

**KRZYSZTOF CHMIELEWSKI, STEFAN BERCZYŃSKI**

**STATYSTYKA MATEMATYCZNA**  
**ĆWICZENIA LABORATORYJNE Z WYKORZYSTANIEM PAKIETU**  
***STATISTICA PL***

Szczecin 2001

Recenzent  
LEON KUKIELKA

Opracowanie językowe

Projekt okładki

Wydano za zgodą rektora Politechniki Szczecińskiej

ISBN

Wydawnictwo Uczelniane Politechniki Szczecińskiej  
al. Piastów 50, 70-311 Szczecin, tel.(91) 449 47 60  
Wydanie I. Nakład 1000 + 39. Arkuszy wydawniczych . Arkuszy druku .....

Druk: Zapol, al. Piastów 42, 71-062 Szczecin, tel. (91) 434 10 21

## Spis treści

Wstęp .....	5
1. Statystyka opisowa. Estymacja parametrów zmiennej losowej .....	7
1.1. Cel ćwiczenia .....	7
1.2. Statystyka opisowa .....	7
1.3. Estymacja parametrów rozkładu zmiennej losowej .....	8
1.4. Obliczenia programem STATISTICA .....	9
2. Weryfikacja hipotez statystycznych .....	19
2.1. Cel ćwiczenia .....	19
2.2. Opis metody .....	19
2.2.1. Testy dla prób niezależnych .....	22
2.2.2. Test dla prób zależnych .....	23
2.3. Obliczenia programem STATISTICA .....	24
2.3.1. Testy dla prób niezależnych .....	24
2.3.2. Test t dla prób zależnych .....	28
3. Badanie zgodności rozkładu zmiennej losowej z rozkładem teoretycznym .....	31
3.1. Cel ćwiczenia .....	31
3.2. Opis metody .....	31
3.2.1. Porównanie kształtów histogramów .....	31
3.2.1.1. Porównanie histogramu częstości z funkcją gęstości .....	31
3.2.1.2. Porównanie histogramu skumulowanej częstości z dystrybuantą .....	32
3.2.2. Testy zgodności .....	33
3.2.2.1. Test chi-kwadrat ( $\chi^2$ ) .....	33
3.2.2.2. Test Kołmogorowa .....	34
3.2.2.3. Test normalności Lillieforsa .....	34
3.2.2.4. Test normalności Shapiro-Wilka .....	35
3.3. Obliczenia programem STATISTICA .....	35
3.3.1. Testy ogólne .....	35
3.3.2. Testy normalności .....	40
4. Regresja liniowa .....	41
4.1. Cel ćwiczenia .....	41
4.2. Wprowadzenie .....	41
4.3. Regresja jednej zmiennej niezależnej .....	44
4.3.1. Opis metody .....	44
4.3.2. Obliczenia programem STATISTICA .....	47
4.4. Regresja wielokrotna - wybór zmiennych .....	59
4.4.1. Opis metody .....	59
4.4.2. Przykład obliczeń programem STATISTICA .....	62

5. Regresja nieliniowa .....	65
5.1. Cel ćwiczenia .....	65
5.2. Wprowadzenie .....	65
5.3. Estymacja współczynników modeli nieliniowych .....	67
5.4. Obliczenia programem STATISTICA .....	74
Tablice statystyczne .....	85
Struktura i zasady działania pakietu Statistica PL .....	89
Literatura .....	93

## WSTĘP

Statystyka znajduje dzisiaj zastosowanie niemal we wszystkich dziedzinach nauki: w naukach technicznych, przyrodniczych, ekonomicznych i wielu innych. Z różnych metod statystycznych korzystają powszechnie naukowcy i praktycy. Umożliwiają one dokonywanie syntetycznego opisu zbieranych informacji, wyszukiwanie związków między zmiennymi oraz umożliwiają prognozowanie. Podstawowym źródłem informacji dla statystyka są dane liczbowe. Odpowiednio zebrane dane należy poddać obróbce statystycznej, aby wyciągnąć z nich jak najwięcej informacji. Dokonuje się tej analizy wykorzystując pakiety statystyczne. Jednym z nich jest pakiet STATISTICA o bardzo dużych możliwościach analitycznych i graficznych.

Niniejszy skrypt powstał w wyniku wieloletnich doświadczeń dydaktycznych w prowadzeniu przedmiotu „Statystyka matematyczna” na Wydziale Mechanicznym PS. Obejmuje on zakres ćwiczeń laboratoryjnych prowadzonych na studiach inżynierskich i magisterskich z przedmiotu „Statystyka matematyczna” oraz doktoranckich w ramach przedmiotu „Statystyka i analiza regresji”. Podzielony jest na pięć części. W pierwszej opisano charakterystyki zmiennej losowej, ocenę opisową rozkładu na podstawie histogramów oraz sposoby estymacji wartości oczekiwanej i wariancji zmiennej o rozkładzie normalnym. W drugiej omówiono zasady weryfikacji hipotez statystycznych oraz testy statystyczne stosowane przy porównywaniu wartości oczekiwanych dwóch zmiennych losowych. W trzeciej przedstawiono podstawowe testy zgodności oraz zasady ich wykorzystania. W czwartej opisano podstawy analizy regresji liniowej. Ostatnią część poświęcono metodom estymacji nieliniowych zależności regresyjnych.

Opis realizacji każdego ćwiczenia poprzedzono krótką charakterystyką podstawowych pojęć i opisem metody stosowanej w obliczeniach w celu prawidłowej interpretacji otrzymanych wyników. Aby poznać głębiej istotę matematycznego opisu zjawisk losowych, a szczególnie takich pojęć jak zmienna losowa, jej rozkład, estymacja parametrów rozkładu czy testowanie hipotez statystycznych, należy skorzystać z literatury uzupełniającej.

# 1. STATYSTYKA OPISOWA. ESTYMACJA PARAMETRÓW ZMIENNEJ LOSOWEJ

## 1.1. Cel ćwiczenia

Celem ćwiczenia jest zapoznanie się ze sposobami opisu istotnych cech zmiennej losowej na podstawie danych z próby, obliczenie podstawowych wielkości charakteryzujących te dane oraz estymacja wartości oczekiwanej i wariancji zmiennej losowej.

## 1.2. Statystyka opisowa

Analiza zebranych danych doświadczalnych powinna umożliwić określenie istotnych właściwości badanej zmiennej losowej na podstawie zebranych danych. Niech  $x_1, x_2, \dots, x_n$  będzie ciągiem  $n$  obserwacji zmiennej losowej. Do podstawowych charakterystyk opisujących zmienną należą:

– średnia arytmetyczna:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , (1.1)

– mediana: wielkość w uporządkowanym ciągu obserwacji, poniżej której leży 50% danych,

– wariancja:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ , (1.2)

– odchylenie standardowe:  $s = \sqrt{s^2}$ , (1.3)

– błąd standardowy średniej:  $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ , (1.4)

– rozstęp:  $R = x_{\max} - x_{\min}$ , (1.5)

– dolny kwartyl: wielkość w uporządkowanym ciągu obserwacji, poniżej której leży 25% danych,

– górny kwartyl: wielkość w uporządkowanym ciągu obserwacji, poniżej której leży 75% danych,

– współczynnik asymetrii (skośność):  $g_1 = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$ , (1.6)

– współczynnik skupienia (kurtoza):  $K = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n s^4} - 3$ . (1.7)

Średnia, mediana i kwartyle należą do grupy charakterystyk nazywanych miarami położenia. Określają one średni lub typowy poziom wartości zmiennej, czyli przedstawiają przeciętny poziom badanej cechy.

Rozstęp, wariancja i odchylenie standardowe charakteryzują zmienność (rozproszenie) badanej cechy i dlatego nazywane są miarami zmienności.

Współczynnik asymetrii (skośność) określa kierunek i siłę asymetrii rozkładu danych. Współczynnik asymetrii równy zero ( $g_1 = 0$ ) wskazuje na symetrię rozkładu zmiennej. Wartość dodatnia ( $g_1 > 0$ ) oznacza asymetrię prawostronną (rozkład ma dłuższy prawy "ogon"), natomiast wartość ujemna ( $g_1 < 0$ ) oznacza asymetrię lewostronną (rozkład ma dłuższy lewy "ogon").

Współczynnik skupienia (kurtoza) opisuje koncentrację wartości badanej cechy wokół średniej. Jeżeli kurtoza jest ujemna ( $K < 0$ ), to rozkład jest bardziej spłaszczony od rozkładu normalnego, a jeżeli kurtoza jest dodatnia ( $K > 0$ ) to rozkład jest bardziej wysmukły niż normalny. Im większa jest wartość kurtozy tym rozkład jest bardziej wysmukły, czyli występuje większa koncentracja cechy wokół wartości średniej.

Wygodnym sposobem przedstawienia danych obserwowanych pozwalających na wizualne poznanie rozkładu danych jest ich prezentacja graficzna. W tym celu, przy dostatecznie dużej liczbie danych ( $n > 30$ ), można je przedstawić w postaci tzw. szeregu rozdzielczego (tabeli liczebności). Tworzy się go, dzieląc przedział zmienności  $[x_{\min}, x_{\max}]$  na zadaną liczbę  $k$  klas o równej długości i obliczając liczbę danych  $n_i$  (liczebność) należących do kolejnych klas. Wybór liczby klas jest w zasadzie dowolny. Należy jednak pamiętać, że zbyt duża liczba klas (tym samym zbyt wąskie przedziały klasowe) nie daje przejrzystego obrazu, ujawniają się przypadkowe odchylenia. Zbyt mała liczba klas zaciera istotne szczegóły zawarte w danych. Można tutaj skorzystać z podawanego w piśmiennictwie [7] wzoru na liczbę klas:

$$k = 1 + 3,3 \log n . \quad (1.8)$$

Graficznym sposobem przedstawienia informacji zawartych w szeregu rozdzielczym jest histogram liczebności (częstości). Jest to wykres słupkowy, w którym wysokość słupka jest proporcjonalna do liczebności. Jeżeli liczby obserwacji w klasach zostaną podzielone przez całkowitą liczbę danych  $n$ , otrzyma się alternatywną formę interpretacji danych zwaną histogramem liczebności względnej. Często też korzysta się z szeregu rozdzielczego w postaci skumulowanej, w którym liczebność danej klasy zastępuje się liczbą obserwacji należących do danej klasy i wszystkich poprzedzających ją. Wówczas dane mogą być przedstawione w postaci histogramu liczebności skumulowanej lub histogramu skumulowanej liczebności względnej.

### 1.3. Estymacja parametrów rozkładu zmiennej losowej

Przyjmuje się założenie, że dane mają rozkład normalny  $N(\mu, \sigma)$  z nieznaną wartością oczekiwaną  $\mu$  i nieznaną wariancją  $\sigma^2$ . Najlepszym estymatorem wartości oczekiwanej (średniej)  $\mu$  jest średnia arytmetyczna z próby  $\bar{x}$ . Jest to estymator zgodny, nieobciążony i o najmniejszej wariancji. Najlepszym estymatorem wariancji  $\sigma^2$  zmiennej losowej jest wariancja z próby  $s^2$ , określona wzorem (1.2). Ten estymator jest też zgodny, nieobciążony i o minimalnej wariancji. Estymatorem odchylenia standardowego  $\sigma$  jest odchylenie standardowe z próby  $s$  (1.3). Ten estymator jest naturalnym estymatorem powstałym z wariancji z próby. Nie ma on jednak tak dobrych własności jak estymator  $s^2$  (np. nie jest estymatorem nieobciążonym), tym niemniej jest powszechnie stosowany. Podane estymatory pozwalają na otrzymanie ocen punktowych nieznanymi parametrów. Obok ocen punktowych innym sposobem estymacji parametrów jest estymacja przedziałowa, polegająca na podaniu przedziałów ufności dla nieznanymi parametrów. Pozwala ona na określenie dokładności uzyskanych ocen parametrów. Niech  $1 - \alpha$  oznacza wybrany poziom ufności. Dwustronny przedział ufności dla wartości oczekiwanej  $\mu$  ma postać:

$$\left( \bar{x} - t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}} \right), \quad (1.9)$$

gdzie  $t_{1-\alpha/2, n-1}$  jest kwantylem rzędu  $1-\alpha/2$  rozkładu t-Studenta z  $n-1$  stopniami swobody.

Przedział ten pokrywa nieznaną wartość oczekiwaną z prawdopodobieństwem równym przyjętemu poziomowi ufności  $1 - \alpha$ . Długość tego przedziału jest wprost proporcjonalna do odchylenia standardowego  $s$  oraz odwrotnie proporcjonalna do pierwiastka z liczby obserwacji  $n$ . Zależy ona również od wartości kwantylu  $t_{1-\alpha/2, n-1}$ , zależnego od przyjętego poziomu ufności  $1 - \alpha$  i liczby stopni swobody  $n - 1$ .

Dwustronny przedział ufności dla wariancji  $\sigma^2$  na poziomie ufności  $1 - \alpha$  ma postać:

$$\left( \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \right), \quad (1.10)$$

gdzie:

$\chi_{1-\alpha/2, n-1}^2$  – jest kwantylem rzędu  $1-\alpha/2$  rozkładu chi-kwadrat z liczbą stopni swobody  $n-1$ ,

$\chi_{\alpha/2, n-1}^2$  – jest kwantylem rzędu  $\alpha/2$  rozkładu chi-kwadrat z liczbą stopni swobody  $n-1$ .

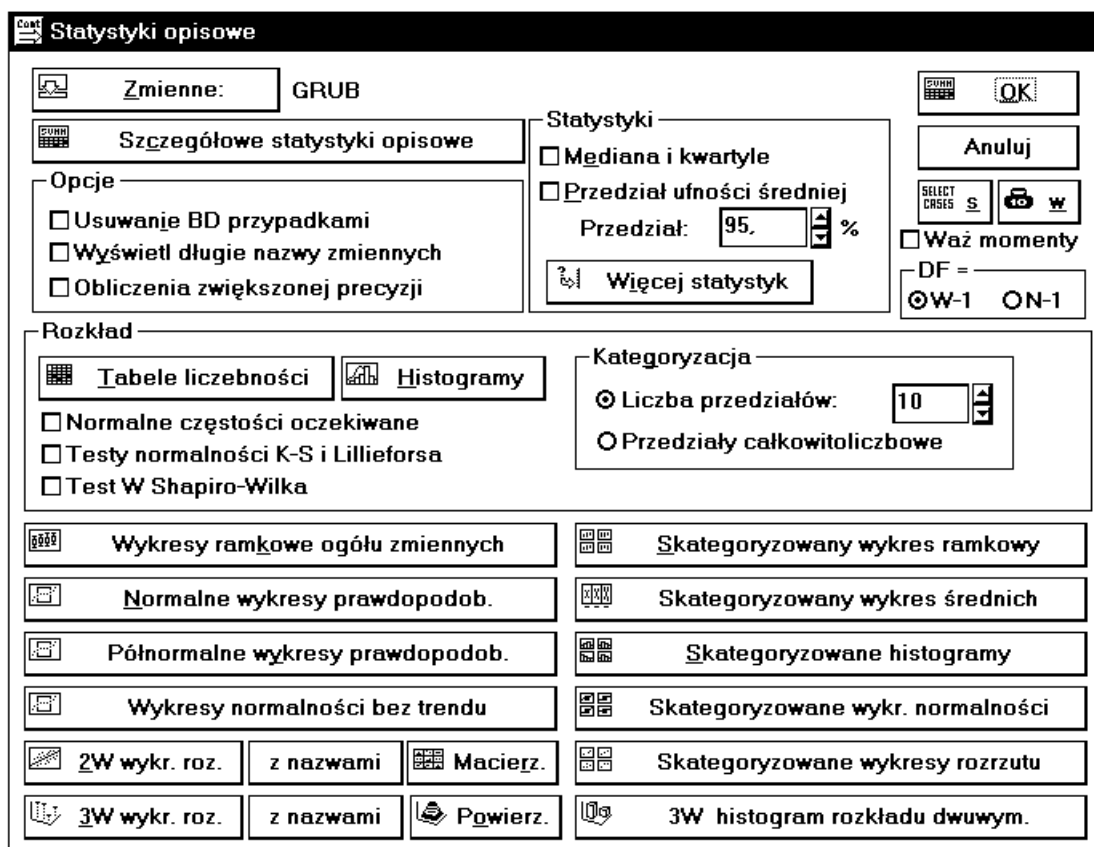
### 1.4. Obliczenia programem STATISTICA

W programie STATISTICA podstawowe wielkości charakteryzujące dane można obliczyć procedurą *Statystyki opisowe* w module *Podstawowe Statystyki*. Po wybraniu tej procedury otwiera się okno *Statystyki opisowe* przedstawione na rys.1.1.



Wyboru zmiennych do analizy dokonuje się po naciśnięciu przycisku **Zmienne**. Wybór odpowiednich parametrów opisowych możliwy jest w polu **Statystyki**. Domyślnie program oblicza średnią arytmetyczną, odchylenie standardowe, minimum i maksimum oraz podaje liczebność każdej wskazanej zmiennej. Można to rozszerzyć włączając opcje:

- **Mediana i kwartyle** – program oblicza dodatkowo medianę i kwartyle;
- **Przedział ufności średniej** – program oblicza dodatkowo przedział ufności dla wartości oczekiwanej (średniej)  $\mu$  dla wskazanego w oknie **Przedział** poziomu ufności (domyślnie 95%).



Rys. 1.1. Okno Statystyki opisowe

Pełny wybór parametrów opisowych otrzymuje się po naciśnięciu przycisku **Więcej statystyk**. Otwiera się wówczas okno, w którym można wskazać odpowiednie parametry opisowe. Wyboru obliczanej wielkości dokonuje się przez kliknięcie na jej nazwie lub zaznaczenie kwadratu obok. Można wybrać wszystkie charakterystyki, przy pomocy przycisku **Wszystkie**, lub wrócić do wyboru domyślnego naciskając przycisk **Domyślne**.

W polu **Rozkład** przyciskiem **Tabele liczebności** uruchamia się tworzenie tabel liczebności dla wskazanych zmiennych. Przycisk **Histogramy** pozwala na graficzne przedstawienie danych w postaci histogramu liczebności. Ponadto w tym polu są dodatkowe opcje:

- **Liczba przedziałów** – umożliwia określenie liczby klas, dla której mają być wykonane tabele liczebności i histogramy, w przypadku zmiennych ciągłych. Program przyjmuje

rzeczywistą liczbę klas w przybliżeniu równą zadeklarowanej, ponieważ przyjęto, że granice przedziałów klasowych powinny być zaokrąglone i zawierać po przecinku liczby 1, 2, 5;

– **Przedziały całkowitoliczbowe** – umożliwia tworzenie tabel liczebności i histogramów dla zmiennych dyskretnych lub interpretowanych jako dyskretne wyrażonych przez wartości całkowite. Wszystkie niecałkowite wartości zmiennej są pomijane;

– **Normalne częstości oczekiwane** – tworzy w tabeli liczebności dodatkowe kolumny zawierające wartości oczekiwane liczebności, skumulowane wartości oczekiwane liczebności oraz wartości oczekiwane liczebności względnej i wartości oczekiwane skumulowanej liczebności względnej obliczone przy założeniu, że zmienna ma rozkład normalny;

– **Testy normalności** – pozwalają na ocenę czy dana zmienna może być traktowana jako zmienna o rozkładzie normalnym. Zostanie to dokładnie opisane w ćwiczeniu "Badanie zgodności rozkładu zmiennej losowej z rozkładem teoretycznym".

W polu **Opcje** można dodatkowo określić:

– **Usuwanie BD przypadkami** – określa sposób postępowania z brakującymi danymi. Po wybraniu opcji **Usuwanie brakujących danych przypadkami** program pomija wszystkie przypadki (wiersze), w których brakuje wartości choćby dla jednej zmiennej. Jeżeli opcja jest wyłączona, to do analizy przyjmowane są wszystkie możliwe wartości dla poszczególnych zmiennych;

– **Wyświetl długie nazwy zmiennych** – oprócz krótkich nazw zmiennych zostaną wyświetlone ich długie nazwy o ile zostały zdefiniowane;

– **Obliczenia zwiększonej precyzji** – program wykonuje obliczenia o zwiększonej precyzji. Zaleca się wybór tej opcji dla zmiennych o bardzo małej względnej wariancji (wariancja podzielona przez średnią).

Dolną część okna **Statystyki opisowe** zajmują przyciski umożliwiające graficzną prezentację danych:

**Wykresy ramkowe ogółu zmiennych** – wykonywane są wykresy ramkowe (skrzynki z wąsami) dla wskazanych zmiennych. Można utworzyć cztery typy wykresów ramkowych w zależności od wyboru jednej z następujących opcji:

– **Mediana/Kwartyle/Rozstęp** – punkt centralny – mediana, ramka – kwartyle, wąsy – rozstęp;

– **Średnia/Bł.std./Odch.std** – punkt centralny – średnia, ramka – błąd standardowy, wąsy – odchylenie standardowe;

– **Średnia/Odch.std./1.96\*Odch.std.** – punkt centralny – średnia, ramka – odchylenie standardowe, wąsy – 1.96 \* odchylenie standardowe;

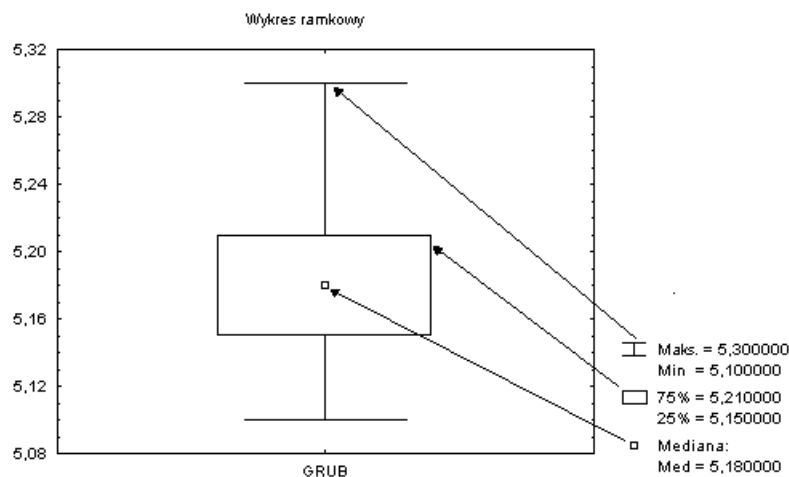
– **Średnia/ Bł.std./1.96\*Bł.std.** – punkt centralny – średnia, ramka – błąd standardowe, wąsy – 1.96 \* błąd standardowe.

Przykładowy wykres ramkowy dla opcji Mediana/Kwartyle/Rozstęp pokazano na rys. 1.2.

**Normalne wykresy prawdopodob.** – wykonywany jest normalny wykres prawdopodobieństwa.

**Półnormalne wykresy prawdopodob.** – wykonywany jest półnormalny wykres prawdopodobieństwa.

**Wykresy normalności bez trendu** – wykonywany jest normalny wykres prawdopodobieństwa po usunięciu trendu liniowego.



Rys. 1.2. Przykładowy wykres ramkowy

**Skategoryzowany wykres ramkowy** – wykonywane są wykresy ramkowe dla skategoryzowanych zmiennych. Można wybrać do trzech zmiennych określających kategorie podziału.

**Skategoryzowany wykres średnich** – wykonywane są wykresy średnich dla skategoryzowanych zmiennych.

**Skategoryzowane Histogramy** – wykonywane są histogramy dla skategoryzowanych zmiennych.

**Skategoryzowane wykry. normalności** – wykonywane są normalne wykresy prawdopodobieństwa dla skategoryzowanych zmiennych.

**Skategoryzowane wykresy rozrzutu** – wykonywane są wykresy rozrzutu wraz z linią regresji dla skategoryzowanych par zmiennych.

**3W histogram rozkładu dwuwym.** – wykonywane są trójwymiarowe histogramy dla wskazanych par zmiennych.

**2W wykry. roz.** – wykonywane są wykresy rozrzutu wraz z linią regresji dla wybranych par zmiennych.

**z nazwami** – wykonywane są wykresy rozrzutu dla wskazanych par zmiennych. Punkty na wykresie oznaczane są za pomocą nazw lub numerów przypadków.

**3W wykry. roz.** – wykonywane są wykresy rozrzutu dla wybranych trzech zmiennych. Jeśli użyje się opcji **z nazwami**, punkty na wykresie będą oznaczone za pomocą nazw lub numerów przypadków.

**Powierz** – umożliwia wykonanie trójwymiarowego wykresu rozrzutu wraz z wykresem powierzchni dla wskazanych trzech zmiennych.

**Macierz** – umożliwia wykonanie macierzowych wykresów rozrzutu dla wybranych zmiennych.

*Przykład 1.1.*

Zmierzono grubość  $n = 60$  podkładek wykonanych na automacie tokarskim. Obliczyć podstawowe wielkości charakteryzujące dane i przedstawić je w postaci graficznej. Dane zapisano w pliku PODKŁADKI. grub

Tabela 1.1  
Wartości podstawowych charakterystyk

Statystyki opisowe (podkładki. sta)	Zmienna GRUB
Nważnych	60
Średnia	5,187500
P. ufn-95,000%	5,175971
P. ufn. +95,000%	5,199029
Mediana	5,180000
Suma	311,2500
Minimum	5,100000
Maksimum	5,300000
Dolny Kwartył	5,160000
Górny Kwartył	5,210000
Rozstęp Kwartył	,200000
Rozstęp	,050000
Warianc.	,001992
Odch. Std	,044631
Błąd standard	,0057620
Skośność	,369864
Bł. std. Skośność	,308694
Kurtoza	,291358
Bł. std. Kurtoza	,608492

*Rozwiązanie*

Po naciśnięciu przycisku **Zmienne** i wskazaniu zmiennej do obliczeń, w polu **Statystyki** naciska się przycisk **Więcej statystyk**, aby obliczyć (oprócz domyślnych) wszystkie możliwe charakterystyki zmiennej. Obliczenie wskazanych statystyk uruchamia się naciskając przycisk **Szczegółowe statystyki opisowe** lub przycisk **OK**. Otrzymane wyniki przedstawiono w tabeli 1.1. W celu wykonania tabeli liczebności i histogramu przyjęto liczbę klas  $k = 7$ , obliczoną z wzoru (1.8). Wyniki obliczeń przedstawiono w tabeli 1.2 i na rys. 1.3. Ocena punktowa wartości oczekiwanej grubości podkładek wynosi  $\bar{x} = 5,1875$  a wariancji  $s^2 = 0,001992$ . Ocena przedziałowa wartości oczekiwanej, 95% przedział ufności wynosi od 5,1760 do 5,199, i oznacza, że ten przedział liczbowy z prawdopodobieństwem 95% pokrywa nieznaną wartość przeciętną grubości

podkładek.

Tabela 1.2

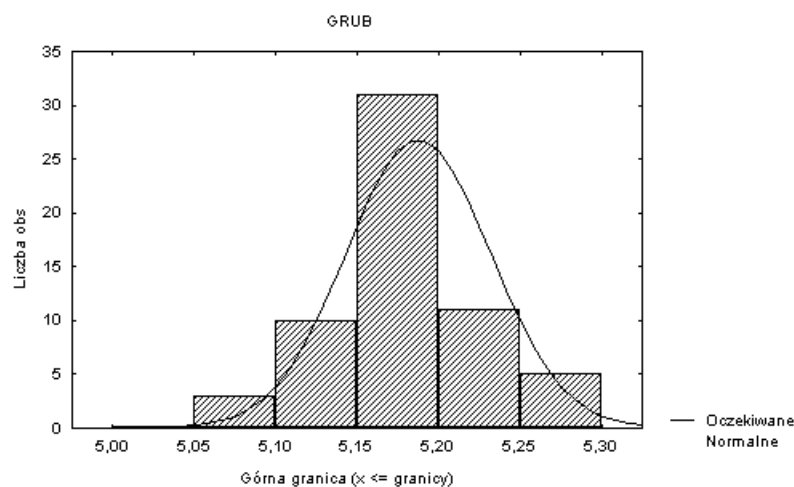
Tabela liczebności

Kategoria	GRUB (podkła~1.sta)					
	Liczność	Skumul Liczność	Procent ważnych	% Skumul ważnych	% ogółu przypad.	% skum ogółu
5,0000 < x <= 5,0500	0	0	0,00000	0,0000	0,00000	0,0000
5,0500 < x <= 5,1000	3	3	5,00000	5,0000	5,00000	5,0000
5,1000 < x <= 5,1500	10	13	16,66667	21,6667	16,66667	21,6667
5,1500 < x <= 5,2000	31	44	51,66667	73,3333	51,66667	73,3333
5,2000 < x <= 5,2500	11	55	18,33333	91,6667	18,33333	91,6667
5,2500 < x <= 5,3000	5	60	8,33333	100,0000	8,33333	100,0000
BD	0	60	0,00000		0,00000	100,0000

Na podstawie obliczonych wartości charakterystyk i histogramu można stwierdzić, że rozkład danych w próbce jest jednomodalny. Wartość mediany jest mniejsza od wartości

średniej oraz skośność jest dodatnia, czyli rozkład jest lekko niesymetryczny o prawostronnej asymetrii.

Kurtoza jest dodatnia, a więc rozkład jest bardziej wysmukły od normalnego.



Rys. 1.3. Histogram liczebności

W procedurze *Statystyki opisowe* można uzyskać tylko tabele liczebności i histogramy wykonane dla przybliżonej liczby klas, wynikającej z przyjętego sposobu wyznaczania granic klas, co może zniekształcać obraz wykresu. Jeśli chcemy otrzymać tabele liczebności i histogramy liczebności z podziałem na dokładną liczbę klas, należy skorzystać z procedury *Tabele liczebności* w module *Podstawowe statystyki*. Okno **Tabele liczebności** przedstawiono na rys. 1.4.

Na górze okna znajduje się przycisk **Zmienne** pozwalający wskazać zmienne do analizy. Poniżej znajdują się trzy przyciski pozwalające na uruchomienie podstawowych obliczeń.

**Tabele liczebności** – uruchamia procedurę tworzenia tabel liczebności według wskazanych w polu **Pokaż opcje** możliwości.

**Histogramy** – uruchamia procedurę tworzenia histogramów, których postać zależy od opcji wybranych w polu **Metoda kategoryzacji dla tabel i wykresów**.

**Statystyki opisowe** – umożliwia obliczenie podstawowych statystyk (średnia, odchylenie standardowe, wartość minimalna i maksymalna itd.).

To, co chce się otrzymać w tabeli liczebności, określa się w polu **Pokaż opcje**.

Poszczególne opcje w tym polu umożliwiają:

- **Liczebności skumulowane** – obliczenie skumulowanych liczebności;
- **Procenty (częstości)** – obliczenie skumulowanych liczebności względnych (częstości);
- **100% minus procenty skumulowane** – obliczenie różnicy między 100% i wartościami skumulowanych liczebności względnych (częstości);

– **Proporcje po transformacji logit** – obliczenie logitowej transformacji skumulowanej liczebności względnej dla każdej kategorii (klasy);

– **Proporcje po transformacji probit** – obliczenie probitowej transformacji skumulowanej liczebności względnej dla każdej kategorii (klasy).

**Tabele liczebności**

Zmienne: GRUB

Tabele liczebności | Histogramy | Statystyki opisowe

Metoda kategoryzacji dla tabel i wykresów

Wszystkie różne wartości  z wart. tekstowymi

Dokładna liczba przedziałów: 10

Przybliżona liczba okrągłych przedz.: 10

Wielkość kroku: 1.

rozpocznij od: 0. lub  od minimum

Kategorie całkowite  z wart. tekstowymi

Kody użytkownika

Kryteria użytkownika

Usuwanie BD przypadkami

Pokaż opcje

Liczebności skumulowane

Procenty (częstości)

Procenty skumulowane

100% minus procenty skumulowane

Proporcje po transformacji logit

Proporcje po transformacji probit

Normalne liczebności oczekiwane

Wliczaj braki danych (BD)  Waż momenty

Braki danych

BD i przypadki pominięte

Testy normalności

test K-S, znana średnia i odch. std.

test Lillieforsa, nieznana średnia/od.std.

test W Shapiro-Wilka

Do dopasowywania innych rozkładów używamy modułu Testy Nieparametryczne i Rozkłady lub wykresów (P-P lub K-K), do dopasowania danych uciętych używamy analiz przeżycia.

Wykres ramkowy ogółu zmiennych (1)

Normalne wykresy prawdopodob. (2)

Półnormalne wykresy prawdopodob. (3)

Wykresy normalności bez trendu (4)

3W histogram rozkładu dwuwym. (5)

OK Anuluj

SELECT CASES s w

Rys. 1.4. Okno Tabele liczebności

Sposób tworzenia tabel liczebności i histogramów określają opcje w polu **Metoda Kategoryzacji dla tabel i wykresów**. Ich znaczenie jest następujące:

– **Wszystkie różne wartości** – tabele liczebności i histogramy tworzone są dla liczby klas równych liczbie danych. Innymi słowy każda wartość zmiennej tworzy nową klasę. Ten sposób kategoryzacji wybierany jest domyślnie;

– **Z wart. tekstowymi** – tabele liczebności i histogramy tworzone są dla każdej wartości tekstowej zmiennej. Czyli każda nowa wartość tekstowa zmiennej tworzy nową klasę;

– **Dokładna liczba przedziałów** – tabele liczebności i histogramy tworzone są dla dokładnie takiej liczby klas, jaka będzie podana w sąsiednim polu;

– **Przybliżona liczba okrągłych przedz.** – tabele liczebności i histogramy tworzone są dla liczby klas w przybliżeniu równej liczbie podanej w sąsiednim polu;

– **Wielkość kroku** – tabele liczebności i histogramy tworzone są dla szerokości przedziałów klasowych podanej w sąsiednim polu;

– **Rozpocznij od** – określa wartość początkową w tworzonej tabeli liczebności. Jeżeli wybierze się opcję **Rozpocznij od minimum** wówczas wartością początkową w tabeli liczebności będzie najmniejsza wartość zmiennej;

– **Kategorie całkowite** – tabele liczebności i histogramy tworzone są dla liczby klas odpowiadającej liczbie wartości całkowitych. Wszystkie liczby niecałkowite są ignorowane;

– **Kody użytkownika** – tabele liczebności i histogramy tworzone są na podstawie liczb całkowitych wskazanych przez użytkownika. Wyboru dokonuje się w oknie, które otwiera się po naciśnięciu przycisku obok. Wszystkie liczby niecałkowite i nie wybrane przez użytkownika są ignorowane;

– **Kryteria użytkownika** – tabele liczebności i histogramy tworzone są w oparciu o kategorie określone przez użytkownika. Zasady selekcji przypadków definiuje się w oknie otwierającym się po naciśnięciu sąsiedniego przycisku. Można wybrać do 16 logicznych warunków selekcji przypadków określających 16 kategorii (klas) w tabeli liczebności.

