Michał Trzęsiok

Uniwersytet Ekonomiczny w Katowicach
Katedra Matematyki
michal.trzesiok@ue.katowice.pl

O JAKOŚCI DANYCH W KONTEKŚCIE OBSERWACJI ODDALONYCH W WIELOWYMIAROWEJ ANALIZIE REGRESJI

Wprowadzenie

Jakość modelu statystycznego zależy bezpośrednio od jakości danych wykorzystanych do jego wyznaczenia. Często w rzeczywistych zbiorach danych występują pewne obserwacje, w których wartości opisujących je zmiennych są nietypowe. Wynika to ze specyfiki badanego zjawiska lub też z różnego rodzaju błędów. Owe obserwacje nietypowe mogą mieć bardzo silny wpływ na wyniki analizy i w związku z tym wymagają szczególnej uwagi.

W artykule zestawiono kilka metod identyfikacji obserwacji oddalonych. Pierwsza z metod wykorzystuje miary kwantylowe. Jest to prosty sposób identyfikowania nietypowych wartości zmiennych obserwowanych. Takie podejście (z graficzną prezentacją w postaci wykresów skrzynkowych) jest często jednym z pierwszych etapów przygotowania danych do właściwej analizy statystycznej. Wadą tego podejścia jest ich bezkontekstowość – każda ze zmiennych jest traktowana w sposób niezależny i bez względu na rodzaj budowanego później modelu. Alternatywą jest definiowanie i identyfikowanie obserwacji oddalonych poprzez porównywanie zmian w postaci modeli zbudowanych na różnych zbiorach obserwacji. Podejście to jest szczególnie często wykorzystywane w analizie regresji, gdzie prowadzi do wyróżnienia dodatkowej klasy – obserwacji wpływowych (odstających, ale „pozytywnych"). Istnieją również metody poszukujące nośnika wielowymiarowego rozkładu badanego zbioru zmiennych. Funkcja identyfikująca, czy dana obserwacja znajduje się w *ogonie* takiego rozkładu, klasyfikuje te obiekty jako oddalone (nietypowe).

W artykule dokonano uporządkowania definicji *obserwacji oddalonych* oraz zweryfikowano empirycznie przydatność zestawionych metod ich identyfikacji na zbiorze danych rzeczywistych. Ponadto na przedstawionym przykładzie empirycznym sprawdzono, w jakim stopniu zbiory zidentyfikowanych obserwacji oddalonych pokrywają się dla różnych metod (reprezentujących bardzo odmienne podejścia).

1. Podstawowe definicje

Pojęcie *obserwacji odstającej* nie jest w literaturze zdefiniowane jednoznacznie. W niniejszej pracy posłużono się dosyć ogólną definicją zaczerpniętą z pracy Hawkinsa [9]:

Definicja 1. *Obserwacja odstająca* (*outlier*) to taka obserwacja, która odchyla się tak bardzo od innych obserwacji, że rodzi to przypuszczenie, że powstała w wyniku działania innego mechanizmu, tj. że pochodzi z innego rozkładu niż pozostałe obserwacje w zbiorze danych.

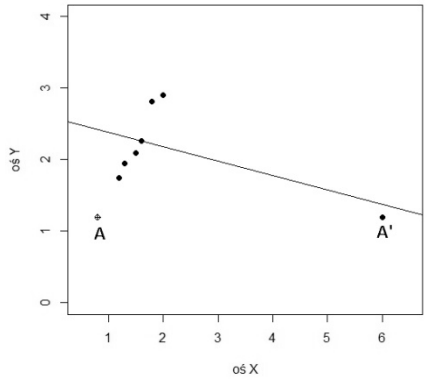
Warto zauważyć, że zgodnie z powyższą definicją, wystąpienie to oznacza brak spełnienia jednego z najbardziej podstawowych założeń metod wielowymiarowej analizy statystycznej. W grupie tych metod na ogół przyjmuje się najbardziej podstawowe założenie dotyczące zbioru danych, że jest to zbiór i.i.d. (*independent and identically distributed*), czyli założenie, że zbiór danych tworzą obserwacje wylosowane w sposób niezależny, o jednakowym, wielowymiarowym rozkładzie określonym przez (nieznaną, ale wspólną) funkcję gęstości [6]. Występowanie obserwacji odstającej oznacza, że pochodzi ona z innego rozkładu i nie powinna być rozpatrywana łącznie z innymi obiektami z analizowanego zbioru danych. W tym znaczeniu *obserwacja oddalona* jest ogólniejszym pojęciem niż *obserwacja odstająca*, gdyż obserwacja oddalona oznacza obiekt, który jest opisany przez rzadkie (nietypowe) wartości zmiennych (lecz mogą to być wartości występujące w ramach rozkładu, w jego *ogonie*, czyli tzw. wartości ekstremalne).

