



WSPOMAGANIE ANALIZY DANYCH ZA POMOCĄ NARZĘDZI STATISTICA

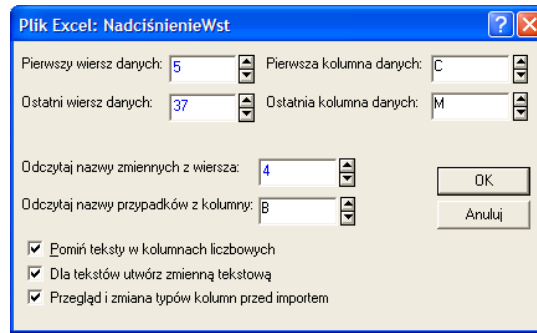
Janusz Wątroba i Grzegorz Harańczyk, StatSoft Polska Sp. z o.o.

Zakres zastosowań analizy danych w różnych dziedzinach działalności biznesowej i badaniach naukowych stale się poszerza. Wynika to w głównej mierze z coraz powszechniejszego przekonania, że przy rozwiązywaniu różnego rodzaju zagadnień poznawczych i praktycznych trzeba opierać się na empirycznych danych, opisujących badane zjawiska i procesy. Osoby i instytucje zajmujące się analizą danych dysponują coraz większą ilością danych, a jednocześnie rośnie nacisk na jak najszybsze otrzymywanie wyników analiz, zwłaszcza w tych dziedzinach, w których rezultaty analiz są wykorzystywane do podejmowania bieżących decyzji. W związku z tym coraz większego znaczenia nabiera dostęp do efektywnych narzędzi wspomaganie analizy danych. Głównym celem tego opracowania jest zaprezentowanie wybranych nowych narzędzi wspomagających prowadzenie analizy danych w programie *STATISTICA 8*. Przy okazji przedstawiono zastosowanie różnych technik statystycznych przy rozwiązywaniu konkretnych zagadnień badawczych na przykładowych danych, pochodzących z badań medycznych.

Przykład analizy danych medycznych z wykorzystaniem nowych możliwości STATISTICA 8

Dane wykorzystywane w opisywanym przykładzie zostały oryginalnie zapisane w arkuszu programu Excel. W programie *STATISTICA 8* można taki arkusz otworzyć bezpośrednio bez konieczności wcześniejszego importowania. Na otwartym w ten sposób arkuszu można następnie wykonywać analizy i tworzyć wykresy dokładnie tak samo jak w środowisku *STATISTICA*. W takiej sytuacji przy definiowaniu pierwszej analizy program wyświetli pokazane poniżej okno (rys. 1), w którym należy określić, jaki obszar arkusza zawiera dane oraz z którego wiersza i której kolumny mają być pobrane nazwy zmiennych i przypadków.

Po podaniu tych informacji program wyświetla jeszcze jedno okno, w którym można przeglądać lub modyfikować typ kolumny z danymi. Po zamknięciu tego okna program pozwala zdefiniować analizę.



Rys. 1. Okno definiowania zakresu danych do analizy, wiersza z nazwami zmiennych i kolumny z nazwami przypadków.

W nowej wersji programu dodano kilka użytecznych statystyk, np. % ważnych obserwacji, współczynnik zmienności, odporne miary tendencji centralnej: średnią przyciętą i średnią Winsora oraz test Grubbsa. Poniżej zamieszczono arkusz z wynikami wybranych statystyk dla kilku zmiennych z pliku danych.

Zmienna	Statystyki opisowe (NadciśnienieWst (E5:AK37))											
	N ważnych	% Waznych	Średnia	Średnia przycięta 5,00 %	Średnia Winsora 5,00 %	Statystyka Grubbsa	p	Mediana	Minimum	Maksimum	Odch.std	Wsp.zmn.
BMI	28	84,85	28,129	28,1254	28,0353	1,858603	1,00000	27,8964	18,6214	37,7370	5,1693	18,37702
Ciśnienie skurczowe	33	100,00	143,697	142,7586	143,2121	2,151475	1,00000	141,0000	118,0000	183,0000	18,2679	12,71283
Cholesterol całkowity	30	90,91	216,800	213,8077	213,4333	3,315422	0,01286	209,5000	140,0000	375,0000	47,7164	22,00941
HDL	28	84,85	40,536	40,0385	40,4286	2,654520	0,25434	37,5000	20,0000	74,0000	12,6065	31,09981
LDL	29	87,88	130,186	129,2593	129,9655	2,417538	0,64042	131,2000	79,8000	205,6000	31,1945	23,96141
Triglicerydy	30	90,91	228,033	200,6538	220,9667	2,902362	0,10275	166,0000	46,0000	806,0000	199,1367	87,32788

Rys. 2. Wyniki obliczeń podstawowych statystyk opisowych.