Sposób postępowania z brakującymi danymi określają trzy opcje:

– **Usuwanie BD przypadkami** – po włączeniu tej opcji, z obliczeń wyłączone są wszystkie przypadki, w których brakuje danych dla którejkolwiek z wybranych zmiennych;

– **Wliczaj braki danych** w polu **braki danych** – po wybraniu tej opcji w tabeli liczebności podawany jest dodatkowy wiersz zawierający informacje o liczbie brakujących przypadków. Liczebności względne i skumulowane liczebności względne liczone są z uwzględnieniem brakujących danych;

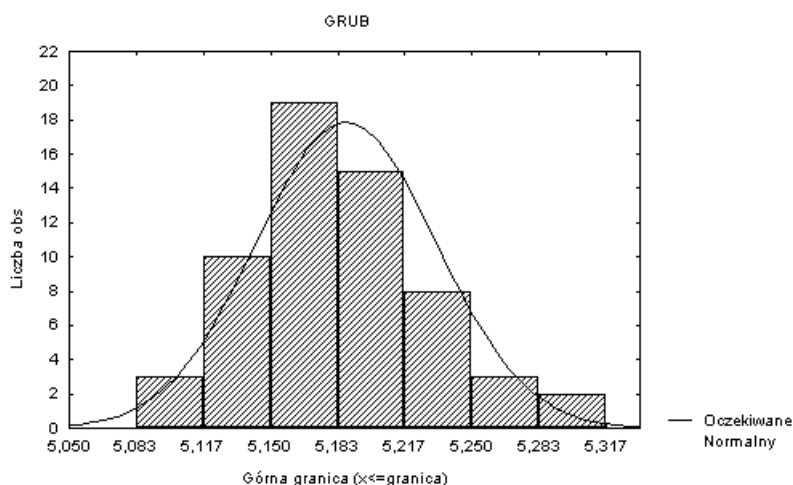
– **Wliczaj BD i przypadki pominięte** w polu **braki danych** – po włączeniu tej opcji w tabeli liczebności podawany jest dodatkowy wiersz zawierający informacje o liczbie brakujących i pominiętych przypadkach. Liczebności względne i skumulowane liczebności względne liczone są z uwzględnieniem brakujących danych.

Znaczenie pozostałych opcji i przycisków jest identyczne jak w procedurze *Statystyki opisowe*. Procedura *Tabele liczebności* pozwala na głębszą analizę danych, stwarzając możliwość ustalenia wielu parametrów mających wpływ na obraz histogramu – liczby klas, rozpiętości klas, dolnej granicy pierwszego przedziału i granic klas. Wymaga to zawsze zastanowienia się jak wyglądają dane, i w jaki sposób je przedstawić. Przy różnych ustawieniach otrzyma się różne postacie histogramów.

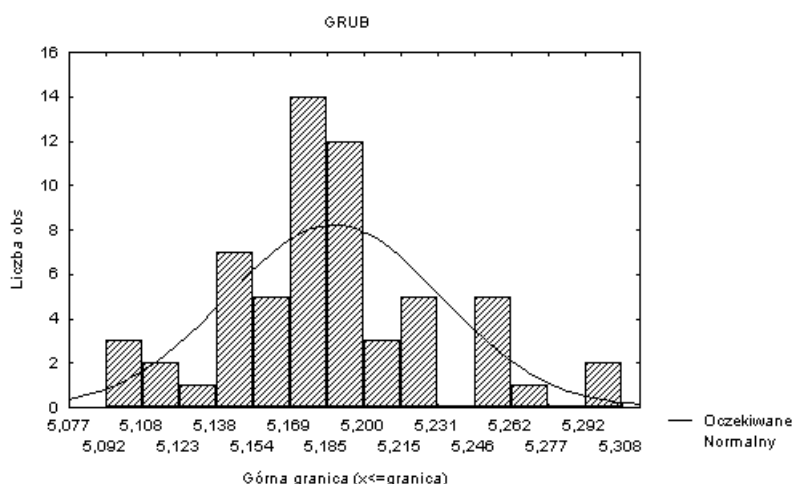
Poniżej przedstawiono dla danych z przykładu 1.1 dwa histogramy, dla liczby klas  $k = 7$  i dwukrotnie większej  $k = 14$ , otrzymanych w procedurze *Tabele liczebności* po włączeniu opcji **Dokładna liczba przedziałów**.

Histogram przedstawiony na rys. 1.5, wykonany dla liczby klas  $k = 7$  różni się od histogramu przedstawionego na rys. 1.3, na którym rzeczywista liczba klas jest równa  $k = 5$ . Na rys. 1.5 lepiej są widoczne istotne cechy badanej zmiennej, przede wszystkim wyraźniej zaznaczona jest asymetria danych i dokładniej określone rozłożenie wartości najmniejszych i największych.

Na rys. 1.6 przedstawiono histogram o dwukrotnie większej liczbie klas niż jest zalecana. Obraz rozkładu danych jest w tym przypadku zniekształcony przez zbyt silne zaznaczenie lokalnych zmian oraz przez pojawiające się klasy puste.



Rys. 1.5. Histogram dla liczby klas  $k = 7$  (dokładnej)



Rys. 1.6. Histogram dla liczby klas  $k = 14$  (dokładnej)

### Przykład 1.2

W celu wyznaczenia wartości przeciętnej długości drogi hamowania samochodu na asfalcie przeprowadzono przy prędkości 40 km/godz. 10. prób i otrzymano następujące wyniki: 22,0; 19,0; 18,5; 20,0; 20,5; 21,0; 19,5; 19,0; 20,5; 20,0 m. Przyjmując, że długość drogi hamowania ma rozkład normalny obliczyć:

- oszacowanie wartości przeciętnej długości drogi hamowania,
- nieobciążoną ocenę wariancji długości drogi hamowania,
- 99% przedział ufności dla wartości oczekiwanej długości drogi hamowania.



Wyniki obliczeń przedstawiono w tabeli 1.3.

Tabela 1.3

Wyniki oszacowania parametrów zmiennej losowej – długość drogi hamowania

STAT. PODST. STATYST.	Statystyki opisowe (predkość sta)								
	Zmienna	Nważnych	Średnia	P.ufn. 99,00%	-P.ufn. +99,00%	Minimum	Maksimum	Warianc.	Odch.Std.
PREDK.	10		20,0000	18,91672	21,08328	18,5000	22,0000	1,11111	1,05409

Ocena punktowa wartości oczekiwanej wynosi  $\bar{x} = 20,000$ , natomiast z oceny przedziałowej wynika, że 99% przedział ufności wynosi od 18,91672 do 21,08328. Nieobciążona ocena wariancji równa jest  $s^2 = 1,11111$ .

## 2. WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

### 2.1. Cel ćwiczenia

Celem ćwiczenia jest zapoznanie się z zasadami weryfikacji hipotez statystycznych.

### 2.2. Opis metody

Każde badanie rozpoczyna się od postawienia hipotezy, czyli przypuszczenia o możliwym rozwiązaniu, na podstawie wiedzy badającego. Przez hipotezę statystyczną rozumie się każde przypuszczenie odnośnie rozkładu zmiennej losowej. Tradycyjnie hipotezy statystyczne dzieli się na:

- hipotezy parametryczne, gdy dotyczą wartości parametrów statystycznych rozkładu zmiennej losowej,
- hipotezy nieparametryczne, gdy dotyczą postaci rozkładu zmiennej losowej.

Procedurę weryfikacji hipotezy statystycznej nazywa się testem statystycznym. Gdy weryfikuje się hipotezę parametryczną, nazywa się to testem parametrycznym. Procedura weryfikacji hipotezy nieparametrycznej nazywana jest testem nieparametrycznym. Hipoteza, która podlega sprawdzeniu, nazywana jest hipotezą zerową i oznaczana jest przez  $H_0$ . Oprócz postawienia hipotezy zerowej test statystyczny wymaga sformułowania hipotezy alternatywnej, to jest takiej hipotezy, którą jesteśmy skłonni przyjąć, gdyby okazało się, że hipotezę zerową należy odrzucić. Przyjęto oznaczać hipotezy alternatywne przez  $H_1$ ,  $H_2$  lub  $H_3$ . W celu weryfikacji hipotezy zerowej  $H_0$  pobiera się z populacji próbę losową  $x_1, x_2, \dots, x_n$ . Na podstawie próby oblicza się statystykę testową, czyli pewną funkcję próby  $T(x_1, x_2, \dots, x_n)$ , na podstawie której wnioskuje się o odrzuceniu lub nie odrzuceniu hipotezy zerowej. Zatem wartość statystyki testowej z danej próby jest podstawą do odrzucenia hipotezy  $H_0$  lub jej nie odrzucania. Oczywiście sposób konkretnego postępowania zależy od konkretnej hipotezy  $H_0$  i hipotezy alternatywnej. Zawsze jednak weryfikacja powinna przebiegać w taki sposób, aby zapewnić jak najmniejszą możliwość popełnienia pomyłki. W trakcie weryfikacji hipotezy  $H_0$  można popełnić dwa błędy:

- błąd pierwszego rodzaju, polegający na odrzuceniu hipotezy zerowej, mimo że jest ona prawdziwa. Prawdopodobieństwo popełnienia błędu pierwszego rodzaju nazywa się poziomem istotności testu i oznacza przez  $\alpha$ . Najczęściej przyjmowane wartości poziomu istotności to 0,1, 0,05, 0,01;

- błąd drugiego rodzaju, polegający na przyjęciu hipotezy zerowej, gdy w rzeczywistości jest ona nieprawdziwa. Prawdopodobieństwo popełnienia błędu drugiego rodzaju oznacza się przez  $\beta$ . Zestawienie omawianych błędów przedstawiono w tabeli 2.1.

Tabela decyzji

Hipoteza zerowa	DECYZJE	
	Nie odrzucać $H_0$	Odrzucić $H_0$
Hipoteza zerowa prawdziwa	decyzja prawidłowa	błąd I rodzaju
Hipoteza zerowa nieprawdziwa	błąd II rodzaju	decyzja prawidłowa

Wartości  $\alpha$  i  $\beta$  są ze sobą powiązane. Należy dążyć do tworzenia takich testów, które dla wybranego poziomu istotności  $\alpha$  zapewniają możliwie najmniejszą wartość błędu II rodzaju  $\beta$ . Przy czym w testach statystycznych większe znaczenie przypisywane jest poziomowi istotności  $\alpha$ . Określa on stopień pewności, co do odrzucenia hipotezy  $H_0$ , im na mniejszym poziomie istotności test odrzucił hipotezę  $H_0$ , tym bardziej można być pewnym, że hipoteza  $H_0$  jest nieprawdziwa.

Proces weryfikacji hipotezy przebiega w następujących pięciu etapach:

1. Formułowanie hipotezy zerowej  $H_0$  i hipotezy alternatywnej  $H_1$ .
2. Przyjęcie poziomu istotności  $\alpha$ .
3. Określenie – stosownie do postawionej hipotezy zerowej  $H_0$  – statystyki testowej i obliczenie jej wartości na podstawie danych z próby losowej.
4. Przy ustalonym poziomie istotności  $\alpha$  określenie obszarów krytycznych.
5. Wnioskowanie o odrzuceniu lub nie odrzucaniu hipotezy zerowej  $H_0$ .

Aby móc wnioskować o odrzuceniu lub o nie odrzucaniu hipotezy zerowej  $H_0$ , należy określić obszar krytyczny (obszar odrzuceń hipotezy zerowej). Statystyka testowa jest zmienną losową, która może przyjmować różne wartości. Przyjmując, że hipoteza zerowa jest prawdziwa, podzbiór obszaru zmienności statystyki testowej, dla którego prawdopodobieństwo odrzucenia hipotezy zerowej  $H_0$  jest nie większe od przyjętego poziomu istotności  $\alpha$ , nazywa się obszarem krytycznym. Jeżeli obszar krytyczny odpowiadający poziomowi istotności oznaczy się przez  $W_\alpha$ , to gdy statystyka testowa  $T \in W_\alpha$ , hipotezę zerową  $H_0$  należy odrzucić, przy czym  $P(T \in W_\alpha | H_0 \text{ prawdziwa}) \leq \alpha$ . W przeciwnym przypadku brak jest podstaw do odrzucenia hipotezy zerowej  $H_0$ .

Lokalizacja obszaru krytycznego zależy od postaci hipotezy alternatywnej. Zostanie to wyjaśnione na przykładzie. Przyjmijmy, że należy sprawdzić hipotezę dotyczącą wartości średniej, czyli hipotezę o prawdziwości przypuszczalnej wartości oczekiwanej zmiennej losowej. Zakłada się, że zmienna ma rozkład normalny  $N(\mu, \sigma)$  z nieznaną wartością oczekiwaną  $\mu$  i nieznaną wariancją  $\sigma^2$ . Zatem zadanie sprowadza się do weryfikacji hipotezy zerowej:

$$H_0: \mu = \mu_0,$$

przeciw jednej z hipotez alternatywnych:

$$H_1: \mu \neq \mu_0 \text{ lub } H_2: \mu < \mu_0 \text{ lub } H_3: \mu > \mu_0.$$

Jeżeli hipoteza zerowa  $H_0$  jest prawdziwa to statystyka:

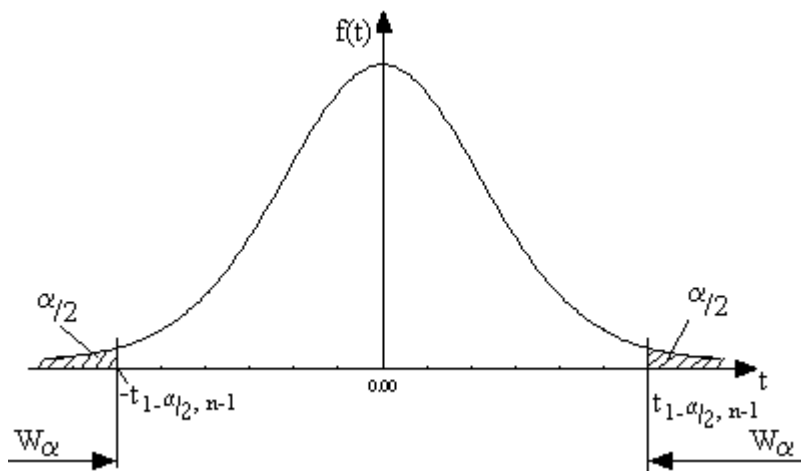
$$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}, \quad (2.1)$$

gdzie:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  ,  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  ,

będąc zmienną losową, ma rozkład t-Studenta o liczbie stopni swobody  $n-1$ . Statystyka  $t$  może przyjmować różne wartości, jeżeli będą to wartości o prawdopodobieństwie zaistnienia równym poziomowi istotności  $\alpha$ , to utworzą one obszar krytyczny. Jeżeli wartość statystyki  $t$ , obliczona z próby, znajdzie się w obszarze krytycznym, to wystąpiło zdarzenie mało prawdopodobne, które może budzić wątpliwość, co do prawdziwości hipotezy zerowej i w konsekwencji odrzuca się hipotezę zerową.

W przypadku hipotezy alternatywnej  $H_1$  obszar krytyczny jest dwustronny (rys. 2.1):

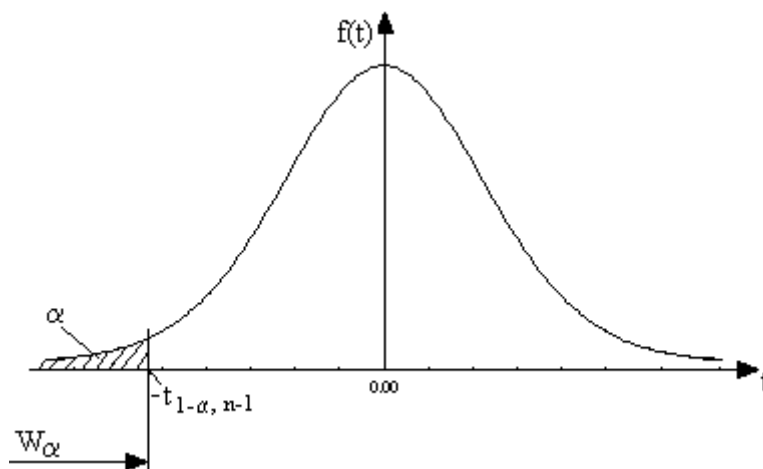
$$W_\alpha \in (-\infty, -t_{1-\alpha/2, n-1}) \cup (t_{1-\alpha/2, n-1}, +\infty).$$



Rys. 2.1. Dwustronny obszar krytyczny przy hipotezie alternatywnej  $H_1$

Przy hipotezie alternatywnej  $H_2$  obszar krytyczny jest lewostronny (rys. 2.2):

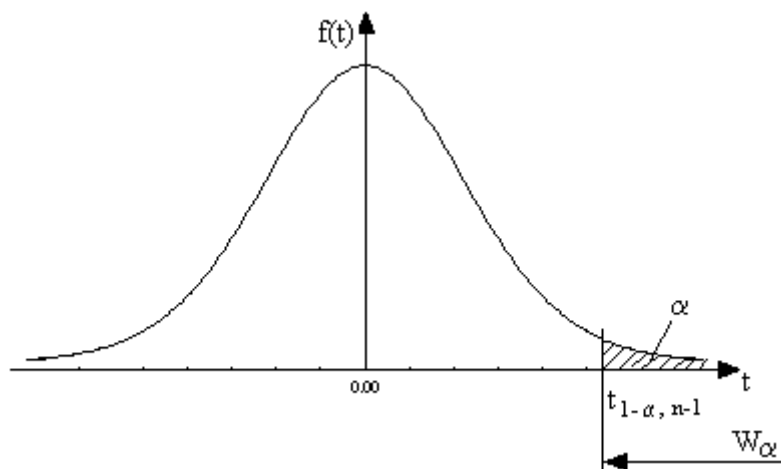
$$W_\alpha \in (-\infty, -t_{1-\alpha, n-1}).$$



Rys. 2.2. Lewostronny obszar krytyczny przy hipotezie alternatywnej  $H_2$

Przy hipotezie alternatywnej  $H_3$  obszar krytyczny jest prawostronny (rys. 2.3):

$$W_\alpha \in \langle t_{1-\alpha, n-1}, +\infty \rangle.$$



Rys. 2.3. Prawostronny obszar krytyczny przy hipotezie alternatywnej  $H_3$

Wartości  $t_{1-\alpha/2, n-1}$  i  $t_{1-\alpha, n-1}$ , będące kwantylami rozkładu t-Studenta odpowiednio rzędu  $1-\alpha/2$  i  $1-\alpha$ , ograniczające obszar krytyczny, nazywane są wartościami krytycznymi i często w literaturze oznaczane są przez  $t_{\alpha/2, n-1}$  i  $t_{\alpha, n-1}$ .

Szczególnie często w praktyce inżynierskiej zachodzi konieczność porównania wartości średnich dwóch próbek. Załóżmy, że pobraliśmy dwie niezależne próbki o licznosciach  $n_1$  i  $n_2$  z dwóch populacji  $X_1$  i  $X_2$ . Chcemy zweryfikować hipotezę zerową polegającą na przypuszczeniu, że  $H_0: \mu_1 = \mu_2$  przeciw jednej z hipotez alternatywnych:

$$H_1: \mu_1 \neq \mu_2 \quad \text{lub} \quad H_2: \mu_1 < \mu_2 \quad \text{lub} \quad H_3: \mu_1 > \mu_2 .$$

Zatem hipoteza zerowa  $H_0$  zakłada, że wartości średnie:

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i} , \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i} , \quad (2.2)$$

pochodzą z tej samej populacji (nie ma między nimi istotnej różnicy).

Testy istotności różnicy między wartościami oczekiwanymi dwu zmiennych można podzielić na dwie grupy:

1. Testy przeznaczone do testowania różnic przy próbach niezależnych.
2. Testy przeznaczone do testowania różnic przy próbach zależnych.

### 2.2.1. Testy dla prób niezależnych

Do weryfikacji hipotezy zerowej stosowane są w zasadzie dwa testy, test t lub test Cochrańa-Coxa. Wybór jednego z nich uzależniony jest od spełnienia warunku równości wariancji  $\sigma_1^2 = \sigma_2^2$  obu populacji. Jeżeli  $\sigma_1^2 = \sigma_2^2$  to stosowany jest test t, w przeciwnym przypadku stosuje się test Cochrańa-Coxa. Oba testy wymagają spełnienia warunku

normalności rozkładu zmiennych  $X_1$  i  $X_2$ , co można sprawdzić odpowiednim testem zgodności (np. testem Shapiro-Wilka lub Lillieforsa).

### Test t

Statystyką testową testu t jest:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{[s_1^2(n_1 - 1) + s_2^2(n_2 - 1)](n_1 + n_2)}{(n_1 + n_2 - 2) n_1 n_2}}}, \quad (2.3)$$

gdzie:  $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$ ,  $s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$ .

Jeżeli weryfikowana hipoteza jest prawdziwa, to statystyka testowa t ma rozkład t-Studenta z  $n_1+n_2-2$  stopniami swobody. Dalsze postępowanie jest identyczne, jak przedstawiono przy omawianiu poprzedniego testu.

### Test Cochran–Coxa.

Statystyką testową tego testu jest:

$$C = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{(n_1 - 1)} + \frac{s_2^2}{(n_2 - 1)}}}. \quad (2.4)$$

Rozkład tej statystyki zależy od licznosci próbek  $n_1$  i  $n_2$  oraz od stosunku  $\sigma_1/\sigma_2$ . Obszary krytyczne wyznacza się analogicznie jak poprzednio, oczywiście odczytując wartości krytyczne z odpowiednich tablic statystycznych.

### Test F

Do weryfikacji hipotezy o równości wariancji  $H_0: \sigma_1^2 = \sigma_2^2$  przeciw jednej z hipotez alternatywnych  $H_1: \sigma_1^2 > \sigma_2^2$  lub  $H_2: \sigma_1^2 < \sigma_2^2$  stosuje się test F. Statystyką testową tego testu jest:

$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_2^2, s_1^2)}. \quad (2.5)$$

Jeżeli hipoteza zerowa jest prawdziwa, to statystyka F ma rozkład F-Snedecora z  $n_1-1$  stopniami swobody licznika i  $n_2-1$  stopniami swobody mianownika, gdzie  $n_1$  jest licznoscią próbki, z której obliczono wariancję licznika, a  $n_2$  – mianownika. Obszarem krytycznym jest przedział  $W_\alpha \in \langle F_{1-\alpha, n_1-1, n_2-1}, +\infty \rangle$ . Jeżeli wartość F należy do obszaru krytycznego, hipotezę o równości wariancji odrzuca się na poziomie istotności  $\alpha$ .

## 2.2.2. Test dla prób zależnych

Test ten stosuje się wówczas, gdy mamy dwie grupy wyników dla tych samych elementów próby. Dla każdego elementu próby losowej otrzymuje się pary liczb  $x_i$  i  $y_i$ . Oblicza się ich różnicę  $d_i = x_i - y_i$  i zakłada, że populacja różnic D ma rozkład normalny. Statystyka testowa ma wówczas postać:

$$t = \frac{\bar{d}}{s_d} \sqrt{n}, \quad (2.6)$$

gdzie:  $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ ,  $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$ .

Jeżeli hipoteza zerowa  $H_0$  jest prawdziwa, statystyka  $t$  ma rozkład t-Studenta o  $n-1$  stopniach swobody. Zasady wyznaczania obszarów krytycznych i podejmowania decyzji są identyczne jak przy wcześniej omówionym teście  $t$ .

**Uwaga:** Tradycyjny sposób wnioskowania oparty o obszary krytyczne wymaga odczytania z odpowiednich tablic statystycznych wartości krytycznych i porównania wartości statystyki testowej z odczytanymi wartościami krytycznymi. W programach komputerowych, aby uniknąć konieczności korzystania z tablic statystycznych, podawana jest wartość prawdopodobieństwa  $p$  zaistnienia statystyki testowej, przy założeniu, że hipoteza zerowa  $H_0$  jest prawdziwa. Im prawdopodobieństwo  $p$  jest mniejsze, tym bardziej świadczy to przeciwko hipotezie  $H_0$ . Jeżeli  $p < \alpha$ , to na danym poziomie istotności  $\alpha$  odrzuca się hipotezę zerową  $H_0$ , natomiast, gdy  $p > \alpha$ , to na przyjętym poziomie istotności nie ma podstaw do odrzucenia hipotezy zerowej  $H_0$ .

### 2.3. Obliczenia programem STATISTICA

W programie STATISTICA do testowania różnic między średnimi służą procedury **Test  $t$  dla prób niezależnych** i **Test  $t$  dla prób zależnych** w module **Podstawowe statystyki**. Pierwsza procedura stosowana jest do weryfikacji hipotezy o równości wartości oczekiwanych w dwóch populacjach na podstawie prób niezależnych, a druga w przypadku prób zależnych.

#### 2.3.1. Testy dla prób niezależnych

Po wybraniu procedury **Test  $t$  dla prób niezależnych** otwiera się okno **Testy dla prób niezależnych** przedstawione na rys. 2.4. Na górze okna znajduje się pole **Plik wejściowy** dla ustalenia sposobu organizacji danych. Możliwe są dwa ustawienia:

1. **Jeden wiersz na przypadek (użyj zmiennej grupującej)** – jest to podstawowy sposób organizacji danych w pliku. Każdy wiersz w zbiorze danych przedstawia jedną obserwację, a każda kolumna reprezentuje jedną badaną zmienną. Czyli dane wpisywane są przypadek za przypadkiem, więc aby przeprowadzić test  $t$ , należy podać, do jakiej grupy zaliczyć każdy z przypadków. Realizowane to jest poprzez wprowadzenie dodatkowej zmiennej zwanej zmienną grupującą.

2. **Każda zmienna zawiera dane dla jednej grupy** – każda zmienna (kolumna) przedstawia jedną grupę danych. Po wybraniu tej opcji okno **Testy dla prób niezależnych** zmieni swoją postać.

W przypadku wyboru pierwszej opcji, należy po naciśnięciu przycisku **Zmienne** określić zmienną grupującą i zmienne zależne. Dla zmiennej grupującej należy podać kody określające grupy (Kod grupy nr 1, Kod grupy nr 2), które mają być porównane. Jeżeli nie zna się kodów zmiennej grupującej, należy kliknąć dwukrotnie pola kodów. Otworzy się okno zawierające wszystkie kody całkowite i ich alfanumeryczne odpowiedniki, które można przenieść do pola kod grupy przez ich dwukrotne kliknięcie.

Rys. 2.4. Okno Testy dla prób niezależnych (opcja podstawowa)

O tym, co zostanie obliczone i podane w wynikach decydują ustawienia w polu **Opcje**, umożliwiając one:

– **Usuwanie braków danych przypadkami** – gdy opcja jest włączona program pomija wszystkie przypadki (wiersze), w których brakuje wartości choćby dla jednej zmiennej zależnej spośród wybranych zmiennych do analizy. Jeżeli opcja jest wyłączona, brakujące dane usuwane są parami dla dwóch aktualnie analizowanych zmiennych;

– **Pokaż długie nazwy zmiennych** – oprócz krótkich nazw zmiennych wyświetlane są ich długie nazwy, o ile zostały zdefiniowane;

– **Test t z oddzielną oceną wariancji** – obliczana jest wartość testu Cochran–Coxa na podstawie oddzielnych ocen wariancji w obu próbach. Opcja ta stosowana jest wtedy, gdy wariancje w próbach różnią się (test F lub test Levene'a odrzuca hipotezę o równości wariancji). Zaleca się użycie tej opcji, gdy licznosci grup znacznie różnią się;

– **Test wielowymiarowy (T<sup>2</sup> Hotellinga)** – jest uogólnieniem testu t na większą liczbę zmiennych;

– **Test Levene'a jednorodności wariancji** – test t opiera się na założeniu, że wariancje w obu próbach są równe (jednorodne). Test Levene'a jest drugim obok testu F mocnym



testem do sprawdzenia hipotezy o równości wariancji. Jeżeli test Levene'a wykaże statystyczną istotność, to hipotezę o jednorodności wariancji należy odrzucić;

– **Browna i Forsytha jednorodności wariancji** – jest to zmodyfikowany test Levene'a, jest on bardziej odporny na niespełnienie warunku o normalności rozkładów.

Porównanie wartości średnich w dwóch grupach można przedstawić graficznie. Służą do tego następujące przyciski:

**Wykres ramkowy** – umożliwia wykonanie wykresów ramkowych dla wybranych zmiennych; jeden wykres na jedną zmienną.

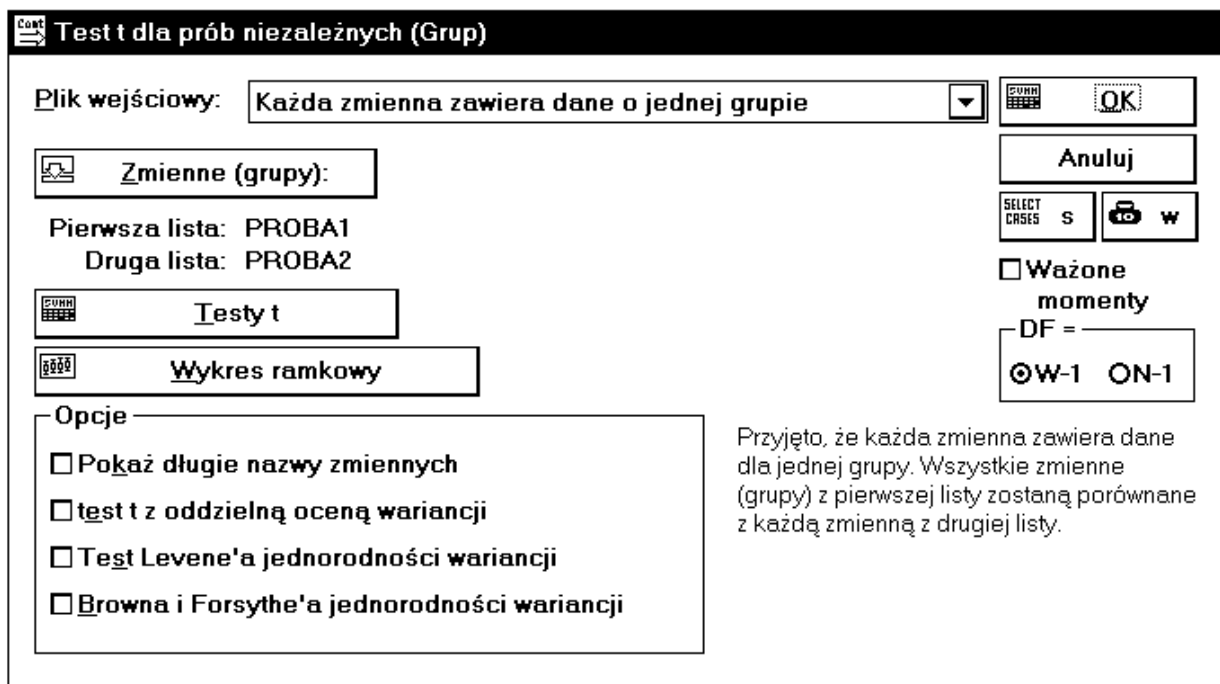
**Skategoryzowane histogramy** – umożliwia wykonanie histogramów dla skategoryzowanych zmiennych.

**Skat. wyk. normalności** – umożliwia wykonanie normalnych wykresów prawdopodobieństwa dla skategoryzowanych zmiennych.

**Skat. wyk. normalności bez trendu** – umożliwia wykonanie normalnych wykresów prawdopodobieństwa z eliminacją trendu liniowego dla skategoryzowanych zmiennych.

**Skat. wykresy rozrzutu** – umożliwia wykonanie wykresów rozrzutu wraz z linią regresji dla skategoryzowanych zmiennych.

W przypadku wyboru w polu **Plik wejściowy** opcji **Każda zmienna zawiera dane o jednej grupie** okno **Testy dla prób niezależnych** zostanie wyświetlone w następującej postaci (rys. 2.5).



Rys. 2.5. Okno Testy dla prób niezależnych (opcja alternatywna)

Po wskazaniu zmiennych i ustawieniu odpowiednich opcji, obliczenie testów realizuje się po naciśnięciu przycisku **Test t** lub przycisku **OK**.

### Przykład 2.1

Przy badaniu stabilności operacji obróbki tulejek na automacie tokarskim pobrano dwie próbki. Pierwszą w okresie początkowym i drugą po określonym czasie. W wyniku pomiaru średnicy zewnętrznej tulejek w pierwszej próbie uzyskano: 20,04; 20,05; 20,06; 19,98; 19,99; 20,06; 19,97; 19,96; 20,00; 20,02 mm. Średnice tulejek w drugiej próbie były następujące: 20,00; 20,05; 20,02; 20,04; 19,99; 20,08; 20,06; 20,10; 19,98; 20,06 mm. Zakładając, że zmiana warunków obróbki w ciągu odcinka czasu pomiędzy dwiema próbkami nie ma wpływu na wariancję oraz że średnica tulejek ma rozkład normalny, sprawdzić na poziomie istotności  $\alpha = 0,05$ , czy zaszła w tym czasie zmiana nastawienia obrabiarki.

### Rozwiązanie

Należy zweryfikować hipotezę  $H_0: \mu_1 = \mu_2$  przy hipotezie alternatywnej  $H_1: \mu_1 \neq \mu_2$ . Ponieważ nie wiemy, czy spełniony jest warunek jednorodności wariancji w obu populacjach, wykonamy wstępnie obliczenia przyjmując, że jest on spełniony. Wyniki obliczeń przedstawiono w tabeli 2.2.

Tabela 2.2

Wyniki obliczeń testu dla prób niezależnych

STAT. PODST. STATYST.	Testy dla prób niezależnych (tulejki.sta)					
	Uwaga: Zmienne traktowane są jako niezależne próby.					
Grupa 1 vs. Grupa 2	Średnia Grupa 1	Średnia Grupa 2	t	df	p	Nważnych Grupa 1
PROBA1 vs.PROBA2	20,01300	20,03800	-1,43885	18	,167355	10
STAT. PODST. STATYST.	Testy dla prób niezależnych (tulejki.sta)					
	Uwaga: Zmienne traktowane są jako niezależne próby.					
Grupa 1 vs. Grupa 2	Nważnych Grupa 2	Odch.Std Grupa 1	Odch.Std Grupa 2	prop.F warianc.	p warianc.	
PROBA1 vs.PROBA2	10	,038020	,039665	1,088394	,901662	

Na początku analizy wyników sprawdza się, czy spełniony jest warunek jednorodności wariancji, czyli weryfikuje się hipotezę zerową  $H_0: \sigma_1^2 = \sigma_2^2$  przeciw hipotezie alternatywnej  $H_1: \sigma_1^2 < \sigma_2^2$ . Obliczona wartość statystyki  $F = 1,088394$  jest mniejsza od wartości krytycznej  $F_{0,95, 9, 9} = 3,18$  (tablica II), nie należy do obszaru krytycznego; świadczy też o tym, wartość poziomu prawdopodobieństwa  $p = 0,901662$ , która jest dużo większa od przyjętego poziomu istotności  $\alpha = 0,05$ , a więc nie ma podstaw do odrzucenia hipotezy zerowej o równości wariancji. Zatem do weryfikacji hipotezy zerowej o równości wartości oczekiwanych w obu populacjach można zastosować test t. Obliczona wartość statystyki testowej  $t = -1,43885$  i jej bezwzględna wartość jest mniejsza od wartości krytycznej  $t_{0,975, 18} = 2,10$  (tablica I), a więc statystyka testowa nie należy do obszaru krytycznego  $W_\alpha \in (-\infty, -2,10) \cup (+2,10, +\infty)$  i nie ma podstaw do odrzucenia hipotezy zerowej. Świadczy też o tym wartość poziomu

prawdopodobieństwa  $p = 0,167355$ , która jest większa od przyjętego poziomu istotności testu  $\alpha = 0,05$ .

