

Stanisław Heilpern

OBSERWACJE NIETYPOWE – PRZYPADEK WIELOWYMIAROWY

1. Wstęp

Praca jest poświęcona badaniu nietypowych obserwacji wielowymiarowych. Jest to praca przeglądowa stanowiąca dalszy ciąg pracy [4] dotyczącej danych jednowymiarowych. W pracy będziemy się zajmować wykrywaniem wielowymiarowych obserwacji nietypowych, dokładnie – obserwacji odstających, oraz opiszemy odporne metody estymacji charakterystyk wielowymiarowych rozkładów opisujących populację. Przypomnijmy, że w [4] obserwacje nietypowe, różniące się od pozostałych, zostały podzielone na obserwacje, których wartości zostały błędnie podane, oraz na obserwacje *odstające*, poprawne, ale pochodzące z populacji innej niż zasadnicza część obserwacji, nazywana rdzeniem. W naszej pracy będziemy się zajmowali obserwacjami odstającymi.

Analiza wielowymiarowych obserwacji odstających jest istotnie trudniejsza niż ta, która dotyczy przypadku jednowymiarowego. Metod wykrywających obserwacje nietypowe stosowanych w przypadku danych jednowymiarowych nie można automatycznie stosować do danych wielowymiarowych. Wszelkie rzuty przestrzeni wielowymiarowych na przestrzenie o mniejszej liczbie wymiarów mogą ukryć występujące obserwacje nietypowe. Powinniśmy więc stosować przede wszystkim metody istotnie uwzględniające wszystkie wymiary danych, a kiedy jesteśmy zmuszeni do redukcji wymiarów, powinniśmy tracić jak najmniej informacji istotnych z punktu widzenia analizy obserwacji odstających. W pracy będziemy badać obserwacje odstające metodami statystycznymi, wykorzystując wnioskowanie statystyczne.

Praca składa się z dwóch części i dodatku zawierającego dwie tablice statystyczne umożliwiające znajdowanie wartości krytycznych przeprowadzanych te-

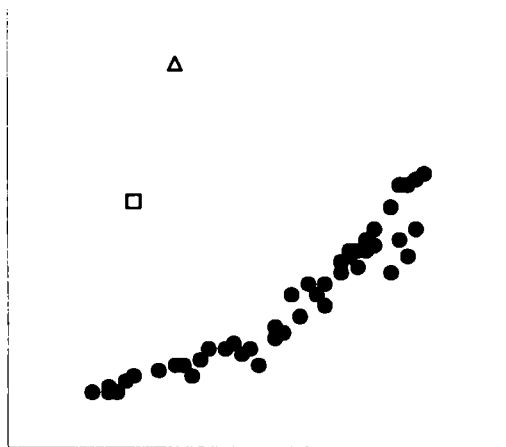
stów statystycznych. W pierwszej części przedstawimy wybrane metody wykrywania obserwacji odstających, a w drugiej opiszemy odporne metody wyznaczania estymatorów charakterystyk rozkładu populacji – metody uwzględniające obserwacje odstające, jednak słabo na nie reagujące.

Zamieszczone w pracy przykłady, oprócz przykładu 4, zawierają dane fikcyjne, służące jedynie do ilustracji przedstawionych metod.

Obliczenia dotyczące wyznaczania odległości Mahalanobisa, testu Shapiro-Wilka oraz metody Campbella zostały wykonane za pomocą pakietu komputerowego opartego na arkuszu kalkulacyjnym EXCEL, zawartego w [7].

2. Metody wykrywania wielowymiarowych obserwacji nietypowych

Jak już we wstępie wspomniano, metod wykrywających odstające obserwacje jednowymiarowe nie powinno się stosować w przypadku wielowymiarowym. Należy mieć na uwadze, że metody oparte na redukcji wymiarów nie wskażą wszystkich tego typu obserwacji. Rzuty przestrzeni wielowymiarowych na przestrzenie o mniejszej liczbie wymiarów mogą ukryć występujące obserwacje odstające.



Rys. 1. Wykres przykładowych danych dwuwymiarowych

Źródło: opracowanie własne.

Zaznaczone na rys. 1 obserwacje dwuwymiarowe tworzą wyraźny rdzeń; można na nim wyróżnić dwie obserwacje odstające. Obserwując jedynie rzuty tych

punktów na oś poziomą, nie możemy wskazać żadnej obserwacji odstającej. Rzuty zaś na oś pionową umożliwiają wychwycenie obserwacji zaznaczonej trójkątem. Obserwacja zaznaczona kwadratem jest odstająca, ale obydwie rzuty jej nie wskażą.

Pomimo tych wad analiza jednowymiarowa może stanowić jednak wstępną analizę poprzedzającą właściwe badanie otrzymanych danych. Jest bowiem znacznie prostsza od analizy wielowymiarowej.

Obserwacje m -wymiarowe będziemy traktować w tym punkcie jako wektory postaci:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}),$$

a zbiór obserwacji będzie tworzyć macierz obserwacji:

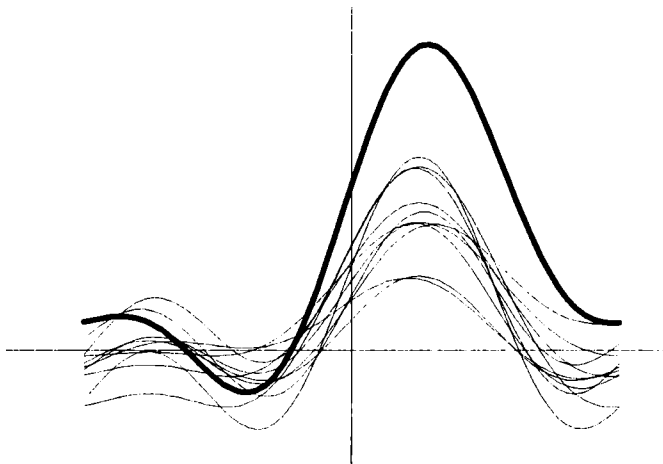
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}.$$

O powyższych uwagach należy więc pamiętać, chcąc metodami graficznymi zlokalizować obserwacje nietypowe. Spośród różnych prezentacji obserwacji wielowymiarowych na płaszczyźnie można wybrać *metodę krzywych Andrews* i na jej podstawie znaleźć tego typu obserwacje. W metodzie tej każda obserwacja m -wymiarowa jest traktowana jako pewna krzywa na płaszczyźnie [5; 11]. Krzywa Andrews wyraźnie różniąca się od pozostałych krzywych może być uważana za krzywą odpowiadającą obserwacji odstającej. Metoda krzywych Andrews może być więc wykorzystywana do wyszukiwania obserwacji tego typu. Jednak wadą tej metody jest zależność wyniku prezentacji od uporządkowania cech. Im mniejsza jest wartość indeksu wskazującego na kolejność danej cechy, tym większy jest jej wpływ na końcowy wynik. W celu znalezienia większej liczby obserwacji odstających powinniśmy stosować tę metodę do różnego uporządkowania cech. Na rys. 2 przedstawiono wykresy krzywych Andrews, które odpowiadają pewnym obserwacjom. Z rysunku można wywnioskować, że krzywa zaznaczona linią pogrubioną wskazuje na obserwację odstającą.

