B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 1 2006

Robert KAPŁON*

ANALIZA DANYCH DYSKRETNYCH W UJĘCIU RETROSPEKTYWNYM – PODSTAWY TEORETYCZNE I ZASTOSOWANIE W MARKETINGU

Przedstawiono historyczny rozwój metod analizy danych dyskretnych – budowa modelu, estymacja i weryfikacja. W obrębie tych zagadnień zaakcentowano wady podejść i historyczne próby ich przezwyciężenia. Następnie podjęto zagadnienie niejednorodności obserwacji, wskazując sposoby radzenia sobie z nią. Omówienie możliwości praktycznego wykorzystania prezentowanych metod ograniczono do zagadnień marketingowych.

Słowa kluczowe: *analiza danych dyskretnych, logit, probit, tabela kontyngencji, największa wiarygodność, niejednorodność obserwacji*

1. Wprowadzenie

Literatura traktująca o statystycznej analizie danych jest bardzo bogata. Prawie w każdej pozycji odnoszącej się do badań marketingowych znajduje się rozdział choć w części dotykający tego problemu (np. [23], [75], [76]). Są również szersze opracowania, zorientowane wyłącznie na metody analizy danych marketingowych [125].

Wiele uwagi poświęca się analizie danych w psychologii i socjologii [10], [19], [56], [69], a jeszcze więcej w biostatystyce [111], [126]. Oprócz tych publikacji pojawiają się osobne, większe opracowania, zorientowane wyłącznie na techniki analityczne. Statystyczna analiza wielowymiarowa zajmuje tu szczególne miejsce [8], [45], [48], [67], [74], [85], [105], [108], [110]. Nierzadko materiał empiryczny stanowi ilustrację poruszanych zagadnień. Większość zaprezentowanych tam technik anali-

* Instytut Organizacji i Zarządzania, Politechnika Wrocławska, ul. Smoluchowskiego 25, 50-372 Wrocław. Robert.Kaplon@pwr.wroc.pl

tycznych wymaga jednak danych, uzyskanych z pomiarów rejestrowanych przynajmniej na skali przedziałowej.

Mniej zainteresowania w literaturze polskiej poświęca się zmiennym dyskretnym, choć w ostatnim czasie zwartych opracowań przybywa [43], [44], [60]. Wiele zagadnień z obszaru analizy danych dyskretnych nie ma tam jednak swojego odbicia. Dlatego celem niniejszego opracowania jest retrospektywna analiza osiągnięć w zakresie problematyki danych dyskretnych i praktyki marketingowej.

2. Modele logit i probit

Funkcja logistyczna została odkryta w dziewiętnastym wieku. Posłużyła do opisu wzrostu populacji, jak i reakcji chemicznych[1]. Przyczynił się do tego Pierre Francois Verhulst. W trzech pracach, opublikowanych między 1838 a 1847 rokiem, przedstawił on propozycję postaci tejże funkcji i jej własności. Pokazał jednocześnie, że krzywa ta dobrze opisuje wzrost populacji kilku krajów. Do takich samych rezultatów doszli Raymond Perl i Lowell J. Reed w 1920 roku. Twierdzi się, że odkryli funkcję logistyczną na nowo, gdyż nie znali prac Verhulsta [27].

Model probitowy ma również długą tradycję. W 1927 roku, poszukując praw rządzących zjawiskami psychologicznymi, L. Thurstone zaproponował binarny model wyboru, oparty na stochastycznej funkcji użyteczności [19], [20]. W modelu tym składniki losowe były niezależne i miały identyczny rozkład normalny, co uprawnia do stwierdzenia, że mamy do czynienia z binarnym modelem probitowym [100]. Odkrycie modelu probitowego przypisuje się również dwóm osobom; są nimi J.H. Gaddum [42] i C.J. Bliss [15], [16].

Powyższe ustalenia nie są bynajmniej do końca uzasadnione. W połowie dziewiętnastego wieku niemiecki naukowiec G.T. Fechner jako pierwszy dokonuje transformacji częstości używając rozkładu normalnego. Nie ma jednak wątpliwości co do tego, że to C.I. Bliss zaproponował nazwę probit. Wspólnie z Gaddumem ustalili standardy estymacji parametrów [27]. Należy tu także wymienić Ronalda A. Fishera, który zaproponował statystyczną procedurę estymacji parametrów w modelu probitowym [100].

W 1938 roku R.A. Fisher wraz z Frankiem Yatesem publikują książkę z tabelami statystycznymi, proponując, jako jedną z wielu, transformację logitową. Jednak termin logit jako analogia do probit – wprowadzony został przez Josepha Bergsona w 1944 roku [4]. Bergson był pierwszym, który dokonywał porównań między mode-

[1] Termin – krzywa logistyczna kojarzy się z Edwardem Wrightem, żyjącym na przełomie szesnastego i siedemnastego wieku, aczkolwiek termin ten nie odnosił się do krzywej logistycznej znanej obecnie.

lem logitowym a modelem probitowym. W kwestii dopasowania rezultaty były bardzo podobne, w przypadku estymacji parametrów probit nastręczał trudności. Między innymi z tego względu Bergson przez długi czas prowadził zażartą walkę i kampanię na rzecz modelu logitowego. Jego argumenty i sugestie nie spotkały się z większym zrozumieniem w środowisku biometryków. Główny zarzut, jaki się pojawiał, to brak rozkładu częstości, który w przypadku modelu probitowego był normalny. Traktowano więc model logitowy jako gorszy [4], [27].

W latach pięćdziesiątych XX wieku dostrzeżono, że transformacja logitowa pozwala na modelowanie dyskretnych, binarnych wyników. Za sprawą Davida R. Coxa [25] regresja logistyczna zyskała na popularności [4]. Dokonano również uogólnienia na przypadek zmiennej dyskretnej o kilku kategoriach. Model taki nazwano wielomianowym modelem logitowym [62], [17], [90], [119]. Pomimo tych propozycji nadal nie istniała podstawa koncepcyjna modeli logitowych. Sytuacja zaczęła się jednak zmieniać pod wpływem rozwoju teorii wyborów dyskretnych.

2.1. Modele oparte na teorii wyborów dyskretnych

W latach sześćdziesiątych dzięki dostępności danych na poziomie indywidualnych zachowań zaczęto skupiać uwagę na zmianach popytu między konsumentami [98]. To wymagało specjalnego podejścia, gdyż zmienna objaśniana zdefiniowana jako wybór dla przyjętej jednostki ma charakter dyskretny. Na przykład, popytu na samochód dla jednego gospodarstwa domowego nie można wyrazić w postaci zmiennej ciągłej, gdyż zmienna ta dotyczy wyboru: pojazd zostanie wybrany lub nie [124].

Duży wpływ na rozwój teorii wyboru miał Duncan Luce, który wprowadził aksjomat niezależności (*IIA*). Formalnie oznacza on, że stosunek prawdopodobieństwa wyboru alternatywy A i B nie zależy od rozpatrywanego zbioru alternatyw. Następnie wykazał, że implikacją tego aksjomatu jest prawdopodobieństwo wyboru alternatywy, równe jej użyteczności podzielonej przez sumę użyteczności wszystkich alternatyw [87], [98]. Jacob Marschak poszedł dalej i pokazał, że aksjomat *IIA* implikuje model zgodny z maksymalizacją użyteczności [123].

