東海大學統計系研究所

碩士論文

指導教授：魏文翔 教授

Robust Model Selection for Correlated Data

相關性資料之強穩模式選取

研究生：邱煒婷

中華民國九十九年六月

Robust Model Selection for Correlated Data

By

Wei-Ting, Chiu

Dept. of Statistics,

Tunghai University,

Taichung, 40704,

Taiwan, R.O.C.

June 29, 2010

謝　誌

在東海大學四年以及研究所兩年的日子, 優靜的校園與專業系所的薰陶，奠定了學生專業的統計知識與培養多角度的思考能力，隨著論文的付梓，將劃上句點，回憶起這幾年來求學過程中所經歷的點點滴滴，充滿許多的不捨與感激。

本論文能順利完成，幸蒙指導教授魏文翔博士對於研究方向、觀念的啟迪，給予我很大的空間，並且不時地討論並指點我正確的方向，使我從中獲益匪淺。老師對學問的認真與嚴謹是我學習的典範，在生活上的關心與鼓勵更讓我倍感窩心。感謝論文口試委林培生老師、張玉媚老師對於論文的指導及許多的建議，讓此篇論文更臻完備，於此獻上最深的敬意與謝意。

在研究所修業期間，感謝系上老師在課業知識上的傳授，感謝系辦的學長姐們，給予我選課上的建議以及行政上的諸多幫助，還有感謝兩年來互相討論與鼓勵的同學們，於研究期間得學習過程更加充實。維芬、阿莓、學長、感謝你們的貼心與陪伴並且給予我信心，讓兩年的研究生活變得絢麗多彩。

最後，將本文獻給我最摯愛的父母與家人，感謝你們無時無刻的關懷呵護與包容，讓我能全力專注於課業研究中，願以此與家人共享。最後，對於所有幫助過我、關懷過我的人，在此謹深致謝忱。

邱煒婷 謹誌於東海大學統計研究所

中華民國九十九年六月

ABSTRACT

In statistical analysis, the data are not usually drew from the same population. This paper is concerned with the problem of regression model selection criteria for the data with correlated data. There are several criteria being widely used in the analysis, such as AIC, T, FPE, nS, and U. We try to conduct some regression model selection simulations to see how the outliers impact on the outcomes. We use the concept of robust methods to improve the performance of these model selection criteria when the data contain outliers. The class of robust section criteria incorporate with the weight functions proposed by Huber, Tukey.

Keywords: Model selection, Robust regression, Generalized cross-validation, Leverage points, Sigma square

摘要

在統計分析中，常常實際的資料並不會都從同一個母體蒐集而來。此篇論文是在關注相關性資料下的迴歸模式選取的準則。有些準則是廣泛的被使用的，例如 AIC，T，FPE，nS，和 U 等。

我們嘗試建造一些模擬，從模式選取中去了解異常點造成的影響。我們利用穩健的方法去解決在模式選取中較差的表現結果。這一系列的穩健準則包含 Huber, Tukey 的方法。在此篇研究中，我們參考 Huber 和 Tukey 的方法去降低 sigma square 和 leverage 值的權重。經過模擬的研究中，我們發現模式選取有較好的表現。