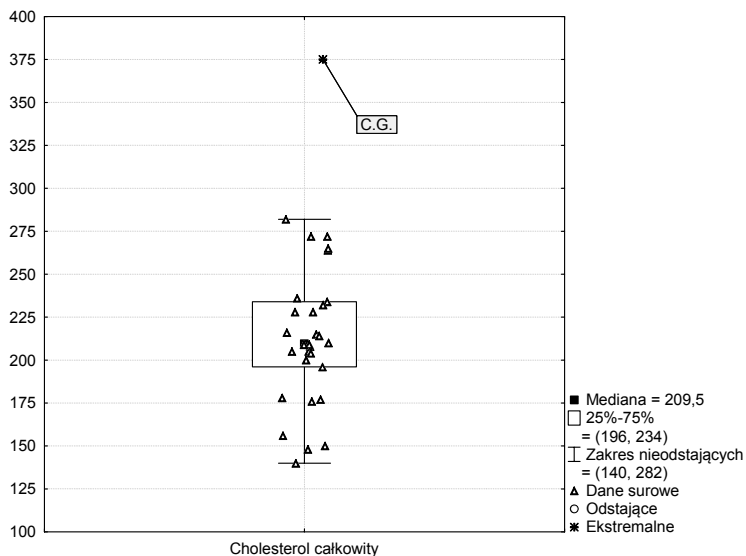
Wyniki testu Grubbsa wskazują na występowanie przynajmniej jednej odstającej obserwacji. Obserwację taką można zidentyfikować na wykresie typu ramka-wąsy z obserwacjami odstającymi i ekstremalnymi. Poniżej (na rys. 3) zamieszczono taki wykres.

Odstającą obserwacją jest wynik pomiaru stężenia cholesterolu całkowitego dla pacjenta o inicjałach C.G. Dodatkowo na wykresie zostały również pokazane surowe dane.

Przy przeprowadzaniu analiz zazwyczaj bardziej efektywnym jest wcześniejsze zaimportowanie danych z pliku źródłowego do formatu arkusza *STATISTICA*. W dalszej części opisywanego przykładu tak właśnie zrobiono.

W nowej wersji programu rozszerzono także możliwości współpracy z programem Word. Wyniki analiz (tabele i wykresy) można kierować do dokumentu tego programu, np. w celu tworzenia raportu z wynikami prowadzonej analizy.

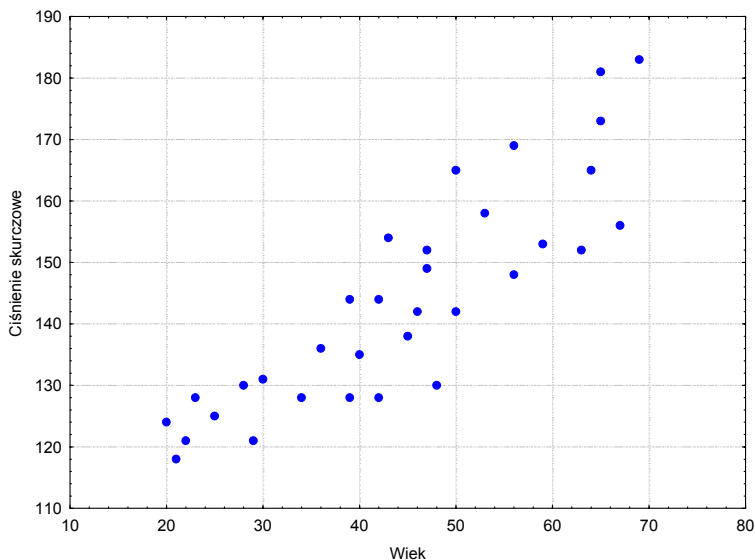
Kontynuując analizę danych, sprawdzimy, czy w badanej grupie pacjentów wiek jest powiązany z ciśnieniem skurczowym krwi (wśród lekarzy wiek jest uważany za jeden z czynników ryzyka wystąpienia choroby nadciśnieniowej). Jeśli okaże się, że tak jest, to wówczas zbudujemy model opisujący to powiązanie.