Znaczny wkład w rozwój wielomianowych modeli logitowych (*MNL*) miał Daniel McFadden [96], gdyż udowodnił coś więcej niż J. Marschak. Pokazał, że model *MNL* implikuje przyjęcie niezależnych – o jednakowych rozkładach Gumbela (znanych też jako dystrybuanta ekstremalnych wartości typu *I*) – składników losowych [123]. To odkrycie dało solidne podstawy modelom logitowym, a Daniel McFadden – za wkład wniesiony do ekonomii wyborów dyskretnych – w 2000 roku otrzymał nagrodę Nobla[2].

[2] McFadden nazwał model warunkowym, wielomianowym modelem logitowym.

Dalszy rozwój modeli logitowych to próba rezygnacji, choć w części, z kłopotliwego założenia o niezależności między alternatywami. Dyskusyjność tego aksjomatu wynika z przyjęcia założenia, że jednostka ma stałe, dobrze znane, wewnętrznie spójne preferencje, czyli zdolności do takiego przetwarzania informacji, które pozwolą jej wyłonić opcję o maksymalnej użyteczności [113]. Wiele jednak przykładów świadczy przeciwko takiej interpretacji. Uważa się, że bardzo często jednostka nie ma dobrze zdefiniowanych preferencji. Konsekwencją tego jest odrzucenie inwariantnego i przyjęcie konstruktywnego jej charakteru, co oznacza, że preferencje kształtowane są w momencie, kiedy się o nie pyta. Sam proces tworzenia, czy ich kształtowania, będzie zależny od samego charakteru decyzji oraz zdolności, jaką posiada jednostka do przetwarzania informacji [117].

Pojawił się więc gniazdowy model logitowy, którego koncepcja skupia się wokół podziału wszystkich alternatyw na rozłączne grupy (tutaj gniazda). Każdy z nich może przynależeć tylko do jednego gniazda. Ten zabieg ma na celu dopuszczenie zależności między alternatywami należącymi do tych samych gniazd i jej braku pomiędzy alternatywami w odrębnych gniazdach [99]. Tego typu model ze względu na swój charakter może być przydatny w ograniczonym zakresie, gdyż powoduje konieczność wcześniejszego podziału marek na odpowiednie gniazda, co zdaniem niektórych może być trudne i nie do końca odzwierciedlać faktyczny proces decyzyjny [106].

Wiele modeli wyboru dyskretnego – jak np.[129]: model uporządkowany, model logitowy par kombinowanych (*PCL*), międzygniazdowy model logitowy (*CNL*), model różnicowania produktów (*PD*), wielomianowy model logitowy (*MNL*), gniazdowy model logitowy (*NL*) – można otrzymać jako specjalny przypadek uogólnionego modelu ekstremalnych wartości (*GEV*). Model ten jest zgodny z zasadą maksymalizacji stochastycznie zdefiniowanej funkcji użyteczności. Wprowadzony przez Daniela McFaddena, bazuje na uogólnionej dystrybuancie rozkładu ekstremalnego [99], [101].

Jak można spostrzec, uogólniony model ekstremalnych wartości daje możliwość tworzenia własnych modeli, które w danych warunkach najlepiej są dopasowane do danych empirycznych. W tym celu należy zdefiniować pewną funkcję. Jednak – i to przemawia na niekorzyść modeli *GEV* – brak ogólnej idei, intuicji przemawiającej za przyjęciem tejże funkcji, może zniechęcać do budowy własnego modelu. Trudno na przykład znaleźć dobre uzasadnienie dla proponowanej funkcji, której przyjęcie prowadzi do wielomianowego modelu logitowego. Z drugiej jednak strony, modele otrzymane w wyniku specyfikacji takiej funkcji są dość elastyczne. Potwierdzeniem tego jest przykład, w którym arbitralnie przyjmuje się jej postać – w tym sensie, że nie przemawia za tym jakaś wiedza na temat badanego zjawiska – i otrzymany model lepiej dopasowany jest do danych empirycznych niż *MNL* czy *NL* [123].

Należałoby wspomnieć o jeszcze jednym rezultacie teoretycznym. Gaudry i Dagenais [46] zaproponowali model DOGIT, który swoją postacią przypomina wielomianowy model logitowy. Jednak w odróżnieniu od niego, a to za sprawą pewnych parametrów, aksjomat *IIA* nie jest spełniony. Nie można pominąć faktu, że alternatywy

w modelu DOGIT zachowują się jak substytuty. Tym samym występujące tam parametry traktuje się jako parametry substytucji. Daje to podstawę do krytyki, gdyż parametry nie wyrażają w żaden sposób zjawiska podobieństwa między markami. Innymi słowy, podobieństwo między atrybutami marek nie jest funkcją rozważanego parametru.

W praktyce badań marketingowych modele logitowe i probitowe, regresje logistyczne są bardzo często wykorzystywane (por. [40], [89], [101]). Nierzadko też stanowią składową budowanych modeli. Przykładowo, wielomianowy model logitowy: posłużył do analizy koszyka produktów nabytych podczas jednego aktu zakupów [115], stał się podstawą segmentacji preferencji, co z kolei umożliwiło ustalenie determinantów przestawiania się na inne marki [77], umożliwił ocenę istotności zmiennych objaśniających mających potencjalny wpływ na dokonywane wybory kawy mielonej [61].

Z kolei w pracy [112] model probitowy stał się podstawą do budowy dynamicznego modelu, pozwalającego analizować sekwencje zachowań z przeszłości. Takie dynamiczne ujęcie problemu wyboru produktów zwiększa możliwości predykcyjne.

P.K. Kannan i G.P. Wright swoją propozycję badania struktury rynku oparli na gniazdowym modelu logitowym. Wykorzystując dane panelowe dotyczące nabywców kawy, potwierdzili istnienie dwóch wyraźnych segmentów [78].

2.2. Wielowymiarowy model probitowy

Prowadzone do tej pory rozważania dotyczące modelu probitowego zorientowane były wyłącznie na przypadek jednowymiarowej zmiennej losowej. Rozszerzenie modelu probitowego na przypadek wielowymiarowy, w wypadku stochastycznej funkcji użyteczności podali w 1970 roku J.R. Ashford i R.R Bowden [7]. Jednocześnie wykorzystując swoje podejście, zastosowali go do przypadku dwuwymiarowej zmiennej.

Atrakcyjność wielowymiarowego podejścia wynika z porzucenia kłopotliwego założenia o niezależności (*IIA*), na którym w większości opierają się modele logitowe. Koncepcja wielowymiarowego modelu probitowego skłania do stwierdzenia – wobec możliwości pozbycia się uciążliwego założenia o niezależności – że model ten jest lepszy od modeli z rodziny logitowej. Jednak, z punktu widzenia możliwości aplikacyjnych, trudno przychylić się do tej opinii. Przyczyną takiego stanu rzeczy są trudności, wynikające z estymacji parametrów modelu. I tak, aby znaleźć estymatory metodą największej wiarygodności, należy wiele razy całkować po obszarach, których wymiar jest o jeden mniejszy od wymiaru zmiennej dotyczącej wyboru produktów. Ponieważ nie można analitycznie obliczyć takich całek, dokonuje się tego numerycznie. Jednak i ten sposób nie gwarantuje sukcesu. Nawet dla kilku wymiarów (np. trzech) jest to uciążliwe.

