International Journal of Probability and Statistics 2018, 7(1): 14-18
DOI: 10.5923/j.ijps.20180701.02

Detection of Outliers in Growth Curve Models: Using Robust Estimators

O. Ufuk Ekiz

Department of Statistics, Gazi University, Ankara, Turkey

Abstract Outliers cause problems in statistical inferences for statistical analysis as well as growth curve model (GCM)s. Hence, robust estimators could be used to construct more significant inferences and this would make it possible to detect outliers more accurately. In this study, the method of least median square (LMS) is adopted to GCM. Then LMS, M, and ML (maximum likelihood) estimators are applied to real data applications and the reasons for the differences in the results are discussed.

Keywords Growth curve model, Outlier, Robust

1. Introduction

Growth curve model (*GCM*) is a generalized multivariate variance model and is defined by Potthof and Roy [1] so as to model longidutional data. This model is studied by several authors in literature as well [2-4]. Applications of this model to especially economic, social, and medicine sciences would give opportunities to investigate the mean growth in a population over a short period of time. Hence, making short term predictions become feasible by employing this model.

Let X and Z be the well-known design matrixes with ranks $m<p$ *and* $r<n$, respectively. p is the number of time points observed on each of n cases. *GCM* is given by

$$Y_{p\times n} = X_{p\times m} B_{m\times r} Z_{r\times n} + \varepsilon_{p\times n} \quad (1)$$

where $Y_{p\times n}$ is the observation matrix and B is the parameter matrix. Moreover, $\varepsilon_{p\times n}$ denotes the error matrix where the columns are p-variate normally distributed independent variables with mean 0 and unknown covariance matrix $\Sigma > 0$, [5]. Hence, $Y \sim N_{p,n}\left(XBZ, \Sigma, I_n\right)$ and I is the identity matrix. However, existing outliers in the data would impact on statistical inferences as they do in statistical analysis. Outlier is an observation that deviates from the rest of the data [6]. To get rid of the negative impacts of these outlying points there are two commonly addressed approaches. First group of methods are the so-called statistical diagnostics [7, 8]. The main purpose of these methods is based on observing the variation that an observation (or a group of observations) do have on the measure so as to point out as effective. However, for the sake of reliability of these approaches masking and swamping problems should not be ignored. The methods, through which the outliers are detected by means of robust estimators conduct the second group approaches [9]. In recent years, these are particularly preferred in statistical analysis. Since they are not likely to account the outliers (or attain minimum weights to them) in calculations of robust estimators, measures used for detection of outliers would not (or minimum) be affected.

Even though studies on the determination of points outlying of the bulk began long ago, it is only after 1990 that it has started to improve [10-14]. The purpose of this paper is to identify outliers in *GCM*s by using robust estimators least median squares (*LMS*) and *M*. Section 2 emphasis on the *M* estimator and the adaptation of *LMS* to *GCM* as well. In Section 3, we explained how to determine outliers by means of residuals. Finally, two real-life applications including outliers are considered to confirm differences on identifying them by means of residuals based on robust and non-robust estimators.

2. Parameter Estimations in Growth Curve Model

The ordinary least square (*OLS*) estimator of parameter B in equation (1) is

$$\hat{B}_{OLS} = (X'X)^{-1}X'YZ'(ZZ')^{-1}. \quad (2)$$

By using $\hat{B}_{OLS}$, the estimator of parameter Σ which is denoted as $\hat{\Sigma}_{OLS}$, [5], is calculated from

* Corresponding author:
ufukekiz@gazi.edu.tr (O. Ufuk Ekiz)
Published online at http://journal.sapub.org/ijps

$$\hat{\Sigma}_{OLS} = \frac{1}{n}\left(Y - X\hat{B}Z\right)\left(Y - X\hat{B}Z\right)'. \quad (3)$$

