Outlier Detection Methods in Multivariate Regression Models

BY JIAQIONG XU

Center for American Indian Health Research,

College of Public Health,

University of Oklahoma Health Sciences Center,

P.O.Box 26901, Rm CHB100, Oklahoma City, OK 73190 U.S.A.

susan-xu@ouhsc.edu

BOVAS ABRAHAM

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON N2L 3G1 Canada.

babraham@uwaterloo.ca

AND **STEFAN H. STEINER**

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON N2L 3G1 Canada.

shsteine@uwaterloo.ca

Corresponding author: Bovas Abraham , Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1 Canada. E-mail: babraham@uwaterloo.ca

Abstract. Outlier detection statistics based on two models, the case-deletion model and the mean-shift model, are developed in the context of a multivariate linear regression model. These are generalizations of the univariate Cook's distance and other diagnostic statistics. Approximate distributions of the proposed statistics are also obtained to get suitable cutoff points for significance tests. In addition, a simulation study has been conducted to examine the performance of these two approximate distributions. The methods are applied to a set of data to illustrate the multiple outlier detection procedure in multivariate linear regression models.

key words: likelihood displacement; likelihood ratio; multivariate regression; outlier detection.

1 Introduction

Most data analysts have come across data which seem to contain some deviant or "outlying" observations (or outliers). This may be the result of unusual and non-repetitive events such as system changes, strikes, special problems, etc. Such data can wreak havoc with the estimation of statistical models and can have undue influence on the conclusions from a statistical analysis.

Consider a univariate linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is a $n \times 1$ vector of observations, X is a $n \times (p+1)$ full rank known matrix, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of unknown parameters and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of normally distributed errors such that $E(\boldsymbol{\epsilon}) = 0$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 I$. Outliers can affect the parameter estimates in this model and many papers have been written on the detection of a single outlier in this context (see, for example, Srikantan (1961) and Barnett and Lewis (1994)) using residuals from least squares fit. Another approach called 'case deletion' studies the effect of deleting an observation on the parameter estimates. Cook (1977) defines a measure of distance between two maximum likelihood estimates, where one is computed with all the observations and the other without a specific observation. Hossain and Naik (1989) and Naik (2003) extend the results of deleting a single observation in univariate regression models to multivariate regression models. Srivastava and von Rosen (1998) develop a formal test for detecting a single outlier in a multivariate linear regression model.

Many authors have considered the problem of detection of multiple outliers in univariate linear regression models. Early work due to Prescott (1975) and Tietjen et al. (1973) focus on repeated application of single case detection methods

to detect multiple outliers. However, it is well known that such methods suffer from the problems of masking and swamping, where the effect of one outlier masks the effect of other outliers (see, for example, Barnett and Lewis, 1994). Hadi and Simonoff (1993) proposed procedures and tests for detection of multiple outliers in univariate linear models. Wei and Fung (1999) consider deleting multiple observations in univariate general weighted regression models. Barrett and Ling (1992) propose general classes of influence measures for multivariate regression based on an analogous form of Cook's influence measure for univariate case. Díaz-García and González-Farías (2004) discuss a generalization of the Cook's distance in multivariate regression under elliptical distributions.

We present multivariate outlier detection methods based on the case-deletion model and the mean shift model in Section 2. The performance of the approximate distributions of the proposed statistics are examined by a simulation study in Section 3, and suggestions for implementation are given in Section 4. The proposed procedure for detecting outliers is illustrated through an example in Section 5. Finally, some concluding remarks are given in Section 6.

2 Multivariate outlier detection methods

Suppose we have m responses and p predictors. We consider the multivariate linear regression model:

$$H_0: \quad Y_{n\times m} = (\mathbf{J}\ X_1)B + E = XB + E, \tag{1}$$

where Y is a $n \times m$ response matrix, $\mathbf{J}$ is a $n \times 1$ unit vector, X_1 is a $n \times p$ design matrix, B is a $(p+1) \times m$ coefficient matrix, and E is a $n \times m$ random error matrix. We assume that rows of E are independent, normally distributed, each with mean vector zero, and $m \times m$ covariance matrix Σ, that is, $Vec(E) \sim MVN(\mathbf{0}, I_n \otimes \Sigma)$, where $Vec(E)$ is a column vector where the first m elements are the entries from the first row of E, the second m elements are the entries from the second row of E, etc., MVN stands for multivariate normal distribution and $\otimes$ stands for the Kronecker product. From now on we simply denote $E \sim MVN(\mathbf{0}, I_n \otimes \Sigma)$.

