Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Politechnika Świętokrzyska
Wydział Zarządzania i Modelowania Komputerowego

Kierunek studiów:
Inżynieria danych

Ewelina Sendek-Matysiak

Materiały dydaktyczne do przedmiotu

MODELOWANIE ZALEŻNOŚCI W DANYCH

opracowane w ramach realizacji Projektu
„Dostosowanie kształcenia w Politechnice Świętokrzyskiej do potrzeb współczesnej gospodarki”
FERS.01.05-IP.08-0234/23

Kielce, 2025

Spis treści

1. Przygotowanie danych do analizy

1.1. Źródła danych i metody pozyskania danych do analizy

Proces analizy danych w inżynierii danych rozpoczyna się od identyfikacji i pozyskania wiarygodnych źródeł informacji.
Dane stanowią fundament wszelkich działań analitycznych, dlatego ich jakość, kompletność i reprezentatywność mają kluczowe znaczenie dla poprawności uzyskiwanych wyników i formułowanych wniosków.
W praktyce inżynierskiej dane mogą pochodzić z wielu różnorodnych źródeł – zarówno pierwotnych, jak i wtórnych – a wybór sposobu ich akwizycji powinien być ściśle uzależniony od celu analizy oraz specyfiki badanego zjawiska.
Do **źródeł pierwotnych** zalicza się dane uzyskane w wyniku bezpośredniego pomiaru, obserwacji lub eksperymentu. W badaniach inżynierskich mogą to być odczyty z czujników pomiarowych, wyniki testów laboratoryjnych, dane telemetryczne, sygnały z systemów IoT (Internet of Things) czy informacje pochodzące z platform monitorujących procesy przemysłowe. Dane pierwotne charakteryzują się wysoką aktualnością i precyzją, jednak ich pozyskanie wymaga odpowiedniego zaplecza technicznego oraz często wiąże się ze znacznymi kosztami. Przykładowo, w zautomatyzowanych zakładach produkcyjnych dane zbierane są w czasie rzeczywistym z wielu urządzeń, co umożliwia analizę dynamiki procesów, identyfikację anomalii lub prognozowanie awarii.
Źródła wtórne obejmują natomiast dane już wcześniej zgromadzone i udostępnione przez inne instytucje lub podmioty. Do tej kategorii należą bazy danych instytucji publicznych, rejestry administracyjne, raporty branżowe, publikacje naukowe, zbiory statystyczne czy ogólnodostępne zasoby portali typu open data, takich jak dane.gov.pl, Eurostat czy OECD Data. W inżynierii danych istotną rolę odgrywają również bazy komercyjne, udostępniane przez przedsiębiorstwa specjalizujące się w agregacji informacji gospodarczych, rynkowych i technologicznych. Korzystanie ze źródeł wtórnych pozwala na szerokie analizy porównawcze, jednak wymaga krytycznej oceny ich wiarygodności, spójności oraz aktualności.
Metody gromadzenia danych obejmują zarówno procesy manualne, jak i zautomatyzowane. W tradycyjnym podejściu dane wprowadzane są ręcznie do arkuszy kalkulacyjnych lub baz danych, jednak współcześnie dominują metody automatyczne, które umożliwiają szybkie i efektywne pobieranie informacji z różnych źródeł. Jedną z najczęściej stosowanych technik jest web scraping, polegający na automatycznym pobieraniu treści z witryn internetowych przy użyciu specjalnych skryptów i narzędzi. Innym powszechnym rozwiązaniem jest wykorzystanie interfejsów API (Application Programming Interface), które zapewniają bezpośredni i ustrukturyzowany dostęp do danych z systemów informatycznych oraz serwisów

internetowych. W przypadku dużych zbiorów danych (big data) stosuje się także platformy strumieniowe, takie jak Apache Kafka, umożliwiające bieżące przetwarzanie napływających informacji.

Ważnym elementem procesu akwizycji danych jest ich wstępna weryfikacja, obejmująca ocenę poprawności, kompletności i spójności uzyskanych informacji. Może ona być przeprowadzana automatycznie – na przykład poprzez implementację reguł walidacyjnych w systemach bazodanowych – lub manualnie, poprzez kontrolę próbki danych przez analityka. Szczególną uwagę należy zwrócić na występowanie wartości odstających, braków danych, duplikatów i błędów typograficznych, które mogą istotnie zaburzyć wyniki analizy. W praktyce inżynierskiej proces ten wspierany jest narzędziami do czyszczenia danych (data cleaning), które pozwalają na ich standaryzację i przygotowanie do dalszego przetwarzania.
W kontekście badań inżynierskich niezwykle istotne jest również dokumentowanie źródeł oraz sposobu pozyskania danych. Każdy zbiór powinien być opisany w sposób umożliwiający jednoznaczną identyfikację jego pochodzenia, daty pobrania, formatu i ewentualnych przekształceń, jakim został poddany. Tego typu opis stanowi część tzw. metadanych, czyli danych o danych, które są kluczowe dla zapewnienia transparentności procesu badawczego oraz powtarzalności wyników.

1.2. Klasyfikacja zmiennych ze względu na charakter danych

Proces analizy danych wymaga nie tylko odpowiedniego pozyskania i wstępnego przygotowania zbioru obserwacji, lecz także właściwego rozpoznania natury i typu każdej zmiennej. Klasyfikacja zmiennych stanowi jeden z najważniejszych etapów przygotowania analizy, ponieważ determinuje zarówno wybór metod statystycznych, jak i sposób wizualizacji oraz interpretacji wyników. Źle sklasyfikowana zmienna może prowadzić do błędnych wniosków, a nawet całkowicie wypaczyć rezultaty modelowania.

1.2.1. Pojęcie obserwacji i zmiennej

W analizie danych zmienną nazywa się każdy atrybut, cechę lub właściwość, która może przyjmować różne wartości w obrębie badanej zbiorowości. Zmienna jest zatem elementem opisującym poszczególne obserwacje (jednostki analizy), który może mieć charakter liczbowy, kategoryczny, porządkowy lub logiczny. Wartości zmiennej mogą być mierzone, obliczane lub deklarowane, w zależności od źródła danych i rodzaju procesu pomiarowego.
Na przykład w analizie dotyczącej zużycia energii przez pojazdy elektryczne zmiennymi mogą być: masa pojazdu (kg), pojemność akumulatora (kWh), rodzaj napędu czy zasięg (km).

W literaturze przedmiotu występują klasyfikacje zmiennych według różnych kryteriów, m.in. sposobu zapisu wartości zmiennych, możliwości ich pomiaru fizycznego, liczebność zbioru wartości.

1.2.2. Klasyfikacja zmiennych

Jednym z najczęściej stosowanych podziałów zmiennych jest ich klasyfikacja z punktu widzenia sposobu zapisu wartości.
W literaturze wyróżnia się cztery podstawowe skale pomiarowe: **nominalną, porządkową, przedziałową** oraz **ilorazową**. Każda z nich charakteryzuje się odmiennym zakresem dopuszczalnych operacji matematycznych i statystycznych.
Skala nominalna - służy do opisu zmiennych jakościowych, które nie mają naturalnego porządku. Wartości w tej skali są jedynie etykietami służącymi do rozróżniania kategorii. Nie można ich dodawać, odejmować ani porównywać pod względem wielkości. Przykładem może być zmienna *rodzaj paliwa* (benzyna, diesel, elektryczny, hybrydowy) lub *kolor pojazdu*.
W analizie takich danych wykorzystuje się często statystyki opisowe oparte na częstościach występowania (np. liczba pojazdów w każdej kategorii), a wyniki prezentuje się przy użyciu wykresów kołowych lub słupkowych. W modelach analitycznych zmienne nominalne koduje się najczęściej metodą **one-hot encoding**, przekształcając kategorie w zmienne binarne.
Skala porządkowa pozwala na uporządkowanie kategorii według określonego kryterium, lecz różnice między kolejnymi wartościami nie są równomierne ani mierzalne. Przykładem może być zmienna *poziom satysfakcji z użytkowania pojazdu* (niski, średni, wysoki).
Dla takich zmiennych można stosować testy nieparametryczne (np. testy rangowe) oraz obliczać medianę czy dominantę, lecz nie średnią arytmetyczną. W wizualizacji często wykorzystuje się wykresy ramkowe lub słupkowe z uporządkowanymi kategoriami.
W skali przedziałowej występuje już wymierna odległość między wartościami, lecz brak naturalnego zera, które oznaczałoby „brak" danej cechy. Typowym przykładem jest *temperatura w stopniach Celsjusza*. Możliwe jest obliczanie różnic i średnich, ale nie ilorazów.
W tej skali można stosować szeroki zakres metod statystycznych — od korelacji po regresję liniową, pod warunkiem że spełnione są odpowiednie założenia dotyczące rozkładu danych.
Skala ilorazowa jest najbardziej zaawansowanym typem pomiaru. Posiada zarówno jednostkę miary, jak i naturalne zero, oznaczające całkowity brak danej wielkości. Przykłady to *masa pojazdu*, *prędkość maksymalna*. Dane w tej skali pozwalają na wykonywanie pełnego zestawu operacji arytmetycznych i statystycznych, w tym

obliczania ilorazów, wskaźników, współczynników korelacji czy regresji. W praktyce analitycznej większość zmiennych ilościowych ma właśnie charakter ilorazowy.

Zbliżona do powyższego podziału jest klasyfikacja zmiennych według możliwości ich pomiaru fizycznego. Według tego kryterium wyróżnia się:
a) zmienne mierzalne, czyli takie których wartości są wynikiem pomiaru fizycznego i wyrażone są w określonych jednostkach miary, np. wzrost w cm, waga w kg,
b) zmienne pośrednio mierzalne, tzn. cechy których wartości wyrażone są liczbowo, ale liczby te są jedynie wynikiem oceny (przyporządkowane są odpowiednim wyrażeniom słownym), np. ocena z egzaminu,
c) zmienne niemierzalne, czyli cechy których wartości są wyrażane słownie, np. płeć, pochodzenie społeczne, marka samochodu.

Ze względu na liczebność zbioru wartości zmiennej wyróżnia się:
a) cechy stałe, dla których zbiór wartości jest jednoelementowy; innymi słowy każdemu obiektowi badanej zbiorowości przypisywana jest ta sama wartość cechy. Niekiedy wśród nich wyróżnia się dodatkowo cechy stałe rzeczowe, czasowe i przestrzenne umożliwiające precyzyjne zdefiniowanie jednostki i zbiorowości dla określonego badania. Cechy te odpowiadają zakresowi rzeczowemu, czasowemu i przestrzennemu badania.
b) cechy zmienne, dla których zbiór wartości jest co najmniej dwuelementowy.
Wśród cech zmiennych dokonuje się zwykle dalszego ich podziału z punktu widzenia podanego wyżej kryterium i wyróżnia się dodatkowo:
1) dla cech opisowych:
a) zmienne dwuwariantowe (zero-jedynkowe) np. płeć,
b) zmienne wielowariantowe, np. poziom wykształcenia, zawód,
2) dla cech liczbowych:
a) zmienne skokowe (dyskretne); posiadają one przeliczalny zbiór wartości zawierający się w zbiorze liczb naturalnych (liczby całkowite nieujemne), np. liczba osób w rodzinie,
b) zmienne ciągłe, których zbiór wartości jest nieprzeliczalny i należy do zbioru liczb rzeczywistych; wartości takiej cechy są nieskończenie podzielne i mogą być wyrażane z dowolnie dużą dokładnością, np. wzrost, wiek, waga.
W modelach analitycznych, zwłaszcza regresyjnych, wyróżnia się dwa zasadnicze typy zmiennych:

- **zmienną objaśnianą (zależną)** – reprezentującą zjawisko, które chcemy wyjaśnić lub przewidzieć (np. zużycie energii, koszt eksploatacji);
- **zmienne objaśniające (niezależne)** – opisujące czynniki mające wpływ na wartość zmiennej zależnej (np. masa pojazdu, rodzaj napędu, pojemność baterii).

Dobór odpowiednich zmiennych objaśniających ma zasadnicze znaczenie dla jakości modelu. W inżynierii danych często stosuje się metody selekcji zmiennych (feature

selection), które pomagają ograniczyć nadmiar informacji i poprawić interpretowalność modelu.
W praktyce analitycznej często spotyka się zmienne o charakterze **binarnym** (zero-jedynkowym), które przyjmują tylko dwie wartości — np. *1 – pojazd elektryczny, 0 – pojazd spalinowy*. Takie zmienne są szczególnie przydatne w modelach klasyfikacyjnych, zwłaszcza w regresji logistycznej, drzewach decyzyjnych i analizie dyskryminacyjnej.
W kontekście kodowania danych zmienne binarne są często tworzone ze zmiennych kategorycznych w wyniku tzw. **dummy coding**. Dzięki temu dane jakościowe mogą być wykorzystane w analizie ilościowej, co pozwala na łączenie różnych typów informacji w jednym modelu.
Poprawna klasyfikacja zmiennych jest nie tylko formalnym wymogiem, lecz przede wszystkim warunkiem rzetelności całego procesu analitycznego. Decyduje ona o tym, jakie statystyki są dopuszczalne (np. średnia czy mediana), jakie testy można przeprowadzić (parametryczne lub nieparametryczne) oraz jakie modele są adekwatne (np. regresja liniowa, regresja logistyczna, analiza wariancji), np. próba zastosowania regresji liniowej do zmiennej nominalnej skutkuje utratą sensu interpretacyjnego współczynników i błędną strukturą modelu. Dlatego każdy etap pracy z danymi — od eksploracji po modelowanie — powinien rozpoczynać się od dokładnej analizy typów zmiennych.

1.3. Weryfikacja i przygotowanie danych do analizy

Po pozyskaniu danych z odpowiednich źródeł i identyfikacji typu zmiennych kolejnym, niezwykle istotnym etapem procesu analitycznego jest ich weryfikacja oraz przygotowanie do dalszej analizy. Dane surowe, niezależnie od sposobu ich pozyskania, bardzo rzadko nadają się bezpośrednio do modelowania czy wizualizacji. Zazwyczaj zawierają błędy, braki, wartości odstające lub niejednolity format zapisu, które mogą prowadzić do zniekształcenia wyników. Dlatego konieczne jest przeprowadzenie szeregu działań mających na celu zapewnienie spójności, wiarygodności i użyteczności danych.
Proces weryfikacji danych rozpoczyna się od oceny ich kompletności, czyli sprawdzenia, czy wszystkie obserwacje i zmienne niezbędne do przeprowadzenia analizy są dostępne. Braki danych (tzw. missing values) mogą wynikać z błędów pomiarowych, nieprawidłowego działania czujników, awarii systemu zbierającego dane lub niedostarczenia informacji przez użytkownika. W zależności od charakteru i zakresu braków możliwe są różne strategie postępowania: usunięcie obserwacji niekompletnych, uzupełnienie ich wartościami średnimi, medianą, interpolacją lub imputacją opartą na modelach statystycznych. Wybór odpowiedniej metody uzupełniania braków powinien być podyktowany zarówno specyfiką danych, jak i celem analizy, aby uniknąć wprowadzenia sztucznych zależności.

Kolejnym elementem weryfikacji jest identyfikacja wartości odstających (outliers), czyli obserwacji znacznie odbiegających od pozostałych. Mogą one wynikać z błędów pomiarowych, ale też stanowić istotne sygnały o nietypowych zjawiskach, dlatego ich automatyczne usuwanie bez wcześniejszej analizy kontekstu nie jest zalecane. W praktyce do wykrywania wartości odstających stosuje się metody statystyczne, takie jak analiza pudełkowa (boxplot), z-score czy odległość Mahalanobisa. W inżynierii danych wykrywanie anomalii często wspierane jest również przez metody uczenia maszynowego, w tym algorytmy izolacji anomalii (Isolation Forest) lub klasteryzacji (np. DBSCAN).
Istotnym etapem przygotowania danych jest ich standaryzacja i normalizacja. Standaryzacja polega na przekształceniu zmiennych w taki sposób, aby miały one średnią równą zeru i odchylenie standardowe równe jeden, natomiast normalizacja sprowadza wartości zmiennych do określonego przedziału, najczęściej [0,1]. Działania te są szczególnie ważne w przypadku metod analizy wielowymiarowej i algorytmów uczenia maszynowego, które są wrażliwe na skalę danych. Przykładowo, w modelu regresji logistycznej czy analizie skupień nieprzeskalowane dane mogą prowadzić do błędnej interpretacji znaczenia poszczególnych zmiennych.
W procesie przygotowania danych istotne znaczenie ma również transformacja zmiennych. Transformacje te mogą mieć charakter matematyczny (np. logarytmiczna, potęgowa, pierwiastkowa) lub kategorialny (np. kodowanie zmiennych jakościowych przy użyciu one-hot encoding czy label encoding). Dzięki transformacjom możliwe jest dostosowanie danych do wymagań określonych metod analitycznych oraz poprawa ich interpretowalności. Na przykład zmienna o silnie skośnym rozkładzie może zostać przekształcona logarytmicznie, aby zbliżyć jej rozkład do normalnego, co ułatwia stosowanie metod parametrycznych.
Kolejnym krokiem jest ujednolicanie formatu danych, zwłaszcza w sytuacji, gdy pochodzą one z wielu różnych źródeł. Różnice w jednostkach miary, formatach zapisu dat, separatorach dziesiętnych czy sposobach reprezentacji wartości tekstowych mogą prowadzić do błędów w analizie, jeśli nie zostaną wcześniej zharmonizowane. W tym celu stosuje się procedury integracji danych, które umożliwiają ich konsolidację w jednym spójnym zbiorze.
W praktyce inżynierii danych coraz częściej stosuje się również metody redukcji wymiarowości, których celem jest ograniczenie liczby zmiennych przy zachowaniu jak największej ilości informacji. Wśród najczęściej wykorzystywanych metod znajdują się analiza głównych składowych (PCA – Principal Component Analysis) oraz analiza czynnikowa. Techniki te pozwalają na uproszczenie struktury danych, ułatwiając wizualizację i przyspieszając proces uczenia modeli predykcyjnych.
Ostatnim, ale nie mniej ważnym etapem jest walidacja jakości przygotowanego zbioru danych. Obejmuje ona zarówno ocenę poprawności przeprowadzonych przekształceń, jak i sprawdzenie spójności pomiędzy poszczególnymi zmiennymi. W tym kontekście szczególnie przydatne są wizualizacje danych – histogramy, wykresy

rozrzutu czy macierze korelacji – które pozwalają na szybkie wykrycie nieprawidłowości lub nielogicznych zależności.
Weryfikacja i przygotowanie danych to proces iteracyjny, który często wymaga wielokrotnego powrotu do wcześniejszych etapów. Tylko dane rzetelnie oczyszczone, ujednolicone i odpowiednio przekształcone mogą stanowić solidny fundament dla dalszych etapów analizy, takich jak modelowanie statystyczne, eksploracja danych czy uczenie maszynowe. Dlatego etap ten, choć często czasochłonny, ma kluczowe znaczenie dla jakości i wiarygodności końcowych wyników.

