

Andrzej Bąk

Akademia Ekonomiczna we Wrocławiu

WYBRANE MODELE ANALIZY HISTORII ZDARZEŃ I ICH ZASTOSOWANIA W BADANIACH PREFERENCJI

1. Wstęp

W badaniach preferencji wykorzystuje się metody i modele reprezentujące tzw. podejście dekompozycyjne. Zalicza się do nich metody analizy łącznej i metody wyborów dyskretnych. W ostatnich latach można zaobserwować szybki rozwój metod opartych na wyborach oraz wzrost liczby ich zastosowań w badaniach empirycznych. Wynika to m.in. ze sposobu gromadzenia danych preferujących modele wyborów dyskretnych, który jest zbliżony do naturalnych zachowań rynkowych konsumentów. Gromadzone są dane o preferencjach wyrażonych konsumentów (wyborach), które mają charakter antycypacyjny i nie zawsze znajdują potwierdzenie w rzeczywistych decyzjach rynkowych.

Preferencje konsumentów nie są zatem stałe, lecz z różnych powodów zmieniają się w czasie (np. w wyniku różnego rodzaju doświadczeń, wpływu otoczenia, mody). Preferencje i ich zmiany można traktować jako pewne zdarzenia obserwowane na rynku, których ciąg tworzy określoną historię zdarzeń (wyborów). Do analizy tego typu danych można wykorzystać modele analizy historii zdarzeń, które są szerzej stosowane na gruncie socjologii, demografii i techniki (analiza niezawodności).

Celem artykułu jest przedstawienie podstawowych założeń i modeli analizy historii zdarzeń oraz wskazanie możliwości ich wykorzystania w badaniach preferencji konsumentów. Zamieszczono także przykład ilustrujący zastosowanie jednego z modeli analizy historii zdarzeń w badaniach empirycznych w obszarze pomiaru i analizy preferencji konsumentów.

2. Analiza historii zdarzeń

Analiza historii zdarzeń (AHZ) jest gałęzią statystyki rozwijaną przede wszystkim na gruncie socjologii i demografii. Opracowane metody i modele są jednak adaptowane i stosowane w innych obszarach badań, także w ekonomii. Dlatego też terminologia dotycząca problematyki związanej z analizą historii zdarzeń jest bardzo zróżnicowana. Spotyka się m.in. następujące określenia (zob. Frątczak [1997]):

- analiza historii zdarzeń (*event history analysis*) – w socjologii matematycznej,
- analiza przeżycia (dożycia) (*survival analysis*) – w demografii, medycynie, biologii,
- analiza trwania (*duration analysis*) – w ekonomii, naukach społecznych,
- analiza przejścia (*transition analysis*) – także w ekonomii i naukach społecznych,
- analiza niezawodności (*reliability analysis*) – w technice, inżynierii,
- analiza czasu niepowodzeń (*failure time analysis*) – także w technice i inżynierii.

Początkowo analiza historii zdarzeń była kojarzona głównie z badaniami zdarzeń interpretowanych w kategoriach porażek lub katastrof (o konotacjach negatywnych), co dobrze oddają niektóre przytoczone określenia. W ostatnich latach obszary zastosowań AHZ rozszerzają się i obejmują także zdarzenia interpretowane w kategoriach pozytywnych. Przykłady takich zdarzeń znaleźć można również na gruncie ekonomii, np. znalezienie pracy, zakup określonego towaru, wygrana w loterii. W tym kontekście określenie „analiza historii zdarzeń” jest pozbawione cechy wartościowania badanych zdarzeń i wydaje się najbardziej uniwersalne.

W literaturze przedmiotu proponowane są różne definicje AHZ. W pracy E. Frątczaka [1997] przytoczone są m.in. następujące:

- definicja Coxa i Oaksa z 1984 r.: „W centrum zainteresowań analizy przeżycia jest grupa bądź grupy jednostek, dla których jest określony czas wystąpienia zdarzenia.”;
- definicja Blossfelda, Hamerlego i Mayera z 1989 r.: „Przez termin analizy historii zdarzeń rozumiemy metody statystyczne stosowane do analizy odstępów czasu pomiędzy sukcesywnymi stanami przejść lub wydarzeniami. Liczba stanów pobytu, w których mogą przebywać jednostki objęte analizą, jest skończona, ale zdarzenia mogą wystąpić w dowolnym momencie. W analizie historii zdarzeń konsekwentnie są wykorzystywane metody statystycznej analizy procesów stochastycznych z dyskretną przestrzenią stanów i czasem ciągłym.”;
- definicja Allisona z 1995 r.: „Analiza dożycia (przeżycia) jest zbiorem metod statystycznych do studiowania wystąpień i rozkładów zdarzeń w czasie.”

