

Klasyfikacja i Metody Odnajdywania Obserwacji Odstających

Na podstawie *"How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration"*
(2019)

Nel Olszewska

Adam Mickiewicz University

1 Wprowadzenie

- Prerejestracja

2 Wiadomości podstawowe

- Definicja obserwacji odstającej
- Klasyfikacja obserwacji odstających
 - ze względu na wymiar
 - ze względu na ich charakterystykę
- Co zrobić z wartościami odstającymi?

3 Standardowe metody znajdowania obserwacji odstających

- Metody
- Odporne wskaźniki obserwacji odstających

4 Proponowane metody

5 Podsumowanie

Autorzy powołują się na badania, które wykazały, że nieusystematyzowane podejście do obserwacji odstających (stosowanie nieodpowiednich metod ich znajdowania, późniejsze zabiegi - np. usuwanie czy przekodowywanie) prowadzi do zniekształconych lub nawet całkiem fałszywych wyników. Uznają to zjawisko jako rosnący problem w badaniach psychologicznych.

Proponują ujednolicenie definicji wartości odstającej, modyfikują istniejące kryteria ich klasyfikacji i przedstawiają lepszy schemat zarządzania wartościami odstającymi.

Proponowana kolejność działania

- 1 Znalezienie obserwacji odstających.

Za pomocą odpowiednich metod matematycznych.

- 2 Podjęcie decyzji dotyczącej traktowania obserwacji odstających (usuwanie, przekodowanie, zachowanie).

Na podstawie niematematycznych danych.

- 3 Przeprowadzenie dalszych działań na bazie danych.

Definicja obserwacji odstającej

Definicja (Obserwacje odstające)

Data points with large residual values.

Obserwacje o dużej wartości rezydualnej.

Definicja (Wartość rezydualna)

The discrepancy between the observed value and the value predicted by the statistical model.

Różnica między wartością obserwowaną a wartością przewidzianą przez model statystyczny.

Podział obserwacji odstających

1 Jednowymiarowe

Wyróżniające się ze względu na wartość jednej cechy.

2 Wielowymiarowe

Wyróżniające się ze względu na wartość dwóch lub więcej cech, przy czym nie muszą się wyróżniać ze względu na każdą z osobna.

Podział obserwacji odstających

Table 1: Adjusted nomenclature of outliers.

Error	<i>e.g., coding error</i>
Interesting	<i>e.g., moderator underlying a potentially interesting psychological process</i>
Random	<i>e.g., a very large value of a given distribution</i>

Leys et al: How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration

1 Błędne/Error

Obserwacje odstające jako rezultat niedopatrzienia.

2 (Potencjalnie) interesujące/Interesting

Obserwacje których nie możemy jasno określić jako błędne, a mogą stanowić dowód istnienia interesującego moderatora.

3 Losowe/Random

Bardzo rzadkie, ale nie niemożliwe przypadki, które nie przynależą do żadnej z poprzednich dwóch kategorii.

Co zrobić z wartościami odstającymi?

1 Błędne

Jeżeli jest to możliwe, powinny być korygowane.

Np. jeśli błąd pojawił się podczas przepisywania danych początkowych, należy powrócić do nich i wstawić poprawną wartość.

2 (Potencjalnie) interesujące i losowe

Strategia 1: Zachowanie.

Strategia 2: Usunięcie.

Strategia 3: Rekodowanie. (np. zastosowanie jakiejś funkcji na wszystkich obserwacjach, aby zmniejszyć ich wariancję, tj. log, arcsin)

Gdy pojawi nam się wiele interesujących obserwacji odstających, które formują wyraźny kształt, może to sugerować istnienie moderatora - wtedy autorzy sugerują usunięcie tych danych i przestudiowanie ich w kolejnych badaniach.

Standardowa metoda dla obserwacji jednowymiarowych

Najbardziej powszechne w badaniach statystycznych w psychologii.

Standardowa metoda dla obserwacji jednowymiarowych

$$X_i \notin (\bar{X} - c \cdot \sigma, \bar{X} + c \cdot \sigma)$$

$\bar{X}$ - średnia,

$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2}$ - odchylenie standardowe,

$c \in \{3, 3.29\}$ - stała.

Wybór wartości c wynika z faktu, że w rozkładzie normalnym 99.7% obserwacji znajduje się w odległości do $3 \cdot \sigma$, a 99.9% do $3.29 \cdot \sigma$ od średniej ($\bar{X}$).

