem m zmiennych, oznaczonych symbolami $X = \{X_1, X_2, \ldots, X_m\}$. Zakładamy, że w zbiorze zmiennych znajdują się wyłącznie zmienne o charakterze stymulant (por. [10; 11; 12]). Kiedy pojawiają się zmienne o charakterze destymulant lub nominant, należy stosować znane formuły przekształceń ich na stymulanty (por. [10, 11, 12]).

Wartości cech obserwowane są w momentach czasowych $t = 1, 2, \ldots, T$. Każdy obiekt zatem może być opisany za pomocą macierzy o postaci:

$$\mathbf{P}_k^t = \begin{bmatrix} x_{k_1}^1 & x_{k_2}^1 & \dots & x_{k_m}^1 \\ x_{k_1}^2 & x_{k_2}^2 & \dots & x_{k_m}^2 \\ \vdots & \vdots & & \vdots \\ x_{k_1}^t & x_{k_2}^t & \dots & x_{k_m}^t \\ \vdots & \vdots & & \vdots \\ x_{k_1}^T & x_{k_2}^T & \dots & x_{k_m}^T \end{bmatrix}_{T \times m}, \quad j = 1,2,\dots,m,\ k = 1,2,\dots,K,\ t = 1,2,\dots,T, \tag{1}$$

gdzie: $x_{k_j}^t$ – wartość j-tej zmiennej $(j = 1,2,\dots,m)$ w k-tym obiekcie badania $(k = 1,2,\dots,K)$ w $t = 1,2,\dots,T$ momencie obserwacji.

Zbiór obiektów $P = \{P_1, P_2, \dots, P_K\}$ może być opisany macierzą blokową o postaci:

$$\mathbf{P}: = \begin{bmatrix} x_{11}^1 & x_{12}^1 & & x_{1m}^1 \\ x_{11}^2 & x_{12}^2 & & x_{1m}^2 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & x_{1j}^t & \vdots \\ \vdots & \vdots & & \vdots \\ x_{11}^T & x_{12}^T & \dots & x_{1m}^T \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & x_{k_j}^t & \vdots \\ \vdots & \vdots & & \vdots \\ x_{k_1}^T & x_{k_2}^T & & x_{k_m}^T \\ \vdots & \vdots & & \vdots \\ x_{K_1}^1 & x_{K_2}^1 & \dots & x_{K_m}^1 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & x_{K_j}^t & \vdots \\ x_{K_1}^T & x_{K_2}^T & \dots & x_{K_m}^T \end{bmatrix}_{K \times m \times T}, \tag{2}$$

gdzie: $x_{k_j}^t$ – wartość j-tej cechy w k-tym obiekcie w t-tym momencie obserwacji $(t = 1,2,\dots,T)$.

3. Klasyfikacja pozycyjna z medianą

Etap I procedury klasyfikacji

Klasyfikację obiektów $P = \{P_1, P_2, \ldots, P_K\}$ przeprowadzamy dla każdego momentu obserwacji $t = 1, 2, \ldots, T$ według następujących kroków.

1) Dla każdej zmiennej X dla każdego momentu $t = 1, 2, \ldots, T$ wyznaczamy medianę według jednego ze wzorów (por. [2]):

$$Me(X_j) = \frac{x_{kj}^{i=K/2} + x_k^{i=K/2+1}}{2} \quad \text{dla parzystej liczby badanych obiektów,} \tag{3}$$

$$Me(X_j) = \frac{x_{kj}^{i=K\cdot m/2} + x_k^{i=K\cdot m/2+1}}{2} \quad \text{dla nieparzystej liczby badanych obiektów.} \tag{4}$$

2) Powstaje wektor median o postaci:

$$\left[Me(X_1)\ \ Me(X_2)\ \ \ldots\ \ Me(X_m)\right]. \tag{5}$$

3) Klasyfikacja obiektów w momentach $t = 1, 2, \ldots, T$.

Proponowana procedura klasyfikacji uwzględnia dwa przypadki. W przypadku pierwszym algorytm klasyfikacji prowadzi do budowy dla każdego momentu $t = 1, 2, \ldots, T$ $(m + 1)$ klas oznaczonych symbolem S_g, gdzie $g = 1, 2, \ldots, G$ $(G = m + 1)$, gdy zbiory opisane są za pomocą m zmiennych. W przypadku drugim algorytm klasyfikacji prowadzi do budowy 2^m (czyli $G = 2^m$) klas możliwych kombinacji z m zmiennych dla każdego momentu $t = 1, 2, \ldots, T$.

Przypadek pierwszy

Krok 1. Do klasy S_1^t (dla $t = 1, 2, \ldots, T$) wchodzą obiekty ze zbioru P, których wartości wszystkich zmiennych X_j^t, czyli m zmiennych, są wyższe (korzystniejsze) od zadanej statystyki pozycyjnej lub jej równe. W naszych rozważaniach przyjmiemy, że statystyką tą jest mediana (Me). Stąd:

$$\bigwedge_j : x_{kj}^t \geq Me\left(X_j^t\right), \tag{6}$$

gdzie: $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, T$.

Krok 2. Do klasy S_2^t wchodzą obiekty ze zbioru P, których wartości tylko $(m - 1)$ zmiennych spełniają warunek:

$$x_{kj}^{1t} \geq Me\left(X_j^t\right). \tag{7}$$

Krok *m*. Do klasy S_g^t $(g = m)$ wchodzą obiekty ze zbioru P, których tylko wartość jednej zmiennej X_j^t ze zbioru X spełnia warunek (7).

Krok (*m* + 1). Do klasy S_{g+1}^t $(g = m + 1)$ wchodzą obiekty P, których wartość x_{kj}^t żadnej zmiennej X_j^t nie spełnia warunku (7).

Przypadek drugi

Krok 1. Klasę S_1^t (dla $t = 1, 2, \ldots, T$) tworzą te obiekty, których wartości wszystkich m zmiennych X_j^t spełniają warunek:

$$x_{kj}^t \geq Me\left(X_j^t\right), \text{ gdzie: } j = 1, 2, \ldots, m. \tag{8}$$

Krok 2. Klasę S_2^t tworzą te obiekty, których wartości jedynie $(m - 1)$ zmiennych tworzących jedną z kombinacji $\binom{m}{m-1}$ zmiennych spełniają warunek (8).

Krok 3. Klasę trzecią S_3^t tworzą te obiekty, których wartości zmiennych kolejnej kombinacji $(m - 1)$-elementowej spełniają warunek (8).

Krok 4. Po wyczerpaniu kombinacji $(m - 1)$-elementowych tworzymy klasy dla kombinacji $(m - 2)$-elementowych i stawiamy warunek (8).

Krok 2^m. Klasę S_g^t $(g = 2^m)$ tworzymy z obiektów, dla których wartości x_{kj}^t wszystkich zmiennych X_j^t nie spełniają warunku (8).

Jak widać, oba przypadki mają wyraźnie odmienne założenia klasyfikacyjne. W przypadku pierwszym przypisujemy identyczne znaczenie wszystkim zmiennym, rozróżniając jedynie klasy obiektów poprzez liczbę zmiennych spełniających zadane warunki. Natomiast w drugim przypadku rozróżniamy grupy obiektów poprzez identyfikację specyfikacji zmiennych spełniających zadane warunki klasyfikacji.

Procedurę klasyfikacji powtarzamy dla każdego momentu $t = 1, 2, \ldots, T$.

4. Klasyfikacja dynamiczna z dominantą

Etap II procedury klasyfikacji

1) Dla każdego obiektu P_k $(k = 1, 2, \ldots, K)$ budujemy wektor, którego elementami są częstości przynależności danego obiektu do poszczególnych klas określonych w I etapie klasyfikacji w badanych momentach $t = 1, 2, \ldots, T$.

2) W wyniku tego pomiaru powstaje macierz, odpowiednio o wymiarach: $(m+1)\times K$ dla przypadku pierwszego oraz $2^m \times K$ dla przypadku drugiego o postaci:

$$\left[a_{g_k}\right]_{(m+1)\times K} \text{ lub } \left[a_{g_k}\right]_{2^m\times K}, \tag{9}$$

gdzie: a_{g_k} – częstość występowania k-tego regionu ($k = 1, 2, \ldots, K$) w t momentach badania ($t = 1, 2, \ldots, T$) w g-tej klasie ($g = 1, 2, \ldots, (m+1)$ lub $g = 1, 2, \ldots, 2^m$.

3) Dla każdego obiektu P_k wyznaczamy dominantę z wektora: $\left[a_{g_k}\right]_{1\times g}$.

4) Klasy dynamiczne (a więc zawierające obiekty obserwowane we wszystkich okresach $t = 1, 2, \ldots, T$) tworzymy w następujących krokach.

Przypadek pierwszy

Krok 1. Do klasy D_1 wchodzą obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie pierwszej.

Krok 2. Do klasy D_2 wchodzą obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie drugiej.

Krok *m*. Do klasy D_g (dla $g = m$) wchodzą obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie m-tej.

Krok (m + 1). Do klasy D_g (dla $g = m + 1$) wchodzą obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie $m + 1$.

Przypadek drugi

Krok 1. Klasę D_1 tworzą te obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie pierwszej.

Krok 2. Klasę D_2 tworzą te obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie drugiej.

Krok 3. Klasę trzecią D_3 tworzą te obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie trzeciej.

Po wyczerpaniu kombinacji $(m-1)$-elementowych tworzymy klasy dla kombinacji $(m-2)$-elementowych.

Krok 2^m. Klasę D_g $(g = 2^m)$ tworzą te obiekty ze zbioru P, dla których dominanta wartości przynależności obiektów do klas utworzonych w I etapie klasyfikacji w momentach $t = 1, 2, \ldots, T$ znalazła się w klasie 2^m.

Pozostają jeszcze do omówienia przypadki szczególne. Jak wiadomo, może na przykład zaistnieć sytuacja, w której brakuje dominanty. Jeżeli zatem częstość występowania obiektów w określonych klasach jest identyczna, przypisujemy dany obiekt do klasy z ostatniego okresu badania.

5. Ilustracja proponowanej procedury

Obiektami badania będą regiony szczebla NUTS-2 w Wielkiej Brytanii, stąd: $P = 1, 2, \ldots, 37$. Obiekty te zostały opisane następującymi cechami:

X_1 – udział pracujących z wyższym wykształceniem w ogólnej liczbie pracujących w regionie,

X_2 – kapitał ludzki w nauce i technologii (*HRST*) jako odsetek aktywnych zawodowo,

X_3 – udział ludności w wieku 25-64 lata uczestniczącej w kształceniu ustawicznym w regionie.

Wartości cech były obserwowane w pięciu latach, tj. w 2001, 2002, 2003, 2004, 2005, stąd $t = 1, 2, 3, 4, 5$.

Zgodnie z procedurą klasyfikacji (według przypadku drugiego) utworzono osiem klas obiektów:

- Klasa I obejmuje obiekty – regiony, dla których wartości wszystkich zmiennych X_1, X_2, X_3 są wyższe od mediany.
- Klasa II obejmuje obiekty – regiony, dla których wartości cech X_1 i X_2 są wyższe od mediany, a wartości cechy X_3 są niższe od mediany.
- Klasa III obejmuje obiekty – regiony, dla których wartości cech X_1 i X_3 są wyższe od mediany, a wartości cechy X_2 są niższe od mediany.
- Klasa IV obejmuje obiekty – regiony, dla których wartości cech X_2 i X_3 są wyższe od mediany, a wartości cechy X_1 są niższe od mediany.
- Klasa V obejmuje obiekty – regiony, dla których wartość cechy X_1 jest wyższa od mediany, a wartości dwóch cech, tj. X_2 oraz X_3, są niższe od mediany.
- Klasa VI obejmuje obiekty – regiony, dla których wartość cechy X_2 jest wyższa od mediany, a wartości dwóch cech, tj. X_2 oraz X_3, są niższe od mediany.
- Klasa VII obejmuje obiekty – regiony, dla których wartość cechy X_3 jest wyższa od mediany, a wartości dwóch cech, tj. X_1 oraz X_2, są niższe od mediany.

- Klasa VIII obejmuje obiekty – regiony, dla których wartości wszystkich cech są niższe od mediany.

Tabela 1. Częstość przynależności badanych regionów do ośmiu klas w latach 2001-2005

Obiekt	1	2	3	4	5	6	7	8	Klasa w dynamicznej klasyfikacji
1. Tees Valley and Durham						5			6
2. Northumberland, Tyne and Wear	1	2				2			6*
3. Cumbria				1		1	1	2	8
4. Cheshire	1	1		2		1			4
5. Greater Manchester	1	1	1		2				5
6. Lancashire				2		3			6
7. Merseyside		1	1		3				5
8. East Riding and North Lincolnshire							1	4	8
9. North Yorkshire			4		1				3
10. South Yorkshire			1		1		2	1	7
11. West Yorkshire			4		1				3
12. Derbyshire and Nottinghamshire				1		4			6
13. Leicestershire, Rutland and Northants				2		3			6
14. Lincolnshire				1		3		1	6
15. Herefordshire, Worcestershire and Warks		1		2		2			6*
16. Shropshire and Staffordshire		1		3		1			4
17. West Midlands	1	2				2			2*
18. East Anglia			1	1	2			1	5
19. Bedfordshire, Hertfordshire	5								1
20. Essex	3				2				1
21. Inner London			5						3
22. Outer London			5						3
23. Berkshire, Bucks and Oxfordshire	5								1
24. Surrey, East and West Sussex	1		4						3
25. Hampshire and Isle of Wight	2			3					4
26. Kent			2		3				5
27. Gloucestershire, Wiltshire and North Somerset	5								1
28. Dorset and Somerset	2			1			2		4*
29. Cornwall and Isles of Scilly				1			1	3	8
30. Devon	1					1	3		7
31. West Wales and The Valleys						2		3	8
32. East Wales	2	1			1	1			1
33. North Eastern Scotland							1	4	8
34. Eastern Scotland			5						3
35. South Western Scotland		2			3				5
36. Highlands and Islands					1		2	2	8*
37. Northern Ireland								5	8

* Klasa z ostatniego okresu badania.

Źródło: obliczenia własne na podstawie danych Eurostatu.

Ze względu na ograniczone ramy artykułu nie podano wyników klasyfikacji dla każdego roku, a w tab. 1 podano tylko częstość występowania każdego regionu w poszczególnych klasach w każdym badanym roku.

Klasyfikacja dynamiczna (za lata 2002-2005) przynosi podział regionów brytyjskich ze względu na przyjęte do analizy cechy, który podano w tab. 2.

Tabela 2. Klasyfikacja regionów Wielkiej Brytanii w ujęciu dynamicznym

Klasa/liczba regionów	$X_1 \geq Me$	$X_2 \geq Me$	$X_3 \geq Me$	Regiony
1/5	+	+	+	1. Bedfordshire, Hertfordshire 2. Essex 3. Berkshire, Bucks and Oxfordshire 4. Gloucestershire, Wiltshire and North Somerset 5. East Wales
2/1	+	+	–	1. West Midlands
3/6	+	–	+	1. North Yorkshire 2. West Yorkshire 3. Inner London 4. Outer London 5. Surrey, East and West Sussex 6. Eastern Scotland
4/4	–	+	+	1. Cheshire 2. Shropshire and Staffordshire 3. Hampshire and Isle of Wight 4. Dorset and Somerset
5/5	+	–	–	1. Greater Manchester, 2.Merseyside 3. East Anglia 4. Kent 5. South Western Scotland
6/7	–	+	–	1. Tees Valley and Durham 2. Northumberland, Tyne and Wear 3. Lancashire 4. Derbyshire and Nottinghamshire 5. Leicestershire, Rutland and Northants 6. Lincolnshire 7. Herefordshire, Worcestershire and Warks
7/2	–	–	+	1. South Yorkshire 2. Devon
8/7	–	–	–	1. Cumbria 2. East Riding and North Lincolnshire 3. Cornwall and Isles of Scilly 4. West Wales and The Valleys 5. North Eastern Scotland 6. Highlands and Islands 7. Northern Ireland

Źródło: obliczenia własne na podstawie danych Eurostatu.

6. Zakończenie

Przedstawiona propozycja klasyfikacji obiektów badania ma charakter wartościujący oraz uwzględnia pozycje obiektów w zadanym okresie, co ma szczególne znaczenie w badaniach dynamicznych. Klasyfikacja wykorzystująca statystyki pozycyjne, w tym medianę i dominantę, pozwala na podział obiektów uwzględniający tendencje relacji, jakie zachodzą między obiektami badania w zadanym przedziale czasowym.

Literatura

[1] Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.
[2] Luszniewicz A., Slaby T., *Statystyka stosowana*, PWE, Warszawa 1998.
[3] Milasewic P., Ducharme G.R., *Uniqueness of the Spatial Median*, „The Annals of Statistics" 1987, vol. 15, no. 3.
[4] Młodak A., *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa 2006.
[5] *Poziom życia w Polsce i krajach UE*, red. A. Zeliaś, PWE, Warszawa 2004.
[6] Strahl D., *Strukturalna miara rozwoju obiektów hierarchicznych*, [w:] Ekonometria 16, red. J. Dziechciarz, Prace Naukowe Akademii Ekonomicznej nr 1100, AE, Wrocław 2006.
[7] Strahl D., *Klasyfikacja regionów z medianą*, [w:] Ekonometria 10, red. J. Dziechciarz, Prace Naukowe Akademii Ekonomicznej nr 950, AE, Wrocław 2002.
[8] Strahl D., *Miejsce Polski w regionalnej przestrzeni UE*, [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, red. A. Zeliaś, AE, Kraków 2005.
[9] Strahl D., Markowska M., *Klasyfikacja dynamiczno-przestrzenna obiektów hierarchicznych z wykorzystaniem statystyk pozycyjnych*, AE, Wrocław (złożone do druku).
[10] *Taksonomia struktur w badaniach regionalnych*, red. D. Strahl, AE, Wrocław 1998.
[11] *Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym*, red. A. Zeliaś, AE, Kraków 2000.
[12] Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, AE, Wrocław 2002.

POSITIONAL CLASSIFICATION IN DYNAMIC ANALYSES

Summary

The article presents the proposal of the studied objects classification in a dynamic perspective. The procedure of objects division is divided into two stages. In the first stage the classification of objects is performed by means of positional statistics. The basic division criterion becomes the evaluation of attributes relations which describe the studied objects up to the median value. The procedure is repeated for each observation moment of values referring to the studied objects' values. The second stage consists in the studied objects classification with reference to positions occupied by

objects in particular moments of the study in the first stage of the classification. The classification criterion in the second stage becomes the frequency dominant of the studied objects' classes, defined in the first stage of each moment of the study.

Danuta Strahl – prof. zw. dr hab., kierownik Katedry Gospodarki Regionalnej Uniwersytetu Ekonomicznego we Wrocławiu – Wydział w Jeleniej Górze.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Alicja Grześkowiak

POŁĄCZENIE KLASYCZNEGO I NIEKLASYCZNEGO PODEJŚCIA W CELU WYKRYWANIA NIEJEDNORODNOŚCI WARIANCJI SKŁADNIKÓW LOSOWYCH MODELU EKONOMETRYCZNEGO

1. Wstęp

Klasyczna metoda najmniejszych kwadratów, stosowana powszechnie do estymacji parametrów liniowych modeli ekonometrycznych, wymaga spełnienia szeregu założeń, dotyczących między innymi rozkładu składników losowych modelu. Przy określonych założeniach estymator metody najmniejszych kwadratów posiada pożądane własności – jest nieobciążony, zgodny i najefektywniejszy w klasie nieobciążonych estymatorów liniowych. Weryfikacja założeń dotyczących składników losowych wiąże się z licznymi trudnościami, wynikającymi z tego, że są one nieobserwowalne. Wnioskowanie o składnikach losowych przeprowadza się za pomocą testów statystycznych na podstawie reszt otrzymanych w wyniku zastosowania metody najmniejszych kwadratów. Z wielu względów nie stanowią one idealnego narzędzia umożliwiającego podjęcie właściwej decyzji co do postaci rozkładu składników losowych. W niniejszej pracy przedstawiono koncepcję połączenia zalet postępowania klasycznego (stosowania testów statystycznych) z korzyściami płynącymi z możliwości wyboru „najlepszego" rozkładu przy zastosowaniu dyskryminacyjnej reguły największej wiarygodności. Problemem, którego rozwiązanie zaproponowano, jest określanie rodzaju heteroskedastyczności składników losowych jednorównaniowego liniowego modelu ekonometrycznego. W pracy przedyskutowano postulowaną koncepcję i zilustrowano ją badaniem empirycznym.

2. Klasyczne podejście do wykrywania niejednorodności wariancji składników losowych modelu ekonometrycznego

Stosując metodę najmniejszych kwadratów do szacowania parametrów modelu ekonometrycznego postaci:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_m X_m + \varepsilon, \tag{1}$$

gdzie: Y – zmienna objaśniana,
$X_1, X_2, \ldots, X_m$ – zmienne objaśniające,
$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m$ – parametry strukturalne,
ε – składniki losowe.

czyni się założenia dotyczące struktury stochastycznej modelu. Zespół założeń dotyczących składników losowych jest następujący:

1. Wartość oczekiwana wynosi zero: $E(\varepsilon_i)$ dla każdego i. U podstaw tego założenia leży przekonanie, iż zakłócenia reprezentowane przez składniki losowe wzajemnie się redukują.

2. Wariancja jest stała: $V(\varepsilon_i) = \sigma^2$ dla każdego i. Stałość wariancji nosi miano homoskedastyczności.

3. Brak autokorelacji, czyli brak skorelowania składników losowych pochodzących z dwóch różnych obserwacji: $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$.

W przypadku wystąpienia różnych wariancji lub istnienia autokorelacji składników losowych estymator metody najmniejszych kwadratów przestaje być estymatorem najefektywniejszym. Preferowaną wówczas metodą szacowania parametrów modelu powinna być uogólniona metoda najmniejszych kwadratów (UMNK), której estymator posiadający pożądane własności podał Aitken. Mimo tego znanego sposobu przezwyciężania trudności związanych z niespełnieniem założeń o składnikach losowych, praktyczne zastosowanie UMNK napotyka wiele trudności, wymaga bowiem wiedzy o kształtowaniu się składników losowych, której badacz raczej nie posiada *a priori*. Nie ma też możliwości, aby bezpośrednio sprawdzić przebiegu procesu, co wynika z samej natury składników losowych, które są nieobserwowalnymi zmiennymi losowymi. Wobec tak silnych ograniczeń poznawczych wnioskowanie o składnikach losowych oparte jest na pewnym ich substytucie, jakim są reszty modelu, które nie będąc realizacjami ε, stanowią tylko pewne ich przybliżenie.

W trakcie procesu modelowania ekonometrycznego badacz zmuszony jest więc do formułowania stwierdzeń dotyczących składników losowych, dysponując bardzo ograniczonym zasobem informacji. Podczas weryfikacji modelu w zakresie badania jednorodności wariancji bardzo ważne wydaje się rozwiązanie następujących problemów:

1. Czy składniki losowe spełniają założenie o homoskedastyczności?
2. Jeżeli występuje w modelu heteroskedastyczność, to jaki jest jej charakter?

Rozstrzygnięcie kwestii zawartej w pierwszym pytaniu pozwala na trafny wybór metody estymacji, natomiast znalezienie odpowiedzi na pytanie drugie otwiera możliwość zastosowania UMNK do szacowania parametrów modelu.

Idealne narzędzie badawcze niosłoby odpowiedzi na oba pytania, umożliwiając tym samym aplikację UMNK w sytuacjach, gdy zostanie stwierdzona niejednorodność składników losowych. Rozwiązanie tych kwestii nie jest jednakże tak proste. W zakresie wykrywania heteroskedastyczności prace ekonometryków (zob. [1; 2; 5; 6; 8; 11; 13]) skupiają się na konstrukcji testów statystycznych i dotychczas nie opracowano skutecznego narzędzia, które jednocześnie wykrywałoby na dużym poziomie ogólności odstępstwo od założenia o homoskedastyczności i postulowało schemat opisujący kształtowanie się wariancji. Proponowane rozwiązania pozwalają albo przebadać możliwość wystąpienia heteroskedastyczności określonego typu, albo stwierdzić istnienie niejednorodności wariancji bez wskazówek, jak się ona kształtuje. W pierwszym przypadku badanie jest bardzo fragmentaryczne, w drugim zaś bardzo ogólne. W literaturze przedmiotu dzieli się testy wykrywające heteroskedastyczność na dwie kategorie:

- testy konstruktywne, w których w hipotezie alternatywnej zdefiniowany jest konkretny model heteroskedastyczności (zob. [4; 10; 13]) – np. test Goldfelda-Quandta, test Bartletta, test Ramseya, test Glejsera,
- testy niekonstruktywne, w których w hipotezie alternatywnej nie specyfikuje się charakteru heteroskedastyczności, tylko ogólnie stwierdza jej występowanie – (zob. [12; 13; 14]) – np. testy klasy Szroetera, nieparametryczny test szczytów Goldfelda-Quandta, test White'a.

Do niewątpliwych zalet testu konstruktywnego można zaliczyć to, że przyjęcie hipotezy alternatywnej oznacza zaakceptowanie wyspecyfikowanego w niej schematu heteroskedastyczności i możliwość zastosowania UMNK. Sytuacja komplikuje się natomiast w przypadku braku podstaw do odrzucenia hipotezy zerowej, gdyż wówczas badacz otrzymuje jedynie informację, że w modelu nie wystąpił określony rodzaj heteroskedastyczności, co nie oznacza, że wariancje są jednorodne – istnieje możliwość, że kształtują się według innego wzorca niż zdefiniowany w zastosowanym teście.

Z kolei zastosowanie testu niekonstruktywnego przezwycięża tę niedogodność. Zastosowanie tego rodzaju postępowania pozwala na przyjęcie jednego ze stwierdzeń: albo składniki modelu są heteroskedastyczne (jeśli odrzucona zostanie hipoteza zerowa na rzecz alternatywnej), albo charakteryzują się homoskedastycznością (jeśli nie ma podstaw do odrzucenia hipotezy zerowej). Grupa testów niekonstruktywnych odznacza się więc dużo większym stopniem ogólności, pozwalając wykryć istnienie niejednorodności wariancji. Niestety, na podstawie wyniku testu nie

można określić, jaki rodzaj schematu heteroskedastyczności wystąpił w modelu, co uniemożliwia zastosowanie UMNK do estymacji parametrów modelu.

3. Kombinacja podejścia klasycznego i dyskryminacyjnej reguły największej wiarygodności umożliwiająca określanie rodzaju heteroskedastyczności

Niedostatki testów statystycznych w zakresie wykrywania istnienia i określania typu heteroskedastyczności mogą zostać zniwelowane poprzez zastosowanie kombinacji podejścia klasycznego z dyskryminacyjną regułą największej wiarygodności, której zastosowanie nie wymaga ani wiedzy o prawdopodobieństwach *a priori* przynależności do klas, ani informacji o kosztach błędnej alokacji. Tego rodzaju niewiedza pojawia się w praktyce modelowania bardzo często, gdyż brakuje przesłanek pozwalających określić prawdopodobieństwa *a priori* wystąpienia poszczególnych schematów heteroskedastyczności, a także niezmiernie trudno sformułować postulaty odnoszące się do kosztów błędnej alokacji. Dlatego też w przedstawianej koncepcji przyjmuje się występowanie równych strat i równych prawdopodobieństw *a priori*, zgodnie z postulatem Bayesa zakładającym, że w przypadku braku informacji przeczącej wszystkie prawdopodobieństwa *a priori* są takie same.

Sugerowana procedura postępowania obejmuje dwa główne etapy:

Etap I: zastosowanie testu niekonstruktywnego, ogólnego, pozwalającego na wykrycie zjawiska niejednorodności wariancji składników losowych modelu ekonometrycznego.

W przypadku stwierdzenia heteroskedastyczności niezbędne staje się określenie jej rodzaju; cel ten jest osiągany poprzez:

Etap II: wykorzystanie dyskryminacyjnej reguły największej wiarygodności do wyboru najlepszego schematu opisującego kształtowanie się wariancji składników losowych.

Jeśli spełnione są założenia dotyczące składników losowych, ich macierz kowariancji jest macierzą skalarną postaci $\sigma^2\mathbf{I}$ ($\mathbf{I}$ – macierz jednostkowa stopnia n). Jeżeli występuje heteroskedastyczność, elementy diagonalne nie są jednakowe, a macierz kowariancji ma wówczas postać $\sigma^2\mathbf{V}$. Przy założeniu normalności rozkładu funkcja gęstości wektora składników losowych ma postać:

$$f(\varepsilon) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|\sigma^2\mathbf{V}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\varepsilon^{\mathrm{T}}\left(\sigma^2\mathbf{V}\right)^{-1}\varepsilon\right\}. \qquad (2)$$

Po zastosowaniu testu niekonstruktywnego nie jest znany rodzaj heteroskedastyczności, nie są więc znane elementy macierzy $\mathbf{V}$. W literaturze można się zetknąć z sugestią, iż wówczas określenie typu heteroskedastyczności zależy od

badacza, który decyzję swoją powinien oprzeć na doświadczeniu i merytorycznej ocenie modelowanych zjawisk. Postulowane rozwiązanie przeciwstawia się tak subiektywnej metodzie wyboru i zmierza ku bardziej obiektywnemu kryterium. Jeśli hipoteza o homoskedastyczności zostanie odrzucona, można zaproponować różne modele opisujące kształtowanie się wariancji i wybrać spośród nich najbardziej adekwatny, stosując dyskryminacyjną regułę największej wiarygodności.