Do znajdowania obserwacji odstających można też wykorzystać to, że obserwacje te charakteryzują się względnie dużą odległością od środka ciężkości rdzenia rozpatrywanych obserwacji. W praktyce stosuje się w tym celu odległość Mahalanobisa D_i między poszczególnymi obserwacjami $\mathbf{x}_i$ a wektorem średnich $\bar{\mathbf{x}}$:

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

gdzie S jest macierzą kowariancji. Wektor średnich $\bar{x}$ jest w tym przypadku przybliżeniem środka ciężkości rdzenia obserwacji – przybliżeniem, ponieważ obliczając średnią z całości danych, bierzemy pod uwagę nie znane nam jeszcze obserwacje odstające [7].



Rys. 2. Krzywe Andrewsa

Źródło: opracowanie własne.

Gdy obserwacje pochodzą z próby losowanej z populacji o wielowymiarowym rozkładzie normalnym $N(\mu, \Sigma)$, możemy, chcąc znaleźć obserwację odstającą, skorzystać ze statystyki

$$F_i = \frac{(n-m)n}{(n^2-1)m} D_i^2.$$

Powyższa statystyka jest zmodyfikowaną odległością Mahalanobisa i ma rozkład F-Snedecora o m i $n-m$ stopniach swobody, czyli $F_i \sim F(m, n-m)$ [3; 5; 7]. W odniesieniu do dużych prób można skorzystać z tego, że odległość Mahalanobisa ma asymptotyczny rozkład chi-kwadrat z m stopniami swobody [8]. Obserwację x_i możemy uznać za odstającą, jeśli dla ustalonego poziomu istotności α wartość krytyczna F_α jest mniejsza niż wartość statystyki F_i . Spełniony jest wtedy warunek

$$P(F_i > F_\alpha) = \alpha.$$

W tym przypadku obserwacja odstająca jest interpretowana jako obserwacja tak daleko oddalona od centrum rdzenia, że praktycznie mało prawdopodobne jest jej uzyskanie w analizowanej próbie. Wielkość tego prawdopodobieństwa ustalamy subiektywnie, określając odpowiedni poziom istotności α .

Przykład 1

W tab. 1 przedstawiono dochody X i wydatki Y w tys. zł 33 losowo wybranych gospodarstw domowych w ustalonym miesiącu. Podano też wartości statystyki F mającej rozkład $F(2; 31)$. Wartość krytyczna na poziomie istotności $\alpha = 0,05$ jest równa $F_\alpha = 3,305$. Widzimy, że w odniesieniu do dwóch ostatnich obserwacji wartość statystyki F_i jest większa niż wartość krytyczna. Obserwacje te można uważać za odstające.

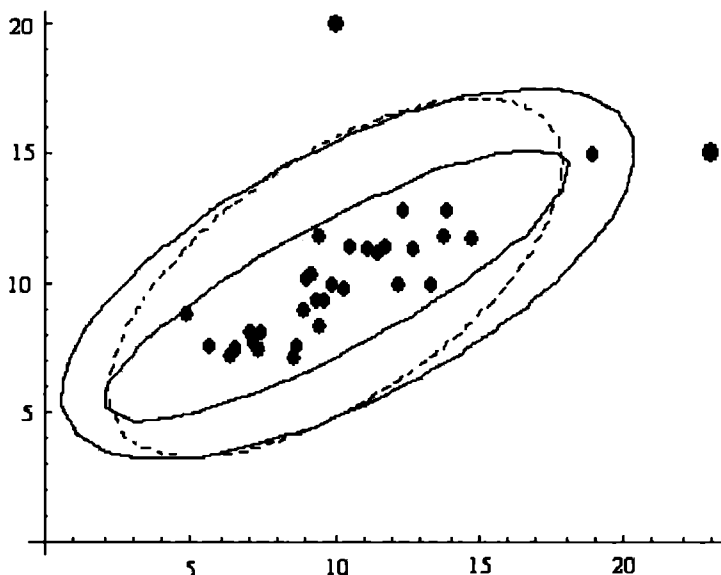
Tabela 1. Dochody i wydatki wybranych gospodarstw domowych

i	X	Y	F	i	X	Y	F	i	X	Y	F
1	9,4	8,4	0,280	13	9,6	9,3	0,073	25	6,5	7,5	0,640
2	10,5	11,5	0,142	14	13,9	12,8	0,487	26	4,8	8,9	1,293
3	12,4	12,9	0,425	15	14,8	11,7	0,739	27	7,4	8,1	0,390
4	5,6	7,6	0,831	16	13,3	10,0	0,643	28	11,5	11,2	0,052
5	12,2	10,0	0,281	17	11,1	11,4	0,075	29	11,7	11,5	0,088
6	8,6	7,6	0,517	18	13,8	11,8	0,402	30	7,3	7,5	0,545
7	6,3	7,2	0,745	19	9,0	10,2	0,104	31	20,0	15,0	2,593
8	8,5	7,1	0,747	20	9,2	10,4	0,100	32	23,0	15,0	5,771
9	8,9	9,0	0,124	21	7,1	7,7	0,509	33	10,0	20,0	11,169
10	10,3	9,8	0,029	22	7,0	8,1	0,449				
11	9,3	9,3	0,074	23	9,9	10,0	0,011				
12	12,7	11,3	0,184	24	9,4	11,9	0,517				

Źródło: opracowanie własne.

Rysunek 3 jest interpretacją graficzną danych z przykładu. Przedstawiona na nim elipsa jest zbiorem punktów, dla których wartość statystyki F jest równa wartości krytycznej $F_\alpha = 3,305$. Punkty leżące wewnątrz tej elipsy możemy interpretować jako obserwacje należące do rdzenia, a dwa punkty leżące na zewnątrz jako obserwacje odstające. Punkt przedstawiający obserwację x_{31} leży wewnątrz elipsy, ale bardzo blisko jej brzegu. Może on być również uważany za obserwację odstającą, ponieważ w tym przypadku występuje efekt przyciągania. Elipsa wyznaczona wartością F_α , określająca rdzeń obserwacji jest „przyciągana” przez obserwację x_{32} .