**Uwaga!** W arkuszu wyników kolorem czerwonym zaznaczane są wyniki istotne statystycznie, to jest te, dla których poziom prawdopodobieństwa  $p$  jest mniejszy od wartości  $0,05$  (domyślne ustawienie poziomu istotności). Wartości poziomu prawdopodobieństwa obliczane są przy założeniu, że hipotezą alternatywną jest hipoteza  $H_1: \mu_1 \neq \mu_2$  i tylko w tym przypadku można przy podejmowaniu decyzji posługiwać się wartościami  $p$ . W przypadku innych hipotez alternatywnych należy korzystać tylko z obliczowych wartości statystyk testowych i porównywać je z wartościami krytycznymi odczytanymi z odpowiednich tablic statystycznych. Gdy nie jest spełniony warunek równości wariancji, stosowany jest test Cochran-Coxa, którego statystyka testowa w arkuszu wyników jest oznaczona literą  $t$ , identycznie jak przy teście  $t$ , lecz nie ma ona rozkładu  $t$ -Studenta, i wartości krytyczne należy odczytywać z innych tablic statystycznych.

### 2.3.2. Test $t$ dla prób zależnych

Po wybraniu procedury **Testy  $t$  dla prób zależnych** otworzy się okno przedstawione na rys. 2.6.

The screenshot shows the 'Testy t dla prób zależnych (skorelowanych)' dialog box. The 'Zmienne:' field contains 'Pierwsza PROBA1' and 'Druga lista: PROBA2'. The 'Testy t' and 'Wykresy ramkowe' buttons are visible. In the 'Wyniki' section, 'Macierz testów t (średnie, różnice)' is selected. In the 'Opcje' section, 'Usuwanie BD przypadkami' and 'Pokaż długie nazwy zmiennych' are unchecked. On the right, 'OK' and 'Anuluj' buttons are present, along with 'SELECT CASES' (with 'S' selected), 'Ważone momenty' (unchecked), 'DF =', and radio buttons for 'W-1' (selected) and 'ON-1'.

Rys. 2.6. Okno Testy  $t$  dla prób zależnych

Wyboru zmiennych do analizy dokonuje się po naciśnięciu przycisku **Zmienne**. Dane w arkuszu danych powinny być rozmieszczone w ten sposób, że każda zmienna jest w oddzielnej kolumnie. Zakres obliczeń i format arkusza wyników zależy od ustawień w polach **Opcje** i **Wyniki**. Znaczenie opcji w polu **Opcje** jest podobne jak we wcześniej opisanych oknach dla prób niezależnych. Natomiast opcje w polu **Wyniki** umożliwiają:

– **Macierz testów t (średnie, różnice)** – tworzony jest ciąg arkuszy wyników. W każdym arkuszu wyników, dla każdej pary wybranych zmiennych, podawane są: różnice średnich, wartości testu t i poziomy prawdopodobieństwa p;

– **Szczegółowa tabela wyników** – w pojedynczym arkuszu wyników podawana jest szczegółowa tabela wyników, w której dla każdej pary zmiennych podawane są: liczebności grup, różnice średnich, odchylenie standardowe różnic, wartości testu t, stopnie swobody statystyki t oraz poziomy prawdopodobieństwa p.

### Przykład 2.2

Dla porównania dwóch mikrometrów wykonano pomiary 9. płytek. Wyniki pomiarów przedstawiono w tabeli poniżej.

Nr płytki	1	2	3	4	5	6	7	8	9
Wynik pomiaru I mikrometrem	3,89	5,91	7,93	9,83	12,96	15,99	18,85	21,90	24,78
Wynik pomiaru II mikrometrem	3,90	5,88	7,90	9,79	12,90	15,97	18,86	21,88	24,76

Zakładając, że wyniki pomiarów mają rozkład normalny, zweryfikować na poziomie istotności  $\alpha = 0,05$  hipotezę zerową, że średnie wskazania obu mikrometrów są jednakowe  $H_0: \mu_1 = \mu_2$ , przy hipotezie alternatywnej, że wskazania mikrometrów różnią się  $H_1: \mu_1 \neq \mu_2$ .

### Rozwiązanie

Po wprowadzeniu danych do arkusza danych (w dwóch kolumnach) wybieramy procedurę **Test t dla prób zależnych** w module **Podstawowe statystyki**. Po wskazaniu zmiennych do analizy i naciśnięciu przycisku **Testy t** otrzymuje się wyniki przedstawione w tabeli 2.3.

Tabela 2.3

Wyniki testu dla prób zależnych

STAT. PODST.	Test T dla prób zależnych (tulejki.sta)							
	Zaznaczone różnice są istotne z $p < 0,05000$							
Zmienna	Średnia	Od.std.	N	Różnica	Od.std Różnica	t	df	p
MIKRM_I	13,56000	7,308678						
MIKRM_II	13,53778	7,310422	9	,022222	,022236	2,998127	8	,017120

Obliczona wartość statystyki  $t = 2,998127$ . Ze względu na hipotezę alternatywną  $H_1: \mu_1 \neq \mu_2$ , obszarem krytycznym jest przedział  $(-\infty, -2,262) \cup (2,262, +\infty)$ . Statystyka testowa należy do obszaru krytycznego, a więc należy odrzucić, na poziomie istotności  $\alpha = 0,05$ , hipotezę

zerową o jednakowych wskazaniach mikrometrów na korzyść hipotezy alternatywnej. Świadczy też o tym wartość poziomu prawdopodobieństwa  $p = 0,01712$ , która jest mniejsza od przyjętego poziomu istotności  $\alpha = 0,05$ .

### **3. BADANIE ZGODNOŚCI ROZKŁADU ZMIENNEJ LOSOWEJ Z ROZKŁADEM TEORETYCZNYM**

#### **3.1. Cel ćwiczenia**

Celem ćwiczenia jest zapoznanie się ze sposobami weryfikacji hipotezy o zgodności rozkładu zmiennej losowej z proponowanym rozkładem teoretycznym (modelem rozkładu).

#### **3.2. Opis metody**

Wielokrotnie w analizie statystycznej zakłada się, że zmienne losowe (badane cechy) mają pewne rozkłady np. normalne, Poissona czy inne. Powstaje, zatem pytanie, czy takie założenie jest zgodne z rzeczywistością. Aby to sprawdzić dokonuje się obserwacji badanej cechy (pobiera próbę losową) i porównuje rozkład obserwacji zawartych w próbie z pewnym teoretycznym rozkładem, który to rozkład można uważać za proponowaną hipotezę zerową. Hipoteza zerowa może być postawiona w różnej postaci, np.:

$H_0$ : zmienna losowa  $X$  ma rozkład normalny o parametrach  $\mu = 5$  i  $\sigma = 0,5$

lub  $H_0$ : zmienna losowa  $X$  jest zmienną Poissona o parametrze  $\lambda = 5$

lub przy braku konkretnej wartości parametru

$H_0$ : zmienna losowa  $X$  podlega rozkładowi wykładniczemu.

Hipoteza alternatywna określająca, że zmienna nie jest taka jak żąda hipoteza zerowa  $H_0$ , jest najczęściej bardzo złożona. Na przykład, jeśli  $X$  nie jest zmienną losową o rozkładzie  $N(5, 0,5)$ , to może mieć rozkład  $N(10, 0,5)$  lub  $N(5, 0,3)$  lub  $N(15, 0,4)$  lub może w ogóle nie mieć rozkładu normalnego, może mieć rozkład logarytmo-normalny lub gamma lub dowolny z nieskończenie wielu, nazwanych lub bez nazwy, ciągłych, dyskretnych lub mieszanych, różnorodnych rozkładów.

##### **3.2.1. Porównanie kształtów histogramów**

Najprostszym sposobem oceny zgodności rozkładu obserwacji próby z rozkładem hipotetycznym jest wizualne porównanie histogramu częstości z funkcją gęstości lub histogramu skumulowanej częstości z dystrybuantą. Porównanie wizualne pozwala na natychmiastowe oszacowanie bliskości rozkładu danych zaobserwowanych z rozkładem hipotetycznym oraz dostarcza cennych informacji o obszarach niezgodności.

###### **3.2.1.1. Porównanie histogramu częstości z funkcją gęstości**

Z powodu konieczności grupowania danych w klasach, potrzebnych dla zbudowania histogramu, najczęściej istotne cechy prawdziwych końców rozkładu nie są przedstawiane na wykresie i ulegają zatraceniu. Wybór węższych przedziałów klasowych zwiększa czytelność wykresów, ale mniejsza liczba obserwacji w klasach powoduje większe zmiany w wysokości

słupków. Należy, zatem odpowiedzieć na pytanie, czy obserwowana zgodność jest wystarczająca czy też należy żądać zgodności ściślejszej lub inaczej mówiąc, czy kształt histogramu częstości istotnie różni się od kształtu proponowanej funkcji gęstości? Aby odpowiedzieć na to pytanie należy dokładniej zrozumieć histogram. Mając dany rzeczywisty rozkład zmiennej losowej i zbiór przedziałów klasowych, obserwowane częstości w przedziałach klasowych są zmiennymi losowymi, wprost proporcjonalnymi do liczby obserwacji z próby należących do tych przedziałów. Liczby te są zmiennymi losowymi o łącznym rozkładzie wielomianowym. Dla ustalonej liczby klas  $k$ , wzrost liczebności  $n$  próby może uczynić zgodność histogramu częstości z funkcją gęstości niemal pewną. Z drugiej strony, zmniejszenie liczby klas  $k$  da ten sam efekt. Jednak w tej sytuacji wiele różnych modeli rozkładu będzie miało niemal identyczną zgodność. Ta niejednoznaczność może być zmniejszona i kształt właściwego rozkładu dokładniej wyznaczony, tylko przez zastosowanie większej liczby klas  $k$ , ale wiąże się to z większą zmiennością w obserwowanych wartościach częstości w poszczególnych klasach konkretnego histogramu. Duża zgodność nie jest w tym przypadku możliwa i dlatego jest mało prawdopodobne, aby można było znaleźć wyraźne wizualne potwierdzenie swojego rozkładu w danych. Zatem porównując kształt histogramu częstości danych z funkcją gęstości rozkładu hipotetycznego, należy tak wybrać liczbę klas histogramu, aby wypośrodkować między większą niejednoznacznością a większą zmiennością. Więc jeżeli uzasadnia się adekwatność swojego rozkładu na podstawie wizualnego porównania, to trzeba pamiętać, że jeśli przyjmie się małą liczbę klas, to inne rozkłady będą miały też prawie identyczną zgodność, a jeśli wybierze się zbyt dużą liczbę klas, to duże podobieństwo kształtów będzie mało prawdopodobne.

### 3.2.1.2. Porównanie histogramu skumulowanej częstości z dystrybuantą

Zamiast porównywać zaobserwowany histogram częstości danych z funkcją gęstości, można porównywać histogram skumulowanej częstości (dystrybuantę empiryczną) z wykresem dystrybuanty. Porównanie kształtu histogramu skumulowanej częstości z dystrybuantą ma wyraźną przewagę nad porównaniem histogramu częstości z funkcją gęstości. W tym przypadku znacznie zmniejsza się niejednoznaczność i zmienność histogramu związana z koniecznością grupowania danych. Należy jednak pamiętać, że można oczekiwać zmienności histogramu skumulowanej częstości w stosunku do dystrybuanty, nawet wtedy, gdy dane mają rozkład określony tą dystrybuantą oraz że inne prawo probabilistyczne może rządzić generowaniem danych nawet, jeśli wydają się one zgodne z proponowanym rozkładem.

W praktyce porównywanie wykresów może być uproszczone przez zmianę skali, to jest dzięki specjalnie wykonanej siatce zwanej siatką prawdopodobieństwa. Skale na osiach współrzędnych siatki są tak dobrane, że wykres dystrybuanty odpowiedniego hipotetycznego rozkładu jest linią prostą. Przy użyciu takiej siatki porównanie hipotetycznego rozkładu z danymi sprowadza się do porównania skumulowanych częstości tych danych (wykreślonych na tej siatce) z linią prostą.

### 3.2.2. Testy zgodności

Dotychczas oceniano zgodność rozkładu obserwowanych danych z rozkładem teoretycznym na podstawie wykresów. Jest to jednak ocena subiektywna zależna od oceniającego. Bardziej obiektywne są oceny ilościowe zgodności na podstawie testów statystycznych. Opracowano wiele testów, które można podzielić na dwie grupy. Do pierwszej należą testy, które mogą być stosowane dla dowolnego hipotetycznego rozkładu. Najbardziej rozpowszechnione są w tej grupie dwa testy: test chi-kwadrat i test Kołmogorowa. Do drugiej grupy należą testy bardziej specjalistyczne służące do badania zgodności z konkretnym rozkładem hipotetycznym. Najliczniejszą grupę stanowią tu testy normalności, czyli testy badające zgodność rozkładu zmiennej losowej z rozkładem normalnym. Wśród nich najczęściej stosowane są: test Shapiro-Wilka i test Lillieforsa.

#### 3.2.2.1. Test chi-kwadrat ( $\chi^2$ )

Jest to najbardziej rozpowszechniony test opracowany przez Karla Pearsona. Hipotetyczny rozkład zmiennej losowej  $X$  może być dowolnym rozkładem i dotyczyć zarówno zmiennej losowej ciągłej, jak i skokowej. Hipoteza zerowa będzie miała postać:

$H_0$ : zmienna losowa  $X$  ma rozkład określony funkcją gęstości  $f(x)$ , przy hipotezie alternatywnej, że rozkład zmiennej  $X$  jest inny niż to określa hipoteza zerowa.

Test  $\chi^2$  wykonuje się w następujący sposób. Dokonuje się  $n$  obserwacji zmiennej losowej  $X$ . Otrzymany zakres zmienności zaobserwowanych wartości dzieli się na  $k$  klas. Oblicza się liczbę obserwacji  $n_i$  należących do każdej klasy. Jeżeli hipoteza zerowa jest prawdziwa, można oczekiwać, że liczba obserwacji w każdej klasie powinna wynosić  $np_i$ . Liczby  $p_i$  są prawdopodobieństwami zaobserwowania zmiennej  $X$  w  $i$ -tej klasie i mogą być obliczone z wzoru:

$$p_i = \int_{\Delta_i} f(x) dx. \quad (3.1)$$

Porównując liczebności zaobserwowane z teoretycznymi otrzymuje się statystykę testową:

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n * p_i)^2}{n * p_i}. \quad (3.2)$$

Statystyka ta posiada asymptotycznie (tzn. dla  $n \rightarrow \infty$ ) rozkład  $\chi^2$  z liczbą stopni swobody  $k - 1$ . Rzadko jednak zdarza się, że można postawić hipotezę zerową, w której podaje się wartości parametrów rozkładu. Najczęściej w hipotezie zerowej podaje się typ rozkładu zmiennej, a wartości parametrów muszą być oszacowane na podstawie zebranych obserwacji (próby). W tym przypadku liczbę stopni swobody statystyki  $\chi^2$  zmniejsza się o liczbę estymowanych parametrów. Zatem liczba stopni swobody statystyki  $\chi^2$  jest równa  $k - r - 1$ , gdzie  $r$  oznacza liczbę oszacowanych parametrów na podstawie zaobserwowanych danych.



Podział zakresu zaobserwowanych wartości zmiennej pomiędzy poszczególne klasy należy wykonać w ten sposób, aby liczba obserwacji  $n_i$  w każdej klasie była dostatecznie duża tak, aby rozkład statystyki  $\chi^2$  można było rozpatrywać jako rozkład  $\chi^2$ . Z drugiej strony liczba klas powinna być dostatecznie duża, aby dobrze odzwierciedlała funkcję gęstości  $f(x)$ . Jako "złoty środek" przyjmuje się zwykle, że  $n_i$  powinno być, co najmniej równe 5. W praktyce przyjmuje się, że jeżeli parametry rozkładu ocenia się na podstawie próby, to liczebność próby powinna wynosić, co najmniej 100, a liczba klas od 10 do 25. Dla rozkładu normalnego, gdy parametry rozkładu są znane, liczba klas może być zmniejszona, ale powinna wynosić od 5 do 10. Następnie wybiera się poziom istotności  $\alpha$ , oblicza wartość statystyki i porównuje ją z wartością krytyczną  $\chi^2_{\alpha, k-r-1}$  dla  $k-r-1$  stopni swobody. Jeżeli zachodzi nierówność  $\chi^2 > \chi^2_{\alpha, k-r-1}$  to hipotezę zerową należy odrzucić. W przeciwnym przypadku wnioskuje się, że nie ma podstaw do odrzucenia hipotezy zerowej.

### 3.2.2.2. Test Kołmogorowa

Przyjmuje się, że zmienna losowa  $X$  typu ciągłego ma rozkład określony dystrybuantą  $F(x)$ . Pobraną próbę porządkuje się w ciągu niemalejącym i wyznacza dystrybuantę empiryczną  $F_n(x)$ . Statystyką testową testu jest:

$$D_n = \max_{1 \leq i \leq n} [|F(x_i) - F_n(x_i)|]. \quad (3.3)$$

Statystyka ta przy prawdziwości hipotezy zerowej ma rozkład Kołmogorowa, którego jedynym parametrem jest liczebność próby. Przyjmując wartość poziomu istotności  $\alpha$ , z tablic wartości krytycznych rozkładu Kołmogorowa odczytuje się wartość krytyczną  $D_{\alpha, n}$ . Jeżeli  $D_n > D_{\alpha, n}$ , to hipotezę zerową odrzuca się na poziomie istotności  $\alpha$ . Przewaga testu Kołmogorowa nad testem  $\chi^2$  polega na tym, że nie wymaga on łączenia danych w grupy i porównywania oddzielnych klas, ale porównuje wszystkie dane w niezmięnionej postaci, a przez to jest niewrażliwy na działanie skrajnych, mało licznych klas. Test zgodności Kołmogorowa może być stosowany dla prób o małej liczności (dla  $n > 5$ ), ale tylko wtedy, gdy hipotetyczny rozkład jest przyjmowany całkowicie niezależnie od danych, tzn. nie zachodzi konieczność szacowania parametrów rozkładu na podstawie próbki. W przypadku, gdy wartości parametrów rozkładu hipotetycznego są estymowane z próby, liczebność próbki powinna wynosić, co najmniej 100, i wówczas można stosować test graniczny  $\lambda$ -Kołmogorowa.

### 3.2.2.3. Test normalności Lillieforsa

Test ten służy do weryfikacji hipotezy zerowej, że zmienna losowa  $X$  ma rozkład normalny  $N(\mu, \sigma)$ . Jeżeli parametry tego rozkładu są znane, to do weryfikacji hipotezy zerowej można stosować opisany wcześniej test Kołmogorowa. Jeżeli zaś parametry te nie są znane to należy stosować zmodyfikowany test Kołmogorowa znany pod nazwą testu Lillieforsa. Statystyka tego testu jest taka sama jak testu Kołmogorowa, z tym, że do wyznaczenia hipotetycznej dystrybuanty  $F(x)$  wykorzystuje się oceny  $\bar{x}$  i  $s^2$  nieznanymi parametrów rozkładu normalnego. Otrzymaną wartość statystyki testowej  $D_n$  porównuje się ze zmodyfikowanymi wartościami krytycznymi testu Kołmogorowa.

### 3.2.2.4. Test normalności Shapiro-Wilka

Drugim testem do weryfikacji hipotezy o normalności rozkładu zmiennej losowej, w przypadku, gdy nieznane są parametry rozkładu hipotetycznego, jest test Shapiro-Wilka. Pobraną próbę należy uporządkować niemalejąco i obliczyć statystykę testową:

$$W = \frac{\left( \sum_{i=1}^l a_{i,n} (x_{n-i+1} - x_i) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.4)$$

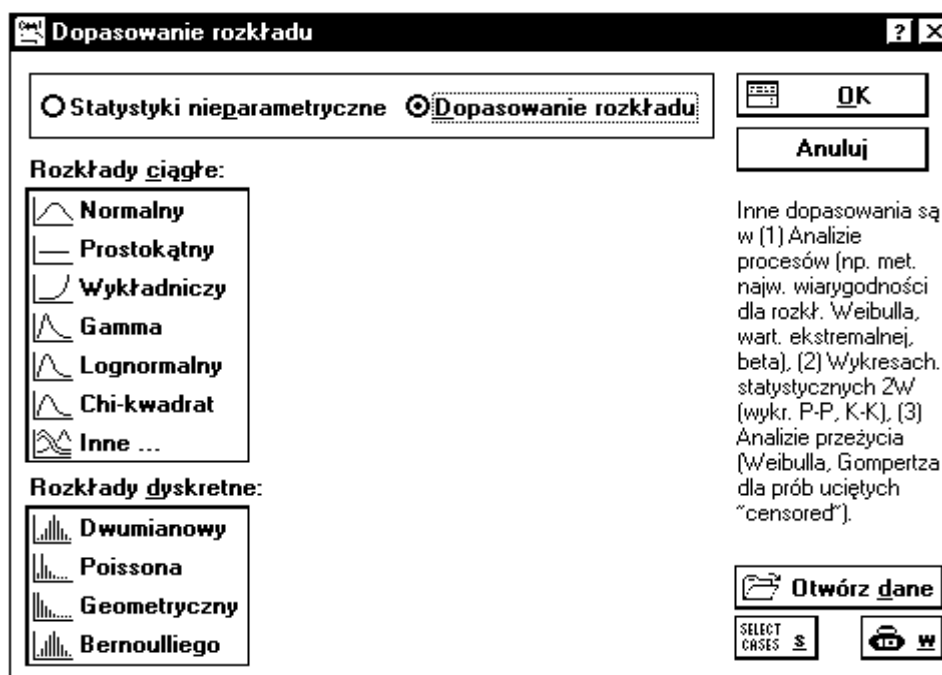
gdzie  $a_{i,n}$  są specjalnymi współczynnikami oraz  $l = n/2$ , dla  $n$  parzystych i  $l = (n - 1)/2$ , dla  $n$  nieparzystych.

W zależności od przyjętego poziomu istotności  $\alpha$  i liczności próbki  $n$  należy odczytać z tablic wartości krytycznych testu Shapiro-Wilka wartość krytyczną  $W_{\alpha, n}$ . Jeżeli  $W < W_{\alpha, n}$  to hipotezę o normalności rozkładu zmiennej losowej  $X$  należy odrzucić.

## 3.3. Obliczenia programem STATISTICA

### 3.3.1. Testy ogólne

W programie STATISTICA do weryfikacji hipotezy o zgodności rozkładu badanej zmiennej losowej z proponowanym rozkładem teoretycznym służy moduł *Statystyki nieparametryczne/Rozkłady*. Po zaznaczeniu opcji **Dopasowanie rozkładów** otwiera się okno przedstawione na rys. 3.1. Możliwe jest badanie zgodności rozkładu zmiennej losowej z każdym z wymienionych rozkładów teoretycznych dla zmiennej losowej ciągłej lub dyskretnej po wyborze odpowiedniego rozkładu teoretycznego i naciśnięciu przycisku **OK**.



Rys. 3.1. Okno Dopasowanie rozkładu

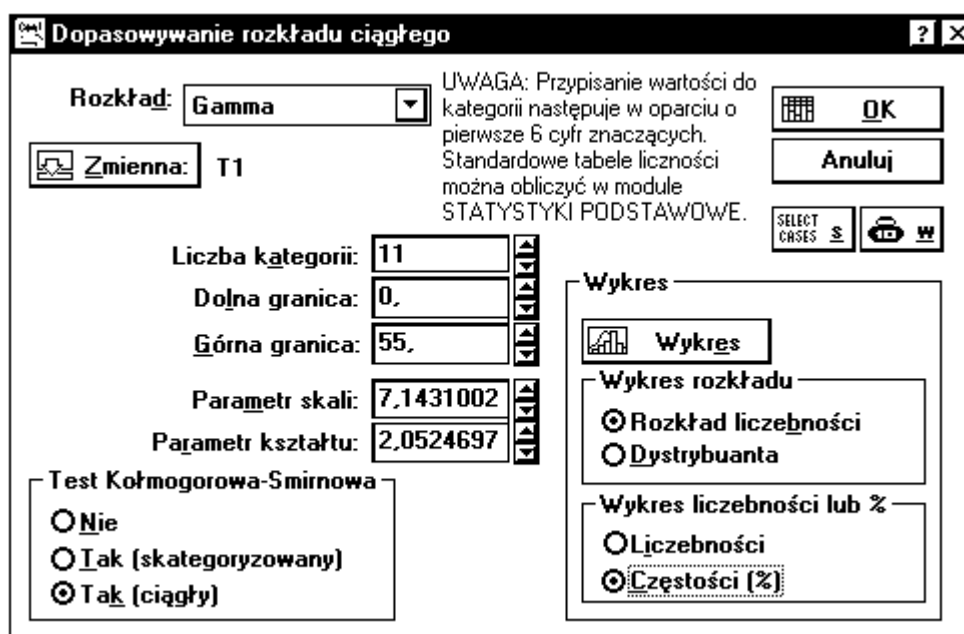
W tabeli 3.1 przedstawiono ogólną charakterystykę proponowanych rozkładów teoretycznych.

Tabela 3.1

Podstawowe informacje o rozkładach

Nazwa rozkładu	Parametry rozkładu	Typ zmiennej
Normalny	wartość oczekiwana (średnia) $\mu$ wariancja $\sigma^2$	ciągła $-\infty < x < +\infty$
Prostokątny	dolna granica a górną granicą b średnia liczebność	ciągła $a \leq x \leq b$
Wykładniczy	lambda	ciągła $x \geq 0$
Gamma	parametr kształtu $\alpha$ parametr skali $\beta$	ciągła $x > 0$
Lognormalny (logarytmo normalny)	wartość oczekiwana (średnia) $\mu_{\ln x}$ wariancja $\sigma_{\ln x}^2$	ciągła $x > 0$
Chi-kwadrat	liczba stopni swobody	ciągła $x \geq 0$
Dwumianowy	liczność próbki n prawdopodobieństwo p	dyskretna $x = 0, 1, 2, \dots, n$
Poissona	lammda	dyskretna $x = 0, 1, 2, \dots, n$
Geometryczny	prawdopodobieństwo p	dyskretna $x = 1, 2, \dots$
Bernoulliego (zero – jedynkowy)	prawdopodobieństwo p	dyskretna $x = 0$ lub $x = 1$

Po wybraniu jednego z wyżej podanych rozkładów teoretycznych otwiera się kolejne okno, które przedstawiono na rys. 3.2.



Rys. 3.2. Okno Dopasowanie rozkładu po wskazaniu rozkładu teoretycznego

Po naciśnięciu przycisku **Zmienna**, wskazuje się zmienną dla której ma być dopasowany rozkład teoretyczny. Na podstawie danych z próby estymowane są parametry wybranego rozkładu teoretycznego i wyświetlane jako wartości domyślne. Opracowujący wyniki ma jednak możliwość ich zmiany i wprowadzania własnych wartości. Można też wybrać inny rozkład teoretyczny w polu **Rozkład**, dla tego samego typu zmiennej (ciągła lub dyskretna). Program oblicza domyślnie wartość statystyki chi–kwadrat, a statystykę Kołmogorowa w zależności od zaznaczonej opcji w polu **Test Kołmogorowa – Smirnowa**. Możliwe są trzy ustawienia:

- **Nie** – nie jest obliczana statystyka Kołmogorowa;
- **Tak (skategoryzowany)** – statystyka Kołmogorowa obliczana jest w oparciu o dane skategoryzowane, jest to szybsza metoda obliczeń i jest zalecana dla dużej liczby danych, wymaga jednak wcześniejszej kategoryzacji zebranych danych;
- **Tak (ciągły)** – statystyka Kołmogorowa obliczana jest w oparciu o dane surowe, nie są one wcześniej kategoryzowane.

Wyniki obliczeń mogą być przedstawione w postaci tabeli liczebności, po naciśnięciu przycisku **OK** lub w postaci histogramu po naciśnięciu przycisku **Wykres**. Sposób obliczenia liczebności będzie zależał od ustawienia wartości w polach: **Liczba kategorii**, **Dolna granica** i **Górna granica**. Program wyświetla domyślne wartości liczby klas oraz dolną i górną granicę klas skrajnych określonych na podstawie wprowadzonych danych. Prowadzący obliczenia może zmieniać te wartości wpływając tym samym na postać tabeli liczebności i histogramu. Sposób przedstawienia histogramu będzie zależał od ustawień opcji w polu **Wykres**:

- **Rozkład liczebności** – histogram przedstawiany jest w postaci histogramu liczebności lub histogramu częstości, zależnie od wybranej opcji w polu **Wykres liczebności lub %**;
- **Dystrybuanta** – histogram przedstawiany jest w postaci histogramu skumulowanej liczebności lub histogramu skumulowanej częstości, zależnie od wybranej opcji w polu **Wykres liczebności lub %**;
- **Liczebności** – histogram przedstawiany jest w postaci histogramu liczebności lub histogramu skumulowanej liczebności, zależnie od opcji wybranej w polu **Wykres rozkładu**;
- **Częstości %** – histogram przedstawiany jest w postaci histogramu częstości lub histogramu skumulowanej częstości, zależnie od wybranej opcji w polu **Wykres rozkładu**.

### *Przykład 3.1*

W wyniku badań eksploatacyjnych narzynek uzyskano następujące okresy trwałości: 4,0; 4,8; 4,8; 4,9; 5,0; 6,4; 8,3; 18,5; 18,5; 19,2; 8,6; 8,9; 9,2; 9,7; 10,0; 12,4; 13,4; 14,2; 14,6; 14,6; 22,4; 24,2; 25,2; 25,5; 16,52; 16,81; 17,30 17,8; 25,6; 25,7; 28,0; 31,0 31,3; 43,8; 47,0; 46,0; 36,0; 45,0; 32,0; 29,0; 27,0; 26,0; 24,0; 23,0; 19,0; 18,0; 17,5; 16,0; 15,0; 14,0; 14,5; 13,0; 12,0; 11,0; 10,4; 10,2; 9,5; 9,0; 8,5; 8,0; 7,9; 7,5; 6,8; 6,2; 6,3; 4,8; 4,7; 4,85; 4,6; 4,5; 4,2; 3,5; 2,0; 7,5; 6,9; 9,5; 11,3; 13,4; 14,2; 15,7; 15,8; 13,5; 23,5; 20,0; 21,0; 37,0. 1,3; 1,6; 2,2; 2,4; 3,3; 3,6; 6,44; 6,5; 7,0; 7,7; 10,5; 10,6; 11,6; 12,4 min.

Na poziomie istotności  $\alpha = 0,05$  zweryfikować hipotezę, że rozkład okresów trwałości jest rozkładem gamma.

### Rozwiązanie

W zadaniu obserwowany jest okres trwałości narzynek, który jest zmienną losową. Należy sprawdzić hipotezę zerową

$H_0$ : zmienna  $X$  ma rozkład gamma,

przy czym nie są znane parametry hipotetycznego rozkładu. W celu weryfikacji hipotezy zastosujemy test chi-kwadrat i test Kołmogorowa.

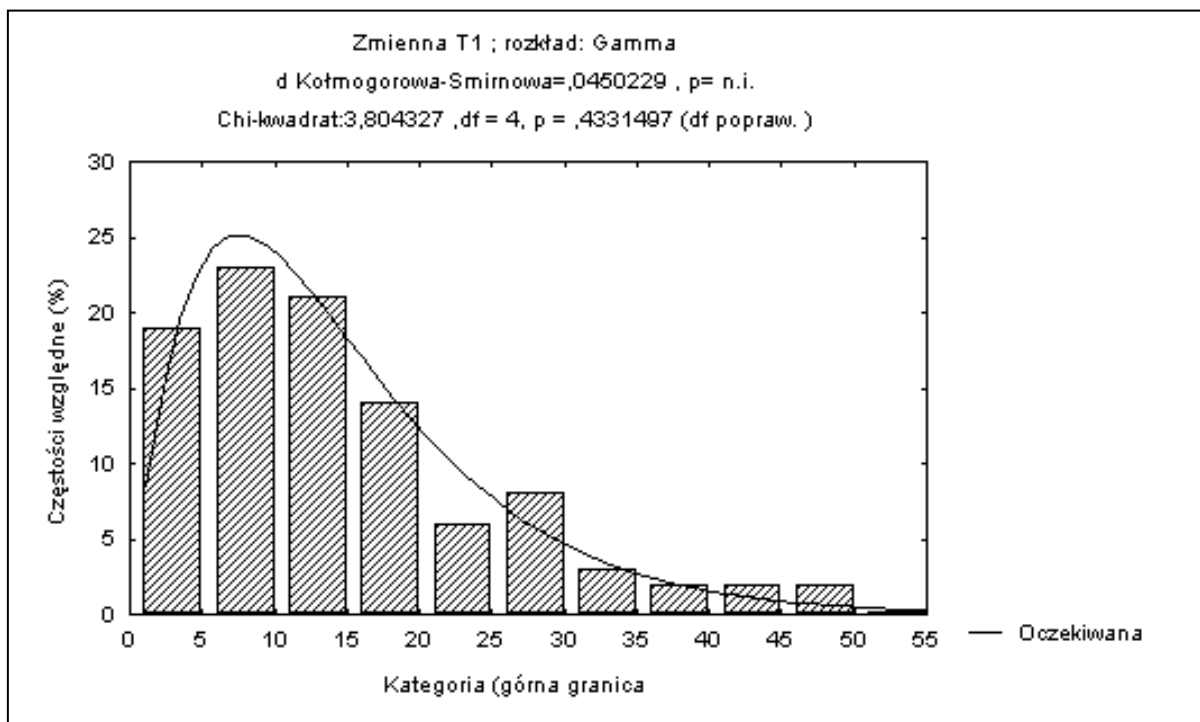
Dane z przykładu zostały zapisane w jednej kolumnie oznaczonej T1 w pliku narzynki.sta. W celu obliczenia wartości statystyk uruchamia się moduł **Statystyki nieparametryczne/ Rozkłady** i wybiera opcję **Dopasowanie rozkładów**. W otwartym oknie wskazuje się typ rozkładu **Gamma** i potwierdza wybór przyciskiem **OK**. Następnie naciska się przycisk **Zmienna** i wskazuje zmienną do analizy. Na podstawie wprowadzonych danych zostają oszacowane parametry hipotetycznego rozkładu gamma: parametr skali  $\beta = 7,1431002$  oraz parametr kształtu  $\alpha = 2,0524697$ . W ten sposób został dokładnie określony rozkład hipotetyczny i można przystąpić do weryfikacji hipotezy zerowej. W oknie **Test Kołmogorowa–Smirnowa** zaznacza się opcję **Tak(ciągły)** naciska przycisk **OK** i otrzymuje się wyniki w postaci tabeli liczebności z podanymi wartościami statystyk (tabela 3.2).

