



EKSPLORACJA DANYCH, TESTOWANIE HIPOTEZ BADAWCZYCH I MODELOWANIE ZALEŻNOŚCI – PRZYKŁADY W STATISTICA 9

Janusz Wątroba, StatSoft Polska Sp. z o.o.

Praktyczne przeprowadzenie analizy zgromadzonych danych składa się zazwyczaj z kilku etapów. Na każdym z tych etapów analityk potrzebuje różnych narzędzi wspomagających przebieg analizy. Przy wstępnej eksploracji danych przydatne są narzędzia do łatwej (analitycznej i graficznej) identyfikacji obserwacji nietypowych oraz możliwość szybkiej oceny ich wpływu na charakterystyki liczbowe rozkładów analizowanych zmiennych. Z kolei na etapie testowania postawionych hipotez badawczych potrzebny jest dostęp do metod sprawdzania założeń, występujących w przypadku określonych testów oraz szeroki zakres wyboru różnych testów parametrycznych i ich nieparametrycznych odpowiedników. W przypadku modelowania związków pomiędzy analizowanymi zmiennymi wymagany jest zarówno dostęp do predefiniowanych modeli, łatwe modyfikowanie modelu (np. dodanie kolejnego predyktora) jak również możliwość estymacji parametrów modelu, zdefiniowanego samodzielnie przez użytkownika.

W trakcie wystąpienia zostaną pokazane przykłady stosowania wspomnianych powyżej narzędzi, wspomagających prowadzenie analiz statystycznych w środowisku najnowszej wersji programu *STATISTICA* [7].

Wstępne badanie (eksploracja) danych

Przed przystąpieniem do właściwej analizy zalecane jest wstępne badanie danych, które obejmuje m.in. różne graficzne techniki, pozwalające na wielostronne spojrzenie na dane pod kątem ustalenia ewentualnych trendów i zależności oraz odróżnienie wyników obserwacji istotnie odbiegających od pozostałych. Metody te są często określane wspólnym terminem *eksploracyjna analiza danych* [1]. Niektóre z proponowanych metod oraz sama nazwa pochodzi od amerykańskiego statystyka Johna Tukeya.

Doświadczenie analityków-praktyków również wskazuje na to, że sprawdzenie danych jest warunkiem koniecznym do uzyskania poprawnych wyników analizy. Chodzi tutaj o dwie ważne kwestie. Po pierwsze badacz (lub analityk) musi sprawdzić, czy wśród zgromadzonych danych nie ma błędnych pomiarów lub obserwacji, gdyż takie dane mogą bardzo mocno wypaczyć wyniki analizy i w dalszej konsekwencji spowodować błędne



wnioskowanie. Po drugie, należy sprawdzić dane pod kątem występowania tzw. obserwacji odstających lub nietypowych (*outliers*). Obserwacje takie mogą dodatkowo zwiększać zmienność badanych cech i w ten sposób powodować zjawisko polegające na obciążeniu (*bias*) oceny rzeczywistego efektu, którego badacz poszukuje. Problem tego typu może się pojawić np. w zastosowaniach statystycznej analizy danych przy opracowywaniu wyników badań klinicznych nowych związków lub procedur medycznych.

Zadaniem analityka jest zidentyfikowanie takich danych oraz ewentualna ocena ich wpływu na wyniki analizy. Możliwe są trzy podejścia:

- ♦ usunięcie takich obserwacji z analizowanego zbioru danych i zastosowanie klasycznych technik analizy,
- ♦ pozostawienie ich w analizowanym zbiorze danych i zastosowanie tzw. metod odpornych (ang. *robust methods*),
- ♦ wykonanie analizy w dwóch wariantach: z pominięciem i z pozostawieniem wątpliwych obserwacji.

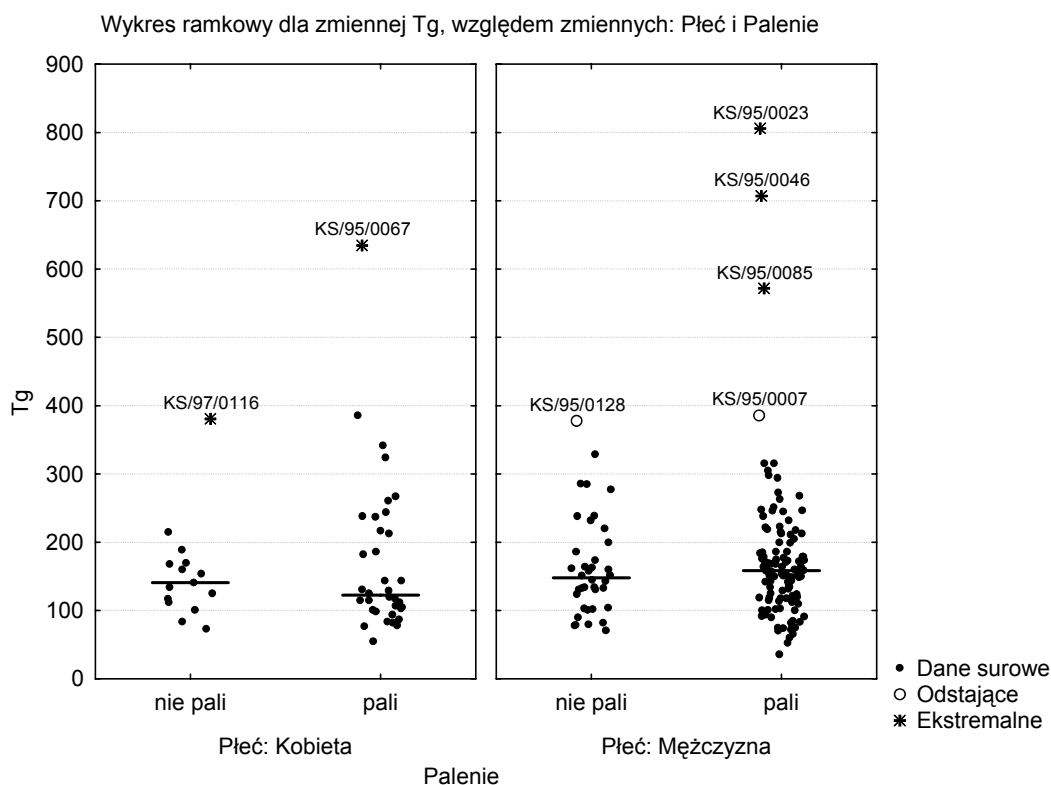
Ostateczna decyzja w tej sprawie powinna być zawsze podejmowana przez badacza. Warto nadmienić, że we wspomnianym wyżej przypadku badań farmakologicznych podawane są specjalne zalecenia, co do postępowania z odstającymi obserwacjami [2, 5].