Rys. 3. Wykres ramka-wąsy z surowymi danymi.

Modelowanie zależności ciśnienia skurczowego od wieku

Dla wstępnej orientacji co do rodzaju i kierunku ewentualnego powiązania został utworzony dwuwymiarowy wykres rozrzutu dla wieku i ciśnienia skurczowego.



Rys. 4. Wykres rozrzutu dla ciśnienia skurczowego i wieku.



Zamieszczony powyżej wykres pozwala zaobserwować stopniowy wzrost przeciętnego ciśnienia skurczowego wraz z wiekiem. 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 zawiera poniższa tabela.

Wyniki analizy regresji dla zmiennej zależnej: Ciśnienie skurczowe						
R=0,8795 R ² =0,7735 Skoryg. R ² =0,7661						
F(1,31)=105,84 p<0,00001 Błąd std. estymacji: 8,8341						
N=33	BETA	Bł. std. BETA	B	Bł. std. B	t(31)	poziom p
W. wolny			94,85560	4,990376	19,00771	0,00000
Wiek	0,879463	0,085486	1,10168	0,107087	10,28776	0,00000

Rys. 5. Wyniki analizy regresji.

Zawarte w tabeli wyniki modelowania umożliwiają statystyczną i merytoryczną ocenę zbudowanego modelu (Quinn i Keough 2002, Sobczyk 2007). 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 wieku o 1 rok podnosi przeciętne ciśnienie skurczowe o 1,1 jednostki. 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 0,7735 i oznacza, że zbudowany model tłumaczy ponad 77 % oryginalnej wariancji zmiennej zależnej. Z tego wynika, że około 23 % wariancji ma charakter losowy lub może zostać wyjaśnione wpływem innych nieuwzględnionych w modelu zmiennych niezależnych.

Warto przypomnieć, że 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 (Afifi i Clark 1996).

W trakcie prowadzenia analizy mogą zdarzyć się sytuacje, w których badacz czy analityk chciałby przerwać analizę i „zamrozić” jej stan, tak aby mógł w dowolnym momencie powrócić do przerwanej analizy w tym samym miejscu. Przykładowo mogą się pojawić wątpliwości co do poprawności zebranych danych lub dobranej metody analizy. W nowej wersji programu *STATISTICA* dodano bardzo użyteczną możliwość zapisywania stanu przeprowadzanych analiz w pliku projektu analizy. Do projektu mogą zostać dołączone arkusze danych, wykresy, skoroszyty, makra, raporty, rozpoczęte analizy oraz wyniki analiz. Po ponownym otwarciu w programie zapisanego projektu można kontynuować analizy w tym samym miejscu, w którym zostały przerwane.

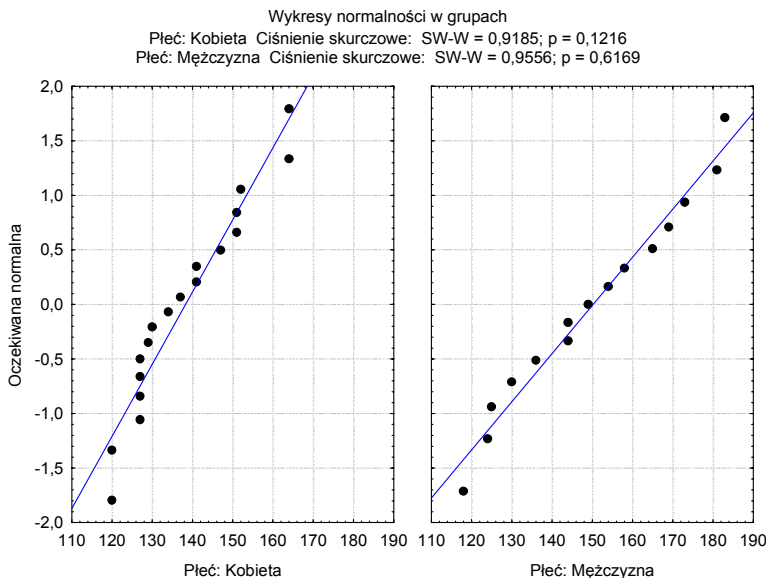
Ocena zróżnicowania poziomu ciśnienia skurczowego krwi pomiędzy kobietami i mężczyznami

