國立交通大學

統計學研究所

碩 士 論 文

基因表現分析之穩健回歸估計量

Robust Regression Estimators in Gene Expression Analysis

研 究 生:張祐華

指導教授:陳鄰安　教授

中 華 民 國 一 百 零 二 年 六 月

基因表現分析之穩健回歸估計量

Robust Regression Estimators in Gene Expression Analysis

研 究 生:張祐華……Student:Yu-Hua Chang

指導教授:陳鄰安……Advisor:Lin-An Chen

國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

A Thesis

Submitted to Institude of Statistics

College of Science

National Chiao Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2013

Hsinchu, Taiwan, Republic of China

中華民國一百零二年六月

摘要

對基因表現分析來說，經由偵測疾病組樣本的離群值來發現對其有影響力的基因，是一個很新而且很重要的方法。不幸的是，我們在文獻裡找到，為了建構回歸模型而發展出的離群值最小平方法估計量，它的影響函數(influence function)無法限制住對獨立變數的影響。為了建構線性回歸模型，我們用 Mallow's type 離群值有界影響最小平方法估計量及離群值回歸分位數的漸進分布，產生出一個影響函數(influence function)在獨立變數空間是有界的統計方法。由蒙地卡羅模擬比較均方差的結果顯示，當過失誤差(gross error)在獨立變數空間發生時，有界影響的估計量比無界影響的更有效。

關鍵字:基因表現分析, 影響函數, 最小平方法估計量, 線性回歸, 回歸分位數

誌謝

韶光荏苒，兩年碩士即將到達尾聲，而我的學生生涯也將到此告一段落。大學及研究所都在新竹度過的我，想必會相當緬懷在這裡遇見的人、事、物，還有那無法令人遺忘的風吧。

首先我要由衷地感謝我的論文指導老師－陳鄰安教授。老師對於教導學生是相當地有一套自己的方法，每次接受老師的指導總是能很有架構的理解新的知識，也非常有耐心地講解其細節的部份，十分慶幸能在大學及研究所都當老師的學生。還有口試期間的三位口試委員，許文郁教授、蕭金福教授以及彭南夫教授，謝謝你們對我要補強的地方的建議與指教。

接著我要感謝交大統計所一百級的同學們，在這兩年的研究所生活中，大家相處十分融洽，一起切磋、一起成長，就好像我們認識了不只兩年一樣。

最後我要感謝我的家人，在我累了的時候，總是有個溫暖的家在等著我、關懷我、照顧我。謝謝，我最愛的家人們。

這篇論文獻給我的家人、朋友、老師及所有曾經幫助我的人。

張祐華　謹誌于

國立交通大學統計學研究所

中華民國一百零二年五月

Content

Robust Regression Estimators in Gene Expression Analysis

Abstract

Discovering the influential genes through the detection of outliers in samples from disease group subjects is a very new and important approach for gene expression analysis. Technique of outlier least squares estimator for regression model has been found in literature that, unfortunately, its influence function can not limit the effect of independent variables. We present asymptotic distributions of the Mallow's type bounded-influence outlier least squares estimator and outlier regression quantile for linear regression models producing statistical techniques with influence functions bounded in the space of independent variables. Monte Carlo simulations comparing mean squared errors show that the bounded-influence ones are more efficient than the unbounded-influence ones when gross errors occur in the independent-variable-space.

Key words. Gene expression analysis; influence function; least squares estimation; linear regression; regression quantile.

1. Introduction

Among the existing techniques in influential genes detection, common statistical methods for two-group comparisons, such as t-test, are not appropriate due to a large number of genes and a limited number of subjects available. Tomlins et al. (2005) observed in a study of prostate cancer that influential genes are over expressed in a small number of disease samples. The problem of constructing statistical procedures based on outlier samples has been attracted considerable recent attention. Tibshirani and Hastie (2007) and Wu (2007) suggested to use an outlier sum, the sum of all the gene expression values in the disease group that are greater than a specified cutoff point and Chen, Chen and Chan (2010) considered the distributional theory of the outlier mean. These methods show desired efficiency for tests based on outliers in detection of influential genes.

Uncertainties of gene expressions also show causal effect upon one or some predictor variables (independent variables) such as age, cell line type or genotype information (see Jin, Si et al. (2006), Huang and Pan (2003), Rambow, Piton et al. (2008), Muller, Chiou and Leng (2008), Vinciotti and Yu (2009) and Zapala and Schork (2006)). Lai, et al. (2013) considered that we have gene expressions for normal group subjects with regression model

$$y_{ai} = x'_{ai}\beta_a + \epsilon_i, i = 1, ..., n_1 \tag{1.1}$$

and those for disease group subjects with regression model

$$y_{bi} = x'_{bi}\beta_b + \delta_i, i = 1, ..., n_2. \tag{1.2}$$

They proposed the outlier least squares estimator (LSE) for influential genes detection by showing that the outlier LSE has an asymptotic representation with influence function of the form

$$M_a x_a \phi_a(\epsilon) + M_b x_b \phi_b(\delta) \tag{1.3}$$

where M_a and M_b are fixed matrices, ϕ_a is a bounded function and ϕ_b measures the tail mean of variable δ. The influence function is not bounded in the independent-variable-space. Therefore, one can conjecture that in small samples the outlier LSE will not be able to handle outliers in the X space. For a general discussion of influence analysis, see Cook and Weisberg (1982). In the literature, consideration has been given to the development of estimators of regression parameters that limit the effects of the error variable and the independent variables. Among them, approaches which simultaneously bound the influence of the design points and the residuals for the linear regression model include Krasker and Welsch(1982) and Krasker(1985). On the other hand, the approach of the Mallow's type bounded-influence trimmed mean is to bound the influence of the design points and the residuals by De Jongh and De Wet (1985) and in the linear regression model by De Jongh, De Wet and Welsh(1988). In a study by Giltinan, Carroll and Ruppert (1986), they found these two approaches are competitive in a way that