The corresponding likelihood function is given by

$$\mathbf{L}(B, \Sigma) = \frac{1}{(2\pi)^{mn/2}} \frac{1}{|\Sigma|^{n/2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}(Y-XB)'(Y-XB)\}}. \tag{2}$$

As in the univariate linear regression model case, if rank $(X) = p+1$, then the maximum likelihood estimate (MLE) of B is given by

$$\hat{B} = (X'X)^{-1}X'Y, \tag{3}$$

and that of Σ is given by

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})'(Y - X\hat{B}). \tag{4}$$

2.1 Likelihood displacement (LD) statistic

Let $\mathcal{A}_k = \{i_1, \cdots, i_k\}$ index a subset of an arbitrary number of k observation, where $i_j \in \{1, \cdots, n\}, j = 1, \cdots, k$ and k is very small compared with the number of observations n. Let $l(\theta)$ be the log likelihood function, where θ are the model

parameters. When considering the effect of k observations on the parameter estimates of a multivariate regression model, we define likelihood displacement $LD_{\mathcal{A}_k}$ for measuring the difference between $\hat{\theta}$ and $\hat{\theta}_{[\mathcal{A}_k]}$ as

$$LD_{\mathcal{A}_k}(\theta) = 2\{l(\hat{\theta}) - l(\hat{\theta}_{[\mathcal{A}_k]})\},$$

where $\hat{\theta}$ denote the MLE of θ with all observations and $\hat{\theta}_{[\mathcal{A}_k]}$ that without the k observations in the set $\mathcal{A}_k$. This definition is directly analogous to the likelihood displacement definition used by Cook & Weisberg (1982) to consider the effect of a single observation on parameter estimates in the univariate case.

When a subset θ_1 of θ is of special interest, the likelihood displacement can be modified as

$$LD_{\mathcal{A}_k}(\theta_1|\theta_2) = 2\{l(\hat{\theta}) - l(\hat{\theta}_{1[\mathcal{A}_k]}, \hat{\theta}_2(\hat{\theta}_{1[\mathcal{A}_k]}))\}, \tag{5}$$

where $l(\hat{\theta}_{1[\mathcal{A}_k]}, \hat{\theta}_2(\hat{\theta}_{1[\mathcal{A}_k]})) = \max_{\theta_2} l(\hat{\theta}_{1[\mathcal{A}_k]}, \theta_2)$ denotes the log likelihood maximized over the parameter space for θ_2 with $\theta_1 = \hat{\theta}_{1[\mathcal{A}_k]}$, which is the MLE of θ_1 when k observations are deleted.

Now we consider deleting k observations given in the set $\mathcal{A}_k$. Let $[\mathcal{A}_k]$ denote the index of the remaining observations and let $Y_{[\mathcal{A}_k]}, X_{[\mathcal{A}_k]}$ denote the response matrix and design matrix when all the row numbers corresponding to the elements of $\mathcal{A}_k$ are deleted. Then, based on the remaining data set, the MLE of B is given by (see Section 7)

$$\hat{B}_{[\mathcal{A}_k]} = \hat{B} - (X'X)^{-1}X'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k}, \tag{6}$$

where $Q_{\mathcal{A}_k} = X_{\mathcal{A}_k}(X'X)^{-1}X'_{\mathcal{A}_k}$, $\hat{E}_{\mathcal{A}_k} = Y_{\mathcal{A}_k} - X_{\mathcal{A}_k}\hat{B}$ are residuals, and $Y_{\mathcal{A}_k}, X_{\mathcal{A}_k}$ are the response matrix and design matrix corresponding to the index $\mathcal{A}_k$.

Similarly, the MLE of Σ is given by

$$\hat{\Sigma}_{[\mathcal{A}_k]} = \frac{n}{n-k}\hat{\Sigma} - \frac{1}{n-k}\hat{E}'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k}. \tag{7}$$

Thus, $\hat{B}_{[\mathcal{A}_k]}$ and $\hat{\Sigma}_{[\mathcal{A}_k]}$ are the regression parameter estimates of B and Σ without the contributions from the k observations in the set $\mathcal{A}_k$, and $\hat{E}_{\mathcal{A}_k}$ are the residuals corresponding to $Y_{\mathcal{A}_k}$.

For the model H_0 given in (1), let $\theta_1 = B$, $\theta_2 = \Sigma$ in (5). Then the likelihood displacement for B given Σ is

$$LD_{\mathcal{A}_k}(B|\Sigma) = 2\{l(\hat{B}, \hat{\Sigma}) - l(\hat{B}_{[\mathcal{A}_k]}, \hat{\Sigma}(\hat{B}_{[\mathcal{A}_k]}))\},$$

where $\hat{\Sigma}(\hat{B}_{[\mathcal{A}_k]})$ is the MLE of Σ when B is estimated by $\hat{B}_{[\mathcal{A}_k]}$ given in (6). Substituting $\hat{B}_{[\mathcal{A}_k]}$ for B in equation (2), the MLE of Σ is given by

$$\begin{aligned}
\hat{\Sigma}(\hat{B}_{[\mathcal{A}_k]}) &= \frac{1}{n}\{\hat{E}'\hat{E} + (\hat{B} - \hat{B}_{[\mathcal{A}_k]})'X'X(\hat{B} - \hat{B}_{[\mathcal{A}_k]})\} \\
&= \hat{\Sigma} + \frac{1}{n}\hat{E}'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}Q_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k}.
\end{aligned}$$