Suppose the covariance matrix Σ is of Rao's simple covariance structure (SCR), i.e., $\Sigma = X\Gamma X' + Q\Theta Q$, where both $\Gamma : m \times m$ and $\Theta : (n-m)\times(n-m)$ are unknown positive definite matrices, and $Q \in \varphi$. φ is the orthogonal matrix space of X defined by

$$\varphi = \{Q \,|\, Q : p \times (p-m),\ \text{rank}(Q) = p-m,\ X'Q = 0\}, \quad (4)$$

[15]. In this case maximum likelihood (*ML*) estimators of parameters B, Γ, Θ and Σ are

$$\hat{B}_{ML} = (X'X)^{-1}XYZ'(ZZ')^{-1}, \quad (5)$$

$$\Gamma_{ML} = \tfrac{1}{n}(X'X)^{-1}X'SX'(X'X)^{-1}, \quad (6)$$

$$\hat{\Theta}_{ML} = \tfrac{1}{n}(Q'Q)^{-1}QYY'Q(Q'Q)^{-1}, \quad (7)$$

$$\hat{\Sigma}_{ML} = X\hat{\Gamma}X' + Q\hat{\Theta}Q', \quad (8)$$

respectively. Here, $S \equiv Y(I_n - P_{Z'})Y'$ and $P_{Z'} = Z(Z'Z)^{-1}Z'$, [12, 16].

Let us now describe the weighted least square (*WLS*) estimator

$$\hat{B}_{WLS} = (X'WX)^{-1}X'WY\Sigma^{-1}Z'(Z\Sigma^{-1}Z')^{-1} \quad (9)$$

which is based on minimizing $\sum_{i=1}^{n} w_i e_i^2$. Here, e_i denotes the residual of the ith observation and

$$e_i^2(B,\Sigma) = \left(y_i - XBZ_i\right)'\Sigma^{-1}\left(y_i - XBZ_i\right). \quad (10)$$

Hence, the covariance matrix weighted estimator is computed from

$$\hat{\Sigma}_{WLS} = \frac{Y'HY}{tr(H)} \quad (11)$$

where $H = W - WZ(Z'WZ)^{-1}Z'W$ [17]. W is a diagonal matrix that consists of weights attained for each observation and "*tr*" denotes the trace of the corresponding matrix.

Define $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ and t as a value that ranges from 1 to $C(n,h)$. $C(n,h)$ denotes the number of h-combinations from a given set of n elements. The *LMS* estimators $\hat{B}_{LMS}$ and $\hat{\Sigma}_{LMS}$ are obtained from equations (9) and (11), respectively, by minimizing the objective function

$$\min_{\hat{\beta}_t}\left(\text{median}(e_i^2(\hat{B}_j, \hat{\Sigma}_j))\right), \quad (12)$$

where $i = 1,...,n$ and $j = 1,...,C(n,h)$, [9]. Moreover, the weight matrix W_t, which will be used for the underlined equations, is determined so that its ith diagonal element $w_i(e_i^2)$ is

$$w_i(e_i^2) = \begin{cases} 1, & \text{if } i\text{th observation exists in the } t\text{th combination} \\ 0, & \text{otherwise} \end{cases}.$$

When $\hat{B}_{LMS}$. is used as the initial point and the value $\hat{B}_k$ is obtained from the kth iteration of $\min_{\hat{\beta}_k}\sum_{i=1}^{n}\rho(e_i^2)$, $\hat{B}_k$ will be the M estimator, $\hat{B}_M$. ρ function has a minimum at "0" for all values of e_i^2. Here, Tukey's ρ function (bi-square), [18, 19], is used to compute the M estimator. Hence, the ith diagonal element $w_i\left(e_i^2\right)$ of the weight function W_k that should be used both in equation (9) and (11) would be

$$w_i\left(e_i^2\right) = \begin{cases} \left(1 - \left(e_i^2/c\right)^2\right)^2, & 0 \le e_i^2 \le c \\ 0, & e_i^2 > c \end{cases}. \quad (13)$$

The constant c in equation (13) is assumed that it will hold $E_{\chi^2_{(p)}}\left[\rho(e_i^2)\right] \Big/ \rho(c) = h/n$, where $E_{\chi^2_{(p)}}[.]$ is the expected value obtained from chi-squared distribution with p degrees of freedom. Here, h/n value is preferred so as to have the same breakdown point as the LMS estimator.