W razie usunięcia tej obserwacji elipsa, oznaczona przerywaną linią, nie obejmuje już obserwacji x_{31} (zob. rys. 3). Po usunięciu odstającej obserwacji x_{33} elipsa istotnie się zmniejszyła i leży wyraźnie wewnątrz elipsy utworzonej przez wszystkie obserwacje. W tej sytuacji punkt x_{31} nadal leży poza tą elipsą, ale dodatkowo dwa punkty: x_{24} i x_{26} , leżą blisko jej brzegu.



Rys. 3. Prezentacja graficzna danych z przykładu 1

Źródło: opracowanie własne.

Statystyka wykorzystywana w powyższej metodzie jest oparta na odległości Mahalanobisa. Podejście to można uogólnić, dopuszczając inne rodzaje odległości w przestrzeni wielowymiarowej. Odległości te porządkują obserwacje od najbliższych położonych od środka ciężkości, będącego w przypadku odległości Mahalanobisa wektorem średnich $\bar{x}$, do najdalszego. Wykrywanie obserwacji odstających sprowadza się wtedy w zasadzie do przypadku jednowymiarowego, ponieważ dana odległość przekształca przestrzeń wielowymiarową na dodatnią część prostej rzeczywistej R_+ . Otrzymujemy więc funkcję

$$\varphi(x_i) = D_i^2,$$

indukującą na wielowymiarowej przestrzeni R^m słaby porządek, relację preferencji $\preceq$, określoną zależnością

$$\mathbf{x}_i \preceq \mathbf{x}_j \Leftrightarrow \varphi(\mathbf{x}_i) \leq \varphi(\mathbf{x}_j).$$

Funkcja φ jest wtedy reprezentacją funkcyjną relacji preferencji $\preceq$. Obserwacje, których odległości od punktu ciężkości są takie same, są nierozróżnialne ze względu na tę relację preferencji.

Procedurę znajdowania obserwacji odstających możemy więc dalej uogólnić, opierając się na samej relacji preferencji porządkującej obserwacje. Nie musi ona wtedy być indukowana przez jakąś odległość, wystarczy jedynie zadać tę relację, np. określając jej reprezentację funkcyjną. Relacja ta porządkuje obserwacje w sposób liniowy z dokładnością do obserwacji nierozróżnialnych:

$$\mathbf{x}_{(1)} \preceq \mathbf{x}_{(2)} \preceq \dots \preceq \mathbf{x}_{(n)},$$

a reprezentacja funkcyjna przenosi obserwacje wielowymiarowe na prostą rzeczywistą. Możemy więc wtedy stosować metody znajdowania obserwacji odstających określone dla przypadku jednowymiarowego.

Reprezentacja funkcyjna φ relacji preferencji $\preceq$ może służyć do konstrukcji testu sprawdzającego, czy ekstremalna obserwacja $\mathbf{x}_{(n)}$ jest odstająca. Należy przy tym pamiętać, że obserwacja $\mathbf{x}_{(1)}$ leży najbliżej środka ciężkości, czyli jest jak najbardziej typowa. W przypadku testowania hipotezy, że cała próba pochodzi z wielowymiarowego rozkładu normalnego $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, przeciwko hipotezie alternatywnej, zakładającej, że istnieje obserwacja zaburzona o rozkładzie $N(\boldsymbol{\mu} + \mathbf{a}, \boldsymbol{\Sigma})$, gdzie $\mathbf{a}$ jest pewnym wektorem różnym od wektora zerowego, wykorzystujemy porządek generowany przez odległość Mahalanobisa $\varphi(\mathbf{x}_i) = D_i^2$. Sprawdzian testu przybiera wtedy postać [1]

$$\varphi_n = \varphi(\mathbf{x}_{(n)}) = \max_{1 \leq i \leq n} (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

Rozkład statystyki φ_n został stablicowany przez Wilksa. Wartości krytyczne dla $\alpha = 0,05$ oraz $\alpha = 0,01$ podano w tabl. I.

Obserwacja $\mathbf{x}_{(n)}$ może zostać uznana za odstającą, gdy hipoteza H_0 zostanie odrzucona na poziomie istotności $1 - \alpha$, tzn. jeśli wartość statystyki φ_n jest większa od wartości krytycznej φ_α odczytanej z tabl. I.

Powyższy test, podobnie jak testy dotyczące obserwacji jednowymiarowych przedstawionych w [4], opiera się na statystykach pozycyjnych. Jest to podejście istotnie odmienne od tych przedstawionych w metodzie opartej na statystyce F_i . Sprawdzamy tutaj, czy ekstremalna obserwacja jest na tyle oddalona od rdzenia,

aby można było uważać ją za odstającą, mimo że jest ekstremalna. Wartości krytyczne, powyżej których obserwacje uznajemy za odstające, zależą istotnie od wielkości próby. Im większa próba, tym trudniej obserwację ekstremalną uznać za odstającą, w tym przypadku nie należącą do rdzenia (zob tab. 2). Jest to zgodne z intuicją [1].

W metodzie opartej na statystyce F_i możemy zaobserwować tendencję odwrotną. Tam w miarę wzrostu próby zmniejsza się wartość krytyczna. W tab. 2 przedstawiono wartości krytyczne dla obydwu metod, gdy $p = 2$ oraz $\alpha = 0,05$. Dla metody opartej na statystykach pozycyjnych, wykorzystujących statystykę φ_n , wartości krytyczne zostały przemnożone przez $\frac{(n-2)n}{2(n^2-1)}$.

Tabela 2. Porównanie wartości krytycznych dla dwóch metod

n	φ	F	n	φ	F	n	φ	F
5	0,991	9,552	14	3,334	3,885	40	5,480	3,245
6	1,371	6,944	16	3,632	3,739	45	5,688	3,214
7	1,717	5,786	18	3,892	3,634	50	5,873	3,191
8	2,027	5,143	20	4,119	3,555	100	6,968	3,089
9	2,303	4,737	25	4,580	3,422	200	7,915	3,042
10	2,554	4,459	30	4,943	3,340	500	9,024	3,014
12	2,979	4,103	35	5,237	3,285			

Źródło: opracowanie własne.

Widzimy, że dla małych prób, gdy $n < 18$, metoda oparta na statystykach pozycyjnych łatwiej uznaje obserwacje za odstające. Natomiast dla prób większych mamy tendencję odwrotną.