Tabela 3.2

Wyniki testu chi-kwadrat i Kołmogorowa

STAT.	Zmienna T1 / rozkład Gamma (narzynki.sta)								
STATYST	d Kołmogorowa – Smirnowa=,0450229, p n.i.								
NIEPAR.	Chi–kwadrat : 3,804327, df = 4, p =,4331497 (df popraw.)								
Górna granica	obserw. liczebność	skumul. obserw.	procent obserw.	skumul. % obserw.	oczekiw. liczebność	skumul. oczekiw.	procent oczekiw.	skumul. % oczekiw.	obsev. – oczekiw.
<= 5	19	19	19,00	19,00	14,49771	14,4977	14,49771	14,4977	4,50229
10,	23	42	23,00	42,00	24,69882	39,1965	24,69882	39,1965	-1,69882
15,	21	63	21,00	63,00	21,33746	60,5340	21,33746	60,5340	-,33746
20,	14	77	14,00	77,00	15,20356	75,7375	15,20356	75,7375	-1,20356
25,	6	83	6,00	83,00	9,87398	85,6115	9,87398	85,6115	-3,87398
30,	8	91	8,00	91,00	6,07133	91,6829	6,07133	91,6829	1,92867
35,	3	94	3,00	94,00	3,60067	95,2835	3,60067	95,2835	-,60067
40,	2	96	2,00	96,00	2,08131	97,3649	2,08131	97,3649	-,08131
45,	2	98	2,00	98,00	1,18023	98,5451	1,18023	98,5451	,81977
50,	2	100	2,00	100,00	,65938	99,2045	,65938	99,2045	1,34062
Niesk.	0	100	0,00	100,00	,79554	100,000	,79554	100,000	-,79554

Dla lepszego zobrazowania rozkładu można zaobserwowane dane przedstawić w postaci histogramu częstości z wykreśloną funkcją gęstości rozkładu gamma. W tym celu należy nacisnąć na przycisk **Wykres** i otrzyma się wykres przedstawiony na rys. 3.3.



Rys. 3.3. Funkcja gęstości rozkładu gamma z parametrami  $\alpha = 2,0524697$  i  $\beta = 7,1431002$  na tle histogramu częstości okresów trwałości narzynek

Otrzymana wartość statystyki  $\chi^2 = 3,804327$  nie przekracza wartości krytycznej rozkładu  $\chi^2$ , która dla 4 stopni swobody na poziomie istotności  $\alpha = 0,05$  wynosi  $\chi^2_{\alpha, 4} = 9,488$  (tablica III), nie ma więc podstaw do odrzucenia hipotezy zerowej. Ten sam wniosek można wyciągnąć na podstawie obliczonej wartości poziomu istotności  $p = 0,4331497$  dla uzyskanej wartości statystyki  $\chi^2$ , która to wartość  $p$  jest większa od przyjętego poziomu istotności  $\alpha$ , a więc nie ma podstaw do odrzucenia hipotezy  $H_0$ .

Podobnie wartość statystyki testowej  $D_n = 0,0450229$  testu Kołmogorowa jest mniejsza od wartości krytycznej testu Kołmogorowa, która dla próby o liczebności  $n = 100$  wynosi  $D_{\alpha, n} = 0,136$  (tablica IV), a więc ten test też wskazuje, że nie ma podstaw do odrzucenia hipotezy zerowej. Prawdopodobieństwo  $p$  nie odrzucenia hipotezy zerowej określone jest jako n.i., co oznacza nie istotny, a więc wartość statystyki  $D_n$  jest nieistotna. Wnioskiem końcowym może być stwierdzenie, że rozkład okresu trwałości narzynek może być opisany rozkładem gamma o wyżej obliczonych parametrach.

Należy tu podkreślić, że gdyby postawiono hipotezę zerową, że badana zmienna ma rozkład logarytmo-normalny lub dowolny inny z możliwych do wyboru, to końcowy wniosek równie dobrze mógłby być taki sam; nie odrzucać hipotezy zerowej. Wynika to z faktu, że test zgodności nie służy do rozróżniania dwóch lub więcej rozkładów lub pomocy w wyborze najlepszego rozkładu spośród zaproponowanych. Celem tego testu jest tylko odpowiedź na pytanie; Czy można przyjąć rozkład, który skłonny jestem uznać za właściwy lub czy dane są na tyle niezgodne z proponowanym rozkładem, że muszę go odrzucić ?

### 3.3.2. Testy normalności

Testy normalności znajdują się w procedurze *Tabele liczebności* modułu *Podstawowe statystyki*, której okno zostało szczegółowo opisane w rozdziale 1.4.

#### Przykład 3.2

Zmierzono średnicę 10., wylosowanych z dostarczonej partii, rur i otrzymano następujące wyniki: 50,0; 50,2; 50,1; 49,8; 50,3; 50,3; 50,1; 50,0; 50,2; 49,9 mm. Sprawdzić na poziomie istotności  $\alpha = 0,05$ , czy można średnicę rur traktować jako zmienną o rozkładzie normalnym.

#### Rozwiązanie

Stawiamy hipotezę zerową,  $H_0$ : X ma rozkład  $N(\mu, \sigma)$  przy czym nie znamy parametrów hipotetycznego rozkładu oraz dysponujemy próbką o małej liczności  $n = 10$ . W celu weryfikacji hipotezy można zastosować jeden z testów normalności: Lillieforsa lub Shapiro–Wilka. Pokażemy jak przeprowadza się weryfikację hipotezy zerowej każdym z tych testów.

Uruchamia się procedurę *Tabele liczebności* i w polu **Testy normalności** zaznacza się opcję **test Lillieforsa nieznaną średnią / od. std.** i **test W Shapiro-Wilka**. Po naciśnięciu przycisku **OK**, otrzymuje się wyniki przedstawione w tabeli 3.3.

Tabela 3.3

Wyniki testu Lillieforsa i Shapiro–Wilka

STAT. PODST. STATYST.	Test k – S, prawdop. Lillieforsa (Średnia i odch. std. wyznaczone z danych)			STAT. PODST. STATYST.	Test W Shapiro – Wilka (Średnia i odch. std. wyznaczone z danych)		
Zmienne	N	maks D	p	Zmienne	N	W	p
śred_rur	10	,145790	p >,20	śred_rur	10	,950668	,660971

Otrzymano wartości statystyk testowych maks D oraz W dla testów Lillieforsa i Shapiro–Wilka. Mamy też podane prawdopodobieństwa p nie odrzucania hipotezy zerowej, gdy jest ona prawdziwa. Ponieważ, wartości tych prawdopodobieństw są wysokie (odpowiednio, co najmniej 0,20 i 0,660971) i przekraczają wartość przyjętego poziomu istotności  $\alpha = 0,05$ , nie ma podstaw do odrzucenia hipotezy zerowej. Zatem rozkład średnic rur może być traktowany jako rozkład normalny o parametrach wyestymowanych na podstawie danych  $\mu = 50,090$  i  $\sigma = 0,16633$ .

## 4. REGRESJA LINIOWA

### 4.1. Cel ćwiczenia

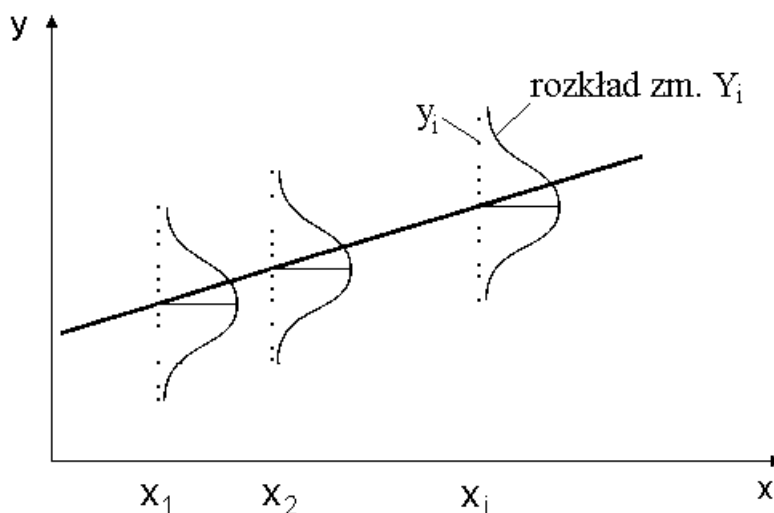
Celem ćwiczenia jest zapoznanie się ze sposobami określania zależności między zmienną zależną i jedną lub wieloma zmiennymi niezależnymi.

### 4.2. Wprowadzenie

W najprostszym przypadku model regresji określa liniową zależność funkcyjną wartości oczekiwanej (średniej) zmiennej losowej  $Y$  od nielosowej zmiennej  $x$ , która może zmieniać się z doświadczenia na doświadczenie. Zależność tę zapisuje się wzorem:

$$E(Y|x) = \alpha + \beta x. \quad (4.1)$$

Powyższe równanie wyznacza linię regresji populacji generalnej (rys. 4.1). Przyjmuje się, że przechodzi ona przez punkty, będące wartościami wielu niezależnych doświadczeń, których wyniki  $y_1, y_2, \dots, y_i$  reprezentują realizacje zmiennych losowych  $Y_1, Y_2, \dots, Y_i$ .



Rys. 4.1. Linia regresji populacji generalnej

W każdym oddzielnym doświadczeniu zmienna niezależna  $x$  przyjmuje pewną wartość  $x_i$ , a zmienna losowa zależna  $Y_i$  wartość  $y_i$ . Można więc, zależność regresyjną przedstawić w postaci:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad (4.2)$$



gdzie  $\varepsilon_i$  jest składnikiem losowym o wartości oczekiwanej zero, nazywanym też odchyleniem losowym. Oprócz tego zakłada się, że wszystkie wyniki doświadczeń  $y_i$  są niezależne oraz podlegają rozkładowi normalnemu o pewnej nieznannej wartości oczekiwanej  $E(Y)$  i pewnej wariancji  $\sigma^2$  niezależnej od wartości zmiennej niezależnej  $x$ .

W innej klasie zagadnień, zainteresowanie skupia się na przewidywaniu wartości jednej zmiennej losowej  $Y$ , na podstawie obserwacji innej zmiennej losowej  $X$ . Przy liniowym modelu zależności, warunkowa wartość oczekiwana zmiennej  $Y$  pod warunkiem, że  $X = x$  jest funkcją liniową  $x$ :

$$E(Y | X = x) = \alpha + \beta x. \quad (4.3)$$

Dla pewnej wartości  $x$  zmiennej losowej  $X$ , zmienna losowa  $Y$  ma pewien warunkowy rozkład wokół wartości oczekiwanej i pewną warunkową wariancję  $\sigma_y^2$ . W wielu analizach statystycznych zakłada się, że ten warunkowy rozkład jest normalny o wariancji zależnej od konkretnej wartości  $x$ . Jeśli zmienne  $X$  i  $Y$  podlegają pewnemu dwuwymiarowemu rozkładowi normalnemu, to:

$$E(Y | X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x). \quad (4.4)$$

Łatwo zauważyć, że:

$$\beta = \rho (\sigma_y / \sigma_x) \quad \text{oraz} \quad \alpha = \mu_y - \beta \mu_x,$$

gdzie:

- $\mu_x$  i  $\mu_y$  – brzegowe wartości oczekiwane zmiennych losowych  $X$  i  $Y$ ,
- $\sigma_x$  i  $\sigma_y$  – brzegowe odchylenia standardowe zmiennych losowych  $X$  i  $Y$ ,
- $\rho$  – współczynnik korelacji między zmiennymi losowymi  $X$  i  $Y$ .

Zauważmy, że jeżeli przyjmiemy, że dla każdego  $x_i$  zmienne  $Y_i$  są zmiennymi losowymi o rozkładzie normalnym, o wartościach oczekiwanych  $\alpha + \beta x_i$  i stałej wariancji  $\sigma^2$ , to wówczas można także założyć, że odchylenia losowe  $\varepsilon_i$  mają identyczny rozkład normalny  $N(0, \sigma)$ .

Z tego, że  $X$  i  $Y$  są zmiennymi losowymi wynika, że można rozpatrywać alternatywny problem prognozy  $E(X | Y = y) = \alpha' + \beta' y$ . Czy jest to celowe zależy od tego w jaki sposób zależność będzie wykorzystywana w praktyce. Czy będzie się prognozować wartość zmiennej  $X$  mając dane wartości  $Y = y$ , czy też odwrotnie trzeba będzie prognozować wartość zmiennej  $Y$  na podstawie  $X = x$ . Zależność zmiennej losowej  $X$  względem zmiennej losowej  $Y$  jest określona wzorem:

$$E(X | Y = y) = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y). \quad (4.5)$$

Równanie to, nie przedstawia tej samej prostej na wykresie, co zależność określająca  $E(Y|X=x)$ . Współczynnik nachylenia do osi  $x$ -ów, w tym przypadku, jest równy  $(1/\rho) * (\sigma_y / \sigma_x)$ , a w poprzednio był  $\rho * (\sigma_y / \sigma_x)$ .

Różnica w obu typach modeli zależności polega na interpretacji  $x$ ; czy przedstawia ona sobą, z góry określoną wartość deterministyczną, czy też reprezentuje wartość obserwowaną zmiennej losowej  $X$ . W każdym przypadku podstawowym założeniem modelu jest fakt, że dla

danych liczb  $x$  wartość oczekiwana  $E(Y)$  jest liniowa względem  $x$ , a więc równa  $\alpha + \beta x$ . W obu przypadkach może być stosowana do estymacji współczynników zależności analiza regresji, chociaż gdy  $X$  i  $Y$  są zmiennymi losowymi, estymacja pięciu parametrów  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ ,  $\sigma_y$  i  $\rho$  również jest właściwa, i taka procedura nazywana jest statystyczną analizą korelacji.

Analiza regresji obejmuje zasadniczo dwie grupy zagadnień: estymację współczynników zależności i weryfikację uzyskanej zależności. W najprostszym modelu zależności liniowej estymacji podlegają trzy parametry: współczynnik  $\alpha$ , współczynnik kierunkowy  $\beta$  i wariancja  $\sigma^2$ . Natomiast przy weryfikacji zależności podstawowe pytanie brzmi, czy dane wskazują na istotną zależność średniej zmiennej losowej  $Y$  od zmiennej  $x$ ? Czyli sprawdza się hipotezę zerową  $H_0: \beta = 0$ . Jeżeli zależność okaże się nieistotna, to model może być uproszczony przez pominięcie zmiennej  $x$  i traktowanie zmiennej  $Y$  jako zwykłej zmiennej losowej. Podobnie może być sprawdzona hipoteza o nieistotności wyrazu wolnego  $\alpha = 0$ .

Naturalnie przy większej liczbie zmiennych niezależnych, model zależności może być przedstawiony w postaci:

$$E(Y | x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k. \quad (4.6)$$

Wówczas metoda analizy statystycznej takiego modelu nosi nazwę analizy regresji wielokrotnej.

Kiedy w analizie regresji mówi się, że model jest liniowy bądź nieliniowy, odnosi się to do liniowości lub nieliniowości względem współczynników. Wartość najwyższej potęgi zmiennej niezależnej modelu nazywa się stopniem modelu. Na przykład, zakładając że związek między zmienną zależną  $Y$  a zmienną niezależną  $x$  jest w postaci modelu nieliniowego multiplikatywnego:

$$Y = \alpha x^\beta \varepsilon, \quad (4.7)$$

gdzie  $\varepsilon$  jest odchyleniem losowym. Wówczas poprzez logarytmowanie można przekształcić ten model do postaci liniowej:

$$\ln Y = \ln \alpha + \beta \ln x + \ln \varepsilon, \quad (4.8)$$

i analizować go przy użyciu metod regresji liniowej.

Podobnie model wykładniczy:

$$Y = \exp(\alpha + \beta x) \text{ przekształca się do } \ln Y = \alpha + \beta x, \quad (4.9)$$

i model odwrotny:

$$Y = 1/(\alpha + \beta x) \text{ przekształca się do } 1/Y = \alpha + \beta x. \quad (4.10)$$

### 4.3. Regresja jednej zmiennej niezależnej

#### 4.3.1. Opis metody

Powszechnie stosowaną metodą estymacji współczynników  $\alpha$  i  $\beta$  w modelu regresji liniowej jest metoda najmniejszych kwadratów. Niech  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  będzie ciągiem wyników obserwacji. Zgodnie z metodą najmniejszych kwadratów oszacowania  $a$  i  $b$  współczynników  $\alpha$  i  $\beta$  minimalizują sumę kwadratów odchyłeń obserwacji od prostej regresji, określoną w następujący sposób:

$$S = \sum_{i=1}^n [y_i - (a + b x_i)]^2 . \quad (4.11)$$

Estymatory  $a$  i  $b$  współczynników regresji  $\alpha$  i  $\beta$  otrzymane metodą najmniejszych kwadratów są określone wzorami:

$$a = \bar{y} - b \bar{x} , \quad (4.12)$$

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} , \quad (4.13)$$

gdzie:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i , \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i , \quad (4.14)$$

są odpowiednio średnimi arytmetycznymi wyników obserwacji  $x_i$  i  $y_i$ .

Sprawdzenie czy zależność między zmienną zależną  $Y$  a zmienną niezależną  $x$  jest istotna statystycznie polega na weryfikację hipotezy  $H_0: \beta = 0$  o nieistotności zależności wobec hipotezy alternatywnej  $H_1: \beta \neq 0$  w oparciu o statystykę:

$$t = b/s_b , \quad (4.15)$$

gdzie  $s_b$  jest odchyleniem standardowym estymatora  $b$  współczynnika  $\beta$  opisanym wzorem:

$$s_b = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2} , \quad (4.16)$$

gdzie  $s^2$  jest oszacowaniem wariancji  $\sigma^2$  zmiennej losowej  $Y$ .

Zwykle, jako tego oszacowania, używa się sumy kwadratów odchyłeń obserwowanych wartości  $y_i$  od wartości estymowanych  $\hat{y}_i$  podzielonych przez liczbę stopni swobody:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 . \quad (4.17)$$

Statystyka  $t$ , przy założeniu prawdziwości hipotezy  $H_0$ , ma rozkład t-Studenta o  $f = n - 2$  stopniach swobody. Hipotezę  $H_0$  odrzuca się, jeśli wartość  $|t|$  przekracza wartość krytyczną  $t_{\alpha/2, f}$ , przy zadanym poziomie istotności  $\alpha$ .

Weryfikację hipotezy  $H_0: \beta = 0$  można też przeprowadzić w oparciu o analizę wariancji zależności regresyjnej. Przyjmując następujące oznaczenia:

$$SG = \sum_{i=1}^n (y_i - \bar{y})^2 \text{ – suma kwadratów poza średnią (zmiennosc całkowita),}$$

$SM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  – suma kwadratów w regresji (zmiennosc wynikajaca z przyjętego modelu),

$$SR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ – suma kwadratów poza regresją (zmiennosc resztowa),}$$

można udowodnić następującą tożsamość:

$$SG = SM + SR . \quad (4.18)$$

Równanie to pokazuje, że wśród zmienności y-ków względem ich średniej wartości część zmienności może być przypisana linii regresji, a część faktowi, że nie wszystkie obserwacje leżą na linii regresji, gdyby bowiem wszystkie leżały, to suma kwadratów poza regresją byłaby równa zero. Z powyższego wynika, że ustalenie jak dalece linia regresji będzie przydatna do prognozowania, sprowadza się do stwierdzenia, jak duża część sumy SG zawarta jest w sumie SM, a jak duża w sumie SR. Będziemy zadowoleni jeśli SM będzie dużo większe od SR, lub co na jedno wychodzi, stosunek:

$$R^2 = SM/SG , \quad (4.19)$$

nie odbiega zbyt wiele od jedności. Stosunek  $R^2$  jest w tym przypadku kwadratem współczynnika korelacji z próby R pomiędzy zmiennymi x i Y:

$$R = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} . \quad (4.20)$$

Z definicji wynika, że współczynnik  $R^2$  może być traktowany jako miara stopnia dopasowania prostej regresji do danych doświadczalnych. Jego wartość należy do przedziału domkniętego  $[0, 1]$ . Gdy  $R^2 = 1$  to przewidywanie jest idealne. Można też wykazać, że jeśli  $\beta = 0$  to statystyka:

$$F = \frac{(n-2) SM}{SR} = (n-2) \frac{R^2}{1-R^2} , \quad (4.21)$$

ma rozkład F-Snedecora z 1 stopniem swobody dla licznika i  $n - 2$  stopniami swobody dla mianownika. Hipotezę  $H_0$  odrzuca się, gdy obliczona wartość statystyki F przekracza wartość krytyczną  $F_{\alpha, 1, n-2}$ , przy zadanym poziomie istotności  $\alpha$ . Dla lepszego zobrazowania zależności między wielkościami omawianymi powyżej, przedstawia się je w postaci tabeli analizy wariancji (tabela 4.1). W tym przypadku przy doborze linii prostej, test F jest dokładnie tym samym, czym test t dla  $\beta = 0$ , podany wcześniej.

Tabela analizy wariancji dla zależności regresyjnej

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średni kwadrat	Wartość F
W modelu regresji	SM	1	SM/1	F = SM/ s <sup>2</sup>
Poza regresją (reszta)	SR	n - 2	s <sup>2</sup> = SR/(n - 2)	
Łącznie (względem średniej)	SG	n - 1		

Ponieważ w wyniku estymacji współczynników zależności regresyjnej, otrzymuje się ocenę punktową  $\hat{Y}$ , celowym jest wyznaczyć przedział ufności dla oczekiwanej wartości  $E(Y)$  przy określonej wartości  $x_0$ , który wyrażony jest wzorem:

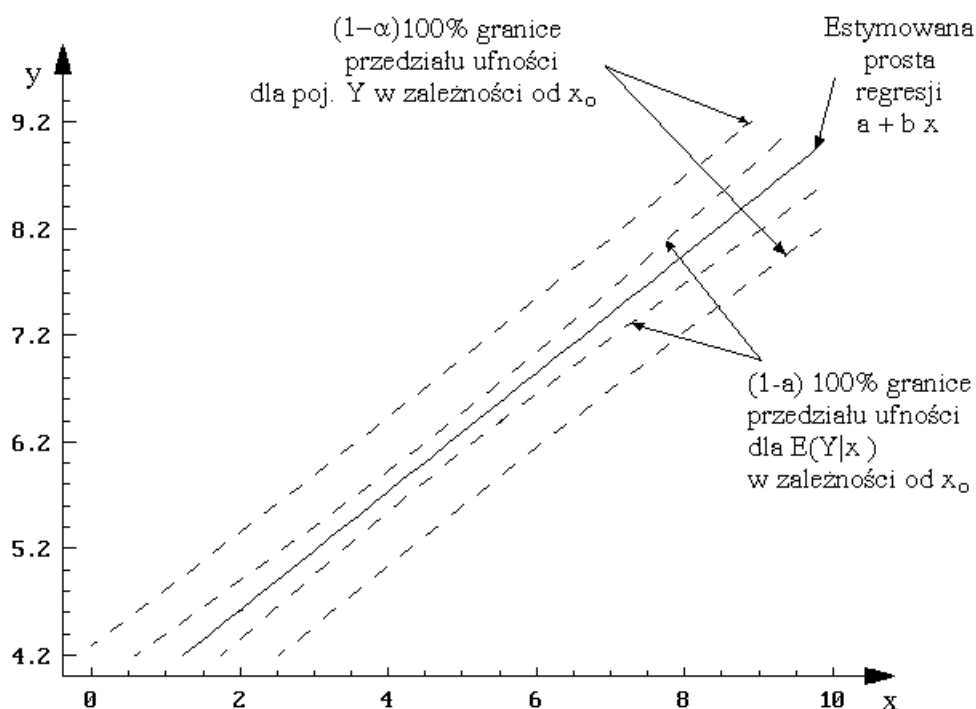
$$\hat{Y}_0 \pm t_{\alpha/2, f} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} s. \quad (4.22)$$

Przedział ten jest najmniejszy, jeśli  $x_0 = \bar{x}$  i zwiększa się przy oddalaniu się  $x_0$  od  $\bar{x}$  w dowolnym kierunku. Więc im większa jest odległość (w dowolną stronę)  $x_0$  od  $\bar{x}$ , tym większy jest przedział ufności dla wartości oczekiwanej, czyli mniejsza jest precyzja prognozy.

Skoro pojedyncza wartość obserwowana  $Y$  może zmieniać się, wokół prawdziwej wartości oczekiwanej, z wariancją  $\sigma^2$ , to przedział ufności dla pojedynczej obserwacji będzie określony wzorem:

$$\hat{Y}_0 \pm t_{\alpha/2, f} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} s. \quad (4.23)$$

Przedział ten jest oczywiście szerszy od przedziału dla wartości oczekiwanej  $E(Y|x_0)$  dla danego  $x_0$ , ponieważ jest to przedział, w którym należy się spodziewać  $(1 - \alpha)$  100% przyszłych obserwacji zmiennej losowej  $Y$  w punkcie  $x_0$ . Na rysunku 4.2 przedstawiono przykładowo granice przedziału ufności dla wartości oczekiwanej  $E(Y|x)$  oraz granice przedziału ufności dla pojedynczej wartości zmiennej losowej  $Y$ .



Rys. 4.2. Estymowana prosta regresji i granice przedziałów ufności

### 4.3.2. Obliczenia programem STATISTICA

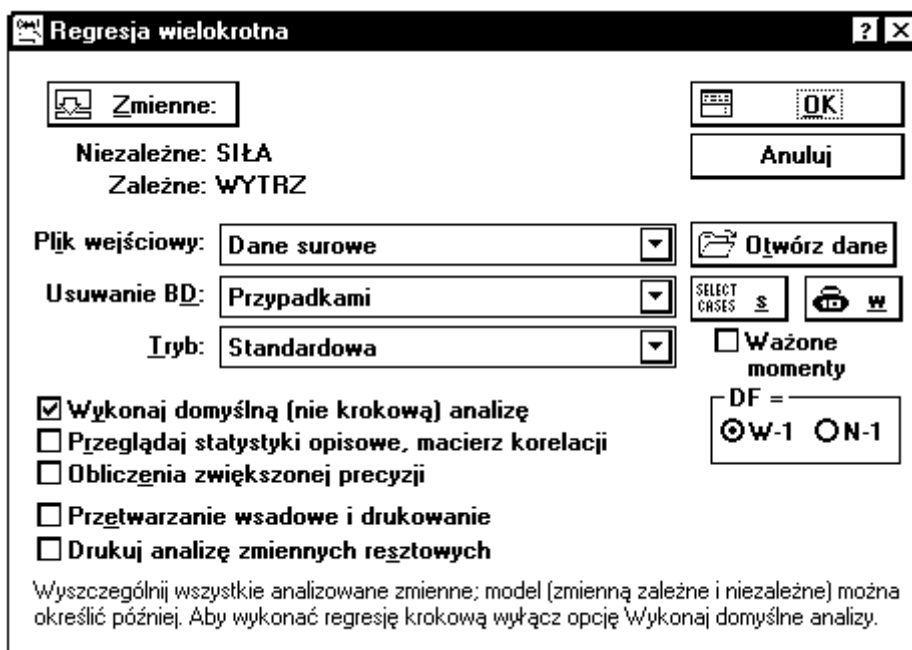
W programie STATISTICA do estymacji zależności regresyjnych służy moduł **Regresja wielokrotna**. Po wybraniu tego modułu otwiera się okno **Regresja wielokrotna** przedstawione na rys.4.3.

Wyboru zmiennych niezależnych i zależnych dokonuje się po naciśnięciu przycisku **Zmienne**. W polu **Plik wejściowy** określa się sposób przygotowania danych w arkuszu danych. Możliwe są dwa ustawienia:

- **Dane surowe** – dane w arkuszu danych przedstawione są w postaci danych surowych;
- **Macierz korelacji** – dane w arkuszu danych przedstawione są jako dane macierzowe.

W polu **Usuwanie BD** określa się sposób postępowania z brakującymi danymi. Pole jest aktywne tylko dla danych surowych. Możliwe są następujące ustawienia:

- **Przypadkami** – program będzie pomijał wszystkie przypadki (wiersze), w których brakuje wartości choćby dla jednej zmiennej spośród zmiennych wybranych do analizy;
- **Zastępowanie średnią** – brakujące dane będą zastępowane wartościami średnimi odpowiednich zmiennych.



Rys. 4.3. Okno Regresja wielokrotna

– **Parami** – program pomija tylko te przypadki, w których dla wybranej pary zmiennych brakuje jednej lub obydwu wartości.

W polu **Tryb** określa się sposób przygotowania zmiennych wybieranych do obliczeń:

– **Standardowa** – zmienne są wprowadzane do obliczeń w tej samej postaci, w jakiej występują w arkuszu danych;

– **Ustalona nieliniowa** – zmienne pierwotne określone w arkuszu danych mogą być różnie przekształcane. Dla każdej z wybranych zmiennych w pamięci komputera zostaną utworzone zmienne będące transformatami zmiennych pierwotnych. Nazwa nowej zmiennej będzie kombinacją numeru zmiennej pierwotnej i wybranej transformacji. Te nowe zmienne mogą być wprowadzone do obliczeń. Dostępne transformacje podane są w oknie **Nieliniowe składniki regresji**, które przedstawiono na rys.4.4.

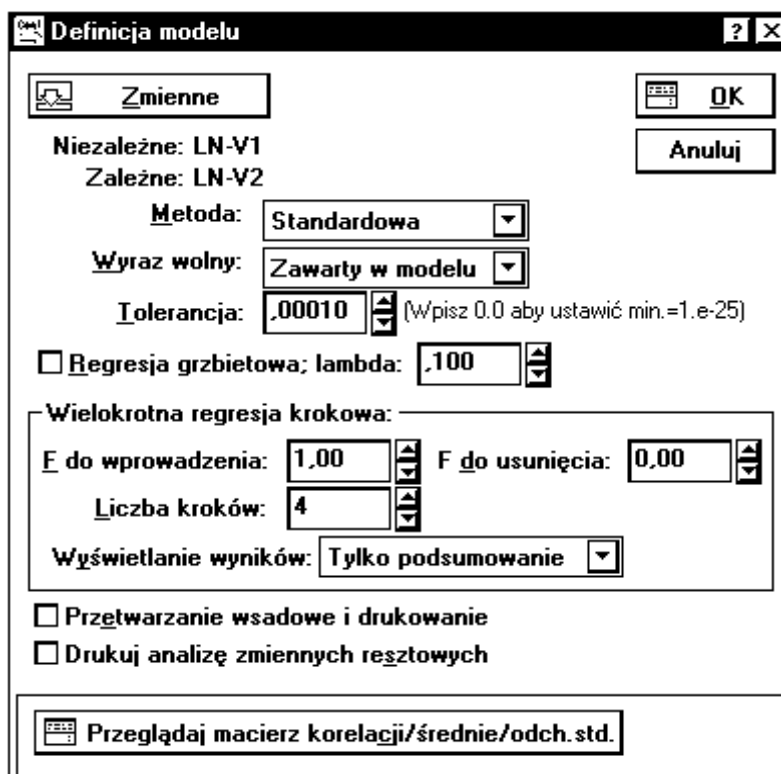
W przypadku, gdy wariancja zmiennej zależnej Y nie jest jednakowa w całym zakresie zmienności zmiennych niezależnych, do estymacji zależności regresyjnej powinna być stosowana metoda ważona najmniejszych kwadratów. W tym celu należy zaznaczyć opcję **Momenty ważne** i **N-1** oraz określić zmienną wagową. Ponadto w oknie **Regresja wielokrotna** można wybrać opcję:

– **Wykonaj domyślną (nie krokową) analizę** – włączenie tej opcji powoduje, że będzie obliczana zwykła, standardowa regresja z uwzględnieniem wyrazu wolnego; zostanie pominięte okno **Definicja modelu**, program wykona obliczenia i przedstawi wyniki w oknie **Wyniki regresji**. Jeżeli wyłączy się tę opcję, program przywoła okno **Definicja modelu**, w którym będzie można szczegółowo określić typ analizy regresji, jaka ma być wykonana;



Rys. 4.4. Okno Nieliniowe składniki regresji

- **Przeglądaj statystyki opisowe, macierze korelacji** – pozwala na obejrzenie szczegółowych statystyk opisowych dla wybranych zmiennych;
  - **Obliczenia zwiększonej precyzji** – należy włączyć tę opcję, gdy macierz korelacji ma być obliczana z podwyższoną dokładnością. Jest to korzystne, jeżeli dane cechują się tzw. małą wariancją względną (mały jest współczynnik zmienności  $s/\bar{x}$ ).
- Dokładne definiowanie analizy regresji odbywa się w oknie **Definicja modelu** (rys.4.5).



Rys. 4.5. Okno Definicja modelu



Zmienne niezależne<sup>1</sup> i zależne określa się po naciśnięciu przycisku **Zmienne**.

Jeżeli wskaże się więcej niż jedną zmienną zależną, obliczenia będą wykonane kolejno dla wszystkich zmiennych zależnych. W polu **Metoda** określa się rodzaj stosowanej analizy regresji, można wybrać:

– **Regresja standardowa** – zostanie wykonana analiza zależności regresyjnej zawierająca wszystkie wskazane zmienne niezależne;

– **Regresja krokowa postępująca** – zostanie wyestymowana zależność regresyjna metodą regresji krokowej z dołączaniem zmiennych. Zmienne niezależne będą wprowadzane do zależności lub z niej usuwane krok po kroku, aż otrzyma się zależność, zawierającą tylko te zmienne niezależne, które w sposób istotny wpływają na zmienną zależną;

– **Regresja krokowa wsteczna** – zależność regresyjna zostanie wyestymowana z zastosowaniem metody regresji krokowej z odrzucaniem zmiennych. Zmienne niezależne będą usuwane z zależności lub do niej wprowadzane krok po kroku, aż otrzyma się zależność zawierającą tylko te zmienne niezależne, które w sposób istotny wpływają na zmienną zależną.

