Wstęp

Niniejsza książka jest monografią poświęconą metodom statystycznej analizy danych jakościowych, nazywanych bardziej precyzyjnie danymi niemetrycznymi, oraz danych symbolicznych o bardziej złożonej strukturze. Wypełnia ona wyraźną lukę na rynku wydawniczym w Polsce, na którym nie ma prac na ten temat.

Celem książki jest przedstawienie podstaw teoretycznych każdej z wybranych metod statystycznej analizy danych jakościowych i symbolicznych wraz z zastosowaniami oraz implementacją w programie **R**. Czytelnik, który nie ma odpowiedniego przygotowania statystycznego lub nie zna dobrze programu **R**, powinien zapoznać się z podręcznikiem *Statystyczna analiza danych z wykorzystaniem programu R*, praca zbior. pod red. M. Walesiaka, E. Gatnara, Wydawnictwo Naukowe PWN, Warszawa 2009.

Praca, którą Czytelnik ma przed sobą, składa się z trzynastu rozdziałów i każdy z nich został poświęcony odrębnej metodzie analizy danych. Struktura rozdziału obejmuje część teoretyczną oraz wybrane zastosowania z wykorzystaniem programu **R**. Dodatkową zaletą książki jest prezentacja oraz wykorzystanie w niej własnych pakietów działających w środowisku **R**. Można tutaj wymienić takie pakiety, jak `clusterSim` oraz `symbolicDA`.

Rozdział pierwszy stanowi wprowadzenie do analizy danych jakościowych i symbolicznych. Omówiono tutaj zagadnienia ważne z punktu widzenia dalszych rozdziałów książki. Wyjaśniono w nim takie podstawowe pojęcia, jak macierz danych i tablica danych. Zaprezentowano miary odległości dla danych porządkowych i danych symbolicznych, zagadnienie dyskretyzacji zmiennych ilościowych, wybrane rozkłady prawdopodobieństwa zmiennych dyskretnych oraz wizualizację danych klasycznych i symbolicznych.

W rozdziale drugim zostały pokazane miary niezależności przeznaczone dla zmiennych o charakterze jakościowym, a także opis i zastosowanie analizy korespondencji dla dwóch oraz wielu zmiennych. Jest to metoda badania współwystępowania zmiennych mierzonych na słabych skalach pomiaru (a raczej ich kategorii), która pozwala na graficzne przedstawienie wyników w postaci mapy percepcji w niskowymiarowej przestrzeni.

W rozdziale trzecim omówiono modele logarytmiczno-liniowe, które są szczególnym przypadkiem uogólnionych modeli liniowych dla zmiennych dyskretnych o rozkładzie Poissona. W modelach logarytmiczno-liniowych obiektem podlegającym modelowaniu są liczebności z poszczególnych komórek tablicy wielodzielczej, które traktujemy jak realizacje pewnej zmiennej losowej. W rozdziale przedstawiono modele pełne (dla dwóch i trzech zmiennych) oraz hierarchiczne. Omówiono metodę wyznaczania

najlepszego modelu logarytmiczno-liniowego przez budowanie wielu modeli, różniących się uwzględnioną w nich liczbą zarówno zmiennych, jak i interakcji między zmiennymi, oraz porównanie tych modeli ze sobą pod względem jakości dopasowania. Następnie zaprezentowano sposoby pozyskiwania wiedzy z modelu końcowego i interpretacji wyników.

Rozdział czwarty został poświęcony modelowaniu i prognozowaniu zmiennych dwumianowych. Przedmiotem modelowania jest sztuczna zmienna jakościowa pełniąca funkcję zmiennej objaśnianej, która przyjmuje dokładnie dwie wartości: zero lub jeden. W rozdziale omówiono liniowy model prawdopodobieństwa (LMP), modele logitowy i probitowy oraz zagadnienie prognozowania na podstawie modeli dwumianowych.

Rozdział piąty poświęcono prezentacji modeli zmiennych wielomianowych o kategoriach nieuporządkowanych. Modele takie znajdują zastosowania w ekonomii, m.in. w badaniach preferencji konsumentów dokonujących wyborów rynkowych. Przedstawiono wielomianowy model logitowy i warunkowy model logitowy oraz możliwości ich estymacji za pomocą funkcji dostępnych aktualnie w programie **R**.

Rozdział szósty został poświęcony analizie wariancji. Metoda ta pozwala ocenić wpływ niezależnego czynnika klasyfikującego x_j ($j = 1, \ldots, m$) o charakterze jakościowym na wartości zmiennej zależnej y o charakterze metrycznym. W rozdziale tym przedstawiono zagadnienia związanie z jedno- i dwuczynnikową analizą wariancji, a także dwuczynnikową analizą wariancji przy uwzględnieniu występowania interakcji rzędu pierwszego. Omówiono tam także problematykę tzw. testów *post hoc* służących sprawdzeniu istotności różnic poszczególnych par średnich na różnych poziomach czynnika klasyfikującego oraz podstawowe schematy badań wykorzystujące technikę analizy wariancji.

W rozdziale siódmym przedstawiono rozwiązania metodyczne pozwalające na przeprowadzanie analizy skupień i porządkowania liniowego dla danych porządkowych. Podstawą do ich zastosowania jest odległość GDM2. W analizie skupień wyróżniono dwie procedury postępowania: klasyczną analizę skupień i klasyfikację spektralną. W procedurze porządkowania liniowego zastosowano nową metodę zamiany nominant na stymulanty właściwą dla danych porządkowych (przy konstrukcji dolnego bieguna rozwoju zachodzi konieczność zamiany nominant na stymulanty).

