Black Sea Journal of Engineering and Science
Open Access Journal
e-ISSN: 2619-8991

BSPublishers

Research Article
Volume 2 - Issue 1: 7-10 / January 2019

A STUDY ON DETERMINATION OF OUTLIER OBSERVATIONS BY USING CHI-SQUARE THRESHOLD VALUE

Fahrettin KAYA[1]*, Esra YAVUZ[2], Şeyma KOÇ[2], Ömer Faruk KARAOKUR[2]

[1]Kahramanmaraş Sütçü Imam University, Andırın Vocational School, Computer Technology Department, 46410, Andırın, Kahramanmaras, Turkey.

[2]Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, 46040, Onikişubat, Kahramanmaraş, Turkey

Received: *October 13, 2018;* ***Accepted:*** *November 18, 2018;* ***Published:*** *January 01, 2019*

Abstract

Outlier observations are observations that are out of the tendency of all observations in a data set. The observations come out in situations such as faulty observation, incorrect data entry. It is important to be able to identify these observations as the results of statistical analysis, for example such as multiple regression analysis, can be quite sensitive against to these observations. Outlier observations are mostly determined by using distance calculation, statistical test and density based approaches. In this study, the distances of each observation vector to the center were calculated with Mahalanobis distance by using R program. For this purpose, the features such as hematokrit (htc), hemoglobin (hgb), mean platelet volume (mpv), platelet distribution width (pdw), nonbacterial prostatitis (nbp) and pulse pressure values measured in the blood of 315 heart patients were examined as data set. As a result of the research, sixteen observations were found as outlier observation. It is thought that the result of this study will help the researchers trying to find out especially the outlier observations.

Keywords: Outlier observation, Mahalanobis distance, Threshold value

***Corresponding author:** Kahramanmaraş Sütçü Imam University, Andırın Vocational School, Computer Technology Department, 46410, Andırın, Kahramanmaras, Turkey.
E mail: fkaya@ksu.edu.tr (F.KAYA)

Fahrettin KAYA	https://orcid.org/0000-0003-1666-4859
Esra YAVUZ	https://orcid.org/0000-0002-5589-297X
Şeyma KOÇ	https://orcid.org/0000-0001-5708-9905
Ömer Faruk KARAOKUR	https://orcid.org/0000-0002-3436-8415

Cite as: KayaF, Yavuz E, Koc S, Karaaokur OF. 2019. A study on determination of outlier observations by using Chi-Square threshold value. BSJ Eng Sci, 2(1): 7-10.

1. Introduction

Researchers generally analyze on multivariate data sets (Hubert and Van Der Veeken, 2008). It is likely that outlier observations may be exist in these data sets. These observations are called in different areas as error, defect, surprise, noise, exception. Such observations are called outlier observations (Gogoi et al., 2011). Outlier observations can have adverse effects on statistical analysis such as regression analysis, clustering analysis, factor analysis. Sometimes, while this data is an observation wanted in security areas, it may also be an observation of a disease in the field of health. Therefore, it is important to identify them. In order to detect outlier observations in multivariate data sets, approaches such as distance calculation, Bayesian, linear regression techniques have been developed (Gupta et al., 2013; Pei and Zaïane, 2006; Singh and Upadhyaya, 2012; Ting et al., 2007a; Ting et al., 2007b). In addition, the filters such as Kalman filter have been used because they deal with huge data that is unknown. (Liu et al., 2004; Rousseeuw and Hubert, 2011). The distance calculation methods are based on the Euclidean distance calculation, which calculates the distance between objects. (De Maesschalck et al., 2000; Hodg and Austin, 2004).

In case of mean and variance-covariance information of the multivariate data is known, such as the distances of observations to each other and the distance of an observation to the center and the distance between groups in the dataset is calculated with the aid of Mahalanobis distance (MD). An observation in a multivariate dataset can be defined as the whole of measured quantities such as height, weight, blood pressure, and level of sugar on blood. In addition, an observation may also contain measurement of a feature (blood sugar) at different measuring times, such as t1, t2, t3. While distance is calculated in such data structures, the distance measure of Mahalanobis becomes important when the correlation is considered. However, these traditional methods of detecting outliers' observations are based on the assumption that the data has the same species and normal distribution (Liu et al., 2004).