The likelihood displacement for B given Σ is therefore given by

$$\begin{aligned}
LD_{\mathcal{A}_k} &\equiv LD_{\mathcal{A}_k}(B|\Sigma) = 2\{l(\hat{B}, \hat{\Sigma}) - l(\hat{B}_{[\mathcal{A}_k]}, \hat{\Sigma}(\hat{B}_{[\mathcal{A}_k]}))\} \\
&= n\log\{\frac{|\hat{\Sigma}(\hat{B}_{[\mathcal{A}_k]})|}{|\hat{\Sigma}|}\} = n\log\{\frac{|n\hat{\Sigma} + \hat{E}'_{\mathcal{A}_k}C_{\mathcal{A}_k}\hat{E}_{\mathcal{A}_k}|}{|n\hat{\Sigma}|}\},
\end{aligned} \tag{8}$$

where $C_{\mathcal{A}_k} = (I - Q_{\mathcal{A}_k})^{-1}Q_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}$.

When we consider a single response ($m = 1$) and delete any single observation (i.e. $k = 1, \mathcal{A}_k = \{i\}$), then $B = \boldsymbol{\beta}, \Sigma = \sigma^2$ and the likelihood displacement statistic in (8) can be written as

$$LD_i(\boldsymbol{\beta}|\sigma^2) = n\log(1 + \frac{p+1}{n}D_i),$$

where $D_i = h_{ii}\hat{e}_i^2/\{(p+1)(1-h_{ii})^2\hat{\sigma}^2\}$ is the standard "Cook's distance" (Cook, 1977). Here h_{ii} is the ith diagonal element of $H = X(X'X)^{-1}X'$ and is usually referred to as the leverage, and $\hat{e}_i$ is the residual corresponding to the ith observation.

In univariate settings, it is usually better to omit the potential outlier(s) to estimate the error variance. This is also true in the multivariate settings where the error variance can be estimated with all the observations or without the possible outlying observations (see (7)). Therefore $LD_{\mathcal{A}_k}$ in (8) is a generalized Cook's statistic for assessing the influence on parameter estimates when deleting an arbitrary set of k observations in the multivariate regression model. The sets of k observations that are influential on $\hat{B}$ are indicated by large values of $LD_{\mathcal{A}_k}$ as given by (8).

When using the $LD_{\mathcal{A}_k}$ statistic to identify outliers and influential observations, we need to find the corresponding critical values which requires the distribution of the $LD_{\mathcal{A}_k}$ statistic under the null model (1). However, it is very difficult to find its exact distribution. It can be shown (Xu, J. (2003). Multivariate outlier detection and process monitoring. Unpublished Ph.D. thesis, the Department of Statistics & Actuarial Science, University of Waterloo, Ontario, Canada. We do not include the proof here to save space.) that $LD_{\mathcal{A}_k}$ converges in distribution to

$$LD_A = \sum_{i=1}^{k} \lambda_i Z_i^2, \tag{9}$$

where $\lambda_i, i = 1, \cdots, k$, are the eigenvalues of $C_{\mathcal{A}_k}$, and Z_i^2 are independent and identically distributed with chi-square distribution having m degrees of

freedom for $k \geq 2$. When $k = 1$, $LD_{\mathcal{A}_k}$ converges in distribution to $\lambda\chi^2_m$, where $\lambda = h_{ii}/(1 - h_{ii})^2$.

We consider two methods to obtain critical values of $LD_{\mathcal{A}_k}$. One method is to extend Field's (1993) method, which obtains tail areas of linear combinations of chi-square random variables with 1 degree of freedom, using a saddlepoint approximation. This method gives more accurate critical values (from our simulation below), but requires solving a non-linear equation. Another useful method in practice is to extend the method of Jensen and Solomon (1972) to approximate a modified version of the likelihood displacement statistic using a $N(0, 1)$ distribution for $k \geq 2$. Using this approximate distribution, the corresponding critical value is

$$LD_{\mathcal{A}_k,\alpha} = \delta_1\{\frac{c_\alpha\sqrt{2\delta_2 f_0^2}}{\delta_1} + \frac{\delta_2 f_0(f_0 - 1)}{\delta_1^2} + 1\}^{1/f_0}$$

where

$$\delta_1 = m\textstyle\sum_1^k \lambda_i, \qquad \delta_2 = m\sum_1^k \lambda_i^2,$$

$$\delta_3 = m\textstyle\sum_1^k \lambda_i^3, \qquad f_0 = 1 - (2\delta_1\delta_3/3\delta_2^2),$$

$$c_\alpha = \begin{cases} z_{1-\alpha} & , \text{ if } f_0 \geq 0 \\ z_\alpha & , \text{ if } f_0 < 0 \end{cases}$$

and λ_i s are the eigenvalues of $C_{\mathcal{A}_k}$ and z_α is the $100\alpha\%$ percentile of the standard normal distribution. We will compare these two approximations in a later section.