W literaturze można spotkać wiele innych definicji obserwacji oddalonych. Bardzo często są to jednak definicje odnoszące się do pojęcia obserwacji oddalonej przez pewien szczególny kontekst. Wyróżnić tu można trzy rodzaje obserwacji oddalonych [15]:

- obserwacje *odstające* (*outliers*) to takie obserwacje, w których wyróżniona jest zmienna objaśniana Y i właśnie wartość tej zmiennej znacząco odchyla się od wartości dla innych obserwacji,

- obserwacje *wysokiej dźwigni* (lub *dźwigniowe*; *leverage*) to takie obserwacje, w których wartość przynajmniej jednej ze zmiennych objaśniających (**X**) znacząco odchyla się od wartości tej zmiennej dla innych obserwacji (rys. 1),
- obserwacje wpływowe (*influential observations*), to takie obserwacje, których wyłączenie ze zbioru danych powoduje istotną zmianę zbudowanego modelu (rys. 1).

Przedstawiona klasyfikacja nie jest rozłączna, np. obserwacja może być jednocześnie odstająca i wpływowa, bądź odstająca i dźwigniowa itp.



Nota:
Poprawne położenie A' oznaczono literą A, lecz wartość zmiennej objaśniającej została błędnie wprowadzona. Wystąpienie obserwacji A' bardzo istotnie wpłynęło na model regresji liniowej. Punkt A' jest obserwacją wpływową oraz dźwigniową.

Rys. 1. Ilustracja konsekwencji wprowadzenia do zbioru danych obserwacji A'

Źródło: Na podstawie [15, rys. 2, s. 5].

Zagadnienie identyfikacji obserwacji oddalonych zawiera w sobie kilka poważnych trudności. Po pierwsze nie zawsze występowanie obserwacji oddalonych jest zjawiskiem negatywnym. Owszem, czasem są rezultatem błędów pomiaru zmiennych, jednak czasem są wynikiem poprawnych pomiarów i obrazują prawdziwe, choć rzadkie i nietypowe zachowanie badanego zjawiska. W tym drugim przypadku zdecydowanie nie należy usuwać tych obserwacji, gdyż na ogół ich zawartość informacyjna jest bardzo duża [18]. W obu przypadkach ważnym jest by zidentyfikować obserwacje oddalone i w odpowiedni sposób je potraktować. Po drugie wiele klasycznych metod identyfikacji obserwacji nietypowych nie potrafi wykrywać mnogich wartości oddalonych (efekt wzajemnego maskowania się dwóch lub więcej obserwacji oddalonych leżących blisko siebie) [por. 11]. Po trzecie niektóre metody są skupione na identyfikowaniu obserwacji

oddalonych, wykorzystując tylko jedną z wielu możliwych konsekwencji ich występowania, np. badając reszty modelu. Tymczasem nie zawsze duża reszta modelu dla danej obserwacji oznacza, że jest to obserwacja oddalona [12], gdyż model może być źle dopasowany do niektórych typowych obserwacji, np. wskutek zakłóceń wywołanych kilkoma innymi obserwacjami, które faktycznie są oddalone.

2. Krótki opis wybranych metod identyfikacji obserwacji oddalonych

2.1. Metody jednowymiarowe – kryterium kwartylowe

Należy podkreślić, że celem stosowania metod identyfikacji obserwacji oddalonych nie jest późniejsze usunięcie tych obserwacji (chyba że przyczyną ich powstania były błędy pomiaru lub błędy przy wprowadzaniu danych), lecz badania empiryczne wskazują, że na ogół znacznie lepsze wyniki niż *usuwanie obserwacji* dają *metody odporne* (*robust methods*) [11].

