Sains Malaysiana 52(9)(2023): 2725-2732
http://doi.org/10.17576/jsm-2023-5209-20

Statistical Method for Finding Outliers in Multivariate Data using a Boxplot and Multiple Linear Regression

(Kaedah Statistik untuk Mencari Data Terpencil dalam Data Multivariat menggunakan Plot Kotak dan Regresi Linear Berganda)

THEERAPHAT THANWISET & WUTTICHAI SRISODAPHOL*

Department of Statistics, Khon Kaen University, 40002 Khon Kaen, Thailand

Received: 1 December 2022/Accepted: 15 August 2023

ABSTRACT

The objective of this study was to propose a method for detecting outliers in multivariate data. It is based on a boxplot and multiple linear regression. In our proposed method, the box plot was initially applied to filter the data across all variables to split the data set into two sets: normal data (belonging to the upper and lower fences of the boxplot) and data that could be outliers. The normal data was then used to construct a multiple linear regression model and find the maximum error of the residual to denote the cut-off point. For the performance evaluation of the proposed method, a simulation study for multivariate normal data with and without contaminated data was conducted at various levels. The previous methods were compared with the performance of the proposed methods, namely, the Mahalanobis distance and Mahalanobis distance with the robust estimators using the minimum volume ellipsoid method, the minimum covariance determinant method, and the minimum vector variance method. The results showed that the proposed method had the best performance over other methods that were compared for all the contaminated levels. It was also found that when the proposed method was used with real data, it was able to find outlier values that were in line with the real data.

Keywords: Boxplot; multivariate data; multiple linear regression; outlier

ABSTRAK

Objektif kajian ini adalah untuk mencadangkan kaedah untuk mengesan data terpencil dalam data multivariat. Ia berdasarkan plot kotak dan regresi linear berganda. Dalam kaedah yang kami cadangkan, plot kotak pada mulanya digunakan untuk menapis data merentas semua pemboleh ubah untuk membahagikan set data kepada dua set: data biasa (kepunyaan pagar atas dan bawah plot kotak) dan data yang boleh menjadi data terpencil. Data biasa kemudiannya digunakan untuk membina model regresi linear berganda dan mencari ralat maksimum baki untuk menandakan titik potong. Untuk penilaian prestasi kaedah yang dicadangkan, kajian simulasi untuk data normal multivariat dengan dan tanpa data tercemar telah dijalankan pada pelbagai peringkat. Kaedah sebelumnya dibandingkan dengan prestasi kaedah yang dicadangkan, iaitu, jarak Mahalanobis dan jarak Mahalanobis dengan penganggar teguh menggunakan kaedah ellipsoid isi padu minimum, kaedah penentu kovarian minimum dan kaedah varians vektor minimum. Keputusan menunjukkan bahawa kaedah yang dicadangkan mempunyai prestasi terbaik berbanding kaedah lain yang dibandingkan untuk semua tahap yang tercemar. Didapati juga apabila kaedah yang dicadangkan digunakan dengan data sebenar, ia dapat mencari nilai data terpencil yang selari dengan data sebenar.

Kata kunci: Data berbilang variasi; data terpencil; plot kotak; regresi linear berganda

INTRODUCTION

In statistics, outliers are data points that are noticeably different from the rest. Outliers are divided into three categories: To begin with, there is intrinsic variability, which refers to variances that exist spontaneously within a group. Experiments with plants produced by soil fertility in experimental plots, for example, are not the same. Second, measurement error is the discrepancy between the measured quantity and its true value that causes the fluctuation. Third, data is incorrectly recorded by selecting a biased sample or including people who are not representative of the group being sampled (Anscombe

& Guttman 1960). Outliers in datasets have an impact on data analysis. Even if the results of the study are not statistically significant, the outliers will render them statistically significant, and vice versa. Outlier troubleshooting: if the resultant outlier is impracticable or the product of erroneous data gathering, it could be excluded from the study. However, if an outlier is confirmed to be caused by an anomaly in the real sample, it will be investigated.