Każdemu z g zaproponowanych możliwych typów heteroskedastyczności odpowiada określona postać macierzy **V**, której elementy można otrzymać na podstawie schematu obrazującego kształtowanie się wariancji, np. uznając, iż wariancja zależy od wartości bezwzględnej pewnej zmiennej objaśniającej X_i, z dokładnością do stałej σ^2:

$$\sigma_j^2 = \sigma^2 \left| X_{ij} \right|, \tag{3}$$

otrzymuje się macierz kowariancji postaci:

$$\sigma^2 \mathbf{V} = \sigma^2 \begin{bmatrix} |X_{i1}| & 0 & \dots & 0 \\ 0 & |X_{i2}| & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & |X_{in}| \end{bmatrix}. \tag{4}$$

W przypadku bardziej skomplikowanych schematów heteroskedastyczności zachodzi potrzeba estymacji ich parametrów i wyznaczenie wartości teoretycznych (oszacowań wariancji). W procesie szacowania stosuje się wartości bezwzględne lub kwadraty reszt uzyskanych przy szacowaniu parametrów modelu metodą najmniejszych kwadratów.

Po ustaleniu g postaci macierzy **V** dla g branych pod uwagę typów heteroskedastyczności można określić g funkcji wiarygodności postaci (2), odpowiadających poszczególnym schematom. Selekcja najlepszego z nich może zostać dokonana przy użyciu metod analizy dyskryminacyjnej, gdyż zadanie sprowadza się do wyboru jednego spośród skończonej i z góry znanej liczby g ($g \geq 2$) rodzajów heteroskedastyczności. Dla każdej z rozpatrywanych możliwości istnieje n-wymiarowy rozkład wektora składników losowych ε. Każdy wariant ($i = 1, 2, \dots, g$) jest scharakteryzowany poprzez funkcję gęstości $f_i(\varepsilon)$, a właściwie wiarygodności $L_i(\varepsilon)$, gdyż wartość jest ustalana przy różnych znanych parametrach rozkładu (elementach macierzy **V**).

Dyskryminacyjna reguła największej wiarygodności przydziela wektor ε do tej populacji, dla której funkcja wiarygodności $L_i(\varepsilon)$ ($i = 1, 2, \dots, g$) jest maksymalna, czyli wybrany zostaje schemat heteroskedastyzności, dla którego funkcja określona wzorem (2) osiąga największą wartość wśród wszystkich g zaproponowanych

możliwości (por. [7; 9]). Ze względu na monotoniczność funkcji logarytmicznej wnioski można formułować również na podstawie wartości $\ln L_i(\varepsilon)$.

4. Zastosowanie postulowanej procedury w badaniu empirycznym

Przedstawiona metoda łącząca zalety testu niekonstruktywnego z korzyściami wynikającymi z zastosowania analizy dyskryminacyjnej została wykorzystana do określania rodzaju heteroskedastyczności w jednorównaniowym modelu ekonometrycznym służącym do opisu relacji pomiędzy oczekiwanymi (Y) a osiąganymi przez respondentów dochodami (X). Prezentowane badanie empiryczne oparte jest na obszernym zbiorze obserwacji pochodzącym z Polskiego Generalnego Sondażu Społecznego[1]. Odpowiadając na pytania ankietera, respondenci wypowiadali się m.in. w kwestiach dotyczących swojej pracy zarobkowej i sytuacji materialnej. Analizie zostały poddane odpowiedzi na następujące pytania zawarte w kwestionariuszu:

Pytanie nr 32: *Biorąc pod uwagę ... rok, proszę powiedzieć, ile wynoszą Pana(-i) przeciętne miesięczne zarobki (dochody) pochodzące z pracy po odjęciu podatków?*

Pytanie nr 34: *A jak Pan(-i) sądzi, na jakie miesięczne wynagrodzenie (dochód z pracy) Pan(i) zasługuje?*

Polski Generalny Sondaż Społeczny przeprowadzany był ośmiokrotnie, począwszy od 1992 r. do 2005 r. Modelowaniu poddano dane dotyczące badania najwcześniejszego oraz najpóźniejszego (1992 r. oraz 2005 r.). W przypadku niekompletnych odpowiedzi respondentów usunięto obserwacje z brakującymi elementami i otrzymano następujące zbiory danych: dla roku 1992 o liczebności 674, dla roku 2005 o liczebności 459. Metodą najmniejszych kwadratów oszacowano parametry modeli liniowych opisujących zależność oczekiwanych dochodów od dochodów osiąganych. W celu sprawdzenia, czy w modelach występuje zjawisko niejednorodności wariancji, zastosowano test White'a bazujący na tzw. modelu pomocniczym, którego parametry estymuje się, korzystając z kwadratów reszt otrzymanych z modelu podstawowego. Zespół hipotez jest następujący (zob. [3, s. 88]):

$H_0: \beta_k = 0$ parametry modelu pomocniczego są równe zeru, czyli wariancja składników losowych modelu podstawowego jest jednorodna,

$H_0: \beta_k \neq 0$ przynajmniej jeden parametr modelu pomocniczego jest różny od zera, czyli wariancja składników losowych modelu podstawowego jest niejednorodna.

[1] Bogdan Cichomski (kierownik programu), Tomasz Jerzyński i Marcin Zieliński. Polskie Generalne Sondaże Społeczne: skumulowany komputerowy zbiór danych 1992-2002. Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa 2003.

Model pomocniczy ma postać:

$$\sigma_{el}^2 = \beta_0 + \sum_{k=1}^{K} \beta_k x_{kl} + \sum_{\substack{k=K+1 \\ i,j=1}} \beta_k x_{i1} x_{j1} + h_i,\ i, j = 1, 2, \ldots, K, \tag{5}$$

gdzie: $\beta_0, \beta_1, \ldots, \beta_K$ – parametry modelu pomocniczego,
h_1 – składnik losowy modelu pomocniczego,
K – liczba zmiennych objaśniających w modelu pomocniczym.

Statystyka służąca weryfikacji hipotez ma postać $W = nR^2$, gdzie R^2 oznacza współczynnik determinacji obliczony dla modelu pomocniczego. Przy prawdziwości hipotezy zerowej statystyka W ma rozkład χ^2 z K stopniami swobody.

Oszacowania parametrów oraz rezultaty testu White'a dla analizowanych modeli przedstawiono w tab. 1.

Tabela 1. Rezultaty estymacji modeli oraz wyniki testu White'a (poziom istotności 0,05)

Rok	Wynik estymacji (MNK) parametrów modelu podstawowego	Wartość statystyki W	Wartość krytyczna testu	Wniosek
1992	$\hat{Y} = \underset{(0,04)}{1,26}\ X + \underset{(12,94)}{102.11}$	21,09	5,99	występuje heteroskedastyczność składników losowych
2005	$\hat{Y} = \underset{(0,04)}{1,27}\ X + \underset{(67,48)}{480.21}$	20,25		występuje heteroskedastyczność składników losowych

Źródło: obliczenia własne.

W obydwu rozpatrywanych modelach wystąpiło zjawisko niejednorodności wariancji składników losowych, właściwą metodą estymacji jest więc uogólniona metoda najmniejszych kwadratów. Ponieważ test White'a nie daje podstaw do sformułowania wniosków o mechanizmie kształtującym wariancje, zaproponowano następujące modele heteroskedastyczności:

Model 1: $\sigma_i^2 = \sigma^2 \left| X_{ij} \right|$.

Model 2: $\sigma_i^2 = \sigma^2 X_{ij}^2$.

Model 3: $\sigma_i = \alpha_0 + \alpha_1 X$.

Model 4: $\sigma_i^2 = \alpha_0 + \alpha_1 X$.

Model 5: $\sigma_i = \alpha_0 + \alpha_1 X + \alpha_2 X^2$.

Model 6: $\sigma_i^2 = \alpha_0 + \alpha_1 X + \alpha_2 X^2$.

Model 7: $\sigma_i^2 = \sigma_0^2 X^{\alpha}$.

Model 8: $\sigma_i^2 = \sigma_0^2 \exp(\alpha X)$.

Po estymacji parametrów sugerowanych modeli metodą najmniejszych kwadratów obliczono wartości teoretyczne, które zastosowano do budowy macierzy **V**.

Oszacowane w ten sposób macierze kowariancji dały podstawę do zastosowania dyskryminacyjnej reguły największej wiarygodności. Ponieważ rozpatrywane zbiory obserwacji były duże, wyznaczając wartości funkcji wiarygodności postaci (2) dla modeli 1-8, przybliżono wektor składników losowych modelu za pomocą wektora reszt. Najlepszym z proponowanych typów heteroskedastyczności jest ten, dla którego wartość funkcji wiarygodności (lub jej logartymu) jest największa. Otrzymane rezultaty zaprezentowano w tab. 2.

Tabela 2. Ocena proponowanych schematów heteroskedastyczności za pomocą logarytmu funkcji wiarygodności

Rok	Wartości logarytmu funkcji wiarygodności							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1992	–67 169,6	**–3779,8**	–4689,5	–5411,3	–4686,1	–4530,4	–5891,8	–6042,6
2005	–180 014,9	–4337,7	–3907,2	–3878,3	–3957,2	**–3761,9**	–5177,6	–4949,4

Tłustym drukiem zaznaczono modele najlepsze spośród zaproponowanych.

Źródło: obliczenia własne.

W przypadku modelu skonstruowanego dla roku 1992 adekwatny jest prosty typ heteroskedastyczności (model 2), w którym wariancja uzależniona jest od kwadratu zmiennej objaśniającej X (przeskalowanej o σ^2). Mechanizm kształtowania się wariancji dla relacji w roku 2005 jest bardziej skomplikowany, model opisujący ma charakter paraboli. Wartości logarytmów funkcji wiarygodności pozwalają również na uporządkowanie proponowanych specyfikacji. W obu przypadkach okazało się, że najgorsze, i znacznie odbiegające *in minus* od pozostałych, jest rozwiązanie najprostsze, uzależniające wartości wariancji bezpośrednio od wartości zmiennej objaśniającej X (model 1).

Po ustaleniu typu heteroskedastyczności możliwe było zastosowanie UMNK do szacowania parametrów modeli. Ostateczne wyniki estymacji są następujące:

- dla roku 1992: $\hat{Y} = \underset{(0.03)}{1{,}30}\ X + \underset{(30{,}98)}{61{,}30}$,
- dla roku 2005: $\hat{Y} = \underset{(0{,}03)}{1{,}25}\ X + \underset{(77{,}81)}{501{,}66}$.

Uwzględnienie faktu występowania niejednorodności wariancji i zastosowanie UMNK prowadzi do nieco odmiennych rezultatów niż aplikacja MNK (por. tab. 1). Współczynnik przy zmiennej X w modelu dla roku 1992 jest większy niż analogiczny współczynnik dla roku 2005, co oznacza, że wraz ze wzrostem dochodów oczekiwania co do wysokości zarobków wzrastały nieco szybciej w 1992 r. Stosowanie MNK prowadzi do przeciwnego wniosku i mniej wyraźnej różnicy. Stosowanie procedury zawierającej niekonstruktywny test heteroskedastyczności oraz dyskryminacyjnej reguły największej wiarygodności stwarza możliwość otrzymania bardziej precyzyjnych wyników.

Literatura

[1] Adijbolosoo S. B-S. K., *Estimation of Parameters of Heteroscedastic Error Models Under Various Hypothesized Error Structures*, „The Statistician" 1993, vol. 42, s. 123-133.
[2] Davidson R., MacKinnon J., *Estimation and Inference in Econometrics*, Oxford University Press 2003.
[3] *Ekonometria współczesna*, red. M. Osińska, Toruń 2007.
[4] Goldfeld S.M., Quandt R.E., *Some Tests for Homoscedasticity*, „Journal of the American Statistical Association" 1965, vol. 60, s. 539-547.
[5] Greene W.H., *Econometric Analysis*, Prentice Hall, Upper Saddle River, New Jersey 1997.
[6] Hayashi F., *Econometrics*, Princeton University Press, Princeton N.J 2000.
[7] Jajuga K., *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa 1990.
[8] Maddala G.S., *Ekonometria*, PWN, Warszawa 2006.
[9] Mardia K.V., Kent J.T., Bibby J.M., *Multivariate Analysis*, Academic Press, London 1979.
[10] Ramsey J.B., *Test for Specification Error in Classical Linear Least-Squares Regression Analysis*, „Journal of the Royal Statistical Society, Ser. B" 1969, vol. 31, s. 350-371.
[11] Ruud P.A., *An Introduction to Classical Econometric Theory*, Oxford University Press, Oxford 2000.
[12] Szroeter J., *A Class of Parametric Tests for Heteroscedasticity in Linear Econometric Models*, „Econometrica" 1977, vol. 46, s. 1311-1327.
[13] Tomaszewicz A., *Jednorównaniowe modele ekonometryczne przy nieklasycznych założeniach*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 1985.
[14] White H., *A Heteroscedasticity – Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity*, „Econometrica" 1980, vol. 48, s. 817-838.

A COMBINATION OF CLASSICAL AND NON-CLASSICAL APPROACH TO DISTURBANCE HETEROSCEDASTICITY DETECTION IN THE ECONOMETRIC MODEL

In this paper a concept of heteroscedasticity character determination in the linear econometric model is presented. The suggested method combines classical approach (hypothesis testing) with the maximum likelihood discriminant rule. The suggested solution is illustrated by the empirical analysis of models describing real and expected incomes of Poles.

Alicja Grześkowiak – dr, adiunkt w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 2008
Ekonometria 22

Marcin Pełka

THE APPLICATION OF SYMBOLIC KERNEL DISCRIMINANT ANALYSIS IN CREDIT RATING

1. Introduction

The symbolic data analysis is an extension of multivariate analysis dealing with data represented in an extended form. Each symbolic variable can contain single quantitative value, categorical value, interval, multivalued variable, and multivalued variable with weights. Besides that symbolic variables can also be taxonomic, hierarchically dependent, and logically dependent. Therefore symbolic data analysis introduces new methods and implements classical methods, where symbolic data is treated as an input. First part of this article presents aims of discriminant analysis with special focus on the non-parametric kernel density estimation method. Second part introduces terms of symbolic objects and symbolic variable. Third part shows how Bayesian discrimination rule can be adapted to deal with data of different symbolic types, using kernel intensity measures for symbolic data [1, pp. 240-242]. The last part of the article presents results of discrimination analysis for symbolic objects in credit rating and compares its results with credit decision made by a credit officer.

2. Discriminant analysis and kernel density estimation

Discriminant analysis assigns objects from test set to an existing structure of classes (training set).

We usually can't make any assumptions concerning density function of data in real life discrimination problems. To solve this problem we can [2, p. 132]:

a) approximate the unknown density function by applying one of well-known density functions as its estimator,

b) apply one of twelve functions proposed by Pearson and solve differential equation (see [6]),

c) estimate unknown density function with non-parametric methods.

One of the most commonly used non-parametric methods of an estimation of distribution density function is kernel density estimation (see: [7, p. 170]). Equation (1) represents general form of kernel density estimator [1, p. 239; 8, p. 27]:

$$\hat{f}_k(z) = \frac{1}{n_k (2h_k)^d} \sum_{i=1}^{n_k} K\left(\frac{z - x_{ki}}{h_k}\right), \; z \in R^d, \tag{1}$$

where: $\hat{f}_k(z)$ – uniform kernel density estimator for object z in the k-th class,
$k = 1, 2, \ldots, g$ – number of classes,
n_k – number of objects in k-th class,
h_k – bandwidth window for k-th class (a parameter),
x_{ki} – i-th object in k-th class,
d – dimension equal to number of variables describing object,
$K\left(\frac{z - x_{ki}}{h_k}\right)$ – uniform kernel.

Uniform kernel can take various forms (see [2, p. 134]). In the simplest case its value is equal 1 if all coordinates of its arguments are smaller than 1, in other cases its value is equal to 0.

3. Symbolic objects and variables

Symbolic data unlike classical data situation are more complex than tables of numeric values, table 1 presents usual data representation with object in rows and variables (attributes) in columns with number in each cell while table 2 presents table of symbolic objects with intervals, sets of categories. In many real-life economic problems we deal with symbolic variables instead of classical ones. We get intervals instead single values (points), set of categories instead single categories and so on.

Table 1. Classical data matrix

Variables / Objects	Income (in PLN)	Seniority (in years)	...	Other collaterals
Client 1	1000	12	...	1
Client 2	2500	1	...	1
Client 3	3000	0.5	...	2
⋮	⋮	⋮		⋮
Client m	675	1	...	3

1 – none; 2 – underwriter; 3 – mortgage.
Source: artificial data.

Table 2. Symbolic data table

Variables / Objects	Income (in PLN)	Seniority (in years)	...	Other collaterals
Client 1	(1000; 1700)	[0; 0.5]	...	{none}
Client 2	(1500; 2200)	(0.5; 1]	...	{insurance, mortgage}
Client 3	(2000; 2700)	(1; 2]	...	{mortgage}
⋮	⋮	⋮		⋮
Client m	(750; 1100)	(2; 3]	...	{insurance, underwriter}

Source: artificial data.

Symbolic data analysis methods were designed to analyze more complex data that is describing either individuals, so called first-order objects, (described by symbolic variables) or groups (classes) of classical individuals, so called second--order objects [1, pp. 18-20].

4. Kernel discriminant analysis for symbolic objects

One cannot discuss the density distribution in the case of a symbolic objects space. The integral operator is not defined in this kind of space and it is not a subspace of Euclidean space as well.

Let us consider the case where the data are symbolic objects described by seven different types of variables (for example 3 are multivalued variable with weights; 2 are quantitative of interval type; and 2 are multivalued variables). The density estimation can be generalized either using one dissimilarity measure or seven different dissimilarity measures (one for each variable) or three dissimilarity measures (one for each of variable types).

Bock and Diday [1] introduced a replacement of kernel density estimator for symbolic objects [1, p. 242; 10, pp. 127-132]:

$$\hat{I}_k(z) = \frac{1}{n_k} \sum_{i=1}^{n_k} \prod_{j=1}^{p} K_{z,h_j}(x_{ki}), \tag{2}$$

where: $\hat{I}_k(z)$ – kernel intensity estimator for the object z and the k-th class,
$k = 1, 2, \ldots, g$ – number of classes,
n_k – number of objects in k-th class,
h_k – bandwidth window for k-th class (a parameter),
$j = 1, 2, \ldots, p$ – number of dissimilarity measures applied,
$K_{z,h_j}(x_{ki})$ – kernel for object z and x-th object in k-th class, defined as follows:

$$K_{z,h_j}(x_{ki}) = \begin{cases} 1 & \text{for } d_j(z, x_{ki}) < h_j \\ 0 & \text{for } d_j(z, x_{ki}) \geq h_j \end{cases}, \tag{3}$$

$d_j(z, x_{ki})$ – dissimilarity measure for symbolic objects.

Many dissimilarity measures are described in [1, pp. 166-183; 9, pp. 473-481]. Posterior probabilities of the class for z-th object are given as [1, p. 244]:

$$q_k(z) = \frac{\hat{p}_k \hat{I}_k(z)}{\sum_{i=1}^{g} \hat{p}_i \hat{I}_i(z)}, \tag{4}$$

where: $\hat{p}_k$ – prior probabilities for the k-th class,
$\hat{I}_k(z)$ – intensity estimator for the z-th object and the k-th class,
$i = 1, 2, \ldots, g$ – number of classes.

Prior probabilities ($\hat{p}_k$) could be equal for each class $\hat{p}_k = \frac{1}{g}$, or they can consider proportions observed in the training set $\hat{p}_k = \frac{n_k}{N}$, or they could be obtained by maximizing the EM-like algorithm $\hat{p}_k(t+1) = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\hat{p}_k \hat{I}_k}{\sum_{i=1}^{g} \hat{p}_i \hat{I}_i}\right)$ for $i = 1, 2, \ldots, g$ number of classes and t steps of iteration for m points to be classified [1, pp. 242-243].

5. Credit rating with application of symbolic kernel discriminant analysis

Training set contains 80 objects describing BGŻ S.A. Department in Kłodzko bank customers in year 2004. It has been divided into two classes. The first one contains 60 objects pre-classified as borrowers and the second contains 20 clients with negative credit decisions (chosen from 45 negative credit decisions). The test set contains 20 objects. Each of the objects has been characterized by fourteen variables:

1. V_1 – average account incomes – quantitative of interval type in thousands,
2. V_2 – borrowers seniority – quantitative of interval type,
3. V_3 – duration of a credit in months – quantitative of interval type,
4. V_4 – borrowers income – quantitative of interval type in thousands,

5. V_5 – applied amount of a credit – quantitative of interval type in thousands,
6. V_6 – credit record – set of categories received from BIK (credit information bureau) and MIG BR (banks list of unreliable clients),
7. V_7 – client seniority in a bank – set of categories,
8. V_8 – underwriter – set of categories,
9. V_9 – underwriters reliability rating – set of categories,
10. V_{10} – other collaterals – set of categories,
11. V_{11} – clients internal rating – set of categories,
12. V_{12} – evaluation of clients loyalty – set of categories,
13. V_{13} – credit information given by a client – set of categories,
14. V_{14} – allocation of a client to a given class – nominal.

For storing information about training set Microsoft Access 2000 has been used and for assigning object from test set to classes Symbolic Official Data Analysis Software (SODAS) modules DB2SO (extracting objects from database to SODAS), DI (distance measurement) and DKS (symbolic kernel discriminant analysis).

Table 3. Posterior probabilities for test set

No. of object in test set	Posterior probabilities for a class		Maximum probability
	Class 1	Class 2	
1	0.7219	0.2781	Class 1
2	0.4248	0.5752	Class 2
3	0.7249	0.2751	Class 1
4	0.5710	0.4290	Class 1
5	0.6357	0.3643	Class 1
6	0.5679	0.4321	Class 1
7	0.4285	0.5715	Class 2
8	0.6327	0.3673	Class 1
9	0.5872	0.4128	Class 1
10	0.6987	0.3013	Class 1
11	0.4261	0.5739	Class 2
12	0.2459	0.7541	Class 2
13	0.4225	0.5775	Class 2
14	0.4395	0.5605	Class 2
15	0.4259	0.5741	Class 2
16	0.4320	0.5680	Class 2
17	0.4329	0.5671	Class 2
18	0.3578	0.6422	Class 2
19	0.2547	0.7453	Class 2
20	0.3658	0.6342	Class 2

Source: own computation (SODAS software).

Ichino-Yaguchi non-standardized dissimilarity measure was applied in the research (see [1, pp. 166-183; 9]).

Prior probabilities have been estimated considering the proportions observed in training set: 0.75 for class 1 and 0.25 for class 2 posterior probabilities are presented in table 3.

Information from table 3 allows us to compare decision made by credit officer and decision resulting from symbolic kernel discriminant analysis. Correctness of classification is presented in table 4.

Table 4. Correctness of classification

No. of object in test set	Decision resulting from discriminant analysis	Bank's decision	Is object correctly classified?
1	Class 1	Class 1	Yes
2	Class 2	Class 1	No
3	Class 1	Class 1	Yes
4	Class 1	Class 1	Yes
5	Class 1	Class 1	Yes
6	Class 1	Class 1	Yes
7	Class 2	Class 1	No
8	Class 1	Class 1	Yes
9	Class 1	Class 1	Yes
10	Class 1	Class 1	Yes
11	Class 2	Class 1	No
12	Class 1	Class 1	Yes
13	Class 2	Class 2	Yes
14	Class 2	Class 2	Yes
15	Class 2	Class 2	Yes
16	Class 2	Class 2	Yes
17	Class 2	Class 2	Yes
18	Class 2	Class 2	Yes
19	Class 2	Class 2	Yes
20	Class 2	Class 2	Yes

Source: own computation.

By analyzing table 4 it can be said, that that 17 out of 20 objects where correctly classified, so the percentage of correct classification is 0.85. This value was reached by selecting a bandwidth parameter at average distance between all objects from training set 0.07420. This bandwidth parameter provides optimal rate of correctly classified objects. Other most used in literature bandwidth parameters (like 1 or 2) provided worse results (rate of correct classification equal to 0.384615 if $h = 1$ or 2).

6. Summary

A relatively small training sample allowed to get a high percentage of the accuracy of borrowers classification. A bigger sample might have provided even more accuracy. It is not a result sampling technique or sample characteristics nor the chosen period. For artificially generated symbolic data with no noisy variables symbolic kernel discriminant analysis gives high percentage of the accuracy (see [4]).

Clients who were denied by a bank to get a credit, would also receive a negative decision in the case of kernel discriminant analysis for symbolic objects.

The highest percentage of correctly classified clients is achieved when a bandwidth parameter h is set on a level of the average distance between the objects from training set.

Three out of four clients, who would not get a credit in the case of applying discriminant analysis for symbolic objects, had problems with the subsequent repayments of a credit.

No comparisons with classical estimators have been made because when we are dealing with symbolic data we need to transform symbolic variables to classical ones and then apply classical methods. Such comparisons are an opened issue for further research.

Literature

[1] Bock H-H., Diday E. (eds.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg 2000.

[2] Domański C., Pruska K., *Nieklasyczne metody statystyczne*, PWE, Warszawa 2000.

[3] Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Prace Naukowe Akademii Ekonomicznej nr 1021, AE, Wrocław 2004, s. 100-106.

[4] Dudek A., Pełka M., *Effectiveness of Symbolic Classification Trees vs. Noisy Variables*. Folia Oeconomica, Acta Universitatis Lodziensis, Łódź (in review).

[5] Dudek A., *Zastosowanie analizy dyskryminacyjnej obiektów symbolicznych do filtrowania poczty elektronicznej*, Folia Oeconomica, Acta Universitatis Lodziensis, Łódź 2005.

[6] Feldman W., *Kryterium wyboru krzywych Pearsona*, „Przegląd Statystyczny" 1975 nr 22/1.

[7] Hand D, Mannila H, Smyth P., *Principles of Data Mining*, MIT Press, Cambridge 2001.

[8] Härdle W., Simar L., *Applied Multivariate Data Analysis*, Springer Verlag, Berlin-Heidelberg 2003.

[9] Malerba D., Esposito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analisys*, [w:] P. Nanopoulos (ed.), *New Technics and Technologies for Statistics and Exchange of Tehnology and Know-how*, (ETK-NTTS'01) Post conference materials, s. 473-481.

[10] Rasson J.F., Lissoir S., *Symbolic Kernel Discriminant Analysis*, „Computational Statistics" 2000 issue 15, s. 127-132.

ZASTOSOWANIE JĄDROWEJ ANALIZY DYSKRYMINACYJNEJ OBIEKTÓW SYMBOLICZNYCH DO OCENY ZDOLNOŚCI KREDYTOWEJ

Streszczenie

Celem artykułu jest przedstawienie możliwości zastosowania jądrowej analizy dyskryminacyjnej obiektów symbolicznych do oceny zdolności kredytowej osób fizycznych. Artykuł pokazuje również, jak „klasyczna" analiza Bayesowska może być zaadaptowana dla różnych typów danych symbolicznych za pomocą jądrowego estymatora intensywności dla obiektów symbolicznych. W części empirycznej dokonano oceny zdolności kredytowej osób fizycznych na podstawie danych uzyskanych z roku 2004 dla banku BGŻ SA Oddział w Kłodzku.

Marcin Pełka – dr, asystent w Katedrze Ekonometrii i Informatyki Uniwersytetu Ekonomicznego we Wrocławiu – Wydział w Jeleniej Górze.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Piotr Michalski

ZASTOSOWANIE WYBRANYCH ESTYMATORÓW MODELU REGRESJI LOGISTYCZNEJ W CREDIT SCORINGU

1. Wstęp

Zadanie oceny zdolności kredytowej rozważa się często w kontekście problemu klasyfikacji pod nadzorem, co oznacza w istocie uznanie tożsamości pojęć metody credit scoringu i metody klasyfikacji pod nadzorem (zob. np. [7; 8]). Umożliwia to zastosowanie w problemach credit scoringu popularnego liniowego modelu regresji logistycznej. W artykule zostanie empirycznie zweryfikowana przydatność w zastosowaniach credit scoringu bardziej zaawansowanych podejść do estymacji modelu logitowego: addytywnego modelu regresji logistycznej oraz boostingu drzew klasyfikacyjnych.