W polu **Wyraz wolny** określa się, czy w zależności ma być uwzględniony wyraz wolny, czy też nie. Możliwe są dwa ustawienia:

– **Zawarty w modelu** – wyraz wolny ma wystąpić w zależności regresyjnej;

– **Równy zero** – zależność ma przechodzić przez początek układu współrzędnych, czyli jest to zależność bez wyrazu wolnego.

W polu **Tolerancja** podaje się najmniejszą tolerancję, jaka powinna być zaakceptowana w obliczeniach. Tolerancja określa wpływ danej zmiennej niezależnej na zmienną zależną, wobec wpływu pozostałych zmiennych, i jest zdefiniowana jako  $1-R^2$  (gdzie  $R^2$  jest to kwadrat współczynnika korelacji wielokrotnej tej zmiennej, z wszystkimi pozostałymi zmiennymi niezależnymi). Jeżeli tolerancja jest bliska zeru, to oznacza, że dana zmienna niezależna jest silnie skorelowana z innymi zmiennymi niezależnymi i nie powinna w zależności wystąpić. Jednocześnie to powoduje, że macierz ( $\mathbf{X}^T\mathbf{X}$ ) jest źle uwarunkowana (wyznacznik jest bliski zeru) i mogą występować duże błędy przy jej odwracaniu. W programie domyślnie ustawiona jest wartość tolerancji 0,001, można ją zmniejszyć nawet do wartości  $1,00 \cdot 10^{-25}$ , jednak należy się wówczas liczyć z tym, że współczynniki, otrzymane w tym przypadku, mogą być mało wiarygodne. Można poprawić stabilność ocen współczynników, stosując zamiast zwykłej metody regresji, metodę regresji grzbietowej. W tej metodzie do wartości diagonalnych macierzy korelacji dodaje się stałą wartość (lambda), a następnie macierz się standaryzuje, przywracając elementom na przekątnej wartość 1,0. Czyli w regresji grzbietowej, sztucznie zmniejsza się wartości współczynników korelacji tak, aby można było otrzymać bardziej stabilne oceny współczynników (lecz są one wówczas obciążone).

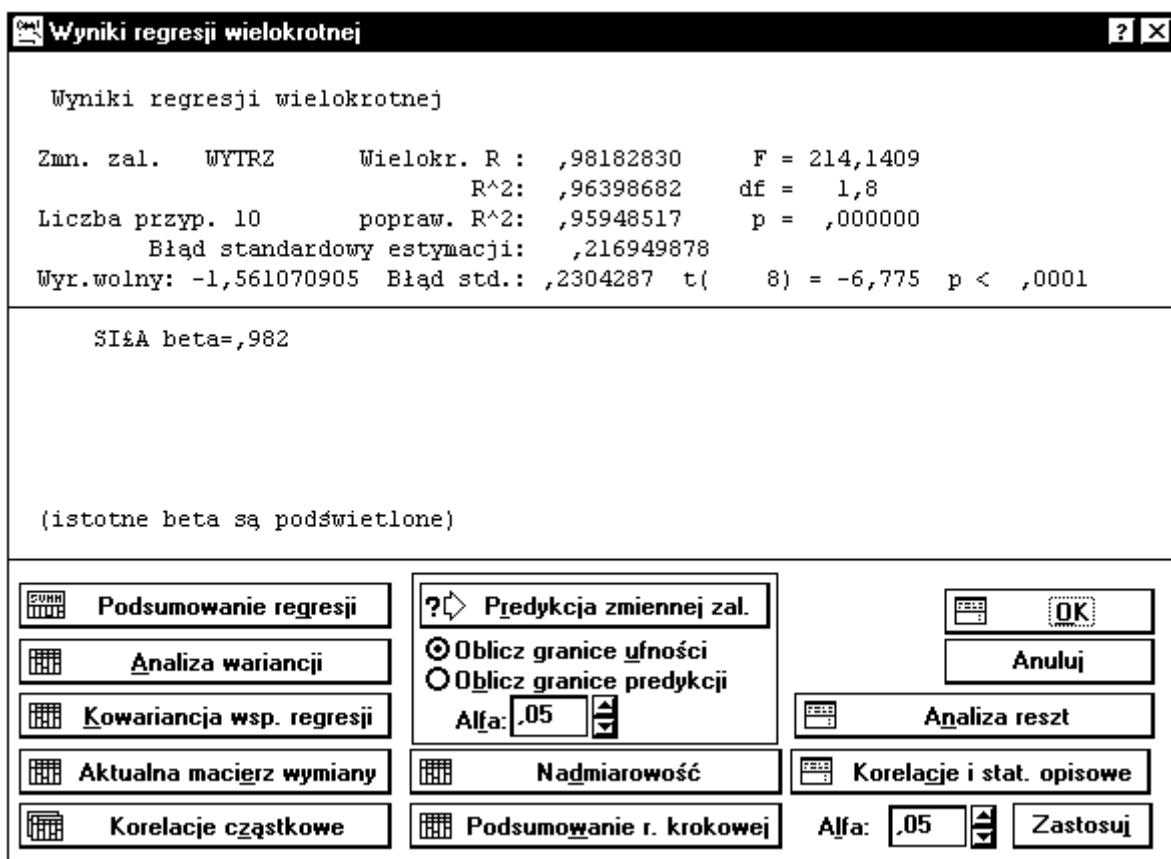
Jeżeli wybrano do analizy regresję krokową, staje się aktywne pole **Wielokrotna regresja krokowa** i możliwe jest określenie wartości progowych *F do wprowadzenia* i *F do usunięcia*. W regresji krokowej zmienna włączana jest do zależności, gdy wartość statystyki *F* dla tej zmiennej jest większa od wartości *F do wprowadzenia*. Analogicznie zmienna jest

---

<sup>1</sup> W analizie regresji przyjęto określać jako zmienną niezależną zarówno zmienną niezależną (pierwotną)  $x$  jak i dowolną funkcję określoną na zmiennych niezależnych  $f(\mathbf{x})$ .

usuwana z zależności, jeżeli odpowiadająca jej wartość  $F$  jest mniejsza od wartości  $F$  do usunięcia. Przy czym zawsze  $F$  do wprowadzenia musi być większe niż  $F$  do usunięcia. Obie wartości powinny być tak ustalone, aby w zależności znalazły się tylko istotne zmienne, i żadna istotna zmienna nie została pominięta. Opcja **Liczba kroków** pozwala na określenie maksymalnej liczby kroków, po wykonaniu których, zostaną wyświetlone wyniki regresji. Sposób wyświetlania wyników regresji krokowej określa się w polu **Wyświetlanie wyników**. Można wybrać:

- **Tylko podsumowanie** – podawane są wyniki końcowe regresji krokowej;
  - **W każdym kroku** – po wykonaniu każdego kroku analizy podawane są wyniki pośrednie.
- Po określeniu modelu, typu analizy regresyjnej i naciśnięciu przycisku **OK**, otrzymuje się okno **Wyniki regresji wielokrotnej** (rys. 4.6), w którym podane są sumaryczne wyniki analizy regresji oraz dostępne są opcje do przeglądania wyników szczegółowych.



Rys. 4.6. Okno Wyniki regresji wielokrotnej

Jako wyniki sumaryczne, podawane są następujące informacje: Zmn. zal. – nazwa zmiennej zależnej; Liczba przyp. – liczebność próby; Wielokr. R – współczynnik korelacji wielokrotnej (jest to pierwiastek z współczynnika wielokrotnej determinacji  $R^2$ );  $R^2$  – współczynnik wielokrotnej determinacji; popraw.  $R^2$  – skorygowany współczynnik wielokrotnej determinacji uwzględniający liczbę stopni swobody; F, df, p – wartość statystyki F, liczba stopni swobody tej statystyki oraz wartość poziomu istotności p do testowania hipotezy o nieistotności zależności regresyjnej; Błąd standardowy estymacji – oszacowanie

odchylenia standardowego zmiennej zależnej na podstawie wariancji zmiennej resztowej  $s^2$ ; Wyr. wolny – wyraz wolny, jeżeli wybrano zależność z wyrazem wolnym; Błąd std. – odchylenie standardowe wyrazu wolnego;  $t(df)$ ,  $p$  – wartość statystyki  $t$  (w nawiasie podana jest liczba stopni swobody) i wartość poziomu istotności  $p$  do testowania hipotezy o nie-istotności wyrazu wolnego.

Szczegółowe obejrzenie wyników umożliwiają następujące opcje:

**Podsumowanie regresji** – otrzymuje się tabelę zawierającą wyestymowane współczynniki zależności regresyjnej dla zmiennych niezależnych standaryzowanych (Beta) i niestandaryzowanych (B) ich odchylenia standardowe, wartości statystyki  $t$  i poziomy istotności  $p$ .

**Analiza wariancji** – otrzymuje się tabelę analizy wariancji dla zależności regresyjnej.

**Kowariancja wsp. regres.** – otrzymuje się macierze korelacji i kowariancji współczynników regresji.

**Aktualna macierz wymiany** – otrzymuje się tzw. macierz wymiany, która określona jest, jako macierz odwrotna macierzy korelacji zmiennych, będących aktualnie w zależności regresyjnej, pomnożona przez  $-1$ .

**Korelacje cząstkowe** – otrzymuje się tabelę z korelacjami cząstkowymi i semi-cząstkowymi. Korelacje cząstkowe są korelacjami między daną zmienną niezależną, a zmienną zależną, po uwzględnieniu dla obu zmiennych, wpływu ich skorelowania z wszystkimi pozostałymi zmiennymi. Korelacje semicząstkowe są korelacjami między daną zmienną niezależną, z uwzględnieniem jej skorelowania z wszystkimi pozostałymi zmiennymi niezależnymi, a zmienną zależną (bez uwzględnienia jej korelacji z innymi zmiennymi). Ponadto podawane są wartości: tolerancji,  $R^2$ , statystyki  $t$  i poziomu istotności  $p$ .

**Predykcja zmiennej zal.** – umożliwia obliczenie wartości prognozowanej zmiennej zależnej, dla zadanych wartości zmiennych niezależnych, oraz granice przedziału ufności (GU) dla wartości oczekiwanej zmiennej zależnej lub granice przedziału predykcji (GP – dla pojedynczej wartości przewidywanej), zależnie od włączonej opcji **Oblicz granice ufności** lub **Oblicz granice predykcji**. Odpowiedni poziom ufności  $1 - \alpha$  określa się polu **Alfa** wprowadzając wartość  $\alpha$ .

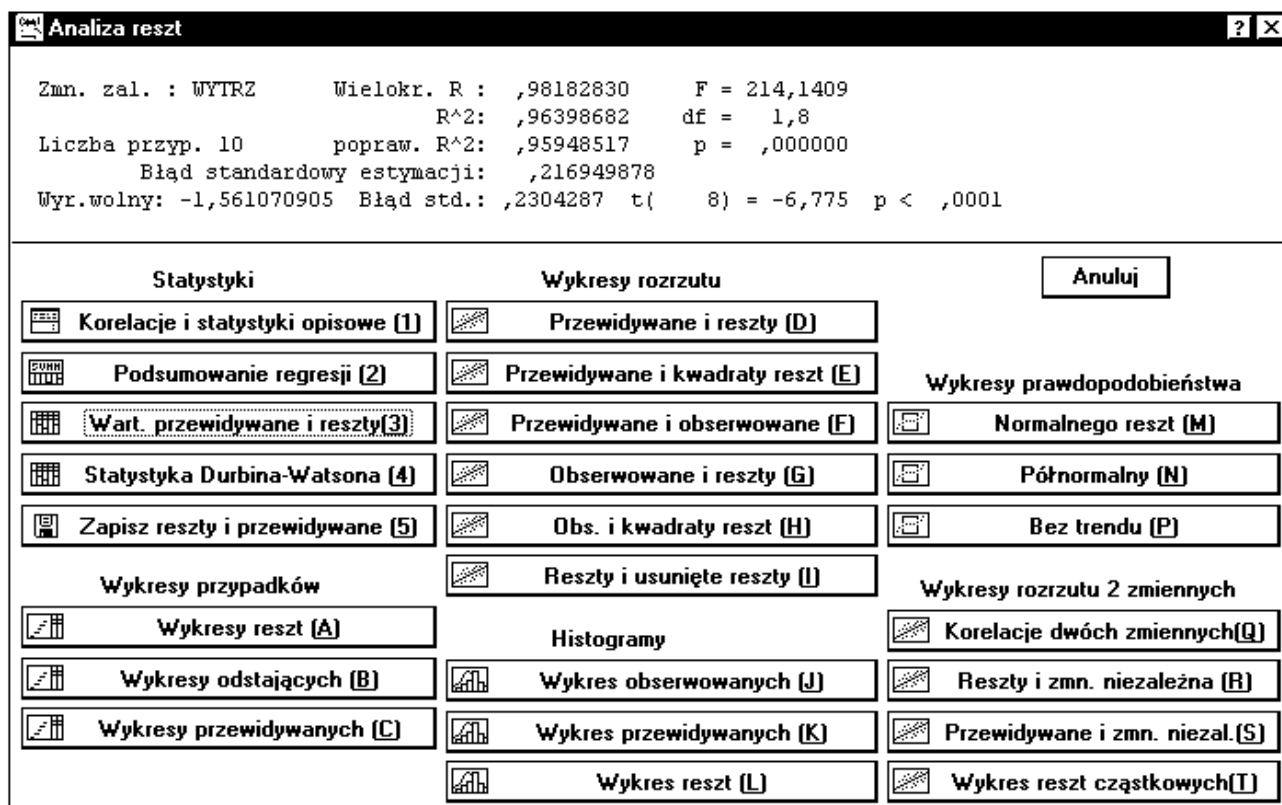
**Nadmiarowość** – podawane są różne wskaźniki nadmiarowości zmiennych niezależnych. Dla każdej zmiennej podawane są: tolerancja,  $R^2$ , korelacja cząstkowa, korelacja semicząstkowa.

**Podsumowanie r. krokowej** – podawane są podstawowe wyniki analizy regresji po każdym kroku.

**Korelacje i stat. opisowe** – umożliwia obejrzenie podstawowych statystyk opisowych dla wybranych zmiennych oraz macierzy korelacji i kowariancji.

– **Alfa** – jest to wartość poziomu istotności  $\alpha$ , wykorzystywanego do oceny istotności statystyk (jeśli wartość  $p$  jest mniejsza od  $\alpha$  to odpowiednie wyniki zostaną wyświetlone w kolorze czerwonym – dla zaznaczenia istotności efektu). Domyślna wartość  $\alpha$  wynosi 0,05. Można ją zmieniać w zakresie od 0,0001 do 0,5, i po naciśnięciu przycisku **Zastosuj**, ta zmiana będzie miała natychmiastowy wpływ na arkusze wyników.

**Analiza reszt** – umożliwia wybór różnych dodatkowych opcji analizy wyników i ich prezentacji graficznych (rys. 4.7).



Rys. 4.7. Okno Analiza reszt

W polu **Statystyki** dostępne są następujące opcje:

**Korelacje i statystyki opisowe (1)** – umożliwia obliczenie podstawowych statystyk opisowych dla wybranych zmiennych oraz macierzy korelacji i kowariancji.

**Podsumowanie regresji (2)** – otrzymuje się tabelę zawierającą wyestymowane współczynniki zależności regresyjnej dla zmiennych niezależnych standaryzowanych (Beta) i niestandardyzowanych (B), ich odchylenia standardowe, wartości statystyki t i poziomy istotności p.

**Wart. przewidywane i reszty (3)** – otrzymuje się tabele wyników, w której podane są wartości obserwowane, przewidywane, reszty, standaryzowane wartości przewidywane, standaryzowane wartości reszt, odchylenia standardowe wartości przewidywanej, odległości Mahalanobisa, usunięte wartości resztowe i odległości Cooka. Odległości Mahalanobisa, Cooka i usunięte wartości resztowe służą do oceny, czy dana obserwacja, może być zaliczona do obserwacji odstających. Odległość Mahalanobisa jest odległością danej obserwacji od centrum, określonego w wielowymiarowej przestrzeni zmiennych niezależnych. Odległość Cooka określa różnicę między obliczoną wartością współczynników (B), a tą samą wartością obliczoną po wyłączeniu danej obserwacji (przypadku) z obliczeń. Wszystkie odległości powinny być tego samego rzędu. Usuniętą wartością resztową jest reszta dla danej obserwacji

obliczona po wyłączeniu tej obserwacji z obliczeń. Jeżeli reszta usunięta znacznie się różni od standaryzowanej reszty, to można sądzić, że dany przypadek jest odstający.

**Statystyka Durбина-Watsona (4)** – statystyka Durбина-Watsona wykorzystywana jest do badania autokorelacji zmiennej resztowej, czyli do sprawdzenia, czy kolejne obserwacje są niezależne.

**Zapisz reszty i przewidywane (5)** – umożliwia zapisanie wyników analizy reszt.

W polu **Wykresy przypadków** można wybrać:

**Wykresy reszt (A)** – otrzymuje się wykresy wartości resztowych. Dostępne są następujące typy wartości resztowych: reszty surowe, reszty standaryzowane, odległości Mahalanobisa, reszty usunięte i odległości Cooka.

**Wykresy odstających (B)** – umożliwia analizę reszt dla przypadków, w których absolutna wartość standaryzowanej reszty jest większa od +2. Można wykreślić wszystkie standaryzowane reszty, które przekraczają przedział, lub w przypadku dużych zbiorów danych, wykreślanych jest 100 najbardziej odstających obserwacji.

**Wykresy przewidywanych (C)** – otrzymuje się wykres wartości przewidywanych lub standaryzowanych wartości przewidywanych.

W polu **Wykresy rozrzutu** można wybrać:

**Przewidywane i reszty (D)** – otrzymuje się wykres rozrzutu wartości reszt (oś Y) względem wartości przewidywanych (oś X). Na podstawie tego wykresu, można ocenić adekwatność zależności regresyjnej. Jeżeli zależność jest adekwatna, reszty powinny rozkładać się równomiernie wokół linii środkowej.

**Przewidywane i kwadraty reszt (E)** – otrzymuje się wykres rozrzutu kwadratów wartości reszt (Oś Y) względem wartości przewidywanych (oś X).

**Przewidywane i obserwowane (F)** – otrzymuje się wykres rozrzutu wartości obserwowanych (oś Y) względem wartości przewidywanych (oś X). Jest on szczególnie przydatny do identyfikowania skupień przypadków, które są źle przewidywane.

**Obserwowane i reszty (G)** – otrzymuje się wykres rozrzutu wartości reszt (oś Y) względem wartości obserwowanych (oś X). Na podstawie tego wykresu można łatwo identyfikować odstające obserwacje.

**Obs. i Kwadraty reszt (H)** – otrzymuje się wykres rozrzutu kwadratów wartości reszt (oś Y) względem wartości obserwowanych (oś X).

**Reszty i usunięte reszty (I)** – otrzymuje się wykres rozrzutu wartości reszt usuniętych (oś Y) względem wartości standaryzowanych reszt (oś X). Na podstawie tego wykresu można łatwo identyfikować odstające obserwacje

W polu **Histogramy** można wybrać:

**Wykres obserwowanych (J)** – histogram liczebności wartości obserwowanych zmiennej zależnej z zaznaczoną krzywą rozkładu normalnego.

**Wykres Przewidywanych (K)** – histogram liczebności wartości przewidywanych lub standaryzowanych przewidywanych z zaznaczoną krzywą rozkładu normalnego.

**Wykres reszt (L)** – histogram liczebności wartości resztowych z zaznaczoną krzywą rozkładu normalnego. Można wykonać wykres dla następujących typów wartości resztowych:

surowe reszty, standaryzowane reszty, odległości Mahalanobisa, usunięte reszty i odległości Cooka.

W polu **Wykresy prawdopodobieństwa** można wybrać:

**Normalnego reszt (M)** – otrzymuje się wykres prawdopodobieństwa normalnego, który pozwala na wizualną ocenę zgodności rozkładu wartości resztowych z rozkładem normalnym. W analizie regresji zakłada się, że zależność jest adekwatna oraz że odchylenia losowe mają rozkład normalny. Niespełnienie tych założeń ujawni się na wykresie w postaci wyraźnych odchyłeń reszt od linii prostej.

**Półnormalny (N)** – otrzymuje się półnormalny wykres prawdopodobieństwa; tworzony jest on identycznie, jak wykres prawdopodobieństwa normalnego, z tą różnicą, że na osi Y przedstawiona jest tylko dodatnia część krzywej rozkładu normalnego.

**Bez Trendu (P)** – otrzymuje się wykres prawdopodobieństwa normalnego z eliminacją trendu, tworzony jest identycznie jak wykres prawdopodobieństwa normalnego, z tą różnicą, że odjęty jest trend liniowy.

W polu **Wykresy rozrzutu 2 zmiennych** można wybrać:

**Korelacje dwóch zmiennych (Q)** – otrzymuje się wykres rozrzutu dwóch dowolnych zmiennych.

**Reszty i zmn. niezależne (R)** – otrzymuje się wykres rozrzutu wartości resztowych względem dowolnej wybranej zmiennej niezależnej. Można wykonać wykres dla następujących typów wartości resztowych: surowe reszty, standaryzowane reszty, odległości Mahalanobisa, usunięte reszty i odległości Cooka.

**Przewidywane i zmn. niezal. (S)** – otrzymuje się wykres rozrzutu wartości przewidywanych lub standaryzowanych przewidywanych (oś Y) względem dowolnej wybranej zmiennej niezależnej (oś X).

**Wykres reszt cząstkowych (T)** – otrzymuje się wykres reszt cząstkowych dla dowolnej zmiennej będącej w zależności regresyjnej. Na osi Y jest reszta plus wkład danej zmiennej do zależności regresyjnej ( $\text{reszta} + b_i * f_i(x)$ ), a na osi X wybrana zmienna niezależna.

#### Przykład 4.1

W celu ustalenia zależności pomiędzy wytrzymałością na rozciąganie [w MPa], a siłą zrywającą [w kN] cięgna laminatowego, zbadano 10. próbek i uzyskano następujące wyniki:

Siła	1,50	0,80	1,20	2,00	1,00	1,80	0,90	1,70	1,40	2,20
Wytrż.	1,88	0,51	0,88	3,15	0,53	2,20	0,48	2,35	1,30	3,30

Wyznaczyć równanie regresji przyjmując: a) model zależności  $Y = \alpha + \beta x$ , b)  $Y = \alpha x^\beta$  oraz zakładając poziom istotności  $\alpha = 0,05$ .

#### Rozwiązanie

Dane do obliczeń zapiszemy w pliku CIEGNA w dwóch kolumnach nadając im nazwy SIŁA i WYTRZ. Z **Przełącznika modułów** programu STATISTICA wybiera się moduł **Regresja wielokrotna**. W zgłaszającym się oknie naciskamy przycisk **Zmienne**

i definiujemy, jako zmienną niezależną zmienną SIŁA a jako zmienną zależną zmienną WYTRZ. Z pozostałych opcji okna **Regresja wielokrotna** wybieramy:

- **Plik wejściowy:** Dane surowe;
- **Usuwanie BD:** Przypadkami;
- **Tryb:** Standardowa (ten tryb oznacza zależność liniową).

Po naciśnięciu przycisku **OK** otrzymuje się ekran wyników z podstawowymi informacjami o rezultatach obliczeń. W celu uzyskania szczegółowych wyników wybiera się przycisk **Podsumowanie regresji**. Otrzymuje się wyniki, które przedstawiono w tabeli 4.2.

Tabela 4.2

Wyniki regresji liniowej dla modelu  $Y = \alpha + \beta x$

Podsumowanie regresji zmiennej zależnej: WYTRZYM_ R=,98182830 R2=,96398682 Popraw. R^2=,95948517 F(1,8)=214,14 p<,00000 Błąd std. estymacji:,21695						
N=10	BETA	Błąd st. BETA	B	Błąd st. B	t(8)	poziom p
W. wolny			-1,56107	,230429	-6,77464	,000141
SIŁA	,981828	,067094	2,22005	,151709	14,63355	,000000

Następnie wybiera się przycisk **Analiza wariancji** i otrzymuje tabelę analizy wariancji dla wyestymowanej zależności (tabela 4.3).

Tabela 4.3.

Tabela analizy wariancji dla zależności regresyjnej

Analiza wariancji ; DV: WYTRZ. (ciegna.sta)					
Efekt	Suma kwadrat	df	Średnia kwadrat.	F	poziom p
Regres.	10,07902	1	10,07902	214,1409	,000000
Resztk.	,37654	8	,04707		
Razem	10,45556				

Z tabeli 4.2 w kolumnie B odczytuje się, że wyraz wolny  $a = -1,56107$  oraz współczynnik kierunkowy  $b = 2,22005$  (w kolumnie BETA podane są współczynniki dla zmiennych standaryzowanych, czyli o wartości średniej 0 i odchyleniu standardowym 1; współczynniki te pozwalają porównać względny wpływ każdej zmiennej niezależnej na zmienną zależną). Estymowane równanie regresji ma postać:

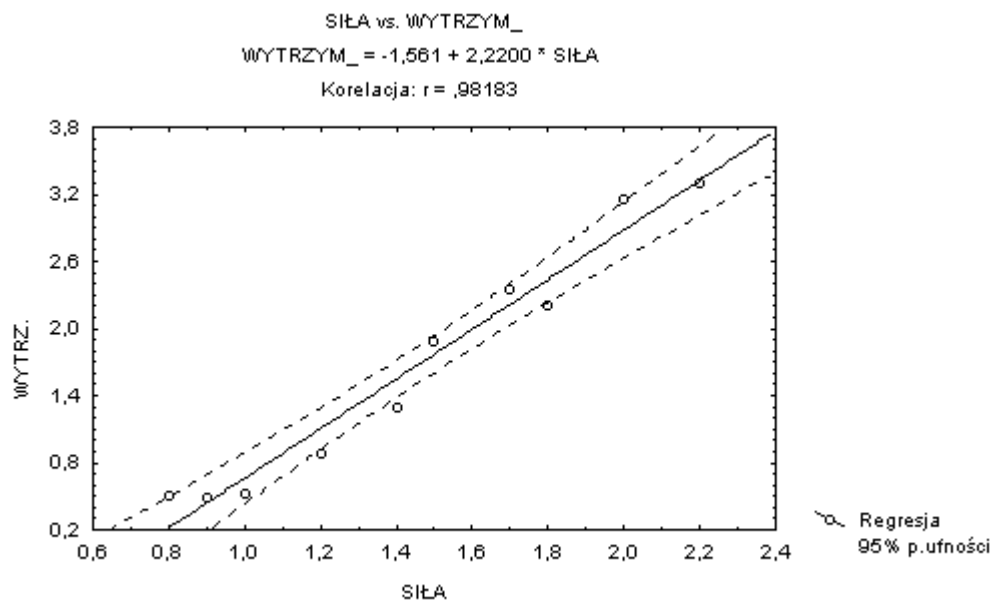
$$\hat{Y} = -1,56107 + 2,22005 x.$$

Kwadrat współczynnika korelacji  $R^2 = 0,9595$  jest wysoki, co świadczy o dobrym dopasowaniu prostej regresji do danych doświadczalnych. Obliczona wartość statystyki t dla współczynnika b  $t = 14,6336$  znacznie przekracza wartość krytyczną  $t_{0,05/2, 8} = 2,306$  dla

ośmiu stopni swobody i przy poziomie istotności  $\alpha = 0,05$  (tablica I). Należy zatem odrzucić hipotezę

o nieistotności współczynnika kierunkowego  $\beta$ , a tym samym, należy odrzucić hipotezę o nieistotności zależności między wytrzymałością na rozciąganie ciągła, a siłą zrywającą. Świadczy też o tym wartość  $p = 0,0000$ , która jest mniejsza od przyjętej wartości poziomu istotności  $\alpha$ . Do podobnego wniosku można dojść biorąc pod uwagę obliczoną wartość statystyki  $F = 214,1409$ , która też znacznie przekracza wartość krytyczną  $F_{0,05, 1, 8} = 5,32$ , przy poziomie istotności  $\alpha = 0,05$  dla 1 stopnia swobody licznika i 8 stopni swobody mianownika (tablica II). Wartość statystyki  $t$  dla wyrazu wolnego równa jest  $t = -6,77464$  i jej bezwzględna wartość znacznie przekracza wartość krytyczną, co świadczy o tym, że wyraz wolny jest istotnie różny od zera.

Obliczoną zależność można przedstawić na wykresie z zaznaczonymi granicami 95% przedziału ufności dla wartości oczekiwanej. W tym celu, w oknie **Wyniki regresji wielokrotnej** należy wybrać przycisk **Analiza reszt**, a w następnym oknie przycisk **Korelacje dwóch zmiennych**. W rezultacie otrzyma się wykres przedstawiony na rys. 4.8.



Rys. 4.8. Wykres obliczonej regresji wytrzymałości na rozciąganie względem siły zrywającej

Otrzymana zależność liniowa dobrze opisuje związek między zmienną zależną  $Y$ , a zmienną niezależną  $x$ . Prowadzący badania zaproponował jednak, jeszcze drugi typ zależności, w postaci regresji nieliniowej  $Y = \alpha x^\beta$ . Zależność ta, może być sprowadzona do postaci liniowej, przez obustronne logarytmowanie, i w tym nowym układzie, mogą być wykonane obliczenia oraz analiza statystyczna, tak jak dla regresji liniowej.

Aby wykonać obliczenia dla tego modelu, należy w oknie **Regresja wielokrotna** w polu **Tryb** wybrać opcję **Ustalona nieliniowa**. Zostanie otwarte nowe okno, w którym należy nacisnąć przycisk **Zmienna** i wskazać, które zmienne mają być przekształcone ( $V1$ –SIŁA i  $V2$ –WYTRZ.). Następnie w oknie **Nieliniowe składniki regresji** należy określić rodzaj przekształcenia, w naszym przypadku zaznacza się pole przy opcji  $LN(X)$ . Otwiera się okno



**Definicja modelu**, w którym po naciśnięciu przycisku **Zmienne**, wskazuje się zmienne do analizy, zależną LN–V2 i niezależną LN–V1, a w polu **Metoda** wybiera opcję **Standardowa**. Jako rezultat obliczeń, otrzymuje się ekran wyników i po wybraniu przycisku **Podsumowanie regresji** uzyskuje się wyniki, które przedstawiono w tabeli 4.4.

Tabela 4.4

Wyniki regresji liniowej dla modelu  $Y = \alpha x^\beta$

Podsumowanie regresji zmiennej zależnej: LN–V2 (cieгна.sta)						
R=,98247266 R2=,96525254 Popraw. R^2=,96090910						
F(1,8)=222,23 p<,00000 Błąd std. estymacji:,15085						
N=10	BETA	Błąd st. BETA	B	Błąd st. B	t(8)	poziom p
W. wolny			–,420884	,066546	–6,32472	,000227
LN–V1	,982473	,065905	2,164311	,145183	14,90747	,000000

Tabela 4.5

Tabela analizy wariancji dla zależności regresyjnej

Analiza wariancji ; DV: LN–V2 (cieгна.sta)					
	Suma kwadrat.	df	Średnia kwadrat.	F	poziom p
Regres.	5,057253	1	5,057253	222,2326	,000000
Resztk. Razem	,182053	8	,022757		
	5,239306				

Z kolumny B tablicy 4.4 odczytuje się wyraz wolny  $\ln a = -0,420884$  oraz współczynnik kierunkowy  $b = 2,164311$ . Zatem otrzymuje się zależność:

$$\ln \hat{Y} = -0,420884 + 2,164311 \ln x .$$

Wartość bezwzględna statystyki  $|t| = 14,90747$  dla współczynnika  $b$  jest dużo większa od wartości krytycznej  $t_{0,05/2, 8} = 2,306$ , co świadczy o istotności współczynnika  $\beta$ . Podobnie wartość bezwzględna statystyki  $|t| = 6,32472$  dla współczynnika  $a$  jest też dużo większa od wartości krytycznej, a więc wyraz wolny jest istotnie różny od zera. W obu przypadkach poziomy  $p$  są dużo mniejsze od przyjętego poziomu istotności  $\alpha$ , co potwierdza wnioski powyższe. Do podobnego wniosku dojdzie się analizując wartość statystyki  $F = 222,2326$ , która też znacznie przekracza wartość krytyczną  $F_{0,05, 1, 8} = 5,32$ .

Kwadrat współczynnika korelacji  $R^2 = 0,9609$  jest nieznacznie większy, niż dla modelu liniowego, co świadczy o nieco lepszym dopasowaniu zależności do danych. Można zatem stwierdzić, że zależność:

$$\hat{Y} = 0,65647 x^{2,164311} ,$$

jest istotna statystycznie oraz dość dobrze dopasowana do danych, i może służyć do opisu związku między zmienną  $Y$ , a zmienną  $x$ .

## 4.4. Regresja wielokrotna – wybór zmiennych

### 4.4.1. Opis metody

Dotychczas rozważane modele regresji były modelami pierwszego stopnia jednej zmiennej niezależnej. Bardziej ogólny typ modelu liniowego zmiennych  $x_1, x_2, \dots, x_k$  może być przedstawiony w postaci:

$$E(Y | x) = \beta_0 + \beta_1 f_1(\mathbf{x}) + \beta_2 f_2(\mathbf{x}) + \dots + \beta_m f_m(\mathbf{x}) . \quad (4.24)$$

Każda funkcja  $f_i(\mathbf{x})$ ,  $i = 1, 2, \dots, m$  jest ogólnie funkcją zmiennych niezależnych  $\mathbf{x}^T = (x_1, x_2, \dots, x_k)$  i może przybierać dowolną postać. W najprostszym przypadku każda  $f_i(\mathbf{x})$  może zawierać tylko jedną zmienną  $x$ . Nieznane współczynniki  $\beta_0, \beta_1, \dots, \beta_m$  nazywa się współczynnikami regresji wielokrotnej. Interesują nas następujące problemy:

- wybór podzbioru  $(f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))$  funkcji zmiennych niezależnych do modelu regresji z pewnego zadanego zbioru,
- oszacowanie współczynników regresji  $\beta_0, \beta_1, \dots, \beta_m$  i weryfikacja hipotez  $H_{0i}: \beta_i = 0$ ,  $i = 0, 1, \dots, m$ ,
- ocena stopnia dopasowania zależności zawierającej wybrane funkcje zmiennych niezależnych do danych.