2. Metody opis zbioru danych

2.1. Opis tabelaryczny

Zgromadzony zbiór danych w swojej pierwotnej postaci jest najczęściej nieuporządkowany, chaotyczny, mało przejrzysty, przez co nie nadaje się bezpośrednio do analizy. Dlatego pierwszym etapem pracy analityka lub inżyniera danych jest jego uporządkowania i przygotowanie danych do dalszego przetwarzania. W przypadku zmiennych jakościowych (opisowych, kwalitatywnych)proces ten polega najczęściej na pogrupowania (agregacji)obserwacji według wariantów danej zmiennej, natomiast w przypadku zmiennych ilościowych(liczbowych, kwantytatywnych) na uporządkowaniu wartości zmiennej od najmniejszej do największej. Prowadzi to do uzyskania uporządkowanego szeregu danych zwanego prostym (szczegółowym). W przypadku dużych zbiorów danych taka forma prezentacji bywa obszerna, nieczytelna a przez to trudna w interpretacji. Problem ten można rozwiązać poprzez budowę szeregów rozdzielczych, w których zbiór obserwacji dzielony jest na określoną liczbę klas (przedziałów)według wartości badanej zmiennej. Każdej klasie przyporządkowuje się następnie liczbę jednostek (rekordów)należących do niej. Szeregi rozdzielcze mogą mieć postać szeregu punktowego bądź przedziałowego.
W szeregu rozdzielczym punktowym każda klasa zawiera tylko jedną wartość (wariant) zmiennej liczbowej(opisowej) i w związku z tym taki szereg posiada tyle klas, ile unikalnych wartości (wariantów)zmiennej.
W szeregu z przedziałami klasowymi dokonuje się grupowania wartości zmiennej w przedziały klasowe określając ich dolną i górną granicę. Szeregi przedziałowe są najczęściej tworzone dla zmiennych liczbowych ciągłych, chociaż należy zaznaczyć, iż przedziały takie można również utworzyć dla zmiennych liczbowych dyskretnych, a nawet opisowych.

Aby zbudować szereg z przedziałami klasowymi, należy określić:
a) liczbę klas,
b) szerokość (rozpiętość, długość) przedziałów,
c) granice poszczególnych przedziałów.
Przy podejmowaniu decyzji o liczbie klas należy mieć na uwadze, iż niema jednej uniwersalnej reguły w tym zakresie. Formułowane są jedynie pewne zalecenia, by liczba klas nie była ani zbyt duża, ani zbyt mała i była uzależniona głównie od liczby obserwacji. Zgodnie z tą sugestią wielu autorów proponuje wyznaczać orientacyjną (przybliżoną)liczbę klas według wzoru:

$$k \approx \sqrt{n} \tag{1}$$

gdzie:
k – przybliżona liczba klas,
n –liczebność badanej zbiorowości.

Uzyskaną według powyższej formuły liczbę klas należy traktować jako orientacyjną, a rzeczywista liczba klas może być o jedną niższa bądź wyższa od wielkości *k*. Takie postępowanie odnosi się do zmiennych liczbowych ciągłych, w przypadku zmiennych liczbowych skokowych bądź opisowych o liczbie klas decyduje głównie liczba wariantów zmiennej.
Podobnie przy określaniu szerokości przedziałów klasowych dąży się do tego, aby miały jednakową rozpiętość, co ułatwia dalszą analizę danych, np. obliczanie parametrów opisowych.
Należy również zwrócić uwagę, by w szeregu nie występowały przedziały „puste", tzn. klasy, dla których nie można przyporządkować żadnej obserwacji.
Uwzględniając powyższe zastrzeżenia przybliżoną szerokość przedziałów klasowych (*h*) można wyznaczyć ze wzoru:

$$h \approx \frac{R(X)}{k} \tag{2}$$

gdzie:
R(*X*) – rozpiętość wartości zmiennej obliczona według wzoru $R(X) = x_{max} - x_{min}$,
k – przyjęta liczba klas.

Taka reguła postępowania będzie miała miejsce w przypadku zmiennych ciągłych. W przypadku zmiennych opisowych bądź liczbowych skokowych równa rozpiętość klas będzie oznaczała jednakową liczbę wartości (wariantów)zmiennej w każdej klasie.
W praktyce inżynierii danych te obliczenia często wykonuje się automatycznie w środowiskach takich jak Python (biblioteki pandas, numpy) czy R, gdzie dostępne są funkcje do generowania histogramów i grupowania danych (pd.cut(), pd.qcut()).
Konieczność określenia zasad zamykania poszczególnych klas odnosi się do cech ciągłych. Przyjmuje się, iż każda z klas musi być jednostronnie zamknięta (z góry

bądź z dołu), co powoduje, że są one rozłączne i umożliwiają jednoznaczne przyporządkowanie poszczególnych obserwacji konkretnym przedziałom. Skonstruowany szereg rozdzielczy (zarówno punktowy jak i przedziałowy) może mieć postać szeregu liczebności (prostej), liczebności skumulowanej, częstości (prostej) bądź częstości skumulowanej.

W szeregu liczebności bądź częstości każdej klasie zostaje przyporządkowana określona liczba jednostek statystycznych lub ich udział (najczęściej wyrażony w procentach) w całym zbiorze danych. W szeregach skumulowanych każdej klasie przyporządkowana jest odpowiadająca jej liczebność bądź częstość skumulowana. Zaprezentowane powyżej szeregi można zaliczyć do tzw. szeregów strukturalnych, umożliwiają one bowiem określenie struktury badanego zjawiska ze względu na analizowaną zmienną w danym momencie. Odmianą takich szeregów strukturalnych są szeregi przestrzenne, w których struktura analizowanego zjawiska określana jest w oparciu o jego terytorialne rozmieszczenie. Ten typ szeregu ilustruje tablica 1.

Tabela1. Banki ogółem w krajach Unii Europejskiej w 2020

Kraj	**Liczba banków**
Niemcy	1508
Polska	621
Austria	492
Włochy	475
Francja	408
Irlandia	301
Finlandia	228
Hiszpania	192
Szwecja	154
Portugalia	144
Luksemburg	129
Dania	100
Holandia	87
Belgia	83
Litwa	81
Rumunia	71
Czechy	57
Łotwa	50
Węgry	42
Estonia	39
Grecja	35
Cypr	29
Słowacja	27
Bułgaria	24
Chorwacja	24
Malta	24
Słowenia	16

W literaturze wyodrębnia się dodatkowo szeregi: dynamiczne (chronologiczne, czasowe) obrazujące kształtowanie się zmiennej w czasie.
Przykład takiego szeregu zawiera tablica 2.

Tabela 2. Liczba pojazdów silnikowych w latach 2015 – 2024

Rok	Liczebność
2015	27 409 106
2016	28 601 037
2017	29 634 928
2018	30 800 790
2019	31 989 313
2020	32 991 083
2021	34 030 267
2022	34 866 137
2023	35 915 484
2024	30 033 036

Podane wyżej przykłady dotyczą prezentacji tabelarycznej badanego zjawiska opisywanego ze względu na jedną zmienną – tabela prosta. W inżynierii danych często analizuje się wiele zmiennych jednocześnie. W takim przypadku opis tabelaryczny będzie polegał na ujęciu zebranych danych w postaci tablicy złożonej (wielowymiarowej), wśród których wyróżnić można tabele zbiorcze (ukazujące relacje między kilkoma zmiennymi) i kombinowane (łącznie prezentujące dane z różnych przekrojów). W tablicy złożonej nie powinno być prezentowane zbyt dużo danych co może wpływać negatywnie na ich analizę.
Właściwie skonstruowana tablica powinna składać się z tytułu, makiety tablicy oraz źródła danych.
Tytuł tablicy powinien precyzyjnie określać badane zjawisko pod względem rzeczowym, czasowym i przestrzennym oraz zawierać ujęte w tablicy zmienne.
Makieta tablicy (zwana również tablicą właściwą) składa się z wierszy i kolumn oraz ich tytułów (tytuły wierszy określa się „boczkiem” tablicy, zaś tytuły kolumn „główką” tablicy).
Wnętrze tablicy, czyli „pola” znajdujące się na skrzyżowaniach poszczególnych wierszy i kolumn są wypełniane zgromadzonymi danymi statystycznym. Należy tu zaznaczyć, iż każde pole tablicy musi być bezwzględnie wypełnione. Jeśli z różnych względów nie ma możliwości wypełnienia pola tablicy danymi liczbowymi wówczas wykorzystywane są odpowiednie znaki umowne. Należą do nich:
„ – „ (kreska) - zjawisko nie występuje,
„ . „ (kropka) - zupełny brak informacji lub brak informacji wiarygodnych,
„0” (zero) - zjawisko występuje w niewielkich ilościach (mniej niż 50% przyjętej jednostki miary),

„×” (ukośny krzyżyk) - wypełnienie danego pola ze względu na układ tablicy jest niemożliwe bądź niecelowe,
„Δ” (pusty trójkąt) - nazwy zostały skrócone w stosunku do obowiązującej klasyfikacji,
„▲” (pełny trójkąt) - dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej,
„w tym” - nie podaje się wszystkich składników sumy.

Bezpośrednio pod tablicą umieszcza się źródło danych zawartych w tablicy. Źródło takie mogą stanowić: rocznik statystyczny, wyniki spisu, badania własne, sprawozdawczość firmy bądź instytucji, itp.
Ilustracją prawidłowo skonstruowanej tablicy opisującej zbiorowość studentów ze względu na trzy zmienne jest tabeli 3.

Tabela3. Studenci Wyższej Szkoły Pedagogicznej w Warszawie według płci, wieku i miejsca zamieszkania (stan na 30.09.2023 r.)

Wyszczególnienie		**Wiek w latach**			
		Razem	**19 – 21**	**21 – 23**	**23 – 25**
Warszawa	Kobiety	717	211	185	321
	Mężczyźni	461	122	252	87
Woj. Mazowieckie	Kobiety	408	187	124	97
	Mężczyźni	285	142	95	48
Inne województwa	Kobiety	102	45	32	25
	Mężczyźni	82	67	15	-
Razem		-	2055	703	578

2.2. Opis graficzny

Opis graficzny stanowi jeden z podstawowych sposobów prezentacji i analizy danych. Polega na przedstawieniu zgromadzonych informacji w formie wykresów, diagramów lub map, które umożliwiają szybkie i intuicyjne rozpoznanie zależności, trendów oraz prawidłowości w zbiorze danych. Wizualizacja danych jest niezwykle istotnym elementem pracy inżyniera danych, ponieważ pozwala nie tylko zrozumieć zjawisko, ale również zweryfikować poprawność danych, wykryć wartości odstające, braki czy błędy pomiarowe.
W praktyce analitycznej wykresy rzadko funkcjonują samodzielnie – najczęściej stanowią uzupełnienie materiału liczbowego przedstawionego w formie szeregów, tabel.
Wybór rodzaju wykresu uzależniony jest od charakteru danych (ilościowych lub jakościowych) oraz celu analizy. Dobór odpowiedniego rodzaju wykresu jest kluczowy, ponieważ niepoprawna wizualizacja może prowadzić do błędnych interpretacji. Przykładowo, wykres liniowy nie powinien być stosowany do zmiennych

jakościowych, natomiast wykres kołowy traci sens, gdy liczba kategorii jest zbyt duża.
Do najczęściej wykorzystywanych form graficznych zalicza się wykresy:
Wykresy kolumnowe i słupkowe – stosowane do prezentacji danych jakościowych lub porównania wartości liczbowych między grupami (np. liczba klientów w poszczególnych regionach).
Wykresy liniowe (diagramy) – używane głównie do przedstawiania zmian wartości zmiennej w czasie, np. dynamiki sprzedaży, temperatury, zużycia energii.
Wykresy punktowe (scatter plot) – umożliwiają analizę zależności pomiędzy dwiema zmiennymi ilościowymi, np. między wiekiem a dochodem, zużyciem paliwa a masą pojazdu.
Wykresy powierzchniowe i bryłowe (3D) – służą do prezentacji danych trójwymiarowych lub dużych zbiorów danych o skomplikowanej strukturze.
Wykresy obrazkowe i piktogramy – stosowane w analizach popularnonaukowych lub prezentacjach, gdzie istotna jest czytelność i atrakcyjność wizualna.
Mapy i wykresy kartograficzne – wykorzystywane w analizie przestrzennej danych, np. rozkładu ludności, natężenia ruchu drogowego czy poziomu emisji zanieczyszczeń.
Wykresy kombinowane – łączą różne typy wizualizacji w jednym układzie, np. wykres słupkowo-liniowy, co pozwala pokazać zarówno strukturę, jak i trend.

Podobnie jak tabela, wykres składa się z kilku istotnych elementów, które wpływają na jego czytelność i wartość informacyjną.
Najważniejsze z nich to:

- tytuł – określa, czego dotyczy wykres, np. „Liczba aktywnych użytkowników w latach 2018–2024”;
- pole wykresu (wykres właściwy) – część zawierająca właściwą prezentację graficzną danych. To element, który decyduje o stopniu realizacji celu wizualizacji, ponieważ przedstawia zjawisko w formie obrazu;
- osie wykresu i ich opisy – zawierają jednostki miary i wartości, które umożliwiają prawidłowy odczyt danych;
- legenda – wyjaśnia stosowane symbole, kolory, znaki i oznaczenia, ułatwiając interpretację wykresu;
- źródło danych – informuje o pochodzeniu materiału empirycznego.

Dobrze skonstruowany wykres powinien w sposób czytelny, komunikatywny i estetyczny przedstawiać dane, ułatwiając ich interpretację bez potrzeby szczegółowego odwoływania się do tabel czy wzorów, stąd przy konstruowaniu wykresów należy przestrzegać kilku zasad:

- skala osi powinna być proporcjonalna i rozpoczynać się od zera (chyba że uzasadnione jest inaczej);
- kolory i symbole należy stosować konsekwentnie i z umiarem, unikając nadmiernej liczby elementów;
- podpisy i legendy muszą być czytelne i jednoznaczne.

Wizualizacja nie może zniekształcać danych – wykres powinien wiernie odzwierciedlać rzeczywiste relacje liczbowe.

2.3. Opis parametryczny

Opis parametryczny stanowi jeden z najczęściej wykorzystywanych sposobów opisu zbioru danych, głównie z uwagi na jego syntetyczny i skondensowany charakter. Ta forma opisu wykorzystuje parametry opisowe, tj. charakterystyki liczbowe, które opisują rozkład wartości badanej zmiennej w danym zbiorze danych. Przy ich pomocy dokonuje się opisu określonych własności analizowanego zbioru wartości zmiennej.
Opis parametryczny może obejmować:
a) określenie średniego poziomu zbioru wartości zmiennej, czyli wybór jednej wartości reprezentatywnej dla całego zbioru danych. W tym celu stosuje się miary tendencji centralnej.
W literaturze spotykany jest najczęściej następujący podział tej grupy miar na:

- miary średnie klasyczne – to takie, które są wyznaczane na podstawie całego zbioru danych (innymi słowy: wszystkie obserwacje w zbiorze danych mają wpływ na wartość tej miary).

Do najczęściej stosowanych miar klasycznych należą: średnia arytmetyczna, średnia geometryczna, średnia harmoniczna oraz średnie potęgowe.