W przykładzie praktycznym, który zostanie zaprezentowany poniżej ograniczono się do stosunkowo prostych metod identyfikacji odstających obserwacji (można jednak wskazać również propozycje wykorzystania bardziej zaawansowanych metod [4]). Pokazano również wpływ obserwacji zidentyfikowanych jako odstające na wartości statystyk określających tendencję centralną oraz charakter rozkładu.

W przykładzie wykorzystano dane dotyczące wybranych parametrów klinicznych i biochemicznych zgromadzonych dla pacjentów ze zdiagnozowaną chorobą niedokrwienną serca. Dane zostały opublikowane w książce Watały [8]. Dla potrzeb niniejszego przykładu przedstawiono wstępne badanie zmiennej Tg (stężenie triglicerydów w osoczu krwi) w grupach pacjentów wyróżnionych przez dwie zmienne jakościowe: *Płeć* i *Palenie*.

W celu czytelnej, graficznej identyfikacji odstających obserwacji w programie *STATISTICA* utworzono skategoryzowane wykresy ramkowe z obserwacjami odstającymi i ekstremalnymi, na których umieszczono surowe dane. Następnie, wykorzystując opcję wyróżniania danych na wykresie, punktom oznaczającym odstające obserwacje przypisano, odpowiadające im etykiety. Na rysunku poniżej widzimy wykres po zmodyfikowaniu paru elementów.

Korzystając z możliwości programu *STATISTICA*, informacje o wyróżnionych obserwacjach można zapisać w zmiennej, a następnie wykorzystywać ją w warunkach selekcji dla analiz i wykresów. Kontynuując przykład, sprawdzono wpływ odstających obserwacji na wartości statystyk opisowych zmiennej Tg w porównywanych grupach pacjentów. Poniżej (rys. 2) tabela z wynikami obliczeń dla wszystkich obserwacji.



Rys. 1. Skategoryzowany wykres ramkowy z surowymi danymi i obserwacjami odstającymi.

Tabela statystyk opisowych w grupach							
Minimalne N (wszystkie zmn): 201							
Płeć	Palenie	Tg Ważnych	Tg Średnie	Tg Mediany	Tg Odch.std	Tg Q25	Tg Q75
Kobieta	nie pali	15	154,93	141,00	73,564	112,00	170,00
Kobieta	pali	36	170,53	122,50	114,951	100,00	227,00
Mężczyzna	nie pali	38	164,53	148,00	73,957	104,00	200,00
Mężczyzna	pali	112	175,64	158,50	108,713	118,00	202,50
Ogół grup		201	171,08	151,00	101,498	112,00	200,00

Rys. 2. Wyniki obliczeń podstawowych statystyk opisowych dla wszystkich obserwacji.

Identyczne obliczenia wykonano również po wykluczeniu z analizy obserwacji odstających. Uzyskane wyniki w różnym stopniu zmieniają sposób wnioskowania na temat różnic w zakresie podstawowych statystyk opisowych charakteryzujących porównywane grupy badanych.

Tabela statystyk opisowych w grupach							
Minimalne N (wszystkie zmn): 194							
Warunek uwzględniania: v15=0							
Płeć	Palenie	Tg Ważnych	Tg Średnie	Tg Mediany	Tg Odch.std	Tg Q25	Tg Q75
Kobieta	nie pali	14	138,79	137,50	40,198	112,00	168,00
Kobieta	pali	35	157,26	120,00	84,118	99,00	217,00
Mężczyzna	nie pali	37	158,76	145,00	65,738	104,00	186,00
Mężczyzna	pali	108	159,27	155,50	61,284	117,00	186,00
Ogół grup		194	157,33	149,50	65,354	110,00	186,00

Rys. 3. Wyniki obliczeń podstawowych statystyk opisowych po wykluczeniu obserwacji odstających.