The fields used for MD are quite extensive. For example, learning machine on computer field is important (Xiang et al., 2008). The researchers, who investigated environmental, biological and natural phenomena and produced solutions, have benefited from this distance measure (Calenge et al., 2008). Moreover, this distance measure is used on time-dependent data, on the modeling of bioclimatic changes, on chemical data. (Egan and Morgan, 1998; Farber and Kadmon, 2003; Teng, 2010). In this study, it was aimed to determine the outliers in the multivariate data set and to display them on the quantile-quantile plot graph (Q-Q) (Url1). This graph is one of the methods used to visually determine whether normality is provided. When it is drawn using two data sets whose values are in ascending order.

2. Material and Method

In this study, hematokrit (htc), hemoglobin (hgb), mean platelet volume (mpv) and platelet distribution width (pdw) values which are measured in the blood of 315 heart patients and consisting of their nonbacterial prostatitis (nbp) and pulse pressure values, data sets with six variables were used as material. This data set was taken from the treated patient data in the medical school.

2.1. Statistical Usage of MD

MD is modeled by using of multivariate normal distribution and Chi-square distribution. In multivariate data analysis, the dataset matrix, which has *nxp* dimension *X* (observation) and *p* column (variable), is shown in Table 1.

Table 1. X data set matrix

	X data set variables			
Observation Number	x1	x1	…	xp
1	x11	x12	…	x1p
2	x21	x22	…	x2p
…	…	…	…	…
n	xn1	xn2	…	xnp

The probability density function of such a data set is as in the following;

$$f(x) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} . e^{-\frac{1}{2}}(x-\mu).\Sigma^{-1}.(x-\mu)^T \qquad (1)$$

is expressed so. In this formula, *Σ* parameter shows the variance-covariance matrix and *μ* parameter shows the mean vector. However, theoretically, as these parameters of a data set are not known in certain, the mean vector $\overline{X}$ is used instead of the vector *μ* and the sampling matrix *S* is used instead of the matrix *Σ*. The value of Mahalanobis distance which has m_j^2 calculated for *j*th observation is given in Equation (2). In this formula, m_j^2 expresses the distance of the *j*th observation vector to the mean vector.

$$m_j^2 = (x_{ij} - \overline{x_\iota})^T S^{-1} (x_{ij} - \overline{x_\iota}) \qquad (2)$$

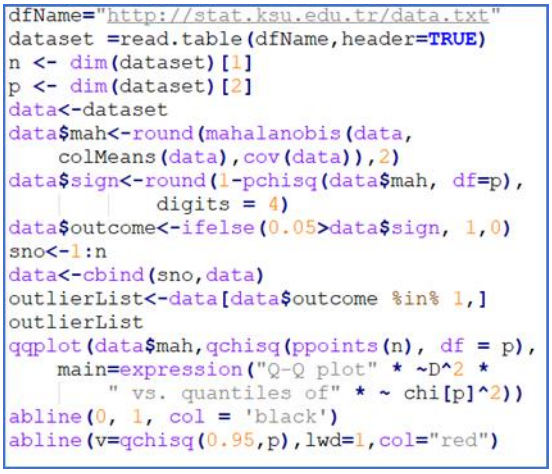
where;

S: Variance-covariance matrix of *pxp* variables, *X*: *px1* observation vector and $\overline{X}$: px1 mean vector.

In fact, in the closed form of equation (1), the m_j^2 value is summed by standardizing *p* random variables each *X* with normal distribution. If an *X* random variable has Standard Normal Distribution, the *Y* random variable shown in Equation (3) has Chi-square distribution with *p*-freedom degree (Url2).