Z przytoczonych definicji wynika, iż przedmiotem badań w analizie historii zdarzeń są zdarzenia i rozkłady zdarzeń w czasie. Pojęcie zdarzenia zaś oznacza przejście badanej jednostki z jednego stanu do drugiego. Przestrzeń stanów jest skończona i dyskretna, a czas jest najczęściej zmienną ciągłą.

W analizie historii zdarzeń wykorzystuje się następujące typy danych:

– dane przekrojowe: $y = [y_i] = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$

– dane przekrojowo-czasowe i dane panelowe: $y = [y_i^t] = \begin{bmatrix} y_1^1 & y_1^2 & \cdots & y_1^T \\ \vdots & \vdots & & \vdots \\ y_i^1 & y_i^2 & & y_i^T \\ \vdots & \vdots & & \vdots \\ y_n^1 & y_n^2 & \cdots & y_n^T \end{bmatrix},$

przy czym z danymi przekrojowo-czasowymi mamy do czynienia, jeżeli i oznacza różne obiekty w okresach $t = 1, \dots, T$; o danych panelowych zaś mówimy wówczas, gdy i oznacza ten sam obiekt we wszystkich okresach $t = 1, \dots, T$.

Wśród dotychczasowych osiągnięć analizy historii zdarzeń znajduje się wiele modeli i metod ich estymacji. W jednej z klasyfikacji modeli analizy historii zdarzeń wyróżnia się następujące grupy [Frątczak 1997]:

- modele nieparametryczne – o nieokreślonej postaci analitycznej rozkładu,
- modele parametryczne – o określonej postaci analitycznej rozkładu (znana jest funkcja gęstości oraz parametry rozkładu),
- modele semiparametryczne – niektóre elementy modelu są wyspecyfikowane parametrycznie, inne natomiast nie (np. funkcja czasu o wartościach nieujemnych).

W grupie modeli semiparametrycznych znajduje się m.in. model proporcjonalnego hazardu Coxa, który został wykorzystany w zamieszczonym dalej przykładzie.

Semiparametryczny model proporcjonalnego hazardu (nazywany często w literaturze przedmiotu regresją Coxa) przybiera postać:

$$h(t|X) = h_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m) = h_0(t) \cdot e^{\sum_{j=1}^m \beta_j X_j}, \quad (1)$$

gdzie: $h(t|X)$ – funkcja hazardu (wskaźnik hazardu) dla zmiennych objaśniających X i czasu t ;

$h_0(t)$ – funkcja hazardu bazowego dla zmiennych objaśniających X przyjmujących wartości równe zero (składnik modelu niewyspecyfikowany parametrycznie – jest funkcją czasu);

$$e^{\sum_{j=1}^m \beta_j x_j} -$$

funkcja wykładnicza, w której parametry β_j reprezentują wpływ zmiennych objaśniających X na wartość wskaźnika hazardu $h(t|X)$ (składnik modelu wyspecyfikowany parametrycznie).

3. Badania preferencji

Pojęcie preferencji (relacji preferencji) odgrywa w ekonomii, zwłaszcza w teorii zachowań konsumentów i teoriach użyteczności, bardzo ważną rolę. Relacja preferencji jest podstawą badań indywidualnych decyzji (wyborów) konsumentów (popytu indywidualnego), co umożliwia z kolei analizę popytu rynkowego. Bezpośredni pomiar użyteczności (rozumianej jako poziom zadowolenia czy satysfakcji konsumenta) jest zadaniem trudnym. W celu jego rozwiązania stosuje się metody mierzące do kwantyfikacji użyteczności na podstawie pojęcia preferencji. Preferencje są relacją binarną odnoszoną do wektorów opisujących wielowymiarowe obiekty, tzw. profile dóbr lub usług, których zbiory tworzą koszyki towarów lub plany konsumpcji. Relacja preferencji umożliwia przypisanie konsumentowi indywidualnej skali preferencji, na której można wartościować profile i optymalizować wybory rynkowe.