Kontynuując przykład analizy, w kolejnym kroku sprawdzimy, czy występuje zróżnicowanie przeciętnego poziomu ciśnienia skurczowego krwi pomiędzy badanymi kobietami i mężczyznami (zgodnie z funkcjonującą hipotezą o podwyższonym poziomie ciśnienia skurczowego krwi u mężczyzn). W tym celu przeprowadzimy analizę porównawczą podstawowych miar położenia i zmienności empirycznego rozkładu zmiennej *Ciśnienie skurczowe* w obu grupach badanych pacjentów (w programie *STATISTICA* służy do tego procedura *Przekroje, prosta ANOVA* dostępna w module *Statystyki podstawowe i tabele*). Przy okazji takie same obliczenia wykonamy dla zmiennej *Wiek*, aby upewnić się, czy nie należy szukać w zróżnicowaniu przeciętnego wieku obu grup ewentualnej przyczyny różnic. Wyniki zawiera poniższa tabela.

	Statystyki opisowe w grupach (N=33)					
	Ciśnienie skurczowe Średnie	Ciśnienie skurczowe N	Ciśnienie skurczowe Odch.std	Wiek Średnie	Wiek N	Wiek Odch.std
Kobieta	138,2778	18	13,88527	44,66667	18	13,73703
Mężczyzna	150,2000	15	21,10924	43,93333	15	16,01993
Ogół grp.	143,6970	33	18,26794	44,33333	33	14,58310

Rys. 6. Wyniki analizy przekrojowej.

Przed przystąpieniem do interpretacji otrzymanych wyników należy przekonać się, czy średnie arytmetyczne dobrze oddają tendencję centralną porównywanych rozkładów. W tym celu można utworzyć skategoryzowane wykresy normalności oraz przeprowadzić test normalności Shapiro-Wilka. Wyniki dla zmiennej *Ciśnienie skurczowe* zostały zaprezentowane na wykresie.



Rys. 7. Wykresy normalności z wynikami testu Shapiro-Wilka.

Na podstawie uzyskanych wyników możemy stwierdzić, że obliczone wartości średnich arytmetycznych prawidłowo odzwierciedlają przeciętny poziom ciśnienia skurczowego krwi w porównywanych grupach pacjentów. Dla zmiennej *Wiek* sytuacja wygląda podobnie. A zatem przeciętny poziom skurczowego ciśnienia krwi jest wyższy o około 12 jednostek w grupie badanych mężczyzn.

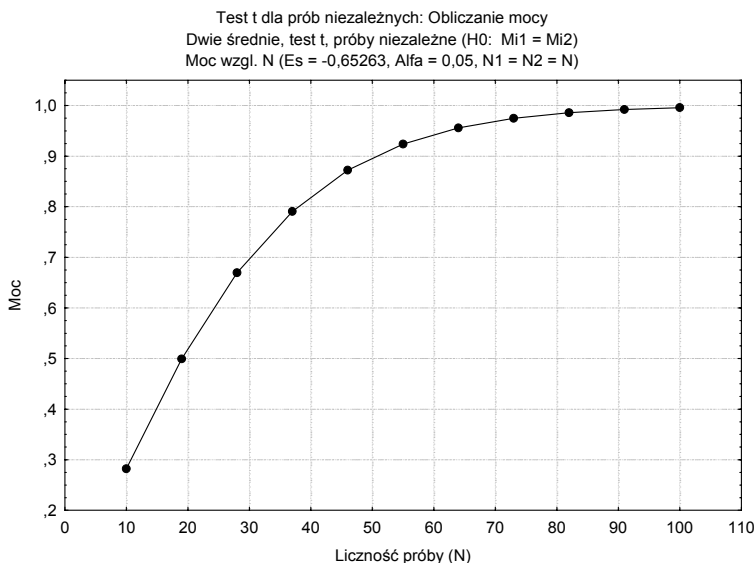
W następnym fragmencie analizy założymy, że badana zbiorowość pacjentów stanowi reprezentatywną próbę dla określonej populacji pacjentów, dla której badacz chce zweryfikować prawdziwość interesujących go hipotez badawczych. W takiej sytuacji jest on zainteresowany przeprowadzeniem odpowiedniego testu statystycznego, który umożliwi mu ocenę stopnia ich wiarygodności. W opisywanym przykładzie badacz jest zainteresowany oceną statystycznej istotności zróżnicowania wartości oczekiwanych dla zmiennej ilościowej (*Ciśnienie skurczowe*) pomiędzy dwoma niezależnymi populacjami, więc powinien do tego celu zastosować albo parametryczny test *t Studenta* (pod warunkiem, że wyniki w obu porównywanych próbach pochodzą z populacji o rozkładach normalnych i ponadto nie różnią się pod względem wariancji) albo test *Cochrana-Coxa* (w przypadku, gdy okaże się, że wyniki z obu porównywanych prób pochodzą z populacji o rozkładach normalnych, ale różnią się pod względem wariancji) albo też nieparametryczny test *Manna-Whitneya* (w przypadku, gdy wyniki przynajmniej jednej z grup nie pochodzą z populacji o rozkładzie normalnym).