- miary pozycyjne – to takie, których wartość zależy wyłącznie od pozycji obserwacji w uporządkowanym zbiorze danych (a więc nie wszystkie obserwacje mają wpływ na ich poziom). Warto podkreślić, że nie wszystkie z nich można traktować jako „średnie” w ścisłym znaczeniu tego słowa. Część z nich ma raczej charakter miar położenia, które opisują rozmieszczenie wartości w zbiorze danych. Podobną rolę pełnią wykorzystywane w opisie parametrycznym momenty rozkładu. Najczęściej stosowane pozycyjne miary średnie to: wartość dominująca (dominanta, moda) oraz wartość środkowa (mediana). Spośród miar należy wymienić przede wszystkim kwartale (pierwszy, drugi i trzeci) oraz decyle i percentyle.

b) określenie poziomu zróżnicowania (zmienności, rozproszenia, dyspersji) w zbiorze danych, co pozwala ocenić, na ile wartości poszczególnych obserwacji różnią się od wartości centralnej. W tym celu stosuje się miary zmienności, tj.: obszar zmienności

(rozstęp), wariancję, odchylenie standardowe, odchylenie przeciętne, współczynnik zmienności.
c) określenie poziomu odchylenia analizowanego rozkładu danych od rozkładu symetrycznego; badanie tej własności dokonywane jest przy pomocy miar skośności bądź asymetrii, wśród których można wyróżnić miary pozycyjne oraz klasyczne.
Do pozycyjnych zalicza się mierniki oparte na badaniu relacji między miarami przeciętnymi, w tym pozycyjnymi. Zalicza się do nich:
a) różnicę między średnią arytmetyczną i dominantą bądź medianą,
b) kwartylowy współczynnik skośności,
c) standaryzowany współczynnik skośności.
Jako klasyczną miarę skośności wykorzystuje się różne postacie momentu centralnego trzeciego rzędu.
a) określenie poziomu skupienia bądź nierównomierności rozłożenia wartości cechy mierzone przy pomocy miar koncentracji,
b) określenie rozmieszczenia realizacji wartości cechy w analizowanym zbiorze poprzez wyznaczenie odpowiednich momentów statystycznych

3. Analiza współzależności zmiennych

Jednostki tworzące zbiorowość można scharakteryzować za pomocą więcej niż jednej zmiennej. Zmienne te wzajemnie się warunkują, co powoduje konieczność odpowiedzi na pytania:

- ✓ czy między badanymi zmiennymi zachodzi jakaś zależność?
- ✓ jaka jest ich siła, kształt i kierunek?

Współzależność między zmiennymi może być:

- ✓ funkcyjna
- ✓ stochastyczna(probabilistyczna).

Zależność funkcyjna polega na tym, że zmiana wartości jednej zmiennej powoduje ściśle określoną zmianę wartości drugiej zmiennej.
Zależność statystyczna występuje wtedy, gdy wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem zależności stochastyczne jest zależność korelacyjne. Polega ona na tym, że określony wartością jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennych.
Jeżeli między badanymi zmiennymi nie występuje zależność stochastyczna, to nie występuje między nimi również zależność korelacyjna. Podkreślenia wymaga fakt, że badanie zależności jest uzasadnione, gdy między zmiennymi istnieje więź

przyczynowo-skutkowa. Najpierw uzasadnić należy logiczne występowanie zjawiska, a dopiero potem przystąpić do analizy zależności.
Istnienie wiele różnorodnych miar korelacji zarówno dla cech ilościowych, jak i jakościowych. Do najważniejszych miar można zaliczyć:

- ✓ współczynnik korelacji liniowej Pearsona,
- ✓ współczynnik kolejnościowy rang Spearmana,
- ✓ współczynnik zbieżności Czuprowa,
- ✓ współczynnik korelacji rang Kendalla,
- ✓ współczynnik kontyngencji C-Pearsona,
- ✓ współczynnik φ Yule'a.

Współczynnik korelacji Pearsona

Współczynnik korelacji liniowej Pearsona (r_{xy}) jest miarą opisową, która określa zarówno siłę, jak i kierunek zależności korelacyjnej pomiędzy dwoma zmiennymi ilościowymi *X* i *Y* gdy związek pomiędzy nimi jest liniowy. Współczynnik korelacji Pearsona oparty jest na
tzw. kowariancji zmiennych – *cov*(*XY*), którą można zapisać w postaci:

$$\boldsymbol{cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}} \tag{3}$$

gdzie:
x_i – wartość cechy *X*,
y_i – wartość cechy *Y*,
$\bar{x}$– średnia arytmetyczna cechy *X*,
$\bar{y}$ – średnia arytmetyczna cechy *Y*,
n – liczba obserwacji.

Współczynnik korelacji Pearsona oblicza się według wzoru:

$$\boldsymbol{r_{xy} = r_{yx} = \frac{cov(X,Y)}{S(X)S(Y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}} \tag{4}$$

gdzie:
$S(X)$ – odchylenie standardowe cechy *X*,
$S(Y)$ – odchylenie standardowe cechy *Y*.

Współczynnik korelacji liniowej Pearsona jest miarą unormowaną, która przyjmuje wartość z przedziału:

$$-1 \leq r_{xy} \leq +1 \tag{5}$$

Kierunek współzależności określa znak współczynnika korelacji: znak (–) oznacza korelację ujemną, z kolei znak (+) korelację dodatnią. Im wartość współczynnika jest bliższa 1 (lub -1) tym zależność korelacyjna cech jest silniejsza, natomiast im wartość współczynnika zbliża się do 0 wówczas siła korelacji jest coraz słabsza. W przypadku braku zależności współczynnik r_{xy} = 0. W sytuacji, kiedy pomiędzy zmiennymi r_{xy} = 1 lub r_{xy} = -1 mówi się o tzw. zależności doskonałej.

Współczynnik rang Spearmana

Współczynnik korelacji rang Spearmana ($\boldsymbol{r_s}$) służy do opisu siły korelacji zarówno cech ilościowych, jak i jakościowych w sytuacji, kiedy jest możliwość uporządkowania ich wariantów. Wzór na obliczenie współczynnika kolejnościowego rang Spearmana można zapisać w postaci:

$$\boldsymbol{r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}} \tag{6}$$

gdzie:
d_i – różnica pomiędzy rangami zmiennych *X* i *Y* tzn. $d_i = x_i - y_i$,
n – liczba par zmiennych *X* i *Y*.

Rangowanie polega na uporządkowaniu wartości cech *X* i *Y* w kolejności rosnącej lub malejącej. Następnie nadaje się im tzw. rangi: 1, 2, 3 …, *n*. W sytuacji, kiedy w uporządkowanym szeregu wystąpią identyczne wartości wówczas oblicza się średnią arytmetyczną z ich kolejnych numerów i przyporządkowuje się im otrzymaną wartość. Współczynnik korelacji rang przyjmuje wartości z przedziału

$$-1 \leq r_s \leq +1 \tag{7}$$

Im współczynnik r_s jest bliższy 1 (lub -1) związek jest silniejszy, z kolei, kiedy jest bliższy 0związek korelacyjny jest coraz słabszy.
Współczynnik korelacji rang Spearmana stosuje się, kiedy liczba obserwacji wynosi n < 30.

Przykład 1

W tablicy 4 przedstawiono dziewięć krajów europejskich o powierzchni powyżej 300 tys. km^2 (*X*) i odpowiadającej tym krajom liczbie ludności w tys. osób (*Y*). Za pomocą współczynnika korelacji rang należy zbadać, czy istnieje zależność między liczbą ludności, a powierzchnią tych krajów.

Tabela 4. Wybrane kraje Europy według powierzchni i ludności w 2005 r. oraz obliczenia pomocnicze

Kraje	*X*	*Y*	Ranga X	Ranga Y	d_i	d_i^2
Finlandia	338,1	5244	7	8	-1	1
Francja	544	60733	2	2	0	0
Hiszpania	506	44079	3	5	-2	4
Niemcy	357	82443	6	1	5	25
Norwegia	385,2	4617	5	9	-4	16
Polska	312,7	38161	8	6	2	4
Szwecja	450,3	9024	4	7	-3	9
Ukraina	603,6	47075	1	4	-3	9
Włochy	301,3	57989	9	3	6	36
			45	45	0	104

Obydwie cechy *X* i *Y* to stymulanty, które uporządkowano od największej wartości do najmniejszej. Tak uporządkowanym wartością nadano rangi dla największej wartości 1 itd. Podstawiając do wzoru (6), obliczamy współczynnik korelacji rang:

$$r_s = 1 - \frac{6 \cdot 104}{9(9^2 - 1)} = 1 - \frac{624}{9 \cdot 80} = 1 - 0,87 \approx 0,13$$

Otrzymany wynik wskazuje, że w badanej zbiorowości dziewięciu krajów europejskich brak współzależności między powierzchnią i liczbą ludności.

3.1.Cechy jakościowe

Analizy oparte na teście chi-kwadrat należą do powszechnie stosowanych i względnie łatwych metod oceny współzależności między zmiennymi jakościowymi. Nazwa testu pochodzi od rozkładu teoretycznego oznaczanego greckim symbolem χ^2.

Algorytm postępowania, wspólna dla badań zależności między zmiennymi jest następująca:

- ✓ formułuje się pewne przypuszczenia co do badanej populacji. Przypuszczenia te muszą się wzajemnie wykluczać. Przypuszczenia te nazywa się hipotezami;
- ✓ zapisuje się częstość występowania badanych zdarzeń w tablicy tzw. wielodzielnych (kontyngencji, krzyżowych, asocjacji)w miejscu przecięcia się odpowiednich wierszy i kolumn. Są to tzw. liczebności zaobserwowane – czyli n_i;
- ✓ na podstawie liczebności zaobserwowanych wyliczane są liczebności teoretyczne – czyli $\hat{n}_i$, które są prawdziwe dla określonych hipotez;
- ✓ na podstawie różnic pomiędzy liczebnościami zaobserwowanymi i oczekiwanymi wylicza się wartość statystyki χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i^2)}{\hat{n}_i} - n \tag{8}$$

✓ porównuje się wartość wyliczonej statystyki z wartością statystyki odczytanej z tablic rozkładu dla *k*- 1 liczby stopni swobody (liczba wariantów, klas)

3.1.1.Tablica czteropolowa

Tablice wielodzielne mogą być kwadratowe lub prostokątne zależnie od liczby pól**.** W przypadku tablicy o wymiarach 2 x 2 do wyznaczenia statystyki χ^2 można zastosować wzór:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{9}$$

Tabela5. Przykład tablicy wielodzielnej kwadratowej – czteropolowej (2 x 2)

X \ Y	y_1	y_2	n_i
x_1	a	b	a + b
x_2	c	d	c + d
n_j	a+c	b+d	*n*

gdzie: x_1 i x_2 – warianty cechy *X*; y_1 i y_2 – warianty cechy *Y*; a – liczba jednostek występująca w wariancie x_1 i $y_1$dla cech *X*i *Y*; b – liczba jednostek występująca w wariancie x_1 i y_2 dla cech *X* i *Y*; c – liczba jednostek występująca w wariancie x_2 i $y_1$dla cech *X* i *Y*; d – liczba jednostek występująca w wariancie x_2 i y_2 dla cech X i Y; *n* – liczebność próby.

Jeżeli w tablicy czteropolowej przynajmniej jedna liczebność jest mniejsza niż 10, to należy zastosować wzór z poprawką Yatesa:

$$\chi^2 = \frac{n(|ad - bc| - 0,5)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{10}$$

Przykład 2

Badano zależność między płcią i uzależnieniem od papierosów. Grupa 100 osób podzielona została w następujący sposób: kobiety - K, mężczyźni - M, palący – P, niepalącym N.

Tabela 6.

X \ Y	**P**	**N**	n_i
K	21	32	53
M	29	18	47
nj	50	50	100

Sformułowano hipotezę zerową, że płeć nie wpływa na uzależnienie.
Obliczono wartość statystyki:

$$\chi^2 = \frac{\mathbf{100(378-928)^2}}{\mathbf{(21+32)(21+29)(32+18)(29+18)}}$$

Liczba stopni swobody w tablicy czteropolowej wynosi 1. Z tablic rozkładu odczytano wartość krytyczną dla α = 0,05. Wartość ta wynosi 3,84. Hipoteza zerowa została więc odrzucona. Istnieje zależność między płcią, a nałogiem nikotynowym.

3.1.1.Tablica wielopolowa

Sytuacja trochę się komplikuje, gdy tablica jest wielopolowa. Idea oczywiście nie ulega zmianie, aczkolwiek obliczenia staja się żmudniejsze. Do weryfikacji hipotezy H_0 o niezależności stochastycznej zmiennych wykorzystuje się w tym przypadku wzór:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}^2}{\hat{n}_{ij}} - N \tag{11}$$

Statystyka ta przy założeniu prawdziwości H_0 – dla każdej próby rozkład χ^2z (r-1)(k-1) stopniami swobody, gdzie k – liczba wierszy, r – liczba kolumn w tablicy.
Liczebność teoretyczna $\hat{n}_{ij}$ oblicza się ze wzoru:

$$\hat{n}_{ij} = \frac{\begin{pmatrix}\textbf{suma liczebności}\\ \textbf{empirycznych}\\ \textbf{i-tego wiersza}\end{pmatrix}\begin{pmatrix}\textbf{suma wartości}\\ \textbf{empirycznych}\\ \textbf{j-tej kolumny}\end{pmatrix}}{\textbf{liczebność próby}} \tag{12}$$

Przykład 3
W 500-osobowej losowo dobranej grupie ludzi przeprowadzano badanie ankietowe, mające na celu uzyskać odpowiedzi na pytanie: czy istnieje zależność między miejscem zamieszkania preferencjami politycznymi. Wyniki przedstawiono w tabeli 7.

Tabela 7. Wyniki badań ankietowych

Miejsce zamieszkania (*X*)	**Ogółem**	**Preferencje polityczne (*Y*)**			
		SLD	**PO**	**PiS**	**LPR**
Ogółem	500	121	135	155	89
Miasto	276	37	97	124	18
Wieś	224	84	38	31	71

Za pomocą testu χ^2 zweryfikować hipotezę o niezależności zmiennych *X* i *Y*. Przyjąć poziom istotności α =0,05.
Przykładowe oczekiwane liczebności wynoszą:

$$\hat{n}_{11} = \frac{121 \cdot 276}{500} = 66,79, \quad \hat{n}_{12} = \frac{135 \cdot 276}{500} = 74,52, \quad \hat{n}_{31} = \frac{155 \cdot 276}{500} = 85,56.$$

Tabela 8. Tablica liczebności oczekiwanych

Zamieszkanie (*X*)	Preferencje polityczne (*Y*)			
	SLD	PO	PiS	LPR
Miasto	66,79	74,5	85,6	49,13
Wieś	54,21	60,5	69,4	39,87

Przystępując do obliczeń wartości χ^2, otrzymuje się:

$$\chi^2 = \left(\frac{37^2}{66,79} + \frac{97^2}{74,5} + \frac{124^2}{85,6} + \cdots + \frac{38^2}{60,5} + \frac{31^2}{69,4} + \frac{71^2}{39,87} + \right) - 500 = 127,4.$$

Dla poziomu istotności α =0,05 i 3 stopni swobody [(2 - 1)(4 - 1) = 3] wartość krytyczna$\chi_k^2 = 7,82$.
Oczywiście: $\chi^2 = 127,4 > \chi_k^2 = 7,82$, co powoduje, że należy odrzucić hipotezę zerową o niezależności zmiennych. Miejsce zamieszkania ma więc wpływ na preferencje polityczne.

3.1.3.Miary zależności

Współczynnik φ Yule'a
Służy do badania związku dwóch cech jakościowych i jest określany wzorem

$$\varphi = \sqrt{\frac{\chi^2}{n}} \tag{13}$$

W celu dokonania obliczeń konieczna jest znajomość statystyki χ^2, jednakże dla tablic o wymiarach 2 x 2 istnieje możliwość bezpośredniego obliczania współczynnika φ. Współczynnik φmoże teoretycznie przyjmować wartości z przedziału od -1 do +1,w przypadku braku zależności między zmiennymi φ = 0.Skrajne wartości, współczynnik Yule'a przyjmuje tylko w przypadku, gdy *a* =*d* = 0 lub *b* = *c* = 0. W innych przypadkach współczynnik nie przyjmuje wartości skrajnych, nawet dla silnego związku cech. Wynika to z faktu, że końcowe wartości współczynnika zależą od uszeregowania liczebności w poszczególnych polach tablicy korelacyjnej.
Znak współczynnika φ nie informuje o kierunku zależności, gdyż zależy od sposobu uporządkowania wariantów zmiennej w tablicy.

Przykład 4

Wiadomo, że istnieje zależność między płcią a uzależnieniem od palenia papierosów. Jak silna jest to zależność?

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{4,86}{100}} = 0,22.$$

Wartość współczynnika wskazuje na słabą współzależność między płcią i uzależnieniem od palenia.

Współczynnik zbieżności *T*-Czuprowa

Miernik ten również oparty jest na teście χ^2. Określa go wzór:

$$T_{XY} = T_{YX} = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}} \tag{14}$$

gdzie:

r – liczba wierszy w tablicy kontyngencji,

k – liczba kolumn w tablicy kontyngencji.

Współczynnik ten przyjmuje wartość przedziału [0, 1], gdy badane zmienne są stochastyczne niezależne. Przy zależności funkcyjnej zmiennych *T* = 1. Im bardziej współczynnik zbieżności jest bliższy zeru, tym słabsza jest zależność między zmiennymi.

Współczynnik *T*-Czuprowa jest symetryczny$T_{XY} = T_{YX}$, czyli nie wskazuje kierunku zależności (jest zawsze dodatni).

Współczynnik zbieżności *V*-Cramera

Współczynnik zbieżności *V*-Cramera oblicza się według wzoru:

$$V = \sqrt{\frac{\chi^2}{n \cdot (min(r,k) - 1)}} \tag{15}$$

Podobnie jak w przypadku współczynnika *T*-Czuprowa współczynnik *V*-Cramera przyjmuje wartości z przedziału [0, 1] oraz im bardziej współczynnik zbieżności *V*jest bliższy jedności, tym silniejsza jest zależność między zmiennymi. Współczynnik ten również nie określa kierunku związku między zmiennymi.