W badaniach preferencji konsumentów wykorzystuje się metody umożliwiające pomiar (kwantyfikację) preferencji oraz modele odwzorowujące zachowania rynkowe konsumentów. Ważną grupę narzędzi badawczych w tym obszarze stanowią tzw. metody dekompozycyjne, do których zalicza się metody analizy łącznej i wyborów dyskretnych (zob. [Bąk 2004]).

Podstawą analizy danych o preferencjach w podejściu dekompozycyjnym jest model dekompozycyjny, reprezentujący zależność między preferencjami konsumentów a atrybutami profilów, którą w sposób ogólny można przedstawić w postaci równania:

$$U_{is} = f_s(\mathbf{X}, \boldsymbol{\beta}, \varepsilon_{is}), \quad (2)$$

gdzie: U_{is} – użyteczność całkowita i -tego profilu dla s -tego konsumenta wyrażająca jego preferencje;

f_s – funkcja preferencji s -tego konsumenta;

$\mathbf{X}$ – macierz zawierająca realizacje zmiennych objaśniających opisujących profile (poziomy atrybutów lub realizacje zmiennych sztucznych reprezentujących układ czynnikowy);

$\boldsymbol{\beta}$ – macierz parametrów (użyteczności częściowych);

ε_{is} – składnik losowy modelu.

Wartości zmiennej objaśnianej U_{is} mogą być mierzone na skalach niemetrycznych lub metrycznych, a wartości zmiennych X objaśniających są mierzone najczęściej na skalach niemetrycznych. Wartości zmiennej objaśnianej są wynikiem bezpośrednich ocen respondentów i reprezentują ich preferencje. Wartości zmiennych objaśniających reprezentują z kolei poziomy atrybutów opisujących oceniane profile.

Na podstawie zgromadzonych ocen respondentów (użyteczności całkowitych profili) dokonuje się podziału (dekompozycji) preferencji całkowitych w celu oszacowania tzw. użyteczności częściowych poziomów atrybutów oraz obliczenia udziałów poszczególnych atrybutów w całkowitej użyteczności każdego profilu (por. [Green, Wind 1975]). Najważniejsze znaczenie w pomiarze preferencji wyrażonych konsumentów mają dwie grupy metod dekompozycyjnych, tj. metody analizy łącznej (*conjoint analysis methods*) oraz metody wyborów dyskretnych (*discrete choice methods*).

W metodach wyborów dyskretnych wykorzystuje się m.in. warunkowy model logitowy, którego estymację można przeprowadzić na podstawie modelu proporcjonalnego hazardu. Funkcja wiarygodności częściowej w modelu proporcjonalnego hazardu jest bowiem taka sama jak w warunkowym modelu logitowym. Do maksymalizacji tej funkcji można wykorzystać metodę optymalizacji Newtona-Raphsona. Zagadnienie to jest szczegółowo omówione w pracy ([Ying, Kuhfeld 1995]).

4. Przykład

W lipcu 2002 r. przeprowadzono badanie preferencji konsumentów piwa jasnego wśród mieszkańców Jeleniej Góry i Lwówka Śląskiego. Badanie zrealizowano metodą wyborów dyskretnych, a do estymacji parametrów wykorzystano model proporcjonalnego hazardu Coxa (zob. [Bąk, Rybicka 2004; Gatnar, Walesiak 2004]).

Celem badania jest identyfikacja czynników, które wywierają decydujący wpływ na wybór określonych gatunków i marek piwa. W badaniu uwzględniono atrybuty i poziomy zamieszczone w tab. 1. Prowadzi to do układu czynnikowego typu L^{MP} (gdzie L oznacza liczbę poziomów każdego z atrybutów, M jest liczbą wszystkich atrybutów, P zaś to liczba profili w każdym zbiorze) o rozmiarze $3^{4.5}$. Przeprowadzono redukcję kompletnego układu czynnikowego za pomocą iteracyjnego algorytmu Fedorova, który umożliwia znalezienie optymalnego nieortogonalnego układu czynnikowego z wytypowanego wstępnie zestawu danych.