There are three main statistical analysis procedures when it comes to the level of analysis. Univariate, bivariate, and multivariate analyses are the three types of analyses. When there is just one variable in the data, the most fundamental statistical data analysis technique is univariate analysis. When compared to univariate analysis, bivariate analysis is slightly more analytical when there are two variables in the data set. Multivariate analysis is a more complicated type of statistical analysis that is performed when a data set has more than two variables.

There are currently several methods for detecting outliers and the most familiar is boxplot (Tukey 1977). It is a method used to detect outliers for univariate data. There are a variety of ways for detecting outliers. Meanwhile, bivariate and multivariate data may or may not have a dependent variable. If the data contains a dependent variable, many works have proposed outlier detection methods, including Cook's distances (Cook 1977), the hat matrix (Hoaglin & Welsch 1978), DFFITS (Belsley, Kuh & Welsch 1980), studentized residuals (Montgomery, Peck & Vining 2012), and R-student (Montgomery, Peck & Vining 2012).

If the data does not contain a dependent variable, the Mahalanobis distance (MD) has been used (Mahalanobis 1936), which is essentially the distance of the vector from the mean with the covariance matrix. The Mahalanobis distance of samples based on the maximum likelihood estimators (MLEs) of the mean vector and covariance matrix follows a chi-square distribution with p degrees of freedom, where p is the number of variables. If the observations have a Mahalanobis distance greater than the quantile value, as $1-\alpha$ of $\chi^2_{1-\alpha,p}$, where $\chi^2_{1-\alpha,p}$ is the $100(1-\alpha)^{th}$ percentile of a chi-square distribution with p degrees of freedom. The cut-off point for detecting outliers is $\sqrt{\chi^2_{1-\alpha,p}}$ usually used $\alpha = 0.05$, they will be considered outliers. Also, robust estimators of the mean vector and covariance matrix were used to find outliers using the Mahalanobis distance. These estimators were made by using the minimum volume ellipsoid (MVE) (Aelst & Rousseeuw 2009), the minimum covariance determinant (MCD) (Hubert & Debruyne 2010), and the minimum vector variance (MVV) (Herdiani, Sari & Sunusi 2019). The MVE is based on the minimum volume ellipsoid, the MCD is based on the minimum covariance determinant, and the MVV is based on the minimum vector variance that covers a subset of observations, respectively.

In this research, only those datasets that do not contain a dependent variable are considered since these datasets with high dimensionality and large sample sizes usually appear in the organizations that collect them. The detection of outliers in these datasets is widely used Mahalanobis distance methods follows a chi-square distribution, which is necessarily denoted by the quantile value. So, the dataset might contain outliers, but some datasets do not necessarily contain outliers. Therefore, to eliminate such a problem, this research aims to propose a method for detecting outliers in multivariate data in another way by combining a boxplot for univariate data with multiple linear regression. The boxplot for univariate data is initially filtered such that the data is in the boxplot for all variables that are denoted as normal data. After that, multiple linear regression is used to find outliers in the rest of the data.

The paper is organized as follows. Next section describes the proposed method. In the following section, an experiment on simulated data with contaminated multivariate normal data to compare the proposed methods with the previous methods. Subsequntly, the behavior with a real dataset example was discussed. Finally, last section provides some conclusions and discussion.

PROPOSED METHOD

This research proposes the outlier detection method in multivariate data in another way by combining a boxplot for univariate data with multiple linear regression. In this method, multiple linear regression analysis is used together with data split with a boxplot to initially filter each variable. This method is called multiple linear regression using data split with the boxplot method (MLRSB) and is explained in the following steps:

Step 1 For a dataset with n observations and p variables, a boxplot is used to filter the data in each variable, splitting the data into 2 sets. The first dataset (n_1 observations and p variables) is an observation with at least one variable outside the upper and lower fences of boxplot, and the second dataset (n_2 observations and p variables) is an observation where all variables are in the upper and lower fences.