Opierając się na macierzy $\mathbf{A}$ postaci

$$\mathbf{A} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

możemy badać też grupy obserwacji. Wyznacznik $|\mathbf{A}|$ nazywamy *wewnętrznym rozproszeniem* próby. Przy założeniu, że dane pochodzą z wielowymiarowego rozkładu normalnego rozpatrujemy ilorazy postaci

$$r_{i_1, i_2, \dots, i_r} = \frac{|\mathbf{A}^{(i_1, i_2, \dots, i_r)}|}{|\mathbf{A}|},$$

gdzie $|\mathbf{A}^{(i_1, i_2, \dots, i_s)}|$ jest wewnętrznym rozproszeniem otrzymanym po odrzuceniu s obserwacji $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_s}$ [1]. Sprawdzian testu jest oparty na statystyce

$$r_s = \min r_{i_1, i_2, \dots, i_s}.$$

W przypadku par obserwacji wykorzystujemy sprawdzian $\sqrt{r_2}$, również stabilizowany przez Wilksa, a jego wartości krytyczne c_α dla $\alpha = 0,05$ i $\alpha = 0,01$ zawiera tabl. II. Pary obserwacji możemy uznać za odstające, gdy wartość sprawdzianu testu będzie mniejsza od wartości krytycznej, czyli wtedy, gdy $\sqrt{r_2} < c_\alpha$. Statystyka r_s opiera się bowiem na najmniejszej wartości ilorazu $r_{i_1, i_2, \dots, i_s}$. Jest to więc test lewostronny.

Przykład 2

Rozpatrzmy dane z przykładu 1. Obserwacja $\mathbf{x}_{33}$ charakteryzuje się największą wartością odległości Mahalanobisa $\varphi_{33} = 23,76$. Wartość ta jest zdecydowanie większa od wartości krytycznej, wynoszącej $\varphi_\alpha = 11,1$ dla 35 obserwacji, gdzie poziom istotności $1 - \alpha = 0,95$. Obserwację $\mathbf{x}_{33}$ możemy więc uważać za odstającą.

Badając parę obserwacji $\mathbf{x}_{32}, \mathbf{x}_{33}$ podejrzanych o nietypowość, otrzymujemy macierz

$$\mathbf{A} = \begin{bmatrix} 444,46 & 216,34 \\ 216,34 & 240,20 \end{bmatrix},$$

której wyznacznik jest równy $|\mathbf{A}| = 59\,958,82$. Natomiast

$$\mathbf{A}^{(32,33)} = \begin{bmatrix} 281,37 & 156,41 \\ 156,41 & 119,00 \end{bmatrix},$$

a jej wyznacznik $|\mathbf{A}^{(32,33)}| = 9019,42$. Wartość sprawdzianu testu wynosi $\sqrt{r_2} = 0,388$ i jest mniejsza od wartości krytycznej $c_{0,05}$, wynoszącej 0,6842. Parę obserwacji $\mathbf{x}_{32}, \mathbf{x}_{33}$ możemy więc uważać za obserwacje odstające – co potwierdza wcześniejsze wyniki.

Grupy obserwacji możemy też badać, stosując testy sekwencyjne i usuwając po kolei ekstremalne obserwacje.

3. Estymacja odporna

W punkcie tym będziemy się zajmować estymacją odporną wektora parametrów położenia $\mu = [\mu_1, \mu_2, \dots, \mu_m]$ oraz macierzy kowariancji $\Sigma = [\sigma_{ij}]_{m \times m}$. Estymatory wektora parametrów położenia mogą być tworzone analogicznie do przypadku jednowymiarowego [4]. Najprostszy jest wektor median

$$\mathbf{Me} = [Me_1, Me_2, \dots, Me_m],$$

gdzie Me_j jest medianą obserwacji $x_{1j}, x_{2j}, \dots, x_{nj}$. W podobny sposób otrzymujemy wektory kwantyli $\mathbf{Q}_q$, obciętych średnich $\bar{x}_{\alpha;\beta}$, średnich Winsora $\mathbf{W}_{\alpha;\beta}$ czy różnego typu L -estymatorów, czyli liniowych kombinacji statystyk pozycyjnych.

Również odporne miary rozproszenia, elementy leżące na przekątnych macierzy kowariancji, można otrzymać, przenosząc na przypadek wielowymiarowy estymatory jednowymiarowe. Uzyskamy wtedy wektor median odchylenia bezwzględnego $\mathbf{MAD}$, odchyłeń ćwiartkowych $\mathbf{Q}$, obciętych standardowych odchyłeń $s_{\alpha;\beta}$, czy różnych L -estymatorów. Przez podzielenie wektora $\mathbf{MAD}$ oraz $\mathbf{Q}$ przez stałą 0,6745 otrzymujemy w przypadku rozkładu normalnego odpowiedniki klasycznego odchylenia standardowego [5]. Jako odporny estymator kowariancji dwóch zmiennych losowych X_i oraz X_j Huber [1] zaproponował statystykę

$$\frac{S(aX_i + bX_j) - S(aX_i - bX_j)}{S(aX_i + bX_j) + S(aX_i - bX_j)} S(X_i) S(X_j),$$

gdzie $S(X)$ jest odpornym estymatorem wariancji, np. obcięta wariancja lub wariancja Winsora, zmiennej X , a a i b są pewnymi stałymi. Często wybiera się

$$a = 1/S(X_i) \quad \text{oraz} \quad b = 1/S(X_j).$$

Do wyznaczania odpornych estymatorów położenia możemy też wykorzystać relację preferencji $\preceq$ porządkującą obserwacje oraz jej reprezentację funkcyjną φ . Odrzucając obserwacje ekstremalne, ze względu na relację $\preceq$, otrzymujemy wielowymiarowe wersje uciętych i typu Winsora estymatorów wektora średnich μ i macierzy kowariancji Σ . Gdy reprezentacja funkcyjna jest kwadratem odległości Mahalanobisa, możemy również skorzystać z działającej na podobnej zasadzie metody Gnanadesikana-Kettenringa [7]. Jest to iteracyjna metoda opisowa nie wymagająca znajomości rozkładu populacji. Analiza obserwacji odstających opiera się na odległości Mahalanobisa.

W pierwszym kroku iteracji obliczamy z próby wektor średnich $\bar{\mathbf{x}}^{(1)}$ oraz macierz kowariancji $\mathbf{S}^{(1)}$. Następnie usuwamy obserwację najdalej oddaloną od wektora $\bar{\mathbf{x}}^{(1)}$ i obliczamy nowe parametry $\bar{\mathbf{x}}^{(2)}$ i $\mathbf{S}^{(2)}$. Procedurę usuwania najbardziej oddalonych obserwacji powtarzamy, aż wyznaczone wielkości $\bar{\mathbf{x}}^{(r)}$ i $\mathbf{S}^{(r)}$ będą niewiele się różnić od wcześniej obliczonych. Otrzymane w r -tej iteracji wielkości uważamy za odporne estymatory wektora średnich $\boldsymbol{\mu}$ i macierzy kowariancji $\mathbf{S}$.