Tabela 1. Atrybuty i poziomy charakteryzujące piwo jasne

Atrybut	Poziomy
Kraj	Polska, Niemcy, Czechy, Holandia, Dania
Cena	do 2 zł, 2-4 zł, powyżej 4 zł
Zawartość alkoholu	do 1,0%, od 1,8-5,0%, powyżej 5,0%
Rodzaj opakowania	butelka, puszka, kufel
Pojemność opakowania	0,33 l, 0,5 l, powyżej 0,5 l

Źródło: opracowanie własne.

Szczegółowa charakterystyka badania przedstawia się następująco:

- eksperyment czynnikowy:
 - kompletny układ czynnikowy: $3^{4.5}$ zbiorów (20 zmiennych po 3 poziomy),
 - wstępny zbiór danych: 19 683 zbioru,
 - minimalna liczba zbiorów: $20 \cdot (3-1)+1 = 41$ zbiorów,
 - przyjęto w badaniu: 45 zbiorów,
 - efektywność układu czynnikowego: $D = 85,96$;
- układ badania:
 - liczba bloków: 3,
 - liczba zbiorów w bloku: 15,
 - liczba profili w zbiorze: 6 (5 plus opcja rezygnacji z wyboru);
- liczba ankiet:
 - rozprawdzonych: 300,
 - wykorzystanych: 235.

Na rysunku 1 przedstawiono jeden ze zbiorów ocenianych przez respondentów w badaniu ankietowym.

Wybierz preferowany profil piwa jasnego lub zrezygnuj z wyboru (zaznacz jedną z 6 opcji):

Kraj	Cena	Alkohol	Opakowanie	Pojemność	Wybieram opcję
Polska	2-4 zł	pow. 5,0 %	kufel	pow. 0,5 l	1
Niemcy	pow. 4 zł	do 1,0 %	puszka	0,5 l	2
Czechy	pow. 4 zł	do 1,0 %	puszka	0,33 l	3
Dania	2-4 zł	1,8–5,0 %	butelka	0,5 l	4
Holandia	do 2 zł	do 1,0 %	kufel	0,33 l	5
Żaden z tych profili					6

Rys. 1. Przykład zbioru ocenianego przez respondentów

Źródło: opracowanie własne.

Każdy z respondentów oceniał 15 zbiorów zawierających po 6 profili (5 profili reprezentujących różne poziomy atrybutów charakteryzujących piwo oraz opcja rezygnacji z wyboru któregośkolwiek z nich). Uzyskano 235 poprawnie wypełnionych ankiet z następującym podziałem na bloki: blok 1 – 75, blok 2 – 78, blok 3 – 82 ankiety. Łącznie zgromadzono 21 150 obserwacji. W wyniku estymacji warunkowego modelu logitowego (modelu proporcjonalnego hazardu) uzyskano oszacowania parametrów przedstawione w tab. 2.

Oszacowane użyteczności cząstkowe poziomów atrybutów zostały wykorzystane do obliczenia użyteczności całkowitych każdego z profili. Następnie obliczono prawdopodobieństwa wyboru profili w każdym ze zbiorów oraz prawdopodobieństwa wyboru każdego profilu z całego ocenianego zestawu. Wybrane wyniki w postaci pięciu profili o największych prawdopodobieństwach wyboru i pięciu profili o najmniejszych prawdopodobieństwach wyboru spośród 270 ocenianych profili zestawiono w tab. 3.