Z powyższych rozważań wynika, że charakterystyki strukturalne pozwalają na dokładniejsze rozpoznanie relacji pomiędzy poszczególnymi elementami lub grupami

elementów analizowanego zbioru danych i pozwalają znaleźć odpowiedź na następujące pytania:

- jaki jest udział poszczególnych elementów w całym zbiorze danych,a w ślad za tym - jaki jest potencjalny wpływ ich kształtowania się na zachowanie analizowanego obiektu?
- na zachowanie których elementów i ich grup, ze względu na ich wpływ na zachowanie całej zbiorowości, należy zwrócić szczególną uwagę?
- w jakim stopniu struktura analizowanego obiektu ze względu na wyróżnione zmienne jest zbliżona lub oddalona od struktury obiektów spełniających podobne funkcje i dalej, czy te różnice mogą być przyczyną odmienności w uzyskiwanych efektach i ponoszonych nakładach?
- czy istnieją i jaka jest siła związków pomiędzy różnymi zmiennymi tej samej lub różnych zbiorów danych, np. pomiędzy stażem pracy poszczególnych pracowników a ich wydajnością, pomiędzy ilością wypożyczeń książek z biblioteki szkolnej a uzyskiwanymi przez uczniów ocenami z poszczególnych przedmiotów itp.?

4. Podstawy modelowania ekonometrycznego

4.1. Wprowadzenie

W dziedzinie eksploracji danych wielokrotnie istnieje potrzeba opisywania (modelowania)związków łączących duże ilości danych, które traktowane są jako wyjściowe (zależne, objaśniane), z dużymi ilościami danych, które traktowane są jako wejściowe (niezależne, objaśniające). Jako przykład można podać duże zbiory danych rejestrowanych w specjalistycznych systemach klasy DCS (ang. Distributed Control System), w które wyposażone są praktycznie wszystkie większe obiekty przemysłowe.
Bardzo wygodnym i efektywnym sposobem opisywania wspomnianych zależności są modele ekonometryczny, a w szczególności najprostsza ich wersja – regresja liniowa.

Z formalnego punktu widzenia model ekonometryczny jest równaniem(układem równań), które w sposób przybliżony przedstawia główne powiązania ilościowe, występujące pomiędzy badanym zjawiskiem a wpływającymi nań czynnikami. Oprócz czynników mających zasadniczy wpływ na badane zjawisko ekonomiczne są też takie, które trudno oszacować i których wpływ na badane zjawisko jest niewielki, trudny do oszacowania czy też czysto losowy, tzw. czynniki uboczne.

Badane czy też modelowane zjawisko nazywane jest zmienną objaśnianą (zależną), natomiast czynniki wpływające na kształtowanie się zmiennej objaśnianej zmiennymi objaśniającymi (niezależnymi).
Złożoność struktury zjawisk powoduje, że w modelu ekonometrycznym może być jedno równanie, wyjaśniające jedną zmienną objaśniającą poprzez wiele zmiennych objaśniających (lub poprzez jedna zmienną objaśniającą), ale może też w modelu znaleźć się wiele równań pokazujących związki pomiędzy zmiennymi objaśnianymi i zmiennymi objaśniającymi, stąd wyróżnia się modele jednorównaniowe oraz wielorównaniowe.

Należy bardzo wyraźnie podkreślić, ze model ekonometryczny buduje się w oparciu o pozyskane dane rzeczywiste (empiryczne) dotyczące wartości poszczególnych zmiennych.

W modelu jednorównaniowym może być uwzględniona jedna zmienna objaśniana (zależna) – pozostałe są zmiennymi objaśniającymi (niezależnymi).

Ekonometryczny model jednorównaniowy można przedstawić w postaci:

$$Y = f(X_1, X_2, \ldots, X_k, \varepsilon) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots \alpha_k X_k + \varepsilon \qquad (16)$$

gdzie:
Y – zmienna objaśniana, reprezentująca modelowane zjawisko,
$X_1, X_2, \ldots, X_k$ – zmienne objaśniające,
f– postać funkcji zmiennych objaśniających, zwykle określana w trakcie budowy modelu ekonometrycznego (np. funkcja liniowa, wielomian, funkcja logarytmiczna),
ε - składnik losowy (odchylenia losowe, zakłócenie, błąd).
Uwzględnienie w modelu ekonometrycznym odchyleń losowych (ε) nadaje mu stochastyczny charakter. Składnik losowy odzwierciedla wpływ zjawisk ubocznych, a jego wprowadzenie do modelu ekonometrycznego jest uwarunkowane poniższymi przyczynami:
– przyjęciem niewłaściwej postaci analitycznej funkcji,
– brakiem możliwości uwzględnienia w modelu wszystkich zmiennych objaśniających opisujących kształtowanie sie zmiennej objaśnianej,
–błędami wynikającymi z niedokładności pomiaru zmiennych,
– losowością postępowania podmiotów ekonomicznych, a zwłaszcza zachowań ludzkich. Przejawia sie to tym, ze ten sam konsument w obliczu tak samo sformułowanego dylematu wyboru każdorazowo może podjąć nieco inną decyzje.

Wyznaczenie postaci funkcji f sprowadza się do:
-estymacji (oszacowania) parametrów $\alpha_0, \alpha_1, \alpha_2, \ldots \alpha_k$ (parametry strukturalne)oraz

-estymacji parametrów struktury stochastycznej (parametrów rozkładu składnika losowego). Jest to model regresji liniowej.

Przykład 5

$Y = \alpha_0 + \alpha_1 X + \varepsilon$

gdzie:

Y – wydatki na osobę w rodzinie, zmienna objaśniana,

X – zarobki w rodzinie, zmienna objaśniająca,

α_0, α_1 – parametry strukturalne modelu,

ε - składnik losowy. Wyraża on tzw. błąd w równaniu, czyli wpływ na Y czynników nie uwzględnionych w modelu w sposób bezpośredni, takich jak warunki klimatyczne, zawartość cukru w burakach, stosowanie nawozów sztucznych itp. opisującym zasadnicze powiązania między określonymi wielkościami ekonomicznymi.

Model wielorównaniowy dotyczy zależności pomiędzy wieloma zmiennymi objaśnianymi i wieloma zmiennymi objaśniającymi.

Przykład 6

$Y = \alpha_0 + \alpha_1 X + \varepsilon_1$

$O = \alpha_2 X + \varepsilon_2$

gdzie:

O – oszczędności w rodzinie, zmienna objaśniana.

4.2. Etapy modelowania ekonometrycznego

Procedura budowy modelu ekonometrycznego ma wieloetapowy charakter. Model jest wynikiem postępowania łączącego w jedną całość metody matematyczne, statystyczne i informatyczne z wiedzą ekonomiczną i intuicją.

W procesie budowy modelu ekonometrycznego można wyróżnić pięć etapów, a mianowicie:

- ✓ specyfikację zmiennych,
- ✓ wybór analitycznej postaci modelu,
- ✓ estymację parametrów,
- ✓ weryfikację modelu,
- ✓ praktyczne wykorzystanie oszacowanego modelu.

4.2.1. Specyfikacja zmiennych, gromadzeniem danych

Specyfikacja zmiennych wraz z gromadzeniem danych obejmuje takie czynności, jak:
- ✓ dobór zmiennych wykorzystanych w badaniu (objaśnianych i objaśniających),
- ✓ zebranie informacji o wartościach zmiennych objaśnianych i objaśniających,
- ✓ graficzną analizę kształtowania się poszczególnych zmiennych oraz zależności zmiennych objaśnianych od zmiennych objaśniających,
- ✓ eliminację zmiennych objaśniających o małym współczynniku zmienności,
- ✓ eliminację liniowo zależnych zmiennych objaśniających.

Zmienne objaśniane wynikają z celów, jakim służyć ma model ekonometryczny. Określając zmienne objaśniające, należy w jak największym stopniu wykorzystać teorię ekonomiczną.
Jeżeli jednak brak teorii, która wyjaśniałaby mechanizmy kształtowania się zjawiska, to jedynie kierując się intuicją i doświadczeniem, można poszukiwać czynników je determinujących. Wstępnie warto wziąć pod uwagę więcej zmiennych, a następnie dokonać selekcji, wykorzystując dostępne metody.
Jeżeli wiadomo, jakie zmienne (przynajmniej potencjalnie) zostaną w modelu wykorzystane, to należy zebrać dane statystyczne. Można w tym celu posłużyć się danymi publikowanymi przez instytucje zajmujące się ich gromadzeniem (np. dane Głównego Urzędu Statystycznego, Narodowego Banku Polskiego, Eurostatu). Często jednak konieczne jest samodzielne poszukiwanie źródeł danych lub opracowanie badania pozwalającego je zdobyć (np. w postaci badania ankietowego). Pamiętać na tym etapie należy, że im więcej danych zgromadzimy, tym opracowany model jest wiarygodniejszy. Chodzi tu przede wszystkim o zapewnienie w miarę szerokiego przedziału wartości wziętych pod uwagę zmiennych.
Jeżeli teoria ekonomii nie określa postaci związków miedzy zmiennymi objaśnianymi a objaśniającymi, to konieczne jest jej założenie. Dane pozwalają zweryfikować (weryfikacja modeli dokonywana jest w etapie 3.),czy decyzja co do postaci zależności była słuszna. Wstępnie zwykle przyjmowana jest zależność liniowa. Jeśli nie jest odpowiednia, to konieczne jest wzięcie pod uwagę związków nieliniowych.

4.2.1.1. Wybór zmiennych objaśniających

Przystępując do badania ekonometrycznego, bierze się zwykle pod uwagę w miarę liczny zbiór zmiennych objaśniających, które przynajmniej potencjalnie wpływają na zjawisko reprezentowane zmienną objaśnianą. Może się jednak okazać, że uwzględnienie wszystkich nadmiernie komplikuje model, nie poprawiając jego jakości. Z tego też względu należy przeprowadzić selekcję, usuwając z wstępnego zbioru zmienne objaśniające, które nie wnoszą do modelu istotnych informacji..

Zmiennym uwzględnionym w modelu ekonometrycznych stawia się trzy podstawowe wymagania:

- ✓ powinna je cechować wysoka zmienność,
- ✓ powinny być silnie skorelowane ze zmienną objaśnianą,
- ✓ powinny być słabo skorelowane między sobą.

Zmienność zmiennej oceniana jest współczynnikiem zmienności:

$$v_k = \frac{S_k}{\bar{x}_k} \tag{17}$$

gdzie:
K – liczba zmiennych objaśniających,
$\bar{x}_k$ - średnia arytmetyczna zmiennej objaśniającej X_k, $k = 1, 2, \ldots, K$,
S_k – odchylenie standardowe zmiennej objaśniającej X_k.

Aby ocenić poziom zmienności, obiera się pewną (większą od zera) wartość krytyczną v^* współczynnika zmienności, np. $v^* = 0{,}1$. Zmienne spełniające nierówność $|v_i| \leq v^*$ uznaje się za quasi- stałe (prawie stałe) i eliminuje się ze zbioru potencjalnych zmiennych objaśniających. Zmienne te nie wnoszą istotnych informacji. Wprowadzenie do modelu zmiennych quasi-stałych oznacza w istocie modyfikacje wyrazu wolnego, co je działaniem zbędnym, zaciemniających obraz zależności.
Na wybór wartości krytycznej nie ma określonych zasad. Przyjmuje się ją według uznania, pamiętając o tym, że im niższa wartość krytyczna, tym więcej zmiennych objaśniających (w tym być może - wiele zmiennych o bardzo mało różniących się od siebie wartościach) wejdzie do opisu zmiennej objaśnianej.
W literaturze podawana jest minimalna wartość graniczna na poziomie 0,1.

Przykład 7

Do opisu produkcji przedsiębiorstwa w mld złoty (Y) zaproponowano wstępnie cztery wielkości X_1 - zatrudnienie w tys.osób, X_2 - wartość maszyn i urządzeń w mld zł, X_3 - czas przestoju maszyn w dniach, X_4 - nakłady inwestycyjne w mln zł. Wartości poszczególnych zmiennych w latach 2001 - 2010 podano w tabeli9.

Tabela 9.Wartości zmiennych modelu opisującego produkcję przedsiębiorstwa w latach 2001–2010

Lata	Y	X_1	X_2	X_3	X_4
'01	10	6	8	14	12
'02	10	6	8	14	12
'03	16	10	12	18	12
'04	16	10	12	18	14
'05	12	8	8	18	10
'06	14	10	8	18	12
'07	20	12	14	24	14
'08	20	12	16	24	12
'09	20	12	16	26	12
'10	22	14	18	26	10

Należy sprawdzić przy założonej wartości krytycznej współczynnika zmienności v^* = 0,15 czy potencjalne zmienne objaśniające odznaczają się odpowiednio wysoką zmiennością.

Rozwiązanie

Dla dowolnej zmiennej X_k, jej wartość średnią wyznacza się ze wzoru:

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \tag{18}$$

$$\bar{x}_1 = \frac{1}{10}(6+6+10+10+8+10+12+12+12+14) = 10,$$

$$\bar{x}_2 = 12,$$

$$\bar{x}_3 = 20,$$

$$\bar{x}_4 = 12,$$

natomiast odchylenie standardowe zmiennej $X_k$ze wzoru:

$$S_k = \left[\frac{1}{N}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2\right]^{1/2} \tag{19}$$

$$S_1 = \sqrt{\frac{1}{10}\cdot 64} = 2{,}53, \quad S_2 = \sqrt{\frac{1}{10}\cdot 136} = 3{,}688,$$

$$S_3 = \sqrt{\frac{1}{10}\cdot 192} = 4{,}382, \quad S_4 = \sqrt{\frac{1}{10}\cdot 16} = 1{,}265,$$

Współczynniki zmienności poszczególnych zmiennych przyjmują wartości:

$$v_1 = \frac{2{,}53}{10} = 0{,}253,\ v_2 = \frac{3{,}688}{12} = 0{,}307, v_3 = \frac{4{,}382}{20} = 0{,}219, v_4 = \frac{1{,}265}{12} = 0{,}105,$$

Ponieważ współczynnik zmienności zmiennej X_4 jest mniejszy od założonej wartości krytycznej v^*=0,15 zmienną tę uznaje się za quasi-stałą i eliminuje ze zbioru zmiennych objaśniających.

Pozostałe dwa wymagania opierają się na współczynnikach korelacji. Dla zmiennych ilościowych stosowany jest współczynnik korelacji Pearsona, który stanowi miarę liniowego związku między zmiennymi. Dla modeli nieliniowych, należałoby oceniać korelacje między zmiennymi transformowanymi.

Siłę zależności pomiędzy potencjalną zmienną objaśniającąXa zmienną objaśnianą Y można obliczyć ze wzoru:

$$\boldsymbol{r_k} = \frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2 \cdot (y_i - \bar{y})^2}} \tag{20}$$

gdzie:
n– liczba obserwacji, i = 1, 2, …, n
Współczynniki korelacji między zmienną zależną w wszystkimi zmiennymi niezależnymi tworzą następujący wektor:

$$\boldsymbol{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_K \end{bmatrix} \tag{21}$$

Natomiast współczynnik korelacji między zmiennymi niezależnymiX_k i X_l, obliczany jest jako:

$$r_{kl} = \frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k) \cdot (x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2 \cdot (x_{il} - \bar{x}_l)^2}} \tag{22}$$

Współczynnik korelacji między wszystkimi zmiennymi niezależnymi umieszcza się w macierzy zwanej macierzą korelacji **R**. Elementy przekątnej tej macierzy są równe 1, ponieważ współczynnik korelacji dowolnej zmiennej z samą sobą jest 1. Macierz korelacji jest ponadto macierzą symetryczną, ponieważ$r_{kl} = r_{lk}$, a zapisuje się ją jako:

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{l2} & \dots & r_{1K} \\ r_{21} & 1 & \dots & r_{2K} \\ \dots & \dots & 1 & \dots \\ r_{K1} & r_{K2} & \dots & 1 \end{bmatrix} \tag{23}$$

Eliminację zmiennych przeprowadzić można tradycyjnymi metodami rekomendowanymi w polskiej literaturze ekonometrycznej. Są to: metoda Nowaka

(metoda analizy współczynników korelacji), metoda Hellwiga (metoda nośników informacyjnych).Można też zastosować algorytmy genetyczne. Ich wykorzystanie wymaga jednak specjalistycznego oprogramowania.

Metoda Nowaka

W metodzie tej wykorzystuje się krytyczną wartość współczynnika korelacji, która. zależy od ilości obserwacji N oraz od poziomu istotności α, który zadajemy (najczęściej α = 0,05 lub α = 0,01). Wartość tę wyznacza się ze wzoru:

$$r^* = \left[\frac{(t_\alpha)^2}{(t_\alpha)^2 + n - 2}\right] \tag{24}$$

gdzie:

t_α- wartość statystyki t Studenta dla zadanego poziomu istotności αoraz dla *n*-2stopni swobody.

Jeżeli krytyczna wartość współczynnika korelacji *r** jest zbyt mała (co nastąpi, jeśli liczba obserwacji jest duża), to można założyć ją z góry.

<u>Postępowanie w metodzie Nowaka jest następujące:</u>

1. Ustalenie wartości krytycznej współczynnika korelacji.
2. Wyeliminowanie ze zbioru potencjalnych zmiennych objaśniających tych zmiennych, które są słabo skorelowane ze zmienną objaśnianą.

Słaba korelacja zmiennych oznacza, że zmiany wartości jednej zmiennej mają mały wpływ na zmiany wartości i drugiej zmiennej. Zatem, jeśli któraś z potencjalnych zmiennych objaśniających jest słabo skorelowana ze zmienną objaśnianą, należy ją wyeliminować.A zatem, jeśli dla którejś ze zmiennych X_k zachodzi związek

$$|r_k| \leq r^* \tag{25}$$

to taką zmienną należy wyeliminować ze zbioru potencjalnych zmiennych objaśniających. Należy zauważyć, że ten etap wystąpi tylko raz, to znaczy, że tylko raz eliminuje się zmienne, słabo skorelowane ze zmienną objaśnianą.