Wart odnotowania jest rezultat osiągnięty przez J. Hausmana i D. Wise [68], pozwalający w efektywny sposób szacować parametry; pokazano to dla trzech i czterech

wymiarów. Z praktycznego punktu widzenia, jak stwierdza Wise, liczba nie powinna przekraczać pięciu (za Lerman i Manski [86]).

Wiele jednak problemów na gruncie klasycznego podejścia – tj. globalne i lokalne maksimum funkcji wiarygodności, wrażliwość na początkowe wartości parametrów, problemy zbieżności – można rozwiązać wykorzystując podejście bayesowskie. Intensywność badań w tym obszarze sprawia, że podejście to daje coraz więcej możliwości przy estymacji skomplikowanych modeli. Przełomowa w tym wypadku okazała się praca [93], w której wykorzystuje się podejście bayesowskie do estymacji wielowymiarowego modelu probitowego. Z jej wyników korzysta DeSarbo, Kim i Fong [32], aby pozycjonować luksusowe samochody, oraz DeSarbo, Kim, Wedel i Fong [36] w procedurze *unfolding* skalowania wielowymiarowego.

2.3. Zależność między wyborami – stochastyczna i deterministyczna

Z punktu widzenia modelowania marketingowego (wyboru produktów) istotne jest uwzględnienie zależności między produktami wybieranymi przez konsumentów. W konsekwencji chodzi o zależność między zmiennymi dyskretnymi. W tym miejscu należy uczynić konstatację, wyrażającą pogląd na temat uwzględnienia owej niezależności.

Przyjmuje się, że zależność może pojawić się na poziomie stochastycznym lub deterministycznym. W pierwszym wypadku definiuje się użyteczność niezależnie od kontekstu wyboru (niezależność na poziomie deterministycznym), ale łagodzi się założenie o niezależności między nieobserwowalnymi częściami użyteczności – a więc miedzy składnikami losowymi. Ten rodzaj zależności reprezentują takie modele jak: gniazdowy model logitowy, model ekstremalnych wartości, model DOGIT i wielowymiarowy probit.

W drugim podejściu, odwrotnie do pierwszego, składniki losowe traktuje się jako niezależne, odchodzi się natomiast od założenia, że użyteczności są niezależne na poziomie deterministycznym. W pracy [79] proponuje się pewną funkcję kontekstu i wykorzystuje ją w stochastycznym modelu skalowania wielowymiarowego. Pokazano tam jednocześnie, że podejście w którym uwzględnia się zależności między markami lepiej odzwierciedla rzeczywiste warunki decyzyjne konsumentów.

3. Analiza zmiennych zgrupowanych w tabeli kontyngencji

Drugą grupę metod i procedur analizy danych dyskretnych reprezentują metody, w których nie wyróżnia się zmiennej zależnej. Skupiając się na nich – i kierując się ustaleniami Agrestiego [4] – należałoby przyjąć, że analiza danych dyskretnych

(w takim ujęciu) miała swój początek w 1900 roku. Wtedy Karl Pearson, zainspirowany problemem losowości wyników ruletki w Monte Carlo, wprowadził statystykę chi-kwadrat. Obowiązujące określenie *tabela kontyngencji* zostało również zaproponowane przez niego. Obok Pearsona wymienia się także G. Udny Yule, który w latach 1900–1912 badał związki w tabelach kontyngencji. Zdefiniował indeks bezpośrednio odnoszący się do liczebności komórek, dziś znany jako współczynnik *Q* Yule'a.

Trzecią znamienitą postacią mającą wkład w analizę danych dyskretnych był Ronald A. Fisher. W 1922 roku wprowadził pojęcie stopni swobody. Zaproponował dokładny test statystyczny. Rozwinął metodę korelacji kanonicznej dla tabel kontyngencji. Należy wspomnieć również o definicji interakcji drugiego rzędu w trójwymiarowej tabeli kontyngencji pochodzącej od Bartleta. Przedstawił on również test statystyczny, pozwalający wykluczyć te interakcje [39].

Kontynuując retrospekcję dokonań w zakresie badania związków w tabeli kontyngencji, należy wymienić prace Goodmana i Kruskala [50]–[53]. Związki te jednak ograniczały się tylko do par zmiennych. Rozszerzenie na przypadek wielowymiarowy daje analiza log-liniowa, której początek datuje się od pracy Bircha [12]. Była ona impulsem do dalszych badań nad modelami log-liniowymi. Następnie należy wymienić prace Habermana z zakresu analizy reszt [63], [64], rozważań nad modelami w których liczba parametrów rośnie wraz ze wzrostem próby, teorii estymacji parametrów [65] oraz pewnych propozycji w wypadku zmiennych porządkowych [66]. Niemniej jednak to za sprawą Goodmana [55], [54] porządkowy model log-liniowy został dalej rozwinięty i zyskał większe zainteresowanie. Od tego czasu zaproponowano wiele różnych podejść i uogólnień (por. [9], [22], [107], [116]).

Okazuje się, że model log-liniowy silnie koresponduje z innymi modelami. Dzięki niemu model analizy klas ukrytych można wyrazić w bardziej ogólny sposób. W konsekwencji prowadzi to do otrzymania modeli, które wychodzą poza klasyczne ujęcie modelu klas ukrytych zaproponowanego przez Lazarsfelda i Henry'ego [70]. Pomimo różnych założeń koncepcyjnych modelu log-liniowego i logitowego istnieje silne powiązanie między nimi. Wprowadzenie przez Neldera i Wedderburna [109] koncepcji uogólnionych modeli liniowych pozwala przedstawić model log-liniowy i logitowy jako jeden z takich uogólnionych modeli. Dzięki tak zwanej funkcji wiążącej, regresje logistyczne, modele logitowe i probitowe są przypadkami szczególnymi uogólnionych modeli.

Intensywność badań w obszarze analizy danych zgrupowanych w wielowymiarowych tabelach kontyngencji może skłaniać do przypuszczeń, że równie intensywnie rozwija się praktyka marketingowa. Okazuje się jednak, że w porównaniu z modelami logitowymi model log-liniowy jest wykorzystywany bardzo rzadko – często natomiast w naukach biologicznych i medycznych oraz w socjologii (por. [6], [13], [41]). Zastanawiające są tego przyczyny, gdyż wiele różnych zagadnień na gruncie badań marketingowych z powodzeniem można rozwiązać za pomocą modelu log-liniowego. Po-

twierdzają to prace, w których wykorzystuje się ten model. Przykłady z akcentem położonym na cel badania przedstawiono poniżej.

Przeznaczając duże środki na reklamę prasową, należy zadać pytanie o jej efektywność. Dość często za jej miarę przyjmuje się wskaźnik ED. Wyraża on proporcję potencjalnych odbiorców reklam, którzy jej nie widzieli, widzieli ją raz, dwa razy itd. Jest on również podstawą do obliczenia innych często wykorzystywanych mierników, takich jak: wskaźnik dotarcia, wskaźnik częstotliwości czy GRPs (*gross rating points*). Choć istnieje kilka podejść do szacowania ED, są one jednak niedokładne i mają pewne obostrzenia [28]. Dlatego Danaher [28] zaproponował podejście oparte na modelu log-liniowym, pozwalające analizować wielowymiarową tabelę kontyngencji. Zmienne w tym modelu reprezentowane są przez różne czasopisma i przyjmują wartości równe liczbie przeglądanych (czytanych) wydań na przestrzeni zadanego okresu. Badanie przeprowadzone na próbie 5201 osób pozwoliło na porównanie i empiryczną weryfikację modelu. Z punktu widzenia reklamodawcy łatwiej i efektywniej można opracować harmonogram kampanii reklamowych, redukując koszty [28], [29].