Na podstawie zawartych w tabeli wyników można zauważyć, że usunięcie odstających obserwacji wpłynęło na ocenę tendencji centralnej badanej zmiennej (mocniej na średnią). Stosunkowo najmniejszy efekt obserwowano w grupie niepalących mężczyzn (wartość średniej obniżyła się o blisko 6 mg/100 ml), a największy u palących mężczyzn i niepalących kobiet (wartość średniej obniżyła się o ponad 16 mg/100 ml). Duży efekt obserwowano również w przypadku miar opisujących rozproszenie danych. Przykładowo w grupie palących mężczyzn odchylenie standardowe zmalało o ponad 47 mg/100 ml.

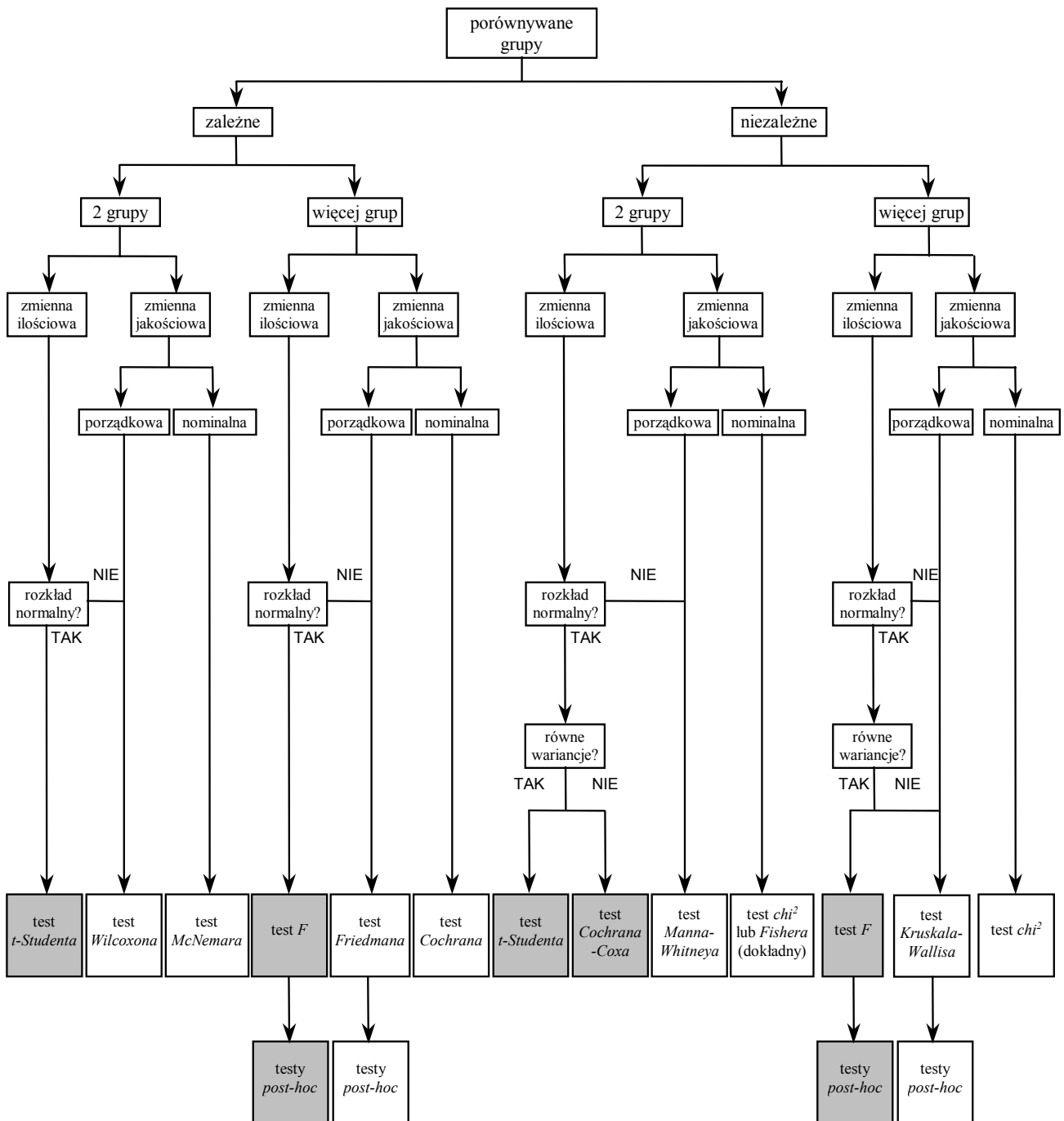
W tym miejscu warto podkreślić, że przy ocenie wyników analizy bardzo wygodnym narzędziem wspomagania analizy są skoroszyty. W programie *STATISTICA* skoroszyty dostarczają nie tylko wygodny mechanizm dokumentowania wszystkich wyników prowadzonej analizy, ale pozwalają również na łatwe powtórzenie analizy na zmienionych zbiorach danych.

Testowanie hipotez badawczych

W badaniach empirycznych prowadzonych w różnych dziedzinach najczęściej mamy do czynienia z sytuacją, w której zebranie wszystkich potencjalnych danych opisujących zjawisko będące przedmiotem zainteresowania nie jest możliwe. W takiej sytuacji o badanym zjawisku można wypowiedzieć się tylko na podstawie zebranych w odpowiedni sposób danych częściowych. Problematyka ta stanowi przedmiot *wnioskowania statystycznego*, które polega na uogólnianiu wniosków wynikających z badania częściowego i wykracza poza informacje wynikające ze zgromadzonych danych.