2. Credit scoring jako zadanie klasyfikacji pod nadzorem

W artykule przyjęto założenie, że doświadczenia banku ze współpracy z klientami pozwalają na wyróżnienie dwóch klas klientów — „dobrych" i „złych". Wtedy zadanie credit scoringu odpowiada problemowi binarnej klasyfikacji pod nadzorem. W ogólności, w klasyfikacji pod nadzorem przyjmuje się istnienie stochastycznej zależności między wielokategorialną cechą objaśnianą Y, której numery kategorii zakodowane są np. jako zbiór $G=\{1,2,\ldots,g\}$, oraz wektorem cech objaśniających $\mathbf{X}=\left[X_1 \;\; X_2 \;\ldots\; X_p\right]$. Wówczas modelowane są prawdopodobieństwa warunkowe $p(Y=k|\mathbf{X})=p_k(\mathbf{X}) \;\; k=1,2,\ldots,g$ lub ich monotoniczne transformacje. Klasyfikacja obserwacji z klasy i do klasy j powoduje stratę (koszt) $L(i,j)$, przy czym zwykle zakłada się, że $L(i,j)\leq 0$, natomiast $L(i,j)>0$

$\forall i, j \in G$. Funkcję strat L można przedstawić za pomocą macierzy strat $\left[L(i,j)\right]_{g\times g}$. Funkcją decyzyjną, minimalizującą oczekiwany błąd predykcji w próbie testowej, jest klasyfikator bayesowski

$$d^*(x) = \arg\min_k \sum_{i=1}^{g} L(i,k)p_i(x), \tag{1}$$

gdzie, zgodnie z twierdzeniem Bayesa, $p_i(x) = \pi_i p(x|i) \Big/ \sum_{r=1}^{g} \pi_r p(x|r)$ oraz $\pi_i = P(Y=i)$, $i = 1, 2, \ldots, g$. Zadaniem zatem jest oszacowanie na podstawie dostępnej próby uczącej $T = \{x_i, y_i\}_1^n$ prawdopodobieństw $p_i(x)$ i podstawienie ich do wzoru (1). W ten sposób uzyskuje się oszacowanie klasyfikatora bayesowskiego, które dalej będzie oznaczane jako $\hat{d}(x)$. Prawdopodobieństwa $p_i(x)$ mogą być bezpośrednio estymowane za pomocą prezentowanych w artykule metod regresji logistycznej.

W rozpatrywanym przypadku dwóch klas klientów banków zbiór G przyjmie postać zbioru {0; 1} dla liniowego i addytywnego modelu logitowego oraz {–1; 1} dla modelu boostingu drzew klasyfikacyjnych. Tak zakodowane kategorie nie tylko będą identyfikować klasy, lecz staną się również przedmiotem obliczeń jako wartości liczbowe.

3. Wybrane modele regresji logistycznej

Model regresji logistycznej można zapisać ogólnie jako funkcję logitową pewnej funkcji warunkowej wartości oczekiwanej cechy Y, wyrażonej jako funkcja wektora cech $\mathbf{X}$:

$$\operatorname{logit}\left[g\left(E\left(Y|\mathbf{X}\right)\right)\right] = h(\mathbf{X}), \tag{2}$$

gdzie: $\operatorname{logit}(x) = \ln\left[1/(1-x)\right]$,

$g(\bullet)$ – pewna monotoniczna funkcja,

$h(\mathbf{X})$ – predyktor.

W teorii uogólnionych modeli liniowych zakłada się, że obserwacje zmiennej objaśnianej pochodzą z rozkładu należącego do rodziny wykładniczych rozkładów prawdopodobieństwa. W przypadku rozpatrywanego problemu klasyfikacji pod nadzorem, w którym występują mikrodane, a kodowanie kategorii jest binarne, cecha Y posiada należący do rodziny rozkładów wykładniczych rozkład zero-jedynkowy, a logitowa funkcja wiążąca jest kanoniczna. W tab. 1 przedstawiono sposób zapisu liniowego modelu regresji logistycznej (GLM-Logit), addytywnego

Tabela 1. Charakterystyki trzech modeli regresji logistycznej

Model	Kodowanie kategorii cechy Y	$E(Y\|\mathbf{X})$	$g(E(Y\|\mathbf{X}))$	$h(\mathbf{X})$
GLM-Logit	{0, 1}	$P(Y=1\|\mathbf{X})$	$E(Y\|\mathbf{X})$	$\alpha_0+\sum_{i=1}^{p}\alpha_i X_i$
GAM-Logit	{0, 1}	$P(Y=1\|\mathbf{X})$	$E(Y\|\mathbf{X})$	$\alpha_0+\sum_{i=1}^{m}\alpha_i X_i+\sum_{j=m+1}^{p} f_j(X_j)$
Real AdaBoost	{−1, 1}	$2P(Y=1\|\mathbf{X})-1$	$[E(Y\|\mathbf{X})+1]/2$	$2\sum_{m=1}^{M} d_m(x)$

Źródło: opracowanie własne.

modelu regresji logistycznej (GAM-Logit) oraz modelu boostingu drzew klasyfikacyjnych Real AdaBoost w konwencji wzoru (2).

W tab. 1 indeksowane wielkości α to pewne parametry, $d(x)$ oznaczają klasyfikatory, natomiast $f_j(\bullet)$ to pewne nieparametryczne, gładkie funkcje. Rozwiązując równanie (2) względem $p_1(x)$, uzyskuje się prawdopodobieństwa *a posteriori* przynależności do klas, wykorzystywane we wzorze (1):

$$p_1(x)=\frac{\exp[h(\mathbf{X})]}{1+\exp[h(\mathbf{X})]}. \tag{3}$$

Model GLM-Logit stał się podstawowym narzędziem analizy regresji w przypadku jakościowej cechy objaśnianej. Z postaci podanej w tab. 1 wynika, że w modelu tym logit prawdopodobieństwa *a posteriori* przynależności do klasy 1 modelowany jest za pomocą tradycyjnego, liniowego predyktora. Uogólniony model addytywny jest rozszerzeniem tego powszechnie wykorzystywanego w credit scoringu narzędzia. W modelu GAM-Logit zakłada się *explicite*, że cechy X_1, X_2, ..., X_m posiadają formę liniowego, parametrycznego predyktora, co może być podyktowane jakościowym bądź dyskretnym charakterem cech X_1, X_2, ..., X_m, albo uprzednią wiedzą o liniowych efektach tych zmiennych. Z kolei ilościowe cechy X_{m+1}, ..., X_p dopasowywane są technikami nieparametrycznymi w nadziei odkrycia ich nieliniowych efektów. Najczęściej wykorzystywane metody nieparametrycznej estymacji jednowymiarowej funkcji regresji $f(\bullet)$ to technika wygładzonej funkcji sklejanej (*smoothing spline*) oraz technika lokalnej regresji, znana w anglojęzycznej literaturze jako *local regression* lub *loess* (zob. np. [2; 6]).

Uogólnione modele addytywne znajdują zastosowanie w wielu problemach klasyfikacji pod nadzorem. Wydaje się, że ocena zdolności kredytowej jest jednym z naturalnych pól zastosowań, szczególnie gdy zbiór danych obejmuje obserwacje cech zarówno ilościowych, jak i jakościowych. Modelowanie nieliniowe cech

ilościowych bowiem pozwala odkrywać w danych nowe wzorce. Możliwe jest przy tym sterowanie stopniem wygładzania, co zapobiega nadmiernemu dopasowaniu do danych, przez co zachowana zostaje korzystna relacja między obciążeniem a wariancją modelu. Poza zastosowaniem predyktywnym, uogólnione modele addytywne służą jako dobre narzędzie eksploracyjne. Uogólniony model addytywny wykorzystuje część narzędzi wnioskowania o uogólnionych modelach liniowych oraz posiada nie mniej użyteczne możliwości interpretacyjne. Problemem pozostaje dobór zmiennych do modelu. Nierozwiązaną kwestią jest też procedura specyfikacji postaci zmiennych w modelu. Problematyczny jest często wybór między parametryczną a nieparametryczną reprezentacją danej cechy. Warunkiem powodzenia zastosowania modelu GAM-Logit w credit scoringu jest zatem umiejętny dobór cech modelowanych nieparametrycznie, wybór nieparametrycznego estymatora oraz ustalenie stopnia wygładzania.

Model GLM-Logit estymowany jest na ogół metodą największej wiarogodności lub jej ekwiwalentem – algorytmem IRWLS (*iteratively re-weighted least squares algorithm*). Algorytm ten zaadaptowano do estymacji modelu GAM-Logit. Krok polegający na zastosowaniu ważonej metody najmniejszych kwadratów (*weighted least squares step*) zastępowany jest procedurą ważonego algorytmu wielokrotnego dopasowania (*weighted backfitting algorithm*) (zob. np. [2; 6]).

Uogólnione modele liniowe są powszechnie oprogramowane (np. pakiet *Statistica* lub funkcja *glm* pakietu *Stats* w środowisku *R*). Model można również oszacować w dowolnym arkuszu kalkulacyjnym, maksymalizując logarytm funkcji wiarogodności (np. za pomocą narzędzia optymalizacyjnego *Solver* w arkuszu kalkulacyjnym *Excel*). Uogólnione modele addytywne oprogramowane są np. pod postacią rozbudowanego modułu w programie *Statistica Data Miner*. W środowisku *R* do wyboru jest kilka pakietów, wśród których najpopularniejszy jest pakiet *gam*. Pakiet *gam* jest implementacją algorytmu IRWLS i umożliwia wybór lokalnej regresji lub wygładzanej funkcji sklejanej w charakterze nieparametrycznego estymatora funkcji regresji.

Przedstawiony tu zostanie algorytm boosting, który został uznany za jedną z najlepszych technik statystycznego uczenia się. Algorytm ten opiera się na pomyśle łączenia decyzji wielu niezależnych klasyfikatorów, przy czym zakłada się, że pojedyncze klasyfikatory są nieco lepsze od klasyfikatora losowego. Algorytm boosting spełnia w przybliżeniu dwa powyższe warunki, dopasowując sekwencyjnie modele klasyfikacyjne do danych ważonych za każdym razem innymi wagami. Obserwacja, która w poprzedzającym kroku została błędnie zaklasyfikowana, otrzymuje większą wagę w kolejnym kroku. Najczęściej wykorzystywanym klasyfikatorem bazowym jest drzewo klasyfikacyjne, a szczególnie jego dwuliściowa wersja. Każdorazowe ważenie obserwacji wypełnia (w sposób niedoskonały) postulat niezależności klasyfikatorów, przy czym najczęściej rozważa się dwie alternatywne koncepcje ważenia:

- wbudowanie wag w mechanizm klasyfikatora,
- szacowanie klasyfikatora na podstawie pseudopróby, powstałej w wyniku losowania ze zwracaniem z próby uczącej ustalonej liczby elementów zgodnie z rozkładem prawdopodobieństwa wyznaczonym przez wagi (konieczność, gdy klasyfikator nie obsługuje obserwacji ważonych).

Należy zauważyć, że każdorazowe ważenie obserwacji nie gwarantuje całkowitej niezależności prób uczących. Mimo to algorytm boostingu prowadzi w wielu przypadkach do znacznego zmniejszenia obciążenia oraz redukcji wariancji i jest obecnie uważany za jedną z najlepszych metod klasyfikacji pod nadzorem.

Najpopularniejszą odmianą algorytmu boosting jest algorytm rzeczywistego boostingu drzew klasyfikacyjnych Real AdaBoost. Algorytm ten opublikowali w 2000 r. w podstawowej wersji Hastie, Tibshirani i Friedman (zob. [5]). Algorytm Real AdaBoost przedstawiono w tab. 2. Przyjęto, że $d_m(x)$ oznacza klasyfikator generujący wartości rzeczywiste.

Tabela 2. Algorytm Real AdaBoost

Krok algorytmu	Procedura
1	Przyjmij wagi początkowe $w_i = 1/n$, $i = 1, 2, \ldots, n$.
2	Powtarzaj dla $m = 1, 2, \ldots, M$: a) oszacuj za pomocą mechanizmu klasyfikacji pod nadzorem prawdopodobieństwa $p_m(x) = \hat{P}_w(Y = 1\|x)$, stosując do próby uczącej wagi w_i; b) podstaw $d_m(x) \leftarrow 0{,}5\ln\left[p_m(x)/(1 - p_m(x))\right]$; c) podstaw $w_i \leftarrow w_i \exp\left[-y_i d_m(x)\right]$, $i = 1, 2, \ldots, n$, i dokonaj renormalizacji (żeby $\sum_i w_i = 1$).
3	Przyjmij za decyzję klasyfikacyjną wielkość $\operatorname{sign}\left[\sum_{m=1}^{M} d_m(x)\right]$.

Źródło: [5, s. 340].

Doniosłe odkrycie Hastiego, Tibshiraniego i Friedmana (zob. [5]) dotyczące statystycznych podwalin algorytmu AdaBoost wskazuje, że algorytm ten jest w istocie pewną procedurą estymacji modelu regresji logistycznej. Rozpatrywana jest funkcja rzeczywista $J(h) = E\left[\exp(-Yh(\mathbf{X}))\right]$. Populacyjną funkcją minimalizującą $J(h)$ w punkcie x jest

$$\tilde{h}(x) = \arg\min_{h(x)}\left\{E_{Y|\mathbf{X}}\left[\exp(-Yh(x))\right]\right\} = \frac{1}{2}\ln\frac{P(Y=1|x)}{P(Y=-1|x)}, \tag{4}$$

co można wykazać, minimalizując funkcję $J(h)$ pod warunkiem $X = x$:

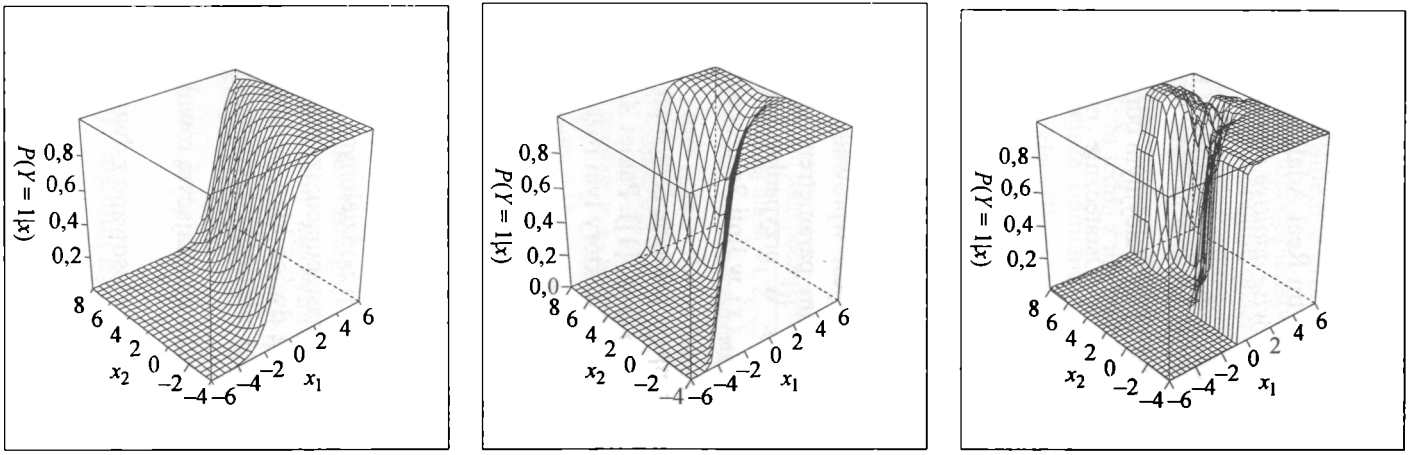
$$E\left[\exp(-Yh(x))|x\right] = P(Y=1|x)e^{-h(x)} + P(Y=-1|x)e^{h(x)}. \quad (5)$$

Obliczając pochodną funkcji (5) względem $h(x)$ i przyrównując ją do zera, otrzymuje się równanie (4). Okazuje się, że algorytm Real AdaBoost dopasowuje w modelu (4) za pomocą procedury *forward stagewise additive modelling* (zob. np. [6]) addytywną funkcję $\tilde{h}(x) = \sum_{m=1}^{M} d_m(x)$, wykorzystując wykładniczą funkcję straty $L(y, h(x)) = \exp(-yh(x))$. Wielkość $yh(x)$ jest klasyfikacyjnym odpowiednikiem reszty, gdy klasy są zakodowane jako $\{-1; 1\}$. Ostatecznie można zapisać $h(x) = 2\tilde{h}(x)$.

Metaparametrem algorytmów boostingu jest wielkość pojedynczego drzewa. Najczęściej stosowane podejście zakłada ustalenie jednakowej wielkości drzewa dla każdej iteracji. Z doświadczeń praktycznych wynika, że optymalna wielkość drzewa zawiera się zwykle w przedziale od 4 do 8 liści i rzadko przekracza 10. Ważną kwestią jest też regularyzacja (ustalenie optymalnej złożoności modelu). W boostingu jednym ze sposobów regularyzacji jest wprowadzenie parametru spowalniającego proces uczenia algorytmu. Takim parametrem jest tzw. stopa szybkości uczenia się $v \in (0; 1]$ (*learning rate*). W przypadku algorytmu Real AdaBoost przez v mnożone jest oszacowanie $d_m(x)$ w linii 2c procedury. Zwykle ustalana jest mała wartość v, np. 0,1.

Algorytmy boostingu oprogramowane są np. w pakiecie *Statistica Data Miner*. W środowisku *R* najpopularniejszy pakiet to *Ada* (zob. [1]). Pakiet *Ada* jest implementacją stochastycznego algorytmu Real AdaBoost, który tym różni się od opisanego, że trenowanie klasyfikatora (linia 2a) odbywa się na podstawie podpróby próby uczącej, powstałej w wyniku losowania z niej bez zwracania nieco mniejszej od n liczby obserwacji. Zamysłem tego zabiegu jest wprowadzenie do algorytmu elementu losowości i obniżenie oczekiwanego błędu predykcji w próbie testowej.

Do celów ilustracyjnych na rys. 1 zaprezentowano oszacowanie prawdopodobieństwa (3) uzyskane za pomocą trzech modeli dla sztucznie wygenerowanej próby uczącej. Próba ta składa się z tysiąca obserwacji podzielonych na dwie klasy. Mechanizm generujący dane przedstawiał się następująco: z prawdopodobieństwem 0,5 przydzielano obserwacje do jednej z dwóch klas. W klasie pierwszej ($Y = 1$) obserwacje pochodzą z dwuskładnikowej mieszanki rozkładów normalnych $N\left(\begin{bmatrix}2\\0\end{bmatrix}, \begin{bmatrix}1 & 1\\1 & 3\end{bmatrix}\right)$, $N\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & -1\\-1 & 2\end{bmatrix}\right)$ o współczynniku mieszającym równym 0,5, natomiast w drugiej klasie ($Y = 0$ lub $Y = -1$) obserwacje losowano z dwuwymiarowego rozkładu normalnego $N\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}4 & 2\\2 & 2\end{bmatrix}\right)$.



Rys. 1. Hiperpowierzchnie logistyczne $\hat{p}_1(x)$ uzyskane modelami GLM-Logit, GAM-Logit oraz Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

Przykłady zastosowania opisanych metod do rzeczywistych zbiorów danych kredytowych zostaną przedstawione w kolejnym punkcie artykułu.

4. Zastosowanie modeli do danych rzeczywistych

Zbiór danych pod nazwą German Credit pochodzi z ogólnodostępnej internetowej bazy danych[1]. Zbiór ten obejmuje obserwacje 20 cech objaśniających pochodzących z 1000 wniosków kredytowych jednego z dużych banków południowych Niemiec. Z ogólnej liczby wniosków 300 to aplikacje klientów uznanych za „złych", a 700 to wnioski klientów „dobrych". Nie wiadomo, czy próba ucząca oddaje rzeczywisty stosunek liczebności dwóch klas klientów banku. Zwykle odsetek klientów mających problemy ze spłatą kredytu jest znacznie mniejszy niż 30%, przeto można wnioskować, że próba ucząca nie jest reprezentatywna. Wśród predyktorów znajduje się 7 cech mierzalnych i 13 cech jakościowych, o liczbach kategorii wynoszących od 2 do 11. Cechy oznaczono symbolami od X_1 do X_{20}. Przyjęto następujący sposób kodowania cechy objaśnianej: $Y=1$, gdy klient jest „dobry", $Y=0$ (lub $Y=-1$) w przeciwnym wypadku. Wśród cech mierzalnych zmienne X_8, X_{11}, X_{16} oraz X_{18} są dyskretne, natomiast cechy X_2 (okres spłaty), X_5 (kwota kredytu) oraz X_{13} (wiek) mają charakter ciągły i tym samym mogą być wykorzystane w nieliniowej części semiparametrycznego modelu GAM-Logit. Opis cech objaśniających zawiera tab. 3.

Zbiór German Credit jest również analizowany w publikacjach [3] i [4]. W pracy [3] zastosowano model regresji logistycznej, drzewo klasyfikacyjne typu CART i jedną z odmian sieci neuronowej. Z kolei w monografii [4] do budowy modelu scoringowego wykorzystano liniowy model prawdopodobieństwa oraz GLM-Logit. W niniejszym artykule zastosowano trzy omawiane modele regresji logistycznej.

Modele ekonometryczne GLM-Logit oraz GAM-Logit wymagają kodowania cech jakościowych i wstępnego doboru zmiennych. Zabiegi te nie są koniecznością w przypadku bazującego na modelu drzew klasyfikacyjnych algorytmu Real AdaBoost. Do wstępnego określenia istotności predyktorów posłużył wskaźnik relatywnej ważności cech (zob. [6]) uzyskany za pomocą algorytmu Real AdaBoost na podstawie oryginalnej próby uczącej (100 iteracji, $\nu = 0{,}1$). Algorytm wskazał cechę X_{20} (pracownik zagraniczny) jako najmniej istotną w zbiorze predyktorów. Cecha ta charakteryzuje się ponadto małą zmiennością, dlatego zdecydowano się na jej wyłączenie z analizy. Na potrzeby analizy ekonometrycznej zakodowano cechy jakościowe i w rezultacie otrzymano zbiór 19 zmiennych zero-jedynkowych (binarnych). Do zbadania zależności cechy objaśnianej Y i zmiennych binarnych zastosowano test niezależności chi-kwadrat. Z analizy wyłączono zmienne binarne nieistotnie skojarzone z cechą objaśnianą Y. Ostatecznego doboru

[1] http://www.niaad.liacc.up.pt/old/statlog (maj 2007).

Tabela 3. Opis predyktorów zbioru German Credit

Cecha	Opis	Rodzaj cechy	Liczba kategorii
X_1	Stan obecnego rachunku rozliczeniowego	Jakościowa	4
X_2	Okres spłaty w miesiącach	Ilościowa	–
X_3	Historia kredytowa	Jakościowa	5
X_4	Przeznaczenie kredytu	Jakościowa	11
X_5	Kwota kredytu	Ilościowa	–
X_6	Konto oszczędnościowe/obligacje	Jakościowa	5
X_7	Okres obecnego zatrudnienia	Jakościowa	5
X_8	Wielkość raty w procentach dochodu rozporządzalnego	Ilościowa	–
X_9	Płeć i stan cywilny	Jakościowa	5
X_{10}	Inni dłużnicy/żyranci	Jakościowa	3
X_{11}	Okres obecnego zamieszkania w latach	Ilościowa	–
X_{12}	Majątek	Jakościowa	4
X_{13}	Wiek kredytobiorcy w latach	Ilościowa	–
X_{14}	Inne zobowiązania ratalne	Jakościowa	3
X_{15}	Mieszkanie	Jakościowa	3
X_{16}	Liczba obecnie spłacanych kredytów w tym banku	Ilościowa	–
X_{17}	Zatrudnienie	Jakościowa	4
X_{18}	Liczba osób na utrzymaniu	Ilościowa	–
X_{19}	Telefon	Jakościowa	2
X_{20}	Pracownik zagraniczny	Jakościowa	2

Źródło: opracowanie własne na podstawie German Credit Data/Dataset Description.

zmiennych dokonano za pomocą procedury postępującej regresji krokowej (w ramach modelu GLM-Logit). Spośród trzech cech ciągłych cecha oznaczająca wiek okazała się nieistotna.

Kolejną kwestią jest ustalenie semiparametrycznej struktury modelu GAM-Logit. W tab. 4 zestawiono wybrane statystyczne charakterystyki konkurencyjnych modeli GLM-Logit oraz GAM-Logit z trzema kombinacjami nieparametrycznych form zmiennych X_2 oraz X_5: *deviance* – odległość między modelem nasyconym a danym; df – liczba stopni swobody; pseudo R^2 – mierzony w kategoriach procentu objaśnienia wariancji zmiennej objaśnianej (nie jest to poprawna miara w przypadku modeli GLM-Logit i GAM-Logit); *deviance explained* – liczba jeden pomniejszona o iloraz *deviance* danego modelu i *deviance* modelu zawierającego wyłącznie wyraz wolny; AIC – wartość kryterium informacyjnego Akaikego. Charakterystyki podano po uśrednieniu dziesięciu realizacji uzyskanych w próbach uczących, losowo pobranych z próby pierwotnej (70% obserwacji), natomiast oszacowania oczekiwanego błędu predykcji (Err) otrzymano po uśrednieniu oszacowań błędu w pozostałych częściach prób (próby testowe, 30% obserwacji). W tab. 4 zawarto również wartość Err wyznaczoną na podstawie całej próby uczącej (resub-

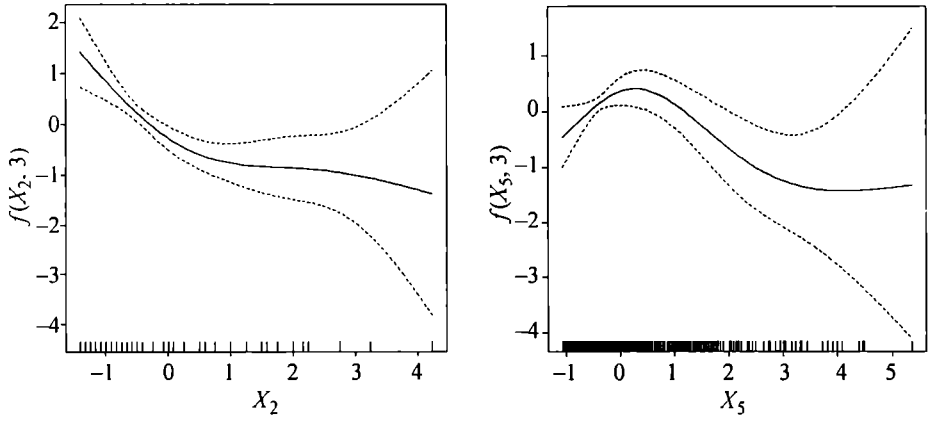
stytucja). Przyjęto zero-jedynkową postać funkcji straty oraz 3 stopnie swobody dla funkcji sklejanych modelu GAM-Logit.

Tabela 4. Zestawienie statystycznych charakterystyk modeli GLM-Logit oraz trzech semiparametrycznych wariantów modelu GAM-Logit (wielkości uśrednione)

Charakterystyki statystyczne modelu	GLM-Logit	GAM-Logit nieparametryczny w:		
		X_2	X_5	X_2, X_5
df	687	684	684	681
Deviance	712,4	708,5	703,3	697,6
AIC	738,4	740,5	735,3	735,6
Pseudo R^2	0,217	0,217	0,225	0,227
Deviance explained	19,1%	19,2%	20,1%	20,4%
p-value X_2	0,000	0,099	0,000	0,027
p-value X_5	0,070	0,079	0,024	0,004
Err Resubstytucja		0,233	0,237	0,235
Err Próba testowa	0,249	0,248	0,246	0,242

Źródło: opracowanie własne wykonane w środowisku R.

Przyjęcie semiparametrycznej postaci zmiennej X_5 lub zmiennych X_2 i X_5 poprawia nieznacznie parametry statystyczne oraz wyniki klasyfikacji. Ostrożnie należy traktować wielkości *p-value*, które w przypadku modelu GAM-Logit ze zmienną nieparametryczną X_2 lub X_5 odnoszą się do różnych testów. Na rys. 2 przedstawiono wykresy cząstkowej predykcji (tj. wykresy funkcji $\hat{f}_j(\bullet)$) modelu nieparametrycznego w zmiennych X_2 oraz X_5.



Rys. 2. Wykresy cząstkowej predykcji zmiennych X_2 oraz X_5 z przedziałem ufności 95%

Źródło: opracowanie własne wykonane w środowisku R.

Po analizie wykresów cząstkowej predykcji oraz charakterystyk konkurencyjnych modeli wskazano jako najlepszy model GAM-Logit nieparametryczny w zmiennych X_2 oraz X_5.