Step 2 After getting 2 sets of data in step 1, the multiple linear regression equations (p equations) are constructed and computed $R_l^2; l = 1, 2, \ldots, p$ from the second dataset. All the variables are thought of as a single dependent variable (Y), while the other variables are called independent variables,

$$\hat{Y}_1 = b_0 + b_2 x_2 + b_3 x_3 + \ldots + b_p x_p$$

$$\hat{Y}_2 = b_0 + b_1 x_1 + b_3 x_3 + \ldots + b_p x_p$$

$$\vdots$$

$$\hat{Y}_p = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_{p-1} x_{p-1}.$$

Step 3 For the $R_l^2; l = 1, 2, \ldots, p$ of multiple linear regression equations in step 2, the multiple linear regression equation with the maximum R_l^2 is selected to denote the cut-off point. The cut-off point is the largest absolute error in this equation.

Step 4 For the multiple linear regression equation with the maximum R_l^2 that is found in step 3, the predicted values are found using the data from the first dataset, and the absolute errors are also found for that prediction.

Step 5 If any observation in the first dataset has an absolute error in step 4 greater than the cut-off point obtained from step 3, it is labeled an outlier.

This proposed outlier detection method is illustrated in Figure 1.

EXPERIMENT ON SIMULATED DATA AND PERFORMANCE COMPARISON

In this section, the data using a multivariate normal distribution with and without contaminated data are simulated to compare the performance of the proposed method with the previous methods.

Step 1 The p -variate normal data with sample sizes n = 100, 300, 500, 800, 1000 and variables p = 2, 3, 4, 6, 8, 10 with and without contaminated data for the multivariate normal distribution are generated using the R program. The contaminated multivariate normal distribution given as a mixture of normal is given as below.

$$(1-\alpha)N(\mathbf{0}, I) + \alpha N(\delta\boldsymbol{\mu}, \lambda I),$$

where $\boldsymbol{\mu}$ denotes the p -dimensional vector of ones and covariance matrix I, the distance of the outliers δ = 5, 10 and the concentration of the contamination λ = 0.1, 1. The contamination levels is α (Cabana, Lillo & Laniado 2021).

The simulated data with and without contaminated data can be considered under the denoting value of α. If $\alpha = 0$, the simulated data is without contaminated data, and vice versa. The contaminated data that is inserted into the generated data are called outliers with α = 0.01, 0.05, 0.10.

Step 2 The proposed method (MLRSB) is used to find outliers in each situation.

Step 3 Steps 1 to 2 are repeated for 1,000 iterations.

Step 4 The proportion of the detected outliers for those without contaminated data and accuracy, precision, recall, and F1-Score for those with contaminated data are calculated, which is the criterion to evaluate the performance of the proposed methods.

The performance of the proposed method and the previous methods (MD, MVE, MCD, MVV) are separated into two cases. Case I: Without contaminated data (α = 0), the proportion of the detected outliers for p = 2, 3, 4, 6, 8, 10 and n = 100, 300, 500, 800, 1000 are showed in Table 1. Case II: With contaminated data (α = 0.01, 0.05, 0.10), the accuracy, precision, recall, and F1-Score for p = 10, n = 500, δ = 5, 10, λ = 0.1, 1 are showed in Tables 2-5. (Another situation can also be available from the corresponding author).