W obrębie problemów występujących we wnioskowaniu statystycznym wyróżnia się dwa obszary: zagadnienie *estymacji* nieznanymi parametrów i zagadnienie *testowania hipotez*. Jeśli chodzi o testowanie hipotez to ważnym praktycznym problemem, jaki pojawia się przed badaczem, jest wybór odpowiedniego testu statystycznego, który pozwoli wiarygodnie ocenić prawdziwość stawianej hipotezy. Wśród testów statystycznych największe znaczenie mają tzw. *testy istotności*. O wyborze odpowiedniego testu decyduje fakt spełnienia bądź niespełnienia kilku założeń dotyczących danych, na podstawie których oceniana jest prawdziwość postawionej hipotezy. Przy wyborze najczęściej stosowanych testów wygodnym rozwiązaniem jest korzystanie z uproszczonego schematu. Poniżej (rys. 4) przedstawiono przykład takiego schematu.

Praktyczne korzystanie ze schematu wymaga przejścia kilku etapów. Jak to zostało wcześniej zasygnalizowane, w każdym kroku sprawdza się, czy zgromadzone dane empiryczne spełniają określone warunki. W przedstawionych dalej przykładach zostanie pokazane praktyczne korzystanie ze schematu oraz wyniki testowania przykładowych hipotez i ich interpretacja na podstawie rzeczywistych danych.



Rys. 4. Schemat wyboru popularnych testów służących do oceny istotności różnic (zaszarzone pola oznaczają testy parametryczne a niezaszarzone testy nieparametryczne).

W pierwszym z prezentowanych przykładów celem analizy będzie dobór odpowiedniego testu dla sprawdzenia hipotezy dotyczącej równości wartości przeciętnych w dwóch porównywanych populacjach. W przykładzie zostaną wykorzystane wyniki badań przeprowadzonych w grupie 33 pacjentów ze zdiagnozowaną chorobą nadciśnieniową. Zebrane dane potraktujemy jako reprezentatywne dla populacji osób z nadciśnieniem. Badana

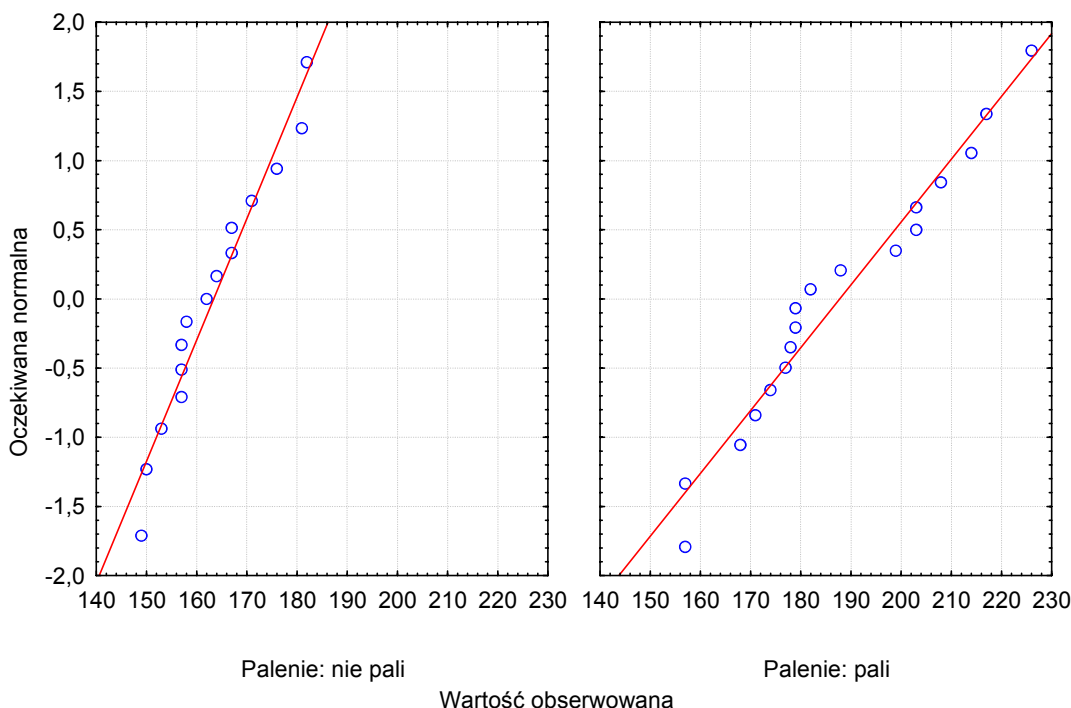


hipoteza dotyczy oceny zróżnicowania przeciętnego poziomu skurczowego ciśnienia krwi pomiędzy osobami palącymi i niepalącymi.

Biorąc pod uwagę postać postawionej hipotezy, można potraktować porównywane grupy jako niezależne. Badana zmienna jest zmienną wyrażoną na skali ilościowej. W związku z tym przy wyborze odpowiedniego testu należy sprawdzić założenia dotyczące normalności rozkładu badanej zmiennej w porównywanych grupach oraz równość wariancji. Poniżej przedstawiono wyniki odpowiednich testów.

