Zarządzanie

Wybrane metody ilościowe w finansach

Dorota Witkowska

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Wybrane metody ilościowe w finansach

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Zarządzanie

Wybrane metody ilościowe w finansach

Dorota Witkowska

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Łódź 2023

Dorota Witkowska (ORCID: 0000-0001-9538-9589)
– Uniwersytet Łódzki, Wydział Zarządzania
Katedra Zarządzania Finansami Przedsiębiorstwa, 90-237 Łódź, ul. Matejki 22/26

RECENZENT
Małgorzata Tarczyńska-Łuniewska

REDAKTOR INICJUJĄCY
Monika Borowczyk

REDAKCJA
Aleksandra Urzędowska

KOREKTA
Małgorzata Mazur

SKŁAD I ŁAMANIE
Munda – Maciej Torz

KOREKTA TECHNICZNA
Anna Jakubczyk

PROJEKT OKŁADKI
efectoro.pl
agencja komunikacji marketingowej

Zdjęcie wykorzystane na okładce: © Depositphotos.com/pressmaster

https://doi.org/10.18778/8142-304-5

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.10213.20.0.M

Ark. wyd. 17; ark. druk. 23,375

ISBN 978-83-8142-304-5
e-ISBN 978-83-8142-305-2

Wydawnictwo Uniwersytetu Łódzkiego
90-237 Łódź, ul. Jana Matejki 34A
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. 42 635 55 77

Pamięci moich Mistrzów, w podzięce moim Bliskim

Spis treści

Wstęp

Metody ilościowe umożliwiają prowadzenie analiz zjawisk społeczno-gospodarczych, tym samym wspomagają podejmowanie decyzji i obiektywną ocenę istniejącej sytuacji. Oprócz tego pozwalają rozpoznać mechanizmy funkcjonowania układów gospodarczych, zidentyfikować istniejące determinanty zjawisk ekonomicznych, a także przewidywać skutki podjętych działań oraz zmiany zachodzące w badanych procesach. Tworzą zatem przesłanki do określania przyszłego rozwoju społeczno-gospodarczego. Metody ilościowe są również wykorzystywane do weryfikacji teorii ekonomicznych, a nawet krążących poglądów i opinii.

Przez ponad 40 lat pracy zawodowej autorka stara się aktywnie wykorzystywać bogatą paletę metod statystycznych, ekonometrycznych i optymalizacyjnych, zarówno w badaniach naukowych, jak i tych realizowanych dla praktyki gospodarczej. Celem tej monografii jest przekazanie jej doświadczeń badawczych z zakresu aplikacji wybranych metod statystycznych.

Kiedy autorka zaczynała pracę jako nauczyciel akademicki, wszyscy studenci kierunków ekonomicznych musieli realizować znaczną liczbę godzin dydaktycznych przeznaczonych na przedmioty ilościowe, takie jak matematyka ekonomiczna, statystyka, ekonometria rozumiana w jej szerokim pojęciu, tzn. obejmująca nie tylko modele opisowe, ale również tzw. modele przepływów międzygałęziowych (stworzone przez Wassily'ego Leontiefa, a dzisiaj zwane analizami *input-output*) i metody programowania liniowego (tzn. modele optymalizacyjne będące elementem badań operacyjnych). Wówczas każdy absolwent posiadał podstawową wiedzę z tego zakresu i rozumiał przynajmniej elementarne pojęcia, a przeszkodę do aplikacji metod ilościowych w praktyce stanowiły możliwości obliczeniowe, ponieważ nie każdy wykształcony ekonomista miał nawet pośrednio dostęp do komputerów. Wynikało to przynajmniej z dwóch powodów. Po pierwsze, komputery były w posiadaniu jedynie dużych i „strategicznych" jednostek, takich jak np. uczelnie, duże przedsiębiorstwa i niektóre agencje rządowe. Wprawdzie istniały tzw. Zakłady Elektronicznej Techniki Obliczeniowej (ZETO), będące komercyjnymi ośrodkami przetwarzania danych, z których usług można było odpłatnie korzystać, ale nie były one powszechnie dostępne, chociażby z powodu

ograniczonych (sprzętowych i pracowniczych) możliwości wykonywania usług. Po drugie, obsługa komputerów, nawet jeśli się miało do nich dostęp, wymagała wiedzy informatycznej[1] daleko bardziej zaawansowanej niż obsługa współczesnych przyjaznych użytkownikom urządzeń.

W okresie gospodarki socjalistycznej nie do końca też wierzono, że obiektywne i racjonalnie prowadzone metodami ilościowymi analizy mogą być przydatne w praktyce gospodarczej, czego wyrazem były wątpliwości pracujących studentów: „co tu optymalizować, jak mam plan do wykonania". Z podobnym podejściem autorka spotkała się również wiele lat później, tj. już w okresie gospodarki rynkowej, kiedy to jeden z dyrektorów dużego banku działającego w Polsce stwierdził, że nie są mu potrzebne nowoczesne metody oceny zdolności kredytowej, bo jego rozliczają z procedur, a straty wynikające z błędnych decyzji kredytowych są wliczone w koszty obsługi klientów.

Ostatnie 30 lat to dla autorki nie tylko przekształcenie polskiej gospodarki z centralnie planowanej w rynkową, ale przede wszystkim przemiany technologiczne, których nieustannie doświadczamy. Jednakże prawdziwą rewolucją tego pokolenia okazał się wręcz niewyobrażalny rozwój technologii, który sprawił, że koncepcje rodem z powieści *science fiction* Lema stały się naszą codziennością. Pojawiły się bowiem komputery osobiste, oprogramowanie rozwiązujące skomplikowane problemy numeryczne poprzez jedno kliknięciem, Internet będący nieograniczoną wręcz skarbnicą wiedzy wszelakiej i komunikatory będące imitacją bliskości. Konieczne stało się zatem nabywanie innych umiejętności – w tym obsługi komputerów, smartfonów i portali społecznościowych. Jednocześnie (niestety) uznano, że technologia i nowe kompetencje zastąpią logiczne rozumowanie.

Szkolnictwo wyższe również ulegało stopniowej zmianie – komercjalizacji, bowiem środki budżetowe dla uczelni „szły za studentem" zatem w „walce konkurencyjnej" pozbywano się z programów nauczania „nieprzyjemnych" i nielubianych przez studentów przedmiotów ilościowych. W konsekwencji od ponad 10 lat absolwenci wielu kierunków ekonomicznych mają znikomy dostęp lub w ogóle nie poznają metod ilościowych, a co najwyżej przechodzą kurs „klikania" bez rozumienia istoty wykorzystywanych metod, a zatem nie wiedzą, jakie informacje otrzymują w wyniku zastosowania profesjonalnego oprogramowania, i nie umieją ani zinterpretować uzyskanych wyników, ani przeprowadzić ich krytycznej analizy. Co więcej, w wielu przypadkach nawet nie potrafią odpowiednio przygotować danych empirycznych, będących „wsadem" do programów komputerowych. Znana reguła

1 Przykładowo, aby przygotować pracę magisterską, musiałam sobie sama napisać program w języku ALGOL, a dostęp do komputera w Ośrodku Obliczeniowym Uniwersytetu Łódzkiego miałam ograniczony do kilku półgodzinnych sesji około północy. Natomiast do doktoratu program (liczący około 2 tys. kart perforowanych) napisał dla mnie programista z Politechniki Warszawskiej, a obliczenia robiłam w Instytucie Badań Jądrowych w Świerku pod Warszawą, bo na Uniwersytecie Łódzkim nie było dostępnych terminów na wielogodzinne przetwarzanie danych.

garbage in, *garbage out* wydaje się ich nie dotyczyć, a brak znajomości metod i zasad ich stosowania, pozostawia pseudoanalitykom nieograniczone pole do aplikacji najbardziej wyszukanych i skomplikowanych metod, które wymagają jedynie kliknięcia w odpowiedni klawisz na klawiaturze lub w ikonę na ekranie monitora.

Autorka jest zdecydowanie przeciwna takim praktykom i choć popiera stosowanie metod ilościowych w analizach społeczno-ekonomicznych, woli, aby używano nieskomplikowanych, ale zrozumiałych (przez analityków) metod, zamiast bełkotania przez ignorantów o „wyklikanych" wynikach. Wszelkiego rodzaju dane pozwalają na wyciągniecie wielu interesujących wniosków, które uzyskać można za pomocą relatywnie prostych narzędzi, a problemem jest jedynie wielkość i gwałtowny rozrost zbiorów danych (tzw. *big data*). Wykorzystywanie bardziej zaawansowanych metod często wiąże się z wieloma rozczarowaniami i koniecznością stosowania wielu podejść w celu uzyskania zadowalających wyników[2], czego też jakby się nie zauważa, pokazując bezkrytycznie (m.in. w artykułach „niby-naukowych") bezsensowne wyniki, które mają świadczyć o „kompetencjach" autora lub braku przydatności metod. Na drugim biegunie niekompetencji zdarza się, że do rangi osiągnięcia naukowego podnosi się obliczenie procentów, prostego indeksu lub estymację funkcji trendu. Innym przykładem może być próba aplikacji dla polskich jednostek gospodarczych modelu predykcji bankructwa, oszacowanego przez Edwarda I. Altmana w 1968 roku dla gospodarki amerykańskiej. Przy czym nie dostrzega się, że minęło już ponad pół wieku od estymacji parametrów tego modelu, a przedsiębiorstwa działające w USA podlegają innemu prawu i regułom funkcjonowania niż polskie firmy. Jednocześnie w pseudonaukowych publikacjach bezmyślnie poszukuje się przyczyn braku skuteczności tego modelu, np. w zaokrągleniach lub przekłamaniach wartości parametru[3], podczas gdy przyczyna jest prosta i dawno odkryta – trzeba ten model dyskryminacyjny oszacować na nowo, uwzględniając istniejące (tu i teraz) realia gospodarcze, o czym zresztą pisał sam Altman.

Rozwój technologii gromadzenia i przetwarzania danych oraz globalizacja rynków finansowych sprawiają, że podejmowanie właściwych decyzji wymaga posługiwania się metodami ilościowymi. Dlatego niniejsze opracowanie zawiera omówienie wybranych metod statystycznych i ekonometrycznych wykorzystywanych w finansach. Zadaniem, które postawiła przed sobą autorka, było pokazanie,

2 Niejednokrotnie zdarzało się, że np. podczas budowania modelu ekonometrycznego akceptowalny okazywał się jego dwudziesty lub pięćdziesiąty wariant, co pokazuje, jak wiele wysiłku wymaga tak – zdawałoby się – proste zadanie.

3 Celowo nie zacytowano tutaj tego typu prac z dwóch powodów: po pierwsze, aby nie wytykać nikogo palcem i nie zawstydzać, zwłaszcza jeśli autor jest na początku kariery, bo istnieje nadzieja, że się dokształci; po drugie, w dzisiejszych czasach ocena „jakości" naukowca dokonywana jest na podstawie liczby cytowań i to niezależnie od tego, w jakim kontekście omawia się daną pracę. Nie warto zatem przysparzać tym autorom dodatkowych punktów w ocenie ich dorobku naukowego, przytaczając ich błędne lub bezwartościowe prace.

że stosowanie nawet relatywnie prostych metod istotnie wspomaga podejmowanie decyzji.

Gromadzenie danych rzeczywistych i prowadzenie zaawansowanych analiz wymaga oczywiście korzystania z komputerów i (najlepiej) profesjonalnego oprogramowania, ale to analityk musi zdecydować, które z metod są adekwatne do charakteru zgromadzonych obserwacji, i dokonać ich świadomego wyboru z katalogu dostępnych w pakiecie. Niezbędna jest również umiejętność wyciągania poprawnych wniosków na podstawie uzyskanych wyników oraz krytyczna ocena ich akceptowalności. Innymi słowy, trzeba wiedzieć, co otrzymuje się w wyniku zastosowania konkretnego narzędzia oraz w jaki sposób można przeprowadzić badanie poprawności uzyskanych wyników.

Monografia składa się z jedenastu rozdziałów, w których omówiono wybrane zagadnienia z zakresu statystyki opisowej i matematycznej, modelowania ekonometrycznego i wielowymiarowej analizy porównawczej. Większość przedstawionych metod została zilustrowana przykładami z zakresu finansów[4]. Prezentowane wyniki badań empirycznych wskazują na pojawiające się problemy i ograniczenia w stosowalności omawianych metod. Kolejność rozdziałów wynika z chęci systematyzacji rozpatrywanych zagadnień.

I tak rozdział pierwszy zawiera omówienie zagadnień związanych z: pomiarem zjawisk społeczno-gospodarczych, metodami pozyskiwania danych oraz ich właściwą prezentacją. Rozdział drugi poświęcono metodom opisu struktury danych, wskazując na ich ograniczenia. W kolejnej części pokazano, jak uogólniać wyniki uzyskane na podstawie badania niepełnego. Czwarty rozdział dotyczy analizy współzależności. Następny zawiera omówienie metod analiz dynamiki zjawisk. Kolejny dedykowany jest modelom ekonometrycznym i to zarówno tym klasycznym, jak i modelom zmiennych binarnych oraz modelom szeregów czasowych. Rozdział siódmy stanowi wprowadzenie do prognozowania, a ósmy do wielowymiarowej analizy porównawczej[5]. Natomiast ostatnie trzy rozdziały stanowią pogłębioną ilustrację aplikacji w finansach większości metod omawianych wcześniej. Tak więc rozdział dziewiąty to opis przeprowadzonych analiz finansowych szeregów czasowych na przykładzie szeregu notowań cen akcji wybranej spółki giełdowej. Natomiast dziesiąty dotyczy dwóch problemów – pierwszy związany jest z zastosowaniem wybranych metod klasyfikacji do oceny wiarygodności kredytowej klientów banku, drugi polega na ocenie kondycji ekonomiczno-finansowej spółek z uwzględnieniem różnych aspektów ich funkcjonowania, przy zastosowaniu mierników taksonomicznych. Ostatni rozdział poświęcono budowie hedonicznych indeksów cen dzieł sztuki, które uwzględniają ich charakterystyki jakościowe.

4 Wszędzie tam, gdzie było to możliwe bez nadmiernej komplikacji obliczeniowej, w rozdziałach 1–8 pokazano rzeczywiste dane.

5 Ilustrację metod omawianych w rozdziale 8 i częściowo 6 przedstawiono w rozdziale 10.

Rozdział 1

Wprowadzenie do analizy danych

Niniejszy rozdział ma charakter wprowadzający, zatem poświęcony będzie podstawowym zagadnieniom związanym z obserwacją, pomiarem i grupowaniem danych wykorzystywanych w finansach. Omówione zostaną podstawowe pojęcia, gdyż każda dziedzina wiedzy posługuje się charakterystycznym dla niej językiem. Rozważania rozpocznie omówienie zagadnień pomiaru, a dalej wskazane zostaną źródła danych i sposób ich prezentacji, a w ostatnim podrozdziale podstawowe pojęcia z rachunku prawdopodobieństwa.

1.1. Podstawowe pojęcia

Statystyka jest nauką o metodach badania prawidłowości występujących w zjawiskach masowych, czyli takich, które mogą zachodzić nieograniczoną liczbę razy i w dużej masie zdarzeń wykazują pewne prawidłowości, jakich nie można zaobserwować w pojedynczym przypadku. Przedmiotem badań statystycznych są zbiorowości statystyczne (populacje), które stanowią zbiór elementów (jednostek) powiązanych ze sobą logicznie i jednocześnie nieidentycznych. Zbiorowość statystyczna to zbiór jednostek (osób, rzeczy lub zjawisk) objętych badaniem statystycznym. Prawidłowe prowadzenie badania wymaga, aby zbiorowość statystyczna była jednoznacznie określona i wyodrębniona. Dokonuje się tego, ustalając cel badania i precyzując, kto lub co należy do badanej zbiorowości, czyli definiuje się jej elementy. Poszczególne jednostki (elementy), które wchodzą w skład badanej zbiorowości statystycznej, posiadają wspólną cechę (właściwość, charakterystykę, atrybut), a jednocześnie różnią się między sobą innymi, sobie właściwymi cechami. Pojedynczą cechę oznacza się przez X, a warianty, jakie może przyjmować, przez x_i (dla $i = 1, 2, \ldots, k$), czyli: $x_1, x_2, \ldots, x_k$. W analizach uwzględnia się tylko te cechy, które są istotne z punktu widzenia celu realizowanego badania.

Analizy zjawisk i obiektów można prowadzić na podstawie:
1) badania pełnego (wyczerpującego lub całkowitego), obejmującego wszystkie jednostki danej zbiorowości statystycznej,
2) badania niepełnego (częściowego), obejmującego jedynie wybrane jednostki badanej zbiorowości statystycznej.

Wybór sposobu badania wynika z celu i zakresu badania, a także wiąże się z zastosowaniem odpowiednich narzędzi badawczych. W pierwszym przypadku będą to metody umożliwiające statystyczny opis badanej zbiorowości. Natomiast w drugim istotny jest sposób doboru elementów do badania, aby możliwe było zastosowanie metod statystyki matematycznej w celu wnioskowania statystycznego o całej zbiorowości na podstawie danych pochodzących z próby[1].

1.2. Pomiar zjawisk, skale pomiarowe

W analizach finansowych wykorzystuje się różnego rodzaju informacje dotyczące zarówno pojedynczych charakterystyk, jak i złożonych zjawisk, obiektów lub procesów społeczno-gospodarczych. Badania prowadzone w naukach społecznych, w tym w szeroko rozumianych finansach, charakteryzują się:
1) złożonością przedmiotów badania (tj. obiektów, procesów i zjawisk),
2) trudnością prowadzenia obserwacji i dokonywania pomiaru,
3) zmiennością w czasie.

Złożoność badań społecznych wynika z konieczności jednoczesnego uwzględnienia w analizach wielu różnych charakterystyk. Innymi słowy, rozpatrywane zjawiska i obiekty są wielowymiarowe (wielocechowe) i do ich opisu wykorzystuje się wiele cech (zmiennych). Przykładem obiektu badania jest przedsiębiorstwo lub kredytobiorca, a zjawiska – notowania cen akcji wybranej spółki. Atrybutami przedsiębiorstwa mogą być: liczba pracowników, osiągany przychód, forma własności, udział w rynku etc. Wiarygodność kredytowa klienta indywidualnego oceniana jest m.in. z punktu widzenia uzyskiwanych dochodów, wykonywanego zawodu, miejsca pracy, płci itp. W przypadku analiz cen akcji bierze się pod uwagę wiele czynników, zarówno związanych ze spółką (np. wskaźniki ekonomiczno-finansowe, charakter działalności, sposób zarządzania), jak i dotyczących jej szeroko rozumianego otoczenia (np. parametry makroekonomiczne, takie jak poziom stóp procentowych, faza wzrostu gospodarczego, sytuacja na rynkach kapitałowych, branża itp.).

1 Szersze omówienie tych problemów znaleźć można w bogatej literaturze przedmiotu, np. w pracach: [Aczel 2000; Luszniewicz, Słaby 2003; Witkowska 2004].

Trudność obserwacji i pomiaru związana jest z faktem, że większość zjawisk nie jest bezpośrednio obserwowalna i występują problemy z ich pomiarem. Przykładowo popyt na dane dobro oznacza chęć nabycia go, ale nie każdy chętny może dokonać zakupu, zatem popytu się nie obserwuje, a jedynie tzw. popyt zrealizowany, który mierzony jest wartością lub wielkością zakupu danego dobra. Innym przykładem jest inflacja, która oznacza wzrost kosztów utrzymania, ale z uwagi na to, że ceny na różne towary i usługi nie rosną w taki sam sposób, to wzrost kosztów utrzymania zależy od koszyka dóbr i usług, będących przedmiotem konsumpcji poszczególnych gospodarstw domowych. Inflacja jest zatem różnie odczuwalna, a więc nie można jej bezpośrednio zaobserwować w skali gospodarki, a jej pomiar zależy od przyjętych założeń, np. odnośnie do koszyka dóbr i usług oraz formuły obliczeniowej wskaźnika zmiany cen. Innym przykładem jest wiarygodność kredytowa, która nie może być bezpośrednio obserwowana ze względu na swoją złożoność, jest natomiast określana za pomocą różnych metod, odmiennych dla klientów indywidualnych i przedsiębiorstw.

Rozwój społeczno-gospodarczy sprawia, że warunki, w jakich funkcjonują jednostki gospodarcze oraz gospodarstwa domowe, ulegają ciągłym zmianom. W konsekwencji badane obiekty i zjawiska również podlegają przeobrażeniom w czasie. Ta zmienność sprawia, że konieczna jest ciągła aktualizacja danych i prowadzenie analiz na bieżąco.

Obserwacje zjawisk i obiektów badania prowadzone są w różnorodny sposób, w różnych miejscach i momentach. W związku z tym uzyskane w wyniku obserwacji statystycznej informacje mogą przyjmować postać danych przekrojowych, czasowych lub panelowych. Dane przekrojowe powstają, gdy w określonym czasie obserwuje się różne obiekty badania, np. w konkretnym okresie analizuje się ofertę różnych banków w celu założenia lokaty lub zaciągnięcia kredytu. Innym przykładem może być analiza wyników finansowych spółek na koniec danego roku obliczeniowego. Dla tego typu danych moment dokonania pomiaru nie jest istotny, a numer każdej obserwacji zazwyczaj oznacza się przez i, ($i = 1, 2, \ldots, n$). O danych czasowych mówi się, kiedy analiza prowadzona jest dla jednego obiektu w różnych momentach lub okresach czasu. Przy czym przez moment rozumie się punkt na osi czasu, a przez okres przedział na osi czasu. Przykładowo badanie dotyczy rozwoju konkretnego przedsiębiorstwa w kilkuletnim okresie, wówczas liczba obserwacji T równa jest liczbie okresów (lub momentów) objętych badaniem, np. kolejnych lat. Szereg czasowy tworzą notowania kursów akcji rejestrowane w trakcie kolejnych sesji giełdowych. W szeregach czasowych numer każdej obserwacji oznacza się przez t, $t = 1, 2, \ldots, T$, a kolejne obserwacje są porządkowane chronologicznie. W przypadku danych panelowych obserwacja prowadzona jest jednocześnie w dwóch wymiarach. Przy czym w praktyce najczęściej wykorzystuje się dane przekrojowo-czasowe oznaczające obserwacje poszczególnych jednostek statystycznych w różnym czasie. Przykładem tego typu danych jest ocena rozwoju wielu spółek w kilkuletnim horyzoncie czasowym w celu budowy portfela

inwestycyjnego, wówczas liczba obserwacji nT równa jest iloczynowi liczby spółek i liczby okresów (lub momentów) objętych badaniem.

Kluczowym pojęciem używanym w statystyce są cechy (statystyczne), czyli właściwości charakteryzujące jednostki wchodzące w skład badanej zbiorowości. Jak wcześniej wspomniano, cechę statystyczną oznacza się przez X, a jej warianty podlegające obserwacji zapisuje się jako $x_1, x_2, \ldots, x_k$, gdzie k, to liczba wyróżnionych wariantów badanej charakterystyki. Wśród cech statystycznych wyróżnia się:

- cechy mierzalne (ilościowe, wymierne), które można wyrazić za pomocą liczb z podaniem odpowiednich uniwersalnych jednostek miary (np. wartość majątku w mln PLN, dochód w tys. PLN, wartość środków trwałych w tys. EUR, kurs euro w złotówkach, liczba rat, wiek w latach, masa w kilogramach, długość w metrach, czas w godzinach, wielkość firmy mierzona liczbą zatrudnionych);
- cechy niemierzalne (jakościowe, niewymierne), których nie można zmierzyć za pomocą uniwersalnych jednostek miary, a jedynie stwierdza się występowanie lub nie określonego wariantu danej cechy (np. forma własności, branża, w jakiej działa przedsiębiorstwo, waluta kredytu, płeć, zawód, wykształcenie, wielkość firmy według podziału na: mikroprzedsiębiorstwa, firmy małe, średnie i duże).

Z terminem „cecha" ściśle związane jest pojęcie zmiennej, którą często określa się jako pomiar cechy. Wśród cech ilościowych wyróżnia się zmienne ciągłe lub skokowe. W pierwszym przypadku zmienna może przyjmować każdą wartość z określonego, skończonego przedziału liczbowego. W drugim przypadku może przyjmować ona jedynie określone wartości z tego przedziału, często należące do zbioru liczb całkowitych lub naturalnych (np. liczba pracowników to cecha skokowa, liczba etatów może być cechą ciągłą). Sposób pomiaru przedmiotu badania pozwala wyróżnić cztery podstawowe skale pomiarowe: nominalną, porządkową, przedziałową i ilorazową (por. [Aczel 2000, s. 36–37; Stanimir 2006, s. 21–22; Sobczyk 2010, s. 15–18]).

Skala nominalna dotyczy cech jakościowych, między którymi można wyróżnić tylko dwie relacje, tj. równości (identyczności) i różności (tj. braku identyczności). Innymi słowy o cechach mierzonych na tej skali w dwóch i więcej jednostkach statystycznych (lub obiektach) można jedynie powiedzieć, czy posiadają taki sam wariant tego atrybutu, lub czy różnią się pod względem tej charakterystyki. W przypadku cech mierzonych na skali nominalnej stosuje się kodowanie binarne za pomocą zmiennych przyjmujących wartość 1, oznaczającą występowanie danego wariantu cechy, lub 0 w sytuacji przeciwnej. Szczególnym przypadkiem skali nominalnej jest skala dychotomiczna (dwudzielna), z której korzysta się w przypadku cechy dwuwariantowej, której przykładem jest płeć lub wyróżnienie wśród spółek tych z udziałem Skarbu Państwa. Kiedy wariantów danej cechy jest więcej niż dwa (np. branża, zawód, forma własności, sektor gospodarki według klasyfikacji PKD, numery tras autobusowych), to mówi się o cechach wie-

lodzielnych (politomicznych). W przypadku cech wielodzielnych do kodowania danych używa się k zmiennych binarnych (gdzie $k > 1$).

Skala porządkowa (rangowa) pozwala określić relacje większości i mniejszości między wyróżnionymi wariantami cech oraz umożliwia porządkowanie obserwacji według wyników pomiaru. Jednakże pomiar i porządkowanie są dokonywane według subiektywnie ustalonych jednostek pomiarowych, zatem działania matematyczne inne niż wymienione są niedopuszczalne. Przykładami pomiarów na tej skali mogą być: grupa ryzyka kredytowego, poziom wykształcenia (podstawowe, średnie, wyższe), ocena produktu (dobry, średni, zły) lub wielkość przedsiębiorstwa (małe, średnie i duże).

Zmienne mierzone na skali przedziałowej (interwałowej) przyjmują wartości liczbowe, które można uporządkować na osi liczbowej z podaniem stałej i obiektywnie istniejącej jednostki pomiarowej. Skala ta nie posiada jednak naturalnego punktu zerowego, który wyznaczany jest arbitralnie na potrzeby konkretnego badania, zazwyczaj według przyjętej konwencji, np. skala temperatury, mierzona stopniami Celsjusza, ma punkt zero w temperaturze zamarzania wody. Umowność punktu zerowego implikuje również nierówność różnic między wartościami na przyjętej skali (np. przy temperaturze 40°C nie jest dwa razy cieplej niż przy 20°C). Skala przedziałowa umożliwia wykonywanie operacji dodawania i odejmowania. Niemożliwe jest natomiast mnożenie i dzielenie dwóch wielkości interwałowych, ich potęgowanie oraz używanie średniej kwadratowej i geometrycznej (por. [Zeliaś 2000, s. 34]). Przykładem pomiaru na skali interwałowej może być wynik finansowy przedsiębiorstwa (zysk lub strata), rok rozpoczęcia działalności firmy. Innym przykładem jest opinia na zadany temat podana w punktach na skali Likerta, która podporządkowana jest zasadom skalowania zrównoważonego, posiada bowiem tyle samo wariantów odpowiedzi pozytywnych, co negatywnych i jest skalą niewymuszoną z neutralną odpowiedzią, kiedy nie ma się zdania na zadany temat. Przyjmuje się, że odległości między kolejnymi wariantami odpowiedzi są identyczne, co pozwala traktować skalę Likerta jako interwałową. Dodatkowo można do poszczególnych odpowiedzi przyporządkować wartości liczbowe, co dopuszcza późniejsze wykonywanie operacji matematycznych (por. [Jajuga, Walesiak 2015, s. 8–10; Zeliaś 2000, s. 34]).

Zmienna jest mierzona na skali ilorazowej (stosunkowej), jeśli zbiór jej wartości można jednoznacznie uporządkować na osi liczbowej z podaniem stałej jednostki pomiarowej i jednocześnie występuje w niej naturalny punkt zerowy. Innymi słowy zmienna jest na skali ilorazowej, jeśli stosunki między dwoma pomiarami na tej skali mają interpretację w świecie rzeczywistym. Skala stosunkowa nie nakłada żadnych ograniczeń dotyczących stosowania operacji matematycznych i metod statystycznych. Zatem skala ilorazowa dodatkowo (tj. w stosunku do skali przedziałowej) pozwala na wykonywanie operacji mnożenia i dzielenia. Przykładem cech mierzonych na tej skali są: temperatura w stopniach Kelvina, liczba zatrudnionych, inflacja, cena złota, kurs wymiany walut, wskaźnik cena/zysk

w stosunku do akcji[2], wiek (w latach), wzrost (w centymetrach), długość (w metrach), dochód kredytobiorcy lub zysk firmy.

Omówione skale pomiarowe przedstawiono w kolejności od najsłabszej do najmocniejszej. Warto odnotować, że mają one właściwości kumulatywne, czyli charakteryzują się narastającym stopniem dokładności pomiaru, a każda silniejsza skala zawiera wszystkie przymioty skali słabszej. Dwie pierwsze skale (tj. nominalną i przedziałową) określa się jako słabe (niemetryczne) i stosuje się je w przypadku cech jakościowych. Natomiast skale przedziałowa i ilorazowa określa się jako mocne (metryczne) i wykorzystuje w przypadku cech ilościowych. W zastosowaniach praktycznych określenie skali pomiaru zmiennych jest bardzo ważne, ponieważ determinuje możliwość zastosowania konkretnych metod analiz. Problem z wykorzystywaniem niektórych metod może się pojawić, gdy w analizowanym zbiorze danych występują zmienne mierzone na skalach różnych typów lub tylko na niemetrycznych. Zazwyczaj przyjmuje się, że metody, które stosuje się do wyników pomiaru w skali słabszej, można też zastosować w przypadku zmiennych mierzonych na skali mocniejszej, ale nie odwrotnie. Oznacza to, że teoretycznie nie jest możliwe wzmacnianie skal. Jednak, jak twierdzą Gatnar i Walesiak [2004, s. 22], w praktycznych badaniach ekonomiczno-społecznych na ogół sztucznie wzmacnia się skale pomiaru, traktując zmienne mierzone na skalach niemetrycznych jak zmienne przedziałowe i ilorazowe.

1.3. Metody pozyskiwania danych do badania, źródła danych

Jak wcześniej wspomniano, badanie może obejmować wszystkie obiekty będące jednostkami danej zbiorowości lub tylko ich część. Zazwyczaj to prowadzący badanie musi dokonać wyboru między badaniem pełnym i częściowym, chociaż zdarzają się przypadki, kiedy badanie całkowite nie jest wykonalne, np. obserwacje niektórych danych ekonomicznych możliwe są jedynie w wybranych punktach pomiarowych, np. zysk przedsiębiorstwa dostępny jest na koniec roku, notowanie kursów akcji podaje się na otwarciu i zamknięciu sesji giełdowej. Przykładem badania pełnego są spisy statystyczne, rejestracje statystyczne i sprawozdawczość statystyczna, np. lista obecności, zapis księgowy majątku przedsiębiorstwa, rejestracja operacji wykonanych na koncie bankowym.

2 Wskaźnik cena/zysk (oznaczany również jako C/Z lub P/E) jest powszechnie używaną przez inwestorów giełdowych miarą opłacalności zakupu akcji danej spółki, którą oblicza się, dzieląc cenę rynkowa jednej akcji przez zysk netto przypadający na jedna akcję.

Wybór konkretnej formy obserwacji jest ściśle związany z celem badania, liczebnością zbiorowości, możliwością dostępu do poszczególnych jednostek statystycznych itp. W praktyce wybór metody warunkuje się terminem (czasem), w jakim należy przeprowadzić badanie, oraz środkami finansowymi przeznaczonymi na ten cel. Badanie pełne, mimo niewątpliwych zalet, jest w wielu przypadkach niemożliwe[3], a czasem – zupełnie niepotrzebne. Wówczas przeprowadza się badania częściowe, co ma miejsce szczególnie w sytuacjach, kiedy:

- przeprowadzenie badania pełnego jest zbyt kosztowne lub wymagałoby zbyt długiego czasu (np. analiza wszystkich przedsiębiorstw działających w Polsce),
- badanie powoduje uszkodzenie lub zniszczenie badanych jednostek (np. kontrola jakości),
- nie jest możliwe przeprowadzenie badania pełnego z powodu braku możliwości przebadanie wszystkich jednostek statystycznych (np. nie jest możliwe zasięgnięcie opinii na zadany temat zarządzających instytucjami finansowymi lub zjawisko obserwowane jest w określonych punktach pomiarowych, np. na koniec każdego miesiąca, i nie ma możliwości pomiaru poziomu zjawiska między tymi punktami),
- celem badania jest uzyskanie wyników przybliżonych – orientacyjnych (np. określenie zainteresowania klientów nowym produktem bankowym, mającym wejść na rynek).

Spośród metod badania częściowego na uwagę zasługuje metoda reprezentacyjna, w której do badania wybiera się jedynie próbę statystyczną zawierającą pewną liczbę jednostek reprezentujących badaną zbiorowość. Metoda ta jest najbardziej prawidłową formą badania częściowego, ponieważ zastosowanie rachunku prawdopodobieństwa przy przenoszeniu wyników z losowej próby na całą zbiorowość umożliwia określenie wielkości popełnianego błędu, czego nie dają inne metody.

Warto zauważyć, że cecha specyficzna analiz zjawisk społecznych to brak możliwości powtórzenia badania w dokładnie takich samych warunkach, zatem wykorzystanie eksperymentu naukowego w naukach społecznych jest dość ograniczone, chociaż takie próby są podejmowane (tzw. ekonomia eksperymentalna). Obserwację procesów występujących w gospodarce prowadzi się jedynie dla udostępnionych (do obserwacji) obiektów badania, w czasie kiedy dokonanie pomiaru jest możliwe. Innymi słowy nie ma możliwości pozyskania informacji dotyczących wszystkich jednostek zbiorowości generalnej, np. z powodu poufności dane o sytuacji finansowej przedsiębiorstw są dostępne tylko w odniesieniu do spółek giełdowych (i to w określonym przez prawo zakresie analizy). Gdyby zatem przyjąć, że zbiorowość statystyczną stanowią wszystkie zarejestrowane przedsiębiorstwa, to obowiązek publicznego informowania o swoim

3 Chociażby dlatego, że pewne dane są niedostępne, gdyż można je zaobserwować tylko w wybranych momentach lub jednostkach pomiarowych lub mogłoby prowadzić do zniszczenia obiektu badania, np. w przypadku statystycznej kontroli jakości.

standingu[4] ma tylko ich niewielka część, a informacje o pozostałych będą udostępniane tylko niektórym instytucjom, np. urzędom skarbowym. Oprócz tego nawet w przypadku występowania procesów ciągłych dane na ten temat pojawiają się w określonych momentach. Przykładowo notowania giełdowe na GPW w Warszawie są realizowane w systemie ciągłym, ale publicznie udostępniane są co 15 minut, a w większości analiz wykorzystuje się ceny zamknięcia z danego dnia sesyjnego. W tym przypadku nie bada się wszystkich zarejestrowanych kursów instrumentu finansowego, a jedynie wybrany, czyli cenę zamknięcia, która reprezentuje cenę waloru z danego dnia. Podobnie jest w przypadku informacji o wielkości zatrudnienia, którą podaje się np. na pierwszy lub ostatni dzień miesiąca, zysku przedsiębiorstwa podawanym na koniec roku itd.

Warto pamiętać, że wykorzystywana w niepełnych badaniach statystycznych próba powinna być reprezentatywna, tzn. winna opisywać strukturę zbiorowości generalnej z przyjętą dokładnością. Reprezentatywność próby zależy od dwóch czynników: sposobu doboru próby oraz wielkości (liczebności) próby. Ogół metod doboru próby do badań można podzielić na dwie grupy – metody doboru losowego i metody doboru nielosowego. Jeżeli próba jest losowa, to wraz ze wzrostem jej liczebności wzrasta stopień jej reprezentatywności.

Informacje gromadzone w toku realizowanych badań mogą pochodzić z tzw. źródeł pierwotnych lub wtórnych. Pierwotne źródła informacji obejmują te wszystkie źródła, które zostały przygotowane specjalnie w celu zbadania wybranego problemu. Podstawowymi pierwotnymi źródłami informacji są studia empiryczne, takie jak obserwacja i badania wykorzystujące kwestionariusze. Innymi słowy, informacja pierwotna powstaje w momencie realizowanego badania, np. zostaje pozyskana przez ankietera prowadzącego badanie. Wtórne źródła informacji obejmują te wszystkie źródła, które nie powstały w trakcie realizacji konkretnych badań. Tego typu dane mogą być dostępne w formie tradycyjnych publikacji (papierowych), plików na nośnikach cyfrowych i publikacji internetowych o wolnym lub ograniczonym dostępie. Głównymi wtórnymi źródłami informacji są:

- publikacje organów państwowych, np. materiały Ministerstwa Finansów, Agencji Rynku Rolnego, Zakładu Ubezpieczeń Społecznych (ZUS), Wojewódzkich Urzędów Pracy etc.;
- publikacje placówek naukowo-badawczych, np. Polskiej Akademii Nauk (PAN), Uniwersytetu Łódzkiego, Szkoły Głównej Handlowej (SGH), Uniwersytetu Marii Curie-Skłodowskiej w Lublinie (UMCS) i innych;
- publikacje wyspecjalizowanych instytucji statystycznych, takich jak Główny Urząd Statystyczny (GUS) i jego oddziały wojewódzkie (WUS), instytucje statystyczne w innych krajach, Eurostat itp.;
- wyspecjalizowane (zazwyczaj płatne) bazy danych, takie jak Notoria Serwis, Thomson Reuters, Amadeus, EMIS (Emerging Markets Information Service) etc.;

4 Standing (lub standing finansowy) oznacza ocenę sytuacji ekonomiczno-finansowej.

- publikacje instytucji finansowych, takich jak Narodowy Bank Polski (NBP), Komisja Nadzoru Finansowego (KNF), Giełda Papierów Wartościowych w Warszawie (GPW), Towarzystwo Funduszy Inwestycyjnych (TFI) itp.;
- serwisy i portale internetowe, np. bankier.pl, stoog.pl i inne;
- publikacje organizacji pozarządowych i międzynarodowych, takich jak Fundacja im. Stefana Batorego, Centrum Informacji i Organizacji Badań Finansów Publicznych i Prawa Podatkowego Krajów Europy Środkowej i Wschodniej przy Wydziale Prawa Uniwersytetu w Białymstoku, Stowarzyszenie Amnesty International, Organizacja Narodów Zjednoczonych (ONZ), OECD (Organization for Economic Co-operation and Development);
- publikacje międzynarodowych instytucji finansowych, takich jak Bank Światowy (World Bank), Europejski Bank Odbudowy i Rozwoju (European Bank for Reconstruction and Development) itp.;
- sprawozdania finansowe spółek publicznych, wynikające z obowiązku informacyjnego lub inne materiały wewnętrzne przedsiębiorstw;
- biuletyny agencji badań opinii publicznej lub badań rynkowych (zazwyczaj płatne).

Na szczególną uwagę zasługują publikacje Głównego Urzędu Statystycznego i urzędów wojewódzkich, w których znaleźć można dane statystyczne dotyczące podstawowych charakterystyk społeczno-ekonomicznych w ujęciu rocznym, kwartalnym i miesięcznym, w odniesieniu do analiz w skali makro, regionalnej i branżowej, a także dotyczące sytuacji międzynarodowej. Dane te dostępne są również w Internecie, czasami w postaci łatwych do przetwarzania plików .xlsx.

Wewnętrzne materiały przedsiębiorstw zawierają podstawowe dane dotyczące m.in. przychodów i kosztów firmy, stanu zatrudnienia, majątku. Są one wykorzystywane w analizach bieżącego funkcjonowania przedsiębiorstwa oraz do planowania jego przyszłego rozwoju, bez których niemożliwe byłoby chociażby uzyskanie kredytu bankowego. Zazwyczaj dane te mają charakter poufny, co oznacza, że dostęp do nich posiadają jedynie pracownicy danej firmy oraz przedstawiciele organów kontrolnych. Szerszy dostęp do tego typu danych jest możliwy jedynie w przypadku spółek publicznych.

Publikacje organów państwowych lub terenowych, a także placówek naukowo-badawczych najczęściej zawierają opracowany już materiał statystyczny. Natomiast dostęp do danych źródłowych ma z reguły dość ograniczony krąg odbiorców. W przypadku wyspecjalizowanych agencji badania rynku lub opinii społecznej korzystanie z wyników ich badań wiąże się z pewnymi (czasem znacznymi) kosztami, a zlecenie badań jest zazwyczaj równie kosztowne. Informacja statystyczna stała się dość drogim „towarem", czego dowodem jest funkcjonowanie wielu instytucji (również prywatnych) zajmujących się gromadzeniem i opracowywaniem danych statystycznych.

Przystępując do zbierania materiału statystycznego, należy określić sposób opisu i pomiaru analizowanych zjawisk, a także rodzaj obserwacji. Innymi słowy trzeba sprecyzować, jakie dane statystyczne zostaną użyte w badaniach, zarówno

w odniesieniu do zakresu przestrzennego wyróżnionych cech, jak i w odniesieniu do częstotliwości pomiaru.

Warto odnotować, że od pewnego czasu w krajach Unii Europejskiej, w tym również w Polsce, prowadzona jest tzw. polityka otwartości nauki, która polega na tym, że zarówno wyniki badań, jak i dane pozyskane w trakcie realizowanych badań mają być publicznie dostępne i archiwizowane w formie cyfrowej w odpowiednich repozytoriach. W szczególności polityka otwartego dostępu (*Open Access*) dotyczy udostępniania w Internecie publikacji naukowych oraz wyników wszelkiego rodzaju badań finansowanych ze środków publicznych. Celem tych działań ma być upowszechnienie dostępu do wiedzy, nauki i danych pozyskiwanych w trakcie badań naukowych ze względu na rosnące ich znaczenie dla rozwoju społeczno-gospodarczego. Niewątpliwie ważnym aspektem otwartego dostępu do danych jest możliwość przeprowadzenia weryfikacji poprawności uzyskanych wyników na podstawie udostępnionych przez innych badaczy i inne ośrodki badawcze danych.

1.4. Grupowanie i prezentacja zebranych danych

W wyniku obserwacji statystycznej otrzymuje się zbiór danych, które należy uporządkować i poddać analizie merytorycznej. Zadaniem tej ostatniej jest usunięcie ze zbioru obserwacji błędnie zarejestrowanych (tzw. błędy pomiaru). Uporządkowanie materiału statystycznego określa się mianem „grupowania". Polega ono na (mniej lub bardziej zróżnicowanym) podziale niejednorodnej zbiorowości na możliwie jednorodne grupy według obranych kryteriów, charakteryzujących poszczególne grupy, i odpowiednim zestawieniu danych statystycznych. Klasyfikację jednostek zbiorowości statystycznej przeprowadza się zazwyczaj według wybranych cech, których prawidłowa analiza jest możliwa dopiero w ramach otrzymanych jednorodnych grup. Podstawowymi czynnościami po dokonaniu segregacji materiału na grupy jest zliczanie danych w poszczególnych grupach oraz prezentacja opracowanego materiału w postaci szeregu statystycznego.

Szeregiem statystycznym nazywa się zbiór wyników obserwacji uporządkowanych według określonych cech (kryteriów). Szereg składa się zazwyczaj z dwóch kolumn, z których jedna podaje wariant lub klasę cechy, a druga informuje o liczbie jednostek przypadających na daną klasę (lub wariant) atrybutu. Najczęściej wyróżnia się dwa kryteria podziału szeregów:

- kryterium formalne, związane z budową szeregu, na podstawie którego można wyodrębnić: szeregi szczegółowe, szeregi rozdzielcze i szeregi kumulacyjne,
- kryterium merytoryczne, wynikające z typu badanej cechy zbiorowości, według którego wyróżnia się szeregi czasowe i szeregi przestrzenne.

Podziały te jednak nie wykluczają się wzajemnie, gdyż np. szereg rozdzielczy może być jednocześnie szeregiem czasowym lub przestrzennym.

Szereg szczegółowy to taki, w którym zawarto opis badanej zbiorowości, podając informację o wartości badanej cechy w każdym punkcie pomiarowym, tj. w każdym momencie lub dla każdej z obserwowanych jednostek wchodzących w skład zbiorowości statystycznej. Przydatność szeregu szczegółowego bierze się stąd, że daje on całkowity materiał statystyczny, od którego można rozpocząć pracę badawczą. Jest on jednak mało przejrzysty, dlatego stosuje się go tylko w tych przypadkach, gdy wymagana jest duża dokładność, a liczba obserwacji jest stosunkowo niewielka. Szeregi czasowe mają zazwyczaj postać szeregów szczegółowych, gdyż w przypadku analizy dynamiki zjawisk istotne są zmiany zachodzące w kolejnych punktach pomiarowych.

Szeregiem rozdzielczym nazywa się uporządkowany i pogrupowany (według przyjętych kryteriów) zbiór informacji dotyczących badanej cechy występującej w określonej zbiorowości lub próbie. Otrzymuje się go, dzieląc zbiorowość statystyczną na klasy zbiorcze według pewnej cechy i podając liczebności każdej z tych klas, zwane liczebnościami klasowymi n_i, $i = 1, 2, \ldots, k$. Szeregi rozdzielcze mogą dotyczyć zarówno cechy jakościowej, jak i ilościowej. Charakteryzują one strukturę danej zbiorowości, stąd nazywane są czasem szeregami strukturalnymi.

W praktyce wyróżnia się również szeregi rozdzielcze kumulacyjne, które powstają przez dodanie liczebności kolejnych przedziałów i obliczanie procentu liczebności kumulowanych w stosunku do liczebności całego zbioru. Szeregi kumulacyjne informują, dla ilu jednostek badanej zbiorowości zmienna przybiera wartości mniejsze od górnej granicy konkretnego przedziału (szereg kumulacyjny rosnący) lub dla ilu jednostek statystycznych zmienna przyjmuje wartości większe od dolnej granicy określonego przedziału (szereg kumulacyjny malejący – kumulacyjny rozkład malejący).

Grupując materiał badawczy, rozpatruje się wszystkie możliwe warianty badanych cech statystycznych x_i. W przypadku cech jakościowych i mierzalnych skokowych, poszczególne warianty badanych cech x_i można wymienić jako: $x_1, x_2, \ldots, x_k$. W przeciwieństwie do cech ciągłych, które mogą przyjmować nieprzeliczalnie wiele wartości $x_i \in <x_{\min}, x_{\max}>$, a więc nie można ich wszystkich wymienić. W takim przypadku tworzy się klasy, dzieląc obszar zmienności ($x_{\min} - x_{\max}$) cechy na tzw. przedziały klasowe, określone przez ich dolną x_{id} i górną x_{ig} granicę. Różnicę między wartością $x_{ig} - x_{id}$, nazywa się rozpiętością przedziału klasowego.

Przedziały klasowe zawierające po kilka wariantów badanej cechy można również utworzyć dla cech skokowych. Tworzenie takich przedziałów może być wynikiem grupowania statystycznego, kiedy z powodu dużej liczby wariantów cechy dokonuje się agregacji niektórych z nich w klasy. Może to także zostać narzucone jeszcze w trakcie zbierania danych statystycznych, kiedy to – zgodnie z celem badania – pytania dotyczą przynależności do z góry określonych przedziałów, a nie konkretnych wariantów cechy.

W przypadku analizowania cech jakościowych zazwyczaj wyróżnia się tyle wariantów, ile jest koniecznych do realizacji celu badania. Czasem jednak okazuje się, z reguły w trakcie opracowywania materiału statystycznego, że pewne warianty cechy występują bardzo rzadko i merytorycznie uzasadniona jest ich agregacja. Wówczas można utworzyć klasy zawierające po kilka wariantów badanej cechy, ponieważ oddzielne ich rozpatrywanie nie ma większego sensu.

Budując szeregi rozdzielcze, należy zdecydować o liczbie klas, ich rozpiętości i sposobie określania granic przedziałów. Należy pamiętać, że dobra klasyfikacja powinna spełniać dwa podstawowe warunki. Po pierwsze, musi być przeprowadzona w sposób rozłączny, co oznacza, że poszczególne jednostki o określonych cechach powinny być w sposób jednoznaczny przydzielone do poszczególnych klas (grup). Po drugie, musi być przeprowadzona w sposób zupełny, co znaczy, że klasy powinny objąć wszystkie cechy występujące w danej zbiorowości. W przeciwnym razie konieczne jest tworzenie klas zbiorczych, ujmujących te cechy, które mają istotne znaczenie z punktu widzenia celu badania. W praktyce wybór liczby klas zależy od liczby obserwacji i od charakteru danych. Należy ustalić takie przedziały klasowe, które obejmują wszystkie dane, oraz zadbać o to, aby każda jednostka mogła trafić tylko do jednej klasy. Również ważną rolę odgrywa liczebność w przedziale klasowym, gdyż zarówno mała liczba obserwacji podzielona na wiele klas, jak i duża liczebność zbioru podzielona na nieliczne klasy nie ujawni obrazu struktury zgodnego z rzeczywistością.

Długość przedziału należy dobierać w taki sposób, aby wartości cechy oscylowały wokół punktu środkowego klasy. Konstruowanie szeregu rozdzielczego polega przede wszystkim na właściwym doborze wielkości przedziału klasowego, przy ustalaniu którego należy wziąć pod uwagę pewne kryteria (nie zawsze jednolite), pozwalające w prawidłowy sposób ustalić strukturę badanej zbiorowości. Przy doborze przedziałów klasowych powinno się dążyć do tego, aby szereg rozdzielczy dawał możliwie szczegółowy i przejrzysty obraz struktury zbiorowości statystycznej z punktu widzenia celu badania.

Poza wielkością przedziałów duży wpływ na rozkład liczebności ma również przyjęcie odpowiedniej wartości dolnej granicy pierwszej klasy. Gdy przedziały klasowe oznaczone są w taki sposób, że górna granica przedziału poprzedzającego jest równa dolnej granicy przedziału następnego (przedziały otwarte), przyjmuje się zwykle, że wartości cechy odpowiadające górnej granicy przedziału zalicza się do przedziału następnego. Innymi słowy przyjmuje się, że przedziały są domknięte z lewej strony i otwarte z prawej (nie musi to dotyczyć przedziałów pierwszego i ostatniego).

Szeregi rozdzielcze można budować zarówno dla statystycznych cech skokowych, jak i ciągłych. W przypadku cech skokowych szereg rozdzielczy uwzględnia poszczególne warianty tych cech. Mówi się wtedy o szeregach z przedziałami klasowymi jednojednostkowymi. Natomiast w przypadku cech ciągłych szereg rozdzielczy uwzględnia przedziały klasowe, grupujące wszystkie warianty cechy

w granicach tych przedziałów. Mówi się wtedy o szeregach z przedziałami klasowymi wielojednostkowymi. Do tej grupy szeregów zalicza się także szeregi rozdzielcze z przedziałami klasowymi obejmującymi po kilka wariantów cech skokowych. W niektórych przypadkach wygodnie jest analizować cechy będące ze swej natury cechami ciągłymi jako cechy skokowe i wtedy, zamiast całego przedziału, podaje się zwykle środek przedziału $\dot{x}_i$ (gdzie $\dot{x}_i = x_{i-1g} + 0{,}5 \cdot (x_{ig} - x_{id})$) albo jego granicę górną x_{ig} (jeśli szereg uporządkowany jest rosnąco) lub dolną x_{id} (dla szeregu malejącego).

Przykład 1.1

W tabeli 1.1 przedstawiono rzeczywiste przychody 81 mikroprzedsiębiorstw działających w różnych branżach na terenie województwa łódzkiego w latach 2016–2019, według kolejności pozyskiwania danych. Dane w tabeli 1.1 tworzą szereg szczegółowy nieuporządkowany. Jak łatwo zauważyć analiza tego szeregu jest dość utrudniona z powodu znacznej liczebności analizowanych firm. Dla ułatwienia analizy należy pogrupować pozyskane dane.

Rozwiązanie

Analizując dane z tabeli 1.1 należy zauważyć, że w procesie anonimizacji nazwy własne firm zakodowano numerycznie. Nadany firmie numer, tak jak jej nazwa własna, jest cechą jakościową mierzoną na skali nominalnej. Natomiast przychód firmy jest cechą mierzalną na skali ilorazowej. Wygodnie jest więc uporządkować analizowany szereg, tworząc przedziały klasowe. W tym celu należy w pierwszej kolejności uporządkować rozpatrywane firmy według wartości przychodu rosnąco lub malejąco.

Tabela 1.1. Przychody mikroprzedsiębiorstw działających w województwie łódzkim

Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN
1	110 236,07	28	122 669,33	55	98 576,61
2	136 578,64	29	1 122 609,63	56	556 378,88
3	886 954,36	30	76 156,78	57	121 479,09
4	754 124,78	31	55 621,01	58	262 143,97
5	1 103 425,62	32	356 784,25	59	186 017,45
6	125 899,03	33	789 556,32	60	77 412,31
7	98 645,12	34	207 556,89	61	211 069,32
8	138 422,33	35	147 077,92	62	3 412 991,56
9	228 323,11	36	978 662,56	63	134 906,72
10	145 562,89	37	86 544,21	64	505 743,94

Tabela 1.1 (cd.)

Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN
11	505 217,15	38	123 156,96	65	202 778,19
12	101 366,25	39	402 677,74	66	774 622,87
13	689 745,65	40	378 691,02	67	92 556,41
14	79 652,32	41	632 014,98	68	51 099,13
15	126 005,63	42	446 731,46	69	464 336,09
16	174 445,01	43	247 756,41	70	134 507,66
17	89 996,47	44	88 411,02	71	252 878,94
18	223 075,67	45	122 265,74	72	708 991,37
19	67 895,75	46	389 176,33	73	4 662 127,49
20	324 551,03	47	65 572,23	74	261 203,65
21	98 477,32	48	845 229,65	75	253 489,71
22	178 542,39	49	177 483,12	76	813 742,33
23	2 562 766,86	50	266 911,20	77	422 110,99
24	89 744,06	51	115 966,36	78	689 116,44
25	209 113,91	52	3 887 412,69	79	302 889,62
26	131 489,55	53	156 733,98	80	416 732,14
27	174 798,64	54	2 143 229,70	81	775 069,01

Źródło: opracowanie własne na podstawie danych [Kokczyński 2019].

Szereg szczegółowy uporządkowany rosnąco według wartości przychodów uzyskanych przez przedsiębiorstwa zaprezentowano w tabeli 1.2.

Tabela 1.2. Przychody mikroprzedsiębiorstw uporządkowane rosnąco

Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN
68	51 099,13	2	136 578,64	39	402 677,74
31	55 621,01	8	138 422,33	80	416 732,14
47	65 572,23	10	145 562,89	77	422 110,99
19	67 895,75	35	147 077,92	42	446 731,46
30	76 156,78	53	156 733,98	69	464 336,09
60	77 412,31	16	174 445,01	11	505 217,15
14	79 652,32	27	174 798,64	64	505 743,94
37	86 544,21	49	177 483,12	56	556 378,88
44	88 411,02	22	178 542,39	41	632 014,98

Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN	Nr firmy	Przychody w PLN
24	89 744,06	59	186 017,45	78	689 116,44
17	89 996,47	65	202 778,19	13	689 745,65
67	92 556,41	34	207 556,89	72	708 991,37
21	98 477,32	25	209 113,91	4	754 124,78
55	98 576,61	61	211 069,32	66	774 622,87
7	98 645,12	18	223 075,67	81	775 069,01
12	101 366,25	9	228 323,11	33	789 556,32
1	110 236,07	43	247 756,41	76	813 742,33
51	115 966,36	71	252 878,94	48	845 229,65
57	121 479,09	75	253 489,71	3	886 954,36
45	122 265,74	74	261 203,65	36	978 662,56
28	122 669,33	58	262 143,97	5	1 103 425,62
38	123 156,96	50	266 911,20	29	1 122 609,63
6	125 899,03	79	302 889,62	54	2 143 229,70
15	126 005,63	20	324 551,03	23	2 562 766,86
26	131 489,55	32	356 784,25	62	3 412 991,56
70	134 507,66	40	378 691,02	52	3 887 412,69
63	134 906,72	46	389 176,33	73	4 662 127,49

Źródło: opracowanie własne na podstawie tab. 1.1.

Jak można łatwo zauważyć, uporządkowanie elementów szeregu ułatwia jego analizę, ale w dalszym ciągu nie jest możliwe uogólnienie wyników przeprowadzonego badania. Kolejną czynnością jest zatem pogrupowanie wszystkich obserwacji w klasy i utworzenie szeregów rozdzielczych, co pokazano w tab. 1.3. W prezentowanym przykładzie przychód firmy jest cechą mierzalną ciągłą, zatem istnieje konieczność utworzenia przedziałów klasowych. Nasuwa się oczywiście pytanie, jak takie przedziały skonstruować – sposobów jest kilka.

Tabela 1.3. Przychody mikroprzedsiębiorstw pogrupowane w przedziały klasowe

Przychody w tys. PLN	Liczebność		Przychody w tys. PLN	Liczebność		Przychody w tys. PLN	Liczebność	
	n_i	n_{isk}		n_i	n_{isk}		n_i	n_{isk}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
50–150	31	31	50–150	31	31	50–100	15	15
150–250	13	44	150–450	27	58	100–150	16	31

Tabela 1.3 (cd.)

Przychody w tys. PLN	Liczebność		Przychody w tys. PLN	Liczebność		Przychody w tys. PLN	Liczebność	
	n_i	n_{isk}		n_i	n_{isk}		n_i	n_{isk}
250–350	7	51	450–950	15	73	150–250	13	44
350–450	7	58	950 i więcej	8	81	250–450	14	58
450–550	3	61	Suma	81	×	450–950	15	73
550–650	2	63				950 i więcej	8	81
650–750	3	66				Suma	81	×
750–850	6	72						
850–950	1	73						
950–1050	3	76						
1050 i więcej	5	81						
Suma	81	×						

Źródło: opracowanie własne na podstawie tab. 1.2.

5

Bodaj najbardziej popularny sposób polega na tworzeniu przedziałów o jednakowej rozpiętości, czyli tak, jak to przedstawiono w kolumnach (1)–(3) tab. 1.3. Utworzono jedenaście przedziałów, z których 10 to przedziały otwarte z prawej strony o rozpiętości 100 tys. PLN, a ostatni, jedenasty jest przedziałem bez określonej górnej granicy (identyczny efekt byłby, gdyby ta granica została ustalona na poziomie wartości maksymalnej, czyli 4663 tys. PLN). Najwięcej przedsiębiorstw charakteryzowało się niskimi przychodami mieszczącymi się w pierwszych dwóch przedziałach dochodowych. Zaobserwować można również stopniowo zmniejszającą się liczbę firm wraz z przechodzeniem do klas o wyższych przychodach (por. dane w kolumnie drugiej, gdzie podano liczbę firm n_i).

Można oczywiście przeprowadzić agregację mniej licznych przedziałów klasowych, jak ma to miejsce w kolumnach (4)–(6), co spowoduje stopniowe zwiększanie się ich liczebności, ale wciąż im większe przychody, tym mniejsza liczebność przedziału. Dlatego w kolumnach (7)–(9) przedstawiono podział na przedziały klasowe, które wprawdzie nie są jednakowej rozpiętości, ale zapewniają podobne liczebności obserwacji w każdej klasie[6].

Warto odnotować, że w kolumnach (3), (6) i (9) tabeli 1.3 przedstawiono liczebności skumulowane (co oznaczono symbolem n_{isk}), które informują, ile obiektów

5 W całej publikacji strzałką oznaczono koniec danego przykładu.

6 Takie postępowanie jest szczególnie zalecane, jeśli będą obliczane mierniki struktury na podstawie danych zawartych w szeregach rozdzielczych z przedziałami klasowymi, chociaż w przypadku wzoru interpolacyjnego na dominantę wymagana jest jednakowa rozpiętość przedziałów sąsiadujących z przedziałem dominanty.

badania (firm) charakteryzowało się wartością cechy nie większą niż górna granica przedziału klasowego. Zatem informacje dotyczące pierwszego przedziału, zawarte w kolumnie (3), oznaczają, że 31 firm miało przychody do 150 tys. PLN, a w kolumnie (9), że 15 firm miało przychody do 100 tys. PLN.

Jednostki badania można pogrupować z punktu widzenia różnych charakterystyk, zarówno ilościowych (jak w przykładzie 1.1), jak i jakościowych. Przy czym decyzja o tym, jakiego rodzaju jest analizowany mierzalny atrybut, należy do prowadzącego badanie. Przykładowo mikroprzedsiębiorstwa można pogrupować w klasy wyznaczone na podstawie wielkości przychodów. W tym przypadku prowadzący badanie określa przedziały, które świadczą np. o małych, średnich i wysokich przychodach, a sama zmienna ilościowa traktowana jest jak zmienna porządkowa.

Innym przykładem zmiany charakteru analizowanej zmiennej ilościowej może być wielkość miejscowości, w jakiej działają analizowane firmy. Wówczas miejscowości można będzie uporządkować według przyjętego kryterium klasyfikacji, np. podział na wsie i miasta o z góry określonej wielkości, tj. o założonej liczbie mieszkańców, nadając odpowiednie rangi (tj. wartości na skali porządkowej) miejscowościom o określonej wielkości. Istnieje również możliwość potraktowania tej charakterystyki jako cechy mierzonej na skali nominalnej, wówczas przynależność do każdej klasy miejscowości traktowana jest jednakowo, a zapis formalny sprowadza się do utworzenia zmiennych zero-jedynkowych reprezentujących poszczególne warianty tej cechy. Oznacza to, że dla n obiektów badania tworzy się zmienne, których wartość wynosi 1, o ile dla i-tego obiektu dany wariant cechy występuje. Przy czym w praktyce zazwyczaj tworzy się $(k - 1)$ zmiennych (gdzie: k to liczba wariantów), przyjmując wariant cechy niereprezentowany przez żadną ze zmiennych binarnych za tzw. wariant referencyjny. Innymi słowy wariant referencyjny to taki, w którym żaden z $(k - 1)$ wyróżnionych wcześniej wariantów cechy nie występuje.

Przykład 1.2

Biorąc pod uwagę miejsce prowadzenia działalności badanych mikroprzedsiębiorstw wszystkie obserwacje pogrupowano jak w tab. 1.4. Jakkolwiek wielkość miejscowości, mierzona liczbą mieszkańców (w tys.), jest cechą ilościową, to można ją przekształcić w cechę jakościową mierzoną na skali porządkowej lub nominalnej.

Rozwiązanie

Tab. 1.4 zawiera szeregi rozdzielcze opisujące strukturę badanych przedsiębiorstw z punktu widzenia wielkości miejscowości. W kolumnach, oznaczonych jako: miejscowość w tys. mieszkańców, cecha ta traktowana jest jako cecha ilościowa. Natomiast oznaczenie tej cechy jako: klasa miejscowości oznacza, że jest ona traktowana jako cecha jakościowa, a ściślej jest to cecha porządkowa.

Tabela 1.4. Grupowanie mikroprzedsiębiorstw według wielkości miejscowości

Lp.	Miejscowość w tys. mieszkańców	Liczebność		Miejscowość w tys. mieszkańców	Liczebność		Miejscowość w tys. mieszkańców	Liczebność	
		n_i	n_{isk}		n_i	n_{isk}		n_i	n_{isk}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	66	13	13	3	6	6	3–20	19	19
2	8	2	15	8	2	8	20–50	13	32
3	57	11	26	14	5	13	50–100	24	56
4	22	7	33	18	6	19	100 i więcej	25	81
5	45	6	39	22	7	26	Klasa miejscowości	n_i	n_{isk}
6	688	25	64	45	6	32	Miejscowość mała	19	19
7	18	6	70	57	11	43	Miejscowość średnia	13	32
8	14	5	75	66	13	56	Miejscowość duża	24	56
9	3	6	81	688	25	81	Miasto powyżej 100 tys.	25	81
	Suma	81	×	Suma	81	×	Suma	81	×

Źródło: opracowanie własne na podstawie danych [Kokczyński 2019].

W kolumnach (2)–(4) przedstawiono szereg rozdzielczy (bez przedziałów klasowych), uporządkowany według kolejności pozyskiwania informacji z różnych miejscowości. W kolejnych trzech kolumnach szereg rozdzielczy został uporządkowany według wielkości miejscowości. Natomiast w kolumnach (8)–(10) w wierszach (1)–(4) pokazano szereg rozdzielczy z przedziałami klasowymi wyznaczonymi według kryterium merytorycznego – arbitralnego podziału na miejscowości małe, średnie, duże i miasta powyżej 100 tys. Ostatni z przedziałów jest podobnie jak pozostałe trzy otwarty z prawej strony, ale można go domknąć liczbą 688 (będącą wartością maksymalną w tym badaniu). Z kolei w wierszach (6)–(9) analizowana cecha: klasy miejscowości stała się cechą porządkową.

Gdyby potraktować analizowaną cechę jako nominalną, należałoby zastosować do jej kodowania zmienne zero-jedynkowe, które w przypadku każdej z 81 obserwacji przyjmowałyby liczby 0, jeśli dany wariant cechy nie występuje, lub 1 – w przeciwnym przypadku. W tab. 1.5 pokazano kodowanie binarne pierwszych 10 obserwacji z tab. 1.1. Pierwsze dwie kolumny tab. 1.5 to fragment tab. 1.1, w kolumnie (3) podano, w jakiej wielkościowo miejscowości znajduje się siedziba firmy, co zostało zakodowane w kolumnach (4)–(6) w postaci trzech zmiennych binarnych Z_1, Z_2, Z_3, odzwierciedlających trzy warianty wielkości miejscowości. Jak można łatwo zauważyć, sumy wszystkich trzech zmiennych dla każdej obserwacji (po wierszach) wynoszą 1, z wyjątkiem obserwacji czwartej, dla której ta suma wynosi 0. Oznacza to, że ostatni wariant (tj. miasta powyżej 100 tys.) jest charakteryzowany samymi 0 wszystkich zmiennych, co oznacza tzw. wariant referencyjny, którym jest w tym przypadku miasto powyżej 100 tys. mieszkańców.

Tabela 1.5. Grupowanie mikroprzedsiębiorstw według klas miejscowości (jako cechy nominalnej)

Nr firmy	Przychody w tys. PLN	Klasa miejscowości	Zmienne binarne		
			Z_1 mała	Z_2 średnia	Z_3 duża
(1)	(2)	(3)	(4)	(5)	(6)
1	110 236,07	Miejscowość mała	0	0	1
2	136 578,64	Miejscowość średnia	0	1	0
3	886 954,36	Miejscowość duża	0	0	1
4	754 124,78	Miasto powyżej 100 tys.	0	0	0
5	1 103 425,62	Miejscowość duża	0	0	1
6	125 899,03	Miejscowość duża	0	0	1
7	98 645,12	Miejscowość mała	1	0	0
8	138 422,33	Miejscowość średnia	0	1	0
9	228 323,11	Miejscowość średnia	0	1	0
10	145 562,89	Miejscowość średnia	0	1	0

Źródło: opracowanie własne na podstawie danych z tab. 1.4.

Przykład 1.3

W tab. 1.6 przedstawiono roczne procentowe stopy zwrotu z akcji 26 spółek notowanych na Giełdzie Papierów Wartościowych w Warszawie w latach 2012–2017. Jest to przykład danych panelowych, przekrojowo-czasowych. Kolejne kolumny zawierają informacje dotyczące rocznych stóp zwrotu odnotowane dla wyróżnionych w badaniu spółek, a dane w wierszach tworzą szeregi czasowe dla każdej spółki oddzielnie.

Tabela 1.6. Roczne procentowe stopy zwrotu z akcji spółek notowanych na GPW w Warszawie

Symbol spółki	Roczna stopa zwrotu (%) w kolejnych latach					
	2012	2013	2014	2015	2016	2017
BUDIMEX	0,06	0,69	0,16	0,35	0,06	0,13
CIECH	0,24	0,34	0,35	0,70	–0,34	–0,01
ENEA	–0,12	–0,12	0,15	–0,26	–0,17	0,21
PGE	–0,01	–0,07	0,17	–0,28	–0,21	0,13
TAURONPE	–0,05	–0,04	0,18	–0,52	–0,01	0,07

Tabela 1.6 (cd.)

Symbol spółki	Roczna stopa zwrotu (%) w kolejnych latach					
	2012	2013	2014	2015	2016	2017
BOGDANKA	0,29	–0,03	–0,22	–1,01	0,74	–0,02
KGHM	0,69	–0,39	–0,04	–0,51	0,40	0,19
ASSECOPOL	–0,02	0,07	0,16	0,16	0,00	-0,15
INTERCARS	0,06	0,79	0,17	0,06	0,16	0,11
ECHO	0,40	0,28	0,05	–0,02	0,74	–0,05
GTC	0,15	–0,28	–0,32	0,30	0,14	0,21
CCC	0,47	0,48	0,15	0,05	0,40	0,35
LPP	0,84	0,69	–0,21	–0,26	0,03	0,46
LOTOS	0,54	–0,15	–0,22	0,06	0,35	0,43
PKNORLEN	0,33	–0,16	0,21	0,35	0,26	0,24
PGNIG	0,24	0,01	–0,11	0,17	0,12	0,14
KERNEL	–0,03	–0,56	–0,29	0,55	0,30	–0,28
CYFRPLSAT	0,20	0,19	0,18	–0,12	0,16	0,02
NETIA	–0,20	0,21	0,14	0,06	–0,07	0,24
ORANGEPL	–0,25	–0,15	–0,11	–0,18	–0,13	0,05
CDPROJECT	0,16	1,04	–0,05	0,28	0,86	0,63
AMREST	0,41	–0,07	0,11	0,63	0,45	0,33
BORYSZEW	–0,05	–0,22	0,14	–0,18	0,53	0,14
EUROCASH	0,40	0,10	–0,21	0,27	–0,19	–0,37
KETY	0,38	0,45	0,31	0,14	0,27	0,13
KRUK	0,02	0,63	0,28	0,47	0,32	0,10
ORIBS	0,04	0,10	0,12	0,37	0,20	0,26

Źródło: opracowanie własne na podstawie [Kuźnik 2019].

Tabela 1.7. Roczne procentowe stopy zwrotu z akcji spółek notowanych na GPW w Warszawie

Symbol spółki	Roczna stopa zwrotu (%) w kolejnych latach					
	2012	2013	2014	2015	2016	2017
ASSECOPOL	–0,02	0,07	0,16	0,16	0,00	–0,15
BOGDANKA	0,29	–0,03	–0,22	–1,01	0,74	–0,02
BORYSZEW	–0,05	–0,22	0,14	–0,18	0,53	0,14
BUDIMEX	0,06	0,69	0,16	0,35	0,06	0,13
CCC	0,47	0,48	0,15	0,05	0,40	0,35
CDPROJECT	0,16	1,04	–0,05	0,28	0,86	0,63

Symbol spółki	Roczna stopa zwrotu (%) w kolejnych latach					
	2012	2013	2014	2015	2016	2017
CIECH	0,24	0,34	0,35	0,70	–0,34	–0,01
CYFRPLSAT	0,20	0,19	0,18	–0,12	0,16	0,02
ECHO	0,40	0,28	0,05	–0,02	0,74	–0,05
ENEA	–0,12	–0,12	0,15	–0,26	–0,17	0,21
EUROCASH	0,40	0,10	–0,21	0,27	–0,19	–0,37
GTC	0,15	–0,28	–0,32	0,30	0,14	0,21
INTERCARS	0,06	0,79	0,17	0,06	0,16	0,11
KERNEL	–0,03	–0,56	–0,29	0,55	0,30	–0,28
KETY	0,38	0,45	0,31	0,14	0,27	0,13
KGHM	0,69	–0,39	–0,04	–0,51	0,40	0,19
KRUK	0,02	0,63	0,28	0,47	0,32	0,10
LOTOS	0,54	–0,15	–0,22	0,06	0,35	0,43
LPP	0,84	0,69	–0,21	–0,26	0,03	0,46
NETIA	–0,20	0,21	0,14	0,06	–0,07	0,24
ORANGEPL	–0,25	–0,15	–0,11	–0,18	–0,13	0,05
ORIBS	0,04	0,10	0,12	0,37	0,20	0,26
PGE	–0,01	–0,07	0,17	–0,28	–0,21	0,13
PGNIG	0,24	0,01	–0,11	0,17	0,12	0,14
PKNORLEN	0,33	–0,16	0,21	0,35	0,26	0,24
TAURONPE	–0,05	–0,04	0,18	–0,52	–0,01	0,07

Źródło: opracowanie własne na podstawie tab. 1.6.

W tab. 1.7 zostały przedstawione dane z tab. 1.6, ale uporządkowane alfabetycznie według nazw spółek. Można je również uporządkować według reprezentowanych branż, wielkości kapitalizacji, roku założenia spółki lub roku pierwszego notowania na GPW w Warszawie, ewentualnie według wielkości osiągniętych stóp zwrotu w jednym lub kolejnych latach. Natomiast nie można zmieniać chronologii obserwacji.

Podsumowując rozważania dotyczące grupowania danych, należy podkreślić, że szeregi statystyczne mogą przedstawiać badaną zbiorowość w układzie statycznym, charakteryzując jej stan w ściśle określonym momencie (np. w określonym dniu, miesiącu, roku) – są to tzw. szeregi przekrojowe. Natomiast dynamikę rozwoju zbiorowości lub zjawisk w pewnym okresie opisują tzw. szeregi

czasowe. Powstają one, gdy podstawą grupowania jest czas, a celem badania analiza zmian pewnego zjawiska w czasie. Z kolei szeregi przestrzenne (geograficzne) opisują rozmieszczenie pewnych zjawisk w przestrzeni (części świata, państwa, regiony gospodarcze, jednostki administracyjne). Budując szereg przestrzenny, można (podobnie jak w przypadku szeregów rozdzielczych) dzielić zbiorowość na większą lub mniejszą liczbę grup (klas) w zależności od przyjętej jednostki terytorialnej.

Należy pamiętać, że w przypadku analiz prowadzonych na podstawie szeregów czasowych obowiązuje porządek chronologiczny. Podczas gdy szereg danych przekrojowych może zostać uporządkowany według dowolnego kryterium, np. według:

- wartości rosnących lub malejących;
- nazw obiektów uporządkowanych według alfabetu (np. spółki w kolejności alfabetycznej) lub innego kryterium je charakteryzującego (np. liczby zatrudnionych);
- kolejności dokonywania obserwacji.

Warto odnotować, że oprócz tabelarycznej formy prezentacji zebranych danych (w postaci szeregów statystycznych) często stosuje się prezentację graficzną w postaci wykresów, diagramów itp. Przykłady graficznej prezentacji danych pokazano na rys. 1.1–1.5.

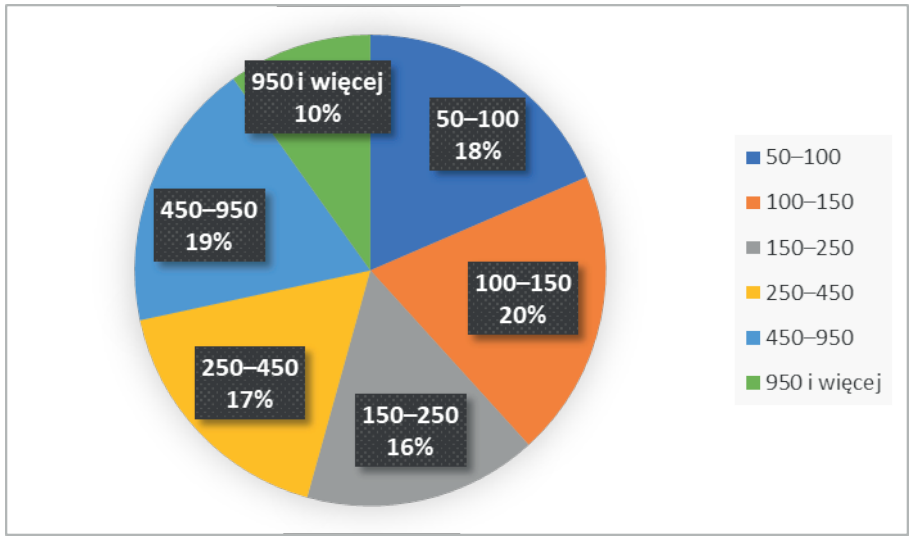
Przykład 1.4

Biorąc pod uwagę strukturę zbioru mikroprzedsiębiorstw rozpatrywanego w przykładzie 1.2 (z punktu widzenia ich przychodów), do jego analizy wygodnie jest wykorzystać tzw. wykres kołowy (diagram).

Rozwiązanie

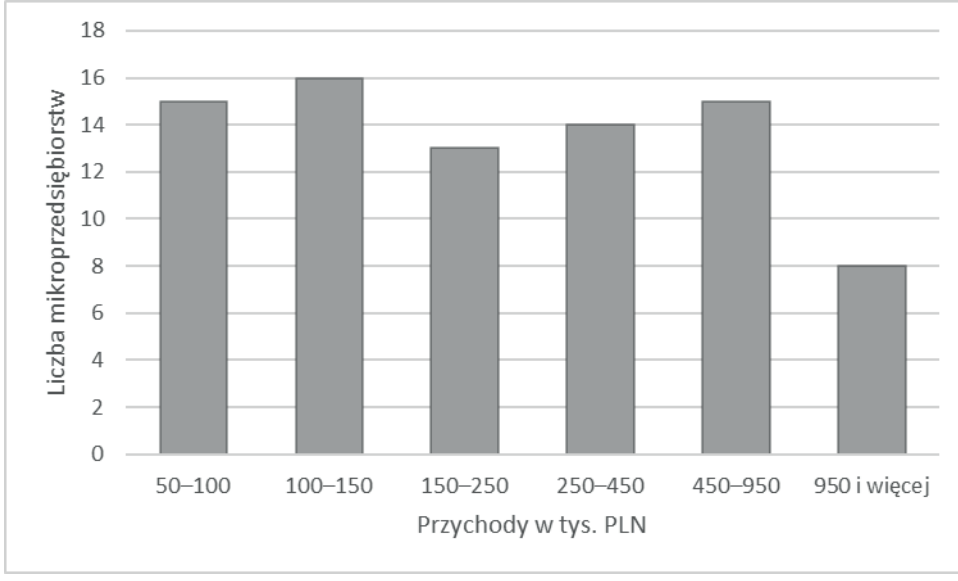
Dane zapisane w kolumnach (7) i (8) w tab. 1.3 przedstawiono na rys. 1.1. Jak można zauważyć, na wykresie poszczególne części koła informują o tym, jaka część wszystkich badanych mikroprzedsiębiorstw ma przychody mieszczące się w przyjętych przedziałach. Na diagramie podano liczebności względne (procentowe) w miejsce liczebności bezwzględnych (z kolumny (8) w tab. 1.3).

Zastosowanie liczebności względnych ułatwia analizę struktury, niezależnie bowiem od liczebności badanej zbiorowości (tzn. nieważne, czy liczba badanych przedsiębiorstw wynosi 15, 90, 150 czy 7000) zawsze suma części składowych jest identyczna i wynosi 100% (lub 1, jeśli korzysta się z liczebności względnych niemianowanych). Warto przy tym zauważyć, że zastąpienie udziałów procentowych liczebnościami bezwzględnymi nie zmieni kształtu wykresu. Innym sposobem prezentacji danych jest zapisanie ich w postaci wykresu słupkowego, co pokazano na rys. 1.2.



Rysunek 1.1. Struktura mikroprzedsiębiorstw z punktu widzenia wielkości przychodów

Źródło: opracowanie własne na podstawie kolumn (7) i (8) w tab. 1.3.



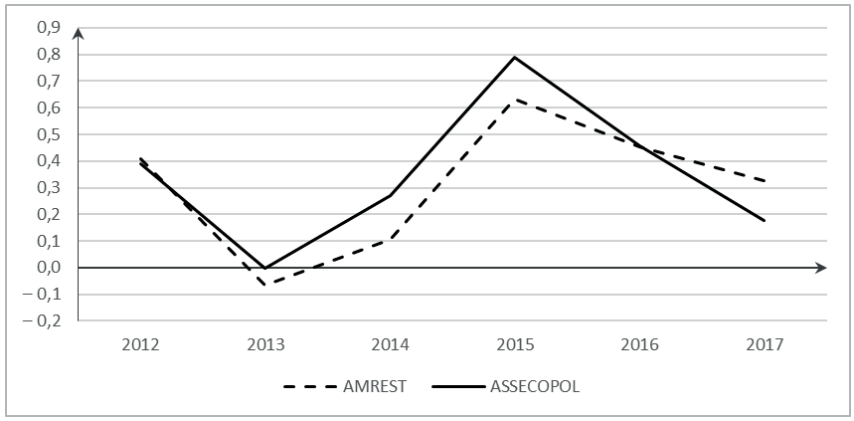
Rysunek 1.2. Liczebności przedsiębiorstw według wartości przychodów

Źródło: opracowanie własne na podstawie kolumn (7) i (8) w tab. 1.3.

Przykład 1.5

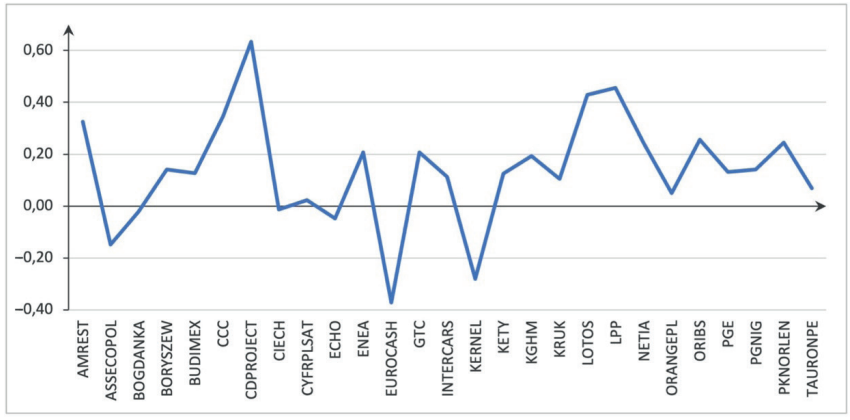
Dane przedstawione w tab. 1.7 można zaprezentować na wykresie liniowym dla każdej omawianej spółki – powstaną w ten sposób wykresy szeregów czasowych stóp zwrotu dla poszczególnych spółek, co pokazano na rys. 1.3 dla dwóch pierwszych (uporządkowanych alfabetycznie) spółek.

Z kolei rys. 1.4 przedstawia szereg danych przekrojowych, są to bowiem stopy zwrotu z akcji wszystkich przedstawionych w tab. 1.7 spółek, obliczone dla 2017 roku. Natomiast szereg przekrojowo-czasowy, zawierający wszystkie dane z tab. 1.7, został przedstawiony na rys. 1.5, na którym różnokolorowe krzywe reprezentują kolejne lata badania.



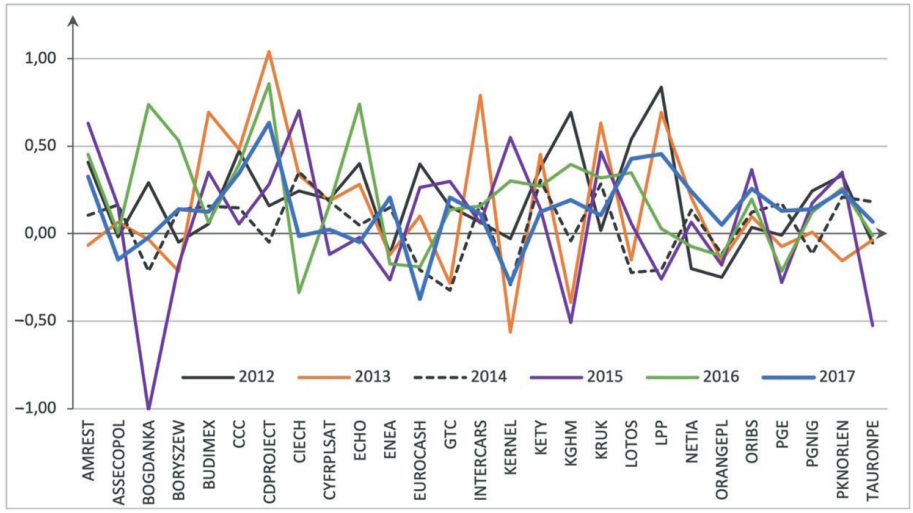
Rysunek 1.3. Stopy zwrotu spółek AMREST i ASSECOPOL w latach 2012–2017

Źródło: opracowanie własne na podstawie tab. 1.7.



Rysunek 1.4. Stopy zwrotu analizowanych spółek w 2017 roku

Źródło: opracowanie własne na podstawie tab. 1.7.



Rysunek 1.5. Stopy zwrotu analizowanych spółek w latach 2012–2017.

Źródło: opracowanie własne na podstawie tab. 1.7.

Podsumowując, należy zauważyć, że istnieją dwie popularne metody prezentowania szeregów danych, tj. w formie tabel lub wykresów. Pierwsza z nich pozwala dokładnie opisać badane zjawisko, co stanowi podstawę do dalszych analiz – chociażby badanie struktury lub dynamiki analizowanego zjawiska. Nie pozwala natomiast na pewną syntetyzację obserwacji, którą wygodnie jest zrobić, korzystając z różnego rodzaju wykresów. Przykładem takiej syntetyzacji jest diagram z rys. 1.1, z którego można łatwo odczytać strukturę firm z punktu widzenia wysokości uzyskiwanego przychodu, lub wykresy szeregów stóp zwrotu z rys. 1.5, wskazujące na to, że najwyższe stopy zwrotu uzyskiwały spółki w 2017 roku, spośród których liderem była spółka: CDPROJEKT.

1.5. Elementy rachunku prawdopodobieństwa

Zmienna losowa jest pojęciem podobnym do cechy statystycznej i do jej oznaczenia używa się identycznego symbolu. W przypadku zmiennej losowej X mówi się o jej realizacjach, które oznacza się przez x_i dla wszystkich $i = 1, 2, \ldots, k$ (skończony zbiór realizacji) lub dla $i = 1, 2, \ldots$ (nieskończony, ale przeliczalny zbiór

realizacji). Zmienna losowa jest pojęciem podstawowym podczas wnioskowania statystycznego.

Warto przy tym zauważyć, że o ile wyróżnione w badaniu warianty cechy są zawsze zbiorem skończonym i występują z określoną częstotliwością empiryczną (co określa się mianem rozkładu empirycznego), to realizacje zmiennej losowej mogą być skończone lub nie i występują z określonym prawdopodobieństwem, tworząc rozkład prawdopodobieństwa zmiennej losowej. W omawianych wcześniej przykładach 1.1–1.5 pokazano rozkłady empiryczne rzeczywistych danych, które prezentowano w tabelach i na wykresach. Zawarte tam liczebności (tab. 1.3–1.4) lub udziały procentowe (rys. 1.1) informują o liczebnościach lub strukturze danych pochodzących z obserwacji.

Przykładami zmiennych losowych mogą być:

- ocena możliwości zbytu produkcji pewnego przedsiębiorstwa ubiegającego się o kredyt, które wyrażane jest w punktach od 1 do 10;
- cena akcji zmieniająca się w kolejnych transakcjach znajdująca się w przedziale (a, b) – są to tzw. dane tikowe (*tick-by-tick*);
- kurs walutowy, np. EUR, notowany przez NBP w danym dniu, wyrażony jako cena otwarcia w PLN.

W pierwszym przykładzie przedstawiono zmienną losową skokową, a w drugim ciągłą. W trzecim przykładzie kurs walutowy można traktować jako zmienną typu skokowego, jeżeli przyjmie się, że wyznaczany jest on z dokładnością do jednego grosza, ponieważ wartości tej zmiennej zmieniają się co pewien interwał. W takim przypadku cena 100 EUR będzie liczbą naturalną. Jeśli natomiast przyjąć kurs walutowy jako liczbę o nieskończonej liczbie cyfr po przecinku, to będzie on zmienną losową ciągłą. W praktyce z uwagi na dużą zmienność kursów walutowych najczęściej traktuje się je jako zmienne ciągłe i wyznacza się prawdopodobieństwa ich realizacji dla pewnych przedziałów.

Rozkład prawdopodobieństwa (zmiennej losowej) przypisuje prawdopodobieństwo, z jakim realizują się pewne warianty (wartości) tej zmiennej i może być przedstawiony:

- dla zmiennej losowej skokowej jako rozkład punktowy przez podanie funkcji prawdopodobieństwa lub dystrybuanty,
- dla zmiennej losowej ciągłej jako rozkład przedziałowy opisany przy pomocy funkcji gęstości prawdopodobieństwa lub dystrybuanty.

Dystrybuanta jest niemalejącą funkcją, której wartość w punkcie x_i równa się prawdopodobieństwu tego, że zmienna losowa X przyjmie wartości mniejsze od x_i. Innymi słowy, dystrybuanta (obliczona w punkcie x_i) oznacza wartość skumulowanego prawdopodobieństwa zmiennej losowej X dla jej realizacji z przedziału $(-\infty; x_i)$. Dla $x \in R$ dystrybuanta wyraża się wzorem:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) \tag{1.1}$$

Zatem w przypadku zmiennej losowej ciągłej dystrybuanta jest postaci:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{1.2}$$

a dla zmiennej losowej skokowej ma wzór:

$$F(x) = \begin{cases} 0 & \text{dla } x < x_{1,} \\ p_1 & \text{dla } x_1 \leq x < x_{2,} \\ p_1 + p_2 & \text{dla } x_2 \leq x < x_{3,} \\ \quad \ldots & \end{cases} \tag{1.2a}$$

Warto odnotować, że w przypadku zmiennej losowej skokowej można wyznaczyć prawdopodobieństwo pojedynczej jej realizacji, natomiast dla zmiennej losowej ciągłej prawdopodobieństwo mierzone w punkcie jest równe 0. Dlatego w tym drugim przypadku wyznacza się prawdopodobieństwa dla założonego przedziału liczbowego realizacji zmiennej losowej, a nie dla jej pojedynczej realizacji.

Wartość oczekiwana, inaczej zwana nadzieją matematyczną, jest oznaczana symbolem $E(X)$. Dla zmiennej losowej skokowej jest definiowana jako suma iloczynu realizacji zmiennej losowej (x_i) i prawdopodobieństw ich wystąpienia (p_i) i wyrażana jest wzorem w przypadku:

- zmiennych osiągających skończoną liczbę wartości jako:

$$E(X) = \sum_{i=1}^{n} x_i p_i \tag{1.3}$$

- dla zmiennych osiągających przeliczalną liczbę wartości jako:

$$E(X) = \sum_{i=1}^{\infty} x_i p_i \tag{1.4}$$

Natomiast wartość oczekiwana zmiennej losowej ciągłej definiowana jest jako całka oznaczona iloczynu realizacji zmiennej losowej (x_i) i funkcji gęstości prawdopodobieństwa $f(x)$:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{1.5}$$

Wartość oczekiwana jest miarą położenia i oznacza przeciętną wartość realizacji zmiennej losowej.

Wariancja zmiennej losowej określa jej zmienność wokół wartości oczekiwanej i jest odpowiednio definiowana dla zmiennych losowych skokowych i ciągłych. Wariancja zmiennych losowych typu skokowego jest postaci:

$$D^2(X) = E(X - E(X))^2 = \sum_i (x_i - E(X))^2 p_i \tag{1.6}$$

a zmiennych losowych typu ciągłego postaci:

$$D^2(X) = E(X - E(X))^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx \tag{1.7}$$

Istotną rolę we wszelkiego rodzaju analizach odgrywa odchylenie standardowe, będące pierwiastkiem kwadratowym z wariancji[7].

Przykład 1.6

Klient agencji nieruchomości sprawdza oprocentowanie kredytów hipotecznych na podstawie informacji pochodzącej z pięciu banków komercyjnych, w których odpowiednie stopy procentowe wynoszą: 7,65, 6,97, 7,25, 8,10 i 7,30%. Należy przeprowadzić analizę oprocentowania tych kredytów, dla wyznaczenia funkcji prawdopodobieństwa, dystrybuanty oraz podstawowych parametrów rozkładu tej zmiennej losowej.

Rozwiązanie

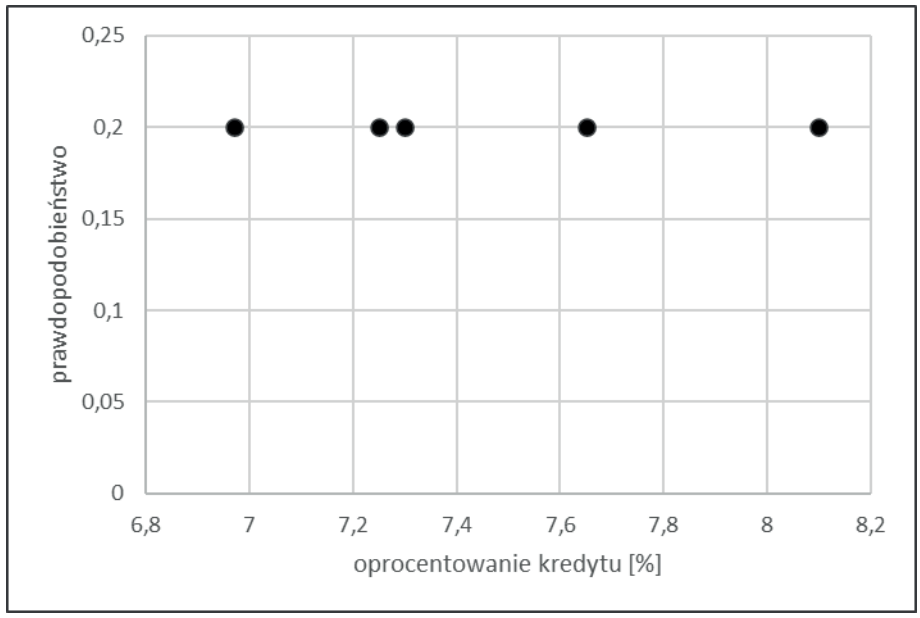
W tym celu należy wyznaczyć wartość oczekiwaną, wariancję i odchylenie standardowe analizowanej skokowej zmiennej losowej. Oprocentowanie kredytów X jest zmienną losową skokową, a ponieważ analizie podlega 5 banków to prawdopodobieństwo realizacji każdej z wymienionych stóp procentowych wynosi p = 1/5 = 0,2. Funkcja prawdopodobieństwa może zostać przedstawiona w postaci tabelarycznej (tab. 1.8) lub na wykresie (rys. 1.6).

Tabela 1.8. Funkcja prawdopodobieństwa

Oprocentowanie x_i [%]	6,97	7,25	7,30	7,65	8,10
Prawdopodobieństwo p_i	0,20	0,20	0,20	0,20	0,20

Źródło: opracowanie własne.

7 Więcej informacji na temat rozkładów zmiennych losowych i własności ich parametrów znaleźć można w wielu publikacjach m.in. [Aczel 2000; Witkowska 2004].



Rysunek 1.6. Wykres funkcji prawdopodobieństwa stóp kredytowych

Źródło: opracowanie własne.

Dystrybuanta tego rozkładu jest postaci:

$$F(x)=\begin{cases}0{,}0 & \text{dla} \quad x<6{,}97\\ 0{,}2 & \text{dla} \quad 6{,}97\le x<7{,}25\\ 0{,}4 & \text{dla} \quad 7{,}25\le x<7{,}30\\ 0{,}6 & \text{dla} \quad 7{,}30\le x<7{,}65\\ 0{,}8 & \text{dla} \quad 7{,}65\le x<8{,}10\\ 1{,}0 & \text{dla} \quad x\ge 8{,}10\end{cases}$$

Nadzieja matematyczna wynosi więc:

$$E(X)=\sum_{i=1}^{5}x_i p_i=0\cdot 0{,}2\cdot(7{,}65+6{,}97+7{,}25+8{,}10+7{,}30)=7{,}454\approx 7{,}45$$

co oznacza, że należy się spodziewać oprocentowania w wysokości 7,45%. Wariancja tej zmiennej wynosi:

$$D^2(X)=\sum_{i=1}^{5}(x_i-7{,}45)^2 p_i=\sum_{i=1}^{5}(x_i-7{,}45)^2\cdot 0{,}2=1{,}4908\approx 1{,}49$$

a odchylenie standardowe jest równe:

$$D(X) = \sqrt{1,49} \approx 1,22$$

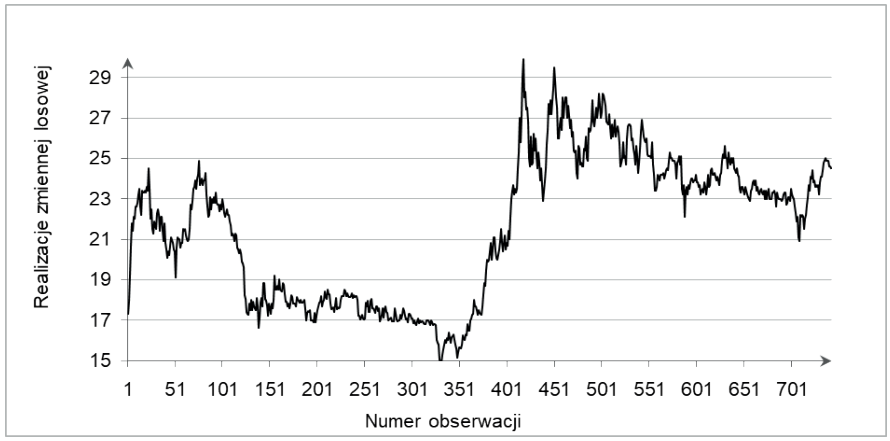
Ostatecznie stwierdza się, że stopa procentowa średnio odchyla się od wartości oczekiwanej o 1,22% i to stanowi 16% wartości oczekiwanej, a zróżnicowanie oprocentowania kredytów hipotecznych w bankach komercyjnych jest niewielkie.

Przykład 1.7

Dzienne ceny otwarcia akcji spółki ORBIS za lata 2002–2004 przedstawiono na rysunku 1.7. W analizowanym okresie ceny te zmieniały się od 14,90 do 29,90 PLN. Należy przeprowadzić analizę notowań cen akcji badanej spółki, w celu opisu funkcji gęstości, wyznaczenia dystrybuanty oraz podstawowych parametrów rozkładu tej zmiennej losowej.

Rozwiązanie

Na podstawie 742 notowań cen akcji spółki Orbis za lata 2002–2004 zostanie przeprowadzona analiza rozkładu tej zmiennej losowej ciągłej.



Rysunek 1.7. Ceny otwarcia akcji spółki Orbis w latach 2002–2004

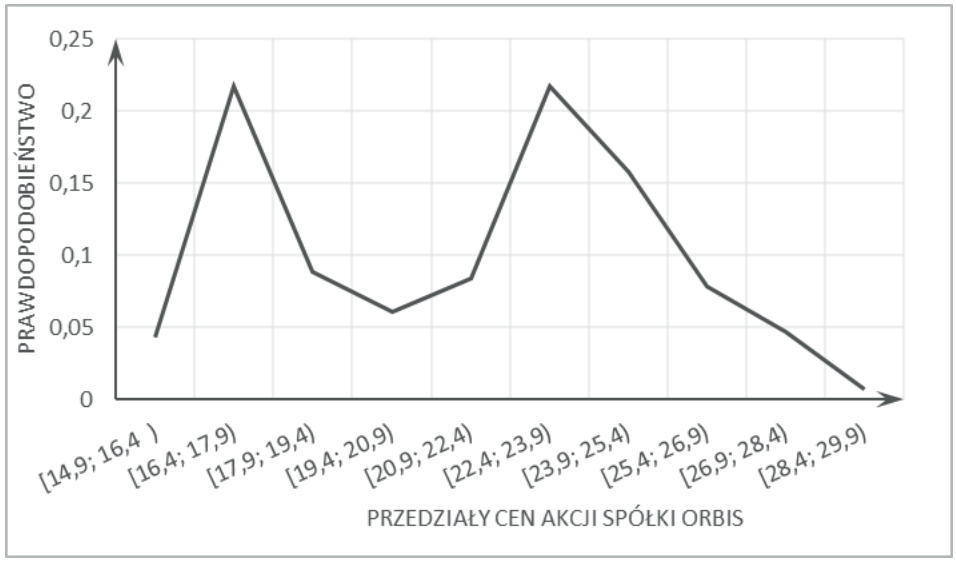
Źródło: opracowanie własne na podstawie danych GPW.

Wszystkie notowania kursu akcji zawierają się w przedziale od 14,9 do 29,9 PLN. W związku z tym dla analizowanej zmiennej losowej utworzono 10 przedziałów klasowych o rozpiętości 1,5 PLN. Po zliczeniu wszystkich obserwacji utworzono szereg rozdzielczy, przedstawiony w tab. 1.9.

Tabela 1.9. Szereg rozdzielczy notowań cen spółki Orbis

Przedziały klasowe $[x_i^d; x_i^g)$	Liczebność n_i	Prawdopodobieństwo p_i	Liczebność skumulowana n_{ski}	Dystrybuanta p_{ski}
[14,9; 16,4)	32	0,0431	32	0,0431
[16,4; 17,9)	161	0,2170	193	0,2601
[17,9; 19,4)	66	0,0889	259	0,3491
[19,4; 20,9)	45	0,0606	304	0,4097
[20,9; 22,4)	62	0,0836	366	0,4933
[22,4; 23,9)	161	0,2170	527	0,7102
[23,9; 25,4)	117	0,1577	644	0,8679
[25,4; 26,9)	58	0,0782	702	0,9461
[26,9; 28,4)	35	0,0472	737	0,9933
[28,4; 29,9)	5	0,0067	742	1
Suma	742	1		

Źródło: opracowanie własne na podstawie danych GPW.



Rysunek 1.8. Rozkład prawdopodobieństwa cen akcji spółki Orbis w latach 2002–2004

Źródło: opracowanie własne na podstawie tab. 1.9.

Na podstawie danych z tab. 1.9 oraz wykresu 1.8 łatwo zauważyć, że funkcja gęstości prawdopodobieństwa badanej zmiennej losowej jest postaci:

$$f(x)=\begin{cases}0{,}000 & \text{dla} & x<14{,}9\\ 0{,}040 & \text{dla} & 14{,}9\le x<16{,}4\\ 0{,}220 & \text{dla} & 16{,}4\le x<17{,}9\\ 0{,}090 & \text{dla} & 17{,}9\le x<19{,}4\\ \quad\cdots & & \\ 0{,}050 & \text{dla} & 26{,}9\le x<28{,}4\\ 0{,}007 & \text{dla} & 28{,}4\le x<29{,}9\\ 0{,}000 & \text{dla} & x>29{,}9\end{cases}$$

Z wykresu (rys. 1.8) wynika także, iż przebieg zmienności notowań cen akcji posiada kilka ekstremów lokalnych. Najczęściej kursy akcji wpadają do przedziału, którego górne granice wynoszą odpowiednio 17,9 i 23,9 PLN.

Dystrybuanta notowań cen może zostać odczytana bezpośrednio z tab. 1.9, z kolumny zawierającej skumulowane prawdopodobieństwa wyznaczone dla wyróżnionych przedziałów zmienności cen akcji p_{ski} i wynosi:

$$F(x)=\begin{cases}0{,}0000 & \text{dla} & x<14{,}9\\ 0{,}0431 & \text{dla} & x\in[14{,}9;16{,}4)\\ 0{,}2601 & \text{dla} & x\in[16{,}4;17{,}9)\\ 0{,}3491 & \text{dla} & x\in[17{,}9;19{,}4)\\ & \cdots & \\ 0{,}9933 & \text{dla} & x\in[26{,}9;28{,}4)\\ 1{,}0000 & \text{dla} & x\in[28{,}4;29{,}9)\end{cases}$$

Najczęstsze, tj. najbardziej prawdopodobne, były notowania kursów akcji spółki ORBIS w przedziałach 16,4–17,9 PLN oraz 22,4–23,9 PLN, przy największym prawdopodobieństwie 0,217. Prawdopodobieństwo, że ceny akcji będą poniżej 19,4 PLN, wyniosło 0,3491, czyli tyle, ile wyniosła wartość dystrybuanty obliczonej dla przedziału [17,9; 19,4).

Wartość oczekiwana, wyznaczona dokładnie na podstawie rozkładu empirycznego, wynosi 21,54 PLN, natomiast odchylenie standardowe $D=\sqrt{12{,}4389}=3{,}53$ PLN, co zostało obliczone w tab. 1.10.

Tabela 1.10. Obliczenie wartości oczekiwanej i wariancji cen akcji spółki Orbis

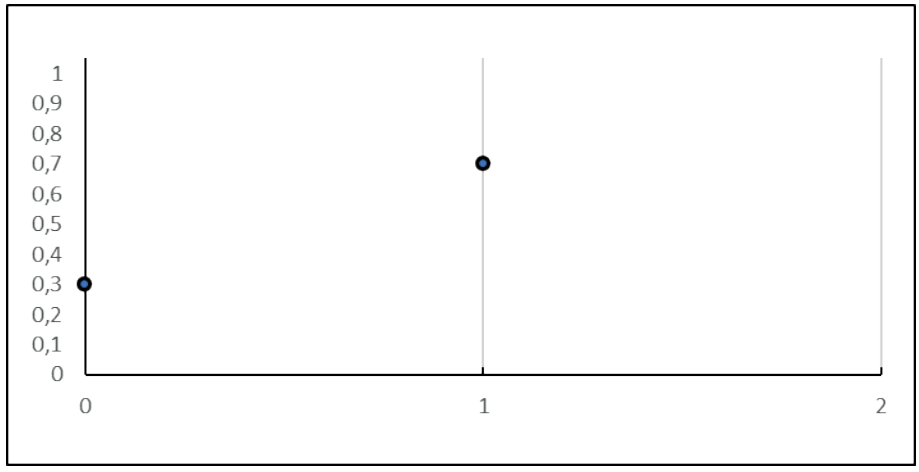
$[x_i^d; x_i^g)$	p_i	$\dot{x}_i$	$\dot{x}_i \cdot p_i$	$\dot{x}_i - E(X)$	$(\dot{x}_i - E(X))^2$	$(\dot{x}_i - E(X))^2 \cdot p_i$
[14,9; 16,4)	0,0431	15,65	0,6749	–5,8908	34,7019	1,4966
[16,4; 17,9)	0,2170	17,15	3,7212	–4,3908	19,2794	4,1833
[17,9; 19,4)	0,0889	18,65	1,6589	–2,8908	8,3569	0,7433

$[x_i^d; x_i^g)$	p_i	$\dot{x}_i$	$\dot{x}_i \cdot p_i$	$\dot{x}_i - E(X)$	$(\dot{x}_i - E(X))^2$	$(\dot{x}_i - E(X))^2 \cdot p_i$
[19,4; 20,9)	0,0606	20,15	1,2220	–1,3908	1,9344	0,1173
[20,9; 22,4)	0,0836	21,65	1,8090	0,1092	0,0119	0,0010
[22,4; 23,9)	0,2170	23,15	5,0231	1,6092	2,5894	0,5619
[23,9; 25,4)	0,1577	24,65	3,8869	3,1092	9,6669	1,5243
[25,4; 26,9)	0,0782	26,15	2,0441	4,6092	21,2444	1,6606
[26,9; 28,4)	0,0472	27,65	1,3042	6,1092	37,3219	1,7605
[28,4; 29,9)	0,0067	29,15	0,1964	7,6092	57,8994	0,3902
Suma	1		21,5408			12,4389

Źródło: opracowanie własne.

W wielu badaniach, w szczególności w procesie wnioskowania statystycznego, wykorzystuje się teoretyczne rozkłady zmiennej losowej jednowymiarowej. Do najczęściej stasowanych należą rozkłady:

- dwupunktowy, który jest rozkładem skokowym (rys. 1.9),
- normalny i t-Studenta będące przykładami rozkładów ciągłych i symetrycznych (rys. 1.10),
- chi-kwadrat (χ^2) i Fishera-Snedecora, które są rozkładami ciągłymi i niesymetrycznymi (rys. 1.12).



Rysunek 1.9. Wykres funkcji prawdopodobieństwa dla rozkładu zero-jedynkowego

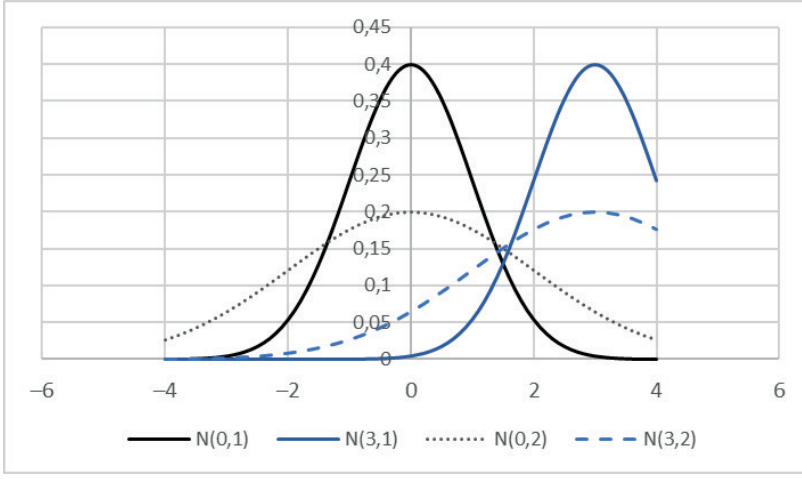
Źródło: opracowanie własne.

Rozkład dwupunktowy charakteryzuje się tym, że zmienna losowa przyjmuje tylko dwie wartości, tj. x_1 i x_2 z prawdopodobieństwem p i $(1 - p)$, co zapisuje się jako:

$$P(X = x_i) = \begin{cases} p & x_i = x_1 \\ (1-p) & x_i = x_2 \end{cases} \tag{1.8}$$

Szczególnym przykładem tego rozkładu jest rozkład zero-jedynkowy, kiedy zmienna losowa przyjmuje wartości 0 lub 1, czyli $x_1 = 1$ oraz $x_2 = 0$.

Najczęściej występującym w przyrodzie i najczęściej wykorzystywanym w analizach statystycznych jest rozkład normalny, zwany też rozkładem Gaussa. Jest to rozkład ciągły i symetryczny w charakterystycznym kształcie dzwonu. Położenie wykresu tego rozkładu zależy od jego wartości oczekiwanej $E(X) = \mu$, a rozpiętość (szerokość) dzwonu determinowana jest przez odchylenie standardowe $D(X) = \sigma$, czyli pierwiastek z wariancji. Na rys. 1.10 pokazano wykresy funkcji gęstości czterech różnych zmiennych losowych o rozkładzie normalnym $N(\mu, \sigma)$. Warto zauważyć, że wykresy zmiennych losowych o rozkładzie N(0, 1) i N(3, 1) są identyczne w kształcie, podobnie jak N(0, 2) z N(3, 2), ale przesunięte na osi poziomej. Natomiast zmienne losowe o rozkładzie N(0, 1) i N(0, 2) oraz N(3, 1) i N(3, 2) znajdują się w tych samych miejscach na osi, ale te drugie są bardziej spłaszczone niż pierwsze wspomniane.



Rysunek 1.10. Funkcje gęstości rozkładu normalnego N(μ, σ) dla różnych wartości μ i σ

Źródło: opracowanie własne.

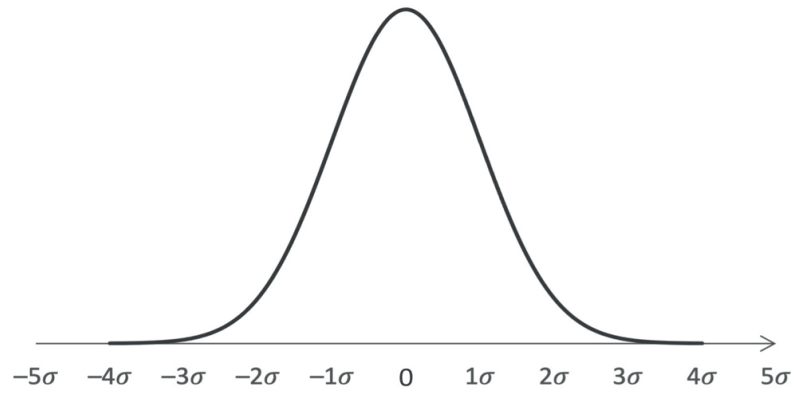
Funkcja gęstości rozkładu normalnego ma następujące własności:

- określoność, gdyż kształt funkcji gęstości zależy od wartości dwóch parametrów: μ i σ (μ decyduje o przesunięciu krzywej, natomiast parametr σ decyduje o „smukłości" krzywej);

- symetryczność względem prostej $x = \mu$, co oznacza, że prawdopodobieństwo, iż zmienna losowa o rozkładzie N(μ, σ) jest mniejsza niż wartość oczekiwana μ, jest identyczne jak prawdopodobieństwo, że zmienna losowa o rozkładzie N(μ, σ) jest większa od μ i wynosi 0,5;
- jednomodalność, co oznacza, że w punkcie $x = \mu$ krzywa rozkładu osiąga wartość maksymalną, czyli wartości zmiennej losowej równe wartości oczekiwanej występują najczęściej;
- zmienność, gdyż ramiona funkcji gęstości mają punkty przegięcia dla $x = \mu - \sigma$ oraz $x = \mu + \sigma$. Własność ta wiąże się z tzw. regułą trzech sigm, według której przyjmuje się, że prawdopodobieństwo, iż realizacje zmiennej losowej ciągłej X nie będą się różniły (*in plus*, *in minus*) od wartości oczekiwanej więcej niż o trzy odchylenia standardowe (rys. 1.11), wynosi w przybliżeniu 1, co można zapisać w postaci relacji:

$$P\{\mu - 3\sigma < X < \mu + 3\sigma\} = 0.9973 \approx 1 \tag{1.9}$$

Zasada trzech sigm wykorzystywana jest do identyfikacji tzw. obserwacji nietypowych (odstających). Wynika z niej bowiem, że jeśli wartość cechy jest mniejsza od $\mu - 3\sigma$ lub większa od $\mu + 3\sigma$, to prawdopodobieństwo, iż obserwacja ta pochodzi z tej samej zbiorowości co pozostałe, jest równe 0 (oczywiście przy założeniu, że rozkład jest normalny).



Rysunek 1.11. Ilustracja reguły trzech sigm

Źródło: opracowanie własne.

W procedurze wnioskowania statystycznego najczęściej korzysta się z rozkładu normalnego standaryzowanego N(0, 1), czyli takiego, dla którego nadzieja matematyczna wynosi 0, a odchylenie standardowe jest równe 1. Procedura standaryzacji

polega na takim przekształceniu zmiennej losowej X o rozkładzie N(μ, σ), aby otrzymać zmienną U o rozkładzie N(0, 1), czyli:

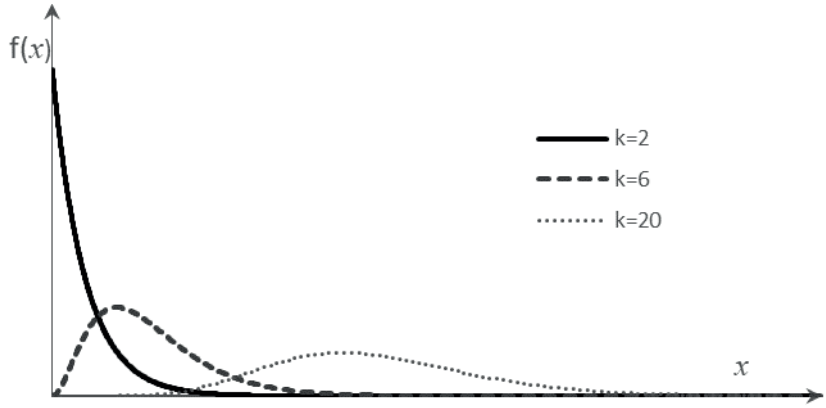
$$u_i = \frac{x_i - \mu}{\sigma} \tag{1.10}$$

gdzie:
u_i, x_i – odpowiednio realizacje zmiennych losowych X i U,
μ, σ – odpowiednio nadzieja matematyczna i odchylenie standardowe zmiennej losowej X.

Warto dodać, że znaczenie rozkładu normalnego związane jest również z faktem, że dla odpowiednio dużej liczby obserwacji zmienne losowe charakteryzujące się różnymi rozkładami prawdopodobieństwa zaczynają mieć rozkład zbieżny do rozkładu normalnego, zgodnie z tzw. centralnym twierdzeniem granicznym.

Innym często używanym w badaniach rozkładem jest rozkład t-Studenta. Uznaje się go za przykład rozkładu ciągłego i symetrycznego, w kształcie zbliżonym do rozkładu N(0, 1). Określa się go jako szybko zbieżny do rozkładu normalnego, ponieważ przyjmuje się, że wystarczającą jest liczebność 30 obserwacji, aby rozkład t-Studenta był zbliżony do rozkładu Gaussa. W przypadku rozkładu t-Studenta ważnym jego parametrem jest tzw. liczba stopni swobody, którą definiuje się jako liczbę niezależnych wyników obserwacji pomniejszoną o liczbę związków, które łączą te wyniki ze sobą.

Kolejnym często stosowanym w badaniach rozkładem jest rozkład chi-kwadrat (χ^2). To rozkład ciągły i niesymetryczny, a jego podstawowym parametrem jest liczba stopni swobody, oznaczona na rys. 1.12 przez k. Rozkład ten jest wolnobieżny do rozkładu normalnego (przyjmuje się, że minimalna liczba stopni swobody wynosi 60). Podobny w kształcie jest tzw. rozkład Fishera-Snedecora, który opisuje się za pomocą dwóch wartości stopni swobody.



Rysunek 1.12. Wykres funkcji gęstości rozkładu chi-kwadrat

Źródło: opracowanie własne.

Rozdział 2

Analiza struktury danych

Jednostki statystyczne, które podlegają badaniu, opisane są za pomocą jednej lub kilku cech charakteryzujących zbiorowość statystyczną. Analizy można zatem prowadzić dla każdej cechy osobno lub badając, czy występują zależności między nimi.

Parametrami opisowymi badanej zbiorowości nazywa się charakterystyki liczbowe umożliwiające sumaryczny i skrócony opis zebranych danych. Parametry opisowe pozwalają na określenie:

- przeciętnej wartości analizowanej cechy statystycznej, przez wyznaczenie pojedynczej wartości będącej miarą przeciętną (miarą położenia),
- zmienności wartości cechy w obserwowanej zbiorowości za pomocą tzw. miar zmienności (dyspersji, rozproszenia),
- w jakim stopniu badany szereg odbiega od idealnej symetrii, do czego wykorzystuje się miary asymetrii.

Analiza parametrów opisowych pozwala wykryć różnice między zbiorowościami, które sprowadzają się do trzech przypadków:

1) rozkłady danych mogą się różnić położeniem, tzn. wartością cechy, w pobliżu której skupiają się obserwacje,
2) obserwacje mogą się skupiać wokół tej samej wartości, lecz różnić obszarem zmienności,
3) rozkłady obserwacji mogą wreszcie różnić się jednocześnie co do obu tych charakterystyk liczbowych.

W opisie właściwości różnych zbiorów danych, istotnym pojęciem jest ich jednorodność (homogeniczność) i jej przeciwieństwo, heterogeniczność. Pojęcia te dotyczą często przyjmowanego założenia, że właściwości statystyczne dowolnej części całego zbioru danych są identyczne jak każdej innej części.

2.1. Wskaźniki struktury

Jak pokazano w poprzednim rozdziale, liczebności przedziałów klasowych mogą być wyrażone w liczbach bezwzględnych (absolutnych) lub w liczbach względnych. W tym drugim przypadku mówi się o liczebności (częstości) względnej, inaczej nazywanej wskaźnikiem struktury w_i, który definiuje się jako stosunek liczebności części zbiorowości (w jakiś sposób wyróżniającej się) do liczebności całej zbiorowości, czyli:

$$w_i = \frac{n_i}{n} \tag{2.1}$$

gdzie:
n_i – liczba jednostek statystycznych charakteryzująca się i-tym wariantem danej cechy ($i = 1, 2, \ldots, k$),
n – liczba wszystkich jednostek statystycznych objętych badaniem, $n = \sum_{i=1}^{k} n_i$.

Wskaźnik struktury informuje, jaki jest udział jednostek statystycznych posiadających i-ty wariant cechy w całej badanej zbiorowości. Wskaźniki struktury są często wyrażane w procentach po przemnożeniu w_i przez 100 (jak to pokazano na rys. 1.1).

Wskaźniki struktury nie zmieniają proporcji pomiędzy liczbami absolutnymi. Wnioski formułowane przy ich użyciu można wyciągnąć na podstawie liczb absolutnych, jednak znacznie łatwiej poznaje się strukturę zjawisk, jeśli jest ona przedstawiona za pomocą liczb względnych. Stanowią one jedną z form opisu rozkładu danej cechy, ułatwiają również porównywanie struktury zbiorowości. W przypadku analizy cech jakościowych wskaźniki struktury są jednym z podstawowych sposobów analizy struktury badanej zbiorowości.

Częstości względne można kumulować, uzyskując w ten sposób udział jednostek statystycznych, charakteryzujących się wariantem cechy sięgającym poziomu górnej granicy wybranego przedziału, dla którego wyznaczono skumulowany wskaźnik struktury. Podawane w szeregach statystycznych liczebności względne, tzw. wskaźniki struktury w_i, odpowiadają prawdopodobieństwu pojawienia się określonego wariantu cechy w badanej zbiorowości (czyli p_i). Natomiast skumulowane wskaźniki struktury, opisujące częstość względną dla wszystkich wariantów cechy mniejszych od górnej granicy wybranego przedziału, czyli $\{x_i: X < x_i\}$, nazywane są dystrybuantą empiryczną.