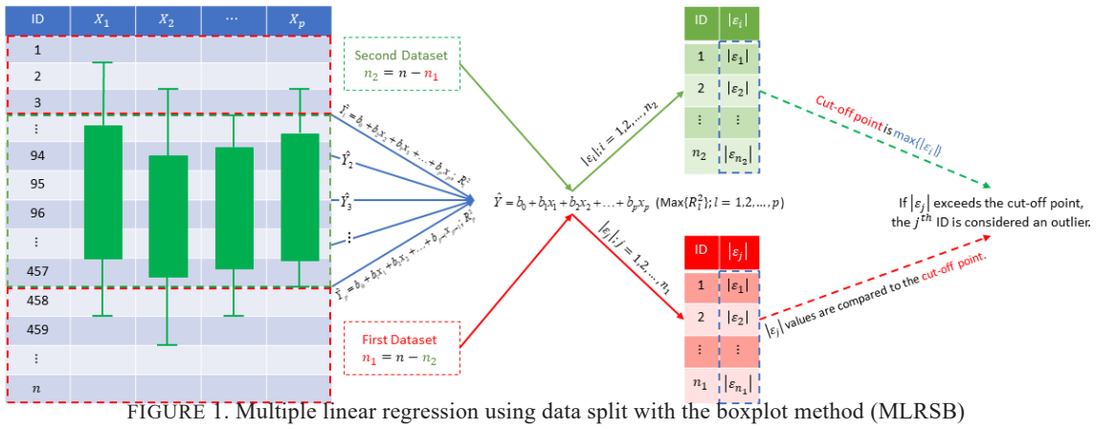

FIGURE 1. Multiple linear regression using data split with the boxplot method (MLRSB)

TABLE 1. Proportion of the detected outliers

n	p	MLRSB	MD	MVE	MCD	MVV
100	2	0.0103	0.0470	0.0546	0.0518	0.4257
	3	0.0100	0.0454	0.0585	0.0550	0.4073
	4	0.0102	0.0454	0.0651	0.0601	0.4081
	6	0.0104	0.0418	0.0770	0.0670	0.4067
	8	0.0119	0.0403	0.0916	0.0787	0.4104
	10	0.0131	0.0382	0.1117	0.0943	0.4100
300	2	0.0076	0.0489	0.0498	0.0490	0.4144
	3	0.0080	0.0487	0.0501	0.0499	0.3899
	4	0.0076	0.0483	0.0505	0.0502	0.3779
	6	0.0075	0.0477	0.0519	0.0506	0.3661
	8	0.0072	0.0468	0.0529	0.0507	0.3669
	10	0.0075	0.0461	0.0551	0.0513	0.3717
500	2	0.0072	0.0494	0.0492	0.0492	0.4119
	3	0.0071	0.0495	0.0495	0.0496	0.3829
	4	0.0069	0.0489	0.0495	0.0494	0.3671
	6	0.0069	0.0486	0.0497	0.0496	0.3492
	8	0.0067	0.0479	0.0496	0.0492	0.3445
	10	0.0071	0.0480	0.0510	0.0497	0.3445
800	2	0.0069	0.0495	0.0492	0.0492	0.4065
	3	0.0069	0.0494	0.0489	0.0494	0.3773
	4	0.0067	0.0493	0.0490	0.0494	0.3581
	6	0.0066	0.0487	0.0485	0.0490	0.3388
	8	0.0065	0.0489	0.0491	0.0492	0.3294
	10	0.0064	0.0486	0.0492	0.0490	0.3250
1000	2	0.0070	0.0497	0.0493	0.0494	0.4057
	3	0.0071	0.0494	0.0490	0.0494	0.3749
	4	0.0068	0.0491	0.0488	0.0492	0.3562
	6	0.0066	0.0493	0.0488	0.0494	0.3357
	8	0.0064	0.0490	0.0488	0.0492	0.3241
	10	0.0062	0.0487	0.0489	0.0489	0.3186

TABLE 2. Accuracy, precision, recall and F1-Score for $p = 10$, $n = 500$, $\delta = 5$, $\lambda = 0.1$

α	Criteria	MLRSB	MD	MVE	MCD	MVV
0.01	Accuracy	0.9940	0.9647	0.9520	0.9535	0.6617
	Precision	0.6274	0.2205	0.1726	0.1769	0.0287
	Recall	0.9932	1.0000	1.0000	1.0000	1.0000
	F1-Score	0.7690	0.3613	0.2944	0.3006	0.0558
0.05	Accuracy	0.9954	0.9401	0.9623	0.9638	0.6900
	Precision	0.9273	0.4273	0.5702	0.5799	0.1389
	Recall	0.9859	0.5846	1.0000	1.0000	1.0000
	F1-Score	0.9557	0.4937	0.7263	0.7341	0.2439
0.10	Accuracy	0.9963	0.8511	0.9727	0.9734	0.5675
	Precision	0.9804	0.0000	0.7856	0.7897	0.0952
	Recall	0.9831	0.0000	1.0000	1.0000	0.3910
	F1-Score	0.9817	N/A	0.8799	0.8825	0.1531