Wykres normalności dla zmiennej Ciśnienie skurczowe; kategorie względem zmiennej Palenie

Palenie: nie pali Ciśnienie skurczowe: SW-W = 0,9383; p = 0,3612
 Palenie: pali Ciśnienie skurczowe: SW-W = 0,9496; p = 0,4191



Zmienna	Testy t; Grupująca: Palenie					
	Grupa 1: nie pali Grupa 2: pali					
	Levene'a F(1,df)	df	p	Brn-Fors F(1,df)	df	p
Ciśnienie skurczowe	8,772244	31	0,005825	4,721320	31	0,037554

Rys. 5. Wyniki testowania normalności rozkładu i równości wariancji.

Tak więc kierując się zaproponowanym schematem do testowania postawionej hipotezy zastosowano test Cochran-Coxa (test z niezależną estymacją wariancji). Jego wyniki przedstawiono w poniższej tabeli.

Zmienna	Testy t; Grupująca: Palenie, Grupa 1: nie pali, Grupa 2 pali							
	Średnia nie pali	Średnia pali	t oddz. est.war.	df	p dwustron	Średnia 1 - Średnia 2	Ufność -95,000%	Ufność +95,000%
Ciśnienie skurczowe	163,4000	187,7778	-4,39551	26,24	0,00016	-24,38	-36,3259	-12,4297

Rys. 6. Wyniki testu Cochran-Coxa wraz przedziałową oceną różnic.



Otrzymane wyniki umożliwiają ocenę statystycznej istotności różnic wartości przeciętnych w porównywanych populacjach. W oparciu o punktową i przedziałową ocenę efektu zróżnicowania badacz może przeprowadzić praktyczną ocenę wielkości występującego zróżnicowania.

Modelowanie związków pomiędzy zmiennymi

Badanie związków zachodzących pomiędzy interesującymi badacza zmiennymi jest zagadnieniem bardzo często podejmowanym w analizach. Znajomość tych związków ma ogromne znaczenie dla przewidywania przyszłego kierunku i tempa rozwoju interesujących badacza wielkości. Wykrywanie, obserwacja i mierzenie natężenia tego rodzaju związków jest również ważnym instrumentem poznania naukowego oraz ułatwia podejmowanie decyzji w różnych dziedzinach działalności praktycznej.

Truizmem jest stwierdzenie, że wiele spośród zjawisk i procesów występujących w otaczającej nas rzeczywistości ma złożony charakter. Efektem tego są trudności, na jakie napotyka się przy próbie ich adekwatnego opisu. Bardzo często stosowanym podejściem przy rozwiązywaniu tego problemu jest uproszczone odwzorowanie rzeczywistości. Zespół metod wykorzystywanych do tego celu jest nazywany *modelowaniem statystycznym* [6].

Uproszczone odwzorowanie rzeczywistych związków występujących pomiędzy badanymi zjawiskami wymaga od badacza umiejętnego wydobycia istoty mechanizmu, który wygenerował dane, i przekształcenia go do postaci umożliwiającej zastosowanie podejścia statystycznego. Najczęściej sprowadza się to do *przyjęcia określonej matematycznej formuły ujmującej powiązania pomiędzy mierzonymi zmiennymi* oraz założeń dotyczących *losowych procesów wpływających na wyniki pojedynczych pomiarów*. W ten sposób powstaje *statystyczny model* zjawiska. Dopasowanie takiego modelu do określonego zbioru danych empirycznych daje podstawę do uogólnienia wyników w szerszym kontekście lub do przewidywania wyników w przyszłości, co często prowadzi do lepszego wyjaśnienia badanego zjawiska.

Model jest pojęciem abstrakcyjnym, swoistym pomostem między abstrakcyjnymi sposobami myślenia a realnie istniejącą rzeczywistością. Przedstawia on pewne wyodrębnione, obiektywnie istniejące relacje, które odwzorowuje za pomocą użytecznych reguł, pozwalających „symulować” zachowanie i własności przedstawionego fragmentu rzeczywistości. Dobrze skonstruowany model w adekwatny sposób odtwarza badane obiekty, zjawiska lub procesy i powinien stanowić kompromis między nadmiernym uproszczeniem rzeczywistości a zbyt dużym nagromadzeniem szczegółów.

Oprócz przyczyn filozoficznych preferowanie prostszych modeli wynika również z powodów czysto praktycznych. Proste modele wymagają zazwyczaj niższych kosztów w przypadku ich praktycznego wykorzystywania oraz kontroli wyników w przyszłości. Ponadto w przypadku prostszych modeli interpretacja wyników modelowania jest zazwyczaj łatwiejsza.

