Annexure – VIII

UNIVERSITY GRANTS COMMISSION

BAHADUR SHAH ZAFAR MARG

NEW DELHI – 110 002

PROFORMA FOR SUBMISSION OF INFORMATION AT THE TIME OF SENDING THE FINAL REPORT OF THE WORK DONE ON THE PROJECT

---

1. **NAME AND ADDRESS OF THE PRINCIPAL INVESTIGATOR**
   Dr. D. N. Kashid,
   Department of Statistics, Shivaji University,
   Kolhapur – 416004. Maharashtra (INDIA).

2. **NAME AND ADDRESS OF THE INSTITUTION**
   Department of Statistics, Shivaji University, Vidyanagar,
   Kolhapur – 416004. Maharashtra (INDIA).

3. **UGC APPROVAL NO. AND DATE**
   F.No. 34-154\2008 (SR) (UGC) dated 30-12-2008.

4. **DATE OF IMPLEMENTATION**
   1 February 2009

5. **TENURE OF THE PROJECT**
   3 Years

6. **TOTAL GRANT ALLOCATED**
   8, 16, 000/-

7. **TOTAL GRANT RECEIVED**
   5, 46, 000/-

8. **FINAL EXPENDITURE**
   6, 41, 504/-

9. **TITLE OF THE PROJECT**
   A Study of Subset Selection Techniques in Regression and Applications.

10. **OBJECTIVES OF THE PROJECT**
    In this study we focused on the following objectives:

    - to take review of various subset selection techniques in normal and non-normal setup
    - to compare performance of existing subset selection methods like Mallow's C_p-statistic in non-normal setup

BARR. BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

- to explore possibilities of providing new / modified subset selection technique(s)
- to develop relevant computer programs.

11. WHETHER OBJECTIVES WERE ACHIEVED (GIVE DETAILS)

Objectives given in the proposal are achieved at our level best.

1. We have reviewed various subset selection methods when error variable follows normal and non-normal distribution. Also, subset selection techniques in the presence of multicollinearity and outliers are reviewed.
2. We studied the performance of subset selection method based on the least squares estimator like Mallows C_p in various situations. It is reported in the 1st chapter of the report.
3. We proposed the new estimator for regression coefficient with multicollinearity and outlier in Y-space.
4. We developed a subset selection technique based on the proposed estimator and compared with Mallow's C_p-statistic, R_p statistic and S_p statistic.
5. We developed a subset selection criterion using rank based parameter estimation for non-normal error variable and compared with N_p criterion.
6. A subset selection criterion in Poisson regression is proposed to tackle the problem of count data when response follows Poisson distribution.
7. Developed relevant computer codes/ programs for implementing the new criteria and for comparison of new criteria with existing methods. These codes are enclosed in the appendix of the work done report.

12. ACHIEVEMENTS FROM THE PROJECT

Two papers are published in international journal and two papers are communicated for possible publication.

A. Published Papers:

1. Sakate, D., M., Kashid, D., N. (2011). Subset Selection in Poisson Regression. Journal of Statistical Theory and Practice, 5 (2): 207-219.
2. Jadhav, N.,H., Kashid, D.,N. (2011). A Jackknifed Ridge M-estimator for Regression Model with Multicollinearity and Outliers. Journal of Statistical Theory and Practice, 5 (4): 659-673.

B. Communicated to Journal for Possible Publications:

1. Jadhav, N., H., Kashid, D., N. Subset Selection in Multiple Linear Regression for Non-normal Symmetric Error Terms. (Communicated to Journal of Statistical Computation and Simulation).
2. Jadhav, N., H., Kashid, D., N. Subset Selection in Multiple Linear Regression in Presence of Outlier and Multicollinearity. (Communicated to Communications in Statistics – Theory and Methods)

13. SUMMARY OF THE FINDINGS (IN 500 WORDS)

In this report, the subset selection problem in multiple linear regression is addressed where the objective is to select the minimum number of regressors which contribute significantly to the response variable and predict response variable as accurate as full model. In classical regression, various subset selection methods are available. The violation of some or more assumptions of regression model affects the existing subset selection procedures. Hence, there would be a scope to modify or develop new methods which deal with the problem of such violations.

The problem of multicollinearity, outlier in Y-space and non-normal distribution of error variable are considered in this report. How well known subset selection method (Mallow's C_p) fails to select proper subset is illustrated with suitable example in Chapter 1.

We consider the problem of multicollinearity and outliers in response variable in Chapter 2. In this context, the new estimator is proposed. An article "A Jackknifed Ridge M-estimator for Regression Model with Multicollinearity and Outliers" is prepared and presented in the National Conference at Bangalore. This article is published in Journal of Statistical Theory and Practice. We thank to Dr. S. R. Kulkarni, Innovative Labs, Bangalore for fruitful discussion that we had in Kolhapur during his visit.

The aim of this Major Research Project is to device a new variable selection technique when the data suffers from multicollinearity problem, outlier problem and non-normal distribution of error variable. The new subset selection method is developed and it is based on above proposed estimator. This subset selection criterion is generalized version of S_p statistic proposed by Kashid and Kulkarni (2002). The equivalence of this method is established with C_p (Mallows, 1973), S_p (Kashid and Kulkarni, 2002) and R_p

(Dorugade and Kashid, 2010) methods. An article "Subset Selection in Multiple Linear Regression in Presence of Outlier and Multicollinearity" is prepared and presented in International Conference at Raleigh (USA). The work related to this article is enclosed in Chapter 3 of work done report.

Non-normal distribution of error variable is one of the important problems of regression model. In most of the practical situations, normality assumption on the response variable is violated or there are some datasets which does not support normality assumption. Very few subset selection methods address this problem. N_p criterion proposed by Kashid and Kulkarni (2003) is one of the methods which use the rank based estimates of regression coefficients. We developed a subset selection method using rank based estimates of regression parameters. The distributional study of proposed statistic is done with the help of QQ plots and chi-square test for goodness of fit. Also, a stepwise procedure is proposed. This work is presented in the International Conference organized by Cochin University of Science and Technology, Cochin. An article "Subset Selection in Multiple Linear Regression for Non-normal Symmetric Error Terms" is prepared.

The Chapter 5 is devoted for count data. In the same Chapter, a new subset selection criterion for Poisson regression is proposed and some results on the same are obtained. A stepwise procedure for selection of subset model is illustration with suitable example. Work of this chapter is presented in National Conference at Jalgaon and published in Journal of Statistical Theory and Practice.

The source codes/ programs are developed using statistical software for all the above developed procedures. These are given in the appendix of the work done report.

14. CONTRIBUTION TO THE SOCIETY (GIVE DETAILS)

Regression analysis is one of the important tools in Statistical data analysis. It is widely used in all most all fields like engineering, chemical sciences, medical, biological sciences, business, economics, social sciences and also in marketing research. Subset selection or variable selection is one of the important problems in regression and important tool for data analysis. Therefore subset selection technique is directly related to

BABA. BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR

all the above fields for developing relationship between the response and subset of predictor variables.

In real life data, one or more following situations may be occur:

- Data contains outlier observation(s)
- Multicollinearity in the data
- Assumption of normality violated
- Linear model is not appropriate

We would like to undertake study related to apply the techniques for the data in which one or more of the above situations exist. The proposed subset selection methods are useful to practitioners for data analysis when the above situations present in the data.

15. WHETHER ANY PH.D. ENROLLED/ PRODUCED OUT OF THE PROJECT

Yes.

16. NO. OF PUBLICATIONS OUT OF THE PROJECT (PLEASE ATTACH RE-PRINTS)

02.

(PRINCIPAL INVESTIGATOR)

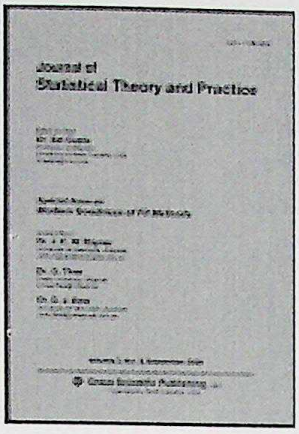
Principal Investigator

UGC Funded Project
"A Study of Subset Selection Techniques in Regression and Application"

(REGISTRAR/PRINCIPAL)

Registrar,
Shivaji University, Kolhapur.

This article was downloaded by: [Shivaji University], [Deepak Sakate]
On: 30 November 2011, At: 21:48
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Statistical Theory and Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ujsp20

Subset Selection in Poisson Regression

D. M. Sakate [a] , D. N. Kashid [a] & D. T. Shirke [a]
[a] Department of Statistics, Shivaji University, Kolhapur, India

Available online: 30 Nov 2011

To cite this article: D. M. Sakate, D. N. Kashid & D. T. Shirke (2011): Subset Selection in Poisson Regression, Journal of Statistical Theory and Practice, 5:2, 207-219

To link to this article: http://dx.doi.org/10.1080/15598608.2011.10412024

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

GSP
©Grace Scientific Publishing

Journal of
Statistical Theory and Practice
Volume 5, No. 2, June 2011

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

Subset Selection in Poisson Regression

D. M. Sakate, *Department of Statistics, Shivaji University, Kolhapur, India. Email: dms.stats@gmail.com*

D. N. Kashid, *Department of Statistics, Shivaji University, Kolhapur, India. Email: dnkashid_in@yahoo.com*

D. T. Shirke, *Department of Statistics, Shivaji University, Kolhapur, India. Email: dts_stats@unishivaji.ac.in*

Received: November 24, 2009 *Revised:* September 18, 2010 & October 9, 2010

Abstract

In this article, we propose a criterion for subset selection in Poisson regression called D_p criterion. This criterion uses the deviance of the full model and subset model to arrive at a decision. Based on the same criterion a stepwise procedure is also developed to select the appropriate subset. The procedure is useful even when the number of regressors is large. The proposed stepwise method is operationally simple to implement. The method is illustrated with examples.

AMS Subject Classification: 62J12.

Key-words: Deviance; Poisson regression; Stepwise procedure; Subset selection.

1. Introduction

The prime motive behind fitting a regression model is predicting the future values of response when value of regressors (X) is known. As a matter of fact, fitting a model with large number of regressors is neither practicable nor economical (Miller, 2002). Naturally, a question arises that which variables should comprise the model? Many researchers have answered this question by proposing various criteria or stepwise methods. Various subset selection methods in linear set up are available in the literature and among these Mallow's C_p (Mallow, 1973) is a well known and widely used criterion. It is defined as,

* 1559-8608/11-2/$5 + $1pp - see inside front cover
© Grace Scientific Publishing, LLC

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p)$$

where, RSS_p is the residual sum of squares for submodel with p parameters; n is the number of data points and $\hat{\sigma}^2$ is an estimate of the error variance obtained from the full model.

Suppose response variable represents a count data of some relatively rare event, such as defects in a unit of manufactured product, bugs in software etc. In this situation, Poisson regression model is suitable for developing the relationship between the observed count and potentially useful regressors or predictor variables. Subset selection problem is of an interest for linear regression model, giving the best subset from the set of regressors. Similarly, for Poisson regression model, one would like to obtain the best subset of regressors. In literature, very few subset selection methods in generalized linear regression models are available like Group least absolute shrinkage and selection operator (Meier *et al.*, 2008), Least Angle Regression (LARS) (Efron *et al.*, 2004) which use a stepwise algorithm to select a parsimonious subset for prediction. Guo and Li (2002) illustrate that the Akaike information Criterion (AIC) and Bayesian information criterion (BIC) work well for selecting the correct model in the error -in-variables Poisson regression.