Przykład 2.1

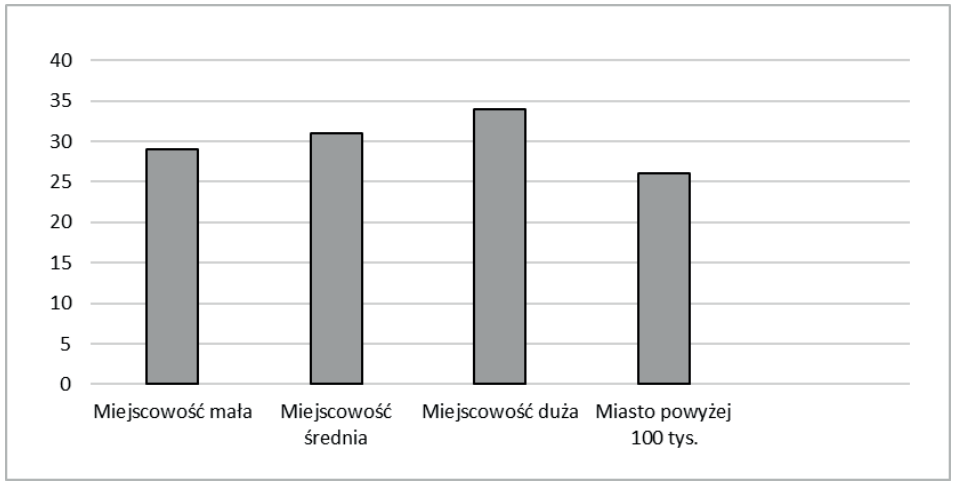
Na danych z przykładu 1.1 i zwiększonej do 120 liczbie przedsiębiorstw pokazane zostanie jakie informacje można uzyskać z analizy rozkładu tych danych. W tab. 2.1 przedstawiono strukturę badanych firm według wielkości miejscowo-

ści w liczebnościach bezwzględnych (tj. w liczbie firm działających w danej klasie miejscowości – kolumna (3)) oraz liczebnościach względnych obliczonych na podstawie wzoru (2.1) jako ułamki całego 120-elementowego zbioru – kolumna (5), i udziały procentowe – kolumna (7). Liczebności skumulowane przedstawiono w kolumnach: (4), (6) oraz (8). Informują one, ile (lub jaka część) przedsiębiorstw funkcjonuje w miejscowościach, w których liczba mieszkańców nie przekracza górnej granicy przedziału. Tak więc połowa firm zlokalizowana jest w miejscowościach małych i średnich (tj. do 50 tys. mieszkańców).

Tabela 2.1. Struktura zbioru przedsiębiorstw według wielkości miejscowości

Klasa miejscowości	Liczba mieszkańców w tys.	Liczebności bezwzględne		Liczebności względne			
		n_i	n_{isk}	w_i	w_{isk}	w_i	w_{isk}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Miejscowość mała	3–20	29	29	0,2417	0,2417	24%	24%
Miejscowość średnia	20–50	31	60	0,2583	0,5000	26%	50%
Miejscowość duża	50–100	34	94	0,2833	0,7833	28%	78%
Miasto powyżej 100 tys.	100 i więcej	26	120	0,2167	1,0000	22%	100%
Suma		120		1,0000		100%	

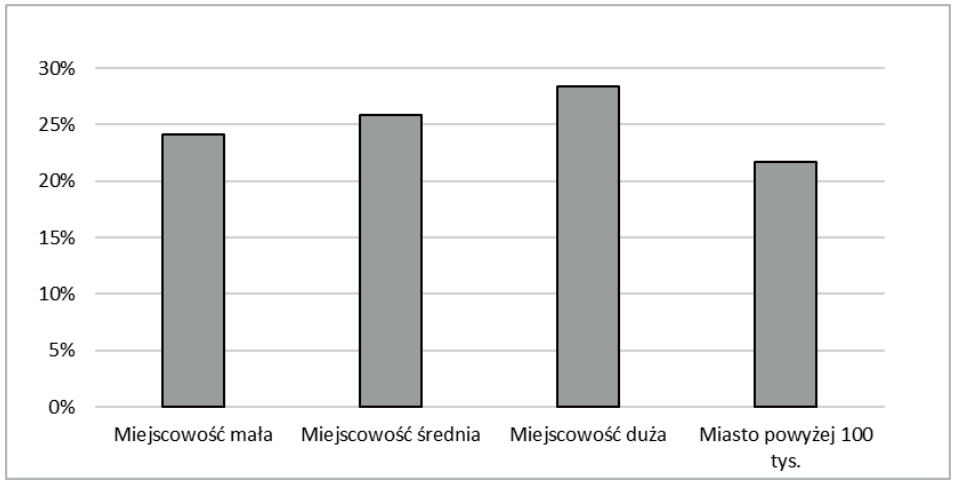
Źródło: opracowanie własne.



Rysunek 2.1. Struktura zbioru przedsiębiorstw według wielkości miejscowości (w liczebnościach bezwzględnych)

Źródło: opracowanie własne.

W celu pokazania, że rozkład empiryczny pozostaje bez zmian niezależnie od tego, czy zastosuje się liczebności bezwzględne i względne, dane z tab. 2.1 przedstawiono na rys. 2.1 i 2.2. Pierwszy z nich pokazuje rozkład wyznaczony według liczby przedsiębiorstw, a na drugim przedstawiono strukturę procentową firm. Jak widać, oba wykresy są identyczne, co oznacza, że struktura danych nie zależy od tego, czy do jej opisu użyje się liczebności bezwzględnych czy względnych.



Rysunek 2.2. Struktura zbioru przedsiębiorstw według wielkości miejscowości (w liczebnościach bezwzględnych, tj. w procentach)

Źródło: opracowanie własne.

2.2. Miary położenia

Miary średnie (zwane też miarami przeciętnymi albo położenia) służą do określania wartości zmiennej, wokół której skupia się większość wartości. Są one najbardziej rozpowszechnionymi miarami statystycznymi używanymi w praktyce, ponieważ za pomocą jednej liczby podają one charakterystykę poziomu zmiennej, czyli centralną tendencję (stąd też często nazywane są również miarami tendencji centralnej). Istnieje kilka różnych typów miar średnich. Ich stosowanie nie jest jednak dowolne, zależy bowiem od celu badania oraz posiadanych danych statystycznych. Wyróżnia się dwa podstawowe rodzaje miar średnich:

- miary klasyczne, do których zalicza się: średnią arytmetyczną, harmoniczną, chronologiczną i geometryczną,

- miary pozycyjne, do których zalicza się: dominantę i kwartyle, przy czym szczególną rolę odgrywa mediana, czyli kwartyl drugi, będąca wartością środkową szeregu. Oprócz kwartylów wyróżnia się również decyle, percentyle itp., zwane ogólnie kwantylami. Kwartyle oznaczają wartości cechy będące punktami podziału uporządkowanego szeregu obserwacji na cztery równoliczne części, decyle – przy podziale na dziesięć części, a percentyle – przy podziale na sto części.

Miary klasyczne są obliczane jako wypadkowe wartości wszystkich wariantów cechy występujących w badanych jednostkach zbiorowości, podczas gdy miary pozycyjne wskazują na określoną pozycję jednostek statystycznych w szeregu. Wszystkie miary przeciętne są mianowane i wyrażane w takich samych jednostkach pomiarowych jak cechy, na podstawie których są wyznaczane.

Średnia arytmetyczna wyraża wartość cechy, którą posiadałaby każda jednostka zbiorowości, gdyby podział sumy wartości był równomierny, tzn. u każdej jednostki zbiorowości występowałaby ta sama wartość cechy. Średnią arytmetyczną definiuje się jako sumę wartości cechy mierzalnej podzieloną przez liczbę jednostek skończonej zbiorowości statystycznej. Wartość średniej arytmetycznej wyznacza się na podstawie wzorów uwzględniających sposób prezentacji danych w szeregu statystycznym. W przypadku szeregu szczegółowego (prostego) korzysta się z relacji:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

Dla szeregu rozdzielczego zawierającego k przedziałów klasowych, w których zmienna reprezentująca badaną cechę statystyczną jest skokowa, a przedziały klasowe jednojednostkowe (punktowe), stosuje się wzór na tzw. średnią ważoną:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \tag{2.3}$$

Omawiając wzory wykorzystywane do wyznaczenia średniej arytmetycznej warto pamiętać, że dotyczą one zarówno liczebności bezwzględnych n_i, jak i liczebności względnych w_i. Wtedy

$$\overline{x} = \sum_{i=1}^{n} x_i \cdot w_i \tag{2.3a}$$

a dla sprawdzenia poprawności tego wzoru wystarczy podstawić (2.1) do (2.3a), co prowadzi do tożsamościowego przekształcenia $\overline{x} = \sum_{i=1}^{n} x_i \cdot w_i = \sum_{i=1}^{n} x_i \cdot \frac{n_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot n_i$ i wzoru (2.3).

Natomiast w przypadku szeregu rozdzielczego o k przedziałach klasowych wielojednostkowych średnią arytmetyczną wyznacza się według wzoru:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i \tag{2.4}$$

gdzie:
x_i – i-ty wariant badanej cechy (wartość zmiennej),
n_i – liczba obserwacji w i-tym przedziale klasowym (tzw. wagi),
n – liczba jednostek objętych badaniem, $n = \sum_{i=1}^{k} n_i$,
k – liczba przedziałów klasowych,
$\dot{x}_i$ – środek i-tego przedziału klasowego wielojednostkowego.

Podobne przekształcenia liczebności bezwzględnych na wskaźniki struktury można przeprowadzić dla szeregu rozdzielczego z przedziałami klasowymi, wtedy wzór (2.4) jest równoważny zapisowi:

$$\overline{x} = \sum_{i=1}^{n} \dot{x}_i \cdot w_i \tag{2.4a}$$

Warto odnotować, że wskaźnik struktury w_i opisuje częstotliwość, tzn. prawdopodobieństwo wystąpienia i-tego wariantu cechy, które w przypadku zmiennych losowych oznacza się przez p_i. Zatem wzór (2.3a) jest równoważny wzorowi (1.3) opisującemu wartość oczekiwaną zmiennej losowej skokowej.

Zwrócić należy uwagę na kilka podstawowych własności średniej arytmetycznej.

1) Jest ona wypadkową wszystkich wartości badanej cechy, w związku z czym nie może być niższa od najmniejszej wartości zaobserwowanej w badaniu ani wyższa od wartości największej, czyli $x_{max} \geq \overline{x} \geq x_{min}$, gdzie: x_{min}, x_{max} to odpowiednio najmniejsza i największa wartość szeregu.
2) Wartość średniej zależy nie od liczebności klas, lecz od ich wzajemnej proporcji, co oznacza, że można ją również wyznaczyć z szeregów, gdzie liczebności bezwzględne zastąpiono względnymi, a otrzymane wyniki będą jednakowe[1].
3) Suma odchyleń wartości badanej cechy od średniej arytmetycznej jest równa 0, czyli: $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.
4) Suma kwadratów odchyleń poszczególnych wartości zmiennych badanej cechy od średniej arytmetycznej rozkładu jest najmniejsza. Oznacza to, że suma kwadratów odchyleń poszczególnych wartości zmiennej rozkładu od wartości różnej od średniej będzie zawsze większa.

1 Porównaj wzory (2.3) i (2.4) z (2.3a) i (2.4a).

Średnia arytmetyczna jest wprawdzie najprostszą z miar średnich, jednak nie zawsze jest ona dobrym miernikiem tendencji centralnej. Chodzi tu o sytuację, kiedy największe liczebności skupiają się wokół najniższych lub najwyższych wartości cechy. Niekiedy średnia arytmetyczna wprowadza po prostu w błąd. Dzieje się tak wówczas, gdy wyznacza się średnią ze zbiorów niejednorodnych[2], ponieważ średnia arytmetyczna zaciera różnice indywidualne. Innym przypadkiem, kiedy średnia arytmetyczna nie powinna być stosowana, jest występowanie zbiorowości z jednostkami odstającymi, ponieważ średnia zawierać będzie wartości nietypowe, a tym samym nie będzie poprawnie opisywać położenia. W tej sytuacji można policzyć średnią arytmetyczną, odrzucając jednostki odstające pod warunkiem, że stanowią one co najwyżej 3% liczebności całej zbiorowości.

Stosując wzór (2.4), można obliczyć średnią arytmetyczną dla zamkniętych przedziałów klasowych, ponieważ tylko wtedy daje się wyznaczyć środki tych przedziałów. W przypadku gdy przedziały klasowe (pierwszy lub ostatni) są otwarte, a ich liczebności są stosunkowo małe[3], można dokonać umownego ich zamknięcia i ustalić wartości środków przedziałów. Nie można jednak tak postąpić w przypadku, gdy udział liczebności otwartych przedziałów w ogólnej sumie liczebności jest znaczny. Innymi słowy, w tej sytuacji nie da się wyznaczyć średniej arytmetycznej.

Kiedy nie można lub nie należy stosować miar klasycznych, zaleca się korzystanie z tzw. miar pozycyjnych. Miarą, która rozdziela całą populację na dwie równe części, jest mediana, tzn. drugi kwartyl, a także piąty decyl i pięćdziesiąty percentyl. Podziału dokonuje się tak, że w jednej grupie znajdują się jednostki o wartościach niższych (lub równych) od mediany, a w drugiej o wartościach wyższych, a każda część zawiera 50% zbiorowości. Wynika z tego, że dla znalezienia mediany trzeba najpierw uporządkować zbiorowość według wielkości jej elementów, tzn. od ich wartości najmniejszej do największej (lub odwrotnie).

W celu wyznaczenia kwantylu oblicza się jego tzw. numer, który dla mediany wynosi $N_{Me} = \frac{n}{2}$. W przypadku gdy numer mediany jest liczbą niecałkowitą, zaokrągla się go w górę do pierwszej liczby całkowitej. Następnie w uporządkowanym szeregu szczegółowym znajduje się jednostkę o numerze mediany i jej wartość jest medianą. Innymi słowy:

$$Me = x_i \quad \text{dla} \quad i = N_{Me} \quad \text{gdzie} \quad N_{Me} = \frac{n}{2} \tag{2.5}$$

Jeśli liczebność szeregu jest liczbą parzystą, istnieją dwie liczby „środkowe". W takiej sytuacji za medianę można przyjąć dowolną z liczb (byle zawsze postępować według tego samego schematu) lub średnią arytmetyczną z obu wartości.

2 Zbiory niejednorodne mają rozkłady z kilkoma ośrodkami dominującymi.

3 Przyjmuje się, że nie przekraczają 3 do 5% liczebności całej zbiorowości.

W podobny sposób jak medianę (będąca kwartylem drugim), wyznacza się wartości pozostałych kwantylów[4] (wartości ćwiartkowych). Kwartyl pierwszy (Q_1) jest wartością, która dzieli szereg w taki sposób, że 1/4 jednostek ma od niego wartości mniejsze lub równe, a 3/4 zbiorowości – większe lub równe. Kwartyl trzeci (Q_3) to taka wartość, od której 3/4 jednostek zbiorowości ma wartości nie większe od Q_3, a 1/4 nie mniejsze. Numery odpowiadające kwartylom znajduje się według wzorów: $N_{Q1} = \frac{n}{4}$ i $N_{Q3} = \frac{3n}{4}$. Z przedstawionych wzorów wynika, że numery kwartylów N_{Qi} mogą być liczbami całkowitymi bądź mieć część ułamkową. W tym drugim przypadku numer kwartylu zaokrągla się w górę do najbliższej liczby całkowitej.

W przypadku liczebnie dużej zbiorowości, ujętej w szereg rozdzielczy, przy poszukiwaniu mediany wykorzystuje się szereg skumulowanych liczebności, a wartość mediany wyznacza się na podstawie wzoru:

$$Me = Q_2 = x_0 + \frac{h_0}{n_0}(N_{Me} - n_{sk-1}) \tag{2.6}$$

gdzie:
x_0 – dolna granica przedziału mediany,
h_0 – rozpiętość przedziału mediany,
n_0 – liczebność przedziału mediany,
N_{Me} – numer mediany,
n_{sk-1} – liczebność skumulowana przedziału klasowego poprzedzającego przedział mediany.

W podobny sposób jak we wzorze (2.6) wyznacza się wartości pozostałych kwartylów dla szeregów rozdzielczych, czyli:

$$Q_1 = x_0 + \frac{h_0}{n_0}(N_{Q_1} - n_{sk-1}) \tag{2.7}$$

$$Q_3 = x_0 + \frac{h_0}{n_0}(N_{Q_3} - n_{sk-1}) \tag{2.8}$$

gdzie oznaczenia jak poprzednio.

Najpopularniejszą wśród miar przeciętnych pozycyjnych jest dominanta, zwana niekiedy wartością dominującą, modalną lub modą. Dominanta to wartość cechy, której odpowiada największa liczba obserwacji. Innymi słowy jest to wartość

4 Kwartyle nie są jednak miarami tendencji centralnej, bo nie wyrażają wartości przeciętnej zjawiska. Zostały tu zamieszczone ze względu na podobieństwo rachunkowe w stosunku do mediany.

najczęściej występująca w szeregu. Można ją stosować zarówno do cech niemierzalnych, jak i mierzalnych. W przypadku indywidualnych informacji, tj. szeregu prostego, łatwo można wyznaczyć dominantę przez zliczenie jednostek o danej wartości cechy. Wariant cechy, który ma największą częstotliwość, jest dominantą. W przypadku szeregu z punktowymi przedziałami klasowymi można łatwo dostrzec, w którym z nich pojawia się największa liczebność. W celu wyznaczenia przybliżonej wartości dominanty na podstawie szeregu rozdzielczego z przedziałami klasowymi wielojednostkowymi korzysta się z następującego wzoru:

$$Do = x_0 + \frac{n_0 - n_{-1}}{(n_0 - n_{-1}) + (n_0 - n_{+1})} \cdot h_0 \tag{2.9}$$

gdzie:
x_0 – dolna granica przedziału dominanty,
n_0 – liczebność przedziału dominanty,
n_{-1} – liczebność przedziału poprzedzającego przedział dominanty,
n_{+1} – liczebność przedziału następującego po przedziale dominanty,
h_0 – rozpiętość przedziału dominanty.

Warto zauważyć, że dla cech ciągłych dominanta powinna być wyznaczana z danych pogrupowanych. Należy podkreślić, że dominantę można wyznaczyć tylko wtedy, gdy są spełnione poniższe warunki:

1) rozkład musi być jednomodalny, tzn. mieć jeden, wyraźnie zaznaczony ośrodek dominujący (wyraźne skupienie największej części jednostek wokół jednej wartości),
2) jeżeli wartości cechy są podane w przedziałach klasowych, to przynajmniej przedział dominanty i dwa z nim bezpośrednio sąsiadujące przedziały muszą być jednakowej rozpiętości,
3) szereg nie może być skrajnie asymetryczny ze skrajnym przedziałem dominującym (ostatnim lub pierwszym).

Niespełnienie warunku (1) oznacza, że szereg nie posiada wartości dominującej i jest np. bimodalny. Natomiast warunki (2) i (3) uniemożliwiają wyznaczenie mody ze wzoru (2.9).

2.3. Miary zróżnicowania

Miary średnie określają przeciętny poziom zjawiska, nie informują jednak o zróżnicowaniu badanej cechy. W tym przypadku niezwykle pomocne okazują się miary dyspersji, zwane inaczej miarami rozrzutu, zróżnicowania lub

rozproszenia. Należą więc one do charakterystyk opisujących rozkład cechy, pozwalają bowiem mierzyć zróżnicowanie wartości zmiennej w ramach badanej zbiorowości, a zatem informują, jak duże są różnice (odchylenia) między poszczególnymi wartościami jednostek zbiorowości a miarą przeciętną. Stopień, w jakim poszczególne wartości odbiegają od wartości średniej, czyli stopień zmienności, decyduje niejednokrotnie o znaczeniu danej miary położenia jako charakterystyki badanego szeregu. Im mniejszy jest stopień zmienności, tym większe jest znaczenie danej miary. Praktycznie wszystkie zbiorowości statystyczne charakteryzuje większa lub mniejsza zmienność, czyli wartości jednostek różnią się od siebie w większym lub mniejszym stopniu. Bywa często tak, że średnie arytmetyczne dwóch szeregów są jednakowe, a mimo to szeregi różnią się bardzo między sobą stopniem zmienności i skupieniem poszczególnych wartości wokół średniej arytmetycznej (porównaj rys. 1.10). Wnioskowanie o badanych zbiorowościach wyłącznie na podstawie ich miar średnich jest więc niewystarczające i jednostronne, a niekiedy może nawet prowadzić do fałszywej oceny badanego zjawiska.

Miary dyspersji dzieli się na dwie podstawowe grupy:

- bezwzględne (absolutne) miary zmienności, które są wielkościami mianowanymi (podobnie jak miary średnie),
- względne (relatywne) miary zmienności, które są wielkościami niemianowanymi lub mogą być wyrażone w procentach.

Jednocześnie, z uwagi na fakt, że miary rozrzutu określają odchylenia od średnich, wyróżnia się wśród nich:

- miary klasyczne, takie jak odchylenie przeciętne i standardowe,
- miary pozycyjne, do których należą rozstęp i odchylenie ćwiartkowe.

Najprostszą miarą rozproszenia jest rozstęp, nazywany inaczej obszarem zmienności, który jest wyznaczany jako różnica między najwyższą i najniższą wartością cechy w badanej zbiorowości statystycznej, czyli:

$$R_x = x_{max} - x_{min} \tag{2.10}$$

gdzie:
x_{min}, x_{max} – to odpowiednio najmniejsza i największa wartość szeregu.

Miara ta używana jest tylko przy wstępnej analizie danych, ponieważ opierając się na dwóch skrajnych wartościach, trudno jest określić rzeczywistą dyspersję występującą w badanej zbiorowości.

Najbardziej popularnym miernikiem dyspersji jest odchylenie standardowe, wyznaczane jako pierwiastek kwadratowy z wariancji $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, będącej średnią arytmetyczną kwadratów odchyleń poszczególnych wartości zbiorowości statystycznej od ich średniej arytmetycznej.

Odchylenie standardowe *S* (a także wariancja), podobnie jak średnia arytmetyczna, obliczane jest według wzorów uwzględniających konstrukcję szeregu statystycznego. W przypadku szeregu prostego odchylenie standardowe wyznacza się jako:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.11}$$

W przypadku szeregu rozdzielczego o punktowej prezentacji wariantów danej cechy w szeregu odchylenie standardowe jest postaci:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i} \tag{2.12}$$

a dla szeregu rozdzielczego z przedziałami klasowymi wielojednostkowymi postaci:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 \cdot n_i} \tag{2.13}$$

gdzie wszystkie oznaczenia jak poprzednio.

Równoważnym, choć rzadziej stosowanym miernikiem dyspersji jest odchylenie przeciętne *d*, które – podobnie jak odchylenie standardowe – określa, o ile wartości wszystkich jednostek badanej zbiorowości średnio różnią się od wartości średniej arytmetycznej. Wyznacza się je jako średnią arytmetyczną bezwzględnych wartości (modułów) odchyleń poszczególnych wartości zbiorowości statystycznej od średniej. Dla szeregu prostego wartość tego odchylenia wyznacza się jako:

$$d = \frac{\sum_{i=1}^{n}\left|x_i - \overline{x}\right|}{n} \tag{2.14}$$

Dla szeregów rozdzielczych stosuje się wzory na odchylenie przeciętne ważone, które w przypadku przedziałów klasowych jednojednostkowych wyraża relacja:

$$\frac{\sum_{i=1}^{k}\left|x_i - \overline{x}\right| \cdot n_i}{n} \tag{2.15}$$

a dla przedziałów wielojednostkowych wzór:

$$d = \frac{\sum_{i=1}^{k}\left|\dot{x}_i - \overline{x}\right| \cdot n_i}{n} \tag{2.16}$$

gdzie wszystkie oznaczenia jak poprzednio.

Można wyróżnić kilka własności odchylenia przeciętnego i standardowego:

1) jako miary klasyczne są one wielkościami obliczanymi na podstawie wszystkich obserwacji i można je poddawać przekształceniom algebraicznym,
2) ich wartość nie ulega zmianie, jeśli liczebności szeregu zostaną wyrażone w liczbach względnych (procentach)[5],
3) dodanie lub odjęcie od wszystkich wartości zmiennej w szeregu jakiejkolwiek (tej samej) liczby nie zmienia ich wartości,
4) jeśli wszystkie wartości szeregu zostaną pomnożone lub podzielone przez jakąkolwiek (tę samą) liczbę $a > 0$, to odchylenia przeciętne i standardowe będą również tylokrotnie większe lub mniejsze,
5) oba odchylenia przedstawiają zróżnicowanie cechy w stosunku do średniej arytmetycznej i są wyrażone w takich samych jednostkach pomiarowych jak cecha,
6) między odchyleniem przeciętnym i standardowym zachodzi relacja:

$$d \leq S \tag{2.17}$$

Niekiedy w analizach statystycznych chodzi o wydzielenie z badanej zbiorowości takiego podzbioru, który składa się z najbardziej typowych jednostek. Przyjmuje się, że typowy przedział zmienności to przedział o długości dwóch odchyleń standardowych, czyli: $[\overline{x} - S, \overline{x} + S]$. Zupełnie inną kwestią jest identyfikacja takich obserwacji, które są nietypowe (inaczej odstające), za które przyjmuje się obserwacje nienależące do przedziału różniącego się od średniej o trzy odchylenia standardowe[6], tj. $[\overline{x} - 3 \cdot S, \overline{x} + 3 \cdot S]$. Występowanie obserwacji odstających powoduje, że zastosowanie klasycznych miar opisowych nie oddaje poprawnie charakteru zbiorowości. Zaleca się wtedy:

- korzystanie wyłącznie z miar pozycyjnych,
- usunięcie lub kalibrację obserwacji nietypowych przed zastosowaniem miar klasycznych.

Pozycyjną, bezwzględną miarą dyspersji jest odchylenie ćwiartkowe Q, będące połową różnicy między kwartylem trzecim a kwartylem pierwszym, czyli:

$$Q = \frac{Q_3 - Q_1}{2} \tag{2.18}$$

Ze względu na to, że kwartyl pierwszy oddziela czwartą część jednostek o wartościach najniższych, a kwartyl trzeci – czwartą część jednostek o wartościach najwyższych, odchylenie ćwiartkowe mierzy rozpiętość połowy najbardziej typowych jednostek zbiorowości. Miarę tę stosuje się wówczas, gdy do opisu tenden-

5 Warto zauważyć, że wyrażenie pod pierwiastkiem wzoru (2.12) jest wariancją, którą można zapisać w podobny sposób jak wariancję (1.6), jako: $S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i = \sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot w_i$

6 Wywodzi się to z reguły trzech sigm, która została zilustrowana na rys. 1.11.

cji centralnej zastosowano medianę, szczególnie jeżeli rozkład cechy jest skrajnie asymetryczny lub występują obserwacje nietypowe. Interpretacja odchylenia ćwiartkowego, wyrażanego w identycznych jednostkach jak mediana, jest podobna do wcześniej omawianych miar, tzn. informuje, o ile wartości cechy przeciętnie różnią się od mediany. Czasami, zwłaszcza w literaturze anglosaskiej, korzysta się z tzw. odchylenia międzykwartylowego, zwanego też rozstępem ćwiartkowym będącym licznikiem wzoru (2.18).

Wprawdzie odchylenia ćwiartkowe, standardowe i przeciętne, są czułymi miarami rozproszenia wartości zmiennej, ale nie mogą być one używane do porównań zmienności poszczególnych cech, obiektów, zbiorowości lub obserwacji poczynionych w różnych momentach pomiarowych, gdyż ich wartości uzależnione są m.in. od absolutnych wielkości obserwacji. W tym przypadku stosuje się względne miary rozrzutu, tzw. współczynniki zmienności. Definiuje się je jako stosunek wartości bezwzględnej miary dyspersji do średniej wyznaczonej dla badanej zbiorowości i wykorzystuje do porównywania dwu lub kilku zbiorowości. Współczynniki zmienności wyraża się wzorami:

$$V_S = \frac{S}{\overline{x}} \tag{2.19}$$

kiedy miernik oparty jest na odchyleniu standardowym, a dla odchylenia przeciętnego jest on postaci:

$$V_d = \frac{d}{\overline{x}} \tag{2.20}$$

Współczynnik zmienności oblicza się również dla odchylenia ćwiartkowego jako:

$$V_Q = \frac{\frac{1}{2}(Q_3 - Q_1)}{Me} = \frac{Q}{Me} \tag{2.21}$$

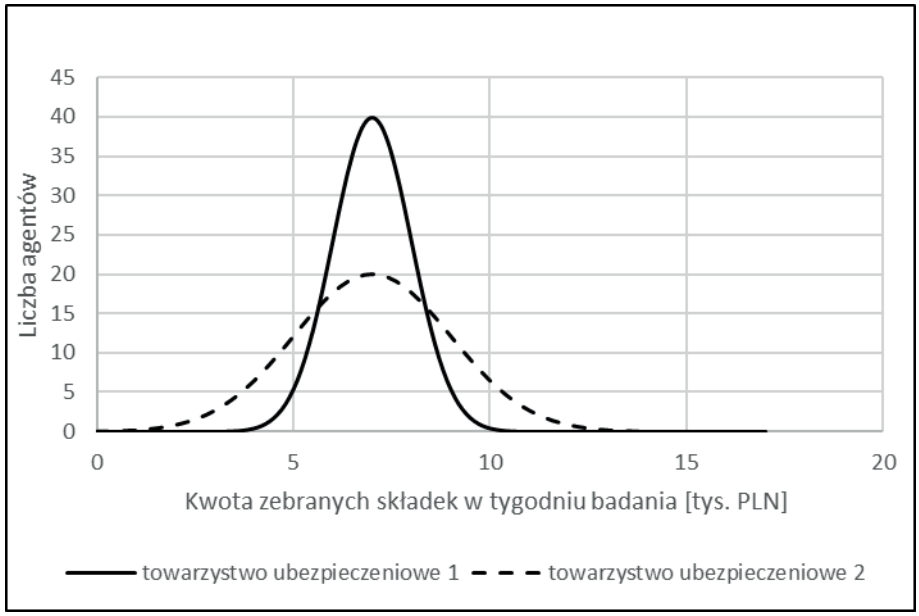
Współczynniki zmienności charakteryzują stosunek nasilenia przyczyn ubocznych do przyczyn głównych. Jako liczby niemianowane (lub wyrażone w procentach) ułatwiają porównywanie zmienności w zbiorowościach, w których wielkości cech są wyrażone w różnych jednostkach. Pozwalają również lepiej ocenić dyspersję występującą w różnych obiektach lub zjawiskach, nawet jeśli dotyczy to tej samej cechy. Zmienność jest jedną z własności wszystkich szeregów, która jest ściśle związana z rodzajem zjawiska, jakie opisują. Przykładowo finansowe szeregi czasowe odzwierciedlające dzienne notowania stóp procentowych (np. WIBOR) charakteryzują się niewielką zmiennością, a ceny akcji charakteryzuje zazwyczaj wysoka zmienność.

Warto zauważyć, że współczynniki zmienności można wykorzystać do wyboru adekwatnej miary położenia analizowanego zjawiska. W szczególności znaczna różnica ich wartości wyznaczonych dla klasycznych i pozycyjnych miar położenia

oznacza zazwyczaj, że lepiej (bardziej poprawnie) strukturę zbiorowości opisują miary pozycyjne. Względne miary zmienności wykorzystywane są również przy wyborze cech reprezentujących rozpatrywane zbiorowości lub zjawiska, pozwalają bowiem ustalić atrybuty różnicujące badaną zbiorowość.

Przykład 2.2

Zbadano po 500 agentów ubezpieczeniowych z 2 towarzystw ubezpieczeniowych oferujących podobny zakres i rodzaj ubezpieczeń. Jedną z analizowanych informacji jest kwota składek wpłaconych przez klientów za zawarte polisy w tygodniu, kiedy prowadzone było badanie. Zebrane dla obu towarzystw dane utworzyły dwa szeregi o tej samej wartości średniej arytmetycznej równej 7 tys. PLN, ale o znakomicie różnych wartościach odchylenia standardowego, które w przypadku pierwszego towarzystwa ubezpieczeniowego wynosi 1 tys. PLN, a w przypadku drugiego jest dwa razy większe (tj. wynosi 2 tys. PLN), co jest widoczne na rys. 2.3.



Rysunek 2.3. Porównanie szeregów wartości zebranych składek

Źródło: opracowanie własne.

Jak widać na rysunku, w obu towarzystwach zebrane w ciągu tygodnia przez agentów ubezpieczeniowych składki najczęściej wynosiły 7 tys. PLN, co stanowi dominantę. Przy czym w pierwszym towarzystwie ubezpieczeniowym było dwa

razy więcej agentów (tj. 40 osób), którzy zebrali składki tej wysokości niż w drugim towarzystwie ubezpieczeniowym (20 osób), co można odczytać z osi opisującej liczbę agentów dla 7 tys. PLN wartości zebranych składek. Łatwo też zauważyć, że w drugim towarzystwie ubezpieczeniowym więcej osób niż w pierwszym towarzystwie ubezpieczeniowym płaciło składki ubezpieczeniowe wyższe (a także niższe) niż 7 tys. PLN, co jest przyczyną większego rozrzutu wysokości składek w tym towarzystwie.

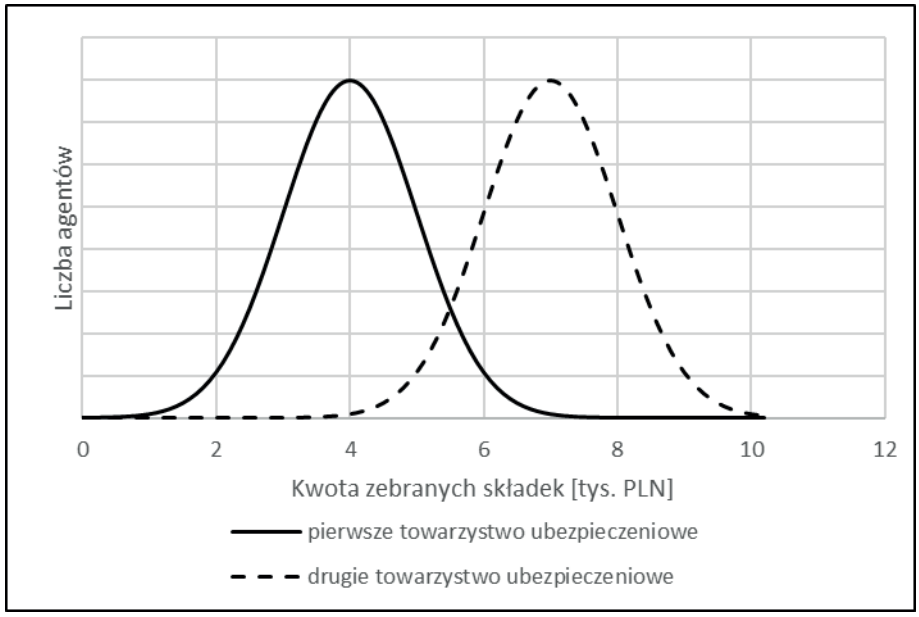
W prezentowanym przykładzie nie tylko średnia i dominanta są równe, ale również mediana, która dzieli oba szeregi na dwie takie same części. Dzięki temu wiadomo, że połowa agentów ubezpieczeniowych (tj. 250) zbiera w badanym tygodniu składki o mniejszej wartości niż wartość środkowa, a połowa o większej wartości w obu towarzystwach ubezpieczeniowych. Natomiast pozostałe dwa kwartyle w obu szeregach są różne. I tak w przypadku pierwszego towarzystwa ubezpieczeniowego 125 agentów zebrało w danym tygodniu mniej niż 6,22 tys. PLN – kwartyl pierwszy (Q_1), a pozostali (tj. 375 agentów) więcej niż Q_1, natomiast wartość kwartylu trzeciego (Q_3) wynosi 7,58 tys. PLN. Zatem 375 agentów zebrało w ciągu tygodnia mniej niż wskazana kwota, a 125 więcej niż wartość kwartylu trzeciego. W przypadku drugiego badanego towarzystwa ubezpieczeniowego wartość kwartylu pierwszego wynosi 5,55 tys. PLN, a trzeciego – 8,25 tys. PLN. Oznacza to, że 125 agentów zebrało mniejsze składki niż Q_1 i tyle samo większe niż Q_3. Zróżnicowana zmienność obu szeregów jest również widoczna, jeśli do jej oceny zastosuje się odchylenia ćwiartkowe, które w przypadku pierwszego towarzystwa wynosi 0,68 tys. PLN, a drugiego 1,35 (czyli dwa razy więcej).

Jak wcześniej wspomniano, porównanie zmienności różnych szeregów powinno się przeprowadzić, korzystając ze wskaźników zmienności. W prezentowanym przykładzie średnia arytmetyczna i mediana obu szeregów są identyczne, zatem wnioski wysnute na podstawie miar bezwzględnych będą identyczne jak w przypadku miar względnych. Do odmiennych wniosków dochodzi się porównując współczynniki zmienności (2.19) i (2.21) – w przypadku pierwszego towarzystwa ubezpieczeniowego wynoszą one odpowiednio dla miar klasycznych $\frac{1}{7} = 14{,}29\%$ i pozycyjnych 9,65%, a w przypadku drugiego: $\frac{2}{7} = 28{,}57\%$ oraz 19,31%, odpowiednio dla miar klasycznych i pozycyjnych.

W przypadku obserwacji dotyczących pierwszego towarzystwa ubezpieczeniowego wykres (rys. 2.3) jest smukły, co świadczy o małej zmienności szeregu. Innymi słowy poszczególne obserwacje nie odbiegają znacząco od wartości średniej. Zupełnie inny charakter ma wykres danych uzyskanych dla drugiego towarzystwa ubezpieczeniowego – jest spłaszczony, co świadczy o znacznym zróżnicowaniu wartości szeregu wokół średniej.

Przykład 2.3

Analiza składek zebranych przez 500 agentów ubezpieczeniowych w ciągu jednego tygodnia charakteryzuje się różnymi wartościami średniej arytmetycznej, wynoszącymi odpowiednio 4 i 7 tys. PLN dla pierwszego i drugiego towarzystwa ubezpieczeniowego, ale jednakową wartością (równą 1 tys. PLN) odchylenia standardowego, co zilustrowano na rys. 2.4.



Rysunek 2.4. Porównanie szeregów wartości zebranych składek

Źródło: opracowanie własne.

Oba wykresy na rys. 2.4 mają ten sam kształt, ale różnią się położeniem[7]. W przeciwieństwie do wykresów zamieszczonych na rys. 2.3, których kształty są zasadniczo różne, ale miary położenia, tj. średnie, mediana i dominanta, są identyczne. Wracając do danych z przykładu 2.3, zauważa się również, że w obu szeregach mediana i dominanta są równe średnim wartościom tygodniowych składek zebranych przez agentów obu towarzystw ubezpieczeniowych, co wraz z obliczonymi miarami zmienności przedstawiono w tab. 2.2.

7 Stąd miary średnie noszą nazwę miar położenia, określają bowiem miejsce wykresu na osi współrzędnych. Jak łatwo odczytać z rysunku, średnia wartość zebranych składek w przypadku agentów pierwszego towarzystwa ubezpieczeniowego wynosi 4 tys. PLN, a drugiego 7 tys. PLN.

Tabela 2.2. Opis struktury wartości składek zebranych w ciągu tygodnia w dwóch towarzystwach ubezpieczeniowych

	Towarzystwo ubezpieczeniowe 1	Towarzystwo ubezpieczeniowe 2
Średnia arytmetyczna	4 tys. PLN	7 tys. PLN
Dominanta	4 tys. PLN	7 tys. PLN
Mediana (kwartyl drugi)	4 tys. PLN	7 tys. PLN
Kwartyl pierwszy	3,22 tys. PLN	6, 22 tys. PLN
Kwartyl trzeci	4,58 tys. PLN	7,58 tys. PLN
Odchylenie standardowe *S*	1 tys. PLN	1 tys. PLN
Odchylenie ćwiartkowe *Q*	0,68 tys. PLN	0,68 tys. PLN
Współczynnik zmienności *S*	25%	14%
Współczynnik zmienności *Q*	17%	10%

Źródło: opracowanie własne.

Jak można wyczytać z tab. 2.2 zmiana położenia obu wykresów skutkuje przesunięciem kwartylów w obu szeregach. Można także zauważyć, że chociaż bezwzględne miary dyspersji są identyczne, to miary względne wskazują na mniejsze zróżnicowanie danych dotyczących drugiego towarzystwa ubezpieczeniowego.

W przypadku analizy cech niemierzalnych najczęściej poprzestaje się na analizie wskaźników struktury i skumulowanych częstości względnych oraz na wyznaczeniu najczęściej występującego wariantu cechy, czyli dominancie, co oczywiście nie wyklucza możliwości graficznej prezentacji danych statystycznych. Zauważyć należy, że dla cech mierzonych na skali nominalnej nie wyznacza się miar dyspersji, bo nie można wyznaczyć miar położenia.

2.4. Miary asymetrii i koncentracji

Pomiar asymetrii można oprzeć na spostrzeżeniu, że w szeregu symetrycznym (porównaj rys. 2.3 i 2.4) średnia arytmetyczna, dominanta i mediana są identyczne, tzn. $\bar{x} = Me = Do$, natomiast w szeregu asymetrycznym kształtują się one na

różnym poziomie. Im większą skośnością charakteryzuje się szereg, tym większe są różnice między tymi miarami. Pozycja dominanty w szeregu nie jest stała. Mediana zajmuje stałe miejsce w szeregu, a średnia arytmetyczna pozostaje pod wpływem wartości skrajnych. Fakty te zostały wykorzystane do określenia kierunku i siły asymetrii, którą można mierzyć np. jako różnicę pomiędzy średnią arytmetyczną a wartością modalną. Jest to najprostsza miara asymetrii nazywana wskaźnikiem asymetrii:

$$M_s = \bar{x} - Do \tag{2.21}$$

Wskaźnik asymetrii (2.21), zwany również miernikiem skośności, dla szeregu symetrycznego jest równy 0. W szeregach asymetrycznych miernik skośności może być większy lub mniejszy od 0, mówi się wówczas o asymetrii prawostronnej (dodatniej) lub lewostronnej (ujemnej). W szeregu o skośności prawostronnej wartości skrajne położone są z prawej strony średniej. Powoduje to przesunięcie średniej arytmetycznej w kierunku prawym w stosunku do dominanty i mediany. Innymi słowy, rozkład prawostronnie skośny to rozkład, którego prawy ogon jest dłuższy, co oznacza, że w szeregu występuje dużo obserwacji, które przyjmują wartości mniejsze od średniej. W szeregu o skośności lewostronnej sytuacja jest odwrotna.

Inna miara skośności została skonstruowana dzięki spostrzeżeniu, iż w szeregu symetrycznym różnica pomiędzy kwartylem trzecim a medianą jest taka sama jak różnica pomiędzy medianą a kwartylem pierwszym. W szeregu asymetrycznym obie różnice będą miały inną wartość. Współczynnik skośności określa zarówno kierunek, jak i siłę asymetrii, a wyznacza się go na podstawie relacji:

$$A_s = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} \tag{2.22}$$

Współczynnik skośności (2.22) jest miarą niemianowaną i unormowaną o wartościach z przedziału <–1, 1> (poza przypadkami skrajnej asymetrii), co umożliwia porównywanie asymetrii różnych rozkładów. Dla szeregu symetrycznego $A_s = 0$.

Przykład 2.4

Szereg rozdzielczy przedstawiony w tab. 2.3 zawiera strukturę rocznych wydatków na ubezpieczenie emerytalne w grupie 100 losowo wybranych pracowników pewnej korporacji.

Tabela 2.3. Dane do przykładu 2.4

Numer klasy	Wydatki roczne w PLN	Liczba pracowników
1	0–100	15
2	100–500	30
3	500–1000	40
4	1000–2000	10
5	2000–5000	5
	Razem	100

Źródło: opracowanie własne.

Należy przeprowadzić pogłębioną analizę statystyczną badanej próby oraz określić, jakie miary statystyczne można zastosować w prowadzonych analizach i które z nich są najbardziej adekwatne.

Rozwiązanie

Rozwiązanie tego zadania sprowadza się do wyznaczenia omawianych wcześniej miar położenia, dyspersji i asymetrii. Tab. 2.4 zawiera dane z tab. 2.3 opatrzone oznaczeniami przyjętymi we wcześniej omawianych wzorach (2.4), (2.13) i (2.16), a w tab. 2.5–2.7 przedstawiono wyniki obliczeń pomocniczych, służących do wyznaczenia średniej arytmetycznej $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_i$ oraz odchylenia standardowego $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\dot{x}_i - \bar{x})^2 \cdot n_i}$ i przeciętnego $d = \frac{\sum_{i=1}^{n} |\dot{x}_i - \bar{x}| \cdot n_i}{n}$.

Średnie roczne wydatki na ubezpieczenie emerytalne (2.3), przypadające na jednego pracownika badanej grupy, wynoszą 722,5 PLN. Natomiast średnie odchylenie od średniej arytmetycznej wynosi 753,57 PLN, jeśli mierzone jest odchyleniem standardowym (2.13), i 455,25 PLN, jeśli zmienność mierzy się odchyleniem przeciętnym (2.16). W pierwszym przypadku współczynnik zmienności (2.19) wynosi 1,04 (104% wartości średniej), a w drugim (2.20) jest równy 0,63 (63%).

Tabela 2.4. Dane do przykładu 2.4

Lp.	Wydatki roczne w PLN x_i	Liczba pracowników n_i	Środki przedziałów $\dot{x}_i$	Rozpiętość przedziałów h_i
1	0–100	15	50	100
2	100–500	30	300	400

Tabela 2.4 (cd.)

Lp.	Wydatki roczne w PLN x_i	Liczba pracowników n_i	Środki przedziałów $\dot{x}_i$	Rozpiętość przedziałów h_i
3	500–1000	40	750	500
4	1 000–2000	10	1 500	1 000
5	2 000–5000	5	3 500	3 000
	Razem	100		

Źródło: opracowanie własne.

Tabela 2.5. Obliczenia do przykładu: miary klasyczne

Lp.	$\dot{x}_i$	n_i	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$	$\lvert\dot{x}_i - \bar{x}\rvert$	$\lvert\dot{x}_i - \bar{x}\rvert \cdot n_i$
1	50	15	750	–672,5	452 256,25	6 783 843,75	672,5	10 087,5
2	300	30	9 000	–422,5	178 506,25	5 355 187,50	422,5	12 675,0
3	750	40	30 000	27,5	756,25	30 250,00	27,5	1 100,0
4	1 500	10	15 000	777,5	604 506,25	6 045 062,50	777,5	7 775,0
5	3 500	5	17 500	2 777,5	7 714 506,25	38 572 531,25	2 777,5	13 887,5
Suma		100	72 250	×	×	56 786 875,00	×	45 525,0
Średnia			722,5	×	wariancja	567 868,75	×	×
Odchylenie					standardowe	753,57	przeciętne	455,25
Współczynniki zmienności						1,04		0,63

Źródło: opracowanie własne.

Tabela 2.6. Obliczenia do przykładu: miary częstości

Lp.	x_i	n_i	h_i	Liczebność skumulowana n_{ski}	Wskaźniki struktury		Wskaźniki struktury skumulowane	
					w_i	w_i [%]	w_{ski}	w_{ski} [%]
1	0–100	15	100	15	0,15	15	0,15	15
2	100–500	30	400	45	0,3	30	0,45	45
3	500–1000	40	500	85	0,4	40	0,85	85
4	1000–2000	10	1 000	95	0,1	10	0,95	95
5	2000–5000	5	3 000	100	0,05	5	1	100
Suma		100	×	×	1,00	100	×	×

Uwaga: kolorem żółtym oznaczono przedział, zawierający Q_1, kolorem zielonym przedział, w którym znajduje się dominanta i pozostałe dwa kwartyle, tj. mediana oraz Q_3.

Źródło: opracowanie własne.

Kolejnym krokiem jest analiza struktury (częstości) badanych pracowników z punktu widzenia kwot przeznaczonych na ubezpieczenie emerytalne. Wyznaczone zostaną zatem wskaźniki struktury (2.1). Z tab. 2.6 wynika, że najwięcej (tj. 40%) pracowników odkłada na ubezpieczenie emerytalne od 500 do 1000 PLN, a 5% pracowników kwoty należące do najwyższego przedziału. Niemal połowa badanych (a dokładnie 45%) przeznacza na przyszłą emeryturę do 500 PLN rocznie, o czym informuje skumulowany wskaźnik struktury w_{isk} (porównaj tab. 2.6).

Na podstawie tak skonstruowanego szeregu można jedynie wyznaczyć przedział dominanty [500; 1000) PLN, ale nie da się obliczyć dominanty ze wzoru interpolacyjnego (2.9), ponieważ rozpiętości tego przedziału i dwóch z nim sąsiadujących nie są równe, co jest warunkiem stosowalności tego wzoru. Natomiast można zastosować wzory interpolacyjne na obliczenie trzech kwartylów (2.6)–(2.8), co pokazano w tab. 2.7.

W celu wyznaczenia kwartylów należy określić przedziały, w jakich się one znajdują na podstawie skumulowanych liczebności. I tak kwartyl pierwszy (tj. obserwacja o numerze $N_1 = 25$) znajduje się w przedziale drugim, dla którego liczebność skumulowana n_{sk2} wynosi 45, a kwartyle drugi (tj. mediana, czyli obserwacja o numerze $N_2 = 50$) i trzeci (o numerze $N_3 = 75$) znajdują się w przedziale trzecim, gdyż liczebność skumulowana n_{sk3} wynosi 85. Obliczone wartości kwartylów oraz odchylenia ćwiartkowego (2.18) przedstawiono w tab. 2.7. Wynika z nich, że połowa badanych wydawała na zabezpieczenie emerytalne do 562,5 PLN, a połowa powyżej tej kwoty. 25% pracowników wydawało do 233,3 PLN rocznie, a 75% powyżej tej kwoty. Z kolei 25% badanych przeznaczało na ten cel ponad 875 PLN, a pozostali mniej niż podaną kwotę. Odchylenie ćwiartkowe informuje o tym, że przeciętnie zróżnicowanie wydatków na zabezpieczenie emerytalne wynosi 320,83 PLN wokół mediany.

Tabela 2.7. Obliczenia do przykładu: miary pozycyjne

Q	Numer kwartylu	Dolna granica	Rozpiętość	Liczebność	Liczebność skumulowana przedziału*/	Wartości miar
		przedziału kwartylu				
	N_Q	x_0	h_0	n_0		
1	25	100	400	30	15	233,33
2	50	500	500	40	45	562,50
3	75	500	500	40	45	875,00
Odchylenie ćwiartkowe			320,83	Współczynnik zmienności		0,57

Uwaga: */ Liczebność skumulowana przedziału poprzedzającego przedział kwartylu.

Źródło: opracowanie własne.

Miara asymetrii (2.22) wynosi:

$$A_s = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{(875 - 562{,}5) - (562{,}5 - 233{,}33)}{312{,}5 + 329{,}17} = \frac{-16{,}67}{641{,}67} = -0{,}026$$

Świadczy to o nieznacznej ujemnej (lewostronnej) asymetrii. Zatem do opisu tego szeregu można korzystać zarówno z miar klasycznych, jak i pozycyjnych, chociaż różnica między wartością środkową (*Me*) i średnią arytmetyczną jest znaczna, a zróżnicowanie obliczone dla miar pozycyjnych wynosi jedynie 57%, podczas gdy dla miar klasycznych 104% w przypadku odchylenia standardowego i 63% dla odchylenia przeciętnego. Zatem bardziej adekwatnymi w analizie badanego szeregu wydają się być miary pozycyjne niż klasyczne.

Współczynnik skośności wyznaczany przez pakiety komputerowe dany jest wzorem:

$$A = \frac{n}{(n-1)\cdot(n-2)} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^3}{\hat{S}^3} \tag{2.23}$$

gdzie:

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1} \cdot S^2 \tag{2.24}$$

pozostałe oznaczenia jak poprzednio.

Ocenę asymetrii uzupełnia tzw. wykres pudełkowy, zwany też wykresem skrzynkowym lub ramkowym (*box-plot*), którego przykład zamieszczono na rys. 2.5. Składa się on z prostokąta, którego wysokość wyznaczają pierwszy i trzeci kwartyl. Wewnątrz pudełka zaznacza się medianę. W rozkładzie symetrycznym mediana dzieli pudełko dokładnie na pół. Gdy asymetria jest prawostronna, odległość między trzecim kwartylem jest większa niż pomiędzy kwartylem pierwszym a medianą. W przypadku asymetrii lewostronnej sytuacja jest odwrotna. Dodatkowo na wykresie zaznacza się tzw. wąsy, które określają wartości:

$$Q_1 - 3Q \quad \text{jeśli} \quad x_{min} < Q_1 - 3Q \tag{2.25}$$

$$Q_3 + 3Q \quad \text{jeśli} \quad x_{max} > Q_3 + 3Q \tag{2.26}$$

Za obserwacje nietypowe przyjmuje się dane spełniające warunek:

$$x_i < Q_1 - 3Q \quad \text{lub} \quad x_i < Q_3 + 3Q \tag{2.27}$$

gdzie oznaczenia jak poprzednio.

Miarą koncentracji jest kurtoza, która charakteryzuje względną spiczastość lub płaskość rozkładu w porównaniu z rozkładem normalnym. Dodatnia kurtoza oznacza rozkład o stosunkowo dużej spiczastości. Ujemna kurtoza oznacza rozkład stosunkowo płaski. Klasyczny współczynnik kurtozy obliczany jest według wzoru:

$$K = \frac{n \cdot (n+1)}{(n-1) \cdot (n-2) \cdot (n-3)} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^4}{\hat{S}^4} - \frac{3 \cdot (n-1)^2}{(n-2) \cdot (n-3)} \tag{2.28}$$

Kurioza równa 0 oznacza rozkład normalny. Zatem rozkład o kurtozie większej od 0 jest rozkładem bardziej wysmukłym niż rozkład normalny, tzn. charakteryzuje się większym skupieniem jednostek wokół średniej. Natomiast rozkład o kurtozie mniejszej od 0 jest rozkładem bardziej spłaszczonym od rozkładu normalnego, czyli charakteryzuje się znacznym rozproszeniem jednostek wokół średniej. Pierwszy z nich nazywany jest rozkładem leptokurtycznym, a drugi – platykurtycznym.

W literaturze przedmiotu podstawowe parametry opisowe zbiorowości określane są mianem momentów statystycznych, wśród których:

- średnia arytmetyczna (2.2) to moment zwykły I rzędu,
- wariancja (2.24) to moment centralny II rzędu,
- skośność (2.23) to moment centralny III rzędu,
- kurtoza (2.28) to moment centralny IV rzędu.

Przykład 2.5

Przeprowadzono badania ankietowe 494 punktów sprzedaży detalicznej i hurtowej produktów chemii gospodarczej i spożywczej w Łodzi. Jedną z uwzględnionych cech jest powierzchnia sprzedaży. Zebrane dane przedstawiono w tab. 2.8. Należy przeprowadzić analizę tych danych.

Tabela 2.8. Analiza powierzchni sprzedaży punktów handlowych

Nr przedziału klasowego k	Powierzchnia [m²] x_i	Liczba przedsiębiorstw n_i	Liczebność skumulowana n_{ski}	Wskaźniki struktury w_i	Wskaźniki struktury skumulowane w_{ski}
1	0–25	93	93	18,8	18,8
2	25–50	131	224	26,5	45,3
3	50–75	105	329	21,3	66,6
4	75–100	39	368	7,9	74,5
5	100–125	47	415	9,5	84,0
6	125–150	5	420	1,0	85,0
7	150–175	12	432	2,4	87,4

Tabela 2.8 (cd.)

Nr przedziału klasowego k	Powierzchnia [m²] x_i	Liczba przedsiębiorstw n_i	Liczebność skumulowana n_{ski}	Wskaźniki struktury w_i	Wskaźniki struktury skumulowane w_{ski}
8	175–200	5	437	1,0	88,4
9	200–400	31	468	6,3	94,7
10	400–600	15	483	3,0	97,7
11	600–800	11	494	2,2	99,9
	Suma	494	×	99,9 ≈ 100	×

Uwaga: kolorem żółtym oznaczono przedział, w którym znajduje się dominanta i Q_1, kolorem niebieskim przedział, gdzie znajduje się mediana, a zielonym – przedział zawierający Q_3.

Źródło: opracowanie własne na podstawie badań ankietowych [Witkowska, Witkowski 1997a].

Rozwiązanie

Dla danych zamieszczonych w tab. 2.8 wyznaczyć należy miary pozycyjne. Z analizy tej tabeli wynika, że pierwsze 8 przedziałów klasowych ma jednakową rozpiętość wynoszącą 25 m². Dominanta znajduje się w drugim przedziale, natomiast przedziały poszczególnych kwartylów to odpowiednio drugi (N_{Q1} = 124), trzeci (N_{Q2} = 247) i piąty (N_{Q3} = 371). Jako pierwszą należy obliczyć dominantę, korzystając ze wzoru interpolacyjnego (2.9):

$$Do = x_0 + \frac{n_0 - n_{-1}}{(n_0 - n_{-1}) + (n_0 - n_{+1})} \cdot h_0 = 25 + \frac{131 - 93}{(131 - 93) + (131 - 105)} \cdot 25 = 39{,}84$$

Z kolei mediana obliczona ze wzoru (2.6) to:

$$Me = Q_2 = x_0 + \frac{h_0}{n_0}(N_{Me} - n_{sk-1}) = 50 + \frac{25}{105} \cdot (247 - 224) = 55{,}48$$

a pozostałe kwartyle wyznacza się ze wzorów (2.7)–(2.8):

$$Q_1 = x_0 + \frac{h_0}{n_0}(N_{Q_1} - n_{sk-1}) = 25 + \frac{25}{131} \cdot (124 - 93) = 30{,}92$$

$$Q_3 = x_0 + \frac{h_0}{n_0}(N_{Q_3} - n_{sk-1}) = 100 + \frac{25}{47} \cdot (371 - 368) = 101{,}60$$

Stąd odchylenie ćwiartkowe obliczone ze wzoru (2.18):

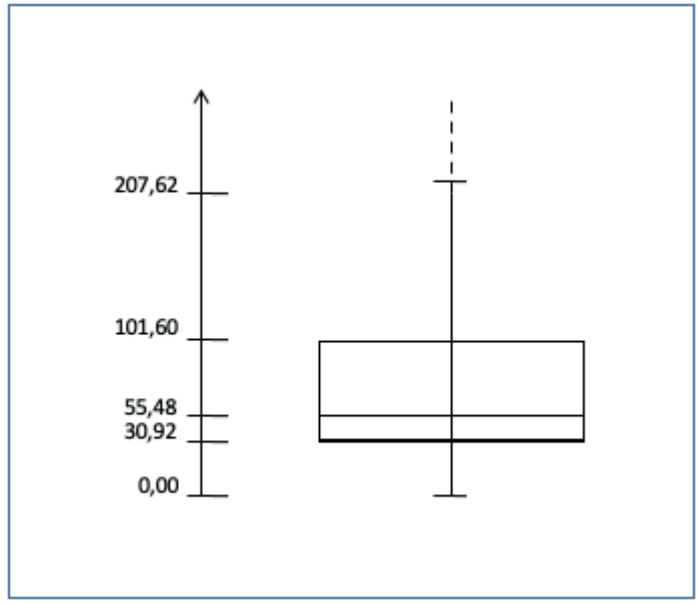
$$Q = \frac{Q_3 - Q_1}{2} = \frac{101{,}60 - 30{,}92}{2} = 35{,}34$$

oraz współczynnik zmienności (2.21):

$$V_Q = \frac{Q}{Me} = \frac{35,34}{55,48} = 0,64$$

W podsumowaniu tej części obliczeń zauważyć należy, że minimalna wartość szeregu to 0, a maksymalna 800. W omawianym badaniu, najczęściej spotyka się punkty sprzedaży o powierzchni 39,84 m^2. Jedna czwarta wszystkich badanych miała powierzchnię do 30,92 m^2, a połowa do 55,48 m^2, natomiast w przypadku 75% punktów sprzedaży ich powierzchnia nie przekraczała 101,6 m^2. Odchylenie ćwiartkowe wynosi 35,34 m^2 wokół mediany i wynosi 64% jej wartości.

Z kolei asymetrię rozkładu można ocenić przy wykorzystaniu wykresu pudełkowego. W celu wyznaczenia długości wąsów i stwierdzenia występowania obserwacji nietypowych, oblicza się warunki (2.26) i (2.27). Pierwszy z wymienionych warunków wskazuje wartość $x_{min} = 0 > Q_1 - 3Q = -75,1$, dlatego na wykresie zaznacza się minimalną wartość szeregu obserwacji tzn. 0. Natomiast z drugiego otrzymuje się: $x_{max} = 800 > Q_3 + 3Q = 207,62$, zatem na wykresie znajdzie się wartość 207,62 i wszystkie wartości obserwacji większe od tej liczby można traktować jako jednostki nietypowe. Wykres pudełkowy przedstawiono na rys. 2.5. Wynika z niego wyraźna asymetria prawostronna (dodatnia), co oznacza znaczną przewagę punktów sprzedaży o mniejszej powierzchni.



Rysunek 2.5. Wykres pudełkowy dla danych z tab. 2.8

Źródło: opracowanie własne.

Wprawdzie wykazano, że szereg z tab. 2.8 jest asymetryczny i adekwatnymi miarami do jego opisu są miary pozycyjne, to wyznaczono dla niego również miary klasyczne w celu ilustracji przedstawionych wcześniej rozważań. Stosowne obliczenia wykonane w Excelu zamieszczono w tab. 2.9.

Stosując wzór (2.4) na średnią arytmetyczną, otrzymuje się jej wartość: $\bar{x}$ = 49412,5/494 = 100,0253, która jest niemal równa kwartylowi trzeciemu, zatem znacząco większa od mediany i dominanty. Obliczone odchylenie standardowe (2.13) wynosi: $S = \sqrt{(281604042 / 494)} = \sqrt{570048,669} = 755,0157$ i stąd współczynnik zmienności (2.19) wyniesie V_S = 755,0157/100,0253 = 7,55. Jak zatem widać, miary klasyczne pokazują znacznie większe zróżnicowanie powierzchni wokół średniej arytmetycznej niż pozycyjne (wartości współczynników zmienności wynoszą odpowiednio 7,55 i 0,64). Świadczy to o asymetrii szeregu i jednoznacznie wskazuje, że do opisu analizowanego szeregu adekwatne są jedynie miary pozycyjne (innymi słowy w tym przypadku nie należy korzystać z miar klasycznych).

Tabela 2.9. Obliczenia miar klasycznych

Nr przedziału klasowego k	Środki przedziałów klasowych $\dot{x}_i$	Liczba przedsiębiorstw n_i	$\dot{x}_i \cdot n_i$	$(\dot{x}_i - \bar{x}) \cdot n_i$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
1	12,5	93	1 162,5	–8 139,8532	712 443,126
2	37,5	131	4 912,5	–8 190,8148	875 708,645
3	62,5	105	6 562,5	–3 940,1569	463 280,828
4	87,5	39	3 412,5	–488,4868	57 733,029
5	112,5	47	5 287,5	586,3107	64 581,490
6	137,5	5	687,5	187,3735	589 828,204
7	162,5	12	1 950,0	749,6964	1 686 133,880
8	187,5	5	937,5	437,3735	3 343 846,430
9	300,0	31	9 300,0	6 199,2156	18 715 263,500
10	500,0	15	7 500,0	5 999,6205	77 270 223,000
11	700,0	11	7 700,0	6 599,7217	177 825 000,000
Suma		494	49412,5	0	281 604 042,392

Źródło: opracowanie własne na podstawie tab. 2.8.

Przykład 2.6

Jak zmieni się opis szeregu zamieszczonego w tab. 2.8, jeśli przedziały klasowe zostaną zagregowane, tak jak przedstawiono w tab. 2.10?

Tabela 2.10. Analiza powierzchni sprzedaży punktów handlowych

Numer przedziału klasowego k	Powierzchnia [m²] x_i	Liczba przedsiębiorstw n_i	Liczebność skumulowana n_{ski}
1	0 – 50	224	224
2	50 –100	144	368
3	100 –150	52	420
4	150 –200	17	437
5	200 –800	57	494
Suma	×	494	×

Źródło: opracowanie własne na podstawie tab. 2.8.

Rozwiązanie

Numery poszczególnych kwartylów pozostają bez zmian i wynoszą: $N_{Q1} = 124$, $N_{Q2} = 247$ i $N_{Q3} = 371$. Jednakże znajdują się one w pierwszych trzech przedziałach klasowych, a ich wartości wyznaczone ze wzorów wynoszą:

$$Me = Q_2 = 50 + \frac{50}{144} \cdot (247 - 224) = 57,99$$

$$Q_1 = 0 + \frac{50}{224} \cdot (124 - 0) = 27,68$$

$$Q_3 = 100 + \frac{50}{52} \cdot (371 - 368) = 102,88$$

Zatem wartość odchylenia ćwiartkowego oblicza się jako (2.18):

$$Q = \frac{102,88 - 27,68}{2} = 37,60$$

a współczynnik zmienności (2.21):

$$V_Q = \frac{37,60}{57,99} = 0,65$$

Jak zatem widać, miary pozycyjne obliczone dla szeregu zawartego w tab. 2.10 różnią się nieznacznie od tych wyznaczonych dla szeregu z tab. 2.8. Należy zatem obliczyć miary klasyczne w tab. 2.11.

Na podstawie wykonanych obliczeń uzyskana zostanie wartość średniej arytmetycznej: $\overline{x} = 54375/494 = 110,0709$, która jest większa od kwartylu trzeciego. Obliczone odchylenie standardowe wynosi: $S = \sqrt{(4401076943 / 494)} = \sqrt{8909062} = 2984,8053$,

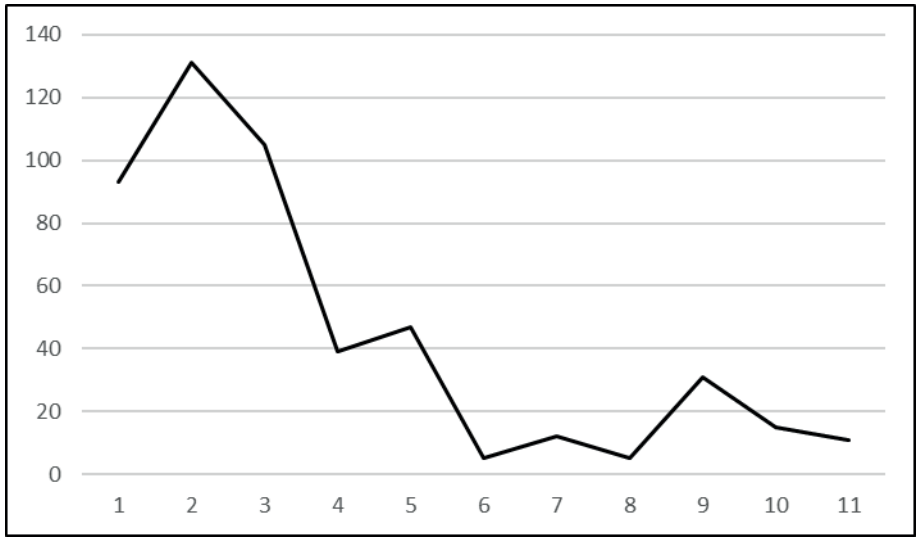
stąd współczynnik zmienności wyniesie $V_S = 2984{,}8053/110{,}0709 = 27{,}12$. Oznacza to, że agregacja przedziałów klasowych spowodowała znaczące zmiany w wyznaczonych miarach klasycznych.

Tabela 2.11. Analiza powierzchni sprzedaży punktów handlowych

Nr przedziału klasowego k	Środki przedziałów klasowych $\dot{x}_i$	Liczba przedsiębiorstw n_i	$\dot{x}_i \cdot n_i$	$(\dot{x}_i - \bar{x}) \cdot n_i$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
1	25	224	5 600	–19 055,87	40 527 477,5
2	75	144	10 800	–5 050,20	13 283 617,0
3	125	52	6 500	776,32	1 448 716,8
4	175	17	2 975	1 103,80	12 541 988,6
5	500	57	28 500	22 225,96	4 333 275 143,1
	Suma	494	54 375	0	4 401 076 943,0

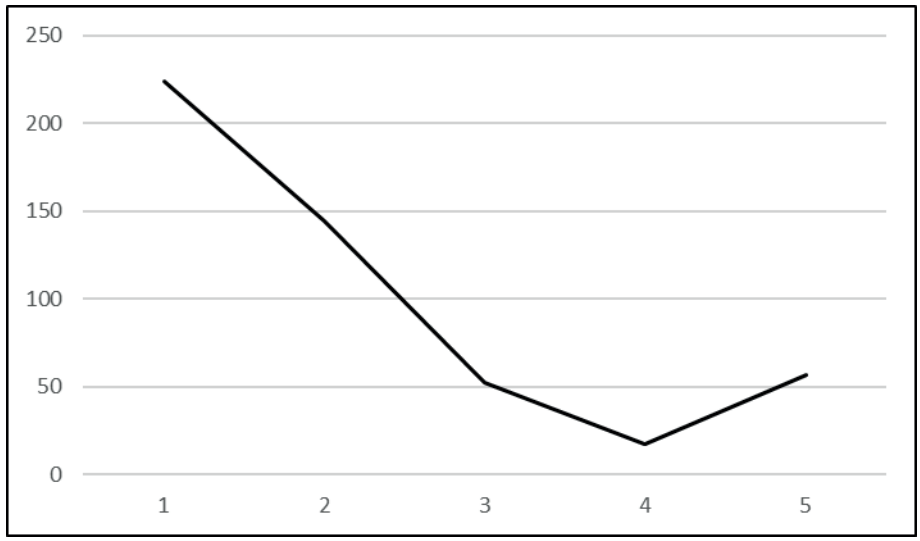
Źródło: opracowanie własne na podstawie tab. 2.10.

W celu zilustrowania, jak zmienił się rozkład empiryczny w szeregu po agregacji wartości cech w przedziałach klasowych, na rys. 2.6. i 2.7 przedstawiono szeregi zapisane w tab. 2.8. i 2.10. Jak widać na prezentowanych wykresach, agregacja danych w znaczący sposób zmieniła rozkład danych.



Rysunek 2.6. Szereg przed agregacją

Źródło: opracowanie własne na podstawie tab. 2.8.



Rysunek 2.7. Szereg po agregacji

Źródło: opracowanie własne na podstawie tab. 2.10.

Rozdział 3

Metody wnioskowania statystycznego

Analiza zjawisk gospodarczych najczęściej prowadzona jest na podstawie badań częściowych. W tym przypadku wartości podstawowych miar można obliczyć jedynie dla jednostek wchodzących w skład próby statystycznej i na tej podstawie można ewentualnie wnioskować o rozkładzie i podstawowych charakterystykach zbiorowości generalnej, wykorzystując w tym celu metody wnioskowania statystycznego, tj.:

- estymację nieznanych parametrów zbiorowości generalnej,
- weryfikację hipotez statystycznych.

Warto przy tym podkreślić, że uogólnianie wyników uzyskanych dla próby statystycznej na całą zbiorowość możliwe jest na podstawie odpowiednio dobranej próby losowej. Przyjmuje się, że próba została wylosowana w sposób nieograniczony i niezależny. Zakłada się też, że jest to próba losowa o strukturze odpowiadającej strukturze populacji.

W poprzednich rozdziałach omówione zostały podstawowe parametry, charakteryzujące zbiorowość statystyczną, takie jak: miary średnie, dyspersji asymetrii, koncentracji i wskaźniki struktury. Niniejszy rozdział poświęcony zostanie metodom wnioskowania statystycznego w odniesieniu do tych parametrów. Podstawową zaletą wnioskowania statystycznego jest możliwość określenia wielkości błędu, z jakim formułowane są wnioski ogólne, czego nie uzyskuje się w przypadku korzystania z innych metod.

3.1. Podstawowe pojęcia związane z estymacją nieznanych parametrów

Parametrem zbiorowości generalnej, oznaczanym jako θ, nazywa się miarę opisową, której wartość nie jest znana. Estymatorem nazywa się statystykę T_n służącą do oszacowania nieznanej wartości parametru w populacji generalnej θ. Inaczej

mówiąc, estymatorem jest zmienna losowa, której rozkład zależy od rozkładu szacowanego parametru. Przez rozkład estymatora rozumie się rozkład prawdopodobieństwa zmiennej losowej T_n, którego parametrami są wartość oczekiwana $E(T_n)$ oraz wariancja $D^2(T_n)$. Konkretną wartość liczbową t_n, jaką przyjmuje estymator T_n (parametru θ) dla określonej próby, nazywa się oceną estymatora parametru θ. Oznacza to, że, wartość t_n, będąca oceną estymatora parametru θ, jest realizacją zmiennej losowej T_n.

W przypadku badania częściowego wartości parametrów zbiorowości generalnej są szacowane (w przeciwieństwie do badania pełnego, kiedy są one dokładnie obliczane) z pewnym błędem. Standardowy błąd szacunku jest odchyleniem standardowym, czyli pierwiastkiem kwadratowym z wariancji $D^2(T_n)$ rozkładu estymatora T_n, za pomocą którego szacuje się parametr θ populacji generalnej. W celu uzyskania możliwie dużej precyzji szacunku od estymatora wymaga się, aby posiadał pewne własności, którymi są między innymi: nieobciążoność, zgodność i efektywność.

Estymator T_n parametru θ jest zgodny, jeśli jest on zbieżny według prawdopodobieństwa do szacowanego parametru θ, tzn. dla każdego $\varepsilon > 0$ i $n \to \infty$ zachodzi:

$$\lim P\{|T_n - \theta| > \varepsilon\} = 0 \tag{3.1a}$$

co jest równoważne z zapisem:

$$\lim P\{|T_n - \theta| < \varepsilon\} = 1 \tag{3.1b}$$

gdzie:
P – symbol prawdopodobieństwa,
T_n – estymator,
θ – nieznany parametr,
ε – dowolnie mała liczba,
n – liczebność próby,
$|\vdots|$ – oznaczenie wartości bezwzględnej.

W przypadku estymatora zgodnego, tj. spełniającego warunek (3.1) dla dostatecznie licznej próby, szansa otrzymania oceny estymatora różnego od parametru jest bliska 0. Innymi słowy, dla estymatora zgodnego wzrost liczebności próby daje dokładniejsze dopasowanie do rzeczywistej wartości nieznanego parametru.

Estymator T_n jest nieobciążonym estymatorem parametru θ, jeśli jego wartość oczekiwana równa jest szacowanemu parametrowi, czyli jest spełniony warunek:

$$E(T_n) = \theta \tag{3.2}$$

Jeśli warunek (3.2) nie zachodzi, to taki estymator nazywany jest obciążonym, a różnicę:

$$o(T_n) = E(T_n) - \theta \tag{3.3}$$

nazywa się obciążeniem estymatora.

Estymatorem najbardziej efektywnym nazywa się ten spośród nieobciążonych estymatorów parametru θ, który ma najmniejszą wariancję $D^2(T_n)$. Stosowanie estymatora najbardziej efektywnego oznacza, że w trakcie estymacji popełniania się najmniejszy błąd szacunku.

Korzystanie z estymatora T_n posiadającego opisane wyżej własności pozwala najlepiej oszacować nieznany parametr θ, ponieważ z dużym prawdopodobieństwem można przyjąć, że wyznaczona ocena estymatora T_n jest bliska rzeczywistej wartości θ. W literaturze i praktyce wyróżnia się dwa rodzaje estymacji:

- estymację punktową,
- estymację przedziałową.

W pierwszym przypadku za nieznaną wartość parametru zbiorowości generalnej przyjmuje się konkretną wartość oceny estymatora wyznaczonego na podstawie n-elementowej próby. Innymi słowy, wartość oczekiwaną szacuje się za pomocą średniej arytmetycznej z próby, wariancję i odchylenie standardowe zbiorowości za pomocą miar dyspersji wyznaczonych dla próby, a frakcję populacji, korzystając ze wskaźnika struktury obliczonego dla danych z próby. Warto przy tym nadmienić, że średnia arytmetyczna z próby jest zgodnym, nieobciążonym i najbardziej efektywnym estymatorem nadziei matematycznej w zbiorowości. Natomiast odchylenie standardowe (zatem i wariancja) wyznaczone ze wzorów (2.11)–(2.13) jest estymatorem obciążonym. Aby uzyskać estymator nieobciążony, korzysta się ze wzoru:

$$\hat{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.4}$$

dla szeregu prostego. Kiedy dane są pogrupowane w przedziały klasowe i korzysta się z szeregów rozdzielczych, to należy wzory (2.12)–(2.13) odpowiednio zmodyfikować w celu uzyskania nieobciążonego estymatora odchylenia standardowego. Warto dodać, że wyznaczony z próby współczynnik skośności według wzoru (2.23) i kurtoza ze wzoru (2.28) są nieobciążonymi estymatorami miar asymetrii i skupienia w populacji generalnej.

W przypadku estymacji przedziałowej wyznacza się przedział liczbowy, który z ustalonym prawdopodobieństwem zawiera nieznaną wartość szacowanego parametru zbiorowości generalnej. Prawdopodobieństwo to nosi nazwę współczynnika ufności i oznaczane jest jako $1 - \alpha$, a znaleziony przedział nazywany jest przedziałem ufności. Informuje on o tym, w jakich granicach należy spodziewać

się wartości poszukiwanego parametru θ z zadanym z góry prawdopodobieństwem. Ogólną postać przedziału ufności można zapisać jako:

$$P\{f_d(T_n) < \theta < f_g(T_n)\} = 1 - \alpha \tag{3.5}$$

gdzie:
P – symbol prawdopodobieństwa,
$1 - \alpha$ – poziom ufności,
T_n – estymator,
θ – szacowany parametr,
$f_d(T_n), f_g(T_n)$ – dolna i górna granica przedziału ufności.

Wyższość estymacji przedziałowej nad punktową polega na tym, że przedział ufności pozwala określić prawdopodobieństwo, z jakim wartości nieznanego parametru znajdą się w oszacowanym przedziale, czego nie daje estymacja punktowa. Należy przy tym pamiętać, że:

- przy zadanym poziomie ufności $1 - \alpha$, im większa jest liczebność, tym krótszy przedział ufności,
- przy ustalonej liczebności próby wraz ze wzrostem poziomu ufności rośnie rozpiętość (długość) przedziału ufności,
- im krótszy przedział ufności, tym mniejszy maksymalny błąd szacunku, co oznacza większą dokładność oszacowania.

Długość przedziału ufności oznaczana jest przez $2d$, a więc:

$$2d = f_g(T_n) - f_d(T_n) \tag{3.6}$$

gdzie:
d – maksymalny błąd szacunku.

Praktyczna użyteczność wyznaczonych przedziałów ufności zależy od maksymalnego błędu szacunku, a więc połowy długości tegoż przedziału. Z kolei długość przedziału jest funkcją współczynnika ufności $1 - \alpha$, który w badaniach społecznych zazwyczaj ustala się na poziomie 0,95, oraz liczebności próby n. Zatem w celu zapewnienia odpowiedniej dokładności estymacji, przy zadanym poziomie ufności, istnieje konieczność obliczania niezbędnej liczebności próby n dla przedziałów ufności konstruowanych dla wybranych parametrów.

Podsumowując, proces estymacji przedziałowej nieznanych parametrów można opisać za pomocą kilku kroków:

1) określenie rodzaju szacowanego parametru;
2) wybór estymatora i określenie jego rozkładu prawdopodobieństwa;
3) ustalenie współczynnika ufności oraz wartości maksymalnego błędu szacunku;
4) wyznaczenie minimalnej liczebności próby;
5) wyznaczenie ocen estymatora;
6) budowa przedziału ufności.

3.2. Estymacja podstawowych parametrów opisowych zbiorowości

Podstawowymi parametrami, szacowanymi dla populacji generalnej, są: wartość oczekiwana (średnia) $E(X) = \mu$, wariancja $D^2(X) = \sigma^2$, odchylenie standardowe σ oraz frakcja (wskaźnik struktury) p. W praktyce stosuje się wiele sposobów estymacji konkretnych parametrów, co jest związane przede wszystkim z rodzajem posiadanej próby, informacją dotyczącą rozkładu zbiorowości generalnej, a także przyjętym stopniem dokładności szacunku.

Najczęściej szacowanym parametrem populacji generalnej jest wartość oczekiwana badanej cechy mierzalnej. Najlepszym estymatorem nadziei matematycznej μ w populacji generalnej jest średnia arytmetyczna z próby $\bar{x}$, którą wyznacza się na podstawie jednego ze wzorów (2.2)–(2.4). Chociaż w wielu przypadkach praktycznych zastosowań wykorzystuje się wyniki estymacji punktowej, to bezpieczniejszą metodą jest stosowanie estymacji przedziałowej.

Konstrukcja przedziału ufności jest ściśle związana z rozkładem estymatora. W przypadku nadziei matematycznej, w zależności od przyjętych założeń, jest to rozkład normalny lub t-Studenta. Wyróżnia się tutaj 3 przypadki:

1) Populacja generalna ma rozkład normalny i znane jest jego zróżnicowanie mierzone odchyleniem standardowym (σ). Jest to sytuacja niezmiernie rzadka, która pozwala na wykorzystanie małej próby[1], a rozkład estymatora również ma rozkład normalny.
2) Dana jest mała próba losowa o nieznanej wariancji populacji generalnej. Jeśli przyjąć, że rozkład cechy w populacji jest normalny, to estymator wartości oczekiwanej ma rozkład t-Studenta.
3) Brak jest wiedzy o rozkładzie cechy w populacji, ale dana jest duża próba losowa, wówczas estymator nadziei matematycznej ma rozkład normalny.

Przedział ufności dla wartości oczekiwanej w przypadku małej (tj. do 30 elementów) próby losowej, pochodzącej z populacji generalnej o rozkładzie normalnym z nieznaną wariancją (tj. przypadek 2 z ww.) wyraża się wzorem:

$$P\left\{\bar{x} - t_\alpha \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_\alpha \frac{S}{\sqrt{n-1}}\right\} = 1 - \alpha \tag{3.7}$$

gdzie:
$\bar{x}$ – średnia wartość cechy pochodząca z n-elementowej próby, wyznaczona na podstawie jednego ze wzorów (2.2)–(2.4),
S – odchylenie standardowe z próby, wyznaczone z jednego ze wzorów (2.11)–(2.13),

1 Pojęcie małej i dużej próby jest umowne i zależy między innymi od celu badania i szacowanego parametru. W niektórych badaniach małą próbą jest próba do 30 elementów, w innych do 60 lub więcej elementów.

n – liczebność próby,
t_α – wartość krytyczna wyznaczona dla $(n - 1)$ stopni swobody z rozkładu t-Studenta dla zadanego poziomu ufności $1 - \alpha$ w taki sposób, aby spełniony był warunek:

$$P(|t| \leq t_\alpha) = 1 - \alpha \tag{3.8}$$

Zatem przedział ufności jest funkcją oceny estymatora punktowego $\bar{x}$ i maksymalnego błędu szacunku $t_\alpha \frac{S}{\sqrt{n-1}}$. Jest to przedział symetryczny, jego długość bowiem wynosi $2 \cdot t_\alpha \frac{S}{\sqrt{n-1}}$. Interpretując informacje przedstawione wzorem (3.7), z prawdopodobieństwem równym poziomowi ufności $(1 - \alpha)$, stwierdza się, że poszukiwana wartość oczekiwana (oznaczana przez $E(X)$ lub μ) znajdzie się w przedziale otwartym:

$$\left(\bar{x} - t_\alpha \frac{S}{\sqrt{n-1}}, \bar{x} + t_\alpha \frac{S}{\sqrt{n-1}} \right) \tag{3.9}$$

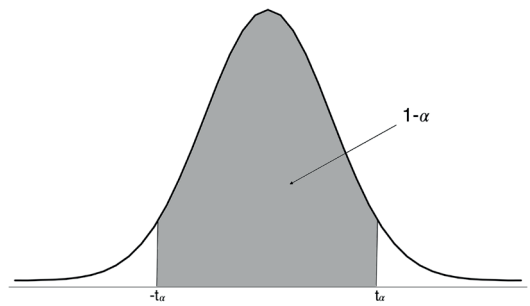
a maksymalny błąd szacunku wynosi:

$$d = t_\alpha \frac{S}{\sqrt{n-1}} \tag{3.10}$$

W sytuacji gdy korzysta się z nieobciążonego estymatora wariancji (3.4.), wzór (3.7) jest postaci:

$$P\left\{ \bar{x} - t_\alpha \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{x} + t_\alpha \frac{\hat{S}}{\sqrt{n}} \right\} = 1 - \alpha \tag{3.7a}$$

Graficzną prezentację przedziału ufności zbudowanego na podstawie rozkładu t-Studenta dla poziomu ufności $(1 - \alpha)$ przedstawiono na rys. 3.1.



Rysunek 3.1. Przedział ufności zbudowany na podstawie rozkładu t-Studenta dla poziomu ufności 1 – α

Źródło: opracowanie własne.

W przypadku dużej próby i braku informacji o rozkładzie cechy (ww. przypadek 3) przedział ufności dla nadziei matematycznej jest postaci:

$$P\left\{\bar{x}-u_{\alpha}\frac{S}{\sqrt{n}}<\mu<\bar{x}+u_{\alpha}\frac{S}{\sqrt{n}}\right\}=1-\alpha \tag{3.11}$$

gdzie:
u_{α} – wartość krytyczna wyznaczona z rozkładu Gaussa dla zadanego poziomu ufności 1 – α aby spełniony był warunek:

$$P(|u| \leq u_{\alpha}) = 1 - \alpha \tag{3.12}$$

Zatem z tablic dystrybuanty standaryzowanego rozkładu normalnego odczytuje się taką wartość, która spełnia równanie:

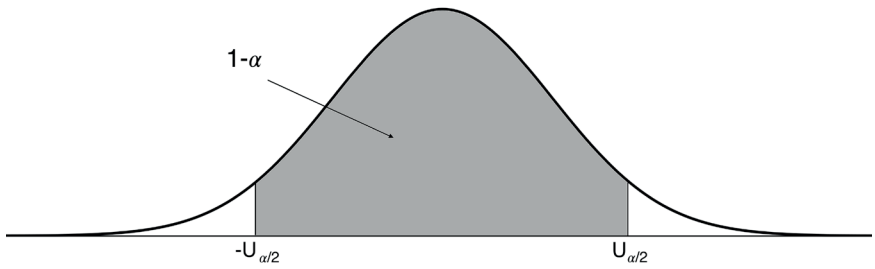
$$\Phi(u_{\alpha})=1-\frac{\alpha}{2} \tag{3.13}$$

gdzie:
$\Phi(u_{\alpha})$ – dystrybuanta rozkładu normalnego standaryzowanego,
pozostałe oznaczenia jak poprzednio.

Innymi słowy, z prawdopodobieństwem (1 – α) można twierdzić, że:

$$E(X)=\mu\in\left(\bar{x}-u_{\alpha}\frac{S}{\sqrt{n}},\bar{x}+u_{\alpha}\frac{S}{\sqrt{n}}\right) \tag{3.14}$$

co przedstawiono na rys. 3.2.



Rysunek 3.2. Przedział ufności zbudowano na podstawie standaryzowanego rozkładu normalnego dla poziomu ufności 1 – α

Źródło: opracowanie własne.

Natomiast dla opisanego wyżej przypadku pierwszego[2], kiedy znane jest odchylenie standardowe populacji, przedział ufności wyznacza się dla estymatora o rozkładzie normalnym w podobny sposób jak (3.11) z tą różnicą, że w miejsce odchylenie standardowego z próby wstawiona zostaje wartość odchylenie standardowego badanej zbiorowości σ, czyli:

$$P\left\{\bar{x}-u_{\alpha}\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+u_{\alpha}\frac{\sigma}{\sqrt{n}}\right\}=1-\alpha \tag{3.15}$$

gdzie:
σ – odchylenie standardowe w populacji,
a pozostałe oznaczenia jak poprzednio.

Ważnymi parametrami opisowymi są miary dyspersji, a wśród nich wariancja oznaczana jako $D^2(X)$ lub σ^2 i odchylenie standardowe σ. W przypadku małej (tj. do 60 elementów) próby losowej pochodzącej z populacji generalnej o rozkładzie normalnym, przedział ufności dla wariancji wyznacza się ze wzoru:

$$P\left\{\frac{nS^2}{\chi_1^2}<\sigma^2<\frac{nS^2}{\chi_2^2}\right\}=1-\alpha \tag{3.16}$$

gdzie:
χ_1^2, χ_2^2 – wartości krytyczne wyznaczone z rozkładu chi-kwadrat (χ_2^2) odczytane dla (n – 1) stopni swobody i zadanego poziomu ufności 1 – α, aby spełnione były równości:

$$P\left(\chi^2\geq\chi_1^2\right)=\frac{\alpha}{2} \tag{3.17a}$$

oraz

$$P\left(\chi^2\leq\chi_2^2\right)=\frac{\alpha}{2} \tag{3.17b}$$

pozostałe oznaczenia jak poprzednio.

Zatem z prawdopodobieństwem równym współczynnikowi ufności przyjmuje się, że wariancja znajdować się będzie w przedziale:

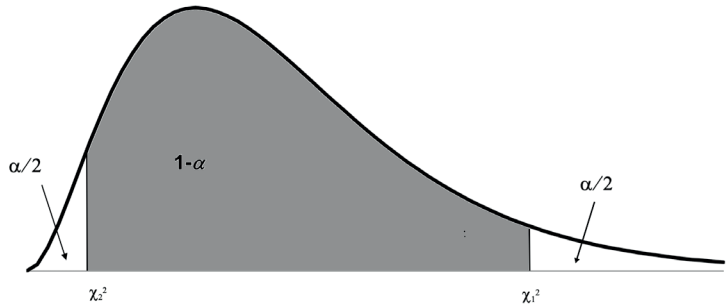
$$\sigma^2\in\left(\frac{nS^2}{\chi_1^2},\frac{nS^2}{\chi_2^2}\right) \tag{3.18a}$$

2 Przypadek ten rzadko pojawia się w praktyce, ponieważ wartość odchylenia standardowego populacji może być znana jedynie w wyniku wcześniejszego przeprowadzenia badania pełnego lub estymacji tego parametru w trakcie innego badania.

lub równoważnie stosując nieobciążony estymator wariancji:

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi_1^2}, \frac{(n-1)\hat{S}^2}{\chi_2^2} \right) \tag{3.18b}$$

Jak pokazano, w przypadku budowy przedziału ufności dla wariancji stosuje się statystykę o rozkładzie chi-kwadrat i sam przedział jest niesymetryczny, co widać na rys. 3.3.



Rysunek 3.3. Przedział ufności zbudowany na podstawie rozkładu chi-kwadrat i poziomu istotności 1 – α

Źródło: opracowanie własne.

Dla dużych prób wyznacza się przedział ufności dla odchylenia standardowego postaci:

$$P\left\{ \frac{S}{1+\frac{u_\alpha}{\sqrt{2n}}} < \sigma < \frac{S}{1-\frac{u_\alpha}{\sqrt{2n}}} \right\} = 1-\alpha \tag{3.19}$$

Innymi słowy w tym przypadku graficzna ilustracja przedziału ufności jest identyczna jak na rys. 3.2. Natomiast dla przyjętego poziomu ufności (1 – α) odchylenie standardowe znajdzie się w przedziale:

$$\sigma \in \left(\frac{S}{1+\frac{u_\alpha}{\sqrt{2n}}}, \frac{S}{1-\frac{u_\alpha}{\sqrt{2n}}} \right) \tag{3.20}$$

Przykład 3.1

Kurs akcji pewnej spółki zmienia się dość często. Na podstawie obserwacji 26 losowo wybranych dziennych kursów zamknięcia w ostatnim kwartale okazało się, że przeciętnie kształtuje się on na poziomie 120 PLN, a współczynnik zmienności wynosi 8%. Przyjmując 95% wiarygodność, trzeba określić, jakiej wielkości kursu akcji można się spodziewać w kolejnym miesiącu, jeśli założy się, że warunki rynkowe będą identyczne jak w minionych 3 miesiącach.

Rozwiązanie

Dane: n = 26, $\bar{x} = 120$, $V = \frac{S}{\bar{x}} \cdot 100\%$, czyli: $S = \frac{V \cdot \bar{x}}{100} = 9{,}6$, a wartość krytyczna odczytana z tablic rozkładu t-Studenta dla $\alpha = 0{,}95$ równa się $t_\alpha = 2{,}06$. Przy założeniu normalnego rozkładu cen akcji, do oszacowania cen w kolejnym miesiącu wykorzystuje się formułę (3.9)

$$P\left\{\bar{x} - t_\alpha \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_\alpha \frac{S}{\sqrt{n-1}}\right\} = 1 - \alpha,$$

otrzymując:

$$P\left\{120 - 2{,}06 \cdot \frac{9{,}6}{\sqrt{25}} < \mu < 120 + 3{,}96\right\} = P\{116{,}04 < \mu < 123{,}96\} = 0{,}95$$

Stąd z prawdopodobieństwem 0,95 można uznać, że kurs akcji będzie w przedziale (116,04; 123,96) PLN.

W badaniach statystycznych, zwłaszcza kiedy dotyczą cech jakościowych, interesujące jest stwierdzenie, jaki udział jednostek populacji generalnej charakteryzuje się określonym wariantem badanej cechy. Służy temu estymacja wskaźnika struktury (frakcji). W tym przypadku wymagana jest duża (przynajmniej 100-elementowa) próba losowa, gdyż zmienna losowa ma rozkład dwupunktowy (jak na rys. 1.6), który jest słabo zbieżny do rozkładu Gaussa. Wówczas przedział ufności dla frakcji wyznacza się ze wzoru:

$$P\left(w_i - u_\alpha \sqrt{\frac{w_i \cdot (1 - w_i)}{n}} < p < w_i + u_\alpha \sqrt{\frac{w_i \cdot (1 - w_i)}{n}}\right) = 1 - \alpha \qquad (3.21)$$

gdzie:
p – frakcja (wskaźnik struktury) w populacji generalnej,
n – liczebność badanej próby,
u_α – wartość odczytana z tablic dystrybuanty standaryzowanego rozkładu normalnego przy poziomie ufności $1 - \alpha$, dla której zachodzi (3.13),
w_i – wskaźnik struktury wyznaczony dla próby ze wzoru (2.1).

Przykład 3.2

W przykładzie 2.4 pokazano dane (w tab. 2.3) o rocznych wydatkach 100 losowo wybranych pracowników pewnej korporacji (zatrudniającej kilka tys. pracowników) na ubezpieczenie emerytalne. Obliczona (tab. 2.5) średnia arytmetyczna wyniosła 722,5 PLN, a odchylenie standardowe 753,57 PLN. Wykazano również (tab. 2.6), że najwięcej pracowników, tj. 40%, przeznacza na ubezpieczenie emerytalne od 500 do 1000 PLN. Należy opisać możliwie najpełniej wydatki pracowników tej korporacji na zabezpieczenie emerytalne, przyjmując, że wybrana próba jest reprezentacyjna.

Rozwiązanie

W tym przykładzie wszyscy pracownicy korporacji tworzą zbiorowość statystyczną, z której wylosowano 100-elementową próbę. Korzystając z wcześniej wykonanych obliczeń, należy zbudować przedziały ufności dla średniej, odchylenia standardowego i frakcji na poziomie ufności 0,95. Próba wynosi 100 pracowników, zatem jest to duża próba i do wyznaczenia przedziałów ufności można skorzystać ze wzorów: (3.11) dla średniej, (3.19) dla odchylenia standardowego i (3.21) dla wskaźnika struktury. Wszystkie przedziały skonstruowane zostaną na podstawie estymatorów o rozkładzie normalnym, dlatego z tablic dystrybuanty rozkładu Gaussa (np. w Excelu) odczytuje się wartość odpowiedniej statystyki dla prawdopodobieństwa $\left(1-\frac{\alpha}{2}\right)=0,975$, która wynosi $u_\alpha = 1,96$. Dla wartości oczekiwanej ze wzoru (3.11) otrzymuje się:

$$P\left\{\bar{x}-u_\alpha\frac{S}{\sqrt{n}}<\mu<\bar{x}+u_\alpha\frac{S}{\sqrt{n}}\right\}=P\left(722,5-1,96\frac{753,57}{\sqrt{100}}<\mu<722,5+1,96\frac{753,57}{\sqrt{100}}\right)=$$
$$=P(722,5-1,96\cdot 75,357<\mu<722,5+147,6997)$$

Stąd $P(574,80 < \mu < 870,20) = 0,95$. Na podstawie zbudowanego przedziału ufności można twierdzić z prawdopodobieństwem 0,95, że pracownicy tej korporacji przeznaczają na zabezpieczenie emerytalne średnio od 574,8 do 870,2 PLN.

Kolejnym szacowanym parametrem jest odchylenie standardowe, dla którego przedział ufności zostanie wyznaczony z (3.19):

$$P\left\{\frac{S}{1+\frac{u_\alpha}{\sqrt{2n}}}<\sigma<\frac{S}{1-\frac{u_\alpha}{\sqrt{2n}}}\right\}=P\left(\frac{753,57}{1+\frac{1,96}{\sqrt{200}}}<\sigma<\frac{753,57}{1-\frac{1,96}{\sqrt{200}}}\right)=$$

$$=P\left(\frac{753,57}{1+\frac{1,96}{14,1421}}<\sigma<\frac{753,57}{1-0,1386}\right)=P\left(\frac{753,57}{1,1386}<\sigma<\frac{753,57}{0,8614}\right)$$

Stąd $P(661{,}84 < \sigma < 874{,}81) = 0{,}95$ i ostateczny wniosek – przeciętne zróżnicowanie oszczędności na starość wśród pracowników badanej korporacji mieści się zatem w przedziale (661,84; 874,81) PLN z prawdopodobieństwem 0,95.

Ostatnim szacowanym parametrem jest frakcja pracowników wydających od 500 do 1000 PLN, dla której przedział ufności wyznaczony zostanie ze wzoru (3.20):

$$P\left(w_i - u_\alpha \sqrt{\frac{w_i \cdot (1-w_i)}{n}} < p < w_i + u_\alpha \sqrt{\frac{w_i \cdot (1-w_i)}{n}} \right) =$$

$$= P\left(0{,}4 - 1{,}96 \sqrt{\frac{0{,}4 \cdot (1-0{,}4)}{100}} < p < 0{,}4 + 1{,}96 \sqrt{\frac{0{,}4 \cdot 0{,}6}{100}} \right) =$$

$$= P(0{,}4 - 1{,}96\sqrt{0{,}0024} < p < 0{,}4 + 1{,}96 \cdot 0{,}049)$$

Stąd ostatecznie przedział ufności wynosi $P(0{,}304 < p < 0{,}496) = 0{,}95$, co upoważnia do wniosku – wśród pracowników tej korporacji najwięcej z nich, tj. od 30 do 50% (co stwierdzić można z 95% prawdopodobieństwem), odkłada na zabezpieczenie emerytalne od 500 do 1000 PLN.

3.3. Wprowadzenie do weryfikacji hipotez statystycznych

Hipoteza statystyczna to osąd (przypuszczenie) odnoszący się do nieznanego poziomu parametrów lub do nieznanej postaci rozkładu zmiennych losowych w zbiorowości generalnej. Wyróżnia się dwa rodzaje hipotez – parametryczne i nieparametryczne. Celem weryfikacji hipotez parametrycznych, tj. sądów dotyczących parametrów rozkładu cechy w populacji generalnej, jest sprecyzowanie wartości parametru rozkładu populacji generalnej znanego typu. Natomiast celem weryfikacji hipotez nieparametrycznych, np. dotyczących kształtu rozkładu populacji generalnej, może być sprecyzowanie typu rozkładu cechy w określonej populacji generalnej. Innymi przykładami hipotez nieparametrycznych mogą być losowość próby lub niezależność cech.

Celem weryfikacji hipotez statystycznych jest sprawdzenie poprawności sformułowanych hipotez statystycznych, czyli podjęcie określonych decyzji statystycznych. Procedura weryfikacji hipotez statystycznych opiera się zawsze na dwóch, konkurencyjnych w stosunku do siebie, hipotezach – zerowej i alternatywnej.

Hipoteza zerowa (H_0) jest podstawową hipotezą statystyczną, będącą przedmiotem weryfikacji. Zatem proces weryfikacji może doprowadzić do jej odrzucenia bądź do stwierdzenia, że nie ma podstaw, aby ją odrzucić. Hipotezę tę formułuje się w często w ten sposób (czasem wbrew rozsądkowi), aby można ją było łatwo odrzucić. Hipoteza alternatywna (H_1) stanowi inaczej niż H_0. Jeśli odrzuca się hipotezę zerową, to przyjmuje się hipotezę alternatywną. Warto w tym miejscu zauważyć, że:

- w przypadku, kiedy hipoteza zerowa jest nieparametryczna, to hipoteza alternatywna jest formułowana w postaci przeciwnej w stosunku do hipotezy zerowej,
- w przypadku, kiedy hipoteza zerowa jest parametryczna, to hipotezę alternatywną można sformułować dwustronnie (tzn. jest różne), prawostronnie (tzn. jest większe od) lub lewostronnie (tzn. jest mniejsze od), przy czym sposób formułowania hipotezy zależy nie tylko od celu badania, ale również od rodzaju informacji statystycznych uzyskanych z próby losowej.

Hipotezy statystyczne weryfikuje się za pomocą testów statystycznych, przy czym w zależności od rodzaju hipotezy wyróżnia się testy parametryczne, które służą do weryfikacji hipotez parametrycznych, i nieparametryczne (np. testy zgodności, losowości, niezależności), służące do weryfikacji hipotez nieparametrycznych.

Przy weryfikacji hipotez statystycznych ważny jest tzw. poziom istotności α oraz obszary odrzucenia hipotez zerowych. Oba te pojęcia wiążą się z teorią błędów popełnianych przy podejmowaniu decyzji w trakcie weryfikacji hipotez statystycznych na podstawie próby, brakuje bowiem całkowitej informacji o zbiorowości, z której została ona pobrana. Tak więc weryfikując hipotezę zerową, można podjąć decyzję poprawną lub popełnić błąd. Rozróżnia się dwa rodzaje błędów, które występują z określonym prawdopodobieństwem (tab. 3.1):

- błąd I rodzaju, który polega na odrzuceniu hipotezy zerowej wtedy, gdy jest ona w rzeczywistości prawdziwa; prawdopodobieństwo jego popełnienia oznacza się przez α,
- błąd II rodzaju, który polega na przyjęciu hipotezy zerowej wtedy, gdy jest ona w rzeczywistości fałszywa, czyli jeśli hipoteza alternatywna jest prawdziwa; prawdopodobieństwo jego popełnienia oznacza się przez β.

Tabela 3.1. Błędy popełniane przy weryfikacji hipotez statystycznych

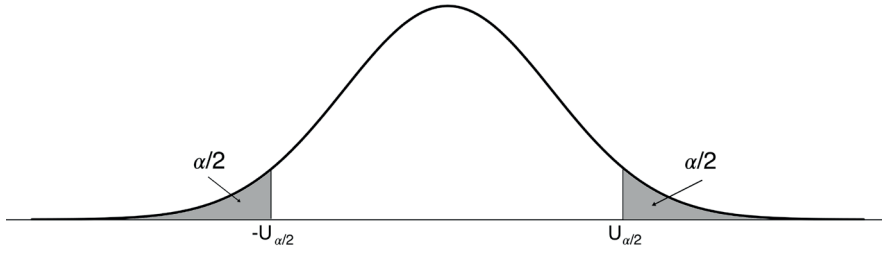
Decyzja	Hipoteza H_0 prawdziwa	Hipoteza H_0 fałszywa
Przyjęcie weryfikowanej hipotezy H_0	Decyzja poprawna	Decyzja błędna – błąd II rodzaju
Odrzucenie weryfikowanej hipotezy H_0	Decyzja błędna – błąd I rodzaju	Decyzja poprawna

Źródło: [Witkowska 2004, s. 187].

Prawdopodobieństwo popełnienia błędu I rodzaju (α) ustalone jest z góry jako dowolnie niskie prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej i nazwane jest poziomem istotności. W naukach ekonomicznych prawdopodobieństwo błędu I rodzaju przyjmuje się zazwyczaj z przedziału <0,001; 0,1>, przy czym najczęściej $\alpha = 0{,}05$. W programach komputerowych podawana jest zazwyczaj wartość prawdopodobieństwa, tzw. *p-value*, która jest minimalną wartością α poziomu istotności, dla którego hipoteza H_0 może zostać odrzucona na podstawie wyników próby. Zatem H_0 odrzuca się, jeśli $p < \alpha$. Mocą testu nazywa się prawdopodobieństwo niepopełnienia błędu II rodzaju, czyli $(1 - \beta)$.

Test statystyczny to reguła postępowania, która na podstawie wyników z próby ma doprowadzić do odrzucenia (lub nie) postawionej hipotezy statystycznej. W tym celu oblicza się wartość tzw. sprawdzianu testu (statystyki testu), będącego zmienną losową o określonym rozkładzie, którego wartość wyznacza się na podstawie danych z próby. Najczęściej używanymi w praktyce są tzw. testy istotności, które umożliwiają odrzucenie hipotezy z ryzykiem równym poziomowi istotności α bez uwzględniania prawdopodobieństwa popełnienia błędu II rodzaju. Stosując testy istotności, podejmuje się decyzję odnośnie do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej albo decyzję o braku podstaw do odrzucenia hipotezy zerowej, co nie jest równoznaczne z jej przyjęciem. W związku z tym w przypadku testów istotności hipotezę zerową dobrze jest formułować w taki sposób, aby można ją było stosunkowo łatwo odrzucić.

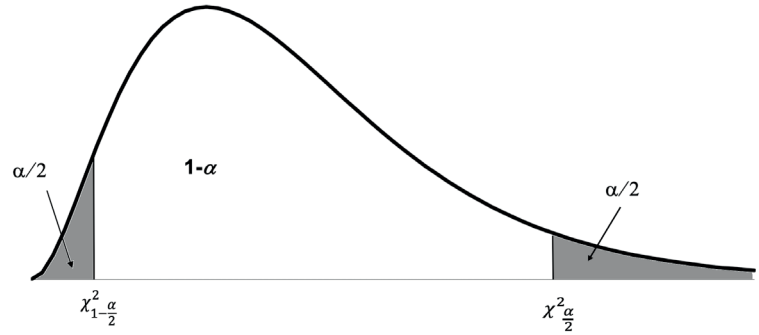
Obszar odrzucenia jest zbiorem wartości zmiennej losowej będącej sprawdzianem testu, który pozwala na odrzucenie hipotezy zerowej[3]. W zależności od rodzaju testu oraz sposobu sformułowania hipotezy alternatywnej wyróżnia się jedno- lub dwustronny obszar odrzucenia. Przykłady obustronnych obszarów odrzucenia przedstawiono na rys. 3.4 i 3.5. Ilustrację przykładowego lewostronnego obszaru odrzucenia pokazano na rys. 3.6 dla rozkładu Gaussa, a prawostronnego na rys. 3.7 dla rozkładu t-Studenta. Natomiast na rys. 3.8 porównano obszary odrzucenia prawostronny i dwustronny.



Rysunek 3.4. Obustronny obszar odrzucenia dla standaryzowanego rozkładu normalnego i poziomu istotności α

Źródło: opracowanie własne.

3 Obszar odrzucenia hipotezy zerowej stanowi dopełnienie tzw. obszaru przyjęcia.



Rysunek 3.5. Obustronny obszar odrzucenia dla rozkładu chi-kwadrat dla poziomu istotności α

Źródło: opracowanie własne.

Obustronny (dwustronny) obszar odrzucenia jest budowany w przypadku testów nieparametrycznych oraz gdy H_1 jest stawiana dwustronnie (w przypadku testów parametrycznych). Jest to zbiór wszystkich wartości zmiennej losowej, która przykładowo jest o rozkładzie normalnym u, dla których zachodzi:

$$|u| \geq u_{\alpha} \tag{3.22}$$

gdzie:
u_{α} – wartość krytyczna odczytana z tablic rozkładu normalnego dla ustalonego z góry poziomu istotności α taka, że zachodzi:

$$P(|u| \geq u_{\alpha}) = \alpha \tag{3.23}$$

Wartość krytyczna testu to wartość zmiennej losowej o określonym rozkładzie, która przy danym α stanowi koniec przedziału odrzucenia. W procesie weryfikacji hipotez sprawdza się, czy obliczona wartość statystyki testowej znalazła się w obszarze odrzucenia czy obszarze przyjęcia. W pierwszym przypadku odrzuca się hipotezę zerową i przyjmuje hipotezę alternatywną. W przypadku drugim na zadanym poziomie istotności α nie ma podstaw do odrzucenia hipotezy zerowej, co oznacza, że test nie rozstrzygnął, czy postawione twierdzenie jest prawdziwe. Zaleca się wtedy zwiększenie liczebności próby lub modyfikację hipotez statystycznych, co jednak nie zawsze jest możliwe.

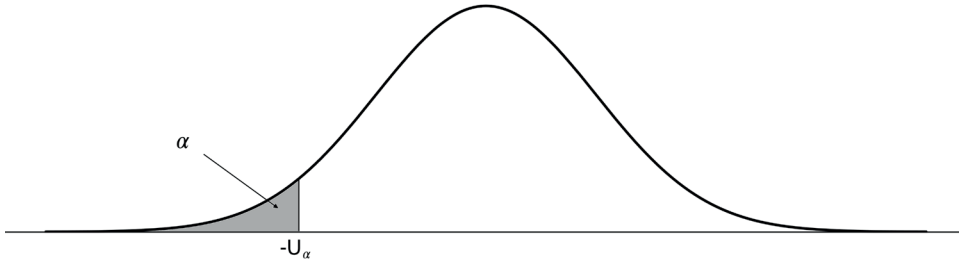
Jednostronny obszar odrzucenia hipotezy jest budowany wtedy, gdy hipoteza alternatywna H_1 jest stawiana jednostronnie. Wówczas:

- lewostronny obszar odrzucenia stanowi zbiór wszystkich wartości zmiennej losowej, np. o rozkładzie normalnym u takich, że:

$$u \leq -u_{\alpha} \tag{3.24}$$

gdzie:
u_α – wartość krytyczna odczytana z tablic rozkładu normalnego dla poziomu istotności α taka, że zachodzi (3.22), zatem:

$$P(u \leq -u_\alpha) = \alpha \tag{3.25}$$



Rysunek 3.6. Lewostronny obszar odrzucenia zbudowany na podstawie standaryzowanego rozkładu normalnego dla poziomu istotności α

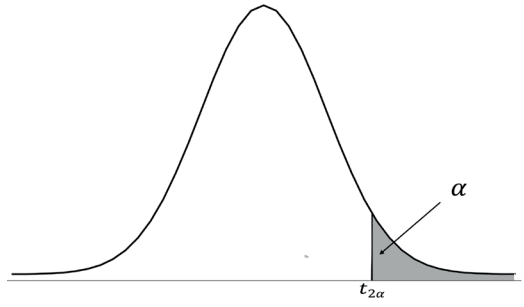
Źródło: opracowanie własne.

- prawostronny obszar odrzucenia stanowi zbiór wszystkich wartości zmiennej losowej, np. o rozkładzie normalnym u takich, że:

$$u \geq u_\alpha \tag{3.26}$$

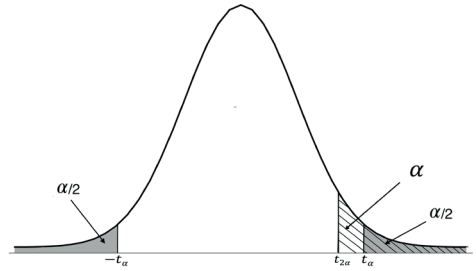
gdzie:
u_α – wartość krytyczna odczytana z tablic dla ustalonego z góry poziomu istotności α taka, że zachodzi:

$$P(u \geq u_\alpha) = \alpha \tag{3.27}$$



Rysunek 3.7. Prawostronny obszar odrzucenia zbudowany na podstawie rozkładu t-Studenta dla poziomu istotności α

Źródło: opracowanie własne.



Rysunek 3.8. Jednostronny i dwustronny obszar odrzucenia zbudowany na podstawie rozkładu t-Studenta dla poziomu istotności

Źródło: opracowanie własne.

Reasumując, w procesie weryfikacji hipotez statystycznych można wyróżnić kilka etapów:

1) sformułowanie hipotezy zerowej (H_0) oraz hipotezy alternatywnej (H_1);
2) wybór testu statystycznego służącego do weryfikacji hipotezy zerowej;
3) wyznaczenie wartości sprawdzianu testu;
4) ustalenie poziomu istotności (α) oraz wyznaczenie obszaru odrzucenia hipotezy zerowej;
5) podjęcie decyzji z określonym prawdopodobieństwem błędu.

3.4. Weryfikacja hipotez dotyczących parametrów opisowych populacji

Testy parametryczne służą do weryfikacji hipotez odnośnie do parametrów rozkładu rozpatrywanej zbiorowości. Przy czym weryfikacja hipotez może dotyczyć:

- jednej populacji generalnej – wówczas celem jest sprawdzenie, czy badany parametr charakteryzujący zbiorowość statystyczną ma określoną wartość,
- dwóch lub więcej zbiorowości – celem jest weryfikacja hipotezy o równości wartości badanego parametru w analizowanych zbiorowościach.

W podrozdziale 3.1 stwierdzono, że estymatorami nieznanych parametrów zbiorowości mogą być adekwatne parametry obliczone na podstawie odpowiednio dobranej próby statystycznej. Jednakże wyznaczona ocena estymatora może nie być wartością poszukiwanego parametru. Stąd istnieje konieczność weryfikacji hipotezy, czy nieznana wartość parametru jest równa lub statystycznie różni się (albo jest większa lub mniejsza) od pewnej (np. przyjętej z góry) wartości.

Weryfikując hipotezy dotyczące wartości oczekiwanej w populacji, rozstrzyga się, czy wartość oczekiwana jest równa konkretnej wartości liczbowej. Stawia się zatem hipotezę zerową:

$$H_0: E(X) = \mu_0 \quad \text{lub} \quad H_0: \mu = \mu_0 \tag{3.28}$$

gdzie:
μ_0 jest zadaną liczbą,

wobec jednej z możliwych hipotez alternatywnych:

$$H_1: E(X) \neq \mu_0 \quad \text{lub} \quad H_1: \mu \neq \mu_0 \tag{3.29a}$$

$$H_1: E(X) > \mu_0 \quad \text{lub} \quad H_1: \mu > \mu_0 \tag{3.29b}$$

$$H_1: E(X) < \mu_0 \quad \text{lub} \quad H_1: \mu < \mu_0 \tag{3.29c}$$

wskazujących odpowiednio na obustronny, prawostronny lub lewostronny obszar odrzucenia, tak jak pokazano na rys. 3.4–3.8. Sformułowanie hipotezy alternatywnej w postaci zaprzeczenia hipotezy zerowej jest zawsze możliwe i nie wymaga dodatkowych wyjaśnień. Natomiast w przypadku jednostronnego obszaru odrzucenia sposób sformułowania hipotezy zależy od wartości parametru wyznaczonego w próbie. Przykładowo, jeśli średnia z próby jest większa od wartości założonej, tj. $\bar{x} > \mu_0$, to hipoteza alternatywna sformułowana zostanie w postaci (3.29b), zatem obszar odrzucenia będzie znajdował się w prawym ogonie rozkładu statystyki testowej. W przypadku kiedy okaże się, że $\bar{x} < \mu_0$, hipoteza alternatywna będzie postaci (3.29c) i obszar odrzucenia będzie lewostronny. Warto przy tym zaznaczyć, że jednostronne sformułowanie hipotezy alternatywnej niesie więcej informacji praktycznej niż w przypadku hipotezy będącej zaprzeczeniem hipotezy zerowej. Co więcej, hipotezy stawiane jednostronnie łatwiej jest odrzucić, co widać na rys. 3.8.

Podobnie jak w przypadku budowy przedziałów ufności, przy weryfikacji hipotez statystycznych wyróżnia się 3 przypadki.

1) Populacja generalna ma rozkład normalny i znane jest jej zróżnicowanie mierzone odchyleniem standardowym (σ). Wówczas sprawdzian testu ma standaryzowany rozkład normalny i jest postaci:

$$u = \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} \tag{3.30}$$

gdzie:
$\bar{x}$ – średnia z próby,
pozostałe oznaczenia jak poprzednio.

2) W przypadku małej próby losowej i braku znajomości wariancji populacji generalnej o zakładanym rozkładzie Gaussa sprawdzian testu ma rozkład t-Studenta o $(n-1)$ stopniach swobody postaci:

$$t = \frac{\overline{x} - \mu_0}{S} \cdot \sqrt{n-1} \tag{3.31a}$$

jeśli korzysta się z odchylenia standardowego z próby. Natomiast jeśli estymator odchylenia standardowego jest nieobciążony, to sprawdzian testu jest postaci:

$$u = \frac{\overline{x} - \mu_0}{\breve{S}} \cdot \sqrt{n} \tag{3.31b}$$

3) Brak jest informacji o rozkładzie cechy w populacji, ale próba jest duża, wówczas statystyka testowa ma rozkład normalny i jest postaci:

$$u = \frac{\overline{x} - \mu_0}{S} \cdot \sqrt{n} \tag{3.32}$$

Dalsze analizy dotyczyć będą dwóch wartości oczekiwanych, np. pochodzących z różnych populacji albo obserwowanych w różnym czasie. W tym przypadku hipoteza zerowa jest postaci:

$$H_0: E(X_1) = E(X_2) \quad \text{lub} \quad H_0: \mu_1 = \mu_2 \tag{3.33}$$

gdzie:
$E(X_1)$, $E(X_2)$ oraz μ_1, μ_2 o oznaczają nadzieje matematyczne odpowiednio w pierwszej i drugiej badanej zbiorowości.

Zakładając, że dopuszczalne są jedno- i dwustronne obszary odrzucenia, hipotezy alternatywne można sformułować jako:

$$H_1: E(X_1) \neq E(X_2) \quad \text{lub} \quad H_1: \mu_1 \neq \mu_2 \tag{3.34a}$$

$$H_1: E(X_1) > E(X_2) \quad \text{lub} \quad H_1: \mu_1 > \mu_2 \tag{3.34b}$$

$$H_1: E(X_1) < E(X_2) \quad \text{lub} \quad H_1: \mu_1 < \mu_2 \tag{3.34c}$$

Podobnie jak wcześniej rozpatruje się kilka sytuacji.

1) Obie badane populacje mają rozkład normalny $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$, a odchylenia standardowe σ_1 i σ_2 są znane. Z obu populacji wylosowano próby o liczebnościach odpowiednio n_1 i n_2. W tym przypadku statystyka testowa jest o rozkładzie normalnym postaci:

$$u = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.35}$$

gdzie:
$\overline{x}_1, \overline{x}_2$ – średnie wartości cechy wyznaczone w obu próbach.

2) Dane są małe próby pobrane z dwu populacji (tj. $n_1 + n_2 < 60$) o rozkładzie normalnym $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$. Odchylenia standardowe σ_1 i σ_2 nie są znane, ale są równe. Wówczas sprawdzian testu ma rozkład t-Studenta o $(n_1 + n_2 - 2)$ stopniach swobody i jest postaci:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{n_1 \cdot S_1^2 + n_2 \cdot S_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{3.36}$$

3) Z dwu populacji o rozkładzie normalnym $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$ pobrano małe próby. Odchylenia standardowe σ_1 i σ_2 nie są znane, ale nie jest spełnione założenie o jednorodności wariancji, tzn. $\sigma_1^2 \neq \sigma_2^2$. Wówczas na podstawie wyników niezależnych prób wyznacza się wartość statystyki testowej o rozkładzie t-Studenta postaci:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{3.37}$$

Test ten nosi nazwę testu Cochrana-Coxa. Istnieje kilka sposobów określania liczby stopni swobody w celu wyznaczenia wartości krytycznej dla tego testu. Przykładowo w [Malarska 2005, s. 139] postuluje się, aby liczbę stopni swobody wyznaczać jako: $\mathbb{E}\left[(T_1 + T_2 - 2) \cdot \left(\frac{S_1^2 \cdot S_2^2}{S_1^2 + S_2^2} + \frac{1}{2}\right)\right]$ gdzie $\mathbb{E}[.]$ oznacza funkcję *entier*.

4) Brak wiedzy o rozkładzie cechy w populacji, ale próby są duże. Wówczas statystyka testowa ma rozkład normalny i jest postaci:

$$u = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{3.38}$$

gdzie:
S_1, S_2 – odchylenia standardowe wyznaczone w obu próbach,
pozostałe oznaczenia jak poprzednio.

Ważnymi parametrami opisowymi są wariancja i odchylenie standardowe, w odniesieniu do których weryfikuje się stosowne hipotezy statystyczne. Po pierwsze bada się, czy odchylenie standardowe (lub wariancja) mają określoną wartość, a po drugie sprawdza się, czy zmienność w dwu populacjach jest identyczna.

Rozważa się zbiorowość o rozkładzie Gaussa, przy czym wartość oczekiwana μ i odchylenie standardowe populacji σ nie są znane. Z badanej populacji wylosowano niezależnie n-elementową próbę. Na podstawie tej próby należy zweryfikować hipotezę zerową o tym, że wariancja ma określoną wartość σ_0^2:

$$H_0: \sigma^2 = \sigma_0^2 \tag{3.39}$$

Hipotezy alternatywne mogą przyjąć jedną z postaci:

$$H_1: \sigma^2 \neq \sigma_0^2 \tag{3.40a}$$

$$H_1: \sigma^2 > \sigma_0^2 \tag{3.40b}$$

$$H_1: \sigma^2 < \sigma_0^2 \tag{3.40c}$$

W przypadku małej próby wartość statystyki testowej ma rozkład chi-kwadrat o $(n - 1)$ stopniach swobody i jest postaci:

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} \tag{3.41}$$

Weryfikacja hipotez o dwóch wariancjach prowadzona jest dla dwóch prób losowych pochodzących z populacji o rozkładzie normalnym, ale o nieznanych parametrach. Formułuje się hipotezę zerową postaci:

$$H_0: \sigma_1^2 = \sigma_2^2 \tag{3.42}$$

i hipotezę alternatywną:

$$H_0: \sigma_1^2 > \sigma_2^2 \tag{3.43}$$

Przy czym przyjmuje się[4], że wariancje z próby spełniają nierówność: $S_1^2 > S_2^2$. Statystyka testowa ma rozkład Fishera-Snedecora o $(n_1 - 1)$ i $(n_2 - 1)$ stopniach swobody i jest postaci:

$$F = \frac{S_1^2}{S_2^2} \tag{3.44}$$

4 Warunek ten wynika z klasycznej budowy tablic rozkładu Fishera-Snedecora.

Przykład 3.3

Dostępne są akcje spółek A i B o średnich zwrotach $r_A = -2\%$ i $r_B = 5\%$ oraz odchyleniach standardowych $S_A = 5\%$ i $S_B = 12\%$ odpowiednio, wyliczonych na próbie $T = 100$. W celu wyboru instrumentu inwestycyjnego należy zbadać, czy stopy zwrotu istotnie różnią się od 0 oraz czy oba wybrane instrumenty mają istotnie różne stopy zwrotu i ryzyka mierzone odchyleniami standardowymi.

Rozwiązanie

Notowania akcji spółek, mimo że są procesem ciągłym, dokonywane są w wybranych momentach pomiarowych $t = 1, 2, \ldots, T$, i dlatego traktuje się je jako obserwacje zmiennej losowej. W konsekwencji, wyznaczone na ich podstawie stopy zwrotu r_{At} i r_{Bt} są traktowane jako realizacje zmiennych losowych R_A i R_B odpowiednio dla spółki A i B. Natomiast wartości średnich zwrotów r_A i r_B pełnią rolę średnich z obu prób statystycznych o 100 elementach, podobnie jak S_A i S_B są odchyleniami standardowymi z tych prób.