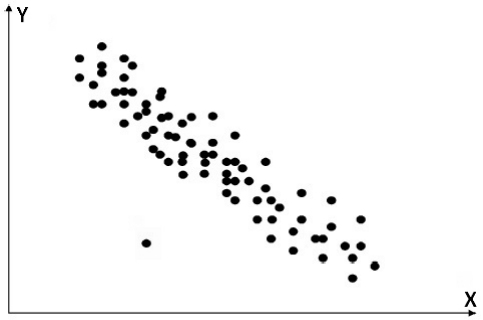
Niech $\mathbf{X} = (X_1, \ldots, X_k)$ będzie wektorem zmiennych objaśniających w n elementowym zbiorze danych. Najprostsze i najstarsze metody identyfikowania obserwacji oddalonych to metody jednowymiarowe, na ogół połączone z prezentacją graficzną wartości zmiennej. Do takich metod należy zaliczyć kryterium kwartylowe wykorzystywane w budowie wykresów pudełkowych wprowadzonych przez Tukeya [17]. Wartość pojedynczej zmiennej jest uznana za oddaloną, jeśli znajduje się poza przedziałem:

$$\left\langle Q_1 - 1{,}5 \cdot IQR, \quad Q_3 + 1{,}5 \cdot IQR \right\rangle, \tag{1}$$

gdzie:
Q_1, Q_3 – odpowiednio pierwszy i trzeci kwartyl,
IQR – rozstęp ćwiartkowy.

Niektórzy autorzy przyjmują nawet dopełnienie przedziału danego wzorem (1) jako definicję obserwacji oddalonej [por. 8, s. 42]. Wykresy pudełkowe są bardzo cennym narzędziem do wstępnego zapoznania się z analizowanym zbiorem danych, lecz łatwo można wykazać, że jednowymiarowe podejście do zagadnienia identyfikacji obserwacji oddalonych jest niewystarczające. Na rys. 2 przedstawiono prosty dwuwymiarowy przykład, w którym zaznaczono obserwację oddaloną, która zarówno ze względu na zmienną objaśniającą, jak i wartość zmiennej objaśnianej z osobna, nie odbiega znacząco od mediany. Kryterium

kwartylowe nie jest skutecznym narzędziem identyfikowania obserwacji oddalonych dla danych wielowymiarowych.



Rys. 2. Przykład zbioru z jedną obserwacją oddaloną, której nie można zidentyfikować jednowymiarowymi metodami kwartylowymi

Źródło: Na podstawie [15, rys. 4, s. 7].

2.2. Graficzna metoda wielowymiarowa – krzywe Andrewsa

Do identyfikacji wielowymiarowych obserwacji oddalonych można wykorzystać metody redukcji wymiaru, np. metodę Andrewsa, która każdą obserwację sprowadza do pewnej krzywej na płaszczyźnie [1]. Andrews zaproponował kilka typów przekształceń wielowymiarowych obserwacji do krzywych. W niniejszej pracy wykorzystano przekształcenie:

$$f(t) = x_1 \cdot \sin t + x_2 \cdot \cos t + x_3 \cdot \sin(2 \cdot t) + x_4 \cdot \cos(2 \cdot t) + \ldots . \qquad (2)$$

Metoda Andrewsa wykorzystuje ideę rozwinięcia funkcji w szereg Fouriera i choć jest elegancka w swojej matematycznej warstwie, to jednak ma ograniczone zastosowanie dla zbiorów danych o dużej liczebności, gdyż otrzymywany rysunek jest nieczytelny (zbyt wiele nakładających się krzywych).