3. Detecting Outliers in Growth Curve Model

It is known that the sum of squared residuals fit to chi-square with p degrees of freedom when the data does not contain outliers [20]. Hence,

$$e_i^2(B,\Sigma) = \left(y_i - XBZ_i\right)'\Sigma^{-1}\left(y_i - XBZ_i\right) \sim \chi^2_{(p)} \quad (14)$$

and if

$$\hat{e}_i^2(\hat{B},\hat{\Sigma}) = \left(y_i - X\hat{B}Z_i\right)'\hat{\Sigma}^{-1}\left(y_i - X\hat{B}Z_i\right) \quad (15)$$

is greater than the critical value $\chi^2_{(p),1-\alpha}$ the ith observation would be identified as an outlier. α denotes the significance level. $\hat{B}$ and $\hat{\Sigma}$ are the estimators of parameters B and Σ, respectively, and are calculated from an any estimation method. However, if estimators are affected by outliers they would cause for determining wrong observations as outliers. Therefore, robust estimators that are less likely to be affected by outlying points should be preferred and this would assure more reliable results.

4. Example Applications

Understanding the importance of selection of estimators in parameter estimates and detection of outliers in *GCM*s, two real data sets are examined.

4.1. Dental Data

This data set, was first considered by Potthoff and Roy, [1], and later analyzed by several authors (see [21-24]). Dental measurements were made on 11 girls and 16 boys at ages 8, 10, 12, and 14 years. Each measurement is the distance in millimetres from the centre of the pituitary to the pterygomaxillary fissure. Sequence number to each measurement is assigned. Then, the *ML*, *LMS*, and *M* estimators of this data set are computed for boys and girls separately, and are given in Table 1 and Table 2. The sequence numbers of detected outliers are summarized in the last column of these tables as well. As it shown in Table 1 by means of *ML* estimator, observation numbered 21 is an outlier. On the other hand, outliers based on *LMS* and *M* estimators are observations numbered as 12, 15, 20, and 24. *ML* estimators are non-robust, so they are greatly affected by outliers in data. Furthermore, by examining the girl's data there are no outliers (see Table 2).

4.2. Mouse Data

This data set is reported by Izenman and Williams, [25], and is analyzed by Roo and Lee as well [22-24]. It consists of weights of 13 male mice measured at intervals of 3 days over the 21 days from birth to weaning. As done to the dental data, sequence number to each measurement is assigned. The *ML*, *LMS*, and *M* estimators of parameters B and Σ for this data set are given in Table 3. The sequence numbers of the detected outliers obtained by using these estimators are also determined. From the table, it is noticeable that robust estimators' performance on detecting outliers differs from *ML* estimators. *ML* estimators have detected only the second observation as an outlier while robust estimators have detected 4th, 11th, and 12th observations as outliers.

Table 1. Parameter Estimations and Detected Outliers from Boys in Dental Data

Estimator	$\hat{B}$	$\hat{\Sigma}$				Outlier's sequence number
ML	15.8283 0.8340	4.7396	3.0635	3.2343	1.5583	21
		3.0635	4.9615	1.3188	3.2167	
		3.2343	1.3188	4.9439	3.0284	
		1.5583	3.2167	3.0284	4.6869	
LMS	14.8412 0.8961	7.4472	2.5500	4.9500	1.7472	15 20 24
		2.5500	5.0111	2.1278	3.4778	
		4.9500	2.1278	9.4000	3.2167	
		1.7472	3.4778	3.2167	4.0028	
M	18.3821 0.6590	4.2235	2.6450	4.8529	3.3202	12 20 24
		2.6450	3.7069	4.0749	3.1501	
		4.8529	4.0749	6.4112	4.7842	
		3.3202	3.1501	4.7842	4.8970	