Z wykresu cząstkowej predykcji dla zmiennej X_2 wynika, że ryzyko niewypłacalności rośnie (coraz wolniej) wraz z wydłużaniem się okresu spłaty kredytu, *ceteris paribus*. Z kolei wykres cząstkowej predykcji zmiennej X_5 ujawnia interesującą zależność, mianowicie do pewnej kwoty kredytu ryzyko niewypłacalności maleje, a po jej przekroczeniu wzrasta, *ceteris paribus*. Można innymi słowy powiedzieć, że kredyty o średniej wartości są mniej ryzykowne od kredytów o niskiej i wysokiej kwocie.

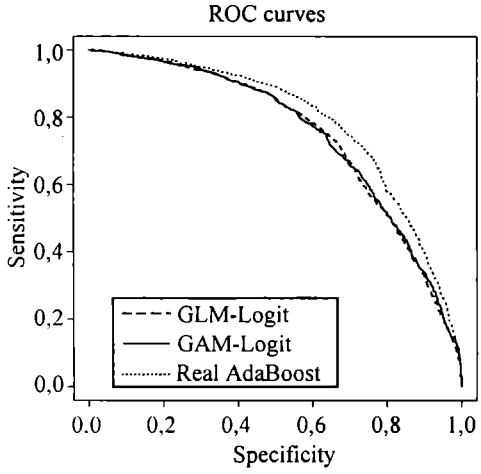
W tab. 5 porównano wyniki klasyfikacji z zastosowaniem modeli GLM-Logit, GAM-Logit nieparametrycznego w zmiennych X_2 oraz X_5 z trzema stopniami swobody oraz algorytmu Real AdaBoost o parametrze $v = 0{,}1$ i 100 iteracjach, opartego na oryginalnym zbiorze cech z wyłączeniem cechy X_{20}. Dodatkowo w tab. 5 uwzględniono drzewo klasyfikacyjne typu CART, będące klasyfikatorem bazowym algorytmu Real AdaBoost. Oszacowania oczekiwanego błędu predykcji podano po uśrednieniu (dziesięciokrotnie losowano z próby uczącej próbę treningową – 70% obserwacji i próbę testową – 30% obserwacji). Przyjęto zero-jedynkową funkcję straty. Obliczono ponadto oszacowania prawdopodobieństw błędnej klasyfikacji (I i II rodzaju).

Tabela 5. Wyniki klasyfikacji pod nadzorem (wielkości uśrednione)

Model	Resubstytucja	Próba testowa (30%)	Oszacowanie prawdopodobieństwa błędu I rodzaju	Oszacowanie prawdopodobieństwa błędu II rodzaju
GLM-Logit	0,235	0,256	0,592	0,112
GAM-Logit	0,233	0,249	0,588	0,103
CART	0,219	0,276	0,544	0,160
Real AdaBoost	0,106	0,237	0,602	0,081

Źródło: opracowanie własne wykonane w środowisku R.

Należy zauważyć, że dane German Credit stanowią dość trudny problem klasyfikacji pod nadzorem. Uzyskane wyniki są bowiem tylko nieco lepsze od metody polegającej na przydzielaniu wszystkich obserwacji do liczniej reprezentowanej klasy. Najlepsze ogólne wyniki klasyfikacyjne uzyskano za pomocą algorytmu Real AdaBoost oraz modelu GAM-Logit. Porównanie metod dla różnych punktów odcięcia (wykres krzywych operacyjno-charakterystycznych ROC) przedstawiono na rys. 3. Punkty (*specificity*, *sensitivity*), tj. dopełnienia do liczby jeden odpowiednio oszacowań prawdopodobieństw popełnienia błędów I i II rodzaju, zostały dziesięciokrotnie uśrednione. Im bardziej krzywa ROC ciąży ku prawemu górnemu rogowi wykresu, tym lepszymi właściwościami predykcyjnymi charakteryzuje się



Rys. 3. Krzywe operacyjno-charakterystyczne liniowego modelu regresji logistycznej (GLM-Logit), addytywnego modelu logistycznego (GAM-Logit) oraz algorytmu Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

rozpatrywany klasyfikator. Algorytm Real AdaBoost dał najlepsze rezultaty dla niemal każdego punktu odcięcia.

Drugi zbiór danych – Australian Credit – pochodzi z tego samego repozytorium co wcześniej rozpatrywany zbiór German Credit. Zbiór składa się z 690 obserwacji 14 cech objaśniających oraz etykiety klasy. Próba ucząca obejmuje 307 obserwacji jednej klasy oraz 383 klasy drugiej. Wśród predyktorów ryzyka znajduje się 6 cech mierzalnych i 8 cech jakościowych, o liczbach kategorii od 2 do 14. Dane związane są ze sferą usług kart kredytowych, jednak ze względów poufności utajnione zostały nazwy cech objaśniających, opisy kategorii cech jakościowych, a do tego nieznana jest interpretacja etykiety klas. Nie jest zatem możliwe nadanie interpretacji ekonomicznej procesowi budowy klasyfikatora. W związku z tym przedstawione zostaną tylko najważniejsze wyniki.

Zmienną – kandydatką do nieparametrycznej reprezentacji w modelu GAM-Logit – była zmienna X33. Efekt wprowadzenia zmiennej X33 do modelu GLM-Logit i GAM-Logit podsumowuje tab. 6. Wyniki uzasadniają nieparametryczne wprowadzenie zmiennej X_{33}.

Ostateczne wyniki klasyfikacji zaprezentowano w tab. 7. Wyniki te uzyskano zgodnie z procedurą opisaną przy analizie zbioru German Credit. Ponownie zwracają uwagę najlepsze rezultaty algorytmu Real AdaBoost oraz dobre wyniki modelu GAM-Logit z nieparametrycznie modelowaną zmienną X_{33}.

Tabela 6. Zestawienie statystycznych charakterystyk modeli GLM-Logit oraz GAM-Logit nieparametrycznego w zmiennej X_{33} (wielkości uśrednienione)

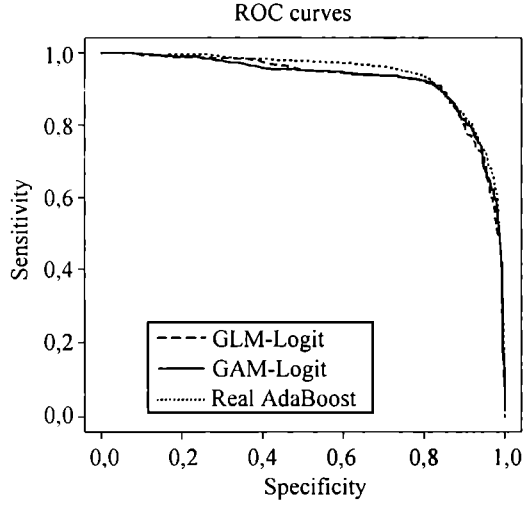
Charakterystyki statystyczne modelu	GLM-Logit bez X_{33}	GLM-Logit z X_{33}	GAM-Logit nieparametryczny w X_{33}
df	467	466	463
Deviance	311,20	306,24	298,89
AIC	329,19	326,24	324,89
Pseudo R^2	0,590	0,601	0,604
Deviance explained	52,4%	53,1%	53,8%
p-value X_{33}	–	0,0725	0,0275

Źródło: opracowanie własne wykonane w środowisku R.

Tabela 7. Wyniki klasyfikacji pod nadzorem (wielkości uśrednione)

Model	Resubstytucja	Próba testowa (30%)	Oszacowanie prawdopodobieństwa błędu I rodzaju	Oszacowanie prawdopodobieństwa błędu II rodzaju
GLM-Logit	0,130	0,137	0,165	0,103
GAM-Logit	0,131	0,133	0,158	0,101
CART	0,101	0,152	0,116	0,198
Real AdaBoost	0,044	0,128	0,126	0,130

Źródło: opracowanie własne wykonane w środowisku R.



Rys. 4. Krzywe operacyjno-charakterystyczne liniowego modelu regresji logistycznej (GLM-Logit), addytywnego modelu logistycznego (GAM-Logit) oraz algorytmu Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

Na rys. 4 zamieszczono wykres krzywych operacyjno-charakterystycznych dla danych Australian Credit.

Widoczna jest ogólna przewaga algorytmu boosting oraz addytywnego modelu regresji logistycznej nad modelem liniowym dla punktów odcięcia, przy których specyficzność i czułość przyjmują jednocześnie wysokie wartości.

Przedstawione przykłady zastosowań pokazują, że uogólnione modele addytywne oraz modele boostingu są interesującym uzupełnieniem/rozszerzeniem liniowej analizy regresji logitowej w problemach credit scoringu. Odpowiednio regularyzowany model GAM-Logit jest ciekawym narzędziem pogłębiającym możliwości interpretacyjne modelu klasyfikacyjnego. Nieparametryczne modelowanie wybranych cech ilościowych pozwala często zastąpić dyskretyzację cechy (w celu uwzględnienia nieliniowości) w ramach liniowego modelu GLM-Logit. Dobrą metodą dokładnej estymacji prawdopodobieństw na potrzeby indywidualnej wyceny kredytu wydaje się z kolei algorytm Real AdaBoost. Boosting redukuje obciążenie i wariancję klasyfikatora bazowego, osiągając zarówno w przedstawionych przykładach, jak i w większości problemów klasyfikacyjnych znakomite wyniki (zob. np. [5; 6]).

Literatura

[1] Culp M., Johnson K., Michailidis G., *Ada: An R Package for Stochastic Boosting*, „Journal of Statistical Software", vol. 17, issue 2, October 2006.

[2] Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Press, London 2006.

[3] Giudici P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons, New York 2003.

[4] Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, SGH, Warszawa 2002.

[5] Hastie T., Tibshirani R., Friedman J., *Additive Logistic Regression: a Statistical View of Boosting*. „The Annals of Statistics" 2000, vol. 28, no. 2, s. 337-407.

[6] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York 2003.

[7] Mester L.J., *What's the Point of Credit Scoring*, „Business Review", September/October 1997, Federal Reserve Bank of Philadelphia.

[8] Migut G., Wątroba J., *Scoring kredytowy a modele data mining*, „Ryzyko w Instytucji Finansowej" 2005 nr 1.

APPLICATION OF SELECTED ESTIMATORS OF LOGISTIC REGRESSION MODELS IN CREDIT SCORING

Summary

The article examines the application opportunities of different logistic regression models in a credit scoring supervised classification problem. The paper covers linear and generalized additive logistic regression model, as well as a classification trees boosting method – Real AdaBoost. The empirical study of two real credit datasets is given.

Piotr Michalski – mgr, doktorant w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 2008
Ekonometria 22

Piotr Michalski

BIAS-VARIANCE TRADE-OFF IN SUPERVISED CLASSIFICATION

1. Introduction

In the last decade of the 20th century much research was devoted to obtaining more accurate approximations of distributions used in classification rules. The main logic behind this was the belief that greater estimation accuracy leads to better predictive properties of classifiers. As was established later, in many cases, enhanced precision of estimation – contrary to intuition – does not necessarily bring better classification results. At issue here is the generalization property of a predictive model, i.e. the ability to retain a predictive power for observations outside a learning sample. It is not uncommon that conceptually simple models, like naive bayesian classifiers or linear probability models outperform some sophisticated regression methods in classification settings. The article presents the decomposition of the expected prediction error in classification introduced by J.H. Friedman [4], which can be used to explain this phenomenon. A simulation example of error calculation via Friedman's decomposition is also given.

2. Prediction models

In a traditional prediction problem it is most often assumed that a continuous dependent variable Y is stochastically associated with non-random explanatory vector $\mathbf{X} = \left[X_1 \;\; X_2 \;\; \ldots \;\; X_p\right]$ through a function $Y = f(\mathbf{X}) + \varepsilon$, $f \in C^1$ where ε is a random component, such that $E(\varepsilon|\mathbf{X}) = 0$. In order to minimise the expected prediction error (with the assumption of the squared loss function) it suffices to estimate a conditional expected value (regression function) $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ point-

wise. Then the prediction problem amounts to the approximation of the function $f(\mathbf{X})$ using a training sample $T=\{x_i, y_i\}_1^n$. In a classification problem, with Y assuming a finite set of values decoded as $G = \{1, 2, \ldots, g\}$, one models conditional probabilities $P(Y=k|\mathbf{X}) = p_k(\mathbf{X})$ $k = 1, 2, \ldots, g$ or their monotonic transformations. The squared loss function is replaced by a matrix-defined loss function of the form $[L(i, j)]_{g\times g}$, where $L(i, j)$ is a cost incurred when predicting $Y = j$ while in reality $Y = i$ It is usually assumed, that $L(i,i) \le 0$ and $L(i, j) > 0$, $\forall i, j \in G$. The optimal decision function is the Bayes classifier (see [1, p. 65]):

$$d^*(\mathbf{x}) = \arg\min_k \sum_{i=1}^{g} L(i,k) p_i(\mathbf{x}),$$

where, according to Bayes theorem, $p_i(\mathbf{x}) = \pi_i p(\mathbf{x}|i) \Big/ \sum_{r=1}^{g} \pi_r p(\mathbf{x}|r)$ and $\pi_i = P(Y = i)$, $i = 1, 2, \ldots, g$. Now the task is to estimate either $p_i(\mathbf{x})$ in a direct fashion, or $p(\mathbf{x}|i)$, π_i, and insert them into Bayes theorem equation. An estimator of the Bayes classifier will be further denoted by $\hat{d}(\mathbf{x})$.

3. Bias-variance trade-off in regression setting

The precision of distribution estimation is a function of model complexity, which depends on the number of parameters (parametric models), or some parameters assuming prespecified values (nonparametric models). The general rule is that an increase in model complexity results in expected prediction error decrease within learning sample. Enhanced data fit does not usually guarantee that the results would be as satisfactory in a test sample – an overfitted model often loses its generalization properties, leading to an increased prediction error. Described mechanism characterizes a phenomenon called bias-variance trade-off, which insists that one should seek an optimal degree of model complexity that minimizes the expected prediction error on independent test sample data. The most favorable point lies somewhere between two extreme models, as depicted in figure 1.

The tool that allows an analitical description of the phenomenon is the decomposition of the expected prediction error into three components: random component, bias and variance (noise-bias-variance decomposition). The expected prediction error will be further denoted by Err and in regression it can be written as

$$Err = E\left[L\left(Y, \hat{f}(\mathbf{X})\right)\right], \tag{1}$$

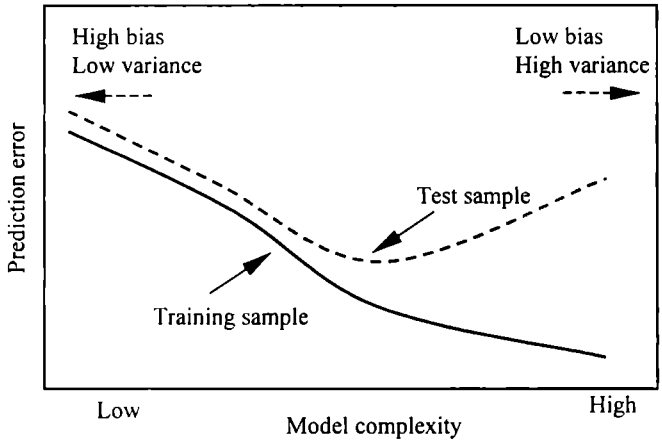


Fig. 1. Prediction error as model complexity function

Source: [5, p. 194].

where $f(\bullet)$ denotes regression function and $L(\bullet)$ is a loss function, that penalizes discrepancies between real value Y and its prediction $\hat{f}(\mathbf{X})$. Squared loss function is the most popular choice of statisticians. In classification one has

$$Err = E\left[L\left(Y, \hat{d}(\mathbf{X})\right)\right]. \tag{2}$$

Let's first consider the well-known case of regression. Suppose that the predictive dependency between features Y and $\mathbf{X}$ can be written as

$$Y = f(\mathbf{X}) + \varepsilon,$$

where $f(\mathbf{X})$ is a deterministic function, ε – random component, such that $E(\varepsilon \mid X) = 0$. The model has the property of $f(x) = E(Y \mid X = x)$. Having at one's disposal a training sample $\mathrm{T} = \{x_i, y_i\}_1^n$ the task is to find the best estimator of $f(X)$:

$$\hat{f}\left(\mathbf{x}|T\right) = \hat{E}\left(Y|\mathbf{X} = \mathbf{x}, \mathbf{x} \in T\right). \tag{3}$$

Let's note, that at point $\mathbf{X} = \mathbf{x}$ the function $\hat{f}\left(\mathbf{x}|T\right)$ is a random variable, as training sample T is also random. It is assumed that the value of the function at every point $\mathbf{x}$ follows a certain distribution $p\left(\hat{f}|\mathbf{x}\right)$ with known expected value and variance:

$$E\hat{f}(\mathbf{x}) = \int \hat{f} p\left(\hat{f}|\mathbf{x}\right) d\hat{f}, \tag{4a}$$

$$Var\hat{f}(\mathbf{x}) = \int \left(\hat{f} - E\hat{f}(\mathbf{x})\right)^2 p\left(\hat{f}|\mathbf{x}\right) d\hat{f}. \tag{4b}$$

The decomposition of the error (1) at a point x with the assumption of the squared loss function can be shown as:

$$\begin{aligned} Err(\mathbf{x}) &= E_{Y,T}\left[\left(Y - \hat{f}(\mathbf{x})\right)^2 |\mathbf{X} = \mathbf{x}\right] = E_Y\left[Y - f(\mathbf{x})\right]^2 + E_T\left[f(\mathbf{x}) - \hat{f}(\mathbf{x})\right]^2 = \\ &= E_Y\left[Y - f(\mathbf{x})\right]^2 + E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2 + \left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2 = \\ &= Var\left(\varepsilon|\mathbf{X} = \mathbf{x}\right) + Var\hat{f}(\mathbf{x}) + Bias^2\left(\hat{f}(\mathbf{x})\right). \end{aligned} \tag{5}$$

The first term in the second and third line of (5) is an irreducible component of the prediction error (i.e. the variance of Y around its mean $f(\mathbf{X})$), resulting from the random nature of Y. The terms $E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2$ and $\left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2$ are dependent only on the real mean $f(\mathbf{X})$ and its estimator $\hat{f}(\mathbf{X})$. $E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2$ is the variance of $\hat{f}(\mathbf{X})$, characterizing sensitivity of $\hat{f}(\mathbf{X})$ to changes in a learning sample (new observations). $\left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2$ – the square of the bias – is the square of a value, by which the mean estimate $\hat{f}(\mathbf{X})$ differs from its actual mean $f(\mathbf{X})$. It is additionally assumed, that learning samples are of the same size and each time drawn from the same distribution $p(\mathbf{X}, Y)$.

For a given bias, increase in a sample size usually leads to a drop in variance. As large samples are common, in practice it is bias that constitutes the main proportion of prediction error. This observation aroused interest in more flexible methods, that aim at bias reduction and simultaneously prevent a model from overfitting (increased variance). Such approach turned out to be successful in regression, but brought disappointing results in classification setting. Startling research results (see [2; 4]), stating that simple methods in a classification problem are no worse or often perform better than more sophisticated ones, encouraged new direction of research and explain their resilience. The next paragraph presents the decomposition of the expected prediction error in classification by J.H. Friedman that provides a coherent conceptual framework elucidating specifics of a supervised classification problem and indicating new ways of improving classifiers. A concise account of other approaches to bias-variance decomposition in supervised classification may be found in [6].

4. Bias-variance trade-off in classification setting – Friedman's decomposition

The decomposition of the expected prediction error was proposed by J.H. Friedman in 1997 (see [4, pp. 55-77]). It concerns the case of two categories decoded as a random variable Y, which assumes two values 0 and 1, $Y \in \{0, 1\}$, and zero-one loss function (the decomposition can be generalized into any loss function). It is assumed that at every point x the variable Y follows a distribution defined by probabilities

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = 1 - P(Y = 0|\mathbf{X} = \mathbf{x}) = p_1(\mathbf{x}). \tag{6}$$

The expected prediction error in independent test sample has the form

$$Err(\mathbf{x}) = \pi_0 P(\hat{d}(\mathbf{x}) = 1|0) + \pi_1 P(\hat{d}(\mathbf{x}) = 0|1) = P(Y \neq \hat{d}(\mathbf{x})|\mathbf{X} = \mathbf{x}). \tag{7}$$

Considering the fact that there are only two categories, Bayes classifier and its approximation can be written as

$$d^*(\mathbf{x}) = I(p_1(\mathbf{x}) > 1/2), \tag{8}$$

$$\hat{d}(\mathbf{x}) - I(\hat{p}_1(\mathbf{x}) > 1/2). \tag{9}$$

The classifier estimate (9) has the full form $\hat{d}(\mathbf{x}|T) = I(\hat{p}(\mathbf{x}|T) > 1/2)$, but the shorter notation will be kept for clarity.

Expanding (7) one obtains

$$\begin{aligned} Err(\mathbf{x}) &= P(Y \neq \hat{d}(\mathbf{x})) = \\ &= P(Y = d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x})) + P(Y \neq d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) = d^*(\mathbf{x})) = \\ &= P(Y = d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x})) + P(Y \neq d^*(\mathbf{x}))\left[1 - P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\right] = \\ &= P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\left[P(Y = d^*(\mathbf{x})) - P(Y \neq d^*(\mathbf{x}))\right] + P(Y \neq d^*(\mathbf{x})) = \\ &= P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\left[1 - 2P(Y \neq d^*(\mathbf{x}))\right] + P(Y \neq d^*(\mathbf{x})). \end{aligned} \tag{10}$$

Let $Err_B(\mathbf{x}) = P(Y \neq d^*(\mathbf{x}))$ denote the irreducible *bayesian error rate* at a point $\mathbf{x}$ – the analogue of the random component's variance in regression decomposition. Then one can write:

$$Err(\mathbf{x}) = P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))[1 - 2Err_B(\mathbf{x})] + Err_B(\mathbf{x}). \tag{11}$$

From (11) it is seen that the bias and the variance influence the expected prediction error not additively, as in regression, but in a multiplicative way. The expected prediction error consists of a noise component and the product of two elements: one which depends on the noise and $P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right)$ – the analogue of $E_T\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2$ from (5).

One can also notice that for a given training sample T the error $Y \neq \hat{d}(\mathbf{x})$ is determined by an agreement between the decision (9) and the bayesian decision (8). In case of an agreement, the Bayesian error $Err_B(\mathbf{x}) = P\left(Y \neq d^*(\mathbf{x}|T)\right) = \min\{p_1(\mathbf{x}), 1 - p_1(\mathbf{x})\}$ is incurred, otherwise one can expect an increased error rate $P\left(Y \neq \hat{d}(\mathbf{x}|T)\right) = \max\{p_1(\mathbf{x}), 1 - p_1(\mathbf{x})\} = |2p_1(\mathbf{x}) - 1| + Err_B(\mathbf{x})$, which can be jointly written as

$$P\left(Y \neq \hat{d}(\mathbf{x}|T)\right) = |2p_1(\mathbf{x}) - 1| I\left(\hat{d}(\mathbf{x}|T)\right) + Err_B(\mathbf{x}). \tag{12}$$

Averaging (12) over all possible samples T one concludes that equivalently to (11) $Err(\mathbf{x})$ may be shown as

$$Err(\mathbf{x}) = |2p_1(\mathbf{x}) - 1| P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right) + Err_B(\mathbf{x}). \tag{13}$$

Friedman argues that it is reasonable to assume a normal distribution of the estimate $\hat{p}_1(\mathbf{x})$. One has therefore $\hat{p}_1(\mathbf{x}) \sim N\left(E\hat{p}_1(\mathbf{x}), Var\hat{p}_1(\mathbf{x})\right)$, where $E\hat{p}_1(\mathbf{x})$ and $Var\hat{p}_1(\mathbf{x})$ can be written in analogy to (4a-b), and the term $P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right)$ can be expressed as

$$P\left(\hat{d} \neq d^*\right) \approx I(p_1 < 1/2) \int_{0,5}^{\infty} c(\hat{p}_1) d\hat{p}_1 + I(p_1 \geq 1/2) \int_{-\infty}^{0,5} c(\hat{p}_1) d\hat{p}_1 . \tag{14}$$

Equation (14) restricts calculation to a specific point $\mathbf{X} = \mathbf{x}$ and $c\left(\hat{p}_1\right)$ denotes a density function of variable $\hat{p}_1$. Alternatively to (14) one has

$$P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right) \approx \Phi\left(\frac{\text{sign}(1/2 - p_1(\mathbf{x}))(E\hat{p}_1(\mathbf{x}) - 1/2)}{\sqrt{Var\,\hat{p}_1(\mathbf{x})}}\right), \tag{15}$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function.

The value $\text{sign}(1/2 - p_1(\mathbf{x}))(E\hat{p}_1(\mathbf{x}) - 1/2)$ can be thought of as the *boundary bias*, for it is dependent on $p_1(\mathbf{x})$ through its location against the boundary 1/2. From (15) it is clear that if both $E\hat{p}_1(\mathbf{x})$ and $\hat{p}_1(\mathbf{x})$ are on the same side of the

boundary 1/2, then the boundary bias is negative and lowering the variance should cause the bias to decrease down to the minimal Bayes error rate. If $E\hat{p}_1(\mathbf{x})$ and $p_1(\mathbf{x})$ are on the opposite sides of the boundary, then the bias is positive and it seems advisable to increase the variance, as it leads to a drop in prediction error (see [5], p. 223). It is naturally preferable that the boundary bias be negative. If this is the case, then the classification error decreases as the value $|E\hat{p}_1(\mathbf{x}) - 1/2|$ increases, irrespective of the bias $p_1 - E\hat{p}_1$. This notice supports conclusion, that the main concern in a two-class supervised classification problem is not keeping the bias small, but retaining low variance, provided the boundary bias is predominantly negative.

Some highly biased methods (in the sense of the squared loss function) produce satisfying results in classification setting. Friedman points at a group of methods, which large bias is caused by excessive smoothing (oversmoothing). It is said, that a method is oversmoothing, when the estimate

$$\hat{p}_1(\mathbf{x}) = (1 - \alpha(\mathbf{x}))p_1(\mathbf{x}) + \alpha(\mathbf{x})\bar{y}$$

has a tendency to assume values close to the mean value of the dependent variable Y, which takes place when $\alpha(\mathbf{x})$ – where $\alpha(\mathbf{x}) \in [0,1]$ is a smoothing parameter – assumes values close to 1. As long as the decision boundary equals $\bar{y}$, the boundary bias remains negative (in our case the sample is balanced, i.e. $\bar{y} = 1/2$). One such method is the nearest neighbours method, in which the approximation of $p_1(\mathbf{x})$ consists in averaging class indicators of the k closest observations in a training sample. When $k \to n$ then $\alpha(\mathbf{x}) \to 1$, which entails $\hat{p}_1(\mathbf{x}) \to \bar{y}$.

The following points recap some general conclusions that can be used in practice.

- Decomposition of the expected prediction error is much more complex in supervised classification setting and unveils complicated, non-additive interplay between its components;
- Friedman's analysis, despite a two-class limitation, helps to explain the competitivness of classifiers, which base on biased probability estimators (ex. naive Bayes classifier, linear probability model);
- More accurate probability estimation does not necessarily lead to better classification results;
- Imposition of a constraint $\hat{p}_1(\mathbf{x}) \in [0,1]$ might decrease estimation bias, but simultaneously it may pose the danger of a boundary bias rise;
- It seems that in practical applications (ex. credit scoring) one should use accurate classification methods to generate a *score*, but if there is only a need for a classification decision, the aforementioned, simple methods would suffice.

The next section gives an example of Friedman's decomposition for three econometric models.

5. Comparison of econometric models with different bias via Friedman's decomposition – simulated data example

In this paragraph Friedman's decomposition of the expected prediction error will be used to compare classification properties of three econometric models: linear probability model (LPM), linear logistic regression model (GLM-Logit) and generalized additive logistic regression model (GAM-Logit). Linear probability models represent a class of simple, highly biased models. Linear logistic regression models are more advanced and less biased models. Generalized additive logistic regression models represent a group of new methods with low bias and adjustable variance. Using the generalized linear models notation $\eta = g\left[E\left(Y|\mathbf{X}\right)\right]$ (in case of binary data $E\left(Y|\mathbf{X}=\mathbf{x}\right)=p_1(\mathbf{x})$), where η is a predictor (function of $\mathbf{X}$) and $g(\bullet)$ is a link function, the three models may be written as follows:

- LPM: $\eta=\alpha_0+\alpha_1 X$, $g\left[E\left(Y|X\right)\right]=E\left(Y|X\right)$ (identity function),
- GLM-Logit: $\eta=\alpha_0+\alpha_1 X$, $g\left[E\left(Y|X\right)\right]=\text{logit}\left[E\left(Y|X\right)\right]=\ln\left[E\left(Y|X\right)/\left(1-E\left(Y|X\right)\right)\right]$,
- GAM-Logit: $\eta=\alpha+f(X)$, $g\left[E\left(Y|X\right)\right]=\text{logit}\left[E\left(Y|X\right)\right]=\ln\left[E\left(Y|X\right)/\left(1-E\left(Y|X\right)\right)\right]$.