Ponieważ przy wyborze odpowiedniej miary tendencji centralnej rozkładu w poprzednim fragmencie analizy sprawdzono, że rozkłady badanej zmiennej w przypadku obydwu porównywanych grup pochodzą z populacji o rozkładzie normalnym, więc przy ocenie statystycznej istotności zastosowano test *t Studenta* (przy przeprowadzaniu tego testu sprawdzono, że jest spełnione również założenie równości wariancji). Wyniki testowania przedstawiono poniżej w tabeli.

Zmienna	Testy t; Grupująca: Płeć		t	df	p	N ważnych	N ważnych
	Średnia Kobieta	Średnia Mężczyzna					
Ciśnienie skurczowe	138,2778	150,2000	-1,94642	31	0,060717	18	15

Rys. 8. Wyniki testu t Studenta dla prób niezależnych.

Zróżnicowanie okazało się statystycznie nieistotne. W takiej sytuacji zaleca się sprawdzenie, czy test ma wystarczającą moc, czyli odpowiednio wysoki poziom prawdopodobieństwa akceptacji prawdziwości hipotezy alternatywnej, jeśli rzeczywiście (w populacji) jest prawdziwa. Zazwyczaj wymaga się, aby moc testu wynosiła przynajmniej 0,8. Korzystając z modułu *Analiza mocy testu*, sprawdzono poziom mocy w opisywanych danych. Okazało się, że moc testu wyniosła 0,44. W literaturze poświęconej zagadnieniu mocy testu podawane są strategie, które pozwalają ją zwiększyć (Bausell i Li 2002). Jedną z nich jest zwiększenie liczebności próby. Moduł *Analiza mocy testu* pozwala również oszacować, jaka liczebność próby byłaby potrzebna do osiągnięcia odpowiedniego poziomu mocy. Ilustruje to poniższy wykres.



Rys. 9. Wykres mocy w funkcji liczebności próby.

Na podstawie wykresu możemy stwierdzić, że dla osiągnięcia mocy testu 0,8 potrzebna była liczebność prób nieco poniżej 40.

W opisywanym przykładzie wykorzystano wyniki badań, w których zwiększono liczebności badanych kobiet i mężczyzn odpowiednio do 32 i 27. W dalszej części analizy przeliczymy wyniki obliczeń dla nowego zbioru danych. Wykorzystamy do tego celu nową funkcjonalność programu *STATISTICA* w wersji 8, umożliwiającą ponowne wykonanie analizy na tym samym lub zmienionym zbiorze danych (pod warunkiem, że w nowym zbiorze danych zostanie zachowana kolejność zmiennych). Poniżej zamieszczono wyniki analizy porównawczej przeciętnego poziomu zmiennych *Ciśnienie skurczowe* i *Wiek* przeprowadzonej na powiększonym zbiorze danych.

	Statystyki opisowe w grupach (N=59)					
	Ciśnienie skurczowe Średnie	Ciśnienie skurczowe N	Ciśnienie skurczowe Odch. std	Wiek Średnie	Wiek N	Wiek Odch. std
Kobieta	136,9063	32	12,95491	43,50000	32	14,22175
Mężczyzna	151,0370	27	21,46102	44,14815	27	16,16669
Ogół grp.	143,3729	59	18,61673	43,79661	59	15,01239

Rys. 10. Wyniki analizy przekrojowej.

Wyniki analizy pokazują, że zróżnicowanie przeciętnego poziomu ciśnienia krwi pomiędzy badanymi kobietami i mężczyznami wzrosło do około 14 jednostek. Ponowna ocena statystycznej istotności różnic pokazuje, że tym razem różnica okazała się statystycznie istotna. Poniżej zamieszczono tabelę z wynikami testu *t Studenta*.