Rozdział ósmy w całości został poświęcony omówieniu metody budowy modeli dyskryminacyjnych i regresyjnych, która umożliwia wykorzystanie zmiennych objaśniających o charakterze jakościowym. Metoda ta opiera się na rekurencyjnym podziale przestrzeni zmiennych i nosi nazwę odnoszącą się do graficznej postaci tego procesu: drzewa klasyfikacyjne i regresyjne. W rozdziale tym pokazano sposoby doboru zmiennych charakterystyczne dla tego rodzaju modeli, oparte m.in. na statystyce χ^2, oraz wyboru modelu w optymalnej postaci.

W rozdziale dziewiątym zaprezentowano modele klas ukrytych, które są przykładem tzw. podejścia modelowego w analizie skupień. W modelach klas ukrytych zmienne obserwowane mają charakter jakościowy. Przedstawiono modele zmiennych binarnych i wielomianowych z uwzględnieniem problemu wyboru modelu i liczby klas. Omówiono

także modele regresji klas ukrytych, w których uwzględnia się dodatkowo zmienne towarzyszące wpływające na przynależność obserwacji do klas.

W rozdziale dziesiątym przedstawiono zastosowanie modeli mieszanek w analizie regresji. Modele mieszanek rozkładów stosowane są wówczas, gdy zbiór obserwacji jest zbiorem niejednorodnym. Celowość podziału badanej zbiorowości na grupy jednorodne, ze względu na przyjęty zestaw cech diagnostycznych, uzasadniona jest istotnymi różnicami relacji pomiędzy zmiennymi (np. wydatkami ogółem względem wybranych cech społeczno-ekonomicznych).

W rozdziale omówiono zagadnienie estymacji parametrów oraz wyboru modelu mieszanek o najlepszej jakości dopasowania. Charakterystyce poddano w szczególności modele najczęściej wykorzystywane w analizie danych jakościowych, tj. modele mieszanek rozkładów dwumianowych oraz rozkładów Poissona.

Rozdział jedenasty poświęcono prezentacji teoretycznych i aplikacyjnych podstaw skalowania wielowymiarowego dla danych niemetrycznych i symbolicznych. Zaprezentowano dwa podejścia optymalizacji funkcji dopasowania, tj. metodę gradientową i metodę majoryzacji. Scharakteryzowano analizę *unfolding*, w której w przeciwieństwie do innych metod skalowania wielowymiarowego danymi wejściowymi nie jest macierz odległości, lecz prostokątna macierz preferencji. W części poświęconej skalowaniu wielowymiarowemu danych symbolicznych przedstawiono modele Interscal, SymScal i I-Scal.

W rozdziale dwunastym przedstawiono rozwiązania metodyczne pozwalające na klasyfikację danych symbolicznych z wykorzystaniem analizy skupień. Spośród metod analizy skupień wyróżniono dwie grupy: metody taksonomii numerycznej i metody taksonomii symbolicznej. Omówiono dwa podejścia w klasyfikacji danych symbolicznych: podejście bazujące na macierzy odległości i podejście bazujące na tablicy danych symbolicznych. Wskazano metody, jakie mają zastosowanie w poszczególnych etapach procedury klasyfikacyjnej w zależności od przyjętego podejścia.

W rozdziale trzynastym przedstawiono podstawy analizy dyskryminacyjnej dla danych symbolicznych. Do metod analizy dyskryminacyjnej, które mogą znaleźć zastosowanie w przypadku danych symbolicznych, zaliczają się przede wszystkim: drzewa klasyfikacyjne, jądrowa analiza dyskryminacyjna oraz metoda K-najbliższych sąsiadów (używana w formie „klasycznej" z wykorzystaniem macierzy odległości obliczonych na podstawie miar symbolicznych). W rozdziale zaprezentowano jądrową analizę dyskryminacyjną opartą na estymatorach intensywności, która jest adaptacją nieparametrycznej analizy dyskryminacyjnej wykorzystującej jądrowe estymatory gęstości, oraz teoretyczne postawy konstrukcji drzew klasyfikacyjnych, które są adaptacją rekurencyjnych drzew klasyfikacyjnych dla danych klasycznych, a także algorytm bayesowskich drzew klasyfikacyjnych, które są rozwiązaniem dostępnym jedynie dla danych symbolicznych.

Ponadto na końcu książki znajduje się dodatek, w którym pokazano sposób przygotowania danych symbolicznych w postaci gotowej do wykorzystania przez procedury i funkcje ujęte w książce dla danych symbolicznych.

Autorzy mają nadzieję, że niniejsza książka okaże się przydatna dla badaczy i praktyków, którzy zajmują się problematyką analizy danych niemetrycznych, nieprecyzyjnych

i nieostrych. Zainteresuje więc z pewnością ekonomistów, psychologów, socjologów, biologów, botaników, archeologów, lekarzy i innych.

Wersję instalacyjną programu **R** oraz dodatkowe pakiety można pobrać ze strony: `http://www.r-project.org/`. Wszystkie skrypty zawarte w książce przetestowano używając wersji 2.13.0 programu **R**.

Na stronie internetowej `http://keii.ue.wroc.pl` znajdują się pliki zawierające wszystkie wykorzystywane dane oraz procedury realizujące zastosowania zamieszczone w książce.