Innym podejściem jest metoda Campbella, przypisująca poszczególnym obserwacjom wagi w_i charakteryzujące ich typowość. Obserwacje należące do rdzenia otrzymują wagę 1, a odstające – wartości bliskie 0. Wagi te możemy interpretować jako stopień przynależności obserwacji do rdzenia traktowanego jako rozmyty zbiór obserwacji.

Stosując metodę Campbella, nie eliminujemy obserwacji odstających [1; 7]. Wszystkie obserwacje biorą udział w dalszej analizie z uwzględnieniem oczywiście przypisanych im wag. Wykorzystujemy wtedy wspomniane już wcześniej estymatory ważone. Metoda ta może też również służyć do eliminacji obserwacji odstających. Za obserwacje odstające możemy wtedy uważać obserwacje, których wagi są mniejsze od ustalonego w sposób subiektywny progu w_0 , np. $w_0 = 0,1$.

Wagi przypisane obserwacjom zależą od odległości Mahalanobisa tych obserwacji od środka ciężkości zbioru. Środek ciężkości wyznaczamy, uwzględniając wagi, według wzoru:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i}.$$

W podobny sposób wyznaczamy macierz kowariancji

$$\mathbf{S} = \frac{\sum_{i=1}^n w_i^2 (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T}{\sum_{i=1}^n w_i^2 - 1}.$$

Metoda Campbella jest metodą iteracyjną. W pierwszym kroku zakładamy, że wszystkie obserwacje należą do rdzenia, czyli $w_i = 1$ dla każdego $i = 1, 2, \dots, n$. W następnych iteracjach wagi są modyfikowane według wzoru:

$$w_i = \frac{w(d_i)}{d_i},$$

gdzie

$$d_i = \begin{cases} \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})}, & \text{gdzie } i = 1, 2, \dots, n, \\ \sqrt{m} + \sqrt{2}, & \text{gdzie } i = 0, \end{cases}$$

oraz

$$w(d_i) = \begin{cases} d_i, & \text{gdzie } d_i \leq d_0, \\ d_0 \exp(-(d_i - d_0)^2 / 2, 5), & \text{gdzie } d_i > d_0. \end{cases}$$

Gdy obserwujemy stabilizację wartości wag, iterację kończymy. Można stosować również inne metody modyfikacji wag, jednak powyższa metoda jest stosowana najczęściej.

Przykład 3

Dane z przykładu 1. Wektor wartości średnich $\bar{\mathbf{x}} = (10,46; 10,36)$, a macierz korelacji wynosi

$$\mathbf{S} = \begin{bmatrix} 14,45 & 6,93 \\ 6,93 & 7,52 \end{bmatrix}.$$

Po zastosowaniu metody Campbella otrzymujemy następujące wagi: $w_i = 1$ dla $i \leq 30$, $w_{31} = 0,74$, $w_{32} = 0,13$ oraz $w_{33} = 0$. Wagi te możemy interpretować jako stopnie przynależenia do rdzenia zbioru obserwacji. Obserwacja w_{33} na pewno do niego nie należy.

Po uwzględnieniu uzyskanych wag otrzymujemy następujący zmodyfikowany wektor średnich $\bar{\mathbf{x}}_w = (10,04; 9,88)$ oraz macierz kowariancji

$$\mathbf{S}_w = \begin{bmatrix} 8,72 & 4,75 \\ 4,75 & 3,65 \end{bmatrix}.$$

Rousseeuw i van Zomeren w [10] przedstawili następujące podejście do odpornej estymacji charakterystyk rozkładów wielowymiarowych. Zadanie polega na wyznaczeniu estymatorów położenia $\mathbf{T}(\mathbf{X})$ i rozproszenia $\mathbf{C}(\mathbf{X})$ n wielowymiaro-

wych danych $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, gdzie $\mathbf{x}_i$ jest m -wymiarowym wektorem. Zakłada się, że estymatory te powinny być *afinicznie niezmienniczymi*, tzn. spełniać zależności:

$$\mathbf{T}(\mathbf{x}_1\mathbf{A} + \mathbf{b}, \dots, \mathbf{x}_n\mathbf{A} + \mathbf{b}) = \mathbf{T}(\mathbf{x}_1, \dots, \mathbf{x}_n)\mathbf{A} + \mathbf{b}$$

oraz

$$\mathbf{C}(\mathbf{x}_1\mathbf{A} + \mathbf{b}, \dots, \mathbf{x}_n\mathbf{A} + \mathbf{b}) = \mathbf{A}\mathbf{C}(\mathbf{x}_1, \dots, \mathbf{x}_n)\mathbf{A}^T$$

dla każdego wektora $\mathbf{b}$ i m -wymiarowej, nieosobliwej macierzy kwadratowej $\mathbf{A}$. Przykładami afinicznie niezmienniczych estymatorów są wektor średnich $\mathbf{T}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ i macierz kowariancji $\mathbf{C}(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{T}(\mathbf{X}))(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))^T$ otrzymane z próby. Nie są to oczywiście estymatory odporne.

Afinicznie niezmienniczymi estymatorami odpornymi są natomiast estymatory otrzymane metodą *elipsoidy o najmniejszej objętości* (MVE) [10]. Są to estymatory położenia $\mathbf{T}$ oraz rozproszenia $\mathbf{C}$ takie, że spośród wszystkich estymatorów afinicznie niezmienniczych spełniających ograniczenia

$$\#\{i: (\mathbf{x}_i - \mathbf{T})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{T}) \leq a^2\} \geq h,$$

macierz $\mathbf{C}$ ma najmniejszy wyznacznik. Tak utworzone estymatory $\mathbf{T}$ i $\mathbf{C}$ wyznaczają elipsoidę o najmniejszej objętości, zawierającą co najmniej h elementów z próby $\mathbf{X}$.

Przyjmując, że próba $X_1, \dots, X_n$ pochodzi z populacji o m -wymiarowym rozkładzie eliptycznym $P_{\mu, \Sigma}$, dobieramy stałą a tak, aby spełniony był warunek [6]

$$P_{\mu, \Sigma} = \left((\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \leq a^2 \right) = 0,5.$$

Nieprecyzyjnie mówiąc, tak utworzona elipsoida zawiera połowę elementów populacji. W przypadku populacji o rozkładzie normalnym przyjmuje się $a^2 = \chi_{0,5}^2[m]$, medianę rozkładu chi-kwadrat z m stopniami swobody. Druga zaś stała h , ograniczająca liczbę elementów należących do tej elipsy, powinna być większa od połowy liczebności próby. Punkt załamania estymatorów MVE jest wtedy równy $[(n - h)/n]$ [9]. Autorzy pracy [10] proponują przyjąć, że $h = [0,5(n + m + 1)]$, gdzie $[q]$ oznacza całkowitą część q . W tym przypadku punkt załamania wynosi $[0,5(n - m + 1)]/n$ i jest największy z możliwych [6].