neither is preferable to the other one. They also note that the Mallow's type estimators should theoretically give more stable inference than the Krasker-Welsch approach. This desired property has been further studied by Chen, Thompson and Hung (2000). In light of the fact that bounded-influence type estimation has not been studied for outliers based estimators, our aim is to study the Mallow's type outlier least squares estimator (LSE) and outlier regression quantile for regression gene expression data sets. The asymptotic theory for the outlier LSE is given in Section 2 for the linear regression model and a simulation study for it is given in Section 3. We introduce the statistical theory and simulation study for the outlier regression quantile in Section 4. Finally the proofs of theorems are displayed in Section 5.

2. Mallows Type Bounded Influence Outlier Least Squares Estimator

For easy expression, let us fix one from thousands of genes for examination. Suppose that there are n_1 subjects in the normal control group and n_2 subjects in the disease group. We assume that this gene expressions for normal group subjects have the regression model

$$y_{ai} = x'_{ai}\beta_a + \epsilon_i, i = 1, ..., n_1 \tag{2.1}$$

where x_{ai} is p-vector with 1 as first element and ϵ_i's are independent and identically distribute (iid) error variables with distribution function F_ϵ and those disease group subjects have the regression model

$$y_{bi} = x'_{bi}\beta_b + \delta_i, i = 1, ..., n_2 \tag{2.2}$$

where x_{bi} is p-vector with 1 as first element and δ_i's are iid error variables with distribution function F_δ. Motivated from Tomlins et al. (2005) we need to construct a cutoff from model (2.1) to identify outlier observations in model (2.2) and develop a statistic based on these outliers as the basis for statistical inferences.

We let the sample Mallow's type bounded-x influence regression γ-quantile of Koenker and Bassett (1978) be a vector $\hat{\beta}_{BIa}(\gamma)$ that solves

$$\text{Min}_{b\in R^p} \sum_{i=1}^{n_1} w_{ai}(y_{ai} - x'_{ai}b)(\gamma - I(y_{ai} \le x'_{ai}b))$$

for defining the cutoff where $w_{ai}, i = 1, ..., n_2$ are weights. The Mallows-type bounded influence outlier LSE (De Jongh, De Wet and Welsh (1988)) is defined as

$$\hat{\beta}_{BIb,out} = (X_b' W_b A_{BI} X_b)^{-1} X_b' W_b A_{BI} y_b \tag{2.3}$$

where trimming matrix $A_{BI} = \text{diag}\{a_{ii} = I(y_{bi} \geq x_{bi}' \hat{\beta}_{BIa}(\gamma)), i = 1, ..., n_2\}$ and W_b is a diagonal matrix of weights w_{bi}'s.

We denote $\lambda_{b,out} = P(\delta \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})$. For the class of Mallows-type bounded influence outlier LSE, we assume that the following assumptions are valid.

Assumption 1: $lim_{n_2,n_1 \to \infty} \frac{n_2}{n_1} = \ell_{ba}$ and $n_2^{-1} \sum_{i=1}^{n_2} x_{bij}^4 = O(1)$ where x_{bij} is the jth element of vector x_{bi}.

Assumption 2: $lim_{n_1 \to \infty} n_1^{-1} X_a' X_a = Q_a$, $lim_{n_2 \to \infty} n_2^{-1} X_b' X_b = Q_b$, $\lim_{n_1 \to \infty} X_a' W_a X_a = Q_{aw}$, $\lim_{n_1 \to \infty} X_a' W_a^2 X_a = Q_{aww}$, $\lim_{n_2 \to \infty} X_b' W_b X_b = Q_{bw}$ and $\lim_{n_2 \to \infty} X_b' W_b^2 X_b = Q_{bww}$, where Q_a, Q_b, Q_{aw}, Q_{aww}, Q_{bw} and Q_{bww} are $p \times p$ positive definite matrices.

Assumption 3: $\beta_{a1} = \beta_{b1}$ where we denote $\beta_b = \begin{pmatrix} \beta_{b0} \\ \beta_{b1} \end{pmatrix}$ and $\beta_a = \begin{pmatrix} \beta_{a0} \\ \beta_{a1} \end{pmatrix}$ with β_{b0} and β_{a0} the intercept parameters and β_{b1} and β_{a1} being vectors of slope parameters.

We denote the outlier proportion $\lambda_{b,out} = P(y_b \geq x' \beta_a(\gamma))$. Under Assumption 3, we see that $\lambda_{b,out} = P(\delta \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})$. We also denote f_ϵ and f_δ the densitity functions, respectively, for F_ϵ and F_δ. For the rest of this paper, we assume that Assumptions 1-4 are true where 4 is listed in Appendix. These assumptions are similar to the standard ones for linear regression models as given in Ruppert and Carroll (1980) and Portnoy and Koenker (1989).