Zmienna	Średnia		t	df	p	N ważnyc Kobieta	N ważnych Mężczyzna	Odch.std Kobieta	Odch.std Mężczyzna
	Kobieta	Mężczyzna							
Ciśnienie skurczowe	136.9063	151.0370	-3.11496	57	0.002878	32	27	12.95491	21.46102

Rys. 11. Wyniki testu t Studenta.

Jednocześnie moc testu zwiększyła się do poziomu 0,82.

Modelowanie zależności ciśnienia skurczowego od wieku i płci

W dalszej części przykładu powrócimy do zagadnienia oceny wpływu wieku na poziom ciśnienia skurczowego krwi i zbudujemy model uwzględniający dodatkowo płeć badanych pacjentów. Zastosujemy technikę regresji liniowej wielorakiej, która umożliwi wprowadzenie do modelu również zmiennych typu jakościowego (Maddala 2006).

Najważniejsze wyniki analizy przedstawiono w tabeli poniżej.

Wyniki analizy regresji dla zmiennej zależnej: Ciśnienie skurczowe R=0,9384 R ² =0,88062 Skoryg. R ² =0,8764 F(2,56)=206,55 p<0,0001 Błąd std. estymacji: 6,5461						
N=59	BETA	Bł. std. BETA	B	Bł. std. B	t(56)	poziom p
W. wolny			77,20156	3,600834	21,43991	0,000000
Płeć	0,362794	0,046181	13,44147	1,711011	7,85586	0,000000
Wiek	0,857616	0,046181	1,06352	0,057269	18,57063	0,000000

Rys. 12. Wyniki analizy regresji.

Parametry strukturalne modelu istotnie różnią się od zera. Obliczone z próby oceny parametrów informują, że zwiększenie wieku o 1 rok podnosi przeciętne ciśnienie skurczowe o około 1,06 jednostki, przy ustaleniu wartości zmiennej *Płeć*. Z kolei jeśli badaną osobą jest mężczyzna, to przeciętny poziom ciśnienia skurczowego rośnie w stosunku do kobiet o około 13,44 jednostki (przy ustaleniu wartości zmiennej *Wiek*). W opisywanym przykładzie wartość współczynnika determinacji jest równa 0,8806 i oznacza, że zbudowany model tłumaczy nieco ponad 88 % oryginalnej wariancji zmiennej zależnej. Z tego wynika, że tylko około 12 % wariancji ma charakter losowy lub może zostać wyjaśnione wpływem innych nieuwzględnionych w modelu zmiennych niezależnych.

Modelując interesujące nas powiązania, mogliśmy postąpić jeszcze inaczej. W kolejnym fragmencie analizy zbudujemy dwa osobne modele dla każdej z badanych płci. Wyniki analizy zamieszczono w poniższych tabelach.

Płeć=Kobieta Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe R=0,8787 R ² =0,7721 Skoryg. R ² =,7645 F(1,30)=101,61 p<0,00001 Błąd std. estymacji: 6,2873						
N=32	BETA	Bł. std. BETA	B	Bł. std. B	t(30)	poziom p
W. wolny			102,0889	3,628421	28,13591	0,000000
Wiek	0,878668	0,087167	0,8004	0,079402	10,08030	0,000000

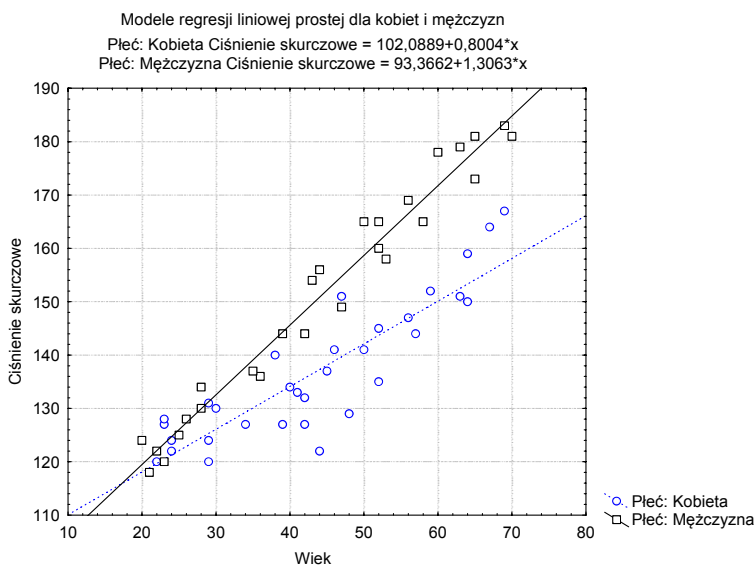
Rys. 13. Wyniki analizy regresji w grupie badanych kobiet.