Tabela 2. Parametry modelu

Atrybut	DF	Parametr	Błąd	χ^2	Pr > χ^2	Iloraz szans
Kraj						
Polska	1	-0,19009	0,08048	5,5779	0,0182	0,827
Niemcy	1	-0,84030	0,08626	94,8857	<,0001	0,432
Czechy	1	-0,66955	0,08359	64,1529	<,0001	0,512
Dania	1	-0,96916	0,09008	115,7588	<,0001	0,379
Holandia	1	-1,37733	0,09538	208,5386	<,0001	0,252
Brak	0	0	.	.	.	.
Cena						
Brak	0	0	.	.	.	.
2-4 zł	1	0,35644	0,05740	38,5609	<,0001	1,428
Do 2 zł	1	0,50510	0,05521	83,7052	<,0001	1,657
Powyżej 4 zł	0	0	.	.	.	.
Alkohol						
Brak	0	0	.	.	.	.
1,8-4,5%	1	-0,02562	0,05012	0,2614	0,6092	0,975
Do 1,0%	1	-0,85350	0,05995	202,7177	<,0001	0,426
Powyżej 5,0%	0	0	.	.	.	.
Opakowanie						
Brak	0	0	.	.	.	.
Butelka	1	0,18360	0,05470	11,2653	0,0008	1,202
Kufel	1	0,25960	0,05625	21,3002	<,0001	1,296
Puszka	0	0	.	.	.	.
Pojemność						
Brak	0	0	.	.	.	.
0,33 l	1	-0,31304	0,05465	32,8142	<,0001	0,731
0,5 l	1	0,00462	0,05182	0,0080	0,9289	1,005
Powyżej 0,5 l	0	0	.	.	.	.

Źródło: opracowanie własne.

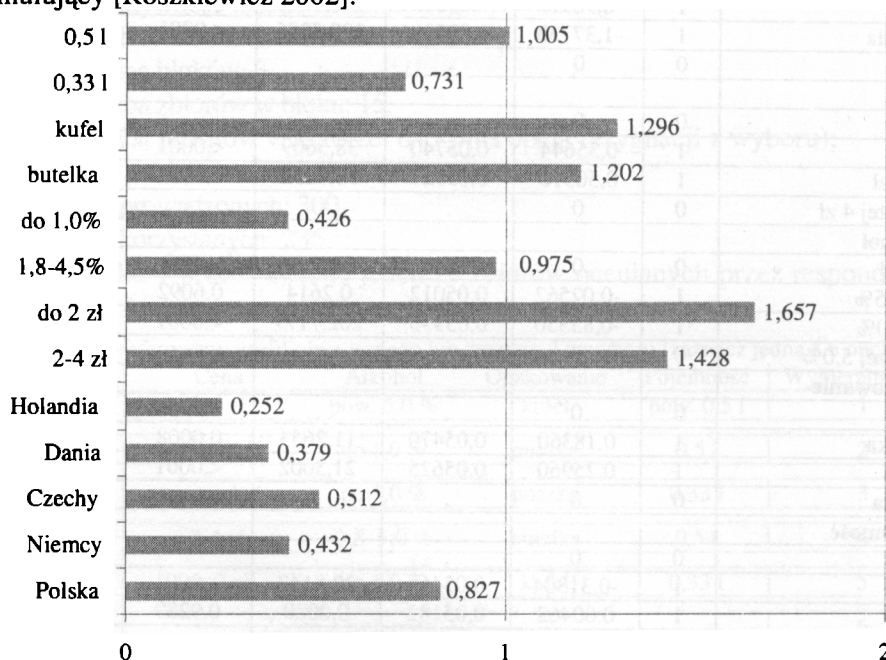
Tabela 3. Wartości prawdopodobieństw wyboru

Kraj	Cena	Alkohol	Opakowanie	Pojemność	Prawdopodobieństwo
największe prawdopodobieństwa wyboru					
Polska	do 2 zł	powyżej 5,0%	kufel	powyżej 0,5 l	0,0104
Polska	do 2 zł	1,8-4,5%	kufel	powyżej 0,5 l	0,0102
Polska	do 2 zł	powyżej 5,0%	butelka	0,5 l	0,0097
Polska	do 2 zł	1,8-4,5%	butelka	powyżej 0,5 l	0,0094
Polska	2-4 zł	powyżej 5,0%	kufel	powyżej 0,5 l	0,0090
najmniejsze prawdopodobieństwa wyboru					
Holandia	powyżej 4 zł	do 1,0%	butelka	powyżej 0,5 l	0,0008
Dania	powyżej 4 zł	do 1,0%	puszka	0,33 l	0,0007
Holandia	2-4 zł	do 1,0%	puszka	0,33 l	0,0007
Holandia	powyżej 4 zł	do 1,0%	puszka	powyżej 0,5 l	0,0006
Holandia	powyżej 4 zł	do 1,0%	butelka	0,33 l	0,0006

Źródło: opracowanie własne.