關鍵字: 模式選取，穩健迴歸，廣義交叉驗證，影響點，標準平方差

CONTENTS

Abstract……..……ii

1. Introduction……..……. 1

1.1 Motivation of Research. …….……1

1.2 Purpose of Research.……...…….…….....2

1.3 Overview . . . . . .……...3

2. Literature Review. …. . ……. ……...……...4

3. Robust Model Selection Methods…….....……. …...…….....6

3.1 Linear Regression…. ……. ……..……. ……6

3.1.1 Linear Regression……. ……..……. ……6

3.1.2 Regression Diagnostics….……...……. ……8

3.1.3 Robust Regression.……..……...…….8

3.2 Robust Model Selection…. …….....…. …...……12

3.2.1 Regression Model Selection.….....…. ……12

3.2.2 Robust Regression Model Selection..…….....…..…13

4. Simulations……...……15

4.1 Weighted Model Selection.……15

4.2 Robust Model Selection.……21

5. Concluding Discussion.……27

References…….....……....28

LIST OF TABLES

Table 4.1 Model selection numbers of true model in 1000 replication.……..…17

Table 4.2 Model selection numbers of model 1 in 1000 replication..…….……..…17

Table 4.3 Model selection numbers of model 2 in 1000 replication.…….……17

Table 4.4 Model selection numbers of model 3 in 1000 replication.……18

Table 4.5 Model selection numbers of model 4 in 1000 replication.……18

Table 4.6 Model selection numbers of model 5 in 1000 replicatio……18

Table 4.7 Model selection numbers of model 6 in 1000 replication.……19

Table 4.8 Model selection numbers of model 7 in 1000 replication.…….....…19

Table 4.9 Model selection in model 4 using Huber's method with c=2…..……21

Table 4.10 Model selection in model 4 using Huber's method with c=3……21

Table 4.11 Model selection in model 4 using Tukey's method with c=3……21

Table 4.12 Model selection in model 5 using Huber's method with c=2……...…22

Table 4.13 Model selection in model 5 using Huber's method with c=3……...…22

Table 4.14 Model selection in model 5 using Tukey's method with c=3……...…22

Table 4.15 Model selection in model 6 using Huber's method with c=2……...…23

Table 4.16 Model selection in model 6 using Huber's method with c=3……...…23

Table 4.17 Model selection in model 6 using Tukey's method with c=3……...…24

Table 4.18 Model selection in model 7 using Huber's method with c=2……...…24

Table 4.19 Model selection in model 7 using Huber's method with c=3……...…24

Table 4.20 Model selection in model 7 using Tukey's method with c=3……...…..25

1. INTRODUCTION

This chapter is divided into three parts including motivation of research, purposes of research, and the overview.

1.1 Motivation of Research

Unusual data are problematic in a least squares regression because they can unduly influence the results of the analysis. Their presence may be a signal that the regression model fails to capture important characteristics of the data.

In the statistical analysis, the data are often not drew from the same distribution. The paper is concerned with the problem of regression model selection criteria for the data with correlated errors. A class of regression model selection criteria has been proposed by Wei (2009). There are several criteria being widely used in the analysis, such as GCV, AIC, T, FPE, nS, and U. We try to conduct some regression model selection simulations to see how the outliers impact on the outcomes. For example, shifting the intercept, or changing in the coefficients of explanatory valuables, or changing the error terms, or adding other explanatory variables.

In robust statistics, robust regression is a form of regression analysis designed to circumvent some limitations of traditional parametric and non-parametric methods. In particular, least squares estimates for regression models are highly non-robust to outliers. We use the robust methods when the data contain outliers. In the presence of

outliers, the ordinary least squares estimation is inefficient and can be not unbiased. In some cases, there is only one outlier. The outlier can be detected easily by Cook's Distance. In some complex cases, there might be multiple outliers in the data. These outliers can not be detected by the residuals plot. It is called masking effect.

Discarding of outlying influential cases that are not clearly erroneous and that cannot be accounted for by model improvements should be done only rarely, such as when the model is not intended to cover the special circumstances related to the outlying cases (see Kutner, Nachtsheim, and Neter 2008, p.437).

1.2 Purpose of Research

The purpose of this research is therefore twofold. The main purpose of the paper is to develop methods that could reduce the influence of potential outliers during the model selection in regression. Another purpose of this paper is to evaluate various selection criteria.

Detecting the outliers by residual plots and leverage values is important. We try to downweight the leverage points. A number of research works have been done in this filed to seek for a function of M-Estimation. The class of robust methods includes the methods of Huber, Tukey, Hampel, Andrew, and Ramsay. We use the methods of Huber and Tukey. Outlying cases that have large residuals or high leverage values are thereby given smaller weights.

Several studies have suggested the benefit of robust regression. The readers are referred to Wei (2007) for an introduction to the above approaches and their applications.

1.3 Overview

In next section, we give the literature review including some extended researches of them. The methods for both regression models and nonparametric models are presented in section 3. Also we discuss the ideas of robust regression and six kinds of model selection criteria. In addition, we also propose our robust model selection criteria. In section 4, we introduce the data sets for simulations and also show some different outcomes. We write our own S-PLUS and R code to implement these methods. Finally, a concluding discussion is given in section 5. We discuss some problems of these methods and directions of further researches.

2. LITERATURE REVIEW

The related work reported in the literature can be classified into two major categories. The first part is about the regression model selection criteria. Another part is about the robust regression.

Nonparametric modeling and semi-parametric modeling have been broadly used techniques in modern years as indicated by Härdle et al. (2004). Nonparametric regression techniques are often sensitive to the presence of correlated errors. The breakdown of several popular data-driven smoothing parameter selection methods was indicated by Opsomer et al. (2001) in nonparametric regression. The practical consequences of this sensitivity are explained in this paper. It discussed that extension to random design, higher dimensional models and adaptive estimation.