In GAM-Logit $f(\bullet)$ is a smooth and nonparametric function, while α, α_0, α_1 are parameters. In this simulation example the ordinary least squares estimator was used for LPM, maximum likelihood estimator for GLM-Logit and iteratively re-weighted least squares algorithm estimator (see [3, p. 240]) with a smoothing spline as a scatterplot smoother for GAM-Logit. Three degrees of freedom were assumed for GAM-Logit smoothing spline. LPM and GLM-Logit models were both estimated using *glm* package in *R-project* environment. *Gam* package was used to estimate generalized additive logit models.

Friedman's expected prediction error decomposition will be used to estimate classification properties of the three classifiers. A mean deviation of residuals will be employed to bias measurement.

Suppose that X is non-random and takes values on the real line [0; 1] at 0,001 intervals starting from 0 (i.e. $x_1=0$, $x_2=0{,}001$, $x_{1000}=0{,}999$), and the Y-generating mechanism has the following form ($Y\in\{0;1\}$):

$$P\left(Y=1|X\right)=0{,}8I\left(X\in[0;0{,}5)\right)+0{,}2I\left(X\in[0{,}5;1]\right),$$

then the best classifier is given by

$$\hat{p}_1(x)>0{,}5 \text{ if } x\in[0;0{,}5) \text{ and } \hat{p}_1(x)\le 0{,}5 \text{ if } x\in[0{,}5;1],$$

and Bayes classifier can be written as $d^*(x) = I\{x \in [0; 0,5)\}$. The calculations of *Err* will be carried out in the proximity of a boundary 1/2, at the point $x = 0,45$. At this point the Bayes error $Err_B(0,45) = 0,2$. The approximate value of $P\left[\hat{d}(0,45) \neq d^*(0,45)\right]$ may be obtained from (14). One then has

$$P\left[\hat{d}(0,45) \neq d^*(0,45)\right] \approx \int_{-\infty}^{0,5} c(\hat{p}_1) d\hat{p}_1 . \tag{16}$$

Following Friedman, the distribution of $\hat{p}_1(0,45)$ will be modelled by normal distribution with expected value and standard deviation estimated in a simulation fashion. The simulation results are based on 10.000 replications.

Figure 2 shows mean estimates $\hat{p}_1$ with two standard deviations confidence bands for the three models, depicting the variability of the estimates $\hat{p}_1$.

Figure 3 shows the plots of the estimated probability density for the three distributions and Table 1 summarizes the simulation results. The last column contains areas under the probability density function plots of $c(\hat{p}_1)$ on $(-\infty; 0,5]$. The distribution of $\hat{p}_1$ is characterized by the parameters given in the first two columns.

Table 1. Simulation results at point $x = 0,45$ (10.000 replicates)

Model	$\hat{E}\hat{p}_1$	$\sqrt{\hat{V}ar\,\hat{p}_1}$	$10^{-4}\sum_{i=1}^{10000}\lvert y_i - \hat{p}_{1i}\rvert$	$P\left[\hat{d}(0,45) \neq 1\right]$
LPM	0,5400	0,0405	0,4721	0,1619
GLM-Logit	0,5499	0,0568	0,4645	0,1903
GAM-Logit	0,6077	0,0821	0,4162	0,0948

Table 2 contains the estimates of the expected prediction error in classification and its components. As an example the error for the linear probability model was calculated using the equation (11):

$$Err(0,45) = 0,1619 \times (1 - 2 \times 0,2) + 0,2 = 0,2971.$$

Table 2. Expected prediction error in classification and its components at point $x = 0,45$

Model	Boundary bias $sign(1/2 - p_1(x))(E\hat{p}_1(x) - 1/2)$	$\sqrt{\hat{V}ar\,\hat{p}_1}$	Err_B	$Err(0,45)$
LPM	−0,0400	0,0405	0,2	**0,2971**
GLM-Logit	−0,0499	0,0568	0,2	**0,3142**
GAM-Logit	−0,1077	0,0821	0,2	**0,2569**

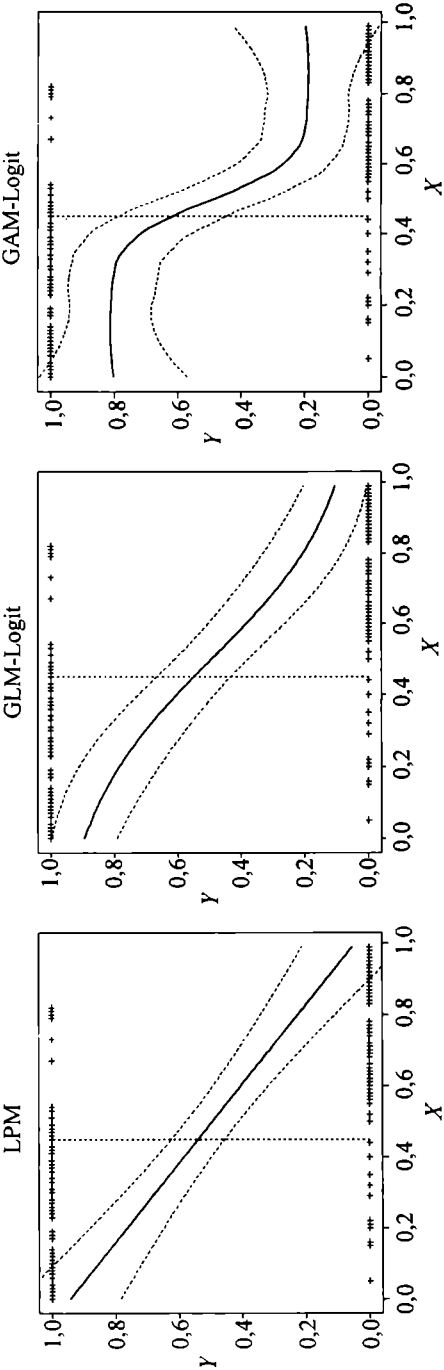


Fig. 2. Mean estimates $\hat{p}_1$ by linear probability model (LPM), linear logit model (GLM-Logit) and additive logit model (GAM-Logit) with two standard deviations confidence bands

From tables 1 and 2 one sees that at point $x = 0{,}45$ the boundary bias is negative in each case. The boundary bias of LPM and GLM-Logit are nearly equal and, as expected, LPM has the highest boundary bias, while GAM-Logit – the smallest. The GAM-Logit estimates are the most volatile, while the estimates of LPM – the most stable.

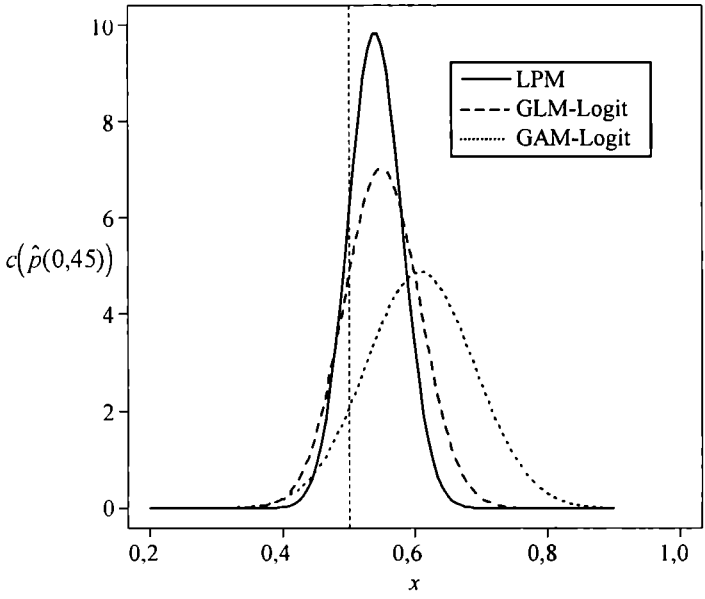


Fig. 3. Probability estimate density function $c(\hat{p}(0{,}45))$ of linear probability model (LPM), linear logistic regression model (GLM-Logit) and additive logistic regression model (GAM-Logit). The values (16) are equal to the areas under the function plot to the left of the dotted line

In the simulation experiment the linear probability model, which has the greatest bias (in the sense of the squared loss function), gave lower classification error rate than less biased linear logit model. This is caused by the smaller variance of the linear probability model, which neutralized the effect of a greater boundary bias. Although the GAM-Logit has the greatest variance, it gives lower boundary bias and the overall prediction error is the lowest of the three models. Figure 4 shows the expected prediction error estimated for the whole realm of X The conclusion is that the Err differences between the three models are immaterial.

Let now change the Y-generating mechanism to the form:

$$P(Y = 1|X) = 0{,}9I(X \in [0; 0{,}340]) + 0{,}2I(X \in [0341; 0{,}669]) + 0{,}9I(X \in [0{,}670; 0{,}999]),$$

so that probability $P(Y=1|X)$ is not a monotonically changing function of X. This is often the case in real classification problems, where some of the explanatory variables are nominants. The expected prediction error estimates, boundary bias and variance for the three models in the new scenario are shown in figure 5.

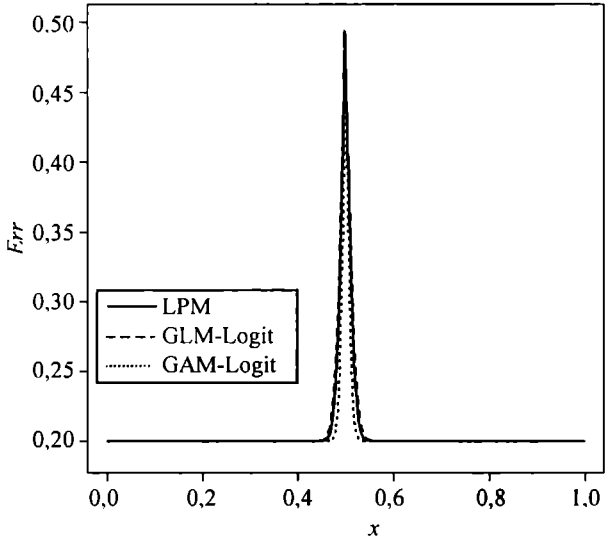


Fig. 4. Expected prediction error estimates for the whole realm of X in the first scenario

The problematic area for LPM and GLM-Logit is $X \in [0{,}341; 0{,}669]$, where the so called „masking" phenomenon occurs – the positive boundary bias is accompanied by low variance of the estimate of $\hat{p}_1$ and $Err = 1 - Err_B$. Two possible solutions to the problem are:

1) discretization of the variable X (most often used in business practice) within LPM or GLM-Logit;

2) using more accurate models like GAM-Logit, which – despite high variability – keep boundary bias negative.

The simulation results are, therefore, consistent with the claim, that there are classification situations, where simple probability estimators remain still competitive (figure 4 is an example). One of the prerequisites is that the variables are stimulants or destimulants. In other case one should resort to more flexible models to obtain negative bias.

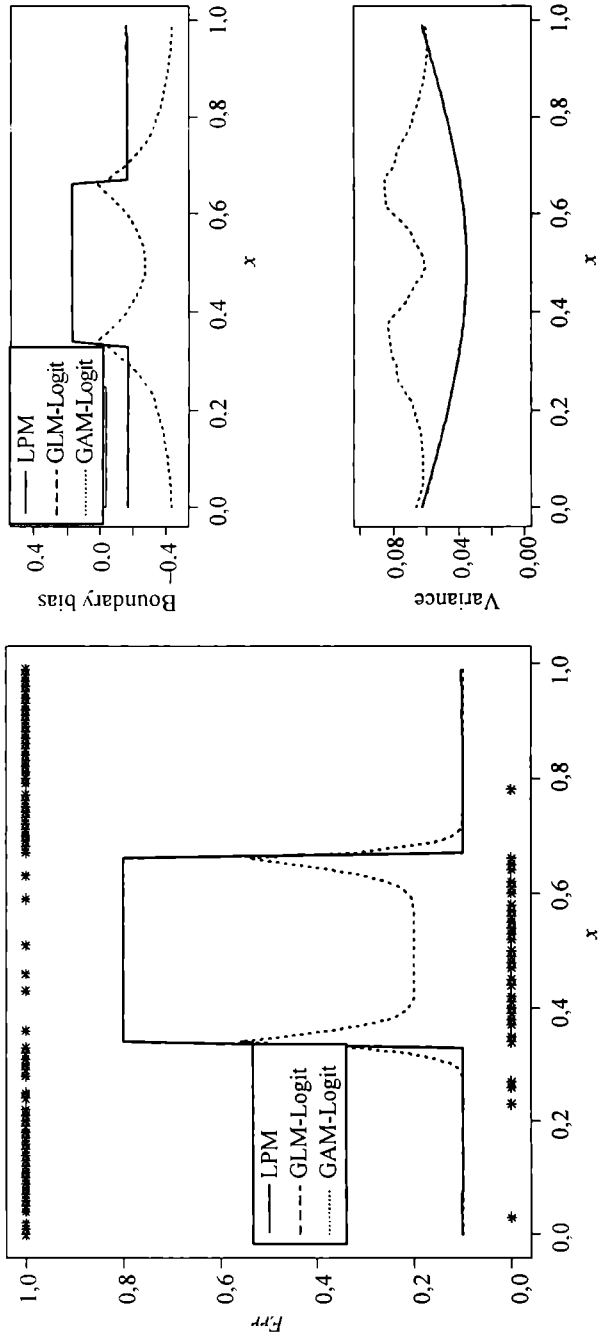


Fig. 5. Expected prediction error estimates for the whole realm of X and its components in the second scenario

Literature

[1] Ćwik J., Koronacki J., *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa 2005.

[2] Domingos P., *A Unified Bias-Variance Decomposition for Zero-One and Squared Loss*, Austin (USA): AAAI Press, Proceedings of the Seventeenth National Conference on Artificial Intelligence 2000, pp. 564-569.

[3] Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Press, London 2006.

[4] Friedman J.H., *On Bias, Variance, 0/1-loss, and the Curse-of-dimensionality*, Kluwer Academic Publishers: Data Mining and Knowledge Discovery 1 1997, pp. 55-77.

[5] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York 2003.

[6] Le Borgne Yann-Ael, *Bias-Variance Trade-off Characterization in a Classification Problem. What Differences with Regression?* Machine Learning Group, Université Libre de Bruxelles 2005.

WYMIENNOŚĆ WARIANCJI I OBCIĄŻENIA W MODELU KLASYFIKACJI POD NADZOREM

Streszczenie

W artykule zaprezentowano podejście do dekompozycji oczekiwanego błędu predykcji w klasyfikacji według J.H. Friedmana. Dekompozycja ta ujawnia multiplikatywną wymienność wariancji i obciążenia w modelu klasyfikacji pod nadzorem oraz pozwala wyjaśnić klasyfikacyjną konkurencyjność prostych, obciążonych modeli, takich jak np. liniowy model prawdopodobieństwa. W artykule przedstawiono również symulacyjny przykład obliczenia oczekiwanego błędu predykcji w klasyfikacji za pomocą dekompozycji Friedmana, porównujący trzy modele ekonometryczne o różnym obciążeniu.

Piotr Michalski – mgr, doktorant w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Aneta Rybicka

OPROGRAMOWANIE KOMPUTEROWE WYKORZYSTYWANE W BADANIACH PREFERENCJI KONSUMENTÓW METODAMI DEKOMPOZYCYJNYMI

W badaniach preferencji konsumentów wykorzystujemy m.in. metody dekompozycyjne. Podejście to reprezentują metody *conjoint analysis* oraz metody wyborów dyskretnych. Na rynku obecnie oferowane są różne oprogramowania komputerowe pozwalające na przeprowadzenie badań preferencji konsumentów z wykorzystaniem tychże metod.

W artykule zamieszczono charakterystykę wybranych programów komputerowych: w badaniach wykorzystujących metody *conjoint analysis* – SPSS® 10.05 for Windows, Conjoint Analysis for Windows, CVA oraz CVA/HB; natomiast w badaniach z wykorzystaniem metod wyborów dyskretnych – SAS/STAT, CBC, CBC Advanced Design Module, CBC/HB, The Latent Class Segmentation Module oraz Individual Choice Estimation (ICE) Module.

W badaniach z wykorzystaniem metod wyborów dyskretnych stosujemy m.in. oprogramowanie SAS/STAT oraz Sawtooth Software.

Oprogramowanie **SAS/STAT** (SAS – *statistical analysis system*) pozwala na wykorzystanie danych pochodzących z różnych źródeł, m.in. z eksperymentów klinicznych, marketingowych baz danych, badań dotyczących zdrowia, analiz preferencji konsumentów, informacji pochodzących z giełd itd. SAS/STAT oferuje techniki statystyczne znajdujące zastosowanie w wielu dziedzinach (*SAS/STAT Software* 2006; *SAS/STAT® Software. Providing the foundation for SAS` analytic intelligence* 2005), jak np.:

- przemysł,
- telekomunikacja,
- zarządzanie,
- badania dotyczące środowiska,

- biotechnologia,
- handel detaliczny,
- skażenie powietrza,
- używanie kart kredytowych.

Niektóre zadania obliczeniowe mogą być realizowane w trybie menu i oknie dialogowym z wykorzystaniem modułów SAS/Base oraz SAS/Enterprise Guide [2, s. 220-221]. Natomiast realizacja pozostałych zadań możliwa jest w trybie programowym, w którym należy skorzystać z języka programowania SAS.

W pakiecie tym oferowane są procedury, które pozwalają m in. na generowanie pełnych i cząstkowych układów czynnikowych oraz na estymację użyteczności cząstkowych.

W pierwszym etapie można wykorzystać trzy rodzaje procedur [2, s. 223; 11]:

- PROC PLAN – jest to procedura umożliwiająca realizację pełnego układu czynnikowego,
- PROC FACTEX – jest to procedura pozwalająca na: realizację ortogonalnego układu czynnikowego (pełnego lub cząstkowego); tworzenie bloków, w której czynniki mogą zawierać taką samą lub różną liczbę poziomów; randomizację układu oraz wybór poziomu rozdzielczości układu,
- PROC OPTEX – procedura, która: generuje optymalny cząstkowy układ czynnikowy, umożliwia wybór kryterium optymalizacji (A, D, G, U, S), umożliwia ustalenie liczby generowanych profilów, umożliwia tworzenie bloków oraz pozwala na wybór poziomu rozdzielczości układu.

Natomiast do estymacji użyteczności cząstkowych wykorzystywane są [2, s. 223; 11]:

- PROC TRANSREG – procedura umożliwiająca kodowanie zmiennych niemetrycznych za pomocą zmiennych sztucznych,
- PROC PHREG (*proportional hazards regression*) – jest to procedura pozwalająca na: dopasowanie modelu proporcjonalnego hazardu Coxa, szacowanie parametrów wielomianowego (warunkowego) modelu logitowego, warstwowanie danych, wykorzystanie układu pełnego lub cząstkowego oraz dopasowanie modelu efektów głównych lub z interakcjami.

Procedury PROC FACTEX oraz PROC OPTEX zostały napisane przez Tobiasa, PROC PHREG napisał Ying So, natomiast Kuhfeld jest autorem procedury PROC PTRANSREG. Są to najczęściej stosowane procedury, dostępne są jednak również inne procedury, m.in.: PROC CATMOD, PROC FORMAT, PROC FREQ, PROC GLM, PROC GPLOT, PROC IML, PROC LOGISTIC.

Sawtooth Software oferuje kilka programów umożliwiających przeprowadzenie badania preferencji konsumentów metodą wyborów dyskretnych. Jednym z takich programów jest **CBC** (*Choice-based Conjoint*)[1].

[1] Fragment przedstawiający system CBC opracowany został na podstawie *Choice-based Conjoint (CBC) Technical Paper* (zob. [3]).

Jedną z zalet przeprowadzenia badania preferencji metodą wyborów dyskretnych na poziomie zagregowanym jest to, że umożliwia ono uwzględnienie interakcji[2]. Większość metod *conjoint* wykorzystuje zazwyczaj modele efektów głównych, które ignorują występowanie interakcji. CBC pozwala na otrzymanie relatywnie precyzyjnych rezultatów, wyników badania, w których wykorzystujemy niewiele atrybutów oraz w którym rozpatrujemy (uwzględniamy) interakcje.

System CBC pozwala na przeprowadzenie badania z wykorzystaniem metod wyborów dyskretnych. System ten charakteryzuje się pewnymi cechami. Pozwala on m.in. na:

- zaprojektowanie i konstrukcję ankiety w formie papierowej bądź też w środowisku Windows,
- administrowanie ankietami respondentów,
- przeprowadzenie analizy otrzymanych danych.

System CBC może zawierać zarówno losowy, jak i stały projekt eksperymentu. W projekcie losowym badacz określa kilka szczegółów, takich jak: liczba zbiorów profilów, które mają być przedstawione każdemu respondentowi, liczba profilów, które ma zawierać każdy ze zbiorów oraz sposób, w jaki będą one ułożone (uporządkowane) na ekranie, po czym ankieta jest automatycznie generowana dla każdego respondenta. Jeśli w badaniu wykorzystujemy stały projekt eksperymentu, badacz musi wyszczególnić ten projekt. Możliwy jest również eksperyment mieszany, w którym pewne zbiory są skonstruowane losowo, inne zaś są stałe.

W systemie CBC badacz musi określić atrybuty oraz ich poziomy, musi także skonstruować tekst objaśniający, który będzie wyświetlany respondentowi na ekranie. Oprócz tych zadań pozostałe mogą być wykonane automatycznie.

Moduł ankiety (*questionnaire module*) charakteryzuje się pewnymi cechami.

- Wywiad może zawierać nielimitowaną liczbę ekranów z informacjami, które wykorzystywane są w dowolnym momencie wywiadu w celu wyjaśnienia atrybutów bądź przedstawienia grupom zbioru. Badacz określa tekst, kolor oraz wzór czcionki edytora na ekranach.
- W wywiadzie może być zawarte do 10 pytań segmentacyjnych (*segmentation questions*). Są to pojedynczo wybrane pytania, które mogą zawierać informacje demograficzne bądź informacje dotyczące przedsiębiorstwa, takie jak rodzaj czy rozmiar przedsiębiorstwa. Każde pojedynczo wybrane pytanie może wykorzystać do 9 kategorii odpowiedzi. Dane pochodzące z tych pytań są automatycznie przeprowadzane przez wszystkie stopnie analizy, do wykorzystania jako filtry bądź też jako zmienne wagowe (*weighting variables*).

[2] Interakcje są brane pod uwagę wtedy, gdy atrybuty są wyrażone w sposób numeryczny, tak jak cena, wartość czy ilości, oraz gdy estymacja jest bardzo precyzyjna, jak w sytuacjach, gdy udzielane są odpowiedzi na wiele pytań i oszacowujemy kilka parametrów. Te oba warunki charakteryzują zazwyczaj badania cen (*pricing studies*). Możliwość uwzględnienia iteracji charakteryzuje przede wszystkim metody, które dostarczają informacji na poziomie zagregowanym (*aggregate methods*).

- Wywiad może zawierać niemal nielimitowaną liczbę zbiorów (jednakże sugeruje się, że liczba zbiorów dużo powyżej 20 może być zbyt duża do rozpatrzenia dla każdego z respondentów[3]). Każdy ze zbiorów przedstawia dwa lub więcej profilów opisanych wybranymi atrybutami i ich poziomami. Zbiór może zawierać do 16 profilów (bądź też do 15 plus opcja rezygnacji z wyboru).
- Różne rozmiary oraz typy zbiorów profilów mogą być wykorzystane w tym samym wywiadzie.
- Profile mogą być opisane liczbą atrybutów nie większą niż 10. Każdy z atrybutów może mieć do 15 poziomów.
- Może być niedozwolone pojawienie się specyficznych kombinacji poziomów atrybutów we wszystkich profilach. Warunek ten jest wprowadzony po to, by mieć pewność, że niezależność atrybutów nie jest naruszona oraz że efekty główne są możliwe do oszacowania.
- Ceny dla otrzymanych profilów mogą być uzależnione od pozostałych poziomów atrybutów.
- Pliki graficzne oraz wideo mogą być również włączone do kwestionariusza, by przedstawić pewne poziomy atrybutów.

Są dwa sposoby prezentowania zbiorów profilów w systemie CBC. Pierwszy z nich pozwala na przedstawienie profilów w poziomie, czyli jedna kolumna – jeden profil (sposób ten charakteryzuje się wykorzystaniem kilku linii do opisu każdego z profilów, lecz tekst w każdej z linii musi być możliwie najkrótszy).

Tabela 1. Pozioma prezentacja profilów

1	2	3	4
Kawa Jacobs	Kawa Tchibo	Kawa Elite	Żadna z wymienionych
Kawa mielona	Kawa rozpuszczalna	Kawa mielona	
Opakowanie próżniowe	Opakowanie szklane	Opakowanie miękkie	
Kawa kofeinowa	Kawa bezkofeinowa	Kawa kofeinowa	
Opakowanie 500 g	Opakowanie 200 g	Opakowanie 250 g	
W zestawie promocyjnym	W zestawie promocyjnym	W zestawie promocyjnym	
15,50 zł	23,20 zł	6,50 zł	
Kawa lekko palona	Kawa mocno palona	Kawa mocno palona	

Źródło: opracowanie własne.

Drugi ze sposobów przedstawia profile w pionie, czyli jeden wiersz – jeden profil (w ten sposób każdy z profilów opisany jest kilkoma liniami, lecz każda z linii może być dość długa) – zob. tab. 2.

[3] Johnson i Orme [9] twierdzą, że ankieta, która składa się minimum z 20 zbiorów nie powoduje utraty jakości danych.

Badacz może wybrać jeden ze sposobów prezentacji profilów, w zależności od liczby profilów, które będą zawarte w każdym zbiorze, oraz od charakterystyk atrybutów, które będą opisane tekstem. Możliwa jest również prezentacja zbiorów profilów w złożone wiersze i kolumny (np. 3 × 3 przedstawia 9 profilów).

Gdy w badaniu wykorzystuje się tylko kilka zbiorów profilów oraz gdy tekst jest krótki, preferowany jest poziomy sposób prezentacji zbiorów profilów. W tej sytuacji łatwiej jest respondentom „przebiegać wzrokiem". Respondent może dokonać wyboru nie tylko z wykorzystaniem klawiatury, lecz również z wykorzystaniem myszy, klikając w obrębie pola profilu.

System CBC powinien być wykorzystany w badaniach, w których rozpatrywanych jest tylko kilka atrybutów oraz gdy będą uwzględniane interakcje (obie te sytuacje charakteryzują często badania cenowe). System ten znajduje zastosowanie również w wielu innych badaniach, tu przede wszystkim w badaniach preferencji.

Tabela 2. Pionowa prezentacja profilów

1	Kawa Jacobs
	Kawa mielona
	Opakowanie próżniowe
	Kawa kofeinowa
	Opakowanie 500 g
	W zestawie promocyjnym
	15,50 zł
	Kawa lekko palona
2	Kawa Tchibo
	Kawa rozpuszczalna
	Opakowanie szklane
	Kawa bezkofeinowa
	Opakowanie 200 g
	W zestawie promocyjnym
	23,20 zł
	Kawa mocno palona
3	Kawa Elite
	Kawa mielona
	Opakowanie miękkie
	Kawa kofeinowa
	Opakowanie 250 g
	W zestawie promocyjnym
	6,50 zł
	Kawa mocno palona

Źródło: opracowanie własne.

Badania opublikowane latem w 2004 r. pokazują, że aż 61% badań[4] z użyciem oprogramowania Sawtooth Software, to badania wykorzystujące system CBC [14, s. 1]. System ACA (*Adaptive Conjoint Analysis*) wykorzystywano w 27% przeprowadzonych badań, system CVA (*Conjoint Value Analysis*) natomiast w 12%.

Innym modułem Sawtooth Software jest **CBC (*Advanced Design Module* – ADM)**. System ten pozwala m in. na [13, s. 2-3; 17, s. 1]:

- plany ze zmiennymi specyficznymi[5] (*alternative-specific plans*),
- format wywiadu z wykorzystaniem profilów częściowych,
- powiększenie rozmiaru badania do 30 atrybutów,
- powiększenie badania do 254 poziomów dla każdego z atrybutu oraz do 100 profilów w zbiorze (tylko w CBC/Web),

[4] Spośród tych badaczy 62% wykorzystuje HB w celu estymacji ostatecznego modelu [14, s. 5].

[5] Projekt z wykorzystaniem zmiennych specyficznych (*alternative-specific*) pozwala na wykorzystanie atrybutów, które opisane są tylko niektórymi zmiennymi, różniącymi się poziomami (np. szybkość drukowania drukarki czarno-białej: 20 stron na minutę, 30 stron na minutę, 40 stron na minutę; natomiast szybkość drukowania drukarki kolorowej: 10 stron na minutę, 15 stron na minutę, 20 stron na minutę) [17, s 2].

- przedstawienie (prezentację) zbiorów profilów w formie „towarów na półkach" (*shelf-facing display*) (tylko CBC/Web).