3. Spośród zmiennych, które pozostały, wybiera się tę, która jest najsilniej skorelowana ze zmienną objaśnianą

Do zbioru zmiennych objaśniających zalicza się tę spośród pozostałych, która jest najsilniej skorelowana ze zmienną objaśnianą, gdyż ta właśnie zmienna ma największy wpływ na zachowanie się zmiennej objaśnianej. Formalnie wybiera się tę zmienną X_k, która spełnia warunek:

$$|r_h| = \max_k\{|r_k|\} \tag{26}$$

4. Ze zbioru pozostałych potencjalnych zmiennych objaśniających eliminuje się te, które są silnie skorelowane z wybraną w kroku 3) zmienną X_h.

Formalnie eliminuje się te zmienne, dla których zachodzi związek

$$|r_h| > r^* \tag{27}$$

Uzasadnienie. Ponieważ w zbiorze zmiennych objaśniających znalazła się zmienna X_h, zatem nie ma sensu uznawać za zmienne objaśniające tych zmiennych, które są silnie skorelowane ze zmienną X_h, gdyż mają one podobny wpływ na zachowanie się zmiennej objaśniającej jak wspomniana zmienna. Można inaczej powiedzieć, że zmienne silnie skorelowane ze zmienną X_h niosą ze sobą takie same informacje, jak sama zmienna X_h.

5. Postępowanie opisane w krokach 3) oraz 4) powtarza się aż do wyczerpania wszystkich potencjalnych zmiennych objaśniających.

Przykład 8

Wartości zmiennej objaśnianej Y oraz potencjalnych zmiennych objaśniających X_1, X_2, X_3, X_4 podano w tabeli 10.

Tabela 10.Wartości zmiennej objaśnianej i potencjalnych zmiennych objaśniających

n	Y	X1	X2	X3	X4
1	11	2,1	11	30	24
2	13	2,2	11	30	26
3	16	2,3	12	31	25
4	24	2,4	13	28	30
5	24	2,5	15	26	25
6	26	2,5	16	24	27
7	27	2,6	16	22	26
8	28	2,7	16	23	29
9	29	2,6	16	20	25
10	33	2,7	16	16	27
11	33	2,7	16	14	26
12	36	2,7	16	12	28

Wykorzystując metodę Nowaka wskazać do modelu zmienne objaśniające.

Rozwiązanie

Wyznaczone średnie poszczególnych zmiennych wyznaczone zgodnie ze wzorem (18) mają wartość:

$\bar{y} = 25, \quad \bar{x}_1 = 2{,}5, \quad \bar{x}_2 = 14{,}5, \quad \bar{x}_3 = 23, \quad \bar{x}_4 = 26{,}5$

natomiast współczynniki korelacji,wyznaczone według (19)

$r_1 = \frac{17,7}{\sqrt{702 \cdot 0,48}} = 0,964,$ $r_2 = \frac{170}{\sqrt{702 \cdot 49}} = 0,917,$ $r_3 = \frac{-524}{\sqrt{702 \cdot 458}} = -0,964,$

$r_4 = \frac{70}{\sqrt{702 \cdot 35}} = 0,447.$

Współczynniki korelacji pomiędzy potencjalnymi zmiennymi objaśniającymi uzyskano ze wzoru (21):

$r_{12} = \frac{4,6}{\sqrt{0,48 \cdot 49}} = 0,949,$ $r_{13} = \frac{-12,8}{\sqrt{0,48 \cdot 458}} = -0,863,$ $r_{14} = \frac{-5241,7}{\sqrt{0,48 \cdot 35}} = 0,415,$

$r_{23} = \frac{-120}{\sqrt{49 \cdot 458}} = -0,801,$ $r_{24} = \frac{12}{\sqrt{49 \cdot 35}} = 0,290,$ $r_{43} = \frac{-30}{\sqrt{458 \cdot}} = -0,237.$

Aby wyznaczyć macierz ***R***, wystarczy znajomość obliczonych powyżej współczynników korelacji, gdyż jest ona symetryczna, a zatem:

$$r = \begin{bmatrix} 0,964 \\ 0,917 \\ -0,924 \\ 0,447 \end{bmatrix}, R = \begin{bmatrix} 1 & 0,949 & -0,863 & 0,415 \\ 0,949 & 1 & -0,801 & 0,290 \\ -0,863 & -0,801 & 1 & -0,237 \\ 0,415 & 0,290 & -0,237 & 1 \end{bmatrix}$$

1. Ustalając wartość krytyczną współczynnika korelacji założono poziom istotności α = 0,05. Dla N=12 $t_\alpha$odczytane z tablic t-Studenta dla 12 – 2=10 stopni swobody wynosi t_α = 2,228.

Zatem

$$r^* = \left[\frac{(2,228)^2}{(2,228)^2 + 12 - 2}\right] = 0,576$$

2. Wyeliminowanie ze zbiory potencjalnych zmiennych objaśniających tych zmiennych, które są słabo skorelowane ze zmienną objaśnianą.

Analizując macierz **R** stwierdzić należy, że jedynie zmienna X_4 jest słabo skorelowana ze zmienną Y, gdyż $|0,447| < 0,576$. Zmienna X_4 nie znajdzie się zatem w zbiorze zmiennych objaśniających.

3. Spośród zmiennych, które pozostały wybiera się tę, która jest najsilniej skorelowana ze zmienną objaśnianą.

W poprzednim kroku wyeliminowana została zmienna *X4*, zatem pozostały zmienne X_1, X_2, X_3. Spośród tych zmiennych wybierana jest zmienna najsilniej skorelowana ze zmienną Y – zatem wybrana została zmienna X_1.

4. Ze zbioru potencjalnych zmiennych objaśniających eliminujemy te, które są silnie skorelowane z wybraną w 3) zmienną X_1. Zatem spośród pozostałych zmiennych (X_2, X_3) eliminowane są te, których współczynnik korelacji ze zmienną $X1$ jest większy co do wartości bezwzględnej od 0,576 (wartość krytyczna). Analizując macierz ***R***

stwierdzić należy, że obie zmienne są silnie skorelowane z X_1, zatem obie zostały wyeliminowane.

5.Postępowanie opisane w krokach 3) i 4) powtarza się aż do wyczerpania wszystkich potencjalnych zmiennych objaśniających.
W tym momencie procedura jest już zakończona, gdyż uwzględnione zostały wszystkie zmienne: zmienna X_4została odrzucona gdyż jest słabo skorelowana ze zmienną Y, natomiast zmienne X_2, X_3 zostały odrzucone jako zbyt silnie skorelowana z wcześniej wybraną do modelu zmienną X_1. Ostatecznie jedynie ta zmienna została wskazana do modelu i tylko ona jest zmienną objaśniającą.

Metoda wskaźników pojemności informacyjnej - metoda Hellwiga
Podobnie jak w poprzedniej metodzie, zakłada się, że dysponuje się pewnym zbiorem potencjalnych zmiennych objaśniającychX$_1$, X_2, …, X_K, które „kandydują” do roli objaśniających zmienną Y. Zakłada się również, że znane są macierze współczynników korelacji ***r*** oraz ***R***.
Każdą z „kandydatek” można uważać za źródło wiedzy o zmiennej Y, zasadne jest więc traktowanie jej jako **nośnika informacji** o zmiennej objaśniającej.
Rozpatruje się wszystkie niepuste kombinacje (oznaczone K_l) zmiennychX$_1$, X_2, …, X_K. Liczba kombinacji wynosi $2^K - 1$. Zbiór numerów zmiennych tworzących l-tą kombinację, $l = 1,2, \ldots, 2^K - 1$, oznacza się przez Z_l. Określa się indywidualną pojemność informacyjną nośnika X_K wchodzącego w skład l-tej kombinacji następująco:

$$h_{lk} = \frac{r_k^2}{\sum_{i \epsilon Z_i} |r_{ik}|} \tag{28}$$

Następnie oblicza się *integralną pojemność informacyjną* l-tej kombinacji:

$$H_l = \sum_{j \in Z_l} h_{lj} \tag{29}$$

Zarówno indywidualna jak integralna pojemność informacyjna przyjmuje wartości z przedziału [0, 1].Za najlepszą kombinację nośników informacji uznaje się ten podzbiór „kandydatek” na zmienne objaśniające, dla których pojemność integralna jest największa, czyli:

$$H_{opt} = max\{H_l : l = 1,2, \ldots, 2^{K-1}\} \tag{30}$$

Przykład 9

W analizie popytu na żywiec wołowy wzięto wstępnie pod uwagę dwie zmienne kandydujące do roli objaśniającej:

X_1 – przeciętne dochody na głowę mieszkańca w poszczególnych województwach,

X_2 - przeciętna cena żywca wołowego w poszczególnych województwach.

Na podstawie zebranych danych oszacowano macierze współczynników korelacji ***r*** oraz ***R***:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0{,}6 \\ 0{,}6 & 1 \end{bmatrix} \boldsymbol{r} = \begin{bmatrix} 0{,}8 \\ -0{,}2 \end{bmatrix}$$

Spośród przedstawionych zmiennych należy dokonać wyboru tych, które niosą najwięcej informacji o modelowanym zjawisku tj. o popycie na żywiec wołowy w Polsce. W rozwiązaniu wykorzystać metodę Helliwga.

Rozwiązanie

Spośród rozpatrywanych wstępnie zmiennych można utworzyć 2^2 -1=3 kombinacje (K=2), zatem l = 1,2,3. Wszystkie kombinacje K_l oraz zbiory numerów Z_l zmiennych tworzących daną kombinację są następujące:

$$K_1 = \{X_1\} \quad Z_1 = \{1\},$$
$$K_2 = \{X_2\} \quad Z_2 = \{2\},$$
$$K_3 = \{X_1, X_2\} Z_3 = \{1,2\},$$

Dla każdej zmiennej w każdej kombinacji oblicza się indywidualną pojemność informacyjną i dla każdej kombinacji oblicza się integralną pojemność informacyjną:

Kombinacja 1

$$h_{11} = \frac{0{,}8^2}{1} = 0{,}64, \quad H_1 = h_{11} = 0{,}64$$

Kombinacja 2

$$h_{22} = \frac{(-0{,}2)^2}{1} = 0{,}04, \quad H_2 = h_{22} = 0{,}04$$

Kombinacja 3

$$h_{31} = \frac{0{,}8^2}{1+0{,}6} = 0{,}4, \quad h_{32} = \frac{(-0{,}2)^2}{1+0{,}6} = 0{,}025, \quad H_3 = h_{31} + h_{32} = 0{,}425$$

Największa wartość pojemności integralnej występuje w przypadku pierwszej kombinacjiK_1={X_1}, zatem za jedyną zmienną objaśniającą należy uznać X_1.

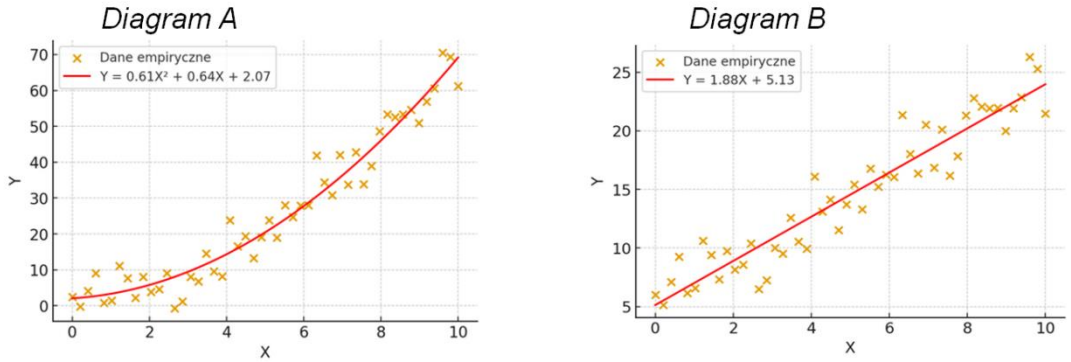
4.2.2. Wybór klasy modelu

Wybór klasy modelu ekonometrycznego wymaga:

- ✓ zdefiniowania postaci analitycznej modelu (liniowa, nieliniowa),
- ✓ określenia liczby funkcji w modelu (modele jedno lub wielorównaniowe),
- ✓ wyznaczenie roli czynnika czasu, modelowaniu (modele statyczne, dynamiczne).

Wyboru postaci analitycznej modelu można dokonać na podstawie m.in. wykresu badania przyrostów zmiennej objaśnianej, metody prób i błędów opartej na analizie miar i testów statystycznych dotyczących różnych postaci modeli.

Dokonując wyboru postaci modelu wykorzystuje się diagramy korelacyjne. I tak na przykład na podstawie poniższych diagramów można przypuszczać, że pierwsza zależność ma charakter liniowy (rysunek 1, diagram A, a druga nieliniowy (rysunek 1, diagram B). Decydując się na badanie przyrostów zmiennej objaśniającej, należy stosować się do ogólnych zasad podanych w Tabeli 11.
[Tekst alternatywny. Wykres dwuwymiarowy przedstawiający dwa diagramy korelacyjne oznaczone jako Diagram A i Diagram B. Opis osi X: wartości zmiennej niezależnej *X* w zakresie od 0 do 10. Opis osi Y: wartości zmiennej zależnej *Y*. Na obu wykresach znajdują się pomarańczowe punkty reprezentujące dane empiryczne oraz czerwona linia trendu dopasowana do obserwacji. *Diagram A* ukazuje nieliniową, rosnącą zależność pomiędzy zmiennymi (funkcja kwadratowa), natomiast *Diagram B* przedstawia liniową zależność dodatnią o równaniu Y = 1,88X + 5,13]



Rys. 1. Przykładowe diagramy korelacyjne.

Tabela 11. Zasady badania przyrostów zmiennej objaśnianej

Funkcja	Wzór	Przyrosty
Liniowa	$Y = a \cdot X + \varepsilon$	$Y_n - Y_{n-1} = const$
Wykładnicza	$Y = a^X + \varepsilon$	$log(Y_n) - log(Y_{n-1}) = const$
Hiperboliczna	$Y = a \cdot X^{-1} + \varepsilon$	$\frac{1}{\log(Y_n)} - \frac{1}{\log(Y_{n-1})} = const$

4.2.3. Estymacja parametrów strukturalnych modelu

Wybór modelu oznacza konieczność skonstruowania pewnej (gdy jest ich więcej – pewnych) funkcji, określającej zależność pomiędzy zmienną objaśnianą (zmiennymi objaśnianymi) a zmiennymi objaśniającymi. Nazywa się to estymacją modelu i polega na wyznaczeniu postaci stałych, które w tym modelu występują.

Estymację parametrów modelu (szacowanie ich wartości na podstawie danych statystycznych) przeprowadzić można właściwymi w danych warunkach metodami, np. metodą momentów, metodą największej wiarygodności, metodą najmniejszych kwadratów, algorytmy optymalizacji nieliniowej.

4.2.3.1. Estymacja modelu liniowego – metoda najmniejszych kwadratów

Dalsze rozważania będą dotyczyły przede wszystkim liniowej zależności pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi.
Wtedy ekonometryczny model jednorównaniowy ma postać:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_K x_{iK} + \varepsilon_i \; i = 1, \ldots, n \;\; k = 1, \ldots, K \tag{31}$$

gdzie:
y_i – *i*-ta obserwacja zmiennej objaśnianej,
x_{ik} – *i*-ta obserwacja *k*-tej zmiennej objaśnianej,
α_k – parametr strukturalny związany z *k*-tą zmienną objaśniającą,
ε_i – *i*-te zakłócenie.
Jego parametry szacuje się klasyczną metodą najmniejszych kwadratów, otrzymując równanie liniowe:

$$\hat{y}_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \ldots + a_K x_{iK} \tag{32}$$

gdzie:
a_k- estymatory nieznanych parametrów α_k podanej funkcji.
W metodzie najmniejszych kwadratów współczynniki a_k dobiera się tak, aby suma kwadratów odchyleń estymowanych wartości zmiennej objaśnianej $\hat{y}$od jej rzeczywistychwartości *Y* była minimalna

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{33}$$

Na to, aby powyższa funkcja, zwana kryterialną (33) osiągnęła minimum musi być spełniony warunek, zwany układem równań normalnych

$$X^T (Y - Xa) = 0 \tag{34}$$

Po przekształceniach otrzymuje się

$$X^T X a = X a Y \tag{35}$$

Jeżeli macierz $X^T X$ jest nieosobliwa (czyli jeśli istnieje macierz do niej odwrotna), co zachodzi wtedy, gdy jej wyznacznik jest różny od zera, otrzymuje się:

$$a = (X^T X)^{-1} X^T Y \tag{36}$$

gdzie:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix}$$ - macierz obserwacji zmiennych objaśniających, (37)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix}$$ – wektor obserwacji zmiennej objaśnianej, (38)

$$a = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_K \end{bmatrix}$$ – wektor estymatorów współczynników równania regresji. (39)

Przykład 10

Tabela 12 przedstawia dane dotyczące wydatków ogółem x_i oraz wydatków na żywność y_i w kilku wybranych rodzinach.

Tabela 12.Wydatki ogółem i wydatki na żywność w wybranych rodzinach

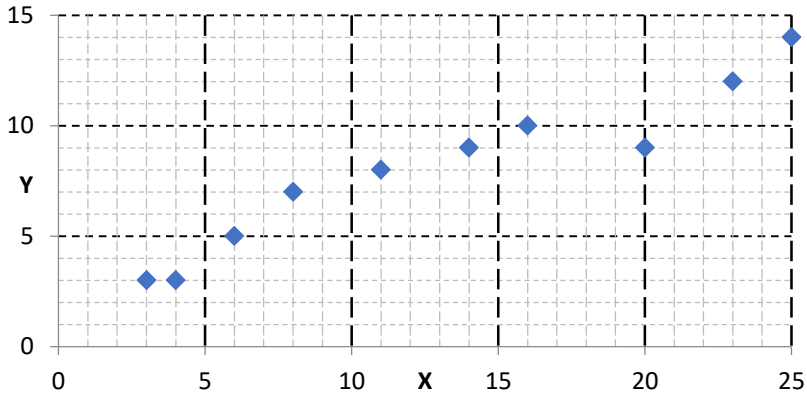
x_i	3	4	6	8	11	14	16	20	23	25
y_i	3	3	5	7	8	9	10	9	12	14

Wyznaczyć liniową postać modelu ekonometrycznego wydatków na żywność *Y* w zależności od wydatków ogółem *X*.