$$Y = X_1^2 + X_2^2 + \cdots + X_p^2 \qquad (3)$$

In this case, m_j^2 values have a distribution of $\chi^2_{(p)}$. The outlier observations in the heart disease dataset that are the subject of this study, is determined using the following R program code. Furthermore, the web application which can run as R codes in Figure 1 is also designed in order to help the researchers to study easily (Url3).



```
dfName="http://stat.ksu.edu.tr/data.txt"
dataset =read.table(dfName,header=TRUE)
n <- dim(dataset)[1]
p <- dim(dataset)[2]
data<-dataset
data$mah<-round(mahalanobis(data,
    colMeans(data),cov(data)),2)
data$sign<-round(1-pchisq(data$mah, df=p),
            digits = 4)
data$outcome<-ifelse(0.05>data$sign, 1,0)
sno<-1:n
data<-cbind(sno,data)
outlierList<-data[data$outcome %in% 1,]
outlierList
qqplot(data$mah,qchisq(ppoints(n), df = p),
    main=expression("Q-Q plot" * ~D^2 *
        " vs. quantiles of" * ~ chi[p]^2))
abline(0, 1, col = 'black')
abline(v=qchisq(0.95,p),lwd=1,col="red")
```

Figure 1. R program sample codes

3. Results and Discussion

All m_j^2 distances were calculated and finalized. As a result of this calculation, list of sixteen outlier observations is shown together with the observation number in Table 2. The " m_i^2 ", "sign" and "outcome" calculation list of the observations is shown in Table 3. While this list is being prepared, chi-square 95% confidence limit has been taken basis. "Outcome" value is marked as "1" in case of $m_j^2 > 12.59$ and in other cases it is marked as "0". Additionally the value "sign" takes the value α which is $1-P\left(\chi^2_{(6)} < m_j^2\right)$. Also, outlier observations out of the chi-square 99% confidence limit is marked as "*". "Object numbers" of these observations are 127,162, 173, 175, 220 and 279.

Table 2. Outlier observations list

Object Number	htc	hgb	mpv	pdw	nbp	pulse
11	51.1	17.4	7.32	18.9	80	50
12	50.1	17.2	7.33	17.8	80	50
96	31.1	9.4	11.90	12.1	100	87
97	31.1	9.4	12.00	12.1	100	87
126	38.0	16.0	11.90	12.2	60	77
*127	37.0	16.0	11.10	19.1	60	77
133	31.1	9.4	11.80	12.1	100	87
*162	39.0	17.1	11.90	12.5	50	66
171	28.9	8.9	8.12	16.9	30	82
*172	38.0	17.9	11.80	13.6	40	66
*173	36.0	17.9	13.20	13.6	40	66
174	56.2	17.9	12.80	19.9	40	66
175	37.0	17.9	12.90	17.7	40	66
220	44.4	12.5	15.50	17.0	65	88
263	41.8	13.6	15.50	17.1	71	91
279	43.7	10.4	12.00	17.2	50	78

Table 3. Outlier observation calculation result list

Object Number	m_i^2	sign	outcome
11	15.57	0.0163	1
12	14.74	0.0224	1
96	14.05	0.0291	1
97	14.13	0.0282	1
126	15.78	0.0150	1
*127	19.09	0.0040	1
133	13.98	0.0299	1
*162	21.32	0.0016	1
171	14.92	0.0209	1
*172	32.92	0.0000	1
*173	43.26	0.0000	1
174	14.96	0.0206	1
*175	39.58	0.0000	1
*220	17.23	0.0085	1
263	12.63	0.0493	1
*279	23.78	0.0006	1

m_i^2:Mahalanobis distance, sign:Chi-square α level of significance, outcome:Outcome of the observation test (0-1)

Some intermediate calculations are given as in the following.

The variance-covariance matrix of the data set (S) is:

	htc	hgb	mpv	pdw	nbp	pulse
htc	20.87	6.49	1.03	1.60	-5.60	7.65
hgb	6.49	2.89	0.38	0.60	-4.29	-3.55
mpv	1.03	0.38	4.17	-1.97	4.34	-0.61
pdw	1.60	0.60	-1.97	8.77	-1.70	1.88
nbp	-5.60	-4.29	4.34	-1.70	258.62	20.72
pulse	-7.65	-3.55	-0.61	1.88	20.72	99.66

The inverse of the variance-covariance matrix is:

$$S^{-1} = \begin{bmatrix} 0.16 & -0.36 & -0.01 & -0.01 & 0.00 & 0.00 \\ -0.36 & 1.20 & -0.04 & -0.03 & 0.01 & 0.01 \\ -0.01 & -0.04 & 0.28 & 0.07 & -0.01 & 0.00 \\ -0.01 & -0.03 & 0.07 & 0.13 & 0.00 & 0.00 \\ 0.00 & 0.01 & -0.01 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.01 \end{bmatrix}$$