2.3. Metoda wykorzystująca odległość Cooka

Bardzo popularną metodą identyfikacji obserwacji nietypowych w analizie regresji wielorakiej jest metoda wykorzystująca odległość Cooka, która to odległość porównuje stopień dopasowania do danych dla dwóch modeli: modelu pełnego, uwzględniającego wszystkie obserwacje ze zbioru uczącego, oraz dla modelu zbudowanego na zbiorze danych, w którym pominięto jedną, wybraną *i-tą* obserwację [5]:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{m \cdot MSE}, \tag{3}$$

gdzie:

$\hat{Y}_j$ – prognozowana wartość zmiennej Y dla obserwacji o numerze j w modelu pełnym, tj. zbudowanym na całym zbiorze uczącym,

$\hat{Y}_{j(i)}$ – prognozowana wartość zmiennej Y dla obserwacji o numerze j w modelu zbudowanym na zbiorze, z którego tymczasowo wyłączono obserwację o numerze i,

MSE – błąd średniokwadratowy modelu,

m – liczba parametrów modelu.

Jako wartość graniczną odległości Cooka, powyżej której należy daną obserwację uznać za odstającą przyjmuję się 1 lub alternatywnie: $\frac{4}{n-m-2}$.

2.4. Metody oparte na odległości Mahalanobisa

Szczególnie w ekonometrii stosuje się metody identyfikacji obserwacji oddalonych wykorzystujące kryterium bazujące na odległości Mahalonobisa [10]:

$$MD^2(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}), \tag{4}$$

gdzie:

$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ – wartość przeciętna,

$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ – macierz wariancji i kowariancji.

Punkty o dużych (w porównaniu z wartościami krytycznymi odczytanymi z rozkładu χ^2) wartościach kwadratu odległości Mahalanobisa są traktowane jako obserwacje oddalone. To podejście ma jednak tę podstawową wadę, że wartość samego kryterium (4) w bezpośredni sposób zależy od statystyk (klasycznych), które są bardzo wrażliwe na występowanie wartości oddalonych. W celu wyeliminowania tej wady zaproponowano modyfikacje wyliczania wartości miernika (4) poprzez zastąpienie średniej $\hat{\boldsymbol{\mu}}$ przez odporny parametr położenia.

Jedna z propozycji to wykorzystanie estymatora *MVE* (*Minimum Volume Ellipsoid Estimator*), tj. estymatora o minimalnej objętości elipsoidy [14]:

$$\hat{\boldsymbol{\mu}} = \text{środek ciężkości elipsoidy o minimalnej objętości zawierającej co najmniej } h \text{ obserwacji danego zbioru,} \quad (5)$$

gdzie:
$h = \lfloor n/2 \rfloor + 1$.

Druga z propozycji [14] to wyznaczenie parametru położenia $\hat{\boldsymbol{\mu}}$ we wzorze (4) według formuły:

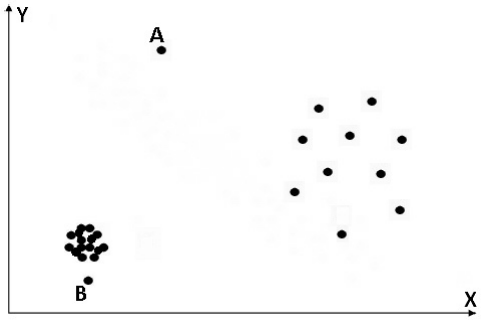
$$\hat{\boldsymbol{\mu}} = \text{średnia z tych } h \text{ obserwacji danego zbioru, dla których wyznacznik macierzy kowariancji jest najmniejszy.} \quad (6)$$

Odporny estymator położenia (6) jest nazywany estymatorem *MCD* (*Minimum Covariance Determinant Estimator*), tj. estymatorem o minimalnym wyznaczniku macierzy kowariancji. Trzecie podejście zasugerowane w pracy [7] wykorzystuje analizę głównych składowych i identyfikuje obserwacje oddalone właśnie po przekształceniu wszystkich obserwacji w przestrzeni głównych przez wyznaczenie w tej przestrzeni wartości kwadratu odległości Mahalanobisa. Autorzy tego podejścia sugerują zastosowanie na etapie przygotowania danych do analizy standaryzacji zmiennych z wykorzystaniem mediany jako parametru położenia oraz MAD, czyli medianowego odchylenia bezwzględnego jako parametru rozproszenia. Po zastosowaniu takiej standaryzacji, obliczanie odległości euklidesowej w przestrzeni głównych składowych jest równoważne obliczaniu odpornego wariantu odległości Mahalanobisa.