In practice, number of regressors may be large, for example say 30. In this case number of all possible subsets is more than one billion. It is impractical to implement an exhaustive search method. It is reasonable to think of a stepwise procedure which is less time consuming and easy to implement. Many researchers have proposed such stepwise procedures to select the best subset in linear regression set up. Some of which are due to Thompson (1978a and b), Efroymson (1960), Yamashita, Yamashita and Kamimura (2007) etc. In the present paper, an effort is made to devise a criterion for subset selection based on the deviance. A stepwise procedure for subset selection has been provided.

This article is organized as follows. In Section 2, we take a brief review of Poisson regression model. In Section 3, a new criterion for subset selection in Poisson regression is proposed and some results on the same are obtained. In Section 4, the proposed method is illustrated with numerical examples. A stepwise procedure with illustration is presented in Section 5. Performance of the proposed method is evaluated through simulation study in Section 6. Concluding remarks are given at the end of the article.

2. Poisson Regression Model

The assumptions and the model that encompass Poisson regression model are as follows. Let $Y_1, Y_2, \ldots, Y_n$ be necessarily independent Poisson variates with respective mean μ_i $(i = 1,2,\ldots,n)$. Let $X_i' = [X_{i0}, X_{i1}, X_{i2}, \ldots, X_{ik}], X_{i0} = 1, i = 1,2,\ldots,n$ be the set of regressors. A set of regressors and μ_i are related via the model

$$\mu_i = exp\{X_i'\beta\}, \quad i = 1,2,\ldots,n \tag{2.1}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$. The log likelihood function for observed values $y_1, y_2, \ldots, y_n$ is given by (Myers, Montgomery and Vining, 2002)

$$lnL(\beta, y) = \sum_{i=1}^{n} (-exp\{X_i'\beta\} + y_i X_i'\beta - ln(y_i!))$$

The score function for the estimation of β is given by

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

$$\sum_{i=1}^{n}(y_i - exp\{X_i'\beta\})X_i = 0$$

That is

$$\sum_{i=1}^{n}(y_i - \mu_i)X_i = 0$$

In matrix notation, we can write the above score function as,

$$X'(Y - \mu) = 0.$$

This is nonlinear in β . Thus, iterative reweighted least squares method can be used to obtain MLE of β (see, Myers, Montgomery and Vining, 2002).

Many statistical tests are available in literature to test the goodness of fit of the model. One of these tests uses likelihood ratio criterion. Deviance theoretically allows one to determine if the fitted regression model is significantly worse than the saturated model (Myers, Montgomery and Vining, 2002). It is defined as

$$D(\hat{\beta}) = -2ln(\frac{L(\hat{\beta})}{L(\hat{\mu})})$$

where, $L(\hat{\beta})$ is the likelihood of fitted model with β replaced by its MLE and $L(\hat{\mu})$ is the likelihood for saturated model in which n parameters $\mu_1, \mu_2, ..., \mu_n$ are estimated without the help of regressors. In Poisson regression, deviance is given by

$$D(\hat{\beta}) = 2\sum y_i\, ln(\frac{y_i}{\hat{\mu}_i})$$

If the model is correct, model deviance $D(\hat{\beta})$ is asymptotically distributed as a chi-square variate with $n - p$ degrees of freedom (Myers, Montgomery and Vining, 2002), where p indicates the number of parameters in the model. The smaller values of $D(\hat{\beta})$ will indicate that Poisson model fits well to the data while large values of $D(\hat{\beta})$ will be an indication of inadequacy of the model. Thus, a model is adequate or not can be determined based on the comparison of $D(\hat{\beta})$ (full set of regressors) with $D(\hat{\beta}_0)$ (without any regressors). Extending this idea further, we give a criterion for subset selection which is based on deviance $D(\hat{\beta})$.

3. A Criterion based on deviance

The full model is given by

$$E(Y) = exp\{X\beta\} \tag{3.1}$$

Let us partition the $X_{(n\times k)}$ matrix as $(X_{1(n\times p)} : X_{2(n\times k-p)})$ and $\beta_{(k\times 1)}$ as $(\beta_{1(p\times 1)} : \beta_{2(k-p\times 1)})$. We rewrite the model in (3.1) as

$$E(Y) = exp\{X_1\beta_1 + X_2\beta_2\}$$

Define the subset model 'A' as

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

$$E(Y) = exp\{X_1\beta_1\}$$

The deviance for the model A is given by

$$D(\hat{\beta}_A) = 2\sum y_i \, ln(\frac{y_i}{\hat{\mu}_{iA}})$$

where, $\hat{\mu}_{iA}$ is MLE of μ_i based on model A.

We compare the deviance $D(\hat{\beta}_A)$ with the deviance $D(\hat{\beta})$ corresponding to the full model by taking the difference

$$D(\hat{\beta}_A) - D(\hat{\beta}) \tag{3.2}$$

Obviously, if the deviance of the subset model A is close to the deviance of full model, then a subset model A will be enough to describe the data. On the other hand, large value of (3.2) means that $D(\hat{\beta}_A)$ is far away from $D(\hat{\beta})$ and the subset A is not as good as full model. If $D(\hat{\beta})$ is 'almost close' to the value of $D(\hat{\beta}_A)$, we can safely deploy the subset model for prediction purpose.

In the light of above discussion, we now define a criterion for subset selection based on deviance function. The statistic D_p is defined as

$$D_p = D(\hat{\beta}_A) - D(\hat{\beta}) - (k - 2p) \tag{3.3}$$

where k and p are the number of parameters in full model and subset model respectively, $D(\hat{\beta})$ and $D(\hat{\beta}_A)$ are deviance of the full model and subset model A respectively.

3.1. Distribution of Statistic D_p

Consider the testing of hypothesis problem $H_0 : \beta_2 = 0$ Vs $H_1 : \beta_2 \neq 0$. Here, the null hypothesis is that the subset model is adequate for the prediction. The difference $D(\hat{\beta}_A) - D(\hat{\beta})$ is always positive because deviance of subset model is always larger than that of full model. Moreover, $D(\hat{\beta}_A)$ has $n - p$ degrees of freedom and $D(\hat{\beta})$ has $n - k$ degrees of freedom. Hence, $D(\hat{\beta}_A) - D(\hat{\beta})$ has $k - p$ degrees of freedom. If the null hypothesis is true and if n is large, the difference in (3.2) has a chi-square distribution with $k - p$ degrees of freedom (Montgomery, Peck and Vining, 2006). To support this statement following simulation study is carried out. A Poisson model comprising of four regressors X_1, X_2, X_3 and X_4 was considered. X_1 was generated from B(1, 0.4), and X_2, X_3 and X_4 from U(0,1). We define two different regression models. The first model is based on X_1 and X_2 while the second model is based on X_1, X_2 and X_3. $Y_i's$ were generated from Poisson distribution with mean $\mu_i = exp\{X_i'\beta\}, i = 1, 2, \ldots, n$. The specifications of the two regression models considered are as follows.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

Model	Model Specification
I	$\mu = exp\{5+X_1+2X_2\}$
II	$\mu = exp\{5+X_1+2X_2+3X_3\}$

Three sets each consisting 100 samples corresponding to Model I and II were generated. In the first set, samples of size 20 were generated. In the second set, sample size was 50 and in the third set it was 100. We compute $D(\hat{\beta}_A) - D(\hat{\beta})$ for each model defined above, where $D(\hat{\beta})$ is the deviance for the model which consist of four regressors X_1, X_2, X_3 and X_4.

Summary. The mean and variance of $D(\hat{\beta}_A) - D(\hat{\beta})$ are computed for each model and reported in the Table 3.1.

Table 3.1. Mean and variance of D_p statistic excluding constant.

Model	Model I				Model II			
Sample size n	Sample Mean	Theoretical Mean	Sample Variance	Theoretical Variance	Sample Mean	Theoretical Mean	Sample Variance	Theoretical Variance
20	1.757614	2	2.27872	4	0.889373	1	1.190349	2
50	1.865858	2	3.284995	4	0.89631	1	1.810907	2
100	2.129	2	3.870	4	0.959122	1	2.117303	2

Chi-square test for goodness of fit. The chi-square test for goodness of fit was carried out for fitting of chi-square distribution with $k-p$ d.f to $D(\hat{\beta}_A) - D(\hat{\beta})$. The results are as shown in the Table 3.2.

Table 3.2. Chi-square test for goodness of fit.

Model	Model I		Model II	
Sample size n	χ^2_{cal}	p-value	χ^2_{cal}	p-value
20	8.674597	0.122769	6.536736	0.257435
50	1.606772	0.952089	2.276621	0.517015
100	1.865771	0.867389	0.991026	0.911154

Table 3.2 reveals that the quantity $D(\hat{\beta}_A) - D(\hat{\beta})$ follows approximately chi-square distribution with $k-p$ d.f.

Q-Q Plot.. A simple graphical tool QQ-plot (D'Agostino and Stephens, 1986) was used to assess distribution of $D(\hat{\beta}_A) - D(\hat{\beta})$. QQ-plots were constructed for each sample and are presented in Fig. 3.1. It is clear that the quantity $D(\hat{\beta}_A) - D(\hat{\beta})$ is distributed as chi square with $k-p$ d.f.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

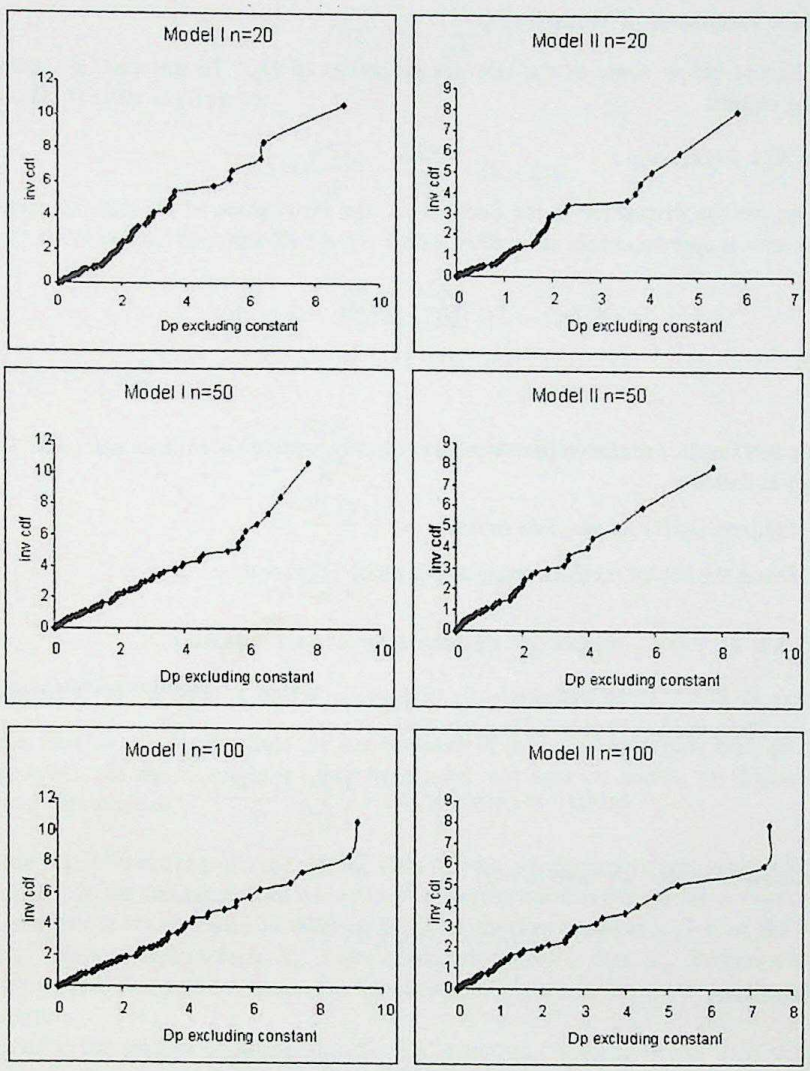


Figure 3.1. QQ-Plots for Model I and II.