Interpretację oszacowanych parametrów warunkowego modelu logitowego ułatwia tzw. iloraz hazardu, który informuje o stymulującym lub destymulującym wpływie oszacowanej wartości na prawdopodobieństwo wyboru danego profilu. Na rysunku 2 przedstawiono ilorazy hazardu dla parametrów, których wartości są różne od zera.

Wartość ilorazu hazardu większa od 1 wskazuje na stymulujący wpływ parametru, natomiast wartość ilorazu szans mniejsza od 1 wskazuje, iż wpływ ten jest destymulujący [Rószkiewicz 2002].



Rys. 2. Wartości ilorazu hazardu dla parametrów różnych od zera

Źródło: opracowanie własne.

Wnioski dotyczące przykładu zastosowania metody dyskretnych wyborów:

- w Polsce dominuje spożycie piwa krajowej produkcji,
- o wyborze profilu piwa decydują w kolejności cechy: cena, rodzaj opakowania, pojemność opakowania, zawartość alkoholu, kraj,
- wpływ stymulujący na prawdopodobieństwo wyboru mają: cena do 2 zł, cena 2-4 zł, opakowanie kufel, opakowanie butelka, pojemność 0,5 l.

Obliczenia do zaprezentowanego przykładu wykonano za pomocą procedur pakietu statystycznego SAS/STAT. W pakiecie tym oferowane są narzędzia programowe wspomagające badania preferencji wyrażonych w zakresie zarówno tradycyjnych metod *conjoint analysis*, jak i metod opartych na wyborach. Dostępne procedury wspierają zwłaszcza etap gromadzenia danych oraz etap estymacji użyteczności cząstkowych.

Jeżeli model wyborów dyskretnych jest szacowany na podstawie regresji Coxa (metodą maksymalizacji wiarygodności cząstkowej), to wybory poszczególnych profilów są zdarzeniami, które wystąpiły w czasie t . W celu zaobserwowania historii takich zdarzeń należałoby badanie preferencji powtórzyć w określonych odstępach czasu. Tak zgromadzone szeregi czasowo-przekrojowe lub panelowe umożliwiłyby ocenę zmienności preferencji konsumentów. Zagadnienia te będą przedmiotem dalszych badań.

Literatura

- Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*. Prace Naukowe Akademii Ekonomicznej nr 1013, Seria: *Monografie i Opracowania* nr 157, AE, Wrocław.
- Bąk A., Rybicka A. (2004), *Application of Discrete Choice Methods in Consumer Preference Analysis*, [w:] D. Baier, K.-D. Wernecke (red.), *Innovations in Classification, Data Science, and Information Systems*, Proc. 27th Annual GfK Conference, University of Cottbus, March 12-14 2003, Springer-Verlag, Heidelberg-Berlin, s. 305-312.
- Frątczak E. (1997), *Analiza historii zdarzeń – elementy teorii, wybrane przykłady zastosowań z wykorzystaniem pakietu TDA*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.
- Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.
- Green P.E., Wind Y. (1975), *New Way to Measure Consumers' Judgments*. „Harvard Business Review” 53 (July-August), s. 107-117.
- Rószkiewicz M. (2002), *Narzędzia statystyczne w analizach marketingowych*, C.H. Beck, Warszawa.
- Ying S., Kuhfeld W.F. (1995), *Multinomial Logit Models*, [w:] SUGI 20 Conference *Proceedings*, <http://support.sas.com/techsup/technote/ts722g.pdf>, Cary, SAS Institute.

SELECTED MODELS OF EVENT HISTORY ANALYSIS AND THEIR APPLICATIONS IN THE RESEARCH OF PREFERENCES

Summary

The main aim of the paper is presentation of the event history analysis models and its applications in the area of consumer preferences research. There are presented:

- basic guidelines of the event history analysis,
- possibilities of the use event history models in the consumer preferences analysis,
- proportional hazard model application in the research of consumer preferences.