Rozwiązanie

Zaznacza się w układzie współrzędnych punkty odpowiadające danym empirycznym.

[Tekst alternatywny. Wykres dwuwymiarowy przedstawiający zależność wydatków na żywność (*Y*) od wydatków ogółem (*X*). Oś X opisuje całkowite wydatki gospodarstw domowych, natomiast oś Y przedstawia odpowiadające im wartości wydatków na żywność. Na wykresie znajdują się niebieskie punkty reprezentujące dane empiryczne.]



Rys. 2. Zależność wydatków na żywność (*Y*) od wydatków ogółem (*X*).

Ponieważ jest jedna zmienna objaśniająca, zatem szukana jest funkcja liniowa postaci:

$$\hat{y} = a_0 + a_1 X$$

W tym przykładzie *K*=1 (jedna zmienna objaśniająca) oraz *n*=10 (ilość obserwacji). Odpowiednie macierze są następujące:

$$Y = \begin{bmatrix} 3 \\ 3 \\ 5 \\ 7 \\ 8 \\ 9 \\ 10 \\ 9 \\ 12 \\ 14 \end{bmatrix} X = \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 11 \\ 1 & 14 \\ 1 & 16 \\ 1 & 20 \\ 1 & 23 \\ 1 & 25 \end{bmatrix} a = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$$

Wykorzystując wzór (36), obliczenia wykonano etapami

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 4 & 6 & 8 & 11 & 14 & 16 & 20 & 23 & 25 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 11 \\ 1 & 14 \\ 1 & 16 \\ 1 & 20 \\ 1 & 23 \\ 1 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 130 \\ 130 & 2252 \end{bmatrix}$$

Zauważyć należy, że otrzymana macierz jest symetryczna.

$$\det(X^T X) = \begin{vmatrix} 10 & 130 \\ 130 & 2252 \end{vmatrix} = 5620 \neq 0$$

Wyznacznik jest różny od zera, zatem istnieje macierz odwrotna, którą liczy się ze wzoru ogólnego:

$$A^{-1} = \frac{1}{\det A} A^D \tag{40}$$

gdzie:
A^D– macierz transponowana dopełnień algebraicznych.
Zatem

$$(X^T X)^{-1} = \begin{bmatrix} \frac{2250}{5620} & -\frac{130}{5620} \\ -\frac{130}{5620} & \frac{10}{5620} \end{bmatrix} = \begin{bmatrix} 0{,}4007117 & -0{,}0231316 \\ -0{,}0231316 & 0{,}0017793 \end{bmatrix}$$

następnie liczy się:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 4 & 6 & 8 & 11 & 14 & 16 & 20 & 23 & 25 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 5 \\ 7 \\ 8 \\ 9 \\ 10 \\ 9 \\ 12 \\ 14 \end{bmatrix} = \begin{bmatrix} 10 & 130 \\ 130 & 2252 \end{bmatrix}.$$

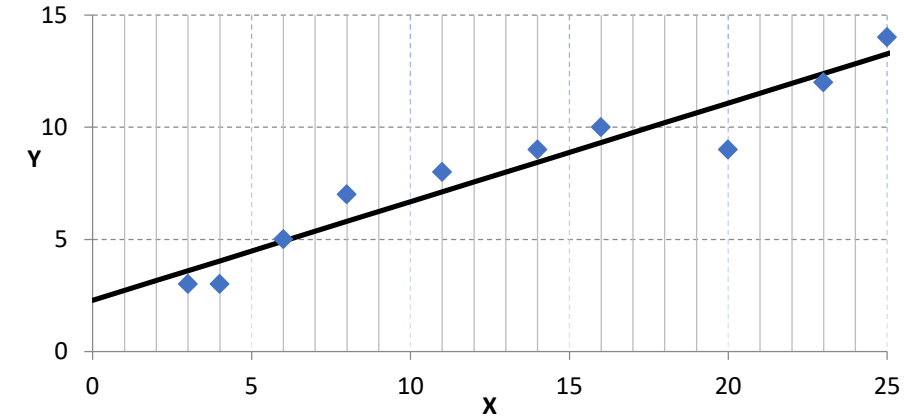
Ostatecznie

$$a = (X^T X)^{-1} \cdot X^T Y = \begin{bmatrix} 0{,}4007117 & -0{,}0231316 \\ -0{,}0231316 & 0{,}0017793 \end{bmatrix} \cdot \begin{bmatrix} 80 \\ 1287 \end{bmatrix} = \begin{bmatrix} 2{,}2865 \\ 0{,}4394 \end{bmatrix}$$

czyli a_0= 2,2865, a_1= 0,4394
Oszacowany metodą najmniejszych kwadratów model liniowy zależności wydatków na żywność *Y* od wydatków ogółem *X* ma postać

$$\hat{y} = 2{,}2865 + 0{,}4394X$$

[Tekst alternatywny. Wykres dwuwymiarowy przedstawiający liniowy model zależności wydatków na żywność (*Y*) od wydatków ogółem (*X*). Opis osi X: wartości zmiennej niezależnej – wydatki ogółem gospodarstw domowych. Opis osi Y: wartości zmiennej zależnej – wydatki na żywność. Na wykresie znajdują się niebieskie punkty reprezentujące dane empiryczne oraz czarna linia trendu dopasowana metodą regresji liniowej. Linia trendu wskazuje dodatnią, liniową zależność pomiędzy analizowanymi zmiennymi – wzrost wydatków ogółem powoduje proporcjonalny wzrost wydatków na żywność.]



Rys. 3. Liniowy model zależności wydatków na żywność (*Y*) od wydatków ogółem (*X*).

W rozważanym przypadku współczynnik kierunkowy prostej jest dodatni czyli uzyskana funkcja jest rosnąca. Oznacza to, że wraz ze wzrostem wydatków ogółem, rosną średnio wydatki na żywność.

Interpolacja

Wśród danych empirycznych nie występowały dane dotyczące wydatków na żywność rodzin, których wydatki ogółem wynoszą $x = 10$. Można je jednak oszacować, korzystając z modelu:

$$\hat{y} = 2{,}2865 + 0{,}4394 \cdot 10 = 6{,}6805$$

Oznacza to, że rodziny te wydają na żywność średnio 6,6805.

Interpolacja polega na szacowaniu wartości zmiennej objaśnianej dla takiego poziomu zmiennej objaśniającej, który mieści się w zakresie obserwowanych danych empirycznych. Oznacza to, że dokonuje się przewidywań w przedziale, w którym znane są dane źródłowe.

Aproksymacja

Rodziny, których wydatki ogółem wynoszą $x = 20$, wydają na żywność średnio y= 11,2 co wynika z danych empirycznych. Tymczasem model przewiduje:

$$\hat{y} = 2{,}2865 + 0{,}4394 \cdot 20 = 11{,}0745.$$

Model nie odtwarza zatem dokładnie wartości empirycznej, lecz ją aproksymuje (przybliża).

Różnica między wartością rzeczywistą a przewidywaną wynosi:

$$e = y - \hat{y} = 11{,}2 - 11{,}745 = 0{,}1255$$

co oznacza, że błąd przybliżenia (reszta) jest stosunkowo niewielki.

Aproksymacja oznacza przybliżenie rzeczywistego przebiegu zależności między zmiennymi za pomocą modelu matematycznego (np. funkcji liniowej). Model nie odwzorowuje dokładnie wszystkich wartości empirycznych, lecz opisuje je w sposób uśredniony, minimalizując odchylenia (np. metodą najmniejszych kwadratów).

Ekstrapolacja

Można postawić pytanie o wydatki na żywność rodzin, których wydatki ogółem wynoszą x =30. Należy zwrócić uwagę, że ich x = 30 wykracza poza zakres danych empirycznych:

$\hat{y} = 2{,}2865 + 0{,}4394 \cdot 30 = 15{,}4685.$

Otrzymujemy zatem, że rodziny te wydają na żywność średnią 15,47.

Ekstrapolacja - polega na przewidywaniu wartości zmiennej objaśnianej dla takiego poziomu zmiennej objaśniającej, który znajduje się poza zakresem obserwowanych danych. W przeciwieństwie do interpolacji, ekstrapolacja obarczona jest większym ryzykiem błędu, gdyż model może nie odzwierciedlać poprawnie zależności w obszarze nieobjętym obserwacjami empirycznymi.

4.2.4. Weryfikacja modelu

Weryfikacja modelu obejmuje zarówno weryfikację merytoryczną, jak i statystyczną.
Weryfikacja merytoryczna służy odpowiedzi na pytanie, czy zastosowanie modelu daje sensowne, z ekonomicznego punktu widzenia, wyniki.
Weryfikacja statystyczna obejmuje wyznaczenie wartości odpowiednich statystyk, które pozwalają rozstrzygnąć prawdziwość hipotez dotyczących modelu i jego elementów.
Jeżeli elementy lub cały model nie zostaną pozytywnie zweryfikowane, to konieczna jest jego przebudowa.
W najprostszym przypadku można zmienić jego postać analityczną.
Czasem konieczne jest uzupełnienie zestawu zmiennych, co wymaga zebrania dodatkowych danych. W obydwu wypadkach należy ponownie wykonać wcześniejsze etapy.

4.2.4.1. Ocena merytoryczna

Ocena merytorycznej poprawności modelu polega na sprawdzeniu zgodności znaków parametrów modelu z teorią ekonomii lub oczekiwaniami badacza.
Negatywna ocena polegająca na niezgodności znaków prowadzi do reastymacji parametrów funkcji, a pozytywna – do przejścia do oceny statystycznej.

4.2.4.2. Weryfikacja statystyczna

Aby otrzymane metodą najmniejszych kwadratów estymatory $a_j$współczynnikówα_j (j = 0, 1, 2,..., k) były efektywne, muszą być spełnione założenia Gaussa–Markowa, a mianowicie:

• Związek między zmienną objaśnianą Y a zmiennymi objaśniającymi X_1, X_2, ..., X_K ma charakter liniowy.

• Wartości zmiennych objaśniających są ustalone (nie są losowe) – losowość wartości zmiennej objaśnianej Y wynika z losowości składnika ε.

• Składniki losowe ε dla poszczególnych wartości zmiennych objaśniających mają rozkład normalny (lub bardzo silnie zbliżony do normalnego) o wartości oczekiwanej zero i stałej wariancji: N(0, δ_ε).

• Składniki losowe nie są ze sobą skorelowane.

Spełnienie założeń Gaussa–Markowa weryfikuje się za pomocą odpowiednich testów statystycznych. Liniowy charakter zależności między zmienną objaśnianą Y a zmiennymi objaśniającymi X_1, X_2, ..., X_K weryfikuje się na podstawie wartości takich statystyk, jak współczynnik determinacji lub współczynnik zbieżności modelu.

Do weryfikacji losowości rozkładu reszt modelu względem równania regresji $\hat{y}$ można zastosować między innymi testy serii (test liczby serii, test maksymalnej długości serii), natomiast do weryfikacji normalności rozkładu składnika losowego: testy zgodności χ^2, λKołmogorowa, Shapiro–Wilka, Dawida–Hellwiga.

Równość wariancji składnika losowego można weryfikować między innymi za pomocą testów: Goldfelda–Quandta, korelacji rangowej Spearmana oraz korelacji modułów składników losowych i czasu.

Zjawisko autokorelacji pierwszego rzędu składników losowych można weryfikować między innymi za pomocą testów Durbina–Watsona, von Neumanna, Durbina, a występowanie autokorelacji dowolnego rzędu testem istotności współczynników autokorelacji.

Dopasowanie modelu do danych empirycznych

Podstawowe miary dopasowania modelu do danych rzeczywistych to:

• odchylenie standardowe składnika resztowego (standardowy błąd szacunku)$S\varepsilon$

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-K-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-K-1}} \tag{41}$$

gdzie:

y_i - rzeczywista wartość zmiennej objaśnianej

$\hat{y}_i$ - wartość zmiennej objaśnianej wyznaczona na podstawie modelu,

$e_i = y_i - \hat{y}_i$ - reszty modelu.

Odchylenie standardowe informuje, o ile średnio wartości rzeczywiste zmiennej objaśnianej różnią się od wartości przewidywanych przez model.

Oczywiście znając jedynie wartość odchylenia standardowego trudno stwierdzić, czy jest ono duże czy też małe. Dlatego liczy się współczynnik zmienności losowej *v*.

$$v = \frac{S_\varepsilon}{\bar{y}} \tag{42}$$

Im bliższe zeru wartości współczynnika zmiennej losowej, tym lepsze dopasowanie modelu do danych empirycznych, tym lepiej opisuje on rzeczywistość. Często przyjmuje się, że powinna być spełniona nierówność $|v| < 0{,}1$.

• współczynnik zbieżności φ^2:

$$\varphi^2 = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{43}$$

gdzie:
$\bar{y}$- wartość średnia zmiennej objaśnianej *Y*.
Współczynnik zbieżności przyjmuje wartości z przedziału [0, 1] i informuje, jaka część całkowitej zmienności zmiennej objaśnianej nie jest wyjaśniana przez model. Dlatego bliższe zeru wartości tego współczynnika informują o lepszym dopasowaniu

• współczynnik determinacji R^2:

$$R^2 = 1 - \varphi^2 \tag{44}$$

Współczynnik determinacji przyjmuje wartości z przedziału [0, 1] i informuje, jaka część całkowitej zmienności zmiennej objaśnianej stanowi zmienność wartości teoretycznych tej zmiennej, tj. część zdeterminowaną przez zmienne objaśniające. Dlatego bliższe jedynki wartości tego współczynnika informują o lepszym dopasowaniu modelu. Arbitralnie ustala się dopuszczalną wartość graniczną R^2 (jest to zazwyczaj wartość około 0,6).

Stosuje się także skorygowany współczynnik determinacji

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-K} \tag{45}$$

Współczynnik ten może przyjmować wartości z przedziału (-∞, 1). Stosowany jest do porównania dopasowania modeli ekonometrycznych z różną liczbą zmiennych objaśniających.

W przypadku modeli nieliniowych, w których zmienna objaśniana *Y* jest transformowana stosuje się także współczynnik

$$"quasi\ \tilde{R}^2" = 1 - \frac{\sum_{i=1}^{n} E_1^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{46}$$

Współczynnik ten ma zastosowanie do porównania dopasowania modeli ekonometrycznych z różnymi kształtami funkcji.

W przypadku modeli ekonometrycznych z wieloma zmiennymi objaśniającymi należy ponadto sprawdzić, czy spełnione są warunki koincydencji:

$$sign\left(r(X_j, Y)\right) = sign(a_j) \tag{47}$$

gdzie:
sign(r(X$_j$, Y))– znak współczynnika korelacji pomiędzy zmienną objaśniającą X_j a zmienną objaśnianą *Y*,
sign(a$_j$) – znak współczynnika a_j w modelu ekonometrycznym przy zmiennej X_j.
Zgodność znaków współczynnika korelacji i współczynnika modelu ekonometrycznego musi zachodzić dla wszystkich zmiennych objaśniających. Jeżeli zmienne objaśniające są liniowo niezależne, to warunek ten jest spełniony.

Istotność układu współczynników regresji

W procesie weryfikacji modelu ekonometrycznego w pierwszej kolejności należy sprawdzić, czy zachodzi zależność liniowa między zmienną objaśnianą *Y*a którąkolwiek ze zmiennych objaśniających X_k modelu.

Test 1 – **istotności układu współczynników regresji**. Hipotezy mają postać:

$$H_0: \sum_{j=1}^{n} \alpha_j^2 = 0$$
$$H_1: \sum_{j=1}^{n} \alpha_j^2 \neq 0 \tag{48}$$

Statystyką testową dla H_0i H_1 jest

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - K - 1}{K} \tag{49}$$

Statystyka ta, przy założeniu prawdziwości hipotezy zerowej, ma rozkład *F* Snedecora o *K* stopniach swobody licznika oraz o (*n* – *K* – 1) stopniach swobody mianownika. Obszar krytyczny tego testu jest prawostronny

$$\theta = \{F: P(F \geq F_\alpha) = \alpha\} \tag{50}$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki *F* jest mniejsza od wartości krytycznej F_α($F<F_\alpha$), to nie ma podstaw do odrzucenia hipotezy H_0 na korzyść hipotezy alternatywnej H_1. Nie zachodzi związek liniowy między zmienną objaśnianą *Y,* a żadną ze zmiennych objaśniających X_j. Oznacza to, iż badany model ekonometryczny jest niepoprawny. W przeciwnym razie, gdy $F \geq F_\alpha$, przyjmuje się hipotezę H_1, a więc uznaje się, że między zmienną *Y* a przynajmniej jedną ze zmiennych uwzględnionych w modelu zachodzi zależność liniowa.

Istotność poszczególnych współczynników regresji

W poprawnym modelu ekonometrycznym zmienna objaśniana *Y* musi istotnie zależeć od każdej ze zmiennych objaśniających X_j modelu. Test weryfikujący ten fakt jest następujący.

Test 1 – istotności poszczególnych współczynników regresji. Dla każdego współczynnika równania regresji (*j*=0, 1, …, *K*) stawiana jest hipoteza

$$H_0: \alpha_j = 0$$
$$H_1: \alpha_j \neq 0 \tag{51}$$

Sprawdzianem zespołu hipotez jest statystyka

$$t = \frac{a_j}{S(\alpha_j)} \tag{52}$$

gdzie:

a_j– estymator współczynnika α_j,

$S(\alpha_j) = \sqrt{d_j}$– estymator dyspersji współczynnika α_j.