Theorem 2.1. (a) The Mallows type bounded influence outlier LSE $\hat{\beta}_{MBb,out}$

has the following representation

$$n_2^{1/2}(\hat{\beta}_{BIb,out} - \beta_{b,out}) = -\lambda_{b,out}^{-1}\ell_{ba}^{1/2}(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})f_\delta(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})$$
$$f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))Q_{aw}^{-1}n_1^{-1/2}\sum_{i=1}^{n_1} w_{ai}x_i(\gamma - I(\epsilon_i \le F_\epsilon^{-1}(\gamma))) + \lambda_{b,out}^{-1}Q_{bw}^{-1}n_2^{-1/2}$$
$$\sum_{i=1}^{n_2} w_{bi}x_i[\delta_i I(\delta_i \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}) - E(\delta I(\delta \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}))] + o_p(1)$$

where $\beta_{b,out} = \beta_b + \mu_{\delta,out}e$ and where e is p vector $(1, 0, ..., 0)'$ and $\mu_{\delta,out} = E(\delta|\delta \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})$.
(b) $n_2^{1/2}(\hat{\beta}_{BIb,out} - \beta_{b,out})$ converges in distribution to a normal random vector with distribution $N_p(0, \sigma_{\delta,c}^2 Q_{aw}^{-1}Q_{aww}Q_{aw}^{-1} + \sigma_{\delta,out}^2 Q_{bw}^{-1}Q_{bww}Q_{bw}^{-1})$ where

$$\sigma_{\delta,out}^2 = var(\lambda_{b,out}^{-1}\delta I(\delta \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})) = \lambda_{b,out}^{-2}\int_{F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0}}^{\infty} \delta^2 dF_\delta(\delta)$$
$$- \mu_{\delta,out}^2, \text{ and}$$
$$\sigma_{\delta,c}^2 = \frac{\gamma(1-\gamma)}{\lambda_{b,out}^2}\ell_{ba}[(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})f_\delta(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})$$
$$f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))]^2.$$

The unbounded outlier LSE of Lai et al. (2013) equals $\hat{\beta}_{b,out} = \hat{\beta}_{BIb,out}$ with $w_{ai} = w_{bi} = 1$ for all i's.

3. Monte Carlo Study

We now compare the efficiencies of the unbounded-influence and the bounded-influence outlier LSE's through a Monte Carlo study. The purpose of the Monte Carlo study is to evaluate the small-sample behavior of these two outlier LSE's. The performance of these two outlier LSE's in presence of outliers and leverage points is of particular interest.

Denote the n observations of the (j-1)-th independent variable by $x_{1j}, ..., x_{nj}$ for $j = 2, 3, ..., p$. Order the n observations $x_{(1)j}, ..., x_{(n)j}$ and define $r_{1j}, ..., r_{nj}$ as the ranks of $x_{1j}, ..., x_{nj}$. Let $L = [\tau n] + 1$ and $U = n + 1 - L$ where τ we call it the Winsorized percentage is specified for 0.15 as it recommended

by De Jongh, De Wet and Welsh (1988). The weights associated with the (j-1)-th independent variable are now defined as

$$w_{ij} = \begin{cases} 1 & \text{if } L \le \gamma_{ij} \le U \\ (x_{(L)j} - x_{(U)j})/D_{ij} & \text{if } \gamma_{ij} < L \\ (x_{(U)j} - x_{(L)j})/D_{ij} & \text{if } \gamma_{ij} > U \end{cases}$$

where $D_{ij} = 2x_{ij} - x_{(U)j} - x_{(L)j}, i = 1, ..., n$. The Mallow's weights are now defined as $w_i = \pi_{j=2}^{p} w_{ij}$. See Denby and Larsen(1977) and De Jongh, De Wet and Welsh (1988) for these settings in regression parameters estimation which also perform well in our quantile study. With sample sizes $n = 50, 100$, the simple linear egression models of (1.1) and (1.2) are considered. The distribution of error variable ϵ is the standard normal (N(0,1)) and contaminated normal distribution

$$CN(\delta, \sigma) = (1 - \delta)N(0, 1) + \delta N(\mu, 1),$$

with $\delta = 0.1$.

The sample of independent variables is considered in the following designs:

D1: $x_{ij}, i = 1, ..., n$ are i.i.d $N(0, 1)$ for $j = 2, ..., p$.

D2: As D1, but one point is moved out 5 units in X space.

D3: As D1, but two points are moved out 5 units in X space.

D4: As D1, but one point is moved out 10 units in X space.

D5: As D1, but two points are moved out 10 units in X space.

Design D1 generates ideal observations x_{ij} and we expect the unbounded-influence outlier LSE to be more efficient than the bounded one no matter what the distribution of the error variable is. On the other hand, influential observations x_{ij} would occur for designs D2 - D5 where we expect that the bounded-influence outlier LSE to be more efficient than the unbounded one; however, it is interesting to see how much more efficient it is.

Table 1. The efficencies of unbounded-influence outlier LSE and bounded influence outlier LSE ($n = 50$)