From Table 3.1, 3.2 and Q-Q plots, we see that the quantity $D(\hat{\beta}_A) - D(\hat{\beta})$ follows approximately chi-square distribution with $k - p$ d.f. when sample size is large. As a final comment, we observe from Table 3.1 and 3.2 as well as the Q-Q plots that for a sample size of 20 or less, the chi-square approximation is not as good as that for a larger sample size. Nevertheless, this is not a serious limitation of the proposed method as in real-life except possibly in situations like the medical data on rare diseases, sample size is invariably large (at least greater than 30 or so) and so the distribution of the deviance can be closely approximated by chi-square distribution.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

BARR. BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

3.2. Some Properties of Statistics D_p

We discuss below some of the relevant properties of D_p. To this end, we prove the following results:

Lemma 3.1. $E(D_p) \approx p$

Proof. As per the discussion in the Section 3.1, the distribution of D_p statistic excluding constant term is approximately chi-square with $k-p$ d.f. Thus,

$$\begin{aligned} E(D_p) &= E[D(\hat{\beta}_A) - D(\hat{\beta})] - (k-2p) \\ &\approx (k-p) - (k-2p) \\ &= p \end{aligned}$$

Based on this result, a selection procedure for selecting appropriate subset can be devised in two steps as follows:

Step I: Compute D_p for all possible subsets.

Step II: Select the subset corresponding to D_p which is close to p.

Lemma 3.2. *D_p and C_p statistic are equivalent for normal responses.*

Proof. Let $Y_1, Y_2, \ldots, Y_n$ be independently distributed, where Y_i is N(μ_i , σ^2) variate, $i = 1, 2, \ldots, n$. Then, the log likelihood is given by

$$lnL(\mu) = -nln(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i - \mu_i}{\sigma})^2 \tag{3.4}$$

For saturated model, we have $\mu_i = Y_i$

$$lnL(\mu) = -nln(\sigma\sqrt{2\pi}) \tag{3.5}$$

For full model,

$$lnL(\hat{\beta}) = -nln(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i - \hat{\mu}_i}{\sigma})^2 \tag{3.6}$$

where $\hat{\mu}_i$ is MLE of μ_i , i= 1, 2,. . . ,n, and $\hat{\mu}_i = exp\{X_i' \hat{\beta}\}$.
Thus, deviance is

$$\begin{aligned} D(\hat{\beta}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \frac{RSS_k}{\sigma^2} \end{aligned} \tag{3.7}$$

On the similar lines for subset model A, deviance under H_0 is given by

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

$$D(\hat{\beta}_A) = \frac{RSS_p}{\sigma^2} \tag{3.8}$$

Thus, D_p statistic is given by

$$D_p = \frac{RSS_p}{\sigma^2} - \frac{RSS_k}{\sigma^2} - (k-2p) \tag{3.9}$$

As σ^2 is unknown, replace it by

$$\hat{\sigma}^2 = \frac{RSS_k}{(n-k)}$$

Thus, (3.9) becomes

$$\begin{aligned} D_p &= \frac{RSS_p}{\hat{\sigma}^2} - (n-k) - (k-2p) \\ &= \frac{RSS_p}{\hat{\sigma}^2} - (n-2p) \\ &= C_p \end{aligned}$$

4. Numerical Examples

In this Section, we demonstrate the performance of proposed procedure through numerical examples. In the Example 4.1, we have used real data set and in the Example 4.2, simulated data is used.

Example 4.1. In this example, the mining data (Myers, Montgomery and Vining, 2002, pp 142) is used. In the dataset, counts were observed on the number of injuries or fractures that occur in upper seam of mines in the coal fields of the Appalachian region in the western Virginia. The variables include, X_1: Inner burden thickness in feet, X_2: Percent extraction of the lower previously mined seam, X_3: Lower seam height and X_4: Time that the mine has been opened.

We define for sake of completeness the AIC criterion (Akaike, 1974). It is defined as $AIC = -2L+2p$, where L is the maximum log-likelihood and p is the number of parameters in the fitted model. The D_p statistic and AIC criterion are applied to the data. The results are presented in the following table.

It is clear from Table 4.1, these two criteria pick up the same subset namely $\{X_1, X_2, X_4\}$.

Alternatively, one can plot the values of D_p against p and select the subset corresponding to D_p which is close to the line $D_p = p$.

In the Fig. 4.1, the point (4, 3.17677) is close to the line $D_p = p$. Thus, the appropriate subset is the subset corresponding to $D_p = 3.17677$. Hence, the best model is based on X_1, X_2, and X_4.

Examples 4.2. In this example, simulated data was used to study the performance of the proposed criterion. 30 observations on X_1 were generated from B(1, 0.4) and those

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

Table 4.1. D_p and AIC for Mining Data.

Model	AIC	D_p
X_1	172.1114	32.97601
X_2	148.8919	9.75651
X_3	174.1132	34.9778
X_4	172.3352	33.19984
X_1X_2	144.3654	5.23009
X_1X_3	173.3420	34.20666
X_1X_4	172.6622	33.52680
X_2X_3	149.8588	10.72348
X_2X_4	143.8988	4.76343
X_3X_4	173.4963	34.36098
$X_1X_2X_3$	146.0221	6.88675
$X_1X_2X_4$	142.3121	3.17677
$X_1X_3X_4$	174.0410	34.90568
$X_2X_3X_4$	145.2918	6.156402
$X_1X_2X_3X_4$	144.1354	5

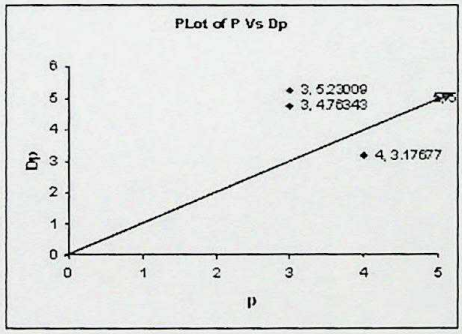


Figure 4.1. Plot of p Vs D_p.

on X_2 and X_3 from U(0,1). Y_i's were generated from Poisson distribution with mean $\mu_i = exp\{2+2X_{1i}+X_{2i}\}, i = 1,2,\ldots,n$. D_p and AIC were computed and are reported in Table 4.2.

Table 4.2. D_p and AIC for simulated data.

Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$
AIC	278.5612	1127.048	1092.168	182.6249	279.5561	1078.483	183.5106
D_p	99.05062	947.5328	912.6572	3.114323	100.0455	898.9727	4

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

The results of Table 4.2 reveal that both the methods pick up the correct subset namely, $\{X_1, X_2\}$. As discussed earlier, in applications, we may have a large number of regressors and it is tedious to compute D_p for all possible $2^k - 1$ subsets of regressors. In the next section we propose a stepwise procedure for subset selection based on D_p criterion.

5. Stepwise procedure for subset selection using D_p criterion

In this method, after each variable is added to the set of selected variables, a test is made to see if any of the previously selected variables can be deleted without appreciably increasing deviance. This procedure incorporates criteria for the addition and deletion of variables as follows.

(a) *Addition Criterion*

Let $D(\hat{\beta}_p)$ be the deviance with p-1 variables. Suppose the smallest deviance which can be obtained by adding another variable to the present set is $D(\hat{\beta}_{p+1})$. The ratio

$$R_1 = \frac{D(\hat{\beta}_p) - D(\hat{\beta}_{p+1})}{D(\hat{\beta}_{p+1})/(n-p-1)} \tag{5.1}$$

is calculated and compared with $F_{1,n-p-1,\alpha}$; the upper α^{th} quantile of F-variate with 1 and $n-p-1$ degrees of freedom. If R_1 is greater than $F_{1,n-p-1,\alpha}$, then the corresponding variable is added to the selected subset of variables.

(b) *Deletion criterion*

With $p-1$ variables in the selected subset, let $D(\hat{\beta}_{p-1})$ be the smallest deviance which can be obtained after deleting any variable from the previously selected variables. The ratio

$$R_2 = \frac{D(\hat{\beta}_{p-1}) - D(\hat{\beta}_p)}{D(\hat{\beta}_p)/(n-p)} \tag{5.2}$$

is calculated and compared with an $F_{1,n-p,\alpha}$. If R_2 is less than $F_{1,n-p,\alpha}$, then the corresponding variable is deleted from the selected subset model.

Based on above addition and deletion criteria, we give an algorithm for selecting the appropriate subset.

(c) *Algorithm*

Let S denotes the set of regressors included in the subset model. Initially, let $S = \{X_0\}$. The stepwise procedure is as follows.

Step I. Regress Y on all the variables and compute deviance of the full model.

Step II. Regress Y on each variable and compute D_p statistic. Add the variable to S that corresponds to minimum D_p. Suppose i^{th} variable is selected at this step. Then $S = \{X_0, X_i\}$.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

Step III. Regress response variable Y on previously selected variables and single variable from set of unselected variables. Obtain D_p statistic. Repeat the same for all unselected variables. Select that subset which corresponds to minimum D_p. For this subset, compute R_1. If $R_1 > F_{1,n-p-1,\alpha}$, include this variable in S and go to next step otherwise terminate.

Step IV. If number of variables in selected model in step III is greater than two then go to the next step otherwise repeat step III.

Step V. Regress response variable Y on currently selected variable and eliminating single variable from set of selected variables and obtain D_p statistic. Repeat this for all previously selected variables. Compute R_2 for the subset corresponding to minimum D_p. If $R_2 < F_{1,n-p,\alpha}$ then remove that variable from S otherwise go to next step.

Step VI. Repeat Steps III to V until no addition is possible or all the variables are included in S.

5.1. Illustration

Let us illustrate the above stepwise procedure by applying it to the simulated data comprising of eight regressors generated from U(0, 1). Sample size was 50. The response variable Y was generated from Poisson distribution with mean $\mu = exp\{2 + X_1 + X_2 + X_3 + 2X_4 + 2X_5\}$. The results are summarized in Table 5.1. The total number of all possible subsets in this case is 255. The number of subsets analyzed to arrive at the decision is 42. We see that this method reduces computations to a greater extent (approximately 84%).

From Table 5.1, we see that stepwise procedure and D_p statistic select a proper subset namely $\{X_1, X_2, X_3, X_4, X_5\}$.

Model selection ability of the criterion D_p is demonstrated by using simulation. In this simulation study we consider two simulation designs as follows:

Table 5.1. Stepwise procedure for subset selection.