2.2 Multivariate leverage

When using Cook's distance to measure the influence of an observation in univariate linear regression models, the leverage h_{ii} measures how extreme the values of the explanatory variables are. Analogously, in multivariate linear regression models, we use the average diagonal element of $Q_{\mathcal{A}_k}$ (ADQ) to measure how extreme the k measurements for the explanatory variables are, that is,

$$ADQ_{\mathcal{A}_k} = \mathrm{tr}(Q_{\mathcal{A}_k})/k.$$

Large values of $ADQ_{\mathcal{A}_k}$ indicate unusual observations in explanatory variables. Belsley et al. (1980) recommend using $\bar{h} = 2(p+1)/n$ as a cutoff value for all h_{ii} in the univariate regression. We use the same value as a cutoff point for $ADQ_{\mathcal{A}_k}$.

2.3 Likelihood ratio (LR) statistic for a mean shift

Now we consider mean shifts in any k observations in the set $\mathcal{A}_k$. We consider the following mean shift model:

$$H_A: \quad Y = \begin{pmatrix} X & Z_{\mathcal{A}_k} \end{pmatrix} \begin{pmatrix} B \\ \Psi \end{pmatrix} + E = X^* B^* + E,$$

where $X^* = (X \;\; Z_{\mathcal{A}_k})$, $Z_{\mathcal{A}_k} = (z_{i_1} \;\cdots\; z_{i_k})$ and $z_j, j = i_1, \cdots, i_k$ denote the $n \times 1$ vector with 1 in row j and zero in all other rows, $B^* = (B \;\; \Psi)^T$, and Ψ is a $k \times m$ shift coefficient matrix corresponding to the observations in the set $\mathcal{A}_k$.

Then the MLE of B^* under model H_A is given by

$$\hat{B}^*_{H_A} = \begin{pmatrix} \hat{B} - (X'X)^{-1}X'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k} \\ (I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k} \end{pmatrix},$$

where $\hat{B}$ is given by (3).

The MLE of Σ under model H_A is given by $\hat{\Sigma}^*_{H_A} = \hat{\Sigma} - \hat{E}'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k}/n$, where $\hat{\Sigma}$ is given by (4).

The possibility that the k observations in the set $\mathcal{A}_k$ are outliers can be assessed by testing the hypothesis that $\Psi = \mathbf{0}$ in the mean shift model H_A. When the hypothesis is true, the model reduces to H_0, as given by (1).

Applying the likelihood ratio test, we get the likelihood ratio as

$$\Lambda = \frac{\max_{H_0} L(B, \Sigma)}{\max_{H_A} L(B, \Sigma)} = \frac{L(\hat{B}, \hat{\Sigma})}{L(\hat{B}^*_{H_A}, \hat{\Sigma}^*_{H_A})} = (\frac{|\hat{\Sigma}^*_{H_A}|}{|\hat{\Sigma}|})^{n/2},$$

i.e.,

$$\Lambda^{2/n} = \frac{|\hat{\Sigma}^*_{H_A}|}{|\hat{\Sigma}|} = \frac{|n\hat{\Sigma}^*_{H_A}|}{|n\hat{\Sigma}^*_{H_A} + n(\hat{\Sigma} - \hat{\Sigma}^*_{H_A})|}.$$

Under $H_0 : \Psi = \mathbf{0}, n\hat{\Sigma}^*_{H_A}$ has a Wishart distribution $W_m(n - p - k - 1, \Sigma)$ and is independent of $n(\hat{\Sigma} - \hat{\Sigma}^*_{H_A})$, which, in turn, has a Wishart distribution $W_m(k, \Sigma)$ (see, for example, Seber, 1984, p.409). The likelihood ratio test is equivalent to rejecting H_0 for large values of

$$-2\log\Lambda = -n\log(\frac{|\hat{\Sigma}^*_{H_A}|}{|\hat{\Sigma}|}) = -n\log\{\frac{|n\hat{\Sigma}^*_{H_A}|}{|n\hat{\Sigma}^*_{H_A} + n(\hat{\Sigma} - \hat{\Sigma}^*_{H_A})|}\}.$$

For n large, and both $n - p$ and $n - m$ large, Box (1949) shows that the statistic -2log Λ approximates to a chi-square distribution with m degrees of freedom. As a result, with adjusted multiplying factors, the likelihood ratio

statistic (LR)

$$LR_{\mathcal{A}_k} = c\log(\frac{|\hat{\Sigma}^*_{H_A}|}{|\hat{\Sigma}|}) = c\log\{\frac{|n\hat{\Sigma} - \hat{E}'_{\mathcal{A}_k}(I - Q_{\mathcal{A}_k})^{-1}\hat{E}_{\mathcal{A}_k}|}{|n\hat{\Sigma}|}\} \qquad (10)$$

is closely approximated by a chi-square distribution with mk degrees of freedom (Johnson and Wichern, 1992), where $c = -\{n - p - k - 1 - (m - k + 1)/2\}$.

Large values of $LR_{\mathcal{A}_k}$ will indicate that the rejection of the hypothesis that there are no outliers.