Niech  $(x_{11}, x_{12}, \dots, x_{1k}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{nk}, y_n)$  będzie ciągiem  $n$  wektorów obserwacji zmiennych niezależnych  $x_1, x_2, \dots, x_k$  oraz zmiennej zależnej  $Y$ . Przyjmuje się, że zbiór funkcji  $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})$  jest zadany, przy czym liczba funkcji w tym zbiorze  $m+1 \leq n$ . Jedną z metod wyboru podzbioru funkcji  $f(\mathbf{x})$  zmiennych niezależnych jest metoda odrzucania. Zasadnicze etapy tej procedury są następujące:

1. Oblicza się oszacowania współczynników regresji w modelu zawierającym wszystkie możliwe funkcje zmiennych niezależnych, stosując metodę najmniejszych kwadratów zdefiniowaną w następujący sposób:

$$S = \sum_{j=1}^n \left[ y_j - \left( \beta_0 + \sum_{i=1}^m \beta_i f_{ji}(\mathbf{x}) \right) \right]^2 . \quad (4.25)$$

2. Niech  $b_0, b_1, \dots, b_m$  będą oszacowaniami nieznanymi współczynników regresji  $\beta_0, \beta_1, \dots, \beta_m$ . Dla każdego współczynnika  $b_i$  oblicza się statystykę:

$$t_i = b_i / s_{b_i} , \quad (4.26)$$

gdzie  $s_{b_i}^2$  jest wariancją oszacowania współczynnika  $\beta_i$ . Statystyka ta służy do weryfikacji hipotezy  $H_{0i}: \beta_i = 0$  określającej, że udział zmiennej  $f_i(\mathbf{x})$  w modelu regresji jest nieistotny. Zakładając, że zmienna zależna  $Y$  ma rozkład normalny i hipoteza  $H_0$  jest prawdziwa, to statystyka  $t_i$  ma rozkład t-Studenta z  $f = n - m - 1$  stopniami swobody. Hipotezę  $H_0$  odrzuca się, jeżeli wartość obliczona  $|t_i|$  przekracza wartość krytyczną  $t_{\alpha/2, f}$ , przy zadanym poziomie istotności  $\alpha$ .

3. Znajduje się najmniejszą wartość  $t_{\min} = \min t_i$  i porównuje się ją z wartością krytyczną  $t_{\alpha/2, f}$  rozkładu t-Studenta.

4. Jeśli  $t_{\min} \geq t_{\alpha/2, f}$  to otrzymane równanie regresji zawiera tylko istotne funkcje zmiennych niezależnych i uważa się je za ostateczne.

5. Jeśli  $t_{\min} < t_{\alpha/2, f}$  to funkcję zmiennej niezależnej  $f_i(\mathbf{x})$  usuwa się z równania, ponownie oblicza oszacowania współczynników równania regresji z pozostałymi funkcjami zmiennych niezależnych, i wraca do etapu 2.

Ocenę stopnia dopasowania wyznaczonego równania regresji do danych, przeprowadza się w oparciu o tabelę analizy wariancji, podobnie jak dla regresji jednej zmiennej niezależnej. Oblicza się więc następujące wielkości:

$$\begin{aligned}
 SG &= \sum_{j=1}^n (y_j - \bar{y})^2 - \text{zmiennosc całkowita,} \\
 SM &= \sum_{j=1}^n [b_0 + \sum_{i=1}^m b_i f_{ji}(\mathbf{x}) - \bar{y}]^2 - \text{zmiennosc wynikajaca z przyjetego modelu regresji,} \\
 SR &= \sum_{j=1}^n [b_0 + \sum_{i=1}^m b_i f_{ji}(\mathbf{x}) - y_j]^2 - \text{zmiennosc resztowa,} \\
 R^2 &= SM / SG. \tag{4.27}
 \end{aligned}$$

Współczynnik  $R^2$  nazywany jest współczynnikiem determinacji i jest on kwadratem współczynnika korelacji wielokrotnej z próby  $R$ . Wartość  $R^2$  należy do przedziału domkniętego  $[0, 1]$  i może być traktowana jako miara stopnia dopasowania powierzchni regresji do danych doświadczalnych.

Hipoteza  $H_0: R^2 = 0$  stwierdza, że udział zmiennych niezależnych w modelu regresji jest nieistotny i jest równoważna hipotezie  $H_0: \beta_1 = \beta_2 = \dots = \beta_m = 0$ . Przy prawdziwości hipotezy  $H_0$  statystyka:

$$F = \frac{(n - m - 1) SM}{m SR} = \frac{(n - m - 1) R^2}{m (1 - R^2)}, \tag{4.28}$$

ma rozkład F-Snedecora z  $m$  stopniami swobody dla licznika i  $n - m - 1$  stopniami swobody dla mianownika. Hipotezę  $H_0$  odrzuca się, jeżeli obliczona wartość  $F$ , przekracza wartość krytyczną  $F_{\alpha}$ , przy przyjętym poziomie istotności  $\alpha$ . Podobnie jak w regresji jednej zmiennej niezależnej, wielkości te przedstawiane są w postaci tabeli analizy wariancji (tabela 4.6).

Tabela 4.6

Tabela analizy wariancji dla regresji wielokrotnej

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średni kwadrat	Statystyka F
W modelu regresji	SM	m	MS=SM/m	
Poza regresją (reszta)	SR	n - m - 1	$s^2 = SR/(n - m - 1)$	$F = MS/s^2$
Łącznie (względem średniej)	SG	n - 1		

Procedura eliminacji zmiennych ma tą niedogodność, że po odrzuceniu kolejnej funkcji, może okazać się, że któraś z wcześniej odrzuconych funkcji, stanie się istotna (z powodu korelacji między funkcjami zmiennych). Powinna być zatem włączona do zależności. Tak zmodyfikowana procedura odrzucania, że po każdym odrzuceniu funkcji, analizowane są funkcje, które wcześniej były odrzucone, i jeśli znajdzie się funkcję istotną, włączana jest ona do zależności, nazywa się regresją krokową z odrzucaniem zmiennych.

Jeżeli przewidywana liczba funkcji zmiennych niezależnych w końcowym równaniu regresji jest znacznie mniejsza od liczby funkcji w zbiorze wyjściowym, bardziej korzystna może okazać się inna metoda wyboru funkcji  $f_i(\mathbf{x})$ , tzw. metoda regresji krokowej z dołączaniem zmiennych. Wówczas obliczenia przebiegają zgodnie z następującymi krokami:

1. Startuje się z modelem  $E(Y | \mathbf{x}) = \beta_0$ , który nie zawiera żadnej z funkcji zmiennych niezależnych  $f_i(\mathbf{x})$ ,  $i = 1, 2, \dots, m$ .

2. Dla każdej funkcji  $f_i(\mathbf{x})$  z postulowanego zbioru oblicza się oszacowania  $b_0$  i  $b_i$  współczynników  $\beta_0$  i  $\beta_i$  modelu regresji  $E(Y | \mathbf{x}) = \beta_0 + \beta_i f_i(\mathbf{x})$ . Następnie oblicza się następujące wielkości:

$$SG = \sum_{j=1}^n (y_j - \bar{y})^2,$$

$$SM_i = \sum_{j=1}^n [b_0 + b_i f_{ji}(\mathbf{x}) - \bar{y}]^2,$$

$$SR_i = \sum_{j=1}^n [b_0 + b_i f_{ji}(\mathbf{x}) - y_j]^2,$$

$$F_i = \frac{(n-2) SM_i}{SR_i}. \quad (4.29)$$

Do równania regresji wybiera się funkcję zmiennej niezależnej, dla której wartość  $F_i$  jest największa i przekracza wartość krytyczną  $F_\alpha$  rozkładu F-Snedecora z 1 i  $n-2$  stopniami swobody. Oczywiście jeżeli dla żadnej z funkcji zmiennych, wartość  $F_i$  nie przekracza wartości krytycznej  $F_\alpha$ , to przyjmuje się model  $E(Y | \mathbf{x}) = \beta_0$ .

3. Jeżeli funkcja  $f_i(\mathbf{x})$  została włączona do równania, to w następnym kroku, poszukuje się kolejnej funkcji  $f_k(\mathbf{x})$ , która mogłaby być włączona do równania. W tym celu oblicza się oszacowania  $b_0$ ,  $b_i$ ,  $b_k$  współczynników  $\beta_0$ ,  $\beta_i$ ,  $\beta_k$  modelu  $E(Y | \mathbf{x}) = \beta_0 + \beta_i f_i(\mathbf{x}) + \beta_k f_k(\mathbf{x})$  oraz:

$$F_{i,k} = \frac{(n-3)(SM_{i,k} - SM_i)}{SR_{i,k}}, \quad (4.30)$$

gdzie:

$$SM_{i,k} = \sum_{j=1}^n [b_0 + b_i f_{ji}(\mathbf{x}) + b_k f_{jk}(\mathbf{x}) - \bar{y}]^2,$$

$$SR_{i,k} = \sum_{j=1}^n [b_0 + b_i f_{ji}(x) + b_k f_{jk}(x) - y_j]^2.$$

Do równania regresji dołącza się funkcję  $f_k(\mathbf{x})$ , dla której wartość  $F_{i,k}$  jest największa, i przekracza wartość krytyczną  $F_\alpha$  rozkładu F–Snedecora z 1 i  $n - 3$  stopniami swobody.

4. Jeżeli dla żadnej z funkcji zmiennych niezależnych, wartość  $F_{i,k}$  nie przekracza wartości krytycznej  $F_\alpha$ , to otrzymane równanie  $E(Y | \mathbf{x}) = \beta_0 + \beta_i f_i(\mathbf{x})$  uważa się za ostateczne.

Dalsze postępowanie polega na poszukiwaniu kolejnej funkcji, która mogłaby być dołączona do podzbioru  $[f_i(\mathbf{x}), f_k(\mathbf{x})]$ , według zasad opisanych w krokach 3 i 4, aż do ustalenia końcowego zbioru funkcji zmiennych niezależnych wchodzących do równania regresji. Należy podkreślić, że w każdym kroku dołączania funkcji do podzbioru może wystąpić konieczność wyeliminowania jednej z wcześniej dołączonych funkcji. Spowodowane to jest korelacją między poszczególnymi funkcjami, w wyniku tego, po dołączeniu pewnej funkcji, inna funkcja będąca już w równaniu, może okazać się nieistotna. W skrypcie pomija się opis tej wstecznej eliminacji funkcji.

#### 4.4.2. Przykład obliczeń programem STATISTICA

##### Przykład 4.2

Wyznaczyć zależność opisującą związek między zmienną zależną  $Y$  i dwoma zmiennymi niezależnymi  $x_1$  i  $x_2$  w postaci wielomianu drugiego stopnia:

$$E(Y | \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2,$$

przyjmując poziom istotności  $\alpha = 0,05$ . W trakcie badań doświadczalnych otrzymano następujące wyniki:

$x_1$	0,33	0,53	0,33	0,53	0,33	0,53	0,43	0,43	0,43
$x_2$	3,00	3,00	5,00	5,00	4,00	4,00	3,00	5,00	4,00
$y$	2,20	2,80	3,90	4,50	2,40	3,27	1,70	3,90	2,30

##### Rozwiązanie

Dane doświadczalne zapiszemy w pliku DOKL w trzech kolumnach  $x_1$ ,  $x_2$ ,  $y$  oraz utworzymy trzy dodatkowe kolumny (funkcje zmiennych niezależnych)  $x_1^2$ ,  $x_2^2$ ,  $x_1 \cdot x_2$ , odpowiadające funkcjom zmiennych występujących w postulowanym modelu zależności. W pierwszym etapie, wykonamy obliczenia metodą standardową regresji wielokrotnej w celu wstępnej oceny istotności zależności i istotności poszczególnych współczynników oraz dopasowania proponowanego modelu do danych. W tym celu, wybiera się w oknie **Regresja wielokrotna** tryb **Standardowy** i zaznacza opcję **Wykonaj domyślną (nie krokową) analizę**. Definiuje się zmienne: zależną  $Y$  i niezależne  $X_1$ ,  $X_2$ ,  $X_1^2$ ,  $X_2^2$ ,  $X_1 X_2$  i otrzymuje wyniki przedstawione w tabeli 4.7.

Tabela 4.7

## Wyniki wstępnej analizy regresji wielokrotnej

Podsumowanie regresji zmiennej zależnej: Y (dokl.sta) R=,99245117 R2=,98495933 Popraw. R^2=,95989154 F(5,3)=39,292 p<,00618 Błąd std. estymacji:,18950						
N=9	BETA	Błąd st. BETA	B	Błąd st. B	t(3)	poziom p
W. wolny			15,3136	3,57658	4,28162	,023406
X1	-3,97396	1,112541	-43,4200	12,15575	-3,57197	,037499
X2	-2,87995	1,051983	-3,1467	1,14941	-2,73764	,071480
X1_2	4,29938	1,057082	54,5000	13,39983	4,06721	,026811
X2_2	3,74388	,983675	,5100	,13400	3,80602	,031874
X1_X2	-,00000	,514189	-,0000	,94751	-,00000	1,000000

Przyjmując poziom istotności  $\alpha = 0,05$  przy  $f = 9-5-1$  stopniach swobody, wartość krytyczna rozkładu t-Studenta  $t_{0,05/2, 3} = 3,182$  (tablica I), zatem nieistotne są współczynniki funkcji  $x_2$  i  $x_1 \cdot x_2$  (wartość poziomu  $p$  jest większa od poziomu istotności  $\alpha = 0,05$ ). Kwadrat współczynnika korelacji wielokrotnej  $R^2 = 0,9599$  jest dość wysoki, co świadczy o dobrym dopasowaniu obliczonego równania do danych doświadczalnych. Z tabeli analizy wariancji (tabela 4.8) wynika, że obliczona wartość statystyki  $F = 39,292$  znacznie przekracza wartość krytyczną, która na poziomie istotności  $\alpha = 0,05$  i przy 5 stopniach swobody dla licznika i 3 stopniach swobody dla mianownika wynosi  $F_{0,05} = 9,01$  (tablica II).

Tabela 4.8

## Tabela analizy wariancji

Analiza wariancji ; DV: Y (dokl.sta)					
Efekt	Suma kwadrat	df	Średnia kwadrat.	F	poziom p
Regres.	7,055067	5	1,411013	39,29183	,006178
Resztk.	,107733	3	,035911		
Razem	7,162800				

Wynika stąd, że cała zależność jest istotna, lecz wzajemne korelacje między wprowadzonymi funkcjami zmiennych niezależnych powodują, że dwie z nich są nieistotne.

Dla doboru odpowiedniej postaci równania zastosujemy metodę regresji krokowej z dołączaniem zmiennych. W tym celu w oknie **Regresja wielokrotna** usuwamy znacznik w opcji **Wykonaj domyślną (nie krokową) analizę**, a następnie po otwarciu się okna **Definicja modelu** w polu **Metoda** wybiera się opcję **Krokowa postępująca**. Po włączeniu opcji regresji krokowej istnieje możliwość określenia wartości *F do wprowadzenia* oraz wartości *F do usunięcia* zmiennej. Wartość *F do wprowadzenia* musi być zawsze większa od wartości

*F do usunięcia* zmiennej. W naszym przykładzie przyjmujemy wartości *F* domyślne i w rezultacie obliczeń otrzymuje się wyniki przedstawione w tabeli 4.9.

Tabela 4.9

Wyniki analizy regresji metodą dołączania

Podsumowanie regresji zmiennej zależnej: Y (dokl.sta) R=,99245117 R2=,98495933 Popraw. R^2=,96991865 F(4,4)=65,486 p<,00067 Błąd std. estymacji:,16411						
N=9	BETA	Błąd st . BETA	B	Błąd st. B	t(4)	poziom p
W. wolny			15,3136	2,75716	5,55410	,005143
X2_2	3,74388	,851888	,5100	,11605	4,39481	,011739
X1_2	4,29938	,915460	54,5000	11,60460	4,69641	,009333
X1	-3,97396	,915460	-43,4200	10,00242	-4,34095	,012244
X2	-2,87995	,851888	-3,1467	,93078	-3,38067	,027768

W wyniku przeprowadzonych obliczeń do modelu regresji zostały wybrane cztery funkcje. Z tabeli 4.9 wynika, że wszystkie współczynniki obliczonego równania regresji:

$$\hat{Y} = 15,3136 - 43,42 x_1 - 3,1467 x_2 + 54,50 x_1^2 + 0,51 x_2^2 ,$$

są istotne; wartość krytyczna rozkładu *t*-Studenta  $t_{0,05/2, 4} = 2,776$  przy 4 stopniach swobody (tablica I). Świadczą też o tym, obliczone dla każdej wartości statystyki *t*, wartości poziomu *p*, które są wszystkie mniejsze od przyjętego poziomu istotności  $\alpha$ . Obliczona wartość statystyki  $F = 65,486$  znacznie przekracza wartość krytyczną  $F_{0,05} = 6,39$  przy 4 stopniach swobody licznika i 4 czterech stopniach swobody mianownika (tablica II), wartość *p* jest mniejsza od poziomu istotności  $\alpha$ , co oznacza, że zależność jest istotna statystycznie i może być wykorzystywana do prognozowania wartości zmiennej *Y* w funkcji zmiennych  $x_1$  i  $x_2$ . Kwadrat współczynnika korelacji wielokrotnej  $R^2 = 0,9699$  jest wysoki, co świadczy o dobrym dopasowaniu powierzchni regresji do danych doświadczalnych.

Dla obliczenia równania regresji metodą regresji krokowej z odrzucaniem zmiennych należy w oknie **Definicja modelu** w polu **Metoda** wybrać opcję **Krokowa wsteczna**. Dla naszego przykładu wyniki obliczeń metodą krokową z odrzucaniem zmiennych są identyczne z wynikami uzyskanymi metodą regresji krokowej z dołączaniem zmiennych.

## 5. REGRESJA NIELINIOWA

### 5.1. Cel ćwiczenia

Celem ćwiczenia jest zapoznanie się ze sposobami estymacji nieliniowych zależności losowych.

### 5.2. Wprowadzenie

W badaniach naukowych zależności opisujące związek między zmienną zależną  $Y$  i zmiennymi niezależnymi  $\mathbf{x}$  często wynikają bezpośrednio z analizy biologicznej, chemicznej czy fizycznej badanych zjawisk i mają charakter nieliniowy względem współczynników. Każdy ze współczynników ma wówczas określone znaczenie, a celem badań jest nie tyle aproksymacja danych doświadczalnych pewną zależnością, co przede wszystkim oszacowanie współczynników z maksymalną dokładnością. Estymacja współczynników funkcji nieliniowych jest znacznie trudniejsza niż funkcji liniowych, ale wraz z dynamicznym rozwojem oprogramowania, zastosowanie postaci nieliniowych stało się dość powszechne. W czasach, gdy dostęp do komputerów był utrudniony, nieliniowa regresja była prawie niedostępna dla większości badaczy. Aby uniknąć trudności związanych z estymacją zależności nieliniowych, stosowano, co najwyżej, takie modele nieliniowe, które poprzez odpowiednie przekształcenie można sprowadzić do postaci liniowej. Jednak przekształcanie danych powoduje zmianę ich wzajemnych relacji i wyniki estymacji zależności liniowych różnią się, niestety często dość znacznie, od wyników otrzymywanych z estymacji nieliniowej.

Funkcję nieliniową w ogólnej postaci można zapisać następująco:

$$y_i = \eta(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad (5.1)$$

gdzie:

$\eta(\mathbf{x}_i, \boldsymbol{\beta})$  – dowolna funkcja różnowartościowa, ciągła i różniczkowalna,

$\mathbf{x}_i = [x_{1i}, x_{2i}, \dots, x_{ki}]^T$  – wektor  $k$  zmiennych niezależnych,

$\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_m]^T$  – wektor  $m$  współczynników,

$\varepsilon_i$  – odchylenie losowe.

Zakłada się, że odchylenia losowe  $\varepsilon_i$  są nieskorelowane o wartości oczekiwanej  $E(\varepsilon) = 0$  i jednorodnej wariancji  $D^2(\varepsilon) = \sigma^2$  oraz zwykle przyjmuje się, że mają one rozkład normalny. Jeżeli dostępnych jest  $n$  obserwacji w postaci:

$$x_{1i}, x_{2i}, \dots, x_{ki}, y_i, \quad i=1, 2, \dots, n,$$

to estymatory współczynników nieliniowej metody najmniejszych kwadratów NMNK (ang. nonlinear least squares, NLS) otrzymuje się minimalizując sumę kwadratów reszt:



$$S(\mathbf{b}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - \eta(\mathbf{x}_i, \mathbf{b})]^2, \quad (5.2)$$

gdzie:

$\mathbf{b} = [b_1, b_2, \dots, b_m]^T$  – wektor estymatorów współczynników  $\beta$ ,  
 $e_i = y_i - \eta(\mathbf{x}_i, \mathbf{b})$  – reszta,

lub w zapisie macierzowym

$$S(\mathbf{b}) = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \boldsymbol{\eta})^T (\mathbf{y} - \boldsymbol{\eta}), \quad (5.3)$$

gdzie:

$\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ ,  $\mathbf{e} = [e_1, e_2, \dots, e_n]^T$ ,  
 $\boldsymbol{\eta} = [\eta(\mathbf{x}_1, \mathbf{b}), \eta(\mathbf{x}_2, \mathbf{b}), \dots, \eta(\mathbf{x}_n, \mathbf{b})]^T$ .

Aby otrzymać estymator metody najmniejszych kwadratów należy zróżniczkować równanie (5.2) względem  $\mathbf{b}$ :

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = 0.$$

Uzyskuje się układ  $m$  nieliniowych równań względem współczynników:

$$-2 \sum_{i=1}^n [y_i - \eta(\mathbf{x}_i, \mathbf{b})] \frac{\partial \eta(\mathbf{x}_i, \mathbf{b})}{\partial \beta_p} = 0, \quad p = 1, 2, \dots, m \quad (5.4)$$

gdzie:  $\frac{\partial \eta(\mathbf{x}_i, \mathbf{b})}{\partial \beta_p} = \left( \frac{\partial \eta(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_p} \right)_{\beta_p = b_p}$ .

Zauważmy, że jeżeli funkcja  $\eta(\mathbf{x}_i, \boldsymbol{\beta})$  jest liniowa względem współczynników, to jej pochodna jest funkcją jedynie  $\mathbf{x}_i$  i nie zawiera żadnych  $\boldsymbol{\beta}$  (jest niezależna od  $\boldsymbol{\beta}$ ).

Oszacowanie przedziału ufności dla współczynników  $\boldsymbol{\beta}$  opiera się zwykle na linearyzacji modelu (5.1) w pobliżu wartości oszacowań  $\mathbf{b}$ . Równanie (5.1) po linearyzacji wokół wartości estymatora  $\mathbf{b}$  wektora współczynników  $\boldsymbol{\beta}$  przyjmie następującą postać:

$$y_i = \eta(\mathbf{x}_i, \mathbf{b}) + \frac{\partial \eta(\mathbf{x}_i, \mathbf{b})}{\partial \boldsymbol{\beta}^T} (\boldsymbol{\beta} - \mathbf{b}) + \varepsilon_i. \quad (5.5)$$

Przyjmując, że odchylenia losowe mają rozkład normalny, estymator  $\mathbf{b}$  wektora współczynników  $\boldsymbol{\beta}$  ma graniczny rozkład normalny o wartości oczekiwanej  $E(\mathbf{b}) = \boldsymbol{\beta}$  oraz o macierzy kowariancji:

$$\mathbf{C} = (\mathbf{Z}^T \mathbf{Z})^{-1} \sigma^2, \quad (5.6)$$

gdzie:

$\mathbf{Z} = \frac{\partial \boldsymbol{\eta}(\mathbf{x}, \mathbf{b})}{\partial \boldsymbol{\beta}^T}$  oznacza  $(n \times m)$  wymiarową macierz wartości pochodnych.

Wówczas elipsoida ufności wektora współczynników  $\boldsymbol{\beta}$  może być w przybliżeniu określona wzorem analogicznym jak dla liniowej regresji:

$$(\boldsymbol{\beta} - \mathbf{b})^T \mathbf{Z}^T \mathbf{Z} (\boldsymbol{\beta} - \mathbf{b}) \leq m s^2 F_{1-\alpha, m, n-m}, \quad (5.7)$$

gdzie:

$s^2 = S(\mathbf{b})(n - m)^{-1}$  – estymator wariancji  $\sigma^2$ .

Jeżeli linearyzacja zależności nie daje dobrego przybliżenia zależności rzeczywistej, wtedy powyższa elipsoida nie będzie prawdziwym obszarem ufności. Dokładny obszar ufności można określić przez wyznaczenie położenia stałych wartości  $S(\mathbf{b}) = \text{const}$ , lecz wtedy nie można dokładnie określić wartości prawdopodobieństwa, ponieważ nie jest znany dokładny rozkład wektora estymatorów  $\mathbf{b}$ . Najczęściej przyjmuje się, że linearyzacja stanowi dość dobre przybliżenie modelu nieliniowego i wtedy analizę statystyczną wyestymowanej zależności wykonuje się wykorzystując sposoby analizy regresji liniowej. Wtedy test statystyczny do weryfikacji hipotezy o nieistotności współczynników  $\beta$  ( $H_0: \beta_p=0, p=1, 2, \dots, m$ ) może być oparty na statystyce  $t_p = \frac{b_p}{s \sqrt{c_{pp}}}$ , która przy założeniu, że hipoteza zerowa jest prawdziwa ma graniczny rozkład t-Studenta o  $(n-m)$  stopniach swobody. Miarą stopnia dopasowania zależności regresyjnej do danych doświadczalnych będzie współczynnik determinacji  $R^2$ , zdefiniowany i interpretowany podobnie jak dla regresji liniowej.

### 5.3. Estymacja współczynników modeli nieliniowych

Układ równań (5.4) najczęściej nie ma rozwiązania analitycznego. Dlatego też rozwiązania poszukuje się za pomocą metod iteracyjnych. Wybierając wektor  $\mathbf{b}_0$  jako wektor startowy dąży się do tego, aby otrzymać ciąg ocen  $(\mathbf{b}_0, \dots, \mathbf{b}_l, \mathbf{b}_{l+1}, \dots)$  spełniający dla każdego  $l$  warunek:

$$S(\mathbf{b}_{l+1}) < S(\mathbf{b}_l). \quad (5.8)$$

Polega to na korygowaniu kolejnych przybliżeń ocen współczynników o pewien wektor:

$$\mathbf{v}_l = \lambda \mathbf{d}_l, \quad (5.9)$$

spełniający warunek:

$$S(\mathbf{b}_l + \lambda \mathbf{d}_l) < S(\mathbf{b}_l), \quad (5.10)$$

gdzie  $\lambda$  jest liczbą określającą długość kroku a  $\mathbf{d}_l$  jest wektorem określającym kierunek poszukiwań. Poszukuje się zatem takiego kierunku wyznaczonego przez wektor  $\mathbf{d}_l$ , aby funkcja  $S(\mathbf{b}_l + \lambda \mathbf{d}_l)$  była funkcją malejącą względem  $\lambda$ . Pochodna funkcji kryterium:

$$\frac{\partial [S(\mathbf{b}_l + \lambda \mathbf{d}_l)]}{\partial \lambda} = \left[ \frac{\partial S}{\partial \mathbf{b}_l} \right]^T \left[ \frac{\partial (\mathbf{b}_l + \lambda \mathbf{d}_l)}{\partial \lambda} \right]$$

musi być wobec tego mniejsza od zera:

$$\left[ \frac{\partial S}{\partial \mathbf{b}_l} \right]^T \mathbf{d}_l < 0. \quad (5.11)$$

Oznaczając gradient funkcji  $\frac{\partial S}{\partial \mathbf{b}_l} = \mathbf{g}_l$ , warunek (5.11) przyjmie postać:

$$\mathbf{g}_l^T \mathbf{d}_l < 0. \quad (5.12)$$

Definiując wektor  $\mathbf{d}_l$  jako:

$$\mathbf{d}_l = -\mathbf{P}_l \mathbf{g}_l, \quad (5.13)$$

gdzie  $\mathbf{P}_l$  jest dowolną macierzą kwadratową dodatnio określoną, forma kwadratowa wektora  $\mathbf{g}_l$  o macierzy  $\mathbf{P}_l$  będzie spełniać relację  $\mathbf{g}_l^T \mathbf{P}_l \mathbf{g}_l > 0$  dla każdego wektora  $\mathbf{g}_l$ .

Kolejne przybliżenie wektora ocen współczynników może być przedstawione w postaci:

$$\mathbf{b}_{l+1} = \mathbf{b}_l - \lambda_l \mathbf{P}_l \mathbf{g}_l, \quad (5.14)$$

gdzie  $\lambda_l$  jest długością kroku w  $l$ -tej iteracji,  $\mathbf{P}_l$  dowolną macierzą dodatnio określoną,  $\mathbf{g}_l$  gradientem funkcji kryterium.

Poszczególne metody gradientowe poszukiwania ekstremum funkcji kryterium różnią się przede wszystkim postacią macierzy  $\mathbf{P}_l$ .

### **Metoda najszybszego spadku**

Jest to najprostsza metoda gradientowa, w której za  $\mathbf{P}_l$  przyjmuje się macierz jednostkową  $\mathbf{I}$ , wówczas:

$$\mathbf{b}_{l+1} = \mathbf{b}_l + \lambda_l \mathbf{g}_l. \quad (5.15)$$

Długość kroku  $\lambda_l$  można określić aproksymując funkcję  $S(\mathbf{b})$  funkcją kwadratową i minimalizując ją na kierunku  $\mathbf{g}_l$ . Wadą metody jest niezmienniczość macierzy  $\mathbf{P}$  w kolejnych iteracjach, co nie pozwala na bieżące korygowanie ocen współczynników w zależności od postaci funkcji kryterium.

### **Metoda Newtona-Raphsona**

Przekształćmy wzór (5.3) w następujący sposób:

$$S(\mathbf{b}) = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \boldsymbol{\eta})^T (\mathbf{y} - \boldsymbol{\eta}) = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \boldsymbol{\eta} - \boldsymbol{\eta}^T \mathbf{y} + \boldsymbol{\eta}^T \boldsymbol{\eta} = \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\eta}^T \mathbf{y} + \boldsymbol{\eta}^T \boldsymbol{\eta}. \quad (5.16)$$

Funkcja  $S(\mathbf{b})$  osiąga ekstremum, gdy jej pochodna jest równa wektorowi zerowemu:

$$\mathbf{g} = \frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2 \mathbf{Z}^T \mathbf{y} + 2 \mathbf{Z}^T \boldsymbol{\eta} = -2 \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\eta}) = 0, \quad (5.17)$$

gdzie  $\mathbf{Z} = \frac{\partial \boldsymbol{\eta}(\mathbf{x}, \mathbf{b})}{\partial \boldsymbol{\beta}^T}$  jest macierzą o wymiarach  $[n * m]$ .

W metodzie Newtona-Rapsona przyjmuje się, że macierz  $\mathbf{P}_l = \mathbf{H}^{-1}$ , gdzie macierz  $\mathbf{H}$  jest hesjanem funkcji kryterium:

$$\mathbf{H} = \frac{\partial S(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^T} = \frac{\partial \mathbf{g}}{\partial \mathbf{b}^T} = 2 \mathbf{Z}^T \mathbf{Z} - 2 \sum_{i=1}^n \left\{ \begin{array}{l} \frac{\partial \eta(\mathbf{x}_i, \mathbf{b})}{\partial b_1 \partial \mathbf{b}^T} (y_i - \eta(\mathbf{x}_i, \mathbf{b})) \\ \dots \\ \frac{\partial \eta(\mathbf{x}_i, \mathbf{b})}{\partial b_m \partial \mathbf{b}^T} (y_i - \eta(\mathbf{x}_i, \mathbf{b})) \end{array} \right\}, \quad (5.18)$$

gdzie  $\frac{\partial^2 \eta}{\partial b_p \partial \mathbf{b}^T}$  jest  $m$  wymiarowym wektorem pochodnych cząstkowych funkcji  $\partial \eta(\mathbf{x}_i, \mathbf{b}) / \partial b_p$  względem  $m$  elementów wektora  $\mathbf{b}^T$ . Macierz w nawiasie klamrowym, której pierwszy i ostatni wiersz pokazano, ma wymiary  $(m \times m)$ .

Kolejne przybliżenie wektora ocen współczynników może być, zatem przedstawione w postaci:

$$\mathbf{b}_{l+1} = \mathbf{b}_l - \lambda_l \mathbf{H}^{-1} \mathbf{g}_l . \quad (5.19)$$

Po uwzględnieniu wzoru (5.17) zależność (5.19) przyjmie postać:

$$\mathbf{b}_{l+1} = \mathbf{b}_l - \lambda_l \mathbf{H}^{-1} * [-2 \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\eta})] . \quad (5.20)$$

Jeśli estymowana funkcja  $\eta(\mathbf{x}_i, \boldsymbol{\beta})$  jest liniowa, to funkcja kryterium jest funkcją kwadratową i powyższy wzór pozwala uzyskać rozwiązanie w jednej iteracji przy długości kroku  $\lambda = 1$ . Jeżeli jednak funkcja  $S(\mathbf{b})$  nie jest kwadratowa to musi być stosowana iteracyjna procedura optymalizacji. Poważnym utrudnieniem w stosowaniu metody Newtona-Raphsona jest konieczność obliczania hesjanu, czyli pochodnych drugiego rzędu funkcji kryterium. Hesjan może nie być dodatnio określony i wówczas proces iteracyjny może prowadzić do punktu, w którym funkcja nie osiąga minimum lub proces może być rozbieżny.

### **Metoda Gaussa-Newtona**

Metoda ta jest uproszczeniem metody Newtona-Raphsona uzyskanym przez aproksymację hesjanu funkcji kryterium za pomocą pierwszego składnika sumy (5.18) i wtedy można uprościć wzór (5.20) do postaci:

$$\mathbf{b}_{l+1} = \mathbf{b}_l - \lambda_l [2 \mathbf{Z}^T \mathbf{Z}]^{-1} * [-2 \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\eta})] . \quad (5.21)$$

Jeżeli w otoczeniu punktu  $\mathbf{b}_l$  funkcja  $\eta(\mathbf{x}_i, \boldsymbol{\beta})$  nie odbiega zasadniczo od funkcji liniowej to taka aproksymacja nie jest obciążona zbyt dużym błędem, a metoda dość efektywna. Z porównania wzorów (5.14) i (5.21) wynika, że macierz  $\mathbf{P}_l$  w tej metodzie ma postać:

$$\mathbf{P}_l = [2 \mathbf{Z}^T \mathbf{Z}]^{-1} . \quad (5.22)$$

### **Metoda Marquardta**

Metoda jest modyfikacją metody Gaussa-Newtona polegającą na wykorzystaniu faktu, że macierz  $\mathbf{P}_l + \gamma_l \bar{\mathbf{P}}_l$  jest zawsze dodatnio określona, jeśli macierz  $\bar{\mathbf{P}}_l$  jest dodatnio określona i stała  $\gamma_l$  jest wystarczająco duża. Macierz  $\mathbf{P}$  będzie miała postać:

$$\mathbf{P}_l = [2 \mathbf{Z}^T \mathbf{Z} + \gamma_l \bar{\mathbf{P}}_l]^{-1} . \quad (5.23)$$

Najczęściej przyjmuje się  $\bar{\mathbf{P}}_l = \mathbf{I}$ . Jeśli stała  $\gamma_l$  jest bliska zeru, to metoda Marquardta jest bliska metodzie Gaussa-Newtona. Im  $\gamma_l$  jest większe, tym bardziej metoda Marquardta odpowiada metodzie najszybszego spadku. Posługując się czynnikiem  $\gamma_l$  można mody-

fikować hesjan funkcji kryterium i rozpatrywać metodę Marquardta jako wariant metody Newtona-Raphsona. Należy podkreślić, że metoda Marquardta pozwala wyznaczyć minimum funkcji kryterium w przypadkach, gdy wcześniej opisane metody zawodzą, np. przy niezbyt dobrze określonych wartościach początkowych.