Atrybuty mogą pełnić różną funkcję w projekcie ze specyficznymi zmiennymi [17, s. 7]:

- atrybut podstawowy, główny (*primary attribute*): jest to atrybut, którego poziomy pojawiają się w każdym profilu,
- atrybut warunkowy (*conditional attribute*): występuje tylko ze szczególnym poziomem bądź poziomami atrybutu podstawowego,
- atrybut wspólny (*common attribute*): występuje ze wszystkimi poziomami atrybutu podstawowego (w tradycyjnym systemie CBC, w projekcie wszystkie atrybuty są wspólne).

Przykładem projektu ze zmiennymi specyficznymi mogą być preferencje dotyczące transportu z domu do pracy [17, s. 6-7]. Można zaproponować atrybuty dla transportu samochodem bądź autobusem:

- samochód: opłata parkingowa 5 dolarów dziennie, opłata parkingowa 8 dolarów dziennie, opłata parkingowa 10 dolarów dziennie;
- autobus: odjeżdżający co 20 minut, odjeżdżający co 15 minut, odjeżdżający co 10 minut, odjeżdżający co 5 minut;
- autobus: 25 centów za bilet jednorazowy, 50 centów za bilet jednorazowy, 75 centów za bilet jednorazowy, 1 dolar za bilet jednorazowy.

Zakładamy, że chcemy dokonać pomiaru również trzech innych (stałych) profilów: spacer piechotą, rower, „wybiorę inny sposób transportu do pracy".

Pytanie zadane respondentowi mogłoby wyglądać następująco:

„Jeśli mieszkałbyś jedną milę od swojego miejsca pracy w centrum miasta, a to by były twoje możliwe alternatywy dotarcia do pracy, którą z nich być wybrał?

1. Transport autobusem, odjeżdżający co 10 minut, 1 dolar za bilet jednorazowy.
2. Rower.
3. Transport własnym samochodem, opłata parkingowa 8 dolarów dziennie.
4. Spacer piechotą.
5. Wybiorę inny sposób transportu do pracy".

Zatem do projektu z wykorzystaniem CBC ADM musimy zdefiniować atrybuty następująco:

Atrybut 1:

- Spacer piechotą
- Rower
- Transport autobusem
- Transport własnym samochodem

Atrybut 2:

- Odjeżdżający co 20 minut

- Odjeżdżający co 15 minut
- Odjeżdżający co 10 minut
- Odjeżdżający co 5 minut

Atrybut 3:

- 25 centów za bilet jednorazowy
- 50 centów za bilet jednorazowy
- 75 centów za bilet jednorazowy
- 1 dolar za bilet jednorazowy

Atrybut 4:

- Opłata parkingowa 5 dolarów dziennie
- Opłata parkingowa 8 dolarów dziennie
- Opłata parkingowa 10 dolarów dziennie

W takim projekcie atrybut 1 jest atrybutem podstawowym. Atrybuty 2, 3 i 4 są atrybutami warunkowymi. Natomiast w projekcie tym nie występuje atrybut wspólny. Zakazy, zależności (*prohibitions*) występują tutaj między wszystkimi poziomami atrybutu (bądź atrybutów) warunkowego (atrybuty 2, 3 i 4) a szczególnym poziomem (bądź poziomami) atrybutu podstawowego (atrybut 1).

Estymacja z wykorzystaniem projektu zawierającego zmienne specyficzne jest możliwa w systemie CBC, w module Logit (ograniczenie do 30 atrybutów), w module Latent Class (ograniczenie maksymalnie do 100 atrybutów), w module ICE (ograniczenie do 10 atrybutów) oraz w module HB (ograniczenie do 1000 atrybutów) [17, s. 10].

W module tym możemy również wykorzystać format wywiadu z użyciem profilów częściowych (w modułach wymienionych powyżej, z takimi samymi ograniczenia liczby atrybutów). W takim projekcie respondent dokonuje wyboru między profilami opisanymi atrybutami w liczbie od 3 do 5 (moduł ADM pozwala wykorzystać w badaniu do 30 atrybutów).

Dostępny jest również system ***Adaptive Choice Based Conjoint Analysis*** (ACBC), który pozwala na łączenie elementów *Adaptive Conjoint Analysis* (ACA) oraz CBC [7]. Badania z wykorzystaniem ACBC przedstawili w swoich pracach m.in. Johnson i in. [8; 10].

Kolejnym narzędziem analitycznym jest moduł pozwalający na przeprowadzenie segmentacji z wykorzystaniem modeli klas ukrytych **The Latent Class Segmentation Module**[6] [18, s. 1]. Moduł ten wykorzystujemy wraz z danymi otrzymanymi np. z badania CBC lub też CBC/Web, w celu przydzielenia respondentów o podobnych preferencjach (uzyskanych z wyborów jakich dokonali np. w ankiecie CBC) do poszczególnych segmentów. Moduł ten pozwala na wykorzystanie anali-

[6] Modele klas ukrytych stały się popularne w połowie lat 90. XX wieku. Jedną z zalet tych modeli jest to, że pozwalają na redukcję negatywnego efektu własności modelu logitowego, tzn. *IIA* [18, s. 1].

zy klas ukrytych w celu estymacji użyteczności cząstkowych każdego z segmentów oraz prawdopodobieństwa przynależności każdego respondenta do segmentu.

Moduł ten spełnia podobną funkcję jak moduł Logit w CBC, nie szuka jednak przeciętnych użyteczności cząstkowych dla wszystkich respondentów razem, ale rozpatruje podgrupy respondentów różniące się od siebie i oszacowuje użyteczności cząstkowe dla każdego segmentu. Podgrupy respondentów charakteryzują się tym, że respondenci wewnątrz grup mają podobne preferencje, różnią się natomiast preferencje respondentów w poszczególnych podgrupach.

Modele klas ukrytych przydzielają użyteczności cząstkowe dla każdego z segmentów. Analiza z wykorzystaniem modeli klas ukrytych nie zakłada, że każdy z respondentów całkowicie przynależy do jednej czy też drugiej grupy. Rozpatrywane jest raczej niezerowe prawdopodobieństwo przynależności każdego z respondentów do każdej z grup. Jeśli rozwiązanie jest bardzo dobrze dopasowane do danych, to te prawdopodobieństwa zbliżają się do 0 bądź też do 1.

W ofercie Sawtooth Software dostępny jest również system pozwalający na estymację użyteczności cząstkowych **CBC/HB**, z wykorzystaniem którego przeprowadzana jest estymacja hierarchiczna Bayesa [19, s. 1]. System ten wykorzystuje dane, które są automatycznie eksportowane z systemu CBC lub też CBC/Web. Można również wykorzystać dane zgromadzone w inny sposób, jeśli dane te są dostosowane do odpowiednich wymogów plików wykorzystywanych w systemie.

Allenby i Ginter (zob. [1]) w 1995 r. jako pierwsi opisali estymację użyteczności cząstkowych z wykorzystaniem metody hierarchicznej Bayesa, a następnie Lenk, DeSarbo, Green i Young w 1996 r. (zob. [12]).

HB pozwala na przeprowadzenie estymacji, wykorzystując informacje o wyborach kilku respondentów, a jednocześnie „pożyczając" informacje od innych respondentów [19, s. 1]. ICE również pozwala na „uzupełnienie" w ten sposób brakujących informacji, jednakże badanie z wykorzystaniem HB dokonuje tego efektywniej i wymaga mniej wyborów od każdego z respondentów.

Sawtooth Software oferuje również inny moduł pozwalający na estymację indywidualnych wyborów ***Individual Choice Estimation* (ICE) *Module for Choice-based Conjoint***.

ICE było skonstruowane jako rozbudowa analizy modeli klas ukrytych, mająca pozwolić na lepszą prognozę poprzez efektywniejsze rozpoznanie heterogeniczności preferencji [6, s. 2]. Moduł ten charakteryzuje się kilkoma zaletami. Jedną z nich jest jego „szybkość". Jeśli modele klas ukrytych są wybrane jako „punkt startowy", to wysiłek potrzebny do obliczenia indywidualnych użyteczności jest minimalny. Nawet jeśli badacz zdecyduje, że nie wykorzysta modeli klas ukrytych jako „punktu startowego", ICE może oszacować użyteczności, zaczynając od początku, w rozsądnym czasie. Dla danych otrzymanych od 300 respondentów oraz z 25 poziomami atrybutów rozwiązanie, które mogłoby być zaakceptowane, można uzyskać w kilka minut. ICE jest modułem szybszym niż pozostałe dwa

rozwiązania stosowane w oszacowaniach heterogeniczności respondentów: jest zdecydowanie szybszy niż analiza klas ukrytych oraz nieporównywalnie szybszy niż analiza hierarchiczna Bayesa.

Inną zaletą modułu ICE jest to, że jest on lepszy w „uchwyceniu" heterogeniczności aniżeli modele klas ukrytych. Doświadczenia wskazują, że również metody hierarchiczne Bayesa są lepsze w oszacowywaniu heterogeniczności aniżeli modeli klas ukrytych.

Trzecią zaletą tego modułu jest to, że pozwala badaczom przechodzić z zagregowanej analizy danych pochodzących z wyborów do analizy danych na poziomie indywidualnym. Jednym z problemów badaczy rynkowych jest to, żeby przewidzieć reakcje rynku na złożone kombinacje iteracji, zróżnicowane efekty krzyżowe oraz zmieniające się podobieństwa między produktami. Wydaje się, że wszystkie te problemy mogą być zredukowane z wykorzystaniem modeli na poziomie indywidualnym [6, s. 23].

Procedury metod wyborów dyskretnych są w podobny sposób realizowane w ramach różnych programów statystycznych. Mogą się różnić np. prezentacją profilów respondentom w badaniu, trybem, w jakim się pracuje (niektóre programy pozwalają na pracę w trybie okienek dialogowych, inne zaś wymagają programowania). Niektóre z nich pozwalają tylko na oszacowanie preferencji na poziomie zagregowanym (np. CBC, CBC Advanced Design Module), niektóre tylko na poziomie segmentowym (The Latent Class Segmentation Module, GLIMMIX, Latent GOLD), inne zaś na poziomie indywidualnym (CBC/HB, Individual Choice Estimation Module). Wybór konkretnego programu zależy od celu badań. Należy też podkreślić, że większość tych pakietów jest dostępna na zasadach komercyjnych, a ceny nie należą do niskich.

W badaniach preferencji konsumentów metodą *conjoint analysis* wykorzystujemy: SPSS® 10.05 for Windows, SYSTAT® 8.0 for Windows, Conjoint Analysis for Windows, CVA oraz CVA/HB.

Program **SPSS® 10.05 for Windows** składa się z modułów programowych, które zawierają algorytmy obliczeń statystycznych oraz niezbędne procedury usługowe [20, s. 127-147]. Obliczenia statystyczne za pomocą programu SPSS® 10.05 for Windows można realizować w jednym z dwóch oferowanych trybów: trybie menu i okien dialogowych (standardowy sposób korzystania z programów w środowisku systemowym Windows) oraz w trybie wsadowym, tzn. z wykorzystaniem języka poleceń SPSS. W programie tym większość obliczeń statystycznych i operacji towarzyszących można wykonać, wybierając odpowiednie polecenia z list opcji (menu) programu. W wyświetlanych na ekranie monitora oknach dialogowych można ponadto wskazać dodatkowe pożądane opcje i ustawić właściwe parametry. Drugim sposobem wykonywania obliczeń statystycznych jest korzystanie z języka poleceń SPSS, który umożliwia przygotowanie procedur obliczeniowych wykonywanych w trybie wsadowym.

Polecenie CONJOINT jest dostępne w module SPSS Conjoint 8.0 i służy do analizy za pomocą metody *conjoint analysis* danych eksperymentalnych zgromadzonych zgodnie z regułami metody pełnych profilów wyboru. W języku poleceń SPSS dostępne są ponadto dodatkowe procedury statystyczne oraz opcje i parametry, z których nie można korzystać w trybie menu i okien dialogowych (m.in. polecenie CONJOINT). Procedurę obliczeniową w języku SPSS można stworzyć, zmodyfikować lub wykonać w oknie SPSS Syntax Editior.

Polecenie ORTHOPLAN umożliwia wygenerowanie ortogonalnego układu eksperymentu na podstawie danych przygotowanych zgodnie z regułami metody pełnych profilów wyboru. W poleceniu tym zaimplementowano model addytywny zależności użyteczności całkowitej od użyteczności cząstkowych. Można określić minimalną pożądaną liczbę generowanych wariantów albo zezwolić programowi na oszacowanie niezbędnej w świetle jakości późniejszego wnioskowania liczby tworzonych profilów. Poza generowanymi przez polecenie wariantami eksperymentalnymi można utworzyć dodatkowo profile w dwóch innych kategoriach: tzw. profile testowe oraz profile symulacyjne.

Program **Conjoint Analysis for Windows** z pakietu Marketing Engineering Applications Version 1.0 umożliwia realizację następujących celów badawczych: zaprojektowanie układu badań metodą *conjoint analysis* (dostępny jest wariant pełnych profilów wyboru) przez specyfikację zmiennych objaśniających i ich poziomów (opisujących produkty lub usługi), zgromadzenie ocen punktowych respondentów odnośnie do wygenerowanego zbioru profilów produktów lub usług, estymację użyteczności cząstkowych poszczególnych poziomów zmiennych objaśniających oraz symulacyjną analizę udziałów w rynku poszczególnych profilów [20, s. 151]. Program ten oferuje dwa podstawowe zestawy poleceń zatytułowane SCENERIO i ANALYSIS.

System **CVA** (*Conjoint Value Analysis*) Sawtooth Software jest oprogramowaniem wykorzystywanym w badaniach preferencji metodą *conjoint analysis* (metodą pełnych profilów) (zob. [4]). System ten okazuje się szczególnie przydatny w badaniach, w których pomiar interakcji nie jest priorytetem. Badanie ankietowe może być przeprowadzone z wykorzystaniem komputera bądź też z wykorzystaniem ankiety papierowej. W badaniu można się posłużyć metodą pojedynczego profilu (ranking lub rating) lub metodą porównywania (prezentacją) profilów parami.

System CVA zawiera: część pozwalającą na zaprojektowanie profilów, które będą zaprezentowane w kwestionariuszu, część pozwalającą na zaprojektowanie kwestionariusza komputerowego bądź papierowego, część pozwalającą na oszacowanie użyteczności (OLS dla danych ratingowych, a MONANOVA dla danych rankingowych) oraz część pozwalającą na modelowanie symulacyjne rynku. Projekt kwestionariusza charakteryzuje się kilkoma ograniczeniami: maksymalna liczba atrybutów wynosi 30 (choć zazwyczaj nie więcej niż 6), maksymalna liczba poziomów każdego z atrybutu wynosi 15, maksymalna zaś liczba pytań wynosi 500,

maksymalna liczba profilów symulacyjnych wynosi 30, natomiast liczba respondentów jest nielimitowana. System CVA działa z programem Microsoft Windows 95 bądź późniejszym. Oszacowane użyteczności cząstkowe mogą być wykorzystane w symulacyjnych segmentacjach rynku (np. z wykorzystaniem systemu CCA – *Convergent Cluster Analysis*).

System **CVA/HB** wykorzystuje estymację hierarchiczną Bayesa do oszacowania użyteczności cząstkowych w badaniu metodą *conjoint analysis* z zastosowaniem metody pełnych profilów (rating) (zob. [5]). System ten wykorzystuje w estymacji modelu metody Monte Carlo i łańcuchy Markowa, jest też obliczeniowo intensywny, dlatego sugerowany jest bardzo szybki procesor (2 GHZ lub większy).

Literatura

[1] Allenby G.M., Ginter J.L., *Incorporating Prior Knowledge into the Analysis of Conjoint Studies*, „Journal of Marketing Research", vol. XXXII, May 1995, pp. 152-162.

[2] Bąk A., *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej nr 1013 AE, Wrocław 2004.

[3] *Choice-based Conjoint (CBC) Technical Paper* (2001). Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/cbctech.pdf.

[4] *Conjoint Value Analysis (CVA) Version 3.0* (1997-2002), Artykuł dostępny w Internecie na stronie:www.sawtoothsoftware.com/download/techpap/cva3tech.pdf.

[5] *CVA/HB Technical Paper* (2002), Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/cvahb.pdf.

[6] *Individual Choice Estimation (ICE) Module for Choice-Based Conjoint*, (2001), Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/icetech.pdf.

[7] Johnson R.M., Huber J., Bacon L (2003), *Adaptive Choice Based Conjoint Analysis*, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/ download/techpap/acbc.pdf.

[8] Johnson R.M., Huber J., Orme B. (2005), *A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs)*, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/acbc2.pdf.

[9] Johnson R.M., Orme B.K. (1996), *How Many Questions Should You Ask in Choice-Based Conjoint Studies?*, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/ download/techpap/howmanyq.pdf.

[10] Johnson R.M., Orme B., Huber J., Pinnell J. (2005), *Testing Adaptive Choice-Based Conjoint Designs*, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/ techpap/acbc3.pdf.

[11] Kuhfeld W.F. (2001), *Multinominal Logit, Discrete Choince Modeleng. An Introdution to Designing Choince Experiments, Collecting, Proccesing, and Analyzing Choice Data with the SAS® System*, Artykuł dostępny w Internecie na stronie: http://ftp.sas.com/techsup/download/ technote/ts643/ts643.pdf, Cary, SAS Institute.

[12] Lenk P., DeSarbo W., Green P., Young P., *Hierarchical Bayes Conjoint Analysis: Recovery of Parthworth Heterogenity from Reduced Experimental Design*, „Marketing Science" 1996 vol. 15, no. 2, pp. 173-191.

[13] Orme B. (2003). *Special Features of CBC Software for Packaged Goods and Beverage Research*, Sawtooth Software Research Paper Series, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/speccbc.pdf.
[14] Pinnell J. (2005), *Comment on Huber: Practical Suggestions for CBC Studies*, Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/pinnell.pdf.
[15] *SAS/STAT® Software. Providing the foundation for SAS' analytic intelligence*, (2005). Artykuł dostępny w Internecie na stronie: www.sas.com/technologies/analytics/statistics/stat/factsheet.pdf.
[16] *SAS/STAT Software*, (2006). Artykuł dostępny w Internecie na stronie: http://support.sas.com/md/app/da/stat.html,.
[17] *The CBC Advanced Design Module (ADM) Technical Paper*, (2005), Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/admtech.pdf.
[18] *The CBC Latent Class Technical Paper (Version 3)*, (2004), Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/lctech.pdf.
[19] *The CBC/HB System for Hierarchical Bayes Estimation Version 4.0 Technical Paper* (2005), Artykuł dostępny w Internecie na stronie: www.sawtoothsoftware.com/download/techpap/hbtech.pdf.
[20] Walesiak M., Bąk A., *Conjoint analysis w badaniach marketingowych*, AE, Wrocław 2000.

COMPUTER APPLICATIONS OF DECOMPOSITIONAL METHODS USED IN MARKETING RESEARCH OF CONSUMER PREFERENCES

Summary

In consumers preferences analysis we apply decompositional methods. This approach is being represented both by conjoint analysis and discrete choice methods. Nowadays there are many different applications that allow to conduct such research.

The paper presents the characteristics of chosen computer applications that are applying conjoint methods – SPSS® 10.05 for Windows, Conjoint Analysis for Windows, CVA and CVA/HB. The SAS/STAT, CBC, CBC Advanced Design Module, CBC/HB, the Latent Class Segmentation Module and Individual Choice Estimation (ICE) Module are used in discrete choice methods research.

Aneta Rybicka – dr, adiunkt w Katedrze Ekonometrii i Informatyki Uniwersytetu Ekonomicznego we Wrocławiu – Wydział w Jeleniej Górze.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Agnieszka Przybylska-Mazur

ROZSTRZYGNIĘCIA TEORII GIER W MODELOWANIU UBEZPIECZENIA OD SKUTKÓW BEZROBOCIA

1. Wstęp

Bezrobocie jest jednym z rodzajów ryzyka społecznego. Ubezpieczenie bezrobocia ma ważne konsekwencje dla działań gospodarczych i dla opieki społecznej. Dlatego ubezpieczenie od skutków bezrobocia jest ubezpieczeniem znanym i stosowanym w wielu krajach.

Duża skala bezrobocia w Polsce zainspirowała niektóre towarzystwa do przygotowania specjalnych produktów. Towarzystwo Ubezpieczeń Generali w połowie 2004 r. wprowadziło do swojej oferty ubezpieczenie od utraty pracy dla osób posiadających kredyty bankowe. Kolejnym towarzystwem, które w lipcu 2004 r. wystąpiło z ofertą ochrony kredytobiorców, było TU Europa. W Towarzystwie Ubezpieczeń COMPENSA SA przedmiotem ubezpieczenia może być utrata stałego źródła dochodu na skutek zaistniałej w okresie ubezpieczenia utraty pracy przez ubezpieczonego. Podobnie w towarzystwie ubezpieczeń CIGNA STU SA – prywatne ubezpieczenia społeczne obejmują ochroną różne rodzaje ryzyka, m.in. bezrobocie. W tym towarzystwie już przed rokiem 2004 zaczęto wdrażać ubezpieczenie kredytobiorców od ryzyka utraty stałego źródła dochodu wskutek utraty pracy.

W opinii specjalistów ten rodzaj ubezpieczenia, jakim jest ubezpieczenie kredytobiorców od utraty stałego źródła dochodów, będzie się rozwijał. Mimo że nie jest to jeszcze klasyczne ubezpieczenie bezrobocia, takie jak w krajach zachodnich, to stanowi przyczynek do pojawienia się na polskim rynku nieznanych dotąd produktów ubezpieczeniowych kompensujących materialne skutki utraty pracy [5].

Zagadnienie ubezpieczenia od skutków bezrobocia jest mało zbadane w polskich realiach. W związku z tym w pracy zaprezentowano jeden z modeli ubezpieczenia bezrobocia. Stosując teorię gier, przedstawiono cele i strategie bezrobotnego

– poszukującego pracy, jak również cele i strategie towarzystwa ubezpieczeń. Przedstawiony model stanowi modyfikację modelu ubezpieczenia bezrobocia zaprezentowaną przez Zuckermana [7] w 1985 r. W założeniach modelu ujęto również dochody bezrobotnego z Funduszu Pracy lub z pomocy społecznej. W prezentowanym przykładzie zmieniono rozkład wynagrodzeń związanych z najlepszą ofertą pracy w danym okresie, dostosowując go do polskich realiów, a na podstawie podanych przykładów numerycznych wysunięto wniosek płynący z wyznaczonych na podstawie tego rozkładu strategii optymalnych.

W prezentowanym modelu zakłada się, że ubezpieczenie bezrobocia ma dwa podstawowe cele:

1) daje zabezpieczenie ekonomiczne ludziom, którzy tymczasowo są bezrobotni i poszukują pracy,

2) stymuluje szukanie pracy, ponieważ intensywne poszukiwanie ofert umożliwia jednostkom bezrobotnym znalezienie zatrudnienia dającego większe wynagrodzenie i dzięki temu większą produktywność.

2. Założenia modelu ubezpieczenia bezrobocia

W tej części pracy zaprezentowano założenia modelu ubezpieczenia bezrobocia.

Zakładamy, że liczba ofert otrzymanych przez poszukującego pracy w ciągu jednego okresu oraz oferowana płaca są wielkościami losowymi. W modelu jako jeden okres został przyjęty miesiąc. Na końcu każdego okresu bezrobotny dokonuje wyboru najlepszej oferty pracy otrzymanej w ciągu ostatniego miesiąca lub kontynuuje poszukiwania w kolejnym okresie. Oferty otrzymane w poprzednich miesiącach, które nie były zaakceptowane przez poszukującego pracy, tracą swoją ważność.

Z poszukiwaniem najlepszej oferty pracy wiążą się pewne koszty. Oznaczmy przez $F_c(n)$ dystrybuantę rozkładu obecnej wartości wynagrodzenia łączonego z najlepszą ofertą pracy otrzymaną w okresie n, przy założeniu, że c jednostek pieniężnych jest przeznaczonych na poszukiwanie zatrudnienia przez bezrobotnego w danym okresie. Ponadto zakładamy, że $F_c(n)$ jest znana dla poszukującego pracy.

Mechanizm ubezpieczenia bezrobocia w omawianym modelu jest następujący: poszukujący pracy zna korzyści z ubezpieczenia na wypadek bezrobocia w przedziale czasu N okresów.

Zakładamy, że świadczenie na wypadek bezrobocia składa się z dwóch składników [7]:

1) W jednostek pieniężnych na miesiąc na pokrycie podstawowych kosztów utrzymania bezrobotnego. W ubezpieczeniu bezrobocia ta kwota jest egzogenicznie określona i jest wypłacana osobie ubezpieczonej tak długo, jak długo pozostaje ona bezrobotna, nawet dłużej niż okres ubezpieczenia;

z najlepszą ofertą pracy otrzymaną w okresie n, przy założeniu, że bezrobotny używa politykę **C** kosztów poszukiwania pracy.

Funkcja celu towarzystwa ubezpieczeniowego jest następująca [7]:

$$\psi_{(\mathbf{C},T)}(\mathbf{U}) = E\left[X_{\mathbf{C}}(T) - \sum_{n=1}^{T}(u_n - c_n) \right]. \tag{2}$$

Celem towarzystwa jest wybranie optymalnej polityki **U**, która maksymalizuje funkcję celu określoną wzorem (2). Przyjmuje się, że towarzystwo zna funkcję odpowiedzi $(\mathbf{C}(\mathbf{U}), T(\mathbf{U}))$ poszukującego pracy.

Na podstawie warunków ubezpieczenia bezrobocia, czyli na podstawie ogłoszonej przez towarzystwo strategii **U** poszukujący pracy wybiera swoją strategię $\left(\mathbf{C}^*(\mathbf{U}), T^*(\mathbf{U})\right)$. Następnie na podstawie optymalnej strategii bezrobotnego towarzystwo wybiera optymalną strategię $\mathbf{U}^*$.

Maksymalny dochód $V_{n-1}(x)$ jednostki w okresie $n-1$ określa następujący związek rekurencyjny:

$$V_{n-1}(x) = \max\left\{ x, \max_{c_n \le u_n}\left\{ (W + u_n - c_n) + \int_0^{\infty} V_n(y) dF_{c_n}(y) \right\} \right\} \tag{3}$$

dla $n = 2, 3, \ldots, N$, przy czym $V_N(x) = x$.

Minimalną do zaakceptowania ofertę (najniższą płacę) ξ_{n-1} w okresie $n-1$ określamy następująco:

$$\xi_{n-1} = \max_{c_n \le u_n}\left\{ (W + u_n - c_n) + \int_0^{\infty} V_n(y) dF_{c_n}(y) \right\} \tag{4}$$

dla $n = 2, 3, \ldots, N$.

Ponadto, ponieważ $T \le N$, przyjmujemy

$$\xi_N = 0. \tag{5}$$

Na podstawie równań (3) – (5) otrzymujemy następujący związek

$$V_n(x) = \max\{x, \xi_n\} \text{ dla } n = 1, 2, \ldots, N. \tag{6}$$

W celu określenia strategii poszukującego pracy zostanie określona funkcja $G_{c_n}(\xi_n)$ mierząca oczekiwany zysk z c_n jednostek pieniężnych, które są przeznaczone na poszukiwania w okresie n, gdy najniższa płaca w tym okresie wynosi ξ_n:

$$G_{c_n}(\xi_n) = \int_{\xi_n}^{\infty} (x - \xi_n) dF_{c_n}(x). \tag{7}$$

Wykorzystując równania (4) – (6), otrzymujemy następujący związek dla $n = 2, 3, \ldots, N$

$$\begin{aligned} \xi_{n-1} &= \max_{c_n \le u_n} \left\{ (W + u_n - c_n) + \int_0^{\infty} \max(x, \xi_n) dF_{c_n}(x) \right\} = \\ &= \max_{c_n \le u_n} \left\{ W + u_n - c_n + \xi_n + G_{c_n}(\xi_n) \right\}. \end{aligned} \tag{8}$$

Funkcja (8) osiąga maksimum w punkcie $c_n^*(\mathbf{U})$ określonym następująco [1]:

$$c_n^*(\mathbf{U}) = \begin{cases} u_n & \text{gdy } \widetilde{c}_n > u_n \\ \widetilde{c}_n & \text{gdy } \widetilde{c}_n \le u_n \end{cases}, \tag{9}$$

gdzie $\widetilde{c}_n$ jest rozwiązaniem następującego równania

$$\frac{\partial G_{c_n}(\xi_n)}{\partial c_n} = 1. \tag{10}$$

Optymalna strategia $\mathbf{C}^*(\mathbf{U})$ poszukującego pracy i minimalna do zaakceptowania płaca ξ_n^* dla $n = 1, 2, \ldots, N$ są opisane rekurencyjnie przy użyciu wstecznej procedury dynamicznej w następujący sposób:

1) $\xi_N^* = 0$.