Innymi estymatorami położenia i rozproszenia są estymatory otrzymane metodą *najmniejszego wyznacznika macierzy kowariancji* (MCD). Metoda ta jest podobna do MVE, a polega na znalezieniu h obserwacji, dla których macierz kowa-

riancji ma najmniejszy wyznacznik. Estymatorem położenia jest wtedy wektor średnich dla wybranych h punktów. Estymatory wybrane tą metodą mają punkt załamania ten sam co w przypadku metody MVE. Jednakże ich zaletą jest asymptotyczna normalność.

Przedstawione powyżej odporne estymatory wyznacza się praktycznie, stosując metody kombinatoryczne. Dwa algorytmy wyznaczające przybliżone rozwiązanie metody MVE podano w [10], a w [9] zawarto algorytm otrzymywania estymatorów MCD. Ten ostatni algorytm opiera się głównie na następującym twierdzeniu.

Twierdzenie 1. [9] Niech $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ będzie zbiorem m -wymiarowych obserwacji, a H_1 h -elementowym podzbiorem $\{1, \dots, n\}$. Dla H_1 wyznaczamy estymatory położenia i rozproszenia: $\mathbf{T}_1 = \frac{1}{h} \sum_{i \in H_1} \mathbf{x}_i$ oraz $\mathbf{S}_1 = \frac{1}{h} \sum_{i \in H_1} (\mathbf{x}_i - \mathbf{T}_1)(\mathbf{x}_i - \mathbf{T}_1)^T$

oraz dla każdego $\mathbf{x}_i$, gdzie $i = 1, \dots, n$, odległości Mahalanobisa d_i wyznaczone przez wektor $\mathbf{T}_1$ i macierz $\mathbf{S}_1$. Otrzymane odległości d_i porządkujemy i wybieramy h najmniejszych $d_{i_1} \leq \dots \leq d_{i_h}$. Niech $H_2 = \{i_1, \dots, i_h\}$, a $\mathbf{T}_2$ i $\mathbf{S}_2$ będą estymatorami wyznaczonymi przez H_2 ; wtedy

$$\det(\mathbf{S}_2) \leq \det(\mathbf{S}_1).$$

Równość zachodzi jedynie, gdy $\mathbf{T}_1 = \mathbf{T}_2$ oraz $\mathbf{S}_1 = \mathbf{S}_2$.

Estymatorami MCD położenia i rozproszenia dla danych z przykładu 1 i $h = 3$ są estymatory $\mathbf{T}$ i $\mathbf{S}$ otrzymane dla zbioru obserwacji po odrzuceniu $\mathbf{x}_{31}$, $\mathbf{x}_{32}$ i $\mathbf{x}_{33}$. Wyznacznik macierzy $\mathbf{S}$ jest najmniejszy spośród wszystkich macierzy kowariancji wyznaczonych dla 30-elementowych zbiorów obserwacji.

Na zakończenie przeprowadzimy analizę danych dotyczących wydatków gospodarstw domowych ze względu na obserwacje odstające.

Przykład 4

W tab. 3 przedstawiono dane dotyczące wydatków 82 losowo wybranych gospodarstw domowych w Polsce w 1993 roku na żywność (X_1), utrzymanie domu (X_2) oraz kulturę i edukację (X_3). Dane pochodzą z raportu RAD Project „Powerty and Targeting od Social Assistance in Eastern Europe and Former Soviet Union” Banku Światowego. Pełna próba liczyła 16 051 gospodarstw domowych.

Nasze badanie zaczniemy od analizy poszczególnych zmiennych, wykorzystując metody przedstawione w [4]. Na początku sprawdzono, że zmienne X_1 oraz X_2 mają rozkład normalny, a zmienna X_3 logarytmiczno-normalny. Dla zmiennych X_1 oraz X_2 nie stwierdzono obserwacji odstających. Sprawdzian testu Grubsa przybiera wartości mniejsze od krytycznych dla $\alpha = 0,05$: $G_{(82)} = 2,72$ i $G_{(1)} = 2,01$ dla

zmiennej X_1 oraz $G_{(82)} = 2,70$ i $G_{(1)} = 2,04$ dla zmiennej X_2 . Zmienna X_3 ma rozkład logarytmiczno-normalny, zbadano więc logarytmy wartości tej zmiennej mające już

Tabela 3. Wydatki na żywność (X_1), utrzymanie domu (X_2) i kulturę i edukację (X_3) w tys. zł (starych)

i	X_1	X_2	X_3	i	X_1	X_2	X_3	i	X_1	X_2	X_3
1	1928,7	546,3	322,2	29	2765,4	570,8	581,6	57	2379,0	880,7	180,6
2	1308,8	343,4	278,7	30	1114,4	338,7	55,8	58	563,0	105,9	23,4
3	1381,1	800,0	86,3	31	2417,4	620,6	132,8	59	2698,7	555,3	122,9
4	2212,4	1657,1	312,7	32	2594,0	564,0	351,0	60	2798,9	593,0	1,2
5	1780,4	467,8	269,3	33	2712,0	1484,0	217,0	61	1465,7	440,3	1510,1
6	1997,9	232,3	161,6	34	1280,9	425,9	82,5	62	832,1	264,0	10,6
7	2200,8	0,0	47,5	35	1945,4	425,9	157,2	63	3497,3	399,5	191,0
8	2133,7	631,7	89,8	36	1330,5	479,8	91,6	64	823,7	528,5	76,4
9	1397,3	348,8	244,4	37	2753,6	712,7	176,8	65	1885,0	478,6	193,7
10	1317,0	852,0	127,0	38	2674,8	687,4	1013,7	66	4250,7	1119,5	187,6
11	3007,8	784,8	165,0	39	2415,0	502,7	750,8	67	2113,0	477,0	808,0
12	1924,2	547,4	1762,5	40	2589,9	597,7	266,3	68	4361,2	837,2	2431,4
13	2498,1	880,9	1055,5	41	2422,7	1239,7	174,4	69	2419,2	1206,2	840,6
14	3690,7	963,1	991,6	42	1352,7	667,2	73,0	70	3401,6	953,9	1184,9
15	3880,8	987,3	1762,4	43	1636,6	830,5	111,5	71	1062,0	244,7	146,6
16	2634,3	964,1	538,8	44	3642,0	757,0	1311,0	72	777,2	265,1	362,2
17	2344,3	1493,7	1341,9	45	2122,0	675,0	464,0	73	1721,3	788,8	426,6
18	2019,9	1130,0	1234,0	46	2829,4	1114,8	334,4	74	1169,5	1000,2	13,4
19	2639,5	1202,6	1201,6	47	1134,9	568,6	238,6	75	1027,1	1013,7	32,0
20	2228,0	1114,0	743,0	48	3041,5	1495,4	1896,5	76	3119,1	758,5	462,4
21	2364,0	874,0	1309,0	49	1722,9	622,1	625,3	77	1776,5	172,4	416,8
22	2800,4	656,6	165,0	50	2603,0	661,0	836,3	78	1297,0	428,0	345,0
23	1559,6	207,4	120,4	51	1926,1	503,7	306,2	79	1601,0	906,0	146,0
24	1535,3	769,8	81,9	52	2869,6	821,7	572,9	80	1955,4	1337,8	80,3
25	1677,6	848,5	3,2	53	3124,1	607,4	149,1	81	2404,7	699,0	278,7
26	1800,5	486,8	1301,0	54	2184,2	1123,5	1357,8	82	2139,8	81,5	83,6
27	1598,8	317,9	216,0	55	3163,0	937,0	120,0				
28	2202,8	572,9	395,4	56	2516,0	1214,0	1393,0				