### Metody zmiennej metryki

W algorytmie Newtona-Raphsona do poszukiwania wartości wektora ocen współczynników  $\mathbf{b}$  zapewniających minimum niekwadratowej funkcji kryterialnej  $S(\mathbf{b})$  należy w każdej iteracji wyznaczyć macierz odwrotną jej hesjanu  $\mathbf{H}^{-1}$ . Jest to dość trudne zwłaszcza w przypadku dużej liczby zmiennych. Metody optymalizacji, w których macierz  $\mathbf{P}$  jest przybliżeniem macierzy odwrotnej hesjanu, zwane są metodami quasi-newtonowskimi lub metodami zmiennej metryki. Stosuje się następujące przybliżenie:

$$\mathbf{H}_{l+1}^{-1} \approx \omega \mathbf{P}_{l+1}, \quad (5.24)$$

przy czym:

$$\mathbf{P}_{l+1} = \mathbf{P}_l + \Delta \mathbf{P}_l, \quad (5.25)$$

zaś  $\omega$  jest współczynnikiem liczbowym zwanym współczynnikiem skali. Poszczególne algorytmy oparte na metodach quasi-newtonowskich różnią się między sobą sposobem wyznaczenia macierzy  $\Delta \mathbf{P}_l$ . Ogólnie przyjmuje się, że macierz:

$$\Delta \mathbf{P}_l = \frac{1}{\omega} \frac{\mathbf{c}_l \mathbf{r}_l^T}{\mathbf{r}_l^T \Delta \mathbf{g}_l} - \frac{\mathbf{P}_l \Delta \mathbf{g}_l \mathbf{z}_l^T}{\mathbf{z}_l^T \Delta \mathbf{g}_l}, \quad (5.26)$$

gdzie:

$$\mathbf{c}_l = \mathbf{b}_{l+1} - \mathbf{b}_l,$$

$$\Delta \mathbf{g}_l = \mathbf{g}_{l+1} - \mathbf{g}_l,$$

$\mathbf{r}_l$  i  $\mathbf{z}_l$  – dowolne niezerowe wektory.

Na przykład, w metodzie Davidona-Fletcher-Powella w pierwszej iteracji stosuje się  $\mathbf{P}_0 = \mathbf{I}$ . W następnych iteracjach przyjmuje się:

$$\omega = 1, \quad \mathbf{r}_l = \mathbf{c}_l, \quad \mathbf{z}_l = \mathbf{P}_l \Delta \mathbf{g}_l.$$

Wówczas zgodnie z wzorem (5.26) macierz korekcyjna:

$$\Delta \mathbf{P}_l = \frac{\mathbf{c}_l \mathbf{c}_l^T}{\mathbf{c}_l^T \Delta \mathbf{g}_l} - \frac{\mathbf{P}_l \Delta \mathbf{g}_l \Delta \mathbf{g}_l^T \mathbf{P}_l}{\Delta \mathbf{g}_l^T \mathbf{P}_l \Delta \mathbf{g}_l}. \quad (5.27)$$

Ważnym elementem algorytmu jest odnowa rozumiana jako podstawienie  $\mathbf{P}_l = \mathbf{I}$ , co prowadzi do wykorzystania kierunku największego spadku. Odnowę przeprowadza się wówczas, gdy wyznaczony w kolejnej iteracji kierunek nie jest kierunkiem poprawy oraz gdy od ostatniej odnowy wykonano  $2m+1$  iteracji.

Metody gradientowe są efektywne, lecz często mogą zawodzić w przypadku nieregularnej funkcji kryterium. Wtedy może być bardziej korzystne, skorzystać z którejś z metod bezgradientowej optymalizacji. W metodach bezgradientowych do wyznaczenia wartości optymalnych wystarczy znajomość funkcji kryterium w poszczególnych punktach.

### **Metoda Hooke'a-Jeevesa**

Na wstępie wyznacza się ortogonalną bazę  $\xi_1, \xi_2, \dots, \xi_m$  oraz długość kroku  $\lambda > 0$ . Kierunki  $\xi_1, \xi_2, \dots, \xi_m$  najczęściej są wersorami układu współrzędnych kartezjańskich. W każdej iteracji występują dwa etapy obliczeń: próbny i roboczy. W etapie próbnym bada się wartości funkcji kryterium w niewielkim otoczeniu wektora ocen  $\mathbf{b}_l$ , poprzez wykonanie próbnego kroku o długości  $\lambda$  we wszystkich kierunkach ortogonalnej bazy i poszukuje się kierunku, w którym występuje minimum funkcji kryterium. W etapie roboczym przechodzi się do nowego punktu  $\mathbf{b}_{l+1}$  będącego kolejnym przybliżeniem wektora ocen współczynników w kierunku wyznaczonym w etapie próbnym, wokół którego będzie realizowany kolejny etap próbny. Jeżeli w trakcie etapu próbnego nie uzyska się zmniejszenia funkcji kryterium, powtarza się etap próbny ze zmniejszoną długością kroku  $\lambda$ , aż zostanie spełniony warunek  $\lambda < \varepsilon$ , gdzie  $\varepsilon$  jest dokładnością obliczeń.

### **Metoda Rosenbrocka**

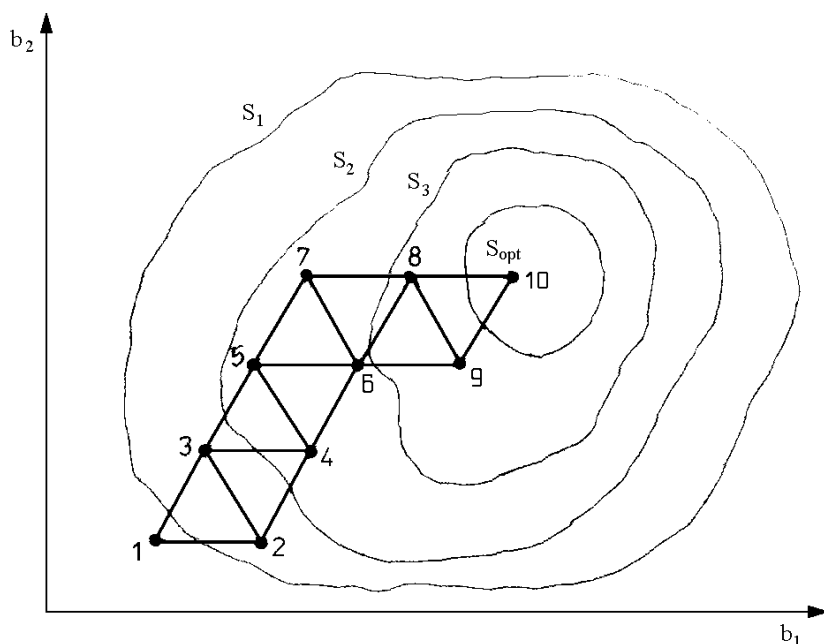
Podobnie jak w metodzie Hooke'a-Jeevesa na początku określa się ortogonalną bazę  $\xi_1, \xi_2, \dots, \xi_m$  oraz wektor współczynników kroku  $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_m]$ . W każdej iteracji wykonuje się kolejne kroki próbne we wszystkich kierunkach ortogonalnej bazy. Długość kroku próbnego w kierunku  $\xi_p$  równa jest odpowiedniej współrzędnej wektora kroku  $\lambda_p$ . Jeżeli w wyniku wykonania kroku w kierunku  $\xi_p$  występuje zmniejszenie wartości funkcji kryterium to odpowiadający mu współczynnik kroku  $\lambda_p$  jest zmieniany zgodnie z wzorem  $\lambda_p = \alpha \lambda_p$ ,  $\alpha > 1$ . W przeciwnym przypadku wartość tego współczynnika przyjmuje się  $\lambda_p = -\gamma \lambda_p$ ,  $\gamma \in (0, 1)$ . Jeżeli w trakcie kroków próbnych znaleziono przynajmniej jeden krok, w którym uzyskano zmniejszenie wartości funkcji kryterium, to wykonywana jest kolejna iteracja, przy tej samej bazie. W przeciwnym przypadku sprawdza się podstawowe kryterium stopu, czy  $\lambda_p = \varepsilon$ , dla  $p = 1, 2, \dots, m$ . W przypadku niespełnienia tego kryterium obraca się układ współrzędnych, w celu otrzymania nowego układu kierunków bazowych  $\xi_1, \xi_2, \dots, \xi_m$ , uwzględniając te kierunki dotychczasowej bazy, dla których otrzymano zmniejszenie wartości funkcji kryterium w poprzednich iteracjach.

### **Metoda sympleksowa**

Metoda sympleksowa zwana inaczej metodą Nelder-Meada polega na przemieszczaniu m-wymiarowego sympleksu określonego przez m+1 punktów  $\mathbf{b}_l$ :

$$\mathbf{b}_l = (b_{1l}, b_{2l}, \dots, b_{ml}),$$

należących do hiperpowierzchni optymalizowanej funkcji, w ten sposób, aby przesuwał się on w kierunku minimum. Decyzje o nowym położeniu punktów podejmowane są na podstawie wartości funkcji kryterium w aktualnych punktach, rys. 5.1.



Rys. 5.1. Ilustracja działania metody sympleksu w przestrzeni dwuwymiarowej,  $m = 2$

W każdej iteracji zmieniane jest położenie tylko jednego punktu, tego, w którym wartość funkcji kryterium jest największa (najgorszego punktu sympleksu). Niech liczby  $F_l$  oznaczają wartości funkcji w punktach  $\mathbf{b}_l$ . Punkt w którym wartość funkcji kryterium jest największa oznaczono przez  $\mathbf{b}_g$ . Punkt, w którym wartość funkcji jest najmniejsza oznaczono przez  $\mathbf{b}_h$ . Środek symetrii sympleksu (środek ciężkości sympleksu z pominiętym punktem  $\mathbf{b}_g$ ) oznaczono przez  $\bar{\mathbf{b}}$ :

$$\bar{\mathbf{b}} = \frac{1}{m-1} \sum_{l \neq g} \mathbf{b}_l . \quad (5.28)$$

Idea metody polega na kolejnych przekształceniach sympleksu (rys. 5.2) za pomocą czterech operacji:

- odbicia, punkt będący odbiciem punktu  $\mathbf{b}_g$  określony jest wzorem:

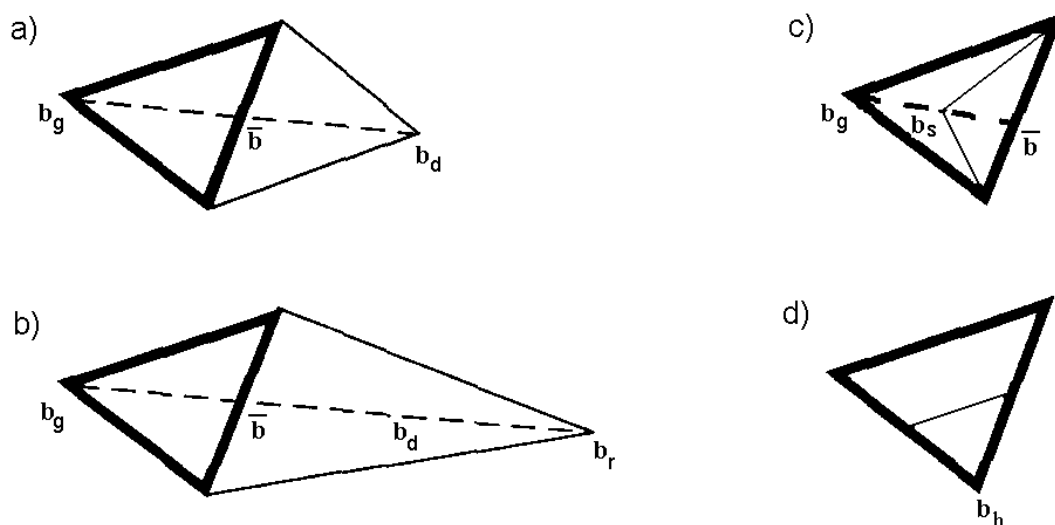
$$\mathbf{b}_d = \bar{\mathbf{b}} + \alpha(\bar{\mathbf{b}} - \mathbf{b}_g) , \quad (5.29)$$

gdzie:  $\alpha > 0$  – współczynnik odbicia,

- rozciągnięcia, punkt  $\mathbf{b}_g$  zastąpiony zostaje przez punkt  $\mathbf{b}_r$ :

$$\mathbf{b}_r = \gamma \mathbf{b}_d + (1 - \gamma) \bar{\mathbf{b}} , \quad (5.30)$$

gdzie:  $\gamma > 1$  – współczynnik rozciągnięcia,



Rys. 5.2. Przekształcenia początkowego sympleksu (oznaczonego grubą linią) za pomocą operacji: a) odbicia, b) rozciągnięcia, c) spłaszczenia, d) kontrakcji

- spłaszczenia, punkt  $\mathbf{b}_g$  zastąpiony zostaje przez punkt  $\mathbf{b}_s$  leżący między punktami  $\mathbf{b}_g$  i  $\bar{\mathbf{b}}$ :

$$\mathbf{b}_s = \beta \mathbf{b}_g + (1 - \beta) \bar{\mathbf{b}}, \quad (5.31)$$

gdzie:  $0 < \beta < 1$  – współczynnik spłaszczenia,

- kontrakcji, wszystkie punkty, oprócz punktu  $\mathbf{b}_h$ , w którym funkcja ma najmniejszą wartość, zostają przesunięte do środków odcinków łączących je z punktem  $\mathbf{b}_h$ :

$$\mathbf{b}_{ci} = (\mathbf{b}_l + \mathbf{b}_h) / 2, \quad l \neq h. \quad (5.32)$$

Początkowy sympleks tworzony jest wokół punktu startowego  $\mathbf{b}_0$  przez zmianę współrzędnych o 10%. Rodzaj dalszych przekształceń określa się wg następujących zasad:

1) jeżeli  $F_d < F_h$ , to najpierw próbuje się przekształcić sympleks za pomocą operacji rozciągania. Jeżeli w jej wyniku otrzyma się  $F_r < F_h$  to zostaje ona wykonana, w przeciwnym przypadku ( $F_r > F_h$ ) wykonuje się operację odbicia,

2) jeżeli  $F_d > F_{g1}$  i  $F_d < F_g$  to wykonuje się przekształcenie odbicia (gdzie  $F_{g1}$  oznacza punkt, w którym funkcja ma drugą co do wielkości wartość). W przeciwnym przypadku sympleks nie ulega zmianie. Następnie próbuje się poddać sympleks operacji spłaszczenia. Jeżeli w jej wyniku wartość funkcji w punkcie  $\mathbf{b}_s$  nie jest mniejsza niż w punktach  $\mathbf{b}_g$  i  $\bar{\mathbf{b}}$  to operacji spłaszczenia nie wykonuje się, a zamiast niej wykonuje się operację kontrakcji.

Przedstawione procedury poszukiwania ekstremum funkcji nieliniowych nie zapewniają w sposób automatyczny uzyskania zbieżności procesu iteracyjnego, ze względu na możliwość istnienia więcej niż jednego minimum funkcji kryterium oraz z powodu zmieniającego się kształtu funkcji w różnych punktach przestrzeni. Dlatego bardzo duże znaczenie ma



wybór wartości początkowych (startowych) współczynników oraz taka modyfikacja macierzy  $\mathbf{P}$  i długości kroku  $\lambda$ , aby zapewnić zbieżność procesu.

Proces iteracyjny optymalizacji można zakończyć, gdy wystąpi:

- 1) ustabilizowanie się wartości funkcji kryterium:

$$|S(\mathbf{b}_{l+1}) - S(\mathbf{b}_l)| < \varepsilon, \quad (5.33)$$

gdzie  $\varepsilon$  jest przyjętą dodatnią liczbą bliską zeru, np. 0,00001 nazywaną współczynnikiem zbieżności,

- 2) ustabilizowanie się wartości ocen współczynników:

$$|\mathbf{b}_{l+1} - \mathbf{b}_l| < \varepsilon, \quad (5.34)$$

gdzie  $\varepsilon$ , podobnie jak w poprzednim przypadku jest przyjętym współczynnikiem zbieżności.

Technicznym warunkiem zakończenia procesu iteracyjnego może być także wykonanie założonej liczby iteracji bez spełnienia żadnego z powyższych kryteriów. Może to oznaczać, że poruszano się w obszarze rozwiązań oddalonych od minimum funkcji kryterium i należy powtórzyć proces iteracyjny zmieniając wartości początkowe ocen współczynników lub zwiększyć liczbę iteracji, jeśli zbieżność jest zbyt wolna.

Stosowanie regresji nieliniowej wymaga wyboru odpowiedniego modelu zależności na podstawie dokładnej analizy teoretycznej badanego zjawiska. Wybór modelu jest decyzją merytoryczną, wynikającą z badanego zjawiska, a nie statystyczną. Podobnie wybór właściwych wartości początkowych współczynników wynika z własności badanego zjawiska i proponowanego modelu. Przy istnieniu trudności w określeniu „prawidłowych” wartości początkowych, zaleca się wykonać estymację dla kilku wartości początkowych rozmieszczonych równomiernie w przestrzeni współczynników. Jeżeli w wyniku tych kilku estymacji otrzyma się różne wyniki, jako rozwiązanie przyjmuje się to, dla którego wartość funkcji kryterium ma najmniejszą wartość oraz otrzymane wartości współczynników mają sens.

Aby uniknąć problemów numerycznych w obliczeniach, zaleca się tak przeskalować dane, aby wartości liczbowe były nie mniejsze niż  $10^{-4}$  i nie większe niż  $10^{+4}$ .

## 5.4. Obliczenia programem STATISTICA

W programie STATISTICA do estymacji nieliniowych zależności regresyjnych służy moduł *Estymacja nieliniowa*. Pozwala on na estymację współczynników różnych typów zależności regresyjnych przy pomocy metody najmniejszych kwadratów lub dowolnej innej określonej przez użytkownika. Najczęściej spotykane modele nieliniowej regresji zostały wstępnie zdefiniowane. Modele te obejmują regresję logit i probit, model regresji wykładniczej oraz regresji segmentowej. Moduł *Estymacja nieliniowa* wykorzystuje domyślnie jako funkcję kryterium (straty) metodę najmniejszych kwadratów. Jednak użytkownik ma możliwość określenia dowolnej funkcji straty i wykonać estymację, np. metodą bezwzględnych odchyłeń, metodą ważoną najmniejszych kwadratów, czy metodą

największej wiarygodności, itd. Istnieje możliwość wyboru odpowiedniej procedury estymacji umożliwiającej otrzymanie stabilnych ocen współczynników. Moduł zgłasza się oknem **Estymacja nieliniowa**, w którym można wybrać odpowiedni model zależności nieliniowej:



Rys. 5.3. Okno Estymacja nieliniowa

– **Regresja użytkownika** – użytkownik może wprowadzić dowolną postać zależności regresyjnej i funkcji kryterium (straty);

– **Regresja logistyczna** – przyjmuje się, że zmienna zależna  $y$  może przybierać wartości z przedziału  $(0, 1)$ . Model zależności określony jest wzorem:

$$y = \frac{\exp(\beta_0 + \beta_1 x_1 + \Lambda + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \Lambda + \beta_k x_k)} ; \quad (5.35)$$

– **Regresja probit** – umożliwia oszacowanie parametrów rozkładu normalnego  $\mu$  i  $\sigma$  na podstawie wartości probitów;

– **Regresja wykładnicza** – estymowana jest następująca zależność regresyjna:

$$y = c + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \Lambda + \beta_k x_k), \quad (5.36)$$

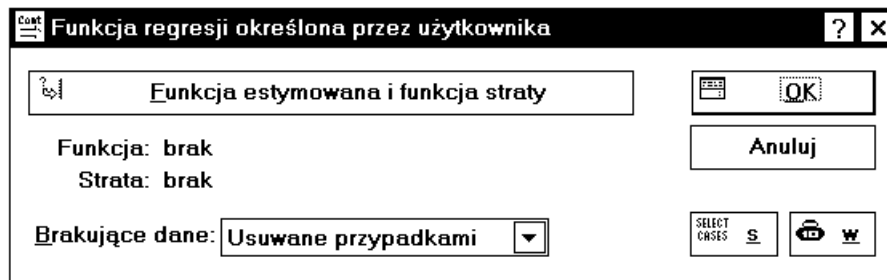
metodą najmniejszych kwadratów.

– **Regresja segmentowa** – będzie estymowana zależność odcinkami liniowa, określona wzorem:

$$y = (\beta_{01} + \beta_{11} x_1 + \beta_{21} x_2 + \Lambda + \beta_{k1} x_k) * (y \leq c) + (\beta_{02} + \beta_{12} x_1 + \beta_{22} x_2 + \Lambda + \beta_{k2} x_k) * (y > c)$$

Zatem będą estymowane dwa oddzielne równania regresji liniowej; jedno dla wartości zmiennej zależnej  $y$ , które są mniejsze lub równe wartości  $c$  (punktowi przełamania), a drugie dla tych wartości  $y$ , które są większe od tego punktu metodą najmniejszych kwadratów. Aby estymować model z punktami nieciągłości dla zmiennych niezależnych należy użyć opcji **Regresja użytkownika**.

Po wyborze tej opcji otwiera się okno **Funkcja regresji określona przez użytkownika** przedstawione na rys. 5.4.



Rys. 5.4. Okno Funkcja regresji określona przez użytkownika

W celu wprowadzenia funkcji regresji i funkcji straty należy nacisnąć przycisk **Funkcja estymowana i funkcja straty**. Otwiera się wówczas okno pokazane na rys. 5.5.

W górnym oknie wpisujemy proponowane równanie regresji według ogólnej zasady, że po lewej stronie znajduje się zmienna zależna, a po prawej stronie podane jest wyrażenie obejmujące zmienne niezależne i współczynniki. Zmienne mogą być wprowadzane przez podanie ich numeru w arkuszu danych (np. v1, v2,...) lub przez podanie ich specyficznych nazw. Przycisk **Zmienne** pozwala na obejrzenie zmiennych występujących w arkuszu danych. Nazwy zmiennych można wpisywać zarówno wielkimi jak i małymi literami. Wszystkie nazwy występujące w wyrażeniu, które nie zostaną przez program rozpoznane jako nazwy zmiennych lub jako zastrzeżone słowa kluczowe, zostaną zinterpretowane jako współczynniki.

Składnia równań regresji jest zgodna ze standardową notacją wyrażeń i obejmuje następujące operatory i funkcje:

– operatory:

arytmetyczne: +, -, \*, /, \*\* lub ^ (potęgowanie)

relacyjne: <, >, >=, <=, <>

logiczne: AND (&), OR (|), NOT (~)

– funkcje:

Abs(x) – wartość bezwzględna x

arcsin(x) – arcus sinus x

cos(x) – cosinus x

exp(x) – e do potęgi x

log(x) – logarytm naturalny z x

log2(x) – logarytm o podstawie 2 z x

log10(x) – logarytm dziesiętny z x

sign(x) – signum x:  $\begin{cases} +1, & \text{jeśli } x > 0 \\ 0, & \text{jeśli } x = 0 \\ -1, & \text{jeśli } x < 0 \end{cases}$

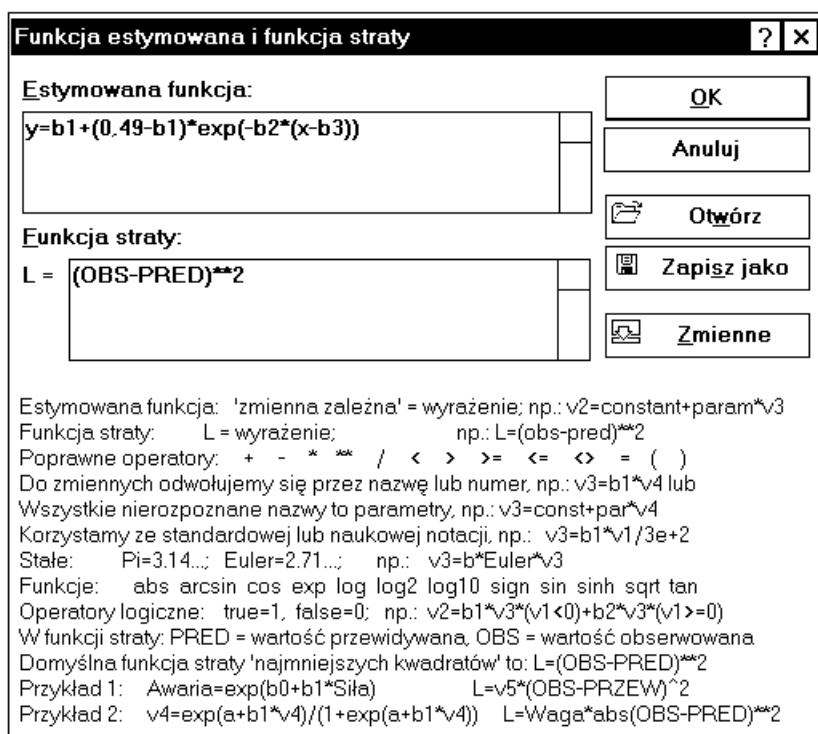
sin(x) – sinus x

sinh(x) – sinus hiperboliczny x

sqrt(x) – pierwiastek kwadratowy z x

tan(x) – tangens x

trunc(x) – część całkowita liczby x



Rys. 5.5. Okno Funkcja estymowana i funkcja straty

W oknie dolnym wpisuje się funkcję straty według tych samych zasad, jakie obowiązują przy wpisywaniu równania regresji. Dodatkowo dostępne są dwa słowa kluczowe **Pred** i **Obs**, które umożliwiają odwołanie się do odpowiednich wartości przewidywanych i obserwowanych zmiennej zależnej. Na przykład domyślna funkcja straty metody najmniejszych kwadratów jest określona jako:

$$L = (Obs - Pred) ** 2.$$

Równania określające funkcję regresji i funkcję straty mogą być zapisane w pliku po naciśnięciu przycisku **Zapisz jako**. Plik zapisany jest w standardowym formacie tekstowym i może być edytowany dowolnym edytorem. Zapisaną funkcję regresji i funkcję straty można wykorzystać po naciśnięciu przycisku **Otwórz**.

W oknie **Brakujące dane** podaje się sposób postępowania z brakującymi danymi. Brakujące dane mogą być usuwane przypadkami lub zastępowane średnimi odpowiednich zmiennych. Po zdefiniowaniu modelu regresji i naciśnięciu przycisku **OK** otworzy się okno **Procedura estymacji** (rys. 5.6), umożliwiające wybór metody estymacji, liczby iteracji, kryterium zbieżności, itd.

W polu **Metoda estymacji** określa się rodzaj procedury estymacji jaka ma być zastosowana. Można wybrać jedną z następujących metod:

- quasi-Newtona,
- sympleksu,
- sympleksu i quasi-Newtona,
- Hooke’a-Jeevesa przemieszczania układu,
- Hooke’a-Jeevesa i quasi-Newtona,

- Rosenbrocka poszukiwania układu,
- Rosenbrocka i quasi-Newtona.

Rys. 5.6. Okno Procedura estymacji

Ponieważ metody sympleksu, Hooke'a-Jeevesa i Rosenbrocka są mniej wrażliwe na minima lokalne, można stosować którąś z tych metod razem z metodą quasi-Newtona. Jest to szczególnie polecane, gdy nie ma pewności, co do właściwych wartości startowych (początkowych) estymacji. W tym przypadku pierwsza metoda pozwala na określenie wstępnych wartości współczynników, które następnie są wykorzystywane w metodzie quasi-Newtona.

Opcja **Asymptotyczne błędy standardowe** udostępnia ocenę odchyłeń standardowych estymatorów współczynników i macierz kowariancji estymatorów współczynników. Macierz kowariancji jest estymowana za pomocą macierzy odwrotnej hesjanu. Hesjan i asymptotyczne odchylenia standardowe współczynników obliczane są przez aproksymację ilorazu różnicowego.

Opcja **Eta dla aproksymacji skończ. różnic** jest dostępna, gdy zaznaczono opcję **Asymptotyczne błędy standardowe**.

**Maksymalna liczba iteracji** – pozwala określić maksymalną liczbę iteracji, która ma być wykonana.

**Kryterium zbieżności** – umożliwia określenie współczynnika zbieżności.

**Wartości początkowe** – umożliwia wprowadzenie wartości startowych (początkowych) współczynników. Domyślne wartości początkowe wszystkich współczynników są równe 0,1.

**Wstępna długość kroku** – umożliwia wprowadzenie wartości początkowych długości kroku dla każdego współczynnika lub jedną wspólną dla wszystkich. Domyślne wartości długości kroku wynoszą:

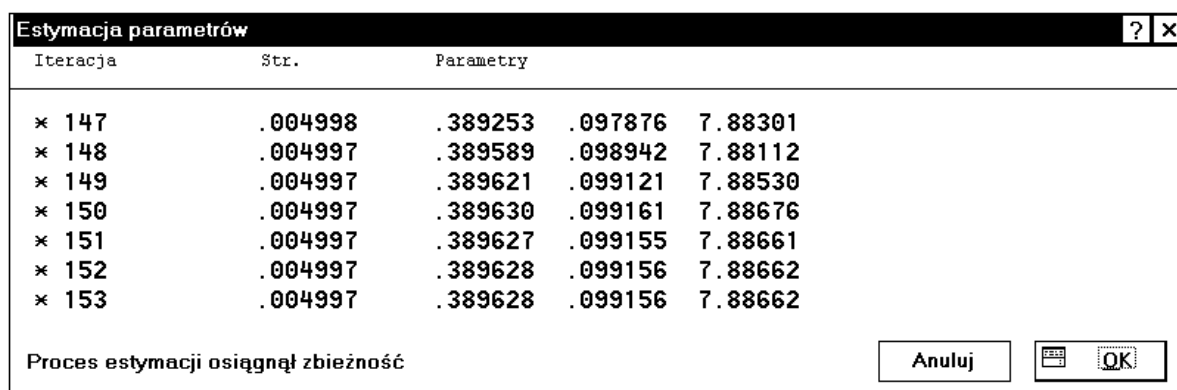
- 0,5 dla metody quasi-Newtona,
- 1,0 dla metody sympleksu i Rosenbrocka,
- 2,0 dla metody Hooke’a-Jeevesa.

**Średnie i odchylenia standardowe** – obliczane są średnie arytmetyczne i odchylenia standardowe dla wybranych zmiennych.

**Wykres macierzowy ogółu zmiennych** – przedstawiona zostaje macierz wykresów rozrzutu wszystkich zmiennych wybranych do analizy.

**Wykres ramkowy ogółu zmiennych** – zostają utworzone wykresy ramkowe z wąsami dla zmiennych wybranych do analizy.

Po wyborze metody estymacji, wartości początkowych oraz pozostałych parametrów i naciśnięciu przycisku **OK** rozpocznie się proces estymacji. W oknie **Estymacja parametrów** podawane są dla każdej iteracji aktualne wartości funkcji straty i wartości ocen współczynników (rys. 5.7).



The screenshot shows a dialog box titled "Estymacja parametrów" with a table of iteration results. The table has three columns: "Iteracja", "Str.", and "Parametry". The "Parametry" column contains five numerical values. The iterations range from 147 to 153. At the bottom of the dialog, there is a message "Proces estymacji osiągnął zbieżność" and two buttons: "Anuluj" and "OK".

Iteracja	Str.	Parametry
* 147	.004998	.389253 .097876 7.88301
* 148	.004997	.389589 .098942 7.88112
* 149	.004997	.389621 .099121 7.88530
* 150	.004997	.389630 .099161 7.88676
* 151	.004997	.389627 .099155 7.88661
* 152	.004997	.389628 .099156 7.88662
* 153	.004997	.389628 .099156 7.88662

Rys. 5.7. Okno *Estymacja parametrów*

Proces estymacji nieliniowej regresji wymaga zebrania obszernej informacji o naturze badanego obiektu i możliwej zależności między zmiennymi. Ułatwi to wybór modeli zależności oraz wartości początkowych współczynników, które będą na tyle bliskie wartościom rzeczywistym, że proces iteracyjny będzie zbieżny. Proces estymacji zależności nieliniowej obejmuje zwykle następujące kroki:

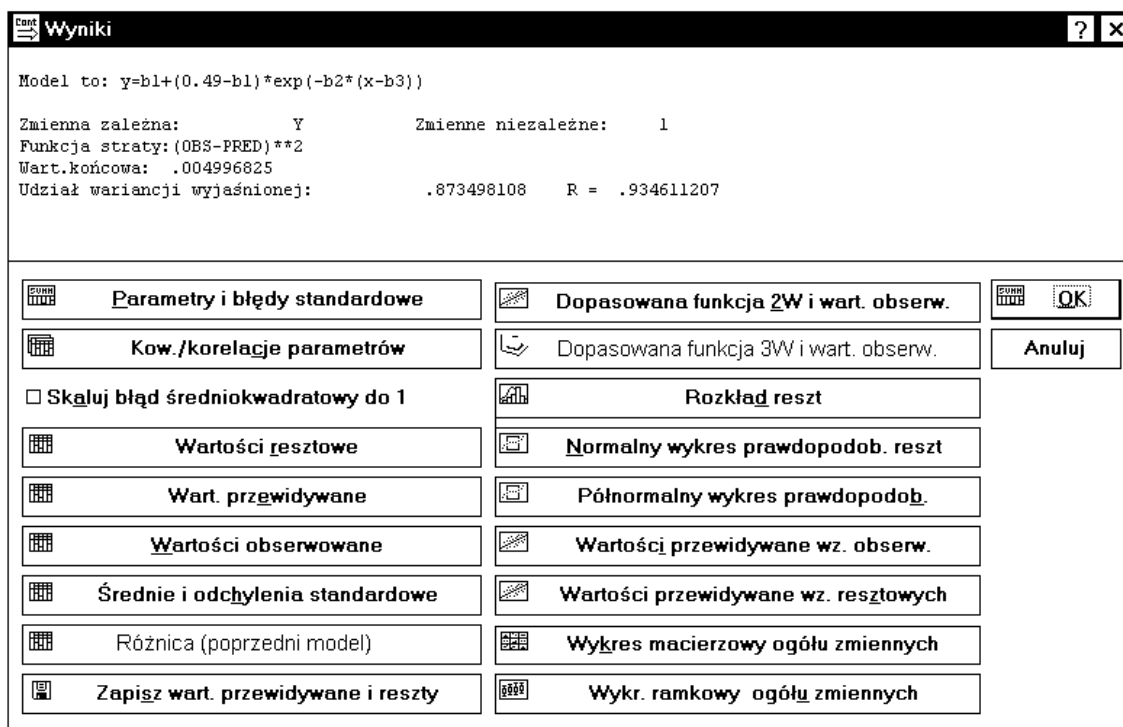
1) wykonuje się estymację metodą sympleks przy zadanych wartościach początkowych współczynników oraz przy domyślnych wartościach długości kroków i kryterium zbieżności.