2) Optymalną politykę wydatków stosowaną przez bezrobotnego $c_N^*(\mathbf{U})$ obliczamy ze wzoru (9), w którym $\widetilde{c}_N$ jest rozwiązaniem następującego równania

$$\frac{\partial G_{c_n}(0)}{\partial c_n} = 1.$$

3) Dla $n = N-1, N-2, \ldots, 1$ optymalne strategie poszukującego pracy $c_N^*(\mathbf{U})$ obliczamy ze wzorów (9) i (10).

A zatem optymalna polityka wydatków stosowana przez bezrobotnego jest następująca: jeżeli strategia towarzystwa u_n jest mniejsza lub równa $\widetilde{c}_n$, to całkowita kwota zasiłku u_n będzie użyta na cele poszukiwań pracy, w przeciwnym wypadku kwota ponad $\widetilde{c}_n$ będzie użyta na cele konsumpcji.

Znając strategię optymalną ubezpieczonego, towarzystwo wyznacza swoją strategię optymalną $\mathbf{U}^* = \left[u_1^* \;\; u_2^* \;\; \ldots \;\; u_N^* \right]$, gdzie $u_n^* = \widetilde{c}_n$; tak więc optymalne kwoty

zasiłku – dodatku do kwoty W powinny być równe kwocie wydatków przeznaczonych przez jednostkę bezrobotną na poszukiwania pracy.

Natomiast optymalną najniższą akceptowalną płacę ξ_n^* dla $n = N-1, N-2, \ldots, 1$ wyznaczamy ze wzoru

$$\xi_n^* = W + \xi_{n+1}^* + G_{c_{n+1}^*(\mathbf{U})}\left(\xi_{n+1}^*\right). \tag{11}$$

Optymalną strategię stopującą – optymalny moment zakończenia poszukiwań pracy przez bezrobotnego określamy jako

$$T^*(\mathbf{U}) = \inf_{1 \le n < N}\left\{n, X_{\mathbf{C}^*(\mathbf{U})}(n) \ge \xi_n\right\}. \tag{12}$$

4. Przykład numeryczny

Zakładamy, że:

1. Rozkład wynagrodzeń związanych z najlepszą ofertą pracy w danym okresie, w którym $c_n > 0$ jednostek pieniężnych jest przeznaczonych na poszukiwania pracy, jest jednostajny w przedziale $\left[100, 150 - 50e^{-c_n}\right]$ (przyjmuje się, że wynagrodzenie możliwe do zaakceptowania należy do przedziału od 1000 do 1500 zł). Zatem

$$dF_{c_n}(x) = \begin{cases} \dfrac{1}{50\left(1-e^{-c_n}\right)} & \text{dla } 100 \le x \le 150 - 50e^{-c_n}, \\ 0 & \text{dla pozostałych } x. \end{cases}$$

2. $W = 5$.

Na podstawie wzorów podanych wcześniej dla arbitralnie przyjętej wartości startowej $N = 6$ obliczamy $\mathbf{C}^*\left(\mathbf{U}_6^*\right) = \mathbf{U}_6^* = \left[\tilde{c}_1\ \tilde{c}_2\ \tilde{c}_3\ \tilde{c}_4\ \tilde{c}_5\ \tilde{c}_6\right]$ oraz ξ_n^* dla $n = 1, 2, \ldots, 6$.

Przyjmujemy $\xi_6^* = 0$.

Wielkość $\tilde{c}_6$ wyznaczamy jako rozwiązanie równania $\dfrac{\partial G_{c_6}(0)}{\partial c_6} = 1$. Ponieważ

$$G_{c_6}(0) = \int_0^{\infty} x dF_{c_6}(x) = \int_{100}^{150-50e^{-c_6}} x dF_{c_6}(x) = \frac{1}{50\left(1-e^{c_6}\right)} \int_{100}^{150-50e^{-c_6}} x dx = \frac{250 - 50e^{-c_6}}{2},$$

zatem $\dfrac{\partial G_{c_6}(0)}{\partial c_6} = 25e^{-c_6}$.

$$\frac{\partial G_{c_6}(0)}{\partial c_6} = 1 \Leftrightarrow 25e^{-c_6} = 1 \Leftrightarrow c_6 = \ln 25 = 3{,}22,$$ a wówczas $\tilde{c}_6 = 3{,}22.$

Na podstawie wzoru (11) obliczamy optymalną najniższą akceptowalną płacę w okresie 5

$$\xi_5^* = W + \xi_6^* + G_{c_6^*(U)}\left(\xi_6^*\right) = 129.$$

Obecnie zostanie obliczona funkcja $G_{c_n}(\xi_n)$.

$$G_{c_n}(\xi_n) = \int_{\xi_n}^{\infty}(x-\xi_n)dF_{c_n}(x) = \int_{\xi_n}^{150-50e^{-c_n}}(x-\xi_n)dF_{c_n}(x) =$$

$$= \frac{1}{50\left(1-e^{c_n}\right)} \cdot \int_{\xi_n}^{150-50e^{-c_n}}(x-\xi_n)dx = \frac{\left(150-50e^{-c_n}-\xi_n\right)^2}{100\left(1-e^{c_n}\right)}$$

oraz pochodna

$$\frac{\partial G_{c_n}(\xi_n)}{\partial c_n} = \frac{e^{-c_n}\left(150-50e^{-c}-\xi_n\right)\left(-50-50e^{-c_n}+\xi_n\right)}{100\left(1-e^{-c_n}\right)}.$$

Przyjmujemy $\xi_5^* = 129$, natomiast $\tilde{c}_5$ wyznaczamy jako rozwiązanie równania $\frac{\partial G_{c_5}\left(\xi_5^*\right)}{\partial c_5} = 1$, otrzymując $\tilde{c}_5 = 2{,}73.$

Obliczamy optymalną najniższą akceptowalną płacę w okresie 4

$$\xi_4^* = W + \xi_5^* + G_{c_5^*(U)}\left(\xi_5^*\right) = 136{,}37.$$

Powtarzając tę procedurę, otrzymujemy optymalne koszty związane z poszukiwaniem pracy i najniższe możliwe do zaakceptowania płace w kolejnych okresach, które zestawiono w tab. 1.

Tabela 1. Optymalne koszty związane z poszukiwaniem pracy i najniższe możliwe do zaakceptowania płace

Okres	1	2	3	4	5	6
$\tilde{c}_n$	–	–	–0,61	1,9	2,73	3,22
ξ_n^*	–	–	141,81	136,37	129	0

Źródło: obliczenia własne.

Ponieważ koszty poszukiwań pracy są nieujemne, zatem $N^* = 3$, więc optymalne koszty poszukiwań pracy przez bezrobotnego w kolejnych okresach wynoszą $\mathbf{C}^*(\mathbf{U}^*) = \mathbf{U}^* = [1{,}9 \ 2{,}73 \ 3{,}22]$.

Przyjmując tę samą gęstość rozkładu wynagrodzeń związanych z najlepszą ofertą pracy w danym okresie, ale wyższą stałą kwotę $W = 10$, otrzymujemy krótszy możliwy okres poszukiwań pracy $N^* = 2$ oraz $\mathbf{C}^*(\mathbf{U}^*) = \mathbf{U}^* = [2{,}41 \ 3{,}22]$.

Ponadto $\xi_6^* = 0$, $\xi_5^* = 134$, $\xi_4^* = 145{,}45$.

Tak więc przyjmując tę samą gęstość rozkładu wynagrodzeń związanych z najlepszą ofertą pracy, można wyciągnąć wniosek, że im wyższa jest kwota przeznaczana przez towarzystwo ubezpieczeń na pokrycie podstawowych kosztów utrzymania jednostki, tym bardziej optymalny okres poszukiwań pracy ulega skróceniu, natomiast najniższa możliwa do zaakceptowania płaca jest wyższa.

5. Zakończenie

Ubezpieczenie bezrobocia ma istotne konsekwencje dla działania gospodarki oraz dla dobrobytu społeczeństwa.

Ubezpieczenie bezrobocia jest ważnym produktem ubezpieczeniowym. Wypłaty z tytułu ubezpieczenia bezrobocia pozwalają zrekompensować skutki materialne utraty pracy, a także pokrywają koszty związane z poszukiwaniem nowej, najkorzystniejszej dla bezrobotnego oferty pracy.

Wykorzystując teorię gier, w pracy przedstawiono teoretyczny model ubezpieczenia bezrobocia.

Literatura

[1] Fichtenholz G.M., *Rachunek różniczkowy i całkowy*, PWN, Warszawa 1995.
[2] de Groot M., *Optymalne decyzje statystyczne*, PWN, Warszawa 1981.
[3] Luce B.D., Raiffa H., *Gry i decyzje*, PWN, Warszawa 1964.
[4] Owen G., *Teoria gier*, PWN, Warszawa 1975.
[5] Raport Roczny 2002 Towarzystwa CIGNA STU SA.
[6] Stackelberg V.H., *The Theory of Market Economy*, Oxford University Press, Oxford 1952.
[7] Zuckerman D., *Optimal Unemployment Insurance Policy*, „Operations Research", vol. 33, no. 2, March-April 1985.

SETTLEMENT OF GAME THEORY IN THE MODELLING OF UNEMPLOYMENT EFFECTS INSURANCE

Summary

Unemployment is one of types of social risk. The insurance of unemployment has the important effect for economy and for the public assistance of society. Therefore the unemployment insurance is well-known in many countries. In last years insurance products which compensate for financial effects of job loss have been introduced also in the Polish market.

In this paper the author presents the model of unemployment insurance. Using techniques and the concepts of game theory she investigates the objectives and the optimal strategy of an unemployed that seeks the job as well as the objectives and the optimal strategy of an insurer.

Agnieszka Przybylska-Mazur – dr, pracownik Katedry Metod Statystyczno-Matematycznych w Ekonomii Akademii Ekonomicznej w Katowicach.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Grzegorz Michalski

VALUE BASED INVENTORY MANAGEMENT

1. Introduction

The basic financial aim of an enterprise is maximization of its value. In the same time, a large both theoretical and practical meaning has the research for determinants increasing the firm value. The financial literature contains information about numerous factors influencing the value. Among those factors is the net working capital, and elements creating it, such as the level of cash tie in account receivable, inventories and operational cash balances. The great part of classic financial models proposals relating to the optimum current assets management was constructed with net profit maximization in view. It is reason, why these models need reconstruction, which make its will be suitable for firms which want to maximize their value. The estimation of the influence of changes in firm decisions in sphere of inventory management, is a compromise between limiting of risk by having greater inventory level and limiting costs of inventory. It is the essential problem of the corporate financial management. The basic financial inventory management aim is holding the inventory on minimal acceptable level because of its costs. Holding inventory ties capital used to finance inventory and links with inventory storage, insurance, transport, obsolescence, wasting and spoilage costs. On the other hand, to low level of inventory, could be source of problems with meeting the supply.

2. Value based inventory management

If advantages from holding inventory on a level defined by the firm will be greater than the negative influence of an alternative costs from its holding, then the firms value will grow. Change of the accounts receivable level affects on the firm value. To measure that, we use a formula, basing on an assumption, that the firm

value is a sum of future free cash flows to firm (*FCFF*) discounted by cost of capital financing the firm:

$$\Delta V_p = \sum_{t=1}^{n} \frac{\Delta FCFF_t}{(1+k)^t}, \tag{1}$$

where: ΔV_p – Firm Value Growth, $\Delta FCFF_t$ – Future Free Cash Flow Growth in Period t, k – Discount Rate[1].

Future free cash flow we have as:

$$FCFF_t = (CR_t - CE_t - NCE) \times (1 - T) + NCE - Capex - \Delta NWC_t, \tag{2}$$

where: CR_t – Cash Revenues on Sales, CE_t – Cash Expenses resulting from fixed and variable costs in time t, NCE – Non Cash Expenses, T – Effective Tax Rate, ΔNWC – Net Working Growth, $Capex$ – Capital Expenses resulting from operational investments growth.

The similar conclusions, about the results of the change inventory management policy on the firm value, can be estimated on the basis of an economic value added, informing about the size of the residual profit (the added value) enlarged the value of the firm in the period:

$$EVA = NOPAT - k \times (NWC + OI), \tag{3}$$

where: EVA – Economic Value Added, NWC – Net Working Capital, OI – Long-Term Operating Investments, $NOPAT$ – Net Operating Profit After Tax, estimated on the basis of the formula:

$$NOPAT = (CR_t - CE_t - NCE) \times (1 - T). \tag{4}$$

The net working capital (NWC) is the part of current assets, financed with fixed capitals. The net working capital (current assets less current liabilities) results from lack of synchronization of the formal rising receipts and the real cash receipts from each sale. Net working capital also results from divergence during time of rising costs and time, from the real outflow of cash when a firm pays its accounts payable.

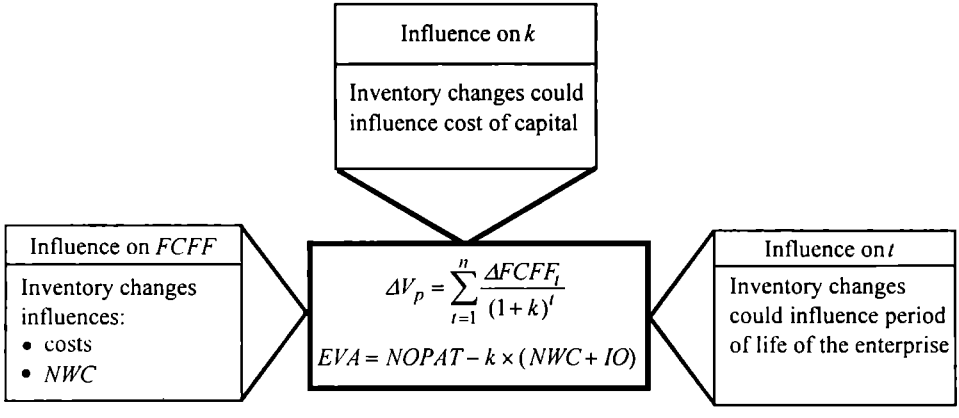
$$NWC = CA - CL = AAR + INV + G - AAP, \tag{5}$$

where: NWC – Net Working Capital, CA – Current Assets, CL – Current Liabilities, AAR – Accounts Receivables, INV – Inventory, G – Cash and Cash Equivalents, AAP – Accounts Payables.

[1] To estimate changes in accounts receivable levels, we accept discount rate equal to the average weighted cost of capital (WACC). Such changes and their results are strategic and long term in their character, although they refer to accounts receivable and short run area decisions [6, pp. 62-63].

During estimation of the free cash flows the holding and increasing of net working capital ties money used for financing it. If net working capital increase, the firm must tie much money and it decrease free cash flows. The production level growth usually makes the necessity of enlargement of cash levels, inventories, and accounts receivable. Part of this growth will be covered with current liabilities. For current liabilities also usually automatically grow up together with the growth of production. The rest (which is noted as net working capital growth) will require other form of financing.

The inventory management policy decisions, create the new inventory level in firm. It has the influence on the firm value. It is result of alternative costs of money tie in inventory and generally of costs of inventory managing. Both the first and the second involve modification of future free cash flows, and in consequence the firm value changes. On figure 1, we have the influence of inventory management decisions on the firm value. These decisions changes the future free cash flows (*FCFF*). These decisions could also influence on life of the firm (t) (by the operational risk, which is the result of possibility to break production cycles if the inventory level is too low), and rate of the cost of capital financing the firm (k). The changes of these three components have influence on the creation the firm value (ΔVp).



where: *FCFF* – Free Cash Flows to Firm; ΔNWC – Net Working Capital Growth; k – cost of the capital financing the firm; t – the lifetime of the firm and time to generate single *FCFF*.

Fig. 1. The inventory management decision influence on firm value

Source: own study.

Inventory changes (resulting from changes in inventory management policy of the firm) affect the net working capital level and as well the level of operating costs of inventory management in a firm. These operating costs are result of storage, insurance, transport, obsolescence, wasting and spoilage of inventory.

3. *EOQ* and *VBEOQ*

Economic order quantity model is a model chich maximizes the firm income by total inventory costs minimization.

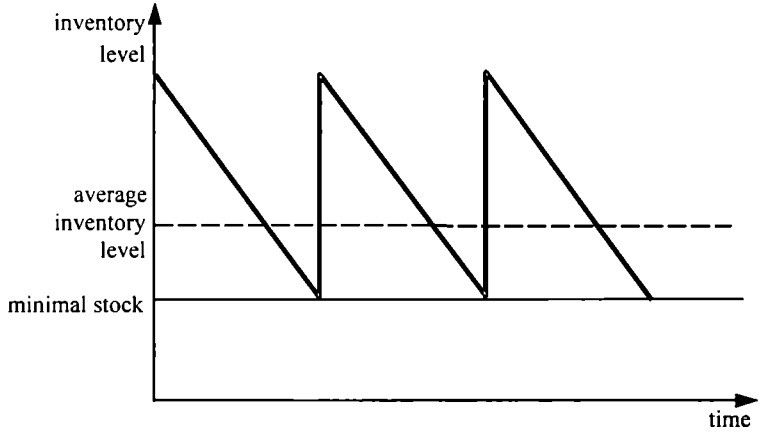


Fig. 2. *EOQ* and *VBEOQ* model

Source: [4, p. 538].

To *EOQ* model we have two equations:

$$EOQ = \sqrt{\frac{2 \times P \times K_z}{C \times v}} = \sqrt{\frac{2 \times P \times K_z}{K_u}}, \tag{6}$$

where: *EOQ* – Economic Order Quantity, P – Demand for the Product/Inventory in period (year, month), K_z – Cost per Order Event, K_u – Holding Cost per unit in period (year, month), C – Holding Cost Factor, v – Purchase Cost per Unit.

Holding cost factor (K_u) is a result of costs [13, p. 112]:

- Alternative costs (price of money tie in inventory),
- Storage, insurance, transport, obsolescence, wasting and spoilage costs.

$$TCI = \frac{P}{Q} \times K_z + \left(\frac{Q}{2} + z_b\right) \times v \times C, \tag{7}$$

where: *TCI*– Total Costs of Inventory, *Q*– Order Quantity, z_b – Minimal Stock.

Example 1. $P = 220\,000$ kg, $K_z = 31$ \$, $v = 2$ \$ / 1 kg, $C = 25\%$. Effective tax rate, $T = 20\%$. Cost of capital financing the firm $WACC = k = 15\%$, $z_b = 300$ kg.

First we estimate *EOQ*:

$$EOQ = \sqrt{\frac{2 \times 220\,000 \times 31}{0{,}25 \times 2}} = 5\,223 \text{ kg}.$$

Next we estimate average inventory level:

$$INV_{EOQ} = \frac{5\,223}{2} + 300 = 2\,912 \text{ kg} \Rightarrow INV_{EOQ} = 2\,912 \times 2 = 5\,824 \$,$$

$$TCI = \frac{220\,000}{5\,223} \times 31 + \left(\frac{5\,223}{2} + 300\right) \times 2 \times 0{,}25 = 2\,762 \$.$$

If we rather will order 5 000 kg than *EOQ* = 5 223 kg:

$$TCI_{5000} = \frac{220\,000}{5\,000} \times 31 + \left(\frac{5\,000}{2} + 300\right) \times 2 \times 0{,}25 = 2\,764 \$.$$

We will have greater *TCI*, but if we check how it influence on the firm value, we will see that if we decide to order less than *EOQ* suggest, we will increase the firms value:

$$\Delta TCI_{5000} = 2\,764 - 2\,762 = 2 \$.$$

$$INV_{5000} = 2 \times \left(\frac{5\,000}{2} + 300\right) = 5\,600 \$.$$

$$\Delta INV_{5000} = 5\,600 - 5\,824 = -224 \$.$$

$$\Delta NWC = \Delta INV.$$

$$\Delta V_{5000} = 224 - \frac{2 \times (1 - 0{,}2)}{0{,}15} = 213{,}33 \$,$$

EOQ model minimize operational inventory costs, but in firm management we also have alternative costs of holding inventories. These costs need that we will order less than *EOQ* if we want maximize the firm value. Knowing that we can use *VBEOQ* model:

$$VBEOQ = \sqrt{\frac{2 \times (1 - T) \times K_Z \times P}{v \times (k + C \times (1 - T))}}, \tag{8}$$

where: k – Cost of Capital financing the Firm (*WACC*); *VBEOQ* – Value Based Economic Order Quantity.

For Alfa data, we have:

$$VBEOQ = \sqrt{\frac{2 \times (1 - 0{,}2) \times 31 \times 220\,000}{2 \times (0{,}15 + 0{,}25 \times (1 - 0{,}2))}} = 3\,948{,}24 \approx 3\,948 \text{ kg}.$$

$$TCI_{3948} = \frac{220\,000}{3\,948} \times 31 + \left(\frac{3\,948}{2} + 300\right) \times 2 \times 0{,}25 = 2\,864{,}46\ \$,$$

$$\Delta TCI_{3948} = 2\,864{,}46 - 2\,762 = 102{,}46\ \$,$$

$$INV_{3948} = 2 \times \left(\frac{3\,948}{2} + 300\right) = 4\,548\ \$,$$

$$\Delta INV_{3948} = 4\,548 - 5\,824 = -1\,276\ \$,$$

$$\Delta V_{3948} = 1\,276 - \frac{102{,}46 \times (1 - 0{,}2)}{0{,}15} = 729{,}55\ \$.$$

4. *POQ* and *VBPOQ*

Production order quantity model (*POQ*) is the *EOQ* modification which we can use, when we have grater production possibilities than market capacity.

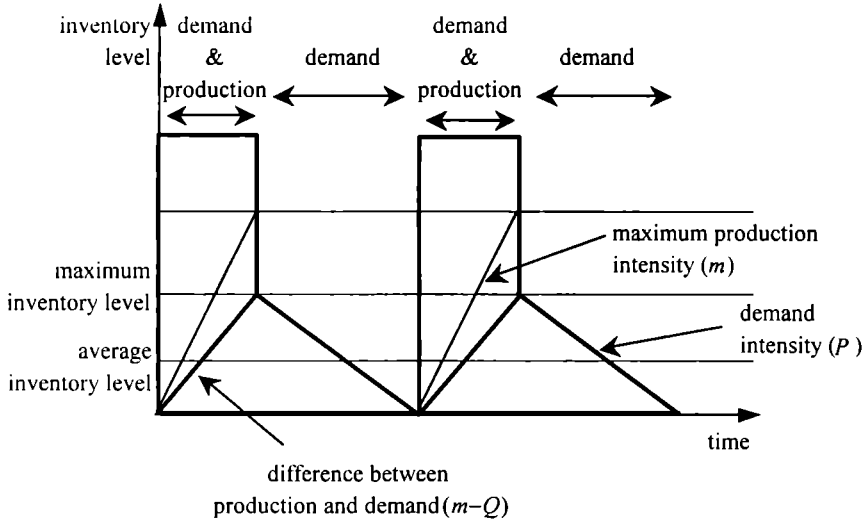


Fig. 3. *POQ* and *VBPOQ*

Source: [10, p. 162].

POQ could be estimated as [10, p. 162]:

$$POQ = \sqrt{\frac{2 \times K_z \times P}{C \times k \times \left(1 - \frac{P}{m}\right)}},\ P < m, \tag{9}$$

where: POQ – Production Order Quantity, K_z – Swith On Production Cost, P – Demand Intensity (how much we can sell annualy), v – Cost per Unit, m – Maximum Annual Production Ability, C – Holding Cost Factor.

$$TCI = \frac{Q}{2} \times \left(1 - \frac{P}{m}\right) \times v \times C + \frac{P}{Q} \times K_z, \tag{10}$$

where: Q – Production Quantity; TCI – Total Costs of Iventories.

$$INV = \frac{Q}{2} \times \left(1 - \frac{P}{m}\right) \tag{11}$$

where: INV – Average Inventory Level.

Example 2. Maximum demand, $P = 2\,500\,000$ kg, $m = 10\,000\,000$ kg annualy. $WACC = k = 15\%$, $C = 25\%$, $T = 19\%$. $K_z = 12\,000$ \$, $v = 0{,}8$ \$.

First we estimate POQ:

$$POQ = \sqrt{\frac{2 \times 12\,000 \times 2\,500}{800 \times 0{,}25 \times \left(1 - \frac{2\,500}{10\,000}\right)}} = 633 \text{ (1000) kg.}$$

$$TCI_{633} = \frac{633}{2} \times \left(1 - \frac{2\,500}{10\,000}\right) \times 800 \times 0{,}25 + \frac{2\,500}{633} \times 12\,000 = 94\,868 \text{ \$.}$$

$$INV_{633} = \frac{633}{2} \times \left(1 - \frac{2\,500}{10\,000}\right) = 237 \text{ (1000) kg} \Rightarrow 237 \times 800 = 189\,600 \text{ \$.}$$

Next, we check how on firm value will influence change of production quantity to 90% POQ, $633\,000 \times 0{,}9 = 570\,000$ kg:

$$TCI_{570} = \frac{570}{2} \times \left(1 - \frac{2\,500}{10\,000}\right) \times 800 \times 0{,}25 + \frac{2\,500}{570} \times 12\,000 = 95\,382 \text{ \$,}$$

$$\frac{-\Delta FCFF_{1\ldots\infty}}{0{,}81} = \Delta TCI_{Q=633 \to Q=570} = 95\,382 - 94\,868 = 514 \text{ \$.}$$

$$INV_{570} = 800 \times INV_{570} = 800 \times \frac{570}{2} \times \left(1 - \frac{2\,500}{10\,000}\right) = 171\,000 \text{ \$,}$$

$$\Delta NWC = \left(-\Delta FCFF_0\right) = \Delta ZAP_{Q=6\,797 \to Q=30\,500} = 171\,000 - 189\,600 = -18\,600 \text{ \$.}$$

$$\Delta V_{Q=633 \to Q=570} = +18\,600 + \frac{-514 \times (1 - 0{,}19)}{0{,}15} = +15\,824 \text{ \$.}$$

As we see, if we will produce less than POQ suggest, it will create additional value. If we want to sign $VBPOQ$, we can use a table:

Table 1. *VBPOQ*

Q	TCI	Δ TCI	INV	Δ INV	Δ V
483	98 337	3 469	144 900	–44 700	25 968
482	98 391	3 523	144 600	–45 000	25 978
481	98 445	3 577	144 300	–45 300	25 984
480	98 500	3 632	144 000	–45 600	25 987
479	98 555	3 687	143 700	–45 900	25 988
478	98 612	3 744	143 400	–46 200	25 985
477	98 668	3 800	143 100	–46 500	25 980

Source: own study.

VBPOQ will be 479 000 kg. From table we see also, that costs TCI for *VBPOQ* will be greater than for *POQ*, but *VBPOQ* tie less Money in inventories what is source of benefits in alternative costs. To estimate *VBPOQ* we also could use a equation:

$$Q_{VBPOQ} = \sqrt{\frac{2 \times P \times K_z \times (1-T)}{v \times \left(1-\frac{P}{m}\right) \times [k + C \times (1-T)]}}, \quad P < m, \tag{12}$$

$$Q_{VBPOQ} = \sqrt{\frac{2 \times 2\,500 \times 12\,000 \times (1-0{,}19)}{800 \times \left(1-\frac{2\,500}{10\,000}\right) \times [0{,}15 + 0{,}25 \times (1-0{,}19)]}} = 479 \text{ (000) kg.}$$

5. Conclusion

Maximization of wealth of his owners is the basic financial aim in management of enterprise. Inventory management must contribute to realization this aim. In article we have seen value based *EOQ* model and value based *POQ* model modifications. Inventory management decisions are complex case. On one side too many money tie in inventory, burdens the enterprise with the high costs of inventory service and additionally high alternative costs. From other side, the higher inventory stock could help enlarge incomes from sales because purchasers have greater flexibility in making purchase decisions. In the article the problem connected with optimal economic order quantity and production order quantity was discussed over. Value based modifications of these two models, could help managers to make better, value creating decisions in inventory management.

Literature

[1] Brigham E.F., Daves P.R., *Intermediate Financial Management*, Thomson, Mason 2004.
[2] Fabozzi F.J., Fong G., *Zarządzanie portfelem inwestycji finansowych przynoszących stały dochód*, PWN, Warszawa 2000.
[3] Jajuga K., *Zarządzanie kapitałem*, AE, Wrocław 1993.
[4] Kalberg J.G., Parkinson K.L., *Corporate liquidity: Management and Measurment*, IRWIN, Homewood 1993.
[5] Luenberger D.G., *Teoria inwestycji finansowych*, PWN, Warszawa 2003.
[6] Maness T.S., Zietlow J.T., *Short-Term Financial Management*, Dryden Press, Fort Worth 1998.
[7] Michalski G., *Leksykon zarządzania finansami*, C.H. Beck, Warszawa 2004.
[8] Piotrowska M., *Finanse spółek. Krótkoterminowe decyzje finansowe*, AE, Wrocław 1997.
[9] Pluta W., Michalski G., *Krótkoterminowe zarządzanie kapitałem*, C.H. Beck, Warszawa 2005.
[10] Sariusz-Wolski Z., *Sterowanie zapasami w przedsiębiorstwie*, PWE, Warszawa 2000.
[11] Sartoris W., Hill N., *A Generalized Cash Flow Approach to Short-Term Financial Decisions*, [w:] A. Pogue, *Cash and Working Capital Management*, „The Journal of Finance", May 1983, p. 349-360.
[12] Scherr F.C., *Modern Working Capital Management. Text and Cases*, Prentice Hall, Englewood Cliffs 1989.
[13] Sierpińska M., Wędzki D., *Zarządzanie płynnością finansową w przedsiębiorstwie*, PWN, Warszawa 2005.