Płeć=Męczyzna						
Podsumowanie regresji dla zmiennej zależnej: Ciśnienie skurczowe						
R=0,9840 R ² =0,9683 Skoryg. R ² =0,9671						
F(1,25)=764.66 p<0,0001 Błąd std. estymacji: 3.8942						
	BETA	Bł. std. BETA	B	Bł. std. B	t(25)	poziom p
N=27						
W. wolny			93,36624	2,216120	42,13050	0,000000
Wiek	0,984043	0,035586	1,30630	0,047240	27,65252	0,000000

Rys. 14. Wyniki analizy regresji w grupie badanych mężczyzn.

Orzymane wyniki pokazują, że w grupie badanych mężczyzn opisywana zależność jest znacznie mocniejsza. Wartość współczynnika determinacji jest bliska 0,97, co oznacza, że tylko około 3 % wariacji ma charakter losowy lub może zostać wyjaśnione wpływem innych nieuwzględnionych w tym modelu zmiennych niezależnych.

Różnice pomiędzy obydwooma modelami dobrze ilustruje zamieszczony poniżej wykres.



Rys. 15. Modele regresji w grupach.

Zamieszczony wykres ułatwia interpretację wyników przeprowadzonej analizy, w szczególności pokazuje różnice w tempie wzrostu ciśnienia skurczowego wraz z wiekiem w porównywanych grupach badanych pacjentów.

Przedstawione w artykule sposoby opracowania tych konkretnych danych mogą zostać z powodzeniem wykorzystane również w przypadku danych pochodzących z innych badań.

Podsumowanie i wnioski

Przedstawione w artykule metody analizy oraz ich wyniki miały przede wszystkim na celu zilustrowanie różnych technik opracowania danych, zarówno tych, które są stosowane przy



eksploracji danych i opisie statystycznym, jak również tych, które wchodzą w zakres wnioskowania statystycznego. Ponadto chodziło o zaprezentowanie wybranych nowych narzędzi wspomaganie analizy danych w programie *STATISTICA 8*.

Zaprezentowana analiza od strony merytorycznej dotyczyła co prawda zagadnień z zakresu medycyny, ale wydaje się, że może również zostać z powodzeniem wykorzystana w przypadku analizy zagadnień pochodzących z innych dziedzin badań empirycznych.

W charakterze podsumowania można przedstawić poniższe wnioski końcowe:

- ◆ przy analizie danych pochodzących z rzeczywistych badań powinno się zawsze przeprowadzić ich eksplorację pod kątem błędnych/nietypowych obserwacji, gdyż obserwacje takie wpływają zarówno na wartości określonych statystyk z próby, jak i oszacowania nieznanymi parametrów populacyjnych,
- ◆ przy stosowaniu testów statystycznych powinno się badać ich moc, aby ocenić, czy użyty test miał wystarczającą zdolność do odrzucenia sprawdzanej hipotezy (gdyby w populacji okazała się ona nieprawdziwa),
- ◆ w przypadku zagadnień, w których występują różnorakie powiązania pomiędzy badanymi zmiennymi, warto stosować różne techniki modelowania,
- ◆ nowa wersja programu *STATISTICA* pozwala znakomicie wspomagać stosowanie różnych technik statystycznej analizy danych.

Literatura

1. Afifi A. A., Clark V., *Computer-Aided Multivariate Analysis*, Third Ed., Chapman & Hall, 1996.
2. Bausell R. B., Li Y.-F., *Power Analysis for Experimental Research. A Practical Guide for the Biological, Medical and Social Sciences*, Cambridge University Press, 2002.
3. Quinn G. P., Keough M. J., *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, 2002.
4. Maddala G. S., *Ekonometria*, Wydawnictwo Naukowe PWN, 2006.
5. Sobczyk M., *Statystyka*, wyd. 5 uzupełnione, PWN, 2007.