When the number of the observations is small, alternatively, we can simulate a cutoff point for the $LR_{\mathcal{A}_k}$ statistic. We will discuss this later.

3 Performance of approximation and a simulation method

3.1 Approximations

In order to study the performance of the approximate distributions of the likelihood displacement statistic LD and the likelihood ratio statistic LR, we conduct a simulation study that uses $p = 5$ predictors, $n = 300$ observations and, as an example, the first k observations. First, we generate a $n \times (p + 1)$ design matrix X and a $(p+1) \times m$ coefficient matrix B. Let $X = (\mathbf{J}\ X_1)$, where $\mathbf{J}$ is a $n \times 1$ unit vector, and X_1 is a $n \times p$ matrix whose elements are generated from a uniform (0, 10) distribution, and the elements of B are generated from a uniform (-5, 5) distribution. For given m, k, desired significance level α and design matrix, we calculate the corresponding critical values for the LR and LD

statistics from their approximate distributions. Then we perform the following steps:

1. Generate an error term E according to multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. In our simulation, we consider the number of responses $m = 1, 2$ and 5. When $m = 1$, we set $\Sigma = 1$; when $m = 2$, we take $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$; and when $m = 5$, we use $\Sigma = \begin{pmatrix} 1 & 0.2 & 0.3 & 0.4 & 0.5 \\ 0.2 & 1 & 0.4 & 0.2 & 0.7 \\ 0.3 & 0.4 & 1 & 0.5 & 0.8 \\ 0.4 & 0.2 & 0.5 & 1 & 0.7 \\ 0.5 & 0.7 & 0.8 & 0.7 & 1 \end{pmatrix}$;
2. Calculate the corresponding responses based on the model $Y = XB + E$;
3. Calculate LD from (8) and LR from (10) for the first k observations;
4. Compare LD to the cutoff point from the saddlepoint and the normal approximations, and LR to the cutoff points determined by the χ^2_{mk} approximate distribution.
5. Repeat the above steps 5,000 times.

We consider the following cases:

Case 1 : m=1, k=1 **Case 2** : m=1, k=5 **Case 3** : m=2, k=2

Case 4 : m=2, k=5 **Case 5** : m=5, k=2 **Case 6** : m=5, k=5

Table 1 gives the proportions of times the LD values and LR values exceed their corresponding critical values for significance levels $\alpha = 0.1, 0.05$, and 0.01. From Table 1, we see that the empirical significance levels are very close to the desired α values in each case. For instance, when $m = 5, k = 2$, and $\alpha = 0.01$, the empirical significance levels for LD are 0.0094 (saddlepoint approximation) and 0.0092 (normal approximation) and the one for LR is 0.0102. The saddlepoint approximation is slightly better than the normal approximation for LD statistic in terms of determining the critical values. However, the difference is small, and it is more convenient to use the normal approximation because the critical value from the saddlepoint approximation requires the solution of a non-linear equation. Hence from now on the critical value of LD statistic is obtained using the normal approximation.

3.2 Critical values using simulation

When the sample size is large, we can use an asymptotic distribution (9) for the LD statistic and the χ^2_{mk} approximate distribution for the LR statistic. Alternatively, when the sample size is small, we can obtain critical values using simulation for each data set. The following procedure illustrates the simulation approach.

Because the statistics LD and LR do not depend on the parameters B and Σ, without loss of generality, we assume $B = \mathbf{0}$ and $\Sigma = I_m$. For given X:

1. Generate Y from $Y = E \sim MVN(\mathbf{0}, I_n \otimes I_m)$;

2. Estimate the parameters B and Σ by fitting the model Y=XB+E to the

data obtained in step 1;

3. Calculate the statistics $LD_{\mathcal{A}_k}$ and $LR_{\mathcal{A}_k}$.

We repeat the above steps M times and order $LD^{(1)}_{\mathcal{A}_k}, \cdots, LD^{(M)}_{\mathcal{A}_k}$ and $LR^{(1)}_{\mathcal{A}_k}, \cdots, LR^{(M)}_{\mathcal{A}_k}$ separately. Then, we take the upper α percentile values as the critical values for any set of k observations. In Section 5 we illustrate the implementation of this procedure in the context of an example.

4 Implementation

In order to detect multiple outliers in multivariate linear regression models, we can combine the use of the LD, LR and ADQ. We propose to calculate LD, LR statistics and ADQ for

1. each single observation $\mathcal{A} = i$;

2. each pair of observations $\mathcal{A} = \{i_1, i_2\}$, where $i_1 < i_2$, and i_1, $i_2 = 1, \cdots, n$;

3. each group of three observations $\mathcal{A} = \{i_1, i_2, i_3\}$, where $i_1 < i_2 < i_3$, and $i_j = 1, \cdots, n, j = 1, 2$, and 3;

4. Continue with 4, 5, $\cdots$ observations at a time. Stop either when the computational burden becomes too large (see comments later), or when no new observations are identified as outliers and the LD and LR values do not change much.