Craven and Wahba (1979) proposed the commonly used GCV (generalized cross validation) criterion. It is nearly an unbiased estimator of the prediction risk in some cases. O'Sullivan (1986) has used the methodology adding to the battery of graphical tools for model building and checking within the generalized linear model structure. In the study, he proposed a more general multivariate smoothing spline-type estimator and developed an explicit cross-validation score to assess the appropriately correct degree of smoothing. It is particularly suited to the analysis of larger data set (which n>50 data points). Shibata (1981) proposed the new method of asymptotically optimal selections of regression variables. The method is called nS. It has been also shown that Mallow's Cp, Akaike's FPE (see Akaike 1970), and Akaike's AIC methods (see

Akaike 1974) are all asymptotically equivalent to this method. Rice (1984) proposed the method which is called T. The paper in Rice (1984) was concerned with the problem choosing a bandwidth parameter for nonparametric regression. It is shown that the bandwidth thus chosen is asymptotically optimal. Wei (2009) introduced a class of regression model selection criteria for the data with correlated errors, including GCV (Craven and Wahba 1979), AIC (Akaike 1974), T (Rice 1984), FPE (Akaike 1970), nS (Shibata 1981), and U (Hocking 1976). A simulation study was conducted to illustrate that all the selection criteria given in those six methods perform well. Furthermore, the numerical results based on the simulated data are quite coherent with the theoretical ones, even in the case of medium-sample. The papers also provided extensive discussions of the applications of weighted predictive mean square error and properties of weighted selection criteria.

3. ROBUST MODEL SELECTION METHODS

3.1 Linear Regression

3.1.1 Linear Regression

Regression analysis is widely used technique for fitting the data. Linear regression refers to any approach to modeling the relationship between one or more variables denoted *y* and one or more variables denoted *X*, such that the model depends linearly on the unknown parameters to be estimated from the data in statistics. Such a model is called a linear model.

Regression model fitting has several implicit assumptions, including the following:

1. The uncorrelated model errors have zero mean and constant variance.
2. The model errors have a normal distribution and are independent.
3. The form of the model, including the specification of the regressors, is correct.

Thus the model takes the form

$$y = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e \ , where \ i = 1, 2, ..., n ,$$

where $e \overset{iid}{\sim} Normal(0, s^2)$.

In matrix notation, the model is

$$y = X b + e$$

where $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$, $e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$,

where y called the dependent variable is an n by 1 vector of observations, and X is an n by p design matrix of the levels of the regressor variables, b is a p by 1 vector of the regression coefficients, and e is an n by 1 vector of random errors. An important objective of regression analysis is to estimate the unknown parameters b_i in the regression model.

Ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model. It minimizes the sum of squared distances between the observed responses in a set of data, and fitted responses from the regression model. The linear least squares computational technique provides simple expressions for the estimated parameters in an OLS analysis, and hence for associated statistical values such as the standard errors of the parameters. The least-squares normal equations are $X'X\hat{b} = X'y$. The value of b which minimizes the sum of squared residuals *S(b)* is called the least squares estimator for *ß*, and the estimator of *ß* is $\hat{b}$, where $\hat{b} = \left(X'X\right)^{-1}X'y$. The $(X'X)^{-1}$ matrix will always exist if the rank of $X'X$ matrix are full rank. The diagonal elements of $X'X$ are the sums of squares of the elements in the columns of X, and the off-diagonal elements are the sums of cross products of the elements in the columns of X. In addition, note that the elements of $X'y$ are the sums of cross products of the columns of X and the observations y_i.

The vector of fitted values $\hat{y}$ corresponding to the observed values y_i is

$\hat{y} = X\hat{b} = X(X'X)^{-1}Xy = Hy$, where $H = X(X'X)^{-1}X'$. Here the n by n matrix H is called the hat matrix. It is symmetric and idempotent. Consequently, the diagonal elements of the hat matrix $h_i = h_{ii}$, are called hat-values. Large hat-values reveal observations that are potentially influential. It turns out that the average of a hat diagonal is p/n. We traditionally assume that any observation with associated values of n_i exceeding the average 2p/n are considered as a leverage points. Observations with large values of h_i and large residuals are likely to be influential. Belsley et al. (1980) suggest that values of h_i exceeding about twice the average p/n are noteworthy.

3.1.2 Regression Diagnostics

Traditional way for regression diagnostics is to check each observation out of the data in turn. Cook's Distance is a measure of how much the estimate changes as each observation is dropped. Cook's Distance is used to measure influence , the extent to which an observation is affecting the location of the regression surface. Cook's distance depends on both its distance and its leverage. Cook suggested that one check observations whose Cook's distance is greater than the median value of *F* statistic with *p* and *n-p* degrees of freedom. Howell (2007, p. 518) suggested investigating any *D* > 1.00 may be influential point. Studentized residuals, DFBETAS, are other regression diagnostics. All of these diagnostics can be performed graphically using the function plot.