Step	Minimum D_p	R_1	R_2	$F_{1,n-p-1,\alpha}$	$F_{1,n-p,\alpha}$	S	Decision
I	{4}* 12780.53	-	-	-	-	{4}	Continue
II	{4, 5} 4866.836	11.461	-	4.0471	-	{4, 5}	Continue
III	{3, 4, 5} 3140.985	24.9473	{3, 4} 102.8366	4.05179	4.05179	{3, 4, 5}	Continue
IV	{2, 3, 4, 5} 1785.858	33.3923	{2, 4, 5} 3211.631	4.05661	4.05661	{2, 3, 4, 5}	Continue
V	{1, 2, 3, 4, 5} 6.062996	1666.76	{1, 2, 3, 5} 1490.833	4.06170	4.06170	{1, 2, 3, 4, 5}	Continue
VI	{1, 2, 3, 4, 5, 6} 6.684952	1.07248	-	4.06704	-	{1, 2, 3, 4, 5}	No addition. Terminate

* $\{i, j, k, \ldots\}$ are the subscripts of $X_i, X_j, X_k, \ldots$.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

A: All the regressors generated from uniform distribution. 500 random samples of size n on regressors X_1, X_2, X_3, X_4 and X_5 were generated from U(0, 1). Y_i's were generated from Poisson distribution with mean as specified below. The D_p criterion was applied to each sample to select the appropriate subset. Thereafter, number of times the criterion selects the correct model was counted. The same procedure was repeated for five different models specified in Table 5.2. The model selection ability is presented in the Table 5.3.

Table 5.2. Submodel Specification.

Model	Submodel Specification
I	$\mu = exp\{2 + 2X_1 + X_2\}$
II	$\mu = exp\{0.5 + X_1 + 2X_2\}$
III	$\mu = exp\{1 + 0.5X_1 + X_3 + X_4\}$
IV	$\mu = exp\{1 + X_1 + 1.5X_4\}$
V	$\mu = exp\{X_1 + X_3\}$

Table 5.3. Model selection ability (%) of D_p.

Sample size n	10		20		50	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
I	77	23	76	24	68	32
II	73	27	68	32	71	29
III	74	26	81	19	80	20
IV	78	22	79	21	69	31
V	69	31	83	17	70	30

Table 5.4. Model selection ability (%) of D_p.

Sample size n	10		20		50	
Model	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
I	80	20	78	22	74	26
II	77	23	82	18	76	24
III	85	15	84	16	89	11
IV	82	18	82	18	81	19
V	80	20	83	17	82	18

B: All the regressors generated from different distributions. The same design in case (A) is followed except only four regressors are considered. X_1 was generated from B(1, 0.4), X_2 was generated from U(0, 1), X_3 was generated from N(0, 1) and X_4 was generated from Poisson(1). Submodel specifications used were same as that given in Table 5.2. Table 5.4 demonstrates the model selection ability of D_p

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

BARR. BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

It is clear that the performance of the proposed method is good irrespective of sub model specification and sample size (n). Further, we have also verified that the performance of stepwise procedure is same as that of D_p statistic.

6. Concluding Remarks

The subset selection method introduced here is based on deviance and it is simple to implement. The performance of the proposed method to select appropriate subset model is more than 70%. The prominent feature of the proposed criterion D_p is to devise a stepwise procedure for subset selection. This procedure helps in reducing the time and cost for computational purpose when large number of variables are in the model. We have implemented the stepwise procedure using *R* software. The code written is available with the first author.

Acknowledgements

The second author wishes to thank the University Grants Commission, New Delhi, India for the financial support under Major Research Project scheme. We wish to thank the Editor and the anonymous referee for their valuable suggestions, which led to improvement in the presentation of this paper.

References

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

D' Agostino, R.B., Stephens, M.A., 1986. *Goodness of Fit Ttechniques.* Marcel Decker, Inc.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. *Annals of Statistics, 32(2), 407–499.*

Efroymson, M.A., 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, Ralston, A. and Wilf, H.S. (Editors), Wiley, New York.

Guo, Jie Q., Li, Tong, 2002. Poisson regression models with errors-in-variables: implementation and treatment. *Journal of Statistical Planning and Inference,* 104(2),391–401.

Mallows, C.L., 1973. Some comments on Cp. *Technometrics, 15, 661–675.*

Meier, L., Sara van de Geer, Bühlmann, P., 2008. Group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 70(1), 53–71.

Miller, A.J., 2002. *Subset Selection in Regression.* Chapman and Hall.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2006. *Introduction to Linear Regression Analysis.* John Wiley and Sons, New York.

Myers, R.H., Montgomery, D.C., Vining G.G., 2002. *Generalized Linear Models: with Applications in Engineering and the Sciences.* John Wiley and Sons, New York.

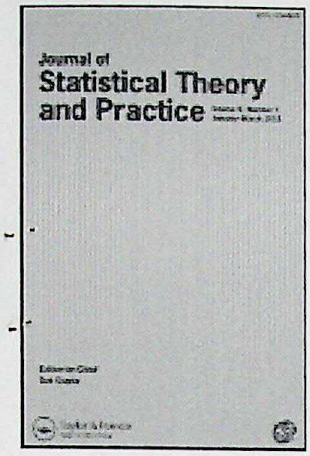
Thompson, M.L., 1978a. Selection of variables in multiple linear regression part I: A review and evaluation. *Inter. Stat. Review.,* 46, 1–9.

Thompson, M.L., 1978b. Selection of variables in multiple linear regression part II: Chosen predictors, computation and examples. *Inter. Stat. Review.,* 46, 129–146.

Yamashita, T., Yamashita, K., Kamimura, R., 2007. A stepwise AIC method for variable selection in linear regression. *Communications in Statistics-Theory and methods,* 36, 2395–2403.

Downloaded by [Shivaji University], [Deepak Sakate] at 21:48 30 November 2011

This article was downloaded by: [Shivaji University], [Nileshkumar Jadhav]
On: 17 February 2012, At: 23:55
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Statistical Theory and Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ujsp20

A Jackknifed Ridge M-Estimator for Regression Model with Multicollinearity and Outliers

N. H. Jadhav [a] & D. N. Kashid [a]

[a] Department of Statistics, Shivaji University, Kolhapur, 416004, India

Available online: 17 Feb 2012

To cite this article: N. H. Jadhav & D. N. Kashid (2011): A Jackknifed Ridge M-Estimator for Regression Model with Multicollinearity and Outliers, Journal of Statistical Theory and Practice, 5:4, 659-673

To link to this article: http://dx.doi.org/10.1080/15598608.2011.10483737

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

GSP
©Grace Scientific Publishing

Journal of
Statistical Theory and Practice
Volume 5, No. 4, December 2011

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

A Jackknifed Ridge M-estimator for Regression Model with Multicollinearity and Outliers

N. H. Jadhav, *Department of Statistics, Shivaji University, Kolhapur - 416004, India.*
Email: n.nil08@gmail.com

D. N. Kashid, *Department of Statistics, Shivaji University, Kolhapur - 416004, India.*
Email: dnkashid_in@yahoo.com

Received: December 5, 2010 *Revised:* June 16, 2011

Abstract

In the multiple regression analysis, most frequently occurring problems are the presence of multicollinearity and outliers. They produce undesirable effects on the least squares estimates of regression parameters. The Jackknifed Ridge Regression estimator and M-estimator have been proposed to overcome multicollinearity and outliers respectively. The Jackknifed Ridge Regression estimator is obtained by shrinking the Ordinary Least Squares estimator. Since the Ordinary Least Squares estimator is sensitive to outliers, the Jackknife Ridge Regression estimator is also sensitive to outliers. To overcome the combined problem of multicollinearity and outliers, we propose a new estimator namely, Jackknifed Ridge M-estimator. This estimator is obtained by shrinking an M-estimator instead of the Ordinary Least Squares estimator.

AMS Subject Classification: 62J05; 62J07.

Key-words: Ridge estimator; Jackknifed ridge estimator; M-estimator; Multicollinearity; Outlier.

1. Introduction

In regression, the least squares estimation method is widely used for parameter estimation. This method performs well under some basic assumptions like errors are independent and follow normal distribution with mean zero and the constant variance. In real life situation, multicollinearity, outliers and departure from the normality assumption are common problems in regression. These produce undesirable effects on the least squares estimator. This fact is pointed out by many researchers.

* 1559-8608/11-4/$5 + $1pp - see inside front cover
© Grace Scientific Publishing, LLC

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

The least squares estimator of regression coefficients produces large variance in the presence of multicollinearity. The problem of multicollinearity has received considerable attention from several researchers. Some of them have developed alternative estimators when the multicollinearity is severe. Among these, Hoerl and Kennard (1970) proposed the ridge estimator which is widely used because of its optimality properties (see Vinod and Ullah, 1981). However, it is a biased estimator. Singh, Chaubey and Dwivedi (1986) have used the Jackknife procedure to reduce the bias and proposed the Jackknifed ridge estimator.

Another important problem which needs attention in regression parameter estimation is the presence of outliers in the data (see Huber, 1981). Outliers affect the efficiency of least squares estimates (see Montgomery, Peck and Vining, 2006). To overcome this problem, robust estimation methods like M-estimator, least median of squares (LMS) estimator, MM estimator (see Huber, 1981; Rousseeuw and Leroy, 1987) are available.

Several procedures which deal with multicollinearity and outliers separately are available. However, there is not much significant work reported in the literature which takes into account the presence of both multicollinearity and outlier problems simultaneously. In this article, we propose a new method of estimation which handles both the problems simultaneously.

The outline of the paper is as follows. In Section 2, we define model and parameter estimation methods with their bias and Mean Square Error (MSE). In Section 3, we propose a new estimator designated as Jackknifed Ridge M-estimator (JRM) and discuss some of its properties. The performance of this estimator is studied in Section 4. In Section 5, a simulation study is carried for justifying the superiority of the suggested estimator. The paper ends with some concluding remarks in Section 6.

2. Model and Estimators

Consider the multiple linear regression model,

$$Y = \beta_0 1 + X\beta + \varepsilon, \tag{2.1}$$

where Y is an n×1 response vector, 1 is column of ones, X is an n×k matrix of predictor variables, β is k×1 vector of unknown regression coefficients, ε is an $n \times 1$ error vector with mean 0 and variance $\sigma^2 I$. For convenience, it is assumed that the X variables are standardized so that $X'X$ has the form of correlation matrix.

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigen values of $X'X$ and $q_1, q_2, \ldots, q_k$ be the corresponding eigen vectors. Let $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $Q = (q_1, q_2, \ldots, q_k)$, such that $X'X = Q\Lambda Q'$. The regression model (2.1) can be written in a canonical form as

$$Y = \beta_0 1 + Z\alpha + \varepsilon, \tag{2.2}$$

where $Z = XQ$ and $\alpha = Q'\beta$.

The ordinary least squares (OLS) estimator of (2.2) is given by

$$\hat{\alpha}_{LS} = \Lambda^{-1} Z'Y. \tag{2.3}$$

Since $\alpha = Q'\beta$ and $Q'Q = I$, the OLS estimator of β is given by

$$\hat{\beta}_{LS} = Q\hat{\alpha}_{LS}. \tag{2.4}$$

Note that, because of the relation $\alpha = Q'\beta$, any estimator $\hat{\alpha}$ of α has a corresponding $\hat{\beta} = Q\hat{\alpha}$ and MSE $(\hat{\alpha})$ = MSE $(\hat{\beta})$. Hence, it is sufficient to consider only the canonical form. Below, we discuss the parameter estimation methods.