Then we compare the $LD_{\mathcal{A}}$ and $LR_{\mathcal{A}}$ to their critical values $LD_{\mathcal{A},\alpha}$ and $LR_{\mathcal{A},\alpha}$ respectively, where for example looking at one observation at a time α

can take 0.05 or 0.01. To accommodate multiple testing we suggest lowering the significance level for the tests for pairs, triples, etc. We use $\bar{h} = 2(p+1)/n$ as the ADQ cutoff value throughout.

In some situations, when n or k or both are large we need to adapt the implementation plan to make it computationally feasible. If, at any stage of the implementation procedure, the number of groups of observations is too large we propose the following general diagnostic plan where we choose a subset of all the observations (the so called basic subset) to look at more closely.

1. Select a basic subset S of potentially influential observations that we wish to consider by utilizing a relaxed (larger) significance level (Belsley et al., 1980, p.31). For instance, for single observations, we may use a significance level of 10% or more instead of holding to the more conventional 5%. If n choose k (the number of possible groups of observations) is too large we do the following:

   - Calculate $\bar{h}^* = 1.5(p+1)/n$ for ADQ; and the critical values corresponding to a relaxed significance level α^* for LD and LR;
   - Consider 1, 2, $\cdots$, $k-1$, observations at a time;
   - Choose all observations for which the values LD and LR exceed their relaxed critical values and ADQ exceed its cutoff value;

2. Calculate LD, LR statistics and ADQ for a subset $\mathcal{A}_k = \{i_1, \cdots, i_k\}$, where $i_j \in S, j = 1, \cdots, k$, and $k = 1, \cdots, N(S)$ (number of elements in S). Compare the $LD_{\mathcal{A}_k}$ and $LR_{\mathcal{A}_k}$ to their critical values $LD_{\mathcal{A}_k,\alpha}$ and

LR_α respectively, where α can take 0.05 or 0.01, and ADQ cutoff value $\bar{h} = 2(p+1)/n$.

Note that in a multivariate process monitoring context where the observations are ordered in time, we are more interested in considering k consecutive observations. Then there are only $n-k+1$ possible choices for k consecutive observations and it is computationally feasible to look at all the combinations to identify outliers.

5 Application

To illustrate the use of the proposed multiple outlier detection methods, we consider an educational research data example in this section.

The data were collected by Dr. W.D. Rohwer of University of California at Berkeley and reproduced in Timm (1975, p. 281, 345) and have been previously studied by Hossain and Naik (1989), Barrett and Ling (1992). The data correspond to 32 randomly selected school children in an upper-class, white residential school. For each of the 32 students, the independent variables are the sum of the number of items correct out of 20 (on two exposures) to five types of paired-associated (PA) tasks. The basic tasks were named (x_1), still (x_2), named still (x_3), named action (x_4), and sentence still (x_5). The goal was to determine if the student's score on these 5 tests could be used to predict the children's performance on three standardized tests, namely, Peabody Picture Vocabulary Test (y_1), Raven Progressive Matrices Test (y_2), and Student

Achievement Test (y_3).

To model this data, we consider the following multivariate linear model:

$$Y_{32\times 3} = X_{32\times 6}B_{6\times 3} + E_{32\times 3},$$

where $E \sim MVN(\mathbf{0}, I_{32} \otimes \Sigma)$, and Σ is a 3×3 covariance matrix.

Let Γ be the matrix B with the first row (intercepts) omitted. Hossain and Naik (1989) gives different statistics for testing $\Gamma = \mathbf{0}$ and all the criteria consistently reject the hypothesis indicating that at least some of the x-variables are important.

We begin our analysis by looking at each single observation at a time. Figure 1 shows the results of using our proposed outlier detection methods on one observation at a time. The solid curve for LD and the solid line for LR are the critical values for $\alpha = 0.05$ based on the approximate distribution. The dashed curve or line are based on simulation (using M=2000) as described in Section 3.2. The solid line in ADQ plot is the cutoff line $\bar{h} = 2(p+1)/n = 0.38$. From Figure 1 we see that the LD value for the 25th observation ($LD = 1.87$) exceeds its corresponding $LD_{0.05}$ critical values (1.62 from the simulation and 1.72 from the approximate distribution), and the LR value for the same observation (LR=9.13) exceeds its corresponding $LR_{0.05}$ critical values (7.69 from the simulation and 7.81 from the approximate distribution), but the corresponding ADQ value is less than its cutoff value $\bar{h}$. In addition, the ADQ values for the 5th and 10th observations are greater than the cutoff value $\bar{h}$. All these indicate that the 25th observation is a Y outlier, but the 5th and 10th observations are X outliers but not influential observations.