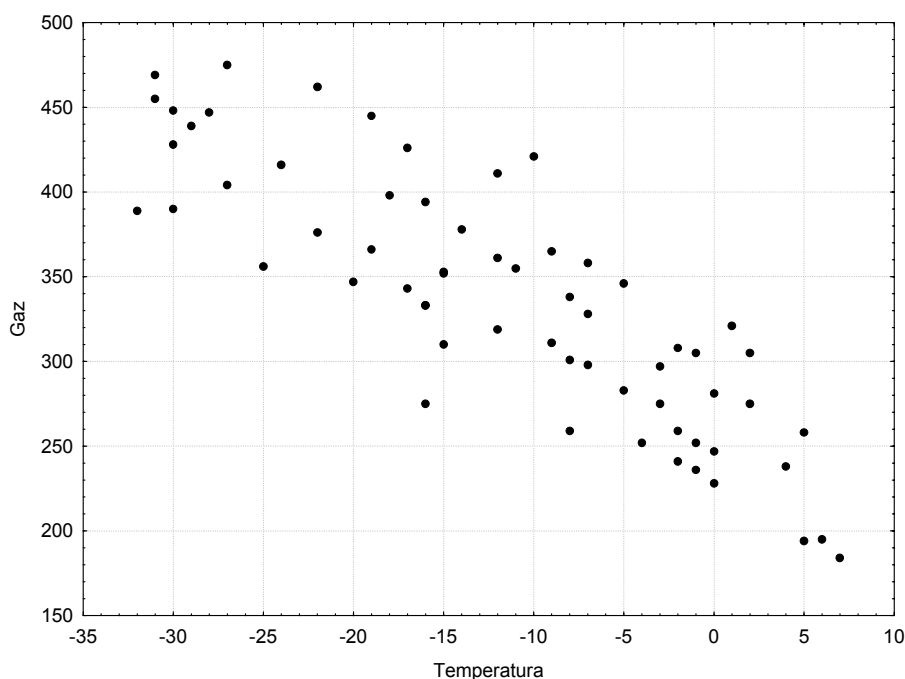
W przykładzie, który zostanie zaprezentowany w dalszej części wykorzystano dane opisujące zużycie gazu w kilku głównych miastach północnej części USA. Dane dotyczą



jednego sezonu grzewczego (zima). Spośród zmiennych mających potencjalny wpływ na wielkość zużycia gazu wykorzystano informacje o średniej temperaturze dobowej oraz średniej dobowej prędkości wiatru.

W procesie modelowania zostanie użyta metoda regresji liniowej. Najpierw zostanie zbudowany model regresji prostej, a następnie model regresji wielorakiej.

W przypadku modelu regresji prostej, gdy w modelu jest uwzględniana tylko jedna zmienna objaśniająca (niezależna lub predyktor), warto rozpocząć od utworzenia zwykłego wykresu rozrzutu dla dwóch zmiennych. Chodzi o to, żeby przekonać się, czy przyjęcie modelu liniowego jest słuszne. Poniżej zamieszczono taki wykres.



Rys. 7. Wykres rozrzutu dla zmiennej *Gaz* względem zmiennej *Temperatura*.

Zamieszczony powyżej wykres pozwala zaobserwować stopniowy spadek przeciętnego zużycia gazu wraz ze wzrostem średniej temperatury dobowej. Wydaje się ponadto, że charakter tej zależności jest zbliżony do liniowej. W związku z tym przy budowie modelu zastosowano technikę regresji liniowej prostej. Najważniejsze wyniki analizy zawiera tabela na rys. 8.

Zawarte w tabeli wyniki modelowania umożliwiają statystyczną i merytoryczną ocenę zbudowanego modelu. Okazuje się, że parametry strukturalne modelu istotnie różnią się od zera (do oceny istotności wykorzystuje się test *t Studenta*). Obliczone z próby oceny parametrów informują, że zwiększenie temperatury o 1 stopień Celsjusza obniża przeciętne zużycie gazu o 5,8 jednostki.

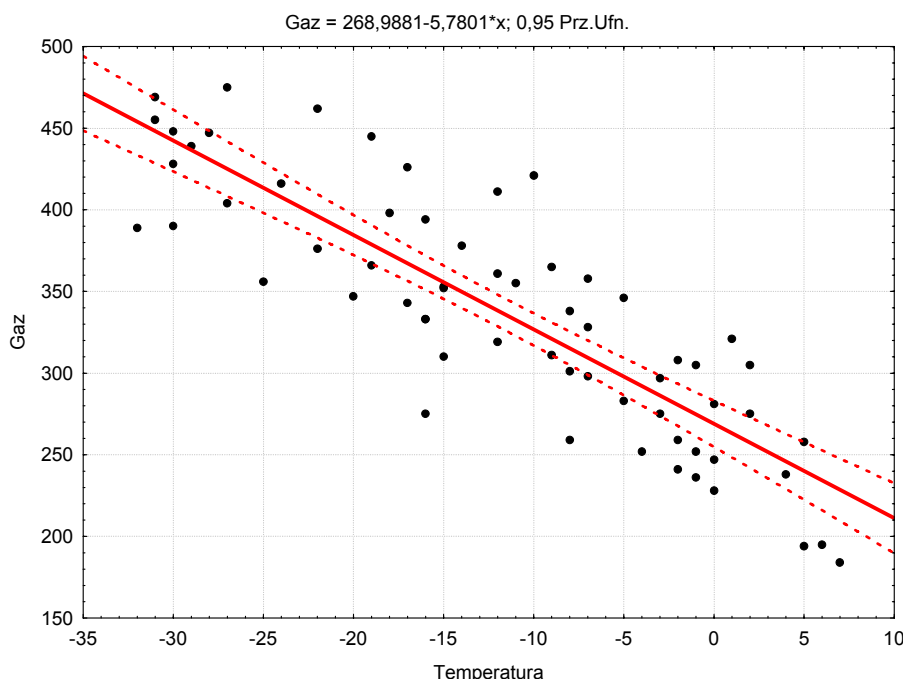


Podsumowanie regresji zmiennej zależnej: Gaz (Zużycie gazu)						
R=0,85736 R ² =0,73507 Skoryg. R ² =0,73073						
F(1,61)=169,25 p<0,0000 Błąd std. estymacji: 38,694						
N=63	b*	Bł. std. b*	b	Bł. std. b	t(61)	p
W. wolny			268,9881	7,131213	37,7198	0,000000
Temperatura	-0,857363	0,065902	-5,7801	0,444298	-13,0096	0,000000

Rys. 8. Tabela z podstawowymi wynikami analizy regresji.