Standardowa metoda dla obserwacji wielowymiarowych

Dystans Mahalanobisa

$$D_M = \sqrt{(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})}$$

$\mathbf{X} = (X_1, X_2, \dots, X_k)$ - wektor zmiennych

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$ - wektor wymiaru k

$\boldsymbol{\Sigma}$ - macierz kowariancji wymiaru $k \times k$

Standardowa metoda dla obserwacji wielowymiarowych

$$\Sigma = [a_{ij} = \text{cov}(X_i, X_j)]_{i,j=1,\dots,k} =$$
$$\begin{pmatrix} \sigma_1^2 & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_k) \\ \text{cov}(X_2, X_1) & \sigma_2^2 & \dots & \text{cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_k, X_1) & \text{cov}(X_k, X_2) & \dots & \sigma_k^2 \end{pmatrix}$$

$$\text{cov}(X_i, X_j) = \frac{1}{k-1} \sum_{a=1}^k (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)$$

Standardowa metoda dla obserwacji wielowymiarowych

W praktyce μ i Σ nie są nam znane. Na podstawie próby $X_1, X_2, \dots, X_n$, estymujemy μ za pomocą $\bar{X}$, a Σ za pomocą $S = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$.

Obserwacja X_i jest odstająca, gdy spełnia

$$\sqrt{(X_i - \bar{X})^T S^{-1} (X_i - \bar{X})} > c_k$$

gdzie c_k jest pewną stałą zależną od k - wymiaru.

Wizualne porównanie zastosowania dystansu Mahalanobisa z zastosowaniem miary Euklidesa. Różnica wynika z tego, że dystans Mahalanobisa bierze pod uwagę strukturę kowariancji danych.

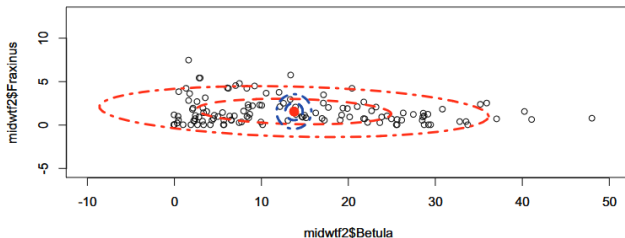


FIG. 3.1: Schematic comparison of the Mahalanobis (ellipse) and Euclidean (circle) distances calculated for a data set. The two lines, circles and ellipses, correspond to the Euclidean and the Mahalanobis distances, of one and two units apart from the center of data

Cykliczność w wykrywaniu wartości odstających

Zauważmy, że obie przedstawione metody opierają się na średniej i standardowym odchyleniu, na które wpływają **wartości obserwacji odstających**, które mają znaleźć.

Tej cykliczności w wykrywaniu wartości odstających możemy uniknąć, stosując **odporne wskaźniki** wartości odstających.

Punkt przegięcia - *breakdown point*

Definicja (Punkt przegięcia)

The proportion of values set to infinity (and thus outlying) that can be part of the dataset without corrupting the estimator used to classify outliers.

Proporcja wartości w zbiorze danych, które można ustawić jako nieskończoność (tym samym robiąc z nich wartości odstające), bez uszkadzania estymatora używanego do klasyfikacji wartości odstających.

Punkt przegięcia - *breakdown point*

Definicja (Punkt przegięcia)

Proporcja wartości w zbiorze danych, które można ustawić jako nieskończoność (tym samym robiąc z nich wartości odstające), bez uszkodzania estymatora używanego do klasyfikacji wartości odstających.

Przykład

Mediana ma punkt przegięcia 0.5, który jest największy możliwy.

$X = \{2, 3, 4, INF, INF, INF\}$, wtedy $Me(X) = \frac{4+INF}{2} = INF$

$Z = \{2, 3, 4, 5, INF, INF\}$, wtedy $Me(Z) = \frac{4+5}{2} = 4.5$

Punkt przegięcia - *breakdown point*

Definicja (Punkt przegięcia)

Proporcja wartości w zbiorze danych, które można ustawić jako nieskończoność (tym samym robiąc z nich wartości odstające), bez uszkodzania estymatora używanego do klasyfikacji wartości odstających.

Przykład

Średnia, odchylenie standardowe i dystans Mahalanobisa mają punkt przegięcia 0, który jest najmniejszy możliwy.

$X = \{2, 3, 4, INF, INF, INF\}$, wtedy $\bar{X} = \sigma(X) = INF$

$Z = \{2, 3, 4, 5, INF, INF\}$, wtedy $\bar{X} = \sigma(X) = INF$

Proponowane metody

Potrzebujemy metod znajdowania obserwacji odstających, które będą miały jak największy punkt przegięcia.