2.5. Metoda uwzględniająca lokalne zagęszczenie obserwacji

Przedstawione w punktach 2.1-2.4 metody identyfikacji obserwacji oddalonych traktują to zagadnienie zero-jedynkowo, czyli albo obserwacja jest oddaloną, albo nie. Odmienne podejście do tego problemu prezentują Breunig, Kriegel, Ng i Sander [4], którzy proponują miernik, wskazujący stopień oddalenia danego obiektu od pozostałych obserwacji ze zbioru danych. Miernik ten nazywają *LOF* (*Local Outlier Factor*) – lokalnym miernikiem stopnia oddalenia obserwacji. Definicja tego miernika ma złożoną postać analityczną oraz zagnieżdżoną strukturę i wymaga zdefiniowania trzech innych pojęć. W tym miejscu podana zostanie jedynie główna idea jego konstrukcji. Miernik *LOF* jest zainspirowany

metodą k najbliższych sąsiadów i wskazuje stopień oddalenia danej obserwacji od pozostałych, uwzględniając zagęszczenie obiektów z k elementowego sąsiedztwa. Takie podejście pozwala identyfikować obserwacje oddalone również w przypadku, gdy zbiór danych tworzą skupienia o różnym stopniu zagęszczenia, czyli różnym poziomie koncentracji wokół środka ciężkości (to, czy pewna odległość punktu od pozostałych jest wystarczająco duża, by uznać punkt za oddalony, jest wszak zależne od stopnia zróżnicowania odległości punktów w danym fragmencie przestrzeni – por. rys. 3).



Nota:
Większość metod zidentyfikuje poprawnie obserwację A jako oddaloną. Zidentyfikowanie obserwacji B jako oddalonej wymaga uwzględnienia stopnia lokalnego zagęszczenia obiektów.

Rys. 3. Przykład zbioru, w którym są dwie klasy o różnym stopniu zagęszczenia oraz dwie obserwacje oddalone (oznaczone: A i B)

2.6. Metoda wyznaczania uogólnionego wielowymiarowego kwantyla rozkładu

Jeden z wariantów metody wektorów nośnych *SVM* (*Support Vector Machines*) pozwala na wyznaczenie *uogólnionego wielowymiarowego kwantyla rozkładu* generującego dane z analizowanego zbioru. Przez uogólniony kwantyl rozkładu rozumieć należy taki obszar $Q \subset \mathbf{R}^k$ wielowymiarowej przestrzeni danych, który spełnia warunek, że niemal wszystkie obserwacje wygenerowane z rozkładu należą do Q, z drugiej strony niemal wszystkie obiekty nie pochodzące z rozkładu generującego dane, należą do dopełnienia zbioru Q. Wykorzystując funkcje jądrowe, określające pewne nieliniowe przekształcenie przestrzeni danych, standardową technikę stosowaną w metodzie wektorów nośnych, poszukiwanie rozwiązania problemu zostaje przeniesione w przestrzeń $\mathbf{Z}$ o znacznie większym wymiarze i w tej nowej przestrzeni cech jest wyznaczana optymalna hiperkula (o najmniejszym możliwym promieniu, tzw. hiperkula Czebyszewa), zawierają-

ca obrazy (niekoniecznie wszystkich) obserwacji ze zbioru uczącego. Tej hiperkuli w przestrzeni **Z** odpowiada (jako przeciwobraz) pewien zbiór w pierwotnej przestrzeni danych. Jest nim poszukiwany uogólniony kwantyl Q. Ze względu na uelastycznienie metody na wypadek wystąpienia w zbiorze danych potencjalnych błędów pomiaru lub obserwacji nietypowych, wyznaczona hiperkula Czebyszewa nie musi zawierać obrazów wszystkich obserwacji ze zbioru danych. Obiekty, które znalazły się poza tą hiperkulą, można łatwo zidentyfikować. Są to obserwacje, znajdujące się poza uogólnionym kwantylem rozkładu i potencjalnie pochodzą z innego rozkładu, czyli mogą zostać zidentyfikowane jako obserwacje oddalone. Szczegóły wraz z formalnym zapisem opisanej metody można znaleźć w pracach Ben-Hur i in. oraz Trzęsiok [3], [16].