W celu wyboru instrumentu inwestycyjnego warto najpierw sprawdzić, czy rozpatrywane akcje przynoszą zyski czy straty, co sprowadza się do rozstrzygnięcia hipotez H_0: $E(R_A) = 0$ vs. H_1: $E(R_A) < 0$ dla spółki A oraz H_0: $E(R_B) = 0$ vs. H_1: $E(R_B) > 0$ dla spółki B. Hipotezy alternatywne są jednostronne, postawione zgodnie z podanymi w zadaniu wartościami średnimi $r_A < 0$ i $r_B > 0$.

Tak sformułowane hipotezy weryfikuje się z wykorzystaniem statystyki (3.32) dla dużych prób. Na potrzeby tego przykładu sprawdzian testu przyjmuje postać: $u = \left(\frac{r}{S}\right)\sqrt{T}$, co prowadzi do wyników: dla spółki A: $u = \left(-\frac{2}{5}\right)\sqrt{100} = -4$, a dla spółki B: $u = \left(\frac{5}{12}\right)\cdot 10 = 4{,}17$. Ponieważ wartość krytyczna testu dla poziomu istotności $\alpha = 0{,}05$ i lewostronnego obszaru odrzucenia wynosi –1,64 oraz 1,64 dla obszaru prawostronnego, to na poziomie istotności $\alpha = 0{,}05$ odrzuca się hipotezy zerowe na rzecz alternatywnych dla obu instrumentów. Zatem oczekiwane zwroty z akcji spółki A są istotnie mniejsze od 0, a akcji spółki B są istotnie większe od 0. Innymi słowy akcje spółki B generują zyski (bo hipoteza alternatywna ma prawostronny obszar odrzucenia i dotyczy zwrotów dodatnich), podczas gdy akcje spółki A – straty (bo hipoteza alternatywna ma lewostronny obszar odrzucenia i dotyczy zwrotów ujemnych). Wniosek – nie opłaca się inwestować w akcje spółki A i dalsze analizy w zakresie porównań stóp zwrotu i ryzyka można pominąć.

Przykład 3.4

Na potrzeby decyzji inwestycyjnych należy porównać stopy zwrotu z akcji spółek A i B o średnich zwrotach $r_A = 2\%$ i $r_B = 4\%$ oraz odchyleniach $S_A = 5\%$ i $S_B = 12\%$ odpowiednio, wyliczonych na próbie $T = 100$.

Rozwiązanie

Analogicznie jak w przykładzie 3.3 rozwiązanie zadania rozpoczyna się do rozstrzygnięcia hipotez o zerowych zwrotach H_0: $E(R_A) = 0$ vs. H_1: $E(R_A) > 0$ dla spółki A oraz H_0: $E(R_B) = 0$ vs. H_1: $E(R_B) > 0$ dla spółki B. Obie hipotezy alternatywne zostały postawione jednostronnie, zgodnie z podanymi wartościami średnich stóp zwrotu $r_A > 0$ i $r_B > 0$. Sprawdzianem testu (dla dużej próby) jest (jak w poprzednim przykładzie) statystyka: $u = \left(\frac{r}{S}\right) \cdot \sqrt{T}$, czyli: dla spółki A: $u = \left(\frac{2}{5}\right) \cdot \sqrt{100} = 4$, a dla spółki B: $u = \left(\frac{4}{12}\right) \cdot 10 = 3,33.$. Ponieważ wartość krytyczna testu dla poziomu istotności $\alpha = 0,05$ i prawostronnego obszaru odrzucenia wynosi $u_\alpha = 1,64$, to odrzuca się obie hipotezy zerowe na rzecz hipotez alternatywnych. Wniosek – stopy zwrotu obu spółek są istotnie dodatnie.

Tym razem jednak nie można zakończyć zadania, bowiem nadal nie wiadomo, które akcje są lepszym instrumentem inwestycyjnym pod względem rentowności i ekspozycji na ryzyko. Najpierw trzeba zatem sprawdzić, czy stopy zwrotu z akcji spółki B są (jak sugeruje treść zadania) istotnie większe od stóp zwrotu spółki A. Zadanie sprowadza się do weryfikacji hipotezy o równości stóp zwrotu H_0: $E(R_A) = E(R_B)$ vs. H_1: $E(R_A) < E(R_B)$.

Korzystając ze statystyki testowej (3.35) dla dużych prób $u = \frac{r_A - r_B}{\sqrt{\frac{S_A^2}{T_A} + \frac{S_B^2}{T_B}}}$, wyliczyć można wartość sprawdzianu $u = \frac{2-4}{\sqrt{\frac{25}{100} + \frac{144}{100}}} = \frac{-2}{\sqrt{\frac{169}{100}}} = -1,54 \cdot |u| < u_\alpha = 1,64.$

co upoważnia do wniosku o braku podstaw do odrzucenia hipotezy zerowej na poziomie istotności $\alpha = 0,05$ i stwierdzenia, że stopy zwrotu obu instrumentów nie różnią się istotnie.

Pozostaje do rozstrzygnięcia ostatnie kluczowe pytanie, tj. czy oba instrumenty charakteryzuje identyczne ryzyko, co sprowadza się do weryfikacji hipotezy o równości wariancji obu instrumentów: H_0: $D^2(R_A) = D^2(R_B)$ *vs.* H_1: $D^2(R_A) > D^2(R_B)$, dla której sprawdzian testu (3.34) wynosi

$$F = \frac{S_B^2}{S_A^2} = \frac{144}{25} = 5,7 \cdot F > F_\alpha = 1,39$$

Wartość krytyczna testu F dla poziomu istotności $\alpha = 0,05$ wynosi 1,39 dla obszaru prawostronnego. W konsekwencji na poziomie istotności $\alpha = 0,05$ odrzuca się hipotezę zerową na rzecz alternatywnej i wnioskuje o (statystycznie) istotnie wyższym ryzyku instrumentu B (mierzonym odchyleniem standardowym) niż

akcji spółki A. Ostateczny wybór instrumentu zależy w tym przypadku od indywidualnej awersji inwestora do ryzyka.

Kolejnym analizowanym parametrem jest frakcja odzwierciedlająca strukturę populacji, tj. udział wybranego wariantu cechy w populacji. W celu sprawdzenia, czy wskaźnik struktury jest na określonym poziomie, stawia się hipotezę zerową postaci:

$$H_0: p = p_0 \tag{3.45}$$

gdzie p_0 – założona wartość frakcji.

Hipotezy alternatywne mogą przyjąć jedną z postaci:

$$H_1: p \neq p_0 \tag{3.46a}$$

$$H_1: p > p_0 \tag{3.46b}$$

$$H_1: p < p_0 \tag{3.46c}$$

Na podstawie dużej (przynajmniej 100-elementowej) próby, wyznacza się wartość statystyki testowej o standaryzowanym rozkładzie Gaussa:

$$u = \frac{w_i - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{3.47}$$

Badanie równości dwóch wskaźników struktury przeprowadza się na podstawie dużych prób pobranych z dwóch populacji o liczebności n_1, n_2 formułując hipotezy:

$$H_0: p_1 = p_2 \tag{3.48}$$

$$H_1: p_1 \neq p_2 \tag{3.49a}$$

$$H_1: p_1 > p_2 \tag{3.49b}$$

$$H_1: p_1 < p_2 \tag{3.49c}$$

gdzie:
p_1, p_2 – wskaźniki struktury odpowiednio w pierwszej i drugiej badanej zbiorowości.

Wówczas sprawdzian testu ma rozkład asymptotycznie normalny N(0, 1) i jest postaci:

$$u = \frac{w_{1i} - w_{2i}}{\sqrt{\frac{\tilde{p} \cdot \tilde{q}}{\tilde{n}}}} \tag{3.47}$$

gdzie:

$$\tilde{p} = \frac{n_{1i} + n_{2i}}{n_1 + n_2} \tag{3.48}$$

$$\tilde{q} = 1 - \tilde{p} \tag{3.49}$$

$$\tilde{n} = \frac{n_1 \cdot n_2}{n_1 + n_2} \tag{3.50}$$

$w_{1i} = \frac{n_{1i}}{n_1}$, $w_{2i} = \frac{n_{2i}}{n_2}$ – wskaźniki struktury odpowiednio w pierwszej i drugiej próbie,

n_{1i}, n_{2i} – liczebność *i*-tego wariantu cechy odpowiednio w pierwszej i drugiej próbie,
n_1, n_2 – liczebność odpowiednio pierwszej i drugiej próby.

Przy badaniu kształtu rozkładu istotne jest stwierdzenie, czy rozkład jest symetryczny oraz czy można go uznać za zbliżony do normalnego. W zasadzie są to hipotezy nieparametryczne, ale do ich weryfikacji można wykorzystać informacje o parametrach rozkładu. W pierwszym przypadku para hipotez ma postać ogólną H_0: rozkład zmiennej jest symetryczny, H_1: rozkład zmiennej nie jest symetryczny.

W celu sprawdzenia, czy rozkład jest symetryczny, bada się współczynniki skośności. Jeżeli wartość współczynnika skośności $A(\vartheta_i)$ jest równa 0, to rozkład zmiennej jest symetryczny. W rzeczywistości wartość tego współczynnika odbiega od 0, z tego względu testuje się jego istotność, stawiając hipotezę zerową:

$$H_0: A(\vartheta_i) = 0 \tag{3.51}$$

wobec hipotezy alternatywnej postaci:

$$H_1: A(\vartheta_i) > 0 \text{ lub } H_1: A(\vartheta_i) < 0 \tag{3.52}$$

gdzie:
$A(\vartheta_i)$ – asymetria procesu generującego szereg obserwacji.

W celu weryfikacji hipotezy zerowej korzysta się z tzw. standaryzowanego współczynnika skośności (por. [Luszniewicz, Słaby 2003, s. 40–41])[5], który pełni rolę sprawdzianu testu:

$$SA = A\sqrt{\frac{n}{6}} \tag{3.53}$$

gdzie:
A – współczynnik skośności wyznaczony według (2.23) dla n-elementowej próby. Przy obliczaniu wartości statystyki A należy uwzględnić założenie, że $n \geq 3$.

W przypadku gdy liczebność próby n wynosi co najmniej 150 obserwacji, statystyka SA ma standaryzowany rozkład normalny. Jeżeli wartość bezwzględna tego współczynnika jest równa lub większa od wartość krytycznej (odczytanej z tablic dystrybuanty standaryzowanego rozkładu normalnego), to hipoteza zerowa, w której zakłada się, że rozkład jest symetryczny, zostaje odrzucona (por. [Dobosz 2004, s. 25 i 64; Luszniewicz, Słaby 2003, s. 40–42]). Standaryzowany współczynnik skośności SA (przy umiarkowanej sile asymetrii) na ogół przyjmuje wartości z przedziału <–3; 3>, co oznacza, że w przypadku znalezienia się statystyki (3.53) w tym przedziale badany rozkład empiryczny można uznać za symetryczny.

W badaniach normalności rozkładu korzysta się zazwyczaj z testów nieparametrycznych, np. testu normalności Shapiro-Wilka i testu Jarque-Bera lub testów zgodności rozkładu, np. chi-kwadrat i Kołmogorowa-Lillieforsa, które implementowane są w popularnych programach komputerowych. Z ich wykorzystaniem rozstrzyga się parę hipotez (nieparametrycznych) postaci H_0: empiryczny rozkład prawdopodobieństwa jest zgodny z rozkładem normalnym, H_1: empiryczny rozkład prawdopodobieństwa nie jest zgodny z rozkładem normalnym.

Przy weryfikacji hipotezy zerowej można również skorzystać ze znacznie prostszego do wykonania testu parametrycznego dotyczącego stopnia spłaszczenia rozkładu analizowanej zmiennej w stosunku do rozkładu normalnego. Dokonuje się tego w trakcie testowania kurtozy, co umożliwia odpowiedź na pytanie, czy rozpatrywany rozkład symetryczny można uznać za rozkład Gaussa. W tym celu hipotezy nieparametryczne zastępuje się hipotezami parametrycznymi postaci:

$$H_0\colon K(\vartheta_i) = 0 \tag{3.54}$$

$$H_1\colon K(\vartheta_i) > 0 \text{ lub } H_1\colon K(\vartheta_i) < 0 \tag{3.55}$$

gdzie:
$K(\vartheta_i)$ – kurtoza procesu generującego szereg obserwacji.

5 Przy obliczaniu wartości statystyki A według (1.16) należy uwzględnić założenie, że $T \geq 3$.

Dla rozkładu normalnego współczynnik K przyjmuje wartość 0. Do statystycznej oceny, czy wartość kurtozy istotnie różni się od 0, wykorzystuje się standaryzowany współczynnik kurtozy SK postaci (por. [Luszniewicz, Słaby 2003, s. 42]):

$$SK = K\sqrt{\frac{n}{24}} \tag{3.56}$$

gdzie:
K – kurtoza wyznaczona według (2.28) dla próby n-elementowej. Przy obliczaniu wartości statystyki SK należy uwzględnić założenie, że $n \geq 4$.

Jeżeli liczba obserwacji n wynosi co najmniej 150, to współczynnik SK ma standaryzowany rozkład normalny. Zatem jeśli wartość bezwzględna tego współczynnika jest równa lub większa od wartość krytycznej, to hipoteza zerowa, w której zakłada się, że wartość współczynnika kurtozy jest równa 0, zostaje odrzucona (por. [Dobosz 2004, s. 25 i 64; Luszniewicz, Słaby 2003, s. 40–43]). Współczynnik SK przy umiarkowanej kurtozie na ogół przyjmuje wartości z przedziału $<-3; 3>$. W przypadku gdy (3.56) < -3, to występuje znaczące spłaszczenie badanego rozkładu empirycznego, natomiast jeśli (3.56) > 3, to występuje znaczna wysmukłość tego rozkładu. Zatem jeśli (3.56) znajdzie się w założonym przedziale, to można uznać rozpatrywany rozkład za normalny, bez konieczności stosowania znacznie bardziej pracochłonnych testów zgodności rozkładu.

Rozdział 4

Analiza współzależności zjawisk

Dotychczasowe rozważania dotyczyły analiz pojedynczych cech. Jednakże w badaniach społecznych rozpatruje się wiele z nich, mierzonych na różnych skalach pomiarowych. Nasuwa się więc naturalne pytanie, czy badane cechy i zjawiska są ze sobą w jakiś sposób powiązane. Analiza współzależności umożliwia weryfikację poglądów dotyczących występujących zależności między zjawiskami oraz identyfikację czynników na nie wpływających. W analizach dwu i więcej cech, czyli inaczej analizach wielowymiarowych (wielocechowych), bada się charakter i siłę oraz (o ile to możliwe) kierunek tych zależności.

4.1. Rozkład dwuwymiarowy

Punktem wyjściowym do badania współzależności cech są dane, w których dla każdej jednostki statystycznej, obiektu lub momentu pomiarowego określono jednocześnie wartości dwóch cech: X i Y. Pojawia się więc zbiór n jednostek i przyporządkowane im obserwacje dotyczące pary cech (x_i, y_i), $i = 1, 2, \ldots, n$, co przedstawiono w tab. 4.1.

Tabela 4.1. Szereg szczegółowy dla dwóch obserwowanych cech

Numer obserwacji i	Obserwacje cechy X: x_i	Obserwacje cechy Y: y_i
1	x_1	y_1
2	x_2	y_2
…	…	…
N	x_n	y_n

Źródło: opracowanie własne.

Jeżeli liczebność zbiorowości jest duża i zachodzi potrzeba pogrupowania danych w szeregi rozdzielcze, to ze względu na dwa różne wymiary grupowania dotyczące k wariantów dla cechy X i l wariantów cechy Y otrzymuje się $(k \cdot l)$ par obserwacji dotyczących obu cech jednocześnie, czyli $(k \cdot l)$ wartości n_{ij}, oznaczających liczebności klas dla i-tego wariantu cechy X ($i = 1, 2, \ldots, k$) i j-tego wariantu cechy Y ($j = 1, 2, \ldots, l$). Takie przyporządkowanie nazywa się dwuwymiarowym rozkładem empirycznym cech (X, Y). Dane pogrupowane umieszcza się zwykle w tablicy, która w przypadku danych ilościowych nosi nazwę tablicy korelacyjnej (tab. 4.2). Kiedy analizie poddawane są cechy jakościowe, tablice te noszą nazwę tablic asocjacji lub kontyngencji.

Tabela 4.2. Tablica korelacyjna

Przedziały klasowe wartości cech	$[y_{1d}, y_{1g})$	$[y_{2d}, y_{2g})$	…	$[y_{ld}, y_{lg})$	$\sum_{j=1}^{l} n_{ij} = n_{i\cdot}$
$[x_{1d}, x_{1g})$	n_{11}	n_{12}	…	n_{1l}	$n_{1\cdot}$
$[x_{2d}, x_{2g})$	n_{21}	n_{22}	…	n_{2l}	$n_{2\cdot}$
…	…	…	…	…	…
$[x_{kd}, x_{kg})$	n_{k1}	n_{k2}	…	n_{kl}	$n_{k\cdot}$
$\sum_{i=1}^{k} n_{ij} = n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot l}$	$\sum_{j=1}^{l} n_{i\cdot} = \sum_{j=1}^{l} n_{\cdot j} = n$

Źródło: opracowanie własne.

Symbole w pierwszej kolumnie i pierwszym wierszu tab. 4.2, tzn.: x_{id}, x_{ig} oraz y_{jd}, y_{jg} oznaczają odpowiednio dolną i górną granicę przedziałów klasowych dla cech X i Y. Natomiast wewnątrz tabeli podano liczebności n_{ij} wyrażające liczbę obserwacji, które charakteryzują się i-tym wariantem cechy X oraz j-tym wariantem cechy Y. Tablice asocjacji i kontyngencji wyglądają podobnie z tą różnicą, że warianty cech są przedstawione w sposób adekwatny do skali pomiarowej każdej z cech. W ostatnim wierszu tablicy korelacyjnej umieszczone zostały sumy liczebności wszystkich klas cechy X dla danego wariantu Y oznaczonego przez $n_{\cdot j}$, tworzące rozkład empiryczny cechy Y w badanej zbiorowości, nazywany tutaj rozkładem brzegowym tej cechy. Podobnie w ostatniej kolumnie tablicy powstał rozkład brzegowy cechy X, oznaczony jako $n_{i\cdot}$. Innymi słowy, w ostatnim wierszu zliczono wszystkie obserwacje (suma po wierszach) dla kolejnych $j = 1, 2, \ldots, l$ wariantów cechy Y, a w ostatniej kolumnie zliczono wszystkie obserwacje (suma po kolumnach) dla kolejnych $i = 1, 2, \ldots, k$ wariantów cechy X.

W analizach wielowymiarowych często bada się tzw. rozkład warunkowy, wyznaczony dla jednej z cech przy założeniu, że cecha druga przyjmuje konkretny wariant. Rozkładem warunkowym cechy Y dla części populacji ($n_{1\bullet}$ jednostek) posiadającej cechę X w pierwszym wariancie[6], tj. dla wartości x_i należącej do przedziału $(x_{1d}; x_{1g})$, przedstawiono w tab. 4.3.

Tabela 4.3. Rozkład warunkowy cechy *Y* dla pierwszego wariantu cechy *X*

Przedziały wartości cechy Y	Liczebność klasy
$[y_{1d}, y_{1g})$	n_{11}
$[y_{2d}, y_{2g})$	n_{12}
…	…
$[y_{ld}, y_{lg})$	n_{1l}

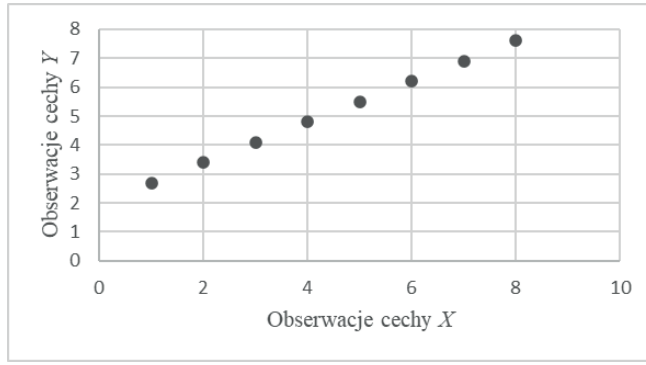
Źródło: opracowanie własne.

4.2. Występowanie współzależności

Współzależność występująca między cechami może być funkcyjna, tj. dokładna lub stochastyczna – probabilistyczna. Pierwsza występuje, kiedy zmiana wartości jednej cechy pociąga za sobą konkretną zmianę drugiej, czyli $y = f(x)$, gdzie x, y – wartości cech, f – symbol pewnej funkcji. Druga, gdy różnym wariantom jednej cechy odpowiadają różne rozkłady drugiej cechy. Przeciwnie, jeżeli dla różnych wariantów jednej cechy druga ma stale taki sam rozkład, to cechy są niezależne stochastycznie. W naukach społecznych w większości przypadków występują zależności stochastyczne. Szczególnym przypadkiem zależności stochastycznej jest zależność korelacyjna (statystyczna). Dwie cechy są zależne korelacyjnie, jeżeli różnym wariantom jednej cechy odpowiadają różne średnie warunkowe drugiej. Przy czym, średnie warunkowe są to średnie liczone dla rozkładów warunkowych. W przykładzie z tab. 4.3 jest to średnia cechy Y wyznaczona dla pierwszego wariantu cechy X.

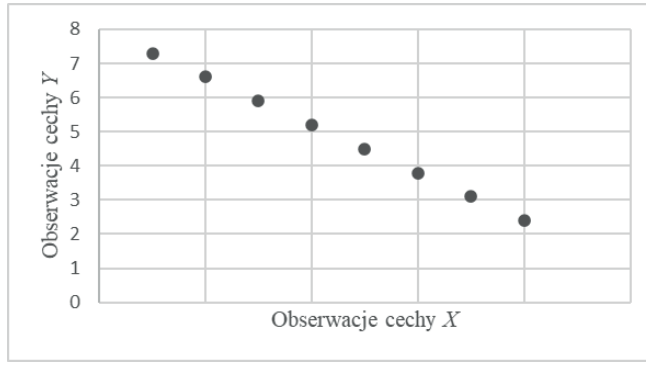
Na rys. 4.1 i 4.2 przedstawiono przykłady zależności funkcyjnej. Jak można zauważyć, każdy punkt na wykresie reprezentuje obserwacje poczynione jednocześnie dla zmiennej X oraz Y. Relacje między zmiennymi mają charakter liniowy, punkty bowiem tworzą proste. Oczywiście zależność funkcyjna może mieć również charakter nieliniowy. Z kolei rys. 4.3 i 4.4 przedstawiają zależności korelacyjne o różnej sile – na rys. 4.3 ta relacja jest silniejsza niż na rys. 4.4.

6 Tab. 4.3 przedstawia przetransponowany pierwszy wiersz tab. 4.2.



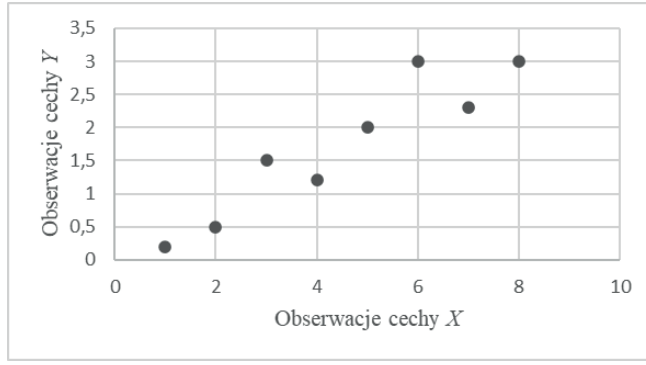
Rysunek 4.1. Przykład zależności funkcyjnej liniowej dodatniej

Źródło: opracowanie własne.



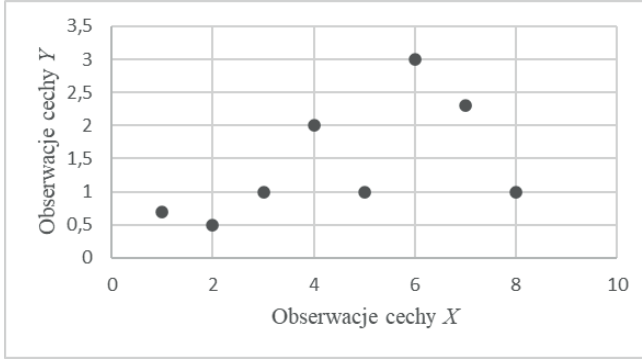
Rysunek 4.2. Przykład zależności funkcyjnej liniowej ujemnej

Źródło: opracowanie własne.



Rysunek 4.3. Przykład silnej zależności korelacyjnej dodatniej

Źródło: opracowanie własne.



Rysunek 4.4. Przykład umiarkowanej zależności korelacyjnej dodatniej

Źródło: opracowanie własne.

Przykład 4.1

Badano występowanie korelacji między powierzchnią biurową a okresem działalności 30 firm deweloperskich, oferujących nieruchomości komercyjne. Obserwacje jednostkowe przedstawiono w tab. 4.4, podając: numer obserwacji (zastępujący nazwę firmy deweloperskiej) według kolejności uzyskania danych, okres działalności firmy w latach oraz powierzchnię biurową w m^2. Natomiast zbiorcze dane w postaci tablicy z danymi pogrupowanymi przedstawiono w tab. 4.5.

Tabela 4.4. Dane z badania firm deweloperskich

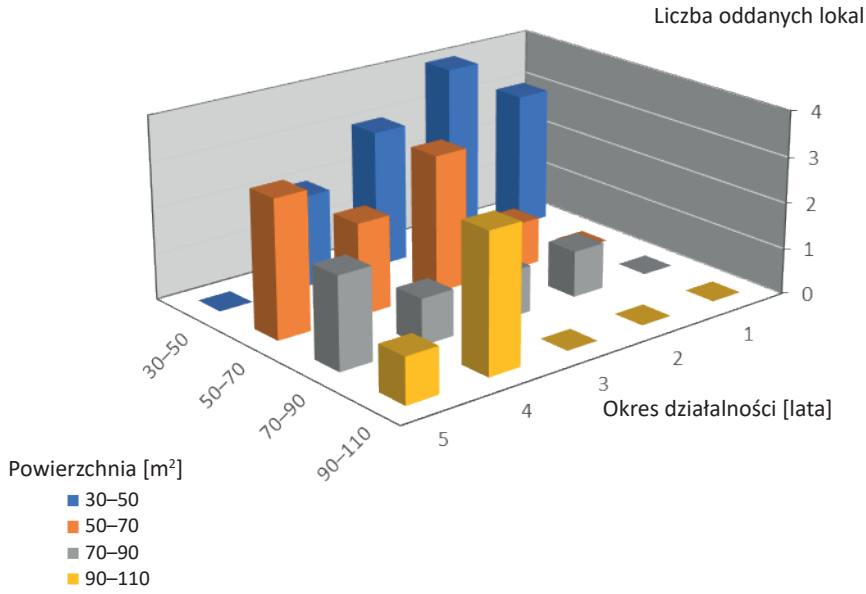
Nr firmy	Okres [Lata]	Powierzchnia [m^2]	Nr firmy	Okres [lata]	Powierzchnia [m^2]	Nr firmy	Okres [lata]	Powierzchnia [m^2]
1	4	42	11	3	58	21	4	64
2	2	48	12	4	72	22	4	56
3	1	37	13	5	96	23	2	30
4	2	56	14	1	38	24	4	93
5	3	46	15	5	64	25	1	49
6	4	102	16	3	75	26	3	66
7	4	33	17	5	68	27	3	56
8	5	74	18	3	46	28	4	104
9	5	63	19	2	74	29	3	43
10	2	42	20	5	85	30	2	39

Źródło: opracowanie własne.

Tabela 4.5. Dane pogrupowane

Powierzchnia użytkowa [m²] – cecha *X*	Okres działalności w latach – cecha *Y*					Razem
	1	2	3	4	5	
30–50	3	4	3	2	×	12
50–70	×	1	3	2	3	9
70–90	×	1	1	1	2	5
90–110	×	×	×	3	1	4
Razem	3	6	7	8	6	30

Źródło: opracowanie własne na podstawie tab. 4.4.

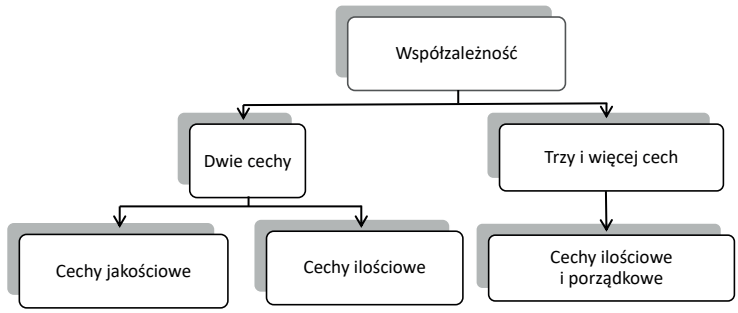


Rysunek 4.5. Dane z tab. 4.5

Źródło: opracowanie własne.

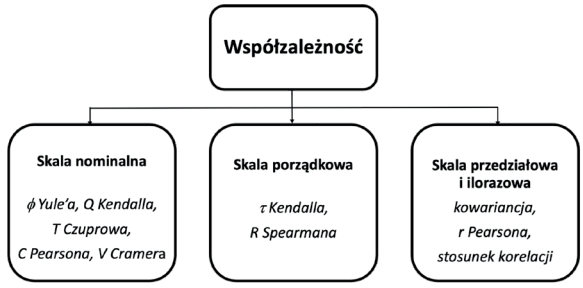
Z danych w tab. 4.5 wynika, że spośród badanych firm deweloperskich 3 działają od roku, 6 z nich funkcjonuje 2 i 5 lat, 7 przez 3 lata i 8 przez 4 lata. Najwięcej jest biur o małej powierzchni, tj. od 30 do 50 m² (w 12 firmach), a najmniej największych biur – o powierzchni 90–110 m². Dane z tab. 4.5 przedstawiono na trójwymiarowym wykresie (rys. 4.5).

Sposób pomiaru współzależności jest ściśle związany z charakterem danych, tzn. zależy od skali pomiarowej oraz od liczby cech uwzględnionych w badaniu, co przedstawiono na rys. 4.6 i 4.7. Jednakże, podobnie jak miało to miejsce przy analizie struktury, w przypadku cech jakościowych istnieje niewiele miar służących do ich pomiaru i miary te są mało „wyraziste". W przypadku analiz wielowymiarowych często pojawia się problem różnych skal używanych do pomiaru cech. Pojawia się wtedy wątpliwość, jakie miary współzależności zastosować. Regułą jest, że zawsze stosuje się miary, które są przeznaczone do najsłabszej skali, w jakiej dokonywany jest pomiar cech.



Rysunek 4.6. Badanie współzależności w zależności od skali pomiarowej i liczby cech

Źródło: opracowanie własne.



Rysunek 4.7. Miary współzależności dwóch cech w zależności od skal pomiarowych

Źródło: opracowanie własne.

Przy badaniu współzależności cech przyjmuje się zwykle jedną cechę za niezależną, której zmienność jest uwarunkowana czynnikami zewnętrznymi, a drugą za zależną, tzn. jej wahania próbuje się wyjaśnić (przynajmniej częściowo) zmiennością cechy niezależnej. Zazwyczaj pierwszą z nich oznacza się przez X, a drugą przez Y, chociaż często bada się relacje obustronne.

4.3. Podstawowe miary korelacji dwóch cech mierzalnych i porządkowych

Najczęściej stosowanym miernikiem współzależności cech mierzalnych (ilościowych) jest współczynnik korelacji liniowej Pearsona, określany w literaturze również jako współczynnik korelacji prostoliniowej. Opiera się on na pojęciu kowariancji, którą dla szeregu szczegółowego (tab. 4.1) wyznacza się według wzoru:

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{4.1}$$

lub równoważnie

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y} \tag{4.2}$$

oraz dla szeregu pogrupowanego w tablicy korelacyjnej (tab. 4.2) wyznacza się ze wzoru:

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}(\dot{x}_i - \overline{x})(\dot{y}_j - \overline{y})n_{ij} \tag{4.3}$$

lub

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}\dot{x}_i\,\dot{y}_j\,n_{ij} - \overline{x} \cdot \overline{y} \tag{4.4}$$

Współczynnik korelacji liniowej Pearsona dany jest wzorem:

$$r_{xy} = r_{yx} = \frac{\mathrm{cov}_{xy}}{S_x S_y} \tag{4.5}$$

gdzie:
x_i, y_i – i-te obserwacje wartości odpowiednio cech X i Y,
$\dot{x}_i$, $\dot{y}_i$ – środki przedziałów klasowych odpowiednio cech X i Y,
$\overline{x}$, $\overline{y}$ – średnie arytmetyczne cech X i Y,
S_x, S_y – odchylenia standardowe zmiennych X i Y,
n_{ij} – liczba obserwacji posiadających jednocześnie i-ty wariant cechy X i j-ty wariant cechy Y,
cov_{xy} – kowariancja cech X i Y,
r_{xy} – współczynnik korelacji liniowej Pearsona dla cech X i Y.

Współczynnik korelacji liniowej r_{xy} jest miarą symetryczną i przyjmuje wartości z przedziału <–1, 1>. Wskaźnik ten informuje o sile oraz kierunku korelacji liniowej między zmiennymi. Wartość współczynnika korelacji Pearsona $r_{xy} = 0$ świadczy o braku korelacji liniowej między badanymi cechami (jednak możliwe jest, że istnieje między nimi korelacja krzywoliniowa). Wartość współczynnika korelacji Pearsona większa od 0 ($r_{xy} > 0$) oznacza korelację dodatnią, tzn. wraz ze wzrostem wartości jednej cechy wzrasta średnia warunkowa drugiej cechy (porównaj rys. 4.3). Wartość współczynnika korelacji Pearsona mniejsza od 0 ($r_{xy} < 0$) informuje, że korelacja jest ujemna, tzn. wzrostowi wartości jednej cechy towarzyszy spadek drugiej. Przy $r_{xy} = 1$ lub –1 występuje liniowa zależność funkcyjna, tzn. $y = ax + b$ (tj. dodatnia, jak na rys. 4.1, i ujemna, jak na rys. 4.2).

Przykład 4.2

Dla danych z przykładu 4.1 obliczyć należy współczynnik korelacji liniowej.

Rozwiązanie

Tablica korelacyjna, zbudowana na podstawie danych pogrupowanych z tab. 4.5, została uzupełniona o dane konieczne do obliczenia S_x, S_y oraz cov_{xy} i przedstawiona w tab. 4.6. Średnie arytmetyczne cech na podstawie przedstawionych danych oblicza się jako:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{1820}{30} = 60,67 \qquad \overline{y} = \frac{\sum_{j=1}^{l} y_j n_j}{n} = \frac{98}{30} = 3,27$$

Kowariancja została obliczona według wzoru (4.4):

$$cov_{xy} = \frac{6380}{30} - 60,67 \cdot 3,27 = 14,49$$

Odchylenia standardowe oblicza się, podobnie jak średnie, na podstawie rozkładów warunkowych:

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 n_i} = \sqrt{\frac{13186,67}{30}} = 20,97$$

$$S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{l}(y_i - \overline{y})^2 n_j} = \sqrt{\frac{47,87}{30}} = 1,26$$

Zatem współczynnik korelacji liniowej jest równy:

$$r_{xy} = r_{yx} = \frac{cov_{xy}}{S_x \cdot S_y} = \frac{14,49}{20,97 \cdot 1,26} = 0,5484 \approx 0,55$$

Tabela 4.6. Obliczenia dla danych z przykładu 4.1

$\dot{x}_i$	Warianty cechy X	Warianty cechy Y					$\sum = n_{i\circ}$	$\dot{x}_i n_{i\bullet}$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 n_{i\bullet}$
		1	2	3	4	5					
40	[30; 50)	3	4	3	2	×	12	480	-20,67	427,11	5 125,33
60	[50; 70)	×	1	3	2	3	9	540	–0,67	0,44	4,00
80	[70; 90)	×	1	1	1	2	5	400	19,33	373,78	1 868,89
100	[90; 110)	×	×	×	3	1	4	400	39,33	1 547,11	6 188,44
$\sum = n_j$		3	6	7	8	6	30	1 820	×	×	13 186,67
$\dot{y}_j n_{\bullet j}$		3	12	21	32	30	98				
$y_j - \bar{y}$		–2,27	–1,27	–0,27	0,73	1,73	×				
$(y_j - \bar{y})$		5,14	1,60	0,07	0,54	3,00	×				
$(y_j - \bar{y})^2 n_{\bullet j}$		15,4	9,63	0,49	4,30	18,00	47,87				
$\sum_{i=1}^{k} \dot{x}_i n_{ij}$		120	300	380	580	440	×				
$\sum_{i=1}^{k} \dot{x}_i n_{ij} y_j$		120	600	1140	2320	2200	6380				

Źródło: obliczenia na podstawie tab. 4.5.

W tym przypadku cechy są wyraźnie dodatnio skorelowane, choć nie jest to silna współzależność. Z przeprowadzonych obliczeń wynika, że czym dłużej działa firma deweloperska, tym wynajmuje większe biuro.

Analiza przykładu prowadzi ponadto do istotnego spostrzeżenia – wartości mierników, podobnie jak to miało miejsce w przypadku obliczania miar opisowych, zależą od sposobu agregacji danych. I tak, wartość współczynnika korelacji dla danych niepogrupowanych, tj. danych z tab. 4.4, będzie się różnić od wartości wyznaczonej dla danych z tab. 4.6. Dzieje się tak ponieważ przy obliczeniu wartości średniej i odchylenia standardowego powierzchni wykorzysta się wzory (2.2) i (2.11) dla szeregów prostych, zamiast wcześniej użytych wzorów (2.4) i (2.13) dla szeregu rozdzielczego z przedziałami klasowymi wielojednostkowymi. W konsekwencji kowariancja zostanie wyznaczona na podstawie wzoru (4.1), zamiast wcześniej zastosowanego wzoru (4.3). A zatem $\bar{x}$ = 60,63; S_x = 20,27 i cov_{xy} = 17,7978, stąd r_{xy} = 0,5779. Jak widać, wyznaczone wskaźniki korelacji liniowej Pearsona różnią się od siebie i to już na drugim miejscu po przecinku.

Inną miarą korelacji jest stosunek korelacji e_{yx} i e_{xy}, nazywany też współczynnikiem korelacji nieliniowej Pearsona. Mierzy on siłę korelacji cech niezależnie od kształtu tej zależności, tj. zarówno wtedy, kiedy zależność jest liniowa, jak i wtedy, gdy jest nieliniowa. Wyznacza się go dla danych pogrupowanych, przy czym cecha zależna musi być cechą mierzalną. Natomiast cecha niezależna może być ilościowa lub jakościowa. Innymi słowy, współczynnik korelacji nieliniowej może być stosowany w przypadku, gdy jedna z cech mierzona jest na skalach słabych (zwłaszcza nominalnej) pod warunkiem, że zmienna ilościowa zostanie uznana za zmienną zależną. Stosunek korelacji cechy *Y* do *X* jest dany wzorem:

$$e_{xy} = \sqrt{\frac{S^2_{\bar{y}_i}}{S^2_y}} \tag{4.6}$$

gdzie $S^2_{\bar{y}_i}$ jest wariancją średnich warunkowych cechy *Y*, czyli:

$$S^2_{\bar{y}_i} = \frac{1}{n}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 n_{i\bullet} \tag{4.7}$$

natomiast $\bar{y}_i$ jest średnią warunkową cechy *Y* dla *i*-tego wariantu cechy *X*, czyli:

$$\bar{y}_i = \frac{\sum_{j=1}^{l} y_j n_{ij}}{n_{i\bullet}} \quad i = 1,2,\ldots,k \tag{4.8}$$

gdzie S^2_y jest wariancją ogólną cechy *Y*.

Analogicznie konstruuje się stosunek e_{xy} korelacji cechy *X* do cechy *Y*. Stosunek korelacji może przyjmować wartości z przedziału <0, 1>, co oznacza, że mierzy jedynie siłę związku między dwiema cechami, ale nie można na jego podstawie określić kierunku zależności. Na ogół jest to miara niesymetryczna, tzn. $e_{xy} \neq e_{yx}$ poza dwoma przypadkami: $e_{xy} = e_{yx} = 0$ (brak korelacji) oraz $e_{xy} = e_{yx} = 1$ (zależność funkcyjna między *Y* a *X*).

Istnieje zależność między e_{xy} i r_{xy} postaci:

$$|r_{xy}| \leq e_{xy} \tag{4.9}$$

która stała się podstawą do utworzenia miary krzywoliniowości związku zmiennych, tzw. miernika krzywoliniowości, który przy badaniu zależności cechy *Y* od *X* jest postaci:

$$m_y = e^2_{yx} - r^2_{yx} \tag{4.10}$$

a dla zależności cechy X od Y jest postaci:

$$m_x = e_{xy}^2 - r_{xy}^2 \tag{4.11}$$

Przykład 4.3

Czy zależność między powierzchnią biurową a długością okresu działalności firmy z przykładu 4.1 można uznać za liniową?

Rozwiązanie

Należy obliczyć stosunek korelacji nieliniowej e_{xy}, tj. zależności powierzchni wynajmowanych biur od okresu działalności firmy deweloperskiej, a następnie wielkość miernika krzywoliniowości (im jego wartość jest mniejsza, tym bardziej zależność można uznać za liniową).

Na podstawie danych z tab. 4.6 można obliczyć wartości średnich warunkowych cechy X – powierzchni biur, dla poszczególnych wariantów cechy Y – liczby lat funkcjonowania na rynku. Tak więc w grupie najmłodszych firm są tylko 3 jednostki i wszystkie mają powierzchnię biur w przedziale 30–50 m^2, zatem

$$\overline{x}_1 = \frac{\sum_{i=1}^{4} \dot{x}_i n_{i1}}{n_1} = \frac{40 \cdot 3 + 60 \cdot 0 + 80 \cdot 0 + 100 \cdot 0}{3} = 40$$

W następnej grupie – firm działających przez 2 lata, średnia powierzchnia biur wynosi $\overline{x}_2 = \frac{\sum_{i=1}^{4} \dot{x}_i n_{i2}}{n_2} = \frac{40 \cdot 4 + 60 \cdot 1 + 80 \cdot 1 + 100 \cdot 0}{6} = 50$ i analogicznie:

$\overline{x}_3 = 54{,}28571$; $\overline{x}_4 = 72{,}5$; $\overline{x}_5 = 73{,}3333$

Uwzględniając obliczoną w przykładzie 4.2 średnią 60,67, wariancję średnich warunkowych oblicza się ze wzoru (4.7) jako:

$$S_{\overline{x}_j}^2 = \frac{1}{n}\sum_{j=1}^{k}(\overline{x}_j - \overline{x})^2 n_j = \frac{1}{30}((40-60{,}67)^2 \cdot 3 + (50-60{,}67)^2 \cdot 6 +$$
$$+(54{,}28571-60{,}67)^2 \cdot 8 + (73{,}3333-60{,}67)^2 \cdot 6) = 144{,}3968$$

Wariancja cechy X na podstawie obliczeń z przykładu 4.2 wynosi $S_x^2 = (13186{,}67/30) = 439{,}5557$. Ostatecznie więc stosunek korelacji nieliniowej (4.6) jest równy:

$$e_{xy} = \sqrt{\frac{S_{\overline{x}_j}^2}{S_x^2}} = \sqrt{\frac{144{,}3968}{439{,}5557}} = 0{,}573155$$

a wartość miernika krzywoliniowości wynosi:

$$m_x = e_{xy}^2 - r_{xy}^2 = (0{,}573)^2 - 0{,}3 = 0{,}33 - 0{,}3 = 0{,}03$$

Ponieważ obliczona wartość m_x jest bardzo bliska 0, zatem zależność z przykładu 4.2 można uznać za liniową. W przykładzie występują dwie cechy mierzalne, co pozwala również wyznaczyć współczynnik korelacji nieliniowej Pearsona e_{xy}.

Dla $x_i = 60$ suma $\sum y_j n_{ij} = 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 3 + 4 \cdot 2 + 5 \cdot 3 = 34$ i średnia warunkowa to:

$$\overline{y}_i = \frac{34}{9} = 3{,}78$$

wariancja warunkowa wynosi:

$$S_{\overline{y}_i}^2 = \frac{23{,}01}{30} = 0{,}767$$

a stosunek korelacji na wartość

$$e_{xy} = \sqrt{\frac{(0.767)}{1{,}6}} = \sqrt{0{,}4795} = 0{,}69.$$

Współczynnik korelacji liniowej Pearsona obliczony w przykładzie 4.2 wyniósł $r_{yx} = 0{,}55$. Spełniony jest zatem warunek (4.9) $|r_{yx}| \leq e_{xy}$, a ponadto zachodzi obliczona wyżej relacja $e_{yx} \neq e_{xy}$. Miernik krzywoliniowości wynosi więc:

$$m_y = e_{yx}^2 - r_{yx}^2 = (0{,}69)^2 - 0{,}3 = 0{,}1761$$

Miarą korelacji, dedykowaną cechom porządkowym, jest współczynnik korelacji kolejnościowej (rang) Spearmana R_{xy}:

$$R_{xy} = R_{yx} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{n^3 - n} \tag{4.12}$$

gdzie:
d_i – różnice między kolejnymi numerami (rangami), nadawanymi w kolejności niemalejącej (lub nierosnącej) osobno dla każdej cechy od 1 do n. Jeżeli kilka elementów w szeregu ma taką samą wartość jednej cechy, to nadaje im się rangi będące średnią arytmetyczną (przypadających na te elementy) rang. Wartość R_{xy} należy do przedziału <–1, 1> i określa siłę oraz kierunek korelacji.

Przykład 4.4

Eksperci finansowi ocenili 14 przedsiębiorstw usługowych i uszeregowali je na pozycjach od 1 do 9, przypisując pozycję 1 najlepszej firmie i 9 najgorszemu zakładowi usługowemu. Każda z firm jest równolegle ewaluowana przez jej klientów, którzy dysponują ocenami od 2 do 5, gdzie 5 to ocena najwyższa. W tab. 4.7 zamieszczono informacje dotyczące obu ocen, tj. pozycji rankingowej nadanej przez ekspertów oraz średniej ocen wystawionych przez klientów. Należy zbadać, czy między tymi ocenami występuje związek i jaka jest jego siła.

Tabela 4.7. Oceny przedsiębiorstw usługowych

Nr firmy	Symbol firmy	Miejsce według oceny ekspertów	Średnia ocen klientów
1	A	1	4,4
2	B	2	3,8
3	C	2	4,0
4	D	3	4,5
5	E	3	3,4
6	F	3	4,2
7	G	4	3,8
8	H	5	4,2
9	I	5	3,3
10	J	5	3,8
11	K	6	4,1
12	L	7	3,9
13	M	8	3,5
14	N	9	3,2

Źródło: opracowanie własne – dane umowne.

Rozwiązanie

Ocena ekspertów i klientów to klasyczny przypadek pomiaru cechy na skali porządkowej. Do rozwiązania zadania należy zatem zastosować współczynnik korelacji rang Spearmana. Należy zwrócić uwagę, że miejsca firm nadane im przez ekspertów nie będą dokładnie równe rangom przypisanym tym miejscom, ponieważ wartości rang są przypisywane od 1 do liczby elementów w próbie, czyli 14, natomiast przyznano miejsca od 1 do 9. Oprócz tego w sytuacji, kiedy wartość cechy porządkowej się powtarza, rangi mogą przyjmować wartości ułamkowe. Przykładowo, firmy B i C mają tę samą ocenę ekspertów, ale zajmują pozycję drugą i trzecią, zatem aby nadać im tę samą rangę, każda z firm uzyska rangę będącą połową sumy pozycji, czyli $0{,}5 \cdot (2 + 3) = 2{,}5$. Z kolei firmy usytuowane na pozycjach czwartej, piątej i szóstej,

tj. D, E i F, uzyskały w ocenie ekspertów trzecią lokatę, zatem otrzymują one rangę równą średniej arytmetycznej przypadających na nich rang, czyli (4 + 5 + 6) / 3 = 5 itd. W przypadku opinii klientów najwyższa średnia ocena wynosi 4,5, a najniższa 3,2. Firmy usługowe zostały uporządkowane w taki sposób, że najgorzej oceniony zakład ma pozycję 1, a najlepiej oceniony ostatnią, tj. 14. Można zauważyć, że firmy F i H oraz B, G i J otrzymały takie same oceny, zatem ich pozycje rankingowe zostały ustalone odpowiednio jako pozycje jedenasta i dwunasta oraz piąta, szósta i siódma. Stąd wyznaczone dla tych firm jednolite rangi to 11,5 dla firm F i H oraz 6 dla firm B, G i J. Tab. 4.8 zawiera wszystkie niezbędne dane do obliczenia współczynnika korelacji kolejnościowej Spearmana, który zgodnie ze wzorem (4.12) wynosi:

$$R_{yx} = 1 - \frac{6 \cdot 658}{14^3 - 14} = -0,446$$

Istnieje zatem wyraźny związek między badanymi cechami: im wyższa średnia ocena klientów, tym wyżej ocenili przedsiębiorstwa eksperci. Zauważyć należy, że im większa liczba oznaczająca średnią, tym mniejsza liczba oznaczająca miejsce w rankingu, dlatego współczynnik korelacji jest ujemny.

Tabela 4.8. Obliczenia do danych z tab. 4.7

Symbol firmy	Miejsce według oceny ekspertów (x_i)	Średnia ocen klientów (y_i)	Rangi dla x (dx_i)	Rangi dla y (dy_i)	Różnica rang	
					(d_i)	d_i^2
A	1	4,4	1,0	13,0	–12,0	144,00
B	2	3,8	2,5	6,0	–3,5	12,25
C	2	4,0	2,5	9,0	–6,5	42,25
D	3	4,5	5,0	14,0	–9,0	81,00
E	3	3,4	5,0	3,0	2,0	4,00
F	3	4,2	5,0	11,5	–6,5	42,25
G	4	3,8	7,0	6,0	1,0	1,00
H	5	4,2	9,0	11,5	–2,5	6,25
I	5	3,3	9,0	2,0	7,0	49,00
J	5	3,8	9,0	6,0	3,0	9,00
K	6	4,1	11,0	10,0	1,0	1,00
L	7	3,9	12,0	8,0	4,0	16,00
M	8	3,5	13,0	4,0	9,0	81,00
N	9	3,2	14,0	1,0	13,0	169,00
					Suma	658,00

Źródło: opracowanie własne.

W praktycznych zastosowaniach często pojawia się dylemat dotyczący określenia granic dla wartości poszczególnych mierników korelacji umożliwiających uznanie danej zależności jako słabej, umiarkowanej lub silnej. Można oczywiście przyjąć za graniczne pewne wartości bezwzględne miernika, np. (0,0; 0,3) oznacza słabą zależność, (0,3; 06) – umiarkowaną oraz (0,6; 1,0) – silną. Jednakże takie podejście oznacza się znaczną dozą subiektywizmu, którego można uniknąć, posługując się testami statystycznymi pozwalającymi na stwierdzenie występowania (lub nie) statystycznie istotnej zależności między cechami. Hipotezę o liniowej niezależności cech, tj. braku korelacji, formułuje się w postaci:

$$H_0: \rho = 0 \tag{4.13}$$

wobec hipotezy alternatywnej wskazującej na występowanie istotnej korelacji, danej w jednej z postaci:

$$H_1: \rho \neq 0 \tag{4.14}$$

$$H_1: \rho > 0 \tag{4.15}$$

$$H_1: \rho < 0 \tag{4.16}$$

gdzie:
ρ korelacja między dwiema cechami.

Sprawdzianem tego testu jest statystyka t-Studenta o $(n - 2)$ stopniach swobody postaci:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{4.17}$$

gdzie:
r – współczynnik korelacji liniowej Pearsona r_{yx} ze wzoru (4.5) lub korelacji rang Spearmana R_{yx} ze wzoru (4.12) obliczone na podstawie obserwacji.

Badanie występowania istotnej korelacji można również przeprowadzić, uwzględniając stosunek korelacji. Wówczas hipoteza zerowa jest postaci (4.13), a alternatywna postaci (4.15), gdyż współczynnik przyjmuje jedynie wartości z przedziału [0; 1]. Sprawdzianem testu jest statystyka o rozkładzie Fishera-Snedecora o $(k - 1)$ i $(n - k)$ stopniach swobody (k oznacza liczbę klas zmiennej niezależnej X) postaci:

$$F = \frac{e_{yx}^2}{1-e_{yx}^2} \cdot \frac{n-k}{k-1} \tag{4.18}$$

gdzie:
e_{yx} – współczynnik korelacji nieliniowej Pearsona wyznaczony ze wzoru (4.6).

Odrzucenie (dla przyjętego poziomu istotności) hipotezy zerowej postaci (4.13) oznacza istotną statystycznie zależność korelacyjną.

Przykład 4.5

Należy sprawdzić, czy analizowana w przykładzie 4.1 zależność pomiędzy wielkością powierzchni biurowej a okresem działalności firm jest istotna statystycznie. Dla 30 obserwacji, w przykładzie 4.2 obliczono wartość współczynnika korelacji liniowej Pearsona, równą 0,55, a w przykładzie 4.3 – wartość stosunku korelacji równą 0,57.

Rozwiązanie

Z treści zadania wynika konieczność weryfikacji hipotezy zerowej o braku zależności między powierzchnią biurową a okresem działalności firmy, czyli:

$$H_0: \rho = 0$$

oraz hipotezy alternatywnej, że między tymi zjawiskami występuje dodatnia korelacja, czyli:

$$H_1: \rho > 0$$

Wartość sprawdzianu testu wyznacza się ze wzoru (4.17):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,55 \cdot \sqrt{30-2}}{\sqrt{1-0,55^2}} = \frac{2,9103}{0,8352} = 3,4847$$

Wartość krytyczna testu dla prawostronnego obszaru odrzucenia, poziomu istotności 0,05 i 28 stopni swobody wynosi 1,7011. Zatem odrzuca się hipotezę zerową na rzecz hipotezy alternatywnej i stwierdza, że istnieje statystycznie istotna korelacja liniowa między okresem działalności przedsiębiorstwa i powierzchnią wynajmowanego przez nie biura. Innymi słowy, na poziomie istotności 0,05 współczynnik korelacji Pearsona jest istotnie większy od 0.

W podobny sposób postępuje się, badając istotność stosunku korelacji, ale tym razem korzysta się ze wzoru (4.18):

$$F = \frac{e_{yx}^2}{1-e_{yx}^2} \cdot \frac{n-k}{k-1} = \frac{0,57^2}{1-0,57^2} \cdot \frac{30-4}{4-1} = \frac{0,3249}{0,6751} \cdot \frac{26}{3} = 0,4813 \cdot 8,6667 = 4,1709$$

Wartość krytyczna testu dla prawostronnego obszaru odrzucenia dla poziomu istotności 0,05 i 3 oraz 26 stopni swobody wynosi 2,9752. Zatem powinno

się odrzucić hipotezę zerową na rzecz hipotezy alternatywnej i należy stwierdzić, że istnieje statystycznie istotna korelacja między okresem działalności przedsiębiorstwa i powierzchnią wynajmowanego przez nie biura (lub inaczej stosunek korelacji jest istotnie większy od 0 na poziomie istotności 0,05).

Przykład 4.6

Należy sprawdzić, czy analizowana w przykładzie 4.4 zależność jest istotna statystycznie.

Rozwiązanie

Rozwiązując zadanie należy sformułować hipotezę zerową o braku zależności między ocenami ekspertów i klientów H_0: $\rho = 0$ oraz hipotezę alternatywną, że między tymi zjawiskami występuje ujemna korelacja H_1: $\rho < 0$, a ponadto obliczyć wartość sprawdzianu testu ze wzoru (4.17):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,446 \cdot \sqrt{14-2}}{\sqrt{1-(-0,446)^2}} = \frac{-1,545}{0,895} = -1,7262$$

Wartość krytyczna testu dla lewostronnego obszaru odrzucenia dla poziomu istotności 0,05 i 12 stopni swobody wynosi –1,7823, dlatego nie ma podstaw do odrzucenia hipotezy zerowej. Nie można zatem stwierdzić, że istnieje statystycznie istotna korelacja między opiniami klientów i ekspertów. Innymi słowy, na poziomie istotności 0,05 nie można twierdzić, że współczynnik korelacji rang Spearmana jest istotnie większy od 0.

4.4. Podstawowe miary współzależności dwóch cech niemierzalnych

W przypadku cech *stricte* jakościowych, tj. mierzonych na skali nominalnej badanie relacji między nimi dokonuje się za pomocą miar asocjacji (porównaj rys. 4.7), które wprawdzie pozwalają ocenić siłę zależności, ale nie pozwalają ocenić jej kierunku. Warto również zauważyć, na podstawie tych miar dość trudno

jest określić czy uzyskana wartość miernika wskazuje na silną czy tylko umiarkowaną zależność. Warto zatem korzystać z testu niezależności chi-kwadrat, który umożliwia weryfikację hipotez o występowaniu statystycznie istotnej korelacji między dwoma zjawiskami lub cechami i to niezależnie od skali pomiarowej, na jakiej mierzone są cechy.

Test niezależności chi-kwadrat (χ^2) jest testem nieparametrycznym pozwalającym na weryfikację hipotezy zerowej postaci:

$$H_0\text{: obie cechy są niezależne} \tag{4.19}$$

wobec hipotezy alternatywnej

$$H_1\text{: obie cechy są zależne.} \tag{4.20}$$

Test niezależności opiera się na tablicy 4.2, w której zamieszczono informacje dotyczące liczebności poszczególnych par wariantów obu zmiennych. Przypomnieć trzeba, że liczebności brzegowe wyznacza się na podstawie zależności:

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} \tag{4.21}$$

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \tag{4.22}$$

$$n = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} \tag{4.23}$$

gdzie:
n_{ij} – liczebność empiryczna i-tego wariantu cechy X i j-tego wariantu cechy Y,
$n_{i\bullet}$ – liczebność brzegowa i-tego warianty cechy X ($i = 1, 2, 3, \ldots, k$),
$n_{\bullet j}$ – liczebność brzegowa j-tego wariantu cechy Y ($j = 1, 2, 3, \ldots, l$).

W pierwszym kroku dla każdej kombinacji (i, j) cech X i Y wyznacza się tzw. liczebności teoretyczne (hipotetyczne) na podstawie liczebności brzegowych tablicy niezależności (4.2), które przy założeniu prawdziwości hipotezy zerowej są postaci:

$$\hat{n}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \tag{4.24}$$

gdzie:
$\hat{n}_{ij}$ – jest liczebnością klasy (i, j) przy teoretycznym założeniu niezależności cech X, Y.

Następnie oblicza się wartość statystyki testowej, która jest zmienną losową o rozkładzie chi-kwadrat o $[(k-1)\cdot(l-1)]$ stopniach swobody i jest postaci:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{4.25}$$

W przypadku tego testu obszar odrzucenia jest prawostronny, zatem jeśli wartość sprawdzianu jest większa od odczytanej z tablic rozkładu (dla zadanego poziomu istotności α) wartości krytycznej χ^2_α, to odrzuca się hipotezę zerową na rzecz hipotezy alternatywnej, co oznacza, że obie cechy lub oba zjawiska są współzależne.

Przykład 4.7

Na podstawie przeprowadzonych badań 1000 mikroprzedsiębiorstw, działających w wybranych 4 branżach, zbadano, czy generowanie zysków jest zależne od rodzaju działalności. Wyniki przeprowadzonych analiz zapisano w tab. 4.9. Należy sprawdzić, czy między wyróżnionymi cechami występuje istotna zależność na poziomie istotności 0,05 oraz 0,01.

Tabela 4.9. Dane: liczebności empiryczne: n_{ij}

Czy firma osiągnęła zysk?	Branża, w jakiej prowadzona jest działalność				Razem
	rolnictwo	handel	usługi	edukacja	$n_{i\bullet}$
Tak	110	500	150	40	800
Nie	40	100	50	10	200
Razem $n_{\bullet j}$	150	600	200	50	1000

Źródło: dane umowne.

Rozwiązanie

Rozpocząć trzeba od sformułowania pary hipotez:

H_0: zysk nie zależy od branży, w jakiej działają mikroprzedsiębiorstwa;

H_1: zysk zależy od branży, w jakiej działają mikroprzedsiębiorstwa.

Na podstawie danych z tab. 4.9 oszacować można liczebności teoretyczne $\hat{n}_{ij}$, korzystając ze wzoru (4.24), które zapisane zostaną w tab. roboczej 4.10. Przykładowo w pierwszej komórce znajdzie się wartość:

$$\hat{n}_{11} = \frac{n_{1\bullet}\cdot n_{\bullet 1}}{n} = \frac{800\cdot 150}{1000} = 120,$$

w kolejnej komórce:

$$\hat{n}_{21} = \frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{200 \cdot 150}{1000} = 30,$$

a w komórce (2, 2):

$$\hat{n}_{22} = \frac{n_{2\bullet} \cdot n_{\bullet 2}}{n} = \frac{200 \cdot 600}{1000} = 120.$$

Tabela 4.10. Liczebności teoretyczne $\hat{n}_{ij}$

Czy firma osiągnęła zysk?	Branża, w jakiej prowadzona jest działalność				Razem
	rolnictwo	handel	usługi	edukacja	$n_{i\bullet}$
Tak	120	480	160	40	800
Nie	30	120	40	10	200
Razem $n_{\bullet j}$	150	600	200	50	1000

Źródło: obliczenia na podstawie tab. 4.9.

Zauważyć należy, że suma wartości teoretycznych po wierszach i kolumnach w tab. 4.10 jest zgodna z wartościami z tab. 4.9. W kolejnych etapach wyznaczone zostaną w tabelach wartości niezbędne do obliczenia statystyki testowej (4.25). Zatem w tab. 4.11 przedstawiono obliczone różnice między liczebnościami empirycznymi i teoretycznymi: $(n_{ij} - \hat{n}_{ij})$, a tab. 4.12 przedstawia kwadraty tych różnic podzielone przez liczebności teoretyczne.

Tabela 4.11. Różnice między liczebnościami empirycznymi i teoretycznymi: $(n_{ij} - \hat{n}_{ij})$

Czy firma osiągnęła zysk?	Branża, w jakiej prowadzona jest działalność				Razem
	rolnictwo	handel	usługi	edukacja	$n_{i\bullet}$
Tak	–10	20	–10	0	0
Nie	10	–20	10	0	0
Razem $n_{\bullet j}$	0	0	0	0	0

Źródło: obliczenia na podstawie tab. 4.9 i 4.10.

Tabela 4.12. Ilorazy $(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

Czy firma osiągnęła zysk?	Branża, w jakiej prowadzona jest działalność				Razem
	rolnictwo	handel	usługi	edukacja	$n_{i\bullet}$
Tak	0,83	0,83	0,63	0,00	2,29
Nie	3,33	3,33	2,50	0,00	9,17
Razem $n_{\bullet j}$	4,17	4,17	3,13	0,00	11,46

Źródło: obliczenia na podstawie tab. 4.10 i 4.11.

W ostatniej rubryce tab. 4.12 znajduje się wartość poszukiwanej statystyki testowej, czyli $\chi^2 = 11{,}46$. Z tablic rozkładu χ^2 odczytać należy wartość krytyczną dla poziomu istotności $\alpha = 0{,}01$ i $\alpha = 0{,}05$ oraz $(2-1)(4-1) = 3$ stopni swobody. Wartości te wynoszą odpowiednio $\chi^2_\alpha = 11{,}3449$ i $\chi^2_\alpha = 7{,}8147$, co prowadzi do wniosku o konieczności odrzucenia hipotezy zerową na rzecz hipotezy alternatywnej dla obu poziomów istotności. W ten sposób wykazano, że osiąganie zysków przez mikroprzedsiębiorstwa jest związane z branżą, w jakiej one działają.

Przedstawiony powyżej przykład wykazał, że test niezależności chi-kwadrat może być z powodzeniem stosowany dla cech mierzonych na skali nominalnej, co oznacza, że można go również wykorzystywać w przypadku zjawisk mierzonych na dowolnych skalach i w dowolnej ich kombinacji. Statystyka testowa χ^2 jest również podstawą do wyznaczenia kilku znanych miar asocjacji.

Przykładem miar asocjacji jest współczynnik zbieżności Czuprowa, który jest miarą zależności stochastycznej cech. Porównuje on bowiem dwuwymiarowy rozkład empiryczny z rozkładem uzyskanym na podstawie rozkładów brzegowych cech i przy założeniu niezależności cech (tzn. równomierności rozkładów warunkowych). Współczynnik zbieżności Czuprowa jest dany wzorem

$$T_{xy} = T_{yx} = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(l-1)}}} \tag{4.26}$$

gdzie:
χ^2 – wartość statystyki obliczona ze wzoru (4.25),
k, l – liczba wariantów odpowiednio cechy X i Y,
n – liczba obserwacji.

Innymi popularnymi miarami asocjacji są:

- współczynnik kontyngencji Youle'a postaci:

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{4.27}$$

- współczynnik kontyngencji Cramera postaci:

$$V = \sqrt{\frac{\chi^2}{n(\varpi - 1)}} \tag{4.28}$$

gdzie:
ϖ – mniejszy z wymiarów tablicy kontyngencji, czyli $\min\{k, l\}$.

Wyznaczenie współczynników Czuprowa i Youle'a wymaga danych pogrupowanych, a oba mierniki należą do przedziału [0; 1].

Przykład 4.8

Przeprowadzone badanie 360 punktów sprzedaży na terenie województwa łódzkiego (z pominięciem Łodzi) dotyczyło zarówno cech ilościowych, jak i jakościowych. Spośród cech mierzonych na skali nominalnej uwzględniono m.in. rodzaj punktu sprzedaży (X) oraz typ własności (Y). Na podstawie danych z ankiet określono po 7 wariantów każdej z cech. Wyróżnione przez ankietowanych rodzaje punktu sprzedaży to:

a) hurtownia specjalistyczna,
b) hurtownia wielobranżowa,
c) specjalistyczny punkt sprzedaży detalicznej,
d) wielobranżowy punkt sprzedaży detalicznej,
e) stoisko w domu handlowym,
f) inne.

Natomiast typ własności występuje w następujących wariantach:

a) spółdzielczy,
b) spółka z o.o.,
c) spółka akcyjna,
d) spółka cywilna,
e) firma prywatna,
f) firma prywatna – rodzinna.

Po uporządkowaniu danych i zliczeniu wszystkich obserwacji wyniki badania przedstawiono w tablicy kontyngencji (tab. 4.13). Zbadać należy, czy występuje istotna zależność między rodzajem punktu sprzedaży i formą własności, obliczając współczynniki Czuprowa, Youle'a i Cramera.

Tabela 4.13. Dane pogrupowane w tablicy niezależności – liczebności empiryczne n_{ij}

		Rodzaj własności y_j						Liczebność brzegowa $n_{i\bullet}$
		a	b	c	d	e	f	
Rodzaj punktu sprzedaży x_i	a	1	0	0	6	0	8	15
	b	0	0	1	1	0	2	4
	c	27	4	0	6	38	135	210
	d	31	8	0	12	27	49	127
	e	1	0	0	0	0	1	2
	f	0	0	0	0	0	2	2
Liczebność brzegowa $n_{\bullet j}$		60	12	1	25	65	197	360

Źródło: opracowanie własne na podstawie [Witkowska, Witkowski 1997b].

Rozwiązanie

W pierwszym kroku zadanie sprowadza się do rozstrzygnięcia hipotez (4.19)–(4.20) testu niezależności chi-kwadrat poprzez obliczenie statystyki testowej. Zgodnie z podanymi wzorami obliczenia rozpoczyna się od wyznaczenia teoretycznych liczebności klas $\hat{n}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. Przy założeniu, że obie cechy są od siebie niezależne, otrzymuje się ciąg liczebności teoretycznych $\hat{n}_{ij}$, którymi uzupełnia się tab. 4.14:

$$\hat{n}_{11} = \frac{n_{1\bullet} n_{\bullet 1}}{n} = \frac{15 \cdot 60}{360} = 2{,}5,\ \hat{n}_{12} = \frac{n_{1\bullet} n_{\bullet 2}}{n} = \frac{15 \cdot 12}{360} = 0{,}5,\ \hat{n}_{13} = \frac{n_{1\bullet} n_{\bullet 3}}{n} = \frac{15 \cdot 1}{360} = \frac{1}{24} \text{ itd.},$$

Tabela 4.14. Liczebności teoretyczne $\hat{n}_{ij}$

x_i / y_j	a	b	c	d	e	f	$n_{i\bullet}$
a	2,5	0,5	$\frac{1}{24}$	$\frac{25}{24}$	$\frac{65}{24}$	$\frac{197}{24}$	15
b	$\frac{2}{3}$	$\frac{2}{15}$	$\frac{1}{90}$	$\frac{5}{18}$	$\frac{13}{18}$	$\frac{197}{90}$	4
c	35	7	$\frac{7}{12}$	$\frac{175}{12}$	$\frac{445}{12}$	$\frac{1379}{12}$	210
d	$\frac{127}{6}$	$\frac{127}{30}$	$\frac{127}{360}$	$\frac{635}{72}$	$\frac{1651}{72}$	$\frac{29019}{360}$	127
e	$\frac{1}{3}$	$\frac{1}{15}$	$\frac{1}{180}$	$\frac{5}{36}$	$\frac{13}{36}$	$\frac{197}{180}$	2
f	$\frac{1}{3}$	$\frac{1}{15}$	$\frac{1}{180}$	$\frac{5}{36}$	$\frac{13}{36}$	$\frac{197}{180}$	2
$n_{\bullet j}$	60	12	1	25	65	197	360

Źródło: opracowanie własne na podstawie tab. 4.13.

Do wyznaczenia wartości statystyki χ^2 należy jeszcze obliczyć wartości $\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$, czyli kolejno:

$$\frac{(n_{11}-\hat{n}_{11})^2}{\hat{n}_{11}}=\frac{(1-2,5)^2}{2,5}=\frac{2,25}{2,5}=0,9$$

$$\frac{(n_{12}-\hat{n}_{12})^2}{\hat{n}_{12}}=\frac{(0-0,5)^2}{0,5}=2 \text{ itd.}$$

Następnie, sumując obliczone wartości według wzoru (4.25), otrzymuje się: $\chi^2 \approx 151,2$. Biorąc pod uwagę, że dla poziomu istotności 0,05 oraz (6 – 1)(6 – 1) = 25 stopni swobody wartość krytyczna wynosi χ^2_α = 37,6525, stwierdzić należy, że między rodzajem punktu sprzedaży i formą własności zależność występuje statystycznie istotna. Jeśli tak, to warto zbadać siłę tej zależności analizując wartości współczynników Czuprowa, Youle'a i Cramera (4.26)–(4.28).

Miara	Wzór	Obliczenia
Czuprowa	$T=\sqrt{\frac{\chi^2}{n\sqrt{(k-1)(l-1)}}}$	$T=\sqrt{\frac{151,2}{360\sqrt{(6-1)(6-1)}}}=\sqrt{\frac{151,2}{1800}}\approx 0,29$
Youle'a	$\phi=\sqrt{\frac{\chi^2}{n}}$	$\phi=\sqrt{\frac{151,2}{360}}=\sqrt{0,42}\approx 0,65$
Cramera	$V=\sqrt{\frac{\chi^2}{n(\varpi-1)}}$	$V=\sqrt{\frac{151,2}{360\cdot 6}}=\sqrt{0072}\approx 0,26$

Z przedstawionych obliczeń wynika, że wartości obliczonych współczynników różnią się od siebie. Można jednak stwierdzić, że istnieje zależność między rodzajem punktu sprzedaży a typem własności, jednakże na podstawie wyliczonych wartości mierników asocjacji trudno jest wyrokować, czy jest ona silna czy słaba. Dlatego ważne jest stwierdzenie, że ta zależność jest statystycznie istotna, co sprawdzono w procedurze weryfikacji hipotez statystycznych (tj. testu niezależności χ^2) na podstawie obliczonej statystyki χ^2.

Przykład 4.9

Badaniem objęto 9 towarzystw ubezpieczeniowych, które scharakteryzowano za pomocą 9 cech (tab. 4.15). Tab. 4.16 stanowi macierz danych zaobserwowanych dla każdego towarzystwa ubezpieczeniowego. Należy sprawdzić, czy wyróżnione charakterystyki badanych towarzystw ubezpieczeniowych są ze sobą skorelowane.

Tabela 4.15. Lista i symbole towarzystw ubezpieczeniowych wraz z listą zmiennych

Towarzystwo ubezpieczeniowe		Zmienne uwzględnione w badaniu	
Nazwa	Symbol	Opis	Symbol
ERGO-HESTIA	ERG	Wynik finansowy netto [tys. PLN]	X_1
FILAR	FIL	Liczba grup ubezpieczeniowych	X_2
GENERALI	GEN	Udział zakładu w rynku ubezpieczeń komunikacyjnych [%]	X_3
PZU	PZU	Kapitały własne [tys. PLN]	X_4
TRYG	TRY	Kapitały podstawowe [tys. PLN]	X_5
UNIQA	UNI	Składka zarobiona na udziale własnym [tys. PLN]	X_6
WARTA	WAR	Składka przypisana brutto [tys. PLN]	X_7
COMPENSA	COM	Świadczenia i odszkodowania wypłacone brutto [tys. PLN]	X_8
LINK4	LIN	Akcjonariusze zagraniczni [%]	X_9

Źródło: opracowanie własne.

Tabela 4.16. Czynniki charakteryzujące towarzystwa ubezpieczeniowe

Towa-rzystwo	Zmienne								
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
ERG	40 690	18	4,57	734 700	167 481	548 291	832 714	534 984	99,40
FIL	700	13	0,61	63 800	73 140	120 769	178 859	82 193	88,71
GEN	–7	13	0,51	38 470	29 000	31 867	77 263	44 894	100,00
PZU	1 248 320	18	64,31	3 551 080	86 352	6 415 295	7 394 195	4 471 756	20,91
TRY	12 000	14	2,40	123 649	55 000	157 578	274 853	142 223	94,14
UNI	–11 260	15	2,96	59 650	95 920	136 392	324 589	191 432	99,16
WAR	112 080	17	11,62	774 120	81 277	1 377 874	1 789 270	999 640	32,58
COM	–12 390	13	1,24	6 186	105 988	80 373	145 240	128 533	99,73
LIN	–35 536	9	0,15	83 000	38 000	44 534	20 165	12 189	100,00

Źródło: opracowanie na podstawie danych z Rocznika Ubezpieczeń i Funduszy Emerytalnych 2000–2002 KNUiFE oraz sprawozdań finansowych dla poszczególnych zakładów ubezpieczeń za lata 2000–2002.

Rozwiązanie

Badanie polega na wyznaczeniu wartości współczynników korelacji liniowej Pearsona między wszystkimi parami zmiennych X_1, X_2, …, X_9 na podstawie obserwacji pochodzących z dziewięciu towarzystw ubezpieczeniowych. Innymi słowy, dla każdej pary zmiennych X_i oraz X_j wyznaczono współczynniki korelacji, któ-

rych jest 81. Jednakże z uwagi na symetryczność współczynników korelacji oraz liniową zależność funkcyjną dla $i = j$ (kiedy $r_{yx} = 1$), można się spodziewać jedynie 36 różnych wartości współczynników korelacji, co jest widoczne w tab. 4.17, która zawiera współczynniki korelacji liniowej Pearsona między parami wszystkich analizowanych cech[7].

Tabela 4.17. Współczynnik korelacji cech charakteryzujących towarzystwa ubezpieczeniowe

	Zmienne								
Zmienne	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1,000	0,524	0,997	0,983	0,077	0,994	0,990	0,992	–0,782
X_2	0,524	1,000	0,559	0,620	0,655	0,570	0,597	0,595	–0,594
X_3	0,997	0,559	1,000	0,989	0,100	0,999	0,998	0,998	–0,818
X_4	0,983	0,620	0,989	1,000	0,192	0,992	0,993	0,994	–0,815
X_5	0,077	0,655	0,100	0,192	1,000	0,108	0,131	0,141	–0,021
X_6	0,994	0,570	0,999	0,992	0,108	1,000	0,999	0,999	–0,839
X_7	0,990	0,597	0,998	0,993	0,131	0,999	1,000	1,000	–0,847
X_8	0,992	0,595	0,998	0,994	0,141	0,999	1,000	1,000	–0,835
X_9	–0,782	–0,594	–0,818	–0,815	–0,021	–0,839	–0,847	–0,835	1,000

Źródło: obliczenia własne przy użyciu programu Excel.

Najsilniejszą zależność między zmiennymi widać w przypadku zmiennych X_1 a X_3, X_4, X_6, X_7, X_8, X_9. Zatem można zauważyć, że wraz ze wzrostem wyniku finansowego następuje wzrost kapitałów własnych, udziału zakładów ubezpieczeń w rynku ubezpieczeń komunikacyjnych, składki przypisanej brutto, składki zarobionej na udziale własnym, świadczeń oraz odszkodowań, jak również zmniejsza się wysokość akcjonariatu zagranicznego. Kolejno największą zależność między badanymi cechami widać pomiędzy cechą X_3 a X_4, X_6, X_7, X_8, X_9, co oznacza, że wraz ze wzrostem udziału towarzystw ubezpieczeniowych w rynku ubezpieczeń komunikacyjnych wzrasta przeciętna wysokość kapitałów własnych, liczba składek przypisanych brutto, składek zarobionych na udziale własnym, świadczeń i odszkodowań wypłacanych brutto, natomiast maleje przeciętnie wysokość inwestycji zagranicznych. W dalszej kolejności wysokie wartości współczynników korelacji obserwuje się między zmienną X_4 a zmiennymi: X_6, X_7, X_8, X_9, co pozwala zauważyć, że wyższym wartościom kapitałów własnych towarzyszy wyższy przeciętny poziom liczby składek przypisanych brutto, składek zarobionych na udziale własnym świadczeń i odszkodowań, lecz niższa jest przeciętna wysokość udziału akcjonariatu zagranicznego. W przypadku składki

7 Są to wartości nad lub pod przekątną utworzoną z jedynek.

zarobionej na udziale własnym widać, że wyższym jej wartościom towarzyszy wyższy przeciętnie poziom świadczeń, odszkodowań i spadek udziału akcjonariatu zagranicznego. Ciekawą rzeczą jest fakt, iż w przypadku zmiennych takich jak świadczenia i odszkodowania wypłacane przez towarzystwo ubezpieczeniowe oraz składki zebranej brutto występuje liniowa zależność funkcyjna. Bardzo silnie skorelowane są również cechy X_7 i X_9 oraz X_8 i X_9, dla których współczynnik korelacji oznacza, że wzrostowi świadczeń i odszkodowań towarzyszy przeciętny spadek wysokości akcjonariatu zagranicznego, jak również wzrostowi składki przypisanej brutto towarzyszy przeciętny spadek udziału akcjonariatu zagranicznego.

4.5. Analiza regresji

Badanie występowania jednoczesnej współzależności między więcej niż dwiema cechami ilościowymi można przeprowadzić za pomocą tzw. analizy regresji, która służy do badania relacji między tzw. cechą zależną i cechami niezależnymi. W tym celu buduje się tzw. model regresji. Można wyróżnić kilka etapów analizy regresji:

- zdefiniowanie zmiennej zależnej i zmiennych niezależnych,
- zebranie danych statystycznych,
- estymacja równania regresji,
- weryfikacja statystyczna modelu regresji,
- wnioskowanie o występowaniu zależności.

Ogólną postać modelu regresji można zapisać jako:

$$y_i = f(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k, x_{1i}, x_{2i}, \ldots, x_{ki}) + \varepsilon_i \tag{4.29}$$

Rozważania w tym podrozdziale zostaną ograniczone do liniowej postaci tej funkcji, którą można zapisać jako:

$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + \ldots + \alpha_k \cdot x_{ki} + \varepsilon_i \tag{4.30}$$

gdzie dla każdego $i = 1, 2, \ldots, n$:
y_i – obserwacje zmiennej zależnej,
$x_{1i}, x_{2i}, \ldots, x_{ki}$ – obserwacje kolejnych j-tych ($j = 1, 2, \ldots, k$) zmiennych niezależnych,
$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$ – nieznane parametry strukturalne równania regresji,
ε_i – składnik losowy.

Przedstawiony zapis (4.29) lub (4.30) oznacza, że zostało wyróżnionych k cech mierzalnych, opisanych za pomocą zmiennych niezależnych (x_1, x_2, …, x_k), które mają wpływ na kształtowanie się zmiennej zależnej y, wyrażającej opisywaną cechę ilościową. Obecność w równaniach regresji (4.29) i (4.30) składnika losowego ε_i tłumaczy się kilkoma przyczynami. Przede wszystkim występujące w zjawiskach społecznych zależności mają zazwyczaj charakter stochastyczny, co oznacza, że nie każda jednostka badania reaguje w identyczny sposób. Może również okazać się, że charakter analizowanego związku między wyróżnionymi cechami nie jest liniowy, w modelu nie uwzględniono wszystkich zmiennych, a także pomiar zmiennych obarczony jest błędem.

Parametry funkcji regresji α_0, α_1, α_2, …, α_k nie są znane (obserwowane), podobnie jak składnik losowy, dlatego równania (4.29) i (4.30) są równaniami stochastycznymi. W celu zbadania wpływu k wyróżnionych zmiennych niezależnych (pełniących w równaniu regresji rolę zmiennych objaśniających) na zmienną zależną (tj. objaśnianą przez model) należy oszacować parametry strukturalne równania regresji (4.30). W wyniku estymacji modelu otrzymuje się równanie:

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_{1i} + \hat{\alpha}_2 \cdot x_{2i} + \ldots + \hat{\alpha}_k \cdot x_{ki} \tag{4.31}$$

gdzie:
$\hat{y}_i$ – wartości teoretyczne zmiennej objaśnianej (zależnej),
$\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, …, $\hat{\alpha}_k$ – oszacowane parametry strukturalne modelu (oceny estymatorów parametrów),
a pozostałe oznaczenia jak poprzednio.

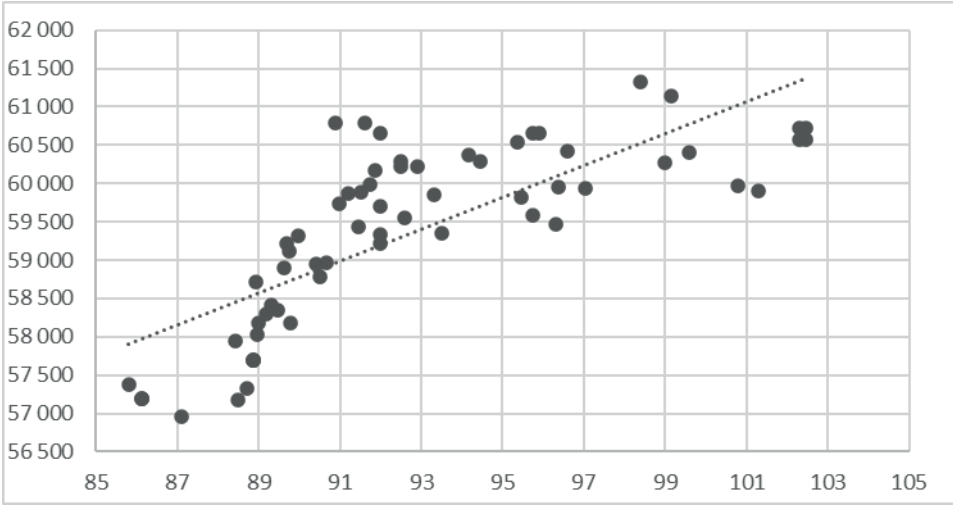
Oceny estymatorów parametrów funkcji regresji $\hat{\alpha}_j$ (j = 1, 2, …, k) informują, jak jednostkowa zmiana j-tej zmiennej niezależnej (objaśniającej) wpłynie na wartość zmiennej zależnej (objaśnianej przez model) przy założeniu, że pozostałe zmienne pozostaną na tym samym poziomie. Innymi słowy, jeżeli wartość zmiennej x_j wzrośnie o jednostkę, to przy założeniu stałego poziomu pozostałych zmiennych objaśniających, wartość zmiennej zależnej y wzrośnie (lub zmaleje) średnio o $\hat{\alpha}_j$. Wyrazu wolnego często się nie interpretuje, chyba że można go traktować jako pewien początkowy stan badanej zmiennej zależnej, tj. poziom, jaki przyjęłaby ta zmienna przy zerowych wartościach zmiennych objaśniających.

Jak łatwo zauważyć, w oszacowanym równaniu (4.31) brak jest składnika losowego, a oceny estymatorów parametrów $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, …, $\hat{\alpha}_k$ są liczbami rzeczywistymi. Zatem oszacowane równanie jest deterministyczne. Opisuje ono zależność funkcyjną występującą między k zmiennymi objaśniającymi x_1, x_2, …, x_k a wartością teoretyczną zmiennej zależnej $\hat{y}_i$. Można zatem powiedzieć, że szacując parametry stochastycznego równania regresji postaci (4.30), otrzymuje się równanie deterministyczne (4.31), w którym znane są parametry (będące ocenami estymatorów nieznanych parametrów), a zależność między wyróżnionymi zmiennymi

jest funkcyjna. Należy jednak zauważyć, że po lewej stronie równania (4.31) nie występują zaobserwowane wartości zmiennej zależnej y_i, ale jej wartości teoretyczne $\hat{y}_i$, wyznaczone na podstawie oszacowanego modelu regresji.

Przykład 4.10

Wykorzystując obserwacje notowań cen akcji spółki KGHM i indeksu giełdowego WIG, oszacowano liniową funkcję regresji[8]. Relacje między wartościami empirycznymi zaobserwowanymi w okresach $t = 1, 2, \ldots, T$, czyli wartościami y_t i wartościami teoretycznymi zmiennej zależnej wyznaczonymi na podstawie oszacowanego modelu z jedną zmienną niezależną, postaci: $\hat{y}_t = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_{1t}$ przedstawiono na rys. 4.8.



Rysunek 4.8. Porównanie wartości empirycznych i teoretycznych na przykładzie funkcji regresji opisującej liniową zależność między indeksem giełdowym WIG i notowaniami akcji KGHM w okresie 24.01–28.02.2019

Źródło: opracowanie własne.

Jak widać na rys. 4.8, dla wielu obserwacji wartości teoretyczne znajdujące się na prostej regresji różnią się od wartości zaobserwowanych, reprezentowanych przez punkty na wykresie. Różnicę między wartościami teoretycznymi i empirycznymi nazywa się resztą empiryczną i jest ona miernikiem błędu, jaki popełnia się, stosując do opisu stochastycznej relacji postaci (4.30) zależność funkcyjną (4.31).

8 Model został oszacowany przy wykorzystaniu funkcji REGLINP w Excelu.

Reszty empiryczne wyznacza się dla każdej obserwacji jako:

$$e_i = y_i - \hat{y}_i \tag{4.32}$$

i są one traktowane jako oceny estymatora składnika losowego ε_i.

W procesie estymacji podstawowym problemem jest zdefiniowanie odpowiedniego estymatora nieznanych parametrów równania regresji, za pomocą którego wyznacza się wartości ocen parametrów α_j. Najczęściej stosowaną metodą estymacji jest metoda najmniejszych kwadratów (MNK), u podstaw której leży minimalizacja sumy kwadratów błędów e_i, stąd nazwa tej metody. Innymi słowy, poszukuje się takich estymatorów nieznanych parametrów strukturalnych, aby uzyskać jak najmniejsze błędy, czyli wyznacza się takie wartości $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, …, $\hat{\alpha}_k$, dla których osiągnie minimum funkcja postaci:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_{1i} + \hat{\alpha}_2 \cdot x_{2i} + \ldots + \hat{\alpha}_k \cdot x_{ki}) \right]^2 \tag{4.33}$$

Estymację nieznanych parametrów przeprowadza się na podstawie danych statystycznych opisujących n obserwacji zmiennej zależnej i zmiennych niezależnych[9]. Otrzymanie zadowalających oszacowań parametrów modelu regresji za pomocą MNK wymaga spełnienia tzw. założeń Gaussa-Markowa, które sformułowano poniżej.

1. Zmienne niezależne są nielosowe.
2. Składniki losowe dla $i = 1, 2, \ldots, n$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym o:
   2a. wartości oczekiwanej równej 0, czyli: $E(\varepsilon_i)=0$,
   2b. stałej wariancji, czyli: $E(\varepsilon_i, \varepsilon_j) = \sigma^2$ dla $i = j$,
   2c. zerowej kowariancji, czyli: $E(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$.
3. Informacje dotyczące obserwacji zmiennej zależnej $\{y_i\}$ i zmiennych niezależnych $\{x_{1i}, x_{2i}, \ldots, x_{ki}\}$ są jedynymi informacjami, na podstawie których oszacowana zostanie funkcja regresji.
4. Nie występuje współliniowość między zmiennymi niezależnymi, a liczba szacowanych parametrów jest mniejsza od liczby obserwacji.

9 W niniejszym opracowaniu ograniczono się do przedstawienia wybranych zagadnień pozwalających na zrozumienie podstawowych kwestii i interpretację uzyskanych wyników. Pominięto natomiast dokładne omówienie zagadnień estymacji za pomocą metody najmniejszych kwadratów, które można znaleźć w wielu podręcznikach m.in. [Aczel 2000; Luszniewicz, Słaby 2003; Witkowska 2012]. Przyjęto bowiem założenie, że przy rozwiązywaniu zagadnień praktycznych korzysta się z profesjonalnego oprogramowania, tj. STATISTICA, STATA, SPSS, SAS, GRETL lub przynajmniej z popularnego pakietu Excel. W związku z tym nie są potrzebne żadne wzory pozwalające wyznaczyć oceny estymatorów parametrów strukturalnych funkcji regresji.

W przypadku spełnienia założeń Gaussa-Markowa (1)–(3) estymatory parametrów modelu oszacowane klasyczną MNK są zgodne, nieobciążone i najbardziej efektywne w danej klasie estymatorów, zatem spełniają wymogi „dobrych" estymatorów. Założenie (4) określa warunek konieczny wykonalności działań w celu oszacowania parametrów funkcji regresji za pomocą metody najmniejszych kwadratów.

Oprócz ocen estymatorów parametrów szacuje się tzw. standardowe błędy szacunku, czyli średnie błędy estymatora $S(\alpha_j)$, które informują o średnim odchyleniu wartości oszacowanej oceny estymatora $\hat{\alpha}_j$ od rzeczywistej (nieznanej) wartości parametru strukturalnego α_j. Standardowe błędy szacunku wykorzystuje się do budowy przedziałów ufności nieznanych parametrów funkcji regresji i weryfikacji hipotezy o istotności parametrów modelu regresji. Przedział ufności dla parametrów funkcji regresji jest postaci:

$$P(\hat{\alpha}_j - t_\alpha \cdot S(\alpha_j) < \alpha_j < \hat{\alpha}_j + t_\alpha \cdot S(\alpha_j)) = 1 - \alpha \tag{4.34}$$

gdzie:
t_α – wartość statystyki t-Studenta o $n - (k + 1)$ stopniach swobody.

Natomiast weryfikacja modelu polega m.in. na sprawdzeniu, czy użyte w modelu zmienne niezależne (objaśniające) mają statystycznie istotny wpływ na zmienną objaśnianą. Test opiera się na zestawie hipotez:

$$H_0: \alpha_j = 0 \tag{4.35}$$

$$H_1: \alpha_j \neq 0 \tag{4.36}$$

$$H_1: \alpha_j > 0 \tag{4.37}$$

$$H_1: \alpha_j < 0 \tag{4.38}$$

a jego sprawdzianem jest statystyka postaci:

$$t = \frac{\hat{\alpha}_j}{S(\alpha_j)} \tag{4.39}$$

Statystyka (4.39) ma rozkład t-Studenta o $[n - (k + 1)]$ stopniach swobody, o ile w modelu występuje wyraz wolny lub o $(n - k)$ stopniach swobody w modelach bez wyrazu wolnego.

Sposób sformułowania hipotezy alternatywnej jednoznacznie określa położenie obszaru odrzucenia, a poziom istotności umożliwia odczytanie wartości krytycznych z tablic rozkładu t-Studenta. Odrzucenie hipotezy zerowej pozwala twierdzić, że badana zmienna x_j istotnie wpływa na zmienną y, ponieważ parametr przy niej

stojący istotnie różni się od zera. Weryfikacja poprawności modelu obejmuje również sprawdzenie stopnia dokładności opisu zmiennej zależnej przez model oraz weryfikację merytoryczną modelu. Badanie stopnia dopasowania modelu do danych empirycznych opiera się na współczynniku determinacji R^2 postaci:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{4.40}$$

gdzie:
$\overline{y}$ – jest średnią arytmetyczną wyznaczoną dla zmiennej zależnej,
a pozostałe oznaczenia jak poprzednio.

Współczynnik determinacji określa, w jakim stopniu zmienność zmiennej zależnej (objaśnianej) została wyjaśniona za pomocą zmian zmiennych niezależnych w oszacowanej funkcji regresji. Innymi słowy określa on wpływ (na zmienną zależną) tych czynników, które w modelu regresji zostały uwzględnione. Model tym lepiej objaśnia kształtowanie się zmiennej zależnej, im wartość współczynnika determinacji jest bliższa jedności[10].

Inną miarą używaną do opisu poprawności modelu jest tzw. wariancja resztowa modelu, będąca nieobciążonym estymatorem wariancji składnika losowego:

$$S_e^2 = \frac{1}{n-(k+1)} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n-(k+1)} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.41}$$

Miara ta informuje o rozrzucie wartości empirycznych wokół wartości teoretycznych.

Modele regresji znajdują szerokie zastosowanie w wielu dziedzinach i dyscyplinach badawczych, w tym również wykorzystywane są w badaniach procesów i zjawisk społeczno-gospodarczych. Warto dodać, że modele postaci (4.29)–(4.31) noszą w literaturze nazwę modeli regresji wielorakiej, ponieważ zawierają wiele (tj. k) zmiennych niezależnych. Najprostszy model regresji zawiera, tak jak w przykładzie 4.10, tylko jedną zmienną niezależną, czyli jest postaci:

$$y_i = \alpha_0 + \alpha_1 \cdot x_i + \varepsilon_i \tag{4.42}$$

a po oszacowaniu jego parametrów otrzymuje się:

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_i \tag{4.43}$$

10 Najogólniej przyjmuje się, że wartości współczynnika determinacji są dodatnie, co jest prawdą w przypadku modeli regresji zawierających wyraz wolny. Jednakże w przypadku modeli regresji bez wyrazu wolnego może zdarzyć się, że współczynnik determinacji jest ujemny, co oczywiście oznacza, że taki model nie nadaje się do wykorzystania w praktyce.

Ilustrację relacji między wartościami empirycznymi opisanymi modelem (4.42) i teoretycznymi (4.43) przedstawiono na rys. 4.8. W liniowych modelach regresji zawierających jedną zmienną niezależną współczynnik determinacji jest kwadratem współczynnika korelacji liniowej Pearsona (4.5) między zmienną zależną i niezależną, tj. $R^2 = r_{yx}^2$.

Warto zauważyć, że w przypadku funkcji regresji z jedną zmienną niezależną postaci (4.42) istnieje możliwość skonstruowania zależności odwrotnej, czyli:

$$x_i = \beta_0 + \beta_1 \cdot y_i + \varepsilon_i \tag{4.44}$$

co po oszacowaniu modelu daje:

$$\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot y_i \tag{4.45}$$

Wówczas współczynnik determinacji wyznaczony dla modeli (4.42)–(4.43) i (4.44)–(4.45) jest identyczny i wynosi:

$$R^2 = \hat{\alpha}_1 \cdot \hat{\beta}_1 \tag{4.46}$$

gdzie:
$\hat{\alpha}_1$, $\hat{\beta}_1$ – są ocenami estymatorów współczynników kierunkowych funkcji regresji odpowiednio (4.42) i (4.44) zapisanymi w równaniach (4.43) i (4.45).

Oczywiście szacowanie obu funkcji regresji (4.42) i (4.44) jest jedynie możliwością, która nie zawsze ma sens i uzasadnienie merytoryczne, budując bowiem model regresji z góry należy ustalić, która ze zmiennych jest zmienną zależną. Innymi słowy w modelu regresji (w przeciwieństwie do liniowej korelacji) zakłada się, że relacja jest jednokierunkowa. Klasycznym przykładem zależności jednokierunkowej jest występowanie związku między zarobkami (zmienna zależna) i stażem pracy (zmienna niezależna). Można jednak znaleźć przykłady relacji dwukierunkowych, chociażby między zarobkami i wydajnością pracy lub w przypadku zależności notowań dwóch instrumentów finansowych.

Przykład 4.11

Na podstawie danych jednostkowych z przykładu 4.1 należy oszacować funkcję regresji z jedną zmienną niezależną postaci (4.42) i przeprowadzić jej analizę. W tym celu wykorzystać należy funkcję REGLINP programu Excel.

Rozwiązanie

W przykładzie 4.2 obliczony został współczynnik korelacji liniowej Pearsona między powierzchnią (X) wynajmowanych biur, a okresem (Y) działalności firmy. Oszacowanie modelu regresji (4.42) dla tak zdefiniowanych zmiennych jest oczywiście możliwe – otrzymuje się wtedy równanie $\hat{y}_i = 1{,}0834 + 0{,}0360 \cdot x_i$; $R^2 = 0{,}3339$.

Jednakże próba oceny merytorycznej tego modelu pokazuje błąd kierunku postawionej zależności. Tak skonstruowany model opisuje bowiem sytuację, w której na okres działalności firmy (w latach) oddziałuje wielkość wynajmowanej powierzchni biurowej, a to jest rozumowaniem niepoprawnym. W związku z tym za zmienną zależną należy uznać zmienną Y opisującą powierzchnię, a okres funkcjonowania przedsiębiorstwa – za zmienną niezależną X. Otrzymuje się wtedy oszacowanie $\hat{y}_i = 30{,}3370 + 9{,}2744 \cdot x_i$; $R^2 = 0{,}3339$, które odczytać można z tab. 4.18, będącej wynikiem obliczeń funkcji REGLINP.

Tabela 4.18. Tabela wynikowa funkcji REGLINP

Opis zawartości niezacienionych elementów tablicy REGLINP	$\hat{\alpha}_1$	$\hat{\alpha}_0$
Oceny estymatorów parametrów	9,2743	30,3370
Standardowe błędy szacunku	2,4753	8,6693
Współczynnik determinacji	0,3339	17,1254
Liczba stopni swobody	14,0386	28
Suma kwadratów reszt empirycznych	4117,2034	8211,7632

Uwaga: w tabeli przytoczono całą tablicę wynikową funkcji REGLINP, przy czym kolorem szarym oznaczono wartości nieistotne w omawianym przykładzie.

Źródło: opracowanie własne.

Jak łatwo można zauważyć oba oszacowane modele, różnią się ocenami estymatorów parametrów funkcji regresji, ale stopień dopasowania tych modeli do danych empirycznych pozostał na tym samym poziomie – w obu przypadkach współczynnik determinacji wynosi tyle samo. Co więcej, jego wartość jest równa iloczynowi ocen estymatorów parametrów stojących przy zmiennej niezależnej obu modeli, czyli: $0{,}0360 \cdot 9{,}2744 = 0{,}3339$. Jednocześnie R^2 jest równy kwadratowi współczynnika Pearsona, obliczonemu dla danych niepogrupowanych w przykładzie 4.2. Wystarczy przypomnieć, że obliczony współczynnik korelacji liniowej wyniósł $r_{xy} = 0{,}5779$ tj. tyle, ile wyniesie pierwiastek ze współczynnika determinacji (z dokładnością do błędów zaokrągleń)[11].

Z oszacowanej funkcji regresji wynika, że wzrost okresu działalności firmy deweloperskiej o jeden rok powoduje wzrost wynajmowanej powierzchni biurowej o 9,27 m^2 (wartość oceny estymatora parametru stojącego przy zmiennej niezależnej). Jednocześnie można przyjąć, że minimalna powierzchnia wynajmowanego

11 *De facto* pierwiastek ze współczynnika determinacji jest w tym przypadku równy modułowi (wartości bezwzględnej) współczynnika korelacji liniowej Pearsona, którego znak ustala się na podstawie wartości współczynnika kierunkowego oszacowanego równania regresji.

przez firmę deweloperską biura to 30,33 m^2 (wartość wyrazu wolnego). Zmiany wielkości powierzchni biurowej wyjaśniane są zmianami okresu funkcjonowania firmy na rynku w 33% (wartość współczynnika determinacji), co oznacza, że na zmienną zależną wpływają inne (tj. nieuwzględnione w modelu) czynniki.

Oceniając poprawność modelu należy oszacować przedziały ufności dla parametrów (4.34) oraz zweryfikować hipotezy dotyczące istotności oszacowanych parametrów (4.35)–(4.38) za pomocą sprawdzianu testu (4.39). W pierwszym przypadku przyjmuje się współczynnik ufności równy 0,9, a w drugim – poziom istotności 0,05. W tym celu z tab. 4.18 odczytać należy wartości standardowych błędów szacunku, które wynoszą $S(\alpha_0) = 8{,}6693$ i $S(\alpha_1) = 2{,}4753$, a z tablic rozkładu t-Studenta – wartość krytyczną dla obustronnego obszaru odrzucenia, $\alpha = 0{,}1$ i 28 stopni swobody[12], która wynosi $t_\alpha = 1{,}7011$. A zatem przedział ufności (4.34) dla wyrazu wolnego:

$$\begin{aligned}&P(\hat{\alpha}_0 - t_\alpha \cdot S(\alpha_0) < \alpha_0 < \hat{\alpha}_0 + t_\alpha \cdot S(\alpha_0)) = \\ &= P(30{,}3370 - 1{,}7011 \cdot 8{,}6693 < \alpha_0 < 30{,}3370 + 1{,}7011 \cdot 8{,}6693) = \\ &= P(15{,}5894 < \alpha_0 < 45{,}0847) = 0{,}9\end{aligned}$$

i współczynnika kierunkowego (tj. parametru stojącego przy zmiennej x):

$$\begin{aligned}&P(\hat{\alpha}_1 - t_\alpha \cdot S(\alpha_1) < \alpha_1 < \hat{\alpha}_1 + t_\alpha \cdot S(\alpha_1)) = \\ &= P(9{,}2744 - 1{,}7011 \cdot 2{,}4753 < \alpha_1 < 9{,}2744 + 1{,}7011 \cdot 2{,}4753) = \\ &= P(5{,}0636 < \alpha_1 < 13{,}4851) = 0{,}9\end{aligned}$$

Jak zatem widać, rozpiętość obu przedziałów ufności jest duża, mimo że poziom ufności wynosi 0,9 (co oznacza, że w każdym z ogonów rozkładów pole pod funkcją gęstości wyniesie $\frac{\alpha}{2} = 0{,}05$). Dalej zweryfikowane zostaną hipotezy (3.35) dla wyrazu wolnego:

$$H_0: \alpha_0 = 0 \text{ wobec hipotezy } H_1: \alpha_0 > 0$$

i parametru stojącego przy zmiennej x:

$$H_0: \alpha_1 = 0 \text{ wobec hipotezy } H_1: \alpha_1 > 0$$

W obu przypadkach występuje prawostronny obszar odrzucenia, bo oceny estymatorów obu parametrów są dodatnie. Zatem odczytana z tablic rozkładu t-Studenta wartość krytyczna dla $\alpha = 0{,}05$ i 28 stopni swobody wynosi $t_\alpha = 1{,}7011$. Obliczone wartości sprawdzianu (4.39) to:

12 Było 30 obserwacji i 2 szacowane parametry funkcji regresji, czyli liczba stopni swobody wynosi 30 – 2 = 28. Wprawdzie obszar odrzucenia jest prawostronny, a poziom istotności wynosi α = 0,05, ale w Excelu funkcja ROZKŁAD PRAWDOPODOBIEŃSTWA dla obszaru dwustronnego wymaga podwojenia α.

$$t = \frac{\hat{\alpha}_0}{S(\alpha_0)} = \frac{30,3370}{8,6693} = 3,4994$$

$$t = \frac{\hat{\alpha}_1}{S(\alpha_1)} = \frac{9,2744}{2,4753} = 3,7468$$

Z obliczeń wynika, że oba parametry są istotnie większe od 0, ponieważ na poziomie istotności 0,05 odrzucić należy hipotezę zerową na rzecz hipotezy alternatywnej.

Warto zauważyć, że w przypadku weryfikacji statystycznej istotności wpływu kilku zmiennych niezależnych (na zmienną zależną w modelu regresji wielorakiej) za pomocą testu (4.36)–(4.39) podejmuje się decyzję o pozostawieniu (jeśli parametr przy niej stojący jest istotnie różny od 0) lub usunięciu (kiedy nie ma podstaw do odrzucenia hipotezy zerowej) testowanej zmiennej z modelu. Nie zawsze jednak podjęcie decyzji o usunięciu konkretnej zmiennej jest merytorycznie uzasadnione i w takim przypadku pozostawia się nieistotną zmienną w modelu, zwłaszcza kiedy jej usunięcie powoduje drastyczny spadek dopasowania modelu do danych empirycznych lub istnieją inne przesłanki uzasadniające takie postępowanie.

Inną istotną kwestią jest posługiwanie się współczynnikiem determinacji (4.40) w modelach regresji wielorakiej[13]. Wiadomo bowiem, że czym więcej zmiennych niezależnych (nawet jeśli ich wpływ na zmienną zależną jest nieistotny), tym większa wartość współczynnika R^2. W związku z tym w przypadku, gdy ocenie podlegają modele regresji wielorakiej zamiast R^2 ze wzoru (4.40), wyznacza się tzw. skorygowany współczynnik determinacji według wzoru:

$$\overline{R}^2 = 1-(1-R^2)\cdot\frac{n-1}{n-(k+1)} \tag{4.47}$$

gdzie oznaczenia jak poprzednio.

Skorygowany współczynnik determinacji $\overline{R}^2$ (4.47) ma tę istotną własność, że jest niezależny od liczby zmiennych w modelu oraz liczby obserwacji. Wykorzystuje się go do sprawdzenia, czy należy dodawać (usuwać) zmienne z modelu. Jeżeli po dodaniu (usunięciu) zmiennej wartość skorygowanego współczynnika $\overline{R}^2$ rośnie (maleje), to oznacza, że należy dodać (usunąć) analizowaną zmienną.

13 Przykłady modeli regresji wielorakiej przedstawione będą w rozdziale 6, dotyczącym modeli ekonometrycznych.

Rozdział 5

Analiza dynamiki zjawisk

Jak wcześniej wspomniano, zjawiska społeczno-ekonomiczne charakteryzuje zmienność w czasie, dlatego istnieje potrzeba badania ich dynamiki. Obserwacje pewnego zjawiska, procesu lub zmiennej dokonywane w różnym czasie tworzą tzw. szeregi czasowe. Szeregiem czasowym (chronologicznym, dynamicznym) nazywa się uporządkowany (według czasu) zbiór obserwacji statystycznych charakteryzujących zmiany zjawiska w czasie. Zatem w szeregach chronologicznych istotna jest kolejność jego poszczególnych elementów, niesie ona bowiem informację o zmianach, jakie zaszły w analizowanym zjawisku w kolejnych punktach pomiarowych. Przykład szeregu czasowego przedstawiono w postaci wykresu liniowego na rys. 1.3.

Wśród szeregów czasowych wyróżnia się dwa podstawowe ich rodzaje, które wynikają z istoty badanego zjawiska. Są to szeregi czasowe momentów i okresów. Przykładem pierwszych są notowania cen akcji na giełdzie, a drugich wolumen obrotów. Szeregi czasowe momentów zawierają informacje o poziomie zjawiska w określonych momentach pomiarowych. Nie wiadomo zatem, jak się ono kształtowało między kolejnymi momentami obserwacji. W związku z tym danych prezentowanych w szeregach tego typu nie można agregować, jeśli np. trzeba wydłużyć okres obserwacji. Istnieje jednak możliwość ominięcia tego problemu poprzez wyznaczenie przeciętnego poziomu zjawiska na podstawie pomiarów przeprowadzonych w kilku momentach należących do okresu przyjętego za nową (zagregowaną) jednostkę czasu. Przykładowo kursy akcji na giełdzie podawane są w różnych momentach, ale szczególne znaczenie mają ceny otwarcia i zamknięcia, a także średni kurs dnia. Można także wydłużyć okresy obserwacji, pomijając część szczegółowych danych, np. można jedynie dokonywać pomiarów cen zamknięcia w kolejne piątki i w ten sposób z danych dziennych uzyska się dane tygodniowe. Szeregi czasowe okresów podają rozmiary danego zjawiska w całym okresie przyjętym za jednostkę czasu. Istnieje zatem możliwość wydłużania okresów, w jakich dokonuje się pomiaru, np. poprzez przechodzenie z danych miesięcznych na kwartalne i roczne, za pomocą agregacji (sumowania) danych.

W celu wyznaczenia przeciętnego poziomu zjawiska dla szeregów czasowych momentów wykorzystuje się średnią chronologiczną:

$$\overline{y}_{ch} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \ldots + \frac{y_{T-1} + y_T}{2}}{T-1} = \frac{1}{T-1}\left(\frac{y_1}{2} + \sum_{t=2}^{T-1} y_t + \frac{y_T}{2} \right) \tag{5.1}$$

gdzie:
y_t – obserwacje szeregu czasowego w kolejnych punktach pomiarowych $t = 1, 2, \ldots, T$.

Średnia chronologiczna daje jedynie ogólną orientację o przeciętnym poziomie badanego zjawiska, bowiem nie wiadomo, jak kształtowało się zjawisko między momentami pomiarów. Wartość średniej chronologicznej (5.1) dla dużych T jest zbieżna do wartości średniej arytmetycznej, postaci:

$$\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{5.2}$$

która jest adekwatną miarą średnią dla szeregów okresów.

5.1. Mierniki dynamiki

W analizie szeregów czasowych interesujące są zmiany, jakim podlega badane zjawisko w kolejnych momentach lub okresach czasu. Najprostszymi miarami są przyrosty absolutne i stosunkowe. Przyrost absolutny (dy_t) to różnica w poziomie zjawiska mierzonego (obserwowanego) w dwóch różnych momentach lub okresach czasu, czyli:

$$dy_{t,\,t-\tau} = y_t - y_{t-\tau} \tag{5.3a}$$

gdzie:
y_t – poziom zjawiska y w okresie lub momencie t,
$y_{t-\tau}$ – poziom zjawiska y w okresie lub momencie poprzedzającym okres t o τ okresów.

Zapis (5.3a) opisuje tzw. przyrosty absolutne łańcuchowe, które mierzą zmiany, jakie zachodzą w bieżącym okresie w stosunku do tzw. okresu bazowego (podstawowego), który we wzorze (5.3a) zmienia się w kolejnych punktach pomiarowych. W analizach dynamiki okres pomiarowy jest określany przez prowadzącego badanie, np. godzina, dzień, tydzień, miesiąc, rok. Dla $\tau = 1$ przyrosty

łańcuchowe mierzą zmiany, jakie zachodzą z okresu na okres. Przyrosty absolutne mogą być również wyznaczane w stosunku do pewnego stałego poziomu, jeżeli w relacji (5.3a) odjemna $y_{t-\tau}$ zostanie zastąpiona przez stałą y_0, oznaczającą poziom zjawiska w okresie bazowym (początkowym), którym może być dowolnie wybrany okres lub moment. Wyznaczone w taki sposób przyrosty nazywane są jednopodstawowymi, ponieważ wyznaczane są w stosunku do stałej wartości zjawiska y_0, czyli:

$$dy_{t,0} = y_t - y_0 \tag{5.3b}$$

Przyrost względny, zwany również stosunkowym (δy_t), definiuje się jako stosunek przyrostu absolutnego do poziomu zjawiska w okresie bazowym, czyli:

$$\delta y_{t,t-\tau} = \frac{y_t - y_{t-\tau}}{y_{t-\tau}} \tag{5.4a}$$

lub

$$\delta y_{t,0} = \frac{y_t - y_0}{y_0} \tag{5.4b}$$

gdzie oznaczenia jak poprzednio.

Przyrosty absolutne są wielkościami mianowanymi, wyrażonymi w tych samych jednostkach co analizowane zjawisko lub cecha. Natomiast przyrosty względne nie są mianowane i są zwykle wyrażane w procentach (po przemnożeniu przyrostów (5.4) przez 100). We wszystkich przypadkach określają one, o ile wzrósł lub zmniejszył się poziom zjawiska w okresie badanym w stosunku do poziomu tegoż zjawiska w okresie bazowym, tj. przyjętym za podstawę porównań, z tą różnicą, że przyrosty (5.3) są wartościami mianowanymi, a (5.4) – stosunkowymi. Warto zauważyć, że przyrosty względne (5.4) są jednym z podstawowych parametrów wykorzystywanych w finansach, określają one bowiem stopę zwrotu z inwestycji, której zmiany wartości w czasie przedstawia szereg czasowy. Jest to tzw. prosta stopa zwrotu. Natomiast w ekonometrii przyrosty względne określane są mianem tempa wzrostu.

Porównania poziomów zjawiska w dwóch momentach lub okresach czasu dokonuje się również za pomocą wskaźników dynamiki zwanych indeksami. Indeks jest stosunkiem wielkości zjawiska w okresie badanym do wielkości tego zjawiska w okresie przyjętym za podstawę. W zależności od stopnia złożoności badanego zjawiska wyróżnia się dwa rodzaje indeksów, tj. indeksy indywidualne, zwane prostymi, i indeksy zespołowe, zwane agregatowymi. Przykładami powszechnie stosowanych w praktyce indeksów indywidualnych są stopa procentowa i dyskontowa, za pomocą których przeszacowuje się wartości w czasie. Przykładem

indeksów agregatowych mogą być stopa inflacji oraz indeksy giełdowe. Indeksy indywidualne stosuje się w przypadku zjawisk jednorodnych i bezpośrednio sumowalnych. Ze względu na przyjętą podstawę porównań dzieli się je, podobnie jak przyrosty, na:

- indeksy jednopodstawowe, czyli wyznaczane w stosunku do stałej podstawy:

$$i_{t,0} = \frac{y_t}{y_0} \tag{5.5}$$

- indeksy łańcuchowe, kiedy ich podstawa się zmienia:

$$i_{t,t-\tau} = \frac{y_t}{y_{t-\tau}} \tag{5.6}$$

wśród których najczęściej porównania prowadzone są w odniesieniu do okresu poprzedniego, czyli:

$$i_{t,t-1} = \frac{y_t}{y_{t-1}} \tag{5.7}$$

Ciąg indeksów jednopodstawowych (5.5) pokazuje, jak zmienia się poziom zjawiska w stosunku do jednego stałego okresu, przyjętego za bazowy. Natomiast indeksy łańcuchowe (5.7) wskazują na zmiany z okresu na okres. Indeksy są miarami niemianowanymi[1], a ich interpretacja polega na analizie ich wartości w stosunku do jedności, gdyż wartość indeksu równa 1 oznacza, że poziom zjawiska w badanym okresie nie uległ zmianie. Wartość indeksu większa od 1 wskazuje na wzrost poziomu zjawiska w porównaniu z okresem przyjętym za podstawę, natomiast mniejsza od 1 na spadek poziomu zjawiska w porównaniu z okresem przyjętym za podstawę. Przykładowo wartość indeksu $i_{t,\,t-1} = 1{,}13$ oznacza, że nastąpił wzrost poziomu zjawiska o 13% w stosunku do okresu poprzedniego, a wartość indeksu $i_{t,0} = 0{,}87$ – spadek o 13% w stosunku do okresu początkowego. Warto odnotować, że interpretacja indeksów statystycznych (5.5)–(5.7) jest *de facto* informacją o przyrostach względnych (5.4), co wynika z prostego przekształcenia

$$i_{t,0} - 1 = \frac{y_t}{y_0} - 1 = \frac{y_t - y_0}{y_0} = \delta y_{t,0}\,.$$

Przeciętne tempo zmian w całym przedziale czasowym wyznacza się za pomocą średniej geometrycznej obliczanej z indeksów łańcuchowych:

1 Czasami wartości indeksów podaje się po ich przemnożeniu przez 100, np. indeksy publikowane przez GUS.

$$\overline{y}_g = \sqrt[T]{i_{1,0} \cdot i_{2,1} \cdot \ldots \cdot i_{T-1,T-2} \cdot i_{T,T-1}} = \sqrt[T]{\frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdot \ldots \cdot \frac{y_{T-1}}{y_{T-2}} \cdot \frac{y_T}{y_{T-1}}} = \sqrt[T]{\frac{y_T}{y_0}} = \sqrt[T]{i_{T,0}} \quad (5.8)$$

lub jako pierwiastek T-tego stopnia z indeksu jednopodstawowego[2]. Interpretacja średniej geometrycznej prowadzona jest (podobnie jak w przypadku indeksów) w odniesieniu do jedności. Zatem interpretuje się wartość:

$$G = \overline{y}_g - 1 \quad (5.9)$$

która informuje o przeciętnym wzroście (jeśli $G > 0$) lub spadku (jeśli $G < 0$) zjawiska z okresu na okres. Jest ona zwykle wyrażana w procentach.

Jak widać w zapisie wzoru (5.8), indeks jednopodstawowy $i_{T,0}$, opisujący zmianę poziomu zjawiska w okresie lub momencie T w stosunku do okresu początkowego ($t = 0$), wyznaczony został jako iloczyn indeksów łańcuchowych $i_{1,0}$, $i_{2,1}$, …, $i_{T,T-1}$ wyrażających zmiany obserwowane z okresu na okres. Zatem wyznaczenie indeksu jednoodstawowego z ciągu indeksów łańcuchowych polega na ich przemnożeniu, a wartości indeksów łańcuchowych wyznacza się jako ilorazy odpowiednich indeksów jednopodstawowych.

Indeksy agregatowe (zespołowe) są wskaźnikami zmian dla wielkości zagregowanych, czyli zawierających kilka (przynajmniej 2) składników. Szczególnie uważnie należy wyznaczać wartości indeksów dla zjawisk niejednorodnych, w których tworzące je elementy mogą nie być bezpośrednio sumowane, np. indeks cen dla grupy towarów o różnym charakterze. W przypadku takich zjawisk, w celu doprowadzenia ich do sumowalności, wprowadza się określone współczynniki przeliczeniowe, spełniające role wag, takie jak np.: ilość, cena, liczba zatrudnionych, czas pracy. Najbardziej znane w zastosowaniach finansowych indeksy to: indeks wartości, cen i indeksy giełdowe. Ogólnie indeks agregatowy wyraża się wzorem:

$$I_t = \frac{\sum_{i=1}^{k} w_{it} \cdot x_{it}}{\sum_{i=1}^{k} w_{i0} \cdot x_{i0}} \quad (5.10)$$

gdzie dla okresu badanego (t) oraz podstawowego (0):
$w_{it} \cdot x_{it}$, $w_{i0} \cdot x_{i0}$ to wartości i-tej składowej, a k – liczba wyróżnionych składowych.

Obliczony indeks informuje, o ile zmienił się poziom zjawiska w okresie badanym w porównaniu z okresem podstawowym. Przykładowo w indeksie wartości obliczonym według (5.10) przyjmuje się następujące oznaczenia: w_{it}, w_{i0} – wagi,

2 Należy zauważyć, że stopień pierwiastka jest taki sam jak liczba indeksów łańcuchowych $i_{t,t-1}$, do wyznaczenia których potrzebnych jest $T + 1$ obserwacji ($t = 0, 1, 2, \ldots, T$).

np. ceny *i*-tej składowej, x_{it}, x_{i0} – ilości *i*-tej składowej. W literaturze przedmiotu wyróżnia się różne formuły wyznaczania indeksów w celu uwzględnienia, który z wyróżnionych składowych (np. cena, waga czy ilość) ma wpływ i z jaką siłą oddziałuje na badane zjawisko. Przykładem są indeksy cen wyznaczane według formuły Laspeyresa (ceny ustalone na poziomie okresu podstawowego) lub Paaschego (ceny ustalone na poziomie okresu badanego).

Indeksy giełdowe obliczane są dla każdej sesji giełdowej i zależą od zmiany kursów papierów wartościowych. Dzięki nim można określić, czy w danym okresie występuje tendencja spadkowa lub zwyżkowa kursów akcji. Szybki wzrost wartości indeksu oznacza, że na giełdzie jest hossa, czyli kursy większości papierów wartościowych idą w górę. Spadek wartości wskaźnika jest charakterystyczny dla bessy, czyli sytuacji odwrotnej.

Większość indeksów giełdowych to agregatowe indeksy wartości (czasem w pewien sposób zmodyfikowane), reprezentujące zmianę wartości rynkowej pewnego koszyka akcji giełdowych w porównaniu do pewnej wartości bazowej, którą zazwyczaj jest wartość portfela indeksu z pierwszego notowania indeksu na danej giełdzie. Indeksy giełdowe pełnią następujące funkcje (por. [Jajuga 2002, s. 96]):

- są wskaźnikami koniunktury na danej giełdzie,
- ułatwiają monitorowanie zmian wartości kapitału na rynku,
- są odbiciem tendencji charakteryzujących daną gospodarkę,
- wykorzystywane są do prognozowania trendów,
- indeks giełdy może być traktowany również jako instrument finansowy, co ma znaczenie przy opcjach i instrumentach pochodnych.

Dobry indeks giełdowy powinien spełniać następujące warunki:

- wskazywać, jakie zaszły zmiany w cenach akcji (papierów wartościowych) na giełdzie w danym dniu w porównaniu do pewnego okresu podstawowego, tzn. rosnąć, gdy rosną ceny większości akcji, a spadać, gdy spadają ceny większości akcji,
- opierać się na stosunkowo dużej liczbie akcji, reprezentujących przeważającą część rynku akcji, którymi obraca się na danej giełdzie, lub sektora, charakteryzowanego przez dany indeks (np. dla indeksów branżowych),
- nie zależeć od samych wartości akcji, a jedynie od zmian ich wartości,
- uwzględniać udziały akcji danej spółki w zbiorze wszystkich akcji, którymi się obraca na danej giełdzie.

Pierwszy indeks giełdowy *Dow Jones Industrial Average* został opracowany i obliczony przez Charlesa Dowa w 1884 roku dla 11 wybranych spółek notowanych na giełdzie w Nowym Jorku. Szacuje się, że obecnie obliczanych jest na świecie ponad 150 tys. indeksów giełdowych, a same indeksy na rynku kapitałowym spełniają coraz więcej funkcji.

Indeks giełdowy może uwzględniać wszystkie lub tylko część akcji notowanych na giełdzie. Wraz ze wzrostem liczby notowanych spółek zazwyczaj zwiększa się liczba notowanych indeksów, które opisują wybrane segmenty rynku, np. spółki

o różnej wielkości lub należące do różnych branż, ewentualnie spółki charakteryzujące się pewnymi cechami, np. mające siedzibę w danym kraju, przestrzegające zasad ładu korporacyjnego (*corporate governance*) lub wypłacające dywidendę akcjonariuszom. Przykładem indeksów notowanych na Giełdzie Papierów Wartościowych, dedykowanych spółkom różnej wielkości, są indeksy WIG20, mWIG40 i sWIG80, a uwzględniających branże, w jakich działają spółki, są indeksy: WIG_{banki}, $WIG_{budownictwo}$, WIG_{chemia}, $WIG_{informatyka}$, $WIG_{spożywczy}$ itp. Z kolei indeks WIG_{ESG} zawiera w swym portfelu spółki uznawane za odpowiedzialne społecznie, a indeks WIG_{div} spółki, które w okresie ostatnich 5 lat obrotowych regularnie dokonywały wypłaty dywidendy swoim akcjonariuszom.