Źródło: RAD Project „Poverty and Targeting of Social Assistance in Eastern Europe and Former Soviet Union”.

rozkład normalny. Obserwację $x_{3,60}$ można uznać za odstającą dolną, wartość statystyki Grubsa wynosi dla niej $G_{(1)} = 3,81$. Dla pary obserwacji $x_{3,60}$ i $x_{3,25}$ sprawdzian testu U przybiera wartość 6,91. Test U odrzuca hipotezę zerową również dla czterech obserwacji: $x_{3,60}$, $x_{3,25}$, $x_{3,62}$ i $x_{3,74}$; otrzymujemy $U = 1,26$. Jednakże sekwencyjny test oparty na kurtozie nie odrzuca hipotezy H_0 dla obserwacji $x_{3,62}$ i $x_{3,74}$.

Przejdźmy teraz, do analizy obserwacji trójwymiarowych $\mathbf{x}_i = (x_{1,i}, x_{2,i}, x_{3,i})$. W tab. 4 przedstawiono 5 obserwacji, dla których wartość statystyki F_i przekracza wartość krytyczną $F_{0,05} = 2,720$. Podano też wartości kwadratu odległości Mahalanobisa D_i^2 oraz wartości F_i .

Tabela 4. Obserwacje podejrzane o nietypowość

i	X_1	X_2	X_3	D_i^2	F_i
4	2212,4	1657,1	312,7	9,747	3,132
12	1924,2	547,4	1762,5	8,935	2,870
48	3041,5	1495,4	1896,5	9,223	2,963
66	4250,7	1119,5	187,6	10,792	3,467
68	4361,2	837,2	2431,4	16,756	5,382

Źródło: opracowanie własne.

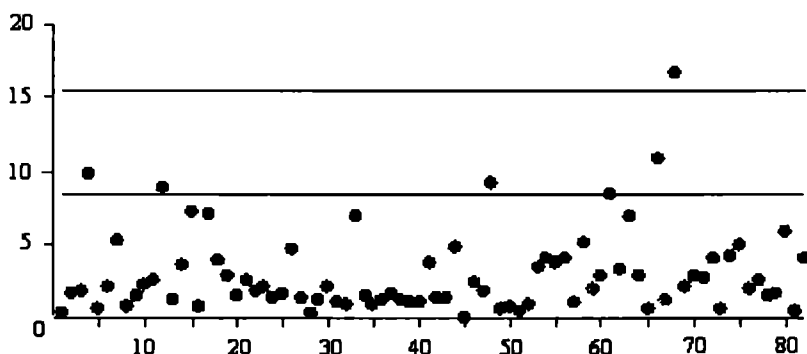
Przeprowadzając test Wilksa, w przypadku obserwacji $\mathbf{x}_{68}$ odrzucamy hipotezę jednorodności próby. Dla próby 50-elementowej wartość krytyczna na poziomie $\alpha = 0,05$ wynosi 14,18, a dla 100-elementowej 16,45. Dla pary obserwacji $\mathbf{x}_{68}$, $\mathbf{x}_{66}$ o największej wartości odległości Mahalanobisa otrzymujemy wartość sprawdzianu $\sqrt{r_2} = 0,8288$, jest on jednak nieznacznie większy od wartości krytycznej $c_{0,05}$, równej w przybliżeniu (aproksymacja liniowa) 0,794. Obserwację $\mathbf{x}_{68}$ możemy więc zaliczyć do obserwacji odstających. Pozostałe wyróżnione obserwacje należy traktować z pewną ostrożnością. Metoda Campbella przyporządkowuje obserwacji $\mathbf{x}_{68}$ wagę 0,330, a obserwacji $\mathbf{x}_{66}$ wagę 0,955; pozostałe obserwacje otrzymują wagę równą 1 – co potwierdza poprzednie rozważania. Przy okazji otrzymujemy odporny estymator wartości średnich $\bar{\mathbf{x}}_w = [2157,5; 711,7; 480,9]$, gdzie zwykły estymator $\bar{\mathbf{x}} = [2176,7; 718,0; 496,7]$, oraz odporny estymator macierzy kowariancji:

$$\mathbf{S}_w = \begin{bmatrix} 592678,8 & 111200,9 & 146024,6 \\ 111200,9 & 123850,1 & 63396,8 \\ 146024,6 & 63396,8 & 248553,1 \end{bmatrix}$$

oraz

$$S = \begin{bmatrix} 643295,1 & 113796,0 & 190582,4 \\ 113796,0 & 122705,0 & 65175,0 \\ 190582,4 & 65175,0 & 287250,1 \end{bmatrix}.$$

Na rys. 4 przedstawiono wykres wartości odległości Mahalanobisa dla poszczególnych obserwacji. Górna linia przedstawia aproksymowaną wartość krytyczną dla testu Wilksa, wynoszącą 15,5, a dolna linia – wartość graniczną $F_{0,05}$ pomnożoną przez wyrażenie $\frac{(n^2 - 1)m}{(n - m)n}$, równą 8,48.



Rys. 4. Wartości odległości Mahalanobisa dla obserwacji z przykładu 4

Źródło: opracowanie własne.

Gospodarstwa domowe odpowiadające obserwacji x_{68} uznanej za odstającą można uważać za istotnie różne od pozostałych gospodarstw rozpatrywanych w próbie i zaliczyć je do innej kategorii. Na pozostałe gospodarstwa domowe odpowiadające obserwacjom podejrzanym o nietypowość, tzn. x_4 , x_{12} , x_{48} oraz x_{66} , należy zwrócić szczególną uwagę w dalszej analizie dotyczącej wydatków.