2.7. Inne możliwe podejścia do identyfikacji obserwacji oddalonych

Do identyfikacji obserwacji oddalonych można również posłużyć się *metodami taksonomicznymi*, licząc, że obserwacje oddalone w wyniku grupowania zostaną wyodrębnione tworząc jednoelementowe klasy. Takie podejście jest jednak krytykowane [4], gdyż metody taksonomiczne mają na celu wyznaczenie skupień i temu podporządkowany jest ich mechanizm (optymalizacyjny), a nie rozpoznawaniu obserwacji oddalonych.

W literaturze przedmiotu można znaleźć bardzo wiele propozycji testów statystycznych do weryfikacji hipotezy, czy dana obserwacja jest obserwacją oddaloną. Obszerny zestaw takich testów można znaleźć w pracy Barnetta i Lewisa [2].

Inne bardzo obiecujące podejście wykorzystuje pojęcie *głębi* [17], lecz niestety w praktyce metoda ta okazuje się mało wydajna dla danych wielowymiarowych z wymiarem $k \geq 4$, gdyż wymaga wyznaczania otoczek wypukłych, co jest bardzo wymagające obliczeniowo [4].

3. Empiryczne porównanie wyników działania wybranych metod

Analiza empiryczna została przeprowadzona na zbiorze danych *Clothing*[1]. W zbiorze tym zebrano informacje na temat sprzedaży odzieży męskiej w sklepach tego typu w Holandii. Zbiór zawiera 400 obserwacji, a zmienne opisujące obiekty to:

X_1 – zysk brutto,

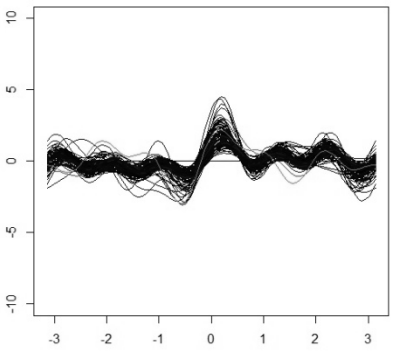
X_2 – liczba właścicieli,

[1] Zbiór danych *Clothing* jest dostępny w bibliotece *Ecdat* programu statystycznego **R**.

X_3 – liczba pracowników pełnoetatowych,
X_4 – liczba pracowników zatrudnionych na część etatu,
X_5 – liczba pracowników okresowych,
X_6 – liczba godzin przepracowanych w roku,
X_7 – liczba godzin przepracowanych w roku przez jednego pracownika,
X_8 – inwestycje w powierzchnię handlową,
X_9 – inwestycje w automatyzację procesów,
X_{10} – powierzchnia handlowa sklepu [w m^2],
X_{11} – rok rozpoczęcia działalności sklepu.
Zmienna objaśniana Y – roczna wartość sprzedaży sklepu.

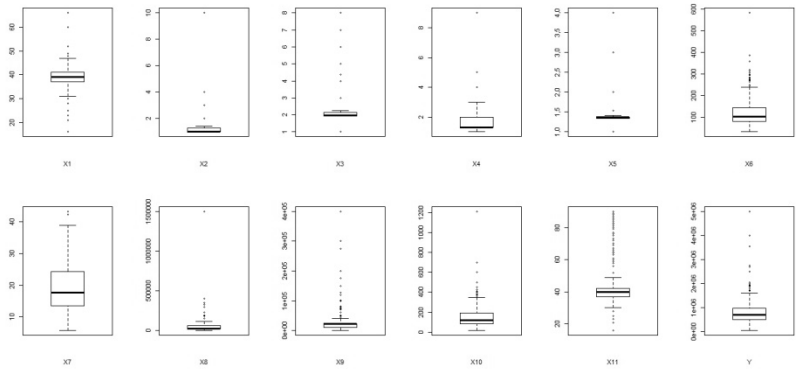
Wszystkie obliczenia zostały przeprowadzone z wykorzystaniem programu statystycznego **R** z dołączonymi bibliotekami oraz autorskimi procedurami programu **R**.