5.2. Dekompozycja szeregu czasowego

W szeregach czasowych wyróżnia się zazwyczaj dwie składowe: systematyczną, będącą efektem oddziaływań stałego zestawu czynników, oraz przypadkową (nieregularną). Składowa systematyczna może występować w postaci:

- tendencji rozwojowej (trendu), odzwierciedlającej długookresową skłonność do jednokierunkowych zmian (wzrostu lub spadku) wartości analizowanej zmiennej,
- stałego (przeciętnego) poziomu analizowanej zmiennej, co czasem (zwłaszcza przez inwestorów giełdowych) nazywane jest trendem horyzontalnym,
- składowej okresowej (periodycznej), która może wystąpić w postaci wahań cyklicznych lub sezonowych.

Wahania cykliczne to długookresowe, rytmiczne wahania wartości zmiennej wokół stałego (przeciętnego) poziomu lub wokół trendu badanej zmiennej. Wahania sezonowe są wahaniami, które powtarzają się w ciągu jednego roku. Odzwierciedlają one wpływ zachowań ludzkich wynikających z pogody i kalendarza (pór dnia i roku, świąt itp.) na kształtowanie się zjawisk gospodarczych. Najczęściej wyróżnia się wahania kwartalne, miesięczne, tygodniowe. W celu identyfikacji poszczególnych składowych szeregu czasowego dla konkretnej zmiennej często wykorzystuje się wykresy szeregów czasowych.

Każda z wymienionych składowych szeregu czasowego wywiera specyficzny wpływ na kształtowanie się analizowanej zmiennej. Jednakże nie są one (tj. każda z osobna) bezpośrednio obserwowalne, dlatego ich pomiar następuje przez dekompozycję szeregu czasowego, co w przypadku występowania tendencji rozwojowej można zapisać jako:

$$y_t = f(t) + g_l(t) + \varepsilon_t \tag{5.11}$$

lub

$$y_t = f(t) \cdot g_l(t) \cdot 10^{\varepsilon_t} \tag{5.12}$$

gdzie:
y_t – poziom badanego zjawiska zaobserwowany w okresie lub momencie t,
$f(t)$ – funkcja trendu,
$g_l(t)$ – funkcje aproksymujące wahania okresowe,
ε_t – składnik losowy (resztowy) modelu odzwierciedlający wahania przypadkowe.

Modele (5.11) i (5.12) nazywa się modelami wahań w czasie odpowiednio typu addytywnego i multiplikatywnego. Modele addytywne (5.11) stosuje się tylko wtedy, kiedy funkcja trendu jest funkcją liniową lub da się do takiej sprowadzić. W pozostałych przypadkach stosuje się model multiplikatywny (5.12) lub modele mieszane. W modelu addytywnym zakłada się, że obserwowane wartości zmiennej czasowej są sumą (wszystkich lub niektórych) składowych szeregu czasowego. Jednocześnie zakłada się, że nie występują interakcje pomiędzy poszczególnymi składowymi szeregu, tzn. składowe są niezależne. W modelu multiplikatywnym przyjmuje się, że obserwowane wartości zmiennej czasowej stanowią iloczyn składowych szeregu czasowego. Model multiplikatywny jest często używanym modelem dekompozycji szeregów czasowych. W modelu tym tylko jedna ze składowych – na ogół trend lub stały (średni) poziom czasowej zmiennej – jest wyrażana w jednostkach analizowanej zmiennej. Pozostałe składowe szeregu są w procesie dekompozycji wyrażane jako względne odchylenia od trendu bądź od stałego (przeciętnego) poziomu zmiennej.

Przedstawione modele (5.11)–(5.12) dekompozycji szeregu czasowego uwzględniają występowanie trendu oraz wahań okresowych i przypadkowych. Można dokonać dalszej dekompozycji modelu, w wyniku czego zostaną wyodrębnione wahania cykliczne i sezonowe, np. dla modelu addytywnego:

$$g_l(t) = g^1_{l1}(t) + g^2_{l2}(t) \tag{5.13}$$

gdzie funkcje: $g^1_{l1}(t)$, $g^2_{l2}(t)$ opisywać będą składowe cykliczne i sezonowe dla wyróżnionych podokresów odpowiednio $l_1 = 1, 2, \ldots, L_1$ oraz $l_2 = 1, 2, \ldots, L_2$.

Warto zauważyć, że w szeregu czasowym mogą jednocześnie występować wahania o różnej częstotliwości, wykazując istnienie cyklu miesięcznego, tygodniowego i dobowego. W innych przypadkach może okazać się, że w szeregu czasowym nie występuje trend, a wahania oscylują wokół pewnego stałego (średniego) poziomu zmiennej, wtedy modele wahań w czasie są postaci:

$$y_t = \text{const} + g_l(t) + \varepsilon_t \tag{5.14}$$

lub

$$y_t = \text{const} \cdot g_l(t) \cdot 10^{\varepsilon_t} \tag{5.15}$$

gdzie: const oznacza stały (średni) poziom zjawiska.

Do opisu wielu szeregów czasowych wystarczające mogą okazać się modele zawierające tylko niektóre składowe szeregu czasowego, np. stały (średni) poziom zmiennej i wahania przypadkowe, albo tendencję rozwojową i wahania przypadkowe lub trend, wahania sezonowe i wahania przypadkowe. Wyodrębnianie poszczególnych składowych szeregu czasowego przeprowadza się krok po kroku, stosując odpowiednie metody. Zanim jednak zastosuje się adekwatne metody, należy przeprowadzić wstępną analizę szeregu czasowego, nazywaną jego wstępnym wyrównywaniem (*pre-adjustment*), które polega na:

1) określeniu, czy związek między poszczególnymi składowymi szeregu ma charakter addytywny (5.11) czy multiplikatywny (5.12),
2) identyfikacji obserwacji nietypowych (odstających) w szeregu,
3) uzupełnieniu brakujących danych szeregu czasowego, np. dniami roboczymi, w których nie odbywały się notowania giełdowe (np. dni świąteczne).

Celem pierwszego działania jest sprawdzenie, czy nie ma potrzeby zastosowania wstępnej transformacji logarytmicznej szeregu obserwacji. Obserwacje nietypowe, które mogą być wynikiem zarówno błędu pomiaru, jak i pojawieniem się sporadycznych zaburzeń, powodują zniekształcenia w prowadzonej analizie i mogą prowadzić do niepoprawnych wniosków. W praktyce stosuje się odpowiednie metody służące wykrywaniu obserwacji odstających[3] oraz eliminowaniu ich wpływu.

Dekompozycję szeregu czasowego zwykle rozpoczyna się od wyodrębnienia tendencji rozwojowej za pomocą metody:

- analitycznej, polegającej na oszacowaniu funkcji trendu lub
- mechanicznej, opartej na średnich.

Modele trendu zwane również modelami tendencji rozwojowej zawierają tylko jedną zmienną objaśniającą, którą jest czas t (zmienna czasowa), czyli:

$$y_t = f(t) + \varepsilon_t \tag{5.16}$$

gdzie:
f – symbol dowolnej funkcji,
t – zmienna czasowa przyjmująca najczęściej wartości $t = 1, 2, \ldots, T$,
y_t – zmienna objaśniana, opisywana przez model,
ε_t – składnik losowy.

W zależności od postaci analitycznej funkcji f wyróżnia się różne rodzaje trendu. Do najczęściej wykorzystywanych należą funkcja liniowa, wykładnicza,

3 Najprostszą metodą jest sprawdzenie czy istnieją dane, które nie mieszczą się w przedziale trzysigmowym, o czym była mowa w podrozdziale 1.5.

potęgowa i paraboliczna (wielomian stopnia drugiego)[4]. Trend liniowy można zapisać jako:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \tag{5.17}$$

gdzie:
β_0, β_1 – nieznane parametry funkcji trendu, przy czym β_0 oznacza wyraz wolny, a β_1 – stały przyrost wartości zmiennej objaśnianej z okresu na okres.

Funkcję (5.17) można, podobnie jak funkcję regresji (4.30) lub (4.42), oszacować za pomocą metody najmniejszych kwadratów. Wartości teoretyczne zmiennej objaśnianej wyznacza się z relacji:

$$\hat{y}_t = b_0 + b_1 t \tag{5.18}$$

gdzie:
$\hat{y}$ – wartości teoretyczne wyznaczone na podstawie oszacowanej funkcji trendu,
b_0, b_1 – oceny MNK estymatorów parametrów trendu liniowego.

W analizach szeregów czasowych można rozpatrywać stały poziom zjawiska wyznaczony jako średnia arytmetyczna (5.2) lub chronologiczna (5.1) liczone na podstawie wszystkich elementów szeregu albo jako średnie ruchome[5] *k*-okresowe. Najłatwiej wyznacza się tzw. prostą średnią ruchomą, która liczona dla nieparzystej liczby obserwacji k ($k < T$) obserwacji jest postaci:

$$\overline{y}_{t-\frac{k-1}{2}} = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i \quad \text{dla } t \geq k \tag{5.19}$$

gdzie:
$\overline{y}_\tau$ – wartości średnich ruchomych obliczone dla kolejnych podokresów,
k – krok uśredniania.

Przedstawiony sposób wyznaczania średniej ruchomej (kroczącej) oznacza, że jej wartość przypisuje się środkowej obserwacji szeregu, na podstawie którego została ona wyznaczona. Należy zauważyć, że korzystanie ze średnich kroczących pozwala wyznaczyć $T - (k - 1)$ średnich w przypadku nieparzystego kroku uśredniania oraz $T - k$ średnich dla parzystego k.

W sytuacji kiedy średnie ruchome mają być wykorzystane do eliminacji wahań okresowych, muszą obejmować ściśle określoną liczbę jednostek czasu, odpowiadającą zaobserwowanemu cyklowi wahań. Na przykład pojawienie się wahań przy danych kwartalnych oznacza potrzebę zastosowania 4-okresowej średniej

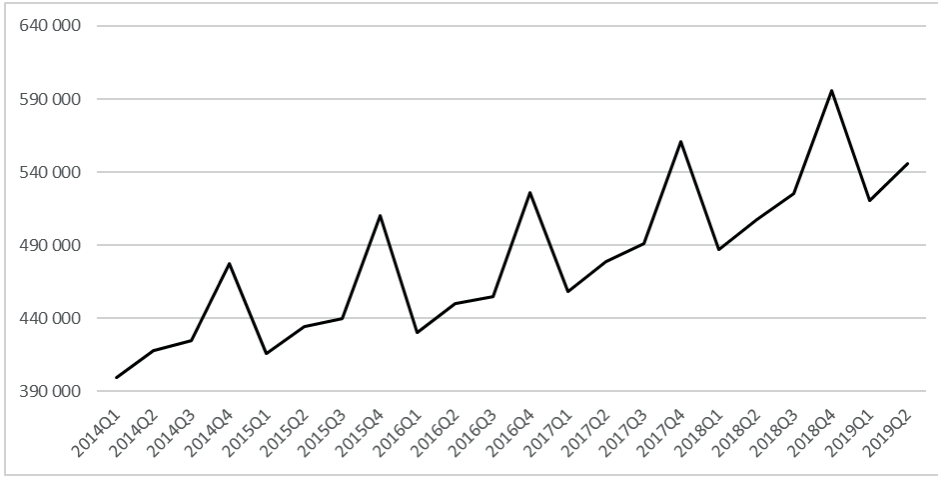
4 Szczegółowo zagadnienia te omówione są m.in. w pracy [Cieślak 1997, s. 77–87].
5 Omawiane będą szczegółowo w dalszej części niniejszego rozdziału.

ruchomej, a wahań dla danych miesięcznych – średniej 12-okresowej. Pojawia się wtedy problem liczenia średnich ruchomych z parzystej liczby okresów, są to tzw. średnie scentrowane, które wyznacza się zgodnie z relacją:

$$\overline{y}_{t-\frac{k}{2}} = \frac{1}{k}\left(\frac{1}{2}\cdot y_{t-k} + \sum_{i=t-k+1}^{t-1} y_i + \frac{1}{2} y_t\right) \text{ dla } t \geq k+1 \tag{5.20}$$

Należy pamiętać, że w przypadku scentrowanych średnich ruchomych, wyznaczanych z k danych, również traci się tyle samo obserwacji ($k/2$) na początku i końcu szeregu.

Wiele zjawisk charakteryzuje się występowaniem wahań okresowych, których szczególnym przypadkiem są wahania sezonowe. Przykładowym szeregiem tego typu jest szereg obserwacji produktu krajowego brutto w ujęciu kwartalnym (rys. 5.1).



Rysunek 5.1. Produkt krajowy brutto (ceny bieżące) w [mln PLN]

Źródło: GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019)

Wyodrębnianie wahań sezonowych przeprowadza się za pomocą tzw. wskaźnika sezonowości, który można wyznaczyć analitycznie w postaci:

$$S_l = \frac{\sum_{i=1}^{l} y_i}{\sum_{i=1}^{l} \hat{y}_i} \tag{5.21}$$

lub mechanicznie jako:

$$S_l = \frac{\sum_{i=1}^{l} y_i}{\sum_{i=1}^{l} \overline{y}_i} \tag{5.22}$$

gdzie:
$\hat{y}_i$ – wartości teoretyczne wyznaczone na podstawie oszacowanej funkcji trendu (5.18),
$\overline{y}_i$ – wartości średnich ruchomych (5.19) lub (5.20) obliczone dla kolejnych podokresów,
S_l – wskaźniki sezonowości,
$t = 1, 2, \ldots, T$, $l = 1, 2, \ldots, L$.

Wskaźniki sezonowości interpretowane są jako względne odchylenia poziomu zjawiska od trendu lub stałej wartości. Wyrażona w procentach wartość $(S_l - 1)$ informuje, o ile procent wartość szeregu w danym „sezonie" (np. kwartale) jest wyższa (jeśli $S_l > 0$) lub niższa (jeśli $S_l < 0$) od poziomu, jaki byłby osiągnięty, gdyby w szeregu nie występowały wahania sezonowe i jego rozwój przebiegałby zgodnie z wyznaczoną tendencją.

Suma wskaźników sezonowości powinna być równa L, czyli:

$$\sum_{l=1}^{L} S_l = L \tag{5.23}$$

Jeśli ta relacja nie zachodzi, to S_l nazywany jest surowym wskaźnikiem sezonowości i należy wyznaczyć wskaźniki skorygowane postaci:

$$S_l^v = v \cdot S_l \tag{5.24}$$

gdzie:
v – współczynnik korygujący, który wyznacza się jako:

$$v = \frac{L}{\sum_{l=1}^{L} S_l} \tag{5.25}$$

Absolutną wielkość odchyleń sezonowych wyznacza się z relacji:

$$g_l(t) = S_l^v \cdot \overline{y} - \overline{y} \tag{5.26}$$

gdzie:
$\overline{y}$ – średni poziom zjawiska y_t, wyznaczony dla wszystkich obserwacji $t = 1, 2, \ldots, T$.

W modelach szeregów czasowych wahania przypadkowe odzwierciedlane są przez składnik resztowy postaci:

$$\varepsilon_t = y_t - \hat{y}_t - g_l(t) = y_t - [\hat{y}_t + g_l(t)] \tag{5.27}$$

Składnik losowy uwzględnia wszystkie czynniki, które nie zostały wyjaśnione funkcją trendu oraz wielkością wahań okresowych. Dla lepszej ilustracji stopnia dopasowania modeli uwzględniających wahania sezonowe (5.12)–(5.15) do danych empirycznych oblicza się procentowy udział tych wahań w średniej wartości szeregu czasowego, czyli:

$$\epsilon_t = \frac{\varepsilon_t}{\bar{y}} \cdot 100\% \tag{5.28}$$

Można również wyznaczyć średni błąd dopasowania obliczony jako średnia wartości bezwzględnych odchyleń utworzonego szeregu od danych empirycznych postaci:

$$M\epsilon = \frac{\sum_{t=1}^{T} |\varepsilon_t|}{T} \tag{5.29}$$

oraz procentowy średni błąd dopasowania postaci:

$$PM\epsilon = \frac{M\epsilon}{\bar{y}} \cdot 100\% \tag{5.30}$$

Przykład 5.1

Jak pokazano na rys. 5.1, produkt krajowy brutto w ujęciu kwartalnym charakteryzuje się trendem wzrostowym i wyraźnymi wahaniami sezonowymi. Szereg zawiera 22 obserwacje kwartalne z lat 2014–2019, przy czym dane z ostatniego roku obejmują jedynie pierwsze 2 kwartały. Należy przeprowadzić dekompozycję tego szeregu czasowego (tab. 5.1), szacując MNK trend liniowy i wyznaczając wskaźniki sezonowości dla poszczególnych kwartałów.

Rozwiązanie

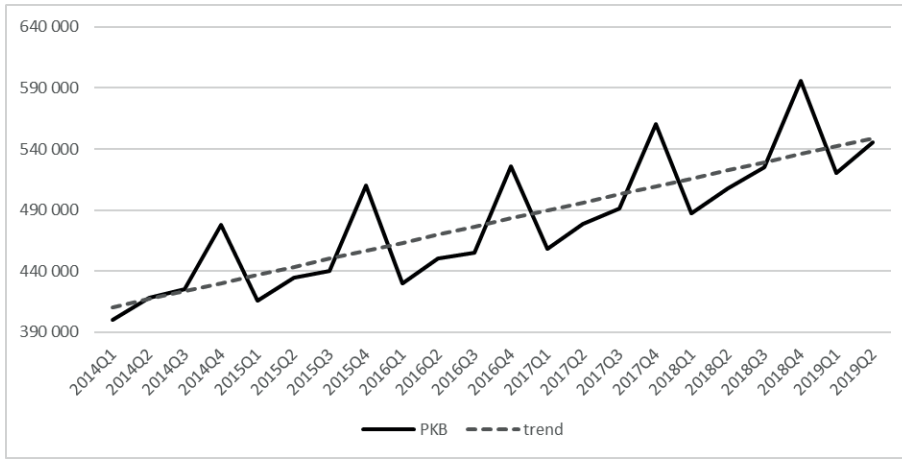
W pierwszym kroku na podstawie 22 obserwacji oszacowana MNK[6] zostaje funkcja trendu liniowego (5.17):

$$\hat{y}_t = 403758{,}245 + 6600{,}107 \cdot t \qquad R^2 = 0{,}6714$$

Oba parametry są statystycznie istotne i większe od 0, bowiem wartości statystyki testowej (4.28) wynoszą odpowiednio $t = 6{,}39$ dla współczynnika kierunkowego i $t = 29{,}78$ dla wyrazu wolnego. Jak można zauważyć na rys. 5.2, funkcja trendu opisuje wyłącznie ogólną tendencję wzrostową wartości PKB w kolejnych kwartałach. Dekompozycja szeregu polega na jego „filtrowaniu", tj. pozbawianiu go zidentyfikowanych składowych.

6 Funkcję trendu można oszacować, korzystając z funkcji REGLINP w Excelu.

W kolejnym kroku wyznacza się wartości teoretyczne PKB, podstawiając do oszacowanej funkcji trendu odpowiednie wartości zmiennej czasowej (zawarte w kolumnie 2), które wpisuje się w kolumnie 4 tab. 5.1. Na tej podstawie oblicza się wskaźniki sezonowości S_l dla każdego kwartału, czyli dla $L = 4$ ($l = 1, 2, 3, 4$). W tym celu obliczane są sumy dla $T/L = 6$ elementów stanowiące licznik i mianownik wzoru (5.21) w przypadku pierwszego i drugiego kwartału oraz dla 5 obserwacji w przypadku ostatnich dwóch kwartałów, co zapisano w kolumnach 5 i 6 tab. 5.1. A poniżej w tych samych kolumnach przedstawiono obliczone, surowe wskaźniki sezonowości oraz ich sumę.



Rysunek 5.2. PKB – dane kwartalne: porównanie danych empirycznych i trendu

Źródło: opracowanie własne.

Wyznaczone wartości surowych wskaźników sezonowych nie sumują się do 4 (ich suma wynosi 4,0135), zatem obliczono współczynnik korygujący (5.25):

$$\nu = \frac{L}{\sum_{l=1}^{L} S_l} = \frac{4}{4{,}0135} = 0{,}9966$$

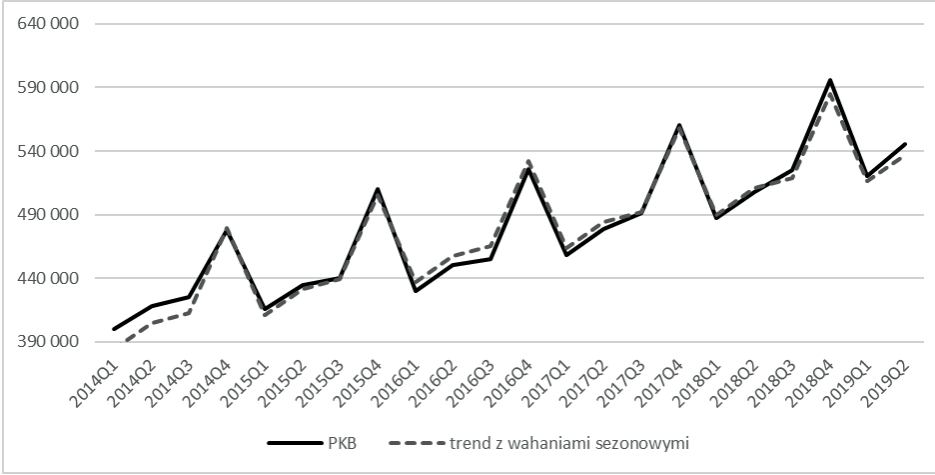
W dalszym postępowaniu wskaźniki sezonowości zostały skorygowane o wyznaczony współczynnik korygujący i w konsekwencji otrzymano wskaźniki skorygowane (5.24), które zapisano w kolumnach 5 i 6. Na ich podstawie (po wcześniejszym wyliczeniu średniej szeregu obserwacji w kolumnie 7) wyznaczono wartości odchyleń sezonowych dla każdego kwartału. Okazuje się, że w przypadku pierwszych 3 kwartałów PKB jest przeciętnie niższe od wartości znajdującej się na prostej trendu o 26 198,64; 12 032,99 i 10 737,18 mln PLN, odpowiednio w pierwszym, drugim i trzecim kwartale, a w czwartym kwartale PKB jest wyższe od trendu o 48 968,80 mln PLN (co podano w kolumnie 7).

Tabela 5.1. PKB dane kwartalne oraz wartości teoretyczne i obliczone wskaźniki i efekty sezonowe

Kwartały	t	y_t	$\hat{y}_t$ z (5.18)	Sumy wartości		Średnia
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2014Q1	1	399 536,9	410 358,35	empirycznych y_t		$\bar{y}$
2014Q2	2	418 230,9	416 958,46	I kw.	2 711 190,9	479 659,5
2014Q3	3	425 004,3	423 558,57	II kw.	2 834 623,6	
2014Q4	4	477 657,9	430 158,67	III kw.	2 336 361,0	
2015Q1	5	415 701,4	436 758,78	IV kw.	2 670 333,1	
2015Q2	6	434 227,3	443 358,89	teoretycznych $\hat{y}_t$		
2015Q3	7	439 967,9	449 959,00	I kw.	2 858 156,6	
2015Q4	8	510 331,0	456 559,10	II kw.	2 897 757,2	
2016Q1	9	430 165,1	463 159,21	III kw.	2 381 797,1	
2016Q2	10	450 116,2	469 759,32	IV kw.	2 414 797,7	
2016Q3	11	454 810,4	476 359,43	Wskaźniki sezonowości		
2016Q4	12	526 019,9	482 959,53	surowe (5.21)		
2017Q1	13	458 461,0	489 559,64	I kw.	0,9486	
2017Q2	14	478 886,8	496 159,75	II kw.	0,9782	Absolutna wielkość odchyleń sezonowych (5.26)
2017Q3	15	491 398,1	502 759,86	III kw.	0,9809	
2017Q4	16	560 568,2	509 359,96	IV kw.	1,1058	
2018Q1	17	487 129,3	515 960,07	Suma (5.23)	4,0135	
2018Q2	18	507 606,2	522 560,18	skorygowane (5.24)		
2018Q3	19	525 180,3	529 160,29	I kw.	0,9454	–26198,64
2018Q4	20	595 756,1	535 760,39	II kw.	0,9749	–12032,99
2019Q1	21	520 197,2	542 360,50	III kw.	0,9776	–10737,18
2019Q2	22	545 556,2	548 960,61	IV kw.	1,1021	48968,80

Źródło: dane w kolumnie y_t pochodzą z GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019), w pozostałych obliczenia własne.

W tab. 5.2 dwa wyróżnione w szeregu czasowym PKB elementy, tj. trend i wahania kwartalne, zostały ze sobą złożone i obliczono wahania przypadkowe w wartościach absolutnych oraz procentowych jako miara dopasowania złożonego z trendu i wahań kwartalnych szeregu do danych empirycznych. Jak można zauważyć, dla wszystkich kwartałów z wyjątkiem pierwszego wahania przypadkowe nie przekraczają 3% średniej obliczonej dla wszystkich obserwacji. Dobre dopasowanie szeregu trendu z wahaniami sezonowymi jest widoczne na rys. 5.3, na którym obie krzywe niemal się pokrywają. Natomiast na rys. 5.4 pokazano porównanie szeregów danych rzeczywistych i wahań przypadkowych. Widać na nim, że wartości niewyjaśnione przez model (zawierający trend i wahania kwartalne), uznane jako wahania przypadkowe, stanowią niewielki ułamek obserwacji PKB w badanym okresie.



Rysunek 5.3. PKB – dane kwartalne: porównanie danych empirycznych i trendu z wahaniami sezonowymi, wyznaczonymi metodą analityczną

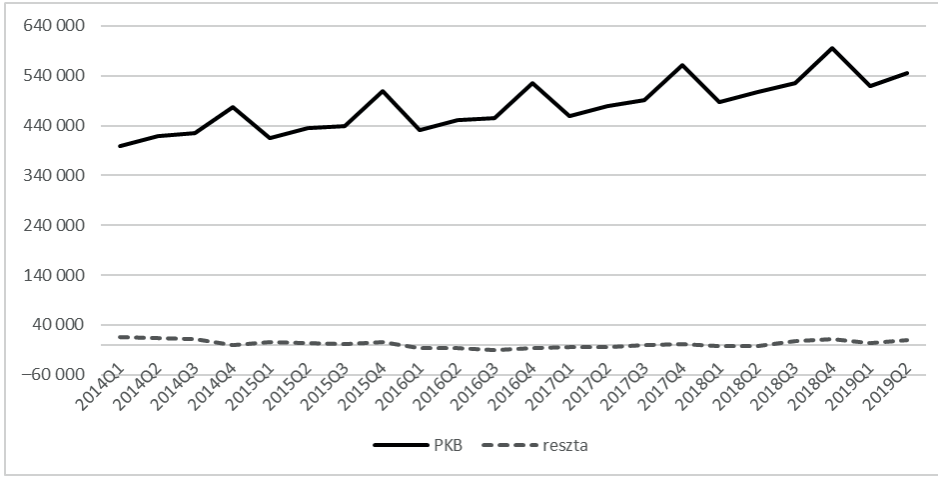
Źródło: opracowanie własne.

Tabela 5.2. PKB – dane kwartalne oraz wartości teoretyczne i odchylenia losowe

Kwartały	y_t	$\hat{y}_t$	$g_l(t)$	$\hat{y}_t + g_l(t)$	ε_t z (5.27)	ϵ_t z (5.28)
2014Q1	399 536,9	410 358,35	–26 198,64	384 159,71	15 377,19	3,2%
2014Q2	418 230,9	416 958,46	–12 032,99	404 925,47	13 305,43	2,8%
2014Q3	425 004,3	423 558,57	–10 737,18	412 821,39	12 182,91	2,5%
2014Q4	477 657,9	430 158,67	48 968,80	479 127,48	–1469,58	–0,3%
2015Q1	415 701,4	436 758,78	–26 198,64	410 560,14	5141,26	1,1%
2015Q2	434 227,3	443 358,89	–12 032,99	431 325,90	2901,40	0,6%
2015Q3	439 967,9	449 959,00	–10 737,18	439 221,82	746,08	0,2%
2015Q4	510 331,0	456 559,10	48 968,80	505 527,91	4803,09	1,0%
2016Q1	430 165,1	463 159,21	–26 198,64	436 960,57	–6795,47	–1,4%
2016Q2	450 116,2	469 759,32	–12 032,99	457 726,33	–7610,13	–1,6%
2016Q3	454 810,4	476 359,43	–10 737,18	465 622,25	–10811,85	–2,3%
2016Q4	526 019,9	482 959,53	48 968,80	531 928,34	–5908,44	–1,2%
2017Q1	458 461,0	489 559,64	–26 198,64	463 361,00	–4899,98	–1,0%
2017Q2	478 886,8	496 159,75	–12 032,99	484 126,76	–5239,95	–1,1%
2017Q3	491 398,1	502 759,86	–10 737,18	492 022,68	–624,59	–0,1%
2017Q4	560 568,2	509 359,96	48 968,80	558 328,77	2239,41	0,5%
2018Q1	487 129,3	515 960,07	–26 198,64	489 761,43	–2632,17	–0,5%
2018Q2	507 606,2	522 560,18	–12 032,99	510 527,19	–2921,03	–0,6%

Kwartały	y_t	$\hat{y}_t$	$g_1(t)$	$\hat{y}_t + g_1(t)$	ε_t z (5.27)	ε_t z (5.28)
2018Q3	525 180,3	529 160,29	–10 737,18	518 423,11	6757,22	1,4%
2018Q4	595 756,1	535 760,39	48 968,80	584 729,20	11 026,90	2,3%
2019Q1	520 197,2	542 360,50	–26 198,64	516 161,86	4035,34	0,8%
2019Q2	545 556,2	548 960,61	–12 032,99	536 927,62	8628,58	1,8%

Źródło: dane w kolumnie y_t pochodzą z GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019), w pozostałych obliczenia własne.



Rysunek 5.4. PKB – dane kwartalne: porównanie danych empirycznych i wahań przypadkowych

Źródło: opracowanie własne.

Podobny (do dekompozycji szeregu czasowego, zilustrowanego przykładem 5.1) efekt modelowania szeregu czasowego z regularnymi wahaniami sezonowymi można uzyskać, stosując tzw. trendy jednoimiennych okresów. Metoda ta polega na szacowaniu kilku funkcji trendu, które opisują zachowanie się szeregu w tzw. jednoimiennych okresach, którymi mogą być np. kwartały, jeśli wahania sezonowe mają charakter odchyleń kwartalnych. Ważnym ograniczeniem tej metody jest konieczność posiadania prób o takiej długości, aby zastosowanie MNK do estymacji modeli trendu było możliwe.

Przykład 5.2

Na podstawie danych z przykładu 5.1 należy oszacować modele trendu dla każdego kwartału i sprawdzić, czy wyznaczone na ich podstawie wartości teoretyczne

szeregu PKB są zgodne z wartościami empirycznymi. Ponadto porównać należy stopień dopasowania wartości teoretycznych do empirycznych w obu przykładach.

Rozwiązanie

W pierwszym kroku tworzy się szeregi jednoimiennych okresów, czyli dane zostają podzielone na 4 szeregi, w których znajdą się obserwacje z wszystkich analizowanych lat, ale dotyczące konkretnego kwartału. Tak pogrupowane dane przedstawiono w tab. 5.3.

Tabela 5.3. Obserwacje kwartalne PKB uporządkowane dla kolejnych kwartałów, oszacowania modeli trendu jednoimiennych okresów i wartości teoretyczne

Lata	y_t	t	Oszacowane modele trendu		Wartości teoretyczne trendów jednoimiennych
Kwartał I					
2014	399 536,9	1	$\hat{y}_t = 367277{,}045 + 24168{,}028 \cdot t$		391 445,07
2015	415 701,4	2	$R^2 = 0{,}9760$		415 613,10
2016	430 165,1	3	Wartość statystyki t-Studenta		439 781,13
2017	458 461,0	4	β_1	β_0	463 949,16
2018	487 129,3	5	12,747	49,739	488 117,19
2019	520 197,2	6			512 285,22
Kwartał II					
2014	418 230,9	1	$\hat{y}_t = 383883{,}893 + 25300{,}963 \cdot t$		409 184,86
2015	434 227,3	2	$R^2 = 0{,}9723$		434 485,82
2016	450 116,2	3	Wartość statystyki t-Studenta		459 786,78
2017	478 886,8	4	β_1	β_0	485 087,74
2018	507 606,2	5	11,853	46,178	510 388,71
2019	545 556,2	6			535 689,67
Kwartał III					
2014	425 004,3	1	$\hat{y}_t = 391737{,}529 + 25178{,}225 \cdot t$		416 915,75
2015	439 967,9	2	$R^2 = 0{,}9572$		442 093,98
2016	454 810,4	3	Wartość statystyki t-Studenta		467 272,20
2017	491 398,1	4	β_1	β_0	492 450,43
2018	525 180,3	5	8,192	38,431	517 628,65
Kwartał IV					
2014	477 657,9	1	$\hat{y}_t = 448136{,}542 + 28643{,}358 \cdot t$		476 779,90
2015	510 331,0	2	$R^2 = 0{,}98763$		505 423,26
2016	526 019,9	3	Wartość statystyki t-Studenta		534 066,62
2017	560 568,2	4	β_1	β_0	562 709,97
2018	595 756,1	5	14,721	69,443	591 353,33

Źródło: opracowanie własne.

W kolejnym kroku dla każdego kwartału szacuje się funkcje trendu liniowego (5.17) na podstawie danych z lat 2014–2019. Można zauważyć, że w przypadku ostatnich 2 kwartałów brakuje danych za rok 2019, zatem szeregi dla kwartałów 3 i 4 są o jedną obserwację krótsze. Innymi słowy, dla dwóch pierwszych kwartałów modele szacowane są na podstawie 6 obserwacji $t = 1, 2, \ldots, 6$, a dla 2 ostatnich dla $t = 1, 2, \ldots, 5$. W tab. 5.3 przedstawiono oszacowane modele trendu liniowego dla każdego z kwartałów. Dla wszystkich 4 jednoimiennych okresów współczynniki determinacji są bliskie jedności i zasadniczo wyższe niż w przypadku funkcji trendu oszacowanej w przykładzie 5.1. Podane wartości statystyk testowych t-Studenta dla wszystkich 4 modeli trendu wskazują na istotne zależności. W związku z tym w kolejnym kroku wyznaczono na podstawie oszacowanych modeli trendu wartości teoretyczne dla kolejnych kwartałów.

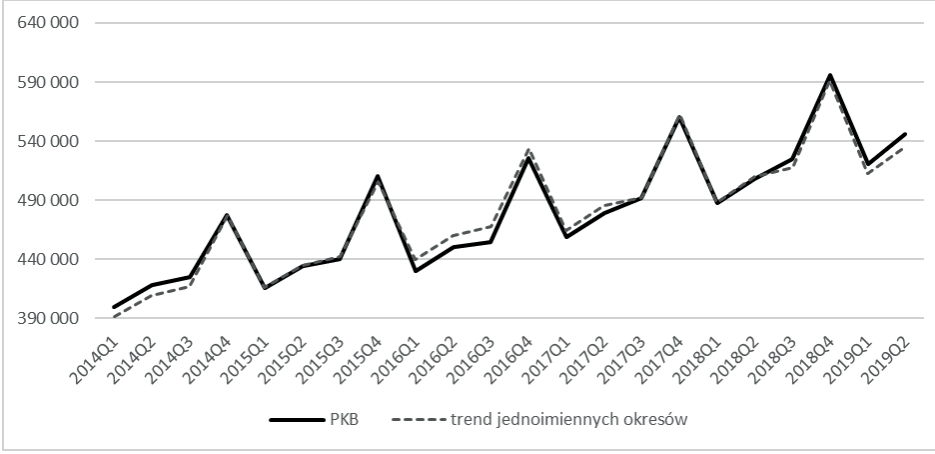
Tabela 5.4. Obserwacje kwartalne PKB i wartości teoretyczne wyznaczone z modeli trendu jednoimiennych okresów

Kwartały	PKB	Wartości teoretyczne trendów jednoimiennych okresów	ε_t z (5.27)	Moduł ε_t z przykładu 5.2	Moduł ε_t z przykładu 5.1
2014Q1	399 536,9	391 445,07	8 091,83	8 091,83	15 377,19
2014Q2	418 230,9	409 184,86	9 046,04	9 046,04	13 305,43
2014Q3	425 004,3	416 915,75	80 88,55	8 088,55	12 182,91
2014Q4	477 657,9	476 779,90	878,00	878,00	1 469,58
2015Q1	415 701,4	415 613,10	88,30	88,30	5 141,26
2015Q2	434 227,3	434 485,82	–258,52	258,52	2 901,40
2015Q3	439 967,9	442 093,98	–2126,08	2 126,08	746,08
2015Q4	510 331,0	505 423,26	4 907,74	4 907,74	4 803,09
2016Q1	430 165,1	439 781,13	–9 616,03	9 616,03	6 795,47
2016Q2	450 116,2	459 786,78	–9 670,58	9 670,58	7 610,13
2016Q3	454 810,4	467 272,20	–12 461,80	12 461,80	10 811,85
2016Q4	526 019,9	534 066,62	–8 046,72	8 046,72	5 908,44
2017Q1	458 461,0	463 949,16	–5 488,16	5 488,16	4 899,98
2017Q2	478 886,8	485 087,74	–6 200,94	6 200,94	5 239,95
2017Q3	491 398,1	492 450,43	–1 052,33	1 052,33	624,59
2017Q4	560 568,2	562 709,97	–2 141,77	2 141,77	2 239,41
2018Q1	487 129,3	488 117,19	–987,89	987,89	2 632,17
2018Q2	507 606,2	510 388,71	–2 782,51	2782,51	2 921,03
2018Q3	525 180,3	517 628,65	7 551,65	7 551,65	6 757,22
2018Q4	595 756,1	591 353,33	4 402,77	4 402,77	11 026,90
2019Q1	520 197,2	512 285,21	7 911,99	7 911,99	4 035,34
2019Q2	545 556,2	535 689,67	9 866,53	9 866,53	8 628,58
			Suma	121 666,71	136 058,00

Źródło: opracowanie własne.

Ostatnim etapem postępowania jest przypisanie wartości teoretycznych kolejnym obserwacjom, poczynając od pierwszego kwartału 2014 roku, a na drugim kwartale 2019 roku kończąc. Przykładowo wartości teoretyczne dla 4 kwartałów 2014 roku wyznaczane są dla $t = 1$, ale według różnych modeli, opisujących kwartały. W ten sposób otrzymuje się szereg wartości teoretycznych uwzględniających już wahania sezonowe, co pokazano w ostatniej kolumnie tab. 5.3, oraz po uporządkowaniu (dla kolejnych kwartałów i lat) w tab. 5.4.

Jak można zauważyć na rys. 5.5, stopień dopasowania szeregu wygenerowanego przez złożenie wartości teoretycznych (wyznaczonych z funkcji trendu jednoimiennych okresów) do danych empirycznych jest bardzo wysoki. Dla porównania wyznaczono wartości procentowego średniego błędu dopasowania dla szeregów wygenerowanych w przykładzie 5.1 i 5.2. Podstawowe obliczenia przedstawiono w ostatnich dwóch kolumnach tab. 5.4. W przypadku dekompozycji szeregu czasowego na trend i wahania sezonowe (przykład 5.1) błąd ten wynosi 1,29%, a stosując metodę trendów jednoimiennych – 1,15%, co oznacza bardzo wysokie dopasowanie w obu przypadkach.



Rysunek 5.5. PKB – dane kwartalne: porównanie danych empirycznych i wartości teoretycznych trendów jednoimiennych okresów

Źródło: opracowanie własne.

Przykład 5.3

Dla danych z przykładu 5.1 należy wyznaczyć wskaźniki sezonowe metodą mechaniczną.

Rozwiązanie

W obliczeniach kwartalnych wskaźników sezonowości metodą mechaniczną wykorzystuje się scentrowaną średnią ruchomą 4-okresową, którą wyznacza się na podstawie wzoru (5.20). Pierwsza ze średnich przypisana zostanie trzeciemu kwartałowi roku 2014 i zostanie obliczona następująco:

$$\bar{y}_3 = \frac{1}{4}\cdot(0{,}5\cdot y_1 + y_2 + y_3 + y_4 + 0{,}5\cdot y_5) =$$
$$= 0{,}25\cdot(0{,}5\cdot 399536{,}9 + 418230{,}9 + 425004{,}3 + 477657{,}9 + 0{,}5\cdot 415701{,}4) =$$
$$= 432128{,}1$$

W podobny sposób oblicza się średnie ruchome dla $t = 4, 5, \ldots, 20$, co przedstawiono w tab. 5.5 i na rys. 5.6.

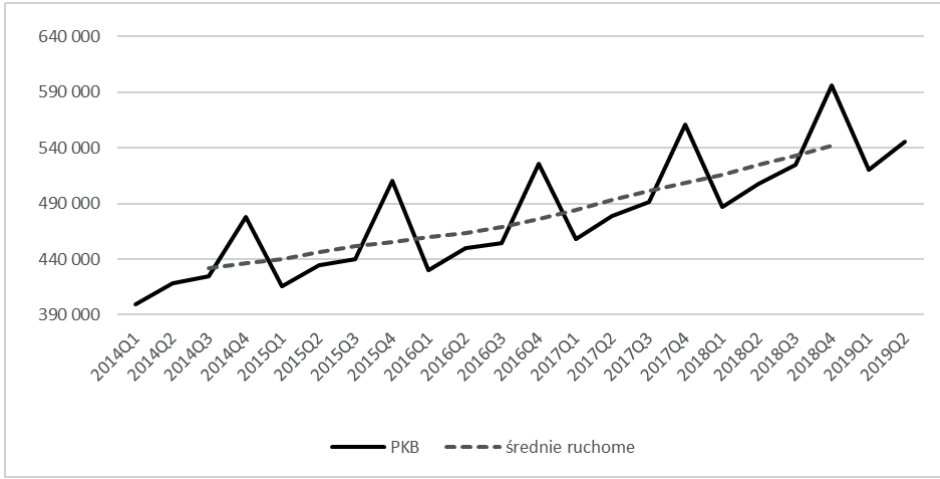
Tabela 5.5. PKB – dane kwartalne, wartości teoretyczne trendu, średnie ruchome i obliczone wskaźniki oraz efekty sezonowe

Kwartały	t	y_t	$\hat{y}_t$ z (5.18)	Średnie ruchome $\bar{y}_t$	Sumy wartości empirycznych y_t		Średnia $\bar{y}$
2014Q1	1	399 536,9	410 358,35				
2014Q2	2	418 230,9	416 958,46		I kw.	1791 456,8	479 659,5
2014Q3	3	425 004,3	423 558,57	432 128,06	II kw.	2289 067,4	
2014Q4	4	477 657,9	430 158,67	436 148,18	III kw.	2336 361,0	
2015Q1	5	415 701,4	436 758,78	440 018,18	IV kw.	2670 333,1	
2015Q2	6	434 227,3	443 358,89	445 972,76	średnich $\bar{y}_t$		
2015Q3	7	439 967,9	449 959,00	451 864,86	I kw.	1899 534,7	
2015Q4	8	510 331,0	456 559,10	455 658,94	II kw.	1926 819,0	
2016Q1	9	430 165,1	463 159,21	459 500,36	III kw.	2386 771,3	
2016Q2	10	450 116,2	469 759,32	463 316,79	IV kw.	2417 769,5	
2016Q3	11	454 810,4	476 359,43	468 814,89	Wskaźniki sezonowości		
2016Q4	12	526 019,9	482 959,53	475 948,21	surowe (5.22)		
2017Q1	13	458 461,0	489 559,64	484 117,99	I kw.	0,9431	
2017Q2	14	478 886,8	496 159,75	493 009,99	II kw.	1,1880	Absolutna wielkość odchyleń sezonowych (5.26)
2017Q3	15	491 398,1	502 759,86	500 912,06	III kw.	0,9789	
2017Q4	16	560 568,2	509 359,96	508 085,50	IV kw.	1,1045	
2018Q1	17	487 129,3	515 960,07	515 898,20	Suma (5.23)	4,2144	
2018Q2	18	507 606,2	522 560,18	524 519,47	skorygowane (5.24)		
2018Q3	19	525 180,3	529 160,29	533 051,46	I kw.	0,8951	–49 980,29
2018Q4	20	595 756,1	535 760,39	541 928,70	II kw.	1,1276	60 781,86
2019Q1	21	520 197,2	542 360,50		III kw.	0,9291	–33 799,57
2019Q2	22	545 556,2	548 960,61		IV kw.	1,0483	22 997,99

Źródło: dane w kolumnie y_t pochodzą z GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019), w pozostałych obliczenia własne.

W dalszym postępowaniu obliczone zostaną wskaźniki sezonowości według wzoru (5.22), przy czym należy pamiętać, że korzystając ze scentrowanych średnich ruchomych 4-okresowych, traci się po dwie obserwacje na początku i końcu szeregu. Dlatego obliczone sumy dotyczyć będą obu szeregów y_t i $\overline{y}_t$ jedynie w odniesieniu do danych, które są wspólne (u nas dla t = 3, 5, …, 20). Dalej należy postąpić podobnie jak w przykładzie 5.1, zatem wyznaczone zostaną wartości odchyleń kwartalnych, które zostaną dodane do obliczonych wcześniej wartości teoretycznych.

Dalej wyznaczone będą błędy dopasowania wygenerowanego szeregu do danych empirycznych. Jak można zauważyć w tab. 5.6, zastosowanie średnich ruchomych do wyznaczenia wskaźników sezonowości wygenerowało wyższe błędy dopasowania niż w przypadku metod analitycznych. Błędy względne przekraczają nawet 16% dla pojedynczych okresów, a średni błąd procentowy wynosi 8,06%, zatem jest wyższy od tych wyznaczonych w przykładach 5.1 i 5.2, co jest widoczne również na rys. 5.7.



Rysunek 5.6. PKB – porównanie notowań kwartalnych i średnich ruchomych

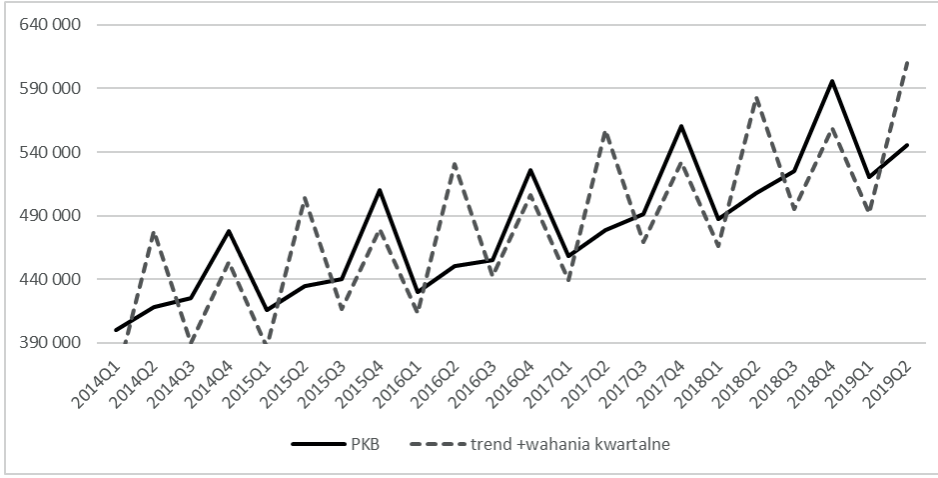
Źródło: opracowanie własne.

Tabela 5.6. PKB – dane kwartalne oraz wartości teoretyczne i odchylenia losowe

Kwartały	y_t	$\hat{y}_t + g_l(t)$	ε_t z (5.27)	ϵ_t z (5.28)	Moduł ε_t
2014Q1	399 536,9	360 378,1	39 158,8	8,2%	39 158,84
2014Q2	418 230,9	477 740,3	–59 509,4	–12,5%	59 509,42
2014Q3	425 004,3	389 759,0	35 245,3	7,4%	35 245,30
2014Q4	477 657,9	453 156,7	24 501,2	5,1%	24 501,23

Kwartały	y_t	$\hat{y}_t + g_l(t)$	ε_t z (5.27)	ϵ_t z (5.28)	Moduł ε_t
2015Q1	415 701,4	386 778,5	28 922,9	6,1%	28 922,91
2015Q2	434 227,3	504 140,8	–69 913,5	–14,7%	69 913,45
2015Q3	439 967,9	416 159,4	23 808,5	5,0%	23 808,47
2015Q4	510 331,0	479 557,1	30 773,9	6,5%	30 773,90
2016Q1	430 165,1	413 178,9	16 986,2	3,6%	16 986,18
2016Q2	450 116,2	530 541,2	–80 425,0	–16,9%	80 424,98
2016Q3	454 810,4	442 559,9	12 250,5	2,6%	12 250,54
2016Q4	526 019,9	505 957,5	20 062,4	4,2%	20 062,37
2017Q1	458 461,0	439 579,3	18 881,7	4,0%	18 881,67
2017Q2	478 886,8	556 941,6	–78 054,8	–16,4%	78 054,80
2017Q3	491 398,1	468 960,3	22 437,8	4,7%	22 437,80
2017Q4	560 568,2	532 358,0	28 210,2	5,9%	28 210,22
2018Q1	487 129,3	465 979,8	21 149,5	4,4%	21 149,49
2018Q2	507 606,2	583 342,0	–75 735,9	–15,9%	75 735,88
2018Q3	525 180,3	495 360,7	29 819,6	6,3%	29 819,61
2018Q4	595 756,1	558 758,4	36 997,7	7,8%	36 997,71
2019Q1	520 197,2	492 380,2	27 817,0	5,8%	27 816,99
2019Q2	545 556,2	609 742,5	–64 186,3	–13,5%	64 186,27
				Suma	844 848,05

Źródło: dane w kolumnie y_t pochodzą z GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019), w pozostałych obliczenia własne.



Rysunek 5.7. Porównanie PKB i trendu z wahaniami sezonowymi, wyznaczonymi metodą mechaniczną

Źródło: opracowanie własne.

5.3. Metody wygładzania szeregów czasowych

Szeregi czasowe są szeroko wykorzystywane w praktyce zarówno do analiz przeszłości i bieżącej sytuacji, jak i prognozowania. Dlatego też istotne znaczenie praktyczne mają metody wygładzania szeregów czasowych. Najczęściej wykorzystywane są do tego celu modele średnich ruchomych (kroczących).

Średnia krocząca wyznaczana jest na podstawie określonego zbioru danych. Wygładzanie szeregu za jej pomocą polega na zastąpieniu pierwotnych wartości zmiennej średnimi arytmetycznymi, obliczanymi sekwencyjnie dla wybranej liczby obserwacji (np. dla 15 kolejnych notowań akcji danej spółki). W dalszym postępowaniu, usuwając ze zbioru danych kolejno wyniki najstarsze i dodając równocześnie coraz to nowsze, wyznacza się średnią ruchomą dla kolejnych okresów. Średnią ruchomą (kroczącą) stosuje się, aby wygładzić wahnięcia (odchylenia) zmiennej, np. kursu akcji. Podstawowymi rodzajami średnich ruchomych są średnie: arytmetyczna, ważona liniowa, ważona potęgowo, ważona wykładnicza i geometryczna.

Średnia arytmetyczna, zwana również prostą średnią kroczącą, jest najczęściej stosowaną i najprostszą do wyznaczenia. Pierwszym elementem przy jej konstrukcji jest wyznaczenie liczby, która określa długość kroku uśrednienia $k < t$ ($t = 1, 2, \ldots$). Model średniej ruchomej prostej (5.19) można zapisać prościej jako:

$$\overline{y}_t = \frac{1}{k} \sum_{i=t-k}^{t-1} y_i \tag{5.31}$$

gdzie:
$\overline{y}_t$ – wartości średnich ruchomych obliczone dla kolejnych podokresów,
y_i – wartość zmiennej w momencie (okresie) i,
k – stała wygładzania.

Wadą średniej ruchomej (zwłaszcza dla dużego k, np. $k = 15$) jest przypisywanie takiego samego znaczenia obserwacjom odległym i najnowszym. W celu uwzględnienia postulatu większego wpływu na średnią obserwacji najnowszych stosuje się tzw. średnią ważoną liniowo, która jest następującej postaci:

$$\overline{y}_t = \sum_{i=t-k}^{t-1} y_i w_{i-t+k+1} \tag{5.32}$$

gdzie:
$w_{i-t+k+1}$ – waga nadana wartościom zmiennej w okresie i, a wagi przypisane kolejnym obserwacjom spełniają warunki $0 < w_1 < w_2 < \ldots < w_k \leq 1$ oraz $\sum_{i=1}^{k} w_i = 1$

Należy zauważyć, że ten rodzaj średniej jest celowy i przydatny tylko w przypadku zmiennych, których wartości zmieniają się ewolucyjnie bez gwałtownych wahań i odchyleń.

Inną metodą stosowaną do wygładzania szeregów czasowych jest średnia wykładnicza postaci:

$$\overline{y}_t = \alpha \cdot y_{t-1} + (1 - \alpha) \cdot \overline{y}_{t-1}, \quad \text{dla } \alpha \in (0; 1] \tag{5.33}$$

którą powinno się stosować w przypadku zmiennych, których wartości podlegają częstym, gwałtownym i losowym wahaniom. Po podstawieniu $q_{t-1} = \overline{y}_{t-1} - y_{t-1}$ wzór (5.33) można zapisać następująco:

$$\overline{y}_t = \overline{y}_{t-1} - \alpha \cdot q_{t-1} \tag{5.34}$$

gdzie:
α – tzw. parametr wygładzania, czyli waga dla ostatniej (najnowszej) obserwacji zmiennej,
q_{t-1} – błąd średniej kroczącej wyznaczonej na okres $t - 1$.

Podstawowym problemem w przypadku stosowania średnich wykładniczych jest ustalenie wartości parametru wygładzania. Dokonuje się tego zazwyczaj eksperymentalnie, tj. przyjmując różne wartości α, i sprawdza się, dla której z nich błąd jest najmniejszy.

Przykład 5.4

Dane kwartalne GUS odnoszące się do wskaźników makroekonomicznych dotyczące działalności finansowej i ubezpieczeniowej za lata 2014–2019 przedstawiono w kolumnie 4 tab. 5.7. Należy wygładzić szereg czasowy za pomocą prostych średnich ruchomych 3- i 5-okresowych oraz 3-okresowej średniej ważonej liniowo dla wag $w_1 = 0{,}2$, $w_2 = 0{,}3$ i $w_3 = 0{,}5$.

Rozwiązanie

Na podstawie wzoru (5.19) lub (5.31) wyznaczyć należy proste średnie ruchome postaci:

- 3-okresową: $\overline{y}_2 = \frac{y_1 + y_2 + y_3}{3}$, $\overline{y}_3 = \frac{y_2 + y_3 + y_4}{3}$, …, $\overline{y}_{T-1} = \frac{y_{T-2} + y_{T-1} + y_T}{3}$

oraz

- 5-okresową: $\overline{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$, $\overline{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$, …, $\overline{y}_{T-2} = \frac{y_{T-4} + y_{T-3} + y_{T-2} + y_{T-1} + y_T}{5}$

Wartości obliczonych średnich zapisano w tab. 5.7 oraz na rys. 5.8.

W dalszym postępowaniu wyznaczone zostaną średnie ważone. Tak więc pierwsza średnia (przypisana drugiej obserwacji) wyznaczona została jako:

$$y_2 = w_1 \cdot y_1 + w_2 \cdot y_2 + w_3 \cdot y_3$$

druga jako:

$$y_3 = w_1 \cdot y_2 + w_2 \cdot y_3 + w_3 \cdot y_4$$

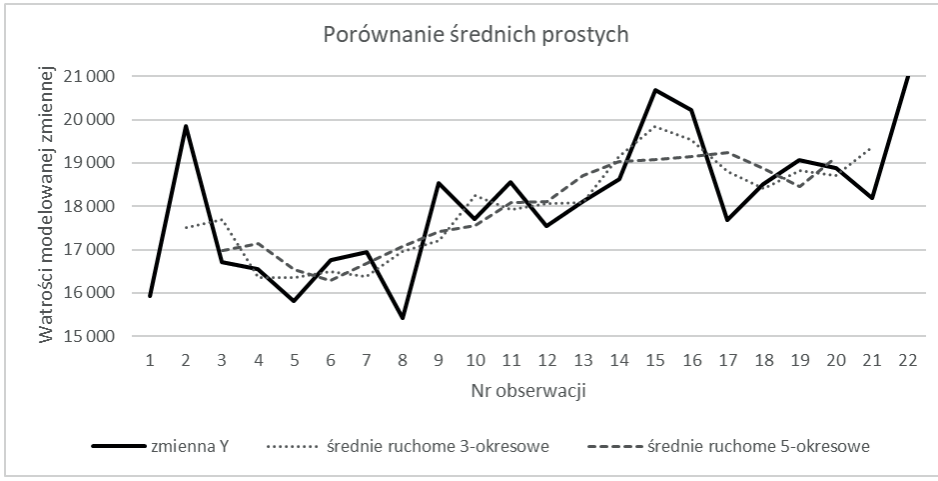
Jak można zauważyć na rys. 5.8, szeregi czasowe średnich ruchomych są bardziej gładkie niż szereg danych empirycznych. Przy czym im dłuższy krok uśredniania k, tym wykres szeregu jest bardziej gładki. Z kolei na rys. 5.9 porównano średnie ruchome 3-okresowe proste i ważone.

Tabela 5.7. Dane dotyczące działalności finansowej i ubezpieczeniowej [w mln. PLN] oraz obliczone proste średnie ruchome 3- i 5-okresowe

Rok	kwartał	t	Zmienna y_t	Średnie ruchome $\bar{y}_t$		Średnie ruchome			Ważone $\bar{y}_t$
(1)	(2)	(3)	(4)	(5)	(6)	Wagi			
				3-okresowe	5-okresowe				
2014	I	1	15 933,4			0,2			
	II	2	19 852,0	17 500,9		0,3	0,2		17 500,93
	III	3	16 717,3	17 703,3	16 969,0	0,5	0,3	0,2	17 255,84
	IV	4	16 540,5	16 353,2	17 131,8	0,2	0,5	0,3	16 206,56
2015	I	5	15 801,9	16 363,3	16 550,4	0,3	0,2	0,5	16 422,42
	II	6	16 747,5	16 498,1	16 290,3	0,5	0,3	0,2	16 657,13
	III	7	16 945,0	16 369,6	16 688,0	0,2	0,5	0,3	16 141,2
	IV	8	15 416,4	16 963,4	17 066,4	0,3	0,2	0,5	17 278,37
2016	I	9	18 528,9	17 213,1	17 427,3	0,5	0,3	0,2	17 489,00
	II	10	17 694,1	18 258,4	17 545,9	0,2	0,5	0,3	18 290,06
	III	11	18 552,1	17 928,0	18 088,1	0,3	0,2	0,5	17 873,35
	IV	12	17 537,8	18 072,5	18 107,3	0,5	0,3	0,2	18 035,61
2017	I	13	18 127,7	18 096,7	18 705,4	0,2	0,5	0,3	18 258,22
	II	14	18 624,7	19 145,6	19 038,6	0,3	0,2	0,5	19 555,20
	III	15	20 684,5	19 842,6	19 068,1	0,5	0,3	0,2	20 039,54
	IV	16	20 218,5	19 529,3	19 143,5	0,2	0,5	0,3	19 044,95
2018	I	17	17 685,0	18 802,7	19 232,2	0,3	0,2	0,5	18 601,50
	II	18	18 504,6	18 419,4	18 869,8	0,5	0,3	0,2	18 622,68
	III	19	19 068,6	18 815,2	18 463,6	0,2	0,5	0,3	18 857,75
	IV	20	18 872,5	18 709,5	19 127,1	0,3	0,2	0,5	18 569,12
2019	I	21	18 187,3	19 354,0		0,5	0,3	0,2	19 731,84
	II	22	21 002,3				0,5		

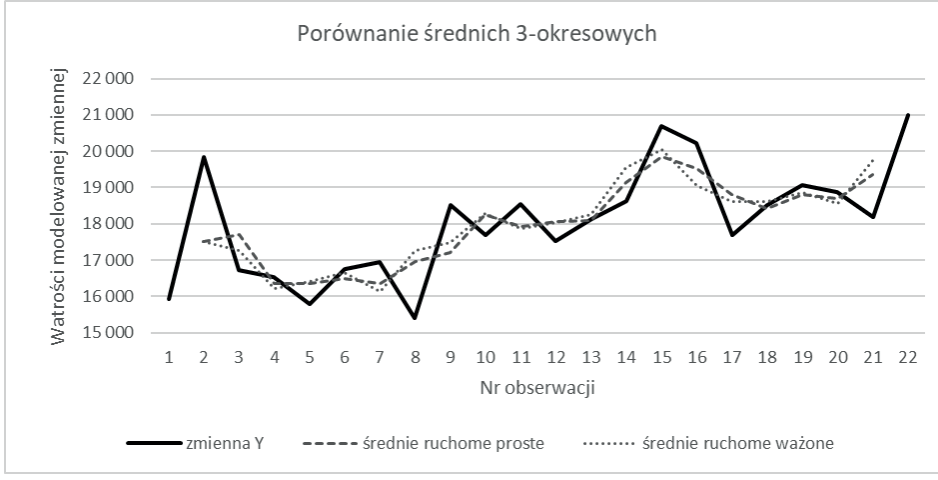
Uwaga: kolorami zaznaczono średnie ruchome oraz zakresy zmiennych, z jakich zostały wyznaczone w przypadku 3 pierwszych i 3 ostatnich średnich.

Źródło: dane w kolumnie 4: GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019); dane w kolumnie 5 i 6: obliczenia własne.



Rysunek 5.8. Porównanie średnich ruchomych prostych o kroku uśredniania 3 i 5 z empirycznymi wartościami oznaczającymi wartość działalności finansowej i ubezpieczeniowej

Źródło: opracowanie własne.



Rysunek 5.9. Porównanie średnich ruchomych 3-okresowych prostych i ważonych z empirycznymi wartościami oznaczającymi wartość działalności finansowej i ubezpieczeniowej

Źródło: opracowanie własne.

Najczęściej stosowanymi przez inwestorów średnimi ruchomymi są średnie arytmetyczne oraz wykładnicze. Ze względu na długość średniej kroczącej dzieli

się je na: krótkoterminowe, średnioterminowe i długoterminowe. Wybór właściwej średniej jest bardzo złożony. Dobór odpowiedniego kroku uśrednienia k jest w tym wypadku głównym problemem i zależy od badanego zjawiska.

Średnie kroczące są m.in. wykorzystywane przez inwestorów giełdowych do wyznaczania sygnałów kupna i sprzedaży instrumentów finansowych. Od długości średniej kroczącej zależy np. jakość i liczba uzyskanych sygnałów. Ogólna zależność jest następująca: im krótszy krok uśrednienia (krótsza średnia), tym liczba sygnałów (zarówno tych, które się następnie potwierdzają, jak i błędnych) jest większa i na odwrót. Długość średniej, jaką inwestor rozpatruje i stosuje w swoich analizach, powinna zależeć od przyjętego horyzontu inwestycyjnego.

Przykład 5.5

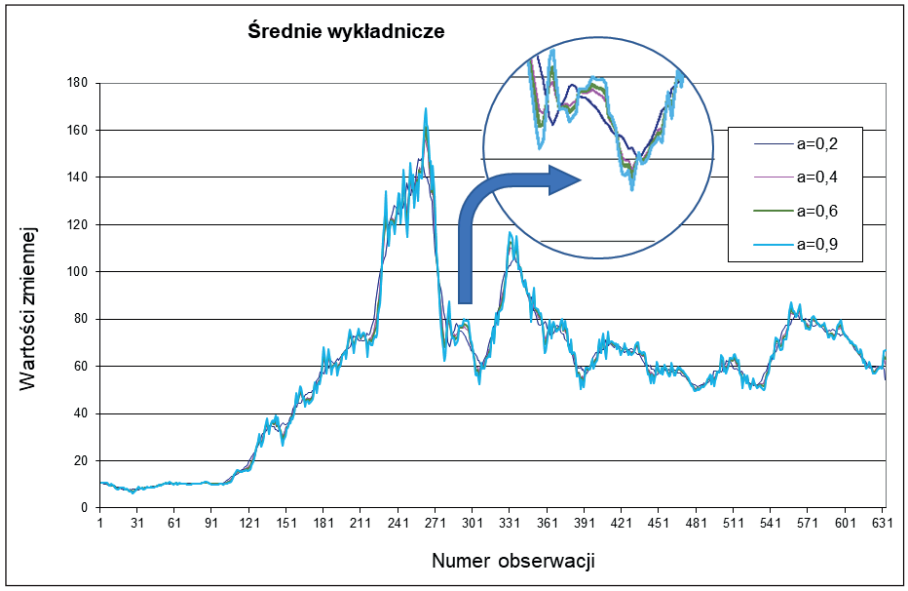
Dla 645 obserwacji kursów akcji spółki Okocim SA notowanych na Giełdzie Papierów Wartościowych w Warszawie wyznaczyć należy średnią kroczącą prostą 15-okresową oraz średnią wykładniczą dla różnych parametrów wygładzania.

Rozwiązanie

Średnie ruchoma prosta i wykładnicze zostały przedstawione na rys. 5.10 i 5.11. W celu bliższej analizy różnic wynikających z zastosowania różnych wartości parametru wygładzania trzeba się przyjrzeć fragmentowi analizowanego na poprzednim rysunku (rys. 5.12) wykresu dotyczącego obserwacji od 222 do 355, był to bowiem okres gwałtownych zmian kursów akcji.

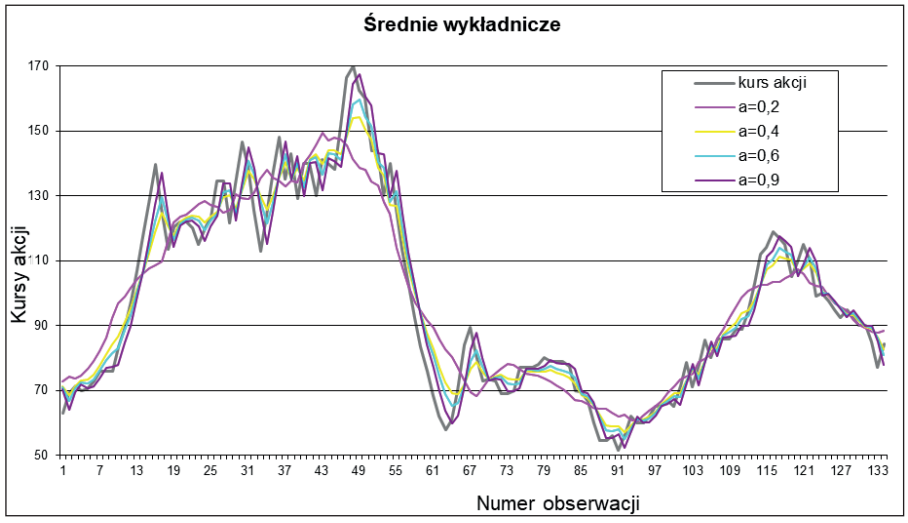
Rysunek 5.10. Porównanie kursów akcji Okocim SA i średniej ruchomej prostej 15-okresowej

Źródło: opracowanie własne.



Rysunek 5.11. Przykłady średnich wykładniczych dla różnych parametrów wygładzania

Źródło: opracowanie własne.



Rysunek 5.12. Porównanie średnich wykładniczych z rzeczywistymi wartościami kursów akcji Okocim SA dla obserwacji od 222 do 355

Źródło: opracowanie własne.

Dla szeregów czasowych, charakteryzujących się tendencją rozwojową i wahaniami przypadkowymi, odpowiednią metodą wygładzania może okazać się liniowy model Holta, który jest postaci:

$$\overline{y}_{t+\tau} = M_{t-1} + (1+\tau)S_{t-1} \tag{5.35}$$

$$M_{t-1} = \alpha y_{t-1} + (1-\alpha)(M_{t-2} + S_{t-2}) \tag{5.36}$$

$$S_{t-1} = \gamma(M_{t-1} - M_{t-2}) + (1-\gamma)S_{t-2} \tag{5.37}$$

gdzie:
M_{t-1} – ocena wartości średniej na moment lub okres $t - 1$; przyjmuje się, że $M_{t-1} = y_1$ lub M_1 jest wyrazem wolnym linowej funkcji trendu, oszacowanej na podstawie próby pilotażowej,
S_{t-1} – wygładzona wartość przyrostu trendu na moment lub okres $t - 1$; przyjmuje się, że $S_1 = y_2 - y_1$ lub S_1 jest współczynnikiem kierunkowym linowej funkcji trendu, oszacowanej na podstawie próby pilotażowej,
α – stała wyrównywania szeregu, $\alpha \in [0; 1]$,
γ – parametr wyrównywania trendu, $\gamma \in [0; 1]$.

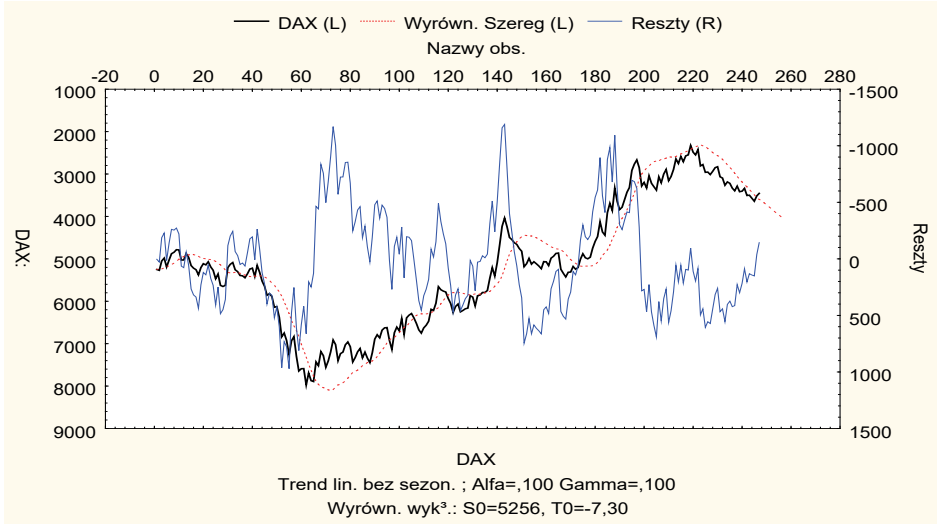
Podstawowym problemem jest wyznaczenie stałej wyrównywania szeregu oraz parametru wygładzania. Znalezienie tych parametrów sprowadza się zazwyczaj do przeprowadzenia szeregu eksperymentów komputerowych.

Przykład 5.6

Dany jest szereg czasowy złożony ze średnich tygodniowych kursów zamknięcia indeksu DAX w ciągu niemal 5 lat (247 obserwacji). Szereg ten poddano wygładzaniu wykładniczemu metodą Holta. W tym celu wykorzystano pakiet STATISTICA dla arbitralnie ustalonych parametrów wygładzania $\alpha = \gamma = 0{,}1$ oraz zastosowano opcję automatycznego poszukiwania najlepszych parametrów.

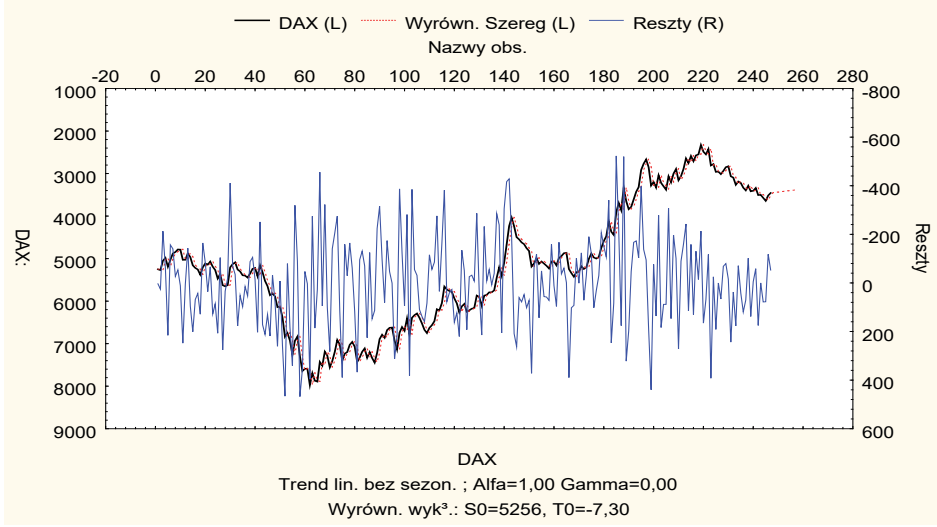
Rozwiązanie

Rys. 5.13 i 5.14 zawierają wykresy danych rzeczywistych i wygładzonych. Można zauważyć, że przedstawione na rysunkach reszty odpowiadają skali zaznaczonej po prawej stronie wykresów, podczas gdy wartości indeksu DAX przedstawiono na skali po lewej stronie wykresu. Jak łatwo można zauważyć, wygładzony szereg znacznie lepiej przystaje do wartości empirycznych, w przypadku gdy parametry wygładzania zostały dobrane eksperymentalnie przez program.



Rysunek 5.13. Porównanie wartości empirycznych i wygładzonych metodą Holta dla przyjętych arbitralnie parametrów wygładzania

Źródło: opracowanie własne.



Rysunek 5.14. Porównanie wartości empirycznych wygładzonych metodą Holta przy automatycznym wyszukiwaniu parametrów wygładzania

Źródło: opracowanie własne.

Rozdział 6

Modelowanie ekonometryczne

Ekonometria zajmuje się ustalaniem, za pomocą metod statystycznych, ilościowych prawidłowości zachodzących w życiu gospodarczym. Według Gregory'ego C. Chowa [1995, s. 15] ekonometria „jest nauką i sztuką stosowania metod statystycznych do mierzenia relacji ekonomicznych". Modele ekonometryczne wywodzą się z modeli regresji, które są wykorzystywane niemal we wszystkich dyscyplinach badawczych. Jednakże specyfika zjawisk społeczno-ekonomicznych sprawia, że pierwotna funkcja regresji musiała zostać zmodyfikowana, a niektóre założenia, np. dotyczące braku w modelu czynników jakościowych, zniesione. Spowodowało to, że ekonometria – jako nauka – wykształciła własne metody badawcze, a rozwój modelowania ekonometrycznego pozwolił na opracowanie bogatego i zróżnicowanego instrumentarium umożliwiającego modelowanie zjawisk na podstawie szeregów czasowych, danych przekrojowych i panelowych oraz mikrodanych. Współcześnie rozwijane metody ekonometryczne pozwalają modelować zjawiska z uwzględnieniem ich swoistej specyfiki, np. przestrzennego charakteru – w ramach ekonometrii przestrzennej czy finansowo-czasowego pochodzenia – w ramach ekonometrii finansowej. Oprócz tego należy wspomnieć o modelach zmiennych jakościowych (klasyfikacyjnych), takich jak liniowe modele prawdopodobieństwa, zwane też modelami wyborów dyskretnych, m.in. wykorzystywane do prognozowania bankructwa, oraz o metodach predykcji ekonometrycznej, stosowanych w prognozowaniu.

6.1. Sformułowanie modelu ekonometrycznego

Podczas konstrukcji modelu ekonometrycznego przyjmuje się szereg założeń, które można podzielić na pięć grup[1]:

1 Taką klasyfikację założeń zaproponował Zbigniew Czerwiński [1973, s. 14].

1) założenia definicyjne, wynikające z określeń pojęć, np. założenia o dopuszczalnych zbiorach wartości zmiennych występujących w modelu, o wartościach parametrów strukturalnych itp.;
2) założenia behawiorystyczne, opisujące powiązania wynikające z typowego zachowania się ludzi biorących udział w procesach gospodarczych, np. założenia o sposobie wydatkowania dochodów ludności;
3) założenia techniczne, odnoszące się do techniki wytwarzania, opisywane za pomocą odpowiednich parametrów technicznych;
4) założenia polityczne, dotyczące reakcji organów władzy wobec procesu rozwoju gospodarczego, uznając pewne stany gospodarki za pożądane, inne zaś za niedopuszczalne;
5) założenie instytucjonalne, opisujące powiązania wynikające z istniejącego w gospodarce systemu organizacji produkcji, podziału i wymiany.

Część z tych założeń wprowadzana jest do modelu w postaci równań lub tożsamości. Pozostałe istnieją niejako poza modelem, tzn. konstruktor modelu, dokonując weryfikacji merytorycznej, podejmuje decyzję o tym, czy wszystkie przyjęte założenia (niezapisane w sposób jawny w modelu) zostały spełnione.

Model ekonometryczny zazwyczaj zapisuje się jako:

$$y_{ji} = f_j(\alpha_{j0}, \alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jk}, x_{j1i}, x_{j2i}, \ldots, x_{jki}) + \varepsilon_{ji} \tag{6.1}$$

gdzie dla każdego $j = 1, 2, \ldots, l$ oraz $i = 1, 2, \ldots, n$:
y_{ji} oznacza i-tą obserwację zmiennej y_j, tj. zmiennej objaśnianej przez j-te równanie modelu,
$x_{j1i}, x_{j2i}, \ldots, x_{jki}$ to i-te obserwacje zmiennych objaśniających odpowiednio: x_{j1}, $x_{j2}, \ldots, x_{jk}$, odwzorowane na dowolnej skali pomiarowej,
$\alpha_{j0}, \alpha_{j1}, \alpha_{j2}$ – nieznane parametry strukturalne j-tego równania,
ε_{ji} – i-ta obserwacja składnika losowego j-tego równania (tj. ε_j),
f_j – postać funkcyjna j-tego równania.

Jak zatem widać, model ekonometryczny, podobnie jak to miało miejsce w przypadku modelu regresji (4.18), opisuje analizowane zjawiska (tj. zmienne objaśniane) za pomocą wyróżnionych czynników reprezentowanych przez zmienne objaśniające. Zapis (6.1) dotyczy j-tego równania ($j = 1, 2, \ldots, l$) modelu o l równaniach, z których (w ogólnym przypadku) każde może mieć inną postać funkcyjną. Przyjmując najczęściej spotykaną w analizach społeczno-gospodarczych funkcję liniową, otrzymuje się model:

$$y_{ji} = \alpha_{j0} + \alpha_{j1} \cdot x_{j1i} + \alpha_{j2} \cdot x_{j2i} + \ldots + \alpha_{jk} \cdot x_{jki} + \varepsilon_{ji} \tag{6.2}$$

którego oszacowanie ma postać:

$$\hat{y}_{ji} = \hat{\alpha}_{j0} + \hat{\alpha}_{j1} \cdot x_{j1i} + \hat{\alpha}_{j2} \cdot x_{j2i} + \ldots + \hat{\alpha}_{jk} \cdot x_{jki} \tag{6.3}$$

gdzie dla każdego $j = 1, 2, \ldots, l$ oraz $i = 1, 2, \ldots, n$:
$\hat{y}_{ji}$ – wartości teoretyczne j-tej zmiennej objaśnianej,
$\hat{\alpha}_{j0}, \hat{\alpha}_{j1}, \hat{\alpha}_{j2}, \ldots, \hat{\alpha}_{jk}$ – oszacowane parametry strukturalne modelu (oceny estymatorów parametrów) w j-tym równaniu modelu,
a pozostałe oznaczenia są jak poprzednio.

Model opisany jednym ze wzorów (6.1)–(6.3) dotyczy l zmiennych objaśnianych, opisanych konkretnymi równaniami zawierającymi odpowiednie zmienne objaśniające. Wszystkie zmienne, zarówno te przez model objaśniane, jak i te, które pełnią rolę zmiennych objaśniających, są obserwowane w n punktach pomiarowych. Oczywiście jeśli przyjąć, że model zawiera jedno równanie, to model ekonometryczny (6.1)–(6.3) przyjmuje postać odpowiednio (4.29)–(4.31).

Podział zmiennych na zmienne objaśniane i objaśniające wynika wprost z konstrukcji modelu – zmienna objaśniana jest zmienną zależną, opisywaną (objaśnianą) danym równaniem, podczas gdy zmienne objaśniające to zmienne niezależne, które opisują (objaśniają) kształtowanie się zmiennej objaśnianej. Jednakże w przypadku modeli zawierających więcej niż jedno równanie, taki podział zmiennych nie jest dostatecznie precyzyjny, gdyż zmienna objaśniana w jednym równaniu może pełnić rolę zmiennej objaśniającej w innym z równań. Stosuje się zatem, odpowiednią do takiej sytuacji, klasyfikację zmiennych, wyróżniając zmienne endogeniczne, które są objaśniane przez poszczególne równania modelu, i zmienne egzogeniczne, które pełnią w modelu rolę zmiennych objaśniających, a same nie są objaśniane przez żadne z równań.

Integralnym elementem modelownia ekonometrycznego są założenia o stochastycznej strukturze modeli (por. [Welfe 1995, s. s. 27–30, 62–63])[2].

1) Model jest niezmienniczy ze względu na obserwacje. Założenie to mówi o stabilności relacji występujących między badanymi zjawiskami. Uchylenie tego założenia prowadzi do modeli o zmiennych (w czasie) parametrach, modeli przełącznikowych, modeli o skomplikowanych konstrukcjach formalnych.
2) Model jest liniowy względem parametrów, czyli dany w postaci (4.30) lub (6.2) dla $l = 1$.
3) Obserwacje zmiennych objaśniających są nielosowe. Uchylenie tego założenia nie pociąga żadnych konsekwencji pod warunkiem, że zmienne objaśniające i składniki losowe są niezależne lub nieskorelowane.
4) Rząd macierzy obserwacji zmiennych objaśniających równy jest liczbie szacowanych parametrów, czyli jeżeli w modelu występuje wyraz wolny, to rząd wynosi $(k + 1)$, a jeśli model jest bez wyrazu wolnego, to rząd wynosi k. Z niniejszego założenia wynika, że macierz obserwacji ma pełen

2 Klasyczny zestaw założeń nosi nazwę założeń Gaussa-Markowa, o których wspomniano w podrozdziale 4.5.

rząd kolumnowy, tzn. nie występuje współliniowość między zmiennymi egzogenicznymi, a więc:
- liczba szacowanych parametrów jest mniejsza od liczby obserwacji,
- żadnej z kolumn macierzy obserwacji nie można przedstawić jako liniowej kombinacji kilku innych kolumn.

Jeżeli warunek ten nie jest spełniony, to nie istnieje możliwość jednoznacznego oszacowania parametrów modelu ekonometrycznego (podobnie jak równania regresji) za pomocą metody najmniejszych kwadratów MNK.

5) Składniki losowe ε_i dla $i = 1, 2, \ldots, n$ są niezależnymi zmiennymi losowymi o rozkładach normalnych. Spełnienie tego założenia ułatwia wnioskowanie statystyczne, ponieważ odpowiednie statystyki mają wówczas pożądane rozkłady.
6) Wartość oczekiwana składnika losowego jest równa 0, czyli:

$$E(\varepsilon_i) = 0. \tag{6.4}$$

Założenie to oznacza, że składnik losowy nie wykazuje tendencji do jednostronnego odchylania się *in plus* lub *in minus*. Uchylenie tego założenia powoduje, że MNK estymatory przestają być nieobciążone.

7) Składnik losowy jest homoskedastyczny – posiada stałą wariancję, co zapisuje się jako:

$$E(\varepsilon_i, \varepsilon_j) = \sigma^2 \text{ dla } i = j \tag{6.5}$$

i nie występuje autokorelacja, co oznacza spełnienie warunku:

$$E(\varepsilon_i, \varepsilon_j) = 0 \text{ dla } i \neq j. \tag{6.6}$$

Zatem macierz wariancji – kowariancji składnika losowego jest macierzą diagonalną o takich samych elementach na głównej przekątnej. Uchylenie tego założenia powoduje, że estymatory MNK tracą efektywność. Modele, dla których nie zostało spełnione założenie o jednorodności wariancji składnika losowego (6.5), nazywa się modelami heteroskedastycznymi (heteroscedastycznymi), a modele bez spełnionego założenia o zerowych kowariancjach składnika losowego (6.6) – modelami z autokorelacją. Estymacja tego typu modeli klasyczną MNK powoduje zazwyczaj zaniżenie standardowych błędów szacunku, co prowadzić może do błędnych decyzji w procesie statystycznej weryfikacji modelu.

8) Informacje zawarte w próbie są jedynymi, na podstawie których szacowane są parametry strukturalne modelu. W przeciwnym przypadku występuje tzw. estymacja przy warunkach pobocznych.

Warto zauważyć, że spełnienie założeń (2) i (4) umożliwia wyznaczenie ocen estymatora MNK. Jednakże jego własności zależą od tego, które z założeń są spełnione. Nieobciążoność, opisana wzorem (3.2), wymaga spełnienia założeń (1), (3) i (6), a zapewnienie największej efektywności – dodatkowo spełnienia założe-

nia (7). Jeżeli estymator jest linowy, nieobciążony i najbardziej efektywny (tj. ma najmniejszą wariancję), to jest to najlepszy z liniowych, nieobciążony estymator (*best linear unbiased estimator* – BLUE).

Założenia dotyczące stochastycznej struktury modelu zapewniają wprawdzie uzyskanie w procesie estymacji MNK – wzbudzających zaufanie – oszacowań parametrów, są jednak często w praktyce zbyt mocne i modele estymuje się mimo braku spełnienia jednego lub więcej założeń. W takiej sytuacji estymatory MNK tracą pożądane własności. Zaleca się wtedy uwzględnienie tego faktu w prowadzonych analizach lub stosowanie innych metod estymacji[3].

6.2. Klasyfikacja modeli ekonometrycznych

Ze względu na różnorodność modeli ekonometrycznych klasyfikuje się je wg różnych kryteriów, z których najczęściej stosowane to:

- liczba równań w modelu,
- postać analityczna równań modelu,
- charakter poznawczy modelu,
- zakres badania,
- specyfika danych.

Biorąc pod uwagę pierwsze z wymienionych kryteriów, modele ekonometryczne dzieli się na modele jednorównaniowe oraz modele wielorównaniowe. Modele jednorównaniowe buduje się na potrzeby analizy ilościowej pojedynczego procesu lub zjawiska. Przykładem tego typu modeli mogą być modele inflacji, modele kosztów w przedsiębiorstwie, modele stóp zwrotu z instrumentu finansowego itp. W sytuacji, gdy celem prowadzonej analizy ekonometrycznej jest badanie zależności występujących między wieloma zjawiskami lub procesami gospodarczymi, stosuje się modele wielorównaniowe. Przykładem takich modeli są modele rynków finansowych, gospodarki narodowej, przedsiębiorstw etc.[4]

Postać analityczna równań pozwala podzielić modele ekonometryczne na liniowe i nieliniowe[5]. Model liniowy to taki, w którym wszystkie równania są liniowe. W sytuacji, gdy przynajmniej jedno z równań modelu jest nieliniowe, model jest nieliniowy. Nieliniowość w modelach może występować względem zmiennych lub względem parametrów modelu. Wśród modeli nieliniowych względem

3 W pracy [Tarczyński, Witkowska, Kompa 2013, s. 57–77] omówiono różne metody estymacji.

4 Zagadnienia ekonometrycznego modelowania gospodarki narodowej przedstawiono w pracy [Welfe 2010], a modelowania w mikroskali w pracy [Dittmann 2017].

5 Omówienie modeli nieliniowych znaleźć można w pracy [Milo 1990b].

parametrów wyróżnia się całą klasę modeli, które za pomocą prostych operacji matematycznych mogą zostać sprowadzone do modeli liniowych. W przypadku nieliniowości względem zmiennych sprowadzenie modelu do postaci liniowej może polegać na odpowiednim przedefiniowaniu zmiennych.

Biorąc pod uwagę charakter poznawczy modelu, istotny jest sposób opisu badanych zależności ekonomicznych, co prowadzi do wyróżnienia trzech podstawowych typów modeli:

- modele tendencji rozwojowej, czyli trendu, uwzględniające zależność analizowanych zjawisk od czasu;
- modele przyczynowo-skutkowe, w których pewne zjawiska (skutki) wyjaśniane są za pomocą innych zjawisk, które traktuje się jako przyczyny tych wcześniej wspomnianych, np. wyniki ekonomiczno-finansowe przedsiębiorstw mają wpływ na wysokość wypłacanych wynagrodzeń, a stopa procentowa, będąca jednym z podstawowych instrumentów polityki gospodarczej, ma istotny wpływ na wielkość inwestycji[6] itp.;
- modele symptomatyczne – budowane są, gdy między zmiennymi nie daje się określić związków przyczynowo-skutkowych, ale wiadomo jest, że zachodzi między nimi zależność korelacyjna. Innymi słowy modele symptomatyczne to modele, w których rolę zmiennych objaśniających pełnią zmienne skorelowane z odpowiednimi zmiennymi objaśnianymi, ale nie wyrażają one źródeł zmienności zmiennych objaśnianych [Gruszczyński 1996, s. 239].

Kryterium zakresu badania pozwala podzielić modele ekonometryczne na modele mikro-, mezo- i makroekonomiczne. Przykładem pierwszych są modele przedsiębiorstw. Przykładem modeli w skali mezo są m.in. modele sektorów gospodarki lub segmentów rynku finansowego. Natomiast modele makroekonomiczne odnoszą się do zjawisk obserwowanych w skali całej gospodarki, w tym krajowego rynku finansowego, dotyczą relacji występujących na rynkach międzynarodowych lub w gospodarce światowej.

Oprócz klasycznych kryteriów podziału, należy wymienić modele uwzględniające specyfikę danych opisujących pewne zjawiska. Należą do nich z pewnością modele szeregów czasowych, modele panelowe, a także modele konstruowane w ramach ekonometrii finansowej, ekonometrii przestrzennej i mikroekonometrii.

Pierwsze z wymienionych dotyczą modelowania zjawisk jedynie na podstawie szeregów czasowych je opisujących, to znaczy bez uwzględniania wpływu dodatkowych czynników. Modele szeregów czasowych[7], zwane też modelami ekono-

---

6 Podniesienie stopy procentowej ma na celu tłumienie inwestycji, a jej obniżanie ma zachęcać do zwiększenia nakładów inwestycyjnych.

7 Fundamentalne opracowania z zakresu analizy szeregów czasowych i ekonometrii dynamicznej w języku polskim to [Box, Jenkins 1983; Charemza, Deadman 1997]. Ośrodkiem akademickim, który od kilku dekad zajmuje się w Polsce modelowaniem szeregów czasowych, jest Uniwersytet Mikołaja Kopernika w Toruniu, a prekursorem tych badań był prof. Zygmunt Zieliński. Porównaj na przykład prace: [Piłatowska 2002; Kufel 2002; Osińska 2007].

metrii dynamicznej, szacowane są na podstawie dużych liczebnie prób (danych o wysokiej zazwyczaj częstotliwości pomiaru) i szczególną uwagę przykłada się do wewnętrznych własności tych szeregów, np. ich stacjonarności. Wymagają one również odpowiednich metod estymacji.

W przypadku danych panelowych próba zawiera obserwacje dotyczące zjawisk lub procesów obserwowanych zazwyczaj w różnym miejscu i czasie. Modele ekonometryczne budowane są przy założeniu występowania różnic między poszczególnymi jednostkami badania. Innymi słowy, w modelach ekonometrycznych, szacowanych na podstawie danych panelowych, z reguły przyjmuje się, że na kształtowanie się zmiennej objaśnianej wpływają, oprócz zmiennych objaśniających, czynniki niemierzalne, stałe w czasie i specyficzne dla danego obiektu, zwane efektami grupowymi i/lub czynniki stałe względem obiektów specyficzne dla danego okresu, zwane efektami czasowymi. Specyfika danych panelowych doprowadziła do opracowania specjalnych, dla nich przeznaczonych, metod analizy i modelowania[8].

Metody wchodzące w zakres ekonometrii finansowej dotyczą szeroko prowadzonych analiz zjawisk zachodzących na rynkach finansowych oraz ich segmentach – zarówno w ujęciu charakterystycznym dla modelowania szeregów czasowych [Osińska 2006], jak i z zastosowaniem wielowymiarowej analizy statystycznej, uwzględniającej czynniki wpływające na decyzje inwestycyjne i ich efektywność[9] [Łuniewska 2012]. W tym pierwszym ujęciu modele ekonometrii finansowej szacowane są na podstawie danych w postaci finansowych szeregów czasowych i zazwyczaj służą weryfikacji hipotez z zakresu teorii finansów lub identyfikacji prawidłowości w danych finansowych[10]. Modele te zostały skonstruowane poprzez adaptację na potrzeby finansowych szeregów czasowych modeli ekonometrii dynamicznej lub są specjalnie tworzone na potrzeby finansów. W tym drugim ujęciu stosuje się metody analiz wielowymiarowych, np. do wyboru instrumentów do portfela.

Ekonometria przestrzenna (por. [Zeliaś 1991; Kopczewska 2007; Suchecki 2010])[11] zajmuje się problemami zależności przestrzennej oraz zróżnicowania przestrzennego, operując m.in. na danych geo-lokalizowanych. Przestrzenne modele ekonometryczne wyjaśniają mechanizm kształtowania się i rozwoju zjawisk ekonomicznych uwarunkowanych terytorialnie (przestrzennie). Modele te wykorzystywane są m.in. w badaniach regionalnych, rynku nieruchomości, zasobów

---

Nie oznacza to jednak, że badania w tym zakresie nie były i nie są prowadzone również w innych ośrodkach naukowych (por. np. [Milo 1990a; Garnczarek-Gamrot 2014]).

8 Omówienie tych zagadnień można znaleźć m.in. w pracy [Dańska-Borsiak 2011].

9 Metody te omawiane są w rozdziale 7 i 9.

10 Jednym z pierwszych opracowań dotyczących polskiego rynku finansowego jest praca [Brzeszczyński, Kelm 2002].

11 Interesująca w kontekście analiz rynków finansowych jest również praca [Szulc, Wleklińska 2021].

naturalnych, lokalizacji przedsiębiorstw itp. Warto zauważyć, że dane przestrzenne i przestrzenno-czasowe mają bardziej skomplikowaną strukturę, zatem nie mogą być one analizowane za pomocą klasycznych metod ilościowych.

Z kolei mikroekonometria obejmuje modele konstruowane na podstawie danych indywidualnych, tzw. mikrodanych [Gruszczyński 2010], aczkolwiek termin ten rozumiany jest również jako zastosowanie metod statystycznych i modeli ekonometrycznych w analizie przedsiębiorstw [Wiśniewski 2009; 2016]. Pewnym kompromisem może być mikroekonometria finansowa rozumiana jako metody mikroekonometrii stosowane do danych o przedsiębiorstwach, instytucjach finansowych, transakcjach, klientach itd. [Gruszczyński 2012, 2020]. Do tych ostatnich zalicza się m.in. liniowe modele prawdopodobieństwa, w tym modele regresji binarnej, logitowe i probitowe[12].

6.3. Budowa modeli ekonometrycznych

Analiza ekonometryczna obejmuje kilka etapów, takich jak:

1) ustalenie celu prowadzonej analizy;
2) specyfikacja modelu ekonometrycznego;
3) zebranie informacji statystycznych;
4) estymacja parametrów strukturalnych modelu;
5) weryfikacja statystyczna i merytoryczna modelu;
6) wykorzystanie praktyczne.

Celem prowadzonej analizy może być:

a) weryfikacja hipotez dotyczących zależności występujących między analizowanymi zjawiskami i opis tychże zależności,
b) budowa prognoz dotyczących przyszłego kształtowania się relacji w gospodarce,
c) prowadzenie symulacji na modelu w celu pogłębienia analizy zależności dynamicznych występujących w rozpatrywanych procesach.

Specyfikacja modelu ekonometrycznego jest ściśle związana z celem, jakiemu dany model ma służyć. Można wyróżnić dwa zasadnicze cele, tj. poznawczy i instrumentalny. Pierwszy obejmuje wyjaśnienie, w jakiej mierze zmienne objaśniające wpływają na kształtowanie się zmiennej objaśnianej, charakteryzującej wyróżnione zjawisko. Drugi cel oznacza praktyczne wykorzystanie modelu, np. do prognozowania.

12 Porównaj także pracę [Bąk 2013], dotyczącą badania preferencji konsumentów na podstawie mikrodanych.

Specyfikacja modelu ekonometrycznego obejmuje ustalenie:
- listy podstawowych zmiennych oraz stopnia ich dezagregacji,
- zależności występujących między zmiennymi,
- postaci analitycznej funkcji wykorzystywanej w modelu.

W trakcie budowy modelu ekonometrycznego, najważniejszą i najtrudniejszą rzeczą jest dobór odpowiednich zmiennych, które przyczyniają się do wyjaśnienia badanego zjawiska. Najczęściej stosuje się trzy podejścia:
- zmienne wybiera się na podstawie wiedzy wynikającej z teorii ekonomii,
- tam, gdzie teoria ekonomiczna nie jest dostatecznie rozwinięta, przy doborze zmiennych wykorzystuje się doświadczenie i intuicję,
- gdy teoria ekonomiczna i analiza empiryczna wraz z doświadczeniem nie są wystarczające, wykorzystuje się odpowiednie metody statystyczne, bazujące na analizie związków korelacyjnych między zmiennymi[13].

Należy przy tym dążyć do uzyskania zestawu zmiennych, które w sposób możliwie pełny charakteryzują badane zjawisko, a przy tym tworzą zespół jak najmniej liczny. Wymagania te są spełnione wtedy, gdy zmienne objaśniające posiadają następujące własności:
- są silnie skorelowane ze zmienną, którą objaśniają,
- są nieskorelowane lub co najwyżej słabo skorelowane między sobą,
- charakteryzują się wysoką zmiennością (zazwyczaj przyjmuje się, że współczynnik zmienności (2.19) jest większy od 0,1),
- nie ulegają wpływom zewnętrznym.

W trakcie gromadzenia danych należy prowadzić bieżącą analizę poprawności i przydatności zebranego materiału empirycznego. Warto zauważyć, że w trakcie realizacji tego etapu badań może okazać się, że nie można pozyskać odpowiednich danych, co w konsekwencji spowoduje konieczność zmiany pierwotnej specyfikacji modelu. Można podać kilka przykładowych sytuacji:
- potrzebne informacje nie są publikowane, a koszty ich pozyskania są wysokie,
- dane dotyczą sektorów gospodarki i nie ma możliwości ich dezagregacji na branże lub przedsiębiorstwa,
- dostępne dane dotyczą podziału na sektory lub jednostki terytorialne inne niż te zdefiniowane w badaniu,
- dane czasowe są o mniejszej niż założona w badaniu częstotliwości[14].

Powinno się dążyć do zebrania możliwie bogatego i dokładnego materiału statystycznego, gdyż im większa liczba obserwacji tym większa dokładność i lepsze

13 Opis metod wyboru zmiennych objaśniających znaleźć można m.in. w [Goryl i in. 2000, s. 16–25]. Zagadnienia te są również omawiane w podrozdziale 7.5.

14 Dane mogą mieć różną częstotliwość pomiaru, np. część zmiennych dostępnych jest w postaci obserwacji dziennych (np. notowania kursów walut), inne mierzone są raz w tygodniu (np. notowania bonów skarbowych), jeszcze inne dostępne są jako dane miesięczne (np. wskaźnik inflacji), kwartalne lub roczne.

oszacowanie modelu. Jednakże nadmierne wydłużanie szeregów czasowych może spowodować utratę porównywalności danych. Zazwyczaj określenie długości szeregu czasowego podyktowane jest koniecznością kompromisu pomiędzy wymaganą liczebnością próby estymacyjnej i wykorzystaniem możliwie aktualnych danych, zwłaszcza jeśli konstruowany model ma zostać wykorzystywany do szacowania prognoz zjawisk o dużej zmienności w czasie.

Oszacowanie modelu polega na znalezieniu wartości ocen estymatorów nieznanych parametrów strukturalnych modelu. Dokonuje się tego na podstawie obserwacji statystycznych dotyczących kształtowania się analizowanych zjawisk gospodarczych, w tym finansowych. Do estymacji wykorzystać można dane przekrojowe, czasowe lub panelowe, a szacowanie parametrów modelu przeprowadza się z wykorzystaniem adekwatnych do charakteru zależności i rodzaju danych metod estymacji[15], z których najbardziej popularną jest MNK. Istotnym pojęciem, które pojawia się na tym etapie, jest identyfikowalność modelu[16], która wiąże się ze stwierdzeniem występowania (lub nie) w modelu pewnych typów zależności między zmiennymi. Dotyczy to takich związków, które uniemożliwiają (lub znacznie utrudniają) estymację parametrów modelu.

Kolejnym etapem jest weryfikacja modelu:

- merytoryczna, polegająca na sprawdzeniu poprawności i sensowności otrzymanych ocen estymatorów parametrów z punktu widzenia teorii ekonomii (w tym finansów), doświadczenia wynikającego z badań własnych i innych badaczy oraz zdrowego rozsądku,
- statystyczna, polegająca m.in. na sprawdzeniu, czy:
  - wpływ zmiennych objaśniających na zmienną objaśnianą jest statystycznie istotny,
  - spełnione zostały różnego rodzaju założenia (np. założenia Gaussa-Markowa, dotyczące stacjonarności szeregu, poprawności postaci analitycznej modelu, stabilności jego parametrów w czasie, występowania efektów grupowych etc.),
  - stopień dopasowania modelu do danych empirycznych jest akceptowalny poprzez badanie różnie zdefiniowanych współczynników determinacji oraz tzw. kryteriów informacyjnych[17].

15 Metody estymacji są ściśle związane z posiadanymi (i zakładanymi) informacjami o stochastycznej strukturze modelu. Można je najogólniej podzielić na trzy grupy: (1) klasa metod najmniejszych kwadratów (MNK), (2) metody zmiennych instrumentalnych (MZI) i (3) metody największej wiarygodności (MNW). W przypadku modeli wielorównaniowych stosuje się zarówno metody szacowania parametrów pojedynczych równań (np. pośrednia i podwójna MNK, MZI oraz MNW z ograniczoną informacją), jak i metody łącznej estymacji parametrów modelu (np. potrójna MNK i MNW z pełną informacją).

16 Szerzej zagadnienia identyfikowalności omówiono m.in. w pracach [Chow 1995, s. 150–161; Charemza, Deadman 1997, s. 145–151; Maddala 2006, s. 396–421].

17 Kryteria informacyjne omówiono w pracy [Charemza, Deadman 1997, s. 236–240], porównaj również prace [Maddala 2006, s. 550, 591 i 615; Osińska 2006, s. 54].

W wyniku weryfikacji modelu, podejmowane są decyzje, czy zostanie on wykorzystany w badaniach lub praktyce albo czy należy go poddać modyfikacjom, np. zmienić specyfikację (respecyfikacja) modelu lub zastosować inne metody estymacji (reestymacja) parametrów strukturalnych. Warto w tym miejscu podkreślić, że proces budowy modelu ekonometrycznego jest z reguły długotrwały i polega na wielokrotnym powtarzaniu ciągu etapów (2)–(5), zanim model będzie spełniać przynajmniej podstawowe oczekiwania jego konstruktora.

Przykład 6.1

Wykorzystując dane z przykładu 4.9, charakteryzujące badane towarzystwa ubezpieczeniowe, oszacować należy model opisujący wynik finansowy netto [tys. PLN] tych towarzystw (w przykładzie 4.9 oznaczony przez (jako) X1). Estymację parametrów należy przeprowadzić na podstawie danych zawartych w tab. 4.16.

Rozwiązanie

W treści przykładu sformułowano cel analizy ekonometrycznej i określono zmienną objaśniana. Otwarta pozostaje kwestia specyfikacji modelu, czyli wyboru zmiennych, które wytłumaczą kształtowanie się zmiennej objaśnianej[18]. Opierając się na posiadanej wiedzy, przyjęto arbitralnie, że odpowiednimi zmiennymi objaśniającymi do opisu wyniku finansowego będą dwie zmienne, tj. wartości składki zarobionej na udziale własnym [tys. PLN] (X_6) oraz wartości świadczeń i odszkodowań wypłaconych brutto [tys. PLN] (X_8). Szacowana będzie zatem funkcja regresji postaci:

$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + \varepsilon_i$$

co zostanie przeprowadzone z wykorzystaniem funkcji REGLINP w Excelu. W wyniku przeprowadzonych obliczeń otrzymano tab. wynikową 6.1, w której zacieniono elementy dalej nieistotne. Tak więc w modelu jest 6 stopni swobody (bo było 9 obserwacji, tj. 9 towarzystw ubezpieczeniowych, i 3 szacowne parametry: α_0, α_1, α_2), a suma kwadratów reszt empirycznych wynosi 16 568 476 560. Oszacowany model jest postaci:

$$\hat{y}_i = -35\,575{,}6205 + 0{,}3312 \cdot x_{1i} - 0{,}1933 \cdot x_{2i} \qquad R^2 = 0{,}9879$$

gdzie dla każdego *i*:
$\hat{y}_i$ – wynik finansowy netto [tys. PLN] (wartości teoretyczne),
x_{1i} – świadczenia i odszkodowania wypłacone brutto [tys. PLN] (obserwacje),
x_{2i} – składka zarobiona na udziale własnym [tys. PLN] (obserwacje).

18 Por. przypis 13.

Tabela 6.1. Wyniki wykonania funkcji REGLINP

Opis zawartości tablicy	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_0$
Oceny estymatorów parametrów	–0,1933	0,3312	–35 575,6205
Standardowe błędy szacunku	0,3505	0,2421	26 421,2959
Współczynnik determinacji	0,9879	52 549	#N/D
Liczba stopni swobody	245,1206	6	#N/D
Suma kwadratów reszt empirycznych	1 353 758 136 474	16 568 476 560	#N/D

Uwaga: 3 ostatnie kolumny tabeli są kopią tablicy wynikowej funkcji REGLINP, a komórki zacienione zawierają informacje nieistotne w tej analizie.

Źródło: obliczenia własne.

Tak oszacowany model podlega weryfikacji statystycznej. Dlatego na podstawie standardowych błędów szacunku:

$$S(\alpha_0) = 26\,421{,}2959;\ S(\alpha_1) = 0{,}2421 \text{ i } S(\alpha_2) = 0{,}3505,$$

wyznaczono wartości sprawdzianu testu (4.39):

$$t = \frac{\hat{\alpha}_0}{S(\alpha_0)} = -1{,}3465;\ t = \frac{\hat{\alpha}_1}{S(\alpha_1)} = 1{,}367765 \text{ i } t = \frac{\hat{\alpha}_2}{S(\alpha_2)} = -0{,}5514,$$

które porównano z wartością krytyczną $t_\alpha = 1{,}9432$, wyznaczoną dla jednostronnego obszaru odrzucenia, współczynnika istotności $\alpha = 0{,}05$ i 6 stopni swobody. Jak widać, nie ma podstaw do odrzucenia hipotezy zerowej. Zatem żaden z parametrów nie jest istotnie różny od 0, co wskazuje, że wyspecyfikowane zmienne objaśniające nie mają istotnego wpływu na zmienną objaśnianą. Mogłoby to oznaczać, że zmienne niezależne zostały źle dobrane, czemu przeczy bardzo wysoki stopień dopasowania modelu do danych empirycznych (współczynnik determinacji $R^2 = 0{,}9879$). Należy zatem przypuszczać, że zachodzi nadmierna korelacja zmiennych objaśniających, zwłaszcza że obliczony dla nich współczynnik korelacji liniowej Pearsona wynosi 0,999 (tab. 4.17, korelacja między X_6 i X_8).

W sytuacji kiedy stwierdzono brak istotnego wpływu wyróżnionych zmiennych niezależnych, zmienne te usuwa się z modelu, o ile ich usunięcie znacząco nie obniży stopnia dopasowania modelu. W omawianym przykładzie na tym etapie trzeba oszacować modele z pojedynczą zmienną objaśniającą. W przypadku modelu z usuniętą zmienną x_2:

$$\hat{y}_i = -59308{,}8539 + 0{,}2858 \cdot x_{1i} \qquad R^2 = 0{,}9841$$

zmienna objaśniająca, tj. świadczenia i odszkodowania wypłacone brutto, jest statystycznie istotną, gdyż statystyka testowa dla zmiennej x_1 wynosi $t = 20{,}84$, a wartość krytyczna przy założeniu jednostronnego obszaru odrzucenia dla $\alpha = 0{,}05$ i 7 stopni swobody wynosi $t_\alpha = 1{,}8946$. Również wyraz wolny jest istotnie mniejszy od 0, ponieważ $t = -2{,}8072$. Zauważyć trzeba, że wartość współczynnika determinacji nadal pozostaje wysoka.

Model, w którym jedyną zmienną objaśniającą jest składka zarobiona na udziale własnym, jest po oszacowaniu postaci:

$$\hat{y}_i = -45330{,}3414 + 0{,}1978 \cdot x_{2i} \qquad R^2 = 0{,}9873$$

i charakteryzuje się nawet wyższym (choć nieznacznie) stopniem dopasowania do danych empirycznych. W tym modelu oba parametry są również istotne, a wartości sprawdzianu testu wynoszą $t = 23{,}32$ i $t = -2{,}43$ odpowiednio dla parametru stojącego przy zmiennej x_2 i wyrazie wolnym.

Przedstawione postępowanie pokazuje, że objaśnienie wyniku finansowego netto [tys. PLN] powinno zostać przeprowadzone z użyciem jednej z wybranych zmiennych, ale nie obu jednocześnie z powodu ich silnego wzajemnego skorelowania.

W celu uzyskania lepszego dopasowania modelu do danych empirycznych można oszacować inny model z dwiema zmiennymi objaśniającymi, tym razem nieskorelowanymi ze sobą. Będą to, arbitralnie wybrane, zmienne X_8 i X_5, opisujące:

x_{1i} – świadczenia i odszkodowania wypłacone brutto [tys. PLN],

x_{3i} – kapitały podstawowe [tys. PLN].

Oszacowany model jest postaci:

$$\hat{y}_i = -8528{,}4012 - 0{,}6480 \cdot x_{1i} + 0{,}2884 \cdot x_{3i} \qquad R^2 = 0{,}9882$$

Jak widać współczynnik determinacji tego modelu jest najwyższy w stosunku do dotychczas oszacowanych modeli, ale jedyną zmienną istotnie wpływającą na wynik finansowy netto jest zmienna x_1, gdyż $t = 22{,}38$. Z kolei dla zmiennej x_3 sprawdzian testu wynosi $t = -1{,}45$, a dla wyrazu wolnego $t = -0{,}21$, co wobec wyznaczonej (dla przyjętego poziomu istotności i 6 stopni swobody) wartości krytycznej $t_\alpha = 1{,}9432$ nie daje podstaw do odrzucenia hipotezy zerowej. Z przedstawionej analizy wynika, że o ile zadowalający jest stopień dopasowania modelu do danych empirycznych na poziomie 98%, to zmienną x_3 należy z tego modelu usunąć. Model powinien zawierać tylko jedną zmienną objaśniającą. Jednakże można uznać za zasadne wzbogacenie tego modelu o dodatkową zmienną, co pozwoli na osiągnięcie wartości współczynnika determinacji na poziomie 0,99.

6.4. Rodzaje zmiennych w modelach ekonometrycznych

Zmienne występujące w modelu ekonometrycznym podlegają różnym klasyfikacjom. Kryteria podziału zmiennych zależą od funkcji, jakie poszczególne zmienne pełnią w modelu, oraz rodzaju informacji, jakie niosą. Wspomnianym już wcześniej kryterium podziału zmiennych jest ich pozycja w modelu, tzn. stwierdzenie, czy znajdują się po lewej czy prawej stronie równania. Na tej podstawie wyróżniono zmienne objaśniane (zależne) i objaśniające (niezależne) w modelach jednorównaniowych oraz zmienne endogeniczne i egzogeniczne w modelach wielorównaniowych.

Modele ekonometryczne, które szacowane są na podstawie szeregów czasowych, zazwyczaj zapisuje się, przy założeniu liniowej postaci funkcyjnej, jako:

$$y_{jt} = \alpha_{j0} + \alpha_{j1} \cdot x_{j1t} + \alpha_{j2} \cdot x_{j2t} + ... + \alpha_{jk} \cdot x_{jkt} + \varepsilon_{jt} \tag{6.7}$$

gdzie:
t ($t = 1, 2, \ldots, T$) – oznacza kolejny numer obserwacji,
pozostałe oznaczenia jak poprzednio.

Modele opisujące zmiany badanych zjawisk w czasie (6.7) często zawierają zmienne opóźnione: $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ i $x_{t-1}, x_{t-2}, \ldots, x_{t-r}$. Zmienne te odzwierciedlają poziom badanego zjawiska w okresach wcześniejszych odpowiednio: o 1, 2, …, p lub r obserwacji[19]. Przykładowo, zapis modelu (6.4) zawierającego opóźnione zmienne objaśniane jest postaci:

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \alpha_2 \cdot x_{2t} + \ldots + \alpha_k \cdot x_{kt} + \alpha_{k+1} \cdot y_{t-1} + \ldots + \alpha_{k+r} \cdot y_{t-r} + \\ \alpha_{k+r+1} \cdot x_{t-1} + \ldots + \alpha_{k+r+p} \cdot y_{t-p} + \varepsilon_t \end{aligned} \tag{6.8}$$

W przypadku modeli szacowanych na podstawie szeregów czasowych specyficzną zmienną jest tzw. zmienna czasowa (oznaczana najczęściej symbolem t), która wyraża systematyczne zmiany w kształtowaniu się wielkości zmiennej objaśnianej w czasie[20]. Wówczas model (6.7) ze zmienną czasową jest postaci:

$$y_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \alpha_2 \cdot x_{2t} + \ldots + \alpha_k \cdot x_{kt} + \alpha_{k+1} \cdot t + \varepsilon_t \tag{6.9}$$

19 W modelach ekonometrycznych mogą pojawiać się także zmienne z wyprzedzeniami czasowymi, oznaczane jako: $y_{t+1}, y_{t+2}, \ldots, y_{t+p}$ i $x_{t+1}, x_{t+2}, \ldots, x_{t+r}$ itp. Opisują one poziom zjawiska rejestrowany w punktach pomiarowych następujących po t-tej obserwacji odpowiednio o jeden, dwa i p (lub r) okresów. Warto zauważyć, że chociaż – z formalnego punktu widzenia – zapisy równań z opóźnieniami i wyprzedzeniami czasowymi są równoważne, np. równanie: $y_t = f(y_{t-1}, x_t)$ jest równoważne zapisowi: $y_{t+1} = f(y_t, x_{t+1})$, to w modelach ekonometrycznych zmienne z wyprzedzeniami czasowymi wykorzystywane są rzadko.

20 Jest to ta sama zmienna, która występuje w omawianych wcześniej modelach trendu (5.16)–(5.17).

Zmienna ta przyjmuje zazwyczaj wartości liczb naturalnych, tj. $t = 1, 2, \ldots, T$, chociaż może przyjmować dowolne wartości[21] byle odległość – „krok" między obserwacjami – była taka sama. Przykładowo notowania w kolejne dni sesji giełdowej numeruje się po kolei z pominięciem sobót i niedziel, tj. dni, kiedy nie ma notowań. Natomiast określając dzienne oprocentowanie kredytu uwzględnia się również obserwacje w dni wolne od pracy. Może też się zdarzyć, że interwał między kolejnymi obserwacjami będzie większy niż jeden, np. podczas analizy danych miesięcznych w następujących po sobie latach uwzględnione zostaną tylko dane z parzystych miesięcy (wtedy zmienna czasowa przyjmować będzie wartości parzyste (np. $t = 2, 4, \ldots$) lub np. dane tylko ze styczniów kolejnych lat, wówczas zmienna czasowa zmieniać się będzie co 12 miesięcy (np. t = 1, 13, 25, …). Te ostatnie tworzą szeregi czasowe jednoimiennych okresów[22].

Modele, które zawierają zmienne z opóźnieniami czasowymi lub dowolną funkcję trendu, nazywa się dynamicznymi. Modele te uwzględniają bowiem dynamiczny charakter analizowanych zjawisk w przeciwieństwie do modeli statycznych, których przykładem mogą być modele (4.30), (6.2) i (6.7).

Czasami przy wyjaśnianiu procesów gospodarczych istotne są cechy jakościowe (niemierzalne), które w modelu mogą być reprezentowane przez tzw. zmienne zero-jedynkowe (binarne). Zmienne te są definiowane w taki sposób, że przyjmują wartość 1 dla tych obserwacji, w których wystąpiło pewne zdarzenie lub pojawił się określony wariant cechy (np. wprowadzenie w życie nowego podatku), natomiast dla obserwacji pozostałych (tzn. kiedy zdarzenie nie ma miejsca) są równe 0. W praktyce modelowania zmienne binarne również wykorzystuje się w celu uwzględnienia sezonowości występującej w zjawiskach ekonomicznych. Innymi słowy, w modelach ekonometrycznych wykorzystuje się zmienne binarne z_i lub z_t, które odzwierciedlają:

- pojawienie się „szczególnej sytuacji", np.:

$$y_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \alpha_2 \cdot x_{2t} + \ldots + \alpha_k \cdot x_{kt} + \alpha_{k+1} \cdot z_t + \varepsilon_t \tag{6.10}$$

  wówczas zadaniem zmiennej zero-jedynkowej (która jest równa 1 tylko dla obserwacji „zakłócających") jest zniwelowanie efektu wynikającego z pojawieniem się „szczególnej sytuacji";
- warianty zmiennych mierzonych na skali nominalnej, np. płeć $z_i = 1$ dla kobiety:

$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + \ldots + \alpha_k \cdot x_{ki} + \alpha_{k+1} \cdot z_i + \varepsilon_t \tag{6.11}$$

21 Zdarza się, że t przyjmuje wartości ujemne, w szczególności kiedy zmienną czasową numeruje się w taki sposób, aby suma wszystkich t była równa 0. Dopuszczalne, aczkolwiek niezmiernie rzadkie jest też wyrażanie zmiennych czasowych w liczbach wymiernych, o ile interwał między nimi jest stały.

22 Omawiane w podrozdziale 5.2.

- w analizach czasowych, kiedy uwzględnia się wahania sezonowe, np. wahania występujące w kolejnych dekadach:

$$y_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \alpha_2 \cdot x_{2t} + \ldots + \alpha_k \cdot x_{kt} + \alpha_{k+1} \cdot z_{1t} + \alpha_{k+2} \cdot z_{2t} + \varepsilon_t \qquad (6.12)$$

gdzie:
zmienne z_{1t} i z_{2t} odzwierciedlają pierwszą i drugą dekadę w miesiącu.

W modelu (6.12) celowo pominięto jedną ze zmiennych, oznaczającą trzecią dekadę – gdyby ją uwzględnić, suma wektorów składających się ze zmiennych zero-jedynkowych utworzyłaby wektor składający się z samych jedynek, a więc zmiennych stojących przy α_0. Wówczas macierz obserwacji zmiennych objaśniających stałaby się macierzą niepełnego rzędu i nie można byłoby zastosować MNK do estymacji tego równania. Wprawdzie można przyjąć, że wyraz wolny w tym modelu wynosi 0 i oszacować równanie MNK, ale wtedy interpretacja parametrów modelu jest nieoczywista. Dlatego zazwyczaj pomija się jedną ze zmiennych zero-jedynkowych w modelach uwzględniających wahania sezonowe lub zawierających zmienne zero-jedynkowe reprezentujące różne warianty cech mierzonych na skali nominalnej[23]. Ta pominięta zmienna nosi nazwę zmiennej referencyjnej, a parametry stojące przy pozostałych zmiennych binarnych informują o relacji zmiennej objaśnianej w warunkach przedstawianych przez dany wariant cechy w stosunku do wariantu referencyjnego.

Podsumowując przedstawione w tym podrozdziale rozważania, należy stwierdzić, że w modelach składających się z jednego równania:

- nieopóźnione zmienne endogeniczne są równoważne zmiennym objaśnianym;
- zmienne egzogeniczne tworzą zbiór zmiennych objaśniających, jeśli w modelu nie występują zmienne objaśniające (endogeniczne) z opóźnieniami czasowymi;
- zmienne egzogeniczne bieżące i opóźnione oraz opóźnione zmienne endogeniczne są zmiennymi objaśniającymi, jeżeli model zawiera zmienne z opóźnieniami czasowymi[24].

Przykład 6.2

Korzystając z danych kwartalnych, dotyczących wartości PKB z przykładu 5.1, oszacować należy model (6.12) uwzględniający wahania sezonowe.

---

23 Zawsze, kiedy w modelach uwzględnia się cechy dwudzielne, np. płeć w modelu (6.11), lub cechy wielodzielne, jak np. sektor gospodarki, forma własności, zawód.

24 W modelach ekonometrycznych zawierających 2 i więcej równań wszystkie zmienne objaśniane są zmiennymi endogenicznymi, natomiast zmiennymi objaśniającymi mogą być zarówno zmienne egzogeniczne, jak i endogeniczne (bieżące i opóźnione). W modelach tych nieopóźnione zmienne endogeniczne nazywane są zmiennymi łącznie współzależnymi. Natomiast zmienne endogeniczne opóźnione razem z bieżącymi i opóźnionymi zmiennymi egzogenicznymi tworzą grupę zmiennych z góry ustalonych.

Rozwiązanie

W szeregu omawianym w przykładzie 5.1 stwierdzono występowanie trendu i wahań kwartalnych, zatem do jego analizy wykorzystać można model postaci:

$$\text{M0: } y_t = \alpha_0 + \alpha_1 \cdot z_{1t} + \alpha_2 \cdot z_{2t} + \alpha_3 \cdot z_{3t} + \alpha_4 \cdot z_{4t} + \alpha_5 \cdot t + \varepsilon_t$$

gdzie:
zmienne z_{1t}, z_2, z_3 i z_{4t} są zmiennymi zero-jedynkowymi i odzwierciedlają efekty sezonowe, tj. odchylenia PKB od trendu w poszczególnych kwartałach. Tab. 6.2 zawiera dane, które zostaną wykorzystane do estymacji modeli.

Tabela 6.2. Dane do estymacji modeli M1 i M2

Okresy badania	PKB	Zmienne					Okresy badania	PKB	Zmienne				
		t	z_1	z_2	z_3	z_4			t	z_1	z_2	z_3	z_4
2014Q1	399 536,9	1	1	0	0	0	2016Q4	526 019,9	12	0	0	0	1
2014Q2	418 230,9	2	0	1	0	0	2017Q1	458 461,0	13	1	0	0	0
2014Q3	425 004,3	3	0	0	1	0	2017Q2	478 886,8	14	0	1	0	0
2014Q4	477 657,9	4	0	0	0	1	2017Q3	491 398,1	15	0	0	1	0
2015Q1	415 701,4	5	1	0	0	0	2017Q4	560 568,2	16	0	0	0	1
2015Q2	434 227,3	6	0	1	0	0	2018Q1	487 129,3	17	1	0	0	0
2015Q3	439 967,9	7	0	0	1	0	2018Q2	507 606,2	18	0	1	0	0
2015Q4	510 331,0	8	0	0	0	1	2018Q3	525 180,3	19	0	0	1	0
2016Q1	430 165,1	9	1	0	0	0	2018Q4	595 756,1	20	0	0	0	1
2016Q2	450 116,2	10	0	1	0	0	2019Q1	520 197,2	21	1	0	0	0
2016Q3	454 810,4	11	0	0	1	0	2019Q2	545 556,2	22	0	1	0	0

Źródło: opracowanie własne.