Figure 1: Plots of LD, LR, and ADQ for looking at one single observation at a time. The ith ($i = 1, 2, \cdots, 32$) dotted points is the corresponding statistic value considering the ith observation in each plot. The solid line for LR and the solid curve for LD are the critical values for $\alpha = 0.05$ based on the approximate distribution; the dashed curve or line are based on the simulation; and the solid line in ADQ plot is the cutoff line $\bar{h} = 2(p+1)/n = 0.38$

Next we consider the effects of all pairs of observations and take $\alpha = 0.01$ to somewhat accommodate multiple testing. The statistic values for the combination of 14th and 25th observation are $LD = 3.76$, $LD_{0.01} = 3.21$ from the simulation, and 3.73 from the approximate distribution; $LR = 17.94$, $LR_{0.01} =$ 16.65 from the simulation, and 16.81 from the approximate distribution; and ADQ=0.14. All other pairs which have ADQ values greater than $\bar{h}$ involve either the 5th observation or the 10th observation. Thus the combination of 14th and 25th observations are Y outliers. It should be noted that if we had only looked at observations one at a time we would not have detected the 14th observation for further investigation.

We can now consider 3 observations at a time. This will lead to too many combinations (32 choose 3) and hence we will try to obtain a basic subset as described in section 4. For this we go through observations one at a time with a relaxed α =0.10 and $\bar{h}$ =0.28, and two at a time with α =0.05 and the same $\bar{h}$. Consideration of observations one at a time with LD and LR leads to picking observations 14 and 25 for the basic subset, while ADQ leads to observations 5, 10, 15, 16, 19, 27 and 29. With observations two at a time with a lower α=0.10 and $\bar{h}$=0.28 we find that observations 3, 7, 8, 9, 12, 13, 17, 20, 21,

23, 31, and 32 need to be added to the basic subset in addition to what was found previously. Thus the basic subset contains 21 observations. Within this basic subset we consider observations three at a time with a smaller α=0.001 to alleviate the multiple testing problem. The LD statistic using the simulation cutoff point picked (13, 14, 25), (14, 23, 25) and (14, 25, 32) (see Table 2) as potential outlier triples. Neither the cutoff value from the approximation nor that from the LR picked any triples. ADQ also did not pick any triples. For reference actual LD and LR values, corresponding cutoff values and ADQ values are given in Table 2.

From the analysis of observations two at a time we found that the pair (14, 25) is a Y outlier and (5, 10) is an X outlier. When we consider observations three at a time it is clear that 14 and 25 are involved in the triples picked by the LD statistic. The differences in the statistic values in going from pairs to triples are small except possibly for 13. Thus we will stop the procedure here.

Thus observations 5, 10, 14, and 25 should be investigated to see why they are different from others. Observation 13 also needs some investigation. Examination of the data set indicate that the 5th observation has the largest x_1 value (20). This is almost twice as large as the x_1 values for the other observations except for the 10th observation. The 5th observation also has the largest value for variable x_3.

It is difficult to formulate an exact procedure for the multiple outlier case in multivariate regression because of the multiple testing and computational problems involved. Very little attention is given to the multiple outlier case

in most papers dealing with outliers. In certain cases some approximate and workable procedures are given as we have done here. The main objective in outlier analysis is to focus attention on potential outliers and investigate why these are different from others. This involves interaction with experimenters, data collectors etc.

6 Concluding remarks

The likelihood displacement statistic and the likelihood ratio statistic are developed to detect multiple outliers in the context of a multivariate linear regression model. The statistics are generalizations of the univariate Cook's distance and other diagnostic statistics. In order to obtain critical values for the two statistics, approximate distributions are obtained to get suitable cutoff points and the performance of the approximate distributions are examined by a simulation study when the sample size is large. When the sample size is small, a procedure is proposed to generate critical values for the two statistics by simulation. An implementation procedure to detect outliers is also proposed and an example demonstrates the procedure for detecting multiple outliers.

7 Proof of equation (6)

For all data without observations in $\mathcal{A}_k$, we have the following multivariate linear model:

$$Y_{[\mathcal{A}_k]} = X_{[\mathcal{A}_k]}B + E_{[\mathcal{A}_k]},$$

where $E_{[\mathcal{A}_k]} \sim MVN(\mathbf{0}, I_{n-k} \otimes \Sigma)$.

With this model, the maximum likelihood estimator of B is given by

$$\hat{B}_{[\mathcal{A}_k]} = (X'_{[\mathcal{A}_k]} X_{[\mathcal{A}_k]})^{-1} X'_{[\mathcal{A}_k]} Y_{[\mathcal{A}_k]}. \tag{11}$$

Now, we want to determine the relationship between $\hat{B}_{[\mathcal{A}_k]}$ and $\hat{B}$, the MLE with all data. Note that

$$X'_{[\mathcal{A}_k]} X_{[\mathcal{A}_k]} = X'X - X'_{\mathcal{A}_k} X_{\mathcal{A}_k}, \text{ and } X'_{[\mathcal{A}_k]} Y_{[\mathcal{A}_k]} = X'Y - X'_{\mathcal{A}_k} Y_{\mathcal{A}_k}. \tag{12}$$