3.1.3 Robust Regression

In the presence of outliers, the ordinary least squares estimation is inefficient and can be not unbiased. Robust regression methods are more insensitive to these outliers than the ordinary least squares method. Robust methods downweight the unusual observations. Essentially, observations that produce large residuals are down-weighted by a robust estimation method.

Robust statistic are resistant to errors in the results, produced by deviations from assumptions (e.g. of normality). This means that if the assumptions are only approximately met, the robust estimator will still have a reasonable efficiency, and reasonably small bias, as well as being asymptotically unbiased, meaning having a bias tending towards 0 as the sample size tends towards infinity. Robust estimators and diagnostics for linear regression models have been widely discussed in the literature. The motivation for much of the work in robust regression was the Princeton robustness study (see Andrews et al. (1973)). There have been several types of robust estimators proposed. Some important basic references include Andrews (1974), Carroll and Rupert (1988), Hogg (1974), Huber (1972), Krasker and Welsch (1982), Rousseeuw (1984), and Rousseeuw and Leroy (1987).

Swamping (inliers appear as outlying) and masking (outliers appear as inliers) effects due to multiple outliers can be revealed and avoided by robust diagnostics. Robust regression diagnostics is discussed for the case when the linear regression model contains outliers. Simulated data and real data analyses illustrate the performance of the resulting approaches.

M-estimators are “maximum likelihood type” estimators. Linear least-squares estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove influential observations from the least-squares fit. Another approach, termed robust regression, is to employ a fitting criterion that is not as vulnerable as least squares to unusual data. Suppose $f(m)$ is a defined function of m , and suppose s is an estimate of s and not necessarily the usual least squares estimate. The M-estimator for b , based on the function $f(m)$ and where $y(m)$ is the partial derivative $\frac{\partial f}{\partial m}$. We define a robust estimator as one that minimizes $\sum_{i=1}^{n} f(\frac{e_i}{s}) = \sum_{i=1}^{n} f\left\{\frac{Y_i - x_i^{'} b}{s}\right\}$. Let the weights $w_{ib} = \frac{y\{(Y - x_i^{'} b)/s\}}{(Y - x_i^{'} b)/s}$. The least squares method gives the highest weight of 1. Some of the suggestions given in the literature for $f(m)$. The purpose of the various $f(m)$ functions is to comparatively down-weight larger residuals in various ways. There are four criteria which we chose including Huber’s method Tukey’s method, and the least squares method. The criterion of $f(m)$ are as follows:

(1) Huber’s with breakpoint $a > 0$

$$f(m) = \begin{cases} \frac{1}{2} m^2, -a \le m \le a \\ a|m| \quad , m \le -a \ \ and \ \ a \le m \end{cases}$$

It increases linearly at a given level $|m| > c$. The 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant *c = 1.345*.

(2) Tukey’s biweight, with breakpoint $a > 0$

$$f(m)=\begin{cases}\frac{1}{2}m^2-\frac{u^4}{4a^2}, -a\le m\le a\\ \frac{1}{4}a^2, m\le -a \ \ and \ \ a\le m\end{cases}$$

(3) Andrew's wave, with breakpoint $ap>0$

$$f(m)=\begin{cases}a\{1-\cos(m/a)\}, \ -a\le m\le a\\ 2a, \ m\le -a \ and \ a\le m\end{cases}$$

(4) The least square method

$$f(m)=\frac{1}{2}m^2, \ -\infty\le m\le\infty$$

3.2 Robust Model Selection

3.2.1 Regression Model Selection

The prediction risk can be used as a criterion for selecting a reasonable number of covariates. For example, the usually used GCV (generalized cross validation) criterion (Craven and Wahba 1979),

$$GCV(l)=\frac{\hat{s}^2(l)}{[1-m_1(l)]^2}$$

is almost an unbiased estimator of the prediction risk, , where the parameter l could be the subset of the discrete index set $\{1,2,\ldots,p\}$, and where

$$\hat{s}^2(l)=y'[I-H(l)]'[I-H(l)]y/n \ ,$$

$[1-m_1(l)]^2$ is a penalty function, $m_1(l)=Tr[H(l)]/n$, and $Tr[H(l)]$ is the trace of the hat matrix $H(l)$ and $H(\{1,2,\ldots,p\})$ is the hat matrix H of the Section 3.1. In addition to GCV, the other one criteria have the form $\frac{\hat{s}^2(l)}{f[m_1(l)]}$, where $f(\cdot)$ is a penalty function. We usually used selection criteria can be obtained by different choices of f, including

1. GCV (Craven and Wahba 1979): $f(m_1)=(1-m_1)^2$,

2. AIC (Akaike 1974): $f(m_1)=\exp(-2m_1)$,

3. T (Rice 1984): $f(m_1)=1-2m_1$,

4. FPE (Akaike 1970): $f(m_1)=\frac{1-m_1}{1+m_1}$,

5. nS(b) (Shibata 1981): $f(m_1)=\frac{1}{1+2-m_1}$,

6. U(b) (Hocking 1976): $f(m_1)=\frac{(1-m_1)(n-1-nm_1)}{(n-1)}$,

(see Eubank 1988, pp. 38–40).