	$\gamma = 0.6$ Eff_b, Eff_{BIb}	$\gamma = 0.7$ Eff_b, Eff_{BIb}	$\gamma = 0.8$ Eff_b, Eff_{BIb}	$\gamma = 0.9$ Eff_b, Eff_{BIb}
D1				
$\mu = 0.5$	100 84	100 85	100 87	100 91
$\mu = 1$	100 84	100 85	100 87	100 90
$\mu = 1.5$	100 84	100 85	100 86	100 89
$\mu = 2$	100 84	100 85	100 86	100 88
$\mu = 2.5$	100 84	100 84	100 85	100 88
D2				
$\mu = 0.5$	19 100	24 100	31 100	48 100
$\mu = 1$	19 100	23 100	30 100	45 100
$\mu = 1.5$	19 100	23 100	29 100	42 100
$\mu = 2$	21 100	24 100	30 100	41 100
$\mu = 2.5$	23 100	26 100	31 100	41 100
D3				
$\mu = 0.5$	30 100	37 100	47 100	64 100
$\mu = 1$	29 100	35 100	45 100	61 100
$\mu = 1.5$	29 100	35 100	43 100	58 100
$\mu = 2$	29 100	35 100	43 100	56 100
$\mu = 2.5$	30 100	36 100	43 100	55 100
D4				
$\mu = 0.5$	63 100	70 100	77 100	85 100
$\mu = 1$	62 100	68 100	76 100	84 100
$\mu = 1.5$	61 100	67 100	74 100	83 100
$\mu = 2$	61 100	67 100	73 100	81 100
$\mu = 2.5$	61 100	67 100	73 100	81 100
D5				
$\mu = 0.5$	78 100	82 100	87 100	90 100
$\mu = 1$	77 100	82 100	86 100	90 100
$\mu = 1.5$	77 100	81 100	85 100	89 100
$\mu = 2$	77 100	81 100	85 100	89 100
$\mu = 2.5$	77 100	81 100	85 100	89 100

A total of $10,000$ replications were performed. Table 1 presents the Monte Carlo results in the form of efficiencies compared with the best of the unbounded-influence outlier LSE and the bounded-influence outlier LSE; that is, the efficiency is equal to the average mean squared error of the best one times 100 divided by the average mean squared error of the outlier LSE

$$Eff_{BIb} = \frac{min\{MSE_b, MSE_{BIb}\}}{MSE_{BIb}} \text{ and } Eff_b = \frac{min\{MSE_b, MSE_{BIb}\}}{MSE_b}$$

where MSE_b is the average of MSE's of the unbounded-influence outlier

LSE and MSE_{BIb} is the average of MSE's of the bounded-influence outlier LSE. In Tables 1 and 2, we consider gross errors appear only on disease group data (x_{bi}).

Table 2. The efficencies of unbounded-influence outlier LSE and bounded-influence outlier LSE ($n_a = n_b = 100$)

	$\gamma = 0.6$ Eff_b, Eff_{BIb}	$\gamma = 0.7$ Eff_b, Eff_{BIb}	$\gamma = 0.8$ Eff_b, Eff_{BIb}	$\gamma = 0.9$ Eff_b, Eff_{BIb}
D1				
$\mu = 0.5$	100 83	100 84	100 84	100 88
$\mu = 1$	100 83	100 84	100 84	100 87
$\mu = 1.5$	100 82	100 83	100 84	100 86
$\mu = 2$	100 82	100 83	100 84	100 86
$\mu = 2.5$	100 83	100 83	100 84	100 85
D2				
$\mu = 0.5$	47 100	55 100	64 100	77 100
$\mu = 1$	47 100	54 100	63 100	76 100
$\mu = 1.5$	47 100	53 100	61 100	73 100
$\mu = 2$	47 100	53 100	62 100	72 100
$\mu = 2.5$	49 100	55 100	62 100	72 100
D3				
$\mu = 0.5$	65 100	71 100	78 100	86 100
$\mu = 1$	64 100	70 100	77 100	85 100
$\mu = 1.5$	64 100	70 100	76 100	84 100
$\mu = 2$	64 100	70 100	76 100	83 100
$\mu = 2.5$	65 100	70 100	76 100	83 100
D4				
$\mu = 0.5$	86 100	88 100	90 100	92 100
$\mu = 1$	86 100	88 100	89 100	91 100
$\mu = 1.5$	86 100	88 100	89 100	91 100
$\mu = 2$	86 100	88 100	89 100	91 100
$\mu = 2.5$	86 100	88 100	89 100	91 100
D5				
$\mu = 0.5$	92 100	93 100	93 100	93 100
$\mu = 1$	92 100	92 100	93 100	93 100
$\mu = 1.5$	92 100	92 100	93 100	94 100
$\mu = 2$	92 100	93 100	93 100	94 100
$\mu = 2.5$	92 100	93 100	93 100	94 100

Several conclusions can be drawn from the simulated results:

(a). In design D1, the regression matricse are well-behaved and the error

variables have distributions with moderate to very heavy tails. The results are as expected, that is, the unbounded-influence outlier LSE is more efficient than the Mallow's type bounded-influence outlier LSE. However, the efficiency of the Mallow's type bounded-influence outlier LSE is quite robust in that its efficiencies are all greater than 84 in Table 1 and 82 in Table 2 in this idea design of the regression matrices.

(b). In designs D2-D5, the error variablles follow the distributions exactly as in design D1, but gross errors are introduced in the regression matrices. The Mallow's type bounded-influence outlier LSE's performed much better than the unbounded-influence outlier LSE's. For the design D2, the unbounded-influence outlier LSE in Table 1 is very poor with efficiency less than 19 in Table 1 and 47 in Table 2.

In the next we consider the simulation that response variables in model (2.1) of control group and model (2.2) of disease group are both simultaneously imposed with gross errors from D1 to D5 to evaluate the efficiencies of Mallows type outlier estimators.

The results also show that the Mallows type bounded influence outlier LSE is much better than the unbounded influence one when gross erros exist in x-space.