W kolejnym badaniu model log-liniowy wykorzystano do zbadania wpływu takich czynników jak dochód, edukacja i mobilność na wybór nowych usług telekomunikacyjnych [58]. Podobną analizę przeprowadzili DeSarbo i Hildebrand [34], stawiając sobie za główny cel budowę modelu predyktywnego, w którym jedną lub kilka zmiennych dyskretnych wybiera się na zmienną zależną. Na jego podstawie szacuje się prawdopodobieństwo tego, że obserwacja pochodzi z ustalonej kategorii zmiennej zależnej.

Na uwagę zasługuje również analiza z udziałem modelu log-liniowego w konfrontacji empirycznej z analizą interakcji detekcji i modelem regresji [14]. Rozważono dwa problemy marketingowe. Pierwszy sprowadzał się do określenia (zdefiniowania) segmentów o największym potencjale. Potencjał wyrażany był procentowo jako wskaźnik struktury osób, które nabyły określony produkt w segmencie. Wzięto pod uwagę takie zmienne demograficzne jak: wiek, edukacja i rejon zamieszkania. Drugi problem badawczy dotyczył firmy zaangażowanej w przesyłki reklamowe. Wykorzystując wiele zmiennych (około 100) opisujących gospodarstwo domowe – takich jak: płeć głowy rodziny, zajęcie głowy rodziny, liczba posiadanych samochodów, dochód itd. – próbowano ustalić zależność między nimi a odpowiedzią na ofertę przesyłaną pocztą. Jako cel postawiono sobie porównanie metod. Okazało się, że model log-liniowy jest lepszy od swoich konkurentów. Jedno z wyjaśnień, jakie formułują Blattberg i Dolan [14], odnosi się do nieliniowego wpływu zmiennych demograficznych na wskaźnik nabycia/odpowiedzi.

Częściej model log-liniowy wykorzystuje się do analizy danych związanych z przestawianiem się na inne marki. Buduje się (macierz) dwuwymiarową tabelę kontyngencji, w której każdy element reprezentuje liczę osób lojalnych i nielojalnych. Analiza takiej tabeli może zostać przeprowadzona z użyciem testu chi-kwadrat. Jednak pozwoli ona tylko stwierdzić, że zależności w tabeli występują lub nie. Aby dociec ich przyczyn, należy dokonać bardziej wnikliwej analizy. Z pomocą przychodzi model log-liniowy.

Przykładowo więc, Iacobucci i Henderson [73] rozważają problem przestawiania się na inne marki samochodów. Budując m.in. modele quasi-niezależne, stwierdzili, że lojalność jest istotnym czynnikiem wyjaśniającym zależności. Jednak jego eliminacja nie pozwala na stwierdzenie, że przestawianie ma charakter losowy. Skłania to do przypuszczeń o występowaniu jakiegoś mechanizmu. Faktycznie, okazało się, że przestawienie ma charakter asymetryczny – co z punktu widzenia menadżera marketingu jest bardzo istotne. Taki wniosek byłby niemożliwy do sformułowania przy użyciu „standardowego podejścia" [73].

4. Estymacja i weryfikacja

W 1922 roku R. Fisher zaproponował funkcję wiarygodności. Od tego czasu metoda największej wiarygodności, przez wzgląd na bardzo dobre własności estymatorów, jest często stosowana na etapie estymacji parametrów modelu. Przykładem mogą być modele wykorzystywane w marketingu (por. [37]). Aby jednak otrzymać dobre rezultaty, zarówno liczebność próby jak i liczba obserwacji przypadających na estymowany parametr powinna być duża.

Praktyka badawcza pokazuje, że rozmiar próby nie zawsze jest zadowalający. Często stoją za tym koszty jej pobrania. W takich sytuacjach użytecznym jest wykorzystanie warunkowej funkcji wiarygodności. Definiuje się ją jako rozkład warunkowy, w którym statystyki dostateczne dla pewnych parametrów są ustalone. Cox [26] zwrócił na to uwagę w wypadku binarnej regresji logistycznej. Niestety, tylko w szczególnych wypadkach zaproponowane przez niego dokładne podejście może być zaimplementowane i wykonane przez komputer. Dlatego Hirji, Mehta i Patel [72] przedstawili rekurencyjny algorytm, rozwiązujący ten problem. Dalsze istotne propozycje w tym zakresie zawierają prace [11], [71].

Zdarza się również, że zbiór danych jest bardzo duży, a mimo to komórki w tabeli kontyngencji są mało liczne lub dane nie są zbilansowane. W tym wypadku, przez wzgląd na zbyt liczną próbę, dokładne podejście nie może zostać wykorzystane. Dlatego Mehta, Patel i Senchaudhuri [104] proponują podejście oparte na sieci i metodzie Monte Carlo.

Korzyści, jakie niesie estymacja oparta na warunkowej funkcji wiarygodności i dokładnych rozkładach są chyba niedoceniane przez praktyków. Na potwierdzenie słuszności tej tezy King i Ryan [84] podają, że w 1999 roku opublikowano 2770 artykułów (według indeksu cytowań) traktujących o regresji logistycznej (tytuł lub słowa kluczowe zawierają zwrot: regresja logistyczna), w których parametry w przeważającej części estymowano metodą największej wiarygodności. King i Ryan pokazują jednak, że takie podejście może dawać nierzetelne wyniki.

Podobne wątpliwości, dotyczące liczebności próby, rodzą się w przypadku kryteriów oceny przyjętego modelu. Wykorzystywane często statystyki: chi-kwadrat Pearsona, oparta na podwojonym logarytmie funkcji wiarygodności, Walda oraz Cressie–Read mają asymptotyczny rozkład chi-kwadrat. Gdy próba jest niewielka lub gdy pojawia się wiele komórek pustych w tabeli kontyngencji, powstaje naturalne pytanie o znaczenie asymptotycznej teorii wnioskowania. W takiej sytuacji wykorzystuje się warunkowe i bezwarunkowe dokładne testy statystyczne.

Propozycji rozwiązań w zakresie dokładnych testów statystycznych pojawia się coraz więcej. McCullagh [92] zaproponował warunkowy rozkład statystyk testowych w przypadku gdy próba jest duża, ale występują małe zaobserwowane częstości. Z jego wyniku korzysta Tang [118] i proponuje warunkowy test, który z kolei pozwala ocenić dopasowanie regresji logistycznej, jeśli dane dyskretne są pogrupowane. McDonald, Smith i Forster [95] podjęli się zadania oceny dopasowania modelu log-liniowego w oparciu o statystykę chi-kwadrat i statystykę opartą na ilorazie funkcji wiarygodności. Ze względu na trudności w oszacowaniu dokładnej, warunkowej *p*-wartości (*p-value*) zaproponowali dwa podejścia, oparte na algorytmie próbkowania Metropolis–Hastings. Na tej podstawie stwierdzono, że asymptotyczna *p*-wartość jest niewiarygodna.

Okazuje się, że nie zawsze dokładne podejście jest lepsze. Agresti i Coull [3] pokazują, że przy estymacji przedziałowej proporcji dokładne testy dają gorsze wyniki. Są również sytuacje, w których testy permutacyjne (dokładne testy) zawodzą i warto wykorzystać podejście bootstrap [49]. Podejście to wykorzystano na przykład jako technikę wspomagającą selekcję materiału statystycznego w modelach wyboru dyskretnego bazujących na koncepcji użyteczności [91].