2) jeżeli procedura iteracyjna nie osiągnie zbieżności zmienia się kryterium zbieżności na 0,1 i próbuje się zrealizować proces estymacji z różnymi wartościami początkowymi oraz zmniejsza się długości kroków. Jeżeli w trakcie obliczeń występują bardzo duże wartości funkcji straty (np.  $1E+37$ ), wskazuje to na rozbieżność procesu estymacji, i wówczas należy zmienić wartości początkowe.

3) Jeżeli poprzednie zabiegi nie przyniosły rezultatu, wykorzystuje się metodę Hooke’a-Jeevesa, a na końcu metodę Rosenbrocka.

4) jeżeli nie ma pewności, co do właściwych wartości początkowych, można łączyć metody estymacji sympleks, Hooke'a-Jeevesa lub Rosenbrocka z metodą quasi-Newtona. W tym przypadku pierwsza metoda określa wstępne wartości współczynników, które zostaną wykorzystane w metodzie quasi-Newtona.

Po pomyślnym zakończeniu procesu estymacji, naciśnięcie przycisku **OK**, spowoduje otwarcie okna **Wyniki** (rys. 5.8), w którym podane są sumaryczne wyniki analizy regresji oraz dostępne są następujące opcje do przeglądania szczegółowych wyników:



Rys. 5.8. Okno Wyniki

**Parametry i błędy standardowe** – otrzymuje się tabelę zawierającą wyestymowane współczynniki oraz, jeżeli w poprzednim oknie dialogowym zaznaczono opcję **Asymptotyczne błędy standardowe**, odchylenia standardowe estymatorów współczynników, wartości statystyk t i poziomy istotności p.

**Kow./Korelacje Parametrów** – otrzymuje się macierz kowariancji i korelacji estymatorów współczynników. Opcja jest dostępna, gdy w oknie **Procedura estymacji** zaznaczono opcję **Asymptotyczne błędy standardowe**.

**Wartości resztowe** – podawane są wartości reszt dla każdego przypadku.

**Wart. Przewidywane** – podawane są wartości przewidywane dla każdego przypadku.

**Wartości obserwowane** – podawane są wartości obserwowane dla każdego przypadku.

**Średnie i odchylenia standardowe** – podawane są średnie arytmetyczne i odchylenia standardowe dla wskazanych zmiennych.

**Różnica (poprzedni model)** – opcja jest dostępna tylko w przypadku regresji logit i probit, gdy aktualny model jest powiązany z modelem poprzednio estymowanym. Powiązanie polega na tym, że aktualny model różni się od poprzedniego tylko tym, że dołączono

lub usunięto z niego jedną lub więcej zmiennych niezależnych. W tym przypadku porównanie stopnia dopasowania tych dwóch modeli ma sens.

**Zapisz wart. przewidywane i reszty** – umożliwia zapisanie wartości przewidywanych i reszt w aktualnym pliku danych.

**Dopasowana funkcja 2W i wart. obserw.** – rysowany jest wykres rozrzutu wartości przewidywanych (oś Y) względem dowolnej wybranej zmiennej niezależnej (oś X).

**Dopasowana funkcja 3W i wart. obserw.** – rysowany jest wykres rozrzutu wartości przewidywanych (oś Z) względem dwóch zmiennych niezależnych (oś X i Y).

**Rozkład reszt** – otrzymuje się histogram liczebności wartości resztowych z zaznaczoną krzywą rozkładu normalnego.

**Normalny wykres prawdopodob. reszt** – wykres prawdopodobieństwa normalnego pozwala na wizualną ocenę zgodności rozkładu wartości resztowych z rozkładem normalnym.

**Półnormalny wykres prawdopodob.** – półnormalny wykres prawdopodobieństwa tworzony jest identycznie jak wykres prawdopodobieństwa normalnego z tą różnicą, że na osi Y przedstawiona jest tylko dodatnia część krzywej normalnej.

**Wartości przewidywane wz. obserw.** – wykres rozrzutu wartości obserwowanych (oś Y) względem wartości przewidywanych (oś X).

**Wartości przewidywane wz. resztowych** – wykres rozrzutu wartości reszt (oś Y) względem wartości przewidywanych (oś X). Na podstawie tego wykresu można ocenić adekwatność zależności regresyjnej. Jeżeli zależność jest adekwatna reszty powinny rozkładać się przypadkowo wokół wartości zerowej.

**Wykres macierzowy ogółu zmiennych** – wykresy rozrzutu między kilkoma wskazanymi zmiennymi przedstawione w formie macierzy wykresów.

**Wykr. ramkowy ogółu zmiennych** – wykresy ramkowe dla kilku wybranych zmiennych.

### Przykład 5.1

Zawartość chloru w pewnym produkcie (zmienna zależna Y) maleje w miarę upływu czasu (zmienna niezależna x). Przyjmuje się, że zależność ta może być opisana wzorem:

$$E(Y | x) = \beta_1 + (0,49 - \beta_1) \exp(-\beta_2(x - \beta_3)).$$

W wyniku obserwacji tego procesu otrzymano następujące dane doświadczalne:

Czas																		
$x_i$	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42
[tyg.]																		
Zaw. chloru	0,49	0,48	0,46	0,45	0,44	0,46	0,42	0,41	0,42	0,41	0,41	0,40	0,41	0,40	0,41	0,40	0,39	0,39
$y_i$		0,48	0,46	0,43	0,43		0,43	0,40	0,40	0,41		0,38			0,38	0,40		
		0,47	0,43															



### Rozwiązanie

Dane zapisano w pliku CHLOR.doc w dwóch kolumnach nadając im nazwy X i Y. Z **Przełącznika modułów** programu STATISTICA wybieramy moduł **Estymacja nieliniowa** a następnie opcję **Regresja użytkownika**. W zgłaszającym się oknie **Funkcja regresji określona przez użytkownika** naciskamy przycisk **Funkcja estymowana i funkcja straty** i wprowadzamy funkcję regresji i funkcję straty (stosujemy metodę najmniejszych kwadratów), rys. 5.5. Przyjmuje się, że brakujące dane będą usuwane przypadkami. Po naciśnięciu przycisku **OK** otwiera się okno **Procedura estymacji**, z podstawowymi informacjami o dotychczas przyjętych ustaleniach, umożliwiające wybór metody estymacji, rys. 5.6.

Procedura estymacji wymaga podania wartości początkowych współczynników  $b_1$ ,  $b_2$ ,  $b_3$ , od których rozpoczyna się poszukiwanie minimum funkcji  $S(\mathbf{b})$ . Te początkowe wartości nie powinny znacznie odbiegać od poszukiwanego rozwiązania. Stąd konieczna jest wstępna analiza danych umożliwiająca odpowiedni wybór wartości początkowych. W rozważanym przykładzie, w oparciu o dane literaturowe, jako wartości  $b_1$ ,  $b_2$ ,  $b_3$  przyjęto 1,0, 0,1, 6,0 i wprowadzono je po naciśnięciu przycisku **Wartości początkowe**. Jako metodę estymacji wybrano metodę sympleksu i quasi-Newtona. Z pozostałych opcji okna wybieramy: maksymalna liczba iteracji – 700; kryterium zbieżności – 0,0001; wstępna długość kroku – wartość domyślną 1,00 jednakową dla wszystkich współczynników oraz zaznaczamy opcję **Asymptotyczne błędy standardowe**.

Po naciśnięciu przycisku **OK** przechodzi się do procesu estymacji, rys. 5.7. Gdy proces estymacji osiągnie zbieżność naciskamy **OK** i przechodzimy do okna **Wyniki** (rys. 5.8), w którym podane są podstawowe informacje o rezultatach obliczeń. W celu uzyskania szczegółowych wyników wybieramy przycisk **Parametry i błędy standardowe** i otrzymuje się wyniki, które przedstawiono w tabeli 5.1.

Tabela 5.1

#### Wyniki nieliniowej regresji

Model: $y=b_1+(0,49-b_1)*\exp(-b_2*(x-b_3))$ (CHLOR.sta)			
Zmn. zal: (OBS-PRED)**2			
Końc. strata .004996825 R=.93461 Wyjaśniona warian.: 87,350%			
	B1	B2	B3
Ocena	0,389628	0,099156	7,886617
Błąd std.	0,004732	0,01603	0,551489
t(41)	82,34432	6,185628	14,30059
poziom p	0	2,36E-07	1,57E-17

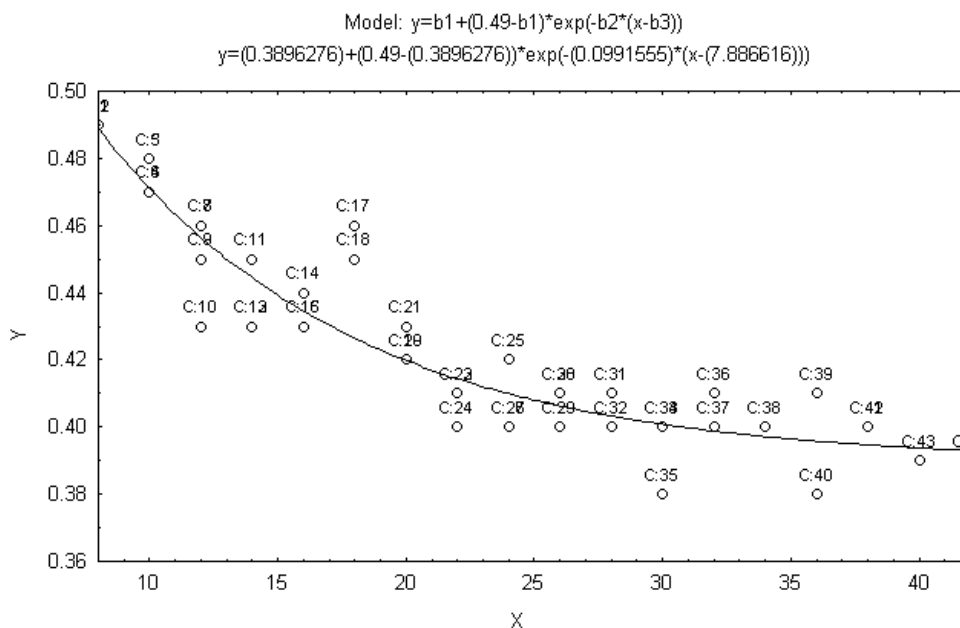
Oszacowaniami współczynników  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  są  $b_1=0,389628$ ,  $b_2=0,099156$ ,  $b_3=7,886617$ . Estymowane równanie regresji ma zatem postać:

$$\hat{Y} = 0,389628 + (0,49 - 0,389628) \exp(-0,099156(x - 7,886617)).$$

Współczynnik determinacji  $R^2=0,8735$  (wyjaśniona wariancja) ma wartość średnią, co świadczy o niezbyt dobrym dopasowaniu zależności do danych doświadczalnych. Może to wynikać z rozrzutu zmiennej zależnej Y lub nieadekwatności proponowanej zależności.

Obliczone wartości statystyki  $|t|$  znacznie przekraczają wartość krytyczną  $t_{0,05/2, 41}=2,019$  dla 41 stopni swobody przy poziomie istotności  $\alpha = 0,05$ . Należy więc odrzucić hipotezę o nieistotności współczynników  $\beta_1, \beta_2, \beta_3$ . Świadczą też o tym wartości  $p$ , które wszystkie są mniejsze od przyjętej wartości poziomu istotności  $\alpha$ . Wyestymowaną zależność dobrze jest przedstawić na wykresie. Wybierając przycisk **Dopasowana funkcja 2W i wart. obserw.** otrzymuje się wykres, rys. 5.9.

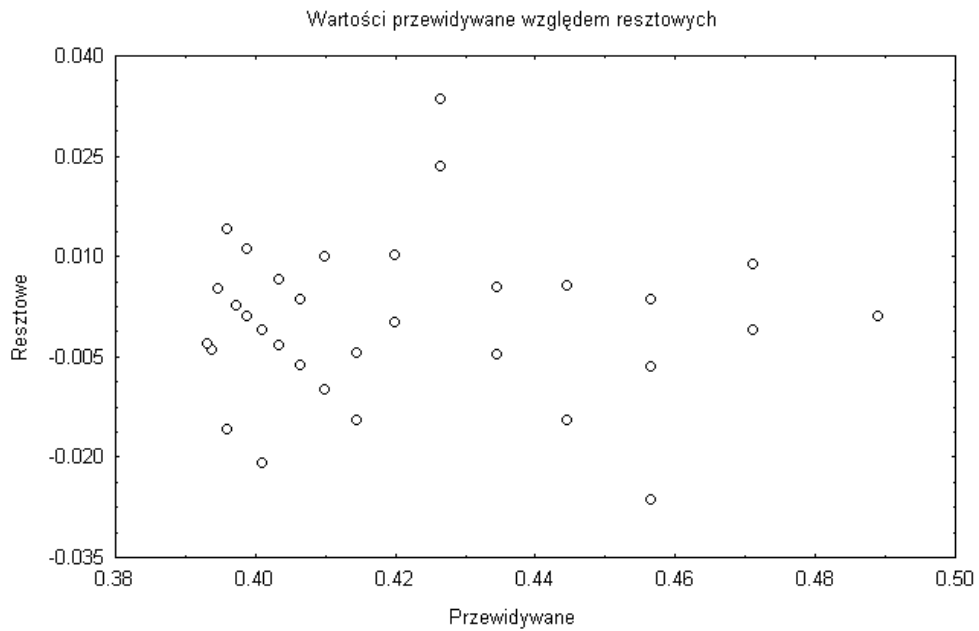
Z wykresu wynika, że punkty odzwierciedlające dane zaobserwowane generalnie rozkładają się losowo wokół wyestymowanej zależności, co może świadczyć o adekwatności zależności. Jednak wydaje się, że punkty oznaczone C17 i C18 odstają znacznie od linii regresji, co może oznaczać, że są to wartości odstające (obarczone błędem).



Rys. 5.9. Wykres zależności regresyjnej

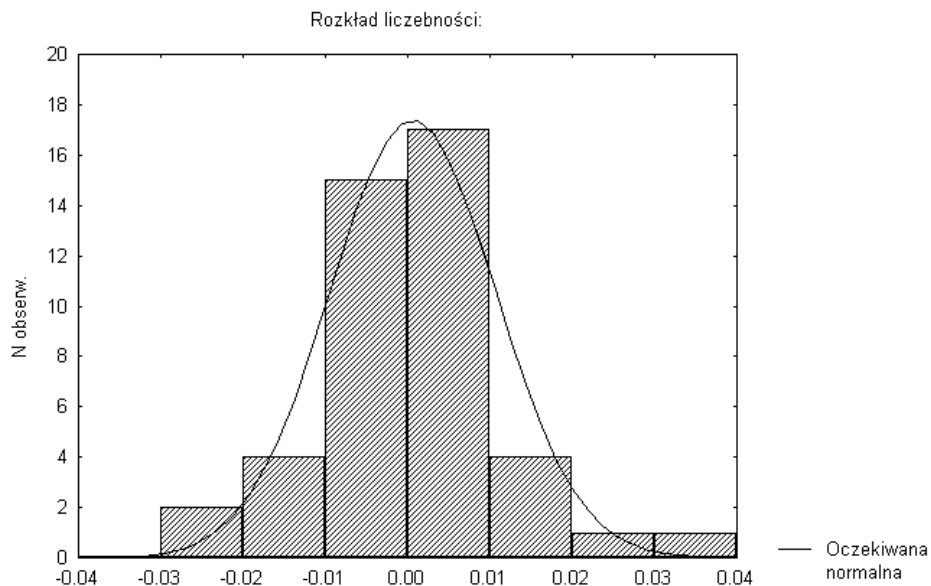
Dokładniejszą analizę można przeprowadzić w oparciu o analizę reszt. W analizie regresji przyjęto założenie o odchyleniach losowych, że są one niezależne, mają zerową wartość oczekiwaną, stałą wariancję i podlegają rozkładowi normalnemu. Jeżeli proponowany model zależności jest adekwatny, to reszty będące realizacjami odchyłeń losowych powinny wykazywać tendencje potwierdzające poczynione założenia lub przynajmniej nie wykazywać sprzeczności z założeniami.

W celu otrzymania wykresu reszt w funkcji wartości przewidywanych należy nacisnąć przycisk **Wartości przewidywane wz. resztowych**.



Rys. 5.10. Wykres reszt względem wartości przewidywanych

Wykres nie wykazuje żadnych szczególnych osobliwości wskazujących na nieadekwatność zależności. Potwierdza też, że punkty C17 i C18 odstają od pozostałych i powinny sprawdzić się czy w trakcie wykonywania doświadczeń nie popełniono błędu. Oceny słuszności założenia o rozkładzie normalnym odchyłek losowych dokonamy na podstawie histogramu liczebności reszt. W tym celu naciskamy przycisk **Rozkład reszt** i otrzymuje się następujący wykres:



Rys. 5.11. Histogram liczebności reszt

Wykres nie wykazuje istotnych różnic, wskazujących na sprzeczność z założeniem o normalności rozkładu.

# TABLICE STATYSTYCZNE

Tablica I

Wartości krytyczne rozkładu t-Studenta

Liczba stopni swobody, f	Poziom istotności $\alpha$	Postać hipotezy alternatywnej		
		H <sub>1</sub> :	H <sub>2</sub> :	H <sub>3</sub> :
3	0,10	2,353	-1,476	1,476
	0,05	3,182	-2,353	2,353
	0,01	5,841	-4,541	4,541
4	0,10	2,132	-1,533	1,533
	0,05	2,776	-2,132	2,132
	0,01	4,604	-3,747	3,747
5	0,10	2,015	-1,638	1,638
	0,05	2,571	-2,015	2,015
	0,01	4,032	-3,365	3,365
6	0,10	1,943	-1,440	1,440
	0,05	2,447	-1,943	1,943
	0,01	3,707	-3,143	3,143
7	0,10	1,895	-1,415	1,415
	0,05	2,365	-1,895	1,895
	0,01	3,499	-2,998	2,998
8	0,10	1,859	-1,397	1,397
	0,05	2,306	-1,859	1,859
	0,01	3,355	-2,897	2,897
9	0,10	1,833	-1,383	1,383
	0,05	2,262	-1,833	1,833
	0,01	3,250	-2,821	2,821
10	0,10	1,812	-1,372	1,372
	0,05	2,228	-1,812	1,812
	0,01	3,169	-2,764	2,764
11	0,10	1,795	-1,363	1,363
	0,05	2,201	-1,795	1,795
	0,01	3,106	-2,718	2,718
12	0,10	1,782	-1,356	1,356
	0,05	2,179	-1,782	1,782
	0,01	3,054	-2,681	2,681
13	0,10	1,771	-1,350	1,350
	0,05	2,160	-1,771	1,771
	0,01	3,012	-2,650	2,650
14	0,10	1,761	-1,345	1,345
	0,05	2,145	-1,761	1,761
	0,01	2,977	-2,624	2,624
15	0,10	1,753	-1,341	1,341
	0,05	2,131	-1,753	1,753
	0,01	2,947	-2,602	2,602
20	0,10	1,725	-1,325	1,325
	0,05	2,086	-1,725	1,725
	0,01	2,845	-2,528	2,528
25	0,10	1,708	-1,316	1,316
	0,05	2,060	-1,708	1,708
	0,01	2,787	-2,485	2,485

Tablica I cd.

Liczba stopni swobody, f	Poziom istotności $\alpha$	Postać hipotezy alternatywnej		
		H <sub>1</sub> :	H <sub>2</sub> :	H <sub>3</sub> :
30	0,10	1,697	-1,310	1,310
	0,05	2,042	-1,697	1,697
	0,01	2,750	-2,457	2,457
40	0,10	1,684	-1,303	1,303
	0,05	2,021	-1,684	1,684
	0,01	2,704	-2,423	2,423
50	0,10	1,676	-1,299	1,299
	0,05	2,009	-1,676	1,676
	0,01	2,678	-2,403	2,403

Tablica II

Wartości krytyczne rozkładu F-Snedecora dla poziomu istotności  $\alpha = 0,05$ 

Liczba stopni swobody mianownika	Liczba stopni swobody licznika													
	1	2	3	4	5	6	7	8	10	15	20	30	50	$\infty$
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,40	19,43	19,45	19,46	19,47	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,79	8,70	8,66	8,62	8,58	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	5,96	5,86	5,80	5,75	5,70	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,74	4,62	4,56	4,50	4,44	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,06	3,94	3,87	3,81	3,75	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,64	3,51	3,44	3,38	3,32	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,35	3,22	3,15	3,08	3,02	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,14	3,01	2,94	2,86	2,80	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	2,98	2,85	2,77	2,70	2,64	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,85	2,72	2,65	2,57	2,51	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,75	2,62	2,54	2,47	2,40	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,67	2,53	2,46	2,38	2,31	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,60	2,46	2,39	2,31	2,24	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,54	2,40	2,33	2,25	2,18	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,49	2,35	2,28	2,19	2,12	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,45	2,31	2,23	2,15	2,08	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,41	2,27	2,19	2,11	2,04	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,38	2,23	2,16	2,07	2,00	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,35	2,20	2,12	2,04	1,97	1,84
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,24	2,09	2,01	1,92	1,84	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,16	2,01	1,93	1,84	1,76	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,08	1,92	1,84	1,74	1,66	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	1,99	1,84	1,75	1,65	1,56	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,91	1,75	1,66	1,55	1,46	1,25
$\infty$	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,83	1,67	1,57	1,46	1,35	1,00

Tablica III

Wartości krytyczne rozkładu chi-kwadrat

Liczba stopni swobody, f	Poziom istotności		
	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$
1	2,706	3,841	6,635
2	4,605	5,991	9,210
3	6,251	7,815	11,345
4	7,779	9,488	13,277
5	9,236	11,071	15,086
6	10,645	12,592	16,812
7	12,017	14,067	18,475
8	13,362	15,507	20,090
9	14,684	16,919	21,666
10	15,987	18,307	23,209
11	17,275	19,675	24,725
12	18,549	21,026	26,217
13	19,812	22,362	27,688
14	21,064	23,685	29,141
15	22,307	24,996	30,578
16	23,542	26,296	32,000
17	24,769	27,587	33,409
18	25,989	28,869	34,805
19	27,204	30,144	36,191
20	28,412	31,410	37,566
21	29,615	32,671	38,932
22	30,813	33,924	40,289
23	32,007	35,172	41,638
24	33,196	36,415	42,980
25	34,382	37,652	44,314
26	35,563	38,885	45,642
27	36,744	40,113	46,963
28	37,916	41,337	48,278
29	39,087	42,557	49,588
30	40,256	43,773	50,892
31	41,422	44,985	52,191
32	42,585	46,194	53,486

Tablica IV

Wartości krytyczne testu Kolmogorowa

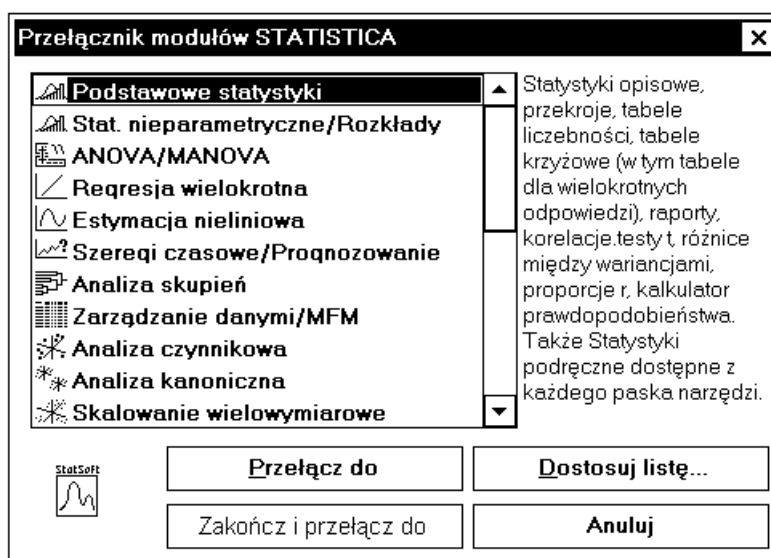
Liczebność próby	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,05$
5	0,50945	0,56328	0,66853
6	0,46799	0,51926	0,61661
7	0,43607	0,48342	0,57581
8	0,40962	0,45427	0,54179
9	0,38746	0,43001	0,51332
10	0,36866	0,40925	0,48893
11	0,35242	0,39122	0,46770
12	0,33815	0,37543	0,44905
13	0,32549	0,36143	0,43247
14	0,31417	0,34890	0,41762
15	0,30397	0,33760	0,40420
20	0,26473	0,29408	0,35241
25	0,23768	0,26404	0,31657
30	0,21756	0,24170	0,28987
35	0,20185	0,22425	0,26897
40	0,18913	0,21012	0,25205
Dla dużych n	$1,22/\sqrt{n}$	$1,36/\sqrt{n}$	$1,63/\sqrt{n}$

## STRUKTURA I ZASADY DZIAŁANIA PAKIETU STATISTICA PL

Program STATISTICA to zintegrowany pakiet programów służący do statystycznej analizy danych, tworzenia wykresów i zarządzania plikami danych. Składa się on z modułów, z których każdy zawiera grupę pokrewnych procedur obliczeniowych zarówno ogólnego zastosowania jak i specjalistycznych.

### *Uruchamianie programu STATISTICA*

Program uruchamia się klikając nazwę (ikonę) programu STATISTICA co powoduje otwarcie okna *Przełącznika modułów programu*:



Okno *Przełącznik modułów STATISTICA*

Przełącznik modułów pozwala na szybki wybór grup procedur obliczeniowych (modułów) dostępnych w danej wersji programu. Po wyborze odpowiedniego modułu zostaje on otwarty wraz z ostatnio używanym plikiem danych.

W każdym module można otworzyć tylko jeden plik danych.

### *Przygotowanie danych do analizy statystycznej*

Dane, dla których mają być wykonane obliczenia statystyczne przechowywane są w plikach z rozszerzeniem \*.sta. Po wywołaniu odpowiedniego modułu (procedury) obliczeniowego należy wskazać plik, który ma być analizowany. Organizacja pliku danych programu jest podobna do tabeli arkusza kalkulacyjnego, gdzie wiersze nazywane są przypadkami a kolumny zmiennymi. W jednym pliku danych może być do 4092 zmiennych i nieograniczona liczba przypadków.



## ***Tworzenie nowego pliku danych***

Pakiet STATISTICA umożliwia tworzenie nowych plików danych na dwa podstawowe sposoby:

- po uruchomieniu modułu obliczeniowego, wybranie menu **Plik**, gdzie znajduje się narzędzie łatwego tworzenia nowego pliku danych z domyślnymi ustawieniami, które potem użytkownik może łatwo modyfikować,
- uruchomienie modułu **Zarządzanie danymi** w przełączniku modułów.

Gdy wybierze się opcję **Nowe dane** w menu **Plik** otworzy się okno dialogu **Nowe dane: Określ nazwę pliku** umożliwiające podanie nazwy pliku. Nowo tworzony plik zgłasza się standardowo z 10 zmiennymi i z 10 przypadkami. Po otwarciu arkusza danych, aby zacząć wpisywać dane, klikamy myszą pierwszą komórkę arkusza. Zostanie ona podświetlona. Gdy zacznie się pisać, podświetlenie znika a pionowa kreska oznacza położenie kursora. Po wpisaniu wartości do komórki należy nacisnąć klawisz **Enter**. Aktywna komórka przesunie się do następnego wiersza. Kiedy w ten sposób dojdziemy do ostatniego wiersza i naciśniemy **Enter** kursor przejdzie do następnej kolumny. Jeśli liczba przypadków jest zbyt mała można ją zwiększyć wybierając opcję **Przypadki** w menu **Edycja** lub naciskając przycisk **Przypadki** na pasku narzędzi i wybrać opcję **Dodaj**. Następnie należy podać liczbę dodawanych przypadków oraz określić po jakiej pozycji mają być wstawione. Analogicznie postępuje się przy zmianie liczby zmiennych z tym, że wybiera się opcję **Zmienne** w menu **Edycja** lub naciska się przycisk **Zmienne** na pasku narzędzi. Po utworzeniu nowego pliku, zmienne przyjmują nazwy domyślne ZMN1, ZMN2 itd., liczby wyświetlane są w formacie 8.3 z przecinkiem oddzielającym miejsca dziesiętne (jeśli wprowadzimy kropkę w miejscu dziesiętnym program potraktuje tą wartość jako tekstową). Można te wartości domyślne zmienić, przez dwukrotne kliknięcie nazwy zmiennej w arkuszu danych dla aktualnie podświetlonej zmiennej, lub wybranie opcji **Specyfikacje zmiennej** z menu **Edycja-Zmienne** lub po naciśnięciu przycisku **Zmienne** na pasku narzędzi i otwarcie okna **Zmienna**. W tym oknie można zmienić nazwę aktualnej zmiennej, format wyświetlania i kategorię (typ) zmiennej. Oprócz zwykłej nazwy zmiennej można w oknie **Długa nazwa** podać bardziej szczegółowy opis zmiennej (uwagi, komentarz) pozwalający na łatwiejszą jej identyfikację. Okno **Długa nazwa** służy też do wprowadzania formuł pozwalających na przekształcenie zmiennej, przekodowywanie zmiennej oraz na tworzenie nowych zmiennych na podstawie już istniejących. Na przykład chcąc utworzyć nową zmienną jako kwadrat pierwszej zmiennej należy w otwartym oknie **Specyfikacje zmiennej** tej nowej zmiennej wpisać formułę:

$$= V1^2$$

zatwierdzić przyciskiem **OK** i wyrazić zgodę na przeliczenie zmiennej. Symbol  $v_i$  oznacza zmienną o numerze  $i$ , przy czym jeśli jako  $i$  poda się 0, to oznacza to numer przypadku. Zamiast używać skrótu  $v_i$  w formułach można stosować pełne nazwy zmiennych. W formułach można też używać funkcji wbudowanych w programie (dostępnych po kliknięciu przycisku **Funkcja**). Dla przykładu chcąc utworzyć zmienną będącą logarytmem naturalnym pierwszej zmiennej należy wpisać formułę:

$$= \log(v1)$$

Formuły muszą zawsze zaczynać się od znaku równości. Wprowadzenie etykiety zaczynającej się od znaku równości powoduje, że program przyjmuje, że jest to formuła i weryfikuje ją ze względu na poprawność składni. W formułach można używać:

– operatorów:

arytmetycznych: +, -, \*, /, \*\* lub ^ (potęgowanie),

relacyjnych: <, >, >=, <=, <> ,

logicznych: AND (&), OR (|), NOT (~),

– stałych: Pi = 3,14, Euler (e) = 2,71.,.

– funkcji:

Abs(x) – wartość bezwzględna x

arcsin(x) – arcus sinus x

cos(x) – cosinus x

exp(x) – e do potęgi x

log(x) – logarytm naturalny z x

log2(x) – logarytm o podstawie 2 z x

log10(x) – logarytm dziesiętny z x

max (x,y) – zwraca większą z x i y

min (x,y) zwraca mniejszą z x i y

rnd(x) – liczba losowa z zakresu 0 do x

sign(x) – signum x: 
$$\begin{cases} +1, & \text{jeśli } x > 0 \\ 0, & \text{jeśli } x = 0 \\ -1, & \text{jeśli } x < 0 \end{cases}$$

sin(x) – sinus x

sinh(x) – sinus hiperboliczny x

sqrt(x) – pierwiastek kwadratowy z x

tan(x) – tangens x

trunc(x) – część całkowita liczby x

Po wpisaniu wszystkich danych do arkusza należy je zapisać na dysk.

W module *Zarządzanie danymi* można także utworzyć nowy plik danych po podaniu szczegółowej specyfikacji pliku: liczby zmiennych, liczby przypadków, formatu itd.

### ***Wybór zmiennych do obliczeń***

Aby wskazać zmienne do analizy należy otworzyć okno wyboru zmiennych w danej procedurze obliczeniowej. Zmienne wybiera się albo przez ich podświetlenie, albo przez wprowadzenie numerów w oknie edycji. Jeśli wybiera się zmienne przez podświetlenie to w celu wskazania kilku zmiennych klikamy każdą zmienną trzymając wciśnięty klawisz CTRL. Jeśli nazwy zmiennych leżą obok siebie (tworzą ciągły blok) można je podświetlić przeciągając po nich kursorem przy wciśniętym lewym przycisku myszy.

## LITERATURA

1. Benjamin J., Cornell C. A.: Rachunek prawdopodobieństwa, statystyka matematyczna i teoria decyzji dla inżynierów, WNT, Warszawa 1977.
2. Bobrowski D.: Probabilistyka w zastosowaniach technicznych, WNT, Warszawa 1986.
3. Brandt S.: Analiza danych, PWN, Warszawa 1999.
4. Chow G. C.: Ekonometria, PWN, 1995.
5. Draper N. R., Smith H.: Analiza regresji stosowana, PWN, Warszawa 1973.
6. Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, Warszawa 2001.
7. Kryszwicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M.: Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, Część II Statystyka matematyczna, PWN, Warszawa 2000.
8. Gajek L., Kałużka M.: Wnioskowanie statystyczne. Modele i metody, WNT, Warszawa 2000.
9. Hellwig Z.: Elementy rachunku prawdopodobieństwa i statystyki matematycznej, PWN, Warszawa 1995.
10. Plucińska A., Pluciński E.: Rachunek Prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne, WNT, Warszawa 2000.
11. PN-ISO 2602:1994 – Estymacja wartości średniej.
12. PN-ISO 2854:1994 – Techniki estymacji oraz testy związane z wartościami średnimi i wariancjami.
13. PN-ISO 3301:1994 – Porównanie dwóch wartości średnich w przypadku obserwacji parami.
14. STATISTICA – dokumentacja pakietu, t. 1÷5, StatSoft Polska 1997.
15. Welfe A.: Ekonometria, PWE, Warszawa 1995.