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład *t* Studenta o (*n* – *K* – 1) stopniach swobody.

Obszar krytyczny testu jest dwustronny

$$\theta = \{t: P(|t| \geq t_\alpha) = \alpha\} \tag{53}$$

Jeżeli zatem wartość statystyki testowej dla parametru $\alpha_j$nie należy do obszaru krytycznego to nie ma podstaw do odrzucenia hipotezy zerowej H_0. Oznacza to, że

parametr ten różni się nieistotnie od zera, a to z kolei oznacza, że zmienna objaśniająca X_k nie wpływa w sposób istotny na zamienną objaśnianą Y.
Jeżeli wartość statystyki testowej dla parametru α_j należy do obszaru krytycznego to, należy odrzucić hipotezę zerową na korzyść hipotezy alternatywnej. Oznacza to, że parametr ten różni się istotnie od zera, a to z kolei oznacza, że zmienna objaśniająca X_k wpływa w sposób istotny na zmienną objaśnianą.

Przykład 11

Dane dotyczące sprzedaży energii elektrycznej i Y(w mln MWh)w pewnym zakładzie energetycznym w zależności od długości linii przesyłowych X_1 oraz liczby odbiorców energii X_2 (w 100 tys.) w ciągu kilku lat przedstawia tabela13.

Tabela 13.Sprzedaż energii elektrycznej (Y) oraz czynniki ją określające (X_1, X_2)

n	Y	X_1	X_2
1	3,2	1,2	3,6
2	3,3	1,3	3,7
3	3,4	1,3	3,8
4	3,5	1,4	3,8
5	3,6	1,4	3,9
6	3,6	1,5	3,9
7	3,7	1,5	4,0
8	3,8	1,6	4,0
9	3,9	1,6	4,1
10	4,0	1,7	4,2

Oszacować parametry liniowego modelu ekonometrycznego zależności sprzedaży energii od długości linii przesyłowych oraz liczby odbiorców. Wyznaczyć parametry dopasowania modelu do danych empirycznych oraz zweryfikować na poziomie istotności 0,05 hipotezy o istotności odpowiednich parametrów.

Rozwiązanie

Są dwie zmienne objaśniające, zatem model ma postać: $\hat{y} = a_0 + a_1X_1 + a_2X_2$.
W przykładzie K=2 oraz n=10.
Odpowiednie macierzy są następujące:

$$Y = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4,0 \end{bmatrix}, X = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4,0 \\ 1 & 1,6 & 4,0 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix}, a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

Parametry strukturalne wyznacza się analogicznie, jak w 4.2.2.1

$$X^T X = \begin{bmatrix} 10 & 14,5 & 39 \\ 14,5 & 21,25 & 56,8 \\ 39 & 56,8 & 152,4 \end{bmatrix}, (X^T X)^{-1} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix}, a = \begin{bmatrix} -0,78 \\ 0,6 \\ 0,9 \end{bmatrix}$$

Ostatecznie model ma postać: $\hat{y} = -0,78 + 0,6X_1 + 0,9X_2$

Odchylenie standardowe składnika resztowego

Wykorzystując wzór (41) otrzymuje się

$S_\varepsilon^2 = \frac{0,006}{10-2-1} = 0,0008571$, $S_\varepsilon = \sqrt{0,0008571} = 0,0293$

Zatem wartości empiryczne różnią się od przewidywanych przez model o 0,0293.

Współczynnik zmienności

Aby wyznaczyć współczynnik zmienności potrzebna jest znajomość średniej zmiennej *Y*, którą można policzyć ze wzoru (42)

$$v = \frac{0,0293}{3,6} = 0,00814$$

Wartość współczynnika zmienności jest bardzo bliska zeru, czyli dopasowanie modelu jest dobre.

Współczynnik zbieżności

Współczynnik zbieżności wyznaczono zgodnie ze wzorem (43)

$$\varphi^2 = \frac{0,006}{0,6} = 0,01$$

Wartość ta oznacza, że jedynie 0,01 całkowitej zmienności zmiennej objaśniającej nie jest wyjaśniana przez model. Oznacza to, że to dopasowanie modelu jest bardzo dobre.

Współczynnik determinacji

Zgodnie ze wzorem (44) R^2 = 1-0,01 =0,99.

Wartość ta oznacza, że 0,99 całkowitej zmienności zmiennej objaśniającej jest zdeterminowana przez zmienne objaśniające. Oznacza to, że dopasowanie modelu jest bardzo dobre.

Istotność układu współczynników regresji

Weryfikowana jest hipoteza czy zachodzi zależność liniowa między zmienną objaśnianą *Y* a którąkolwiek ze zmiennych objaśniających X_k modelu na poziomie istotności 0,05:

$$H_0: \sum_{j=1}^{n} \alpha_j^2 = 0$$
$$H_1: \sum_{j=1}^{n} \alpha_j^2 \neq 0 \qquad (54)$$

Przyjmując poziom istotności 0,05 odczytana z tablic *F* Snedecora wartość dla 2 stopni swobody licznika i 10-2-1=7 mianownika wynosi 4,74.

Wartość statystyki testowej wyznaczona ze wzoru (32) to $F = \frac{0,99}{1-0,99} \cdot \frac{10-2-1}{2} = 346,5$.

Ponieważ wartość statystyki testowej należy do obszaru krytycznego odrzucić należy hipotezę zerową, uznając za prawdziwą alternatywną. Oznacza to, że między zmienną *Y*, a przynajmniej jedną ze zmiennych uwzględnionych w modelu zachodzi zależność liniowa.

Istotność ocen parametrów strukturalnych modelu

Hipotezy dla parametrów mają zgodnie ze wzorem (34) następującą postać:

dla α_0	dla α_1	dla α_2
$H_0: \alpha_0 = 0$	$H_0: \alpha_1 = 0$	$H_0: \alpha_2 = 0$
$H_1: \alpha_0 \neq 0$	$H_1: \alpha_1 \neq 0$	$H_1: \alpha_2 \neq 0$

Przyjmując poziom istotności 0,05odczytuje się z tablic rozkładu *t* studenta liczbę dla 10-2- 1=7 stopni swobody - jest to wartość 2,365. Obszar krytyczny ma zatem we wszystkich przypadkach postać $\theta = (-\infty; 2,368) \cup (2,368; \infty)$.

Zgodnie ze wzorem (37) otrzymano dla poszczególnych parametrów:

dla α_0	dla α_1	dla α_2
$T = \frac{-0,78}{0,4584} = -1,7 \notin \theta$	$T = \frac{0,6}{0,2267} = 2,65 \in \theta$	$T = \frac{0,9}{0,1916} = 4,7 \in \theta$

Dla parametru α_0 wartość statystyki testowej nie należy do obszaru krytycznego, czyli nie ma podstaw, aby odrzucić hipotezę zerową. Oznacza to, że parametr ten różni

się nieistotnie od zera, a to z kolei oznacza, że wyraz wolny występujący w modelu nie ma istotnego znaczenia.
Dla parametrów α_1 oraz α_2 wartość statystyki testowej należy do obszaru krytycznego. Należy odrzucić hipotezę zerową na korzyść hipotezy alternatywnej. Oznacza to, że parametry te różnią się istotnie od zera, a to z kolei oznacza, że zmienne objaśniające X_1 jeden oraz X_2 wpływają w sposób istotny na zmienną objaśnianą Y.

Własności składników losowych

Trzeci i czwarty warunek Gaussa-Markowa formułują własności składnika losowego modelu ekonometrycznego, których spełnienie jest wymagane dla zapewnienia efektywności estymatorów współczynników, tj.:

- składniki losowe dla poszczególnych wartości zmiennych objaśniających mają rozkłady normalne o wartości oczekiwanej zero i stałej wariancji: $N(0, \delta_\varepsilon)$. Wybór testu służącego do oceny normalności rozkładu zależy od liczby obserwacji. W przypadku dużej próby hipotezę o normalności składników losowych weryfikuje się testem zgodności $\chi 2$ lub testem λ Kołmogorowa. Dla małych prób można zastosować test Shapiro–Wilka lub test Dawida–Hellwiga;
- składniki losowe nie są ze sobą skorelowane.
  Autokorelacja to współzależność składników losowych i w sposób oczywisty niejest pożądana. Podstawowe przyczyny występowania autokorelacji to:

➢ niewłaściwie dobrana postać modelu ekonometrycznego,
➢ nieuwzględnienie w modelu istotnej zmiennej (objaśnianej, objaśniającej),w szczególności opóźnionej w czasie,
➢ cykliczność analizowanego zjawiska.

Stopień autokorelacji τ można ustalić na podstawie analizy właściwości badanego zjawiska lub można przyjąć τ odpowiadające największej wartości współczynnika korelacji $\rho(\varepsilon_t - \varepsilon_{t-\tau})$:

$$\rho_\tau = \rho(\varepsilon_t, \varepsilon_{t-\tau}) = \frac{cov(\varepsilon_t, \varepsilon_{t-\tau})}{\sqrt{D^2(\varepsilon_t) D^2(\varepsilon_{t-\tau})}} \tag{55}$$

Współczynnik autokorelacji $\rho(\varepsilon_t - \varepsilon_{t-\tau})$ nosi nazwę współczynnika autokorelacji rzędu τ.
Opracowano wiele testów, które umożliwiają wykrycie autokorelacji składników losowych. W przypadku autokorelacji pierwszego rzędu ($\tau = 1$) hipotezę o braku autokorelacji składników losowych można weryfikować za pomocą testów Durbina–Watsona, von Neumanna, Durbina. Natomiast obecność autokorelacji dowolnego rzędu ($\tau > 1$) bada wykorzystując test istotności autokorelacji składników losowych rzędu τ składników lub test ogólnej istotności autokorelacji.

Przykład 12

Wyniki estymacji modelu liniowego *czas* = $\alpha_0 + \alpha_1 droga + \varepsilon$ przedstawiono na rysunku 4. Utworzony na ich podstawie model ekonometryczny $\widehat{czas} = 2{,}426929 + 0{,}008885 droga$ zweryfikować na poziomie istotności $\alpha = 0{,}05$.

[Tekst alternatywny. Zestawienie tabelaryczne przedstawiające wyniki estymacji liniowego modelu ekonometrycznego określającego zależność czasu przejazdu od długości drogi. Na rysunku znajdują się trzy tabele:
– pierwsza tabela prezentuje podstawowe statystyki regresji, w tym współczynnik determinacji (R^2), współczynnik korelacji, błąd standardowy oraz liczbę obserwacji;
– druga tabela zawiera analizę wariancji (ANOVA) z wartościami sum kwadratów, stopniami swobody, średnimi kwadratami, statystyką *F* i poziomem istotności *p*;
– trzecia tabela przedstawia oszacowane parametry modelu, czyli wyraz wolny i współczynnik kierunkowy wraz z błędami standardowymi, statystykami *t* i wartościami *p*.]

Statystyki regresji	
Wielokrotność *R*	0,986784
R kwadrat	0,973743
Dopasowany *R* kwadrat	0,972102
Błąd standardowy	1,319274
Obserwacje	18

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	1032,72	1032,72	593,3524	4,49E-14
Resztkowy	16	27,84773	1,740483		
Razem	17	1060,568			

	Współczyn-niki	*Błąd standardowy*	*Statystyka t Studenta*	*Wartość p*	*Dolne 95%*	*Górne 95,0%*
Przecięcie	2,426929	0,585748	4,143296	0,000764	1,185198	3,66866
Odległość	0,008885	0,000365	24,35883	4,49E-14	0,008111	0,009658

Rys. 4. Wyniki estymacji liniowego modelu ekonometrycznego zależności czasu przejazdu od długości drogi.

Dopasowanie modelu do danych empirycznych. Współczynnik determinacji modelu wynosi R^2 = 0,973743 (współczynnik zbieżności φ^2 =2,6%).
A zatem model wyjaśnia 97,4%zmienności badanej cechy. Świadczy to o dobrym dopasowaniu modelu do danych empirycznych.

Istotność układu współczynników regresji. Stawiamy hipotezy:

$$H_0: \sum_{j=1}^{n} \alpha_j^2 = 0$$

$$H_1: \sum_{j=1}^{n} \alpha_j^2 \neq 0$$

Sprawdzianem zespołu hipotez jest statystyka$F = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$

Statystyka ta, przy prawdziwości hipotezy zerowej, ma rozkład F Snedecora o 1 stopniu swobody licznika i 16 stopniach swobody mianownika.
Wyznaczona wartość empiryczna statystyki wynosi *F* = 593,3524, a odpowiadający jej krytyczny poziom istotności (istotność *F*) wynosi 4,49E-14 i jest mniejszy od przyjętego poziomu istotności *α* = 0,05. Należy zatem odrzucić hipotezę H_0 na korzyść H_1.
Wniosek. Nie ma podstaw do odrzucenia hipotezy o zależności czasu podróży od odległości.

Istotność poszczególnych współczynników regresji. Dla każdego współczynnika modelu regresji (*j* = 0,1) stawiamy hipotezy:

$H_0: \alpha_j = 0$

$H_1: \alpha_j \neq 0$

Sprawdzianem zespołu hipotez jest statystyka$t(a_j) = \frac{a_j}{S(a_j)}$

Statystyka ta, przy prawdziwości hipotezy zerowych, ,a rozkład *t* Studenta o 16 stopniach swobody.
Wyznaczone empiryczne wartości statystyk *t* Studenta wynoszą odpowiednio:
t(α_0) = 4,14,
t(α_1) = 24,36.
Odpowiadające im wartości krytycznego poziomu istotności (*p*-wartość) 0,000764oraz 4,491E-14 są mniejsze od przyjętego poziomu istotności *α* = 0,05.
Wniosek. Nie ma podstaw do odrzucenia hipotezy o istotności obu współczynników modelu.

Analiza składników losowych modelu

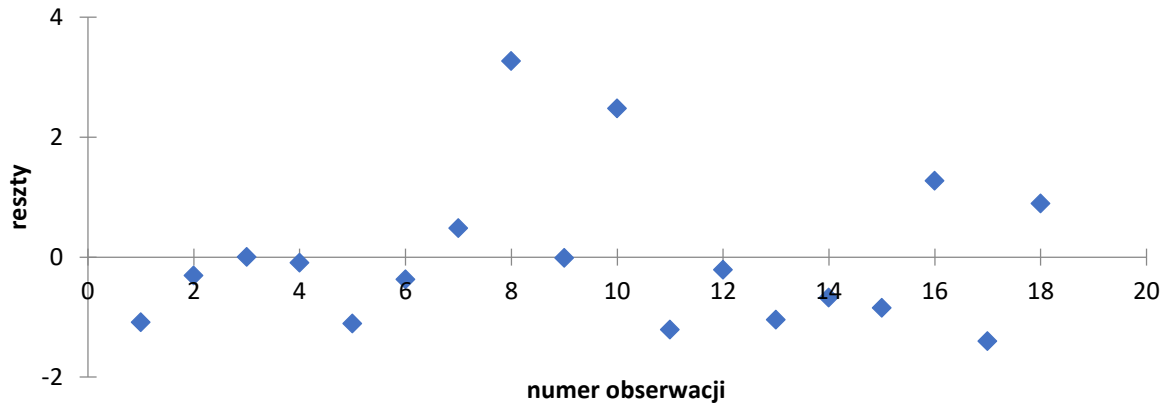
Tabela 14.Składniki losowe (reszty) i przewidywane wartości modelu ekonometrycznego

Obserwacja	Przewidywany czas	Składniki resztowe	Std. składniki resztowe
1	23,01343	1,269907	0,992205
2	7,631525	0,001808	0,001413
3	8,45957	-1,10957	-0,86693
4	8,570627	0,479373	0,374544
5	16,6254	-1,20874	-0,94441
6	11,03433	3,265675	2,55154
7	32,62478	0,891884	0,696848
8	16,7951	-0,21176	-0,16546
9	5,74266	-0,30933	-0,24168
10	28,42148	-1,40481	-1,09761
11	13,50602	2,477313	1,935577
12	20,12949	-0,84615	-0,66112
13	16,87861	-1,04528	-0,8167
14	8,02689	-0,09356	-0,0731
15	18,31259	-0,67925	-0,53071
16	8,488	-0,37133	-0,29013
17	5,488561	-1,08856	-0,85052
18	11,58428	-0,01762	-0,01376

Tabela 15.Składniki resztowe uporządkowane według przewidywanych wartości zmiennej czas

Obserwacja	Przewidywany czas	Składniki resztowe	Std. składniki resztowe
17	5,488561	-1,08856	-0,85052
9	5,74266	-0,30933	-0,24168
2	7,631525	0,001808	0,001413
14	8,02689	-0,09356	-0,0731
3	8,45957	-1,10957	-0,86693
16	8,488	-0,37133	-0,29013
4	8,570627	0,479373	0,374544
6	11,03433	3,265675	2,55154
18	11,58428	-0,01762	-0,01376
11	13,50602	2,477313	1,935577
5	16,6254	-1,20874	-0,94441
8	16,7951	-0,21176	-0,16546
13	16,87861	-1,04528	-0,8167
15	18,31259	-0,67925	-0,53071
12	20,12949	-0,84615	-0,66112
1	23,01343	1,269907	0,992205
10	28,42148	-1,40481	-1,09761
7	32,62478	0,891884	0,696848

[Tekst alternatywny. Wykres dwuwymiarowy przedstawiający rozkład reszt modelu liniowego opisującego zależność czasu podróży od odległości. Opis osi X: wartości przewidywane czasu podróży uzyskane z modelu liniowego. Opis osi Y: wartości reszt, czyli różnice pomiędzy wartościami empirycznymi a teoretycznymi. Na wykresie znajdują się niebieskie punkty reprezentujące poszczególne obserwacje. Punkty rozmieszczone są losowo wokół osi poziomej, co wskazuje na brak systematycznych odchyleń i potwierdza spełnienie założenia o losowości składnika resztowego w modelu regresji liniowej.]