2.1. Ordinary Ridge Regression (ORR) Estimator in the Presence of Multicollinearity

Hoerl and Kennard (1970) proposed the following Ordinary Ridge Regression (ORR) estimator

$$\begin{aligned}\hat{\alpha}_R(r) &= (\Lambda + rI)^{-1}Z'Y \\ &= (I - rA^{-1})\hat{\alpha}_{LS},\end{aligned} \tag{2.5}$$

where $r > 0$ is the shrinkage or ridge parameter and $A = (\Lambda + rI)$.

The Mean Square Error (MSE) matrix of ORR estimator is given by

$$\begin{aligned}\text{MSE}(\hat{\alpha}_R(r)) &= \text{Var}(\hat{\alpha}_R(r)) + [\text{bias}(\hat{\alpha}_R(r))][\text{bias}(\hat{\alpha}_R(r))]' \\ &= \sigma^2(I - rA^{-1})\Lambda^{-1}(I - rA^{-1})' + r^2A^{-1}\alpha\alpha'A^{-1},\end{aligned} \tag{2.6}$$

where Var $(\hat{\alpha}_{LS}) = \sigma^2\Lambda^{-1}$.

2.2. Ordinary Jackknifed Ridge Regression (OJR) Estimator

Singh, Chaubey and Dwivedi (1986) extended the Jackknife procedure to ridge regression and proposed Ordinary Jackknifed Ridge Regression (OJR) estimator as

$$\begin{aligned}\hat{\alpha}_J(r) &= (I + rA^{-1})\,\hat{\alpha}_R(r) \\ &= (I - r^2A^{-2})\hat{\alpha}_{LS}.\end{aligned} \tag{2.7}$$

The OJR estimator is obtained by shrinking the OLS estimator $\hat{\alpha}_{LS}$ using the matrix $(I - r^2A^{-2})$. The matrix of MSE value of this estimator is

$$\begin{aligned}\text{MSE}(\hat{\alpha}_J(r)) &= \text{Var}(\hat{\alpha}_J(r)) + [\text{bias}(\hat{\alpha}_J(r))][bias(\hat{\alpha}_J(r))]' \\ &= \sigma^2(I - r^2A^{-2})\Lambda^{-1}(I - r^2A^{-2})' + r^4A^{-2}\alpha\alpha'A^{-2}.\end{aligned} \tag{2.8}$$

2.3. M-Estimator when Data contains Outliers

M-estimator for parameters of regression model is obtained by minimizing the sum of the function of the scaled residuals

$$\sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right), \tag{2.9}$$

where s is an estimate of scale. Differentiation of Eq. (2.9) with respect to each parameter leads to $k + 1$ equations of the form

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

$$\sum_{i=1}^{n} v_{ij}\,\psi\left\{\frac{Y_i-\beta_0-z_i'\alpha}{s}\right\}=0,\quad j=0,1,\ldots,k, \tag{2.10}$$

where $\psi(.)$ is the partial derivative of ρ with respect to parameters and v_{ij} is the jth entry in the ith row of matrix $V=[1, Z]$. The solution is obtained by iterative reweighted least squares method (Huber, 1973; Draper and Smith, 1998). Huber (1973) has shown that asymptotically $\hat{\alpha}_M$ has an approximate normal distribution with covariance matrix

$$cov(\hat{\alpha}_M)=\Omega=\sigma^2\frac{E\psi^2(\varepsilon/s)}{[E\psi'(\varepsilon/s)]^2}(V'V)^{-1}. \tag{2.11}$$

Notice that OJR estimator is robust to multicollinearity and M-estimator is robust to outliers. However, if the multicollinearity and outliers occur simultaneously in the data, individual performance of these estimators is unsatisfactory. To deal with the combined problem of multicollinearity and outliers, we propose therefore a new estimator called the Jackknifed Ridge M-estimator (JRM).

2.4. Choice of Ridge parameter r

The theoretical choice of r suggested by Hoerl, Kennard and Baldwin (1975) is

$$r=\frac{k\sigma^2}{\alpha'\alpha}, \tag{2.12}$$

which minimizes the MSE of $\hat{\alpha}_R(r)$. Usually, the value of r is obtained by substituting σ^2 and α by their OLS estimates. To develop the proposed method, we estimate the value of r by, say $\hat{r}$ by replacing σ^2 and α in Eq. (2.12) with their robust estimates. More specifically, the estimate of r is obtained by

$$\hat{r}=\frac{ks^2}{\hat{\alpha}_M'\hat{\alpha}_M}, \tag{2.13}$$

where $s=1.4826\,median\,|e_i-median(e_i)|$ and e_i is ith residual obtained by using OLS estimator. We use s as a robust estimate of σ.

3. Proposed Estimator - Jackknifed Ridge M-estimator

Let $(\hat{\beta}_{0M},\ \hat{\alpha}_M)$ be an M-estimator of $(\beta_0,\ \alpha)$. It is obtained by solving the equations $\sum_{i=1}^{n} v_{ij}\,\psi\left\{\frac{Y_i-\beta_0-z_i'\alpha}{s}\right\}=0$, where $\psi(.)$ is some function (see Huber, 1981; Hampel *et al*, 1986). Thus, Jackknifed Ridge M-estimator (JRM) is defined as,

$$\hat{\alpha}_{\mathrm{JRM}}(r)=(I-r^2A^{-2})\hat{\alpha}_M. \tag{3.1}$$

Notice that how (3.1) combines the ridge estimator and the robust estimator to minimize the problem of multicollinearity and outliers simultaneously.

Having proposed the new estimator, now we proceed to obtain the expressions for bias, variance and MSE of JRM estimator.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

3.1. Bias

The bias in $\hat{\alpha}_{\mathrm{JRM}}(r)$ is given by

$$\begin{aligned}\mathrm{Bias}\,(\hat{\alpha}_{\mathrm{JRM}}(r)) &= E\,[\hat{\alpha}_{\mathrm{JRM}}(r) - \alpha] \\ &= -r^2A^{-2}\alpha. \end{aligned} \tag{3.2}$$

3.2. Variance

The variance of $\hat{\alpha}_{JRM}(r)$ is as follows:

$$\begin{aligned}\mathrm{Var}\,(\hat{\alpha}_{\mathrm{JRM}}(r)) &= E\left[(\hat{\alpha}_{\mathrm{JRM}}(r) - \alpha)\,(\hat{\alpha}_{\mathrm{JRM}}(r) - \alpha)'\right] \\ &= (I - r^2A^{-2})\,\Omega\,(I - r^2A^{-2})', \end{aligned} \tag{3.3}$$

where Ω is covariance matrix of $\hat{\alpha}_M$, and we assume that it is finite.

3.3. Mean Square Error (MSE)

$$\begin{aligned}\mathrm{MSE}\,(\hat{\alpha}_{\mathrm{JRM}}(r)) &= Var((\hat{\alpha}_{\mathrm{JRM}}(r)) + [\mathrm{Bias}\,(\hat{\alpha}_{\mathrm{JRM}}(r))][\mathrm{Bias}\,(\hat{\alpha}_{\mathrm{JRM}}(r))]' \\ &= (I - r^2A^{-2})\,\Omega\,(I - r^2A^{-2})' + r^4A^{-2}\alpha\,\alpha'A^{-2}. \end{aligned} \tag{3.4}$$

4. Performance of JRM Estimator

We need a suitable measure which is a scalar function of the MSE matrix to compare meaningfully the performance of the above estimators. To do this, note that the diagonal elements of MSE matrix represent the mean square error of individual estimates. Further, the Trace of MSE measures the 'total variation' in the estimator. We use the Trace of MSE as a measure for evaluation of performance of various estimators. What it means is that suppose $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are two estimators of α with $\mathrm{m}(\hat{\alpha}_1) = \mathrm{trace}(\mathrm{MSE}(\hat{\alpha}_1))$ and $\mathrm{m}(\hat{\alpha}_2) = \mathrm{trace}(\mathrm{MSE}(\hat{\alpha}_2))$ respectively where $\mathrm{MSE}(\hat{\alpha}_1)$ and $\mathrm{MSE}(\hat{\alpha}_2)$ are the matrices of MSE. We prefer $\hat{\alpha}_2$ to $\hat{\alpha}_1$, if $\mathrm{m}(\hat{\alpha}_1) - \mathrm{m}\,(\hat{\alpha}_2) \geq 0$ (see Sakallioğlu and Kaçiranlar, 2008). Using the Trace criterion, we compare the JRM estimator with ORR estimator, OJR estimator, and M-estimator.

4.1. Comparison between the JRM and ORR estimator

We write the difference between the MSE matrices as

$$\begin{aligned}\mathrm{MSE}\,(\hat{\alpha}_R(r)) - \mathrm{MSE}\,(\hat{\alpha}_{\mathrm{JRM}}(r)) &= \sigma^2(I - rA^{-1})\Lambda^{-1}(I - rA^{-1})' + r^2A^{-1}\alpha\alpha'A^{-1} \\ &\quad - (I - r^2A^{-2})\,\Omega\,(I - r^2A^{-2})' - r^4A^{-2}\alpha\,\alpha'A^{-2}. \end{aligned} \tag{4.1}$$

Equivalently, we can write

$$\begin{aligned}m(\hat{\alpha}_R(r)) - m(\hat{\alpha}_{\mathrm{JRM}}(r)) &= \sum_{i=1}^{k}\left(1 - \frac{r}{\lambda_i + r}\right)^2\left[\frac{\sigma^2}{\lambda_i} - \left(1 + \frac{r}{\lambda_i + r}\right)^2\Omega_{ii}\right] \\ &\quad + \sum_{i=1}^{k}\left(\frac{r}{\lambda_i + r}\right)^2\left(1 - \frac{r^2}{(\lambda_i + r)^2}\right)\alpha_i^2, \end{aligned} \tag{4.2}$$

where Ω_{ii} is the ith diagonal element of Ω.
Hence

$$m(\hat{\alpha}_R(r)) - m(\hat{\alpha}_{\mathrm{JRM}}(r)) > 0,$$

whenever,

$$\frac{\sigma^2}{\lambda_i} > \left(1 + \frac{r}{\lambda_i + r}\right)^2 \Omega_{ii} \quad \forall i = 1, 2, \ldots, k.$$

Thus, we may state the following theorem.

Theorem 4.1. *If*

$$\frac{\sigma^2}{\lambda_i} > \left(1 + \frac{r}{\lambda_i + r}\right)^2 \Omega_{ii} \quad \forall \lambda_i, i = 1, 2, \ldots, k \text{ and } r > 0, \tag{4.3}$$

then $m(\hat{\alpha}_R(r)) > m(\hat{\alpha}_{\mathrm{JRM}}(r))$.