As $Var(e)=s^2V^{-1}(a)$, the vector of fitted values is $\hat{f}_v=H_v(h)y$, where $a=(a_1,a_2,\ldots,a_m)$ is a set of correlation parameters and. $h=(a,l)$.

3.2.2 Robust Regression Model Selection

The class of weighted selection criteria is $\frac{\hat{s}_v^2(h)}{f[m_{1v}(h)]}$, where $m_{1v}(h)=\frac{Tr[H(h)]}{y[V(a)]}$ is a bounded function and

$$\hat{s}_v^2(h)=\frac{y^t\left[I-H_v(h)\right]^t V(a)\left[I-H_v(h)\right]y}{y[V(a)]}.$$

We define that $V(a)^{\frac{1}{2}}\left[I-H_v(h)\right]y=V(a)^{\frac{1}{2}}\left[y-\hat{y}(h)\right]=e^{*}(h)$

$$\Rightarrow\left[e^{*}(h)\right]^{t}\left[e^{*}(h)\right]=\sum_{i=1}^{n}(e_i^{*}(h))^2$$

$$\Rightarrow \hat{s}_v^{\,2}(h)=\frac{\sum_{i=1}^{n}(e_i^{*}(h))^2}{y\,[V(a)]}$$

First we downweight the criterion of $\sum_{i=1}^{n}(e_i^{*}(h))^2$ .There are three downweight criteria:

(1) Downweight Huber's with breakpoint

$$(e_i^{*}(h))^2=\begin{cases}(e_i^{*}(h))^2, & -c\le e_i^{*}\le c\\ c\left|e_i^{*}(h)\right|-\frac{1}{2}c^2, & e_i^{*}\le -c \text{ and } c\le e_i^{*}\end{cases}$$, where $c=2$, for i=1,2,3...,n

(2) Downweight Huber's with breakpoint

$$(e_i^{*}(h))^2=\begin{cases}(e_i^{*}(h))^2, & -c\le e_i^{*}\le c\\ c\left|e_i^{*}(h)\right|-\frac{1}{2}c^2, & e_i^{*}\le -c \text{ and } c\le e_i^{*}\end{cases}$$, where $c=3$, for i=1,2,3...,n

(3) Downweight Tukey's with breakpoint

$$(e_i^{*}(h))^2=\begin{cases}\frac{1}{2}(e_i^{*}(h))^2-\frac{c^4}{4c^2}, & -c\le e_i^{*}\le c\\ \frac{1}{4}c^2, & e_i^{*}\le -c \text{ and } c\le e_i^{*}\end{cases}$$, where $c=2$, for i=1,2,3...,n

Secondly we downweight the criterion of leverage points.

$$hi^{*} = \begin{cases} hi \ , hi^{*} \leq \frac{2p}{n} \\ \\ \frac{2p}{n} \ , \frac{2p}{n} \leq hi^{*} \end{cases}$$, for i=1,2,3...,n. Where p is columns number of X matrix, n is row number of X matrix. If the leverage points over than 2p/n, we can downweight the leverage points. Let the new leverage points be 2p/n.

4. SIMULATIONS

4.1 Weighted Model Selection

In this simulation, the values of four variables, X_1, X_2, X_3, X_4 were in $U[0,1]$ and 110 observations were generated from the model

$$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i \text{ , } i = 1, \ldots, 110,$$

where e_i are zero mean random errors. The errors were generated from Gaussian AR(1) processes with the standard deviations of uncorrelated Gaussian errors, s_g , equal to 2. The AR(1) model is $y_i = u + py_{i-1} + e_i$,where u is a constant, e_i is a white noise process with zero mean and variance s_g^2. The autocorrelation values at lag 1, for the Gaussian AR(1) processes were p=0.8. Assume that $V(a) = V$ is known. We choose the sample size and retained number of principal components of the matrix V by including enough components to explain 90% amount of the variance for $y(V)$. We generated the random errors 1000 replicates. These selection criteria produce very similar results.

There were 8 models including model for generating the normal data and model 1 to model 7 for generating the data with 10 discordant observations.