From the identity (see, for example, Muirhead, 1982, p580)

$$(A + CBD)^{-1} = A^{-1} - A^{-1}CB(B + BDA^{-1}CB)^{-1}BDA^{-1},$$

where A and B are nonsingular $r \times r$ and $q \times q$ matrices respectively, and C is $r \times q$ and D is $q \times r$ matrices, we let $A = X'X, B = I, C = -X'_{\mathcal{A}_k}$, and $D = X_{\mathcal{A}_k}$, and get

$$\begin{aligned} (X'_{[\mathcal{A}_k]} X_{[\mathcal{A}_k]})^{-1} &= (X'X - X'_{\mathcal{A}_k} X_{\mathcal{A}_k})^{-1} \\ &= (X'X)^{-1} + (X'X)^{-1} X'_{\mathcal{A}_k} (I - X_{\mathcal{A}_k} (X'X)^{-1} X'_{\mathcal{A}_k})^{-1} X_{\mathcal{A}_k} (X'X)^{-1}. \end{aligned} \tag{13}$$

Now from (11), (12) and (13) we obtain

$$\hat{B}_{[\mathcal{A}_k]} = \hat{B} - (X'X)^{-1} X'_{\mathcal{A}_k} (I - Q_{\mathcal{A}_k})^{-1} \hat{E}_{\mathcal{A}_k}.$$

Acknowledgement

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada. We also would like to thank the referees for their helpful comments on an earlier version of this paper.

References

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd Edition. Wiley Series in Probability and Mathematical Statistics.

Barrett BE, Ling RF (1992) General classes of influence measures for multivariate regression. Journal of the American Statistical Association 87: 184-191.

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. John Wiley & Sons.

Box GEP (1949) A general distribution theory for a class of likelihood criteria. Biometrika 36: 317-346.

Cook RD (1977) Detection of influential observation in linear regression. Econometrics 19: 15-18.

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall.

Díaz-García JA, González-Farías G (2004). A note on the Cook's distance. Journal of Statistical Planning and Inference 120: 119-136.

Field C (1993) Tail areas of linear combinations of chi-squares and non-central chi-squares. Journal of Statistical Computation and Simulation 45: 243-248.

Hadi AS, Simonoff JS (1993) Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association 88: 1264-1272.

Hossain A, Naik DN (1989) Detection of influential observations in multivariate regression. Journal of Applied Statistics 16: 25-37.

Jensen DR, Solomon H (1972) A Gaussian approximation for the distribution of definite quadratic forms. Journal of the American Statistical Association 67: 898-902.

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, third edition. Prentice Hall.

Naik DN (2003) Diagnostic methods for univariate and multivariate normal data. In Handbook of Statistics 22 Statistics in Industry, edited by Khatree R and Rao CR.

Prescott P (1975) A simple alternative to Student's t. Applied Statistics 24: 210-217.

Seber GAF (1984) Multivariate observations. New York: John Wiley.

Srikantan KS (1961) Testing for the single outlier in a regression model. Sankhy$\bar{a}$, Series A 23: 251-260.

Srivastava MS, von Rosen D (1998) Outliers in multivariate regression models. Journal of Multivariate Analysis 65: 195-208.

Tietjen GL, Moore RH, Beckman RJ (1973) Testing for a single outlier in simple linear regression. Technometrics 15: 737-752.

Timm N (1975) Multivariate analysis with applications in education and psychology. Brooks/Cole, California.

Wei WH, Fung WK (1999) The mean-shift outlier model in general weighted regression and its applications. Computational Statistics & Data Analysis 30: 429-441.

Table 1: Comparison between the empirical significance levels of LD and LR from the approximations.

m	k	α	LD		LR
			Saddlepoint	Normal	Chi-square
1	1	0.10	0.1046	0.1046	0.1050
		0.05	0.0572	0.0572	0.0574
		0.01	0.0098	0.0098	0.0104
1	5	0.10	0.0964	0.0956	0.1068
		0.05	0.0460	0.0446	0.0548
		0.01	0.0102	0.0090	0.0122
2	2	0.10	0.0984	0.0976	0.1016
		0.05	0.0464	0.0452	0.0518
		0.01	0.0100	0.0092	0.0114
2	5	0.10	0.1044	0.1044	0.0974
		0.05	0.0512	0.0494	0.0470
		0.01	0.0102	0.0088	0.0104
5	2	0.10	0.0964	0.0960	0.0956
		0.05	0.0500	0.0496	0.0466
		0.01	0.0094	0.0092	0.0102
5	5	0.10	0.0940	0.0926	0.1074
		0.05	0.0450	0.0436	0.0478
		0.01	0.0098	0.0090	0.0102

Table 2: Identified Y outliers by the proposed methods: Rohwer's data, where LD_{sim} and LR_{sim} are corresponding cutoff values obtained from the simulation

Observation	LD	LD_{sim}	LR	LR_{sim}	$ADQ(\bar{h}=0.38)$
25	1.87	1.62	9.13	7.69	0.16
(14, 25)	3.76	3.21	17.94	16.65	0.14
(13, 14, 25)	5.35	4.72	22.51	25.17	0.13
(14, 23, 25)	4.08	4.03	21.32	26.52	0.11
(14, 25, 32)	4.60	4.20	23.37	29.39	0.12