Szczególnego znaczenia nabiera problematyka komputerowych możliwości w tym zakresie. Okazuje się bowiem, że pewne analizy dokładnych testów statystycznych są niewykonalne. Ostatnie postępy w tym zakresie zaowocowały jednak metodami Monte Carlo i Monte Carlo z łańcuchami Markowa oraz podejściem bayesowskim [2]. Warto nadmienić, że to ostatnie z powodzeniem jest stosowane zarówno na etapie wyboru zmiennych w modelu logitowym (por. [21]), jak i diagnostyki opartej na resztach modelu [47]. Podkreśla to szczególne znaczenie i możliwości bayesowskiego podejścia.

5. Niejednorodność obserwacji

Bardzo często konwencjonalne metody i procedury analizy danych marketingowych opierają się na założeniu, że badana próba jest homogeniczna. Rozważając prosty model logitowy, za pomocą którego próbuje się uchwycić, jak zmiana ceny wpły-

nie na wybór, można otrzymać uśrednione wyniki, mimo że wcześniejsze badania wyraźnie wyłaniają dwie skrajne grupy konsumentów – jedna jest bardzo wrażliwa na zmianę ceny, druga natomiast prawie wcale. Odwołując się do innego przykładu, zaczerpniętego z pracy [37], można wyciągnąć wniosek, że nieuwzględnienie heterogeniczności jest przyczyną bardzo słabego dopasowania modelu do danych. W przykładzie tym przedstawiono dane, dotyczące ilości nabywanych słodkości w ciągu 7 dni. Posłużono się rozkładem Poissona. Jedyny estymowany parametr to średnia. Model był źle dopasowany do danych empirycznych, dlatego powtórzono procedurę, uwzględniając heterogeniczność. Wyodrębniono sześć segmentów i otrzymano znacznie lepsze wyniki.

W świetle powyższych wywodów nie może dziwić, że uwzględnienie heterogeniczności przyczynia się do lepszego odzwierciedlenia struktury badanych. Wobec tego pojawia się naturalna potrzeba segmentacji. Jednak, co należy podkreślić, jedno z tradycyjnych zagadnień segmentacji polegające na doborze kryteriów segmentacji, w tym wypadku może okazać się niewłaściwie. Wynika to z tego, że przyjęte zmienne klasyfikacji niekoniecznie muszą być zgodne z przyjętym modelem. Innymi słowy, odwołując się do wcześniejszego przykładu, różnica w średniej ilości nabywanych słodkości nie musi korespondować z wydzielonymi segmentami na podstawie uprzednio określonych zmiennych. Również liczba segmentów, w zależności od przyjętych zmiennych, może być różna.

Jeśliby przyjąć właściwy wybór kryteriów segmentacji, to nasuwa się pytanie: czy problem, przed jakim stoi badacz wobec niejednorodnych danych może być rozwiązany dwuetapowo? Intuicja podpowiada, że w pierwszym kroku należy wykorzystać odpowiednią metodę klasyfikacji; w konsekwencji pozwoli to otrzymać jednorodne grupy. Wtedy drugi krok polegałby na właściwej analizie. Jednakże takie dwuetapowe podejście jest w literaturze z zakresu psychometrii i klasyfikacji krytykowane [20], [35], [38], [57], [127]. W pewnych sytuacjach jest praktycznie niemożliwe. Jako przykład rozważmy zagadnienie budowy mapy percepcji produktów. Do tego celu dość często wykorzystuje się analizę czynnikową. Jednak otrzymana w ten sposób przestrzeń jest wspólna dla wszystkich badanych. Wynika to z wcześniej przyjętego założenia, że konsumenci identycznie postrzegają produkty. Ponieważ mapa powstała w ten sposób daje ograniczone możliwości interpretacyjne, wskazane byłoby więc uwzględnienie heterogeniczności. Pojawia się pytanie, jak podzielić badanych na grupy, aby później wykorzystać analizę czynnikową? Trudno zaproponować dwuetapową procedurę. Problem ten i opisany wcześniej można rozwiązać, wykorzystując podejście oparte na modelach klas ukrytych i mieszankach rozkładów. Co ważne, w tych modelach jawnie nie definiuje się kryteriów segmentacji, gdyż reprezentują one zmienne ukryte. Dzięki dodatkowo wprowadzonej tam zmiennej ukrytej możliwy jest podział badanej zbiorowości na wzajemnie rozłączne, bezpośrednio nieobserwowalne klasy [24], [30], [80], [94], [103], [114], [122]. Własności te przyczyniają się do tego, że analiza klas ukrytych i mieszanki

rozkładów są z powodzeniem wykorzystywane w modelach marketingowych (por. np. [31], [33], [59], [81], [82], [83], [128].

Warty odnotowania jest fakt, że idea mieszanek rozkładów nie jest nowa. Jedną z pierwszych analiz wymagających potraktowania materiału statystycznego jako zbioru niejednorodnych obserwacji, pochodzących z dwóch populacji, przeprowadził w 1894 roku Karl Pearson. Zaproponował model, oparty na mieszance dwóch rozkładów normalnych o różnych średnich i wariancjach. Takie ujęcie zagadnienia pozwoliło, na podstawie danych dotyczących długości 1000 krabów, wyprowadzić wniosek o istnieniu dwóch podgatunków tych skorupiaków.

Chociaż upłynął wiek od propozycji Pearsona, to jednak stosunkowo niedawno, bo od czasu pojawienia się pracy Dempstera, Laida i Rubina w 1978 roku, metody uwzględniające heterogeniczność obserwacji, takie jak mieszanki rozkładów (*FMD*) i analiza klas ukrytych (*LCA*), zyskały duże zainteresowanie. Wspomniani autorzy przezwyciężyli problem estymacji parametrów metodą największej wiarygodności, wykorzystując koncepcję algorytmu EM. Tym samym możliwości aplikacyjne znacznie wzrosły i trudno wskazać obszar dziedziny naukowej, w której wskazane metody pozwalające rezygnować z założenia o jednorodności próby byłyby pomijane.

Wspomniane metody nie są jedynymi sposobami radzenia sobie z niejednorodnością obserwacji. Taką możliwość dają również modele mieszane, a więc modele, w których występują efekty stałe i losowe. Pomimo ich długiej tradycji, stosunkowo niedawno zaczęto tę koncepcję wykorzystywać w modelach analizy danych dyskretnych [1]. Przyczynił się do tego niewątpliwie rozwój technik symulacyjnych oraz wzrost możliwości obliczeniowych. Mieszany model logitowy został na przykład wykorzystany przez Boyda i Melomana [18] do badania popytu na samochody. Niemniej jednak dopiero pewne propozycje McFaddena i Traina [102] pozwoliły wykorzystać w praktyce bardziej skomplikowany mieszany wielomianowy model logitowy.