TABLE 3. Accuracy, precision, recall and F1-Score for $p = 10$, $n = 500$, $\delta = 5$, $\lambda = 1$

α	Criteria	MLRSB	MD	MVE	MCD	MVV
0.01	Accuracy	0.9936	0.9668	0.9527	0.9539	0.6625
	Precision	0.6158	0.2312	0.1745	0.1782	0.0288
	Recall	0.9568	1.0000	1.0000	1.0000	1.0000
	F1-Score	0.7493	0.3756	0.2971	0.3025	0.0560
0.05	Accuracy	0.9932	0.9689	0.9619	0.9635	0.6908
	Precision	0.9213	0.6193	0.5674	0.5778	0.1392
	Recall	0.9449	0.9790	1.0000	1.0000	1.0000
	F1-Score	0.9330	0.7587	0.7240	0.7324	0.2444
0.10	Accuracy	0.9913	0.9082	0.9722	0.9731	0.7261
	Precision	0.9790	0.5620	0.7822	0.7879	0.2675
	Recall	0.9330	0.3711	1.0000	1.0000	1.0000
	F1-Score	0.9554	0.4470	0.8778	0.8814	0.4221

TABLE 4. Accuracy, precision, recall and F1-Score for $p = 10$, $n = 500$, $\delta = 10$, $\lambda = 0.1$

α	Criteria	MLRSB	MD	MVE	MCD	MVV
0.01	Accuracy	0.9939	0.9676	0.9520	0.9536	0.6625
	Precision	0.6208	0.2359	0.1724	0.1772	0.0288
	Recall	1.0000	1.0000	1.0000	1.0000	1.0000
	F1-Score	0.7660	0.3817	0.2941	0.3011	0.0560
0.05	Accuracy	0.9963	0.9608	0.9620	0.9634	0.6902
	Precision	0.9306	0.5613	0.5680	0.5771	0.1390
	Recall	1.0000	0.9909	1.0000	1.0000	1.0000
	F1-Score	0.9641	0.7167	0.7245	0.7318	0.2441
0.10	Accuracy	0.9981	0.8519	0.9727	0.9736	0.5821
	Precision	0.9813	0.0000	0.7856	0.7914	0.1099
	Recall	1.0000	0.0000	1.0000	1.0000	0.4480
	F1-Score	0.9906	N/A	0.8799	0.8836	0.1765

TABLE 5. Accuracy, precision, recall and F1-Score for $p = 10$, $= 500$, $\delta = 10$, $\lambda = 1$

α	Criteria	MLRSB	MD	MVE	MCD	MVV
0.01	Accuracy	0.9938	0.9690	0.9522	0.9539	0.6630
	Precision	0.6171	0.2438	0.1730	0.1783	0.0288
	Recall	0.9998	1.0000	1.0000	1.0000	1.0000
	F1-Score	0.7632	0.3920	0.2950	0.3026	0.0560
0.05	Accuracy	0.9962	0.9706	0.9624	0.9634	0.6917
	Precision	0.9290	0.6301	0.5706	0.5777	0.1395
	Recall	0.9997	0.9993	1.0000	1.0000	1.0000
	F1-Score	0.9631	0.7729	0.7266	0.7323	0.2448
0.10	Accuracy	0.9980	0.9101	0.9725	0.9735	0.7259
	Precision	0.9808	0.5747	0.7840	0.7905	0.2673
	Recall	0.9998	0.3885	1.0000	1.0000	1.0000
	F1-Score	0.9902	0.4636	0.8789	0.8830	0.4218