Jak łatwo zauważyć w tab. 6.2, suma wektorów zmiennych zero-jedynkowych tworzy wektor składający się z samych jedynek. Zatem w przypadku modelu M0 macierz obserwacji staje się macierzą niepełnego rzędu i nie można estymować tego modelu MNK. W takim przypadku można zastosować kilka przekształconych wariantów tego modelu.

Po pierwsze, można zrezygnować z wyrazu wolnego i wtedy model będzie postaci:

$$\text{M1: } y_t = \alpha_1 \cdot z_{1t} + \alpha_2 \cdot z_{2t} + \alpha_3 \cdot z_{3t} + \alpha_4 \cdot z_{4t} + \alpha_5 \cdot t + \varepsilon_t$$

Po drugie, można zrezygnować z jednej zmiennej binarnej, co da nam model:

$$\text{M2: } y_t = \alpha_0 + \alpha_1 \cdot z_{1t} + \alpha_2 \cdot z_{2t} + \alpha_3 \cdot z_{3t} + \alpha_4 \cdot t + \varepsilon_t$$

Po trzecie, zaleca się odjęcie obserwacji jednej ze zmiennych binarnych, np. ostatniej, od pozostałych zmiennych, redukując w ten sposób liczbę zmiennych, które *nota bene* nie będą już zmiennymi zero-jedynkowymi *sensu stricte*, co dla odróżnienia oznaczone zostanie przez $\tilde{z}_{it}$, ($i = 1, 2, 3$), a sam model przyjmie postać:

$$\text{M3: } y_t = \alpha_0 + \alpha_1 \cdot \tilde{z}_{1t} + \alpha_2 \cdot \tilde{z}_{2t} + \alpha_3 \cdot \tilde{z}_{3t} + \alpha_4 \cdot t + \varepsilon_t$$

W tabeli 6.3 zamieszczono zbiór obserwacji zmiennych $\tilde{z}_{it}$, niezbędnych do oszacowania modelu M3 oraz wartości teoretyczne PKB, a w tabeli 6.4 – wyniki estymacji[25] metodą najmniejszych kwadratów 3 modeli: M1, M2 i M3. Przedstawiono podstawowe informacje o oszacowanych modelach: oceny estymatorów parametrów $\hat{\alpha}_i$, wartości sprawdzianu testu t dotyczącego istotności zmiennej i współczynnika determinacji R^2. Analizując wyniki, można zauważyć, że modele oznaczone jako M2 i M3 są *de facto* identycznie skonstruowane i mimo różnych wartości ocen estymatorów parametrów, mają ten sam stopień dopasowania modelu do danych empirycznych. Co więcej, wszystkie 3 modele generują te same wartości teoretyczne[26] (porównaj rys. 6.2), których stopień dopasowania do danych empirycznych mierzony za pomocą średniego błędu dopasowania (5.30) wynosi 1,25%, czyli porównywalnie z wynikami uzyskanymi w przykładzie 5.1 i 5.2.

Tabela 6.3. Dane do estymacji modelu M3 oraz wartości teoretyczne PKB

Okresy badania	PKB	$\widehat{\text{PKB}}$	Zmienne				Okresy badania	PKB	$\widehat{\text{PKB}}$	Zmienne			
			t	$\tilde{z}_1$	$\tilde{z}_2$	$\tilde{z}_3$				t	$\tilde{z}_1$	$\tilde{z}_2$	$\tilde{z}_3$
2014Q1	399 536,9	388 050,5	1	1	0	0	2016Q4	526 019,9	534 066,6	12	–1	–1	–1
2014Q2	418 230,9	408 622,6	2	0	1	0	2017Q1	458 461,0	464 628,1	13	1	0	0
2014Q3	425 004,3	416 220,5	3	0	0	1	2017Q2	478 886,8	485 200,2	14	0	1	0
2014Q4	477 657,9	483 014,9	4	–1	–1	–1	2017Q3	491 398,1	492 798,1	15	0	0	1
2015Q1	415 701,4	413 576,3	5	1	0	0	2017Q4	560 568,2	559 592,5	16	–1	–1	–1
2015Q2	434 227,3	434 148,4	6	0	1	0	2018Q1	487 129,3	490 154,0	17	1	0	0
2015Q3	439 967,9	441 746,3	7	0	0	1	2018Q2	507 606,2	510 726,1	18	0	1	0
2015Q4	510 331,0	508 540,7	8	–1	–1	–1	2018Q3	525 180,3	518 324,0	19	0	0	1
2016Q1	430 165,1	439 102,2	9	1	0	0	2018Q4	595 756,1	585 118,4	20	–1	–1	–1
2016Q2	450 116,2	459 674,3	10	0	1	0	2019Q1	520 197,2	515 679,8	21	1	0	0
2016Q3	454 810,4	467 272,2	11	0	0	1	2019Q2	545 556,2	536 252,0	22	0	1	0

Uwaga: $\widehat{\text{PKB}}$ – wartości teoretyczne PKB wyznaczone z oszacowanych modeli.

Źródło: opracowanie własne.

25 Modele zostały oszacowane przy wykorzystaniu funkcji REGLINP w Excelu.

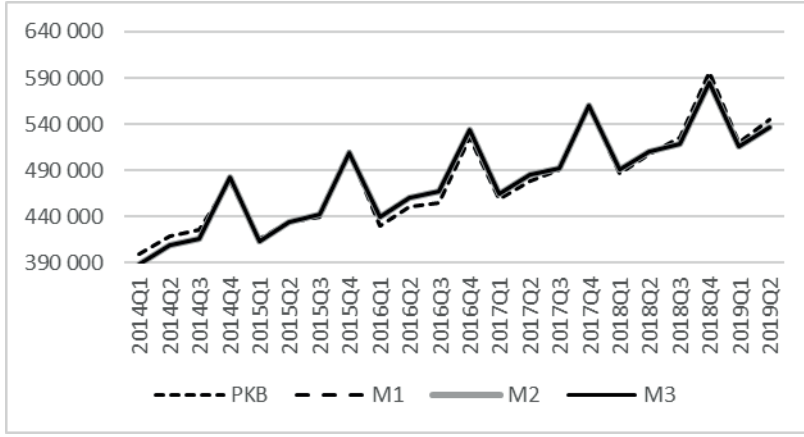
26 Wartości teoretyczne wyznaczono przy wykorzystaniu funkcji REGLINW w Excelu.

Tabela 6.4. Charakterystyki oszacowanych modeli

Model		Oceny estymatorów parametrów i wartości statystyk testowych					
		$\hat{\alpha}_4$	$\hat{\alpha}_3$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_5$	$\hat{\alpha}_0$
M1	$\hat{\alpha}_i$	457 488,99	397 076,04	395 859,63	381 668,99	6 381,47	
$R^2 = 0{,}9998$	t	**94,35**	**85,00**	**85,66**	**86,07**	**23,54**	
M2	$\hat{\alpha}_i$		–60 412,94	–61 629,35	–75 820,00	6 381,47	457 488,99
$R^2 = 0{,}9809$	t		**–11,86**	**–12,66**	**–15,55**	**23,54**	**94,35**
M3	$\hat{\alpha}_i$		–10 947,37	–12 163,78	–26 354,43	6 381,47	408 023,41
$R^2 = 0{,}9809$	t		**-3,56**	**–4,20**	**–9,11**	**23,54**	**114,59**

Uwaga: pogrubiono wartości statystyk testowych pozwalających na odrzucenie hipotezy zerowej o braku istotności zmiennej na poziomie istotności 0,05. Oznaczenia parametrów są takie jak w modelu M0.

Źródło: obliczenia własne.



Rysunek 6.1. PKB – porównanie wartości teoretycznych modeli M1–M3 z danymi empirycznymi

Źródło: opracowanie własne.

Zastosowanie modelu trendu z sezonowością pozwala na dekompozycję szeregu czasowego w znacznie mniej pracochłonny sposób, niż to pokazano w poprzednim rozdziale. Warto zauważyć, że parametry modelu M3 interpretuje się jako stałe efekty sezonowe dla trzech pierwszych kwartałów (porównywalne do tych wyznaczonych w przykładzie 5.1, tab. 5.1). Innymi słowy, w pierwszych kwartałach każdego roku PKB było średnio niższe o 60 412,94 mln PLN od wartości tej zmiennej wyznaczonej na podstawie liniowej funkcji trendu. Natomiast efekt sezonowy ostatniego kwartału wyznacza się jako $(-\sum_{i=1}^{3}\hat{\alpha}_i)$, co w analizowanym przykładzie daje

wartość 49 465,57 mln PLN. Zatem jest to wartość, o jaką PKB jest średnio większe w ostatnich kwartałach od wartości wynikającej z trendu liniowego. W modelu M2, oceny parametrów strukturalnych, stojących przy zmiennych zero-jedynkowych, interpretowane są jako różnica efektu sezonowego w danym kwartale w stosunku do kwartału czwartego (tj. zmiennej referencyjnej). Przykładowo, z podanych w tab. 6.4 wartości oszacowanych parametrów wynika, że w pierwszym kwartale efekt sezonowy był mniejszy o 10 947,37 mln PLN niż w kwartale czwartym.

Przykład 6.3

Przeprowadzono identyfikację czynników[27] wpływających na roczne stopy zwrotu z akcji 15 wybranych (spośród 30) spółek tworzących portfel indeksu Dow Jones Industrial Average (DJIA) w latach 2000–2017. Podstawowym kryterium wyboru spółek do badania była ich nieprzerwana obecność w składzie indeksu w okresie 2000–2017 oraz dostępność danych. Jest to przykład danych panelowych (przekrojowo-czasowych), obserwacje bowiem dotyczą 15 różnych spółek w 18-letnim okresie – razem 270 obserwacji.

Identyfikację czynników przeprowadzono, budując modele (6.2) opisujące relacje między rocznymi stopami zwrotu wybranych spółek (y) i stopami zwrotu z indeksu DIJA oraz wybranymi czynnikami wykorzystywanymi w analizie fundamentalnej (x_j). Badanie przeprowadzono w 3 etapach:

- wybór zmiennych do modeli,
- estymacja MNK modeli oraz
- weryfikacja modeli.

Rozwiązanie

Analizie poddano 21 zmiennych finansowych, zazwyczaj wykorzystywanych w analizie fundamentalnej, z których wybrano 6 o największej wartości współczynnika korelacji liniowej Pearsona z rocznymi stopami zwrotu z akcji wybranych spółek. Oprócz zmiennych charakteryzujących poszczególne spółki w badaniu uwzględniono ogólną sytuację na Giełdzie Papierów Wartościowych w Nowym Jorku, reprezentowaną przez indeks giełdowy DJIA. Podstawowe informacje dotyczące średnich wartości zmiennych, ich dyspersji i korelacji zmiennych objaśniających ze zmienną objaśnianą przedstawiono w tab. 6.5. Jak widać, wszystkie zmienne charakteryzują się znaczącą zmiennością (współczynniki zmienności znacząco większe od 0,1). Najsilniej skorelowaną (ze stopami zwrotu z akcji badanych spółek) zmienną objaśniającą jest stopa zwrotu z indeksu DIJA, która charakteryzuje się mniejszą zmiennością niż zwroty z akcji badanych spółek, chociaż w przypadku obu zmiennych jest ona znaczna (DIJA – 272%, spółki 320% wokół średniej).

27 Podstawowe badanie zostało przeprowadzone w pracy [Styś 2019].

Tabela 6.5. Podstawowe informacje o zmiennych wykorzystanych w modelach

Zmienne	Średnia	Współczynnik zmienności	Odchylenie standardowe	Współczynnik Pearsona
Stopa zwrotu z 15 spółek	0,08	3,20	0,26	×
Marża zysku netto	12,74	0,60	7,68	0,0917
Zadłużenie długoterminowe/ zadłużenie całkowite	0,74	0,31	0,23	0,1478
ROE	35,50	2,53	89,67	0,1895
Stopa dywidendy	2,19	0,50	1,09	0,1641
Cena/zysk	21,73	0,72	15,70	0,1762
Cena/wartość księgowa	7,05	4,11	28,96	0,2042
Stopa zwrotu z indeksu DIJA	0,06	2,72	0,15	0,5929

Źródło: opracowanie własne.

Tabela 6.6. Oszacowane MNK modele

Zmienne	Parametry	Statystyka t-Studenta	Parametry	Statystyka t-Studenta
	Model M1		Model M2	
Współczynnik determinacji R^2	0,4084	×	0,4062	×
Marża zysku netto	–0,0017	–0,9765	×	×
Zadłużenie długoterminowe/ zadłużenie całkowite	0,0793	1,2763	0,1020	**1,7716**
ROE	0,0002	0,5804	0,0002	0,5506
Stopa dywidendy	–0,0313	**–2,5317**	–0,0333	**–2,7303**
Cena/zysk	0,0014	**1,6727**	0,0015	**1,8121**
Cena/wartość księgowa	0,0006	0,4891	0,0006	0,5411
Stopa zwrotu z indeksu DIJA	0,9742	**11,3304**	0,9640	**11,2965**
Wyraz wolny	0,0174	0,2653	–0,0189	–0,3501

Uwaga: pogrubioną czcionką oznaczono wartości statystyk testowych, wskazujących na odrzucenie hipotezy o zerowych wartościach parametrów strukturalnych modeli na poziomie istotności 0,05.

Źródło: opracowanie własne.

Skonstruowano modele zawierające różne zestawy zmiennych objaśniających. Oszacowane MNK modele[28] przedstawiono w tab. 6.6. Jak widać, wartości

28 Modele zostały oszacowane przy wykorzystaniu funkcji REGLINP w Excelu.

współczynników determinacji nie przekraczają 0,41, czyli nie są zbyt wysokie. Spośród 7 zmiennych objaśniających tylko 3 z nich: stopa dywidendy, stosunek ceny do zysku i stopa zwrotu z indeksu DIJA są statystycznie istotne w obu modelach. W modelu M2, z którego usunięto marżę zysku netto, stosunek zadłużenia długoterminowego do całkowitego stał się statystycznie istotny. W obu modelach stopa zwrotu z indeksu DIJA, odzwierciedlająca sytuację na rynku kapitałowym, w znacznym stopniu wpływa na stopy zwrotu z akcji analizowanych spółek. Wzrost stopy zwrotu z portfela indeksu DIJA o 1% przyczynia się do wzrostu stopy zwrotu badanych spółek o 0,97 i 0,96% odpowiednio w modelu M1 i M2. Model, w którym ta zmienna jest jedyną zmienną objaśniającą, tłumaczy zmiany stóp zwrotu z akcji 15 wybranych spółek w 35% (współczynnik korelacji liniowej Pearsona dla tych 2 zmiennych wynosi 0,5929 – tab. 6.5)[29].

Tabela 6.7. Oszacowane MNK modele

Zmienne	Parametry	Statystyka t-Studenta	Parametry	Statystyka t-Studenta
	Model M3		Model M4	
Współczynnik determinacji R^2	0,1185	×	0,3864	×
Marża zysku netto	0,0007	0,3324	×	×
Zadłużenie długoterminowe/ zadłużenie całkowite	0,2232	**3,0139**	0,1083	1,8640
ROE	0,0003	0,5278	×	×
Stopa dywidendy	–0,0390	**–2,5942**	–0,0333	**–2,7078**
Cena/zysk	0,0026	**2,4906**	0,0015	**1,7624**
Cena/wartość księgowa	0,0010	0,6973	×	×
Stopa zwrotu z indeksu DIJA	×	×	0,9880	**11,4847**
Wyraz wolny	–0,0781	–0,9843	–0,0117	–0,2149

Uwaga: pogrubioną czcionką oznaczono wartości statystyk testowych, wskazujących na odrzucenie hipotezy o zerowych wartościach parametrów strukturalnych modeli na poziomie istotności 0,05.

Źródło: opracowanie własne.

Istotny wpływ sytuacji rynkowej na zwroty z akcji potwierdza model M3, który jedynie w 12% wyjaśnia zmiany stóp zwrotu zmianami 6 wskaźników finansowych, co jest znacząco gorszym wynikiem w porównaniu z modelem zawiera-

29 W modelach z jedną zmienną objaśniającą współczynnik determinacji jest równy kwadratowi współczynnika korelacji liniowej Pearsona.

jącym jedynie stopy zwrotu z DIJA (zmienna pominięta w modelu M3). Model M4 zawiera jedynie 3 statystycznie istotne na poziomie istotności 0,5 wskaźniki finansowe. Warto odnotować, że R^2 wynosi 0,39 wyłącznie dzięki obecności stopy zwrotu z portfela indeksu DIJA (tab. 6.7).

Należy zwrócić uwagę, że wprawdzie w przykładzie 6.3 modele oszacowane zostały na podstawie danych panelowych, ale dane te potraktowano je jak dane przekrojowe. Oznacza to, że w analizowanym zjawisku założono brak występowania efektów indywidualnych i zmian w czasie. W konsekwencji do estymacji parametrów strukturalnych modelu zastosowano MNK. Jest to najprostszy sposób analizy danych panelowych, który w literaturze przedmiotu nazywany jest *pooled regression*.

6.5. Modele zmiennych jakościowych

W badaniach ekonomicznych zmiennymi objaśnianymi mogą być również zmienne o charakterze jakościowym, czyli takie, dla których można jedynie stwierdzić występowanie konkretnego wariantu cechy. Zmienne mierzone na skali nominalnej można zakodować, tworząc odpowiednie zmienne zero-jedynkowe (binarne). W przypadku cechy o więcej niż dwóch wariantach definiuje się wiele zmiennych binarnych (k wariantów atrybutu implikuje występowanie w modelu (k-1) zmiennych zero-jedynkowych). Przykładem jakościowych zmiennych objaśnianych może być przynależność do określonej grupy dochodowej lub ryzyka kredytowego, sytuacja finansowa przedsiębiorstwa, aktywność zawodowa (zatrudniony czy bezrobotny). Jeśli, jak to pokazano w przykładzie 1.2, cecha przyjmuje tylko dwa warianty (np. przy kodowaniu binarnym dla wiarygodnego kredytobiorcy zmienna przyjmuje wartość 1, a dla niewiarygodnego 0), to taka zmienna nazywana jest dychotomiczną (dwudzielną).

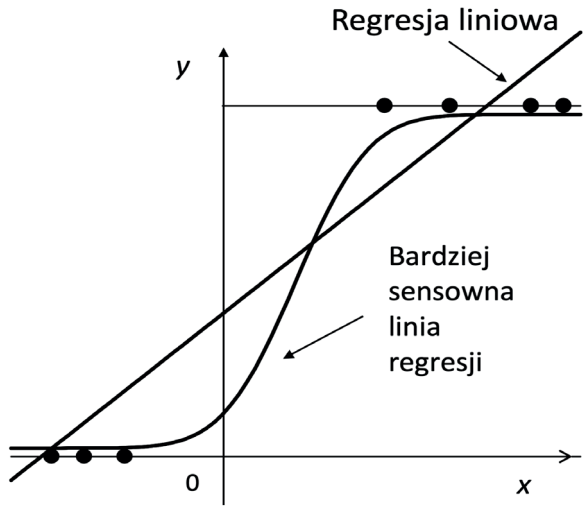
Model, w którym zmienna objaśniana jest mierzona na skali nominalnej, nazywa się modelem zmiennych jakościowych, modelem dyskretnych wyborów (*discrete* lub *qualitative choice models*) lub modelem regresji logistycznej. Modele tego typu umożliwiają dokonanie dyskretnych (jakościowych) wyborów między dwiema alternatywnymi lub większą liczbą możliwości. W modelach regresji logistycznej opisywany jest związek między zmiennymi objaśniającymi, które mogą być zarówno ilościowe, jak i jakościowe (tj. mierzone na dowolnej skali pomiarowej), a pojedynczą zmienną binarną, pełniącą rolę zmiennej objaśnianej. Modele

zmiennych jakościowych należą do mikroekonometrii, ponieważ zazwyczaj wykorzystuje się je w analizach danych indywidualnych (mikrodanych) dotyczących przedsiębiorstw, klientów banków, pracowników itp. W teorii i praktyce rozważa się różne postacie tych modeli, którym dedykowane są różne metody estymacji ich parametrów.

W dalszych rozważaniach (w celu ich uproszczenia) założone zostanie, że cechy są dwudzielne (dychotomiczne), tj. przyjmują dwa stany, np. zatrudniony i bezrobotny, bankrut i przedsiębiorstwo dalej funkcjonujące, kredyt obarczony wysokim i niskim ryzykiem. Binarnie zakodowaną zmienną traktuje się jako zmienną losową dwumianową opisaną wzorem (1.8) i przedstawioną na rys. 1.9. Jeśli zmienna zero-jedynkowa znajdzie się po lewej stronie modelu regresji (4.29) (tzn. jest zmienną objaśnianą), to otrzymuje się:

$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + ... + \alpha_k \cdot x_{ki} + \varepsilon_i \tag{6.13}$$

gdzie dla każdego $i = 1, 2, \ldots, n$:
y_i – zmienna zależna przyjmująca dla i-tej obserwacji wartość 0 lub 1, co oznacza brak lub występowanie analizowanego wariantu cechy jakościowej,
$x_{1i}, x_{2i}, \ldots, x_{ki}$ – obserwacje zmiennych niezależnych,
$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$ – nieznane parametry strukturalne modelu,
ε_i – składnik losowy.



Rysunek 6.2. Zależności występujące w liniowym modelu prawdopodobieństwa

Źródło: [Maddala 2006, s. 369].

Model (6.13) nazywa się liniowym modelem prawdopodobieństwa lub modelem Goldbergera (por. [Maddala 2006, s. 367–371; Gruszczyński 2010, s. 57–62; Huterska, Zdunek-Rosa 2016, s. 123]) i jest on najprostszym modelem, w którym zmienna objaśniana jest zmienną jakościową (mierzoną na skali nominalnej). Zmienne objaśniające w modelu (6.13) mogą być zarówno ciągłe, jak i dyskretne, w tym również zmienne zero-jedynkowe, będące szczególnym przypadkiem zmiennych skokowych. W tym przypadku modeluje się prawdopodobieństwo realizacji wariantu zmiennej y równego 1, czyli $P(y_i = 1) = p_i$. Jednakże model ten nie zapewnia, że wartości teoretyczne $\hat{y}_i$ znajdą się w przedziale [0; 1], co uniemożliwia ich interpretację w kategoriach prawdopodobieństwa. Pojawiają się też problemy z estymacją MNK modelu (6.13), wariancja składnika losowego jest bowiem niejednorodna (składnik losowy jest heteroskedastyczny). Warto w tym miejscu zauważyć, że w przypadku liniowych modeli prawdopodobieństwa wartość współczynnika determinacji R^2 (4.40) jest z reguły niska, zwłaszcza kiedy oszacowane prawdopodobieństwa nie osiągają wartości z krańca przedziału [0; 1]. Zatem nawet niska (tj. bliska 0, np. 0,10) wartość R^2 może świadczyć o poprawności oszacowanego modelu, co jest doskonale widoczne na rys. 6.2.

Interpretacja ocen estymatorów parametrów modelu (6.13) jest odmienna dla zmiennych objaśniających ilościowych oraz zero-jedynkowych. W pierwszym przypadku stwierdza się, że jeśli wartość zmiennej objaśniającej x_j ($j = 1, 2, \ldots, k$) wzrośnie o jednostkę, to prawdopodobieństwo wystąpienia zdarzenia opisywanego przez y wzrośnie o $\hat{\alpha}_j$, o ile pozostałe parametry pozostaną na niezmienionym poziomie (*ceteris paribus*). Innymi słowy $\hat{\alpha}_j$ określa wpływ jednostkowej zmiany zmiennej niezależnej na wielkość prawdopodobieństwa sukcesu (tj. zdarzenia opisywanego przez y). W drugim przypadku $\hat{\alpha}_j$ określa różnicę między prawdopodobieństwem wystąpienia sukcesu dla zmiennej binarnej równej 0 i równej 1, przy pozostałych zmiennych ustalonych na poziomach średnich.

Należy jednak pamiętać, że omówione interpretacje parametrów są prawdziwe jedynie przy zachowaniu zasady *ceteris paribus*, zgodnie z którą interpretuje się pojedynczą zmianę przy zachowaniu wszystkich pozostałych warunków na stałym poziomie. Innymi słowy, porównuje się np. dwa badane obiekty, które mają wszystkie cechy identyczne z wyjątkiem wartości jednej – interpretowanej zmiennej.

Jak wcześniej wspomniano, wartości teoretyczne modelu (6.13) mogą znajdować się poza przedziałem [0; 1]. W celu pozbycia się tej wady liniowego modelu prawdopodobieństwa, przeprowadza się transformację prawdopodobieństwa w taki sposób, aby zmienna objaśniana należała do przedziału $(-\infty; \infty)$. Dlatego zaleca się funkcję logitową do analizy zmiennych jakościowych lub zmiennych ilościowych i jakościowych, która obok przekształcenia probitowego jest najczęściej wykorzystywaną metodą transformacji prawdopodobieństwa.

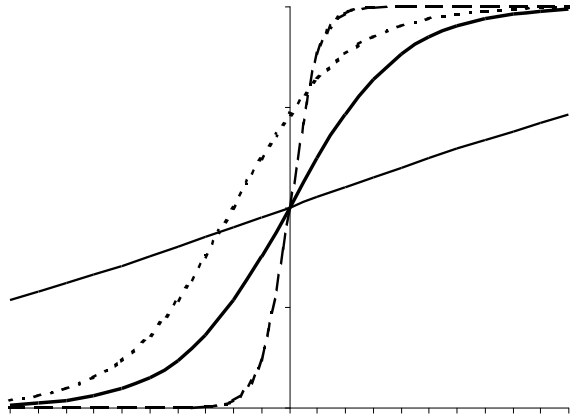
Dystrybuantę rozkładu logistycznego zazwyczaj zapisuje się w postaci:

$$p_i = \varphi(y_i) = \frac{e^{y_i}}{e^{y_i}+1} = \frac{1}{1+e^{-y_i}} = \frac{1}{1+\exp(-y_i)} \tag{6.14}$$

dla:

$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + ... + \alpha_k \cdot x_{ki} + \varepsilon_i \tag{6.15}$$

gdzie:
e – liczba Eulera będąca podstawą logarytmu naturalnego. Dla $\varphi(y_i) = \frac{1}{1+\exp(-\beta \cdot x_i)}$ dystrybuantę rozkładu logistycznego pokazano na rys. 6.3.



Rysunek 6.3. Przykład dystrybuanty rozkładu logistycznego

Źródło: [Witkowska 2002, s. 7].

Użyteczność modelu logitowego wynika z faktu, że funkcję odwrotną do (6.14) $\varphi^{-1}(y)$ można zapisać jako logit, czyli logarytm naturalny ilorazu prawdopodobieństwa danego zdarzenia i zdarzenia do niego przeciwnego:

$$y_i = ln\frac{p_i}{1-p_i} \tag{6.16}$$

Po podstawieniu (6.16) do (6.14) otrzymuje się [Gruszczyński 2010, s. 62]:

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + \ldots + \alpha_k \cdot x_{ki} + \varepsilon_i \tag{6.17}$$

Model (6.17) nosi nazwę modelu logitowego (logistycznego) lub regresji logistycznej, ponieważ zmienna losowa Y ma rozkład logistyczny. Logitowa transformacja prawdopodobieństwa pozwala zastąpić miarę p_i, liczbą z przedziału $(-\infty; \infty)$, co jest główną korzyścią ze stosowania modelu logitowego. Modele logitowe opisują relację między dychotomiczną zmienną objaśnianą a zmiennymi objaśniającymi[30].

W modelu tym nie interpretuje się bezpośrednio wartości parametrów, a jedynie efekty krańcowe (cząstkowe). Zatem w przypadku ocen estymatorów parametrów modelu logitowego $\hat{\alpha}_j$ można jedynie stwierdzić, że:

- jeżeli $\hat{\alpha}_j > 0$, to czynnik opisywany przez zmienną x_j działa stymulująco na prawdopodobieństwo wystąpienia badanego zjawiska,
- jeżeli $\hat{\alpha}_j < 0$, to czynnik opisywany przez zmienną x_j działa destymulująco na prawdopodobieństwo wystąpienia badanego zjawiska,
- jeżeli $\hat{\alpha}_j = 0$, to czynnik opisywany przez zmienną x_j nie wpływa na prawdopodobieństwo wystąpienia badanego zjawiska [Gruszczyński 2010, s. 60].

Innymi słowy, znak parametru przy zmiennej egzogenicznej x_j pokazuje kierunek jej wpływu na zmienną y. Jeśli ten parametr jest dodatni, wówczas wzrost wartości zmiennej x_j wiąże się ze wzrostem szans na to, że $y = 1$, a dla ujemnego parametru wyższym wartościom x_j odpowiada niższe prawdopodobieństwo tego, że $y = 1$.

W modelu (6.17) zmienną objaśnianą jest zmienna $ln\left(\frac{p_i}{1-p_i}\right)$ nazywana logitem, gdzie $p_i = P(y_i = 1)$, czyli p_i jest prawdopodobieństwem, że zmienna losowa y_i przyjmie wartość 1. Stąd logit jest logarytmem tzw. ilorazu szans (*odds ratio*), czyli ilorazu prawdopodobieństwa przyjęcia i nieprzyjęcia przez zmienną losową Y wartości jeden. Jeśli te prawdopodobieństwa (szanse) są jednakowe, tj. $p_i = 0{,}5$, to logit wynosi 0[31], jeśli $p_i > 0{,}5$ to logit jest dodatni, a dla $p_i < 0{,}5$ logit jest ujemny. Interpretując parametry modelu logitowego, można zauważyć, że wzrost wartości zmiennej x_j ($j = 1, 2, \ldots, k$) o jednostkę prowadzi do wzrostu ilorazu szans o $e^{\check{\alpha}_j} = \exp(\check{\alpha}_j)$. Innymi słowy, ilorazy szans są krotnościami, o jakie zmienią się one (ilorazy szans) przy wzroście każdej ze zmiennych o jednostkę. Wartości $\exp(\hat{\alpha}_j) - 1$ można interpretować jako zwiększenie szansy na podjęcie decyzji w odniesieniu do $y = 1$, w przypadku jednostkowej zmiany zmiennej x_j, co można wyrazić w procentach, czyli interpretować wyrażenie $[\exp(\hat{\alpha}_j) - 1] \cdot 100\%$.

Wpływ zmian wartości zmiennej na wartość prawdopodobieństwa przyjęcia przez zmienną objaśnianą wartości 1 definiuje się jako efekt krańcowy, który wynosi:

$$\hat{\alpha}_j \cdot p_i(1-p_i) = \hat{\alpha}_j \cdot \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_{1i} + \hat{\alpha}_2 \cdot x_{2i} + \ldots + \hat{\alpha}_1 \cdot x_{ki})}{\left[1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot x_{1i} + \hat{\alpha}_2 \cdot x_{2i} + \ldots + \hat{\alpha}_1 \cdot x_{ki})\right]^2} \tag{6.18}$$

30 Podobne w konstrukcji są modele probitowe, których opis można znaleźć w literaturze, np. [Gruszczyński 2010, s. 69–90].

31 Wynika to z wykonania działania $ln\left(\frac{p_i}{1-p_i}\right) = ln\left(\frac{0{,}5}{1-0{,}5}\right) = \ln(1) = 0$.

Efekty krańcowe zależą zatem od wartości zmiennej objaśnianej i parametru strukturalnego. Można je interpretować jako przyrost prawdopodobieństwa związanego z jednostkowym wzrostem każdej ze zmiennych „w okolicy" wybranej jednostki obserwacji. W związku z tym dobrze jest je wyznaczać dla jednostki będącej charakterystyczną w próbie badawczej. Dlatego efekty krańcowe są najczęściej liczone dla średnich wartości zmiennych objaśniających.

Model regresji logistycznej (6.17) jest modelem nieliniowym, a prawdopodobieństwa p_i są nieobserwowalne, znane są bowiem jedynie realizacje zmiennej losowej Y, czyli y_i, które przyjmują wartości binarne. Właściwą metodą estymacji modelu logistycznego jest metoda największej wiarygodności (MNW). W modelu logitowym, ze względu na jego nieliniowość, nie można stosować do oceny poprawności modelu klasycznego współczynnika determinacji R^2, zatem w tym przypadku stosuje się tzw. *pseudo* R^2 McFaddena[32] (*likelihood-ratio index*), który wyznacza się jako:

$$pseudo\ R^2 = 1 - \frac{lnL_{pełny}}{lnL_0} \tag{6.18}$$

gdzie:
$lnL_{pełny}$ – logarytm funkcji wiarygodności dla pełnego modelu,
lnL_0 – logarytm funkcji wiarygodności dla modelu zawierającego jedynie wyraz wolny, co odpowiada sytuacji, w której prawdopodobieństwo podjęcia każdej decyzji jest takie samo.

Zatem *pseudo* R^2 McFaddena pokazuje, o ile lepiej dopasowany do danych jest oszacowany model w porównaniu z modelem naiwnym, zawierającym jedynie wyraz wolny. Okazuje się przy tym, że im większa liczba obserwacji jest w próbie, na podstawie której estymuje się parametry modelu, tym niższą wartość przyjmuje *pseudo* R^2 McFaddena, który dodatkowo nigdy nie osiąga wartości 1. Jednocześnie wartość miernika McFaddena, podobnie jak klasycznego współczynnika determinacji R^2, zwiększa się wraz ze wzrostem liczby zmiennych objaśniających. Dlatego korzysta się również z miernika (6.18) skorygowanego o liczbę zmiennych w modelu, który jest postaci:

$$pseudo\ \overline{R}^2 = 1 - \frac{lnL_{pełny} - k}{lnL_0} \tag{6.19}$$

gdzie:
k – liczba parametrów szacowanych w modelu.

32 W literaturze przedmiotu omawia się różne rodzaje mierników określanych jako *pseudo* R^2 (por. [Maddala 2006, s. 377–379]).

Natomiast do oceny jakości dopasowania modelu do danych empirycznych wykorzystuje się tzw. zliczeniowy (liczebnościowy) współczynnik determinacji R^2 (*count*-R^2), który określa udział liczby trafnie wyznaczonych wartości teoretycznych w ogólnej liczbie obserwacji[33]. Innymi słowy, zliczeniowy współczynnik determinacji określa odsetek przypadków, w których wartość teoretyczna zmiennej losowej Y, czyli $\hat{y}_i$ była równa wartości empirycznej y_i w stosunku do liczebności próby estymacyjnej. Zatem zliczeniowy współczynnik determinacji jest postaci:

$$zliczeniowy\ R^2 = \frac{\hat{n}}{n} \tag{6.20}$$

gdzie:
$\hat{n}$ – liczba przypadków, w których wartość teoretyczna zmiennej losowej Y, czyli $\hat{y}_i$, jest równa wartości empirycznej y_i, czyli jest to liczba prawidłowo „rozpoznanych" przez model przypadków,
n – liczba wszystkich obserwacji w próbie estymacyjnej.

Wartość współczynnika (6.20) może być jednak myląca. Przykładowo, jeśli wszystkie wartości teoretyczne binarnej zmiennej objaśnianej $\hat{y}_i$ będą równe najczęściej występującemu wariantowi tej zmiennej, wówczas zliczeniowy R^2 będzie równy udziałowi liczby częściej obserwowanego wariantu w liczebności całej próby (tj. przynajmniej równy 0,5). Dlatego zaleca się korzystanie ze skorygowanego współczynnika zliczeniowego postaci:

$$zliczeniowy\ \overline{R}^2 = \frac{\hat{n} - n_{max}}{n - n_{max}} \tag{6.21}$$

gdzie:
n_{max} – liczba obserwacji wariantu częściej występującego.

Miara (6.21) jest co najwyżej równa jeden, ale jest dodatnia tylko wtedy, kiedy udział przypadków, w których wartość teoretyczna $\hat{y}_i$ jest równa wartości empirycznej y_i, przekracza udział najczęściej obserwowanego w próbie wariantu zmiennej.

Przykład 6.4

Przykład pochodzi z pracy [Gruszczyński 2010, s. 59, 60 i 64], z której zaczerpnięto średnie wartości zmiennych objaśniających oraz oszacowania liniowego modelu prawdopodobieństwa i modelu logitowego, przedstawione w tab. 6.8–6.9. Zmienną zależną jest zakup dodatkowego ubezpieczenia zdrowotnego (oznaczony jako

33 Więcej na temat tej miary zostanie powiedziane w rozdziale 8.

$y = 1$, a sytuacja przeciwna $y = 0$), a zmiennymi objaśniającymi są indywidualne cechy klientów. Próba badawcza wynosi 3206 osób.

Tabela 6.8. Charakterystyka zmiennych

Cechy badanych	Rodzaj zmiennej	Zakres zmienności	Średnia
Źródło dochodów – x_1	Dychotomiczna	Emeryt: $x = 1$	0,6247661
Własna ocena stanu zdrowia – x_2	Dychotomiczna	Dobre i bardzo dobre: $x = 1$	0,7046163
Liczba ograniczeń codziennej aktywności – x_3	Skokowa	$x = 1, 2, \ldots, 5$	0,3016220
Roczny dochód [tys. USD] – x_4	Ciągła	$x = \in [0; 1312{,}124]$	45,2639100
Liczba lat nauki – x_5	Skokowa	$x = 1, 2, \ldots, 17$	11,8986300
Stan cywilny – x_6	Dychotomiczna	W związku małżeńskim: $x = 1$	0,7330006

Źródło: opracowanie własne na podstawie danych z pracy [Gruszczyński 2010, s. 59].

Korzystając z danych przedstawionych w tab. 6.8–6.9 dla modelu logitowego należy wyznaczyć efekty krańcowe i ilorazy szans oraz przedstawić interpretację uzyskanych wyników estymacji modeli (6.13) i (6.17)

Rozwiązanie

Liniowy model prawdopodobieństwa, którego oszacowania MNK przedstawiono w drugiej kolumnie tab. 6.9, jest postaci:

$$\hat{y} = -0{,}0730 + 0{,}0310 \cdot x_1 + 0{,}4099 \cdot x_2 - 0{,}3284 \cdot x_3 + 0{,}0005 \cdot x_4 + \\ + 0{,}0263 \cdot x_5 + 0{,}1186 \cdot x_6$$

Wynika z niego, że emeryci ($x_1 = 1$) w porównaniu z pozostałymi respondentami ($x_1 = 0$) charakteryzują się większym średnio o 0,031 prawdopodobieństwem tego, że mają wykupione dodatkowe ubezpieczenie zdrowotne. Również osoby, które oceniają swój stan zdrowia jako dobry i bardzo dobry ($x_2 = 1$) mają to prawdopodobieństwo wyższe o 0,4099 w porównaniu do tych, którzy nie oceniają swojego stanu zdrowia tak wysoko ($x_2 = 0$). Z kolei osoby pozostające w związku małżeńskim ($x_6 = 1$) mają dodatkowe ubezpieczenie zdrowotne z większym o 0,1186 prawdopodobieństwem niż pozostali badani ($x_6 = 0$). Zwiększenie dochodów (x_4) o 1 tys. USD rocznie zwiększa prawdopodobieństwo posiadania dodatkowego ubezpieczenia o 0,0005, a każdy dodatkowy rok nauki (x_5) zwiększa to prawdopodobieństwo o 0,0263. Z kolei dodatkowe zwiększenie liczby ograniczeń codziennej aktywności (x_3) zmniejsza prawdopodobieństwo, że badany jest dodatkowo ubezpieczony, o 0,0328.

Tabela 6.9. Dane do analiz

Zmienne	Oszacowania parametrów		Efekty krańcowe dla średnich wartości zmiennych objaśniających	Ilorazy szans
	modelu (6.13)	modelu (6.17)		
x_1	0,0309614	0,149676	0,0350616	1,161458
x_2	0,0409946	0,1816156	0,0425435	1,199153
x_3	–0,0328428	–0,1957039	–0,0458437	0,822256
x_4	0,0004717	0,0022105	0,0005178	1,002213
x_5	0,0263493	0,1273481	0,0298313	1,135812
x_6	0,1185766	0,5589380	0,1309313	1,748814
wyraz wolny	–0,0730228	–2,7004320		

Źródło: oszacowania parametrów modeli pochodzą z pracy [Gruszczyński 2010, s. 60, 64]; efekty krańcowe i ilorazy szans – obliczenia własne.

Warto również przypomnieć, że w przypadku liniowego modelu prawdopodobieństwa niektóre wartości teoretyczne mogą znaleźć się poza przedziałem (0; 1). Przykładowo wartość teoretyczna oszacowanego modelu dla respondenta[34], który: (1) nie jest emerytem, (2) nisko ocenia swój stan zdrowia, (3) odnotował 5 ograniczeń w codziennej aktywności, (4) posiada roczny dochód 10 tys. USD, (5) uczył się przez 5 lat i (6) jest samotny, wynosi –0,1007733, czego nie da się interpretować w kategorii prawdopodobieństwa. Warto również odnotować, że wyznaczony dla tego modelu współczynnik determinacji wyniósł $R^2 = 0{,}0810$, a współczynnik skorygowany $\overline{R}^2 = 0{,}0793$. Spośród uwzględnionych w modelu zmiennych objaśniających tylko zmienna opisująca źródło dochodów okazała się być statystycznie nieistotna na poziomie istotności 0,05 (*p-value* wynosi 0,077).

Omawiając wartości ocen estymatorów parametrów modelu logitowego, oszacowanych metodą największej wiarygodności (MNW), postaci:

$$ln\left(\frac{p_i}{1-p_i}\right) = -2{,}7004 + 0{,}1497 \cdot x_1 + 0{,}1816 \cdot x_2 - 0{,}1957 \cdot x_3 + 0{,}0022 \cdot x_4 + \\ +0{,}1273 \cdot x_5 + 0{,}5589 \cdot x_6$$

można jedynie powiedzieć, że wszystkie wyróżnione cechy badanych z wyjątkiem liczby ograniczeń codziennej aktywności wpływają dodatnio na prawdopodobieństwo wykupienia dodatkowego ubezpieczenia zdrowotnego. Innymi słowy, emeryci i małżonkowie z większym prawdopodobieństwem wykupią dodatkowe

34 W pracy [Gruszczyński 2010, s. 60], podano zakres zmienności wartości teoretycznych dla badanej próby [–0,1685; 1,1843]. W prezentowanym przykładzie wartości zmiennych wynoszą: $x_1 = x_2 = x_6 = 0$, $x_3 = x_6 = 5$ i $x_4 = 10$.

ubezpieczenie zdrowotne niż osoby mające inne niż emerytura źródła utrzymania oraz samotni (formalnie) respondenci. Podobnie jak osoby, które oceniają swój stan zdrowia jako dobry i bardzo dobry w porównaniu z osobami, które gorzej oceniają swój stan zdrowia. Można również stwierdzić, że czym wyższe roczne dochody badanych oraz wyższe wykształcenie (mierzone liczbą lat nauki), tym prawdopodobieństwo wykupienia dodatkowego ubezpieczenia zdrowotnego jest większe. Natomiast liczba ograniczeń codziennej aktywności wpływa ujemnie, czyli czym więcej zgłoszonych ograniczeń tym prawdopodobieństwo wykupienia dodatkowego ubezpieczenia zdrowotnego jest mniejsze.

W tab. 6.9 przedstawiono obliczone ilorazy szans $e^{\hat{\alpha}_j} = \exp(\hat{\alpha}_j)$ oraz wyznaczone na podstawie (6.18) efekty krańcowe dla średnich wartości zmiennych objaśniających. Obliczenia wykonano w Excelu dla danych o takiej samej dokładności, jak zostały zacytowane w [Gruszczyński 2010, s. 59, 60 i 64].

Efekty krańcowe modelu logitowego, obliczone dla średnich wartości zmiennych objaśniających, interpretuje się podobnie jak oceny estymatorów parametrów liniowego modelu prawdopodobieństwa. Jak widać w tab. 6.9, wartości efektów krańcowych[35] są bliskie wartościom parametrów liniowego modelu prawdopodobieństwa. Interpretacja efektów krańcowych pozwala stwierdzić, że emeryci w porównaniu z pozostałymi badanymi ($x_1 = 0$) mają większe średnio o 0,0351 prawdopodobieństwo wykupienia dodatkowego ubezpieczenia zdrowotnego. Prawdopodobieństwo bycia dodatkowo ubezpieczonym w przypadku osób wysoko oceniających swój stan zdrowia ($x_2 = 1$) jest większe o 0,0423 w porównaniu do tych, którzy nie oceniają swojego stanu zdrowia tak wysoko ($x_2 = 0$). Małżonkowie ($x_6 = 1$) korzystają z dodatkowego ubezpieczenia zdrowotnego z prawdopodobieństwem większym o 0,1309 niż pozostali respondenci ($x_6 = 0$). Zwiększenie dochodów (x_4) o 1 tys. USD rocznie zwiększa prawdopodobieństwo posiadania dodatkowego ubezpieczenia o 0,0005, a każdy dodatkowy rok nauki (x_5) zwiększa to prawdopodobieństwo o 0,0298. Z kolei zwiększenie liczby ograniczeń codziennej aktywności (x_3) o jeden zmniejsza prawdopodobieństwo, że badany jest dodatkowo ubezpieczony, o 0,0458.

Interpretując wartości ilorazów szans, można powiedzieć, że szansa podjęcia decyzji o dodatkowym ubezpieczeniu zwiększa się u emerytów o 16,1% (iloraz szans wynosi 1,161), u osób z dobrym i bardzo dobrym stanem zdrowia o 19,9%, a u pozostających w związku małżeńskim o 74,9% w porównaniu z pozostałymi badanymi. Zwiększenie dochodów o 1 tys. USD zwiększa szansę podjęcia decyzji o 0,2%, a kolejny rok nauki o 13,6%. Natomiast kolejne ograniczenie codziennej aktywności zmniejsza szansę podjęcia decyzji o 17,8%, ponieważ $\exp(\hat{\alpha}_3) - 1 = 0{,}822 - 1 = -0{,}178 = -17{,}8\%$.

35 Przedstawione w tab. 6.9 wartości efektów krańcowych różnią się nieco od tych obliczonych przez pakiet STATA i przedstawionych w [Gruszczyński 2010, s. 66], co może wynikać z przyjętych zaokrągleń.

W modelu logitowym niemal wszystkie zmienne objaśniające okazały się statystycznie istotne na poziomie istotności 0,05 z wyjątkiem źródła dochodów i własnej oceny stanu zdrowia (*p-value* odpowiednio równe 0,064 i 0,059), a obliczony *pseudo* $R^2 = 0{,}0654$. Można zatem powiedzieć, że wnioski płynące z oszacowań obu modeli (tj. liniowego modelu prawdopodobieństwa i modelu logitowego) są podobne.

W wielu sytuacjach jakościowa zmienna objaśniana ma postać dyskretną, lecz przyjmuje więcej niż dwie wartości, np. grupy ryzyka kredytowego, których jest z reguły więcej niż dwie, ocena sytuacji finansowej przedsiębiorstwa wyróżniająca różne stany zagrożenia. W takim przypadku modele binarne zastępowane są modelami wielomianowymi (*multiple outcome*, *multiple response* lub *multiresponse model*), wśród których w zależności od rodzaju zmiennej objaśnianej rozróżnia się modele:

- uporządkowane (*ordered logit model*), w których warianty zmiennej objaśnianej zachowują naturalny porządek,
- nieuporządkowane (*multinomial* lub *polynomial logit models*, czasem *unordered logit model*), w których nie można ustalić logicznej hierarchii poszczególnych wariantów zmiennej objaśnianej[36].

Liniowy model prawdopodobieństwa, w którym wszystkie zmienne są dychotomiczne, nosi nazwę modelu regresji binarnej (*binary regression*). W modelach regresji binarnej oceny MNK dają wartości teoretyczne z przedziału (0; 1), a interpretacje ocen estymatorów parametrów tego modelu są takie, jak w modelu logitowym [Gruszczyński 2010, s. 84–85].

6.6. Modele szeregów czasowych

W modelach regresji i w wywodzących się od nich modelach ekonometrycznych przyjmuje się, że zjawiska opisywane poszczególnymi równaniami modelu zależą od pewnych czynników natury zewnętrznej. Alternatywą dla nich są prostsze w konstrukcji modele szeregów czasowych. W przypadku tego typu modeli zakłada się, że zjawiska ekonomiczne, będące procesami stochastycznymi, są generowane przez siebie same, a więc na podstawie analiz opóźnień w przebiegu procesów gospodarczych można wnioskować o ich przebiegu w przyszłości. Podstawowym

36 Modele wielomianowe omawiane są m.in. w [Gruszczyński 2010, s. 103–192].

problemem w stosowaniu modeli szeregów czasowych jest dostępność dostatecznie dużej[37] próby statystycznej.

Przy modelowaniu długich szeregów, w szczególności tych o dużej częstotliwości pomiaru, pojawia się problem stacjonarności procesów[38], których obserwacje tworzą analizowane szeregi czasowe. Procesy stacjonarne charakteryzują się tym, że mają stałą wariancję, a ich wartości w poszczególnych momentach oscylują wokół pewnego, względnie stałego poziomu, który jest poziomem średnim dla całego badanego okresu. W przypadku występowania bardziej gwałtownych zmian procesy stacjonarne potrafią się ustabilizować na nowym poziomie. Do opisu tego typu procesów szczególnie przydatny okazuje się model autoregresji AR (*autoregressive model*). W modelu tym bieżąca wartość zmiennej wyrażona jest przez skończoną kombinację jej wartości poprzednich. Modele tego typu wywodzą się z szerszej klasy modeli regresji, szacowanych na podstawie szeregów czasowych, czyli[39]:

$$y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \alpha_2 \cdot y_{t-2} + \ldots + \alpha_k \cdot y_{t-r} + \varepsilon_t \tag{6.22}$$

Model (6.22) jest modelem autoregresyjnym r-tego rzędu – AR(r), ponieważ maksymalne opóźnienie zmiennej wynosi r. Jeśliby, zgodnie ze wzorem (6.22), zapisać każdą z opóźnionych wartości za pomocą tej relacji, to otrzyma się:

$$y_{t-1} = \alpha_0 + \alpha_1 \cdot y_{t-2} + \alpha_2 \cdot y_{t-3} + \ldots + \alpha_k \cdot y_{t-r-1} + \varepsilon_{t-1} \tag{6.23a}$$

$$y_{t-2} = \alpha_0 + \alpha_1 \cdot y_{t-3} + \alpha_2 \cdot y_{t-4} + \ldots + \alpha_k \cdot y_{t-r-2} + \varepsilon_{t-2} \tag{6.23b}$$

$$\ldots$$

$$y_{t-r} = \alpha_0 + \alpha_1 \cdot y_{t-r-1} + \alpha_2 \cdot y_{t-r-2} + \ldots + \alpha_k \cdot y_{t-2r} + \varepsilon_{t-r} \tag{6.23c}$$

Podstawiając kolejne relacje (6.23) w miejsce zmiennych $y_{t-\tau}$ (dla $\tau = 1, 2, \ldots, r, \ldots$), zauważyć można, że proces autoregresji wyraża odchylenia y_t jako nieskończoną sumę ważoną impulsów ε_t w różnych okresach. Jeżeli tę sumę ograniczyć do kilku pierwszych składników, to otrzyma się – mający duże znaczenie praktyczne – tzw. proces średniej ruchomej MA (*moving average model*). W modelu średniej ruchomej zakłada się, że y_t zależy liniowo od skończonej liczby q poprzednich wartości ε_t.

---

37 Wprawdzie Box i Jenkins [1983] określili minimalną liczbę obserwacji na 50, jednak w praktyce zaleca się próby przynajmniej 100-elementowe.

38 Zagadnienia stacjonarności procesów omawiane są m.in. w pracach: [Box, Jenkins 1983, s. 18, 34–38; Chow 1995, s. 234–236; Charemza, Deadman 1997, s. 103–113; Osińska 2006, s. 47–49].

39 W wielu pracach (por. m.in. [Box, Jenkins 1983; Witkowska 2003; Witkowska i in. 2012]) model AR(r) formalnie zapisuje się w postaci: $\tilde{y}_t = \alpha_0 + \alpha_1 \cdot \tilde{y}_{t-1} + \alpha_2 \cdot \tilde{y}_{t-2} + \ldots + \alpha_k \cdot \tilde{y}_{t-r} + \varepsilon_t$, gdzie: $\tilde{y}_t = y_t - \mu$ jest odchyleniem obserwacji stacjonarnego procesu stochastycznego od średniej ciągu obserwacji μ.

Przykładowo, dla procesu AR(1) będzie:

$$y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \varepsilon_t = \alpha_0 + \alpha_1 \cdot (\alpha_0 + \alpha_1 \cdot y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \\ = \gamma_0 + \alpha_1 \cdot \varepsilon_{t-1} + \alpha_1^2 \cdot \varepsilon_{t-2} + \ldots + \alpha_1^q \cdot \varepsilon_{t-q} + \varepsilon_{t-q} \tag{6.24}$$

gdzie:
γ_0 – wyraz wolny procesu utworzony poprzez zsumowanie kolejnych wyrazów wolnych. Zatem dla procesu AR(*r*) parametry stojące przy zmiennej ε_t będą w postaci wielomianów.

W celu osiągnięcia większej elastyczności w dopasowaniu modelu do rzeczywistego szeregu czasowego uzasadnione jest włączenie do modelu zarówno członów autoregresji AR*(p)*, jak i średniej ruchomej MA(*q*). Takie ujęcie prowadzi do mieszanego modelu autoregresji i średniej ruchomej (*autoregressive moving average model*) ARMA(*p*, *q*), którego używa się, podobnie jak modeli autoregresyjnych i średniej ruchomej, do opisu stacjonarnych szeregów czasowych.

Jednakże wiele zjawisk gospodarczych, a szczególnie te odwzorowane w finansowych szeregach czasowych, ma charakter niestacjonarny, co oznacza, że nie można wyróżnić w ich przebiegu wartości średniej, wokół której zjawisko oscyluje. Dobrym tego przykładem jest kształtowanie się kursów akcji na giełdzie. Niemniej jednak często własności szeregów, będących realizacjami procesów niestacjonarnych, są w pewien sposób jednorodne. Może się okazać, że mimo iż poziom, względem którego zachodzą wahania, zmienia się w czasie, to ogólne zachowanie się szeregu, po uwzględnieniu różnic w poziomie, będzie podobne (por. [Chow 1995])[40]. Korzysta się wtedy np. z różnicowego przekształcenia szeregu, którego celem jest osiągnięcie stacjonarności. Innymi słowy, po *d*-krotnym różnicowaniu szeregu niestacjonarnego powstaje szereg stacjonarny, do którego można zastosować model ARMA. Model ten nazywany jest scałkowanym procesem autoregresji i średniej ruchomej (*autoregressive integrated moving average model*) rzędu (*p*, *d*, *q*), w skrócie ARIMA(*p*, *d*, *q*).

Rozwój technik komputerowych spowodował, że coraz częściej do analizy rynków finansowych wykorzystywane są modele nieliniowe. Do najpopularniejszych należą modele ARCH i ich uogólniona wersja GARCH wraz z ich licznymi modyfikacjami, np. SV, TAR, STAR, SETAR, modele przełącznikowe, modele ze zmiennymi parametrami oraz sztuczne sieci neuronowe[41]. Jednym z najpopularniejszych modeli nieliniowych, stosowanych w analizach finansowych szeregów

40 W pracy [Charemza, Deadman 1997] przedstawiono przykłady procesu błądzenia przypadkowego i błądzenia przypadkowego z dryfem (*random walk with drift*), będących procesami niestacjonarnymi.

41 Jedną z pierwszych na polskim rynku pozycji z zakresu aplikacji sztucznych sieci neuronowych w modelowaniu finansów była monografia [Witkowska 2002].

czasowych o dużej częstotliwości pomiaru, jest model opisujący zmiany zachodzące w wariancji warunkowej, tzw. model ARCH (*autoregressive conditional heteroscedastic model*), zaproponowany przez Roberta Engle'a[42].

Model ARCH można zapisać w postaci dwóch równań opisujących warunkową średnią (6.25) i warunkową wariancję (6.26):

$$y_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \alpha_2 \cdot x_{2t} + \ldots + \alpha_k \cdot x_{kt} + \varepsilon_t \tag{6.25}$$

$$\varepsilon_t = \vartheta_t \sqrt{h_t} \ ; \ h_t = f(\varepsilon_t^2) \tag{6.26}$$

gdzie:
$x_{1i}, x_{2i}, \ldots, x_{ki}$ to zmienne objaśniające w modelu (6.20),
$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$ – parametry strukturalne modelu,
ε_i – składnik losowy,
h_t – funkcja wariancji warunkowej,
ϑ – szereg zakłóceń o rozkładzie normalnym N(0; 1),
f – symbol funkcji.

Funkcja (6.25) przyjmuje zwykle postać modelu ARMA lub modelu autoregresyjnego z rozłożonymi opóźnieniami (*autoregressive distributed lags model*) ARDL. Stosowanie modelu ARCH do opisu szeregu wymaga potwierdzenia istnienia tak zwanego efektu ARCH z wykorzystaniem odpowiednich testów.

Przykład 6.5

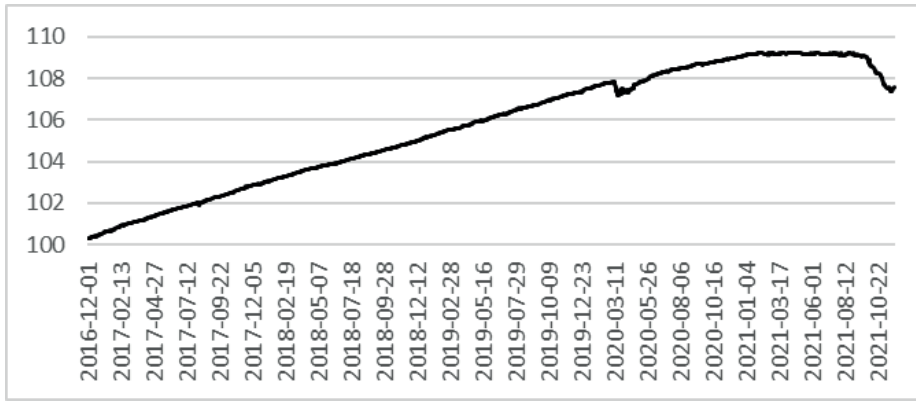
Analizie poddano notowania dwóch funduszy inwestycyjnych, zarządzanych przez Towarzystwo Funduszy Inwestycyjnych (TFI) Aviva Investors Poland, realizujących różną politykę inwestycyjną. Pierwszy z nich, Fundusz Stabilnego Dochodu, jest kierowany do inwestorów o niskiej skłonności do ryzyka. Drugi, Fundusz Polskich Akcji, przeznaczony jest dla inwestorów akceptujących nawet znaczne ryzyko. Badania dotyczą dziennych notowań jednostek udziałowych tych funduszy w okresie 1.12.2016–1.12.2021 (1250 notowań). Wykresy notowań obu funduszy przedstawiono na rys. 6.4 i 6.5.

Jak widać, notowania obu funduszy różnią się nie tylko wartością jednostek udziałowych, ale i dynamiką zmian ich wartości. Opisane zatem zostaną notowania obu funduszy za pomocą 5 modeli autoregresji i funkcji trendu:

- AR(1), AR(5) i AR(15), w którym notowanie bieżące zależy od notowania poprzedniego oraz opóźnionego o 5 i 15 okresów;
- AR(5), w którym notowanie bieżące zależy od 5 poprzednich notowań;

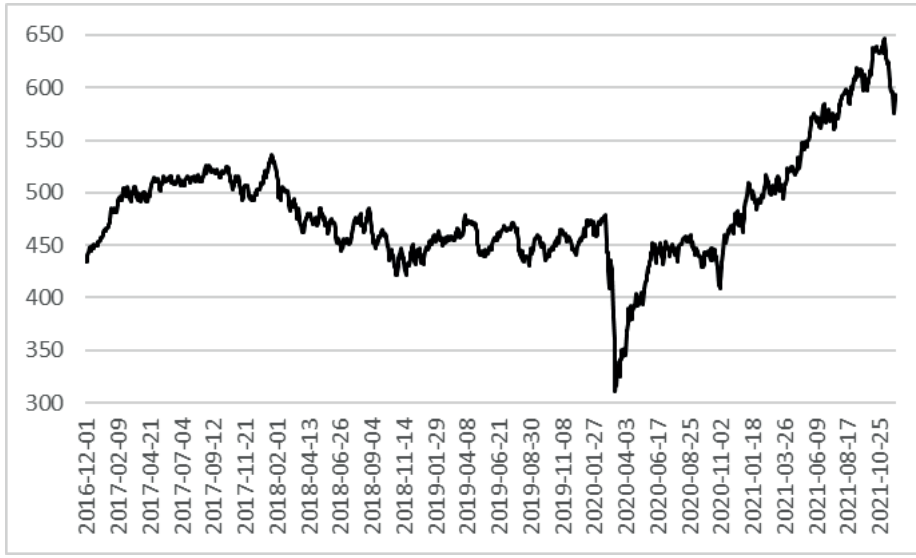
42 Szczegółowe omówienie tych zagadnień znaleźć można m.in. w [Osińska 2006, s. 79–116], Natomiast rozwinięciu tej problematyki i aplikacjom tej klasy modeli w finansach poświęcono m.in. opracowania: [Fiszeder 2009; Górka 2012].

- AR(15), w którym notowanie bieżące zależy od kilku wybranych poprzednich notowań (co piąte), czyli (w przybliżeniu) notowania dotyczące tego samego dnia tygodnia oraz
- modelu trendu.



Rysunek 6.4. Dzienne notowania jednostek uczestnictwa Funduszu Stabilnego Dochodu

Źródło: opracowanie własne na podstawie danych z Aviva Investors Poland, https://pl.investing.com/equities/aviva-historical-data (dostęp 15.12.2001).



Rysunek 6.5. Dzienne notowania jednostek uczestnictwa Funduszu Polskich Akcji

Źródło: opracowanie własne na podstawie danych z Aviva Investors Poland, https://pl.investing.com/equities/aviva-historical-data (dostęp 15.12.2001).

Rozwiązanie

Modele autoregresji rzędu r, dla $r = 1, 5$ i 15, są zgodnie ze wzorem (6.22) postaci:

AR(1): $y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \varepsilon_t$
AR(2): $y_t = \alpha_0 + \alpha_2 \cdot y_{t-5} + \varepsilon_t$
AR(5): $y_t = \alpha_0 + \alpha_2 \cdot y_{t-15} + \varepsilon_t$
AR(5): $y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \alpha_2 \cdot y_{t-2} + \alpha_3 \cdot y_{t-3} + \alpha_4 \cdot y_{t-4} + \alpha_5 \cdot y_{t-5} + \varepsilon_t$
AR(15): $y_t = \alpha_0 + \alpha_5 \cdot y_{t-5} + \alpha_{10} \cdot y_{t-10} + \alpha_{15} \cdot y_{t-15} + \varepsilon_t$

Tabela 6.10. Oszacowania modeli autoregresyjnych

Parametry modeli autoregresyjnych z jedną zmienną objaśniającą						
Opóźnienia zmiennych	$t-1$	const	$t-5$	const	$t-5$	const
Fundusz Stabilnego Dochodu	0,9986	0,1521	0,9930	0,7683	0,9784	2,3677
Współczynnik determinacji R^2	0,9999		0,9995		0,9972	
Fundusz Polskich Akcji	**0,9973**	1,4327	**0,9830**	**8,7917**	0,9516	25,2246
Współczynnik determinacji R^2	0,9919		0,9540		0,8533	
Parametry modeli autoregresyjnych z kilkoma zmiennymi objaśniającymi						
Opóźnienia zmiennych	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$	const
Fundusz Stabilnego Dochodu	–0,1440	0,0278	0,0393	**0,1677**	**0,9082**	**0,1177**
Współczynnik determinacji R^2	0,9999					
Fundusz Polskich Akcji	–0,0294	–0,0072	-0,0012	0,0083	**1,0262**	1,6923
Współczynnik determinacji R^2	0,9919					
Opóźnienia zmiennych	$t-15$	$t-10$	$t-5$	const		
Fundusz Stabilnego Dochodu	0,1172	–0,4214	1,2982	0,6540		
Współczynnik determinacji R^2	0,9995					
Fundusz Polskich Akcji	**–0,0597**	0,0023	**1,0366**	**10,5446**		
Współczynnik determinacji R^2	0,9542					
Model trendu						
Parametry	t	const	Współczynnik determinacji R^2			
Fundusz Stabilnego Dochodu	**0,0075**	**101,0109**	0,9558			
Fundusz Polskich Akcji	**0,0323**	**463,0647**	0,0487			

Uwaga: pogubione wartości parametrów świadczą o istotności zmiennej (na poziomie $\alpha = 0,05$, przy której stoją)

Źródło: obliczenia własne.

Po oszacowaniu MNK parametrów tych modeli[43] dla notowań obu analizowanych funduszy inwestycyjnych otrzymano wyniki przedstawione w tab. 6.10.

43 Modele zostały oszacowane przy wykorzystaniu funkcji REGLINP w Excelu.

Jak widać, modele autoregresyjne bardzo dobrze opisują notowania jednostek udziałowych obu badanych funduszy inwestycyjnych. Podane wartości współczynników determinacji R^2 można ze sobą bezpośrednio porównywać, ponieważ duża liczba obserwacji i mała liczba szacowanych parametrów sprawia, że liczba stopni swobody we wszystkich modelach waha się od 1239 do 1248.

W przypadku Funduszu Stabilnego dochodu zmienność notowań jest niewielka (rys. 6.4) i obserwowany jest nieznaczny trend wzrostowy, zatem model trendu liniowego równie dobrze opisuje zmiany cen jednostek udziałowych tego funduszu. Odnotować również można, że niemal wszystkie zmienne uwzględnione w modelach autoregresji są statystycznie istotne.

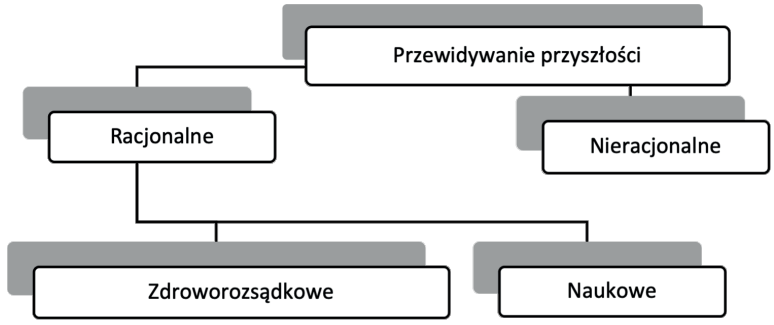
W przeciwieństwie do notowań Funduszu Stabilnego Dochodu notowania Funduszu Polskich Akcji zmieniają się dynamicznie (rys. 6.5) i dlatego model trendu całkowicie się nie sprawdził (R^2 poniżej 5%). Zarazem jednak modele autoregresyjne poprawnie opisują kształtowanie się cen jednostek tego funduszu, chociaż w tych zawierających więcej niż jedną zmienną opóźnioną część jest statystycznie nieistotna.

Rozdział 7

Prognozowanie na podstawie szeregów czasowych

Przewidywanie w szerszym rozumieniu oznacza wnioskowanie o nieznanych zjawiskach i zdarzeniach odnoszących się zarówno do przeszłości czy teraźniejszości, jak i przyszłości. Jeśli ograniczyć się do przewidywania na podstawie modelu matematycznego, to przewidywanie oznacza ekstrapolację modelu poza obszar próby estymacyjnej. Może być zatem realizowane zarówno w przypadku modeli oszacowanych na podstawie szeregów czasowych, jak i danych przekrojowych. W pierwszym przypadku mówi się o prognozowaniu przyszłości, a w drugim dokonuje się oceny modelowanego zjawiska, np. przyporządkowując (zaliczając) je do jakiejś klasy[1]. W niniejszym podrozdziale omawiane będą jedynie zagadnienia związane z prognozowaniem przyszłych zdarzeń na podstawie szeregów czasowych.

Przewidywanie przyszłości można realizować na wiele sposobów, które zazwyczaj dzieli się na dwie grupy, tj. przewidywanie racjonalne i nieracjonalne (irracjonalne), a wśród przewidywań racjonalnych wyróżnia się przewidywanie zdroworozsądkowe i naukowe, co przedstawiono na rys. 7.1.



Rysunek 7.1. Sposoby przewidywania przyszłości

Źródło: opracowanie własne.

1 W literaturze anglojęzycznej używa się terminu *prediction* dla obu rodzajów modeli. Zagadnienia związane z klasyfikacją i grupowaniem omówione zostaną w rozdziale 8.

7.1. Podstawowe pojęcia

Prognozowanie[2] to rodzaj wnioskowania, którego celem jest:

- przewidywanie prawdopodobnego przebiegu przyszłych zjawisk i procesów zarówno pod względem ilościowym, jak i jakościowym;
- wspomaganie procesu podejmowania decyzji poprzez uwzględnianie lub pobudzanie pożądanych zmian lub zapobieganie zjawiskom niepożądanym.

Prognozowanie to racjonalne naukowe przewidywanie przyszłych zdarzeń (por. [Cieślak 1997, s. 18]). Innymi słowy, prognoza będąca wynikiem procesu prognozowania, to osąd o zajściu określonego zdarzenia w czasie określonym z dokładnością do momentu (punktu na osi czasu) lub okresu (przedziału na osi czasu), należącego do przyszłości. Przyjmuje się, że prognozy formułuje się z wykorzystaniem dorobku nauki i są one weryfikowalne empirycznie[3]. Ze względu na sposób określania zdarzenia, prognozy dzieli się na:

- ilościowe, kiedy prognoza jest wyrażona liczbą, np. w przyszłym roku spodziewany jest 4% przyrost PKB;
- jakościowe, gdy w wyniku procesu prognozowania otrzymuje się słowny opis sytuacji, np. w przyszłym tygodniu spodziewane jest obniżenie stóp procentowych przez Radę Polityki Pieniężnej.

Przedmiotem dalszych rozważań będą jedynie prognozy ilościowe, wśród których wyróżnia się prognozy punktowe i przedziałowe. W pierwszym przypadku prognoza jest pojedynczą wartością liczbową, a w tym drugim buduje się przedział ufności dla zmiennej prognozowanej. W praktyce częściej korzysta się z prognoz punktowych, ponieważ prognozy przedziałowe wymagają znajomości rozkładu zmiennej prognozowanej, czym zazwyczaj się nie dysponuje.

Każda prognoza spełnia następujące funkcje (por. [Cieślak 1997, s. 20–23]):

- preparacyjną – według której prognoza jest działaniem przygotowującym inne działania;
- informacyjną – polegająca na oswajaniu ludzi z nadchodzącymi zmianami i zmniejszaniu lęku przed przyszłością;
- aktywizującą – polegająca na pobudzaniu do podejmowania działań sprzyjających realizacji prognozy, gdy zapowiada ona zdarzenia korzystne oraz przeciwstawiających się jej realizacji, kiedy przewidywane zdarzenia są oceniane negatywnie.

Istotnym pojęciem przy wyznaczaniu prognoz jest horyzont prognozy, czyli maksymalna liczba okresów lub momentów objętych prognozą. Z uwagi na to,

2 W niniejszym podrozdziale zagadnienia związane z prognozowaniem potraktowane zostały bardzo skrótowo. Szeroką dyskusję nt. metod prognozowania można znaleźć m.in. w pracach [Cieślak 1997; Witkowska 2012].

3 Wynik procesu prognozowania jest niepewny, ale akceptowalny.

że zjawiska społeczno-gospodarcze podlegają zmianom w czasie, wyróżnia się prognozy:

- krótkoterminowe, kiedy dotyczą okresu, w którym w prognozowanym zjawisku zachodzą tylko zmiany ilościowe,
- średniookresowe, kiedy w okresie objętym prognozowaniem oczekuje się, że oprócz dominujących zmian ilościowych wystąpią również zmiany jakościowe,
- długookresowe, odnoszące się do okresu, w którym w prognozowanym zjawisku wystąpią istotne zmiany jakościowe.

Klasyfikacja prognoz dokonywana jest z punktu widzenia różnych kryteriów podziału, takich jak (por. [Zeliaś, Pawełek, Wanat 2003, s. 18–19]):

1) horyzont prognozy,
2) charakter lub struktura prognozy,
3) stopień szczegółowości i zakres ujęcia prognozy,
4) zasięg terenowy (przestrzenny) prognozy,
5) metoda opracowania,
6) cel lub funkcja prognozy.

Pierwsze, bodaj najważniejsze kryterium, według którego oprócz prognoz krótko-, średnio- i długoterminowych, wskazuje się na prognozy operacyjne, wykorzystywane w planowaniu bieżącej działalności (są to zatem najczęściej prognozy krótkookresowe), i strategiczne, które dostarczają przesłanek do podejmowania długofalowych decyzji (są to zatem najczęściej prognozy długookresowe). Drugie z wymienionych kryteriów, oprócz prognoz ilościowych i jakościowych, pozwala wyróżnić prognozy:

- jednorazowe, jeśli wyznaczane są jednokrotnie, oraz powtarzalne, kiedy są wyznaczane wielokrotnie wraz z napływem nowych (uaktualnionych) informacji;
- proste, jeśli dotyczą pojedynczej zmiennej, i złożone, kiedy odnoszą się do złożonego zjawiska gospodarczego;
- kompleksowe, które całościowo opisują przyszłość złożonego zjawiska, oraz sekwencyjne, odnoszące się do kilku okresów;
- samosprawdzające się, czyli takie, których ogłoszenie sprzyja realizacji tego przewidywania, oraz destruktywne, których ogłoszenie zmniejsza prawdopodobieństwo realizacji przewidywanych zdarzeń.

Kolejne kryterium prowadzi do wyróżnienia prognoz ogólnych (całościowych, globalnych), dotyczących całego zjawiska oraz prognoz szczegółowych (częściowych, odcinkowych), odnoszących się do pewnego aspektu analizowanego zjawiska. Biorąc pod uwagę metodę opracowania prognoz, można podzielić je na:

- minimalne, średnie i maksymalne, jeśli zostały przyjęte na poziomie odpowiednio: najmniejszym, średnim i największym z wyznaczonych,
- czyste, tj. uzyskane w wyniku ekstrapolacji trendu, weryfikowane (powtarzalne) i modelowe, tj. opracowane na podstawie modelu,
- prognozy wynikające z zastosowanych reguł prognozowania (np. nieobciążone, według największego prawdopodobieństwa).

Ostatnie z kryteriów pozwala wyróżnić prognozy ostrzegawcze, które zwracają uwagę na niekorzystne kształtowanie się zjawisk, a także prognozy aktywne (pobudzające do działania) i pasywne (zniechęcające do działania).

Wyznaczenie prognoz jest ściśle związane z celem i wybraną metodą prognozowania. Ze względu na rodzaj posiadanej informacji wyróżnia się prognozy:

- *ex post*, w przypadku których wartości zmiennych objaśniających są znane, a prognoza może być porównana z wartościami zaobserwowanymi,
- *ex ante*, w przypadku których dostępność danych, dotyczących wartości zmiennych (objaśniających) niezbędnych do wyznaczenia prognozy, zależy od długości i struktury występujących opóźnień oraz charakteru zjawisk. Dane te są zazwyczaj nieznane, a wartości zmiennych objaśniających są wyznaczane z pewnym prawdopodobieństwem. Dlatego tak otrzymana prognoza nosi nazwę warunkowej (względem zmiennych objaśniających).

Prognozy wyznaczane na okresy, dla których znane są wartości zmiennej prognozowanej, nazywane są prognozami wygasłymi. Wyznacza się je głównie w celu oceny poprawności metody prognozowania, którą będzie można zastosować do budowy prognoz *ex ante*.

Prognozowanie jest procesem, na który składa się kilka etapów:

1) sformułowanie zadania prognostycznego,
2) określenie przesłanek prognostycznych,
3) zebranie i analiza danych prognostycznych,
4) wybór metody prognozowania,
5) konstrukcja prognozy,
6) ocena dokładności prognozy.

Sformułowanie zadania prognostycznego polega na sformułowaniu celu budowy prognozy, określeniu prognozowanego zjawiska i ustaleniu horyzontu prognozy oraz sprecyzowaniu wymagań dopuszczalności prognoz. Formułując przesłanki prognostyczne opisuje się mechanizm generujący prognozowaną zmienną oraz listę czynników, które na ten proces wpływają. W trzecim etapie następuje zbieranie danych potrzebnych do sformułowania modelu i konstrukcji prognozy.

Metoda prognozowania obejmuje sposób przetworzenia danych używanych do opisu zmiennej prognozowanej oraz ich ekstrapolacji poza próbę. Wybór metody prognozowania zależy od zdefiniowanego problemu prognostycznego i rodzaju zebranych danych. Metody prognozowania można podzielić na subiektywne i sformalizowane, a należą do nich:

- prognozowanie na podstawie modelu formalnego,
- metody heurystyczne, np. burza mózgów, metoda delficka, wpływów krzyżowych itp.,
- prognozowanie analogowe, które polega na przewidywaniu przyszłości zmiennej prognozowanej przez wykorzystanie informacji o innych zmiennych, o podobnych równoczesnych zmianach w czasie (jednakże zależności

między zmiennymi są zbyt słabe, aby można było przypuszczać, że są one powiązane przyczynowo ze zmienną prognozowaną),
- prognozowanie scenariuszowe, polegające na opracowaniu różnych scenariuszy dotyczących rozwoju zjawisk wpływających na zmienną prognozowaną.

Ocenę dokładności prognoz przeprowadza się na podstawie wyznaczonych błędów prognoz.

W teorii prognozy przyjmuje się szereg założeń, będących warunkami wstępnymi umożliwiającymi wnioskowanie o przyszłości w uzasadniony sposób[4].

1. Znane są oszacowania parametrów modelu, w którym zmienna prognozowana pełni rolę zmiennej objaśnianej.
2. Model jest stabilny w czasie (tzn. nie ulega zmianom jego postać analityczna i wartości parametrów strukturalnych są stałe w czasie, na jaki wyznaczane są prognozy) bądź też następujące zmiany są powolne i na tyle regularne, że można je z dostateczną dokładnością przewidzieć.
3. Rozkład składnika losowego modelu jest stacjonarny bądź zmiany tego rozkładu są powolne i na tyle regularne, że umożliwiają odpowiednio dokładne ekstrapolacje.
4. Znane są wartości zmiennych objaśniających w okresie prognostycznym.
5. Dopuszczalna jest ekstrapolacja modelu poza obszar zmienności zmiennych objaśniających zaobserwowany w próbie, na podstawie której oszacowano model.

7.2. Sformalizowane metody prognozowania

Do sformalizowanych metod prognozowania można zaliczyć zarówno metody proste, jak i dość skomplikowane modele ekonometryczne, w których zmienna prognozowana jest zmienną objaśnianą. Modele wykorzystywane do prognozowania zazwyczaj dzieli się na dwie grupy, tj. metody bezpośrednie i pośrednie. W przypadku tych pierwszych korzysta się jedynie z informacji o dotychczasowym kształtowaniu się zmiennej prognozowanej. Do grupy metod bezpośrednich zalicza się wszystkie metody prognozowania na podstawie szeregów czasowych, wśród których wymienić można:

1) metody naiwne,
2) metody średniej ruchomej (5.19),
3) metody wygładzania wykładniczego (5.33)–(5.34),
4) modele tendencji rozwojowej (5.16)–(5.18),
5) modele szeregów czasowych, tj. modele AR (6.22) oraz ARMA, ARIMA.

4 Ich omówienie znaleźć można m.in. w pracy [Witkowska 2012, s. 182].

Wśród modeli naiwnych wyróżnia się:

- model błądzenia przypadkowego postaci:

$$y_t^* = y_{t-1} \tag{7.1}$$

- model dla szeregu czasowego z trendem:

$$y_t^* = y_{t-1} + (y_{t-1} - y_{t-2}) \tag{7.2}$$

- model dla założonych zmian (tj. wzrostu lub spadku) zjawiska w czasie:

$$y_t^* = (1 + c) \cdot y_{t-1} \tag{7.3}$$

gdzie:
y_t^* – prognoza (wartość) zmiennej prognozowanej w okresie t,
y_{t-1}, y_{t-2} – wartości zmiennej prognozowanej w okresach poprzednich w stosunku do okresu, na jaki wyznacza się prognozę, tj. odpowiednio $t - 1$, $t - 2$,
c – wskaźnik stałego wzrostu (spadku) z okresu na okres.

Jak łatwo zauważyć, korzystając z metod naiwnych (7.1)–(7.3), bez problemu wyznaczyć można prognozę na jeden okres do przodu, ponieważ do jej wyznaczenia zawsze potrzebna jest obserwacja poprzedzająca okres prognozowany t. Zatem jeśli horyzont prognozy jest dłuższy niż jeden okres, to korzystając z tych wzorów, należy prognozy wyznaczać okres po okresie, każdorazowo dokonując bieżącej obserwacji wartości zmiennej prognozowanej. Przykładowo, jeśli prognozy mają zostać zbudowane w marcu na kilka kolejnych miesięcy (mając dane od stycznia do marca), to bez trudu wyznaczyć można prognozę na kwiecień na podstawie danych z marca według wzorów (7.1) i (7.3) oraz danych z lutego i marca według wzoru (7.2). Natomiast, aby wyznaczyć prognozę na maj, trzeba odczekać do kwietnia, kiedy dostępne będą obserwacje z tego miesiąca, i w kwietniu wyznaczyć prognozę na maj na podstawie danych dotyczących rzeczywistej wartości zmiennej prognozowanej w kwietniu. Prognozy na dalsze miesiące wyznacza się w podobny sposób, każdorazowo odczekując miesiąc i obliczając wartość zmiennej prognozowanej na następny miesiąc. W podobny sposób wyznacza się prognozy za pomocą średnich ruchomych oraz modeli, w których występują zmienne opóźnione o jeden okres, np. modele AR(1).

Jednakże takie wyznaczanie prognoz nie jest wygodne, a w wielu przypadkach jest niesatysfakcjonujące, kiedy istnieje potrzeba wyznaczenia całego ciągu prognoz lub w sytuacji, kiedy w momencie budowy prognoz nie są dostępne dane dotyczące okresu bezpośrednio poprzedzającego pierwszy okres, na jaki się prognozuje[5]. Wówczas prognozy wyznacza się rekurencyjnie, korzystając z prognoz obliczonych dla

5 Ta ostatnia przyczyna związana jest z opóźnieniem obserwacji statystycznej, która – w szczególności w szeregach czasowych dużej częstotliwości – sprawia, że informacje na temat zjawiska w okresie t pojawiają się dopiero w okresach $t + 1$ lub później, np. dane o sumie transakcji finansowych w danym miesiącu będą znane nie wcześniej jak pierwszego kolejnego miesiąca.

wcześniejszych okresów. Zatem prognozy dla pierwszego okresu prognostycznego wyznacza się według wzorów (7.1)–(7.2), a na kolejne okresy, korzystając ze wzorów:

- model (7.1): $y^*_{t+1} = y^*_t$ dla okresu $t + 1$,
  $y^*_{t+2} = y^*_{t+1}$ dla okresu $t + 2$ itd.;
- model (7.2): $y^*_{t+1} = y^*_t + (y^*_t - y_{t-1})$ dla okresu $t + 1$,
  $y^*_{t+2} = y^*_{t+1} + (y^*_{t+1} - y^*_t)$ dla okresu $t + 2$ itd.;
- model (7.3): $y^*_{t+1} = (1 + c) \cdot y^*_t$ dla okresu $t + 1$,
  $y^*_{t+2} = (1 + c) \cdot y^*_{t+1}$ dla okresu $t + 2$ itd.

Przykład 7.1

Na podstawie danych dotyczących działalności finansowej i ubezpieczeniowej, przedstawionych w przykładzie 5.4, należy wyznaczyć prognozy *ex post* dla okresu obserwacji oraz prognozy *ex ante* na kolejne 4 kwartały, tj. na okresy 2019Q3–2020Q2, stosując metody naiwne (7.1)–(7.3), które oznacza się symbolami odpowiednio: N1, N2 i N3.

Rozwiązanie

W tab. 7.1 zawarto wszystkie niezbędne dane, obliczenia pomocnicze oraz wyznaczone prognozy. W przypadku modelu stałych zmian w czasie (7.3) wartość parametru c określono na poziomie średniej geometrycznej (5.8) opisującej przeciętne zmiany zjawiska z okresu na okres, obliczonej dla szeregu „historycznych obserwacji" zmiennej prognozowanej, tj. dla danych 2014Q1–2019Q2. Indeks jednopodstawowy (5.5), wyznaczony dla ostatniego okresu 2019Q2 w stosunku do pierwszego badanego kwartału 2014Q1, wynosi 1,3181, a pierwiastek 21 stopnia[6] z tej liczby równa się 1,0132. Oznacza to, że zakładane jednakowe przyrosty wartości działalności finansowej i ubezpieczeniowej z okresu na okres wyniosły 1,32% kwartalnie, stąd $c = 0{,}0132$.

Prognozy wyznaczone dla kwartałów 2014Q1–2019Q2 są prognozami *ex post*, natomiast te obliczone na następne 4 kwartały prognozami *ex ante*. Prognozy na trzeci kwartał 2019 roku zostały wyznaczone na podstawie danych rzeczywistych pochodzących z okresu 2019Q2, natomiast na kolejne okresy z wykorzystaniem wyznaczonych prognoz na okres 2019Q3 i dalsze.

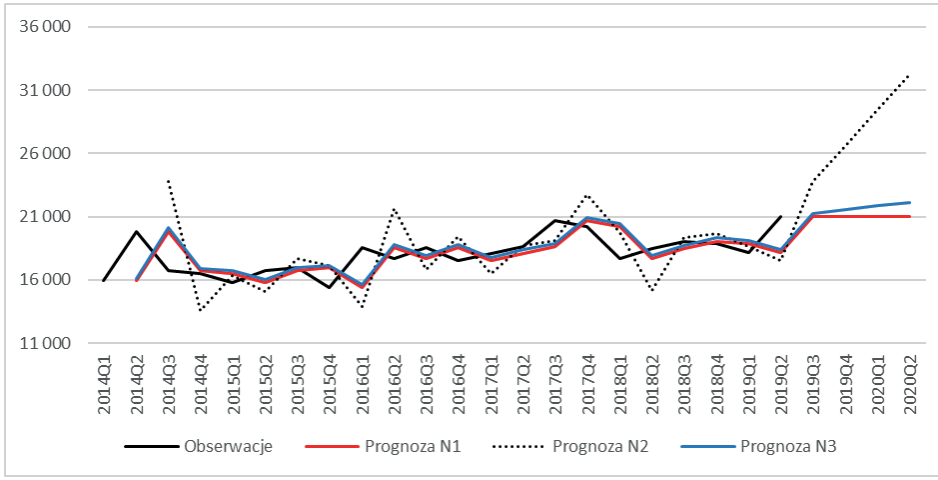
Analizując prognozy wyznaczone za pomocą modelu błądzenia losowego (kolumna 3, tab. 7.1) zauważa się, że wszystkie prognozy *ex ante* są identyczne, ponieważ „powielają" ostatnią obserwację. Z kolei prognozy wyznaczone na podstawie (7.2) zmieniają się z okresu na okres dokładnie o tę sama wartość, będącą różnicą między dwiema ostatnimi obserwacjami (kolumna 4). Natomiast prognoza wyznaczona na podstawie modelu (7.3) wskazuje systematyczny wzrost i poczynając od obserwacji 2029Q4, jest obliczana poprzez wykorzystanie (wcześniej) wyznaczonych prognoz dotyczących poprzednich kwartałów.

6 Stopień pierwiastka wynika z liczby indeksów łańcuchowych, których dla 22 obserwacji można wyznaczyć 21.

Tabela 7.1. Dane do przykładu 7.1 i wyznaczone prognozy

Okresy	Obserwacje	Prognozy			
		N1	N2		N3
		$y_t^* = y_{t-1}$	$y_t^* = y_{t-1} + (y_{t-1} - y_{t-2})$		$y_t^* = (1 + c) \cdot y_{t-1}$
			$y_{t-1} - y_{t-2}$	$y_t^* = y_{t-1} + (y_{t-1} - y_{t-2})$	
2014Q1	15 933,4				
2014Q2	19 852,0	15 933,4	×	×	16 144,36
2014Q3	16 717,3	19 852,0	3 918,6	23 770,6	20 114,84
2014Q4	16 540,5	16 717,3	–3 134,7	13 582,6	16 938,64
2015Q1	15 801,9	16 540,5	–176,8	16 363,7	16 759,50
2015Q2	16 747,5	15 801,9	–738,6	15 063,3	16 011,12
2015Q3	16 945,0	16 747,5	945,6	17 693,1	16 969,24
2015Q4	15 416,4	16 945,0	197,5	17 142,5	17 169,35
2016Q1	18 528,9	15 416,4	–1 528,6	13 887,8	15 620,51
2016Q2	17 694,1	18 528,9	3 112,5	21 641,4	18 774,22
2016Q3	18 552,1	17 694,1	–834,8	16 859,3	17 928,37
2016Q4	17 537,8	18 552,1	858,0	19 410,1	18 797,73
2017Q1	18 127,7	17 537,8	–1 014,3	16 523,5	17 770,00
2017Q2	18 624,7	18 127,7	589,9	18 717,6	18 367,71
2017Q3	20 684,5	18 624,7	497,0	19 121,7	18 871,29
2017Q4	20 218,5	20 684,5	2 059,8	22 744,3	20 958,36
2018Q1	17 685,0	20 218,5	–466,0	19 752,5	20 486,19
2018Q2	18 504,6	17 685,0	–2 533,5	15 151,5	17 919,15
2018Q3	19 068,6	18 504,6	819,6	19 324,2	18 749,60
2018Q4	18 872,5	19 068,6	564,0	19 632,6	19 321,07
2019Q1	18 187,3	18 872,5	–196,1	18 676,4	19 122,37
2019Q2	21 002,3	18 187,3	–685,2	17 502,1	18 428,10
Prognozy *ex ante* wyznaczone rekurencyjnie					
		N1		N2	N3
2019Q3		21 002,3	2 815,0	23 817,3	21 280,37
2019Q4		21 002,3	2 815,0	26 632,3	21 562,12
2020Q1		21 002,3	2 815,0	29 447,3	21 847,60
2020Q2		21 002,3	2 815,0	32 262,3	22 136,86

Źródło: dane w kolumnie 4: GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019); dane w kolumnach 5–6: obliczenia własne.



Rysunek 7.2. Porównanie prognoz z obserwacjami

Źródło: opracowanie własne.

Porównanie prognoz uzyskanych różnymi metodami przedstawiono na rys. 7.2. Jak można zauważyć, prognozy wygasłe (dla obserwacji z okresu 2014Q1–2019Q2) w miarę poprawnie odwzorowują dane rzeczywiste w tym okresie. Przy czym prognozy wyznaczone za pomocą metod N1 i N3 praktycznie się pokrywają i są bliższe obserwacjom niż prognozy obliczone za pomocą modelu (7.2). Zróżnicowanie wartości wyznaczonych prognoz jest wyraźnie widoczne dla prognoz *ex ante*, z których te wyznaczone metodą N2 wykazują stały i dość znaczny wzrost.

Jak wcześniej wspomniano, średnie ruchome mogą być wykorzystywane do prognozowania tzn. wyznaczania nieznanych – przyszłych wartości szeregu czasowego. W tym celu wartości średnich przypisuje się obserwacji następnej po ostatnim okresie uwzględnionym w obliczeniu średniej kroczącej. Przykładowo prognozę dla okresu $(T + 1)$ wyznacza się jako wartość średniej wyznaczonej dla okresów $(T - k, T - k + 1, T - k + 2 + \ldots, T - 1, T)$. Natomiast prognozy na kolejne okresy wyznacza się w podobny sposób, podstawiając za nieznane wartości szeregu ich wyznaczone wcześniej prognozy. I tak dla okresu $(T + 2)$ prognoza jest wartością średnią obliczoną dla okresów $(T - k + 1, T - k + 2 + \ldots, T - 1, T, T + 1)$, przy czym ostania dana potrzebna do wyznaczenia tej średniej nie pochodzi z szeregu obserwacji, a jest prognozą wyznaczoną dla okresu $(T + 1)$. W kolejnym okresie, tj. $(T + 3)$ prognoza będzie średnią z okresów $(T - k + 2 + \ldots, T - 1, T, T + 1, T + 2)$, a ostatnie dwie wartości niezbędne do obliczenia średniej są wcześniej wyznaczonymi prognozami itd.

Przykład 7.2

Korzystając z 3-okresowej prostej średniej kroczącej, obliczonej w przykładzie 5.4, należy wyznaczyć prognozy działalności finansowej i ubezpieczeniowej na kolejne 4 kwartały.

Rozwiązanie

Wyznaczone wartości prognoz przedstawiono w tab. 7.2. Prognozę $\bar{y}_t^*$ na pierwszy kwartał 2019 roku wyznaczono jako średnią $\bar{y}_{19}$, w tab. 5.7, (czyli średnią obliczoną na podstawie y_t dla $t = 18, 19$ i 20), co zapisać należy jako: $\bar{y}_{21}^* = \bar{y}_{19}$. W dalszej kolejności, na podstawie danych pochodzących z obserwacji, wyznaczone zostały prognozy na drugi i trzeci kwartał 2019 roku odpowiednio jako $\bar{y}_{22}^* = \bar{y}_{20}$ i $\bar{y}_{23}^* = \bar{y}_{21}$. Obliczenie prognoz na kolejne okresy wymaga zastosowania procedury rekurencyjnej i skorzystania z już wcześniej wyznaczonych prognoz. Prognozę PKB na czwarty kwartał 2019 roku obliczono jako: $\bar{y}_{23}^* = (y_{20} + y_{21} + \bar{y}_{22}^*)/3$ na pierwszy kwartał 2020 roku wyznaczono jako: $\bar{y}_{24}^* = (y_{21} + \bar{y}_{22}^* + \bar{y}_{23}^*)/3$, a na drugi kwartał 2020 roku jako: $\bar{y}_{25}^* = (\bar{y}_{22}^* + \bar{y}_{23}^* + \bar{y}_{24}^*)/3$.

Tabela 7.2. Prognozy działalności finansowej i ubezpieczeniowej w [mln PLN] do przykładu 7.2

Rok	Kwartał	t	Zmienna y_t	Średnie ruchome przyporządkowane obserwacji	
				środkowej ($\bar{y}_t$)	następnej po ostatniej ($\bar{y}_t^*$)
(1)	(2)	(3)	(4)	(5)	(6)
2018	II kw.	18	18 504,6		
	III kw.	19	19 068,6	18 815,2	
	IV kw.	20	18 872,5	18 709,5	
2019	I kw.	21	18 187,3	19 354,0	18 815,2
	II kw.	22	21 002,3		18 709,5
	III kw.	23			19 354,0
	IV kw.	24			19 514,5
2020	I kw.	25			19 957,0
	II kw.	26			19 608,5

Uwaga: wartości w kolumnach 4 i 5 pochodzą z tab. 5.7.

Źródło: dane w kolumnie 4: GUS, https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019); dane w kolumnach 5–6: obliczenia własne.

Jak zatem widać, proces wyznaczania prognoz jest rekurencyjny i obserwacje są stopniowo zastępowane przez wyznaczone wcześniej prognozy. Nasuwa się oczywiście pytanie o jakość wyznaczonych prognoz, ale to można stwierdzić dopiero po dokonaniu obserwacji zmiennej, którą prognozowano[7].

7 Będzie o tym mowa w dalszych rozważaniach.

W podobny sposób, jak w przypadku modeli naiwnych i średnich ruchomych, wyznacza się prognozy na podstawie modeli szeregów czasowych, w których jedynymi zmiennymi objaśnianymi są opóźnione zmienne objaśniane. Teoretycznie można wyobrazić sobie, że w modelu AR(r) najkrótsze opóźnienie zmiennej prognozowanej będzie przynajmniej o takiej długości, jak założony horyzont prognozy. Przykładowo przy założonym pięciookresowym horyzoncie prognozy korzysta się (tak jak to pokazano w przykładzie 6.5) z modelu postaci:

$$y_t = \alpha_0 + \alpha_5 \cdot y_{t-5} + \alpha_{10} \cdot y_{t-10} + \alpha_{15} \cdot y_{t-15} + \varepsilon_t$$

Wówczas wszystkie prognozy na okresy $t = 1, 2, \ldots, 5$ wyznaczyć można, korzystając z danych rzeczywistych, będących obserwacjami opóźnionymi o odpowiednią liczbę okresów. Jednakże tego typu sytuacje nie są zbyt częste i w praktyce prognozy wyznacza się albo z okresu na okres, na podstawie danych pochodzących z obserwacji, albo rekurencyjnie na podstawie wcześniej wyznaczonych prognoz zmiennej prognozowanej.

Przykład 7.3

Korzystając z danych dotyczących PKB z przykładu 5.1, oszacowane zostaną MNK modele autoregresji i na ich podstawie trzeba będzie wyznaczyć prognozy dla kolejnych 4 kwartałów[8].

Rozwiązanie

W pierwszym kroku należy oszacować model, który zostanie wykorzystany do budowy prognoz. Zdecydowano zbudować model autoregresji zawierający opóźnienia do 4 okresów, ponieważ posiadane dane są obserwacjami kwartalnymi. Zatem model (6.22) jest w tym przypadku postaci:

- M1: $y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \alpha_2 \cdot y_{t-2} + \alpha_3 \cdot y_{t-3} + \alpha_4 \cdot y_{t-4} + \varepsilon_t$

Warto w tym miejscu przypomnieć, że przy szacowaniu modeli autoregresji traci się tyle obserwacji, ile wynosi maksymalne opóźnienie zmiennych. Zatem w przypadku modelu AR(4) z 22 obserwacji utraci się 4 z powodu najdłuższego opóźnienia (o 4 okresy), co oznacza, że model będzie szacowany *de facto* na podstawie 18 danych, poczynając od pierwszego kwartału 2015 roku.

M1 jest modelem autoregresji rzędu czwartego AR(4), którego oszacowania przedstawiono w tab. 7.3. Jak widać, model M1 dobrze opisuje obserwacje ($R^2 = 0{,}99$), ale poza wyrazem wolnym i zmienną opóźnioną o 4 kwartały pozostałe zmienne są statystycznie nieistotne w tym modelu.

8 W niniejszym przykładzie estymację MNK parametrów modeli przeprowadzono, korzystając z funkcji REGLINP, a wartości teoretyczne wyznaczono za pomocą funkcji REGLINW w Excelu.

Tabela 7.3. Parametry oszacowanych modeli

		Zmienne opóźnione				const
		y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	
Model M1						
Liczba obserwacji		18	Liczba stopni swobody			13
$R^2 = 0{,}9873$	$\hat{\alpha}_i$	1,0874	0,0446	0,0319	0,0444	–72 710,4006
$\bar{R}^2 = 0{,}9834$	t	**25,8286**	1,0343	0,8450	1,1804	**–3,2825**
Model M2						
Liczba obserwacji		18	Liczba stopni swobody			16
$R^2 = 0{,}9816$	$\hat{\alpha}_i$	1,1375				–38 022,0691
$\bar{R}^2 = 0{,}9805$	*t*	29,1930				–2,0909
Model M3						
Liczba obserwacji		18	Liczba stopni swobody			16
$R^2 = 0{,}1646$	$\hat{\alpha}_i$				0,4215	285 435,4094
$\bar{R}^2 = 0{,}1124$	*t*				1,7755	2,4581
Model M3*						
Liczba obserwacji		21	Liczba stopni swobody			19
$R^2 = 0{,}2618$	$\hat{\alpha}_i$				0,5011	244 702,8252
$\bar{R}^2 = 0{,}2229$	*t*				2,5961	2,6460

Uwaga: pogrubiono wartości statystyk testowych pozwalających na odrzucenie hipotezy zerowej o braku istotności zmiennej na poziomie istotności 0,05.

Źródło: obliczenia własne.

W związku z tym usunięto z modelu zmienne nieistotne i oszacowano model AR(4) postaci:

- M2: $y_t = \alpha_0 + \alpha_4 \cdot y_{t-4} + \varepsilon_t$

którego stopień dopasowania do danych empirycznych jest równie wysoki, jak modelu M1 ($R^2 = 0{,}98$). Należy przy tym zauważyć, że oba modele estymowane są na podstawie prób o tej samej długości ($T = 18$), ale liczba stopni swobody w modelu M2 jest większa. Zatem do modelowania szeregu kwartalnych obserwacji PKB wystarczy użyć modelu AR(4) z jedną zmienną. W celach porównawczych dodatkowo oszacowano model AR(1) zawierający zmienną opóźnioną o 1 kwartał postaci:

- M3: $y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \varepsilon_t$

do oszacowania którego wykorzystano dwie próby estymacyjne, tj. próbę o tej samej długości co w modelach M1 i M2, co oznaczono jako model M3, oraz próbę, z której usunięto tylko jedną obserwację (z powodu opóźnienia o 1 okres), co oznaczono jako model M3*.

W celu prawidłowej oceny modeli, które różnią się specyfikacją i liczbą stopni swobody obliczono dla nich skorygowane współczynniki determinacji (4.47). Jak widać w tab. 7.3, najlepiej dopasowanymi do danych empirycznych są oba modele AR(4). Wprawdzie w obu modelach AR(1) zmienna opóźniona o 1 okres stała się statystycznie istotna, ale współczynniki determinacji R^2 są bardzo niskie, co oznacza nieuwzględnienie w tych modelach istotnych zmiennych objaśniających. Dlatego modele te nie nadają się do wyznaczenia prognoz, a prognozy PKB na kolejne 4 kwartały zostaną wyznaczone na podstawie modeli M1 i M2.

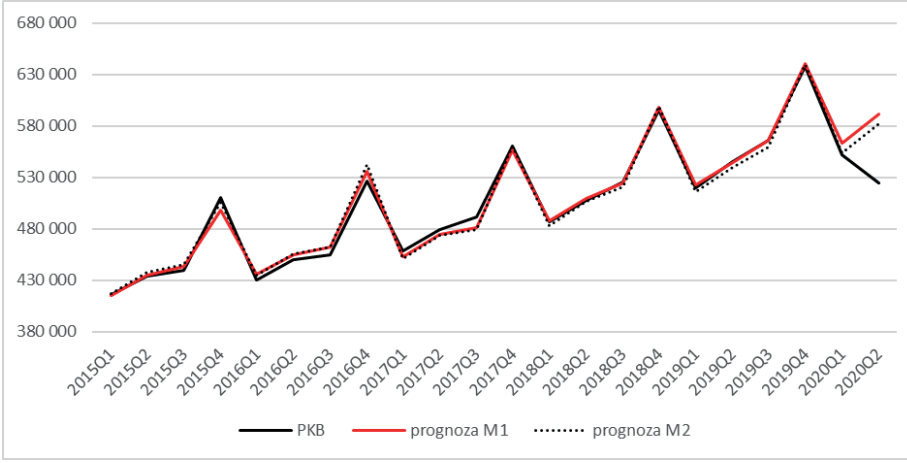
Tabela 7.4. Prognozy i dane do oszacowania modelu i wyznaczenia prognoz

Okresy	Obserwacje PKB	Wartości zmiennych opóźnionych				Wartości teoretyczne wyznaczone z modelu	
		y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	M1	M2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2015Q1	415 701,4	399 536,9	418 230,9	425 004,3	477 657,9	415 137,23	416 468,20
2015Q2	434 227,3	418 230,9	425 004,3	477 657,9	415 701,4	434 698,61	437 733,43
2015Q3	439 967,9	425 004,3	477 657,9	415 701,4	434 227,3	443 256,43	445 438,46
2015Q4	510 331,0	477 657,9	415 701,4	434 227,3	439 967,9	498 596,82	505 334,18
2016Q1	430 165,1	415 701,4	434 227,3	439 967,9	510 331,0	435 354,31	434 856,01
2016Q2	450 116,2	434 227,3	439 967,9	510 331,0	430 165,1	454 444,05	455 930,01
2016Q3	454 810,4	439 967,9	510 331,0	430 165,1	450 116,2	462 150,51	462 460,19
2016Q4	526 019,9	510 331,0	430 165,1	450 116,2	454 810,4	535 935,89	542 501,22
2017Q1	458 461,0	430 165,1	450 116,2	454 810,4	526 019,9	452 959,90	451 309,09
2017Q2	478 886,8	450 116,2	454 810,4	526 019,9	458 461,0	474 138,99	474 004,32
2017Q3	491 398,1	454 810,4	526 019,9	458 461,0	478 886,8	481 168,49	479 344,17
2017Q4	560 568,2	526 019,9	458 461,0	478 886,8	491 398,1	556 798,06	560 348,01
2018Q1	487 129,3	458 461,0	478 886,8	491 398,1	560 568,2	487 711,17	483 496,90
2018Q2	507 606,2	478 886,8	491 398,1	560 568,2	487 129,3	509 429,00	506 732,12
2018Q3	525 180,3	491 398,1	560 568,2	487 129,3	507 606,2	524 682,71	520 964,23
2018Q4	595 756,1	560 568,2	487 129,3	507 606,2	525 180,3	598 058,72	599 648,16
2019Q1	520 197,2	487 129,3	507 606,2	525 180,3	595 756,1	522 803,91	516 108,26
2019Q2	545 556,2	507 606,2	525 180,3	595 756,1	520 197,2	544 753,77	539 401,60
						Prognozy	
2019Q3		525 180,3	595 756,1	520 197,2	545 556,2	565 724,54	559 392,97
2019Q4		595 756,1	520 197,2	545 556,2	565 724,5	640 805,41	639 675,92
2020Q1		520 197,2	545 556,2	565 724,5	609 802,1	563 745,48	553 724,44
2020Q2		545 556,2	565 724,5	609 802,1	656 089,0	591 196,96	582 571,39

Uwaga: zacienione komórki zawierają prognozy wyznaczone rekurencyjne dla kolejnych okresów.

Źródło: kolumna (2) wg https://stat.gov.pl/obszary-tematyczne/inne-opracowania/informacje-o-sytuacji-spoleczno-gospodarczej/biuletyn-statystyczny-nr-92020,4,104.html (dostęp 20.12.2020), pozostałe – obliczenia własne.

Warto odnotować, że o ile korzystając z modelu M2 można wyznaczyć prognozy na wszystkie 4 okresy, korzystając z obserwacji (tj. z danych rzeczywistych) opóźnionej zmiennej prognozowanej, tj. PKB[9], to na podstawie modelu M1, opierając się na danych rzeczywistych można wyznaczyć jedynie prognozę na kolejny kwartał[10], tj. 2019Q3. Zatem dla kolejnych 3 okresów prognozy wyznacza się rekurencyjnie, trzeba bowiem uzupełniać brakujące wartości zmiennych: y_{t-1}, y_{t-2}, y_{t-3} w kolejnych kwartałach, tj. 2019Q4, 2020Q1 i 2020Q2, podstawiając wcześniej wyznaczone prognozy. W tab. 7.4 przedstawiono zestaw danych do estymacji modeli i budowy prognoz, a na rys. 7.3 – porównanie wartości teoretycznych i prognoz wyznaczonych na podstawie obu modeli AR(4) z danymi rzeczywistymi. Jak widać, prognozy wyznaczone na podstawie modeli M1 i M2 dobrze odzwierciedlają kształtowanie się PKB, z wyjątkiem drugiego kwartału 2020 roku. W tym okresie, z powodu pandemii COVID-19, rzeczywista wartość PKB znacząco spadła w porównaniu z analogicznym okresem roku poprzedniego, co nie mogło zostać uwzględnione w opracowywanych prognozach, bo nie było można tego przewidzieć w drugim kwartale 2019 roku, kiedy je wyznaczano.



Rysunek 7.3. Porównanie wartości rzeczywistych i prognoz wyznaczonych na podstawie modeli autoregresyjnych

Źródło: obliczenia własne.

9 Zauważyć trzeba, że w kolumnie (3) dane rzeczywiste dotyczą zmiennej opóźnionej y_{t-4} dla całego horyzontu prognozowania, tj. wszystkich obserwacji do 2020Q2.

10 W modelu M1 obecne są zmienne opóźnione o mniej niż 4 okresy. Zatem dla zmiennej y_{t-1} z kolumny (6) brakuje danych rzeczywistych już dla okresu 2019Q4. Obserwację tę trzeba uzupełnić prognozą wyznaczoną dla tego okresu na podstawie modelu M1. W kolejny okresie, tj. 2020Q1, brakuje danych rzeczywistych zarówno dla zmiennej y_{t-1}, jak i zmiennej y_{t-2} z kolumny (5). Zatem należy uzupełnić prognozami dane dla tych zmiennych. Natomiast dla ostatniego okresu prognozy należy dodatkowo uzupełnić dane dla zmiennej y_{t-3}.

W przypadku metod pośrednich, w procesie prognozowania wykorzystuje się informacje dotyczące zarówno zmiennej prognozowanej, jak i innych zmiennych, wpływających na kształtowanie się zmiennej prognozowanej. W procesie prognozowania pośredniego najpierw określa się wartości zmiennych oddziałujących na zmienną prognozowaną, a dopiero na ich podstawie wyznacza się wartości zmiennej prognozowanej. Przykładem metod pośrednich mogą być prognozy budowane z wykorzystaniem opisowych modeli ekonometrycznych (6.7), w których przyszłe wartości, objaśnianych przez model, zmiennych prognozowanych, są wyznaczane na podstawie założonych lub oszacowanych (dla przyszłych okresów) wartości zmiennych objaśniających – deskryptorów.

W przypadku wykorzystania modeli ekonometrycznych wnioskowanie nosi nazwę predykcji ekonometrycznej, która jest pewnym typem wnioskowania statystycznego. Polega na szacowaniu mającej się zrealizować (w przyszłości) zmiennej losowej[11]. Wartość wyznaczonej prognozy punktowej jest wartością teoretyczną obliczoną dla zadanego przyszłego okresu na podstawie oszacowanego modelu, który musi spełniać, wcześniej sformułowane, założenia teorii predykcji.

Warto przy tym zdawać sobie sprawę z wielu problemów, które pojawiają się w praktyce prognozowania. Po pierwsze, obiekty gospodarcze nie są ani w pełni sterowalne, ani całkiem autonomiczne. Oznacza to, że nawet najlepszy model opisujący badany obiekt lub zjawisko, np. procesy zachodzące w przedsiębiorstwie, tj. model kosztów może okazać się bezużyteczny, jeśli nastąpią gwałtowne zmiany w otoczeniu, np. zmianie ulegnie stawka lub sposób rozliczania podatku VAT. Po drugie, dobre dopasowanie oszacowanego modelu do danych obserwowanych w przeszłości nie oznacza, że będzie on równie dobrze opisywał badane zjawisko w przyszłości. Zasadą jest, że modele ekonometryczne budowane są dla określonych celów, tzn. są inaczej konstruowane, jeśli mają służyć badaniu mechanizmów zjawisk społeczno-gospodarczych, a inaczej – jeśli mają zostać wykorzystane do prognozowania. W tym ostatnim przypadku zaleca się szacowanie modeli na podstawie szeregów czasowych i małą liczbę zmiennych objaśniających współbieżnych ze zmienną objaśnianą. Po trzecie, w teorii prognozy zakłada się niezmienność postaci analitycznej, zbioru zmiennych objaśniających i wartości parametrów w przyszłości, co określa się mianem stabilności modelu[12]. Jednakże ciągłe zmiany zachodzące w zjawiskach społeczno-gospodarczych mogą sprawiać szybką dezaktualizację modeli. Dlatego w praktyce przeprowadza się aktualizację

11 Predykcja obejmuje zjawiska czysto losowe lub zjawiska kontrolowane z losowymi zakłóceniami. Zagadnienia prognozowania na podstawie opisowych modeli ekonometrycznych nie są omawiane w niniejszym opracowaniu. Omawianie tego tematu znaleźć można m.in. w pracach: [Cieślak 1997; Dittmann 2017; Dunis 2001; Witkowska 2012; Zeliaś, Pawełek, Wanat 2003].

12 Stabilność modelu podobnie jak jego właściwości prognostyczne są testowane w procesie weryfikacji modelu za pomocą odpowiednich testów statystycznych oraz mierników.

modelu poprzez ponowną estymację jego parametrów każdorazowo po pojawieniu się nowych obserwacji i wyznacza się kolejne prognozy na podstawie reestymowanego modelu. W niektórych przypadkach, konieczna bywa respecyfikacja modelu. Praktyką prognozowania jest również wyznaczanie prognoz scenariuszowych, uwzględniających różne założenia dotyczące kształtowania się w przyszłości wartości zmiennych objaśniających lub zróżnicowany wpływ otoczenia na zmienną prognozowaną, przykładowo przyjmując warianty optymistyczny, neutralny i pesymistyczny.

Przykład 7.4

Korzystając z danych dotyczących PKB z przykładu 5.1 i oszacowanego w przykładzie 6.2 modelu (M1) trendu z wahaniami sezonowym, należy wyznaczyć prognozy na kolejne 4 kwartały[13].

Rozwiązanie

Do wyznaczenia prognoz posłuży oszacowany model M1 postaci:

$$\hat{y}_t = \hat{\alpha}_1 \cdot z_{1t} + \hat{\alpha}_2 \cdot z_{2t} + \hat{\alpha}_3 \cdot z_{3t} + \hat{\alpha}_4 \cdot z_{4t} + \hat{\alpha}_5 \cdot t$$

dla którego oceny estymatorów parametrów zamieszczono w tab. 6.2. W celu wyznaczenia prognoz należy określić wartości deskryptorów w tym modelu. Model został oszacowany na podstawie danych kwartalnych od 2014Q1 do 2019Q2, tj. 22 obserwacji. Stąd wartości zmiennej czasowej dla okresu estymacji wynoszą $t = 1, 2, 3, \ldots, 22$, a dla kolejnych 4 kwartałów: $t^* = 22, 23, 24$ i 25. W modelu tym, oprócz zmiennej t, występują 4 zmienne zero-jedynkowe, których wartości oznaczają kolejne kwartały, więc nietrudno je wyznaczyć dla kolejnych okresów. Dane niezbędne do oszacowania prognoz przedstawiono w tab. 7.5.

Tabela 7.5. Prognozy i dane do wyznaczenia prognoz

Okres prognozy	Prognozy wyznaczone z modelu trendu M1	t	Zmienne zero-jedynkowe			
			z_{1t} I kw.	z_{2t} II kw.	z_{3t} III kw.	z_{4t} IV kw.
2019Q3	543 849,832	23	0	0	1	0
2019Q4	610 644,244	24	0	0	0	1
2020Q1	541 205,713	25	1	0	0	0
2020Q2	561 777,829	26	0	1	0	0

Źródło: obliczenia własne.

13 Wszystkie oszacowane modele w podobny sposób opisują prognozowane zjawisko i mają tyle samo stopni swobody (tab. 6.2), zatem do prognoz można wybrać dowolny model.

Podstawową zaletą modeli trendu jest łatwość określenia przyszłych wartości jedynej zmiennej objaśniającej, przyjmuje ona bowiem wartości okresu, na jaki wyznacza się prognozę. Również uzupełnienie tego modelu o efekty sezonowe, reprezentowane przez zmienne zero-jedynkowe, nie sprawia żadnych problemów z określeniem ich przyszłych wartości. Znacznie trudniej jest budować prognozy w modelach, w których wartości zmiennych objaśniających nie są tak oczywiste. Należy wówczas najpierw wyznaczyć wartości tych zmiennych, wykorzystując na przykład metody omawiane wcześniej, a w drugim etapie na podstawie oszacowanych wartości zmiennych objaśniających wyznaczyć prognozy zmiennej prognozowanej.

Przykład 7.5

Dysponując kwartalnymi wartościami PKB i eksportu towarów za okres 2014Q1–2019Q2 – dane z przykładu 5.1 – należy wyznaczyć prognozy eksportu dla kolejnych 4 kwartałów.

Rozwiązanie

Do wyznaczenia prognoz wykorzystać należy model, w którym wartość eksportu jest wyjaśniana wielkością produktu krajowego brutto. Oszacowanie MNK tego modelu jest postaci[14]:

$$\text{ME1: } \hat{y}_t = -39677{,}011 + 0{,}681 \cdot x_t$$

gdzie:
y_t – eksport towarów i usług w mln PLN (ceny bieżące),
x_t – produkt krajowy brutto w mln PLN (ceny bieżące).

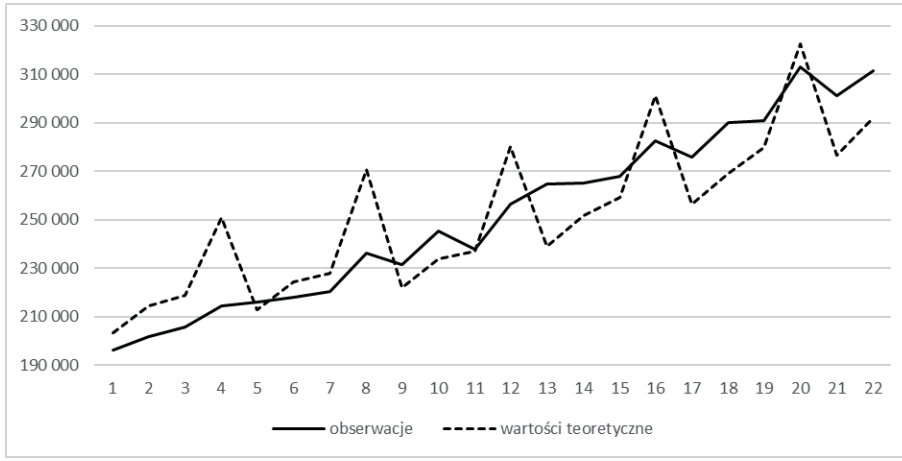
Współczynnik determinacji wynosi $R^2 = 0{,}7509$, a wartość statystyki t-Studenta dla współczynnika kierunkowego wynosi 7,7649. Upoważnia to do stwierdzenia, że PKB istotnie wpływa na wartość eksportu, ponieważ na poziomie istotności 0,05 hipoteza o zerowej wartości parametru stojącego przy tej zmiennej (4.35) została odrzucona. Porównanie wartości rzeczywistych z teoretycznymi przedstawiono na rys. 7.4. Jak widać na rysunku, wartości teoretyczne odbiegają od rzeczywistych najbardziej w pierwszym i czwartym kwartale.

Dlatego model opisujący eksport wzbogacono o 2 zmienne zero-jedynkowe przyjmujące wartość 1 dla pierwszego i czwartego kwartału. Po oszacowaniu ten model jest postaci:

$$\text{ME2: } \hat{y}_t = -152157{,}12 + 0{,}8566 \cdot x_t + 12663{,}64 \cdot z_{1t} - 44714{,}92 \cdot z_{2t}$$

gdzie:
z_{1t}, z_{2t} – zmienne zero-jedynkowe przyjmujące wartość jeden odpowiednio dla pierwszego i czwartego kwartału, a pozostałe oznaczenia jak poprzednio.

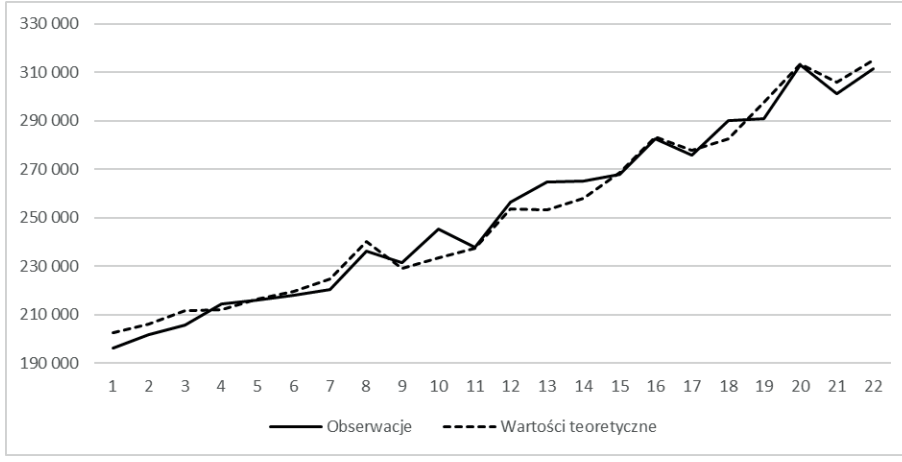
14 Estymację MNK parametrów modeli przeprowadzono, korzystając z funkcji REGLINP, a wartości teoretyczne wyznaczono za pomocą funkcji REGLINW w Excelu.



Rysunek 7.4. Porównanie wartości empirycznych i teoretycznych wyznaczonych dla modelu ME1

Źródło: opracowanie własne.

Tym razem współczynnik determinacji wynosi $R^2 = 0{,}9779$, a wartości statystyki t-Studenta dla wszystkich parametrów (odpowiednio –10,48 dla wyrazu wolnego, 27,94 dla x_t, 4,16 dla z_{1t}, oraz –11,96 dla z_{2t}) wskazują, że na poziomie istotności 0,05 hipotezy o zerowej wartości parametrów stojących przy tej zmiennej (4.35) zostały odrzucone. Zatem oszacowany model można przyjąć za dostatecznie poprawny do budowy prognoz, co potwierdza porównanie wartości rzeczywistych z teoretycznymi (rys. 7.5).



Rysunek 7.5. Porównanie wartości empirycznych i teoretycznych wyznaczonych dla modelu ME2

Źródło: opracowanie własne.

W kolejnym kroku należy wyznaczyć wartości PKB na kolejne 4 okresy, aby można było oszacować prognozy eksportu. Prognozy PKB wyznaczono na podstawie modeli autoregresji M1 i M2 (przykład 7.3) oraz modelu trendu z sezonowością (przykład 7.4). Korzystając z wcześniej wyznaczonych prognoz (w przykładach 7.3 i 7.4, tab. 7.4 i 7.5) produktu krajowego brutto, oszacowano prognozy eksportu dla kolejnych kwartałów na podstawie modelu ME2, który przedstawiono w tab. 7.7 i na rys. 7.6.