6. Teoria a praktyka

W konkluzji należałoby stwierdzić, że przeważająca większość cytowanych prac pojawiła się w specjalistycznych czasopismach statystycznych. Praktycy, chcąc czerpać korzyści z tych prac, przy analizie dyskretnych danych marketingowych muszą sprostać bardzo wysokim wymaganiom w zakresie znajomości zagadnień statystycznych. Dlatego słusznie zauważa Agresti [2, s. 18], że „głównym wyzwaniem dla statystyków jest wyjaśnienie metodologii tym, którzy nie są specjalistami w tym obszarze, a mogliby skorzystać z tych metod. Statystycy nie powinni niedoceniać tego wyzwania. Przykładowo, modele dla danych porządkowych wciąż wydają się być wykorzystywane w niewielkim stopniu w wielu dyscyplinach, takich jak nauki społeczne, cho-

ciaż pomiary porządkowe są powszechne. [...] zadanie dla statystyka dotyczące wyjaśnienia interpretacji modelu i różnic pomiędzy modelami [...] nie jest łatwe".

Bibliografia

[1] AGRESTI A., BOOTH J.G., HOBERT J.P., CAFFO B., *Random-effects modeling of categorical response data sociological methodology*, Sociological Methodology, 2000, Vol. 30, Issue 1, s. 27–80.

[2] AGRESTI A., *Challenges for categorical data analysis in the twenty-first century*, [w:] C.R. Rao, G.J. Szekely *Statistics for the 21st Century: methodologies for applications of the future*, Marcel Dekker 2000.

[3] AGRESTI A., COULL B.A., *Approximate is better than `exact' for interval estimation of binomial proportions*, American Statistician, 1998, Vol. 52, Issue 2, s. 119–126.

[4] AGRESTI, A., *An introduction to categorical data analysis*, New York, John Wiley & Son 1996.

[5] AGRESTI, A., *Categorical data analysis*, (2-nd edition) Wiley-Interscience Publication 2002.

[6] AGRESTI, A., *Categorical data analysis*, NY, Wiley, 1990.

[7] AMEMIYA T., *Qualitative response models: A survey*, Journal of Economic Literature, 1981(13), s. 105–111.

[8] ANDERSON T.W., *An introduction to multivariate statistical analysis*, Wiley-Interscience (3rd edition) 2003.

[9] ANDERSON, C.J., *The analysis of three-way contingency tables by three-mode association model,* Psychometrika, 1996(61), s. 465–483.

[10] ARON A., ARON E.N., *Statistics for psychology*, Prentice Hall, 2002.

[11] BAGLIVO J., PAGANO M., *Permutation distributions via generating functions with applications to sensitivity analysis of discrete data*, Journal of the American Statistical Association, 1996, Vol. 91, Issue 435, s. 1037–1036.

[12] BIRCH M.W., *Maximum likelihood in three way contingency tables*, Journal of the Royal Statistical Society, 1963, ser. B (25), s. 220–223.

[13] BISHOP Y.M., FIENBERG STR.E., P HOLLAND P.W., *Discrete multivariate analysis theory and practice*: *theory and practice*, MIT Press, Cambridge, Massachusetts 1975.

[14] BLATTBERG R.C., DOLAN R.J., *An assessment of the contribution of log linear models to marketing research*, Journal of Marketing, 1981, Vol. 45, Issue 2, s. 89–97.

[15] BLISS, C. I., *The method of probits*, Science, 1934(79), s. 38–39.

[16] BLISS, C. I., *The method of probits*, Science, 1934(79), s. 409–410.

[17] BLOCH D.A, WATSON, G.S., *A bayesian study of the multinomial distribution*, The Annals of Mathematical Statistics, 1967(38), s. 1423–1435.

[18] BOYD, J., MELLMAN J., *The effect of fuel economy standards on the U.S. automotive market: A hedonic demand analysis*, Transportation Research A, 1980(**14)**, s. 367–378.

[19] BRZEZIŃSKI J., *Metodologia badań psychologicznych*, Wydawnictwo Naukowe PWN, Warszawa, 2003.

[20] CHANG, W.C., *On using principal components before separating a mixture of Two multivariate normal distributions*, Journal of the Royal Statistical Society – Applied Statistics (Series C), 1983(32), s. 267–275.

[21] CHEN MING-HUI, DEY D.K., *Variable selection for multivariate logistic regression models*, Journal of Statistical Planning & Inference, 2003, Vol. 111, Issue 1/2, s. 37–55.

[22] CHOULAKIAN V., *Exploratory analysis of contingency tables by log-linear formulation and generalizations of correspondence analysis*, Psychometrika, 1988(53), s. 235–250.

[23] CHURCHILL G.A., IACOBUCCI D., *Marketing research: methodological foundations*, South-Western College Publication, 2004.
[24] CLOGG C.C., *Latent class models*, [w:] Arminger G., Clogg C.C., & Sobel M.E., *Handbook of statistical modelling for social and behavioural science*, 1995, s. 311–359, New York, Plenum.
[25] COX D.R., *The regression analysis of binary sequences*, Journal of the Royal Statistical Society. Series B (Methodological), 1958,20, s. 215–242.
[26] COX, D. R., *Analysis of binary data*, London: Chapman and Hall, 1970.
[27] CRAMER J.S., *Logit models from economics and other fields*, Cambridge University Press, 2nd Ed 2003.
[28] DANAHER P., *A log-linear model for predicting magazine audiences*, Journal of Marketing Research, 1988, Vol. 25, Issue 4, s. 356–362.
[29] DANAHER P., *An approximate log-linear model for predicting magazine audiences*, Journal of Marketing Research, 1989, Vol. 26, Issue 4, s. 473–479.
[30] DAYTON C.M., *Latent class scaling analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-126, Thousand Oaks, CA, Sage 1998.
[31] DE SOETE G., HEISER W.J., *A latent class unfolding model for analyzing single stimulus preference ratings*, Psychometrika, 1992, Vol. 58, No. 4, s. 545–565.
[32] DESARBO W., KIM Y., FONG D., *A Bayesian multidimensional scaling procedure for the spatial analysis of revealed choice data*, Journal of Econometrics, 1999, **89**(1–2), s. 79–108.
[33] DESARBO W.S, MANRAI A., MANRAI L., *Latent class multidimensional scaling: A review of recent developments in the marketing and psychometric literature*, [w:] R.P. Bagozzi, *Advanced Methods of Marketing Research*, str. 190–222, Oxford, Blackwell 1994.
[34] DESARBO W.S., HILDEBRAND D.K., *A marketer's guide to log-linear models for qualitative data analysis*, Journal of Marketing, 1980, Vol. 44, Issue 3, s. 40–51.
[35] DESARBO W.S., JEDIDI K., COOL K., SCHENDEL D., *Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups*, Marketing Letters, 1990(3), s. 129–146.
[36] DESARBO W.S., KIM Y., WEDEL M., FONG D.K.H., *A Bayesian approach to the spatial representation of market structure from consumer choice data*, European Journal of Operational Research, 1998(111), s. 285–305.
[37] DILLON W., KUMAR A., *Latent structure and other mixture models in marketing: an integrative survey and overview*, [w:] R.P. Bagozzi, *Advanced methods of marketing research*, s. 295–351. Oxford, Blackwell, 1994.
[38] DILLON WILLIAM R., MULANI N., FREDERICK D.G., *On the use of component scores in the presence of group structure*, Journal of Consumer Research, 1989(16), s. 106–112.
[39] FIENBERG S.E., *Contingency tables and log-linear models: basic results and new developments*, Journal of the American Statistical Association, 2000, Vol. 95, Issue 450, s. 643–647.
[40] FRANSES P.H., PAAP R., *Quantitative models in marketing research*, Cambridge University Press, Cambridge, 2001.
[41] FIENBERG S.E., *The analysis of cross-classified categorical data*, MIT Press, Cambridge, Massachusetts, 1980.
[42] GADDUM, J. H., *Reports on biological standard III. Methods of biological assay depending on a quantal response*, London 1933, Medical Research Council. Special Report Series of the Medical Research Council, No. 183.
[43] GATNAR E. (red.), *Analiza i prognozowanie zjawisk rynkowych o charakterze niemetrycznym*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice, 2003.
[44] GATNAR E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa, 1998.
[45] GATNAR E., WALESIAK M. (red), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 2004.