Tables 1-5 correspond to all simulation scenarios for multivariate normal data with and without contaminated data. For Table 1, the results showed that for the proposed method, the MLRSB had the proportion of the detected outliers at or near the contaminated level of 0, or in the case of no contamination, but the other methods (MD, MVE, MCD, MVV) had the proportion of the detected outliers far from the contaminated level of 0. It means that the MLRSB indicates there are no outliers when some datasets do not contain outliers, as the preferred method should have the proportion of detected outliers at 0 in this instance. For Tables 2-5, the results showed that the MLRSB had higher accuracy, precision, recall, and F1-Score than the other methods. The preferred method should have accuracy, precision, recall, and F1-Score as high as possible and higher than other methods. When the precision and recall of the MLRSE are considered, the results indicate that it has a high actual number of correctly predicted outliers that came out to be outliers and actual outliers that were correctly predicted, respectively.

REAL DATA ANALYSIS

In this section, a real dataset is used to show the efficiency of the proposed method. Hepatitis C virus (HCV) data, which contains laboratory values of blood donors and Hepatitis C patients (Lichtinghagen, Klawonn & Hoffmann 2020), there are 589 records and 11 variables. The 11 variables are explained as follows, X_1: Age (in years), X_2: Albumin (ALB) is the amount of protein that floats in the bloodstream and is produced by the liver, X_3: Alkaline phosphatase (ALP) is the amount of enzyme produced by proteins in diseased or dysfunctional organs, such as the liver, X_4: Alanine transaminase (ALT) is the amount of an enzyme that floats in the bloodstream that can be caused by damage to any organ, such as the liver, X_5: Aspartate aminotransferase (AST) is the amount of an enzyme used to help diagnose liver disease, X_6: Bilirubin (BIL) is a blood breakdown value that is used as an important indicator of liver disease, X_7: Cholinesterase (CHE) is a value used to diagnose the degree of intoxication, clarify the condition and assess liver function, X_8: Cholesterol (CHOL) is the amount of cholesterol that comes from food or is made by the liver, X_9: Creatinine (CREA) is the amount of waste produced by muscles. This value is used to see if the kidneys are able to filter waste products from the blood and excrete urine normally, X_{10}: Gamma-glutamyl transferase (GGT) is the amount of an enzyme produced by the liver that aids in detoxification, and X_{11}: Protein (PROT) is the amount of protein in the bloodstream used as an indicator of liver function.

The efficient method that was obtained in the simulation study for detecting outliers, that is, the MLRSB, will be applied to real data. In this dataset, blood donors and hepatitis C patients were originally divided into five categories in order of severity, from least to greatest: 0 (526 blood donors), 0s (7 suspect blood donors), 1 (20 hepatitis), 2 (12 fibrosis), and 3 (24 cirrhosis). To make it easier to study, these five categories are broken as follows: non-morbid (0) in green points, moderate (0s, 1, 2) in gray points, and severe (3) in red points.

Figure 2 shows that the MLRSB can detect outliers (patients with severe symptoms), with individuals identified as having severe symptoms over the cut-off point value (blue line). 9 green points, 27 gray points, and 24 red points were detected by this method. This means that, based on real data, the MLRSB can find all patients with severe symptoms. Therefore, this method will be used to identify outliers that are confirmed to be caused by a strange thing in the real data.