4.2. Comparison between the JRM and OJR estimator

Using equations (2.8) and (3.3), we write the difference of MSE matrices as

$$\begin{aligned} MSE(\hat{\alpha}_J(r)) - MSE(\hat{\alpha}_{\mathrm{JRM}}(r)) &= \sigma^2 (I - r^2 A^{-2}) \Lambda^{-1} (I - r^2 A^{-2})' + r^4 A^{-2} \alpha \alpha' A^{-2} \\ &\quad - (I - r^2 A^{-2})\, \Omega\, (I - r^2 A^{-2})' - r^4 A^{-2} \alpha\, \alpha' A^{-2} \\ &= \sigma^2 (I - r^2 A^{-2}) \Lambda^{-1} (I - r^2 A^{-2})' \\ &\quad - (I - r^2 A^{-2})\, \Omega\, (I - r^2 A^{-2})'. \end{aligned} \tag{4.4}$$

Equivalently,

$$m(\hat{\alpha}_J(r)) - m(\hat{\alpha}_{\mathrm{JRM}}(r)) = \sum_{i=1}^{k} \left(1 - \frac{r^2}{(\lambda_i + r)^2}\right)^2 \left[\frac{\sigma^2}{\lambda_i} - \Omega_{ii}\right]. \tag{4.5}$$

The MSE of JRM is less than OJR if $\frac{\sigma^2}{\lambda_i} > \Omega_{ii} \quad \forall i = 1, 2, \ldots, k$.
Thus, we have the following theorem:

Theorem 4.2. *If*

$$\frac{\sigma^2}{\lambda_i} \geq \Omega_{ii}. \qquad \forall \lambda_i, i = 1, 2, \ldots, k, \tag{4.6}$$

then $m(\hat{\alpha}_J(r)) \geq m(\hat{\alpha}_{\mathrm{JRM}}(r))$.

4.3. Comparison between the JRM and M-estimator

The MSE expressions of $\hat{\alpha}_M$ and $\hat{\alpha}_{\mathrm{JRM}}(r)$ are

$$\mathrm{MSE}(\hat{\alpha}_M) = \sum_{i=1}^{k} \Omega_{ii}. \tag{4.7}$$

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

$$f(r) = \mathrm{MSE}(\hat{\alpha}_{\mathrm{JRM}}(r)) = \sum_{i=1}^{k} \left(\left(1 - \frac{r^2}{(\lambda_i + r)^2} \right)^2 \Omega_{ii} - \frac{r^4}{(\lambda_i + r)^4} \alpha_i^2 \right), r > 0. \quad (4.8)$$

Since $f(0) = \mathrm{MSE}(\hat{\alpha}_M)$, we find the condition on r such that $f(r)$ is decreasing function in r. If we take derivative of $f(r)$ with respect to r, we obtain

$$f'(r) = \sum_{i=1}^{k} \left[\frac{-4r\lambda_i}{(\lambda_i + r)^3} \Omega_{ii} + \frac{4r^3 \lambda_i}{(\lambda_i + r)^5} \left(\Omega_{ii} + \alpha_i^2 \right) \right]. \quad (4.9)$$

Hence, $f'(r) < 0$ if

$$\frac{-4r\lambda_i}{(\lambda_i + r)^3} \Omega_{ii} + \frac{4r^3 \lambda_i}{(\lambda_i + r)^5} \left(\Omega_{ii} + \alpha_i^2 \right) < 0, \quad \forall i = 1, 2, \ldots, k.$$

The function $f(r)$ will be a decreasing function of r if

$$r < \frac{\lambda_i}{\sqrt{\frac{\Omega_{ii} + \alpha_i^2}{\Omega_{ii}}} - 1} \quad \forall i = 1, 2, \ldots, k. \quad (4.10)$$

From this result, we may state the following theorem.

Theorem 4.3. *If*

$$r < \frac{\lambda_i}{\sqrt{\frac{\Omega_{ii} + \alpha_i^2}{\Omega_{ii}}} - 1} \quad \forall i = 1, 2, \ldots, k, \quad (4.11)$$

then $\mathrm{m}(\hat{\alpha}_M) > \mathrm{m}(\hat{\alpha}_{\mathrm{JRM}}(r))$.

It can be seen from the above results that in case of more severe multicollinearity and more influential outliers, the values of last few λ_i's will be small and consequently the $m(\hat{\alpha}_{\mathrm{JRM}}(r))$. It indicates that in the presence of multicollinearity and outlier, performance of JRM estimator is better than ORR, OJR and M-estimator.

Note: When there is no multicollinearity and no outliers in the data, the JRM, M-estimator, OJR and ORR estimators reduce to the least squares estimator. Theoretically, the MSE of each estimator is equal.

5. A Simulation Study

In this section, we present a simulation study to evaluate the performance of JRM estimator. To achieve required degree of multicollinearity, we used a simulation design given by McDonald and Galarneau (1975). The regressor variables are generated by

$$x_{ij} = \left(1 - \rho^2\right)^{1/2} z_{ij} + \rho z_{i5}, \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, p. \quad (5.1)$$

where $p = 4$, z_{ij} are independent standard normal pseudo random numbers, ρ^2 is the correlation between any two explanatory variables and error term is generated from $N(0, \sigma^2)$. Response variable Y is generated using the regression model

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

$$Y = 2 + 5X_1 + 3X_2 + 4X_3 + 8X_4 + \varepsilon.$$

Note that the choice of regression model given above is arbitrary and for the sake of illustration, it is used here. Experiments with some other choices of regression coefficients have produced similar results. In this study, $\rho = 0.9, 0.99, 0.999$, sample size $(n) = 20, 30, 50, 100$ and error variance $(\sigma^2) = 1, 9, 25, 100, 400, 900$ are considered.

Let $e_i^*, i = 1, 2, \ldots, n$ be the absolute value of ith residual obtained from the least squares fit. Next, arrange these values in increasing order of their magnitude such that $e_{(1)}^* \leq e_{(2)}^* \leq \ldots \leq e_{(n)}^*$. We introduce one outlier by multiplying actual value of Y corresponding to $e_{(n)}^*$ by twenty. Similarly, two outliers are introduced in response variable Y corresponding to $e_{(n)}^*$, $e_{(n-1)}^*$ and so on.

5.1. Frequency of smaller MSE

The above simulation experiment is repeated 10000 times and the MSE of JRM, M-estimator, OJR, ORR and OLS are computed. The frequency with which each estimator has lowest MSE is reported in Table 5.1. This experiment is replicated for all possible combinations of ρ, n and σ^2.

It is observed that from Table 5.1;

- Smaller MSE frequency of ORR estimator is greater than that of the OLS, OJR, M-estimator and JRM estimators in the presence of only multicollinearity.
- In the presence of multicollinearity and outlier, the performance of JRM estimator is better than OLS, ORR, OJR and M-estimators.
- Smaller MSE frequency of JRM estimator increases with increase in degree of multicollinearity and error variance.
- More than 80 % times MSE of JRM estimator is small as compared to other estimators in the presence of severe multicollinearity and large variance for outlier data.

Table 5.1. Smaller MSE frequencies of JRM, M-estimator, OJR, ORR and OLS estimators.

ρ	σ^2	JRM	M-estimator	OJR	ORR	OLS	JRM	M-estimator	OJR	ORR	OLS
		without outlier					with outlier				
		n = 20									
0.9	1	1875	1755	202	3978	2190	6146	3788	2	33	31
	9	1568	986	215	6052	1179	7845	1926	6	175	48
	25	1340	543	218	7341	558	8730	822	4	416	28
	100	904	132	194	8611	159	8869	205	18	898	10
	400	571	34	230	9125	40	9049	62	9	878	2
	900	471	18	184	9308	19	9282	22	8	686	2
0.99	1	1630	977	174	6100	1119	8194	1724	2	65	5
	9	858	154	167	8662	159	9436	223	3	333	5
	25	604	58	133	9157	48	9371	57	6	565	1
	100	440	12	116	9419	13	8947	15	5	1033	0
	400	302	2	63	9633	0	8930	5	4	1061	0
	900	273	0	58	9666	3	9237	1	0	762	0
0.999	1	872	135	146	8685	162	9711	184	1	102	2
	9	435	18	110	9417	20	9601	14	2	382	1
	25	314	7	94	9577	8	9348	4	1	647	0
	100	264	1	49	9686	0	8930	0	0	1070	0
	400	243	0	29	9728	0	8956	0	0	1044	0
	900	213	0	16	9771	0	9255	0	0	745	0

Table 5.1 contd. ...

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

BABU BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR

ρ	σ^2	JRM	M-estimator without outlier	OJR	ORR	OLS	JRM	M-estimator with outlier	OJR	ORR	OLS
						n = 30					
0.9	1	1984	1815	191	3715	2295	5914	4026	3	38	19
	9	1648	1208	230	5471	1443	7319	2416	10	212	43
	25	1422	755	215	6785	823	8217	1302	5	437	39
	100	934	202	246	8382	236	8763	326	16	883	12
	400	580	49	253	9066	52	9097	79	11	807	6
	900	464	32	224	9250	30	9435	46	5	514	0
0.99	1	1713	1139	216	5476	1456	7554	2365	1	67	13
	9	981	262	162	8304	291	9292	350	5	345	8
	25	621	96	177	8993	113	9225	98	6	666	5
	100	364	26	130	9451	29	8841	34	8	1115	2
	400	300	7	104	9585	4	8997	6	5	991	1
	900	256	4	71	9669	0	9377	2	0	621	0
0.999	1	921	220	202	8405	252	9570	291	2	135	2
	9	402	28	155	9387	28	9522	22	7	448	1
	25	297	12	113	9567	11	9225	12	6	756	1
	100	209	6	71	9710	4	8790	3	1	1206	0
	400	163	0	38	9798	1	8987	0	2	1011	0
	900	152	0	24	9823	1	9308	0	1	691	0
						n = 50					
0.9	1	2067	1928	165	3514	2326	5583	4343	3	40	31
	9	1884	1398	191	4732	1795	6792	2904	13	208	83
	25	1621	943	230	6009	1197	7613	1840	16	435	96
	100	1093	324	245	7939	399	8510	574	23	857	36
	400	546	97	280	8969	108	9116	146	18	716	4
	900	520	47	283	9098	52	9437	77	2	483	1
0.99	1	1865	1375	223	4823	1714	6876	3016	1	78	29
	9	1117	402	211	7784	486	8905	681	6	387	21
	25	676	157	180	8820	167	9077	178	10	722	13
	100	359	52	191	9357	41	8618	47	18	1316	1
	400	277	6	147	9558	12	8996	15	5	983	1
	900	223	3	117	9654	3	9380	7	2	611	0
0.999	1	1123	347	197	7919	414	9294	562	1	137	6
	9	423	47	188	9295	47	9351	60	5	583	1
	25	310	20	163	9484	23	9047	23	9	920	1
	100	216	8	108	9660	8	8616	3	8	1372	1
	400	153	0	56	9790	1	8887	3	3	1107	0
	900	138	0	44	9818	0	9360	1	0	639	0
						n = 100					
0.9	1	2092	1922	135	3289	2562	5383	4514	2	61	40
	9	1972	1649	187	4112	2080	6059	3585	9	223	124
	25	1825	1306	193	5042	1634	6808	2604	30	419	139
	100	1336	593	234	7060	777	7889	1123	37	866	85
	400	768	183	259	8543	247	9032	254	12	692	10
	900	619	86	299	8899	97	9352	135	4	504	5
0.99	1	1972	1610	171	4167	2080	6265	3599	2	85	49
	9	1458	696	222	6790	834	8197	1296	18	442	47
	25	979	276	230	8176	339	8646	468	16	834	36
	100	521	86	197	9101	95	8375	128	21	1464	12
	400	327	22	179	9438	34	8875	32	10	1082	1
	900	290	10	152	9540	8	9255	10	1	734	0
0.999	1	1411	661	203	6894	831	8686	1163	5	132	14
	9	546	84	201	9064	105	9157	121	15	697	10
	25	367	28	176	9389	40	8739	52	11	1197	1
	100	244	14	152	9582	8	8296	8	14	1682	0
	400	201	2	95	9700	2	8894	2	1	1103	0
	900	165	1	69	9764	1	9287	1	1	711	0

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

5.2. Comparison of MSE

The simulation experiment given in 5.1 is repeated 10,000 times and the MSE of each estimator is computed as

$$MSE(\hat{\alpha}_i) = \frac{1}{10000}\sum_{j=1}^{10000}(\hat{\alpha}_{ij} - \alpha_i)^2 \tag{5.2}$$

where, $\hat{\alpha}_{ij}$ denote the estimate of the ith parameter in jth replication and α_i is a true parameter value. MSE ratio of JRM estimator over M-estimator, OJR, ORR and OLS estimator is tabulated in Table 5.2.