Podjęto próbę identyfikacji obserwacji oddalonych metodą krzywych Andrewsa, ale ze względu na dużą liczbę obserwacji, umieszczenie 400 krzywych na jednym wykresie, wykres jest nieczytelny. Jednak dla zilustrowania metody na rys. 4 przedstawiono 30 krzywych Andrewsa dla pierwszych 30 obserwacji.

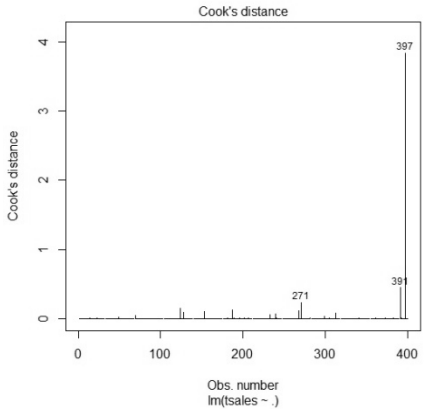


Rys. 4. Krzywe Andrewsa dla 100 obserwacji ze zbioru *Clothing*

W dalszej części przeprowadzono identyfikację obserwacji oddalonych pięcioma metodami: jednowymiarową metodą kwartylową (rys. 5), metodą wykorzystującą odległość Cooka (rys. 6), metodą MD^* opartą na odległości Mahalanobisa z poprawkami zaproponowanymi przez Filzmosera i in. (identyfikacja w przestrzeni głównych składowych), metodą *LOF*, uwzględniającą lokalne zagęszczenie obserwacji oraz metodą wektorów nośnych *SVM*. Liczba obserwacji oddalonych zidentyfikowanych przez każdą z metod nie pozwala na ich prezentację tabelaryczną, ale dla metody *SVM* kilka wybranych obserwacji zidentyfikowanych jako oddalone przedstawiono w tabeli 1.



Rys. 5. Wykresy pudełkowe dla każdej ze zmiennych z zaznaczonymi obserwacjami oddalonymi na podstawie metody kwantylowej



Rys. 6. Wyniki identyfikacji obserwacji oddalonych z wykorzystaniem odległości Cooka

Tabela 1

Fragment wyników identyfikacji oddalonych dla metody *SVM*

Nr	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Y
124	28	1	1	2,2	1,4	117	20,8	1500000	300000	350	28	976817
391	48	2	5	5	3	385	25,7	62269	16624	1214	86	1993529
268	40	10	4	2,2	1,4	582	33,0	292857	71570	100	40	1926395
392	42	1	1	1	1	151	37,8	292857	1687	180	75	695000
397	40	1	8	9	1	120	6,3	292857	126800	300	76	5000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

W tabeli 2 przedstawiono zgodność klasyfikacji zastosowanych metod parami, tj. liczbę obserwacji, które zgodnie zostały zidentyfikowane przez dwie metody jako oddalone.

Tabela 2

Zgodność klasyfikacji zastosowanych metod parami, tj. liczba obserwacji, które zgodnie zostały zidentyfikowane przez dwie metody jako oddalone

Metoda	*Cook*	*MD **	*LOF*	*SVM*
Cook	4	4	*2*	*4*
*MD **		39	17	11
LOF			70	9
SVM				22

Z tabeli 2 widać, że przedstawione metody znacząco różnią się w podejściu do zagadnienia, a w konsekwencji również wyznaczają w niewielkim stopniu pokrywające się zbiory obserwacji oddalonych. Nadmienić również należy, że liczba obserwacji zidentyfikowanych jako oddalona jest zależna od parametrów metody, które ustalane były symulacyjnie. Oznacza to jednak, że przedstawione wyniki są tylko jednym z możliwych wariantów. Brak jednoznaczności rozwiązania zagadnienia oraz rozbieżności między metodami wynikają z natury zagadnień klasyfikacji bezwzorcowej. W celu zredukowania subiektywizmu w doborze wartości parametrów wykorzystanych metod można zbudować wiele modeli dla różnych kombinacji parametrów i np. zastosować regułę majoryzacyjną.