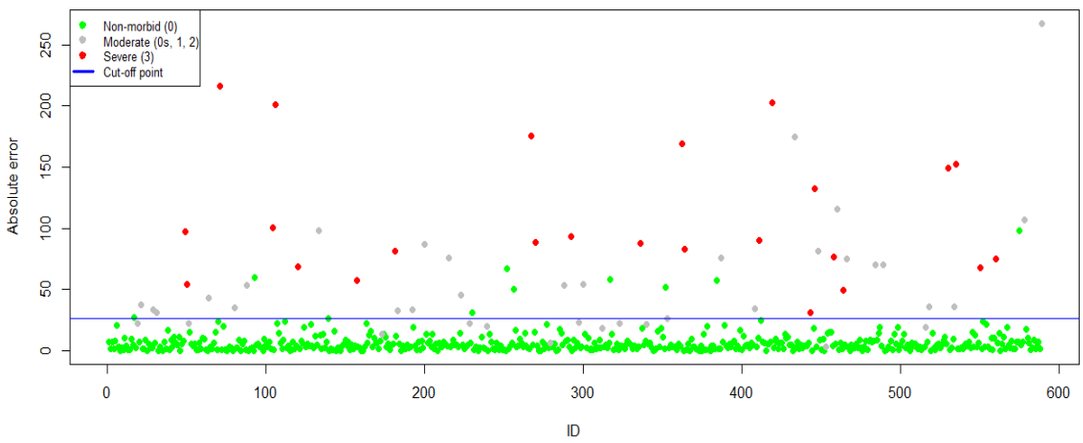

FIGURE 2. The MLRSB detected outliers for Hepatitis C virus (HCV) data

Conclusion and discussion

The datasets with high dimensionality and large sample sizes that usually appear in organizations are considered. The Mahalanobis distance and the Mahalanobis distance with robust approaches: MVE (Aelst & Rousseeuw 2009), MCD (Hubert & Debruyne 2010), and MVV (Herdiani, Sari & Sunusi 2019) were used to detect outliers. In the simulation study, they still label outliers in the dataset even if there is no contamination. The proposed method (MLRSB) could solve a problem in cases where some datasets do not have to have outliers if there is no contamination. The MLRSE and other methods that were compared have a high number of actual outliers that were correctly predicted. But the MLRSE has a higher actual number of correctly predicted outliers that came out to be outliers than other methods. This superiority was evident in various key metrics such as accuracy, precision, recall, and the F1-Score. Therefore, the MLRSB is suggested to be used to check whether there are outliers or not and to find the outliers. The standout feature of MLRSB was its ability to accurately detect outliers while maintaining a proportion of detected outliers close to the contaminated level of 0. This characteristic is crucial in cases where the absence of outliers needs to be accurately identified. The method effectively minimized false positives and false negatives, making it a reliable tool for distinguishing between data with and without outliers. Furthermore, the MLRSB method demonstrated exceptional performance in identifying patients with severe symptoms in a real dataset of blood donors and hepatitis C patients. This suggests that the method has the potential to uncover instances that are truly caused by unusual factors or anomalies in real-world data. In practical applications, this capability could lead to a better understanding of rare or extreme cases, especially in domains such as healthcare, where identifying such cases is of paramount importance.

Acknowledgements

The Development and Promotion of Science and Technology Talents Project (DPST) is acknowledged.

References

Aelst, S.V. & Rousseeuw, P. 2009. Minimum volume ellipsoid. *WIREs Computational Statistics* 1: 71-82.

Anscombe, F.J. & Guttman, I. 1960. Rejection of outliers. *Technometrics* 2(2): 123-147.

Belsley, D.A., Kuh, E. & Welsch, R.E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Cabana, E., Lillo, R.E. & Laniado, H. 2021. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Stat Papers* 62: 1583-1609.

Cook, R.D. 1977. Detection of influential observations in regression. *Technometrics* 19: 15-18.

Herdiani, E.T., Sari, P.P. & Sunusi, N. 2019. Detection of outliers in multivariate data using minimum vector variance method. *Journal of Physics: Conference Series* 1341(9): 092004.

Hoaglin, D.C. & Welsch, R.E. 1978. The hat matrix in regression and ANOVA. *The American Statistician* 32: 17-22.

Hubert, M. & Debruyne, M. 2010. Minimum covariance determinant. *WIREs Computational Statistics* 2: 36-43.

Lichtinghagen, R., Klawonn, F. & Hoffmann, G. 2020. *UCI Machine Learning Repository*. Irvine: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/datasets/HCV+data

Mahalanobis, P.C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 2(1): 49-55.

Montgomery, D.C., Peck, E.A. & Vining, G.G. 2012. *Introduction to Linear Regression Analysis*. 3rd ed. New York: John Wiley & Sons.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Massachusetts: Addison Wesley.

*Corresponding author; email: wuttsr@kku.ac.th