4. Mallows Type Outlier Regression Quantile

Regression quantile, introduced by Koenker and Bassett (1978), plays the role of order statistics for the linear regression model that is useful in constructing broad class of L-estimators (Koenker and Zhao (1994) and Portnoy and Koenker (1989)) as different measures of central tendency and statistical dispersion and also measures of other distributional characteristics. A regression outlier α-quantile $\beta_{b,out}(\alpha)$ models the relationship between covariates and variable y_b with $\alpha = P(y_b \leq x'\beta_{b,q}(\alpha)|y_b \geq x'\beta_a(\gamma))$ that could be seen in the form

$$\beta_{b,q}(\alpha) = \beta_b + F_\delta^{-1}(1 - \lambda_{b,out}(1-\alpha))e.$$

Following Koenker and Bassett (1978), we define the sample bounded-influence

regression outlier α-quantile as

$$\hat{\beta}_{BIb,q}(\alpha) = \arg_{b\in R^p}\min \sum_{i=1}^{n} w_{bi}(y_{bi}-x'_{bi}b)[\alpha-I(y_{bi} \le x'_{bi}b)]I(y_{bi} \ge x'_{bi}\hat{\beta}_{BIa}(\gamma))$$

Table 3. The efficencies of outlier LSE and bounded influence outlier LSE ($n = 30$)

	$\gamma = 0.6$ Eff_{oq}, Eff_{boq}	$\gamma = 0.7$ Eff_{oq}, Eff_{boq}	$\gamma = 0.8$ Eff_{oq}, Eff_{boq}	$\gamma = 0.9$ Eff_{oq}, Eff_{boq}
D1				
$\mu = 0.5$	100 83	100 83	100 83	100 85
$\mu = 1$	100 82	100 83	100 84	100 85
$\mu = 1.5$	100 83	100 82	100 83	100 84
$\mu = 2$	100 83	100 83	100 83	100 84
$\mu = 2.5$	100 83	100 83	100 83	100 84
D2				
$\mu = 0.5$	47 100	53 100	61 100	69 100
$\mu = 1$	47 100	53 100	60 100	68 100
$\mu = 1.5$	48 100	53 100	59 100	66 100
$\mu = 2$	50 100	54 100	59 100	65 100
$\mu = 2.5$	53 100	57 100	61 100	65 100
D3				
$\mu = 0.5$	57 100	62 100	67 100	75 100
$\mu = 1$	56 100	61 100	67 100	74 100
$\mu = 1.5$	57 100	60 100	66 100	74 100
$\mu = 2$	58 100	61 100	66 100	73 100
$\mu = 2.5$	59 100	62 100	66 100	73 100
D4				
$\mu = 0.5$	51 100	60 100	66 100	71 100
$\mu = 1$	49 100	59 100	65 100	69 100
$\mu = 1.5$	49 100	58 100	63 100	69 100
$\mu = 2$	49 100	57 100	62 100	67 100
$\mu = 2.5$	50 100	57 100	61 100	67 100
D5				
$\mu = 0.5$	64 100	68 100	72 100	75 100
$\mu = 1$	63 100	67 100	71 100	75 100
$\mu = 1.5$	63 100	66 100	70 100	74 100
$\mu = 2$	62 100	66 100	70 100	75 100
$\mu = 2.5$	64 100	66 100	69 100	74 100

The following theorem gives $\hat{\beta}_{BIb,out}(\alpha)$ the asymptotic representation and asymptotic distribution.

Theorem 4.1. (a) A Bahadur representation for the bounded-influence outlier regression quantile is

$$n_2^{1/2}(\hat{\beta}_{BIb,q}(\alpha) - \beta_{b,q}(\alpha)) = f_\delta^{-1}(F_\delta^{-1}(1-\lambda_{b,out}(1-\alpha)))f_\delta(\beta_{a0}-\beta_{b0}+F_\epsilon^{-1}(\gamma))f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))$$
$$\ell_{ba}^{1/2}Q_{aw}^{-1}n_1^{-1/2}\sum_{i=1}^{n_1}w_{ai}x_{ai}[\gamma - I(\epsilon_i \le F_\epsilon^{-1}(\gamma))] + f_\delta^{-1}(F_\delta^{-1}(1-\lambda_{b,out}(1-\alpha)))Q_{bw}^{-1}$$
$$n_2^{-1/2}\sum_{i=1}^{n_2}w_{bi}x_{bi}[\alpha - I(\delta_i \le F_\delta^{-1}(1-\lambda_{b,out}(1-\alpha)))]I(\delta_i \ge \beta_{a0}-\beta_{b0}+F_\epsilon^{-1}(\gamma)) + o_p(1)$$

(b) $n_2^{1/2}(\hat{\beta}_{BIb,q}(\alpha) - \beta_{b,q}(\alpha))$ coverges to normal distribution with mean 0_p and covariance matrix

$$\rho_{\delta,q}^2 Q_{aw}^{-1}Q_{aww}Q_{aw}^{-1} + \rho_{\delta,out}^2 Q_{bw}^{-1}Q_{bww}Q_{bw}^{-1}$$

where

$$\rho_{\delta,q}^2 = \gamma(1-\gamma)\ell_{ba}(f_\delta^{-1}(F_\delta^{-1}(1-\lambda_{b,out}(1-\alpha)))f_\delta(\beta_{a0}-\beta_{b0}+F_\epsilon^{-1}(\gamma))$$
$$f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma)))^2, \text{ and}$$
$$\rho_{\delta,out}^2 = \lambda_{b,out}\alpha(1-\alpha)(f_\delta^{-1}(F_\delta^{-1}(1-\lambda_{b,out}(1-\alpha))))^2.$$