[46] GAUDRY M., DAGENAIS M., *The dogit model*, Transportation Research B, 1978(13), s. 105–111.
[47] GELMAN A., GOEGEBEUR Y., TUERLINCKX F., VAN MECHELEN I., *Diagnostic checks for discrete data regression models using posterior predictive simulations*, Journal of the Royal Statistical Society, Series C (Applied Statistics), 2000, Vol. 49, Issue 2, s. 247–268.
[48] GIRI N.C., *Multivariate statistical analysis*, Marcel Dekker, New York, 2003.
[49] GOOD PH., *Permutation Test. A practical guide to resampling methods for testing hypotheses*, Springer-Verlag, New York, 1994.
[50] GOODMAN L.A., KRUSKAL W.H., *Measures for association for cross-classification, IV: Simplification of Asymptotic Variances*, Journal of the American Statistical Association, 1972, Vol. 67, s. 415–421.
[51] GOODMAN L.A., KRUSKAL W.H., *Measures for association for cross-classification*, II: *Further Discussion and References*, Journal of the American Statistical Association, 1959, Vol. 54, s. 123–163.
[52] GOODMAN L.A., KRUSKAL W.H., *Measures for association for cross-classification I*, Journal of the American Statistical Association, 1954, Vol. 49, s. 732–764.
[53] GOODMAN L.A., KRUSKAL W.H., *Measures for association for cross-classification*, III: *Approximate Sampling Theory*, Journal of the American Statistical Association, 1963, Vol. 58, s. 310–364.
[54] GOODMAN L.A., *Measures, models, and graphical displays in the analysis of cross-classified data (with discussion)*, Journal of the American Statistical Association, 1991(86), s. 1085–1111.
[55] GOODMAN, L.A., *Simple models for the analysis of association in cross-classifications having ordered categories*, Journal of the American Statistical Association, 1979, Vol. 74, Issue 367, s. 537–553.
[56] GRAVETTER F.J., WALLNAU L.B., *Statistics for the behavioral sciences*, Wadsworth Publishing, 2003.
[57] GREEN P.E., KRIEGER A., *Alternative approaches to cluster-based market segmentation*, Journal of the Market Research Society, 1995, 37(3), s. 231–239.
[58] GREEN P.E., CARMONE F.J., WACHSPRESS D.P., *On the analysis of qualitative data in marketing research*, Journal of Marketing Research, 1977, Vol. 14, Issue 1, s. 52–59.
[59] GROVER R., SRINIVASAN V., *A simultaneous approach to market segmentation and market structuring*, Journal of Marketing Research, 1987(May), s. 139–153.
[60] GRUSZCZYŃSKI M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa, 2002.
[61] GUADAGNI P.M., LITTLE D.C., *A logit model of brand choice calibrated on scanner data*, Marketing Science, 1983, **2**(3), s. 203–238.
[62] GURLAND J., LEE I., DAHM P.A., *Polychotomous quantal response in biological assay*, Biometrics, 1960(16), s. 382–398.
[63] HABERMAN S.J., *The analysis of residuals in cross-classified tables*, Biometrics, 1973(29), s. 205–220.
[64] HABERMAN S.J., *Generalized residuals for log-linear models*, In Proceedings of the Ninth International Biometrics Conference 1, 1976, s. 104–172.
[65] HABERMAN S.J., *Log-linear models for frequency data: Sufficient statistics and likelihood equations,* The Annals of Statistics, 1973(1), s. 617–632.
[66] HABERMAN S.J., *Log-linear models for frequency tables with ordered classifications*, Biometrics, 1974(30), s. 589–600.
[67] HAIR J.F., TATHAM R.L., ANDERSON R.E., BLACK W., *Multivariate data analysis*, Prentice Hall (5th Edition), 1998.
[68] HAUSMAN J., WISE D., *A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences*, Econometrica, 1978, **48**(2), s. 403–426.
[69] HEALEY J.F., *Statistics: a tool for social research*, Wadsworth Publishing, 2004.
[70] HEINEN T., *Latent class and discrete latent trait models: Similarities and differencestr*, Thousand Oaks, California, Sage, 1996.
[71] HIRJI K.F., *Computing exact distributions for polytomous response data*, Journal of the American Statistical Association, 1992, Vol. 87, Issue 418, s. 487–492.

[72] HIRJI K.F., MEHTA C.R., PATEL N.R., *Computing distributions for exact logistic regression*, Journal of the American Statistical Association, 1987, Vol. 82, Issue 400, s. 1110–1117.
73] IACOBUCCI D., HENDERSON G., *Log linear models for consumer brand switching behavior: What a manager can learn from*, Advances in Consumer Research, 1997, Vol. 24, Issue 1, s. 375–380.
[74] JAJUGA K., *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa, 1993.
[75] KACZMARCZYK S., *Badania marketingowe – Metody i techniki*, Państwowe Wydawnictwo Ekonomiczne, Warszawa, 2003.
[76] KACZMARCZYK S., *Badania marketingowe: metody i techniki*, Polskie Wydawnictwo Ekonomiczne, Warszawa, 1999.
[77] KAMAKURA W.A., RUSSELL G.J., *A probabilistic choice model for market segmentation and elasticity structure*, Journal of Marketing Research, 1989, **26**(November), s. 379–390.
[78] KANNAN, P., WRIGHT G., *Modeling and testing structured markets: a nested logit approach*, Marketing Science, 1991, **10**(Winter), s. 58-82.
[79] KAPŁON R., *Pozycjonowanie produktów – próba metodologicznej interpretacji*, rozprawa doktorska, Wrocław, 2003.
[80] KAPŁON R., *Analiza danych dyskretnych za pomocą metody LCA*, [w:] Jajuga K., Walesiak M., *Klasyfikacja i analiza danych – teoria i zastosowania*. Taksonomia nr 9, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 2002.
[81] KAPŁON R., *Estymacja parametrów modelu czynnikowego wykorzystującego klasy ukryte*, [w:] Jajuga K., Walesiak M., *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia nr 11, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 2004.
[82] KAPŁON R., *Mapy pozycjonowania wyrobów przy wykorzystaniu analizy czynnikowej klas ukrytych*, [w:] Jajuga K., Walesiak M., *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia nr 10, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 2003.
[83] KAPŁON R., *Model logitowy z klasami ukrytymi*, Ekonometria XV (w druku) 2004.
[84] KING E.N., RYAN TH.P., *A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression*, American Statistician, 2002, Vol. 56, Issue 3, s. 163–170.
[85] KRZYŚKO M., *Wielowymiarowa analiza statystyczna*, UAM, Poznań, 2000.
[86] LERMAN S., MANSKI C., *On the use of simulated frequencies to approximate choice probability*, [w:] C. Manski D. McFadden (red.), *structural analysis of discrete data with econometric applications*, s. 305–319, MIT Press Cambridge 1981.
[87] LUCE D., *Individual choice behavior*, John Wiley and Sons, New York, 1959.
[88] MAGIDSON Y., *Some common pitfalls in causal analysis of categorical data*, Journal of Marketing Research, 1982, Vol. 19, Issue 4, s. 461–471.
[89] MALHOTRA N.K., *The use of linear logit models in marketing research*, Journal of Marketing Research, 1984, Vol. 21, Issue 1, s. 20–31.
[90] MANTEL, N., *Models for complex contingency tables and polychotomous dosage response curve*, Biometrics, 1966(22), s. 83–95.
[91] MAZURKIEWICZ M., MERCIK J. W., DOBROWOLSKI W., *Verifcation of ideological classifications – a statistical approach*, Control and Cybernetics, 2001, Vol. 30, Issue 4, s. 451–465.
[92] MCCULLAGH P., *The conditional distribution of goodness-of-fit statistics for discrete data*, Journal of the American Statistical Association, 1986, Vol. 81, Issue 393, s. 104–107.
[93] MCCULLOCH R., ROSSI P.E., *An exact likelihood analysis of the multinomial probit model*, Journal of Econometrics, 1994(**64)**, s. 207–240.
[94] MCCUTCHEON A. L., *Latent Class Analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-064, Thousand Oaks, CA, Sage 1987.
[95]MCDONALD J.W., SMITH P.W.F., FORSTER J.J., *Exact tests of goodness of fit of log-linear models for rates*, Biometrics, 1999, Vol. 55, Issue 2, s. 620–624.