Tabela 7.6. Wartości rzeczywiste i teoretyczne wyznaczone na podstawie modeli ME1 i ME2

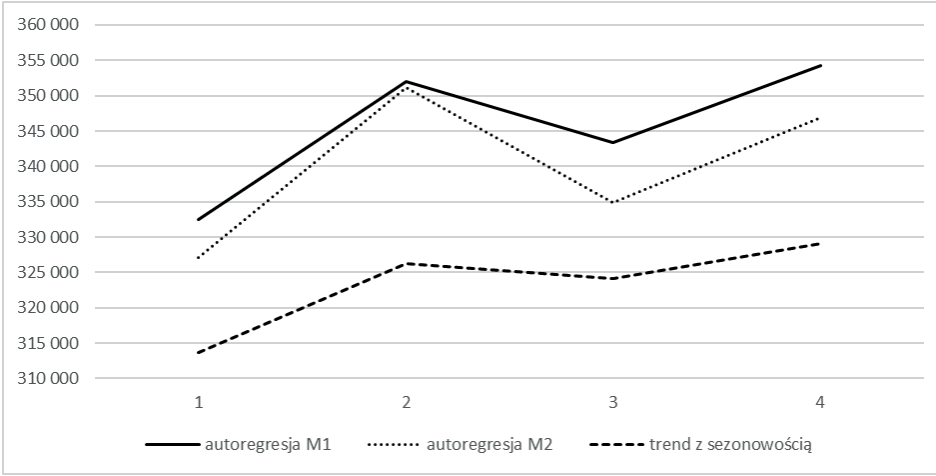
Okresy	Obserwacje	Wartości teoretyczne		Okresy	Obserwacje	Wartości teoretyczne	
		ME1	ME2			ME1	ME2
2014Q1	196 189,9	203 299,6	202 762,6	2016Q4	256 455,3	280 219,6	253 733,4
2014Q2	201 895,7	214 668,6	206 112,8	2017Q1	264 686,5	239 134,0	253 238,8
2014Q3	205 816,5	218 787,46	211 915,1	2017Q2	265 385,3	251 555,8	258 072,6
2014Q4	214 488,1	250 808,5	212 304,9	2017Q3	268 138,4	259 164,5	268 790,1
2015Q1	216 159,5	213 129,9	216 609,6	2017Q4	282 746,3	301 230,0	283 328,5
2015Q2	218 225,4	224 396,3	219 815,8	2018Q1	275 837,6	256 568,4	277 797,0
2015Q3	220 407,6	227 887,5	224 733,4	2018Q2	290 022,0	269 021,4	282 674,5
2015Q4	236 282,9	270 678,5	240 293,8	2018Q3	291 146,9	279 709,0	297 729,1
2016Q1	231 470,0	221 925,9	228 999,6	2018Q4	313 159,6	322 629,4	313 471,6
2016Q2	245 541,4	234 059,1	233 426,8	2019Q1	301 188,1	276 678,5	306 124,0
2016Q3	237 932,6	236 913,9	237 448,0	2019Q2	311 390	292 100,5	315 183,7

Źródło: obliczenia własne.

Tabela 7.7. Wartości zmiennych objaśniających i eksportu

Wartości zmiennych objaśniających					Prognozy eksportu wyznaczone dla prognoz zmiennych objaśniających		
PKB(M1)	PKB(M2)	PKB(trend)	z_{1t}	z_{2t}	PKB(M1)	PKB(M2)	PKB(trend)
565 724,54	559 392,97	543 849,83	0	0	332 460,54	327 036,72	313 721,97
640 805,41	639 675,92	610 644,24	0	1	352 062,29	351 094,73	326 225,28
563 745,48	553 724,44	541 205,71	1	0	343 428,86	334 844,52	324 120,58
591 196,96	582571,39	561 777,83	0	0	354 281,03	346 892,09	329 079,66

Źródło: obliczenia własne.



Rysunek 7.6. Porównanie prognoz eksportu wyznaczonych dla prognoz PKB oszacowanych na podstawie modeli autoregresyjnych M1 i M2 oraz trendu z sezonowością

Źródło: opracowanie własne.

7.3. Błędy prognoz

Ocenę jakości prognoz przeprowadza się na podstawie błędów prognoz, wśród których wyróżnić należy:

- ocenę *ex post*, opartą na zrealizowanym błędzie prognozy, przeprowadzaną po dokonaniu obserwacji zmiennej prognozowanej (tj. gdy prognoza wygaśnie),
- ocenę *ex ante*, opartą na oczekiwanym błędzie prognozy i wyrażającą przypuszczenie co do dokładności prognozy (wyrażone zanim prognoza wygaśnie).

Innymi słowy, błędy *ex post* dla prognoz *ex ante* można wyznaczyć dopiero po dokonaniu obserwacji zmiennej prognozowanej, a błędy *ex ante* wyznaczyć można już w momencie sporządzania prognozy.

Najprostsze do wyznaczenia są błędy *ex post*, które wyznacza się podobnie jak reszty empiryczne (4.32) dla każdego okresu prognozowania t, tzn. jako:

$$e_t^* = y_t^* - y_t \tag{7.4}$$

gdzie:
y_t^* – wartość wyznaczonej prognozy,
y_t – wartość zrealizowana, tj. obserwacja empiryczna.

Błędy prognoz (7.4) informują o odchyleniu (*in plus* lub *in minus*) prognozy od wartości rzeczywistej (zaobserwowanej). W praktyce, zwłaszcza dla dłuższych horyzontów prognozowania, oblicza się również błędy średnie różnej postaci, wśród których najczęściej używany jest błąd średniokwadratowy *RMSE* (*root mean square error*) wyrażany wzorem:

$$RMSE = \sqrt{\frac{1}{T^*} \cdot \sum_{t=1}^{T^*} (y_t^* - y_t)^2} \tag{7.5}$$

gdzie:
T^* – horyzont prognozy.

Błędy (7.4) i (7.5) wyrażane są w takich samych jednostkach co zmienna prognozowana. Czasem wygodnie jest ocenić jakość prognoz na podstawie błędów względnych. Najprościej jest odnieść wartość błędu *RMSE* do średniej wartości zmiennej prognozowanej wyznaczonej dla zadanego horyzontu prognozy, czyli wyznaczyć względny błąd prognozy *RMSER* postaci:

$$RMSER = \frac{RMSE}{\frac{1}{T^*} \cdot \sum_{t=1}^{T^*} y_t} \tag{7.6}$$

Błąd *RMSEP* jest niemianowany lub można go wyrazić w procentach. Innym, często wykorzystywanym w praktyce, błędem względnym jest błąd *RMSEP* postaci:

$$RMSEP = \sqrt{\frac{1}{T^*} \cdot \sum_{t=1}^{T^*} \left(\frac{y_t^* - y_t}{y_t} \right)^2} \tag{7.7}$$

Mniej popularny jest błąd prognoz wyznaczony na podstawie wartości bezwzględnych odchyleń, czyli średni błąd absolutny MAE (*mean absolute error*):

$$MAE = \frac{1}{T^*} \cdot \sum_{t=1}^{T^*} \left| y_t^* - y_t \right| \tag{7.8}$$

oraz średni absolutny błąd względny MAPE (*mean absolute percentage error*):

$$MAPE = \frac{1}{T^*} \cdot \sum_{t=1}^{T^*} \left| \frac{y_t^* - y_t}{y_t} \right| \tag{7.9}$$

Interpretacja błędów (7.8)–(7.9) jest podobna jak błędów średniokwadratowych, tzn. informują one, o ile wyznaczone prognozy średnio różnią się od wartości rzeczywistych. Należy jednak zauważyć, że błędy przeciętne są z reguły niższe (z definicji nie są większe) niż błędy średniokwadratowe. W praktyce oprócz

błędów średnich analizuje się również maksymalne błędy prognoz, ponieważ wyznaczona prognoza jest użyteczna jedynie w sytuacji, gdy błąd maksymalny nie przekroczy pewnej akceptowanej wartości. Błędy maksymalne:

$$MAX = \max_t |e_t^*| = \max_t |y_t^* - y_t| \tag{7.10}$$

wyrażane są w tych samych jednostkach co zmienna prognozowana. Można również wyznaczyć względne maksymalne błędy procentowe postaci:

$$MAXP = \max_t \left| \frac{y_t^* - y_t}{y_t} \right| \tag{7.11}$$

W przypadku gdy w szeregach danych występują wartości bliskie 0, np. dane są zdefiniowane jako stopy zwrotu lub logarytmy danych rzeczywistych, zaleca się korzystanie ze skorygowanych błędów względnych postaci:

$$RMSEPS = \sqrt{\frac{\sum_{t=1}^{T^*} (y_t^* - y_t)^2}{T^* \cdot \sum_{t=1}^{T^*} y_t^2}} \tag{7.12}$$

$$MAPES = \frac{\sum_{t=1}^{T^*} |y_t^* - y_t|}{T^* \cdot \sum_{t=1}^{T^*} |y_t|} \tag{7.13}$$

$$MAXPS = \max_t \left| \frac{y_t^* - y_t}{\sum_{t=1}^{T^*} |y_t|} \right| \tag{7.14}$$

Błędy *ex post* wyznacza się w odniesieniu do prawdziwych wartości zmiennej prognozowanej, tj. po dokonaniu obserwacji w zadanym horyzoncie prognostycznym. Dlatego można je obliczyć niezależnie od zastosowanej metody prognozowania. Natomiast błędów *ex ante* nie da się wyznaczyć, np. dla metod naiwnych lub średnich ruchomych. Błędy *ex ante*, zwane również błędami predykcji lub predyktora, można wyznaczyć już w momencie sporządzania prognoz. Informują one, o ile oszacowana wartość zmiennej prognozowanej średnio odchyla się od nieznanej wartości rzeczywistej. Do ich wyznaczenia używa się między innymi ocen estymatora wariancji składnika losowego (tzw. wariancji resztowej) oszacowanych wcześniej modeli ekonometrycznych (w tym modeli szeregów czasowych)[15].

15 Zagadnienie to jest omawiane m.in. w pracy [Witkowska 2012, s. 196–199], ale zostało ominięte w niniejszej pracy z uwagi na wcześniejsze pominięcie wykładu nt. estymacji MNK nieznanych parametrów modeli regresji i ekonometrycznych.

Przykład 7.6

Na podstawie wartości PKB z przykładu 5.1, obserwowanych od trzeciego kwartału 2018 do drugiego kwartału 2019 (tj. w ostatnich 4 kwartałach, traktowanych jako dane do budowy modeli trendu i autoregresji) wyznaczyć należy prognozy na kolejne 4 kwartały za pomocą 3- i 4-okresowej średniej ruchomej[16]. Ponadto trzeba obliczyć błędy prognoz wyznaczonych w tym przykładzie oraz w poprzednich przykładach (tj. 7.3 i 7.4).

Rozwiązanie

Najpierw wyznaczyć trzeba prognozy, korzystając ze średnich kroczących 3- i 4-okresowych (tab. 7.8). Jak widać, do wyznaczenia prognoz potrzebne są ostatnie 3 lub 4 obserwacje, na podstawie których rekurencyjnie wyznacza się prognozy okres po okresie, uzupełniając brakujące dane rzeczywiste wyznaczonymi wcześniej prognozami.

Tabela 7.8. Prognozy dotyczące działalności finansowej i ubezpieczeniowej w [mln. PLN]

Rok	Kwartał	Obserwacje PKB	Prognozy wyznaczone jako średnie ruchome	
			3-okresowe	4-okresowe
2018Q3	III	525 180,3		
2018Q4	IV	595 756,1		
2019Q1	I	520 197,2		
2019Q2	II	545 556,2		
2019Q3	III		553 836,500	546 672,458
2019Q4	IV		539 863,300	552 045,489
2020Q1	I		546 418,667	541 117,837
2020Q2	II		546 706,156	546 347,996

Źródło: obliczenia własne.

Następnie oblicza się błędy prognoz (7.4) dla każdego okresu (tab. 7.9). W celu syntetyzacji analiz wyznaczono też błędy średniokwadratowe (7.5) *RMSE* oraz błędy względne *RMSER* (7.6) i *RMSEP* (7.7), a także błędy przeciętne i maksymalne (7.8)–(7.11).

Porównując prognozy wyznaczone za pomocą oszacowanych modeli i średnich kroczących, dostrzec można lepsze dopasowanie prognoz oszacowanych na podstawie modeli estymowanych MNK. Biorąc pod uwagę błędy średniokwadratowe i maksymalne, należy stwierdzić, że najlepiej dopasowanymi do danych

16 Zastosowanie drugiej średniej jest podyktowane występowaniem kwartalnych odchyleń sezonowych.

empirycznych okazały się prognozy wyznaczone na podstawie modelu trendu z sezonowością (por. przykłady 6.2 i 7.4). Niskie średnie błędy prognoz uzyskano również w przypadku modeli autoregresyjnych, oszacowanych w przykładzie 7.3. Przy czym model M2 okazał się lepszym modelem prognostycznym niż M1, w przypadku którego prognozy na ostanie trzy kwartały wyznaczano rekurencyjnie, wykorzystując wcześniej oszacowane prognozy. Wynika z tego, że prognozy wyznaczane na podstawie danych rzeczywistych są obarczone mniejszymi błędami niż te wyznaczone rekurencyjnie na podstawie wcześniej wyznaczonych prognoz. Warto odnotować, że w przypadku tych 3 modeli największe błędy dotyczą ostatniego okresu, w którym PKB odnotowało znaczący spadek z powodu pandemii.

Prognozy wyznaczone (rekurencyjnie) za pomocą średnich kroczących 3- i 4-okresowych są obarczone wyższymi błędami niż prognozy oszacowane na podstawie modeli, a błędy maksymalne dotyczą drugiego okresu prognozowania. Oceniając wszystkie wyznaczone prognozy, należy zauważyć, że wysokie błędy prognoz, w szczególności błędy maksymalne, dyskwalifikują prognozy wyznaczone za pomocą średnich ruchomych.

Tabela 7.9. Błędy prognoz wyznaczonych w przykładzie 7.6

	Modele autoregresyjne		Trend z sezonowością	Średnia ruchoma	
Okres	M1	M2		3-okresowa	4-okresowa
2019Q3	–108,86	–6440,43	–21 983,6	–11 996,9	–19 160,9
2019Q4	2896,71	1767,22	–27 264,5	–98 045,4	–85 863,2
2020Q1	11 517,58	1496,54	–11 022,2	–5809,2	–11 110,1
2020Q2	66 332,16	57 706,59	36 913,0	21 841,4	21 483,2
RMSE	33 693,52	29 055,52	26 032,1	50 664,6	45 619,6
RMSER	5,9%	5,1%	4,6%	8,9%	8,0%
RMSEP	6,4%	5,5%	4,7%	8,0%	7,3%
MAE	20 213,83	16 852,70	24 295,81	34 423,22	34 404,35
MAPE	3,8%	3,2%	4,3%	5,7%	5,7%
MAX	66 332,16	57 706,59	36 913,0	98 045,4	85 863,2
MAXP	12,64%	10,99%	7,03%	15,37%	13,46%

Źródło: obliczenia własne.

Przykład 7.7

Dzienne logarytmiczne stopy zwrotu z akcji spółki KGHM w okresie od 2 listopada do 30 grudnia 2020 roku przedstawiono w tab. 7.10. Należy oszacować model AR(5) dla danych do 14.12.2020 roku oraz wyznaczyć prognozy na kolejne okresy i ocenić jakość tych prognoz.

Tabela 7.10. Notowania cen akcji spółki KGHM oraz logarytmiczne stopy zwrotu

Nr	Stopy zwrotu	Nr	Stopy zwrotu	Nr	Stopy zwrotu	Nr	Stopy zwrotu
1	0,0172	12	–0,0161	23	0,0476	34	0,0110
2	0,0857	13	–0,0112	24	0,0130	35	–0,0051
3	–0,0197	14	–0,0365	25	0,0503	36	–0,0212
4	0,0299	15	0,0471	26	0,0102	37	–0,0049
5	0,0411	16	–0,0106	27	0,0377	38	0,0227
6	–0,0029	17	0,0484	28	0,0054	39	0,0000
7	–0,0206	18	0,0286	29	0,0232	40	0,0000
8	0,0000	19	0,0221	30	–0,0542	41	0,0150
9	0,0173	20	–0,0019	31	–0,0135	42	–0,0082
10	–0,0107	21	–0,0421	32	0,0445	43	–0,0189
11	0,0668	22	0,0523	33	–0,0011		

Źródło: obliczenia własne.

Rozwiązanie

Model AR(5) opisujący stopy zwrotu z akcji KGHM jest postaci:

$$y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \alpha_2 \cdot y_{t-2} + \alpha_3 \cdot y_{t-3} + \alpha_4 \cdot y_{t-4} + \varepsilon_t$$

Model ten oszacowano na podstawie 26 stóp zwrotu (z okresu od 2 listopada do 14 grudnia 2020 roku), a oceny estymatorów parametrów i wartości statystyk t-Studenta przedstawiono w tab. 7.11. Model opisuje zmiany stóp zwrotu w 16% ($R^2 = 0{,}1526$), co wydaje się całkiem dobrym wynikiem wobec znacznej zmienności zmiennej objaśnianej (współczynnik zmienności 294%).

Tabela 7.11. Oceny estymatorów parametrów i wartości statystyk t-Studenta

Zmienna	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	const
Ocena parametru	–0,2392	–0,1719	–0,3265	–0,0436	–0,1390	0,0228
Statystyka t-Studenta	1,1576	–0,8076	–1,4219	–0,1873	–0,6594	2,0922

Źródło: obliczenia własne.

W kolejnych krokach wyznaczono prognozy i błędy prognoz (7.4), *MAE* (7.8), *MAX* (7.10) oraz błędy względne *MAPES* (7.13) i *MAXPS* (7.14), ponieważ nie można wyznaczyć błędów względnych (7.9) i (7.11) z powodu zerowych stóp zwrotu w obserwacjach o numerach 39 i 40. Dane do wyznaczenia prognoz,

oszacowane prognozy, błędy (7.4) i obliczenia pomocnicze do obliczenia błędów (7.13) i (7.14) przedstawiono w tab. 7.12.

Tabela 7.12. Dane do wyznaczenia prognoz, oszacowane prognozy, błędy i obliczenia pomocnicze

Nr	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}	y_t^*	e_t^*	$\|e_t^*\|$	$\|e_t^*\|/y_t$
32	0,0445	–0,0135	–0,0542	0,0232	0,0054	0,0377	0,0095	–0,0350	0,0350	0,2292
33	–0,0011	0,0095	–0,0135	–0,0542	0,0232	0,0054	0,0345	0,0355	0,0355	0,2328
34	0,0110	0,0345	0,0095	–0,0135	–0,0542	0,0232	0,0257	0,0147	0,0147	0,0964
35	–0,0051	0,0257	0,0345	0,0095	–0,0135	–0,0542	0,0299	0,0350	0,0350	0,2290
36	–0,0212	0,0299	0,0257	0,0345	0,0095	–0,0135	0,0078	0,0291	0,0291	0,1903
37	–0,0049	0,0078	0,0299	0,0257	0,0345	0,0095	0,0038	0,0087	0,0087	0,0572
38	0,0227	0,0038	0,0078	0,0299	0,0257	0,0345	–0,0005	–0,0232	0,0232	0,1520
39	0,0000	–0,0005	0,0038	0,0078	0,0299	0,0257	0,0088	0,0088	0,0088	0,0578
40	0,0000	0,0088	–0,0005	0,0038	0,0078	0,0299	0,0118	0,0118	0,0118	0,0775
41	0,0150	0,0118	0,0088	–0,0005	0,0038	0,0078	0,0184	0,0034	0,0034	0,0224
42	–0,0082	0,0184	0,0118	0,0088	–0,0005	0,0038	0,0160	0,0242	0,0242	0,1588
43	–0,0189	0,0160	0,0184	0,0118	0,0088	–0,0005	0,0145	0,0334	0,0334	0,2190

Źródło: obliczenia własne.

Wyznaczone prognozy średnio różnią się od wartości rzeczywistych o $MAE = 0{,}0202$, co daje 13% średni absolutny błąd względny *MAPES*. Natomiast maksymalny błąd prognozy wynosi $MAX = 0{,}0355$ i $MAXPS = 0{,}2328$, czyli 23,28%.

Rozdział 8

Elementy wielowymiarowej analizy statystycznej

Analizy wielowymiarowe dotyczą zjawisk lub obiektów mierzonych w więcej niż jednym wymiarze. Innymi słowy, zakłada się, że opis zjawisk, procesów, systemów lub obiektów dokonuje się poprzez obserwacje przynajmniej dwóch cech jednocześnie. W tab. 8.1 przedstawiono *N* obiektów, z których każdy opisany został za pomocą *K* cech.

Tabela 8.1. Obserwacje obiektów wielocechowych

Obiekt	Cecha 1	Cecha 2	…	Cecha *K*-ta
1	x_{11}	x_{12}	…	x_{1K}
2	x_{21}	x_{22}	…	x_{2K}
3	x_{31}	x_{32}	…	x_{3K}
…	…	…	…	…
N	x_{N1}	x_{N2}	…	x_{NK}

Źródło: opracowanie własne.

Metody prowadzenia analiz zależą od celu badania, którym może być:

1) opis zjawisk i znalezienie czynników je kształtujących,
2) prognozowanie polegające na wyznaczeniu przyszłych wartości,
3) porządkowanie – ranking obiektów,
4) klasyfikacja lub grupowanie obiektów.

Pierwsze dwa cele można zrealizować za pomocą miar korelacji i odpowiednio zbudowanych modeli ekonometrycznych, o czym była już mowa w poprzednich rozdziałach. Natomiast ostanie dwa cele badawcze realizowane są za pomocą metod taksonomicznych, które pozwalają odpowiedzieć na pytania: „czy badane obiekty są do siebie podobne (i w jakim stopniu)?", „czy też do siebie podobne nie są?"

Znajomość odpowiedzi na te pytania pozwala zaliczyć obiekty do homogenicznych klas (w szczególności – tej samej klasy), wnioskować przez analogie o ich

własnościach, itp. Przez taksonomię rozumie się dyscyplinę naukową zajmującą się zasadami i procedurami klasyfikacji (tj. porządkowania, grupowania, dyskryminacji czy podziału). Porządkowanie obiektów wielocechowych przeprowadza się z wykorzystaniem tzw. mierników taksonomicznych, a pozostałe – z wykorzystaniem odpowiednich metod klasyfikacji i grupowania[1].

8.1. Podstawowe pojęcia

Zadanie klasyfikacji polega na określeniu, do jakiej klasy (podzbioru, grupy) należy obiekt opisany za pomocą pewnej liczby cech. Przez obiekty rozumie się jednostki badania podlegające klasyfikacji. Zbiór obiektów będących przedmiotem klasyfikacji można zapisać w postaci ogólnej jako:

$$\Omega = \{O_1, O_2, \ldots, O_N\} \tag{8.1}$$

Jednocześnie przyjmuje się, że każdy z obiektów badania jest opisany za pomocą K wybranych cech diagnostycznych:

$$X = \{X_1, X_2, \ldots, X_K\} \tag{8.2}$$

których dobór ma kluczowe znaczenie dla jakości klasyfikacji. Dlatego muszą to być takie charakterystyki obiektów, które mają na nie istotny (merytorycznie) wpływ. Warto odnotować, że zmienne diagnostyczne powinny spełniać następujące kryteria:

- uniwersalność – która oznacza, iż zmienne opisują tę samą kategorię – cechę we wszystkich obiektach badania;
- mierzalność – polegającą na pomiarze zmiennych według uniwersalnej skali pomiarowej;
- dostępność – możliwość zgromadzenia potrzebnych danych – obserwacji.

Dodatkowym kryterium jest prawdziwość danych, które powinny pochodzić z zaufanych źródeł. Warto również zadbać, aby dane dotyczące pojedynczych cech były brane z jednego źródła dla wszystkich obiektów badania. W takim przypadku bowiem można oczekiwać, że zostały one wygenerowane w jednakowy sposób, a ewentualne błędy pomiaru stanowią taki sam procent wartości pomiaru danych we wszystkich obiektach.

1 Przykłady wykorzystania omawianych w niniejszym rozdziale metod przedstawiono w rozdziale 10.

Zadanie klasyfikacji można przedstawić jako problem podziału N – elementowego zbioru Ω, na p ($1 \leq p \leq N$) podzbiorów (grup, klas) oznaczonych jako $\{A_1, A_2, \ldots, A_p\}$ w taki sposób, aby spełnione były następujące warunki:

$$A_1 \cup A_2 \cup \ldots \cup A_p = \Omega \tag{8.3}$$

$$A_i \cap A_j = \varnothing \text{ dla } i, j = 1, 2, \ldots, p; i \neq j \tag{8.4}$$

$$A_i \neq \varnothing \text{ dla } i = 1, 2, \ldots, p \tag{8.5}$$

Warunek addytywności (zupełności), zapisany jako (8.3), oznacza, że suma wszystkich wyodrębnionych klas równa jest zbiorowi podlegającemu podziałowi. Innymi słowy warunek (8.3) zapewnia kompletność procesu podziału, czyli każdy obiekt znajdzie się w jakiejś klasie, żaden nie zostanie pominięty w procesie klasyfikacji. Warunek rozłączności grup typologicznych (8.4) oznacza, że poszczególne grupy nie zawierają żadnych elementów wspólnych. Zatem każdy obiekt zostanie zaklasyfikowany tylko do jednej z grup, czyli wyklucza się przypadki zaliczenia jednego obiektu do dwóch i więcej klas równocześnie. Natomiast warunek (8.5) oznacza, że nie ma zbiorów pustych, czyli w każdej klasie znajduje się przynajmniej jeden obiekt.

Zagadnienia klasyfikacji rozwiązywane są za pomocą metod taksonomicznych, które w szerokim rozumieniu tego pojęcia można podzielić na:

- metody taksonomii wzorcowej,
- metody taksonomii bezwzorcowej.

Pierwsze z wymienionych określane są również mianem metod dyskryminacyjnych lub metod rozpoznawania z nauczycielem, np. analiza dyskryminacyjna lub drzewa klasyfikacyjne, dzielą zbiór danych na klasy na podstawie znanych *a priori* wzorców tych klas. Drugą grupę metod, określaną również jako metody rozpoznawania bez nauczyciela lub metody taksonomiczne w wąskim rozumieniu, wykorzystuje się do rozwiązywania zadań grupowania w przypadkach, kiedy brak jest *a priori* informacji o wzorcach klas. Do najpopularniejszych metod grupowania należą metody *k*-średnich i Warda.

Podstawowymi pojęciami używanymi w taksonomii są: podobieństwo, odległość, jednorodność i klasa. Przez podobieństwo rozumie się wspólność (zbieżność) wybranych właściwości dwu lub więcej obiektów. Jednorodność jest właściwością zbioru obiektów składającego się z jednostek podobnych. Natomiast klasa jest podzbiorem zawierającym obiekty podobne, tj. wyróżnione na podstawie podobieństwa wspólnych właściwości. Jedną z często stosowanych miar podobieństwa cech jest współczynnik korelacji liniowej Pearsona.

Do ważniejszych pojęć w taksonomii należy tzw. odległość taksonomiczna definiowana jako odległość między obiektami w wielowymiarowej przestrzeni ich cech. Wymiar tej przestrzeni jest określony przez liczbę zmiennych charakteryzujących jednostki badanej zbiorowości. Najbardziej znaną i najczęściej stosowaną odległością jest, powszechnie znana z geometrii analitycznej, odległość euklidesowa.

8.2. Mierniki taksonomiczne

Badane będą obiekty opisane zgodnie z zapisem (8.1)–(8.2) za pomocą kilku zmiennych (charakterystyk, cech diagnostycznych). Należy obiekty te ze sobą porównać i uporządkować według ustalonego kryterium, uwzględniającego wszystkie cechy ich opisu jednocześnie. W tym celu wykorzystać można mierniki taksonomiczne, zwane również miernikami agregatowymi lub syntetycznymi, umożliwiające porównywanie obiektów wielocechowych (wielowymiarowych) za pomocą jednej miary.

Ze względu na kierunek wpływu zmiennych diagnostycznych na rozwój badanego zjawiska, wyróżnia się trzy ich rodzaje.

a) Stymulanty to zmienne mające pozytywny wpływ na badane kryterium ogólne. Ich wyższe wartości decydują o lepszej pozycji (ocenie) badanego obiektu (zjawiska).
b) Destymulanty to te zmienne o przeciwnym kierunku oddziaływania co stymulanty. Wzrost ich wartości powoduje pogorszenie się pozycji (oceny) rozpatrywanego obiektu (zjawiska).
c) Nominanty, czyli zmienne, które przyjmują pewną ustaloną (najkorzystniejszą z punktu widzenia oceny obiektów) wartość lub wartości unormowane w pewnym przedziale. Rozróżnia się ponadto nominanty symetryczne i niesymetryczne (lewostronne i prawostronne) w zależności od tego, jaki wpływ na badane zjawisko mają wartości zmiennej zarówno powyżej, jak i poniżej wartości optymalnej (lub przedziału dopuszczalnego).

W badaniach często wykorzystuje się pojęcie wzorca (lub antywzorca), przez który rozumie się taki obiekt, dla którego wartości wszystkich charakterystyk są optymalne (lub przeciwne do optymalnych). Optymalne wartości stymulant to wartości maksymalne, a destymulant – minimalne.

Konstrukcja miernika agregatowego prowadzona jest w kilku etapach.

- Wybór miernika oznacza decyzję, czy będzie to miara ze wzorcem lub bez wzorca oraz jaki konkretnie typ miernika zostanie zastosowany w badaniach. Dodatkowo w przypadku, gdy wybrany zostanie miernik ze wzorcem, trzeba będzie zdefiniować wzorzec (rzeczywisty lub teoretyczny), w odniesieniu do którego będą porównywane poszczególne obiekty badania.
- Wybór zmiennych diagnostycznych polega na sporządzeniu listy potencjalnych zmiennych opisujących w sposób możliwie pełny analizowane obiekty, zebraniu obserwacji ich dotyczących oraz sprawdzeniu, czy nie są one quazi-stałe (zaleca się, aby współczynniki zmienności (2.19) wynosiły przynajmniej 10%) oraz nadmiernie skorelowane ze sobą. Oprócz tego należy opisać, w jaki sposób każda ze zmiennych diagnostycznych wpływa na badane obiekty, aby prawidłowo określić ich rolę (tj. czy są to stymulanty, destymulanty, czy nominanty).
- Normalizacja zmiennych polega na pozbyciu się mian i sprowadzeniu wszystkich zmiennych do porównywalności.

- Wyznaczenie wartości miernika.
- Porządkowanie według wartości danego miernika.

Normalizacja zmiennych diagnostycznych pozwala na uwolnienie poszczególnych zmiennych od ich jednostek pomiarowych (mian), a także na ujednolicenie rzędów wielkości i/lub rozrzutu zmiennych. W ogólnym przypadku normalizacja zmiennych diagnostycznych polega na przekształceniu zgodnie ze wzorem:

$$z_{ik} = \left(\frac{x_{ik} - a}{b} \right)^p \tag{8.6}$$

gdzie dla $i = 1, 2, \ldots, N$; $k = 1, 2, \ldots, K$:
z_{ik} – znormalizowana wartość k-tej zmiennej w i-tym obiekcie,
x_{ik} – wartość (obserwacja) k-tej zmiennej diagnostycznej dla i-tej jednostki (obiektu),
a, b, p – parametry normalizacyjne $b \neq 0$.

Popularnymi metodami normalizacji zmiennych są: standaryzacja, unitaryzacja i przekształcenie ilorazowe.

Celem standaryzacji jest otrzymanie zmiennych o zerowej wartości średniej i odchyleniu standardowym równym 1. Przeprowadza się ją zgodnie ze wzorem[2]:

$$z_{ik} = \frac{x_{ik} - \overline{x}_k}{S_k} \tag{8.7}$$

gdzie dla $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, K$:
$\overline{x}_k$ – średnia arytmetyczna k-tej zmiennej diagnostycznej x_k,
S_k – odchylenie standardowe k-tej zmiennej diagnostycznej x_k,
pozostałe oznaczenia jak poprzednio.

Unitaryzacja polega na takim przekształceniu, które pozwoli na uzyskanie zmiennych o ujednoliconym zakresie zmienności, definiowanym przez różnicę pomiędzy ich wartościami maksymalnymi i minimalnymi, czyli

$$z_{ik} = \frac{x_{ik} - \min\limits_i \{x_{ik}\}}{\max\limits_i \{x_{ik}\} - \min\limits_i \{x_{ik}\}} \tag{8.8}$$

gdzie dla $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, K$:
$\max\limits_i \{x_{ik}\}, \min\limits_i \{x_{ik}\}$ – to odpowiednio maksymalna i minimalna wartość k-tej zmiennej diagnostycznej dla i-tej jednostki (obiektu),
a pozostałe oznaczenia jak poprzednio.

2 Jest to ta sama procedura, którą opisano wzorem (1.9), omawiając standaryzację rozkładu normalnego.

Natomiast przekształcenie ilorazowe pozwala na odniesienie zmiennej diagnostycznej do pewnej stałej i jest postaci:

$$z_{ik} = \frac{x_{ik}}{\overline{x}_k} \tag{8.9}$$

Odległość między obiektami jest jedną z miar wykorzystywaną do porównywania obiektów wielocechowych. Jeśli oznaczyć odległości między obiektami jako:

$$d_{ij} = d(O_i, O_j) \tag{8.10}$$

to odległości pomiędzy wszystkimi parami obiektów O_i oraz O_j $i, j = 1, 2, ..., N$ zapisać można w postaci macierzy:

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1N} \\ d_{21} & 0 & \cdots & d_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ d_{N1} & d_{N2} & \cdots & 0 \end{bmatrix} \tag{8.11}$$

gdzie dla $i = 1, 2, \ldots, n$:
d_{ij} – odległość obiektu i-tego od obiektu j-tego,
D – macierz odległości.

Odległość między obiektami posiada następujące własności:

- jest nieujemna, tzn. odległość między dwoma obiektami jest dodatnia dla różnych obiektów i równa 0 dla tego samego obiektu, co można zapisać jako:

$$d_{ij} \geq 0 \tag{8.12}$$

$$d_{ii} = 0 \tag{8.13}$$

- jest symetryczna, tzn. odległość między obiektami i oraz j jest identyczna jak między obiektami j oraz i, czyli:

$$d_{ij} = d_{ji} \tag{8.14}$$

- spełnia tzw. zasadę trójkąta, która polega na tym, że odległość między obiektami i oraz k jest niemniejsza niż suma odległości między obiektami i oraz j i j oraz k, czyli zachodzi:

$$d_{ik} \leq d_{ij} + d_{jk} \tag{8.15}$$

Najczęściej odległości między obiektami[3], opisywanymi przez cechy ilościowe, wyznacza się, korzystając z metryki euklidesowej. Jest to tzw. odległość euklidesowa, którą zapisuje się jako:

$$d_{ij} = \sqrt{\sum_{k=1}^{K}\left(z_{ik} - z_{jk}\right)^2} = \left[\sum_{k=1}^{K}\left(z_{ik} - z_{jk}\right)^2\right]^{0,5} \tag{8.16}$$

gdzie:
k – oznacza numery cech diagnostycznych, po których przebiega sumowanie, $k = 1, 2, \ldots, K$,
z_{ik}, z_{jk} – oznaczają znormalizowane wartości tych cech dla obiektów $i, j = 1, 2, ..., N$

Przedstawiona miara odległości (8.16) pozwala określić odległość dwóch obiektów między sobą. W praktyce wygodnie jest porównywać obiekty w relacji do jednego uniwersalnego obiektu będącego wzorcem[4] dla wszystkich porównywanych. Odległość euklidesowa od pewnego wyznaczonego wzorca jest postaci:

$$d_i = d_{i0} = \sqrt{w_k \cdot \sum_{k=1}^{K}(z_{ik} - z_{0k})^2} \tag{8.17}$$

gdzie:
d_i – odległość obiektu i-tego od wzorca,
z_{ik} – zmienne po normalizacji opisujące i-ty obiekt,
z_{0k} – zmienne opisujące wzorzec,
w_k – wagi określające „ważność" k-tej cechy (zmiennej diagnostycznej).

W badaniach często przyjmuje się, że wagi dla poszczególnych zmiennych diagnostycznych nie są zróżnicowane i wynoszą 1 lub $1/K$ dla każdej zmiennej, gdzie K oznacza liczbę zmiennych diagnostycznych. Oczywiście w przypadku wag równych jedności wzór (8.17) na odległość od wzorca redukuje się do:

$$d_i = d_{i0} = \sqrt{\sum_{k=1}^{K}\left(z_{ik} - z_{0k}\right)^2} \tag{8.18}$$

3 W literaturze przedmiotu, oprócz opisanych wzorami (8.16)–(8.18) i (8.20) wyróżnia się także inne odległości, tj. odległość Minkowskiego, Czebyszewa czy Mahalanobisa. Ta ostania jest postaci: $d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)\right]^{0,5}$, gdzie: **W** – macierz dodatnio określona. W praktyce jako **W** najczęściej przyjmuje się macierz odwrotną do macierzy kowariancji wektora losowego $\mathbf{x}_n$.

4 Można oczywiście porównywać obiekty do antywzorca.

Wzorzec może być obiektem rzeczywiście istniejącym albo hipotetycznym. W tym drugim przypadku konstruuje się go jako obiekt opisany maksymalnymi wartościami stymulant i minimalnymi wartościami destymulant, czyli:

$$z_{0k} = \begin{cases} \min\limits_{i}\{z_{ik}\} & dla \quad x_{ik}^{D} \\ \max\limits_{i}\{z_{ik}\} & dla \quad x_{ik}^{S} \end{cases} \tag{8.19}$$

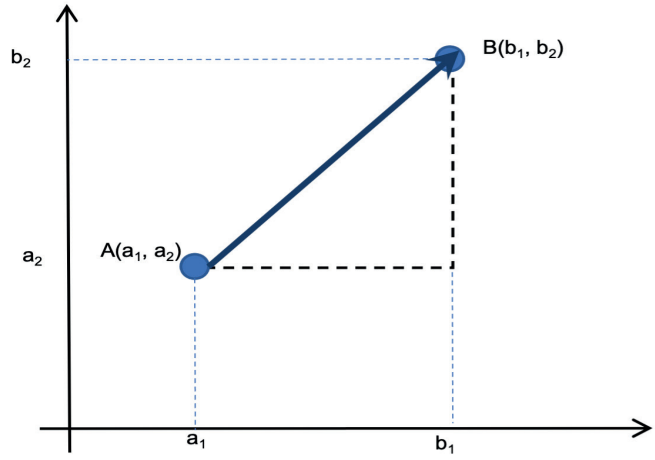
gdzie:
x_{ik}^{S}, x_{ik}^{D} – odpowiednio stymulanta lub destymulanta.

Innym sposobem pomiaru odległości jest tzw. odległość miejska, określana też odległością Manhattan:

$$d_{ij} = \sum_{k=1}^{K} \left| z_{ik} - z_{jk} \right| \tag{8.20}$$

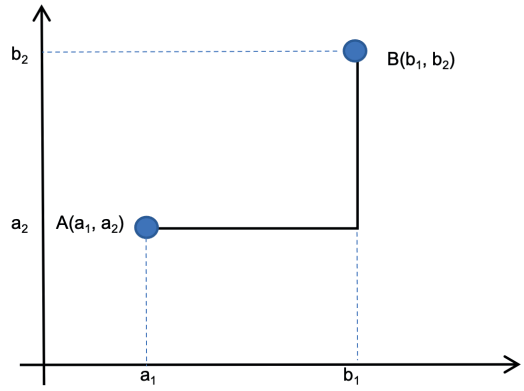
Stosując miary odległości do porównań obiektów, można dostrzec, że są one tym bardziej do siebie podobne im odległość między nimi jest mniejsza. W szczególności obiekt jest tym wyżej oceniany, im jego odległość od wzorca jest mniejsza.

Jak widać we wzorach (8.16)–(8.18) i (8.20), w przypadku stosowania odległości „miejskiej" lub euklidesowej trzeba najpierw przeprowadzić normalizację (najczęściej standaryzację (8.7)) zaobserwowanych wartości zmiennych. Dzięki czemu możliwe będzie wykonanie działań opisanych tymi wzorami. Porównanie odległości euklidesowej i miejskiej przedstawiono na rys. 8.1 i 8.2.



Rysunek 8.1. Odległości punktów na płaszczyźnie według metryki euklidesowej

Źródło: opracowanie własne.



Rysunek 8.2. Odległości punktów na płaszczyźnie według metryki miejskiej

Źródło: opracowanie własne.

W niniejszym podrozdziale opisane zostaną dwa rodzaje mierników syntetycznych:

1) ze wzorcem, które bazują na odległości taksonomicznej (od przyjętego wzorca rozwoju), wyznaczonej dla wystandaryzowanych zmiennych oraz
2) bez wzorca, które wyznacza się jako sumę unormowanych zmiennych diagnostycznych.

Syntetyczny miernik rozwoju, opracowany przez Zdzisława Hellwiga [1968], jest miarą porządkowania liniowego ze wzorcem. Zazwyczaj jest on zapisywany w postaci:

$$SMR_i = 1 - \frac{d_i}{\overline{d} + 2 \cdot S_d} \tag{8.21}$$

gdzie dla $i = 1, 2, \ldots, N$:
d_i – odległość i-tego obiektu od wzorca wg metryki euklidesowej (8.17) lub (8.18),
$\overline{d}$ – średnia odległości d_i:

$$\overline{d} = \frac{1}{N} \sum_{i=1}^{N} d_i \tag{8.22}$$

S_d – odchylenie standardowe odległości d_i:

$$S_d = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (d_i - \overline{d})^2} \tag{8.23}$$

Miernik (8.21) jest unormowany i zazwyczaj przyjmuje wartość z przedziału [0, 1], im obiekt jest położony bliżej wzorca, tzn. wartość miernika bliższa jest jedności, tym pozycja obiektu w rankingu jest lepsza. Jako przykład aplikacji syntetycznego miernika rozwoju w analizach finansowych można podać zaproponowaną przez Tarczyńskiego [1994] taksonomiczną miarę atrakcyjności inwestycji (TMAI) do kwantyfikacji siły fundamentalnej spółek giełdowych.

W przypadku mierników bez wzorca istotne jest sprowadzenie wszystkich zmiennych diagnostycznych nie tylko do porównywalności, ale również konieczne jest ujednolicenie wpływu zmiennej na obiekt. Służy temu transformacja destymulant na stymulanty (in. stymulacja destymulant), którą przeprowadza się przed wyznaczeniem miernika. Najprostsze transformacje tego typu to przekształcenia:

- ilorazowe:

$$x_{ik}^{S} = c \cdot \frac{1}{x_{ik}^{D}} \tag{8.24}$$

- różnicowe:

$$x_{ik}^{S} = a - b \cdot x_{ik}^{D} \tag{8.25}$$

gdzie:
x_{ik}^{S} – wartość k-tej zmiennej stymulanty w i-tym obiekcie,
x_{ik}^{D} – wartość k-tej zmiennej destymulanty w i-tym obiekcie,
a, b, c – arbitralnie przyjmowane parametry przekształceń ($c > 0$), przy czym najczęściej: $a = 0$ lub $a = \max_i x_{ik}^{D}$, $b = 1$ i $c = 1$.

Dla obiektów, które opisywane są za pomocą nominant, możliwa jest tzw. stymulacja nominant, przeprowadzana za pomocą transformacji:

- ilorazowej:

$$x_{ik}^{S} = \frac{\min_i \left\{ x_k^N ; x_{ik}^N \right\}}{\max_i \left\{ x_k^N ; x_{ik}^N \right\}} \tag{8.26}$$

- różnicowej:

$$x_{ik}^{S} = -\left| x_{ik}^{N} - x_k^{N} \right| \tag{8.27}$$

gdzie:
x_{ik}^{N} – wartość k-tej zmiennej nominanty w i-tym obiekcie,
x_k^{N} – nominalna (pożądana) wartość k-tej zmiennej.

Absolutne mierniki rozwoju budowane są jako ważone sumy znormalizowanych zmiennych diagnostycznych i do ich wyznaczenia nie wykorzystuje się żad-

nego wzorca. W konsekwencji zmienne muszą charakteryzować się takim samym kierunkiem oddziaływania na badane obiekty. Najczęściej sprowadza się je do postaci stymulant. Wtedy wartość absolutnego miernika rozwoju wynosi:

$$AMR_i = \sum_{k=1}^{K} w_k \cdot z_{ik}^S \tag{8.28}$$

gdzie:
z_{ik}^S – znormalizowane wartości stymulanty k-tej w i-tym obiekcie,
w_k – waga przypisana k-tej stymulancie.

Dla wag równych jedności absolutny miernik rozwoju (8.28) nosi również nazwę sum standaryzowanych i jest postaci:

$$AMR_i = \sum_{k=1}^{K} z_{ik}^S = \sum_{k=1}^{K} \frac{x_{ik}^S - \overline{x}_k}{S_k} \tag{8.29}$$

W przypadku mierników tworzonych na podstawie stymulant zachodzi oczywista relacja – czym wyższa wartość miernika dla danego obiektu badania, tym lepsza pozycja obiektu w rankingu. Zastosowanie różnych metod normalizacji (8.6)–(8.9) prowadzi do różnych postaci miernika. I tak zastosowanie unitaryzacji prowadzi do znanej formuły[5]:

$$UMR_i = \frac{1}{K} \sum_{k=1}^{K} \frac{x_{ik}^S - \min(x_{ik})}{\max(x_{ik}) - \min(x_{ik})} \tag{8.30}$$

Inne przykłady często używanych mierników to:

$$ASM_i = \sum_{k=1}^{K} \frac{x_{ik}^S}{S_k} \tag{8.31}$$

$$BSM_i = \sum_{k=1}^{K} \frac{x_{ik}^S}{\max(x_{ik})} \tag{8.32}$$

Po wyznaczeniu wartości mierników dla wszystkich analizowanych obiektów można przejść do:

1) tworzenia rankingu obiektów (porządkowania obiektów), polegającego na ich uszeregowaniu według wartości mierników,
2) grupowania obiektów w homogeniczne klasy.

5 Na podstawie miernika (8.30) wyznaczany był przez ONZ do 2010 roku wskaźnik rozwoju społecznego (*Human Development Index*), na podstawie trzech kategorii: „długie i zdrowe życie" (*long and healthy life*), „wiedza" (*knowledge*) i „dostatni standard życia" (*decent standard of living*).

W celu grupowania obiektów badania należy utworzyć kryteria przynależności do poszczególnych klas. Istnieje wiele możliwości dotyczących zarówno liczby klas, jak i wyboru kryterium klasyfikacji do poszczególnych grup, które są związane z celem badania. Klasycznym podejściem jest podział na cztery grupy typologiczne pozwalające na ocenę poszczególnych obiektów. Przykładowo:

- klasa I (jednostki bardzo dobre), jeśli:

$$S_i \geq S + SS \tag{8.33}$$

- klasa II (jednostki dobre), jeśli:

$$S + SS > S_i \geq S \tag{8.34}$$

- klasa III (jednostki przeciętne), jeśli:

$$S > S_i \geq S - SS \tag{8.35}$$

- klasa IV (jednostki słabe), jeśli:

$$S_i < S - SS \tag{8.36}$$

gdzie:
S_i – wartość (dowolnego) miernika syntetycznego wyznaczona dla i-tego obiektu,
S i SS – odpowiednio średnia i odchylenie standardowe wartości mierników syntetycznych obliczonych dla wszystkich obiektów.

Przedstawiony kierunek nierówności (8.33)–(8.36) odnosi się do sytuacji, kiedy większa wartość miernika syntetycznego świadczy o lepszej pozycji obiektu. W sytuacji przeciwnej wszystkie relacje są odwrócone.

W innym podejściu (do podziału na cztery klasy) dokonuje się wstępnego podziału wszystkich obiektów na dwie klasy według wartości średniej arytmetycznej wyznaczonego miernika, a następnie wyznacza się średnie arytmetyczne każdej z podgrup.

Warto w tym miejscu podkreślić, że prezentowane w niniejszym podrozdziale mierniki zostały zbudowane na podstawie klasycznych miar służących do opisu obiektów badania. Jednakże często okazuje się, że poszczególne cechy diagnostyczne mają rozkłady empiryczne skrajnie asymetryczne i wtedy, w miejsce miar klasycznych, zaleca się stosowanie miar pozycyjnych (tj. kwartylów i odchylenia ćwiartkowego). Wówczas również podział na klasy przeprowadzany jest na podstawie miar pozycyjnych.

Zachodzi pytanie, czy za pomocą danego miernika można poprawnie rozpoznać obiekty z punktu widzenia ich rozwoju. Do tego służy miara właściwości dyskryminacyjnych mierników syntetycznych (*MWDMS*), którą wykorzystuje się do oceny przydatności mierników taksonomicznych przy klasyfikacji obiektów. Wyznaczana jest ona dla uporządkowanych malejąco mierników syntetycznych jako [Sokołowski 1985]:

$$MWDMS = 1 - \sum_{i=1}^{N-1} \min\left\{\frac{S_i - S_{i+1}}{R}, \frac{1}{N-1}\right\} = 1 - \sum_{i=1}^{N-1} \min\{W_1, W_2\} \tag{8.37}$$

gdzie dla $i = 1, 2, \ldots, N-1$:
S_i, S_{i+1} – kolejne, uporządkowane wartości miernika taksonomicznego,
R – rozstęp (2.10) liczony jako różnica między maksymalną S_{max} a minimalną S_{min} wartością wyznaczonych mierników:

$$R = S_{max} - S_{min} \tag{8.38}$$

Wartość wskaźnika *MWDMS* zawiera się w przedziale $\left[0,\ 1 - \frac{1}{N-1}\right]$.

Przyjmuje wartość 0, gdy dla każdego i-tego obiektu różnice są jednakowe. Wartość maksymalna wystąpi, gdy dla $(N-1)$ obiektów wartości miernika będą równe, i tylko jeden obiekt przyjmie inną wartość, różniąc się od pozostałych. Dlatego też postulowana wartość tego wskaźnika znajduje się w środku jego przedziału zmienności. Przyjmuje się, że im wartość miernika (8.38) większa, tym wyższa zdolność miernika taksonomicznego do wyodrębniania grup typologicznych dla analizowanych obiektów (por. [Kompa 2009; Kądziołka 2021]). Do porównań zdolności dyskryminacyjnych mierników taksonomicznych można również użyć współczynnika zmienności postaci:

$$V_s = \frac{SS}{S} \tag{8.39}$$

gdzie oznaczenia jak poprzednio.

Przykład 8.1

Badaniu poddano działalność 10 firm usługowych, które oceniano wziąwszy pod uwagę 4 kryteria: (a) średnią ocenę klientów za świadczone usługi, (b) liczbę zgłoszonych reklamacji, (c) liczbę nowo pozyskanych klientów, którzy podpisali długoterminową umowę i (d) najwyższą kwotę prowizji uzyskaną za pojedynczą usługę. W tab. 8.2 podano charakterystyki każdego przedsiębiorstwa. Określić należy ranking firm z punktu widzenia pojedynczych cech i wszystkich cech jednocześnie.

Tabela 8.2. Cechy badanych przedsiębiorstw

Firmy	Średnia ocena klientów (a)	Liczba zgłoszonych reklamacji (b)	Liczba nowo pozyskanych klientów (c)	Najwyższa kwota prowizji (d)
A	4,56	11	25	1 375
B	4,67	7	28	546
C	4,23	8	14	345

Tabela 8.2 (cd.)

Firmy	Średnia ocena klientów (a)	Liczba zgłoszonych reklamacji (b)	Liczba nowo pozyskanych klientów (c)	Najwyższa kwota prowizji (d)
D	4,32	13	35	2 345
E	4,90	23	17	123
F	5,00	17	23	765
G	4,87	15	19	456
H	4,12	5	29	256
I	4,01	4	37	2 901
J	4,98	12	18	1 234

Źródło: dane umowne.

Rozwiązanie

Zauważyć trzeba, że cechy (a), (c) i (d) są stymulantami, a (b) jest destymulantą. Zatem ranking firm usługowych został przeprowadzony w taki sposób, że pierwsze miejsce w rankingu zajmuje to przedsiębiorstwo, które charakteryzuje się najwyższą wartością jednej ze stymulant i najniższą liczbą reklamacji. W tab. 8.3 przedstawiono pozycje rankingowe każdej z firm, biorąc pod uwagę wartości pojedynczej cechy.

Tabela 8.3. Pozycje rankingowe poszczególnych przedsiębiorstw ustalone dla każdej cechy oddzielnie

Firmy	Cecha (a)	Cecha (b)	Cecha (c)	Cecha (d)	Suma rang	Pozycja w rankingu
A	6	5	5	3	19	4
B	5	3	4	6	18	2–3
C	8	4	10	8	30	9
D	7	7	2	2	18	2–3
E	3	10	9	10	32	10
F	1	9	6	5	21	6
G	4	8	7	7	26	8
H	9	2	3	9	23	7
I	10	1	1	1	13	1
J	2	6	8	4	20	5

Źródło: opracowanie własne.

Jak można łatwo zauważyć w tab. 8.3, w zależności od analizowanej cechy, kolejność poszczególnych przedsiębiorstw jest zróżnicowana. Innymi słowy,

mając jednocześnie kilka kryteriów, trudno jest dokonać jednoznacznej oceny. Można jednak zauważyć, że firma I znalazła się na pierwszej pozycji aż w 3 kategoriach oceny, chociaż na ostatniej pozycji z punktu widzenia cechy (a). Natomiast firma E jest na ostatnim miejscu w dwóch kategoriach (b) i (d). Jeśli zsumować pozycje rankingowe we wszystkich kategoriach, jak to zrobiono w przedostatniej kolumnie, to można zbudować ranking firm, uwzględniając wszystkie kategorie oceny. Zapisano to w ostatniej kolumnie tab. 8.3, przyjmując, że najlepszym przedsiębiorstwem jest to o najniższej sumie rang. Zgodnie z wcześniejszymi analizami firma E uplasowała się na ostatniej pozycji jako najgorsza, a firma I na pierwszej jako najlepsza. Jednocześnie dwie z badanych firm (B i D) mają identyczną sumę rang i znalazły się *ex aequo* na tych samych pozycjach w rankingu.

W celu oceny przedsiębiorstw z punktu widzenia wszystkich cech jednocześnie dla wszystkich firm obliczone zostaną 4 mierniki taksonomiczne: (a) sumy standaryzowane (8.29), (b) sumy zmiennych po zastosowaniu unitaryzacji jako reguły normalizacyjnej (8.30), (c) odległości euklidesowe od wzorca (8.18) i (d) syntetyczne mierniki rozwoju (8.21). Trzeba rozpocząć od standaryzacji wszystkich zmiennych, do czego wykorzystane zostaną obliczone średnie arytmetyczne i odchylenia standardowe (tab. 8.4).

Tabela 8.4. Wyznaczenie średnich i odchyleń standardowych cech

Firmy	Cecha (a)	Cecha (b)	Cecha (c)	Cecha (d)
A	4,56	11	25	1 375
…	…	…	…	…
J	4,98	12	18	1 234
Średnia	4,5660	11,5000	24,5000	1 034,6000
Odchylenie standardowe	0,3544	5,5543	7,3519	889,6742

Źródło: obliczenia własne.

Można teraz, korzystając ze wzoru (8.7), obliczyć wartości zmiennych standaryzowanych. Tak więc dla średniej ocen w firmie A wartość wystandaryzowana tej zmiennej wyniesie: $z_{ik} = \frac{x_{ik} - \overline{x}_k}{S_k} = \frac{4,56 - 4,566}{0,3544} = -0,0169$, co wpisano w tab. 8.5 w drugiej kolumnie (dla cechy (a)) i drugim wierszu (dla firmy A). W ten sam sposób przekształceniu ulegną wartości tej cechy dla pozostałych firm i dalej standaryzacji poddane zostaną kolejne zmienne. Zauważyć trzeba, że średnie wystandaryzowanych zmiennych wynoszą 0, a odchylenie standardowe 1, co zostało zapisane w odpowiednich wierszach tabeli. Warto także odnotować, że zmienne po standaryzacji są pozbawione mian, więc można je poddawać sumowaniu.

W kolejnym kroku wyznaczony zostanie wzorzec według (8.19), czyli określone będą wartości maksymalne stymulant i minimalna wartość destymulanty, co pokazano w dwóch wierszach tab. 8.5, oznaczonych jako max i min. Wartości wzorca zaznaczono pogrubioną czcionką i zapisano w ostatnim wierszu. Innymi słowy, na podstawie wartości ekstremalnych wszystkich cech utworzono hipotetyczny obiekt, który charakteryzuje się najlepszymi wartościami cech i stanowi wzorzec do porównań wszystkich firm. Zatem w dalszym postępowaniu wyznaczone zostaną odległości od wzorca według wzoru (8.18). Tab. 8.6 zawiera różnice wartości wystandaryzowanych zmiennych i wzorca, zamieszczonych w tab. 8.5. W tab. 8.7 różnice te zostały podniesione do kwadratu i zsumowane. W kolejnej kolumnie wyznaczono odległości jako pierwiastek kwadratowy z obliczonych sum i podano pozycje rankingowe firm.

Tabela 8.5. Zmienne po standaryzacji

Firmy	Cecha (a)	Cecha (b)	Cecha (c)	Cecha (d)
A	–0,0169	–0,0900	0,0680	0,3826
B	0,2934	–0,8102	0,4761	–0,5492
C	–0,9481	–0,6301	–1,4282	–0,7751
D	–0,6941	0,2701	1,4282	1,4729
E	0,9424	2,0705	–1,0201	–1,0246
F	1,2246	0,9902	–0,2040	–0,3030
G	0,8578	0,6301	–0,7481	–0,6504
H	–1,2584	–1,1703	0,6121	–0,8752
I	–1,5688	–1,3503	1,7002	2,0978
J	1,1681	0,0900	–0,8841	0,2241
Średnia	0,0000	0,0000	0,0000	0,0000
Odchylenie standardowe	1,0000	1,0000	1,0000	1,0000
Max	**1,2246**	2,0705	**1,7002**	**2,0978**
Min	–1,5688	**–1,3503**	–1,4282	–1,0246
Wzorzec	1,2246	–1,3503	1,7002	2,0978

Źródło: obliczenia własne.

Wykorzystując odległości od wzorca do porządkowania obiektów, należy pamiętać, że ten obiekt jest na lepszej pozycji w rankingu, dla którego odległość jest mniejsza. Zatem przedsiębiorstwa zostały ponownie uporządkowane od najlepszego (pierwsza pozycja) do najgorszego (pozycja dziesiąta). Jak widać w tab. 8.7,

wprawdzie na ostatniej pozycji wciąż znajduje się firma E, ale pierwszą pozycję w tym rankingu zajmuje firma D, a firma I znalazła się na pozycji drugiej. W stosunku do poprzedniego uszeregowania przedsiębiorstw zaszło zatem nieznaczne „przetasowanie" na pozycjach rankingowych.

Tabela 8.6. Różnice wartości wystandaryzowanych zmiennych od wzorca

Firma	Cecha (a)	Cecha (b)	Cecha (c)	Cecha (d)
A	–1,2415	1,2603	–1,6322	–1,7152
B	–0,9311	0,5401	–1,2242	–2,6470
C	–2,1726	0,7202	–3,1285	–2,8730
D	–1,9187	1,6204	–0,2720	–0,6249
E	–0,2822	3,4208	–2,7204	–3,1225
F	0,0000	2,3405	–1,9043	–2,4009
G	–0,3668	1,9805	–2,4484	–2,7482
H	–2,4830	0,1800	–1,0882	–2,9730
I	–2,7934	0,0000	0,0000	0,0000
J	–0,0564	1,4403	–2,5844	–1,8737

Źródło: obliczenia własne.

Tabela 8.7. Wyznaczenie odległości euklidesowych

Firma	Kwadraty różnic od wzorca dla cech:				Suma kwadratów	Odległość euklidesowa	Pozycja w rankingu
	(a)	(b)	(c)	(d)			
A	1,5414	1,5883	2,6642	2,9420	8,7359	2,9557	3
B	0,8670	0,2917	1,4986	7,0068	9,6642	3,1087	4
C	4,7204	0,5186	9,7872	8,2539	23,2802	4,8250	9
D	3,6814	2,6256	0,0740	0,3906	6,7716	2,6022	1
E	0,0796	11,7018	7,4006	9,7500	28,9319	5,3788	10
F	0,0000	5,4781	3,6263	5,7642	14,8686	3,8560	6
G	0,1345	3,9222	5,9944	7,5526	17,6038	4,1957	8
H	6,1654	0,0324	1,1841	8,8387	16,2206	4,0275	7
I	7,8031	0,0000	0,0000	0,0000	7,8031	2,7934	2
J	0,0032	2,0746	6,6790	3,5108	12,2676	3,5025	5

Źródło: obliczenia własne.

W celu wyznaczenia syntetycznego miernika rozwoju SMR należy unormować miarę odległości. W tym celu wyznacza się średnią (8.22) i odchylenie standardowe (8.23) odległości obliczonych i wpisanych do tab. 8.7. Wynoszą one odpowiednio: 3,7245 i 0,8617. W tab. 8.8 przedstawiono obliczone wartości miernika SMR i pozycje rankingowe poszczególnych przedsiębiorstw. Należy zauważyć, że ten ranking firm jest tożsamy z przedstawionym w tab. 8.7 z tą różnicą, że w przypadku SMR większa wartość wskaźnika świadczy o lepszej pozycji rankingowej firmy. Jednocześnie korzystając z wyznaczonych wartości SMR, skonstruowano 4 homogeniczne klasy firm zgodnie ze wzorami (8.33)–(8.36). Jak widać, w pierwszej klasie znalazły się 2 najlepsze firmy D i I, a w ostatniej firmy C i E.

Tabela 8.8. Wartości mierników SMR i grupowanie firm

Klasyfikacja	Firma	Odległość euklidesowa	SMR	Pozycja w rankingu	Grupowanie	Firma	SMR	Klasa
	A	2,9557	0,4575	3		D	0,5223	I
	B	3,1087	0,4294	4		I	0,4873	
	C	4,8250	0,1143	9		A	0,4575	II
	D	2,6022	0,5223	1		B	0,4294	
	E	5,3788	0,0127	10		J	0,3571	
	F	3,8560	0,2922	6		F	0,2922	III
	G	4,1957	0,2299	8		H	0,2607	
	H	4,0275	0,2607	7		G	0,2299	
	I	2,7934	0,4873	2		C	0,1143	IV
	J	3,5025	0,3571	5		E	0,0127	
	Średnia	3,7245	0,3163	×				
	Odchylenie standardowe	0,8617	0,1582	×				

Źródło: obliczenia własne.

W kolejnym etapie analiz obliczyć należy absolutne mierniki rozwoju (sumy standaryzowane) AMR (8.29). Zanim jednak do tego dojdzie, trzeba wystymulować destymulantę, stosując transformacje (8.24) lub (8.25). Tu wykorzystano uproszczone przekształcenia różnicowe i ilorazowe postaci odpowiednio: (*) $z_{ik}^S = -z_{ik}^D$ lub (**) $z_{ik}^S = \frac{1}{z_{ik}^D}$, gdzie: z_{ik}^S, z_{ik}^D oznaczają wystandaryzowane zmienne, odpowiednio stymulanty i destymulanty. Otrzymane wyniki zapisano w tab. 8.9 dla transformacji (*) oraz w tab. 8.10 dla przekształcenia (**). W obu tabelach podano obliczone wartości sum standaryzowanych i utworzone na tej podstawie rankingi spółek.

Tabela 8.9. Obliczanie sum standaryzowanych dla różnicowego przekształcenia destymulanty

Firma	Wystandaryzowane wartości stymulant, tj. cech:				Miernik AMR(*)	Pozycja w rankingu
	(a)	(b*)	(c)	(d)		
A	–0,0169	0,0900	0,0680	0,3826	0,5237	4
B	0,2934	0,8102	0,4761	–0,5492	1,0305	3
C	–0,9481	0,6301	–1,4282	–0,7751	–2,5212	9
D	–0,6941	–0,2701	1,4282	1,4729	1,9369	2
E	0,9424	–2,0705	–1,0201	–1,0246	–3,1728	10
F	1,2246	–0,9902	–0,2040	–0,3030	–0,2727	6
G	0,8578	–0,6301	–0,7481	–0,6504	–1,1708	8
H	–1,2584	1,1703	0,6121	–0,8752	–0,3512	7
I	–1,5688	1,3503	1,7002	2,0978	3,5796	1
J	1,1681	–0,0900	–0,8841	0,2241	0,4181	5

Źródło: obliczenia własne.

Tabela 8.10. Obliczanie sum standaryzowanych dla ilorazowego przekształcenia destymulanty

Firma	Wystandaryzowane wartości stymulant, tj. cech:				Miernik AMR(**)	Pozycja w rankingu
	(a)	(b**)	(c)	(d)		
A	0,0169	–11,1086	0,0680	0,3826	–10,6749	10
B	0,2934	–1,2343	0,4761	–0,5492	–1,0140	7
C	–0,9481	–1,5869	–1,4282	–0,7751	–4,7383	9
D	–0,6941	3,7029	1,4282	1,4729	5,9098	2
E	0,9424	0,4830	–1,0201	–1,0246	–0,6194	6
F	1,2246	1,0099	–0,2040	–0,3030	1,7274	3
G	0,8578	1,5869	–0,7481	–0,6504	1,0462	5
H	–1,2584	–0,8545	0,6121	–0,8752	–2,3760	8
I	–1,5688	–0,7406	1,7002	2,0978	1,4887	4
J	1,1681	11,1086	–0,8841	0,2241	11,6167	1

Źródło: obliczenia własne.

Analizując otrzymane wyniki, można zauważyć, że ranking przedsiębiorstw z tab. 8.9 jest niemal identyczny z porządkiem w tab. 8.3. Natomiast ranking prezentowany w tab. 8.10 jest całkowicie różny, co jedynie świadczy o tym, że w analizach wielowymiarowych nie tylko istotny jest wybór miernika, ale i sposób przekształcania zmiennych.

Istnieje oczywiście możliwość wyznaczenia mierników dla innego sposobu normalizacji zmiennych, np. stosując unitaryzację (8.8) i miernik (8.30). W tym przypadku konieczne jest sprowadzenie zmiennych do postaci stymulant, co wynika

z bezwzorcowego charakteru miernika i konieczności sumowania wszystkich zmiennych. W analizowanym przykładzie destymulantą jest jedynie zmienna (b), która przekształcona zostanie w stymulantę wg transformacji różnicowej (*) $z_{ik}^{S} = - z_{ik}^{D}$. Tak utworzony zbiór wartości zmiennych poddaje się unitaryzacji (8.8) i dalej dla każdej firmy oblicza się wartość miernika (8.30), który będzie podstawą porządkowania firm, co przedstawiono w tab. 8.11.

Tabela 8.11. Budowa miernika – sumy zunitaryzowanych zmiennych (8.30)

Firma	Zunitaryzowane stymulanty, tj. cechy				Miernik UMR	Pozycja w rankingu
	(a)	(b*)	(c)	(d)		
A	0,5556	0,6316	0,4783	0,4507	0,5290	5
B	0,6667	0,8421	0,6087	0,1523	0,5674	3
C	0,2222	0,7895	0,0000	0,0799	0,2729	9
D	0,3131	0,5263	0,9130	0,7999	0,6381	2
E	0,8990	0,0000	0,1304	0,0000	0,2574	10
F	1,0000	0,3158	0,3913	0,2311	0,4845	6
G	0,8687	0,4211	0,2174	0,1199	0,4068	8
H	0,1111	0,9474	0,6522	0,0479	0,4396	7
I	0,0000	1,0000	1,0000	1,0000	0,7500	1
J	0,9798	0,5789	0,1739	0,3999	0,5331	4

Źródło: obliczenia własne.

Tabela 8.12. Porównanie rankingów utworzonych za pomocą różnych miar

Firma	Numery tabel, w których utworzono rankingi					
	8.3. Suma rang	8.7. Odległość	8.8. SMR	8.9. AMR(*)	8.10. AMR(**)	8.11. UMR
A	4	3	3	4	10	5
B	2,5	4	4	3	7	3
C	9	9	9	9	9	9
D	2,5	1	1	2	2	2
E	10	10	10	10	6	10
F	6	6	6	6	3	6
G	8	8	8	8	5	8
H	7	7	7	7	8	7
I	1	2	2	1	4	1
J	5	5	5	5	1	4

Uwaga: ponieważ w tab. 8.2 dwie firmy znajdowały się na tych samych pozycjach, zapisano je na pozycji 2,5.

Źródło: obliczenia własne.

Utworzony za pomocą miernika (8.30) ranking firm (tab. 8.11) jest zbliżony do większości wcześniejszych rankingów. Ich porównanie przedstawiono w tab. 8.12. Jedynie miernik syntetyczny (sumy standaryzowane dla ilorazowego przekształcenia destymulanty) wygenerował zasadniczo inną kolejność analizowanych przedsiębiorstw. W przypadku miary odległości i miernika SMR kolejność firm jest identyczna, bowiem oba mierniki bazują na odległości euklidesowej, a jedyną różnicę stanowi sposób porządkowania obiektów. Nasuwa się w tym miejscu pytanie, który z mierników jest najlepszy, co sprowadza się do zbadania ich właściwości dyskryminacyjnych za pomocą miernika *MWDMS* (8.37) oraz współczynnika zmienności (8.39).

Tabela 8.13. Badanie własności dyskryminacyjnych wyznaczonych mierników

Pozycja w rankingu i	Numery tabel, w których wyznaczono mierniki					
	8.3. Suma rang	8.7. Odległość	8.8. SMR	8.9. AMR(*)	8.10. AMR(**)	8.11. UMR
	Obliczone wartości wyrażenia: $W_1 = (S_i - S_{i+1})/R$					
1	0,1053	0,1995	0,0689	0,2433	0,2560	0,2272
2	0,2105	0,2266	0,0584	0,1342	0,1876	0,1434
3	0,1579	0,0606	0,0551	0,0751	0,0107	0,0696
4	0,1053	0,0618	0,1418	0,0156	0,0198	0,0084
5	0,0526	0,1273	0,1273	0,1023	0,0747	0,0903
6	0,0526	0,1418	0,0618	0,0116	0,0177	0,0912
7	0,0526	0,0551	0,0606	0,1214	0,0611	0,0667
8	0,0000	0,0584	0,2266	0,2000	0,1060	0,2717
9	0,2632	0,0689	0,1995	0,0965	0,2663	0,0316
max	32	5,3788	0,5223	3,5796	11,6167	0,7500
min	13	2,6022	0,0127	–3,1728	–10,6749	0,2574
Rozstęp	19	2,7766	0,5097	6,7524	22,2916	0,4926
Suma	0,7018	0,7492	0,7492	0,7456	0,6234	0,6911
MWDMS	0,2982	0,2508	0,2508	0,2544	0,3766	0,3089
V	25,23%	23,14%	50,00%	2,67E+15	2387,07%	29,68%

Uwaga: $Suma = \sum_{i=1}^{N-1} \min\left\{\frac{S_i - S_{i+1}}{R}, \frac{1}{N-1}\right\}$, kolor szary wskazuje na sytuację, w której $W_1 > \frac{1}{N-1}$

Źródło: obliczenia własne.

Z przedstawionych w tab. 8.13 obliczeń wynika, że biorąc pod uwagę ocenę właściwości dyskryminacyjnych wyznaczonych mierników syntetycznych *MWDMS*,

można dostrzec, że większość z nich ma podobne zdolności do tworzenia grup typologicznych, choć największe wykazuje miernik AMR(**). Z kolei porównując współczynniki zmienności, zauważyć trzeba, że najwyższą wartość osiągnęły sumy standaryzowane AMR(*)[6] oraz AMR(**).

8.3. Metody klasyfikacji

Rozwiązanie zadania klasyfikacji przebiega w kilku etapach.

1) Wybranie obiektów badania O_i i cech je charakteryzujących X_j.
2) Dobór zmiennych diagnostycznych – zaleca się, aby zmienne diagnostyczne były słabo skorelowane między sobą i silnie skorelowane ze zmienną grupującą.
3) Zebranie danych statystycznych. Podział zebranego materiału empirycznego na próby uczącą i testową, np. proporcja obu prób to 70:30 lub 80:20.
4) Wybór metody klasyfikacji – np. liniowa funkcja dyskryminacji zbudowana dla wybranego zestawu cech diagnostycznych.
5) Estymacja funkcji klasyfikacyjnej na podstawie próby uczącej.
6) Ocena jakości klasyfikacji na podstawie błędów klasyfikacji (w próbie uczącej).
7) Testowanie oszacowanego modelu na podstawie danych testujących.
8) Ocena jakości klasyfikacji na podstawie błędów klasyfikacji wyznaczonych dla próby testowej (zawierającej inne obiekty niż w próbie uczącej).
9) Jeśli jakość klasyfikacji jest akceptowalna, to stosuje się oszacowany model do klasyfikacji obiektów o nieznanej przynależności do klasy.

Do najpopularniejszych metod klasyfikacji ze wzorcem należą modele dyskryminacyjne i drzewa klasyfikacyjne. Główna idea leżąca u podstaw analizy dyskryminacyjnej to rozstrzyganie, czy wyróżnione *a priori* klasy różnią się ze względu na średnią pewnej zmiennej, a następnie wykorzystanie tej zmiennej do przewidywania przynależności do grupy (np. nowych przypadków). W rzeczywistości pod uwagę bierze się pewną liczbę zmiennych w celu sprawdzenia, które najlepiej przyczyniają się do dyskryminowania grup. Analiza dyskryminacyjna jest więc statystyczną techniką pozwalającą badać różnice pomiędzy dwiema lub więcej grupami, analizując kilka zmiennych jednocześnie. Zmienne użyte do rozróżnienia grup nazywają się zmiennymi dyskryminacyjnymi.

6 Wartość współczynnika zmienności jest tak duża, bo średnia wartości tego miernika jest bliska 0.

Budowę funkcji dyskryminacji należy poprzedzić wielowymiarowymi analizami zmiennych diagnostycznych i weryfikacją podstawowych założeń[7].

1) Przyjmuje się, że zmienne dyskryminacyjne reprezentują wielowymiarowy rozkład normalny. Chociaż badania dotyczące aplikacji wielowymiarowej funkcji dyskryminacji potwierdzają, że jest ona dobrym klasyfikatorem, mimo naruszenia tego założenia. Do weryfikacji założenia o wielowymiarowym rozkładzie normalnym wykorzystuje się odpowiednie testy statystyczne, np. Kołmogorowa-Smirnowa, lub Shapiro-Wilka.
2) Zakłada się podzielność zmiennych, która przejawia się w systematycznej różnicy wartości średnich między grupami. Do wyeliminowania zmiennych niepodzielnych korzysta się z testu U-Mann-Whitneya (będącego wielowymiarową odmianą testu t-Studenta).
3) Przyjmuje się, że macierze kowariancji zmiennych diagnostycznych są równe w grupach. Badania empiryczne wykazują, że można pominąć to założenie.

Liniowa funkcja dyskryminacyjna jest ważoną sumą analizowanych wskaźników, czyli:

$$y(x_i) = a_0 + a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + a_k \cdot x_{ki} = a_0 + \boldsymbol{a}^T \boldsymbol{x_i} \tag{8.40}$$

gdzie:
$\boldsymbol{a} = [a_j]$ – wektor [K x 1] parametrów funkcji dyskryminacyjnej ($j = 1, 2, \ldots, K$),
$\boldsymbol{x_i} = [x_{ij}]$ – wektor [K x 1] zmiennych diagnostycznych (dyskryminacyjnych) opisujących cechy diagnostyczne X_j w i-tym obiekcie.

Analiza dyskryminacyjna (por. [Kolonko 1980, s. 153; Maddala 2006, s. 370–371]) jest metodą klasyfikacji ze wzorcem (nauczycielem), gdyż wartość funkcji dyskryminacyjnej, wyznaczoną dla rozpatrywanych obiektów, przyrównuje się do pewnego wzorca i w ten sposób określa przynależność obiektów do określonej klasy. Wśród parametrów funkcji dyskryminacyjnej wyróżnia się współczynniki surowe, standaryzowane i czynnikowe. Surowe parametry (8.40) wykorzystywane są do obliczenia wartości funkcji dyskryminacji, jednak nie nadają się do interpretacji. Natomiast współczynniki standaryzowane określają siłę wpływu danej zmiennej dyskryminacyjnej na zróżnicowanie grup, przy czym im większa jest wartość parametru, tym silniejszy jest wpływ dyskryminacyjny określonej zmiennej. Z kolei współczynniki czynnikowe wskazują na siłę powiązań pomiędzy zmiennymi dyskryminacyjnymi[8].

W przypadku dyskryminowania wśród posiadanych danych 2 klas, wyznaczenie funkcji dyskryminacji jest dość proste. Na podstawie danych empirycznych

7 Podobnie sformułowano założenia w [Krzyśko 1990, s. 19].

8 Jednakże w przypadku pierwszych z nich uwzględnia się udziały pozostałych zmiennych, a przy drugich nie (por. [Strzelecka 2017]).

pochodzących z próby (zbioru treningowego) zawierającej obiekty, opisane za pomocą wybranych K zmiennych diagnostycznych, oblicza się średnie poziomy poszczególnych cech w obu grupach typologicznych oraz macierz kowariancji[9]. Wektor parametrów funkcji (8.40) wyznacza się ze wzoru:

$$\boldsymbol{a} = \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1) \tag{8.41}$$

gdzie:
$\boldsymbol{S}$ – macierz [K x K] kowariancji,
$\overline{\boldsymbol{x}}_2$, $\overline{\boldsymbol{x}}_1$ – wektory [K x 1] średnich wartości zmiennych diagnostycznych odpowiednio w grupie pierwszej i drugiej,
pozostałe oznaczenia jak poprzednio.

W dalszym postępowaniu określa się wartość krytyczną funkcji dyskryminacyjnej, na podstawie której dokonuje się klasyfikacji badanych obiektów. W tym celu wyznacza się przeciętne wartości funkcji dyskryminacyjnej dla obu grup typologicznych za pomocą relacji:

$$\overline{y}_p = (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_p \tag{8.42}$$

gdzie:
p – numer grupy typologicznej: pierwszej lub drugiej,
$\overline{y}_p$ – przeciętna wartość funkcji dyskryminacyjnej w p-tej grupie,
pozostałe oznaczenia jak poprzednio.

Po wyznaczeniu współczynników dyskryminacyjnych określa się wartość krytyczną funkcji dyskryminacyjnej, którą wyznacza się na podstawie wzoru:

$$a_0 = -[\alpha \cdot (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_1 + (1 - \alpha) \cdot (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_2] \tag{8.43}$$

gdzie:
α – prawdopodobieństwo wystąpienia elementów klasy 1,
$1 - \alpha$ – prawdopodobieństwo wystąpienia elementów klasy 2.

Kiedy wartość krytyczna znajduje się w połowie odległości między $\overline{y}_1$ i $\overline{y}_2$, czyli gdy $\alpha = 0{,}5$, wtedy relacja (8.43) redukuje się do postaci:

$$a_0 = -0{,}5 \cdot (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}}_2 + \overline{\boldsymbol{x}}_1) \tag{8.44}$$

Mając wyznaczoną wartość krytyczną dokonuje się podziału obiektów do grup typologicznych. Obiekty, dla których obliczone wartości funkcji dyskryminacji są od wartości krytycznej mniejsze, zalicza się do grupy pierwszej, a pozostałe zalicza się do grupy drugiej.

9 Chodzi tutaj o oszacowanie współczynników dyskryminacji i statystyki Walda-Andersona (por. [Jajuga 1993, s. 130]).

W sytuacji, gdy pewna liczba obiektów z próby treningowej zostaje błędnie zaklasyfikowana, czyli znajdzie się w innej grupie typologicznej niż rzeczywiście przynależy, powstaje tzw. strefa niepewności, w której nie można podjąć decyzji o zaklasyfikowaniu obserwacji do żadnej z klas. Wówczas proponuje się wyznaczanie tzw. punktu krytycznego postaci[10]:

$$y_c = \frac{p_1 \cdot k_1}{p_2 \cdot k_2} \tag{8.45}$$

gdzie:
y_c – wartość punktu krytycznego,
p_1, p_2 – prawdopodobieństwo zakwalifikowania *i*-tego obiektu do klasy odpowiednio: A_1 i A_2,
k_1, k_2 – koszt błędnej klasyfikacji do grupy odpowiednio: A_1 i A_2.

Przeprowadzając klasyfikację ze wzorcem, często korzysta się z tzw. funkcji klasyfikacyjnych[11] postaci (8.40), które formułowane są dla każdej grupy typologicznej. W praktyce przyjmuje się, że wszystkie funkcje klasyfikacyjne mają tę samą postać analityczną, a różnią się jedynie współczynnikami. Funkcje te są tak generowane, aby maksymalizować wartość ilorazu zmienności międzygrupowej i wewnątrzgrupowej. Liniowa funkcja dyskryminacji opisuje hiperpłaszczyznę rozdzielającą zbiory obiektów w taki sposób, aby je jak najlepiej odseparować. W przypadku nierównych macierzy kowariancji szacuje się kwadratową lub inną (np. logistyczną[12]) postać funkcji dyskryminacji.

Modele dyskryminacyjne (klasyczne) są modelami klasyfikacji dychotomicznej. Pierwsze prace nad ich zastosowaniem do prognozowanie bankructwa prowadził Altmana (1968). W klasycznym podejściu zmienna zależna przyjmuje wartość 0 dla bankruta i 1 dla nie-bankruta. Zmienne dyskryminacyjne (opisowe, niezależne) to zazwyczaj zmienne opisujące kondycję finansową przedsiębiorstwa, np. wskaźniki finansowe. Modele dyskryminacyjne są systemami wczesnego ostrzegania. Powstały po to, aby zawczasu zidentyfikować zagrożenia dotyczące efektywności gospodarowania i związanej z tym utraty zdolności płatniczej przedsiębiorstwa. Można również budować modele do klasyfikacji wielodzielnej – politomicznej, tj. kiedy klas jest więcej niż 2. Przykładem tego typu modeli może być ocena zdolności kredytowej i klasyfikacja do różnych grup ryzyka.

---

10 Problem ten został omówiony w pracy [Altman 1993, s. 254–264].

11 Metody oparte na funkcjach dyskryminacji stanowią część pakietów komputerowych realizujących analizy statystyczne, np. STATISTICA, SAS, SPSS. Wyznacza się tam nie funkcje dyskryminacyjne, tylko klasyfikacyjne.

12 W przypadku klasyfikacji dychotomicznej model dyskryminacji logistycznej jest równoważny modelowi regresji logistycznej.

Przykład 8.2

Model Altmana (*Zeta Model*, *Z-score* lub *zeta score model*) prognozowania upadłości przedsiębiorstw jest oszacowaną funkcją dyskryminacji postaci (por. [Altman 1968]):

$$D = 1{,}2 \cdot x_1 + 1{,}4 \cdot x_2 + 3{,}3 \cdot x_3 + 0{,}6 \cdot x_4 + 1{,}0 \cdot x_5$$

gdzie:

$$x_1 = \frac{kapitał\ pracujący}{aktywa\ ogółem}$$

$$x_2 = \frac{skumulowane\ zyski\ reinwestowane}{aktywa\ ogółem}$$

$$x_3 = \frac{zysk\ brutto + odsetki}{aktywa\ ogółem}$$

$$x_4 = \frac{wartość\ rynkowa\ kapitału\ własnego}{wartość\ księgowa\ kapitału\ obcego}$$

$$x_5 = \frac{przychód\ ze\ sprzedaży}{aktywa\ ogółem}$$

Prezentowany wyżej model Altman oszacował na danych pochodzących z 66 przedsiębiorstw, z których połowa miała stabilną sytuację finansowa, a 33 firmy były zagrożone upadłością. Wyznaczona wartość dyskryminacyjna wyniosła 1,81. Dla analizowanej próby Altman wyznaczył prawdopodobieństwa upadku analizowanych korporacji, np. korporacja, dla której wartość funkcji dyskryminacji wynosiła –0,55, upadnie z prawdopodobieństwem 75%, natomiast prawdopodobieństwo bankructwa korporacji, dla której funkcja dyskryminacji przyjmie wartość 2,30 wyniosło zaledwie 1%.

W 1984 roku Altman (por. [Altman 1984]) wykazał, że modele służące do przewidywania upadku przedsiębiorstw powinny być stosowane tylko w kraju, w którym zostały zebrane dane służące do ich opracowania. Przyczyną są między innymi różnice w systemach rachunkowości obowiązujących w różnych krajach. Autorami pierwszych modeli dyskryminacyjnych szacowanych na podstawie danych z polskich przedsiębiorstw byli: Gajdka i Stos [1996], Hołda [2000], Homrola, Czajka i Piechocki [2004], Prusak [2005] oraz Mączyńska i Zawadzki [2006].

Drzewa klasyfikacyjne należą do nieparametrycznych metod badania zależności występujących w zbiorach danych, czyli ich stosowanie nie wymaga ani znajomości rozkładów zmiennych opisujących obiekty, ani postaci analitycznej związku między nimi. Drzewa klasyfikacyjne dokonują podziału wielowymiarowej przestrzeni zmiennych $\boldsymbol{X}^m$, w której znajdują się klasyfikowane obiekty $(x_1, x_2, \ldots, x_n)$ w taki sposób, aby uzyskać rozłączne segmenty tej przestrzeni. W tych segmentach znajdują się obiekty, które należą do tej samej klasy. Drzewo składa się z korzenia (znajdującego się zwykle u góry rysunku) oraz gałęzi prowadzących z korzenia do kolejnych węzłów. W każdym węźle sprawdzany jest pewien warunek dotyczący danej obserwacji i na jego podstawie wybiera się jedną z gałęzi prowadzącą do kolejnego węzła piętro niżej. Na dole znajdują się liście, w których odczytuje się, do której z klas należy przypisać daną obserwację. Klasyfikacja danej obserwacji polega na przejściu od korzenia do liścia i przypisaniu do tej obserwacji klasy zapisanej w danym liściu. Procedura podziału ma charakter rekurencyjny i składa się z pięciu kroków (por. [Gatnar, Walesiak 2004, s. 108; Gatnar 1998, s. 169–170]).

1) Sprawdzenie czy przestrzeń $\boldsymbol{X}^m$, w której znajduje się zbiór uczący $\boldsymbol{S}$, jest jednorodna pod względem wartości zmiennej zależnej y. Jeżeli tak, to należy zakończyć postępowanie[13], jeżeli nie – przejść do kroku (2).
2) Rozpatrzenie wszystkich możliwych sposobów podziału przestrzeni $\boldsymbol{X}^m$ na rozłączne segmenty $R_1, R_2, \ldots, R_k$, uwzględniając wartości kolejno wybieranych zmiennych objaśniających.
3) Ocena jakości każdego z tych podziałów zgodnie z przyjętym kryterium homogeniczności oraz wybór najlepszego z nich.
4) Podzielenie przestrzeni $\boldsymbol{X}^m$ wybranym sposobem.
5) Powtórzenie kroków (1)–(4) rekurencyjnie dla każdego z segmentów $R_1, R_2, \ldots, R_k$.

Podział przestrzeni obiektów w (4) kroku jest dokonywany na podstawie wartości cech opisujących obiekty. Cecha, ze względu na którą dokonuje się podziału, nie może być wybierana losowo, ponieważ wtedy drzewo klasyfikacyjne mogłoby mieć liczbę liści równą liczbie obiektów. Dlatego w celu oceny jakości działania drzewa stosuje się miary jednorodności lub miary zróżnicowania klas uzyskanych w wyniku podziału. Wszystkie te miary opierają się na założeniu, że zachodzi związek między wartościami cech obiektów, a ich przynależnością do danej grupy.

Optymalny model drzewa klasyfikacyjnego to taki, który daje mały błąd klasyfikacji i jednocześnie charakteryzuje się niewielką złożonością. W celu uzyskania takiego drzewa stosuje się takie metody, jak np. przycinanie krawędzi czy skracanie długości krawędzi, a następnie ocenia się błędy klasyfikacji dla zbioru uczącego (koszty resubstytucji) oraz zbioru testowego (koszty sprawdzianu krzyżowego).

13 Procedurę można też zakończyć, gdy jest spełniony inny warunek stopu, np. minimalna liczba obiektów w powstałych segmentach.

W programach komputerowych do budowy drzew klasyfikacyjnych najczęściej wykorzystuje się algorytmy *CART* (*Classification And Regression Trees*) i *QUEST* (*Quick Unbiased Efficient Statistical Tree*). *CART* (por. [Gatnar 1998, s. 182]) jest algorytmem, który polega na rekurencyjnym przeszukiwaniu wszystkich możliwych podziałów jednowymiarowych. W przypadku dużej liczby zmiennych predykcyjnych (diagnostycznych) z wieloma poziomami, działanie tego algorytmu może być czasochłonne, a także obciążone w kierunku wyboru zmiennych predykcyjnych, które mają więcej poziomów do podziałów. Dlatego często otrzymuje się najlepsze podziały (z punktu widzenia najlepszych wyników klasyfikacji) w zbiorze uczącym, jednak niekoniecznie w próbach sprawdzianu krzyżowego. Algorytm *QUEST*, zamiast funkcji liniowych, wykorzystuje kwadratowe funkcje dyskryminacyjne do wyczerpującego poszukiwania podziałów jednowymiarowych [Gatnar, Walesiak 2004, s. 106]. Algorytm ten jest szybki i nieobciążony, jeżeli chodzi o wybór podziałów ze względu na liczbę poziomów predyktorów (predyktory z wieloma poziomami mają tendencję do generowania przypadkowych rozwiązań, które dobrze pasują do danych, ale mają słabą trafność predykcyjną).

W przypadku klasyfikacji ze wzorcem istnieje możliwość oceny jakości klasyfikacji na podstawie odpowiednio sformułowanych błędów. Ogólny błąd klasyfikacji jest postaci:

$$E = \frac{BN_1 + BN_2}{N} \cdot 100 \tag{8.46}$$

Natomiast błąd klasyfikacji do wyróżnionej klasy $p = 1, 2$:

$$E_p = \frac{BN_p}{N_p} \cdot 100 \tag{8.47}$$

gdzie:
N – liczba wszystkich obiektów podlegających klasyfikacji,
N_p – liczba obiektów należących do p-tej klasy,
BN_p – liczba błędnie zaklasyfikowanych elementów do p-tej klasy.

Warto odnotować, że wyrażony w procentach *zliczeniowy* R^2 (6.20), określając odsetek poprawnie zaklasyfikowanych obiektów, jest przeciwieństwem ogólnego błędu klasyfikacji i zachodzi relacja: *zliczeniowy* $R^2 = (100 - E)/100$.

Błędy klasyfikacji oblicza się dla obiektów o znanej przynależności do klasy. Wyznacza się je zarówno dla próby uczącej, jak i testującej. Jednak o jakości – poprawności metody klasyfikacji decydują wartości błędów wyznaczonych dla próby testowej.

8.4. Metody grupowania

Jak wcześniej wspomniano, jednymi z częściej stosowanych metod klasyfikacji bez wzorca są metoda *k*-średnich i metoda Warda. Zastosowanie obu metod do rozwiązania praktycznych problemów wymaga zazwyczaj użycia profesjonalnego oprogramowania. Poniżej przedstawiony zostanie zarys postępowania w obu przypadkach.

Metoda Warda należy do metod aglomeracyjnych, w których wstępnym założeniem jest przypuszczenie, że każdy obiekt badania O_i tworzy początkowo jedną klasę A_i, ($i = 1, 2, \ldots, p$). Realizacja procedury klasyfikacyjnej przebiega następująco.

1) Tworzy się p skupień, czyli każdy obiekt badania stanowi jedną klasę.
2) W macierzy odległości szuka się pary klas najbardziej podobnych (tj. najmniej odległych od siebie). Niech będą to klasy A_i oraz A_j.
3) Redukuje się liczbę klas o 1, łącząc klasy A_i oraz A_j w nową klasę.
4) Przekształca się odległości stosownie do metody między połączonymi klasami A_i i A_j oraz pozostałymi klasami.
5) Powtarza się kroki (2)–(4) do chwili, gdy wszystkie obiekty znajdą się w jednej klasie.

W przypadku metod aglomeracyjnych, gdy liczba klas jest znana i wynosi l, należy proces grupowania przerwać po $n - l$ etapach.

Metoda *k*-średnich należy do grupy metod klasyfikacji bez wzorca, tj. dokonuje podziału całego zbioru przypadków na z góry określoną liczbę k (u nas $k = p$) różnych[14], możliwie odmiennych klas bez znajomości wzorców tych klas. Algorytm metody polega na przenoszeniu obiektów między klasami tak, aby uzyskać małą zmienność obiektów wewnątrz skupień i zmaksymalizować zmienności między skupieniami. W każdym skupieniu powinny znajdować się zatem obiekty najbardziej podobne do siebie pod względem opisujących je atrybutów, przy jednoczesnym znacznym zróżnicowaniu pomiędzy skupieniami. Metoda *k*-średnich opiera się na oszacowaniu odległości między skupieniami a obiektami i przebiega zgodnie z opisanym poniżej algorytmem.

1. Ustala się maksymalną liczbę iteracji[15] oraz liczbę grup p, na jakie ma być podzielony analizowany zbiór obiektów, przy czym $p \in [2; N - 1]$, gdzie: N jest liczbą obiektów.
2. Ustala się wyjściową macierz środków ciężkości grup postaci:

$$\mathbf{B} = [b_{sj}] \ (s = 1, 2, \ldots, p; j = 1, 2, \ldots, k) \tag{8.48}$$

14 Z reguły przyjmuje się, że liczba grup w populacji jest równa k, stąd nazwa metody. Tutaj przyjmuje się, że liczba klas wynosi p.

15 Omawiana metoda optymalizacji wstępnych podziałów zbioru nie jest zbieżna. Dlatego, aby nie dopuścić do powstania pętli w obliczeniach, należy z góry ustalić dopuszczalną liczbę iteracji.

gdzie: k jest liczbą zmiennych oraz przyporządkowuje poszczególne obiekty do takich grup, dla których ich odległość od środka ciężkości danej grupy jest najmniejsza.

3. Wyznacza się wartość wyjściowego błędu podziału obiektów między p grup:

$$\omega = \sum_{i=1}^{N} d_{is}^2 \tag{8.49}$$

gdzie: d_{is}^2 jest odległością Euklidesa (7.16) między i-tym obiektem a najbliższym s-tym środkiem ciężkości:

$$d_{is}^2 = \sum_{j=1}^{k} (x_{ij} - b_{sj})^2 \quad (i = 1, 2, \ldots, N) \tag{8.50}$$

4. Dla pierwszego obiektu określa się zmiany błędu podziału wynikające z przyporządkowania go kolejno do wszystkich aktualnie występujących grup:

$$\Delta\omega_s^1 = \frac{n_s \cdot d_{1s}^2}{n_s + 1} - \frac{n_{s_1} \cdot d_{1s_1}^2}{n_{s_1} + 1} \tag{8.51}$$

gdzie:
n_s – liczebność s-tej grupy;
d_{1s}^2 – odległość pierwszego obiektu od środka ciężkości s-tej grupy;
n_{s1} – liczebność grupy zawierającej pierwszy obiekt;
d_{1s1}^2 – odległość pierwszego obiektu od najbliższego środka ciężkości.

Jeżeli minimalna wartość wyrażenia (8.51) dla wszystkich $s \neq s_1$ jest ujemna, to pierwszy obiekt przypisuje się do grupy, dla której $\Delta\omega_s^1 = min$. W dalszym postępowaniu przelicza się środki ciężkości grup **B**, uwzględniając dokonaną transformację obiektu oraz wyznacza się aktualną wartość błędu podziału (8.49). Jeżeli minimalna wartość tego wyrażenia jest dodatnia lub równa 0, to nie dokonuje się żadnych zmian.

5. Operacje opisane w punkcie (4) powtarza się dla każdego następnego obiektu, co kończy pierwszą iterację procedury.
6. Jeżeli w danej iteracji nie obserwuje się żadnych przesunięć obiektów z grupy do grupy, to postępowanie się kończy. W przeciwnym wypadku rozpoczyna się następną iterację, aż do momentu, w którym liczba iteracji nie przekroczy ustalonej wartości.

Przedstawiony algorytm (1)–(6) umożliwia znalezienie lokalnego optimum błędu podziału ω, przy czym zazwyczaj liczba niezbędnych iteracji jest mniejsza od 15.

8.5. Metody doboru cech diagnostycznych

W przypadku metod statystycznych stosowanych do rozpoznawania obiektów istotną kwestią jest wybór odpowiedniego zestawu zmiennych diagnostycznych, ponieważ ostateczne wyniki badania w znacznym stopniu od niego zależą. Zmienne diagnostyczne powinny w sposób możliwie pełny charakteryzować najważniejsze aspekty badanego zjawiska. Zatem muszą to być zmienne niosące informacje ogólne o poszczególnych jednostkach, a nie informacje unikatowe. Jednakże wiele zmiennych ekonomicznych wykazuje silną wzajemną korelację, co oznacza powielanie tych samych informacji przez kilka zmiennych. Pojawia się wówczas konieczność redukowania liczby zmiennych diagnostycznych, co może również wynikać z faktu posiadania mało licznej próby, a dużej liczby szacowanych parametrów.

Dobór zmiennych diagnostycznych odbywa się w drodze przetwarzania i analizy danych w oparciu o kryteria merytoryczne i metody statystyczne. Kryterium merytoryczne jest oceną jakościową, a statystyczne oparte jest na miernikach ilościowych, które wyznaczane są za pomocą procedur formalnych. Podstawą do wyboru zmiennych diagnostycznych jest tzw. wstępna lista cech zaproponowana przez badacza na podstawie ogólnej znajomości zjawiska.

Do wyznaczenia zbioru zmiennych diagnostycznych stosuje się odpowiednie procedury w celu uzyskania zestawu zmiennych, które w sposób możliwie pełny charakteryzują badane jednostki, a przy tym tworzą zespół jak najmniej liczny. Wymagania te są spełnione wtedy, gdy zmienne diagnostyczne posiadają następujące własności[16]:

- są nieskorelowane lub co najwyżej słabo skorelowane między sobą,
- są silnie skorelowane ze zmienną grupującą (o ile taka istnieje),
- posiadają zdolność dyskryminacji badanych jednostek, tj. charakteryzują się wysoką zmiennością wśród wszystkich jednostek zbioru, a niską wśród jednostek wydzielonych grup,
- nie ulegają wpływom zewnętrznym.

Procedura doboru zmiennych diagnostycznych realizowana jest w kilku etapach.

1) Na podstawie wiedzy merytorycznej sporządza się zestaw tzw. pierwotnych zmiennych diagnostycznych, którymi są wszystkie najważniejsze wielkości oddziałujące na zmienną decyzyjną.
2) Przeprowadza się obserwacje x_{ij} ($i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, K$), będące realizacjami potencjalnych zmiennych diagnostycznych oraz y_i zmiennej grupującej (w przypadku jej występowania). Zebrane dane statystyczne tworzą

16 Należy zauważyć, że podobne postulaty dotyczą zmiennych użytych do konstrukcji modeli ekonometrycznych.

macierz $\mathbf{X} = [x_{ij}]$, której wiersze charakteryzują badane obiekty O_i, a kolumny pierwotne zmienne diagnostyczne oraz wektor $\mathbf{y} = [y_i]$ zmiennej decyzyjnej (klasyfikacyjnej).

3) Bada się zmienność wszystkich zmiennych zawartych w pierwotnym zestawie zmiennych diagnostycznych. Najczęściej wykorzystuje się w tym celu współczynnik zmienności.
4) Wyznacza się współczynniki korelacji między wszystkimi rozpatrywanymi zmiennymi diagnostycznymi i (o ile istnieje) zmienną grupującą.
5) Z pierwotnego zbioru zmiennych diagnostycznych usuwa się zmienne charakteryzujące się małą zmiennością (zazwyczaj eliminuje się zmienne, dla których współczynnik zmienności $V_j < 0{,}1$), a także zmienne diagnostyczne silnie skorelowane między sobą oraz słabo skorelowane ze zmienną grupującą.

Podstawowe metody statystyczne doboru zmiennych diagnostycznych to metody: analizy macierzy współczynników korelacji, analizy czynnikowej i analizy głównych składowych.

Jedna z formalnych procedur doboru cech diagnostycznych, w analizie obiektów wielocechowych opiera się na macierzy korelacji **R** między potencjalnymi cechami diagnostycznymi[17], czyli:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1K} \\ r_{21} & 1 & r_{23} & \dots & r_{2K} \\ \dots & \dots & \dots & \dots & \dots \\ r_{K1} & r_{K2} & r_{K3} & \dots & 1 \end{bmatrix} \tag{8.52}$$

gdzie dla $i, j = 1, 2, \dots, K$:
r_{ij} – to współczynnik korelacji liniowej (4.5) między cechami X_i i X_j.

Kryterium doboru zmiennych diagnostycznych stanowi, wyznaczana w sposób formalny lub przyjmowana arbitralnie, krytyczna wartość współczynnika korelacji r^*. Na jej podstawie z tablicy (8.52) wyróżnia się tzw. skupienia, czyli podzbiory cech, dla których wartości bezwzględne współczynników korelacji nie są większe od wartości krytycznej. W skupieniach wyróżnia się jedną tzw. cechę centralną oraz pewną liczbę tzw. cech satelitarnych. Cechy należące do skupień nazywane są systemowymi, a pozostałe – izolowanymi. Cechy centralne i izolowane tworzą zbiór cech diagnostycznych. Postępowanie w tej metodzie składa się z kilku etapów:

17 Metodę tę zaproponował Zdzisław Hellwig, a jej opis można znaleźć m.in. w pracach: [Urbanek 1995; Witkowska 2002, s. 73–75], w których przedstawiono przykłady aplikacji tej metody w analizach sytuacji finansowej przedsiębiorstwa.

1) w macierzy korelacji **R** wyznacza się sumę elementów każdej kolumny, czyli:

$$R_j = \sum_{i=1}^{K} \left| r_{ij} \right| \tag{8.53}$$

2) wyszukuje się kolumnę p, dla której suma jest największa, a cechę odpowiadającą wyróżnionej kolumnie uważa się za cechę centralną,
3) w wyróżnionej kolumnie wyszukuje się elementy spełniające nierówność:

$$\left| r_{pj} \right| \geq r^* \tag{8.54}$$

cechy odpowiadające wyróżnionym wierszom uważa się za cechy satelitarne,
4) z macierzy **R** skreśla się wyróżnione kolumny i wiersze, uzyskując w ten sposób tzw. zredukowaną macierz korelacji,
5) postępowanie (1)–(4) powtarza się, otrzymując dalsze punkty skupienia i nowe zredukowane macierze korelacji, aż do momentu wyczerpania zbioru cech,
6) przedstawione wyżej postępowanie może być wykorzystywane nie tylko przy doborze cech diagnostycznych, ale również do klasyfikacji obiektów wielocechowych.

Rozdział 9

Metody badania finansowych szeregów czasowych

Badania dotyczące instrumentów finansowych i tworzonych z nich portfeli inwestycyjnych prowadzone są zazwyczaj na podstawie obserwacji ich zachowań w czasie, tj. na podstawie szeregów czasowych. W analizach finansowych uwzględnia się dwa podstawowe parametry charakteryzujące inwestycje, tzn.

- dochody inwestorów, które zazwyczaj reprezentowane są przez różnie definiowane stopy zwrotu oraz
- ryzyko, opisywane za pomocą różnych mierników, odzwierciedlających zmienność stóp zwrotu.

Dlatego też wiele uwagi poświęca się zarówno badaniu wewnętrznych własności szeregów czasowych obserwacji notowań instrumentów finansowych lub portfeli inwestycyjnych, jak i wzajemnych relacji występujących między różnymi instrumentami lub portfelami.

Celem niniejszego rozdziału jest pokazanie, w jaki sposób powszechnie znane metody statystyczne i modele ekonometryczne są wykorzystywane w praktyce analiz instrumentów finansowych. Rozważania ogólne zostaną zilustrowane badaniem notowań cen akcji spółki z GPW w Warszawie. W analizach wykorzystane zostaną:

- statystyczne metody opisowe,
- metody wnioskowania statystycznego,
- modele Sharpe'a i CAMP,
- klasyczny miernik efektywności inwestycyjnej Jensena.

9.1. Stopy zwrotu

Dochód z tytułu posiadania akcji pochodzi z różnicy ceny sprzedaży i zakupu akcji oraz z dywidendy. Stąd też stopa zwrotu z akcji wyrażana jest jako:

$$R_{it}^{*} = \frac{y_{it} - y_{it-k} + D_t}{y_{it-k}}$$

gdzie dla każdego $t = 1, 2, \ldots, T$:
R_{it}^{*} – stopy zwrotu z instrumentu ($t = 1, 2, \ldots, T$),
T – liczba okresów, w których dokonywano pomiaru zjawiska,
y_{it}, y_{it-k} – obserwacje szeregu czasowego dotyczące notowania i-tego instrumentu finansowego, poczynione odpowiednio w okresie (lub momencie) t oraz opóźnione o k okresów,
D_t – dywidenda wypłacona w okresie t.

Należy zauważyć, że wypłacanie akcjonariuszom dywidendy jest obwarowane pewnymi warunkami[1]. Zatem nie wszystkie spółki i nie w każdym roku dywidendę wypłacają, a oprócz tego należy się ona tylko tym inwestorom, którzy są posiadaczami akcji w momencie sporządzania stosownej listy akcjonariuszy. Dlatego w analizach często pomija się jej wpływ. Przyjmując, że $D_t = 0$, można stopę zwrotu z dowolnego instrumentu finansowego zapisać jako:

$$R_{it}^{*} = \frac{y_{it} - y_{it-k}}{y_{it-k}} \tag{9.1}$$

Warto odnotować, że t i k jednoznacznie określają interwał, z jakiego są wyznaczane stopy zwrotu, będące wielkościami niemianowanymi, lub są wyrażane w procentach.

W praktyce wyróżnia się dwa rodzaje stóp zwrotu z instrumentów finansowych, tj. proste (zwykłe) i logarytmiczne stopy zwrotu. Pierwsze ze wspomnianych wyznacza się ze wzoru (9.1) i stanowią one *de facto* względny przyrost (5.4) wartości waloru w zadanym okresie. Drugi rodzaj stóp zwrotu to stopy logarytmiczne będące logarytmem naturalnym ieksu (5.6) wyznaczonego dla cen instrumentu finansowego, obserwowanych w dwóch momentach lub okresach, postaci:

$$R_{it} = ln\left(\frac{y_{it}}{y_{it-k}}\right) = ln(y_t) - ln(y_{t-k}) \tag{9.2}$$

gdzie dla każdego $t = 1, 2, \ldots, T$:
R_{it}^{*}, R_{it} – stopy zwrotu z instrumentu ($t = 1, 2, \ldots, T$),
ln – symbol logarytmu naturalnego,
pozostałe oznaczenia jak poprzednio.

1 W Polsce o wypłacie dywidendy, której kwota jest pochodną zysku netto spółki, decyduje walne zgromadzenie akcjonariuszy.

Stopy zwrotu (9.1)–(9.2) wyrażają zysk (jeśli są dodatnie) lub ratę (gdy są ujemne), jakie osiągnie inwestor w momencie (lub okresie) t po zainwestowaniu y_{it-k} środków w $(t - k)$-tym momencie (lub okresie). Prosta stopa zwrotu (9.1) wykorzystuje koncepcję kapitalizacji okresowej, a stopa logarytmiczna (9.2) – kapitalizacji ciągłej. Rozkład zwykłej stopy zwrotu jest tej samej postaci co rozkład ceny (z dokładnością do parametrów położenia i skali). Natomiast rozkład logarytmicznej stopy zwrotu ma inną postać niż rozkład kursu, jest bowiem logarytmicznym przekształceniem rozkładu ceny. Wartości stóp zwrotu (9.2) należą do przedziału $(-\infty; +\infty)$ i są addytywne, co oznacza, że logarytmiczna stopa zwrotu w danym okresie jest równa sumie logarytmicznych stóp z poszczególnych jego odcinków i można z nich obliczać średnią arytmetyczną, która odpowiada kapitalizacji ciągłej (zakłada się, że wszystkie dochody są reinwestowane) (por. [Jajuga 2000, s. 71; Fiszeder 2009, s. 22–23]). Skumulowana dla k okresów stopa zwrotu dla stóp logarytmicznych jest postaci:

$$R_{it}^{sk} = \sum_{t=1}^{k} R_{it} = ln\left(\frac{y_{ik}}{y_{i1}}\right) \tag{9.3}$$

podczas gdy dla stóp prostych skumulowane zwroty wynoszą[2]:

$$R_{it}^{sk^*} = \prod_{t=1}^{k}\left(R_{it}^{*} + 1\right) - 1 \tag{9.4}$$

W związku z powyższym średnie stopy zwrotu ze stóp logarytmicznych wyznacza się jako średnią arytmetyczną (5.2), a ze stóp prostych[3] – jako średnią geometryczną (5.8).

W badaniach często wykorzystuje się logarytmiczne stopy zwrotu ze względu na ich niektóre własności statystyczne[4]. Przede wszystkim jest to związane z założeniem o normalności rozkładu szeregu stóp zwrotu, które jest często przyjmowane przy estymacji nieznanych parametrów i weryfikacji hipotez statystycznych, a formalna możliwość przyjęcia założenia o normalności pojawia się wyłącznie w odniesieniu do stóp logarytmicznych.

---

2 gdzie: $\prod_{t=1}^{k} a_t$ – symbol iloczynu k czynników a_t.

3 Średnie zwroty w przypadku prostych stóp zwrotu wyznacza się jako: $R_i^* = \sqrt[k]{R_{it}^{sk^*}} = \sqrt[k]{\prod_{t=1}^{k}(R_{it}^* + 1)}$.

4 Użycie logarytmicznych stóp zwrotu często jest bardziej uzasadnione niż wykorzystanie prostych (zwykłych) stóp zwrotu. Dyskusję na ten temat można znaleźć m.in. w: [Tarczyński i in. 2013, s. 26–27]. Dotyczy to w szczególności założenia o normalności rozkładu stóp zwrotu, ponieważ rozkład zwykłych zwrotów jest „ucięty" w punkcie odpowiadającym (–1) lub (–100%), co wyklucza rozkład normalny takiej zmiennej (por. [Jajuga 2000, s. 71]).

Warto jednak zauważyć, że logarytmiczne stopy zwrotu mają również swoje wady lub ich własności nie zawsze są pożądane. Na przykład:

- stopy zwrotu z portfela instrumentów w przypadku stóp prostych stanowią ważoną średnią prostych stóp zwrotu z pojedynczych aktywów, natomiast stopy logarytmiczne tej własności nie posiadają;
- stopy logarytmiczne „spłaszczają" zmiany w przypadku dużych wahań cen. Innymi słowy, kiedy następują duże zmiany wartości instrumentów, stopy logarytmiczne są znacząco niższe niż proste stopy zwrotu, co może mieć wpływ na wyniki analiz, zwłaszcza kiedy istotne są wartości z ogonów rozkładu, np. przy szacowaniu wartości zagrożonej.

9.2. Metody analizy własności szeregów stóp zwrotu

W analizie finansowych szeregów czasowych zazwyczaj przyjmuje się, że są one próbami statystycznymi, chociażby dlatego, że notowania kursów obserwowane są w określonych punktach pomiarowych, np. ceny otwarcia czy zamknięcia. W tej sytuacji parametry populacji generalnej nie są znane, a wyznaczane na podstawie obserwacji miary pełnią rolę ocen estymatorów nieznanych parametrów. Podstawowymi miernikami wykorzystywanymi w analizach własności finansowych szeregów czasowych są, omówione w rozdziale 2, miary:

1) położenia, opisujące przeciętny poziom zjawiska, traktowane jako mierniki dochodu;
2) dyspersji określające zmienność (rozproszenie, rozrzut), które wykorzystuje się do pomiaru ryzyka;
3) asymetrii i skupienia (koncentracji).

W badaniach finansowych najczęściej wykorzystuje się mierniki klasyczne, zatem średnia (2.2) lub (5.2) jest najczęściej stosowaną miarą określania przeciętnego poziomu zjawiska w szeregach czasowych okresów

$$R_i = \frac{1}{T}\sum_{t=1}^{T} R_{it} = \frac{1}{T} R_{it}^{sk} \tag{9.5}$$

gdzie dla każdego i-tego instrumentu i t-tego momentu (okresu):
R_{it} – stopy zwrotu z instrumentu ($t = 1, 2, \ldots, T$),
R_i – średnia stopa zwrotu i-tego instrumentu,
R_{it}^{sk} – skumulowane zwroty wyznaczone jako (9.3),
pozostałe oznaczenia jak poprzednio.

W przypadku szeregów czasowych momentów, do jakich zalicza się notowania cen instrumentów finansowych, powinno się wykorzystywać średnią chronologiczną (5.1). Jednakże w praktyce, zwłaszcza kiedy liczba elementów szeregu T jest znaczna, zamiast średniej chronologicznej korzysta się ze średniej arytmetycznej[5], zresztą notowania stóp zwrotu tworzą szeregi okresów, dla których adekwatną miarą jest średnia (5.2). Warto zauważyć, że klasyczne miary średnie wyznaczane są na podstawie wszystkich elementów szeregu, co – w sytuacji występowania danych odstających, ekstremalnych lub szeregów skrajnie asymetrycznych – powoduje pewne „przekłamania" w opisie zjawiska. Dlatego w takich przypadkach zaleca się stosowanie miar pozycyjnych, takich jak mediana i pozostałe kwartyle oraz dominanta (por. [Tarczyński 1997, s. 25–26])[6], aczkolwiek ich wykorzystanie w praktyce jest mniej popularne niż miar klasycznych, tj. średniej arytmetycznej, która (jeśli jest wyznaczona na podstawie odpowiedniej próby statystycznej) jest nieobciążonym estymatorem wartości oczekiwanej w populacji.

Ryzyko, które jest istotne w analizach finansowych szeregów czasowych, jest związane ze zmiennością. Stąd do pomiaru ryzyka wykorzystywane są zazwyczaj popularne miary rozrzutu, tj. wariancja i odchylenie standardowe (2.11), odchylenie przeciętne (2.14) czy współczynniki zmienności (2.19) i (2.20). Spośród wymienionych mierników najbardziej popularnym (bezwzględnym) miernikiem ryzyka jest odchylenie standardowe stóp zwrotu, postaci:

$$S_i = \sqrt{\frac{\sum_{t=1}^{T}(R_{it} - R_i)^2}{T}} \tag{9.6}$$

Przyjmując, że obserwacje $\{R_{it}\}$ są elementami próby statystycznej, zauważa się, że kwadrat S_i^* z (9.6) jest obciążonym estymatorem nieznanej wariancji z populacji. Natomiast nieobciążony estymator odchylenia standardowego, szacowany na dstawie próby, wyznacza się jako:

$$S_i^* = \sqrt{\frac{\sum_{t=1}^{T}(R_{it} - R_i)^2}{T-1}} \tag{9.7}$$

Warto jednak zauważyć, że przy znacznej liczebności próby różnice między wartościami ocen estymatorów obciążonych i nieobciążonych są nieistotne i dążą do 0 wraz ze wzrostem liczby elementów w szeregu czasowym.

5 Im liczba T jest większa, tym różnice między obu średnimi są mniejsze.

6 Miary pozycyjne są odporne na asymetrię szeregu i obserwacje odstające, dlatego są chętnie wykorzystywane w analizach szeregów o takich własnościach.

W przypadku porównań ryzyka różnych instrumentów finansowych należy korzystać ze względnych miar dyspersji, z których najczęściej stosowany jest współczynnik zmienności postaci:

$$V_i = \frac{S_i}{R_i} \tag{9.8}$$

Istnieją również inne miary ryzyka, do których zalicza się m.in. odchylenie przeciętne stóp zwrotu (2.14), a w przypadku stosowania pozycyjnych miar średnich do oceny rozrzutu wykorzystuje się tzw. odchylenie ćwiartkowe (2.18) lub międzykwartylowe. Można też wyznaczyć odchylenie standardowe w stosunku do mediany.

Warto odnotować, że wymienione mierniki ryzyka, m.in. (9.6)–(9.8), traktują odchylenia od oczekiwanych zwrotów *in plus* i *in minus* w jednakowy sposób. Dlatego ciekawymi miernikami ryzyka są semiodchylenie przeciętnie lub standardowe stóp zwrotu, które – w przeciwieństwie do wcześniej wymienionych – nie są symetryczne i wyznaczane są jako odchylenie dla ujemnych zwrotów. Semiodchylenie standardowe wyznacza się jako:

$$SV_i = \sqrt{\frac{\sum_{t=1}^{T} d_{it}^2}{T}} \tag{9.9}$$

gdzie:

$$d_{it} = \begin{cases} R_{it} - R_i & \text{gdy} \quad R_{it} - R_i < 0 \\ 0 & \text{gdy} \quad R_{it} - R_i \geq 0 \end{cases} \tag{9.10}$$

Do innych miar ryzyka można zaliczyć tzw. *tracking error*, który jest postaci:

$$TE_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left[R_{it} - R_{Bt} - (R_i - R_B)\right]^2} \tag{9.11}$$

gdzie:
R_{Bt} – stopy zwrotu z benchmarku ($t = 1, 2, \ldots, T$),
R_B – średnia stopa zwrotu z benchmarku, wyznaczona dla T okresów (momentów),
TE_i – *tracking error*, czyli odchylenie standardowe różnicowych stóp zwrotu z inwestycji i benchmarku (*differential return*),
pozostałe oznaczenia jak poprzednio.

Inwestorzy oraz zarządzający funduszami są istotnie zainteresowani tym, jakie można osiągnąć dochody oraz jakie jest ryzyko inwestycji w konkretny instrument finansowy (lub portfel), chociażby w porównaniu z inwestycjami innego

typu. W tych badaniach często wykorzystuje się (opisane w rozdziale 3) testy statystyczne, które polegają na weryfikacji hipotez dotyczących oczekiwanych stóp zwrotu oraz ich zmienności (por. m.in. [Witkowska i in. 2012, s. 110–116]).

W przypadku analiz stóp zwrotu z inwestycji kluczowym jest stwierdzenie, czy zainwestowane środki przynoszą zyski, czy też generują straty. Innymi słowy weryfikuje się hipotezę o wartości oczekiwanej szeregu czasowego (3.28) dla sprawdzenia, czy istotnie różni się od 0. Zatem hipoteza zerowa formułowana jest w postaci:

$$H_0: E(\vartheta_i) = 0 \tag{9.12}$$

wobec hipotezy alternatywnej postaci:

$$H_1: E(\vartheta_i) > 0 \text{ lub } H_1: E(\vartheta_i) < 0 \tag{9.13}$$

gdzie: $E(\vartheta_i)$ oznacza wartość oczekiwaną stóp zwrotu z i-tego instrumentu finansowego.

Do weryfikacji poprawności hipotezy zerowej dla dużych prób stosuje się statystykę testową, która ma rozkład normalny i jest postaci (3.32), czyli:

$$u_i = \frac{R_i}{S_i}\sqrt{T} \tag{9.14}$$

w przypadku małych prób korzysta się ze sprawdzianu testu (3.31) o rozkładzie t-Studenta:

$$t_i = \frac{R_i}{S_i}\sqrt{T-1} \tag{9.15}$$

gdzie oznaczenia jak poprzednio.

Czasami interesujące jest nie tylko sprawdzenie czy inwestycje przynoszą dochód, ale również porównanie stóp zwrotu do jakiegoś przyjętego benchmarku, albo do wymaganej przez inwestora stopy zwrotu wynoszącej B. Wówczas stawiane hipotezy są nieco innej postaci. Hipoteza zerowa:

$$H_0: E(\vartheta_i) = B \tag{9.16}$$

i hipotezy alternatywne:

$$H_1: E(\vartheta_i) > B \text{ lub } H_1: E(\vartheta_i) < B \tag{9.17}$$

a statystyka testowa dla dużych prób ma rozkład normalny i jest postaci:

$$u_i = \frac{R_i - B}{S_i}\sqrt{T} \tag{9.18}$$

a dla małych prób ma rozkład t-Studenta:

$$t_i = \frac{R_i - B}{S_i}\sqrt{T-1} \tag{9.19}$$

Inną ważną kwestią jest pytanie dotyczące występowania statystycznie istotnych różnic w dwóch szeregach stóp zwrotu. Może to dotyczyć zarówno stóp zwrotu z dwóch różnych instrumentów finansowych, jak i stóp zwrotu z tego samego instrumentu, ale w różnych okresach. Należy wtedy poddać weryfikacji hipotezy postaci (3.33), czyli:

$$H_0: E(\vartheta_1) = E(\vartheta_2) \tag{9.20}$$

oraz postaci (3.34), tzn.

$$H_1: E(\vartheta_1) > E(\vartheta_2) \text{ lub } H_1: E(\vartheta_1) < E(\vartheta_2) \tag{9.21}$$

W celu ich weryfikacji wykorzystuje się różne statystyki testowe, których wybór zależy od rozkładu populacji, znajomości wariancji i liczebności próby statystycznej[7]. Stąd w przypadku małych prób:

- jeśli wariancje: σ_1^2 oraz σ_2^2 w populacji są znane, to korzysta się ze statystyki (3.35), która jest teraz postaci:

$$u = \frac{R_1 - R_2}{\sqrt{\frac{\sigma_1^2}{T_1} + \frac{\sigma_2^2}{T_2}}} \tag{9.22}$$

- jeśli wariancje w populacji nie są znane, ale są równe $\sigma_1^2 = \sigma_2^2$, to korzysta się ze wzoru (3.36) na sprawdzian testu, co daje:

$$t = \frac{R_1 - R_2}{\sqrt{\frac{T_1 \cdot S_1^2 + T_2 \cdot S_2^2}{T_1 + T_2 - 2} \cdot \left(\frac{1}{T_1} + \frac{1}{T_2}\right)}} \tag{9.23}$$

- przy założeniu braku równości wariancji w obu populacjach: $\sigma_1^2 \neq \sigma_2^2$ wykorzystuje się test Cochrana-Coxa o równości średnich ze statystyką testową (3.37), czyli:

$$t = \frac{R_1 - R_2}{\sqrt{\frac{S_1^2}{T_1} + \frac{S_2^2}{T_2}}} \tag{9.24}$$

7 Tak jak to opisano w rozdziale 3 – wzory (3.35)–(3.38).

W przypadku dużych prób statystyka testowa (3.38) jest o rozkładzie normalnym, czyli:

$$u = \frac{R_1 - R_2}{\sqrt{\frac{S_1^2}{T_1} + \frac{S_2^2}{T_2}}} \tag{9.25}$$

gdzie:
$E(\vartheta_1), E(\vartheta_2)$ – wartości oczekiwane stóp zwrotu w obu porównywanych szeregach,
R_1, R_2 – średnie arytmetyczne obu porównywanych szeregów stóp zwrotu,
S_1^2, S_2^2 – wariancje wyznaczone dla obu szeregów stóp zwrotu,
σ_1^2, σ_2^2 – wariancje populacji, z których pochodzą szeregi czasowe,
T_1, T_2 – liczba obserwacji w każdym z szeregów.

Przyjęcie statystyk testowych o rozkładzie normalnym wynika z faktu, że zazwyczaj finansowe szeregi czasowe zawierają znaczną liczbę elementów. Można wtedy przyjąć, że dla dużych prób przestaje być istotne założenie o normalności rozkładu elementów szeregu, którego nie trzeba weryfikować[8]. W przypadku mniejszej liczby obserwacji w szeregu należy stosować inaczej zdefiniowane statystyki testowe. Pojawia się jednak wtedy problem konieczności weryfikacji założeń o normalności rozkładu i dylematy związane z faktem, że zazwyczaj rozkładu prawdopodobieństwa finansowych szeregów czasowych nie można uznać za rozkład Gaussa. Jednakże „większość statystyk z próby jest dość odporna na drobne odchylenia od tych założeń" [Malarska 2005, s. 141].

Pewnym rozwiązaniem jest stosowanie w tych przypadkach testów nieparametrycznych[9], które wprawdzie nie są pozbawione pewnych wymagań formalnych, ale zazwyczaj łatwiejszych do spełnienia w porównaniu z wymaganiami formułowanymi dla testów parametrycznych. Jednakże – co należy podkreślić – testy parametryczne są bardziej wrażliwe niż nieparametryczne w tym sensie, że dają bardziej wyraziste wyniki, chociaż wymagają spełnienia czasem dość restrykcyjnych założeń. Przykłady zastosowania obu rodzajów testów do analiz stóp zwrotu można znaleźć np. w pracy [Foo, Witkowska 2017].