Let $\hat{\beta}_{b,q}(\alpha)$ be the unbounded-influence outlier regression α-quantile of Lai et al. (2013). We perform a simulation study of replications $1,000$. Let MSE_{BIbq} and MSE_{bq} be the average MSE's of $\hat{\beta}_{BIbq,q}(\alpha)$ and $\hat{\beta}_{b,q}(\alpha)$, respectively. We define efficiencies of these two unbounded-influence and bounded-influence regression quantiles as

$$Eff_{bq} = \frac{min\{MSE_{bq}, MSE_{BIbq}\}}{MSE_{bq}} \text{ and } Eff_{BIbq} = \frac{min\{MSE_{bq}, MSE_{BIbq}\}}{MSE_{BIbq}}.$$

Table 4. The efficencies of unbounded-influence outlier quantile and bounded-influence outlier quantile ($\alpha = 0.8$)

	$\gamma = 0.6$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.7$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.8$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.9$ Eff_{bq}, Eff_{BIbq}
D1				
$\mu = 0.5$	100 90	100 88	100 89	100 96
$\mu = 1$	100 90	100 88	100 89	100 97
$\mu = 1.5$	100 90	100 88	100 89	100 96
$\mu = 2$	100 91	100 88	100 89	100 96
$\mu = 2.5$	100 91	100 89	100 89	100 96
D2				
$\mu = 0.5$	77 100	39 100	19 100	61 100
$\mu = 1$	79 100	42 100	20 100	58 100
$\mu = 1.5$	80 100	45 100	24 100	57 100
$\mu = 2$	82 100	50 100	31 100	61 100
$\mu = 2.5$	85 100	58 100	43 100	65 100
D3				
$\mu = 0.5$	44 100	22 100	28 100	84 100
$\mu = 1$	48 100	23 100	28 100	80 100
$\mu = 1.5$	53 100	28 100	30 100	77 100
$\mu = 2$	60 100	38 100	38 100	77 100
$\mu = 2.5$	70 100	52 100	48 100	78 100
D4				
$\mu = 0.5$	20 100	11 100	12 100	48 100
$\mu = 1$	23 100	13 100	12 100	45 100
$\mu = 1.5$	27 100	16 100	15 100	43 100
$\mu = 2$	35 100	22 100	19 100	46 100
$\mu = 2.5$	46 100	33 100	29 100	50 100
D5				
$\mu = 0.5$	20 100	13 100	23 100	88 100
$\mu = 1$	23 100	14 100	21 100	84 100
$\mu = 1.5$	27 100	17 100	23 100	80 100
$\mu = 2$	35 100	23 100	29 100	79 100
$\mu = 2.5$	46 100	35 100	39 100	89 100

Table 5. The efficencies of unbounded-influence outlier quantile and bounded-influence outlier quantile ($\alpha = 0.9$)

	$\gamma = 0.6$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.7$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.8$ Eff_{bq}, Eff_{BIbq}	$\gamma = 0.9$ Eff_{bq}, Eff_{BIbq}
D1				
$\mu = 0.5$	100 94	100 92	100 91	100 97
$\mu = 1$	100 94	100 92	100 91	100 97
$\mu = 1.5$	100 94	100 92	100 91	100 96
$\mu = 2$	100 95	100 92	100 91	100 97
$\mu = 2.5$	100 95	100 93	100 91	100 96
D2				
$\mu = 0.5$	91 100	68 100	34 100	65 100
$\mu = 1$	92 92	72 100	38 100	62 100
$\mu = 1.5$	100 92	83 100	60 100	67 100
$\mu = 2$	94 100	83 100	60 100	67 100
$\mu = 2.5$	100 92	89 89	71 100	71 100
D3				
$\mu = 0.5$	75 100	53 100	42 100	85 100
$\mu = 1$	78 100	58 100	43 100	81 100
$\mu = 1.5$	83 100	68 100	51 100	79 100
$\mu = 2$	89 100	80 100	63 100	79 100
$\mu = 2.5$	$100_9 0$	89 89	73 100	81 100
D4				
$\mu = 0.5$	48 100	34 100	22 100	51 100
$\mu = 1$	52 100	39 100	25 100	47 100
$\mu = 1.5$	61 100	48 100	31 100	46 100
$\mu = 2$	73 100	62 100	43 100	50 100
$\mu = 2.5$	84 100	77 100	56 100	55 100
D5				
$\mu = 0.5$	48 100	35 100	32 100	89 100
$\mu = 1$	52 100	39 100	32 100	86 100
$\mu = 1.5$	61 100	49 100	39 100	83 100
$\mu = 2$	73 100	63 100	50 100	83 100
$\mu = 2.5$	85 100	78 100	64 100	84 100

Several conclusions can be drawn from the simulated results:

(a). In design D1, the regression matricse are well-behaved and the error variables have distributions with moderate to very heavy tails. The results are as expected, that is, the unbounded-influence outlier regression quantile is more efficient than the Mallow's type bounded-influence outlier regression quantile. However, the efficiency of the Mallow's type bounded-influence outlier regression quantile is quite robust in that its efficiencies are all greater than 88 in Table 1 and 92 in Table 2 in this idea design of the regression

matrices.