Przy ocenie stopnia dopasowania modelu do rzeczywistych danych stosowane są różne miary. Jedną z nich jest współczynnik determinacji (R^2). Jego wartość mówi o tym, w jakim stopniu oszacowany model wyjaśnia oryginalną wariancję wartości zmiennej zależnej. W opisywanym przykładzie jego wartość jest równa 0,7351 i oznacza, że zbudowany model tłumaczy około 73,5 % oryginalnej wariancji zmiennej zależnej. Z tego wynika, że około 26,5 % wariancji ma charakter losowy lub może być wyjaśnione wpływem innych nieuwzględnionych w modelu zmiennych niezależnych.

Budowanie modelu regresji ma zazwyczaj dwa cele. Pierwszy z nich to lepsze poznanie badanego zjawiska poprzez ilościowy opis charakteru i siły powiązania pomiędzy interesującymi badacza zmiennymi. Drugi cel jest bardziej praktyczny. Jeśli model dobrze pasuje do rzeczywistych danych, wówczas może zostać użyty do przewidywania lub symulacji wartości zmiennej zależnej przy określonych wartościach zmiennej lub zmiennych niezależnych. Miarą przeciętnego błędu prognozy jest standardowy błąd estymacji. W prezentowanym przykładzie jego wartość wyniosła 38,7 jednostki (co stanowi ok. 11,5 % średniej dla zmiennej zależnej). Graficzna postać modelu została przedstawiona na rysunku poniżej.



Rys. 9. Wykres rozrzutu z dopasowaną linią regresji i 95% przedziałami ufności.



W przypadku, gdy celem modelowania jest wykorzystanie modelu do prognozowania i przeprowadzania symulacji, dąży się to tego, aby przeciętny błąd prognozy był jak najmniejszy. Można to osiągnąć poprzez dobór innej postaci modelu (np. próba dopasowania modelu nieliniowego) lub wprowadzenie do modelu jednego lub większej liczby predyktorów. W opisywanym przykładzie zastosowano drugie z proponowanych podejść. Zbudowano model uwzględniający dodatkowo średnią dobową prędkość wiatru (zmienna *Wiatr*). Najważniejsze wyniki dla oszacowanego modelu zostały przedstawione w tabeli poniżej.

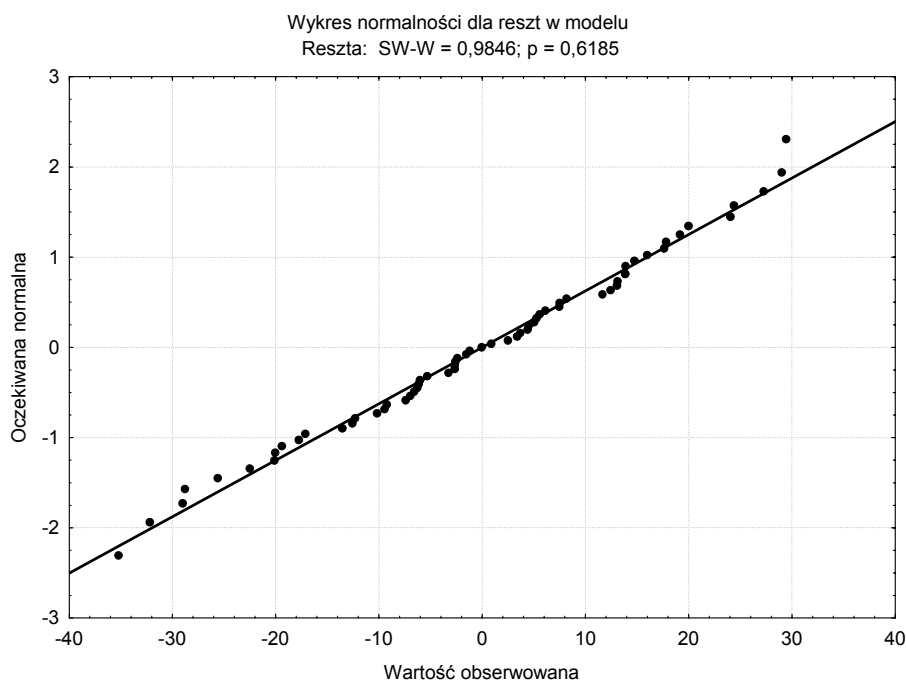
Podsumowanie regresji zmiennej zależnej: Gaz (Zużycie gazu)						
R=0,97789 R ² =0,95627 Skoryg. R ² =0,95482						
F(2,60)=656,10 p<0,0000 Błąd std. estymacji: 15,850						
N=63	b*	Bł. std. b*	b	Bł. std. b	t(60)	p
W. wolny			132,1401	8,380383	15,7678	0,000000
Temperatura	-0,563618	0,031828	-3,7998	0,214578	-17,7082	0,000000
Wiatr	0,554517	0,031828	12,6194	0,724329	17,4222	0,000000

Rys. 10. Wyniki modelowania po dodaniu predyktora *Wiatr*.