(0) True model

The true model is $y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, $i = 1, \ldots, 110$,

where e_i are zero mean random errors.

(1)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, $i = 1, \ldots, 100$,

$y_i = 5 + 2x_{i1} + 4x_{i3} + e_i$, $i = 101, \ldots, 110$,

where e_i are zero mean random errors.

(2)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100, s_g equal to 2

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 101, . . . ,110, s_g equal to 0.2

where e_i are zero mean random errors.

(3)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100,

$y_i = 1 + 3x_{i1} + 5x_{i3} + e_i$, i = 101, . . . ,110,

where e_i are zero mean random errors.

(4)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100, p=0.8

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 101, . . . ,110, p=-0.8

where e_i are zero mean random errors.

(5)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100,

$y_i = 1 + 2x_{i1} + 3x_{i2} + 4x_{i3} + e_i$, i = 101, . . . ,110,

where e_i are zero mean random errors.

(6)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100,

$y_i = 1 + 2x_{i1} + 4x_{i3} + 5x_{i4} + e_i$, i = 101, . . . ,110,

where e_i are zero mean random errors.

(7)

$y_i = 1 + 2x_{i1} + 4x_{i3} + e_i$, i = 1, . . . ,100,

$y_i = 1 + 2x_{i1} + 3x_{i2} + 4x_{i3} + 5x_{i4} + e_i$, i = 101, . . . ,110,

where e_i are zero mean random errors.

In the models, we ***choose the minimum of values in the same criteria and count the numbers in 1000 replication.***

Table 4.1 ***Model selection numbers of true model in 1000 replication***

Model selection numbers of true model in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	860	71	67	2
AIC	845	75	77	3
Rice's T	878	60	61	1
FPE	848	75	75	2
nS	813	90	89	8
U	861	70	67	2

Table 4.2 ***Model selection numbers of model 1 in 1000 replication***

Model selection numbers of model 1 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	948	22	29	1
AIC	937	25	37	1
Rice's T	954	20	26	0
FPE	938	24	37	1
nS	919	33	45	3
U	949	22	28	1

Table 4.3 ***Model selection numbers of model 2 in 1000 replication***

Model selection numbers of model 2 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	872	64	61	3
AIC	849	73	74	4

Rice's T	891	58	50	1
FPE	850	73	73	4
nS	814	86	93	7
U	874	62	61	3

Table 4.4 ***Model selection numbers of model 3 in 1000 replication***

Model selection numbers of model 3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	841	70	76	13
AIC	830	75	80	15
Rice's T	858	63	70	9
FPE	830	75	80	15
nS	789	93	101	17
U	843	69	75	13

Table 4.5 ***Model selection numbers of model 4 in 1000 replication***

Model selection numbers of model 4 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	795	97	98	10
AIC	782	101	106	11
Rice's T	808	90	94	8
FPE	782	101	106	11
nS	743	123	120	14
U	796	97	98	9

Table 4.6 ***Model selection numbers of model 5 in 1000 replicatio***

Model selection numbers of model 5 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	269	634	22	75

AIC	257	636	24	83
Rice's T	285	634	21	60
FPE	257	636	24	83
nS	228	646	26	100
U	273	631	22	74

Table 4.7 ***Model selection numbers of model 6 in 1000 replication***

Model selection numbers of model 6 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	185	13	727	75
AIC	173	14	729	83
Rice's T	193	13	736	58
FPE	175	14	728	83
nS	154	13	733	100
U	185	13	728	74

Table 4.8 ***Model selection numbers of model 7 in 1000 replication***

Model selection numbers of model 7 in 1000 replication				
criteria\model	**True model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	309	109	433	149
AIC	286	109	439	166
Rice's T	339	106	425	130
FPE	287	109	438	166
nS	246	105	436	213
U	313	106	433	148

In these simulation, we select the 10 percent data from different populations. In the Table 4.1, all criteria attain 80 percent of selecting the correct model. Not only are there also above 90 percent of selecting the correct model in the Table 4.2, but also are there also above 80 percent of selecting the correct model in the Table 4.3.

Changing the intercept in model 1, the proportions of selecting the correct model increase 10 percent. Various standard deviations of uncorrelated Gaussian errors in model 2, it is a little decrease in selecting the correct model. We can realize that changing the intercept and various standard deviations of uncorrelated Gaussian errors are not lead to selecting the wrong model in model selections. In the Table 4.4, changing the slopes results in the decreased proportions of selecting the correct model. It is similar to the Table 4.1. In the Table 4.5, we can see that decreased proportions of selecting the correct model occur as the discordant data were generated by different Gaussian processes. However, the decreases are not significant. There are also above 80 percent proportion of selecting the correct model.