The mean vector of the data set is:

$$\begin{array}{ccccccc} & \text{htc} & \text{hgb} & \text{mpv} & \text{pdv} & \text{nbp} & \text{pulse} \\ \bar{X} = & \big[42.15 & 14.00 & 9.93 & 15.68 & 57.99 & 74.05\big] \end{array}$$

Additionally, the Q-Q plot of outlier observations is shown in Figure 2. In this graph, the threshold value has been drawn with a red line. The outlier observations are located to the right of this line.

Researchers can identify outlier observations because of mistakes while examining them. In this case, they can analyze these observations by separating from the data set. However, if the outlier observation is the real observation value, it need to be careful to distinguish this observation from the data set. Because, in this case, information may be lost. Therefore, in both cases, the result of the statistical analysis should be examined and decided.

4. Conclusions

As the sample size increases in a statistical study, it is known that population parameters are to be approached. However, when there is a wrong entry at the observation value, it is moved away from the parameter values. Therefore, it has been shown in the present study whether there is any outlier observations in the multivariate data set beyond the 95% and %99 confidence limit to detect such observations.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This research was presented as an oral presentation at the International Congress on Domestic Animal Breeding Genetics and Husbandry (ICABGEH-2018) held on 26-28 September 2018 in Antalya.

References

Calenge C, Darmon G, Basille M, Loison A, Jullien JM. 2008. The factorial decomposition of the Mahalanobis distances in habitat selection studies. Ecol, 89(2): 555–566.

Egan WJ, Morgan SL. 1998. Outlier detection in multivariate analytical chemical data. Anal Chem, 70(11): 2372–2379.

Farber O, Kadmon R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. ECMOD, 160 (1-2): 115-130.

Gogoi P, Bhattacharyya DK, Borah B, Kalita JK. 2011. A survey of outlier detection methods in network anomaly identification. Computer J, 54(4): 570-588.

Gupta M, Gao J, Aggarwal C, Han J.2013. Outlier detection for temporal data : A survey. IEEE TKDE, 26(9): 2250-2267.

Hodge VJ,Austin J. 2004. A survey of outlier detection methodologies. Artif Intell Rev, 22: 85-126.

Hubert M, Van Der Veeken S. 2008. Outlier detection for skewed data. J Chemomet, 22(3-4): 235-246.

Liu H, Shah S, Jiang W. 2004. On-line outlier detection and data cleaning. CCEND, 28(9): 1635-1647.

Maesschalck RD, Jouan-Rimbaud D, Massart DL. 2000. The Mahalanobis distance. Chemomet Intel Lab Syst, 50: 1-18.

Pei Y, Zaïane O. 2006. A synthetic data generator for clustering and outlier analysis. Department of Computing science, University of Alberta.URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.5133&rep=rep1&type=pdf (accesess date: 10.09.2018).

Rousseeuw PJ, Hubert M. 2011. Robust statistics for outlier detection. WIREs Data Mining Knowl Discov, 1(1): 73-79.

Singh, K, Upadhyaya S.2012. Outlier detection: applications and techniques. IJCSI, 9(1): 307-323.

Teng M. 2010. Anomaly detection on time series. 2010. IEEE International Conference on Progress in Informatics and Computing, 1:603-608.

Ting JA, D'Souza A, Schaal S. 2007a. Automatic outlier detection: A Bayesian approach. IEEE International Conference on Robotics and Automation. 2489-2494.

Ting JA, Theodorou E, Schaal S. 2007b. A Kalman filter for robust outlier detection. IEEE International Conference on Intelligent Robots and Systems, 1514-1519.

Url1: http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html (access date: 09.10.2018).

Url2: http://rstat.web.tr (access date: 09.10.2018)

Url3: https://onlinecourses.science.psu.edu/stat414/node/154/ (access date: 09.10.2018).

Xiang S, Nie F, Zhang C. 2008. Learning a Mahalanobis distance metric for data clustering and classification. Pattern Recog, 41(12): 3600-3612.