Table 2. Parameter Estimations and Detected Outliers from Girls in Dental Data

Estimator	$\hat{B}$	$\hat{\Sigma}$				Outlier's sequence number
ML	17.4220 0.4823	3.4771	3.4477	3.6405	3.6111	-
		3.4477	3.9492	3.7841	4.2855	
		3.6405	3.7841	4.5942	4.7378	
		3.6111	4.2855	4.7378	5.4123	
LMS	17.6716 0.4921	5.0111	3.6778	4.7778	4.8167	-
		3.6778	3.4139	4.0417	4.2333	
		4.7778	4.0417	5.9028	5.8611	
		4.8167	4.2333	5.8611	6.4556	
M	17.5299 0.4708	4.0428	3.0500	3.8061	3.8727	-
		3.0500	3.3058	3.5543	3.6296	
		3.8061	3.5543	4.8836	4.8152	
		3.8727	3.6296	4.8152	5.3432	

Table 3. Parameter Estimations and Detected Outliers from Mouse Data

Estimator	$\hat{B}$	$\hat{\Sigma}$							Outlier's sequence number
ML	0.0222 0.2084 -0.0108	0.0009	0.0001	-0.0003	-0.0006	-0.0008	-0.0008	-0.0007	2
		0.0001	0.0017	0.0026	0.0035	0.0040	0.0042	0.0041	
		-0.0003	0.0026	0.0051	0.0067	0.0078	0.0083	0.0081	
		-0.0006	0.0035	0.0067	0.0092	0.0108	0.0115	0.0114	
		-0.0008	0.0040	0.0078	0.0108	0.0128	0.0139	0.0140	
		-0.0008	0.0042	0.0083	0.0115	0.0139	0.0153	0.0159	
		-0.0007	0.0041	0.0081	0.0114	0.0140	0.0159	0.0171	
LMS	0.0196 0.2115 -0.0110	0.0007	0.0010	0.0008	0.0013	0.0009	0.0015	0.0013	2 11 12
		0.0010	0.0015	0.0017	0.0028	0.0027	0.0035	0.0031	
		0.0008	0.0017	0.0030	0.0048	0.0059	0.0066	0.0058	
		0.0013	0.0028	0.0048	0.0105	0.0131	0.0144	0.0119	
		0.0009	0.0027	0.0059	0.0131	0.0198	0.0204	0.0166	
		0.0015	0.0035	0.0066	0.0144	0.0204	0.0226	0.0190	
		0.0013	0.0031	0.0058	0.0119	0.0166	0.0190	0.0179	
M	0.0190 0.2102 -0.0109	0.0007	0.0010	0.0007	0.0012	0.0007	0.0012	0.0013	2 4 11
		0.0010	0.0015	0.0016	0.0026	0.0024	0.0030	0.0028	
		0.0007	0.0016	0.0028	0.0047	0.0053	0.0058	0.0053	
		0.0012	0.0026	0.0047	0.0100	0.0120	0.0125	0.0104	
		0.0007	0.0024	0.0053	0.0120	0.0177	0.0174	0.0142	
		0.0012	0.0030	0.0058	0.0125	0.0174	0.0185	0.0159	
		0.0013	0.0028	0.0053	0.0104	0.0142	0.0159	0.0154	

5. Conclusions

ML estimators in *GCM*s both using for comparison of groups and making short term predictions could be badly affected by outliers in data. This affection can lead to bad estimates of parameters for the assumed distribution of the data. Moreover, utilizing non-robust *ML* in hypothesis tests to determination of outliers will give misleading results as well. When the investigation of outliers is based on robust test statistics, it is well-known that the obtained results could reflect the reality much better. In Section 4, two data sets that are used in literature for different purposes are analyzed. Accordingly, the results differ to a great extend when using robust or non-robust estimators. Moreover, the variance of robust *M* estimator is smaller than *LMS*'s. Hence, obtaining results with robust *M* estimator will be more convenient.

REFERENCES

[1] R. F. Potthoff and S. N. Roy, "A generalized multivariate analysis of variance model useful especially for growth curve problems", Biometrika, vol 51, pp. 313-326, 1964.