W analizach własności szeregów stóp zwrotu porównuje się również ryzyko z inwestycji, konfrontując zmienność dwóch szeregów finansowych, co formułuje

8 W przypadku dostatecznie dużej próby statystycznej można przyjąć, że rozkład prawdopodobieństwa jest asymptotycznie normalny. Dyskusje na temat rozkładów stóp zwrotu z instrumentów polskiego rynku kapitałowego znaleźć można w [Piasecki, Tomkiewicz 2013; Tarczyński, Witkowska, Kompa 2013, s. 79–95; Kompa, Witkowska 2020, s. 45–94].

9 Przykładami testów pozwalających weryfikować hipotezę o równości stóp zwrotu są testy: sumy rang Wilcoxona i Manna-Whitneya dla porównań dwóch oraz test Kruskala-Wallisa – dla więcej niż dwóch szeregów. Natomiast do porównań wariancji dwóch i więcej szeregów finansowych można zaproponować test Levene'a. Opis wymienionych testów znaleźć można m.in. w [Domański 1990; Malarska 2005; Zieliński 1997].

się w postaci pary hipotez (3.42)–(3.43), czyli: H_0: $\sigma_1^2 = \sigma_2^2$ i H_0: $\sigma_1^2 > \sigma_2^2$. Sprawdzian testu (3.44) ma rozkład Fishera-Snedecora o $(T_1 - 1)$ i $(T_2 - 1)$ stopniach swobody i jest postaci: $F = \frac{S_1^2}{S_2^2} = \frac{S_{max}^2}{S_{min}^2}$. Statystykę testową tej postaci można również wykorzystać do porównań więcej niż dwóch szeregów.

Weryfikacja hipotez o zgodności rozkładów stóp zwrotu z rozkładem normalnym polega na zastosowaniu testów nieparametrycznych, takich jak test Kołmogorowa-Lillieforsa, Jarque'a-Bera, Doornika-Hansena, czy Shapiro-Wilka, który może być stosowany również do małych prób[10]. Rozkład Gaussa jest rozkładem symetrycznym o kurtozie równej 0, zatem jeśli szeregi nie spełniają jednego z tych warunków, to nie mogą mieć rozkładu normalnego. Dlatego wystarczy przeprowadzić znacznie prostsze testy parametryczne, aby określić, czy kształt rozkładu jest zbliżony do krzywej normalnej.

W celu sprawdzenia symetrii rozkładu bada się współczynniki skośności. Jeżeli wartość współczynnika skośności $A(\vartheta_i)$ jest równa 0, to rozkład szeregu stóp zwrotu jest symetryczny. W rzeczywistości wartość tego współczynnika odbiega od 0, z tego względu testuje się jego istotność, stawiając hipotezę zerową postaci (3.51) wobec hipotezy alternatywnej postaci (3.52). Sprawdzianem testu jest standaryzowany współczynnik skośności SA_i (3.53). Sprawdzenia stopnia spłaszczenia rozkładu analizowanej zmiennej w stosunku do rozkładu normalnego dokonuje się w trakcie testowania kurtozy. Hipotezy zerowa i alternatywne są formułowane jako (3.54) i (3.55), a do oceny, czy wartość kurtozy istotnie różni się od 0 wykorzystuje się standaryzowany współczynnik kurtozy (3.56).

Warto dodać, że przedstawione w tym podrozdziale testy statystyczne stanowią tylko drobny przykład badań, jakie prowadzone są w celu weryfikacji statystycznych własności stóp zwrotu. Przykładowo w pracach [Zamojska 2012, s. 166–171; Czekaj 2014, s. 64–70] dodatkowo bada się stacjonarność, ergodyczność, normalność i własności ogonów rozkładu stóp zwrotu.

9.3. Modele opisujące kształtowanie się stóp zwrotu

W praktyce inwestowania istnieją dwa najbardziej znane modele, za pomocą których wyjaśnia się kształtowanie stóp zwrotu z pojedynczego instrumentu finansowego lub portfela inwestycyjnego w istniejących warunkach rynkowych. Są to

10 Omówienie tych testów można znaleźć w wielu opracowaniach, wszystkie wyżej wymienione przedstawiono w pracach [Tarczyński i in. 2013, s. 21–23; Witkowska 2016b, s. 27–50; Kompa, Witkowska 2020, s. 45–94], w których również przedstawiono wyniki ich aplikacji do wnioskowania o finansowych szeregach czasowych.

model jednowskaźnikowy, określany w literaturze również jako model Sharpe'a, oraz model wyceny aktywów kapitałowych, skrótowo nazywany CAPM (*capital asset pricing model*). Oba, mimo często pojawiającej się w literaturze przedmiotu krytyki[11], odgrywają istotną rolę przy:

- testowaniu teorii wyceny aktywów,
- szacowaniu kosztu kapitału,
- ocenie efektywności portfeli inwestycyjny, a zatem i efektowności funduszy inwestycyjnych lub emerytalnych.

Jednym z ważniejszych narzędzi wspomagających podejmowanie decyzji inwestycyjnych na rynku kapitałowym stał się jednowskaźnikowy model rynku (*single-index model*), opracowany przez Sharpe'a[12]. Jest on najprostszym modelem, opisującym powiązania zmian wartości instrumentów finansowych (np. akcji) lub portfela inwestycyjnego z zachowaniem całego rynku. Rynek jest w modelu reprezentowany przez tzw. indeks lub czynnik rynku, którym w przypadku rynku kapitałowego jest zazwyczaj jeden z indeksów giełdowych. W sytuacji, kiedy badania dotyczą innych instrumentów rynku finansowego konstruuje się portfele charakteryzujące cały rynek finansowy lub jego segmenty[13].

W modelu Sharpe'a przyjmuje się szereg założeń (zob. [Tarczyński 1997, s. 104; Witkowska i in. 2012, s. 207–208]). Jednakże sama konstrukcja modelu jednowskaźnikowego opiera się na obserwacji, że ceny akcji losowo wybranych spółek w większości przypadków rosną w czasie dobrej koniunktury na giełdzie, a kiedy sytuacja na rynku się pogarsza – spadają. Stąd w modelu Sharpe'a stopa zwrotu z danego aktywu jest uzależniona od stopy zwrotu z indeksu rynku. Zatem stopę zwrotu z inwestycji i-tej w t-tym okresie (lub momencie) opisuje się za pomocą modelu regresji (4.30) lub modelu (6.7) z jedną zmienną objaśniającą, czyli (por. [Tarczyński 1997, s. 107; Witkowska i in. 2012, s. 207–221; Tarczyński i in. 2013; Witkowska 2016b, s. 42–46]):

$$R_{it} = \alpha_i + \beta_i R_{rt} + \varepsilon_{it} \tag{9.26}$$

gdzie dla każdego t ($t = 1, 2, \ldots, T$):
R_{it} – stopa zwrotu z i-tego instrumentu lub portfela,
R_{rt} – stopa zwrotu z tzw. czynnika rynku, np. reprezentowanego przez indeks giełdowy,
α_i – wyraz wolny modelu, będący składnikiem stopy zwrotu z waloru i, niezależnym od sytuacji na rynku,

---

11 Niektóre argumenty, będące podstawą krytyki wspomnianych modeli, przedstawiono m.in. w pracy [Tarczyński i in. 2013, s. 44–45].

12 Omówienie modelu Sharpe'a znaleźć można m.in. w pracach [Tarczyński 1997, s. 103–111; Witkowska i in. 2012, s. 207–213], a praca [Tarczyński i in. 2013] jest niemal w całości poświęcona temu modelowi.

13 W wielu badaniach korzysta się z odpowiednio skonstruowanych portfeli odniesienia stanowiących benchmarki lub pełniących rolę indeksu rynku (por. [Witkowska 2016b, s. 87–89]). Szerzej na temat konstrukcji benchmarków znaleźć można w [Borowski 2011].

β_i – stały w czasie współczynnik kierunkowy równania, który mierzy oczekiwaną zmianę R_i przy danej zmianie R_r,
ε_{it} – składnik losowy równania,
t – numer obserwacji w szeregu czasowym.

W równaniu (9.26) najważniejszym dla inwestora jest parametr β_i, nazywany współczynnikiem agresywności waloru, np. akcji, lub po prostu współczynnikiem beta. W zależności od wartości, jakie przyjmuje współczynnik beta, objaśniane walory charakteryzują się określonymi cechami.

- Jeżeli $\beta < 0$, to stopa zwrotu z waloru (portfela) reaguje na zmiany przeciwnie do zachowań rynku. Jest to stosunkowo rzadki przypadek, bardzo pożądany w sytuacji spodziewanego spadku stóp zwrotu większości walorów na rynku.
- Gdy $\beta = 0$, stopa zwrotu z instrumentu (portfela) nie reaguje na zmiany rynku. Taki walor (portfel) jest pozbawiony ryzyka rynku, może nim być instrument finansowy gwarantowany przez Skarb Państwa, np. bon skarbowy lub obligacja rządowa.
- Dla $0 < \beta < 1$ stopa zwrotu z instrumentu (portfela) w małym stopniu reaguje na zmienność rynku. Taki walor (portfel) określany jest mianem defensywnego, gdyż w okresie hossy dochody są mniejsze niż z indeksu rynku, ale w okresie bessy straty są również mniejsze.
- Jeśli $\beta = 1$, to stopa zwrotu z aktywu (portfela) zmienia się w takim samym stopniu jak stopa zwrotu indeksu giełdowego, czyli reaguje jak portfel rynkowy.
- Gdy $\beta > 1$, stopa zwrotu z instrumentu (portfela) silnie reaguje na zmiany zachodzące na rynku. Instrument (lub portfel) określany jest mianem waloru agresywnego (portfela agresywnego). Posiadacz takich walorów może liczyć na większy niż w przypadku portfela rynkowego zysk w okresie wzrostów, ale i większą stratę w okresie spadków.

W najprostszym ujęciu w okresach znacznego wzrostu cen zaleca się kupowanie instrumentów finansowych o jak najwyższym poziomie współczynnika beta. Z kolei przy znacznych spadkach cen aktywów na giełdzie najlepsza jest inwestycja w instrumenty o współczynniku beta mniejszym od 0.

Model Sharpe'a pozwala również oszacować ryzyko i przeprowadzić jego dekompozycję, tj. wyróżnić z ryzyka całkowitego modelowanego waloru (lub portfela) część dywersyfikowaną i niedywersyfikowaną. Warto dodać, że beta jest traktowana jako miara ryzyka instrumentów i portfeli inwestycyjnych, dlatego poświęca się jej wiele uwagi w literaturze przedmiotu[14].

14 Przykładowo praca [Tarczyński i in. 2013] została w całości poświęcona problemom teoretycznym i praktycznym związanym z wyznaczaniem tego parametru i ocenie jego stabilności.

Do oszacowania modelu jednowskaźnikowego wykorzystuje się najczęściej klasyczną metodę najmniejszych kwadratów (MNK), wówczas otrzymuje się:

$$\hat{R}_{it} = \hat{\alpha}_i + \hat{\beta}_i R_{rt} \tag{9.27}$$

gdzie dla każdego t ($t = 1, 2, \ldots, T$):
$\hat{R}_{it}$ – wartości teoretyczne (tj. wyznaczone z oszacowanego modelu) stóp zwrotu,
$\hat{\alpha}_i$, $\hat{\beta}_i$ – oceny estymatorów parametrów modelu.

Model wyceny aktywów kapitałowych CAPM należy do grupy modeli równowagi rynku kapitałowego. Model ten jest wynikiem przeprowadzonych w latach 60. ubiegłego stulecia badań nad rynkiem finansowym i wywodzi się z podejścia Markowitza[15]. CAPM jest najczęściej stosowanym modelem równowagi rynku kapitałowego.

Model CAPM obrazuje, jak będą się kształtowały ceny oraz stopy zwrotu z instrumentu lub portfela na rynku, gdy działają na nim inwestorzy postępujący w myśl teorii portfelowej (czyli inwestujący w portfele efektywne znajdujące się na granicy tzw. pocisku Markowitza). Model CAMP został zbudowany przy założeniach, które są w większości identyczne z tymi, jakie przyjmuje się przy budowie modelu Sharpe'a.

Model wyceny aktywów kapitałowych składa się z dwóch równań opisujących linię rynku kapitałowego (*capital market line*) CML i linię rynku papierów wartościowych (*security market line*) SML. Interesujące jest zwłaszcza równanie linii rynku papierów wartościowych, będące rozwinięciem modelu jednowskaźnikowego:

$$R_{it} = R_{ft} + \beta_i(R_{rt} - R_{ft}) + \varepsilon_{it} \tag{9.28}$$

Parametry modelu szacuje się (najczęściej MNK) na podstawie nieco zmienionej postaci tego modelu, a mianowicie:

$$R_{it} - R_{ft} = \alpha_i + \beta_i(R_{rt} - R_{ft}) + \varepsilon_{it} \tag{9.29}$$

gdzie dla każdego t ($t = 1, 2, \ldots, T$):
R_{ft} – instrument wolny od ryzyka,
$(R_{it} - R_{ft})$ – premia za ryzyko, czyli nadwyżkowy dochód z inwestycji ponad ten, jaki można osiągnąć, inwestując w instrument wolny od ryzyka,
$(R_{rt} - R_{ft})$ – premia za ryzyko rynkowe, czyli różnica w dochodzie uzyskanym z czynnika rynkowego i instrumentu wolnego od ryzyka,
pozostałe oznaczenia jak poprzednio.

15 Markowitz zaproponował sposób konstruowania krzywej portfeli inwestycyjnych. Każdy portfel na tej krzywej charakteryzuje się najwyższą oczekiwaną stopą zwrotu, jaką można osiągnąć przy danym poziomie ryzyka (związanym z danym portfelem). Ze względu na skomplikowanie obliczenia została opracowana uproszczona wersja tego modelu przez Williama Sharpe'a. Jest to model jednowskaźnikowy, który był punktem wyjściowym do CAPM. Dzięki Sharpe'owi teoria portfelowa mogła zaistnieć w praktyce.

Linia rynku papierów wartościowych SML pokazuje, czy badany portfel jest niedoszacowany (*underpriced*) czy też przeszacowany (*overpriced*). Zestawiając interpretacje poszczególnych elementów linii rynku papierów wartościowych, można powiedzieć, że stopa zwrotu z portfela jest równa sumie ceny czasu oraz ceny ryzyka. Cena ryzyka to iloczyn współczynnika beta badanego portfela i premii za ryzyko. Zatem przy różnych wartościach współczynnika beta instrumentu lub portfela stopa zwrotu z waloru może się różnie zachowywać. W modelu CAPM oprócz indeksu rynku pojawia się instrument wolny od ryzyka, który jest zazwyczaj reprezentowany przez obligacje skarbowe, indeks obligacji lub stopy procentowe.

Model wyceny aktywów kapitałowych jest, podobnie jak model Sharpe'a, liniową funkcją regresji. Po oszacowaniu parametrów modelu (9.29) otrzymuje się:

$$\widehat{(R_{it} - R_{ft})} = \hat{\alpha}_i + \hat{\beta}_i (R_{rt} - R_{ft}) \tag{9.30}$$

gdzie dla każdego t ($t = 1, 2, \ldots, T$):
$\widehat{(R_{rt} - R_{ft})}$ – teoretyczna (tj. wyznaczona z modelu) wartość premii za ryzyko,
pozostałe oznaczenia jak poprzednio.

Oszacowane modele Sharpe'a i CAPM podlegają, jak wszystkie modele ekonometryczne, weryfikacji statystycznej, m.in. na podstawie wzorów (4.35)–(4.39) i ocenie stopnia dopasowania modeli do wartości empirycznych (4.40). Literatura z zakresu ekonometrii dostarcza bogatej metodologii w tym zakresie, jak również wskazuje na inne metody estymacji tych modeli w przypadku, kiedy nie są spełnione założenia stosowalności klasycznej MNK. Dodatkowo wiele niepożądanych własności, jakimi charakteryzują się szeregi stóp zwrotu, skłania do korzystania z bardziej zaawansowanych modeli, np. klasy GARCH. Jednakże w pracy [Tarczyński i in. 2013], w której omówiono podstawowe zagadnienia związane z estymacją modelu jednowskaźnikowego, wykazano, że w większości przypadków stosowanie innych metod estymacji niż klasyczna MNK nie zmienia w istotny sposób wartości parametru beta.

W analizach współczynnika β istotne jest również sprawdzenie, czy parametr jest mniejszy lub większy od jedności, co wiąże się z wcześniej omówionymi charakterystykami walorów, które na podstawie wartości beta klasyfikowane są jako agresywne lub defensywne. Służy temu rozwinięcie testu (4.35)–(4.39), który jest postaci:

$$H_0: \beta_i = \beta^* \tag{9.31}$$

$$H_1: \beta_i > \beta^* \text{ lub } H_1: \beta_i < \beta^* \tag{9.32}$$

Sprawdzianem hipotezy zerowej jest statystka o rozkładzie t-Studenta postaci:

$$t = \frac{\hat{\beta}_i - \beta^*}{S(\hat{\beta}_i)} \tag{9.33}$$

gdzie:
β^* – wartość parametru założona w hipotezie zerowej, przy czym w celu sprawdzenia charakteru waloru przyjmuje się $\beta^* = 1$.

Wartości ocen estymatorów parametrów równań regresji, do których należą oba omawiane modele, zależą m.in. od:

- długości okresu analizy,
- sposobu pomiaru zmiennych, tj. rodzaju stóp zwrotu,
- specyfikacji modelu.

Pojawia się zatem pytanie, na ile zmiany ocen estymatorów parametrów świadczą o „prawdziwej" ich zmianie. Aby na nie odpowiedzieć, można wykorzystać wcześniej omówiony test (9.31)–(9.33), podstawiając w miejsce β^* oszacowane wartości parametrów. Innymi słowy $\beta^* = \hat{\beta}_i^m$. Weryfikowane hipotezy są wtedy postaci:

$$H_0: \beta_i^n = \hat{\beta}_i^m \tag{9.34}$$

$$H_1: \beta_i^n > \hat{\beta}_i^m \text{ lub } H_1: \beta_i^n < \hat{\beta}_i^m \tag{9.53}$$

Sprawdzianem jest natomiast statystyka t-Studenta postaci:

$$t = \frac{\hat{\beta}_i^n - \hat{\beta}_i^m}{S(\hat{\beta}_i^n)} \tag{9.36}$$

gdzie:
n, m oznaczają próby badawcze,
$\hat{\beta}_i^n, \hat{\beta}_i^m$ – bety pochodzące z dwóch porównywanych modeli oszacowanych dla i-tego waloru.

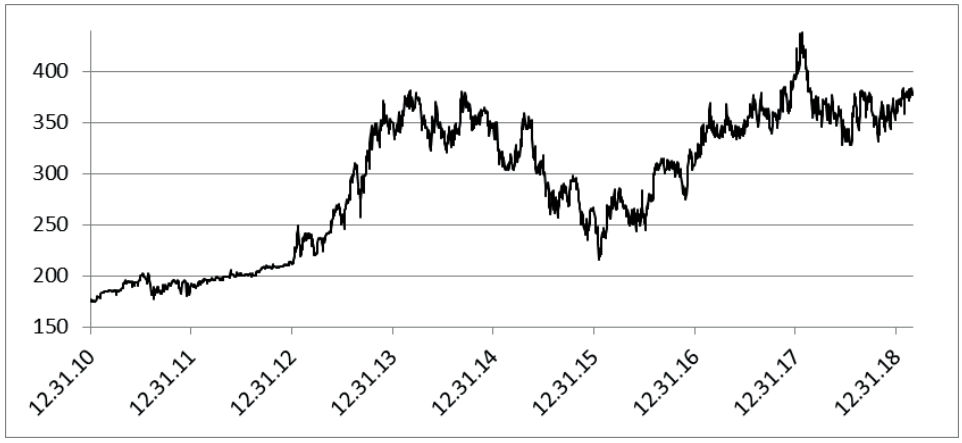
Jednym z istotnych elementów analiz inwestycji oraz zarządzania kapitałem przez instytucje zbiorowego inwestowania jest badanie tzw. persystencji (*persistence*), czyli sprawdzenie, czy osiągane przez nie wyniki inwestycyjne są powtarzalne albo czy fundusze inwestycyjne utrzymują swoje (czołowe lub w końcówce) pozycje rankingowe w różnych warunkach. W zasadzie badanie persystencji sprowadza się do oceny stabilności pozycji rankingowych, która może dotyczyć różnych aspektów inwestowania, zarówno w odniesieniu do stóp zwrotu, jak i ryzyka oraz mierników efektywności inwestycyjnej. Jedną z metod badania persystencji jest korzystanie ze współczynnika korelacji rang Spearmana (4.12) oraz testu weryfikującego jego istotność (4.13)–(4.17).

9.4. Wykorzystanie metod statystycznych w analizach finansowych szeregów czasowych

W niniejszym podrozdziale przedstawiona zostanie analiza inwestycji w akcje przykładowej spółki notowanej na Giełdzie Papierów Wartościowych w Warszawie. Wybraną spółką jest Santander Bank Polska SA, a badanie dotyczy kształtowania się cen i stóp zwrotu z akcji w okresie 31.12.2010–28.02.2019 na podstawie cen zamknięcia pochodzących z portalu stooq.pl.

9.4.1. Analiza notowań wybranej spółki

Analizy zazwyczaj rozpoczyna się od badania szeregu notowań cen. Z uwagi na dalsze badania dane oryginalne (2035 obserwacji) zostały uzupełnione w dniach, kiedy nie było notowań (np. w dni świąteczne). Imputacja danych polegała na wpisaniu kursu akcji z poprzedniego (ostatniego) dnia notowania, w ten sposób powstał szereg danych uzupełnionych liczący 2130 obserwacji. Wykres notowań oryginalnych przedstawiono na rys. 9.1, a porównanie obu szeregów, tj. oryginalnego i uzupełnionego w tab. 9.1.



Rysunek 9.1. Wykres notowań spółki Santander Bank Polska SA w okresie 31.12.2010–28.02.2019

Źródło: opracowanie własne.

Jak łatwo zauważyć w tab. 9.1, zastosowana metoda imputacji nie spowodowała istotnej zmiany podstawowych parametrów szeregu czasowego. Z przedstawionej analizy wynika, że w badanym, ponad ośmioletnim okresie cena akcji zmieniała

się znacząco, o czym świadczy rozstęp 263,38 PLN, chociaż średnia zmienność notowań mierzona odchyleniem standardowym wynosi niecałe 68 PLN, czyli 23% średniej arytmetycznej. Warta odnotowania jest relatywnie mała różnica między średnią a medianą.

Analizując przebieg szeregu na rys. 9.1, zauważa się jego znaczną zmienność, zatem w celu wygładzenia szeregu i prognozowania jego dalszego przebiegu wykorzystuje się różne metody w tym średnie ruchome. Warto dodać, że inwestorzy przy podejmowaniu decyzji (zwłaszcza ci stosujący analizę techniczną) często korzystają z różnego rodzaju średnich ruchomych i modeli trendu[16]. Na rys. 9.2 porównano wykresy prostych średnich ruchomych (5.19) lub (5.31) o kroku uśredniana 15, 25 i 35 wyznaczonych dla kursów akcji. Dla lepszej ilustracji zagadnienia na rysunku zamieszczono wykres danych z okresu 25.01–14.06.2011 roku (100 notowań).

Tabela 9.1. Porównanie podstawowych parametrów danych oryginalnych i uzupełnionych

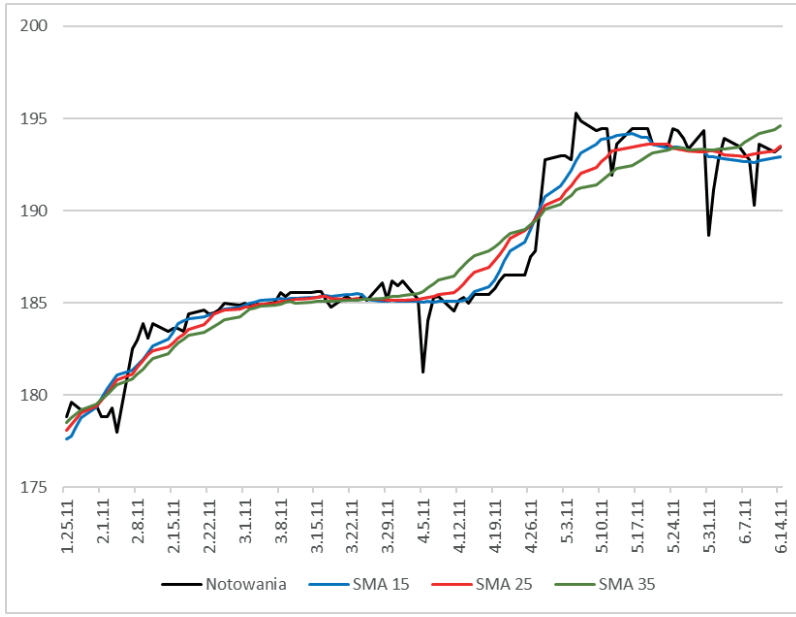
Dzienne notowania indeksu				
Podstawowe miary			Dane	
			oryginalne	uzupełnione
Wartość minimalna	(Max)		174,69	174,69
Wartość maksymalna	(Min)		438,07	438,07
Średnia arytmetyczna	($\bar{y}$)	(2.2)	290,85	290,89
Mediana	(Me)	(2.5)	305,59	305,90
Rozstęp	(R)	(2.10)	263,38	263,38
Odchylenie standardowe	(S)	(2.11), (3.4)	67,98	67,99
Współczynnik zmienności	(V)	(1.19)	0,23	0,23
Skośność	(A)	(2.23)	–0,26	–0,26
Kurtoza	(K)	(2.28)	–1,34	–1,35

Uwaga: w nawiasach podano numery wzorów i oznaczenia stosowane w tabelach 9.2.– 9.10.

Źródło: opracowanie własne.

Analiza kursów akcji umożliwia zaobserwowanie zmian zachodzących w notowaniach i wskazanie określonych tendencji rynkowych, co pokazano, nanosząc na wykres notowań oszacowane funkcje trendu (5.18) dla wyróżnionych podokresów (zob. rys. 9.3). Na tej podstawie określono punkty przegięcia funkcji trendu, czyli sześć okresów o jednorodnej tendencji rynkowej ceny akcji badanej spółki, które zapisano w tab. 9.2.

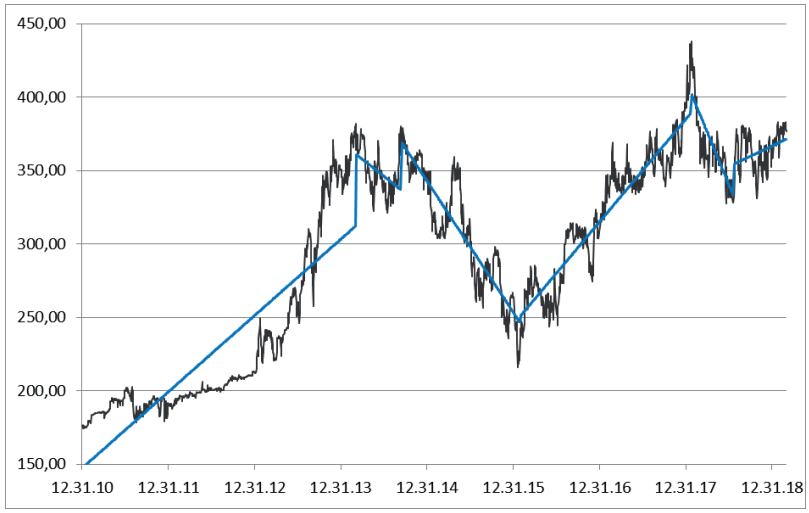
16 Porównaj chociażby pracę [Borowski 2017], poświęconą analizie technicznej.



Rysunek 9.2. Porównanie notowań i średnich ruchomych

Uwaga: symbole SMA15, SMA25 i SMA35 oznaczają średnie ruchome o krokach uśredniania odpowiednio: 15, 25 i 35.

Źródło: opracowanie własne.



Rysunek 9.3. Funkcje trendu wyznaczone dla wyróżnionych podokresów

Źródło: opracowanie własne.

Czasami analizy kursów giełdowych prowadzone są wokół istotnych wydarzeń gospodarczych i politycznych, które mogą mieć wpływ na rynek finansowy. W badaniach podjęto próbę oceny, czy i na ile wybory prezydenckie w Polsce w 2015 roku wpłynęły na notowania banku Santander. Oszacowano zatem funkcje trendu w obu podokresach i otrzymano następujące wyniki:

$y = -5466{,}2 + 0{,}1385t \qquad R^2 = 0{,}8274$

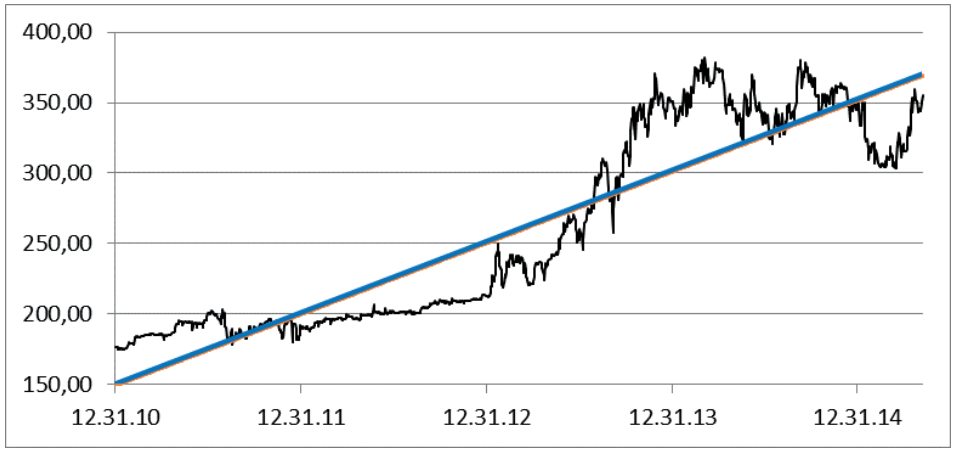
$y = -3581{,}1 + 0{,}0912t \qquad R^2 = 0{,}6500$

Jak można zauważyć, widoczna jest zmiana trendu, który wprawdzie pozostaje rosnący, ale o innym – mniejszym[17] kącie nachylenia, co ilustrują rys. 9.4 i 9.5.

Tabela 9.2. Wyróżnione okresy o jednorodnej tendencji rynkowej

Daty	Liczba obserwacji	Rodzaj tendencji rynkowej
31.12.2014–05.03.2014	829	hossa
06.03.2014–10.09.2014	135	stagnacja
11.09.2014–26.01.2016	359	bessa
27.01.2016–17.01.2018	516	hossa
18.01.2018–13.07.2018	127	bessa
14.07.2018–28.02.2018	164	stagnacja

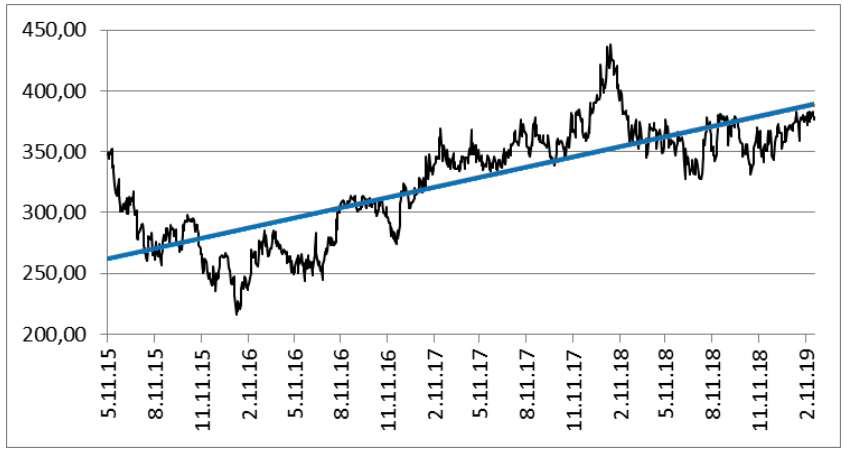
Źródło: opracowanie własne.



Rysunek 9.4. Kurs akcji przed wyborami prezydenckimi

Źródło: opracowanie własne.

17 Współczynnik kierunkowy prostej jest tangensem kąta jej nachylenia.

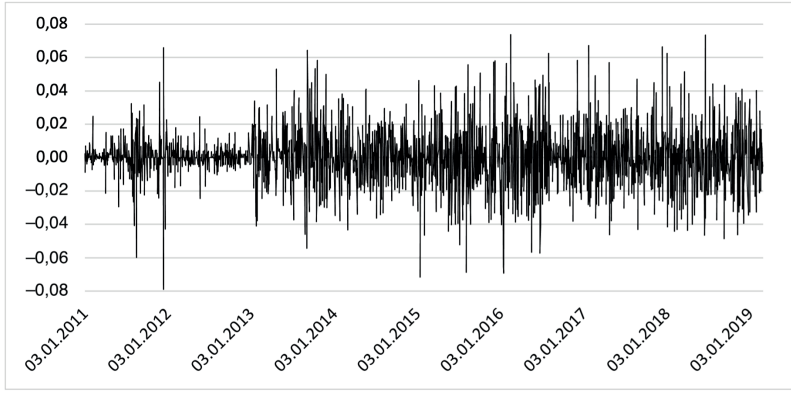


Rysunek 9.5. Kurs akcji po wyborach prezydenckich

Źródło: opracowanie własne.

9.4.2. Analiza szeregu stóp zwrotu

Dalsze analizy prowadzono dla logarytmicznych stóp zwrotu (9.2), pokazanych na rys. 9.6, który jest charakterystycznym wykresem zmienności, ilustrującym dzienne zwroty. Widać na nim okresy mniejszej (np. w drugiej połowie 2012 roku) i zwiększonej (np. w końcu 2011 roku lub na początku 2016 roku) zmienności kursów akcji badanej spółki, a także tzw. efekt skupiania (gromadzenia) zmienności, będący typową własnością finansowych szeregów czasowych. Analizę dziennych stóp zwrotu uzupełniają dane zamieszczone w tab. 9.3.



Rysunek 9.6. Logarytmiczne dzienne stopy zwrotu z akcji spółki Santander

Źródło: opracowanie własne.

Tabela 9.3. Miary opisowe wyznaczone dla procentowych dziennych logarytmicznych stóp zwrotu

Podstawowe miary	Dzień tygodnia					
	Wszystkie	Poniedziałek	Wtorek	Środa	Czwartek	Piątek
Liczba obserwacji	2129	426	426	426	426	425
Min (%)	–7,91	–6,94	–5,23	–7,91	–7,17	–6,52
Max (%)	7,38	5,83	6,25	7,38	6,64	6,45
$\bar{y}$ (%)	0,04	0,03	0,06	0,10	0,07	–0,09
Me (%)	0,00	0,00	0,00	0,00	0,00	0,00
Q1 (%)	–0,77	–0,64	–0,82	–0,67	–0,74	–0,89
Q3	0,83	0,84	0,84	0,80	0,88	0,72
R (%)	15,29	12,77	11,48	15,29	13,81	12,97
S (%)	1,69	1,60	1,64	1,73	1,90	1,56
V	47,44	48,35	25,85	17,75	25,95	17,46
Q (%)	0,80	0,74	0,83	0,74	0,81	0,81
A	0,11	–0,08	0,22	0,24	0,17	–0,13
K	2,23	1,65	1,56	3,03	2,14	2,05
Weryfikacja hipotez dotyczących: (a) zerowych zwrotów, (b) skośności i (c) kurtozy						
(a): statystyka (9.14)	0,9727	0,4269	0,7983	1,1631	0,7954	–1,1805
(b): statystyka (3.53)	**2,1550**	–0,6589	**1,8677**	**1,9879**	1,4042	–1,1141
(c): statystyka (3.56)	**42,0452**	**13,8689**	**13,1683**	**25,5563**	**18,0095**	**17,2914**

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05. Weryfikowane hipotezy dotyczą zerowych wartości parametrów opisujących stopy zwrotu: (a) nadziei matematycznej, (b) współczynnika skośności i (c) kurtozy.

Min oznacza wartość minimalną, Max – wartość maksymalną, $\bar{y}$ – średnią arytmetyczną, S – odchylenie standardowe, V – współczynnik zmienności, Q – odchylenie ćwiartkowe, Q1 – kwartyl I, Q3 – kwartyl III, a reszta oznaczeń jak w tab. 9.1.

Źródło: opracowanie własne.

Co można odczytać z tab. 9.3, stopy zwrotu charakteryzują się znaczną zmiennością, która jest różna w poszczególne dni tygodnia. Najmniejsza (mierzona współczynnikiem zmienności) jest w środy i piątki, a największa w poniedziałki. Widać również, że największe średnie stopy zwrotu obserwowane są w środy, a najmniejsze w piątki. Mediana jest równa 0, co oznacza, że połowa zwrotów jest dodatnich, a druga połowa ujemnych. Potwierdza to analiza pozostałych kwartylów. Tab. 9.5–9.7 pokazują kształtowanie się tygodniowych i miesięcznych stóp zwrotu.

Tabela 9.4. Wartości statystyk testowych wyznaczone dla dziennych stóp zwrotu oraz wariancji z poszczególnych dni tygodnia

Dni tygodnia	Porównanie stóp zwrotu (9.25)				Porównanie ryzyka (9.34)			
	Wtorek	Środa	Czwartek	Piątek	Wtorek	Środa	Czwartek	Piątek
Poniedziałek	–0,2739	–0,5628	–0,3341	1,1301	1,0535	1,1639	**1,4147**	1,0517
Wtorek		–0,2924	–0,0806	1,3920		1,1048	**1,3428**	1,1080
Środa			0,1923	**1,6545**			**1,2155**	**1,2240**
Czwartek				1,3638				**1,4878**

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05. Na czerwono oznaczono sytuacje, kiedy wariancja w dniu opisanym w kolumnie jest mniejsza od tej w dniu opisanym w wierszu.

Źródło: opracowanie własne.

Na potwierdzenie tych spostrzeżeń warto skorzystać z testów statystycznych, pozwalających nie tylko na sprawdzenie, czy oczekiwane zwroty istotnie różnią się od 0, weryfikując hipotezę (9.12), ale i sprawdzić, czy są one różne w sąsiednich dniach tygodnia, stawiając hipotezę zerową (9.20). Zbadano również ryzyko obserwowane w różnych dniach tygodnia, wykorzystując test (3.42)–(3.44). Przedstawione w tab. 9.3 i 9.5 wartości statystyk nie dają podstaw do odrzucenia hipotez zerowych na poziomie istotności 0,05 w odniesieniu do stóp zwrotu. Oznacza to, że zarówno dzienne, jak i tygodniowe stopy zwrotu wyznaczone dla poszczególnych dni tygodnia nie różnią się istotnie od 0. Jedynie oczekiwane tygodniowe zwroty wyznaczone na podstawie wszystkich obserwacji są istotnie dodatnie. W przypadku dziennych zwrotów porównano również oczekiwane zwroty i ryzyko w sąsiednich dniach tygodnia (wartości statystyk znajdują się na przekątnych w tab. 9.4).

Weryfikacja wspomnianych hipotez statystycznych znajduje również zastosowanie w sytuacji badania tzw. informacyjnej efektywności rynku[18] w formie słabej do sprawdzania czy w szeregach stóp zwrotu lub w analizie ryzyka nie pojawiają się tzw. efekty dnia tygodnia, polegające na tym, że w niektóre dni oczekiwane stopy zwrotu lub ryzyko są istotnie większe (lub mniejsze) niż w pozostałe. Tak więc, jak pokazano w tab. 9.4, ryzyko jest istotnie większe w czwartki w stosunku do pozostałych dni tygodnia oraz w środy w stosunku do piątku. Zidentyfikowano również istotnie większe zwroty w środy niż w piątki (tab. 9.4), co upoważnia do stwierdzenia występowania efektu dnia tygodnia, a co zatem idzie – braku informacyjnej efektywności.

18 Jest to znana w teorii finansów hipoteza rynku efektywnego (*efficient market hypothesis*) sformułowana przez Eugene'a E. Famę (1970).

Tabela 9.5. Miary opisowe wyznaczone dla procentowych tygodniowych logarytmicznych stóp zwrotu

	Dzień tygodnia					
Podstawowe miary	Wszystkie	Poniedziałek	Wtorek	Środa	Czwartek	Piątek
Liczba obserwacji	2550	425	425	425	425	425
Min (%)	–12,37	–10,82	–12,35	–11,63	–12,37	–10,08
Max (%)	14,35	9,64	12,79	13,08	14,35	12,44
$\bar{y}$ (%)	0,18	0,18	0,18	0,18	0,18	0,18
Me (%)	0,12	0,13	0,04	0,13	0,21	0,21
Q1 (%)	–1,49	–1,22	–1,34	–1,68	–1,50	–1,58
Q3 (%)	1,69	1,46	1,93	1,65	1,78	1,72
R (%)	26,73	20,46	25,15	24,71	26,73	22,51
S (%)	3,22	3,09	3,27	3,36	3,43	3,22
V	18,31	16,99	18,13	18,46	19,01	17,74
Q (%)	1,59	1,34	1,64	1,67	1,64	1,65
A	0,14	0,17	–0,21	0,13	0,23	0,22
K	1,37	1,41	1,61	1,16	1,72	1,07
Weryfikacja hipotez dotyczących: (a) zerowych zwrotów, (b) skośności i (c) kurtozy						
(a): statystyka (9.14)	**2,7164**	1,2092	1,1332	1,1130	1,0805	1,1583
(b): statystyka (3.53)	**2,9353**	1,4189	**–1,7445**	1,1230	**1,9231**	**1,8421**
(c): statystyka (3.56)	**27,8005**	**11,8670**	**13,5224**	**9,7806**	**14,4877**	**9,0003**

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05.

Min oznacza wartość minimalną, Max – wartość maksymalną, $\bar{y}$ – średnią arytmetyczną, S – odchylenie standardowe, V – współczynnik zmienności, Q – odchylenie ćwiartkowe, Q1 – kwartyl I, Q3 – kwartyl III, a reszta oznaczeń jak w tab. 9.1.

Źródło: opracowanie własne.

Tygodniowe stopy zwrotu, liczone dla kolejnych dni jako zwroty w danym dniu w relacji do notowania z tego samego dnia tygodnia zaobserwowane w poprzednim tygodniu (tab. 9.5) są zdecydowanie wyższe od dziennych zwrotów, ponieważ są dodatnie w każdym dniu tygodnia i charakteryzują się mniejszą zmiennością mierzoną współczynnikiem zmienności wyznaczonym dla odchylenia standardowego. W tab. 9.6 przedstawiono zwroty miesięczne, liczone jako jednookresowe logarytmiczne stopy zwrotu porównujące notowania na ostatni dzień miesiąca. Z kolei tab. 9.7 zawiera miesięczne zwroty wyznaczone jako średnie z dziennych stóp zwrotu zaobserwowanych w ciągu miesiąca. Warto odnotować, że przeprowadzona analiza miesięcznych stóp zwrotu nie dała podstaw do stwierdzenia, że różnią się one istotnie od 0 i to niezależnie od tego, jak zostały wyznaczone.

Stwierdzono natomiast istotnie większe ryzyko w lipcu niż w czerwcu w przypadku jednookresowych miesięcznych stóp zwrotu.

Tabela 9.6a. Miary opisowe wyznaczone dla procentowych jednookresowych miesięcznych logarytmicznych stóp zwrotu

Podstawowe miary	Wszystkie	Styczeń	Luty	Marzec	Kwiecień	Maj	Czerwiec
Liczba obserwacji	98	9	9	8	8	8	8
Min (%)	–17,91	–8,60	–11,55	–6,90	–9,69	–8,94	–4,20
Max (%)	14,89	8,63	7,30	11,16	7,48	10,46	6,03
$\bar{y}$ (%)	0,76	1,16	1,13	0,32	0,85	–1,62	0,28
Me (%)	1,08	1,89	1,89	–0,54	3,05	–2,50	0,62
Q1 (%)	–2,92	1,31	0,44	–2,66	–2,27	–3,98	–0,71
Q3 (%)	4,28	4,26	5,20	2,34	4,74	–0,80	1,13
R (%)	32,80	17,23	18,85	18,06	17,17	19,40	10,22
S (%)	5,95	5,92	5,72	5,40	6,45	5,74	3,25
V	7,79	5,10	5,07	17,07	7,61	3,55	11,66
Q (%)	3,60	1,47	2,38	2,50	3,50	1,59	0,92
A	–0,22	–0,95	–1,44	1,02	–0,86	1,31	0,16
K	0,52	0,01	2,56	1,87	–0,62	2,83	0,68
Weryfikacja hipotez dotyczących: (a) zerowych zwrotów, (b) skośności i (c) kurtozy							
(a): statystyka (9.14)	1,2705	0,5877	0,5921	0,1657	0,3715	–0,7972	0,2425
(b): statystyka (3.53)	–0,8863	–1,1602	–1,7598	1,1820	–0,9971	1,5168	0,1879
(c): statystyka (3.56)	2,1201	–0,5848	2,3575	1,6463	–1,4096	2,5798	–1,0199

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05.

Min oznacza wartość minimalną, Max – wartość maksymalną, $\bar{y}$ – średnią arytmetyczną, S – odchylenie standardowe, V – współczynnik zmienności, Q – odchylenie ćwiartkowe, Q1 – kwartyl I, Q3 – kwartyl III, a reszta oznaczeń jak w tab. 9.1.

Źródło: opracowanie własne.

Omawiając własności stóp zwrotu wyznaczonych dla różnych interwałów, badaniu poddano również miary asymetrii i koncentracji, weryfikując hipotezy o zerowych wartościach miernika skośności (3.51) i kurtozy (3.54). W przypadku dziennych zwrotów w szeregu wszystkich obserwacji, a także wtorkowych i środowych stóp zwrotu obserwuje się istotną asymetrię dodatnią, również kurtoza jest istotnie dodatnia dla wszystkich analizowanych w tab. 9.3 stóp zwrotu. W przypadku zwrotów tygodniowych sytuacja jest podobna, tj. nie ma podstaw do odrzucenia hipotezy zerowej jedynie w przypadku poniedziałkowych i środowych stóp zwrotu (tab. 9.5).

Warto przy tym zauważyć, że analizując miesięczne stopy zwrotu (tab. 9.7), należy stwierdzić, że niemal wszystkie statystyki testowe nie upoważniają do odrzucenia hipotezy o symetrii rozkładu szeregów stóp zwrotu (z wyjątkiem stóp lutowych i grudniowych), również w pięciu (na 13) szeregach kurtoza nie jest istotnie różna od 0. To spostrzeżenie potwierdza znany z literatury fakt, że wydłużanie przedziału, z jakiego wyznaczane są stopy zwrotu, powoduje, że ich rozkłady stają się bardziej zbliżone do rozkładu normalnego (por. [Tarczyński i in. 2013, s. 79–95]).

Podsumowując analizę miesięcznych zwrotów badanej spółki, zauważa się, że niemal dla wszystkich miesięcy są one dodatnie. Jedynie w trzech miesiącach odnotowuje się przeciętne zwroty ujemne (w marcu mediana, w maju mediana i średnie oraz w listopadzie średnie) i to niezależnie od sposobu wyznaczania miesięcznych stóp zwrotu (tab. 9.6 i 9.7).

Tabela 9.6b. Miary opisowe dla procentowych jednookresowych miesięcznych logarytmicznych stóp zwrotu

Podstawowe miary	Lipiec	Sierpień	Wrzesień	Październik	Listopad	Grudzień
Liczba obserwacji	8	8	8	8	8	8
Min (%)	–12,53	–7,33	–5,65	–9,84	–17,91	–4,10
Max (%)	14,89	13,66	10,96	7,47	7,96	13,66
$\bar{y}$ (%)	3,23	1,05	1,16	1,09	–1,35	1,77
Me (%)	3,81	1,08	0,59	2,26	0,75	0,32
Q1 (%)	–1,13	–3,12	–3,16	–1,18	–2,99	–0,70
Q3 (%)	9,73	2,93	3,66	4,86	3,02	2,93
R (%)	27,42	21,00	16,60	17,31	25,87	17,76
S (%)	8,88	6,36	5,49	5,63	8,14	5,49
V	2,75	6,09	4,72	5,15	6,02	3,10
Q (%)	5,43	3,02	3,41	3,02	3,00	1,81
A	–0,57	0,93	0,70	–1,05	–1,33	1,60
K	–0,09	1,64	–0,08	0,93	1,83	3,27
Weryfikacja hipotez dotyczących: (a) zerowych zwrotów, (b) skośności i (c) kurtozy						
(a): statystyka (9.14)	1,0286	0,4646	0,5996	0,5497	–0,4701	0,9131
(b): statystyka (3.53)	–0,6545	1,0792	0,8032	–1,2119	–1,5323	**1,8430**
(c): statystyka (3.56)	–0,1002	**1,8970**	–0,0963	1,0734	**2,1151**	**3,7808**

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05.

Min oznacza wartość minimalną, Max – wartość maksymalną, $\bar{y}$ – średnią arytmetyczną, S – odchylenie standardowe, V – współczynnik zmienności, Q – odchylenie ćwiartkowe, Q1 – kwartyl I, Q3 – kwartyl III, a reszta oznaczeń jak w tab. 9.1.

Źródło: opracowanie własne.

Tabela 9.7a. Miary opisowe dla procentowych średnich miesięcznych logarytmicznych stóp zwrotu

Podstawowe miary		Miesiąc						
		Wszystkie	I	II	III	IV	V	VI
Liczba obserwacji		98	9	9	8	8	8	8
Min	(%)	–0,85	–0,41	–0,58	–0,33	–0,46	–0,43	–0,21
Max	(%)	0,65	0,39	0,37	0,49	0,34	0,45	0,27
$\bar{y}$	(%)	0,04	0,05	0,05	0,01	0,04	–0,08	0,01
Me	(%)	0,06	0,09	0,09	–0,03	0,15	–0,11	0,03
Q1	(%)	–0,14	0,06	0,02	–0,12	–0,10	–0,17	–0,03
Q3	(%)	0,20	0,19	0,25	0,11	0,23	–0,04	0,05
R	(%)	1,50	0,80	0,94	0,81	0,80	0,88	0,48
S	(%)	0,27	0,27	0,28	0,24	0,30	0,26	0,15
V	(%)	0,17	0,06	0,11	0,11	0,17	0,07	0,04
Q		6,45	5,64	5,19	24,74	7,57	–3,35	16,03
A		–0,34	–0,97	–1,44	0,87	–0,91	1,16	0,04
K		0,58	0,08	2,60	1,62	–0,55	2,65	0,54

Źródło: opracowanie własne.

Tabela 9.7b. Miary opisowe dla procentowych średnich miesięcznych logarytmicznych stóp zwrotu

Podstawowe miary		Miesiąc					
		VII	VIII	IX	X	XI	XII
Liczba obserwacji		8	8	8	8	8	8
Min	(%)	–0,54	–0,32	–0,27	–0,15	–0,85	–0,21
Max	(%)	0,65	0,59	0,52	0,59	0,38	0,35
$\bar{y}$	(%)	0,15	0,05	0,05	0,18	–0,06	0,05
Me	(%)	0,18	0,05	0,01	0,16	0,03	0,03
Q1	(%)	–0,05	–0,14	–0,14	0,06	–0,14	–0,08
Q3	(%)	0,45	0,13	0,18	0,29	0,15	0,16
R	(%)	1,19	0,91	0,79	0,74	1,23	0,56
S	(%)	0,39	0,28	0,26	0,23	0,38	0,21
V	(%)	0,25	0,13	0,16	0,11	0,14	0,12
Q		2,59	5,94	4,93	1,26	–6,14	4,16
A		–0,58	0,89	0,76	0,46	–1,36	0,35
K		–0,25	1,43	0,11	0,65	2,02	–0,93

Źródło: opracowanie własne.

Interesująca może być również analiza rocznych stóp zwrotu, liczonych jako jednokresowe stopy wyznaczone na początek następnego roku (tab. 9.8), z której wynika, że ujemne stopy zwrotu odnotowano z rocznych inwestycji poczynionych w latach 2014, 2015 i 2018. Natomiast rok 2013 był bardzo owocny dla inwestorów banku Santander.

Tabela 9.8. Miary opisowe wyznaczone dla procentowych rocznych jednookresowych logarytmicznych stóp zwrotu

Lata	Zwroty (%)	Podstawowe miary opisowe			
2010	0,24	Liczba obserwacji	9	R	77,85%
2011	7,15	Min	–27,79%	S	21,76%
2012	11,35	Max	50,05%	V	2,76
2013	50,05	$\bar{y}$	7,88%	Q	8,61%
2014	–0,52	$\bar{y}_{ch}$	9,43%	A	0,43
2015	–27,79	Me	7,15%	K	1,32
2016	16,70	Q1	–0,52%		
2017	22,94	Q3	16,70%		
2018	–9,20				

Uwaga: $\bar{y}$ – oznacza średnią chronologiczną, a reszta oznaczeń jak w tab. 9.3.

Źródło: opracowanie własne.

Tabela 9.9. Miary opisowe wyznaczone dla procentowych dziennych logarytmicznych stóp zwrotu w wyróżnionych okresach koniunktury spółki

Podstawowe miary		Hossa I	Stagnacja I	Bessa I	Hossa II	Bessa II	Stagnacja II
Liczba obserwacji		828	134	358	515	126	163
Min	(%)	–7,91	–3,64	–7,17	–5,73	–4,66	–4,88
Max	(%)	6,59	4,12	5,80	7,38	7,35	4,42
$\bar{y}$	(%)	0,09	0,00	–0,14	0,13	–0,18	0,08
Me	(%)	0,00	0,00	–0,01	0,00	0,00	0,05
Q1	(%)	–0,38	–1,20	–1,20	–1,07	–1,34	–0,94
Q3	(%)	0,45	1,04	0,82	1,21	0,83	0,97
R	(%)	14,50	7,76	12,97	13,11	12,01	9,30
S	(%)	1,35	1,55	1,92	1,92	1,99	1,78
V		14,43	–786,05	–13,32	14,86	–10,77	22,21
Q	(%)	0,42	1,12	1,01	1,14	1,09	0,96

Tabela 9.9 (cd.)

Podstawowe miary	Hossa I	Stagnacja I	Bessa I	Hossa II	Bessa II	Stagnacja II
A	0,02	0,07	–0,16	0,41	0,30	–0,04
K	5,39	–0,21	1,69	1,14	1,34	0,36
Weryfikacja hipotez						
Statystyka (9.14)	**1,9945**	-0,0147	–1,4206	1,5267	–1,0426	0,5748
Statystyka (9.24)		0,6731	0,8469	**–2,0690**	1,5993	–1,1745
Statystyka (3.44)		**1,3195**	**1,5379**	1,0029	1,0695	1,2396
Wartości krytyczne F-S		1,2560	1,2580	1,1725	1,2741	1,3165

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05. Zacienienie oznacza sytuację, kiedy wariancja w późniejszym okresie jest mniejsza. Kolejność okresów ustawiono w porządku chronologicznym, tak jak w tab. 9.2.

Źródło: opracowanie własne.

Przedstawione analizy zostały uzupełnione badaniem dziennych stóp zwrotu w wyróżnionych w tab. 9.2 okresach koniunktury zaobserwowanych dla badanej spółki, co przedstawiono w tab. 9.9. Naturalnym jest sprawdzenie, czy wyróżnione podokresy zostały poprawnie zdefiniowane. Zakłada się bowiem, że w okresach koniunktury oczekiwane zwroty są dodatnie, a w okresach dekoniunktury – ujemne, co można stwierdzić na podstawie przeprowadzonej weryfikacji hipotez (9.12)–(9.14). Drugim testem, który pozwoli stwierdzić, czy punkty przegięcia funkcji trendu zostały dobrze określone, jest test (9.20)–(9.25), sprawdzający równość oczekiwanych stóp zwrotu w sąsiadujących ze sobą podokresach. Dodatkowo można również zweryfikować, czy ryzyko w sąsiednich okresach jest takie samo. Na podstawie przeprowadzonych badań (tab. 9.9) stwierdzono, że w żadnym z analizowanych podokresów oczekiwane zwroty nie różniły się istotnie od 0 na poziomie istotności 0,05. Wyjątkiem jest pierwszy z podokresów – hossa I, kiedy to oczekiwane zwroty są istotnie większe od 0. Nie stwierdzono również istotnych różnic w oczekiwanych stopach zwrotu w sąsiadujących okresach z wyjątkiem porównywanych okresów hossy II i bessy II, kiedy to na poziomie istotności 0,05 odrzuca się hipotezę zerową o równości stóp zwrotu w tych okresach. Z kolei porównanie ryzyka w sąsiadujących ze sobą podokresach prowadzi do wniosku, że w okresie stagnacji I ryzyko było większe niż w czasie hossy I, ale mniejsze niż w czasie bessy I. Przeprowadzono również analizy dotyczące porównania sytuacji spółki w okresach przed i po wyborach prezydenckich w 2015 roku. Co widać w tab. 9.10, w obu podokresach nie było podstaw do odrzucenia hipotezy zerowej o zerowych zwrotach i o istotnych różnicach w uzyskiwanych zwrotach z akcji spółki Santander. Natomiast po wyborach prezydenckich istotnie wzrosło ryzyko.

Tabela 9.10. Porównanie stóp zwrotu w okresach przed wyborami prezydenckimi i po nich

Podstawowe miary		Okres przed wyborami	Okres po wyborach
Liczba obserwacji		1135	993
Wartość minimalna	(%)	–7,91	–6,94
Wartość maksymalna	(%)	6,59	7,38
Średnia arytmetyczna	(%)	0,06	0,01
Odchylenie standardowe	(%)	1,41	1,97
Współczynnik zmienności		22,75	253,30
Statystyka (9.14)		1,48	0,12
Statystyka (9.24)		0,72	×
Statystyka (3.44)		**1,96**	×

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05.

Źródło: opracowanie własne.

9.4.3. Modele wyceny aktywów kapitałowych

Modele wyceny aktywów kapitałowych pełnią istotną rolę zarówno w teorii, jak i praktyce. W omawianych badaniach oszacowano dla spółki Santander Bank Polska S.A. modele Sharpe'a i CAPM dla całego okresu badania i w wyróżnionych podokresach zgodnych z sytuacją rynkową, tzn. ustalonych na podstawie przebiegu szeregu czasowego indeksu giełdowego WIG. Innymi słowy dokonano podziału całego rozpatrywanego okresu na 8 podokresów, charakteryzujących się zróżnicowaną koniunkturą na polskim rynku kapitałowym, co opisano w tab. 9.11. Dodatkowo oszacowano modele dla dwóch wybranych okresów zwyżki i zniżki cen akcji banku Santander (wyróżnionych w tab. 9.2). W modelach przyjęto indeks giełdowy WIG za indeks rynku oraz indeks obligacji TBSP za instrument wolny od ryzyka.

Warto odnotować, że przeprowadzone badania dotyczące stóp zwrotu spółki Santander w różnych okresach koniunktury polskiego rynku kapitałowego wprawdzie wskazały na dodatnie średnie zwroty w okresach trzech hoss i stagnacji oraz ujemne w obu wyróżnionych na rynku okresach bess, jednakże weryfikacja odpowiednich hipotez statystycznych pozwoliła na odrzucenie hipotez o zerowych zwrotach jedynie w okresie oznaczonym jako hossa 2, a spółka odnotowała wyższe stopy zwrotu niż indeks WIG jedynie w okresie hossy 1. Natomiast w porównaniach zwrotów uzyskanych w sąsiednich okresach istotną zmianę odnotowano jedynie między bessą 1 i hossą 3.

Tabela 9.11. Podokresy badawcze

Okresy	Nr okresu	Oznaczenie okresu	Data	
			rozpoczęcia	zakończenia
według przebiegu WIG	1	hossa 1	31.12.2010	5.09.2011
	2	stagnacja 1	8.09.2011	5.06.2012
	3	hossa 2	6.06.2012	31.10.2013
	4	stagnacja 2	4.11.2013	30.04.2015
	5	bessa 1	4.05.2015	15.01.2016
	6	hossa 3	18.01.2016	23.01.2018
	7	bessa 2	24.01.2018	29.06.2018
	8	stagnacja 3	30.06.2018	28.02.2019
według sytuacji spółki	A	hossa II	27.01.2016	17.01.2018
	B	bessa II	18.01.2018	13.07.2018

Źródło: opracowanie własne.

Tabela 9.12. Oszacowania modeli Sharpe'a w całym okresie badawczym i w wyróżnionych podokresach

	beta	alfa	beta	alfa	beta	alfa	beta	alfa
Wartości	Cały okres		Hossa 1		Stagnacja 1		Hossa 2	
Parametru	0,9524	0,0003	0,1840	0,0006	0,3434	0,0004	0,7072	0,0007
t-Studenta	**29,7750**	0,8201	**2,4341**	1,0017	**6,5632**	0,5527	**9,4066**	1,0403
R^2	0,2942		0,0373		0,1669		0,1951	
Wartości	Stagnacja 2		Bessa 1		Hossa 3		Bessa 2	
Parametru	1,2575	–0,0001	1,4283	–0,0002	1,4481	0,0000	1,4970	0,0003
t-Studenta	**15,9162**	–0,1868	**12,5696**	–0,1606	**18,2249**	–0,0665	**9,3742**	0,2403
R^2	0,3950		0,4620		0,3875		0,4419	
Wartości	Stagnacja 3				Hossa A		Bessa B	
Parametru	1,3069	0,0002			1,4391	0,0001	1,4791	–0,0001
t-Studenta	**12,7685**	0,1867			**17,8169**	0,1359	**9,6236**	–0,0825
R^2	0,4866				0,3818		0,4256	

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05. Na żółto oznaczono wartości beta powyżej jedności, a kolorem zielonym i czerwonym odpowiednio najwyższą i najniższą wartość współczynnika determinacji.

Źródło: opracowanie własne.

Oceniając oszacowane modele jednowskaźnikowe i wyceny aktywów kapitałowych, należy zauważyć, że parametr beta w obu modelach jest zawsze istotnie większy od 0, ale w okresach dwóch pierwszych trendów zwyżkowych oraz w pierwszym okresie stagnacji, tj. w okresie od 31.12.2010 do 31.10.2013, oraz dla całego okresu analizy spółka ma charakter defensywny, podczas gdy w pozostałych podokresach jest spółką agresywną (o czym świadczy wartość parametru odpowiednio poniżej i powyżej jedności). Warto też odnotować, że w kolejnych podokresach wartości współczynników determinacji są, podobnie jak wartości ocen estymatorów bety, dość zróżnicowane – najsłabiej dopasowane do danych empirycznych są modele szacowane w okresie hossy 1, a najlepiej modele z ostatniego omawianego okresu wyznaczonego dla całego rynku kapitałowego, tj. stagnacji 3. Można zauważyć, że szacowanie modeli dla indywidualnie wyróżnionych w rozpatrywanej spółce podokresów (A i B) pozwala na osiągnięcie dość wysokiego stopnia ich dopasowania do danych empirycznych. Jak widać współczynniki korelacji Pearsona między stopami zwrotu z instrumentu i indeksu rynku znajdują się w przedziale od 0,1931 do 0,6976.

W przypadku modeli CAPM interpretacji podlega również wyraz wolny – tzw. alfa Jensena[19], który można wykorzystać do oceny efektywności informacyjnej badanego instrumentu (lub portfela). We wszystkich modelach wartości ocen estymatora tego parametru są nieistotnie różne od 0, chociaż w większości modeli oceny są dodatnie, co jest pozytywnym sygnałem dla inwestorów. Ujemne wartości odnotowano jedynie w okresach bessy 1 i stagnacji 2, wyznaczonych dla całego rynku kapitałowego oraz w czasie bessy B wyróżnionej dla spółki.

Z uwagi na istotną rolę parametru beta w praktyce istnieje potrzeba rozstrzygnięcia wielu istotnych kwestii[20]. Jedną z nich jest metoda szacowania modeli wyceny aktywów kapitałowych oraz stabilność bety w czasie. Wprawdzie najczęściej do estymacji modeli stosuje się MNK, ale w większości przypadków nie są spełnione założenia Gaussa-Markowa i proponuje się estymację innymi metodami, miedzy innymi korzystając z modeli ARCH (6.20)–(6.21). Innym problemem jest niestabilność parametru beta, co sprawdza się m.in. za pomocą testu (9.34)–(9.36), który pozwala zweryfikować hipotezę o równości parametru beta szacowanego w różnych okresach. Przykładowo dla obu typów modeli wyznaczono wartości statystyki (9.36) dla sąsiadujących okresów oraz porównano największą i najmniejszą wartość parametru beta, którą zaobserwowano odpowiednio w ostatnim i pierwszym podokresie. Zastosowano test dwustronny, tj. obliczono wartości statystyki testowej dla błędów standardowych z obu okresów. Co widać w tab. 9.14, w większości przypadków należy odrzucić hipotezy zerowe odnośnie do niezmienności wartości parametru beta w czasie. Ostatecznie można przyjąć, że w początkowym

19 Alfa Jensena jest jednym z mierników oceny efektywności inwestycyjnej i jest wykorzystywana do badania efektywności informacyjnej w formie silnej.

20 Szerokie omówienie tych zagadnień znaleźć można w [Tarczyński, Witkowska, Kompa 2013].

okresie, tj. od 31.12.2010 do 30.04.2015, zmiany wartości parametru z okresu na okres są istotne, a w końcowym okresie, tj. od 4.05.2015 do 28.02.2019, nie ma podstaw do odrzucenia hipotezy zerowej o równości parametru beta.

Tabela 9.13. Oszacowania modeli CAPM w całym okresie badawczym i w wyróżnionych podokresach

	beta	alfa	beta	alfa	beta	alfa	beta	alfa
Wartości	Cały okres		Hossa 1		Stagnacja 1		Hossa 2	
Parametru	0,9557	0,0002	0,1916	0,0004	0,3419	0,0003	0,7503	0,0006
t-Studenta	**29,7991**	0,7916	**2,5215**	0,6860	**6,4395**	0,3414	**10,2383**	0,8670
R^2	0,2945		0,0399		0,1617		0,2231	
Wartości	Stagnacja 2		Bessa 1		Hossa 3		Bessa 2	
Parametru	1,2504	–0,0001	1,4899	–0,0001	1,4443	0,0000	1,4584	0,0003
t-Studenta	**15,6241**	–0,0874	**12,5189**	–0,0577	**18,0810**	0,0013	**9,3327**	0,2392
R^2	0,3862		0,4600		0,3837		0,4397	
Wartości	Stagnacja 3				Hossa A		Bessa B	
Parametru	1,2891	0,0002			1,4386	0,0001	1,4476	–0,0001
t-Studenta	**12,7067**	0,2398			**17,8251**	0,2038	**9,5684**	–0,0669
R^2	0,4842				0,3820		0,4228	

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05. Na żółto oznaczono wartości beta powyżej jedności, a kolorem zielonym i czerwonym odpowiednio najwyższą i najniższą wartość współczynnika determinacji.

Źródło: opracowanie własne.

Tabela 9.14. Wartości statystyk (9.36) obliczone dla porównywanych okresów

Okresy	1 : 8	1 : 2	2 : 3	3 : 4	4 : 5	5 : 6	6 : 7	7 : 8
Model Sharpe'a	**–14,8532**	**–2,1083**	**–6,9516**	**–7,3204**	**–2,1622**	–0,1741	–0,6148	1,1902
	–10,9658	**–3,0462**	**–4,8385**	**–6,9656**	–1,5033	–0,2489	–0,3059	1,8570
Model CAPM	–14,4408	**–1,9782**	**–7,6935**	**–6,8244**	**–2,9927**	0,3833	–0,1770	1,0835
	–10,8128	**–2,8310**	**–5,5734**	**–6,2490**	**–2,0125**	0,5710	–0,0905	1,6690

Uwaga: pogrubieniem wyróżniono wartości statystyk testowych upoważniających do odrzucenia hipotezy zerowej na poziomie istotności 0,05.

Źródło: opracowanie własne.

Rozdział 10

Analiza obiektów wielowymiarowych

Każde przedsiębiorstwo może być rozpatrywane jako obiekt w wielowymiarowej przestrzeni cech, branych pod uwagę przy jego charakterystyce i ocenie. W szczególności podejście wielowymiarowe ma zastosowanie w analizie fundamentalnej, kiedy jednym z etapów jest ocena spółki na podstawie wielu jej parametrów lub analizy portfelowej, kiedy wybiera się spółki do kompozycji portfela inwestycyjnego. Innymi obszarami zastosowań mogą być np. prognozowanie bankructwa, czy badanie zdolności kredytowej klientów (firm, różnego typu instytucji, czy osób fizycznych), z których każdy charakteryzuje się wielością cech jednocześnie branych pod uwagę. W pierwszym przypadku dokonuje się grupowania obiektów do homogenicznych klas (skupisk), a w drugim przeprowadza się klasyfikację wg ustalonego *a priori* wzorca. W niniejszym rozdziale przedstawione zostaną zagadnienia związane z zastosowaniem metod analiz wielowymiarowych do rozwiązania wybranych problemów praktycznych[1].

10.1. Ocena zdolności kredytowej

Ocena zdolności kredytowej sprowadza się do stwierdzenia, czy podmiot zaciągający zobowiązanie będzie miał możliwość, w dysponowanym czasie i na określonych z góry warunkach, zobowiązanie to spłacić. Innymi słowy to szacowanie prawdopodobieństwa, że (potencjalny) kredytobiorca spłaci dług w określonym terminie, czyli przypisanie kredytobiorcy do odpowiedniej klasy klientów, o z góry zdefiniowanej charakterystyce. Można zatem ocenę zdolności kredytowej potraktować to jako zadanie klasyfikacji i zastosować do jego rozwiązania adekwatne

1 Wszystkie prezentowane w niniejszym rozdziale zadania rozwiązano za pomocą pakietu STATISTICA.

metody. W dalszej części przedstawione zostaną przykłady rozwiązania tego typu zadań bazujących na danych rzeczywistych[2].

Decyzje kredytowe opierają się zazwyczaj na przeprowadzonej przez inspektorów kredytowych analizie zdolności kredytowej potencjalnych kredytobiorców. Analiza wniosków kredytowych może być również prowadzona na podstawie metod statystycznych. Należy jednak rozróżnić procedury stosowane w przypadku podmiotów gospodarczych oraz w przypadku klientów indywidualnych. W pierwszym przypadku ocena przedsiębiorstwa prowadzona jest najczęściej na podstawie analizy wskaźnikowej, a więc dotyczy cech ilościowych. Natomiast w drugim przypadku wiele charakterystyk klientów indywidualnych jest natury jakościowej, wymaga zatem aplikacji innych metod rozwiązania zadania klasyfikacji.

W przypadku decyzji kredytowych wśród błędów klasyfikacji należy wyróżnić ogólny błąd klasyfikacji postaci (8.46) oraz rozróżnić błędy klasyfikacji do określonej klasy (8.47). Przy przypadku klasyfikacji dychotomicznej utożsamia się je z błędami pierwszego i drugiego rodzaju. Błąd pierwszego rodzaju – E_1, pojawia się wtedy, gdy odrzucony przez bank klient zostanie zaklasyfikowany jako wiarygodny, co może prowadzić do strat na kapitale banku. Błąd drugiego rodzaju – E_2, powstaje wówczas, gdy klient posiadający zdolność kredytową zostanie rozpoznany jako niewiarygodny. W tej sytuacji instytucja kredytowa traci korzyści związane z udzieleniem kredytu. Innymi słowy ogólny błąd klasyfikacji jest postaci:

$$E = \frac{BN}{N} \cdot 100\% \tag{10.1}$$

podczas gdy błąd klasyfikacji pierwszego rodzaju wyraża się wzorem:

$$E_1 = \frac{BN_1}{N_1} \cdot 100\% \tag{10.2}$$

a drugiego rodzaju zapisuje się jako:

$$E_2 = \frac{BN_2}{N_2} \cdot 100\% \tag{10.3}$$

2 Omawiane badania zostały przeprowadzone na podstawie informacji udostępnionych przez banki. Warto jednak odnotować, że – zgodnie z istniejącymi przepisami i zwyczajami – banki udostępniają jedynie te dane, które uniemożliwiają identyfikację ich klientów oraz są neutralne dla instytucji kredytowych, tzn. nie można na ich podstawie wnioskować o realizowanym przez nie modelu biznesowym oraz o faktycznej efektywności prowadzonej polityki kredytowej. Zatem w wielu przypadkach uzyskane informacje dotyczą podjętych decyzji kredytowych, a nie tego, czy klienci spłacają bądź nie kredyt, co jednoznacznie określa poprawność podjętej decyzji.

gdzie:
N_1, N_2 – liczba klientów uznanych odpowiednio za wiarygodnych i niewiarygodnych, przy czym $N_1 + N_2 = N$,
BN_1, BN_2 – liczba klientów błędnie zaklasyfikowanych odpowiednio do klasy wiarygodnych i niewiarygodnych klientów, przy czym $BN_1 + BN_2 = BN$,
N – liczba wszystkich klientów ubiegających się o kredyt (tj. podlegających klasyfikacji),
BN – liczba wszystkich błędnie rozpoznanych klientów.

Warto przy tym zauważyć, że informacje udostępniane przez instytucje kredytowe charakteryzują się pewnymi specyficznymi cechami, które mają wpływ na jakość klasyfikacji.

1) Dane dotyczące kredytobiorców często mają charakter jakościowy. Zatem w celu wykorzystania ich w analizach ilościowych konieczne jest ich kodowanie, a sposób przekształcania cech opisowych w quasi-ilościowe może mieć wpływ na efektywność klasyfikacji [Gruszczyński 2002].
2) Zmienne wykorzystywane do klasyfikacji zawarte we wnioskach kredytowych lub dane w postaci wskaźników finansowych są silnie ze sobą skorelowane. Powoduje to występowanie wielu zmiennych powielających te same informacje, co negatywnie wpływa na jakość grupowania. Potrzebne jest więc przetwarzanie wstępne (preprocessing) danych wejściowych, które umożliwia dobór zmiennych diagnostycznych do modelu i redukcję ich liczby.
3) Liczebności zbiorów danych, które są wykorzystywane przy szacowaniu i testowaniu modeli jest z reguły niewielka (a ściślej mówiąc – jest niewystarczająca). Wynika to z konieczności posiadania danych jednorodnych[3] z punktu widzenia analizowanego produktu finansowego (np. leasing, kredyty inwestycyjne, konsumpcyjne, na zakup środków obrotowych itp.) oraz okresu (zmiana sytuacji gospodarczej ma istotny wpływ na politykę instytucji finansowych, zmieniają się nie tylko wysokości stóp procentowych, ale również zasady przyznawania kredytów). Należy przypomnieć, że liczba szacowanych parametrów modelu musi być znacząco mniejsza od liczby obserwacji, zatem redukcja wymiarów zadania przeprowadzana w procesie doboru zmiennych diagnostycznych jest niezwykle istotna.
4) Rzadko zdarza się, nawet przy klasyfikacji dychotomicznej, aby zbiory obiektów należących do różnych klas były równoliczne (albo o przybliżonej liczebności). Jest to związane z faktem, że instytucje kredytowe większą wagę przywiązują do klientów, których uznają za wiarygodnych niż do klientów należących do

3 Danych dotyczących różnych form kredytowania, okresów lub nawet różnych instytucji finansowych nie można łączyć.

grup podwyższonego i wysokiego ryzyka[4]. Konsekwencją tego jest znacznie większa reprezentacja klientów, którym udzielono kredyt, niż klientów, którym odmówiono przyznania kredytu. Brak „symetrii" obiektów należących do różnych klas powoduje, że wykorzystywane metody ilościowe łatwiej rozpoznają klientów pochodzących z większych liczebnie grup (por. np. [Mazur, Staniec, Witkowska 1999, s. 167–172; Witkowska 2002, s. 173–176]).

5) Dane udostępniane przez instytucje kredytowe nie są kompletne w tym sensie, że inspektorzy kredytowi posiadają znacznie bogatszą bazę danych, na podstawie której podejmują swoje decyzje, niż zbiór informacji wykorzystywanych w procesie klasyfikacji. Jest to związane z ochroną danych o klientach (faktyczną lub formalną), uniemożliwiającą przekazanie informacji, np. o charakterze prowadzonej działalności, historii współpracy klienta z bankiem, rodzajach posiadanych zabezpieczeń etc.
6) Ocena wyników klasyfikacji jest przeprowadzana w stosunku do „wzorca", którym najczęściej są decyzje inspektorów kredytowych, a nie stan faktyczny, jakim w omawianych przypadkach jest fakt spłaty (lub nie) kredytu wraz z odsetkami w terminie. Niestety instytucje finansowe bardzo niechętnie ujawniają informacje o własnej efektywności, a wiadomo jest, że nie wszystkie kredyty są spłacane w terminie i istnieje pewna liczba kredytów, które w ogóle nie zostają spłacone lub zostają spłacone tylko częściowo. Odsetek niespłacanych kredytów jest *de facto* błędem klasyfikacji pierwszego rodzaju.

Specyfika danych wykorzystywanych w ocenie zdolności kredytowej sprawia, że przy zastosowaniu metod ilościowych do klasyfikacji należy rozwiązać kilka problemów natury praktycznej, a w szczególności:

1) dobór zmiennych diagnostycznych,
2) sposób kodowania cech jakościowych,
3) dobór postaci funkcji progowych, pozwalających na zaklasyfikowanie konkretnego obiektu do danej grupy na podstawie uzyskanego wyniku,
4) wybór miernika oceny jakości metody klasyfikacji.

Warto dodać, że poprawne zastosowanie metod ilościowych w praktyce oceny zdolności kredytowej[5] jest podobne do tych stosowanych przy prognozowaniu na podstawie szeregów czasowych i opisanych w rozdziale 7. Procedura bowiem polega na:

1) budowie i estymacji modelu klasyfikacyjnego, zgodnie z opisanymi w podrozdziale 6.3 etapami,
2) weryfikacji poprawności modelu i jego zdolności dyskryminacyjnych,
3) zastosowaniu modelu w praktyce oceny obiektów podlegających klasyfikacji lub powrocie do etapu (1).

4 Przechowuje się informacje o wszystkich klientach, którym udzielono kredytu. Natomiast tylko nieliczne informacje o klientach, których wnioski kredytowe zostały odrzucone.

5 W podobny sposób postępuje się przy prognozowaniu bankructwa.

Estymacja modelu klasyfikacyjnego dokonywana jest na podstawie danych ze zbioru uczącego, a weryfikacja modelu opiera się na błędach klasyfikacji, wyznaczanych zarówno dla zbioru uczącego, jak i testowego. Ideę błędów obliczonych dla próby treningowej można porównać z oceną modelu ekonometrycznego przeprowadzoną na podstawie wartości reszt empirycznych dla próby estymacyjnej. Natomiast błędy obliczone dla próby testowej są koncepcyjnie podobne do błędów prognoz *ex post* wyznaczanych dla prognoz wygasłych. Decyzję o aplikacji modelu podejmuje się przede wszystkim na podstawie oceny jego poprawności w zbiorze testowym.

10.1.1. Analiza dyskryminacyjna

W omawianych dalej badaniach wykorzystano dane pochodzące z wniosków kredytowych 75 przedsiębiorstw ubiegających się o przyznanie kredytu (55 z nich kredytu udzielono). Wykorzystując dane z wniosków kredytowych, zdefiniowano zbiór 9 potencjalnych zmiennych diagnostycznych, opisujących każde z analizowanych przedsiębiorstw:

a) zysk brutto za okres sprawozdawczy,
b) zysk netto za okres sprawozdawczy,
c) średni miesięczny zysk netto,
d) miesięczna rata spłaty wnioskowanego kredytu,
e) wielkość maksymalnych miesięcznych odsetek od wnioskowanego kredytu,
f) inne obciążenia miesięczne (np. spłata innych zobowiązań),
g) pozostały średni zysk miesięczny po odliczeniu rat i odsetek,
h) ocena sytuacji finansowo-ekonomicznej firmy (od 0 do 4),
i) ocena zdolności kredytowej (od 0 do 3).

Zmienną grupującą w tym badaniu jest dychotomiczna decyzja banku dotycząca przyznania lub odmowy przyznania kredytu, przyjmująca wartości binarne.

Podstawowym problemem przy konstrukcji funkcji dyskryminacji jest dobór cech diagnostycznych. Korzystając z danych o przedsiębiorstwach ubiegających się o finansowanie, przeprowadzono analizę współzależności między potencjalnymi czynnikami, opisującymi analizowane firmy, a decyzją banku. Stwierdzono występowanie istotnej zależności dla cech: (b), (c), (g), (h) oraz (i). Przeprowadzono również badanie korelacji między poszczególnymi cechami diagnostycznymi (a)–(i), co przedstawiono w tab. 10.1. Jak widać, niemal wszystkie zmienne są ze sobą silnie skorelowane, co oznacza powielanie tych samych informacji. Jedynie zmienne (d) i (e) wykazują brak istotnej korelacji ze zmiennymi (f), (g), (h) oraz (i).

Tabela 10.1. Współczynniki korelacji liniowej Pearsona (4.5) wyznaczone dla cech diagnostycznych

	a	b	c	d	e	f	g	h	i
a	1								
b	**0,974**	1							
c	**0,904**	**0,913**	1						
d	**0,502**	**0,505**	**0,457**	1					
e	0,339	**0,372**	**0,392**	**0,480**	1				
f	**0,671**	**0,769**	**0,752**	0,188	0,246	1			
g	**0,782**	**0,740**	**0,888**	0,152	0,081	**0,524**	1		
h	**0,614**	**0,630**	**0,638**	0,227	0,059	**0,446**	**0,645**	1	
i	**0,663**	**0,677**	**0,709**	0,190	0,133	**0,481**	**0,730**	**0,850**	1

Uwaga: pogrubieniem wyróżniono statystycznie istotną korelację zweryfikowaną testem (4.13)–(4.17) na poziomie istotności 0,05.

Źródło: [Witkowska 1999b; 2002, s. 132].

W pierwszym podejściu spośród analizowanych cech wybrano jako zmienne diagnostyczne jednie te, które są słabo skorelowane między sobą i istotnie powiązane ze zmienną grupującą, reprezentującą decyzję banku. Należy zatem zauważyć, że istotną zależność ze zmienną grupująca wykazują zmienne: (b), (c), (g), (h) oraz (i). Zysk netto za okres sprawozdawczy (b) jest silnie skorelowany z pozostałymi zmiennymi diagnostycznymi, w szczególności ze zmienną (c) – wartość współczynnika korelacji 0,913 oznacza, że obie zmienne niosą podobne informacje. Dodatkowo poziom skorelowania z decyzją banku zmiennej (c) jest nieco wyższy niż zmiennej (b). Dlatego ze zbioru zmiennych diagnostycznych usunięto tę ostatnią. Przy czym zestaw zmiennych dyskryminacyjnych wzbogacono o te zmienne, które są istotne dla podejmujących decyzje o przyznaniu kredytu, tj. o zmienne (d) i (e). Stąd nowy zbiór zawierał sześć cech diagnostycznych: (c), (d), (e), (g), (h) oraz (i). Ten zestaw oznaczono cech jako A.

W drugim podejściu wykorzystano opisaną w podrozdziale 8.5 metodę Hellwiga grupowania cech, arbitralnie przyjmując krytyczną wartość współczynnika korelacji $r^* = 0{,}85$. Zgodnie z przedstawionym algorytmem postępowania najpierw zsumowano elementy poszczególnych kolumn w tab. 10.2 według wzoru (8.53) i stwierdzono, że cechą centralną jest zmienna (c), czyli średni miesięczny zysk netto. Następnie na podstawie (8.54) zidentyfikowano cechy satelitarne, którymi są cechy (a) i (b), reprezentujące zysk brutto i netto ze sprzedaży. Utworzono zredukowaną macierz cech (tab. 10.3) i ponowne wyznaczono cechę centralną charakteryzującą się maksymalną sumą (8.53). Tym razem jest to zmienna (i), będąca oceną zdolności kredytowej firmy. Dalej, korzystając ze wzoru (8.54), ustalono, że cechą satelitarną jest zmienna (h), będąca oceną sytuacji finansowo-ekonomicznej przedsiębiorstwa. Usunięcie z macierzy tych dwóch cech pozwoliło na utworzenie kolejnej zredukowanej macierzy

cech diagnostycznych, zawierającej jedynie cztery zmienne (tab. 10.4), dla której cechą centralną okazała się zmienna (f), informująca o innych obciążeniach miesięcznych. Ta zmienna nie posiada już żadnych cech satelitarnych, gdyż wszystkie współczynniki korelacji są mniejsze od wartości krytycznej r*, co oznacza, że pozostałe zmienne są zmiennymi izolowanymi. Zatem postępowanie uznaje się za zakończone.

Tabela 10.2. Pierwotna macierz korelacji

Cechy	a	b	c	d	e	f	g	h	i	Rodzaj cechy
a	1,000	0,974	0,904	0,502	0,339	0,671	0,782	0,614	0,663	satelitarna
b	0,974	1,000	0,913	0,505	0,372	0,769	0,74	0,63	0,677	satelitarna
c	0,904	**0,913**	1,000	0,457	0,392	0,752	0,888	0,638	0,709	
d	0,502	0,505	0,457	1,000	0,480	0,188	0,152	0,227	0,19	
e	0,339	0,372	0,392	0,480	1,000	0,246	0,081	0,059	0,133	
f	0,671	0,769	0,752	0,188	0,246	1,000	0,524	0,446	0,481	
g	0,782	0,74	0,888	0,152	0,081	0,524	1,000	0,645	0,730	
h	0,614	0,630	0,638	0,227	0,059	0,446	0,645	1,000	0,850	
i	0,663	0,677	0,709	0,190	0,133	0,481	0,730	0,850	1,000	
Suma	6,449	6,580	6,653	3,701	3,102	5,077	5,542	5,109	5,433	
Cecha			centralna							

Uwaga: kolorem żółtym oznaczono kolumnę zawierającą cechę centralną, a szarym wiersze zawierające cechy satelitarne. Redukcja wymiaru macierzy polega na wykreśleniu cech zaznaczonych kolorami.

Źródło: opracowanie własne.

Tabela 10.3. Zredukowana macierz korelacji

Cechy	d	e	f	g	h	i	Rodzaj cechy
d	1,000	0,480	0,188	0,152	0,227	0,190	
e	0,480	1,000	0,246	0,081	0,059	0,133	
f	0,188	0,246	1,000	0,524	0,446	0,481	
g	0,152	0,081	0,524	1,000	0,645	0,730	
h	0,227	0,059	0,446	0,645	1,000	0,850	satelitarna
i	0,190	0,133	0,481	0,730	0,850	1,000	
Suma	2,237	1,999	2,885	3,132	3,227	3,384	
Cecha						centralna	

Uwaga: kolorem żółtym oznaczono kolumnę zawierającą cechę centralną, a szarym wiersz zawierający cechę satelitarną. Redukcja wymiaru macierzy polega na wykreśleniu cech zaznaczonych kolorami.

Źródło: opracowanie własne.

Tabela 10.4. Zredukowana macierz korelacji

Cechy	d	e	f	g
d	1,000	0,480	0,188	0,152
e	0,480	1,000	0,246	0,081
f	0,188	0,246	1,000	0,524
g	0,152	0,081	0,524	1,000
Suma	1,820	1,807	1,958	1,757
Cecha			centralna	

Uwaga: kolorem żółtym oznaczono kolumnę zawierającą cechę centralną.

Źródło: opracowanie własne.

Wśród analizowanych potencjalnych cech diagnostycznych wyróżniono trzy cechy centralne: średni miesięczny zysk netto (c), ocenę zdolności kredytowej (i) oraz inne obciążenia miesięczne (f), a także trzy cechy izolowane, którymi są: miesięczna rata spłaty wnioskowanego kredytu (d), wielkość maksymalnych odsetek (e) i pozostały średni zysk miesięczny po odliczeniu rat i odsetek (g). W konsekwencji, na podstawie analizy korelacji oraz wskazówek merytorycznych, utworzono osiem zestawów zmiennych diagnostycznych, zbudowanych w różny sposób. Zbiory zmiennych zawierały:

A. sześć cech diagnostycznych: (c), (d), (e), (g), (h) oraz (i), wyselekcjonowanych poprzez eliminację cech, dla których stwierdzono brak istotnej zależności ze zmienną grupującą (tj. z przyznaniem/nieprzyznaniem kredytu) lub silnie ze sobą skorelowanych,
B. tylko dwie cechy centralne, tj. (c) oraz (i), obie istotnie współzależne – powiązane ze zmienną grupującą,
C. wszystkie cechy centralne i izolowane, tj. (c), (d), (e), (f), (g) oraz (i),
D. zestaw A, w którym średni miesięczny zysk netto (c), zastąpiono zyskiem netto za okres sprawozdawczy (b), obie bowiem zmienne istotnie wpływają na zmienną grupującą z podobną siłą,
E. cechy z tab. 10.4, opisujące sytuację firmy, gdyby kredyt został jej przyznany, z których tylko cechy (d) i (e) są ze sobą istotnie skorelowane,
F. zestaw E uzupełniony oceną zdolności kredytowej (i), tj. zmienną statystycznie istotnie wpływającą na zmienną grupującą,
G. zestaw E uzupełniony oceną sytuacji finansowo-ekonomicznej (h), tj. zmienną silnie skorelowaną ze zmienną (i),
H. zestaw F uzupełniony oceną sytuacji finansowo-ekonomicznej firmy (h).

Warto w tym miejscu przypomnieć, że badanie jakości klasyfikacji powinno być dokonywane na podstawie zbioru testowego, zawierającego obiekty, które nie uczestniczyły w procesie wyznaczania parametrów funkcji dyskryminacji. W celu

podziału całego zbioru przedsiębiorstw na próbę treningową (uczącą) i testową (weryfikującą) poszczególne firmy ponumerowano za pomocą generatora liczb pseudolosowych, a cały zbiór podzielono na dwa podzbiory rozłączne[6] (tab. 10.5), które wykorzystano do oszacowania parametrów funkcji dyskryminacji.

Tabela 10.5. Struktura zbioru danych

Próba	Liczba wniosków kredytowych rozpatrzonych		Liczba obiektów w próbie
	pozytywnie	negatywnie	
Treningowa	32	14	46
Testowa	23	6	29
Razem	55	20	75

Źródło: opracowanie własne na podstawie [Witkowska 1997].

Parametry różnie wyspecyfikowanych funkcji dyskryminacji oszacowano na podstawie danych dotyczących przedsiębiorstw, które znalazły się w próbie treningowej (uczącej). Przykładowo, dla zestawu zmiennych diagnostycznych {c, d, e, g, h, i}, parametry funkcji dyskryminacji (7.40) wynoszą:

$$D(x) = -0{,}000078 \cdot x_c - 0{,}00062 \cdot x_d - 0{,}00011 \cdot x_e - 0{,}00010 \cdot x_g + \\ + 0{,}450257 \cdot x_h + 1{,}435595 \cdot x_i \qquad (10.4)$$

Wartości parametrów funkcji dyskryminacyjnej (10.4) są współczynnikami surowymi, zatem nie mają interpretacji. Mogą być jednak wykorzystane do wyznaczenia wartości funkcji dla średnich wartości zmiennych dyskryminacyjnych w każdej grupie przedsiębiorstw. Dla funkcji dyskryminacyjnej (10.4) wartości te wynoszą:

- dla grupy przedsiębiorstw, którym nie przyznano kredytu: –0,0925
- dla grupy przedsiębiorstw, którym kredyt przyznano: 3,3733.

Punktem leżącym w połowie odległości między tymi wartościami (7.42) jest $\bar{y} = 1{,}6404$. Wyznaczona reguła klasyfikacyjna przyjmuje zatem postać:

obiekty, dla których zachodzi:

- $Z < (\bar{y} = 1{,}6404)$ należy zaklasyfikować do grupy klientów niewiarygodnych,
- $Z > (\bar{y} = 1{,}6404)$ należy zaklasyfikować do grupy klientów wiarygodnych.

Następnie oceniono poprawność modelu zarówno dla danych z próby treningowej, jak i testowej, wyznaczając ogólne błędy klasyfikacji (10.1) i błędy klasyfikacji pierwszego i drugiego rodzaju (10.2) i (10.3). Wymienione błędy klasyfikacji przedstawiono w tab. 10.6.

6 Należy przy tym zapewnić reprezentację przedsiębiorstw uznanych przez inspektorów kredytowych za wiarygodne i niewiarygodne w obu próbach, tj. uczącej i testowej.

Tabela 10.6. Błędy klasyfikacji przeprowadzonej za pomocą funkcji dyskryminacyjnej

Symbol zbioru	Zmienne diagnostyczne	Rodzaje błędów					
		E	E_1	E_2	E	E_1	E_2
		Próba treningowa			Próba testowa		
A	c, d, e, g, h, i	2,17	7,14	0,00	20,69	0,00	26,09
B	c, i	15,21	0,00	21,87	24,14	0,00	30,43
C	c, d, e, f, g, i	2,17	7,14	0,00	13,79	0,00	17,39
D	b, i	2,17	7,14	0,00	20,69	0,00	26,09
E	d, e, f, g	23,91	0,00	34,38	27,59	0,00	34,8
F	d, e, f, g, i	4,35	7,14	3,13	17,2	0,00	21,74
G	d, e, f, g, h	8,70	14,29	6,25	20,69	0,00	26,09
H	d, e, f, g, h, i	2,17	7,14	0,00	20,69	0,00	26,09

Źródło: opracowanie własne na podstawie [Witkowska 1999b; 2002, s. 144].

Oceniając uzyskane wyniki klasyfikacji, zauważa się, że błędy ogólne są większe w przypadku próby testowej niż uczącej. Jest to szczególnie widoczne w przypadku błędów drugiego rodzaju E_2, co po części wynika z zasadniczej asymetrii obu prób – próba weryfikacyjna stanowi niewiele ponad 60% próby uczącej[7]. Niewątpliwie jednak, z punktu widzenia potencjalnego użytkownika, dla którego najbardziej istotnym jest błąd klasyfikacji pierwszego rodzaju E_1, wszystkie prezentowane modele należy ocenić pozytywnie, ponieważ wszyscy niewiarygodni klienci zostali poprawnie rozpoznani w próbie testowej.

W tab. 10.6 widać również, że modele o niskich błędach klasyfikacji uzyskanych w próbie treningowej nie zawsze dają niższe błędy w próbie testowej. Wystarczy porównać modele zbudowane na podstawie zestawów zmiennych diagnostycznych: A, C, D i H, z których tylko A i H generują podobne wyniki w obu próbach (co potwierdzają informacje zawarte w tab. 10.7). Pozostałe dwa modele dają zupełnie różne wyniki. Co więcej, model skonstruowany na podstawie zbioru danych F, mimo dwa razy wyższych ogólnych błędów klasyfikacji generowanych w próbie uczącej, daje w próbie weryfikacyjnej lepsze wyniki niż model, w którym zmienne dyskryminacyjne oznaczono jako zestaw D.

Jak zatem można zauważyć, wybór zmiennych diagnostycznych ma kluczowe znaczenie dla jakości klasyfikacji. W prezentowanych analizach najlepsze wyniki uzyskano dla próby testowej w przypadku zestawu zmiennych diagnostycznych C, który zawierał cechy centralne i izolowane wyznaczone metodą Hellwiga. W tym przypadku błędnie zaklasyfikowano jedynie cztery wnioski kredytowe i to

7 Skutkiem tego przykładowo 4 błędnie zaklasyfikowane firmy w próbie testowej stanowią 13,8%, a w próbie treningowej 8,7% w przypadku ogólnego błędu klasyfikacji, zaś w przypadku błędu drugiego rodzaju – odpowiednio 66,7 oraz 28,6%.

wszystkie do grupy przedsiębiorstw niewiarygodnych. Często istotne informacje niesie analiza błędnie zaklasyfikowanych obiektów, gdyż w kolejnych modelach klasyfikacyjnych błędnie rozpoznawane są te same firmy (tab. 10.7, w której podano numery błędnie rozpoznanych firm).

Tabela 10.7. Wykaz błędnie zaklasyfikowanych przedsiębiorstw

Symbol zbioru	Listy przedsiębiorstw klasyfikowanych wg rodzajów błędów		
	E_2	E_1	E_2
	Próba treningowa		Próba testowa
A		10	53, 57, 68, 70, 72, 74
B	4, 8, 12, 16, 18, 22, 31		49, 53, 57, 68, 70, 72, 74
C		10	49, 57, 70, 72
D		10	49, 53, 57, 70,72, 74
E	4, 8, 12, 16, 18, 19, 22, 31, 34, 44, 46		49, 55, 57, 62, 68, 70, 72, 74
F	46	10	49, 57, 68, 70, 72
G	4, 26	10, 32	53, 66, 68, 70, 72, 74
H		10	53, 57, 68, 70, 72, 74

Źródło: opracowanie własne na podstawie [Witkowska 1999b; 2002, s. 144].

Oceniając wyniki klasyfikacji, można dostrzec, że błędy klasyfikacji często dotyczą tych samych obiektów badania, które zostają niepoprawnie rozpoznane przez model. Przykładowo wymienione w tab. 10.7 przedsiębiorstwa zostały w większości przypadków błędnie zaklasyfikowane przynajmniej dwukrotnie. Jest to szczególnie widoczne w próbie testowej, w której jedynie trzy firmy (55, 62 i 66) zostały błędnie rozpoznane tylko w jednej procedurze klasyfikacyjnej. Natomiast przedsiębiorstwa o numerach 70 i 72 zostały niepoprawnie zaklasyfikowane jako niewiarygodni klienci przez wszystkie 8 modeli, 57 – przez 7 modeli, 68 i 74 – przez 6 modeli oraz 49 i 53 przez 5 modeli. Z kolei firma oznaczona jako 10, należąca do próby uczącej, została błędnie rozpoznana jako wiarygodny klient przez sześć modeli dyskryminacyjnych.

Może to oznaczać, że te wielokrotnie błędnie rozpoznawane przedsiębiorstwa są nietypowe. Dla wzmocnienia tego stwierdzenia porównano wyniki klasyfikacji przeprowadzonej za pomocą liniowej funkcji dyskryminacji z wynikami dla tych samych przedsiębiorstw uzyskanymi przy wykorzystaniu procedury nieliniowej (tab. 10.8), tj. z wykorzystaniem sztucznych sieci neuronowych trenowanych metodą wstecznej propagacji błędów[8]. Jak można zauważyć, aplikacja sztucznych

8 Tematyce sztucznych sieci neuronowych i ich zastosowaniach w finansach poświęcono zbiorcze prace [Witkowska 1999a; 2002].

sieci neuronowych dała znakomicie lepsze efekty klasyfikacji, bowiem wszystkie błędy klasyfikacji przedsiębiorstw z próby testowej (tab. 10.8) są mniejsze od tych z tab. 10.6. Co więcej, w próbie uczącej połowa sieci dokonała bezbłędnej klasyfikacji, a w pozostałych jedyną błędnie rozpoznaną firmą była ta oznaczona numerem 10. W przypadku sieci neuronowych najlepsze wyniki klasyfikacji osiągnięto dla zestawu zmiennych diagnostycznych C i F, co potwierdza wcześniejszy wniosek o kluczowej roli odpowiedniego doboru zmiennych w procedurach klasyfikacyjnych. Sztuczne sieci neuronowe błędnie rozpoznały firmy oznaczone numerami 57 – 8 razy, 10 – 4 razy, 59 i 72 po 3 razy, a także pojedynczo obiekty 49, 53, 71 i 74. Zatem za pomocą obu metod najczęściej błędnie klasyfikowane były przedsiębiorstwa 57 (15 razy na 16 klasyfikacji), 10 i 72 (po 10 razy), 74 (7 razy), 49 (6 razy) i 53 (4 razy).

Tabela 10.8. Błędy klasyfikacji przeprowadzonej za pomocą sztucznych sieci neuronowych (z wykorzystaniem algorytmu wstecznej propagacji błędów)

Nr	Zmienne diagnostyczne	Błędy klasyfikacji					
		w próbie treningowej			w próbie testowej		
		E	E_1	E_2	E	E_1	E_2
A	c, d, e, g, h, i	0,00	0,00	0,0	6,90	0,0	8,70 (57, 59)
B	c, i	2,17	7,14 (10)	0,0	6,90	0,0	8,70 (57, 72)
C	c, d, e, f, g, i	2,17	7,14 (10)	0,0	3,45	0,0	4,35 (57)
D	b, i	2,17	7,14 (10)	0,0	17,24	0,0	21,74 (49, 53, 57, 71, 72)
E	d, e, f, g	0,00	0,00	0,0	6,90	0,0	8,70 (57, 72)
F	d, e, f, g, i	2,17	7,14 (10)	0,0	3,45	0,0	4,35 (57)
G	d, e, f, g, h	0,00	0,00	0,0	10,34	0,0	13,04 (57, 59, 74)
H	d, e, f, g, h, i	0,00	0,00	0,0	6,90	0,0	8,70 (57, 59)

Uwaga: w nawiasach podano numery błędnie rozpoznanych przedsiębiorstw.

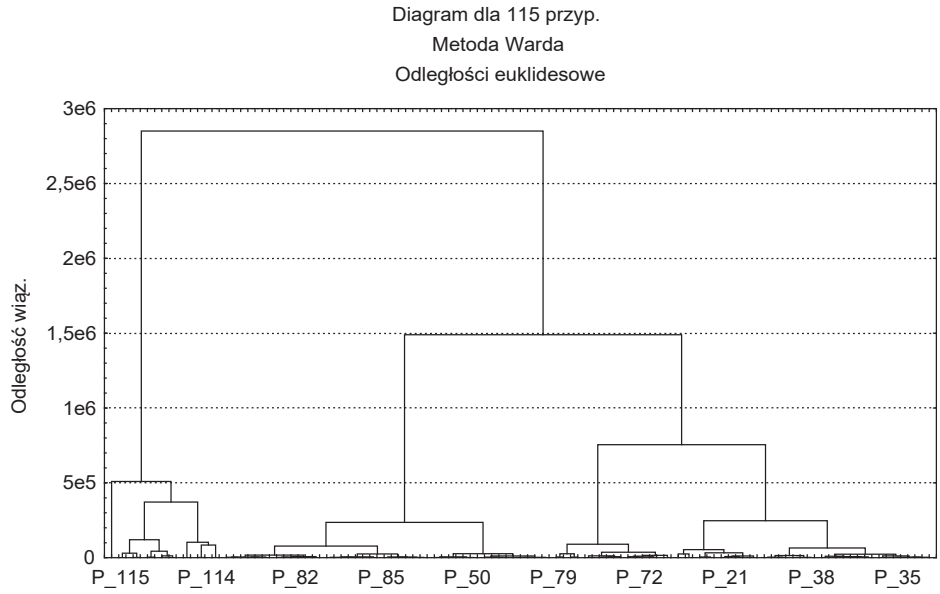
Źródło: opracowanie własne na podstawie [Witkowska 1999b].

10.1.2. Klasyfikacja bezwzorcowa

W badaniach[9] wykorzystano dane pochodzące z wniosków kredytowych 115 przedsiębiorstw ubiegających się o przyznanie kredytu w dwóch bankach łódzkich (75 pierwszych z jednego banku i 40 z drugiego). Na podstawie oceny przeprowadzonej przez inspektorów kredytowych, wszystkie rozpatrywane wnioski kredytowe podzielono na dwie klasy, tj. takie, na podstawie których udzielono

9 Opisanych w [Witkowska, Staniec 2000].

kredytu (89 firm), i takie, na podstawie których odmówiono przyznania kredytu (26 firm). Zebrane dane dotyczyły tych samych zmiennych diagnostycznych, opisanych w poprzednim podrozdziale z wyjątkiem cechy (f), która została pominięta w tym badaniu.



Rysunek 10.1. Wyniki klasyfikacji przy użyciu metody Warda

Źródło: [Witkowska, Staniec 2000].

W przypadku klasyfikacji bezwzorcowej następuje grupowanie do założonej liczby klas, które nie są jednak precyzyjnie zdefiniowane. Innymi słowy, w metodzie klasyfikacji bez nauczyciela pomija się istotną informację dotyczącą przyznania lub odmowy przyznania przedsiębiorstwu finansowania. Poszukuje się natomiast wspólnych cech firm zakładając, że oba zbiory klientów uznanych przez bank za wiarygodnych i niewiarygodnych zasadniczo różnią się między sobą, tworząc homogeniczne klasy. W tym przypadku jednoczesnemu grupowaniu podlegają wszystkie badane obiekty[10]. Przyjęto, że rozkład utworzonych grup odpowiada strukturze próby. W tym przypadku bardziej liczna grupa odpowiada przedsiębiorstwom, którym kredyt przyznano, a mniej liczna tym, którym

10 To znaczy bez podziału na zbiór treningowy i testowy, jak ma to miejsce w przypadku klasyfikacji ze wzorcem.

finansowania odmówiono. W analizach wykorzystano dwie najbardziej popularne metody, tj. metodę Warda i metodę *k*-średnich[11].

Wykorzystując metodę Warda, do klasyfikacji zastosowano odległość euklidesową jako miarę odległości pomiędzy skupieniami. Wyniki klasyfikacji przedstawiono na rysunku 10.1. W przypadku metody Warda nie zakłada się *a priori* liczby utworzonych klas, które odczytuje się z diagramu. Jak widać na rysunku, przedsiębiorstwa zostały pogrupowane w kilka skupień, które łączą ze sobą ramiona diagramu. Przy czym dla założonych dwóch klas klientów mniej liczna grupa znajduje się po lewej stronie diagramu, co utożsamiono z firmami, którym odmówiono kredytu. Natomiast znacznie bardziej liczne skupienie znajdujące się po prawej stronie diagramu potraktowano jako zbiór firm, którym kredyt przyznano.

W tab. 10.9 przedstawiono wyniki klasyfikacji, prezentując zbiór wszystkich analizowanych firm z podziałem na utworzone skupienia wraz z informacją o rzeczywistej przynależności tych obiektów do zbioru klientów wiarygodnych i niewiarygodnych. Innymi słowy, według opisu w główce tabeli jej kolumny zawierają numery przedsiębiorstw, które zostały zakwalifikowane przez metodę Warda do dwu klas:

1) pierwszej (czyli do skupienia po prawej stronie diagramu), zawierającej przedsiębiorstwa, którym przyznano kredyt,
2) drugiej (czyli do skupienia po lewej stronie diagramu), zawierającej przedsiębiorstwa, którym kredytu nie przyznano.

Natomiast wiersze tabeli, zgodnie z opisem w jej boczku, zawierają stan faktyczny, czyli numery firm, którym kredyt przyznano lub nie.

Jak widać w tab. 10.9, klasyfikując metodą Warda otrzymano w pierwszej grupie 95 obiektów, a w drugiej 20. Można zauważyć, że wszystkie przedsiębiorstwa, którym odmówiono przyznania kredytu, nie zostały rozpoznane, stąd błąd pierwszego rodzaju wynosi 100%. Natomiast błąd drugiego rodzaju polegający na tym, że wiarygodnych klientów zaklasyfikowano jako niewiarygodnych, wynosi 22,5%. Zatem ogólny błąd klasyfikacji wynosi 40%.

Tabela 10.9. Przedsiębiorstwa zakwalifikowane do obu skupień otrzymanych metodą Warda

Stan faktyczny	Numery firm, które znalazły się w klasie pierwszej, a którym	
	przyznano kredyt	nie przyznano kredytu
Przyznanie kredytu	2–5, 7–8, 11–16, 18, 19, 22, 25, 27, 31, 33–36, 8, 42, 44, 48, 49, 51, 53, 55, 57, 59–62, 64–66, 68, 70–72, 74–76, 78–81, 82, 85–91, 93–97, 100, 101, 106, 108–111, 113	9, 20, 24, 26, 29, 39, 46, 56, 58, 63, 69, 77, 81, 83, 84, 99, 102, 104, 114, 115
Odmowa przyznania kredytu	1, 6, 10, 17, 21, 23, 28, 30, 32, 37, 40, 41, 43, 45, 47, 50, 52, 54, 67, 73, 92, 98, 103, 105, 107, 112	

Źródło: opracowanie własne na podstawie [Witkowska, Staniec 2000].

11 Ogólny opis algorytmów obu metod przedstawiono w rozdziale 7.

Podobne wyniki grupowania uzyskano za pomocą metody *k*-średnich (tab. 10.10). Ponownie niewiarygodni klienci nie zostali rozpoznani (zatem błąd E_1 = 100%), a w drugim skupieniu znalazło się 16 firm wiarygodnych, czyli błąd drugiego rodzaju wynosi 18%, a ogólny błąd klasyfikacji 36,5%. Przeprowadzone analizy jednoznacznie dowodzą, że informacja o przynależności do klas jest istotna w procesie rozpoznawania obiektów.

Tabela 10.10. Obiekty zakwalifikowane do poszczególnych klas otrzymanych metodą *k*-średnich

Stan faktyczny	Numery firm, które znalazły się w klasie pierwszej	
	którym przyznano kredyt	którym nie przyznano kredytu
Przyznanie kredytu	3–5, 7–8, 11–16, 18, 19, 22, 25–27, 31, 33–36, 38, 42, 44, 48, 49, 51, 53, 55–57, 59–66, 68, 70–72, 74–76, 78–82, 85–91, 93–97, 100–102, 106, 108–111, 113	2, 9, 20, 24, 29, 39, 46, 8, 69, 77, 83, 84, 99, 104, 114, 115
Odmowa przyznania kredytu	1, 6, 10, 17, 21, 23, 28, 30, 32, 37, 40, 41, 43, 45, 47, 50, 52, 54, 67, 73, 92, 98, 103, 105, 107, 112	

Źródło: [Witkowska, Staniec 2000].

10.1.3. Porównanie regresji logistycznej i modeli dyskryminacyjnych

W dalszych rozważaniach porównane zostaną wyniki klasyfikacji uzyskane z wykorzystaniem modeli regresji logistycznej i modeli dyskryminacyjnych. Badania przeprowadzono na zbiorze klientów indywidualnych, w przypadku których wiele cech ma charakter jakościowy, utrudniający dodatkowo prowadzenie analiz. Udostępniona baza danych zawiera informacje o 2576 kredytobiorcach indywidualnych, z których 419 nie spłaca kredytu hipotecznego[12]. Do badań wylosowano próbę złożoną z 740 obserwacji o identycznej liczbie klientów spłacających i niespłacających zaciągnięte zobowiązania (tzw. próba symetryczna – zbilansowana). Korzystanie z prób zbilansowanych z reguły daje lepsze wyniki klasyfikacji, gdyż modele łatwiej rozpoznają te obiekty, które należą do klasy o większej liczebności (tj. częściej reprezentowane w próbie). Próba zawiera 514 obserwacji, które stanowić będą zbiór uczący, i 226 obserwacji zbioru testowego. Zmienną grupującą jest: spłacanie bądź nie kredytu przez kredytobiorcę (zmienna przyjmuje wartość 1, kiedy kredytobiorca spłaca kredyt w terminie, i 0 w przeciwnym przypadku). Wszystkich analizowanych klientów opisano za pomocą następujących potencjalnych zmiennych diagnostycznych:

12 Udostępnione przez bank informacje dotyczą kredytów hipotecznych, udzielanych pod zastaw posiadanej nieruchomości. Zobowiązania tego rodzaju spłaca się najczęściej do 15–20 lat.

x_1 – grupa klienta (1 – klient standard; 0 – klient VIP),
x_2 – płeć (0 – mężczyzna; 1 – kobieta),
x_3 – waluta kredytu (0 – inne; 1 – PLN),
x_4 – wiek (w latach),
x_5 – miejsce udzielenia kredytu (zmienna wielodzielna: miasto M1 – duże; M2 – średnie; M3 – małe),
x_6 – procentowy udział aktualnie spłaconej kwoty obliczony w stosunku do całej kwoty kredytu,
x_7 – okres spłaty kredytu (w latach),
x_8 – wartość kredytu (PLN),
x_9 – posiadane lokaty (0 – do 1 roku, 1 – powyżej 1 roku),
x_{10} – relacja bieżącej stopy procentowej do poprzedniej stopy procentowej kredytu.

W celu zbadania, które z potencjalnych zmiennych diagnostycznych mają wpływ na decyzję o przyznaniu kredytu, czyli są istotnie skorelowane ze zmienną grupującą, zastosowano test niezależności chi-kwadrat (4.19)–(4.25). Wybór tego testu jest spowodowany tym, że zmienna grupująca jest mierzona na skali nominalnej, podobnie jak część analizowanych cech, będących potencjalnymi zmiennymi diagnostycznymi. Badanie, przeprowadzone za pomocą testu niezależności chi-kwadrat wykazało, występowanie statystycznie istotnej współzależności między wszystkimi potencjalnymi zmiennymi diagnostycznymi i zmienną grupującą, informującą o terminowym lub nieterminowym spłacaniu kredytu. W dalszych badaniach do klasyfikacji klientów banku wykorzystano modele logitowe i analizę dyskryminacyjną z wykorzystaniem tzw. funkcji klasyfikacyjnych[13].

Specyfikację modelu dyskryminacyjnego przeprowadzono, korzystając z krokowej analizy wstecznej, która polega na tym, że model jest budowany w kolejnych etapach, poczynając od modelu zawierającego wszystkie potencjalne zmienne diagnostyczne. W każdym kroku analizy szacuje się model i przeprowadza ocenę wszystkich zmiennych diagnostycznych, eliminując tę z nich, która najmniej wnosi do przewidywania przynależności badanych obiektów do określonych klas. W wyniku przeprowadzonej procedury uzyskuje się model zawierający jedynie te zmienne, które w istotny sposób przyczyniają się do dyskryminacji grup. Innymi słowy, analiza krokowa wsteczna pozwala na wyłonienie spośród wszystkich potencjalnych zmiennych diagnostycznych takich, które w największym stopniu przyczyniają się do rozpoznawania obiektów. W rezultacie oszacowano model składający się z dwóch funkcji klasyfikacyjnych postaci:

$$D_1(x) = 0{,}4652 \cdot x_7 + 0{,}2370 \cdot x_{10} - 12{,}0682$$
$$D_2(x) = 0{,}6095 \cdot x_7 + 0{,}2134 \cdot x_{10} - 11{,}7258$$

13 W programach komputerowych często zamiast klasycznej funkcji dyskryminacyjnej (7.40) wyznaczane są tzw. funkcje klasyfikacyjne budowane oddzielnie dla każdej z klas.

Konkretny obiekt (kredytobiorca) przypisywany jest do tej grupy, dla której uzyskuje wyższą wartość funkcji klasyfikacyjnej. Przedstawione wartości parametrów funkcji klasyfikujących są współczynnikami standaryzowanymi i wskazują, jak silny jest wpływ danej zmiennej dyskryminacyjnej na różnicowanie grup typologicznych. W tym konkretnym przypadku okres spłaty kredytu w latach (tj. zmienna x_7) ma większy wpływ na rozróżnienie kredytobiorców niż relacja bieżącego oprocentowania kredytu do poprzedniej stopy procentowej (zmienna x_{10}).

W przypadku modelu logitowego (6.17), którego siedem wariantów oszacowano metodą największej wiarygodności, najlepiej dopasowany okazał się model postaci:

$$ln\left(\frac{p_i}{1-p_i}\right) = 0{,}1829 + 0{,}1439 \cdot x_2 - 0{,}6324 \cdot x_3 + 0{,}0217 \cdot x_4 + 0{,}014 \cdot x_6 +$$
$$+0{,}1382 \cdot x_7 + 0{,}0211 \cdot x_{10}$$

zliczeniowy $R^2 = 67{,}5\%$;
logarytm wiarygodności = –295,971;
$\chi^2 = 102{,}15$;
R^2 McFaddena = 17,5%

Wprawdzie wartość statystyki t-Studenta dla zmiennych x_2 (t = 0,7) i x_3 (t = –0,92) wskazują na ich statystyczną nieistotność, jednak usunięcie tych zmiennych ze zbioru zmiennych objaśniających znacznie obniżyło jakość klasyfikacji klientów banku, prowadzonej za pomocą tego modelu, zatem zdecydowano o pozostawieniu tych zmiennych.

Interpretując ilorazy szans modelu, należy zauważyć, że szansa na terminową spłatę kredytu zwiększa się, jeśli kredytobiorca jest kobietą o 15,4% w stosunku do mężczyzn, ale zmniejsza się o 44,9%, jeśli walutą kredytu jest PLN w stosunku do kredytów w walutach obcych. Pozostałe zmienne, tj. wiek, procentowy udział aktualnie spłaconej kwoty obliczony w stosunku do całej kwoty kredytu, długość okresu spłaty i relacja bieżącej stopy procentowej do poprzedniej stopy procentowej kredytu zwiększają szanse spłaty kredytu w terminie odpowiednio o 2,2, 0,1, 14,8 i 2,1% dla jednostkowego wzrostu wartości zmiennej objaśniającej, przy założeniu, że wszystkie pozostałe zmienne pozostaną na niezmienionym poziomie.

Korzystając z obu wymienionych wyżej metod klasyfikacji ze wzorcem, tj. funkcji klasyfikacyjnych i modelu logitowego, przeprowadzono klasyfikację klientów banku, a jej wyniki pokazano w tab. 10.11–10.12. W pierwszej z nich porównano rzeczywiste liczebności klientów terminowo spłacających i niespłacających kredyt z liczbą rozpoznanych obiektów w każdej klasie. Jak widać, ani w próbie uczącej, ani w testowej nie została zachowana nawet zbliżona rzeczywista

proporcja liczby kredytobiorców należących do obu klas, gdyż w przypadku obu modeli rozpoznana liczba klientów niespłacających swoich zobowiązań w terminie jest zawsze większa.

Tabela 10.11. Klasyfikacja kredytobiorców na podstawie funkcji klasyfikacyjnych i modelu logitowego

Zaobserwowana liczba klientów spłacających kredyt	Próba treningowa		Próba testowa	
	Liczba klientów spłacających kredyt			
	terminowo	nieterminowo	terminowo	nieterminowo
	Klasyfikacja za pomocą modelu dyskryminacyjnego			
terminowo	148	109	75	38
nieterminowo	70	187	32	81
razem	218	296	107	119
	Klasyfikacja za pomocą modelu logitowego			
terminowo	159	98	65	48
nieterminowo	69	188	25	88
razem	228	286	90	136

Źródło: opracowanie własne na podstawie [Chrzanowska, Witkowska 2008].

Spostrzeżenie to potwierdzają obliczone błędy klasyfikacji (10.1–10.3). Oba modele rozpoznają obie grupy kredytobiorców dość słabo, generując wysokie błędy klasyfikacji. Jak można zauważyć, model logitowy wprawdzie lepiej klasyfikuje w próbie treningowej, ale ocenę jakości modelu przeprowadza się na podstawie próby testowej, dla której ogólny błąd klasyfikacji oraz błąd drugiego rodzaju są niższe w przypadku modelu dyskryminacyjnego niż logistycznego. Jednakże ten ostatni generuje w próbie testowej niższy błąd pierwszego rodzaju, który jest dla banku bardziej istotny.

Tabela 10.12. Błędy klasyfikacji przeprowadzonej za pomocą funkcji klasyfikacyjnych i modelu logitowego

	Rodzaje błędów					
	E	E_1	E_2	E	E_1	E_2
Próba / Model	treningowa			testowa		
dyskryminacyjny	34,8	27,2	42,4	31,0	28,3	33,6
logitowy	32,5	26,8	38,1	32,3	22,1	42,5

Źródło: opracowanie własne na podstawie [Chrzanowska, Witkowska 2008].

10.1.4. Porównanie regresji binarnej, bayesowskiej analizy dyskryminacyjnej

Badania realizowano na podstawie danych omówionych w paragrafie 10.1.3, ale tym razem wylosowano symetryczne próby o nieco zmienionej liczebności – treningowa zawiera 560 elementów, a testowa 278 obserwacji. Ograniczono również liczbę uwzględnionych w analizach cech diagnostycznych, wybierając te, które zdaniem ekspertów mają znaczący wpływ na zmienną grupującą. Wybrano trzy zmienne opisujące:

- x_3 – waluta kredytu (0 – inne; 1 – PLN),
- x_4 – wiek (kodowany binarnie z uwzględnieniem podziału na cztery klasy),
- x_7 – czas spłaty kredytu (kodowany binarnie z uwzględnieniem podziału na trzy klasy),

które są silnie powiązane ze zmienną grupującą i jednocześnie nie są istotnie współzależne między sobą. Przy czym wszystkie te zmienne zostały zamienione na zmienne binarne po uprzednim pogrupowaniu ich w klasy (tab. 10.13). Tak więc walutę kredytu potraktowano jako zmienną dychotomiczną, a kredyt inny niż złotówkowy jest wariantem referencyjnym. W przypadku zmiennej opisującej wiek kredytobiorcy utworzono cztery klasy, przy czym najmłodsza grupa wiekowa jest wariantem referencyjnym przy kodowaniu binarnym. Okres spłaty kredytu podzielono na trzy przedziały klasowe o jednakowej rozpiętości, przyjmując pierwszy z nich za wariant referencyjny.

Tabela 10.13. Definicja zmiennych binarnych

Zmienne wielowariantowe	Definicja zmiennych binarnych	
x_3 – Waluta kredytu	$x_{31} = 1$ dla $x_3 = PLN$	$x_{31} = 0$ dla $x_3 \neq PLN$
x_4 – Wiek kredytobiorcy [w latach]	$x_{41} = 1$ dla $x_4 \in \langle 35, 45\rangle$ lat	$x_{41} = 0$ dla $x_4 < 35$ lat
	$x_{42} = 1$ dla $x_4 \in \langle 46, 55\rangle$ lat	$x_{42} = 0$ dla $x_4 < 35$ lat
	$x_{43} = 1$ dla $x_4 \geq 56$ lat	$x_{43} = 0$ dla $x_4 < 35$ lat
x_7 – Okres spłaty kredytu [w latach]	$x_{71} = 1$ dla $x_7 \in \langle 6, 10\rangle$ lat	$x_{71} = 0$ dla $x_7 \in \langle 1, 5\rangle$ lat
	$x_{71} = 1$ dla $x_7 \in \langle 11, 15\rangle$ lat	$x_{72} = 0$ dla $x_7 \in \langle 1, 5\rangle$ lat

Uwaga: waluta kredytu jest cechą dychotomiczną; okres spłaty ma 3 warianty cechy, a wiek kredytobiorcy – 4; każdemu wariantowi cechy odpowiada jedna zmienna dychotomiczna.

Źródło: opracowanie własne na podstawie [Chrzanowska, Witkowska 2007].