4. Zakończenie

Podobnie jak w [4], w naszej pracy przedstawiono jedynie wybrane metody wykrywania wielowymiarowych obserwacji odstających oraz odporne metody estymacji. Obszerne omówienie tych metod znajduje się w pracach [1; 3; 8; 9; 10]. Autor zamierza w przyszłości przedstawić dalszą część rozważań dotyczącą nietypowych obserwacji, rozpatrując zagadnienia regresyjne dotyczące zarówno klasycznej regresji liniowej, jak i uogólnionych modeli liniowych.

Dodatek – tablice statystyczne

Tablica I. Rozkład Wilksa

n	$m = 2$		$m = 3$		$m = 4$		$m = 5$	
	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
5	3,17	3,19						
6	4,00	4,11	4,14	4,16				
7	4,71	4,95	5,01	5,10	5,12	5,14		
8	5,32	5,70	5,77	5,97	6,01	6,09	6,11	6,12
9	5,85	6,37	6,43	6,76	6,80	6,97	7,01	7,08
10	6,32	6,97	7,01	7,47	7,50	7,79	7,82	7,98
12	7,10	8,00	7,99	8,70	8,67	9,20	9,19	9,57
14	7,74	8,84	8,78	9,71	9,61	10,37	10,29	10,90
16	8,27	9,54	9,44	10,56	10,39	11,36	11,20	12,02
18	8,73	10,15	10,00	11,28	11,06	12,20	11,96	12,98
20	9,13	10,67	10,49	11,91	11,63	12,93	12,62	13,81
25	9,94	11,73	11,48	13,18	12,78	14,40	13,94	15,47
30	10,58	12,54	12,24	14,14	13,67	15,51	14,95	16,73
35	11,10	13,20	12,85	14,92	14,37	16,40	15,75	17,73
40	11,53	13,74	13,36	15,56	14,96	17,13	16,41	18,55
45	11,90	14,20	13,80	16,10	15,46	17,74	16,97	19,24
50	12,23	14,60	14,18	16,56	15,89	18,27	17,45	19,83
100	14,22	16,95	16,45	19,26	18,43	21,30	20,26	23,17
200	15,99	18,94	18,42	21,47	20,59	23,72	22,59	25,82
500	18,12	21,22	20,75	23,95	23,06	26,37	25,21	28,62

Źródło: [1].

Tablica II. Rozkład Wilksa dla par obserwacji

n	$m = 2$		$m = 3$		$m = 4$		$m = 5$	
	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
5	0,0025	0,0005	0,0000					
6	0,0337	0,0150	0,0011	0,0002				
7	0,0860	0,0498	0,0202	0,0090	0,0006	0,0001		
8	0,1417	0,0937	0,0580	0,0335	0,0136	0,0060	0,0004	0,0001
9	0,1942	0,1393	0,1024	0,0674	0,0425	0,0245	0,0098	0,0043
10	0,2419	0,1831	0,1470	0,1049	0,0788	0,0518	0,0327	0,0189
12	0,3229	0,2616	0,2288	0,1791	0,1549	0,1163	0,0966	0,0686
14	0,3879	0,3276	0,2982	0,2460	0,2246	0,1804	0,1631	0,1270
16	0,4410	0,3828	0,3563	0,3040	0,2853	0,2389	0,2242	0,1838
18	0,4850	0,4295	0,4054	0,3542	0,3376	0,2908	0,2782	0,2360
20	0,5221	0,4694	0,4472	0,3976	0,3828	0,3366	0,3257	0,2830
25	0,5935	0,5472	0,5288	0,4839	0,4722	0,4290	0,4211	0,3798
30	0,6451	0,6041	0,5882	0,5478	0,5380	0,4984	0,4923	0,4537
35	0,6842	0,6475	0,6335	0,5969	0,5885	0,5523	0,5473	0,5116
40	0,7150	0,6818	0,6693	0,6360	0,6285	0,5953	0,5911	0,5580
45	0,7399	0,7097	0,6982	0,6677	0,6610	0,6304	0,6267	0,5961
50	0,7605	0,7328	0,7222	0,6941	0,6880	0,6596	0,6564	0,6270
100	0,8629	0,8477	0,8417	0,8260	0,8225	0,8065	0,8047	0,7883
200	0,9232	0,9152	0,9118	0,9035	0,9015	0,8929	0,8918	0,8830
500	0,9650	0,9618	0,9602	0,9568	0,9558	0,9523	0,9517	0,9480

Źródło: [1].

Literatura

- [1] Barnett V., Levis T., *Outliers in Statistic Data* (3 wyd.), J. Wiley & Sons, Chichester 1995.
- [2] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., *Robust Statistics*, J. Wiley & Sons, New York 1986.
- [3] Hebak P., Hustopecky J., *Vicerazmerne statistické metody s aplikacemi*, SNTL/Alfa, Praha 1987.
- [4] Heilpern S., *Nietypowe realizacje jednowymiarowych zmiennych losowych*, Prace Naukowe Akademii Ekonomicznej nr 1097, AE, Wrocław 2005.
- [5] Jajuga K., *Statystyczna analiza wielowymiarowa*, Wyd. Nauk. PWN, Warszawa 1993.

-
- [6] Lopuhaa H.P., Rousseeuw P.J.: *Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariate Matrices*, „The Annals of Statistics”, 19 (1991), s. 229-248.
 - [7] Miszczak W., *Statystyczne metody analizy danych. Materiały do ćwiczeń*. AE, Wrocław 1999.
 - [8] Rocke D.M., Woodruff D.L.: *Identification of Outliers in Multivariate Data*. „JASA”, 91 (1996), s. 1047-1061.
 - [9] Rousseeuw P.J., van Driessen K., *A Fast Algorithm for the Minimum Covariance Determinant Estimator*. „Technometrics”, 41 (1999), s. 212-223.
 - [10] Rousseeuw P.J., van Zomeren B.C., *Unmasking Multivariate Outliers and Leverage Points*. „JASA”, 85 (1990), s. 633-639.
 - [11] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1998.

OUTLIERS – MULTIDIMENSIONAL CASE

Summary

The paper is devoted to the outliers in multivariate data. It reviews selected methods of the detection of the multidimensional outliers. These are the methods based on the statistical inference using Mahalanobis distance and the preference relations and based on the descriptive and graphical methods. The robust estimation of the vector location and the covariance matrix, weakly reacting to the outliers, are studied. The Huber's, Gnanadesikan-Kettenring's and Cambell's methods are presented and the affine invariant estimators obtained by the minimum volume ellipsoid and the minimum covariance matrix determinant methods are studied.

The paper contains the examples illustrating presented methods. There are two tables with the critical values of the statistical tests used to the detection of the multidimensional outliers in the appendix.