We may selecte the wrong model frequently by adding a variable X_2. In the Table 4.6, the wrong model which includes X_0, X_1, X_2, X_3 variables is selected more often than the other models. It is surprising that we selecte the wrong models over 63 percent. The proportions of selecting the correct models are very small. In the Table 4.7, the wrong model which includes X_0, X_1, X_3, X_4 variables is selected more often than the other models. It is surprising that the proportion of selecting the wrong model is over 70 percent. The proportion of selecting the correct model is less than 20 percent. In the Table 4.8, the wrong model which includes X_0, X_1, X_3, X_4 variables is selected more often than the other models. The proportion of selecting the wrong models is 43 percent

4.2 Robust Model Selection

Table 4.9 ***Model selection in model 4 using Huber's method with c=2***

Model selection numbers of model 4 using Huber's method with c=2 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	802	86	107	5
AIC	789	92	113	9
Rice's T	821	82	82	3
FPE	786	93	113	8
nS	745	103	135	17
U	803	86	106	5

Table 4.10 ***Model selection in model 4 using Huber's method with c=3***

Model selection numbers of model 4 using Huber's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	817	88	90	5
AIC	805	91	99	5
Rice's T	831	81	84	4
FPE	805	91	99	5
nS	778	103	108	11
U	817	88	90	5

Table 4.11 ***Model selection in model 4 using Tukey's method with c=3***

Model selection numbers of model 4 using Tukey's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	926	33	40	1
AIC	915	38	46	1
Rice's T	934	27	38	1

FPE	915	38	46	1
nS	897	47	53	3
U	927	32	40	1

Table 4.12 ***Model selection in model 5 using Huber's method with c=2***

Model selection numbers of model 5 using Huber's method with c=2 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	913	39	46	2
AIC	899	46	52	3
Rice's T	925	32	41	2
FPE	899	46	52	3
nS	881	54	59	6
U	915	39	44	2

Table 4.13 ***Model selection in model 5 using Huber's method with c=3***

Model selection numbers of model 5 using Huber's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	951	22	27	0
AIC	939	26	33	2
Rice's T	955	20	25	0
FPE	940	26	32	2
nS	924	34	40	2
U	951	22	27	0

Table 4.14 ***Model selection in model 5 using Tukey's method with c=3***

Model selection numbers of model 5 using Tukey's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**

GCV	983	2	15	0
AIC	981	2	17	0
Rice's T	989	2	9	0
FPE	982	2	16	0
nS	977	2	21	0
U	983	2	15	0

Table 4.15 ***Model selection in model 6 using Huber's method with c=2***

Model selection numbers of model 6 using Huber's method with c=2 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	972	18	10	0
AIC	972	18	10	0
Rice's T	975	15	10	0
FPE	972	18	10	0
nS	962	25	13	0
U	973	17	10	0

Table 4.16 ***Model selection in model 6 using Huber's method with c=3***

Model selection numbers of model 6 using Huber's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	988	9	3	0
AIC	984	12	4	0
Rice's T	994	4	2	0
FPE	984	12	4	0
nS	978	15	6	1
U	988	9	3	0

Table 4.17 ***Model selection in model 6 using Tukey's method with c=3***

Model selection numbers of model 6 using Tukey's method with c=3 in 1000 replication

criteria\model	Correct model (including variables) (X0,X1,X3)	(including variables) (X0,X1,X2,X3)	(including variables) (X0,X1,X3,X4)	(including variables) (X0,X1,X2,X3,X4)
GCV	995	5	0	0
AIC	994	6	0	0
Rice's T	996	4	0	0
FPE	994	6	0	0
nS	989	11	0	0
U	995	5	0	0

Table 4.18 ***Model selection in model 7 using Huber's method with c=2***

Model selection numbers of model 7 using Huber's method with c=2 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	987	6	7	0
AIC	984	7	9	0
Rice's T	990	5	5	0
FPE	984	7	9	0
nS	978	12	10	0
U	987	6	7	0

Table 4.19 ***Model selection in model 7 using Huber's method with c=3***

Model selection numbers of model 7 using Huber's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	993	3	4	0
AIC	991	4	5	0
Rice's T	997	2	1	0
FPE	991	4	5	0
nS	989	6	5	0
U	993	3	4	0

Table 4.20 ***Model selection in model 7 using Tukey's method with c=3***

Model selection numbers of model 7 using Tukey's method with c=3 in 1000 replication				
criteria\model	**Correct model (including variables) (X0,X1,X3)**	**(including variables) (X0,X1,X2,X3)**	**(including variables) (X0,X1,X3,X4)**	**(including variables) (X0,X1,X2,X3,X4)**
GCV	996	3	1	0
AIC	996	3	1	0
Rice's T	997	3	0	0
FPE	996	3	1	0
nS	995	3	2	0
U	996	3	1	0

In the Table 4.9, we use Huber's method with c=2 in model 4 and the proportions of selecting the correct model is about 79 percent. In the Table 4.10, we use Huber's method with c=3 and the proportions of selecting the correct model is about 81 percent. In Table 4.11, the proportions of selecting the correct model is about 92 percent as using Tukey's method.