[96] MCFADDEN D., *Conditional logit analysis of qualitative choice behavior*, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, s. 105–142, Academic Press, New York, 1974.
[97] MCFADDEN D., *Econometric models of probabilistic choice*, [w:] C. Manski D. McFadden, *Structural analysis of discrete data with econometric applications*, s. 198–272, MIT Press Cambridge, 1981.
[98] MCFADDEN D., *Economic choices*, American Economic Review, 2001(**91**), s. 351–378.
[99] MCFADDEN D., *Modeling the Choice of Residential Location*, [w:] A. Karlqvist, L. Lundqvist, F. Snickars, J. Weibull (red.), *Spatial Interaction Theory And Planning Models*, s. 75–96, North Holland, Amsterdam, 1978.
[100] MCFADDEN D., *Quantal choice analysis: A survey*, Annals of Economic and Social Measurement, 1976, **5**(4), s. 363–390.
[101] MCFADDEN D., *The choice theory approach to market research*, Marketing Science, 1986, **5**(4), s. 275–297.
[102] MCFADDEN D., TRAIN K., *Mixed MNL models for discrete response*, Journal of Applied Econometrics, 2000, Vol. 15, Issue 5, s. 447–470.
[103] MCLACHLAN, G., BASFORD, K., *Mixture models: inference and applications to clustering*, New York, Marcel Dekker, 1988.
[104] MEHTA C.R., PATEL N.R., SENCHAUDHURI P., *Efficient monte carlo methods for conditional logistic regression*, Journal of the American Statistical Association, 2000, Vol. 95, Issue 449, s. 99–108.
[105] MERITER C., *Advanced and multivariate statistical methods*, Pyrczak Publication, 2001.
[106] MEYER R.J., KAHN B.E., *Probabilistic models of consumer choice behavior*, [w:] T.STR. Robertson, H.H. Kassarjian (red.), *Handbook of consumer behaviour*, s. 85–123, Englewood Cliffs, NJ, Prentice-Hall, 1991.
[107] MOOIJAART A., *Three-factor interaction models by log-trilinear terms in three-way contingency tables*, Statistica Applicata, 1992(4), s. 669–677.
[108] MORRISON D.F., *Multivariate statistical methods*, McGraw-Hill Companies (3rd edition), 1990.
[109] NELDER J.A., WEDDERBURN R.W.M., *Generalized linear models*, Journal of the Royal Statistical Society, 1972, ser. A (135), s. 370–384.
[110] OSTASIEWICZ W. (red.), *Statystyczne metody analizy danych*, AE we Wrocławiu, Wrocław, 1999.
[111] PAGANO M., GAUVREAU K., *Principles of biostatistics*, Duxbury Press, 2000.
[112] PAPATLA P., KRISHNAMURTHY L., *A Probit model of choice dynamics*, Marketing Science, 1992, **12**(Spring), s. 189–206.
[113] PAYNE J.W., BETTMAN J.R., JOHNSON E.J., *Behavioral decision research: A constructive processing perspective*, Annual Review of Psychology, 1992, **43**(1), s. 87–131.
[114] ROST J., LANGEHEINE R., *A guide through latent structure models for categorical data*, [w:] Rost J., Langeheine R. (eds.), *Applications of latent trait and latent class models in the social science*, Berlin, Waxmann, 1997.
[115] RUSSELL G.J., PETERSEN A., *Analysis of cross category dependence in market basket selection*, Journal of Retailing, 2000, **76**(3), s. 367–392.
[116] SICILIANO R., MOOIJAART A., *Three-factor association models for three-way contingency tables*, Computational Statistics & Data Analysis, 1997, Vol. 24, Issue 3, s. 337–356.
[117] SLOVIC P., *Construction of preference*, American Psychologist, 1995, **50**(5), s. 364–371.
[118] TANG MAN-LAI, *Exact goodness-of-fit test for binary logistic model*, Statistica Sinica, 2001(11), s. 199–211.
[119] THEIL H., *A multinomial extension of the linear logit model*, International Economic Review, 1969(10), s. 251–259.
[120] THURSTONE L., *A law of comparative judgment*, Psychological Review, 1927(34), s. 273–286.
[121] THURSTONE L., *Psychological analysis*, American Journal of Psychology, 1927(38), s. 368–389.

[122] TITTERINGTON D.M., SMITH A.F.M., MARKOV U.E., *Statistical analysis of finite mixture distributions*, John Wiley & Son, 1985.
[123] TRAIN K.E., *Discrete choice methods with simulation*, Cambridge University Press, Cambridge 2003.
[124] TRAIN K.E., *Qualitative choice analysis*, MIT Press, Cambridge, 1986.
[125] WALESIAK M., *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa 1996.
[126] WATAŁA C., *Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych*, Alfa Medica Press, Bielsko-Biała 2002.
[127] WEDEL M., KAMAKURA W., *Market segmentation: conceptual and methodological foundations*, Dordrecht, Kluwer Academic Publisher, 1999.
[128] WEDEL M., DESARBO W.S., BULT J.R., RAMASWAMY V., *A latent class poisson regression model for heterogeneous count data*, Journal of Applied Econometrics, 1993, **8**(4), s. 397–411.
[129] WEN C., KOPPELMAN F., *The generalized nested logit model*, Transportation Research B, 2001, **35**(7), s. 627–641.

A retrospective review of categorical data analysis – theory and marketing practice

The paper presents historical development of the categorical data analysis for models with explicit response variables defined as well as models without such a distinction. Besides difficulties in model building we focus on methods and procedures for model testing and for the estimation of model parameters. Within these issues we emphasize the drawbacks of the models and historical trials to overcome them. The problem of data heterogeneity and methods that help to handle it were considered. Discussion of practical usefulness of categorical data analysis is limited to marketing problems.

Keywords: *categorical data analysis, logit, probit, contingency table, maximum likelihood estimation, data heterogeneity*