ZARZĄDZANIE ZAPASAMI UKIERUNKOWANE NA MAKSYMALIZACJĘ WARTOŚCI PRZEDSIĘBIORSTWA

Streszczenie

Podstawowym celem finansowym działania przedsiębiorstwa jest maksymalizacja bogactwa jego właścicieli. Zarządzanie zapasami powinno także przyczyniać się do realizacji tego celu. Większość funkcjonujących w literaturze modeli dotyczących zarządzania aktywami bieżącymi nie uwzględnia jednak tego celu i jest ukierunkowane na maksymalizację zysku księgowego. Niniejszy artykuł prezentuje propozycję ukierunkowanej na maksymalizację wartości przedsiębiorstwa modyfikacji dwóch najpopularniejszych w literaturze modeli do zarządzania zapasami.

Grzegorz Michalski – dr, adiunkt w Katedrze Finansów Przedsiębiorstwa i Zarządzania Wartością Uniwersytetu Ekonomicznego we Wrocławiu.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Marcin Wojciel

PROPOZYCJA USPRAWNIENIA PROCESU PLANOWANIA SPRZEDAŻY USŁUG UBEZPIECZENIOWYCH W SEKTORZE MSP

Jednym z kluczowych zadań firmy ubezpieczeniowej jest planowanie sprzedaży usług ubezpieczeniowych dla małych i średnich przedsiębiorstw. Dokładność planowania jest wyzwaniem, które powoduje ciągłe poszukiwanie możliwości jego doskonalenia. Celem artykułu jest prezentacja propozycji usprawnienia planowania sprzedaży w aspekcie regionalnym. Narzędziem doskonalenia jakości planu jest użycie zmiennej makroekonomicznej odzwierciedlającej zróżnicowanie popytu na ubezpieczenia dla małych i średnich przedsiębiorstw. Zmienną, która wydaje się najlepiej realizować powyższe zadania, jest miernik produktu krajowego brutto (PKB) wytworzonego w regionach. Ponieważ PKB nie jest rejestrowane na dostatecznie niskim poziomie regionalnym, jego wartość musi być szacowana. Okazuje się, że miernik ten może być aproksymowany z dużą dokładnością przez dochody ludności brutto. Zmienna ta jest dostępna dla dostatecznie małych jednostek terytorialnych, ponadto jest publikowana z mniejszym opóźnieniem niż regionalne PKB. W opracowaniu pokazano propozycję rozwiązania problemu wraz z algorytmem tworzenia planu.

Planowanie sprzedaży można rozpatrywać w dwu podstawowych aspektach. Pierwszy z nich to właściwe określenie wolumenu sprzedaży w kolejnym okresie rozliczeniowym. Można go umownie nazwać *czasowym wymiarem planowania.* W tym przypadku planuje się przyszłą sprzedaż w określonym przedziale czasowym. Aby zrobić to dobrze, trzeba wziąć pod uwagę przewidywany rozwój rynku, przeanalizować działania konkurencji oraz własne cele i priorytety. Wypadkową tych wszystkich elementów powinien być plan, który maksymalizuje zadany efekt biznesowy – np. wolumen lub stopę zysku. Jednocześnie jakość planu oceniana jest stopniem jego realności przy dostępnych środkach (zasobach).

Drugi aspekt planowania sprzedaży usług ubezpieczeniowych to podział planu krajowego na terytorialne jednostki organizacyjne wchodzące w skład działu sprzedaży – innymi słowy, wyznaczenie celów sprzedaży dla zespołów odpowiadających za sprzedaż na poszczególnych terytoriach. Można to nazwać *terytorialnym wymiarem planowania*. Wydaje się on być nie mniej ważny niż poprzedni ze względu na skutki, jakie dla wyników działalności ma niewłaściwie zaplanowana wielkość sprzedaży w poszczególnych obszarach terytorialnych. Łatwo można sobie wyobrazić sytuację, w której przez zastosowanie prostego mnożnika jako podstawy wyznaczania planu (np. każdy region powinien sprzedać o 30% więcej niż sprzedał w roku ubiegłym) doprowadza się do stworzenia nierealistycznego planu. Dla regionów, które mają bardzo wysoki udział w rynku, plan sprzedaży może okazać się niewykonalny, regiony zaś, które mają niewielki udział w rynku i istnieje tam rezerwa niewykorzystanych możliwości – łatwo realizują plan. Nie zostaną jednak osiągnięte dodatkowe efekty wynikające z motywującego działania ambitnego, realistycznego planu. Ponadto może się pojawić oportunistyczne dążenie do niepodejmowania wysiłków, które zabezpiecza przed nadmiernym obciążeniem zadaniami planowymi wynikającymi z zastosowania metody mnożnikowej w kolejnych latach. Powoduje to zwiększenie ryzyka niepowodzenia w dążeniu do realizacji założonego poziomu sprzedaży. W części regionów personel sprzedaży może być sfrustrowany niemożnością realizacji wygórowanego planu, w innych regionach personel sprzedaży może być niewystarczająco zmotywowany.

Propozycja zmierzająca do poprawy jakości procesu planowania sprzedaży ubezpieczeń na rynku małych i średnich przedsiębiorstw w rozbiciu na jednostki terytorialne wymaga ustalenia definicji przedsiębiorstw uważanych za należące do sektora małych i średnich (MSP). Zwykle przyjmuje się w Polsce, że górną granicą jest zatrudnienie do 50 osób i obrót roczny nieprzekraczający 4 mln euro. Przekroczenie którejkolwiek z granic powoduje przesunięcie przedsiębiorstwa poza sektor MSP.

Propozycja modyfikacji procesu podziału krajowego planu sprzedaży ubezpieczeń dla firm z grupy MSP, polegająca na wprowadzeniu do analizy regionalnych zmiennych makroekonomicznych, odchodzi od dotychczas stosowanej logiki procesu planowania opartego głównie na planowaniu oddolnym, gdzie jednostki terenowe, będące najbliżej klientów, proponowały możliwą do osiągnięcia wartość sprzedaży, mierzoną najczęściej wartością składki przypisanej. Zebrane w ten sposób dane stanowiły podstawę negocjacji z jednostkami terenowymi w celu uzyskania sumarycznego, planowanego na kolejny rok wzrostu sprzedaży. Ten sposób planowania ma bardzo istotną zaletę, mianowicie osiąga się wyższy niż w przypadku planu narzuconego z góry stopień identyfikacji jednostek z tak wypracowanymi celami sprzedażowymi. Wadami takiego postępowania są duża czasochłonność oraz fakt, że utrwala ono dotychczasową terytorialną strukturę sprzedaży. Innymi słowy, regiony o wysokim udziale w rynku muszą go powiększać szybciej niż regiony o niskim udziale w rynku.

Ważnym elementem planowania sprzedaży jest pomiar poziomu udziału w rynku w poszczególnych regionach. Stąd konieczność zbudowania modelu, który pozwoliłby określić wielkość wolumenu rynku ubezpieczeniowego w regionach, a dzięki temu zmierzyć (oszacować) udział w rynku, jaki mają poszczególne jednostki terytorialne towarzystwa.

W celu określenia potencjału rynku ubezpieczeń MSP w regionach najlepszą metodą byłoby zbudowanie pełnego modelu ekonometrycznego opartego na wytypowanych zmiennych. Brak danych odpowiedniej jakości dotyczących reprezentatywnej próby podmiotów sprawia, że stosuje się odmienne podejście oparte na porównaniu względnej siły ekonomicznej regionów. Pomiar odbywa się za pomocą zmiennej/zmiennych, które w sposób wystarczająco dokładny charakteryzują region, a jednocześnie wykazują się silną korelacją z popytem na usługi ubezpieczeniowe zgłaszanym przez klientów z MSP. Po wstępnej analizie najbardziej obiecującą zmienną wydaje się być PKB w regionie (aproksymowane przez całkowite dochody ludności brutto).

W literaturze brakuje opracowań dotyczących ekonometrycznego modelowania rynku ubezpieczeń MSP. Istniejące konstrukcje modelowe odnoszą się zwykle do segmentu klientów indywidualnych. Specyfika klientów z sektora MSP powoduje, że nie można bezpośrednio zastosować metod przyjętych do analizy popytu na ubezpieczenia zgłaszanego przez klientów indywidualnych ani technik stosowanych przy badaniu popytu zgłaszanego przez duże firmy. Dlatego należało przeprowadzić nowe rozważania, opierając się na dotychczasowych rozwiązaniach.

Wobec niemożliwości oparcia procesu modelowania, który prowadziłby do precyzyjnego planu sprzedaży w regionach, na danych opisujących potencjał rynku, należało zmienić podejście i zastąpić miary absolutne zastępczymi oszacowaniami. Podstawą stało się określenie względnych potencjałów w poszczególnych regionach. Stworzenie takiego modelu, w pełni weryfikowalnego, wymagałoby informacji o zgłaszanym w regionach popycie. Najbardziej obiecującą i łatwo dostępną zmienną jest miernik PKB wytworzonego w regionie. Jest to dostatecznie duży agregat gospodarczy, ogólnie przyjęty w opisie poziomu rozwoju gospodarczego. Widoczne jest również silne powiązanie tego miernika z innymi potencjalnymi zmiennymi – PKB jest silnie skorelowany z liczbą pracujących, wartością środków trwałych, dochodami ludności itp.

Poważnym problemem związanym z wykorzystaniem tej zmiennej jest niedostępność bieżących danych statystycznych. Występuje tu prawie dwuletnie opóźnienie w publikacji potrzebnych danych statystycznych. Jest to zbyt duża zwłoka, dlatego konieczne jest poszukiwanie zmiennej, która publikowana jest z mniejszym opóźnieniem. Uzasadnionym wyborem wydaje się być całkowity dochód ludności brutto, który jest wysoce skorelowany z PKB wytworzonym w regionie[1], a jed-

[1] Dla wojewódzkich danych z 2005 r. współczynnik korelacji liniowej Pearsona r wynosi 0.99679.

nocześnie występuje mniejsze, niż w przypadku PKB, opóźnienie publikacji. Dane o dochodach ludności są dostępne na poziomie powiatów, co jest zgodne z najniższym poziomem podziału na jednostki terytorialne w towarzystwie ubezpieczeniowym. Jednostki terytorialne działu sprzedaży firm ubezpieczeniowych często obejmują obszar jednego lub kilku powiatów (wyjątkiem mogą być wielkie miasta).

Porównanie podziału regionalnego dochodów ludności z wynikami sprzedaży w 2005 r. (por. tab. 1) pokazuje, że aktualne terytorialne zróżnicowanie poziomu sprzedaży usług ubezpieczeniowych w sektorze MSP wykazuje niską korelację ze zróżnicowaniem poziomu dochodów ludności w regionach (współczynnik korelacji $r < 0{,}33$). Z tego wynika, że istnieją regiony, których udział w rynku jest dużo wyższy od przeciętnego, dlatego w tych regionach trudniej będzie zwiększać udział w rynku równie szybko jak w regionach o niskim udziale w rynku. Z drugiej strony są regiony, które mają niewykorzystany potencjał zwiększenia sprzedaży. Zidentyfikowanie regionalnego udziału w rynku doprowadzi do ustalenia planu sprzedaży lepiej wykorzystującego potencjał każdego regionu.

Wykorzystanie dochodów ludności brutto jako zmiennej zastępczej przybliżającej rozkład terytorialny popytu na usługi ubezpieczeniowe wydaje się być rozwiązaniem prostym, zrozumiałym i łatwym do implementacji. Jest jednak obarczone wadami. Wybór PKB jako zmiennej użytej do modelowania rynku ma charakter wyboru *a priori*, który nie może być obiektywnie zweryfikowany z powodu braku danych opisujących ten rynek. Wybrana zmienna nie jest cechą charakterystyczną wyłącznie dla sektora MSP, zawiera informację o całym rynku, co może wpływać na jakość oszacowania. Jednakże, w związku z dużym zróżnicowaniem klientów MSP, określenie zmiennej charakterystycznej wyłącznie dla tego sektora jest niezwykle trudne. Zmienna taka, poza warunkiem odzwierciedlenia rozwoju sektora MSP, musiałaby być mierzona i publikowana na odpowiednio niskim poziomie terytorialnym oraz z niewielkim opóźnieniem. Tylko spełnienie powyższych warunków pozwalałoby jej użyć w omawianym zagadnieniu. Wobec braku takiej zmiennej zwykle przyjmuje się założenie, że rozwój sektora MSP koreluje z całkowitym rozwojem gospodarczym mierzonym za pomocą PKB[2].

Zwiększenie wartości dochodów ludności brutto w regionie w oczywisty sposób przekłada się na zwiększenie siły nabywczej ludności, a tym samym staje się czynnikiem generującym zwiększony popyt konsumpcyjny w regionie. Przedsiębiorstwa z segmentu MSP są najczęściej podmiotami niewielkimi, więc bardziej mobilnymi i elastycznymi w porównaniu z dużymi przedsiębiorstwami. Dlatego szybciej i łatwiej przystosowują się do nowej sytuacji ekonomicznej regionu (przede wszystkim zmian popytu ludności). Obserwacje pokazują, że liczba nowo powstałych przedsiębiorstw MSP jest większa w regionach o większym popycie

[2] Przyjęcie takiego rozwiązania oznacza, że zakłada się stały udział sektora MSP w wytwarzaniu PKB w każdym z regionów.

generowanym przez mieszkańców. Użycie zmiennej, jaką jest dochód ludności brutto skumulowany w regionie, nie tylko staje się wyznacznikiem obecnej sytuacji w segmencie MSP, ale zawiera także element prognostyczny – informujący o potencjale rozwoju rynku. To oznacza, że zawiera także informację o potencjalnym wzroście liczby przedsiębiorstw z tego sektora w bliskiej przyszłości, a także o możliwości rozwoju istniejących firm. Przedstawiona argumentacja wskazuje, że proponowany wybór zmiennej ma więcej zalet niż wad. Jednocześnie proponowane rozwiązanie umożliwia szybkie wprowadzenie w życie usprawnienia procesu planowania sprzedaży.

W tab. 1 zaprezentowano porównanie struktury regionalnej przypisu składki (jako miary sprzedaży usług ubezpieczeniowych w sektorze MSP) i struktury dochodu ludności (jako miary potencjału rynku w regionie). Znaczne różnice pomiędzy wartościami w obu kolumnach pokazują miejsca, w których wynik sprzedaży wyraźnie odbiega od potencjału rynku. Tam, gdzie udział regionu w strukturze składki przypisanej jest niższy niż w strukturze dochodów ludności, konieczna jest dodatkowa praca nad lepszym wykorzystaniem potencjału rynku. Przedstawione dane pokazują przykładowe regionalne zróżnicowanie struktury sprzedaży pewnej firmy ubezpieczeniowej i dochodów ludności w wybranym roku.

Tabela 1. Regionalne zróżnicowanie struktury sprzedaży i dochodów ludności (w %)

Region	Oddział	Udział regionu w sprzedaży ogółem	Udział dochodów ludności regionu w dochodach ludności Polski ogółem
R1	O1	5,48	8,55
	O2	4,40	5,64
	O3	2,84	2,27
	O4	9,30	7,63
R2	O5	2,98	3,83
	O6	5,79	6,09
	O7	4,57	4,25
	O8	3,49	3,98
R3	O9	6,89	5,82
	O10	3,45	2,89
	O11	5,80	3,76
	O12	2,84	2,22
R4	O13	8,37	5,92
	O14	6,10	5,01
	O15	5,93	4,38
	O16	5,49	4,61
R5	O17	3,83	4,78
	O18	5,68	11,94
	O19	3,78	3,70
	O20	2,99	2,76

Źródło: dane przykładowe.

Powyższe dane odzwierciedlają udział regionów w popycie na usługi ubezpieczeniowe zgłaszanym przez małe i średnie przedsiębiorstwa na poziomie oddziałów (istnieją jeszcze mniejsze jednostki terytorialne towarzystwa obejmujące pojedyncze powiaty). Określenie modelu popytu wyznacza zakończenie pierwszego etapu proponowanego usprawnienia planowania. Następny etap, czyli zastosowanie stworzonego modelu popytu do zaplanowania sprzedaży usług ubezpieczeniowych, można przeprowadzić dwoma sposobami.

Pierwszy z nich to bezpośrednie wykorzystanie obserwowanej struktury rynku do wyliczenia udziału regionu i w planie krajowym, np. jako kombinacji liniowej istniejącej regionalnej struktury sprzedaży i przyjętej regionalnej struktury popytu – aproksymowanego poziomem dochodów ludności. Wtedy

$$P_i = \alpha SP_i + (1 - \alpha) UR_i,$$

gdzie: P_i – udział regionu i w planowanej sprzedaży dla sektora MSP,
SP_i – udział regionu i w strukturze sprzedaży dla sektora MSP,
UR_i – udział regionu i w strukturze całego rynku ubezpieczeń dla MSP,
α – ustalany arbitralnie współczynnik z przedziału <0; 1>.

Takie rozwiązanie jest bardzo proste, zawiera jednak element uwzględnienia udziału regionu w rynku. Mankamentem takiego rozwiązania jest trudność znalezienia istotnie różnego od jedności współczynnika α, takiego, który gwarantuje strukturę planu przewidującą wzrost sprzedaży we wszystkich regionach. Ponieważ zwykle zakłada się niewielki wzrost sprzedaży na poziomie krajowym, część regionów może mieć przewidziany niższy plan niż osiągnięta już sprzedaż.

Drugi sposób polega na dalszej rozbudowie modelu planowania o elementy związane z rynkiem ubezpieczeń w sektorze MSP (m.in. oszacowanie udziałów regionów w rynku) oraz na próbie znalezienia funkcji, która pozwoli określić możliwości wzrostu sprzedaży w zależności od osiągniętego udziału w rynku. W dalszej części opracowania opisane jest rozwinięcie struktury modelu.

Opis proponowanego rozwiązania. Pierwszym etapem ustalenia planu według opisywanej propozycji jest oszacowanie udziałów w rynku dla poszczególnych regionów. Aby znaleźć wartość udziału towarzystwa ubezpieczeniowego w rynku, należy oszacować wartość całego rynku w każdym terytorium. Można tego dokonać, szacując wartość całego rynku usług ubezpieczeniowych dla sektora MSP. Całkowitą wartość rynku ubezpieczeń MSP w Polsce można oszacować na podstawie średniej składki przypadającej na klienta MSP w towarzystwie oraz liczby aktywnych podmiotów gospodarczych w segmencie MSP w kraju w danym okresie[3]. Specyfikacja modelu przyjmuje następującą postać:

[3] Dane na podstawie raportu o stanie sektora MSP w Polsce w latach 2005-2006 (wyd. PARP 2007), dostępnego w wersji elektronicznej na stronie http://www.parp.gov.pl/index/more/1583.

$$WR = ST \times LM,$$

gdzie: WR – całkowita oszacowana wartość rynku ubezpieczeń MSP w Polsce,
LM – liczba aktywnych przedsiębiorstw w segmencie MSP w Polsce,
ST – przeciętna składka ubezpieczenia w segmencie MSP w towarzystwie.

Jakość oszacowania wolumenu rynku zależy od reprezentatywności badanej zbiorowości firm. Przyjmuje się, że badana zbiorowość to ubezpieczane przez towarzystwo przedsiębiorstwa z sektora MSP. Korzystając z danych towarzystwa, przeciętną składkę wyliczono ze wszystkich zawartych przez towarzystwo umów z przedsiębiorstwami zaliczanymi do MSP. W tym przypadku przyjęcie zbioru firm za reprezentatywny było uzasadnione wielkością udziału towarzystwa w rynku, który znacznie przekracza 10%.

Oszacowany wolumen rynku (WR) dzieli się pomiędzy regiony zgodnie z opracowanym algorytmem, na podstawie wytworzonego w regionie PKB (aproksymowanych poziomem dochodu ludności brutto). W ten sposób dla każdego regionu otrzymuje się potencjalną wartość rynku ubezpieczeń MSP w i-tym regionie (WR_i).

Udział w rynku poszczególnych oddziałów regionalnych szacuje się z wykorzystaniem wartości zebranej przez towarzystwo składki oraz oszacowanej wartości rynku każdego regionu

$$UT_i = FS_i / WR_i,$$

gdzie: UT_i – udział towarzystwa w rynku ubezpieczeń MSP w regionie i,
FS_i – faktyczna wartość składki zebranej w regionie i,
WR_i – oszacowana wartość rynku ubezpieczeń MSP w regionie i.

Należy przy tym pamiętać, aby do wyliczeń wziąć dane za ten sam okres.

Oszacowany udział oddziałów regionalnych w regionalnym rynku jest podstawą do ustalenia udziału poszczególnych regionów w planie firmy. Podstawą wyliczenia wspomnianego udziału jest możliwy przyrost sprzedaży wynikający z oszacowanego udziału regionu w rynku (WS_i). W proponowanym rozwiązaniu związek pomiędzy tymi wielkościami opisuje się za pomocą funkcji wykładniczej, której argumentem jest udział w rynku, a wartości określają możliwy przyrost sprzedaży w regionie. Funkcja ma postać:

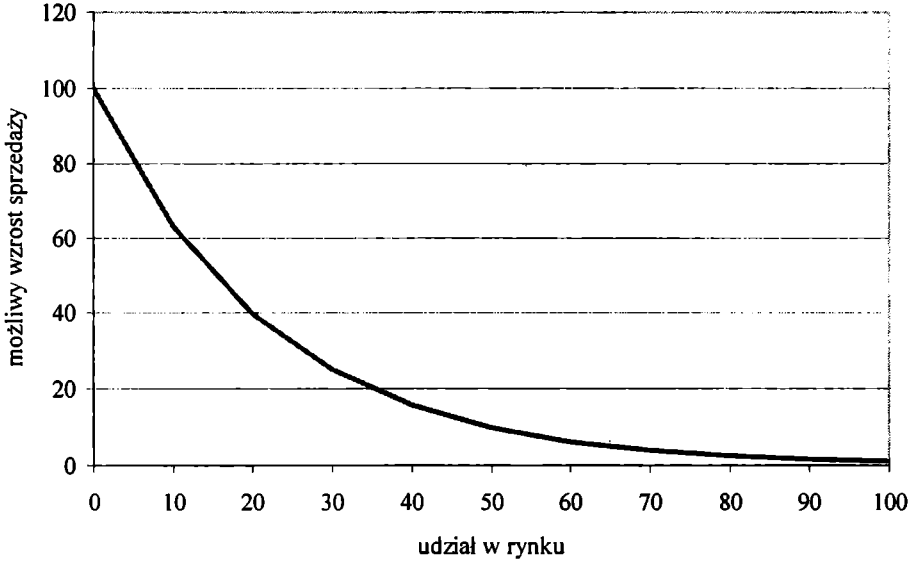
$$WS = a^{UT},$$

gdzie: WS – możliwy do osiągnięcia wzrost sprzedaży.

Wartość podstawy funkcji wykładniczej (a) ustala się w taki sposób, aby spełniała warunek, że $a^{UT_{\text{kraju}}} = WS_{\text{krajowe}}$, czyli że $\log_a\left(WS_{\text{krajowe}}\right) = UT_{\text{kraju}}$.

Funkcja wykładnicza dobrze realizuje kryteria, jakie powinna spełniać zależność pomiędzy udziałem w rynku a możliwym przyrostem składki przypisanej, jest

silnie, ale nie jednostajnie malejąca w rozpatrywanym przedziale (udział w rynku jest wartością z przedziału <0, 1>) oraz osiąga wartość bardzo bliską zeru dla argumentów bliskich jedności. Ponadto szybciej maleje dla niskich wartości udziału w rynku niż dla wysokich. Taka funkcja pozwala silnie zaktywizować regiony o niskim udziale w rynku, natomiast planowany wzrost sprzedaży dla regionu o wysokim udziale w rynku jest tym niższy, im bardziej ten udział przewyższa przeciętny dla kraju. Przykładem funkcji spełniającej opisane wyżej wymagania jest funkcja wykładnicza o podstawie $a = 0{,}01$. Na rys. 1 pokazano tę funkcję.



Rys. 1. Funkcja wiążąca potencjalny wzrost sprzedaży z udziałem w rynku

Źródło: opracowanie własne.

Algorytm wyliczenia udziału poszczególnych jednostek terytorialnych w planie krajowym polega na tym, że wylicza się wartości bezwzględne planowanej sprzedaży dla każdego oddziału regionalnego, a następnie przelicza się je na wartości względne.

$$TS_i = (WS_i + 1) \times PS_i,$$

gdzie: TS_i – teoretyczna wartość przypisanej składki w i-tym regionie,
WS_i – współczynnik wzrostu przypisanej składki w i-tym regionie,
PS_i – przypis składki w regionie w poprzednim okresie.

Natomiast udział poszczególnych regionów w planie (P_i) określony jest jako:

$$P_i = \frac{TS_i}{\sum_i TS_i} \times 100\%.$$

Plan sprzedaży regionu powstaje jako iloczyn planu krajowego oraz wskaźnika P_i dla tego regionu.

Algorytm planowania można podsumować następująco:

1. Za pomocą danych towarzystwa szacuje się średnią wielkość składki ubezpieczeniowej pozyskiwanej od jednej firmy MSP w Polsce.

2. Wylicza się całkowitą potencjalną wartość rynku ubezpieczeń MSP w Polsce jako iloczyn średniej składki i liczby aktywnych podmiotów MSP (obliczenie *WR*).

3. Dzieli się wartość rynku na poszczególne regiony za pomocą wskaźnika udziału w całkowitym dochodzie ludności (obliczenie WR_i).

4. Określa się dla poszczególnych regionów ich udział w rynku (udział faktycznej sprzedaży towarzystwa w oszacowanej wartości rynku w regionie – obliczenie UT_i).

5. Określa się za pomocą funkcji wykładniczej możliwy do osiągnięcia przyrost sprzedaży w regionie (w %) na podstawie osiągniętego udziału w rynku (obliczenie WS_i).

6. Wylicza się teoretyczną wartość wolumenu sprzedaży na podstawie wielkości sprzedaży w poprzednim okresie i obliczonego powyżej współczynnika wzrostu (obliczenie TS_i).

7. Na podstawie teoretycznych wartości przypisanej składki w regionach tworzona jest struktura podziału planu pomiędzy regiony (obliczenie P_i).

8. Oblicza się wartości absolutne planu w poszczególnych regionach jako iloczyn P_i i planowanej sprzedaży dla kraju w kolejnym okresie.

Proponowane rozwiązanie jest jednym z wielu możliwych rozwiązań problemu planowania terytorialnego. Jego zaletą jest prostota zastosowania i łatwość dostępu do danych, na których rozwiązanie zostało oparte. Dane o dochodach ludności w ujęciu powiatowym są publikowane przez Główny Urząd Statystyczny. Oznacza to, że rozwiązanie to można zastosować wtedy, gdy granice oddziałów regionalnych pokrywają się z granicami powiatów lub województw.

Elementem zmniejszającym dokładność planowania opartego na strukturze terytorialnej jest brak możliwości jednoznacznego przypisania firm MSP do administracyjnych jednostek podziału terytorialnego. W praktyce jednostka organizacyjna towarzystwa zazwyczaj nie odmawia ubezpieczenia klientowi, którego siedziba znajduje się poza obsługiwanym przez nią terenem. Zdarza się, że przedsiębiorca posiada kilka filii w różnych powiatach lub województwach. Czynnik ten nie zaburza poważnie jakości modelu.

Wydaje się, że przedstawiona propozycja może stosunkowo niewielkim kosztem poprawić jakość planowania i wpłynąć na poprawę wyników sprzedaży. Zwiększenie sprzedaży jest skutkiem lepszego wykorzystania potencjału runku i nadawania regionom ambitnych, ale osiągalnych planów sprzedaży.

A PROPOSAL OF IMPROVEMENT IN INSURANCE SALES PLANNING FOR SMALL AND MEDIUM ENTERPRISES

Summary

The quality of territorial planning of sales in an insurance company can be improved by implementing a market model which includes variables reflecting the economic power of regions. A proposal described here leading to sales planning improvement is devoted to small and medium enterprises' insurance market. The most convenient and useful variable which reflects differences among insurance demand in different regions is regional gross domestic product. The proposal shows the way how to implement this variable into sales planning to make it more accurate.

Marcin Wojciel – mgr, doktorant w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.