[2] C. R. Rao, "Least squares theory using an estimated dispersion matrix and its application to measurement of signals", In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Univ. of California Press, pp. 355-372, 1967.

[3] C. G. Khatri, "A note on a manova model applied to problems in growth curve", Annals of the Institute of Statistical Mathematics, vol 18, pp. 75-86, 1966.

[4] D. Rosen, "Maximum likelihood estimators in multivariate linear normal models", Journal of Multivariate Analysis, vol 31:2, pp. 187-200, 1989.

[5] J. X. Pan and K. T. Fang, Growth Curve Models and Statistical Diagnostics, Springer Science & Business Media, New York, Springer-Verlag, 2002.

[6] V. Barnett and T. Lewis, Outliers in Statistical Data, New York, Wiley & Sons, 1984.

[7] R. D. Cook and S. Weisberg, Residuals and Influence in Regression, New York, Chapman and Hall, 1982.

[8] O. U. Ekiz and M. Ekiz, "The role of outliers in growth curve models: a case study of city-based fertility rates in Turkey", International Journal of Statistics and Applications, vol 7:3, pp. 178-185, 2017.

[9] P. J. Rousseeuw and A. M. Leroy, Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987.

[10] E. P. Liski, "Detecting influential measurements in a growth curves model", Biometrics, vol. 47:2, pp. 659-668, 1991.

[11] J. X. Pan and K. T. Fang, "Multiple outlier detection in growth curve model with unstructured covariance matrix", Annals of the Institute of Statistical Mathematics, vol 47, pp. 137-153, 1995.

[12] J. X. Pan and K. T. Fang, "Detecting influential observations

in growth curve model with unstructured covariance", Computational Statistics and Data Analysis, vol 22, pp. 71-87, 1996.

[13] J. X. Pan, K. T. Fang, and D. von Rosen, "Local influence assessment in the growth curve model with unstructured covariance", Journal of Statistical Inference and Planning, vol 62, pp. 263-278, 1997.

[14] X. Tong and Z. Zhang, "Outlying observation diagnostics in growth curve modeling", Multivariate Behavioral Research, vol 52:6, pp. 768-788, 2017.

[15] C. R. Rao, "Least squares theory using an estimated dispersion matrix and its application to measurement of signals," In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Univ. of California Press, pp. 355-372, 1967.

[16] J. X. Pan, "Discordant outlier detection in the growth curve model with Rao's simple covariance structure", Statistics & Probability Letters, vol. 69, pp. 135-142, 2004.

[17] J. F. Pendergast and J. D. Broffitt, "Robust estimation in growth curve models", Communications in Statistics - Theory and Methods, vol 14: 8, pp. 1919-1939, 1985.

[18] R. A. Maronna, D. R. Martin, and V. J. Yohai, Robust Statistics Theory and Methods, New York, John Wiley & Sons, 2006.

[19] M. Ekiz and O. U. Ekiz, "Outlier detection with Mahalanobis square distance: incorporating small sample correction factor", Journal of Applied Statistics, vol 44:13, pp. 2444-2457, 2017.

[20] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, New York, John Wiley & Sons, 2003.

[21] J. C.-S. Lee and S. Geisser, "Applications of Growth Curve Prediction", Sankhyā: The Indian Journal of Statistics, Series A, vol. 37:2, pp. 239-256, 1975.

[22] C. R. Roo. "Prediction of future observations in growth curve models", Statistical Science, vol 2, pp. 434-471, 1987.

[23] J. C. Lee, "Prediction and estimation of growth curve with special covariance structure", Journal of the American Statistical Association, vol 83, pp. 432-440, 1988.

[24] J. C. Lee, "Tests and model selection for the general growth curve model", Biometrics, vol 47, pp. 147-159, 1991.

[25] A. J. Izenman and J. S. Williams, "A class of linear spectral models and analyses for the study of longitudinal data", Biometrics, vol 45, pp. 831-849, 1989.