From Table 5.2,

- When there is no outlier in the data, MSE ratio of JRM over OJR and ORR estimator is greater than one.
- When there is no outlier in the data, efficiency of ORR estimator over the JRM estimator increases when the degree of multicollinearity and error variance increases.
- For outlier data, JRM estimator is more efficient than OLS, ORR and OJR estimators.
- For fixed sample size, efficiency of JRM estimator over M-estimator increases with increase in the degree of multicollinearity and error variance.

Table 5.2. Ratio of MSE of JRM, M-estimator, OJR, ORR and OLS estimators.

ρ	σ^2	$\frac{JRM}{M-est}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$	$\frac{JRM}{M-est}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$
		without outlier				with outlier			
		n = 20							
0.9	1	0.9996	1.0697	1.0891	1.0693	0.9996	0.0002	0.0002	0.0002
	9	0.9828	1.0700	1.2079	1.0510	0.9844	0.0017	0.0020	0.0017
	25	0.9395	1.0731	1.3387	1.0063	0.9472	0.0043	0.0056	0.0039
	100	0.8288	1.0688	1.5526	0.8784	0.8528	0.0114	0.0181	0.0084
	400	0.7416	1.0768	1.7004	0.7855	0.7607	0.0206	0.0360	0.0118
	900	0.7188	1.0879	1.7508	0.7686	0.7329	0.0233	0.0412	0.0126
0.99	1	0.9801	1.0662	1.2157	1.0441	0.9831	0.0002	0.0002	0.0002
	9	0.8396	1.0805	1.5643	0.8998	0.8682	0.0018	0.0027	0.0014
	25	0.7641	1.0814	1.6891	0.8144	0.8099	0.0047	0.0078	0.0031
	100	0.7140	1.0931	1.7726	0.7649	0.7587	0.0141	0.0246	0.0076
	400	0.6994	1.0918	1.7999	0.7470	0.7201	0.0219	0.0394	0.0109
	900	0.6958	1.0933	1.8024	0.7432	0.7004	0.0251	0.0456	0.0122
0.999	1	0.8309	1.0766	1.5707	0.8867	0.8669	0.0002	0.0003	0.0002
	9	0.7135	1.0947	1.7828	0.7652	0.7744	0.0021	0.0034	0.0012
	25	0.6963	1.0891	1.7985	0.7416	0.7604	0.0053	0.0090	0.0030
	100	0.6921	1.0957	1.8149	0.7401	0.7399	0.0136	0.0242	0.0070
	400	0.6901	1.0993	1.8175	0.7388	0.7103	0.0222	0.0404	0.0108
	900	0.6949	1.0917	1.7975	0.7413	0.7022	0.0243	0.0443	0.0120
		n = 30							
0.9	1	0.9999	1.0585	1.0701	1.0583	0.9999	0.0003	0.0003	0.0003
	9	0.9932	1.0645	1.1586	1.0572	0.9933	0.0021	0.0024	0.0021
	25	0.9663	1.0578	1.2682	1.0215	0.9688	0.0050	0.0063	0.0048
	100	0.8641	1.0653	1.5212	0.9161	0.8749	0.0132	0.0207	0.0106
	400	0.7420	1.0731	1.7342	0.7864	0.7517	0.0223	0.0404	0.0134
	900	0.7004	1.0822	1.8223	0.7441	0.7085	0.0259	0.0484	0.0134
0.99	1	0.9922	1.0634	1.1639	1.0549	0.9924	0.0002	0.0003	0.0002
	9	0.8767	1.0663	1.5100	0.9307	0.8904	0.0022	0.0033	0.0018
	25	0.7780	1.0719	1.6926	0.8263	0.8048	0.0057	0.0099	0.0039
	100	0.6914	1.0779	1.8437	0.7319	0.7191	0.0161	0.0302	0.0084
	400	0.6659	1.0842	1.8929	0.7073	0.6764	0.0244	0.0469	0.0116
	900	0.6556	1.0877	1.9132	0.6982	0.6594	0.0275	0.0530	0.0127
0.999	1	0.8646	1.0622	1.5234	0.9139	0.8827	0.0003	0.0004	0.0002
	9	0.6979	1.0877	1.8469	0.7469	0.7464	0.0027	0.0049	0.0015
	25	0.6730	1.0808	1.8708	0.7139	0.7109	0.0066	0.0119	0.0034
	100	0.6586	1.0819	1.9021	0.6979	0.6874	0.0164	0.0312	0.0079
	400	0.6575	1.0854	1.9035	0.6982	0.6613	0.0248	0.0483	0.0113
	900	0.6571	1.0854	1.9080	0.6972	0.6538	0.0278	0.0537	0.0126

Table 5.2 contd. ...

ρ	σ^2	$\frac{JRM}{M-est}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$	$\frac{JRM}{M-est}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$
		without outlier				with outlier			
					n = 50				
0.9	1	1.0000	1.0590	1.0660	1.0589	1.0000	0.0004	0.0004	0.0004
	9	0.9977	1.0535	1.1098	1.0511	0.9977	0.0030	0.0032	0.0030
	25	0.9869	1.0557	1.1942	1.0417	0.9873	0.0069	0.0081	0.0068
	100	0.9132	1.0557	1.4278	0.9622	0.9169	0.0161	0.0238	0.0141
	400	0.7753	1.0698	1.7102	0.8237	0.7751	0.0261	0.0472	0.0173
	900	0.7027	1.0690	1.8198	0.7422	0.7057	0.0322	0.0621	0.0177
0.99	1	0.9973	1.0548	1.1175	1.0520	0.9974	0.0004	0.0004	0.0004
	9	0.9257	1.0612	1.4119	0.9811	0.9295	0.0030	0.0043	0.0027
	25	0.8208	1.0640	1.6392	0.8688	0.8328	0.0072	0.0124	0.0054
	100	0.6962	1.0717	1.8618	0.7367	0.7143	0.0196	0.0381	0.0108
	400	0.6471	1.0798	1.9497	0.6860	0.6453	0.0314	0.0641	0.0141
	900	0.6318	1.0825	1.9875	0.6703	0.6374	0.0356	0.0723	0.0160
0.999	1	0.9175	1.0574	1.4300	0.9685	0.9211	0.0004	0.0005	0.0003
	9	0.7053	1.0703	1.8439	0.7453	0.7261	0.0035	0.0068	0.0020
	25	0.6554	1.0763	1.9402	0.6934	0.6812	0.0088	0.0172	0.0045
	100	0.6312	1.0736	1.9709	0.6651	0.6464	0.0213	0.0431	0.0095
	400	0.6271	1.0805	1.9879	0.6636	0.6314	0.0333	0.0681	0.0146
	900	0.6211	1.0797	2.0016	0.6565	0.6275	0.0367	0.0750	0.0158
					n = 100				
0.9	1	1.0000	1.0559	1.0593	1.0559	1.0000	0.0006	0.0006	0.0006
	9	0.9995	1.0486	1.0763	1.0481	0.9995	0.0052	0.0054	0.0052
	25	0.9965	1.0521	1.1270	1.0484	0.9965	0.0113	0.0124	0.0113
	100	0.9653	1.0596	1.3011	1.0224	0.9653	0.0241	0.0315	0.0229
	400	0.8452	1.0592	1.5891	0.8915	0.8423	0.0381	0.0647	0.0293
	900	0.7503	1.0616	1.7547	0.7904	0.7545	0.0449	0.0847	0.0282
0.99	1	0.9994	1.0546	1.0847	1.0539	0.9994	0.0006	0.0006	0.0006
	9	0.9713	1.0515	1.2718	1.0211	0.9722	0.0050	0.0064	0.0048
	25	0.8981	1.0556	1.4864	0.9463	0.9019	0.0114	0.0177	0.0098
	100	0.7415	1.0637	1.8023	0.7824	0.7435	0.0291	0.0564	0.0179
	400	0.6472	1.0728	1.9759	0.6843	0.6433	0.0467	0.0973	0.0220
	900	0.6190	1.0697	2.0193	0.6530	0.6207	0.0532	0.1123	0.0233
0.999	1	0.9671	1.0564	1.2945	1.0211	0.9681	0.0006	0.0008	0.0006
	9	0.7519	1.0621	1.7827	0.7923	0.7640	0.0055	0.0104	0.0036
	25	0.6639	1.0659	1.9371	0.6998	0.6802	0.0144	0.0293	0.0075
	100	0.6136	1.0754	2.0454	0.6490	0.6225	0.0335	0.0707	0.0145
	400	0.6008	1.0706	2.0533	0.6320	0.6018	0.0503	0.1070	0.0207
	900	0.6030	1.0754	2.0633	0.6365	0.5996	0.0534	0.1138	0.0221

5.3. Comparison of MSE for more than one outlier

The same simulation experiment is repeated 10,000 times for all possible combinations of ρ and σ^2 for sample size 30 by introducing one, two and three outliers. The MSE ratio of JRM over M-estimator, OJR, ORR and OLS estimators is computed and reported in the following table.

From Table 5.3 it is clear that

- JRM estimator becomes more and more efficient as compared to OJR, ORR and OLS estimators as the number of outliers in the data increases.
- JRM estimator is better than M-estimator, OJR, ORR and OLS estimators for different degrees of multicollinearity and number of outliers in the data.

The same fact is also depicted graphically. Figure 5.1 to 5.6 show MSE ratio of JRM to OJR, ORR and OLS for σ^2 = 1 and 9, with different degrees of multicollinearity and number of outliers.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

Table 5.3. Ratio of MSE of estimators in presence of multicollinearity and outliers.