Table 6. The efficencies of outlier quantile and bounded influence outlier quantile ($\alpha = 0.8, n = 50$)

	$\gamma = 0.6$ Eff_{oq}, Eff_{boq}	$\gamma = 0.7$ Eff_{oq}, Eff_{boq}	$\gamma = 0.8$ Eff_{oq}, Eff_{boq}	$\gamma = 0.9$ Eff_{oq}, Eff_{boq}
D1				
$\mu = 0.5$	100 94	100 92	100 91	100 97
$\mu = 1$	100 94	100 92	100 91	100 96
$\mu = 1.5$	100 94	100 92	100 91	100 95
$\mu = 2$	100 95	100 92	100 91	100 95
$\mu = 2.5$	100 94	100 92	100 91	100 94
D2				
$\mu = 0.5$	72 100	53 100	59 100	86 100
$\mu = 1$	75 100	55 100	59 100	86 100
$\mu = 1.5$	77 100	60 100	62 100	85 100
$\mu = 2$	80 100	67 100	68 100	85 100
$\mu = 2.5$	84 100	75 100	76 100	85 100
D3				
$\mu = 0.5$	61 100	59 100	77 100	91 100
$\mu = 1$	63 100	59 100	75 100	90 100
$\mu = 1.5$	68 100	62 100	74 100	90 100
$\mu = 2$	75 100	67 100	74 100	89 100
$\mu = 2.5$	81 100	74 100	77 100	88 100
D4				
$\mu = 0.5$	40 100	31 100	48 100	81 100
$\mu = 1$	44 100	33 100	48 100	79 100
$\mu = 1.5$	49 100	37 100	50 100	79 100
$\mu = 2$	57 100	47 100	56 100	78 100
$\mu = 2.5$	66 100	59 100	64 100	79 100
D5				
$\mu = 0.5$	42 100	50 100	76 100	88 100
$\mu = 1$	45 100	48 100	73 100	88 100
$\mu = 1.5$	50 100	50 100	72 100	87 100
$\mu = 2$	58 100	56 100	72 100	85 100
$\mu = 2.5$	68 100	65 100	76 100	84 100

(b). In designs D2-D5, the error variablles follow the distributions exactly as in design D1, but gross errors are introduced in the regression matrices. The Mallow's type bounded-influence outlier regression quantile's performed much better than the unbounded-influence outlier regression quantile's. For

the design D2, the unbounded-influence outlier regression quantile in Table 1 is very poor with efficiency less than 11 in Table 1 and 22 in Table 2.

In the next we consider the simulation that response variables in model (2.1) of control group and model (2.2) of disease group are both simultaneously imposed with gross errors from D1 to D5 to evaluate the efficiencies of Mallows type outlier quantile estimators.

The results also show that the Mallows type bounded influence outlier regression quantile is much better than the unbounded influence one when gross erros exist in x-space.

5. Appendix

It requires one more assumption for the proofs of theorems in this paper.

Assumption 4: Pobability density functions f_ϵ and f_δ are bounded away from zero, respectively, in neighborhoods of $F_\epsilon^{-1}(\alpha)$ and $F_\delta^{-1}(\alpha)$ for $\alpha \in (0,1)$.

Proof of Theorem 2.1. From the expression of $\hat{\beta}_{BIb,out}$ of (2.3) and model (2.2), we have

$$\begin{aligned} n_2^{1/2}(\hat{\beta}_{BIb,out} - \beta_{b,out}) &= n_2^{1/2}(\sum_{i=1}^{n_2} w_{bi}x_{bi}x_{bi}' I(y_{bi} \geq x_i'\hat{\beta}_{aw}(\gamma)))^{-1}\{\sum_{i=1}^{n_2} w_{bi}x_{bi}\delta_i[I(\delta_i \geq F_\epsilon^{-1}(\gamma) \\ &+ \beta_{a0} - \beta_{b0} + n_1^{-1/2}x_i'T_a) - I(\delta_i \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})] \\ &+ \sum_{i=1}^{n_2} w_{bi}x_{bi}\delta_i I(\delta_i \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})\} + o_p(1) \end{aligned} \tag{5.1}$$

where $T_a = n_1^{1/2}(\hat{\beta}_{BIa}(\gamma) - \beta_a(\gamma))$.

With Assumption (4) and Jureckova and Sen (1987) extension of Billingsly's Theorem (see also Koul (1992)), the first term on the right hand side of (5.1) may be expressed as

$$\begin{aligned} &n_2^{-1/2}\sum_{i=1}^{n_2} w_{bi}x_{bi}\delta_i[I(\delta_i \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0} + n_1^{-1/2}x_i'T_n) - I(\delta_i \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})] \\ &\quad = -(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})\ell_{ba}^{1/2} f_\delta(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})Q_{bw}T_n + o_p(1) \end{aligned} \tag{5.2}$$

for any sequence T_n with $T_n = O_p(1)$.

We know that, from Chen, Thompson and Chuang (2000),

$$n_1^{1/2}(\hat{\beta}_{BIa}(\gamma)-\beta_a(\gamma)) = Q_{a,w}^{-1} f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma)) n_1^{-1/2} \sum_{i=1}^{n_1} w_{ai} x_{ai}(\gamma - I(\epsilon_i \le F_\epsilon^{-1}(\gamma))) + o_p(1). \tag{5.3}$$

By the same rational, we can derive

$$\begin{aligned} & n_2^{-1/2} \sum_{i=1}^{n_2} w_{bi} x_{bi} x_{bi}' I(\delta_i \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0} + n_1^{-1/2} x_{bi}' T_a) \\ & \quad = n_2^{-1/2} \sum_{i=1}^{n_2} w_{bi} x_{bi} x_{bi}' I(\delta_i \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}) + o_p(1), \end{aligned}$$

for any sequence $T_a = O_p(1)$. This indicates

$$n_2^{-1} \sum_{i=1}^{n_2} w_{bi} x_{bi} x_{bi}' I(y_{bi} \ge x_i' \hat{\beta}_{aw}(\gamma)) = \lambda_{b,out} Q_{bw} + o_p(1). \tag{5.4}$$

By letting $T_a = T_n$ and combining the results in (5.1)-(5.4), result (a) of the theorem is followed.