Podsumowanie

Zaprezentowano wybrane metody identyfikacji obserwacji oddalonych. Określenie liczby zidentyfikowanych obserwacji oddalonych wymaga użycia heurystyk (jest wysoce subiektywna). Wybrane metody w różny sposób realizują cel identyfikacji obserwacji oddalonych, co przekłada się również na odmienne rezultaty ich działania (zbiory zidentyfikowanych obserwacji oddalonych dla różnych metod w niewielkim stopniu się pokrywają). Nie oznacza to jednak, że niektóre metodą są gorsze, tylko że metody te można traktować jako komplementarne.

Problem identyfikacji obserwacji oddalonych ma być jedynie narzędziem wstępnej poprawy jakości danych – zwróceniem uwagi na występujące w zbiorze anomalie. Wszystkie przedstawione metody spełniają ten postulat, choć każda w nieco inny sposób.

Literatura

[1] Andrews D.F., *Plots of High-Dimensional Data*, „Biometrics" 1972, Vol. 28, No. 1, s. 125-136.

[2] Barnett V., Lewis T., *Outliers in Statistical Data*, 3rd Edition, John Wiley & Sons, New York 1998.

[3] Ben-Hur A., Horn D., Siegelman H.T., Vapnik V., *Support Vector Clustering*, „Journal of Machine Learning Research" 2001, Vol. 2, s. 125-137.

[4] Breunig M.M., Kriegel H.-P., Ng R.T., Sander J., *LOF: Identifying Density-Based Outliers*, Proceedings of the 29th ACM SIDMOD International Conference on Management of Data (SIGMOD 2000), Dallas 2000, s. 93-104.

[5] Cook R.D., *Detection of Influential Observations in Linear Regression*, „Technometrics" 1977, 19 (1), s. 15-18.

[6] Duda R.O., Hart P.E., Stork D.G., *Pattern Classification*, John Wiley & Sons, New York 2001.

[7] Filzmoser P., Maronna R.A., Werner M., *Outlier Identification in High Dimensions*, „Computational Statistics & Data Analysis" 2008, Vol. 52, s. 1694-1711.

[8] Giudici P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons, New York 2003.

[9] Hawkins D., *Identification of Outliers*, Chapman and Hall, London 1980.

[10] Healy M.J.R., *Multivariate Normal Plotting*, „Applied Statistics" 1968, Vol. 17, s. 157-161.

[11] Huber P.J., Ronchetti E.M., *Robust Statistics*, 2nd Edition, John Wiley & Sons, Hoboken, NJ 2009.

[12] Maddala G.S., *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa 2006.

[13] Maronna R.A., Martin R.D., Yohai V.J., *Robust Statistics: Theory and Methods*, John Wiley & Sons, Chichester 2006.

[14] Rousseeuw P.J., *Least Median of Squares Regression*, „Journal of the American Statistical Association" 1984, Vol. 79, s. 871-880.

[15] Rousseeuw P.J., Leroy A.M., *Robust Regression and Outlier Detection*, John Wiley & Sons, New York 2003.

[16] Trzęsiok M., *Identyfikacja obserwacji oddalonych z wykorzystaniem metody wektorów nośnych*, [w:] *Taksonomia 14. Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wydawnictwo Naukowe Akademii Ekonomicznej, Wrocław 2007, s. 350-357.

[17] Tukey J.W., *Exploratory Data Analysis*, Addison-Wesley, Boston 1977.

[18] Webb A.R., *Statistical Pattern Recognition*, Second Edition, John Wiley & Sons, New York 2002.

ON SELECTED DATA QUALITY ISSUES IN MULTIVARIATE REGRESSION ANALYSIS

Summary

The paper presents different definitions of outliers. We also collate selected outlier detection techniques, which represent very different approaches to outliers identification: classical univariate method embodied in boxplots, Andrews' curves, methods based on Cook's distance and Mahalonobis' distance, local outlier factor method, support vector machines. Moreover we empirically examine the agreement between the results of outlier detection methods on the benchmarking, real world dataset.