ρ	σ^2	With 1 outlier				With 2 outliers				With 3 outliers			
		$\frac{JRM}{Mest}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$	$\frac{JRM}{Mest}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$	$\frac{JRM}{Mest}$	$\frac{JRM}{OJR}$	$\frac{JRM}{ORR}$	$\frac{JRM}{OLS}$
0.9	1	0.9999	0.0003	0.0003	0.0003	0.9999	0.0001	0.0001	0.0001	0.9999	0.0001	0.0001	0.0001
	9	0.9933	0.0021	0.0024	0.0021	0.9933	0.0012	0.0013	0.0012	0.9933	0.0009	0.0010	0.0009
	25	0.9688	0.0050	0.0063	0.0048	0.9712	0.0030	0.0037	0.0028	0.9719	0.0022	0.0027	0.0021
	100	0.8749	0.0132	0.0207	0.0106	0.8859	0.0081	0.0124	0.0066	0.8951	0.0064	0.0096	0.0054
	400	0.7517	0.0223	0.0404	0.0134	0.7629	0.0147	0.0260	0.0090	0.7766	0.0121	0.0212	0.0076
	900	0.7085	0.0259	0.0484	0.0134	0.7135	0.0172	0.0320	0.0092	0.7146	0.0141	0.0258	0.0077
0.99	1	0.9924	0.0002	0.0003	0.0002	0.9926	0.0001	0.0002	0.0001	0.9923	0.0001	0.0001	0.0001
	9	0.8904	0.0022	0.0033	0.0018	0.8974	0.0012	0.0018	0.0010	0.9057	0.0009	0.0013	0.0008
	25	0.8048	0.0057	0.0099	0.0039	0.8218	0.0031	0.0052	0.0022	0.8382	0.0024	0.0039	0.0018
	100	0.7191	0.0161	0.0302	0.0084	0.7480	0.0095	0.0172	0.0054	0.7702	0.0073	0.0128	0.0044
	400	0.6764	0.0244	0.0469	0.0116	0.6957	0.0163	0.0307	0.0080	0.7130	0.0132	0.0242	0.0068
	900	0.6594	0.0275	0.0530	0.0127	0.6704	0.0181	0.0348	0.0086	0.6865	0.0156	0.0295	0.0075
0.999	1	0.8827	0.0003	0.0004	0.0002	0.8923	0.0001	0.0002	0.0001	0.8988	0.0001	0.0002	0.0001
	9	0.7464	0.0027	0.0049	0.0015	0.7680	0.0014	0.0024	0.0008	0.7926	0.0011	0.0018	0.0007
	25	0.7109	0.0066	0.0119	0.0034	0.7476	0.0036	0.0063	0.0021	0.7626	0.0026	0.0045	0.0016
	100	0.6874	0.0164	0.0312	0.0079	0.7121	0.0097	0.0178	0.0050	0.7449	0.0077	0.0137	0.0043
	400	0.6613	0.0248	0.0483	0.0113	0.6867	0.0165	0.0313	0.0079	0.6990	0.0135	0.0250	0.0067
	900	0.6538	0.0278	0.0537	0.0126	0.6642	0.0188	0.0365	0.0085	0.6795	0.0156	0.0296	0.0074

Figure 5.7 to 5.12 present the graph of MSE ratio of JRM estimator with M-estimator versus degree of multicollinearity for one, two and three outliers. It shows that for various values of σ^2, efficiency of JRM estimator over M-estimator increases as degree of multicollinearity increases. It is also clear that, as error variance increases, JRM estimator becomes more efficient than M-estimator.

6. Conclusion

In this article, we have introduced a new estimator for estimating regression coefficients of a linear regression model in the presence of multicollinearity and outliers. Theoretically, we have derived some conditions for comparison of the performance of proposed estimator with respect to the other three estimators ORR, OJR and M-estimators in the MSE sense and this study indicates that the performance of JRM estimator is better than the OLS, ORR, OJR and M-estimator in presence of multicollinearity and outliers.

It should be pointed here that we proposed the new estimator in the presence of multicollinearity and outliers present in the response variable, called Y-outlier. As is well-known (Montgomery, Peck and Vining, 2006), there are three types of outliers that may be present in the data. One type is Y-outlier, another one is X-outlier (outliers present in predictors) and the third is XY-outlier (outliers present in both response and predictors). Huber's landmark book on Robust Statistics (1981) focused on Y-outlier and he proposed M-estimator as a robust estimator of regression parameters. Thereafter, major work in this area was done by Rousseeuw and Leroy (1987). He proposed several robust estimators like least median squares (LMS), least trimmed squares (LTS) and so on which are robust to Y as well as to X-outliers.

This article assumes the presence of Y-outlier and so we considered Huber's M-estimator. If there are outliers present in the predictors' space, then one can use LTS type estimator in place of M-estimator and with some suitable modification, further analysis can be done.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

BARR. BALASAHEB KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

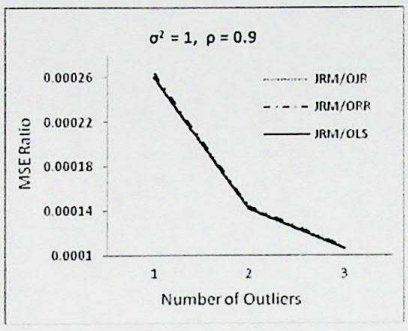


Figure 5.1

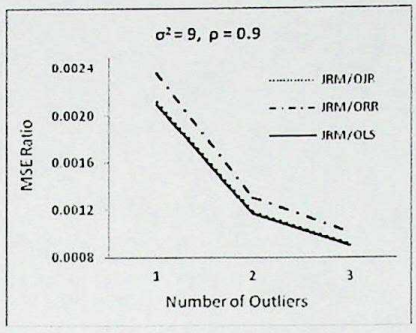


Figure 5.2

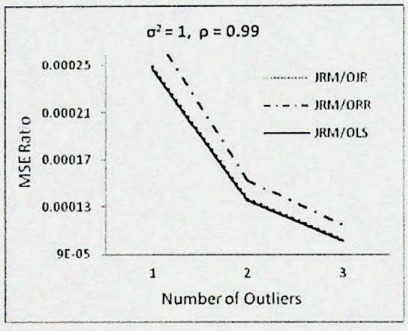


Figure 5.3

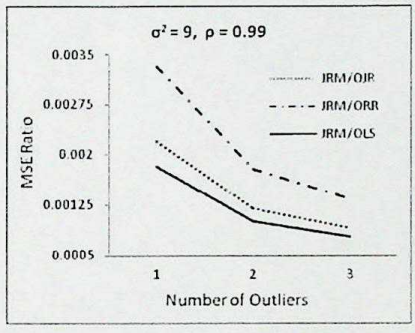


Figure 5.4

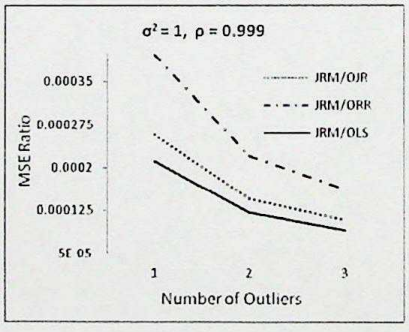


Figure 5.5

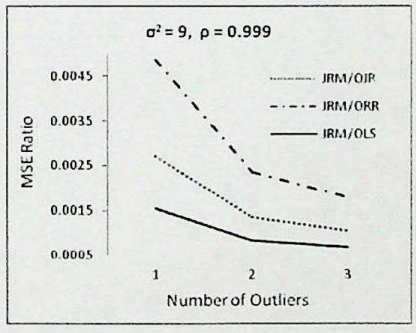


Figure 5.6

Acknowledgements

The authors wish to express their gratitude to the editor and the referees for their critical reading of the paper, useful comments and suggestions. The authors also thank Dr. S. R. Kulkarni for his helpful suggestions and encouragement. Authors acknowledge support from UGC Major Research Project.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

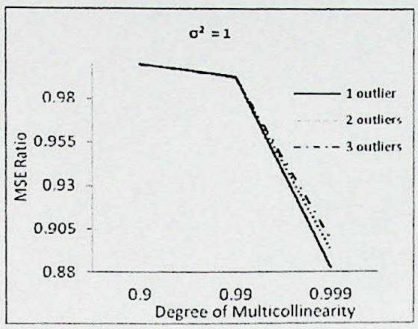


Figure 5.7

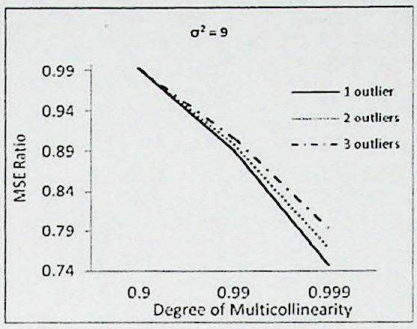


Figure 5.8

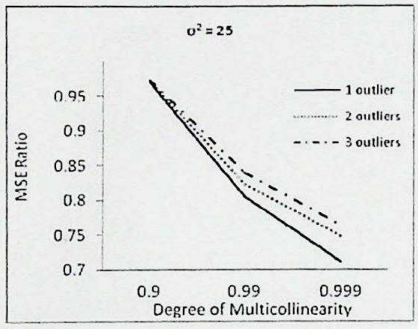


Figure 5.9

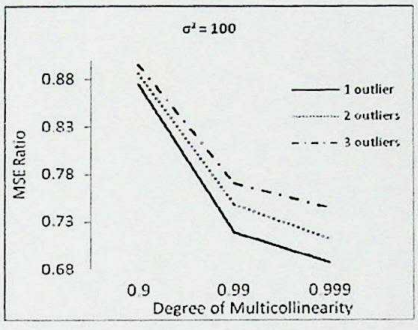


Figure 5.10

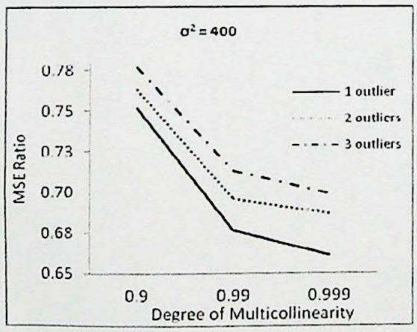


Figure 5.11

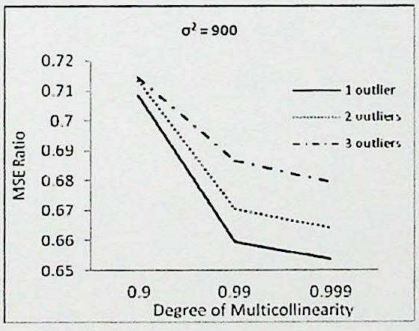


Figure 5.12

References

Draper, N.R., Smith, H., 1998. *Applied Regression Analysis, Third edition.* John Wiley, New York.

Hampel, F.R., Ronchetti, E.M., Rousseeuvw, P.J., Stahel, W.A., 1986. *Robust Statistics: The Approach Based on Influence Function.* John Wiley, New York.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics,* 12(1), 55–67.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012

Hoerl, A.E., Kennard, R.W., Baldwin, K.F., 1975. Ridge regression: some simulations. *Communications in Statistics – Theory and Methods*, 4(2), 105–123.

Huber, P.J., 1973. Robust Regression: asymptotics, conjectures, and Monte Carlo. *Ann. Stat.*, 1, 799–821.

Huber, P.J., 1981. *Robust Statistics*. John Wiley, New York.

McDonald, G.C., Galarneau, D.I., 1975. A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350), 407–416.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2006. *Introduction to Linear Regression Analysis, Third edition*. John Wiley and sons, New York.

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

Sakallioğlu, S., Kaçiranlar, S., 2008. A new biased estimator based on ridge estimation. *Stat. Papers*, 49, 669–689.

Singh, B., Chaubey, Y.P., Dwivedi, T.D., 1986. An almost unbiased ridge estimators. *Sankhya*, 48(3), 342–346.

Vinod, H.D., Ullah, A., 1981. *Recent Advance in Regression Methods*. Marcel Dekker, New York.

Downloaded by [Shivaji University], [Nileshkumar Jadhav] at 23:55 17 February 2012