Do dalszych analiz zastosowano bayesowską analizę dyskryminacyjną, wykorzystującą informacje *a priori*, dotyczące rozpatrywanego problemu (por. [Jajuga 1998]). Otrzymane funkcje klasyfikacyjne są postaci:

$$D_1(x) = -0{,}01 \cdot x_{31} - 2{,}93 \cdot x_{41} + 1{,}37 \cdot x_{42} - 0{,}37 \cdot x_{43} - 2{,}88 \cdot x_{71} + 0{,}9 \cdot x_{72} - 3{,}29$$
$$D_2(x) = -0{,}08 \cdot x_{31} - 1{,}75 \cdot x_{41} + 0{,}54 \cdot x_{42} - 0{,}16 \cdot x_{43} - 1{,}91 \cdot x_{71} + 0{,}34 \cdot x_{72} - 4{,}33$$

W przypadku tak oszacowanego modelu dyskryminacji, poza zmianą wylosowanej próby badawczej[14], istotny jest również inny (niż w paragrafie 10.1.3) sposób wyboru[15] i pomiaru zmiennych dyskryminacyjnych. Tym razem zmienne, opisujące wiek kredytobiorcy (zmienna x_4) i okres spłaty kredytu (x_7), są mierzone na skali nominalnej. Dlatego w modelach klasyfikacyjnych, zawierających *de facto* trzy zmienne diagnostyczne, znajduje się ich sześć, ponieważ zmienne x_4 i x_7 są wielodzielne. Interpretacja uzyskanych standaryzowanych wskaźników dyskryminacyjnych wskazuje, że na klasyfikację kredytobiorców do pierwszej klasy największy wpływ ma zmienna x_{41} (wiek kredytobiorcy od 36 do 45 lata), a w drugiej kolejności zmienna x_{72} (okres spłaty kredytu od 6 do 10 lat). Natomiast podczas klasyfikacji kredytobiorców do klasy drugiej wspomniane zmienne mają wprawdzie największy wpływ, ale ich kolejność jest odwrotna.

Innym modelem wykorzystanym do klasyfikacji klientów jest model regresji binarnej, tj. model (6.13), w którym również zmienne objaśniające, jako mierzone na skali nominalnej, są kodowane zmiennymi zero-jedynkowymi. Po oszacowaniu model ten jest postaci:

$$\hat{y}_i = 0{,}246 \cdot x_{31} + 0{,}03 \cdot x_{41} + 0{,}01 \cdot x_{42} + 0{,}042 \cdot x_{43} + 0{,}443 \cdot x_{71} + 0{,}402 \cdot x_{72} + 0{,}197$$

przy $R^2 = 0{,}167$. Jednocześnie przyjęto, że obiekt zostaje przypisany do grupy spłacających kredyt, jeśli $\hat{y}_i \geq 0{,}5$.

W prezentowanym modelu regresji binarnej zauważa się niską wartość współczynnika determinacji, co jest często spotykaną własnością modeli tej klasy (porównaj rys. 6.2). Zmienne opisujące okres spłaty kredytu i waluta kredytu mają istotny wpływ na zmienną objaśnianą. Z kolei nieistotnie różne od 0 okazały się parametry stojące przy wszystkich trzech wariantach zmiennej odzwierciedlającej grupy wiekowe (wartości statystyk testowych są na poziomie odpowiednio: $t = 0{,}04$, 0,23 i 0,12), co świadczy o braku wpływu wieku na spłacalność kredytu. Jednakże pozostawiono pierwotną specyfikację modelu w celu porównania skuteczności klasyfikacji prowadzonej różnymi metodami.

Oceniając uzyskane wyniki (tab. 10.14 i 10.15) należy stwierdzić, że ogólne błędy klasyfikacji są akceptowalne w próbie testowej jedynie w przypadku regresji binarnej, co oznacza, że bayesowska analiza dyskryminacyjna się w tym przypadku nie sprawdziła. Warto również odnotować, że skuteczność klasyfikacji zależy od rodzaju cech użytych do budowy modeli. Z reguły łatwiej jest uzyskać pożądaną jakość klasyfikacji, jeśli wykorzystuje się raczej zmienne ilościowe niż jakościowe charakterystyki.

---

14 W badaniach omawianych w paragrafie 10.1.3 i 10.1.4 próby losowane są wprawdzie z tego samego zbioru kredytobiorców i są symetryczne, ale nie są to takie same próby.

15 W paragrafie 10.1.3 dobór zmiennych został przeprowadzony za pomocą regresji krokowej wstecz, a w 10.1.4 dokonano arbitralnego wyboru.

Tabela 10.14. Wyniki klasyfikacji przeprowadzonej za pomocą bayesowskiej analizy dyskryminacyjnej i regresji binarnej

Klasyfikacja rzeczywista	Próba treningowa		Próba testowa	
	Klasyfikacja za pomocą modelu: klient			
	spłaca kredyt	nie spłaca kredytu	spłaca kredyt	nie spłaca kredytu
Klient	Model dyskryminacyjny			
spłaca kredyt	178	102	15	124
nie spłaca kredytu	50	230	27	112
razem	228	332	42	236
Klient	Regresja binarna			
spłaca kredyt	244	36	127	12
nie spłaca kredytu	142	138	66	73
razem	386	174	193	85

Źródło: opracowanie własne na podstawie [Witkowska, Chrzanowska 2006].

Tabela 10.15. Błędy klasyfikacji przeprowadzonej za pomocą bayesowkiej analizy dyskryminacyjnej i regresji binarnej

		Rodzaje błędów					
		E	E_1	E_2	E	E_1	E_2
		Próba treningowa			Próba testowa		
Model	dyskryminacyjny	27,1	17,9	36,4	54,3	19,4	89,2
	logitowy	31,8	50,7	12,9	28,1	47,5	8,6

Źródło: opracowanie własne na podstawie [Witkowska, Chrzanowska 2006].

10.1.5. Klasyfikacja indywidualnych kredytobiorców za pomocą drzew klasyfikacyjnych

W analizach kredytobiorców indywidualnych (opisanych w punkcie 10.1.3) zastosowano również drzewa klasyfikacyjne budowane za pomocą algorytmu CART dla 540 wylosowanych obiektów. Wykorzystano dane dotyczące trzech zmiennych:

x_5 – miejsca udzielenia kredytu,

x_7 – okresu spłaty kredytu,

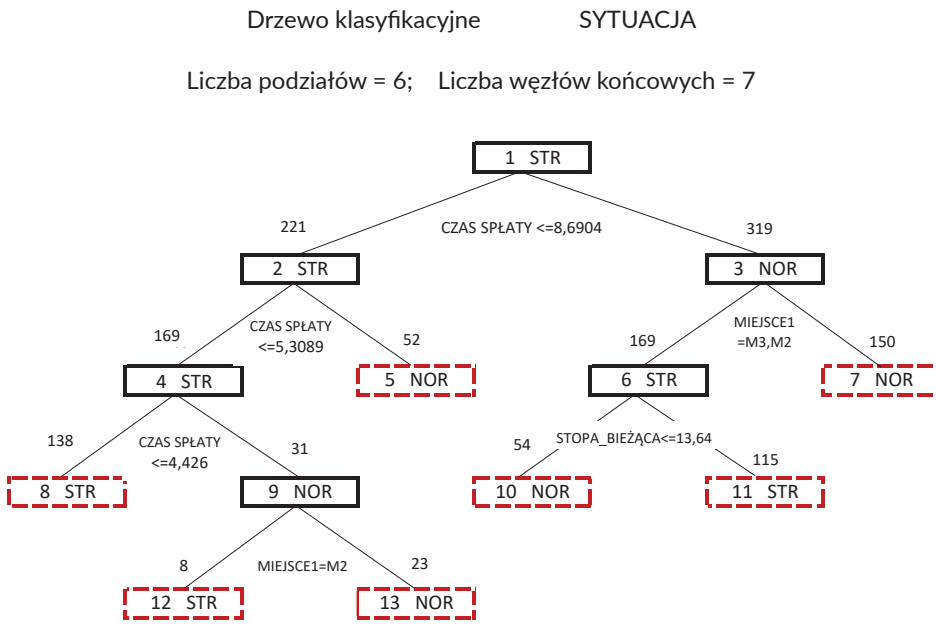
x_{11} – bieżącej stopy procentowej kredytu.

W wyniku przeprowadzonej klasyfikacji uzyskano drzewo jak na rys. 10.2.

Na początku wszystkie kredyty zostały przypisane do węzła nr 1 jako kredyty stracone. Węzeł ten uległ następnie podziałowi na węzły o numerach 2 i 3.

Kredyty, których *czas spłaty* jest krótszy niż 8,6904 roku (221 zobowiązań) zostały przypisane do węzła nr 2 jako kredyty niespłacone. Pozostałe 319 kredytów, których *czas spłaty* jest dłuższy niż 8,6904 roku zostały przyporządkowane do węzła nr 3 jako kredyty normalne, tzn. spłacane terminowo. Węzeł nr 3 uległ dalszemu podziałowi. Kredyty, które zostały zaciągnięte w małych i średnich miastach (zmienna *miasto*; warianty M2 i M3), przypisane zostały do węzła nr 6 i określone jako kredyty „stracone". Pozostałe zobowiązania (kredyty zaciągnięte w dużych miastach) zostały przypisane do węzła nr 7 i sklasyfikowane jako kredyty normalne. Węzeł nr 7 jest węzłem końcowym (liściem) i nie ulega dalszemu podziałowi. Analizując tę ścieżkę drzewa, można twierdzić, że spłacane terminowo zobowiązania (150 kredytów) to kredyty zaciągnięte w dużych miastach (M1), dla których *czas spłaty* jest dłuższy niż 8,6904 roku.

Węzeł nr 6 ulega dalszemu podziałowi na dwa węzły: nr 10 i 11. Kredyty, dla których *stopa bieżąca* jest niższa niż 13,64, zostały rozpoznane jako normalne i przydzielone do węzła nr 10. Pozostałe kredyty o *aktualnej stopie procentowej* wyższej niż 13,64% zostały sklasyfikowane jako kredyty stracone. Analizując tę gałąź drzewa, można stwierdzić, że zobowiązania spłacane terminowo to kredyty o *bieżącej stopie procentowej* niższej niż 13,64%, dla których *czas spłaty* jest dłuższy niż 8,6904 roku i zostały zaciągnięte w małych i średnich miastach.



Rysunek 10.2. Schemat drzewa klasyfikacyjnego

Źródło: obliczenia własne.

Węzeł nr 2, w którym kredyty zostały rozpoznane jako kredyty stracone, ulega dalszemu podziałowi. Kredyty, dla których *czas spłaty* był wyższy niż 5,3089 roku, zostały przypisane do węzła nr 5 i rozpoznane jako kredyty normalne (kredyty spłacane terminowo). Na podstawie tej ścieżki można stwierdzić, że kredyty normalne to kredyty, dla których czas *spłaty* jest w przedziale między 5,31 a 8,69 roku.

Węzeł nr 4 (gałąź od węzła nr 2) to kredyty, dla których *czas spłaty* jest krótszy niż 5,3089 roku. Węzeł ten uległ dalszemu podziałowi na węzły nr 8 i 9. Kredyty, dla których *czas spłaty* jest krótszy niż 4,426 roku, zostały rozpoznane jako kredyty stracone. Pozostałe kredyty przypisano do węzła nr 9 i określono jako normalne. Węzeł nr 9 uległ dalszemu podziałowi na węzły 12 i 13. Zobowiązania zaciągnięte w średnich miastach, zostały przypisane do grupy kredytów straconych; pozostałe kredyty – normalne – zostały przypisane do węzła nr 13. Podsumowując tę gałąź drzewa, można twierdzić, że kredyty spłacane to zobowiązania zaciągnięte w małych i dużych miastach, dla których *czas spłaty* należy do przedziału (4,426; 5,3089) roku.

W prezentowanym modelu najsilniejszy wpływ na zmienną grupującą ma zmienna *czas spłaty*. W dalszej kolejności wyraźny wpływ na klasyfikację mają zmienne *stopa bieżąca* oraz *miejsce* (*zaciągnięcia kredytu*).

Tabela 10.16. Wyniki klasyfikacji kredytobiorców w próbie treningowej

Liczba zaklasyfikowanych kredytobiorców do każdej z klas			Błędy klasyfikacji		
Klasyfikacja rzeczywista	Klasyfikacja modelowa				
Klient	spłaca kredyt	nie spłaca kredytu	E	E_1	E_2
spłaca kredyt	220	50	20,2	21,9	18,5
nie spłaca kredytu	59	211			
razem	279	261			

Źródło: obliczenia własne.

Biorąc pod uwagę wyniki klasyfikacji przeprowadzone za pomocą drzewa przedstawionego na rys. 10.2, wydedukować można, że do klasy:

- terminowo spłacających kredyt zaliczono 279 kredytobiorców (suma obiektów zapisanych do „liści drzewa" wyrażających sytuację normalną, tj. 150 w węźle 7, 54 w węźle 10, 52 w węźle 5 i 23 w węźle13),
- niespłacających kredyt – 261 klientów banku (suma obiektów zapisanych do „liści drzewa" wyrażających sytuację straconą, tj. 115 w węźle 11, 138 w węźle 8 i 8 w węźle 12).

Porównanie obiektów znajdujących się w obu klasach z danymi rzeczywistymi należącymi do próby treningowej daje wyniki zapisane w tab. 10.16, a w próbie testowej w tab. 10.17.

Tabela 10.17. Wyniki klasyfikacji kredytobiorców w próbie testowej

Liczba zaklasyfikowanych kredytobiorców do każdej z klas			Błędy klasyfikacji		
Klasyfikacja rzeczywista	Klasyfikacja modelowa				
Klient	spłaca kredyt	nie spłaca kredytu	E	E_1	E_2
spłaca kredyt	112	27	30,2	41,0	19,4
nie spłaca kredytu	57	82			
razem	169	109			

Źródło: obliczenia własne.

10.2. Ocena sytuacji finansowej spółek za pomocą mierników syntetycznych

Ocena sytuacji finansowej przedsiębiorstw jest m.in. elementem analizy fundamentalnej, której celem jest wybór spółek będących przedmiotem inwestycji na rynku kapitałowym. W trakcie prowadzenia analizy fundamentalnej bada się szereg aspektów działania przedsiębiorstw, oceniając takie charakterystyki jak:

- rodzaj produktu lub usługi,
- technologię wytwarzania,
- kwalifikacje pracowników,
- sposób zarządzania,
- pozycję rynkową,
- relacje z dostawcami i odbiorcami, a także
- sytuację finansową spółki.

Spółka jest więc opisana za pomocą wielu charakterystyk, które są mierzone na różnych skalach pomiarowych, tzn.:

- nominalnej (np. forma własności, sektor gospodarki, rodzaj produktu lub usługi, relacje z dostawcami i nabywcami),
- porządkowej (np. wielkość przedsiębiorstw: mikro-przedsiębiorstwa, małe, średnie i duże; ocena kierownictwa spółki i sposobu zarządzania),
- interwałowej (np. wielkość produkcji mleka [hektolitry] lub samochodów [sztuki], majątek przedsiębiorstwa [tys. PLN], wysokość zadłużenia [tys. PLN], cena akcji [PLN], cena USD w relacji do euro [euro]),
- ilorazowej (np. wskaźniki finansowe, wydajność pracy).

Ważne jest stwierdzenie, że w zależności od tego, czemu służy prowadzona analiza, w badaniu wykorzystuje się różne charakterystyki spółek. Co więcej, czasami wykorzystując w nich te same cechy opisujące przedsiębiorstwo, przyjmuje się za pożądane inne wartości badanych cech. Przykładowo popularny wskaźnik rynko-

wy wyrażający stosunek ceny akcji do wartości księgowej P/BV z punktu widzenia emitenta akcji powinien być jak najwyższy, a z puntu widzenia inwestora – jak najmniejszy, bo wtedy akcja jest niedowartościowana i warto w nią inwestować.

Istotą analizy wielowymiarowej jest sprowadzenie badanych obiektów wielocechowych (tj. opisywanych za pomocą wielu cech) do – wyrażonej jedną liczbą – ich charakterystyki. Innymi słowy, za pomocą metod wielowymiarowej analizy porównawczej każdy badany obiekt, np. spółka, uzyskuje pomiar w postaci miary agregującej różne cechy tworzące przestrzeń wielowymiarową. Ta miara nazywana jest miernikiem taksonomicznym, syntetycznym lub agregatowym. Mierniki taksonomiczne są niemianowane i *de facto* nie mają interpretacji ekonomicznej. Pozwalają natomiast uporządkować obiekty badania od najlepszych do najgorszych oraz przeprowadzić ich grupowanie[16].

Ocena standingu finansowego, stanowiąca jeden z etapów analizy fundamentalnej, przeprowadzana jest najczęściej na podstawie analizy wskaźników finansowych. Indykatory wykorzystywane w tej analizie mierzą sytuację finansową spółki, a przy ich obliczaniu bierze się pod uwagę przede wszystkim dane rynkowe spółki i jej sprawozdania finansowe (bilans, rachunek zysków i strat, rachunek przepływów finansowych – *cash flow*).

Warto zauważyć, że już sama ocena sytuacji finansowej spółki wymaga zazwyczaj zbadania przynajmniej kilku wskaźników finansowych, które – jeśli są rozpatrywane pojedynczo (indywidualnie) – dają zróżnicowane rankingi uwzględnionych w badaniu obiektów. Zatem, stosując wielowymiarową analizę porównawczą w miejsce kilku wskaźników finansowych, które pozycjonują zbiór analizowanych spółek w różnej kolejności, uzyskuje się jeden miernik zagregowany. W celu wyboru spółek, będących przedmiotem inwestycji na rynku kapitałowym, należy przeprowadzić ranking spółek i wybrać te, które są najlepsze z punktu widzenia odpowiednio skonstruowanego miernika syntetycznego.

W dotychczasowych badaniach wielokrotnie wykorzystywano mierniki syntetyczne do oceny siły fundamentalnej spółek i ich wyboru do budowy portfela inwestycyjnego. Jednym z popularnych podejść tego typu są różnie skonstruowane taksonomiczne miary atrakcyjności inwestycji[17] (TMAI). Przy czym do budowy TMAI w większości przypadków wykorzystywano wskaźniki finansowe lub mierniki ich dynamiki. Najczęściej były to wskaźniki płynności, zadłużenia, rotacji zapasów, należności i zobowiązań, ROA, ROE, a także wskaźniki rynkowe, np. cena do wartości księgowej[18].

16 Por. przykład 8.1.

17 Miernik ten opisany został po raz pierwszy w pracy [Tarczyński 1994].

18 TMAI jest miernikiem syntetycznym często wykorzystywanym w analizach polskiego rynku kapitałowego (por. np. [Tarczyński 1995; Tarczyński, Łuniewska 2004; Tarczyńska-Łuniewska 2013; Tarczyńska-Łuniewska, Tarczyński 2006; Dmitruk 2012; Węgrzyn 2013; Juszczyk 2015; Witkowska, Kuźnik 2019; Witkowska, Staszak 2021]).

10.2.1. Opis danych

Wybór wskaźników finansowych, które uwzględnia się w analizach, zależy od celu badania i może być dokonywany arbitralnie, tj. metodą ekspercką, lub na podstawie mniej lub bardziej skomplikowanych analiz statystycznych[19]. Ważnym jednak jest, aby zbiór zmiennych diagnostycznych spełniał, podobnie jak w omawianych wcześniej zadaniach klasyfikacji, określone wymogi, m.in. uniwersalności, kompletności, odpowiedniej zmienności, braku korelacji między zmiennymi. Aby to zapewnić, prowadzi się analizy w odniesieniu do wskaźników należących do różnych grup, to jest płynności finansowej, rentowności, sprawności działania itp. Często w badaniach należy uwzględnić specyfikę działalności, jaką prowadzą analizowane przedsiębiorstwa. Przykładowo badania realizowane są oddzielnie dla spółek finansowych i niefinansowych z uwzględnieniem ich specyfiki, w tym również dla zróżnicowanych zbiorów wskaźników finansowych.

Prezentowane w dalszej części analizy dotyczą 20 spółek giełdowych, mających w drugim kwartale roku 2020 najwyższe udziały w czterech indeksach giełdowych: WIG 20, WIG spożywczy, WIG energia i WIG informatyka. Lista spółek wraz z informacją o ich procentowym udziale w portfelach indeksów giełdowych została zamieszczona w tab. 10.18. W badaniu wykorzystano syntetyczny miernik rozwoju (8.21)–(8.23) oraz sumy standaryzowane postaci (8.29).

Tabela 10.18. Wartości wskaźników w 2018 roku

Indeks	Spółka	Udział w indeksie (%)	Wskaźniki			
			płynności bieżącej	ROE	ogólnego zadłużenia	rotacji należności w dniach
WIG 20	PKNORLEN	14,43	1,7879	0,1568	0,4428	33,2658
	CDPROJEKT	16,54	6,297	0,1090	0,1100	60,3883
	KGHM	7,06	1,2391	0,1063	0,4439	15,7483
	DINOPL	5,23	0,5998	0,2538	0,6314	1,8787
	LPP	4,99	1,3629	0,1766	0,4684	8,0718
WIG energia	PGE	40,39	0,7231	3,161	0,3703	53,6260
	ENERGA	16,96	1,9312	7,1842	0,5205	64,1758
	TAURONPE	13,73	0,6252	1,1235	0,5032	42,9235
	ENEA	13,39	1,4414	4,7793	0,4978	59,1313
	CEZ	5,33	1,0406	4,3881	0,6618	121,4723

19 Opis szeroko zakrojonych badań nad wyborem odpowiednich wskaźników finansowych w celu konstrukcji miernika syntetycznego umożliwiającego wybór spółek do portfela inwestycyjnego przedstawiono w pracy [Witkowska, Kompa, Staszak 2021].

Indeks	Spółka	Udział w indeksie (%)	Wskaźniki			
			płynności bieżącej	ROE	ogólnego zadłużenia	rotacji należności w dniach
WIG informatyka	ASSECOPOL	47,17	1,6068	0,0775	0,164	95,1842
	COMARCH	12,57	1,7227	0,0597	0,1288	121,9478
	ASSECOSSE	11,45	1,7887	0,0852	0,0494	71,9353
	LIVECHAT	9,70	10,9118	1,0679	0,0000	2,358
	ASSECOBS	7,08	1,2122	0,2096	0,0415	66,4378
WIG spożywczy	KERNEL	62,81	26,1888	0,1326	2,7837	11,1154
	WAWEL	14,01	3,9897	0,1153	0,1337	119,6564
	AMBRA	5,17	1,1727	0,1171	0,1543	61,5556
	KRUSZWICA	4,36	4,1124	0,1384	0,2759	12,4316
	OVOSTAR	4,17	4,898	0,1375	0,1289	2,9660
Charakter zmiennej			S	S	D	D
Średnia			3,7326	1,1790	0,4255	51,3135
Odchylenie standardowe			5,6969	1,9823	0,5779	39,6336

Źródło: opracowanie własne na podstawie danych z raportów spółek[20] i GPW (dostęp 30.05.2020).

Ocenę sytuacji finansowej oparto na danych z raportów rocznych analizowanych spółek za rok 2018. Do budowy mierników taksonomicznych wykorzystano cztery nieskorelowane (współczynnik korelacji liniowej mniejszy niż 0,8) zmienne (porównaj dane w tab. 10.19):

1) wskaźnik płynności bieżącej,
2) wskaźnik rentowności kapitału ROE,

20 www.stockwatch.pl/gpw/indeks/wig-spozyw,sklad.aspx; www.kernel.ua/wp-content/uploads/2019/10/Kernel_FY2019_Annual_Report_.pdf; www.test2.wawel.com.pl/relations.php/pl/raporty/raporty-okresowe.html; www.ambra.com.pl/assets/RI/Raporty/okresowe/20182019/AMBRA_raport_roczny_2018-2019.pdf; https://ztkruszwica.pl/pl/relacje-inwestorskie/raporty-gieldowe/raporty-roczne/2019; https://www.GPW.pl; https://www.gkpge.pl/relacje-inwestorskie/Materialy-do-pobrania?rok=2018; https://ir.energa.pl/pr/426660/skonsolidowane-wyniki-finansowe-za-2018-rok; https://www.tauron.pl/tauron/relacje-inwestorskie/raporty-okresowe; https://ir.enea.pl/pr/427942/skonsolidowany-raport-roczny-rs-2018; http://www.cezpolska.pl/pl/cez-w-polsce/raporty-roczne; https://raportzintegrowany2018.orlen.pl/pl/skonsolidowane-sprawozdanie-finansowe; https://www.cdprojekt.com/pl/wp-content/uploads-pl/2019/03/2018_rs_spraw_finansowe.pdf; https://kghm.com/pl/inwestorzy/prezentacje/raporty-finansowe; https://grupadino.pl/wp-content/uploads/2019/03/RS_2018_Grupa_Grupa_Dino_Polska_Sprawozdanie_Finansowe.pdf; https://www.lppsa.com/wp-content/uploads/2019/04/GK-LPP-skonsolidowany-roczny-raport-za-2018-rok.pdf; https://gpwbenchmark.pl/karta-indeksu?isin=PL9999999987#Portfolio; opracowanie własne na podstawie informacji z bazy danych EMIS (dostęp 30.05.2020).

3) wskaźnik ogólnego zadłużenia,
4) wskaźnik rotacji zapasów w dniach,

z których dwie pierwsze to stymulanty (oznaczone w tab. 10.18 przez S), a dwie ostatnie to destymulanty (oznaczone w tab. 10.18 przez D).

Tabela 10.19. Współczynniki korelacji liniowej Pearsona (4.5)

Wskaźniki	Wskaźniki		
	ROE	ogólnego zadłużenia	rotacji należności w dniach
płynności bieżącej	–0,17978	0,7637	–0,3219
ROE		0,0693	0,2302
ogólnego zadłużenia			–0,2637

Źródło: obliczenia własne.

10.2.2. Budowa mierników syntetycznych

Korzystając z danych zawartych w tab. 10.18, w której określono również rodzaj wpływu danej zmiennej na sytuację finansową spółki, przeprowadzono ich standaryzację za pomocą wzoru (8.7), a następnie wyznaczono mierniki syntetyczne. W przypadku sum standaryzowanych do stymulacji destymulant zatasowano standaryzację ilorazową (8.24) przy $c = 1$, a skonstruowany w ten sposób miernik oznaczono jako AMR(1) oraz stymulację różnicową (8.25) dla $a = 0$ i $b = 1$, a miernik oznaczono przez AMR(2). Po wyznaczeniu wartości mierników wszystkie spółki uporządkowano od najlepszej do najgorszej i zaklasyfikowano je do czterech grup opisanych wzorami (8.33)–(8.36). W klasie pierwszej znalazły się spółki najlepsze, a w ostatniej najsłabsze. Uporządkowanie spółek po klasach przedstawiono w tab. 10.20.

Jak można zauważyć w tab. 10.20, rankingi spółek odzwierciedlające ich syntetyczną ocenę dokonaną za pomocą miernika ze wzorcem SMR i miernika bez wzorca AMR(2) są bardzo zbliżone, chociaż zawartości poszczególnych klas nieco się różnią. Natomiast ranking utworzony na podstawie miernika AMR(1) jest znacząco różny. Zróżnicowanie ocen sytuacji finansowej spółek (wyznaczonej na podstawie różnych indykatorów) pokazano w tab. 10.21, zawierającej porównanie pozycji rankingowych poszczególnych spółek w zależności od skonstruowanego miernika syntetycznego. Jest to szczególnie widoczne dla spółek: CEZ, KERNEL, KGHM, KRUSZWICA, LIVECHAT i OVOSTAR. Warto zauważyć, że w przypadku inwestycji giełdowych zaleca się inwestowanie w spółki należące do pierwszej, ewentualnie drugiej grupy (por. np. [Witkowska, Staszak 2021]). Analizując częstość pojawiania się spółek w tych dwóch klasach, zauważa się, że we wszystkich trzech rankingach w najwyższych klasach znalazły się spółki: ENEA, ENERGA i PGE, czyli spółki z WIG energia, a dwukrotnie (na trzy sporządzone rankingi) obecne były spółki CDPROJEKT z WIG 20, KRUSZWICA i OVOSTAR z WIG spożywczy oraz LIVECHAT z WIG informatyka.

Tabela 10.20. Ranking spółek

	Spółki	AMR(1)		AMR(2)		SMR
Grupa 1.	PKNORLEN	30,3824	LIVECHAT	3,1756	LIVECHAT	0,3814
	KGHM	29,3418	ENERGA	2,2243	ENERGA	0,2998
Grupa 2.	ENEA	14,4789	OVOSTAR	1,4123	ENEA	0,2716
	ENERGA	11,8790	ENEA	1,0917	PGE	0,2252
	LPP	11,6381	KRUSZWICA	0,7817	OVOSTAR	0,2187
	PGE	7,1435	PGE	0,5088	CDPROJEKT	0,2116
Grupa 3.	CEZ	4,1572	KERNEL	0,3478	KRUSZWICA	0,2000
	KERNEL	2,6731	CDPROJEKT	0,2274	TAURONPE	0,1532
	CDPROJEKT	2,4461	LPP	0,0952	CEZ	0,1454
	TAURONPE	2,1421	KGHM	–0,1133	PKNORLEN	0,1423
	DINOPL	0,9887	DINOPL	–0,1256	LPP	0,1417
	AMBRA	0,7538	PKNORLEN	–0,4315	KGHM	0,1354
	ASSECOBS	0,1841	TAURONPE	–0,4962	DINOPL	0,1229
	ASSECOSSE	–0,5076	ASSECOBS	–0,6486	ASSECOBS	0,1143
	LIVECHAT	–0,9636	ASSECOSSE	–0,7625	ASSECOSSE	0,1137
	WAWEL	–1,8920	AMBRA	–0,7742	AMBRA	0,1134
	ASSECOPOL	–2,2353	CEZ	–1,0327	KERNEL	0,0932
	COMARCH	–2,3041	ASSECOPOL	–1,5832	WAWEL	0,0822
	OVOSTAR	–3,0890	WAWEL	–1,7109	ASSECOPOL	0,0784
Grupa 4.	KRUSZWICA	–5,3404	COMARCH	–2,1862	COMARCH	0,0359

Uwaga: cieniowaniem oznaczono spółki należące do poszczególnych grup.

Źródło: obliczenia własne.

Tabela 10.21. Pozycje poszczególnych spółek w rankingach utworzonych na podstawie mierników syntetycznych

Spółki	Pozycja w rankingu			Spółki	Pozycja w rankingu		
	AMR(1)	AMR(2)	SMR		AMR(1)	AMR(2)	SMR
AMBRA	12	16	16	KERNEL	8	7	17
ASSECOBS	13	14	14	KGHM	2	10	12
ASSECOPOL	17	18	19	KRUSZWICA	20	5	7
ASSECOSSE	14	15	15	LIVECHAT	15	1	1
CDPROJEKT	9	8	6	LPP	5	9	11
CEZ	7	17	9	OVOSTAR	19	3	5
COMARCH	18	20	20	PGE	6	6	4
DINOPL	11	11	13	PKNORLEN	1	12	10
ENEA	3	4	3	TAURONPE	10	13	8
ENERGA	4	2	2	WAWEL	16	19	18

Uwaga: cieniowaniem oznaczono spółki, których pozycje były najbardziej zróżnicowane.

Źródło: obliczenia własne.

Rozdział 11

Badanie zmian cen na rynku dóbr heterogenicznych

Inwestycje w instrumenty finansowe zaliczane są do „klasycznych" w odróżnieniu od inwestycji zwanych alternatywnymi, do których należą m.in. inwestycje w nieruchomości, numizmaty, dzieła sztuki, dobra luksusowe lub wina. Szczególne zainteresowanie tego typu inwestycjami pojawia się w okresach rozczarowania klasycznymi instrumentami rynku finansowego. Jednakże, aby można było wymienione wyżej dobra traktować jako instrumenty inwestycyjne, konieczne jest chociażby przybliżone oszacowanie wartości możliwych do zrealizowania stóp zwrotu. Wobec heterogeniczności obiektów występujących na wymienionych rynkach i braku możliwości obiektywnej ich wyceny (lub tzw. *fair value*) niezbędne jest konstruowanie specyficznych indeksów cen dla określonych rynków i/lub ich segmentów. Bez uszczerbku dla ogólności i możliwych zastosowań dalsze rozważania ograniczone zostaną do wybranych zagadnień z analizy rynku sztuki.

11.1. Dzieła sztuki jako przedmioty inwestycji alternatywnych

Dzieła sztuki, które są uznawane przez historyków sztuki i krytyków, były i są od zawsze – niezależnie od ich walorów estetycznych oraz wzmacniania pozycji w hierarchii społecznej ich właścicieli – postrzegane jako jedna z form tezauryzacji lub inwestycji. Postrzegając dzieła sztuki jako przedmiot inwestycji[1] (*art investing*), należy zauważyć, że są one wysoce heterogeniczne, co uniemożliwia ich bezpośrednie porównanie, a ich wartość (cena) zależy od wielu subiektywnych czynników. Aktualna wartość dzieła sztuki jest trudna do oszacowania,

1 Szersze omówienie tych zagadnień znaleźć można m.in. w pracach [Goetzmann i in. 2011; Frey, Cueni 2013; Renneboog, Spaenjers 2013; Witkowska 2014; Sawicki, Borowski 2018].

a cena konkretnej pracy jest limitowana jedynie kwotą, jaką inwestor jest gotów i może za nie zapłacić. Inwestycje na rynku sztuki są niepodzielne i niepłynne w porównaniu z klasycznymi instrumentami finansowymi, a okresy, na jakie zamraża się kapitał w dzieła sztuki, są z reguły długie. Prace uznanych artystów są często drogie, a koszty transakcyjne dochodzą do 25% ceny wylicytowanego dzieła, czyli są znacznie wyższe niż w przypadku inwestycji w klasyczne instrumenty finansowe. Oprócz tego, co warto odnotować, rynek dzieł sztuki jest wysoce nieprzejrzysty, bowiem większość dokonywanych transakcji nie jest rejestrowana, a rynek aukcyjny obejmuje według szacunków jedynie 30% obrotów. A nawet w tym przypadku obie strony, to jest kupujący i sprzedający, mogą pozostać anonimowi (oczywiście wtedy, kiedy sobie tego życzą).

Inwestowanie w sztukę jest dodatkowo obarczone szeregiem ryzyk, charakterystycznych dla tego typu inwestycji i niewystępujących przy klasycznie pojmowanych inwestycjach. Po pierwsze, kupujący nigdy nie ma pewności, czy zakupiony obiekt jest oryginalny, tzn. nie jest kopią lub falsyfikatem. A nawet w przypadku, kiedy zakupione dzieło jest oryginalne, nie jest pewne, kto jest jego autorem: sam mistrz, ktoś z uczniów, czy zostało ono jedynie przygotowane „w stylu" danego twórcy. Po drugie, jakość zakupionego dzieła może być różna, gdyż mogło zostać ono podmalowane, przemalowane, uszkodzone lub przechowywane w niewłaściwych warunkach, co ma istotny wpływ na jego wartość. Po trzecie, wartościowe dzieła sztuki mogą zostać skradzione lub uszkodzone, co naraża ich właścicieli na znaczne straty. Konieczność ich przechowywania i zabezpieczenia w odpowiedni sposób, ubezpieczenia (np. od kradzieży) lub opłacenia podatków (od zakupu lub sprzedaży) powoduje dodatkowe koszty takich inwestycji. Co więcej, niektóre dzieła sztuki mogą zostać uznane za elementy dziedzictwa narodowego i zostać skonfiskowane lub może pojawić się zakaz ich wywozu za granicę. Po czwarte, moda i gust odbiorców sztuki zmieniają się, co powoduje, że nigdy nie jest wiadomo, który z twórców przejdzie do historii, a kto zostanie zupełnie zapomniany.

Pierwszą usługę dotyczącą inwestowania w sztukę zaoferował nowojorski Citibank w 1979 roku. Motywacją do wprowadzenia oferty była chęć wyróżnienia elitarnych klientów. Wówczas nie przypuszczano, że dzieła sztuki, antyki i biżuteria będą instrumentami inwestycyjnymi. Liderami wśród instytucji zajmujących się obsługą inwestycji dotyczących sztuki są Deutsche Bank, UniCredit i UBS. Podobne usługi są oferowane również przez ABN Amro, Bank of America, J.P. Morgan Chase & Co i inne banki. Większość poważnych instytucji finansowych posiada ekskluzywną ofertę art bankingu lub funduszy dzieł sztuki (*art funds*). Jest to doskonały pomysł na reklamę i umocnienie wizerunku, ponadto stale przybywa zamożnych i wymagających interesującej oferty klientów. Na rynku inwestowania w sztukę spotykane jest również podejście czysto ekonomiczne, w którym dzieła kupowane wyłącznie w celach inwestycyjnych nie posiadają tak zwanej wartości dodanej. W takim podejściu sztuka jest alternatywną inwestycją, nieróżniącą się

od złota czy zboża. Kolejne publikacje oraz analizy rynkowe stale napędzają modę na inwestycje w dzieła sztuki (por. [Borowski 2007])[2].

Dominującą strategią inwestycyjną na rynku dzieł sztuki jest „kup i trzymaj", gdyż posiadacze dzieł sztuki z reguły wystawiają je na licytację dopiero po dłuższym czasie. W ostatnich 125 latach średnia długość posiadania obrazu przez jednego właściciela wynosiła około 30 lat. Przy czym dzieła sztuki jako aktywa są najrzadziej usuwane z portfela inwestycyjnego (por. [Brewster 2006]). Najkrótszym okresem inwestycji w sztukę, jaki powinno się brać pod uwagę, jest 5 lat. Jednakże w Polsce średnie zwroty z prac artystów trzymanych dłużej niż 15 lat wynoszą 46,6%, a tych trzymanych krócej niż 5 lat tylko 0,2% (por. [Deloitte 2013]).

Przeszkodami w rozwoju inwestycji w sztukę na szeroką skalę są brak podzielności dzieł i wysoki koszt jednostkowy. Oprócz tego, o czym była mowa, dzieła sztuki są wysoce heterogeniczne, uniemożliwiając ich bezpośrednie porównanie i charakteryzują się niską płynnością. Brak możliwości sprzedania dzieła przez dłuższy czas po wyznaczonej przez właściciela cenie jest częstym zjawiskiem, zwłaszcza w sytuacji nabycia obrazu w okresie szczytu cenowego na dzieła sztuki. Wiadomo też, że nawet nieznaczny spadek dochodów decyla najlepiej zarabiających powoduje istotne obniżenie cen na rynku dzieł sztuki. Niska płynność stwarza dodatkowo problem z dywersyfikacją portfela, ponieważ nie ma możliwości kupna części obrazu, jak w przypadku niektórych instrumentów[3]. Przy czym płynność obrotu dziełami sztuki jest niemal dwukrotnie wyższa w przypadku najdroższych dzieł sztuki niż w przypadku prac, których cena nie przekracza 10 tys. USD (por. [Finkel 2004]).

11.2. Indeksy hedoniczne cen

Indeksy są uniwersalnym narzędziem wykorzystywanym w analizach różnych rynków do syntetycznego przedstawienia zmian, jakie na nich zaszły. Istnieje wiele powszechnie stosowanych metod konstrukcji indeksów cenowych, które są indeksami agregatowymi o ogólnej postaci (5.10). Problematyczne pozostaje zagadnienie dokładności, z jaką skonstruowane indeksy odzwierciedlają rzeczywiste zmiany, zważywszy na fakt, że wiele z nich nie rozpatruje problemu zmian jakościowych. Wśród metod konstrukcji indeksów cenowych wymienia się (por. [Tomczyk, Widłak 2010])[4]:

2 W pracy [Sawicki, Borowski 2018, s. 86–101] zawarto komentarze na temat ekonometrycznych badań rynku dzieł sztuki.

3 Pewną nadzieję na zmianę tej sytuacji stwarza tokenizacja dzieł sztuki, choć podejmowane próby ciągle mają charakter nowinek FinTech, bez istotnego wpływu na skalę obrotów.

4 Klasyfikację i omówienie niektórych indeksów rynku sztuki znaleźć można w pracy [Bernaś 2016].

- metody średniej lub mediany,
- metodę śledzenia ceny reprezentatywnej,
- metodę stratyfikacji i średniej ważonej,
- metodę powtórnej sprzedaży (bez hedonicznej korekty zmian jakości) i
- metodę hedoniczną.

Indeksy hedoniczne[5] to takie indeksy cenowe, które wykorzystują funkcję hedoniczną. Model hedoniczny to relacja pomiędzy cenami różnych rodzajów danego dobra (np. dzieł sztuki lub nieruchomości) a cechami, które to dobro charakteryzują. Model ten przyjmuje zazwyczaj postać (6.11) lub ogólnie (6.1), w której większość zmiennych objaśniających (niezależnych) jest zero-jedynkowa, ponieważ reprezentuje cechy mierzone na skali nominalnej. Indeksy hedoniczne są szeroko wykorzystywane jako wskaźniki cen na rynkach dóbr heterogenicznych (np. nieruchomości lub dzieł sztuki) lub dóbr o często zmieniających się charakterystykach jakościowych (np. komputerów). Mają odzwierciedlać ruchy cen skorygowane o pojawiające się zmiany w jakości (*quality adjustment*) dóbr uwzględnionych we wskaźniku dynamiki cen. Hedoniczny indeks cenowy wykorzystuje oceny estymatorów parametrów modelu hedonicznego postaci (6.11) lub ogólnie (6.1).

Konstruktorzy modeli hedonicznych zakładają, że zmiana cen dóbr heterogenicznych (np. dzieła sztuki, mieszkania, komputera) jest związania z ich jakością, na którą wpływają wartości pisujących je charakterystyk. Przykładowo w przypadku dzieł sztuki podstawowymi atrybutami są:

- rodzaj dzieła, np. rzeźba, obraz, numizmat;
- autor dzieła (nazwisko lub pseudonim), w szczególności ważne jest, czy jest to uznany twórca, czy też nieznany oraz czy autor żyje w momencie wyceny dzieła;
- okres, z jakiego dzieło pochodzi, np. współcześnie powstałe dzieło, dzieło powstałe przed drugą wojną światową, w XIX wieku;
- technika wykonania, przykładowo w przypadku obrazów może to być: olej na płótnie, akwarela, gwasz;
- wielkość dzieła sztuki, gdyż mniejsze obiekty są z reguły łatwiejsze do sprzedania niż wielkie, monumentalne prace;
- podmiot wystawiający obiekt do sprzedaży, co może stanowić gwarancję autentyczności dzieła.

Dla mieszkań istotne są m. in. lokalizacja, wyposażenie mieszkania lub technologia wykonania, a w przypadku komputera jego podstawowe charakterystyki, które są ważne dla użytkownika, np. pamięć RAM, rodzaj procesora, karta graficzna czy pojemność dysku. Wymienione cechy wpływające na cenę badanych obiektów nazywane są zmiennymi hedonicznymi.

Dwufazowy hedoniczny indeks cen jest metodą szacowania przybliżonej wartości obiektu, która pozwala obliczyć indeks cen, korygujący średnią cenę dobra o cechy jakościowe, będące zmiennymi objaśniającymi w modelu hedonicznym.

5 Szerokie omówienie indeksów hedonicznych znaleźć można w pracy [Triplett 2006].

W pierwszym kroku szacuje się MNK parametry modelu regresji hedonicznej postaci (por. [Kräussl, van Elsland, 2008])[6]:

$$ln(P_{it}) = \alpha_0 + \sum_{j=1}^{k} \alpha_j \cdot x_{ijt} + \sum_{t=1}^{\tau} \beta_t \cdot z_{it} + \varepsilon_{it} \tag{11.1}$$

gdzie dla każdego i (i = 1, 2, …, m) lub (i = 1, 2, …, n), t (t = 1, 2, …, τ) oraz (j = 1, 2, …, k):
$ln(P_{it})$ – logarytm naturalny ceny i-tego obiektu sprzedanego w okresie t,
α_0, α_j, β_t – parametry regresji hedonicznej,
x_{ijt} – zmienne hedoniczne,
z_{it} – zmienne binarne opisujące okres, w jakim zarejestrowano cenę i-tego obiektu,
ε_{it} – składnik losowy modelu,
i – numer sprzedanego obiektu w okresie t – 1 (i = 1, 2, …, m) oraz w okresie t (i = 1, 2, …, n),
t – numer okresu,
j – numer zmiennej hedonicznej.

W drugim kroku wyznacza się hedoniczny indeks cen na podstawie wzoru:

$$HPI_t = \frac{NPI_t}{HQA} \tag{11.2}$$

gdzie:
HPI_t – hedoniczny indeks cen,
NPI_t – surowy, „naiwny" indeks cen (*naive price index*), postaci:

$$NPI_t = \frac{\prod_{i=1}^{n} \left(P_{it}\right)^{\frac{1}{n}}}{\prod_{i=1}^{m} \left(P_{it-1}\right)^{\frac{1}{m}}} \tag{11.3}$$

gdzie:
P_{it}, P_{it-1} – ceny obiektów odpowiednio w okresie t i t – 1,
HQA – współczynnik korygujący (*hedonic quality adjustment*), który wyznacza się na podstawie oszacowanych w modelu (11.1) parametrów jako:

$$HQA = exp\left[\sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} \frac{x_{ijt}}{n} - \sum_{i=1}^{m} \frac{x_{ijt-1}}{m}\right)\right] \tag{11.4}$$

gdzie:
$\hat{\alpha}_j$ – oceny MNK estymatorów parametrów regresji hedonicznej (11.1),
m, n – liczba obiektów sprzedanych odpowiednio w roku t – 1 i t.

6 Szacowany model jest postaci (6.11).

Warto odnotować, że oprócz modeli hedonicznych do obliczania indeksów dóbr heterogenicznych używa się również modeli uwzględniających powtórną sprzedaż (*repeat-sales regression*)[7]. Jednakże to indeksy hedoniczne umożliwiają porównanie cen dóbr w kolejnych okresach przy uwzględnieniu wszystkich sprzedanych obiektów, a zatem i tych, które pojawiają się w obrocie tylko jeden raz, co skutkuje zwiększeniem liczebności próby.

Najbardziej znanym indeksem dzieł sztuki jest Mei Moses Fine Art Index, który powstał w 2002 roku na podstawie analizy danych o wylicytowanych (przez domy aukcyjne Sotheby's i Christie's od 1950 roku) dziełach. Indeks ten jest oparty na modelu powtórnej sprzedaży i co pół roku jest poszerzany o kolejne dzieła z rynku aukcyjnego. Aktualnie jest to jeden z wielu indeksów, opisujących różne segmenty rynku dzieł sztuki. Przykładowo konstruowane są indeksy cen dla obrazów starych mistrzów, impresjonistów, sztuki nowoczesnej, rzeźb, czy indeksy obejmujące prace pojedynczych artystów, np. Picassa. W przypadku wyznaczania indeksów cen dzieł sztuki pojawia się ryzyko polegające na tym, że uzyskanie przez jedną z prac znanego artysty wysokiej ceny pociąga za sobą wzrost wartości takiego indeksu. Skutkiem tego zjawiska może okazać się nieuwzględnienie przy obliczaniu indeksu wielu transakcji dotyczących dzieł mniej znanych malarzy. Aby zniwelować ten efekt, niektórzy specjaliści stosują odrzucanie 5% najdroższych transakcji lub nieuwzględnianie niektórych transakcji. Wadami indeksów jest również to, że opierają się tylko na licytacjach domów aukcyjnych i cenie aukcyjnej, całkowicie pomijając ceny uzyskiwane przez dealerów, podczas gdy znani artyści rzadko korzystają z usług domów aukcyjnych w celu sprzedaży dzieła. Szacuje się, że dzieła sztuki przynoszą średnie zyski na poziomie 1,6% w skali roku (por. [Borowski 2007]).

11.3. Hedoniczne indeksy cen obrazów najbardziej popularnych polskich malarzy

Podstawowym atrybutem dzieła sztuki jest jego niepowtarzalność, dlatego bardzo trudno jest zmierzyć zmiany cen występujące na tym rynku, o ile nie następuje wielokrotna sprzedaż tych samych prac. Warto przy tym odnotować, że w przypadku różnych prac, nawet jednego autora, pojawia się brak porównywalności, ponieważ cena dzieła zależy od wielu subiektywnych czynników, wśród których wymienić można indywidualne cechy dzieła i jego autora, w tym okres powstania dzieła lub jego przynależność do jakiegoś cyklu prac. W ba-

7 Taką metodą zastosowano między innymi w pracy [Mei, Moses 2002].

daniu wyznaczono hedoniczne indeksy cen, ponieważ umożliwią one porównanie cen obrazów w kolejnych okresach przy uwzględnieniu wszystkich prac, również tych, które pojawiają się na rynku jako przedmiot transakcji tylko jeden raz. Zatem można przyjąć, że indeksy hedoniczne w szerszym zakresie opisują analizowany rynek sztuki niż indeksy zbudowane na podstawie powtórnej sprzedaży.

Przyjmuje się, że budując model hedoniczny, cenę obrazu można przedstawiać w zależności od takich cech jak: autor pracy oraz informacja o tym, czy wciąż tworzy (żyje), technika wykonania (np. obraz olejny), wielkość pracy, tematyka (np. pejzaż, martwa natura, portret), styl i okres reprezentowany przez pracę (np. impresjonizm, abstrakcja, symbolizm), stan dzieła (np. wymagający renowacji, w dobrym stanie), sprzedawca obiektu (np. dom aukcyjny, osoba prywatna), informacje o występowaniu sygnatury autora lub innych oznaczeniach wskazujących na autentyczność lub historię pracy.

11.3.1. Baza danych i wybór zmiennych hedonicznych

Badania realizowano na podstawie danych dotyczących obrazów sprzedanych na aukcjach dzieł sztuki, które obywały się w Polsce w latach 2007–2010. Pierwotna baza danych[8] zawiera informacje dotyczące 10 400 transakcji zrealizowanych na rynku malarstwa w latach 2007–2010 przez 41 domów aukcyjnych i fundacji mających swoją siedzibę w Polsce. Prace wystawiane na aukcjach w analizowanym okresie były dziełem 2938 (w większości polskich) twórców, reprezentujących różne okresy i style. Łączna suma transakcji wyniosła ponad 160 mln PLN, przy czym ceny poszczególnych dzieł wahały się od 30 PLN do ponad 1 mln PLN.

W tab. 11.1 przedstawiono ranking malarzy, będących autorami najdroższych prac oraz tych, których suma wszystkich sprzedanych w latach 2007–2010 dzieł osiągnęła najwyższą wartość. Jak widać, w obu grupach znaleźli się jednocześnie tylko dwaj malarze – Józef Brandt i Józef Chełmoński. Można też zauważyć, że wszyscy malarze wymienieni w tab. 10.1, z wyjątkiem Jerzego Nowosielskiego, nie żyli w czasie trwania aukcji. Dodatkowo, niemal wszyscy oni, z wyjątkiem Meli Muter, zmarli nie później niż w 1929 roku.

8 Pierwotna baza danych została opisana w pracy [Lucińska 2012], zawiera informacje dotyczące ponad 10 tys. transakcji zrealizowanych na rynku malarstwa w latach 2007–2010 przez 41 domów aukcyjnych i fundacji mających swoją siedzibę w Polsce. Prace wystawiane na aukcjach w analizowanym okresie były dziełem 2938 (w większości polskich) twórców. Pierwsze, zbudowane z wykorzystaniem tej bazy danych, indeksy hedoniczne cen malarstwa polskiego opublikowano w [Kompa, Witkowska 2013], natomiast bardziej szczegółowy opis prac, będących przedmiotem transakcji na aukcjach w latach 2007–2010 zamieszczono w [Kompa, Witkowska 2014a].

Tabela 11.1. Ranking malarzy według największej wartości transakcyjnej z pierwotnej bazy danych

Lp.	Autorzy (okres życia)	Wartość wszystkich sprzedanych na polskich aukcjach dzieł [PLN]
1	Malczewski Jacek (1854–1929)	9 401 300
2	Nowosielski Jerzy (1923–2011)	5 706 700
3	Brandt Józef (1841–1915)	4 596 000
4	Muter Mela (Mutermilch Maria Melania) (1876–1967)	4 470 700
5	Chełmoński Józef (1849–1914)	4 120 000
		Średnia wartość pojedynczych dzieł
1	Czachórski Władysław (1850–1911)	1 100 000
2	Renoir Pierre-Auguste (1841–1919)	700 000
3	Matejko Jan (1838–1893)	448 667
4	Podkowiński Władysław (1866–1895)	389 000
5	Brandt Józef (1841–1915)	383 000
6	Chełmoński Józef (1849–1914)	343 333
7	Kleinman Fryderyk (1897–1943)	310 000
8	de Laveaux Ludwik Stanisław (1868–1894)	240 000
9	Wierusz-Kowalski Alfred (1849–1915)	234 118
10	Gierymski Maksymilian (1846–1876)	227 000

Źródło: opracowanie własne.

Podstawowym zagadnieniem przy modelowaniu cen jest utworzenie bazy danych zawierającej obiekty, które utworzą „koszyk dóbr". Warto zauważyć, że nie jest możliwe uwzględnienie w jednym modelu wszystkich dzieł, co zresztą nie byłoby poprawnym działaniem z uwagi na ogromne zróżnicowanie prac, będących przedmiotem transakcji. Dlatego do dalszych analiz utworzono poprzez arbitralny dobór obiektów próbę badawczą. Oczywiście już sama kwestia, których twórców uznać za polskich malarzy może być dyskusyjna, nie ma bowiem jednoznacznych kryteriów takiej klasyfikacji[9], zwłaszcza jeśli artyści żyli w okresie, kiedy Polska pozbawiona była państwowości. Dodatkowo można podać w wątpliwość czy praca, którą na aukcji sprzedano za 30 PLN, jest dziełem sztuki, jeśli nie jest się ekspertem w tej dziedzinie. Dlatego do próby badawczej wybrano malarzy, dla których licz-

9 Może być to kryterium narodowościowe, miejsce urodzenia lub zamieszkania, deklaracji artysty albo tematyki prac. Przykładowo Mela Muter urodziła się w Warszawie przed uzyskaniem przez Polskę niepodległości, ale zmarła w Paryżu i była narodowości żydowskiej. W opisach uznawana jest za polską malarkę żydowskiego pochodzenia. Innym przykładem może być Nikifor Krynicki (właściwie Epifaniusz Drowniak), który całe życie mieszkał w Polsce, ale był Łemkiem.

ba sprzedanych dzieł w latach 2007–2010 wyniosła przynajmniej 40 prac, a minimalna średnia cena pracy była większa niż 2400 PLN. W rezultacie powstała baza zawierająca 750 dzieł (7,2% wszystkich prac znajdujących się w pierwotnej bazie danych), wykonanych przez 11 polskich malarzy, które zostały sprzedane za 26,76 mln PLN (16,2% całkowitych obrotów w ciągu analizowanych czterech lat). Lista malarzy wraz z podstawowymi informacjami przedstawiona została w tab. 11.2.

Tabela 11.2. Lista artystów, których prace stały się obiektami realizowanych badań

Lp.	Malarze wybrani do badania	Liczba [szt.]	Wartość [PLN]	Rok urodzenia	Rok śmierci
		sprzedanych prac		autora prac	
1	Chmieliński (Stachowicz) Władysław	55	648 200	1911	1979
2	Dominik Tadeusz*	46	608 000	1928	2014
3	Dwurnik Edward*	63	431 300	1943	2018
4	Erb Erno	58	816 500	1890	1943
5	Hofman Wlastimil (właśc. Hofmann Vlastimil)	85	1 817 050	1881	1970
6	Kossak Jerzy	91	1 261 000	1886	1955
7	Kossak Wojciech	60	2 027 500	1856	1942
8	Malczewski Jacek	71	9 401 300	1854	1929
9	Nikifor (albo Nikifor Krynicki) (właśc. Drowniak Epifaniusz)	79	196 400	1895	1968
10	Nowosielski Jerzy*	81	5 706 700	1923	2011
11	Wyczółkowski Leon	61	3 848 300	1852	1936
	Suma	750	26 762 250	×	×

Uwaga: * oznaczono malarzy żyjących w latach, kiedy odbywały się aukcje ich prac.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014a; 2014b].

Większość wybranych do próby artystów urodziła się w XIX wieku i tylko dwóch z nich żyło i tworzyło w momencie realizacji aukcji, choć najmłodszy z nich Edward Dwurnik urodził się w 1943 roku (zmarł w 2018 roku). Wszyscy wyróżnieni malarze mieli w momencie realizacji badań ugruntowaną pozycję na rynku, co zapewne jest przyczyną tak dużego zainteresowania ich pracami na aukcjach.

Mając skompletowaną listę obiektów badania, należy dokonać wyboru cech hedonicznych. Wprawdzie każda praca wystawiana na aukcji jest opisana przez wystawiającego za pomocą kilku (czasem kilkudziesięciu) atrybutów, ale zbiory charakterystyk uwzględniane w opisach przez organizatorów aukcji mogą się różnić. W związku z tym do dalszych badań wybrano wspólne dla wszystkich obiektów cechy charakteryzujące analizowane dzieła sztuki.

Zmienną objaśnianą we wszystkich konstruowanych modelach jest logarytm naturalny z ceny sprzedaży [PLN] i jest to, poza powierzchnią obrazu, jedyna cecha ilościowa. Pozostałe cechy mają charakter jakościowy i zostały zakodowane jako zmienne binarne. Wyróżnionymi w badaniach zmiennymi hedonicznymi są:

1) *malarz* – zmienna zero-jedynkowa oznaczająca autora pracy, identyfikowanego nazwiskiem (dodatkowo, jeśli potrzeba, inicjałem imienia), przy czym cechą referencyjną jest WYCZOLKOWSKI, co oznacza, że interpretacja parametrów stojących przy pozostałych zmiennych, reprezentujących poszczególnych twórców, przeprowadzona będzie w odniesieniu do prac autorstwa Leona Wyczółkowskiego;
2) *dom aukcyjny* – zmienna binarna identyfikowana skrótem nazwy instytucji organizującej aukcję, wariant referencyjny to: INNE DOMY;
3) *technika* – zmienna zero-jedynkowa oznaczająca technikę wykonania obrazu, a wariant referencyjny to: INNE TECHNIKI;
4) *rok* – zmienna zero-jedynkowa, która przyjmuje wartość 1 dla kolejnych lat, w jakich odbywały się transakcje (aukcje): ROK_1 oznacza obserwacje z 2007 roku, ROK_2 – obserwacje z 2008 roku, ROK_3 – obserwacje z 2009 roku oraz ROK_4 – obserwacje z 2010 roku, ten ostatni wariant stanowi zmienną odniesienia;
5) *powierzchnia* – cecha ilościowa wyrażająca powierzchnię obrazu w [cm^2];
6) *sygnatura* – zmienna dychotomiczna, która informuje, czy obraz jest podpisany przez autora ($X_i = 1$ gdy obraz jest sygnowany, $X_i = 0$ w przeciwnym przypadku);
7) *zgon* – zmienna dychotomiczna informująca o tym, czy autor dzieła żyje w momencie realizacji aukcji ($x_i = 0$ dla żyjących artystów, $x_i = 1$ w przeciwnym przypadku);
8) *wartość* – zmienna dychotomiczna informująca o tym, czy cena oferowana przez kupujących nie jest mniejsza od ceny wywoławczej ($x_i = 0$, gdy wyższa jest cena oferowana, $x_i = 1$ w przeciwnym przypadku). Wprowadzenie tej zmiennej podyktowane jest faktem występowania tzw. sprzedaży warunkowej w sytuacji, kiedy nie zostanie osiągnięta cena wywoławcza. Innymi słowy, w takim przypadku może w ogóle nie dojść do transakcji.

W badaniach przeanalizowano podstawowe charakterystyki cen prac będących dziełami wyróżnionych malarzy, które zostały sprzedane na aukcjach w latach 2007–2010. Wyniki analiz przeprowadzonych dla podstawowych zmiennych hedonicznych przedstawiono w tab. 11.3–11.4. Jak można zauważyć, szeregi osiągniętych na aukcjach cen sprzedaży są mocno zróżnicowane, co przedstawiają zarówno dane pogrupowane według autorów prac, jak i domów aukcyjnych oraz według techniki wykonania.

Z grona wybranych do dalszych badań twórców największą liczbę sprzedanych prac odnotowano dla Jerzego Kossaka, Wlastimila Hofmana i Jerzego Nowosielskiego. Z kolei największą średnią wartość dla kupujących przedstawiały prace

Jacka Malczewskiego, na drugim miejscu był Jerzy Nowosielski, a na trzecim Leon Wyczółkowski. Natomiast najniższe średnie ceny pojedynczej pracy zaobserwowano dla malarza prymitywisty Nikifora Krynickiego. Z kolei największe zróżnicowanie cen osiągniętych na aukcjach odnotowano dla Leona Wyczółkowskiego (173% średniej), Jacka Malczewskiego i Wojciecha Kossaka (odpowiednio 118 i 115% średniej). Dla wymienionych malarzy odnotowano najwyższe współczynniki asymetrii. Najmniejsze zróżnicowanie cen obserwuje się w przypadku prac Nikifora oraz Erno Erba (41 i 47% średniej ceny).

Tabela 11.3. Podstawowe parametry opisowe cen dzieł sztuki wg zmiennej oznaczającej wyróżnionych malarzy

Zmienna: *malarz*	Średnia	Odchylenie standardowe	Współczynnik zmienności	Skośność	Kurtoza
CHMIELINSKI	11 785	6425	0,55	0,36	–0,73
DOMINIK	13 217	7499	0,57	0,75	1,49
DWURNIK	6846	5823	0,85	1,75	3,24
ERB	14 078	6582	0,47	1,05	2,28
HOFMAN	21 377	17 286	0,81	2,32	9,33
KOSSAK_J	13 857	11 050	0,80	3,00	12,28
KOSSAK_W	33 792	38 703	1,15	3,57	17,08
MALCZEWSKI	132 413	156 276	1,18	1,24	0,47
NIKIFOR	2486	1021	0,41	2,01	7,29
NOWOSIELSKI	70 453	65 808	0,93	1,09	0,80
WYCZOLKOWSKI	63 087	108 969	1,73	3,63	13,54

Uwaga: pogrubiono wariant referencyjny cechy.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014a; 2014b].

Biorąc pod uwagę dzieła sprzedane przez poszczególne domy aukcyjne, należy podkreślić, że spośród 9 wyróżnionych, najwięcej prac sprzedał dom aukcyjny REMPEX, a następnie AGRA-ART, RYNEK SZTUKI i DESA UNICUM. Natomiast najmniej sprzedanych (spośród 750 uwzględnionych w badaniu) dzieł odnotowano dla domu aukcyjnego AUKCJE ON-LINE, co oznacza, że uwzględnione w referencyjnym wariancie tej zmiennej inne domy aukcyjne sprzedały mniej obrazów niż 7. Najdroższe prace sprzedano za pośrednictwem domu aukcyjnego DESA UNICUM, a następnie POLSWISS ART oraz AGRA-ART i OKNA SZTUKI, ale już o znacznie niższych cenach. DESA UNICUM i POLSWISS ART charakteryzują się też największym zróżnicowaniem cen osiągniętych na aukcjach, a największą skośnością – DESA.

Tabela 11.4. Podstawowe parametry opisowe cen dzieł sztuki wg zmiennej domy aukcyjne i techniki

Zmienna: *dom aukcyjny*	Średnia	Liczba prac	Odchylenie standardowe	Współczynnik zmienności	Skośność	Kurtoza
AGRAART	48 627	220	111 443	2,29	5,82	42,54
AUKCJE_ON_LINE	3057	7	1513	0,49	–0,05	–1,80
DESA	23 825	61	65 896	2,77	7,35	56,06
DESA_UNICUM	115 866	105	241 391	2,08	4,10	19,45
OKNA_SZTUKI	44 665	20	57 487	1,29	1,42	1,01
OSTOYA	13 061	50	11 816	0,90	2,81	10,71
POLSW	87 564	73	126 193	1,44	2,32	6,08
REMPEX	21 948	270	34 653	1,58	4,28	22,67
RYNEK_SZTUKI	3385	114	6885	2,03	3,12	10,32
INNE DOMY	4044	48	3894	0,96	1,84	2,95
Zmienna: *technika*						
AKRYL	13 408	53	28 499	2,13	3,49	13,33
AKWARELA	9370	148	13 939	1,49	3,03	10,71
GWASZ	18 056	53	17 645	0,98	0,94	-0,24
OLEJ	54 890	596	135 080	2,46	6,22	51,57
OLOWEK	8920	15	8621	0,97	1,93	4,48
PASTEL	47 627	33	104 314	2,19	5,07	27,46
TEMPERA	27 431	16	28 519	1,04	1,80	3,80
TUSZ	13 033	9	8184	0,63	–0,04	–0,96
INNE TECHNIKI	16 725	45	32 772	1,96	3,79	17,78

Uwaga: pogrubiono warianty referencyjne cech.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014a; 2014b].

Kolejną zmienną hedoniczną jest technika wykonania. Jak można zauważyć, zdecydowanie najwięcej (spośród uwzględnionych w badaniu prac) sprzedano obrazów olejnych, a następnie akwareli. Natomiast najmniej prac wykonanych tuszem, zatem pozostałe techniki, reprezentowane przez wariant referencyjny tej zmiennej, zostały wykorzystane rzadziej niż liczba dzieł wykonanych tuszem (tj. mniej niż 9). Najdroższe prace to obrazy olejne, pastele i wykonane temperą. Dzieła wykonane za pomocą dwóch pierwszych technik charakteryzują się największym zróżnicowaniem i skośnością.

11.3.2. Modele regresji hedonicznej

Wykorzystanie modeli hedonicznych do określania korekty jakościowej indeksu cen wymaga rozstrzygnięcia sposobu oceny modeli. Z dotychczasowych rozważań wynika, że nawet po dokonaniu wyboru próby estymacyjnej konstruktor modelu musi zdecydować o wyborze zmiennych objaśniających, zwłaszcza wtedy, kiedy istnieje znaczna ich różnorodność. Tworzy się wtedy zmienną referencyjną „inne warianty", która np. dla zmiennej *wystawca* obejmuje 5% i *technika* – 2% obiektów przy najbardziej rozbudowanych wersjach tych zmiennych. Pojawia się zatem pytanie, co zrobić w sytuacji, gdy niektóre (lub wszystkie) warianty danej zmiennej są statystycznie nieistotne. Pozostaje zazwyczaj redukcja liczby wariantów i wzbogacanie zmiennej referencyjnej „inne warianty" lub w ostateczności usunięcie całej zmiennej.

Innym problemem pojawiającym się przy modelowaniu cen dzieł sztuki jest współliniowość zmiennych, co wynika z faktu, że znakomita większość zmiennych hedonicznych jest kodowanymi binarnie cechami jakościowymi. W takim przypadku wartości parametrów stojących przy danym wariancie zmiennej zero-jedynkowej (dychotomicznej lub wielodzielnej) interpretowane są w odniesieniu do wariantu referencyjnego tej zmiennej (tj. tego wariantu zmiennej, który nie jest widoczny w modelu). Należy zauważyć, że lista dostępnych zmiennych hedonicznych (tzn. danych z aukcji niosących informacje o uwzględnionych w badaniu pracach) jest w zasadzie zamknięta i można jedynie tworzyć sztuczne zmienne, które opisują np. reputację autora lub wystawcy dzieła i wprowadzać je w miejsce nieistotnych zmiennych binarnych. W analizach uzupełniono listę zmiennych hedonicznych o dodatkowo utworzone zmienne charakteryzujące cenę obrazu i okres powstania dzieła.

Utworzona na potrzeby tego badania zmienna *klasa* odzwierciedla wartość wystawionego na aukcji obrazu. Utworzono cztery klasy cenowe, uwzględniając średnie ceny wyznaczone dla poszczególnych twórców: KLASA_1 to obrazy najdroższe w cenie powyżej 73 tys. PLN, KLASA_2 – dzieła w cenie od 16 750 PLN do 73 000 PLN, KLASA_3: od 5817 PLN do 16 749 PLN oraz KLASA_4, zawierająca najniżej wycenione prace, tj. poniżej 5816 PLN i stanowiąca wariant odniesienia. Kolejną, subiektywnie zdefiniowaną zmienną jest *epoka*, za pomocą której dokonano klasyfikacji twórców na podstawie roku urodzenia. Wyróżniono dwa warianty tej zmiennej przyjmującej wartość 1 dla malarzy urodzonych przed 1900 rokiem i 0 w pozostałych przypadkach, co jest oczywiście dyskusyjne.

W realizowanych badaniach zbudowano wiele rożnych modeli hedonicznych, uwzględniających zarówno omówione wcześniej zmienne, jak i zmienne dodatkowe [Kompa, Witkowska 2013, 2014b; Witkowska 2014, 2016a]. Tutaj pokazanych zostanie 16 przykładowych modeli (oznaczonych w dalszych rozważaniach symbolami M1–M16), zawierających omówione wcześniej zmienne hedoniczne. Modele te różnią się specyfikacją i zawierają od 22 do 36 zmiennych, co pokazano w tab. 11.5–11.6, które dodatkowo zawierają podstawowe charakterystyki tych oszacowanych modeli. W przypadku cech wielodzielnych obok oznaczenia

występowania danej zmiennej w modelu (co oznaczono przez +) podano w nawiasach liczbę wariantów zmiennych użytych w konkretnej specyfikacji. Zmniejszenie liczby wariantów danej zmiennej oznacza, że pominięte warianty weszły w skład zmiennej referencyjnej. Szczegółowa specyfikacja modeli została pokazana w tab. 11.7–11.12, w których podano oceny estymatorów parametrów.

Tabela 11.5. Porównanie specyfikacji modeli hedonicznych i ich charakterystyk

Zmienne wielodzielne	Oznaczenia modeli							
	M1	M2	M3	M4	M5	M6	M7	M8
Rok	+ (3)			+ (3)				+ (3)
Dom aukcyjny	+ (8)	+ (8)	+ (5)	+ (8)	+ (5)	+ (5)	+ (5)	+ (8)
Malarz	+ (10)	+ (10)	+ (10)	+ (10)	+ (10)	+ (10)	+ (10)	+ (10)
Technika	+ (8)	+ (8)	+ (6)	+ (8)	+ (4)	+ (2)	+ (2)	+ (8)
Klasa	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	
Zmienna ilościowa i zmienne dychotomiczne								
Powierzchnia	+	+	+	+	+	+	+	+
Sygnatura	+	+	+	+	+	+		+
Wartość	+	+	+	+	+	+		+
Zgon	+	+	+		+	+	+	
$\bar{R}^2$	0,9342	0,9343	0,9329	0,9338	0,9330	0,9330	0,9329	0,8114
F	296,40	324,11	374,30	374,30	402,01	435,35	471,65	101,68
Liczba zmiennych	36	33	28	35	26	24	22	32
Kryterium Akaikego	482,96	478,25	487,42	486,07	487,26	485,45	483,80	1269,1
D-W	1,8206	1,8207	1,7823	1,9284	1,7853	1,7816	1,7861	1,9311

Uwaga: w nawiasach podano liczbę wariantów cechy.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Analizując własności modeli (tab. 11.5–11.6) należy stwierdzić, że wszystkie dość dobrze opisują zlogarytmowane ceny obrazów, najniższym bowiem skorygowanym współczynnikiem determinacji (wyznaczonym według wzoru (4.36) i oznaczonym w tabelach jako $\bar{R}^2$) charakteryzują się modele M16 z $R^2 = 0{,}55$ oraz M8 i M9, dla których wynosi on ponad 81%. Również wartości statystyki F[10] są

10 Statystyka testowa Fishera-Snedecora F nie była omawiana w niniejszym opracowaniu. W ocenie jakości modeli ekonometrycznych wykorzystuje się ją w celu sprawdzenia łącznego wpływu wszystkich wyróżnionych zmiennych objaśniających na zmienną objaśnianą (por. [Witkowska 2012, s. 102–103]).

we wszystkich modelach bardzo wysokie, co świadczy o łącznej istotności zmiennych, chociaż pojedyncze zmienne (lub ich warianty) mogą być w poszczególnych modelach statystycznie nieistotne. Zróżnicowanie jakości modeli jest najbardziej widoczne z punktu widzenia kryterium informacyjnego Akaikego[11], które jest dla większości modeli dodatnie, a w przypadku modeli M8 i M9 wyjątkowo wysokie. Jedynie modele M13, M14 i M15 charakteryzują się ujemną wartością tej miary i dodatkowo odznaczają się najwyższym skorygowanym R^2 wynoszącym ponad 99%. Jednakże modele te cechuje dodatnia autokorelacja reszt, o czym świadczą wartości statystyki Durbina-Watsona[12] (D-W). Z kolei w przypadku modeli M8 i M9 można uznać, że autokorelacja składnika losowego nie występuje.

Tabela 11.6. Porównanie specyfikacji modeli hedonicznych i ich charakterystyk

Zmienne wielodzielne	Oznaczenia modeli							
	M9	M10	M11	M12	M13	M14	M15	M16
Rok	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)	+ (3)
Dom aukcyjny	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)
Malarz	+ (10)				+ (10)	+ (10)	+ (10)	
Technika	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)	+ (8)
Klasa		+ (3)	+ (3)	+ (3)				
Zmienna ilościowa i zmienne dychotomiczne								
Powierzchnia	+	+	+	+	**+**	**+**	**+**	+
Sygnatura	+	+	+	+	+	+	+	+
Wartość		+		+	+		+	
Zgon		+	+	+			+	+
Epoka				+				+
$\bar{R}^2$	0,8115	0,9165	0,9207	0,9209	0,9953	0,9953	0,9953	0,5507
F	105,07	317,30	335,58	312,25	4910,11	5071,98	4910,11	40,91
Liczba zmiennych	31	26	25	27	32	31	33	23
Kryterium Akaikego	1267,3	651,88	613,17	613,83	–1492,91	–1494,38	–1492,91	1911,41
D-W	1,9359	1,7314	1,7032	1,7171	1,4808	1,4780	1,4808	0,9174

Uwaga: szare pola w przypadku zmiennej powierzchni oznaczają, że do modelu wprowadzono kwadrat jej logarytmu; w nawiasach podano liczbę wariantów cechy.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

11 Kryterium informacyjne Akaikego służy do wyboru najlepszego modelu spośród kilku opisujących tę samą zmienną objaśnianą. Przy czym najlepszy jest model o najniższej wartości kryterium informacyjnego.

12 Jednym z założeń Gaussa-Markowa jest brak autokorelacji reszt modelu ekonometrycznego (6.6), które weryfikuje się m.in. za pomocą testu Durbina-Watsona (por. np. [Witkowska 2012, s. 122–125]).

Podejmując się interpretacji uzyskanych ocen estymatorów parametrów, na podstawie modelu M1 (tab. 11.7), będącego najbogatszą wersją modelu, można stwierdzić, że AGRA-ART, OKNA SZTUKI, OSTOYA, REMPEX i RYNEK SZTUKI uzyskiwały niższe ceny ze sprzedaży obrazów niż nieuwzględnione w modelu domy aukcyjne. Spośród 11 malarzy jedynie obraz Malczewskiego uzyskał wyższą cenę niż Wyczółkowskiego. Obrazy akrylowe, akwarele, gwasze, olejne, pastele i te malowane temperą miały wyższe ceny niż te przygotowane z pomocą nieuwzględnionych w modelu technik. Zgon malarza wpływa dodatnio na cenę dzieła, podobnie jak wielkość obrazu mierzona jego powierzchnią.

Tabela 11.7. Oszacowane modele

Zmienne	Parametry M1	Współczynnik M2	Parametry M3	Parametry M4
const	5,6804 ***	5,7067 ***	5,7731 ***	5,8257 ***
ROK_1	0,0075			–0,0018
ROK_2	0,0231			0,0076
ROK_3	0,0374			0,0092
AGRAART	–0,1599 **	–0,1544 **	–0,0890 **	–0,1674 **
DESA	–0,1151	–0,1118		–0,1283
DESA_UNICUM	–0,1211	–0,1064		–0,1292
OKNA_SZTUKI	–0,2795 ***	–0,2726 **	–0,2191 **	–0,2576 **
OSTOYA	–0,1904 **	–0,1856 **	–0,1272 **	–0,2022 **
POLSW	0,0460	0,0553		0,0409
REMPEX	–0,2252 ***	–0,2208 ***	–0,1546 ***	–0,2318 ***
RYNEK_SZTUKI	–0,2733 ***	–0,2615 ***	–0,1910 ***	–0,2749 ***
KOSSAK_J	–0,5372 ***	–0,5405 ***	–0,5014 ***	–0,5352 ***
KOSSAK_W	–0,3312 ***	–0,3343 ***	–0,3112 ***	–0,3254 ***
CHMIELINSKI	–0,4340 ***	–0,4331 ***	–0,3964 ***	–0,4334 ***
DWURNIK	–0,9170 ***	–0,9180 ***	–0,8372 ***	–0,9172 ***
ERB	–0,3546 ***	–0,3565 ***	–0,3348 ***	–0,3498 ***
HOFMAN	–0,3503 ***	–0,3532 ***	–0,3203 ***	–0,3498 ***
MALCZEWSKI	0,1506 **	0,1509 **	0,1626 **	0,1490 **
NIKIFOR	–0,4730 ***	–0,4762 ***	–0,4680 ***	–0,4698 ***
NOWOSIELSKI	–0,1383 **	–0,1426 **	–0,0852	–0,1412 **
DOMINIK	–0,7095 ***	–0,7103 ***	–0,6552 ***	–0,7115 ***
Sygnatura	0,0641	0,0606	0,0646	0,0654
AKWARELA	0,2566 ***	0,2537 ***	0,1592 **	0,1709 *
AKRYL	0,2798 **	0,2730 **	0,1781 *	0,2835 **

Zmienne	Parametry M1	Współczynnik M2	Parametry M3	Parametry M4
GWASZ	0,2364 **	0,2375 **	0,1499	0,2014 *
OLEJ	0,3449 ***	0,3465 ***	0,2296 ***	0,3514 ***
OLOWEK	0,0414	0,0463		0,0522
PASTEL	0,2363 **	0,2363 **	0,1472	0,2423 **
TEMPERA	0,4014 ***	0,4080 ***		0,4088 ***
TUSZ	–0,1522	–0,1486	–0,2435 *	–0,1408
WARTOŚĆ	–0,0404	–0,0419	–0,0341	–0,0391
KLASA_1	2,9213 ***	2,9186 ***	2,9446 ***	2,9247 ***
KLASA 2	1,5779 ***	1,5787 ***	1,5911 ***	1,5806 ***
KLASA_3	0,84983 ***	0,8472 ***	0,8476 ***	0,8497 ***
Powierzchnia	0,2326 ***	0,2322 ***	0,2258 ***	0,2314 ***
Zgon	0,1265 **	0,1214 **	0,1358 **	
$\bar{R}^2$	0,9342	0,9343	0,9329	0,9338

Uwaga: * oznacza istotność na poziomie 0,1; ** – na poziomie 0,05 i *** – 0,01.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Analizując poszczególne modele, należy stwierdzić, że w przypadku modeli M1 i M4 nieistotne były zmienne *rok* (wszystkie warianty), *sygnatura* i *wartość* oraz niektóre warianty zmiennych opisujących *malarzy*, *domy aukcyjne* i *technikę*. W modelach M2–M3 i M5–M6 (tab. 11.7–11.8) pominięto pierwszą z wymienionych nieistotnych zmiennych, ale pozostawiono *sygnaturę* i *wartość*. Przy czym w kolejnych modelach usuwano pojedyncze warianty zmiennych *dom aukcyjny* i *technika*, które były nieistotne. Oprócz tego *rok*, *sygnatura* i *wartość* usunięto w modelu M7, natomiast model M4, w porównaniu z modelem M1 różni się brakiem zmiennej *zgon*.

Tabela 11.8. Oszacowane modele

Zmienne	Parametry M5	Parametry M6	Parametry M7
const	5,8385 ***	5,8752 ***	5,9258 ***
AGRAART	–0,0943 ***	–0,0984 ***	–0,0844 **
OKNA_SZTUKI	–0,2115 **	–0,1987 **	–0,1928 **
OSTOYA	–0,1274 **	–0,1339 **	–0,1238 **
REMPEX	–0,1531 ***	–0,1548 ***	–0,1606 ***
RYNEK_SZTUKI	–0,1854 ***	–0,1829 ***	–0,1889 ***
KOSSAK_J	–0,5182 ***	–0,4981 ***	–0,4951 ***
KOSSAK_W	–0,3283 ***	–0,3088 ***	–0,3076 ***

Tabela 11.8 (cd.)

Zmienne	Parametry M5		Parametry M6		Parametry M7	
CHMIELINSKI	–0,4039	***	–0,3806	***	–0,3796	***
DWURNIK	–0,8548	***	–0,8314	***	–0,8369	***
ERB	–0,3510	***	–0,3305	***	–0,3209	***
HOFMAN	–0,3381	***	–0,3194	***	–0,3127	***
MALCZEWSKI	0,1453	**	0,1596	**	0,16178	**
NIKIFOR	–0,4618	***	–0,4467	***	–0,4895	***
NOWOSIELSKI	–0,1204	*	–0,1115	*	–0,1204	**
DOMINIK	–0,6730	***	–0,6434	***	–0,6467	***
Sygnatura	0,0621		0,0634			
AKWARELA	0,0688					
AKRYL	0,1075					
OLEJ	0,1569	***	0,1172	***	0,11304	**
TUSZ	–0,3083	**	–0,3355	***	–0,3405	***
Wartość	–0,0367		–0,0350			
KLASA 1	2,9536	***	2,95941	***	2,9617	***
KLASA_2	1,5968	***	1,5983	***	1,6021	***
KLASA 3	0,8494	***	0,8486	***	0,8558	***
Powierzchnia	0,2289	***	0,2299	***	0,2297	***
Zgon	0,1231	**	0,0940	**	0,0930	**
$\bar{R}^2$	0,9330		0,9330		0,9329	

Uwaga: * oznacza istotność na poziomie 0,1; ** – na poziomie 0,05 i *** – 0,01.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Modele M8 i M9 (tab. 11.9) zostały pozbawione zmiennej *klasa*, przy czym jeden z wariantów zmiennej *rok* stał się istotny na poziomie 0,05, ale zmienne *sygnatura* i *wartość* pozostały w dalszym ciągu statystycznie nieistotne. *Sygnatura* jest istotną zmienną w modelach M10–M12 (tab. 11.10), w których pominięto zmienne charakteryzujące poszczególnych twórców. Chociaż w podobnie skonstruowanym modelu M16 nie było podstaw do odrzucenia hipotezy o zerowej wartości parametru stojącego przy tej zmiennej. Natomiast w wymienionych modelach *wartość* i *epoka* nie są istotne w żadnym modelu. Z kolei modele M13–M15 charakteryzują się tym, że powierzchnia obrazu jest w modelu uwzględniona jako kwadrat logarytmu powierzchni.

Reasumując, należy stwierdzić, że spośród uwzględnionych w badaniach zmiennych nieistotną we wszystkich modelach okazała się zmienna *wartość* i *epoka* (oszacowano tylko jeden model z tą zmienną), a także w większości przypadków zmienna *rok*. Można też wskazać na niektóre warianty pozostałych zmiennych, które były nieistotne.

Tabela 11.9. Oszacowane modele

Zmienne	Parametry M8		Parametry M9		Parametry M16	
const	2,7877	***	2,7934	***	1,0188	*
ROK_1	0,0934		0,0971		0,9419	***
ROK_2	0,0758	**	0,0768	**	0,2479	**
ROK_3	0,0145		0,0154		0,0270	
AGRAART	0,2945	**	0,2989	**	0,8689	***
DESA	0,1990		0,2013		0,6997	***
DESA_UNICUM	0,4084	***	0,4026	***	1,2088	***
OKNA_SZTUKI	0,4798	***	0,4816	***	1,1270	***
OSTOYA	0,0998		0,1032		0,2577	
POLSW	0,8052	***	0,7968	***	1,5653	***
REMPEX	0,0895		0,0807		0,4768	***
RYNEK_SZTUKI	0,0172		0,0080		0,2697	
KOSSAK_J	–1,5906	***	–1,5896	***		
KOSSAK_W	–0,8769	***	–0,8780	***		
CHMIELINSKI	–1,2274	***	–1,2266	***		
DWURNIK	–2,2824	***	–2,2810	***		
ERB	–1,0908	***	–1,0864	***		
HOFMAN	–1,0883	***	–1,0862	***		
MALCZEWSKI	0,3115	***	0,3125	***		
NIKIFOR	–1,3326	***	–1,3319	***		
NOWOSIELSKI	–0,1185		–0,1186			
DOMINIK	–1,9053	***	–1,9050	***		
Sygnatura	–0,0435		–0,0457		0,1118	
AKWARELA	0,1968		0,1991		-0,4595	**
AKRYL	0,6975	***	0,6998	***	–0,2321	
GWASZ	0,2849		0,2918		–0,3334	
OLEJ	0,8856	***	0,8869	***	0,2444	
OLOWEK	–0,2460		–0,2453		–0,4840	
PASTEL	0,4502	**	0,4525	**	0,5824	**
TEMPERA	0,6350	***	0,6377	***	0,5159	
TUSZ	–0,5984	**	–0,5986	**	–0,5053	
Powierzchnia	0,5646	***	0,5636	***	0,5027	***
Wartość	–0,0273					
Zgon					0,3548	***
EPOKA_1					0,9771	***
$\bar{R}^2$	0,8114		0,8115		0,5507	

Uwaga: * oznacza istotność na poziomie 0,1; ** – na poziomie 0,05 i *** – 0,01.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Tabela 11.10. Oszacowane modele

Zmienne	Parametry M10		Parametry M11		Parametry M12	
const	6,3626	***	5,5577	***	5,4910	***
ROK_1	–0,0343		–0,0206		–0,0279	
ROK_2	–0,0034		0,0035		0,0017	
ROK_3	0,0185		0,0136		0,0119	
AGRAART	–0,1679	**	–0,2437	***	–0,2415	***
DESA	–0,1112		–0,1822	**	–0,1996	**
DESA_UNICUM	–0,1148		–0,1989	**	–0,1947	**
OKNA_SZTUKI	–0,3267	***	–0,3944	***	–0,4000	***
OSTOYA	–0,2513	***	–0,3061	***	–0,3026	***
POLSW	–0,0130		–0,0976		–0,0805	
REMPEX	–0,2337	***	–0,3160	***	–0,3127	***
RYNEK_SZTUKI	–0,2915	***	–0,3719	***	–0,3677	***
Sygnatura	0,1474	***	0,1615	***	0,1605	***
AKWARELA	0,0436		0,0809		0,0831	
AKRYL	0,0090		0,0926		0,0951	
GWASZ	0,0645		0,1046		0,0967	
OLEJ	0,0955		0,1167		0,1125	
OLOWEK	0,1458		0,0733		0,1006	
PASTEL	0,1968		0,2292	*	0,2102	*
TEMPERA	0,2989	**	0,2512	*	0,2794	**
TUSZ	0,0238		–0,0792		–0,0586	
Powierzchnia	0,1343	***	0,1945	***	0,2011	***
Wartość	–0,0361				–0,0382	
KLASA_1	3,7026	***	3,5083	***	3,5223	***
KLASA_2	1,9900	***	1,8907	***	1,8925	***
KLASA_3	1,0717	***	1,0170	***	1,0184	***
Zgon	0,1905	***	0,3879	***	0,3442	***
EPOKA_1					0,0580	
$\bar{R}^2$	0,9165		0,9207		0,9209	

Uwaga: * oznacza istotność na poziomie 0,1; ** – na poziomie 0,05 i *** – 0,01.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Zauważyć należy, że model M15 ma identyczne oceny estymatorów parametrów oraz statystyki opisowe co model M13. Jest to związane ze specyfikacją obu modeli – z powodu współliniowości w M13 pominięto zmienną *zgon*, a w modelu

M15, pomijając wyraz wolny, wprowadzono brakującą zmienną. Mimo identycznych parametrów obu modeli, korekty hedoniczne znacząco się różnią, co wpływa na zróżnicowanie indeksów cen (tab. 11.12).

Tabela 11.11. Oszacowane modele

Zmienne	Parametry M13	Parametry M14	Parametry M15
const	5,0405 ***	5,0400 ***	
ROK_1	0,0067	0,0075	0,0067
ROK_2	0,0007	0,0010	0,0007
ROK_3	0,0019	0,0021	0,0019
AGRAART	0,0834 **	0,0843 ***	0,0834 **
DESA	0,0844	0,0848 ***	0,0844
DESA_UNICUM	0,0550 ***	0,0536 **	0,0550 ***
OKNA_SZTUKI	0,0701 **	0,0707 **	0,0701 **
OSTOYA	0,0642 ***	0,0649 ***	0,0642 ***
POLSW	0,0717 ***	0,0697 ***	0,0717 ***
REMPEX	0,0606 ***	0,0584 ***	0,0606 ***
RYNEK_SZTUKI	0,0504 **	0,0481 **	0,0504 **
KOSSAK_J	–0,0566 ***	–0,0565 ***	–0,0566 ***
KOSSAK_W	–0,0318	–0,0321 *	–0,0318
CHMIELINSKI	–0,0601 ***	–0,0600 ***	–0,0601 ***
DWURNIK	–0,1413 ***	–0,1413 ***	–0,1413 ***
ERB	–0,0420 **	–0,0409 **	–0,0420 **
HOFMAN	–0,0484 **	–0,0479 **	–0,0484 **
MALCZEWSKI	–0,1007 ***	–0,1003 ***	–0,1007 ***
NIKIFOR	–0,2556 ***	–0,2553 ***	–0,2556 ***
NOWOSIELSKI	–0,0471 ***	–0,0471 ***	–0,0471 ***
DOMINIK	–0,0594 ***	–0,0596 ***	–0,0594 ***
Sygnatura	–0,0038	–0,0043	–0,0038
AKWARELA	–0,0155	–0,0149	–0,0155
AKRYL	0,0448	0,0452	0,0448
GWASZ	–0,0038	–0,0021	–0,0038
OLEJ	0,0561 **	0,0563 **	0,0561 **
OLOWEK	–0,0704 **	–0,0701 **	–0,0704 **
PASTEL	0,0336	0,0341	0,0336
TEMPERA	0,0296	0,0303	0,0296
TUSZ	–0,0171	–0,0172	–0,0171

Tabela 11.11 (cd.)

Zmienne	Parametry M13	Parametry M14	Parametry M15
POWIERZCHNIA^2	0,0484 ***	0,0484 ***	0,0484 ***
WARTOŚĆ	–0,0065		–0,0065
Zgon			5,0405 ***
$\bar{R}^2$	0,9953	0,9953	0,9953

Uwaga: * oznacza istotność na poziomie 0,1; ** – na poziomie 0,05 i *** – 0,01.

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

11.3.3. Hedoniczne indeksy cen

Kolejnym etapem badania było wyznaczenie korekt hedonicznych (11.4) oraz indeksów cen dla prac uwzględnionych w badaniu. W tab. 11.12 przedstawiono surowe indeksy cen wyznaczone według wzoru (11.3), które nie uwzględniają korekty hedonicznej, a także indeksy hedoniczne (11.2), wyznaczone dla wybranych postaci modeli hedonicznych. Indeksy zostały wyliczone zarówno jako indeksy łańcuchowe wyznaczone w stosunku do roku poprzedniego, jak i indeksy o stałej podstawie z roku 2007. Dodatkowo dla indeksów jednopodstawowych obliczono procentowe zmiany cen na rynku aukcyjnym w odniesieniu do roku 2007.

Tabela 11.12. Indeksy hedoniczne

Model	Rok	HQA (11.4)	Indeks (11.2) liczony jako łańcuchowy (5.7)	Indeks (11.2) liczony jako jednopodstawowy (5.5)	% zmiana ceny
Surowy indeks (wz. 11.3)	2008		1,4984	1,4984	49,8
	2009		0,6163	0,9235	–7,7
	2010		0,9441	0,8718	–12,8
M1	2008	1,4716	1,0182	1,0182	1,8
	2009	0,6150	1,0021	1,0204	2,0
	2010	0,9607	0,9827	1,0027	0,3
M2	2008	1,4743	1,0163	1,0163	1,6
	2009	0,6156	1,0011	1,0175	1,7
	2010	0,9576	0,9859	1,0031	0,3
M3	2008	1,4616	1,0251	1,0251	2,5
	2009	0,6152	1,0018	1,0269	2,7
	2010	0,9472	0,9967	1,0235	2,4
M4	2008	1,4732	1,0171	1,0171	1,7
	2009	0,6087	1,0125	1,0299	3,0
	2010	0,9705	0,9727	1,0018	0,2

Model	Rok	HQA (11.4)	Indeks (11.2) liczony jako łańcuchowy (5.7)	Indeks (11.2) liczony jako jednopodstawowy (5.5)	% zmiana ceny
M8	2008	1,4137	1,0599	1,0599	6,0
	2009	0,6867	0,8975	0,9512	–4,9
	2010	0,9860	0,9575	0,9108	–8,9
M9	2008	1,4160	1,0582	1,0582	5,8
	2009	0,6862	0,8981	0,9504	–5,0
	2010	0,9887	0,9549	0,9075	–9,2
M10	2008	1,4607	1,0258	1,0258	2,6
	2009	0,5934	1,0387	1,0655	6,5
	2010	0,9643	0,9790	1,0431	4,3
M11	2008	1,4575	1,0280	1,0280	2,8
	2009	0,5930	1,0394	1,0685	6,8
	2010	0,9821	0,9613	1,0271	2,7
M13	2008	1,1029	1,3586	1,3586	35,9
	2009	0,6596	0,9344	1,2695	27,0
	2010	1,1313	0,8345	1,0594	5,9
M15	2008	1,3256	1,1303	1,1303	13,0
	2009	0,8552	0,7206	0,8145	–18,5
	2008	0,9088	1,0389	0,8462	–15,4

Źródło: opracowanie własne na podstawie [Kompa, Witkowska 2014b].

Na uwagę zasługuje fakt, że korekta hedoniczna w istotny sposób zmienia wartości indeksów cen dzieł sztuki, które zostały również wyznaczone jako indeksy surowe z relacji (11.3). Co więcej, korekta wyznaczona na podstawie modeli M8 i M9 w znaczący sposób zmienia wartość indeksu sztuki, chociaż zachowana jest tendencja zaobserwowana dla indeksów surowych. Dotyczy to również modelu M15, w którym pominięto istotną zmienną *malarz*. Z kolei indeks wyznaczony na podstawie oszacowanego modelu M13 wskazuje na wysokie dodatnie zwroty z inwestycji. W przypadku pozostałych indeksów hedonicznych wprowadzona korekta znacząco spłaszczyła zmiany cen. Nasuwa się zatem pytanie, jakie własności powinien posiadać model hedonicznej regresji (11.1), aby można było wykorzystać go do budowy indeksu cen. Pytanie to nie znalazło do tej pory zadowalającej odpowiedzi.

Warto odnotować, że surowy indeks cen wskazuje na niemal 50% wzrost cen analizowanych prac w 2008 roku w porównaniu z rokiem poprzednim, a następnie na istotne spadki cen spowodowane zapewne kryzysem finansowym i spowolnieniem gospodarki. Natomiast indeksy hedoniczne w przypadku większości modeli nie wykazywały tak drastycznych zmian w badanym okresie i jedynie w przypadku modeli M8, M9 i M15 widoczny jest spadek cen.

Zakończenie

Zgodnie z założonym we wstępie celem monografii omówiono w niej wybrane metody statystyczne i ekonometryczne wykorzystywane w finansach. Początkowo wydawało się, że można pominąć omawianie tematyki pierwszych rozdziałów, ale niestety nie da się prowadzić dyskusji w nieznanym języku. Okazuje się bowiem, że dla wielu praktykujących finansistów, doktorantów, czy nawet pracowników uczelni większość zagadnień z zakresu statystyki i ekonometrii jest znana jedynie z nazwy. Dlatego w celu ułatwienia zrozumienia istoty pokazanych analiz empirycznych zamieszczono rozdziały metodologiczne.

Zamiarem moim było również pokazanie, że wiele stosowanych w finansach metod, np. modele wyceny aktywów kapitałowych, to *de facto* aplikacja znanych ze statystyki i ekonometrii modeli, a modele regresji hedonicznej znajdują zastosowanie w analizach dzieł sztuki. Przyznam, że – uczestnicząc w seminariach organizowanych przez prof. Stephana Klasena w Georg August University w Getyndze w 2012 roku – sama otwierałam usta ze zdziwienia, jak wiele niemierzalnych i z pozoru nieobserwowalnych zjawisk można analizować, modelować i prognozować za pomocą metod ilościowych. Świadczy to jedynie o tym, że wciąż jest wiele możliwości aplikacyjnych, a dążenia do rozwiązywania problemów badawczych nie są niczym ograniczone.

Literatura

Aczel, A. D. (2000). *Statystyka w zarządzaniu*. Wydawnictwo Naukowe PWN.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (1984). A further empirical investigation of the bankruptcy cost question. *The Journal of Finance*, *39*(4), 1067–1089. https://doi.org/10.1111/j.1540-6261.1984.tb03893.x

Altman, E. I. (1993). *Corporate financial distress and bankruptcy: A complete guide to predicting & avoiding distress and profiting from bankruptcy (Wiley Finance)*. J. Wiley & Sons.

Bąk, A. (2013). *Mikroekonometryczne metody badania preferencji konsumentów z wykorzystaniem programu R*. Wydawnictwo C.H. Beck.

Bernaś, P. (2016). Klasyfikacja indeksów na rynku sztuki. *Studia i Prace Kolegium Ekonomiczno-Społeczne / Szkoła Główna Handlowa*, *1*(25), 157–178.

Borowski, K. (2007). Nowe kierunki bankowości inwestycyjnej – rynek dzieł sztuki. *Studia i Prace Kolegium Zarządzania i Finansów / Szkoła Główna Handlowa*, *78*, 20–37.

Borowski, K. (2011). *Teoria i praktyka benchmarków*. Difin.

Borowski, K. (2017). *Analiza techniczna: Średnie ruchome, wskaźniki i oscylatory*. Difin.

Box, G. E. P., Jenkins, G. M. (1983). *Analiza szeregów czasowych*. Państwowe Wydawnictwo Naukowe.

Brewster, D. (2006, 14 sierpnia). Art – a classy asset or a new asset class? *Financial Times*, 1–6.

Brzeszczyński, J., Kelm, R. (2002). *Ekonometryczne modele rynków finansowych. Modele kursów giełdowych i kursów walutowych*. IG-Press.

Charemza, W., Deadman, D. F. (1997). *Nowa ekonometria*. Polskie Wydawnictwo Ekonomiczne.

Chow, G. C. (1995). *Ekonometria*. Wydawnictwo Naukowe PWN.

Chrzanowska, M., Witkowska, D. (2007). Zastosowanie wybranych metod klasyfikacji do rozpoznawania indywidualnych kredytobiorców. W: K. Jajuga, M. Walesiak (red.), *Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Taksonomia 14*. *Klasyfikacja i analiza danych – teoria i zastosowania*, *1169* (s. 108–114). Akademia Ekonomiczna we Wrocławiu.

Cieślak, M. (1997). *Prognozowanie gospodarcze*. Wydawnictwo Naukowe PWN.

Ciżkowicz, P. (2020). *Ekonometria stosowana* [slajdy]. Player.Slideplayer.Pl. http://player.slideplayer.pl/download/33/10169472/3gqBKZHh5eqW3k9e_4fq-g/1661467498/10169472.ppt (dostęp 12.09.2020).

Czekaj, J. (red.). (2014). *Efektywność giełdowego rynku akcji w Polsce z perspektywy dwudziestolecia*. Polskie Wydawnictwo Ekonomiczne.

Czerwiński, Z. (1973). *Podstawy matematycznych modeli wzrostu gospodarczego*. Państwowe Wydawnictwo Ekonomiczne.

Dańska-Borsiak, B. (2011). *Dynamiczne modele panelowe w badaniach ekonomicznych*. Wydawnictwo Uniwersytetu Łódzkiego. https://doi.org/10.18778/7525-549-2

Deloitte (2013, kwiecień). *Rynek sztuki. Sztuka rynku*. Deloitte Poland. https://rynekisztuka.pl/wp-content/uploads/2013/04/pl_artbanking_pl.pdf (dostęp 12.11.2013).

Dittmann, P. (2017). *Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie*. Wolters Kluwer Polska/Wydawnictwo Nieoczywiste.

Dmitruk, J. (2012). Taksonomiczna miara atrakcyjności inwestowania (TMAI) na przykładzie spółek giełdowych. *Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie. Ekonomika i Organizacja Gospodarki Żywnościowej*, *99*, 161–173.

Dobosz, M. (2004). *Wspomagana komputerowo statystyczna analiza wyników badań*. Akademicka Oficyna Wydawnicza EXIT.

Domański, C. (1990). *Testy statystyczne*. Państwowe Wydawnictwo Ekonomiczne.

Dunis, C. L. (2001). *Prognozowanie rynków finansowych*. Dom Wydawniczy ABC/Oficyna Ekonomiczna.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Finkel, J. (2004, sierpień). Where the blue chips fall. *Art & Auction*, 77–83.

Fiszeder, P. (2009). *Modele klasy GARCH w empirycznych badaniach finansowych*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Foo, J., Witkowska, D. (2017). A comparison of global financial market recovery after the 2008 global financial crisis. *Folia Oeconomica Stetinensia*, *17*(1), 109–128. https://doi.org/10.1515/foli-2017-0009

Frey, B. S., Cueni, R. (2013). Why invest in art? *The Economist's Voice*, *10*(1), 1–6.

Gajdka, J., Stos, D. (1996). Wykorzystanie analizy dyskryminacyjnej w ocenie kondycji finansowej przedsiębiorstw. W: R. Borowiecki (red.), *Restrukturyzacja w procesie przekształceń i rozwoju przedsiębiorstw* (s. 56–65). Wydawnictwo Akademii Ekonomicznej w Krakowie.

Ganczarek-Gamrot, A. (2014). *Analiza szeregów czasowych*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Gatnar, E. (1998). *Symboliczne metody klasyfikacji danych*. Wydawnictwo Naukowe PWN.

Gatnar, E., Walesiak, M. (red.). (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu.

Goetzmann, W., Renneboog, L., Spaenjers, C. (2011). Art and money. *The American Economic Review*, *101*(3), 222–226. http://www.jstor.org/stable/29783743 (dostęp 23.11.2019).

Goryl, A., Jędrzejczak, Z., Kukuła, K., Osiewalski, J., Walkosz, A. (2000). *Wprowadzenie do ekonometrii w przykładach i zadaniach*. Wydawnictwo Naukowe PWN.

Górka, J. (2012). *Modele klasy Sign RCA GARCH*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Gruszczyński, M. (red.). (1996). *Ekonometria*. Oficyna Wydawnicza SGH.

Gruszczyński, M. (1999). Scoring logitowy w praktyce bankowej a zagadnienie koincydencji. *Bank i Kredyt*, *5*, 57–63.

Gruszczyński, M. (2001). *Modele i prognozy zmiennych w jakościowych finansach i bankowości*. Oficyna Wydawnicza SGH.

Gruszczyński, M. (2002). Kondycja finansowa przedsiębiorstw. Prognozy ekonometryczne. W: D. Zarzecki (red.), *Czas na pieniądz. Zarządzanie finansami. Klasyczne zasady – nowoczesne narzędzia* (s. 101–111). Uniwersytet Szczeciński, Wydział Nauk Ekonomicznych i Zarządzania.

Gruszczyński, M. (red.). (2010). *Mikroekonometria*. Wolters Kluwer Polska.

Gruszczyński, M. (2012). *Empiryczne finanse przedsiębiorstw. Mikroekonometria finansowa*. Difin.

Gruszczyński, M. (2020). *Financial microeconometrics. A research methodology in corporate finance and accounting*. Springer.

Hamrol, M., Czajka, B., Piechocki, M. (2004). Upadłość przedsiębiorstw – model analizy dyskryminacyjnej. *Zeszyty Teoretyczne Rachunkowości*, *76*, 37–48.

Hellwig, Z. (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. *Przegląd Statystyczny*, *4*, 307–327.

Hołda, A. (2001). Prognozowanie bankructwa jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej ZH. *Rachunkowość*, *5*, 306–310.

Huterska, A., Zdunek-Rosa, E. (2016). Zastosowanie modelu probitowego i uciętego liniowego modelu prawdopodobieństwa do analizy kondycji ekonomiczno-finansowej wybranych przedsiębiorstw z indeksu mWIG40. *Finanse, Rynki Finansowe, Ubezpieczenia*, *5*(83), cz. 2, 121–130.

Jajuga, K. (1993). *Statystyczna analiza wielowymiarowa*. Państwowe Wydawnictwo Naukowe.

Jajuga, K. (1998). Metoda analizy dyskryminacyjnej w przypadku zmiennych dyskretnych. *Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Ekonometria t. 1*, *765*, 24–32.

Jajuga, K. (red.). (2000). *Metody ekonometryczne i statystyczne a analizie rynku kapitałowego*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Jajuga, K. (2002). *Podstawy inwestowania na rynku papierów wartościowych*. Giełda Papierów Wartościowych w Warszawie S.A.

Juszczyk, M. (2015). Powiązanie kondycji finansowej spółek giełdowych określonej syntetycznym miernikiem atrakcyjności inwestowania (TMAI) z kształtowaniem się kursów ich akcji. *Zeszyty Naukowe SGGW. Ekonomika i Organizacja Gospodarki Żywnościowej*, *111*, 81–95. https://doi.org/10.22630/eiogz.2015.111.36

Kądziołka, K. (2021). Propozycja metody wspomagającej wybór miernika taksonomicznego na przykładzie oceny atrakcyjności giełd kryptowalut. *Zeszyty Naukowe ZPSB „Firma i Rynek"*, *1*(59), 65–76. https://www.zpsb.pl/wp-content/uploads/2021/12/FIR_2021_260-6-KKadziolka-1.pdf?x71584 (dostęp 23.11.2019).

Kokczyński, B. (2019). *Identyfikacja czynników wpływających na ocenę ryzyka kredytowego mikroprzedsiębiorstw* (praca magisterska napisana pod kierunkiem D. Witkowskiej). Uniwersytet Łódzki.

Kolonko, J. (1980). *Analiza dyskryminacyjna i jej zastosowania w ekonomii*. Państwowe Wydawnictwo Naukowe.

Kompa, K. (2009). Budowa mierników agregatowych do oceny poziomu rozwoju społeczno-gospodarczego. *Zeszyty Naukowe SGGW. Ekonomika i Organizacja Gospodarki Żywnościowej*, *74*, 5–26. https://sj.wne.sggw.pl/pdf/EIOGZ_2009_n74_s5.pdf (dostęp 23.11.2019).

Kompa, K., Witkowska, D. (2013). Indeks rynku sztuki. Badania pilotażowe dla wybranych malarzy polskich. *Zarządzanie i Finanse*, *3*(2), 33–50. http://bazekon.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171313015 (dostęp 23.11.2019).

Kompa, K., Witkowska, D. (2014a). Indeks hedoniczny malarstwa polskiego dla najbardziej popularnych autorów na rynku aukcyjnym w latach 2007–2010. *Acta Universitatis Nicolai Copernici. Ekonomia*, *XLV*(1), 7–26. https://doi.org/10.12775/AUNC_ECON.2014.001

Kompa, K., Witkowska, D. (2014b). Prices of paintings on Polish art market in years 2007–2010 – hedonic price index application. *Acta Scientiarum Polonorum, Oeconomia*, *13*(3), 127–142. http://acta_oeconomia.sggw.pl/pdf/Acta_Oeconomia_13_3_2014.pdf (dostęp 23.11.2019).

Kompa, K., Witkowska, D. (2020). *Efektywność i ryzyko polskiego rynku funduszy inwestycyjnych w świetle zmian warunków funkcjonowania funduszy emerytalnych*. Wydawnictwo Uniwersytetu Łódzkiego.

Kopczewska, K. (2007). *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*. CeDeWu.

Kräussl, R., Elsland, N. van (2008, listopad). *Constructing the true art market index: A novel 2-step hedonic approach and its application to the German art market*. CFS Working Paper Series. https://www.econstor.eu/bitstream/10419/25546/1/577545752.PDF (dostęp 23.11.2019).

Krzyśko, M. (1990). *Analiza dyskryminacyjna*. Wydawnictwa Naukowo-Techniczne. https://www.researchgate.net/profile/Miroslaw-Krzysko/publication/292148698_Analiza_dyskryminacyjna/links/5d2cb463458515c11c3359ea/Analiza-dyskryminacyjna.pdf (dostęp 23.11.2019).

Kufel, T. (2010). *Ekonometryczna analiza cykliczności procesów gospodarczych o wysokiej częstotliwości obserwowania*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Kuźnik, P. (2019). *Efektywność inwestycyjna spółek a ocena ich kondycji ekonomiczno-finansowej* (praca magisterska napisana pod kierunkiem D. Witkowskiej). Uniwersytet Łódzki.

Luszniewicz, A., Słaby, T. (2008). *Statystyka z pakietem komputerowym STATISTICA PL*. *Teoria i zastosowania.* Wydawnictwo C.H. Beck.

Łuniewska, M. (2012). *Ekonometria finansowa*. Wydawnictwo Naukowe PWN.

Maddala, G. S. (2006). *Ekonometria*. Wydawnictwo Naukowe PWN.

Malarska, A. (2005). *Statystyczna analiza danych wspomagana programem SPSS*. SPSS Polska.

Mazur, A., Staniec, I., Witkowska, D. (1999). Klasyfikacja podmiotów gospodarczych według grup ryzyka kredytowego. W: T. Trzaskalik (red.), *Modelowanie Preferencji a Ryzyko ,99* (s. 247–258). Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach.

Mączyńska, E., Zawadzki, M. (2006). Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw. *Ekonomista*, *2*, 205–235.

Mei, J., Moses, M. (2002). Art as an investment and the underperformance of masterpieces. *American Economic Review*, *92*(5), 1656–1668. https://www.jstor.org/stable/3083271 (dostęp 23.11.2019).

Milo, W. (1990a). *Nieliniowe modele ekonometryczne*. Państwowe Wydawnictwo Naukowe.

Milo, W. (1990b). *Szeregi czasowe*. Państwowe Wydawnictwo Ekonomiczne.

Osińska, M. (2006). *Ekonometria finansowa*. Polskie Wydawnictwo Ekonomiczne.

Osińska, M. (red.). (2007). *Procesy STUR modelowanie i zastosowanie do finansowych szeregów czasowych*. Towarzystwo Naukowe Organizacji i Kierowania „Dom Organizatora".

Piasecki, K., Tomasik, E. (2013). *Rozkłady stop zwrotu z instrumentów polskiego rynku kapitałowego*. Wydawnictwo edu-Libri.

Piłatowska, M. (red.). (2002). *Analiza szeregów czasowych na początku XXI wieku*. Wydawnictwo Uniwersytetu Mikołaja Kopernika w Toruniu.

Prusak, B. (2005). *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstwa*. Difin.

Renneboog, L., Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. *Management Science*, *59*(1), 36–53. https://doi.org/10.1287/mnsc.1120.1580

Rocznik KNF (2002). *Rocznik ubezpieczeń i funduszy emerytalnych 2002*. Komisja Nadzoru Ubezpieczeń i Funduszy Emerytalnych. https://www.knf.gov.pl/knf/pl/komponenty/img/uwagi_ogolne_02_2091.pdfD (dostęp 16.05.2020).

Sawicki, Ł., Borowski, K. (2018). *Rynek dzieł sztuki. Podejście inwestycyjne i behawioralne.* Difin.

Sobczyk, M. (2010). *Statystyka opisowa*. Wydawnictwo C.H. Beck.

Sokołowski, A. (1985). Wybrane zagadnienia pomiaru i ważenia cech w taksonomii. *Zeszyty Naukowe / Akademia Ekonomiczna w Krakowie*, *203*, 41–53.

Stanimir, A., Błaczkowska, A. (2006). *Analiza danych marketingowych. Problemy, metody, przykłady*. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego.

Strzelecka, A. (2017). Zastosowanie analizy dyskryminacyjnej do oceny zdolności separacyjnej determinant efektywności technicznej szpitali. *Zeszyty Naukowe Politechniki Śląskiej. Organizacja i Zarządzanie*, *113*, 457–467. https://www.polsl.pl/Wydzialy/ROZ/ZN/Documents/z113/Strzelecka.pdf (dostęp 23.11.2019).

Styś, D. (2019). *Identyfikacja czynników wpływających na stopy zwrotu spółek z indeksu Dow Jones Industrial Average* (praca magisterska napisana pod kierunkiem D. Witkowskiej). Uniwersytet Łódzki.

Suchecki, B. (red.). (2010). *Ekonometria przestrzenna: metody i modele analizy danych przestrzennych*. Wydawnictwo C.H. Beck.

Szulc, E., Wleklińska, D. (2021). *Przestrzenno-czasowa analiza powiązań rynków papierów wartościowych z uwzględnieniem odległości ekonomicznej vs. geograficznej*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika. https://doi.org/10.12775/978-83-231-4639-1

Tarczyńska-Łuniewska, M. (2013). *Metodologia oceny siły fundamentalnej spółek (giełdowych i pozagiełdowych)*. Przedsiębiorstwo Produkcyjno-Handlowe ZAPOL Dmochowski, Sobczyk.

Tarczyńska-Łuniewska, M., Tarczyński, W. (2006). *Metody wielowymiarowej analizy porównawczej na rynku kapitałowym*. Wydawnictwo Naukowe PWN.

Tarczyński, W. (1994). Taksonomiczna miara atrakcyjności inwestycji w papiery wartościowe. *Przegląd Statystyczny*, *3*, 275–300.

Tarczyński, W. (1995). O pewnym sposobie wyznaczania składu portfela papierów wartościowych. *Przegląd Statystyczny*, *42*(1), 91–106.

Tarczyński, W. (1997). *Rynki kapitałowe. Metody ilościowe, cz. II*. Wydawnictwo Placet.

Tarczyński, W., Łuniewska, M. (2004). *Dywersyfikacja ryzyka na polskim rynku kapitałowym*. Wydawnictwo Placet.

Tarczyński, W., Witkowska, D., Kompa, K. (2013). *Współczynnik beta. Teoria i praktyka*. Wydawnictwo Pielaszek Research.

Tomczyk, E., Widłak, M. (2010). Konstrukcja i własności hedonicznego indeksu cen mieszkań dla Warszawy. *Bank i Kredyt*, *41*(1), 99–128. https://www.bankandcredit.nbp.pl/content/2010/01/bik_01_2010_05_art.pdf

Triplett, J. (2006). *Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products*. Van Haren Publishing/OECD Publishing.

Urbanek, M. (1995). Klasyfikacja i selekcja wskaźników finansowych. W: E. Nowak, M. Urbanek (red.), *Ekonometryczne modelowanie danych finansowo-księgowych* (s. 113–123). Norbertinum.

Welfe, A. (1995). *Ekonometria*. Polskie Wydawnictwo Ekonomiczne.

Welfe, W. (2010). *Zarys historii ekonometrycznego modelowania gospodarki narodowej*. Wydawnictwo Uniwersytetu Łódzkiego.

Węgrzyn, T. (2013). Dobór spółek do portfela z wykorzystaniem wskaźników finansowych i ich względnego tempa przyrostu. Analiza w latach 2001–2010. *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, *174*, 63–74.

Wiśniewski, J. W. (2009). *Mikroekonometria*. Wydawnictwo Naukowe UMK.

Wiśniewski, J. W. (2015). *Microeconometrics in business management*. Wiley.

Witkowska, D. (1997). Analiza wniosków kredytowych za pomocą algorytmu wstecznej propagacji błędu. W: *Sztuczna Inteligencja i Inżynieria Finansowa* (s. 235–242). CIR-12'97.

Witkowska, D. (1999a). Porównanie wyników klasyfikacji przedsiębiorstw za pomocą liniowej funkcji dyskryminacji i sztucznych sieci neuronowych. W: J. Lewandowski (red.), *Materiały Konferencyjne VI Międzynarodowej Konferencji Naukowej: Zarządzanie Organizacjami Gospodarczymi* (s. 775–785). Instytut Organizacji i Zarządzania Politechniki Łódzkiej.

Witkowska, D. (1999b). *Sztuczne sieci neuronowe w analizach ekonomicznych*. INTRO-GRAPH.

Witkowska, D. (2000). *Sztuczne sieci neuronowe w analizach ekonomicznych*. Wydział Organizacji i Zarządzania Politechniki Łódzkiej/Wydawnictwo Menadżer.

Witkowska, D. (2002). *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe.* Wydawnictwo C.H. Beck.

Witkowska, D. (red.). (2004). *Statystyka w zarządzaniu*. Wydawnictwo AND Adam Domagała.

Witkowska, D. (2012). *Podstawy ekonometrii i teorii prognozowania*. Wolters Kluwer Polska.

Witkowska, D. (2014). An application of hedonic regression to evaluate prices of Polish paintings. *International Advances in Economic Research*, *20*(3), 281–293. https://doi.org/10.1007/s11294-014-9468-x

Witkowska, D. (2016a). Evaluation of the individual hedonic art price indexes for the Polish painters representing the auction market in Poland. *Transformations in Business & Economics*, *15*(2(38)), 61–73. http://www.transformations.knf.vu.lt/38/article/eval

Witkowska, D. (2016b). *Zmiana warunków funkcjonowania a efektywność inwestycyjna otwartych funduszy emerytalnych*. Wydawnictwo Uniwersytetu Łódzkiego.

Witkowska, D., Kuźnik, P. (2019). Does fundamental strength of the company influence its investment performance? *Dynamic Econometric Models*, *19*, 85–96. https://doi.org/10.12775/DEM.2019.005

Witkowska, D., Staniec, I. (2000). Dychotomiczna klasyfikacja kredytobiorców za pomocą wybranych metod klasyfikacji. *Zeszyty Naukowe Wydziału Mechanicznego Politechniki Koszalińskiej. Polioptymalizacja i Komputerowe Wspomaganie Projektowania*, *Mielno 2000*, 349–356.

Witkowska, D., Staszak, M. (2021). Metody doboru spółek do portfela. Analiza porównawcza ich skuteczności w odniesieniu do spółek notowanych na GPW w latach 2002–2019. W: S. Franek, A. Adamczyk (red.), *Finanse jako katalizator przemian współczesnej gospodarki* (s. 119–132). Wydawnictwo Uniwersytetu Szczecińskiego.

Witkowska, D., Witkowska, J. (1997a). Analiza sieci dystrybucji produktów chemii gospodarczej i spożywczej w Łodzi. *Studia Prawno-Ekonomiczne*, *LV*, 247–259.

Witkowska, D., Witkowski, J. (1997b). Analiza sieci dystrybucji produktów spożywczych i chemicznych w miastach województwa łódzkiego (bez Łodzi). *Studia Prawno-Ekonomiczne*, *LVI*, 311–323.

Witkowska, D., Kompa, K., Staszak, M. (2021). Indicators for the efficient portfolio construction. The case of Poland. *Procedia Computer Science*, *192*, 2022–2031. https://doi.org/10.1016/j.procs.2021.08.208

Witkowska, D., Matuszewska, A., Kompa, K. (2012). *Wprowadzenie do ekonometrii dynamicznej i finansowej*. Wydawnictwo SGGW.

Zamojska, A. (2012). *Efektywność funduszy inwestycyjnych w Polsce*. Wydawnictwo C.H. Beck.

Zeliaś, A. (1991). *Ekonometria przestrzenna*. Państwowe Wydawnictwo Ekonomiczne.

Zeliaś, A. (2000). *Metody statystyczne*. Polskie Wydawnictwo Ekonomiczne.

Zeliaś, A., Pawełek, B., Wanat, S. (2003). *Prognozowanie ekonomiczne*. *Teoria, przykłady, zadania.* Wydawnictwo Naukowe PWN.

Zieliński, W. (1997). *Wybrane testy statystyczne*. Fundacja „Rozwój SGGW".

Źródła danych

https://gpwbenchmark.pl/karta-indeksu?isin=PL9999999987#Portfolio (dostęp 16.05.2020).

https://grupadino.pl/wp-content/uploads/2019/03/RS_2018_Grupa_Grupa_Dino_Polska_Sprawozdanie_Finansowe.pdf (dostęp 16.05.2020).

https://ir.enea.pl/pr/427942/skonsolidowany-raport-roczny-rs-2018 (dostęp 16.05.2020).

https://ir.energa.pl/pr/426660/skonsolidowane-wyniki-finansowe-za-2018-rok (dostęp 16.05.2020).

https://kghm.com/pl/inwestorzy/centrum-wynikow/raporty-finansowe (dostęp 16.05.2020).

https://ovostar.ua/uploads/investors/files/5e8ecb114fb54.pdf (dostęp 14.04.2020).

https://pl.investing.com/equities/aviva-historical-data (dostęp 18.05.2022).

https://raportzintegrowany2018.orlen.pl/pl/skonsolidowane-sprawozdanie-finansowe (dostęp 16.05.2020).

http://test2.wawel.com.pl/relations.php/pl/raporty/raporty-okresowe.html (dostęp 14.04.2020).

https://stat.gov.pl/obszary-tematyczne/inne-opracowania/informacje-o-sytuacji-spoleczno-gospodarczej/biuletyn-statystyczny-nr-92020,4,104.html (dostęp 20.12.2020).

https://stat.gov.pl/wskazniki-makroekonomiczne/ (dostęp 23.11.2019).
https://www.ambra.com.pl/assets/RI/Raporty/okresowe/20182019/AMBRA_raport_roczny_2018-2019.pdf (dostęp 14.04.2020).
https://www.cdprojekt.com/pl/wp-content/uploads-l/2019/03/2018_rs_spraw_finansowe.pdf (dostęp 16.05.2020).
https://www.cezpolska.pl/pl/cez-w-polsce/raporty-roczne (dostęp 16.05.2020).
https://www.gkpge.pl/relacje-inwestorskie/Materialy-do-pobrania?rok=2018 (dostęp 16.05.2020).
https://www.GPW.pl oraz opracowanie własne na podstawie informacji z bazy danych EMIS (dostęp 30.05.2020).
https://www.kernel.ua/wp-content/uploads/2019/10/Kernel_FY2019_Annual_Report_.pdf (dostęp 07.04.2020).
https://www.lppsa.com/wp-content/uploads/2019/04/GK-LPP-skonsolidowany-roczny-raport-za-2018-rok.pdf (dostęp 16.05.2020).
https://www.stockwatch.pl/gpw/indeks/wig-spozyw,sklad.aspx (dostęp 14.04.2020).
https://www.stooq.pl (dostęp 18.05.2022).
https://www.tauron.pl/tauron/relacje-inwestorskie/raporty-okresowe (dostęp 16.05.2020).
https://ztkruszwica.pl/pl/relacje-inwestorskie/raporty-gieldowe/raporty-roczne/2019 (dostęp 14.04.2020).

W monografii przedstawiono wiele zagadnień związanych z prowadzeniem analiz finansowych za pomocą metod ilościowych. Walorem opracowania jest omówienie problemów, z jakimi spotykają się analitycy finansowi oraz prezentacja przykładów aplikacji różnych metod zarówno ułatwiających ocenę i opis omawianych sytuacji, jak i wspomagających podejmowanie decyzji w określonych warunkach. Większość ukazanych metod poparto przykładami zawierającymi rzeczywiste dane finansowe i opatrzono bogatymi komentarzami, zawierającymi interpretację uzyskanych wyników analiz. Publikacja jest dedykowana zarówno praktykom działającym na szeroko rozumianym rynku finansowym, jak i studentom. Do jej studiowania nie jest wymagana znajomość zaawansowanych metod ilościowych, wszystkie zagadnienia omawiane są bowiem w przystępny sposób.

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

wydawnictwo.uni.lodz.pl
ksiegarnia@uni.lodz.pl
(42) 665 58 63

Książka dostępna również
jako e-book

ISBN 978-83-8142-304-5