Rys. 5. Rozkład reszt modelu liniowego czasu podróży od odległości.

Normalność

Stawia się hipotezę

H_0: składniki losowe mają rozkład $N(0, 1{,}319274)$,

która zostanie zweryfikowana za pomocą test Dawida-Hellwiga.

Test Dawida-Hellwiga wykorzystuje to,że każda dystrybuanta rozkładu ciągłego ma rozkład jednostajny na odcinku [0, 1]. Procedura testowania jest następująca:

- ✓ konstruuje się cele, dzieląc odcinek [0, 1] na n rozłącznych odcinków o długości $1/n$,

$$\left[0,\frac{1}{n}\right),\left[\frac{1}{n},\frac{2}{n}\right),\left[\frac{2}{n},\frac{3}{n}\right),\ldots,\left[\frac{n-1}{n},1\right],$$

- ✓ wyznacza się wartości dystrybuanty hipotetycznej dla wszystkich wartości reszt modelu $F(e_i)$ (dla $i = 1, 2, ..., n$),
- ✓ należy sprawdzić, do których cel należą wyznaczone wartości dystrybuanty,
- ✓ wyznacza się liczbę k pustych celi, do których nie wpadła żadna wartość $F(e_i)$.

Obszar krytyczny testu jest dwustronny:

$$\theta = \left\{k: P(k \le k_1) = \frac{\alpha}{2}\right\} \cup \left\{k: P(k \ge k_1) = \frac{\alpha}{2}\right\}. \tag{56}$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki k nie wpada do obszaru krytycznego ($k \in (k_1, k_2)$), to nie ma podstaw do odrzucenia hipotezy H_0 o normalności rozkładu składników losowych.
Cele, w analizowanym przypadku, to 16 odcinków o długości 1/16.

Tabela 16. Wyznaczone cele w teście Hellwiga

Nr celi	Początek	Koniec
1	0,000	0,063
2	0,063	0,125
3	0,125	0,188
4	0,188	0,250
5	0,250	0,313
6	0,313	0,375
7	0,375	0,438
8	0,438	0,500
9	0,500	0,563
10	0,563	0,625
11	0,625	0,688
12	0,688	0,750
13	0,750	0,813
14	0,813	0,875
15	0,875	0,938
16	0,938	1,000

Reszty modelu, standaryzowane reszty, wartość dystrybuanty oraz nr celi, do której „wpada” dystrybuanta przedstawiono w tabeli 17.

Tabela 17. Reszty i dystrybuanta dla weryfikowanego modelu ekonometrycznego

Nr obserwacji	Składniki resztowe	Std. składniki resztowe	Dystrybuanta	Cela
10	-1,40481	-1,09761	0,136187412	3
5	-1,20874	-0,94441	0,172480087	4
3	-1,10957	-0,86693	0,192990182	4
17	-1,08856	-0,85052	0,197518023	4
13	-1,04528	-0,8167	0,207049946	4
12	-0,84615	-0,66112	0,25426768	5
15	-0,67925	-0,53071	0,297809878	6
16	-0,37133	-0,29013	0,385858393	7
9	-0,30933	-0,24168	0,404514064	8
8	-0,21176	-0,16546	0,434290966	8
14	-0,09356	-0,0731	0,470863271	9
18	-0,01762	-0,01376	0,494510727	9

2	0,001808	0,001413	0,500563705	10
4	0,479373	0,374544	0,646000186	12
7	0,891884	0,696848	0,757051038	14
1	1,269907	0,992205	0,839451233	16
11	2,477313	1,935577	0,973540235	18
6	3,265675	2,55154	0,994637599	18

Puste cele to cele o numerach: 1, 2, 11, 13, 15, 17. Liczba pustych cel K=6. Krytyczne liczby pustych cel dla 18 obserwacji dla przyjętego poziomu istotności α=0,05 wynosząK_1 = 3 oraz K_2 = 9. Nie ma zatem podstaw do odrzucenia hipotezy H_0 że składniki losowe mają rozkład normalny N(0;1,319274).

Autokorelacja

W kolejnym kroku, zbadano czy wraz ze wzrostem długości trasy występuje autokorelacja składników losowych rzędu pierwszego. Hipotezę o braku autokorelacji składników losowych zweryfikowano testem Durbina-Watsona:

$H_0 : \rho(\varepsilon_t,\varepsilon_{t-1}) = 0$

$H_0 : \rho(\varepsilon_t,\varepsilon_{t-1}) < 0$ (57)

Sprawdzianem zespołu hipotez jest statystyka:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad (58)$$

Empiryczna wartość statystyki d = 2,15911. Wartość krytyczne d_l = 4-1,39 =2,61 oraz d_U = 4 – 1,16 = 2,84. Nie ma zatem podstaw do odrzucenia hipotezy o braku autokorelacji składników rzędu pierwszego

Symetria

Do sprawdzenia symetrii składnika losowego zastosujemy test symetrii składników losowych. Stawiamy hipotezy:

$$H_{0:} \quad p_+ = \frac{1}{2}, \qquad H_{1:} \quad p_+ \neq \frac{1}{2}. \quad (59)$$

$$t = \frac{\frac{m}{n} - \frac{1}{2}}{\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n-1}}} \quad (60)$$

Statystyka ta, przy prawdziwości hipotezy H_0, ma rozkład t Studenta o 17 stopniach swobody. Empiryczna wartość statystyki wynosi –1,45774. Wartość krytyczna2,11.

Nie ma zatem podstaw do odrzucenia hipotezy, że rozkład składników losowych jest symetryczny.

Losowość

Stawiamy hipotezę: H_0: reszty modelu są losowe. Zweryfikujemy ją testem liczby serii.

Należy zatem, uporządkować reszty chronologicznie lub zgodnie z rosnącymi wartościami jednej ze zmiennych objaśniających, a następnie wyznaczyć liczbę serii L reszt tych samych znaków. Przy prawdziwości hipotezy H_0 zmienna losowa L podlega rozkładowi liczby serii dla m elementów jednego rodzaju (reszty dodatnie) oraz ($n - m$) elementów drugiego rodzaju (reszty ujemne).

Obszar krytyczny testu jest dwustronny:

$$\theta = \left\{L: P(L \leq L_1) = \frac{\alpha}{2}\right\} \cup \left\{L: P(L \geq L_2) = \frac{\alpha}{2}\right\} \tag{61}$$

Jeżeli zatem wyznaczona wartość empiryczna statystyki nie wpada do obszaru krytycznego $L \in (L_1, L_2)$, to nie ma podstaw do odrzucenia hipotezy H_0 o losowości reszt modelu.

W analizowanym przykładzie reszty zostały uporządkowane względem rosnących wartości długości tras, a następnie zliczono liczbę serii, która wynosi $L = 10$. Krytyczne wartości liczby serii dla 6 reszt dodatnich i 12 reszt ujemnych, na przyjętym poziomie istotności $\alpha = 0{,}05$ wynoszą 4 i 12. Empiryczna wartość statystyki nie wpada w obszar krytyczny $– 4 < L = 10 < 12$.

Nie ma podstaw do odrzucenia hipotezy o losowości reszt modelu.

Homosedastyczność

Stałość wariancji składnika losowego zbadano testem Spearmana. Testem tym można sprawdzić, czy wariancja składników losowych rośnie (maleje) wraz ze wzrostem wartości zmiennej objaśniającej X.

$$H_0 : \rho(|\varepsilon_x|, x) = 0$$
$$H_0 : \rho(|\varepsilon_x|, x) \neq 0 \tag{62}$$

Sprawdzenie zespołu hipotezy jest statystyka

$$r = r(|\varepsilon|, x) = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \tag{63}$$

gdzie: D_i – różnica rang zmiennej i modułu reszty *e dla i-tej obserwacji.*

Tabela 18. Obliczenia testu korelacji rang Spearmana

Odległość *X*	**Ranga *X***	**Składniki resztowe**	**Moduł *e***	**Ranga $\|e\|$**	***D***	D^2
suma						612
344,60	1,00	-1,09	1,09	12,00	-11,00	121,00
373,20	2,00	-0,31	0,31	5,00	-3,00	9,00
585,80	3,00	0,00	0,00	1,00	2,00	4,00
630,30	4,00	-0,09	0,09	3,00	1,00	1,00
679,00	5,00	-1,11	1,11	13,00	-8,00	64,00
682,20	6,00	-0,37	0,37	6,00	0,00	0,00
691,50	7,00	0,48	0,48	7,00	0,00	0,00
968,80	8,00	3,27	3,27	18,00	-10,00	100,00
1030,70	9,00	-0,02	0,02	2,00	7,00	49,00
1247,00	10,00	2,48	2,48	17,00	-7,00	49,00
1598,10	11,00	-1,21	1,21	14,00	-3,00	9,00
1617,20	12,00	-0,21	0,21	4,00	8,00	64,00
1626,60	13,00	-1,05	1,05	11,00	2,00	4,00
1788,00	14,00	-0,68	0,68	8,00	6,00	36,00
1992,50	15,00	-0,85	0,85	9,00	6,00	36,00
2317,10	16,00	1,27	1,27	15,00	1,00	1,00
2925,80	17,00	-1,40	1,40	16,00	1,00	1,00
3398,90	17,00	0,89	0,89	10,00	7,00	49,00

Rangi (1, 2, ..., *n*) przypisujemy kolejno wartościom zmiennej *X* (reszt *e*) uporządkowanym w ciąg niemalejący. Jeżeli wystąpią takie same wartości zmiennej *X* (reszt *e*),to przypisujemy im rangę równą średniej arytmetycznej odpowiadających im pozycji w ciągu.
Na podstawie obliczeń z tabeli 18 wyznaczono wartość empiryczną statystyki równą $r = 0{,}1$. Obszar krytyczny testu jest dwustronny. Na poziomie istotności $\alpha = 0{,}05$ wartość krytyczna statystyki Spearmana wynosi 0,399.Nie ma zatem podstaw do odrzucenia hipotezy H_0 o homoskedastyczności składników losowych.

Podsumowując, model ekonometryczny możemy zatem uznać $\widehat{czas} = 2{,}43 + 0{,}009droga$ za poprawny.

4.2.5 Wnioskowanie na podstawie modelu

Skonstruowany model może być stosowany między innymi do budowy prognoz. Wyróżnia się trzy rodzaje prognoz (predykcji ekonometrycznych).

Prognoza punktowa. Jest to prognoza warunkowej wartości oczekiwanej zmiennej objaśnianej *Y* dla ustalonych wartości zmiennych objaśniających $x_0 = (x_{01}, x_{01}, \dots, x_{0k})$na podstawie zbudowanego równania regresji$\hat{y}_0 = a_0 + a_1 x_{01} + a_2 x_{02} + \cdots + + a_k x_{0k}$.

Prognoza przedziałowa wartości zmiennej objaśnianej *Y*. Jest to przedział losowy mający następującą postaci:

$$\left(\hat{y}_0 - t_\alpha S_\varepsilon \sqrt{1 + x_0^T (X^T X)^{-1} x_0}, \hat{y}_0 + t_\alpha S_\varepsilon \sqrt{1 + x_0^T (X^T X)^{-1} x_0}\right) \quad (64)$$

gdzie: t_α– wartość krytyczna rozkładu *t* Studenta o *n* – *k* – 1 stopniach swobody odpowiadająca przyjętemu poziomowi ufności 1 – *α* taka, że:

$$\{P(|t| \geq t_\alpha) = \alpha\} \quad (65)$$

S_ε– estymator odchylenia standardowego składnika losowego modelu ekonometrycznego.

Prognoza przedziałowa wartości oczekiwanej zmiennej objaśnianej *Y*. Dla ustalonego poziomu ufności 1 – *α* jest to przedział losowy postaci:

$$\left(\hat{y}_0 - t_\alpha S_\varepsilon \sqrt{x_0^T (X^T X)^{-1} x_0}, \hat{y}_0 + t_\alpha S_\varepsilon \sqrt{x_0^T (X^T X)^{-1} x_0}\right) \quad (66)$$

5. Regresja logistyczna

W analizie danych często spotyka się zmienne dychotomiczne, przyjmujące wartości 0 lub 1, które wskazują wystąpienie lub brak określonego zdarzenia, np. zmienna *Wystąpienie_Wypadku* (*Y* = 1 – wystąpił wypadek, *Y* = 0 – brak wypadku), czy *Przeżycie* (*Y* = 1– pacjent przeżył, *Y* = 0 – pacjent zmarł).

W takich przypadkach istotne jest określenie czynników (zmienne objaśniające X_1, X_2, …, X_k), które w istotny sposób wpływają na prawdopodobieństwo wystąpienia danego zdarzenia. Do tego celu stosuje się model regresji logistycznej, który umożliwia modelowanie zależności między zmiennymi objaśniającymi a zmienną dychotomiczną.

W klasycznej regresji liniowej zakłada się, że zmienna zależna *Y* jest ilościowa i może przyjmować dowolne wartości rzeczywiste. W regresji logistycznej natomiast *Y* przyjmuje tylko dwie wartości (0 lub 1), dlatego model opisuje nie wartość *Y* bezpośrednio, lecz **prawdopodobieństwo** wystąpienia zdarzenia *Y*= 1.

Model ma postać:

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \quad (67)$$

gdzie:
$P(Y=1)$ – prawdopodobieństwo wystąpienia zdarzenia,
β_0 – wyraz wolny,
β_i – współczynniki regresji,
X_i – zmienne objaśniające (predyktory).
Model można również zapisać w formie logitowej, czyli po przekształceniu funkcji logistycznej:

$$ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \quad (68)$$

Lewa strona równania, czyli **logit**, stanowi logarytm ilorazu szans (*odds ratio*), a prawa to liniowa kombinacja predyktorów.
Każdy współczynnik β_i informuje o wpływie danej zmiennej X_i na logarytm ilorazu szans.
W praktyce jednak częściej interpretuje się **iloraz szans (odds ratio)**, wyrażony jako:

$$OR_i=e^{\beta_i} \quad (69)$$

jeśli $OR_i > 1$, wzrost zmiennej X_i zwiększa prawdopodobieństwo wystąpienia zdarzenia $Y = 1$,
jeśli $OR_i < 1$, wzrost zmiennej X_i zmniejsza to prawdopodobieństwo,
jeśli $OR_i = 1$, zmienna X_i nie ma wpływu na wystąpienie zdarzenia.

Przykład 13
Zakłada się, że analizowane jest prawdopodobieństwo zakupu produktu przez klienta w zależności od jego **dochodu (X_1)** oraz **posiadania karty lojalnościowej (X_2)**.
Zmienna zależna Y przyjmuje wartość:

- 1 – klient dokonał zakupu,
- 0 – klient nie dokonał zakupu.

Tabela 19. Wyniki estymacji modelu logistycznego

Zmienna	Współczynnik β_i	Wartość e^{β_i} (OR)	Interpretacja
Stała ($\beta 0$)	-3.2	–	–
Dochód (X_1)	0.004	1.004	Wzrost dochodu o 1 jednostkę zwiększa szansę zakupu o 0,4%.
Karta	1.8	6.05	Klienci posiadający kartę mają 6-krotnie

lojalnościowa (X_2)			większe szanse zakupu niż ci bez karty.

Na podstawie wyników można zapisać równanie modelu:

$$ln\left(\frac{p}{1-p}\right) = -3{,}2 + 0{,}004X_1 + 1{,}8X_2$$

Przykładowo, dla klienta o dochodzie 3000 zł posiadającego kartę lojalnościową (X_2= 1)

$$z = -3{,}2 + 0{,}004 \cdot 3000 + 1{,}8 \cdot 1 = 10{,}6$$

$$p = \frac{1}{1 + e^{-10{,}6}} \approx 0{,}999975$$

Oznacza to, że prawdopodobieństwo zakupu przez takiego klienta wynosi około 99,9%.

Bibliografia

1. Aczel A.D. (2000). Statystyka w zarządzaniu. Pełny wykład, PWN, Warszawa.
2. Biecek P. (2014). Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce prezentowania danych, Fundacja Naukowa SmarterPoland.pl, Warszawa.
3. Borkowski B., Dudek H., & Szczęsny W. (2004). Ekonometria: wybrane zagadnienia, PWN, Warszawa.
4. Bowerman B.L., & O'Connell R.T. (2007). Business Statistics in Practice (4th ed.), McGraw-Hill, Irwin.
5. Chow G.C. (1995). Ekonometria, PWN, Warszawa.
6. Goryl A., Jędrzejczyk Z., Kukuła K., Osiewalski J., & Walkosz A. (1996). Wprowadzenie do ekonometrii w przykładach i zadaniach, PWN, Warszawa.
7. Górecki T. (2011). Podstawy statystyki z przykładami w R, BTC, Legionowo.
8. Grysa K., & Maciąg A. (1997). Podstawy ekonometrii, Wydawnictwo Wyższej Szkoły Handlowej, Kielce.
9. Hastie T., Tibshirani R., & Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, Berlin.
10. Koronacki J., & Mielniczuk J. (2001). Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, Warszawa.
11. Maciąg A., Pietroń R., & Kukla S. (2013). Prognozowanie i symulacja w przedsiębiorstwie, Polskie Wydawnictwo Ekonomiczne, Warszawa.
12. Nowak E. (1994). Zarys metod ekonometrii. Zbiór zadań, WNT, Warszawa.
13. Welfe W., & Welfe A. (1996). Ekonometria stosowana, Państwowe Wydawnictwo Ekonomiczne, Warszawa.