Na podstawie zamieszczonych w tabeli wyników można stwierdzić zdecydowaną poprawę dopasowania modelu do danych. Obliczona wartość współczynnika determinacji dla modelu przekracza 95,6%. Oznacza to, że tylko nieco ponad 4% wariacji zmiennej zależnej to wpływ zmienności losowej lub innych, nieuwzględnionych w modelu zmiennych. Jednocześnie błąd standardowy estymacji uległ obniżeniu do wartości 15,85 jednostki (stanowi to około 4,7% średniej modelowanej zmiennej zależnej). Interpretację ocen parametrów strukturalnych modelu dokonuje się zgodnie z tzw. zasadą *ceteris paribus* [3]. W przypadku otrzymanych wyników można oczekiwać, że zwiększenie prędkości wiatru o 1 km/godz. skutkuje zwiększeniem teoretycznego, czyli wynikającego z modelu, zużycia gazu o nieco ponad 12,6 jednostki, przy niezmiennych wartościach pozostałych zmiennych objaśniających. Ocena wyrazu wolnego równa 132,1 jednostki oznacza oszacowaną wartość zmiennej zależnej dla zerowych wartości predyktorów. W związku z tym ewentualna interpretacja tej wartości jest możliwa tylko w sytuacji, gdy sensownym jest założenie, że wszystkie uwzględnione w modelu predyktory przyjmują jednocześnie zerową wartość.

Przed przystąpieniem do interpretacji wyników modelowania powinno się sprawdzić spełnienie założeń. Większość z nich dotyczy składnika losowego modelu. W opisywanym przykładzie sprawdzono założenie normalności rozkładu reszt. Wyniki przedstawia wykres na rys. 11.

Na podstawie wykresu łatwo zobaczyć, że punkty oznaczające położenie obserwacji w stosunku do nałożonej linii prostej (oznaczającej zgodność empirycznych danych z rozkładem normalnym) w niewielkim stopniu od niej odbiegają. Na dobrą zgodność z rozkładem normalnym wskazują także wyniki analitycznego testu Shapiro-Wilka (prawdopodobieństwo testowe p jest mocno odległe od przyjmowanego standardowo granicznego poziomu istotności $\alpha=0,05$).



Rys. 11. Wykres ilustrujący rozkład reszt wraz z wynikami testu normalności Shapiro-Wilka.

Podsumowanie i wnioski

W artykule zaprezentowano wybrane zagadnienia związane z praktyczną stroną prowadzenia statystycznej analizy danych. Zwrócono uwagę na duże znaczenie *wstępnej analizy danych*. Jej staranne przeprowadzenie pozwala wyeliminować błędy w danych oraz zidentyfikować odstające obserwacje, które mogą prowadzić do niepoprawnych wniosków. W zakresie metod wnioskowania statystycznego podkreślono kwestię wyboru poprawnych metod. I wreszcie przy okazji przykładów modelowania związków pomiędzy zmiennymi pokazano korzyści z prowadzenia analizy od modeli podstawowych do bardziej zaawansowanych.

Przy okazji praktycznych przykładów zaprezentowano kilka wybranych, nowych narzędzi wspomagania analizy danych w programie *STATISTICA 9*.

Poniżej podsumowano omawiane kwestie w postaci kilku wniosków ogólnych:

- ◆ przed przystąpieniem do właściwej analizy zawsze warto przeprowadzić wstępne badanie danych,
- ◆ przy testowaniu hipotez statycznych należy zwracać uwagę na wybór odpowiednich testów statystycznych,
- ◆ przy modelowaniu związków pomiędzy zmiennymi warto stosować zarówno metody zaawansowane, jak i metody bardziej podstawowe,
- ◆ Stosowanie narzędzi programu *STATISTICA* wspomagających prowadzenie analizy umożliwia szybkie otrzymywanie wyników oraz ułatwia ich merytoryczną interpretację.



Literatura

1. Aczel A.D. (2000). Statystyka w zarządzaniu. Pełny wykład, PWN.
2. Chow S.C., Liu, J. P. (2004). Design and Analysis of Clinical Trials. Concepts and Methodologies, (2nd ed.), John Wiley and Sons, NJ: Erlbaum.
3. Ekonometria i badania operacyjne. Podręcznik dla studiów licencjackich, pod red. naukową M. Gruszczyńskiego, T. Kuszewskiego i M. Podgórskiej, PWN 2009.
4. Hadi A.S., Rahmatullah A.H.M., Werner M. (2009). Detection of Outliers, WIREs Comp Stat 1, Wiley.
5. ICH (1998). International Conference on Harmonization. Guideline E9 on Statistical Principles for Clinical Trials.
6. Krzanowski W.J. (1998). An Introduction to Statistical Modelling, Arnold.
7. StatSoft, Inc. (2009). *STATISTICA* (data analysis software system), version 9. www.statsoft.com.
8. Watała C. (2002). Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α -medica press.