The asymptotic normality of (b) is a direct consequence of the representation and the central limit theorem. □

Proof of Theorem 4.1. Let $U(t_1, t_2) = n_2^{-1/2} \sum_{i=1}^{n_2} w_{bi} x_{bi} I(\delta_i \le F_\delta^{-1}(1 - \lambda_{b,out}(1-\alpha)) + n_2^{-1/2} x_{bi}' t_2) I(\delta_i \ge \beta_{a0} - \beta_{b0} + F_\epsilon^{-1}(\gamma) + n_1^{-1/2} x_{bi}' t_1)$. From Jureckova and Sen's (1987) extension of Billingsley's Theorem (see also Koul (1992)), we have

$$\begin{aligned} U(T_1, T_2) - U(0,0) = & Q_{bw} f_\delta(F_\delta^{-1}(1 - \lambda_{b,out}(1-\alpha))) T_2 - Q_{bw} f_\delta(\beta_{a0} \\ & - \beta_{b0} + F_\epsilon^{-1}(\gamma)) \ell_{ba}^{1/2} T_1 + o_p(1) \end{aligned} \tag{5.5}$$

for any sequences $T_1 = O_p(1)$ and $T_2 = O_p(1)$. Following the proof of Lemma 3.3 of Chen and Chiang (1996) (see also Ruppert and Carroll (1980)), it can see that

$$\begin{aligned} & U(n_1^{1/2}(\hat{\beta}_{BIa}(\gamma) - \beta_a(\gamma)), n_2^{1/2}(\hat{\beta}_{BIb,q}(\alpha) - \beta_{b,q}(\alpha))) \\ & \quad = n_2^{-1/2} \sum_{i=1}^{n_2} w_{bi} x_{bi} [\alpha - I(y_{bi} \le x_{bi}' \hat{\beta}_{b,q}(\alpha))] I(y_{bi} \ge x_{bi}' \hat{\beta}_a(\gamma)) \\ & \quad = o_p(1). \end{aligned} \tag{5.6}$$

Also, using the method of Jureckova (1977, Lemma 5.2) and (5.5), one can show that for $\lambda > 0$ there exists, η, k and N_0 such that

$$P\{\inf_{|t_2|\geq k} n_2^{-1/2} | \sum_{i=1}^{n_2} w_{bi} x_{bi} [\alpha - I(\delta_i \leq F_\delta^{-1}(1 - \lambda_{b,out}(1-\alpha) + n_2^{-1/2} x_{bi}' t_2)] \\ I(\delta_i \geq \beta_{a0} - \beta_{b0} + F_\epsilon^{-1}(\gamma) + n_1^{-1/2} x_{bi}' T_3)| \leq \eta\} \leq \lambda \qquad (5.7)$$

where T_3 is any sequence of random vector with $T_3 = O_p(1)$. Then the weak consistency of $\hat{\beta}_{BIb,out}(\alpha)$ can be obtained from the root-consistency of $\hat{\beta}_{BIb,out}(\alpha)$ given by

$$n_2^{1/2}(\hat{\beta}_{BIb,q}(\alpha) - \beta_{b,q}(\alpha)) = O_p(1)$$

which is induced from (5.6) and (5.7). Result (a) in Theorem 4.1 is followed from (5.5) and (5.7) by setting $T_1 = n_1^{1/2}(\hat{\beta}_{BIa}(\gamma) - \beta_a(\gamma))$ and $T_2 = n_2^{1/2}(\hat{\beta}_{BIb,q}(\alpha) - \beta_{b,q}(\alpha))$. □

REFERENCES

10. Chen, L.-A., Thompson, P. and Chuang, H.-C. (2000). Mallow's type bounded influence regression quantile for linear regression model and simultaneous equations model. *Sankhya Ser. B.* 62, 217-232.

Chen, L.-A., Chen, D.-T. and Chan, W. (2010). The p Value for the Outlier Sum in Differential Gene Expression Analysis. *Biometrika*, **97**, 246-253.

Chen, L.-A. and Chiang, Y. C. (1996). Symmetric type quantile and trimmed means for location and linear regression model. *Journal of Nonparametric Statistics.* **7**, 171-185.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression, Chapmanand Hall, NewYork.

De Jongh, P. J. and De Wet, T. (1985). Trimmed mean and bounded influence estimators for the parameters of the AR(1) process, *Communications in Statistics - Theory and Methods*, 14,1361-1357.

De Jongh, P. J., De Wet, T. and Welsh, A. H. (1988). Mallows-type bounded-influence-regression trimmed means. *Journal of the American Statistical Association* **83**, 805-810.

Giltinan, D. M., Carroll, R. J. and Ruppert, D. (1986). Some new estimation methods for weighted regression when there are possible outliers. *Technometrics,* 28, 219-230.

Koenker, R. and Bassett, G.J. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

Koenker, R. W. and Portnoy, S. (1987). L-estimation for linear model. *Journal of the American Statistical Association*, 82, 851-857.

Krasker, W. S. (1985). Two stage bounded-influence estimators for simultaneous equations models. *Journal of Business and Economic Statistics*, 4, 432-444.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded influence regression estimation. *Journal of the American Statistical Association*, 77, 595-604.

Lai, Y.-H., Chen, H.-C., Chen, L.-A. and Chen, D.-T. (2013). Statistical inferences based on outliers for gene expression analysis. Unpublished paper.

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of American Statistical Association* **75**, 828-838.

Tibshirani, R. and Hastie, T. (2007). Outlier sums differential gene expression analysis. *Biostatistics*, **8**, 2-8.

Tomlins, S. A., Rhodes, D. R., Perner, S., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644-648.