Proponowana metoda dla obserwacji jednowymiarowych: zastosowanie MAD

$$x_i \notin (\text{Me}(X) - k \cdot \text{MAD}, \quad \text{Me}(X) + k \cdot \text{MAD})$$

$\text{MAD} = \text{Me}(|x_i - \text{Me}(X)|)$ - Median Absolute Deviation

k - stała, domyślnie przyjmująca wartość 1.4826

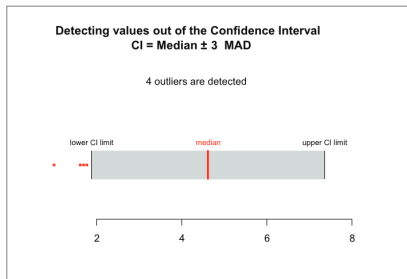


Figure 3: Univariate extreme values of sense of coherence (Antonovsky, 1987) detected by the MAD method on a sample of 2077 subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016).

Leys et al: How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration

Proponowana metoda dla obserwacji wielowymiarowych: zastosowanie MCD

Idea polega na odnalezieniu pewnej h liczby obserwacji, które nie są obserwacjami odstającymi, i wyliczeniu na ich podstawie częściowej średniej i kowariancji.

Proponowana metoda dla obserwacji wielowymiarowych: zastosowanie MCD

Dla n obserwacji $h \in \{\frac{n}{2}, \dots, n\}$. Iterujemy po wszystkich możliwych wartościach h i wybieramy tę, dla której wyznacznik macierzy kowariancji jest minimalny. Naszym celem jest znalezienie najbardziej "centralnej" takiej próbki h elementów, i minimalny wyznacznik zapewni nam minimalną zmienność między obserwacjami.

Stąd pochodzi nazwa MCD (Minimal Covariance Determinant).

Proponowana metoda dla obserwacji wielowymiarowych: zastosowanie MCD

Estymatory lokalizacji i rozrzutu oznaczone $\hat{X}$ i $\hat{\Sigma}_{MCD}$, odpowiadają średniej i macierzy kowariancji wybranej w ten sposób próbki.

$$D_{MMCD} = \sqrt{(X - \hat{X})^T \hat{\Sigma}_{MCD}^{-1} (X - \hat{X})}$$

$$\sqrt{(X_i - \hat{X})^T \hat{\Sigma}_{MCD}^{-1} (X_i - \hat{X})} > c_k$$

$c_k = \sqrt{\chi_k^2(1 - \alpha)}$ co wynika z tego, że kwadrat D_{MMCD} możemy przybliżyć za pomocą rozkładu χ_k^2 .

Ponadto $1 - \alpha$ to najczęściej 0.9, 0.95, 0.975, 0.99, 0.999.

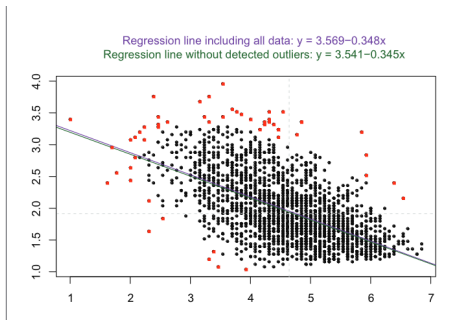


Figure 4: Bivariate extreme values when considering the combination of sense of coherence (Antonovsky, 1987) and anxiety and depression symptoms (Derogatis et al. 1974) detected by the MCD method on a sample of 2077 subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016).

Leys et al: How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration

Podsumowanie

- Poprawa lub usunięcie obserwacji odstających błędne.
- Unikanie stosowania metod, które wykorzystują wariancję lub średnią.
- Stosowanie zaproponowanych metod.
- Prerejestracja.

Bibliografia

- [1] Hamid Ghorbani. “Mahalanobis Distance and Its Application For Detecting Multivariate Outliers”. W: *Facta Universitatis* 34.3 (2019), s. 583–595. DOI: 0.22190/FUMI1903583G.
- [2] Christophe Leys i in. “Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance”. W: *Journal of Experimental Social Psychology* 74 (2018), s. 150–156. DOI: 10.5334/irsp.289.
- [3] Christophe Leys i in. “How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration”. W: *International Review of Social Psychology* 32(1).5 (2019), s. 1–10. DOI: 10.5334/irsp.289.