In the Table 4.12, we use Huber's method with c=2 in model 5 and the proportions of selecting the correct model is about 91 percent. In Table 4.13, we use Huber's method with c=3 and the proportions of selecting the correct model is about 94 percent. In Table 4.14, the proportions of selecting the correct model is about 98 percent as using Tukey's method.

In the Table 4.15, we use Huber's method with c=2 in model 6 and the proportions of selecting the correct model is about 97 percent. In the Table 4.16, we use Huber's method with c=3 and the proportions of selecting the correct model is about 98 percent. In Table 4.17, the proportions of selecting the correct model is about

99 percent as using Tukey's method.

In the Table 4.18, we use Huber's method with c=2 in model 7 and the proportions of selecting the correct model is about 98.6 percent. In the Table 4.19, we use Huber's method with c=3 and the proportions of selecting the correct model is about 99.2 percent. In Table 4.11, the proportions of selecting the correct model is about 99.6 percent as using Tukey's method.

Overall, even we down-weight in model 4, model 5, model 6, and model 7, robust model selection has all good performance. Using Huber's method with c=2 has large proportions of selecting the correct model than using Huber's method with c=3.

In addition, we can see that using Tukey's method to downweight is better than Huber's method to down-weight. Even we down-weight in using Huber's method or in Tukey's method , the model selection has all good performance.

5. CONCLUDING DISCUSSION

All criteria, including GCV, AIC, Rice's T, FPE, nS, and U, have similar optimal properties. Four of these findings are worth summarizing. The first one is that we may choose the wrong model by the data draw from different populations. The second one is that these model selection criteria might select the wrong model as the data are from different populations. The proportions of selecting wrong models are high as the discordant observations are generated from the models with additional variables or less variables. The third one is that the proportions of selection correct models are increasing as these discordant observations are downweighted. The last one is that the Huber's and Tukey's methods are both effective.

In my simulation, the larger c in Huber's method, the bigger proportion in choosing correct models. We down-weight in model 4, model 5, model 6, and model 7, robust model selection has all good performance. Using Huber's method with c=2 has large proportions of selecting the correct model than using Huber's method with c=3. Not only Huber's method with c=2 but also Huber's method with c=3, the proportions of selecting the correct model is over 90 pencent.

The proportions of selecting correct models by Tukey's method are about 99 percent.

There is still space for future research in regression model selection with correlated errors. The proposed criteria have different sensitivities to the changes of the selection parameters or correlation parameters. Sensitivity analysis and a complete

robust for different selection criteria could be helpful. The study does suggest that the detailed study of using robust in nonparametric regression in correlated errors is a promising line of inquiry.

REFERENCES

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics, 22,* **203–217**.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* **716–723**.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics, 16,* **523-531**.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location*. Princeton University Press, Princeton, N.J.

Belsley, D. A., kuh, E., and Welsh, R.E (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* New York: John Wiley.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik, 31,* **377–403**.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. Marcel Dekker, New York.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32,* **1–49**.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Associatio., 69,* **909-925**.

Howell, D. C. (2007). *Statistical Methods for Psychology (6th Ed.)*. Thomson Wadsworth. Belmont, CA.

Huber, P. J. (1972). Robust statistics: A review. *The Annals of Mathematical Statistics.*, *43*, **1041-1067**.

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

Kutner, M. H., Nachtsheim, C., Neter, J. (2008). *Applied Linear Regression Models.* McGraw-Hill, New York.

Opsomer, J.,Wang,Y.,Yang,Y. (2001). Nonparametric regression with correlated errors. *Statistical Science, 16*, **134–153**.

O'Sullivan, F., Yandell, B., Raynor,W. (1986). Automatic smoothing of regression function in generalized linear models. *Journal of the American Statistical Association, 81*, **96–103**.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics, 12*, **1215–1230**.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*, **871-880**.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*, **45–54**.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association, 93*, **341–348**.

Wei, W. H. (2009). On regression model selection for the data with correlated errors. *Annals of the Institute of Statistical Mathematics, 61*, **291–308**.