Chapter 0

A Brief Outline of the Project

0.1 Introduction:

Francis Galton introduced the concept of regression in a paper on "Typical Laws of Heredity" presented in England in 1877. At present, it is most widely used statistical technique for investigation and modeling the relationship between variables. One of the primary objectives of regression analysis is to develop an equation that will help an investigator to predict the response variable as accurately as possible for given values of regressors. For example, prediction of rainfall in a particular region is an important problem. But accurate prediction is based on a set of weather variables which are large in number. Some of them may be relevant, while some may not be. Suppose, all the variables are included in the prediction equation, then cost of data collection and analysis goes up and hence the resulting model will not be economical. Many researchers have pointed out that large number of regressors will decrease prediction accuracy. Hence, subset selection becomes a vital task while modeling the data

Many researchers have reported variable selection techniques when response variable is continuous and normally distributed. The main objective of the project entitled

"A STUDY OF SUBSET SELECTION TECHNIQUES IN REGRESSION AND APPLICATIONS"

is to introduce new subset selection criteria in the regression. In the literature, various subset selection methods are available which perform well under some basic assumptions like errors are normal with constant variance. Violation of such assumptions may affect the subset selection criteria.

So we have studied techniques for subset selection in regression under the different situations such as: data contains outlier observations and multicollinearity, error follows non-normal distributions, response variable follows Poisson distribution etc.

In this study, we have focused on the following objectives

(i) to take review of various subset selection techniques in normal and non-normal setup.

(ii) to compare performance of existing subset selection methods like Mallows' C_p-statistic in various situations.

(iii) to explore possibilities of providing new/ modified subset selection technique(s).

(iv) to develop relevant computer programs.

0.2 An Outline of the Work Done:

In this report, the subset selection problem in multiple linear regression is addressed. 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. Most of them are based on ordinary least squares estimation method. These methods fail to select proper subset when some or more assumptions of regression model are violated. Hence, there would be a scope to modify or develop new methods which deal with the problem of such violations.

In this study, the problem of multicollinearity, outlier and non-normal distribution of error variable is traced and how-well known subset selection method (Mallows'C_p) fails to select proper subset is illustrated with suitable example.

We consider the problem of multicollinearity and outliers in response variable in Chapter 2. A new estimator is proposed to deal with this problem. An article "A Jackknifed Ridge M-estimator for Regression Model with

Multicollinearity and Outliers" is prepared and presented in the National Conference organized by Bangalore University, Bangalore. This article is published in Journal of Statistical Theory and Practice. We thank 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 devise a new variable selection technique when the data with multicollinearity or outliers or both. We have defined the new estimator to combat the problem of multicollinearity and outlier, a subset selection method is developed based on this estimator. The developed 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, 2010a) methods. An article "Subset Selection in Multiple Linear Regression in Presence of Outlier and Multicollinearity" is prepared and presented in International Conference organized by North Carolina State University, Raleigh (USA). This article will be communicated to for possible publication. The work related to this article is enclosed in Chapter 3 of this report.

Non-normal distribution of error variable is another problem for subset selection in regression model. In most of the practical situations, normality assumption is violated or there are some real datasets which does not support normality assumption. Very few subset selection methods address the non-normality of error variable. The 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 carried out with the help of Q-Q plots and chi-square test for goodness of fit. Also, a stepwise procedure is proposed when the number of variables is considerably large. This work is presented in the International Conference organized by Cochin University of Science and Technology,

Cochin and is given in Chapter 4 of this report. An article "Subset Selection in Multiple Linear Regression for Non-normal Symmetric Error Terms" is prepared and will be communicated for possible publication.

In Chapter 5, 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 illustrated with suitable example. Work of this chapter is presented in National Conference organized by North Maharashtra University, Jalgaon. An article "Subset Selection in Poisson Regression" is prepared and published in Journal of Statistical Theory and Practice.

0.3 Visits:

A. <u>Reference Work and Discussions</u>:

1. Indian Statistical Institute, Kolkata: 20.03.2010 to 27.03.2010.
2. Bangalore University, Bangalore: 23.10.2010 to 25.10.2010.
3. Indian Statistical Institute, Pune: 11.10.2011 to 12.10.2011.
4. Solapur University, Solapur: 04.11.2011 to 05.11.2011.
5. Dr. Babasaheb Ambedkar Marathwada University Aurangabad: 29.01.2012 to 30.01.2012.

B. <u>Discussion with Invited Persons:</u>

1. Dr. S.V. Kakade, Krishna Institute of Medical Science, Karad: 22.03.2011 to 26.03.2011.
2. Dr. S.R. Kulkarni, Innovative Labs, Bangalore: 05.04.2011 to 08.04.2011 and 27.11.2011 to 30.11.2011.
3. Dr. A.V. Dorugade, Yashvantrao Chavan College, Halkarni, Kolhapur: 01.11.2011 to 02.11.2011.

C. <u>Paper Presentations at Conferences / Seminars</u>:

1. National Conference on 'Industrial Statistics and Computational Mathematics'. *Jalgaon, January 21-23, 2010.*

BARR. B. [illegible] KHA DEK[illegible] LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

2. International Conference on 'Probability and Data Mining'. *Raleigh, USA, April 21-24, 2011.*
3. National Conference on 'Advances in Statistical Theory and Applications'. *Bangalore, November 16-18, 2011.*
4. International Conference on 'Statistics Probability and Related Areas'. *Cochin, December 19-22, 2011.*
5. National Conference on 'Interface between Statistics and Other Sciences'. *Aurangabad, February 24-25, 2012.*

0.4 Purchase of Equipments and Books:

A. Equipments:

Sr. No.	Equipment Title	Quantity	Price
1	Minitab – 15 Single User **Software**	1	16, 920/-
2	Microsoft Office 2010 **Software**	2	7, 588/-
3	Ricoh Aficio SP 1100s **Printer**	1	14, 756/-
4	Acer type-IV Desktop **Computer**	1	33, 500/-
5	Dell – 15 **Laptop**	1	49, 920/-
6	320 GB **Hard Drive Disc**	1	3, 417/-
7	**Speaker**	1	1, 598/-
8	**Wireless Key board with Mouse**	1	1, 700/-
9	**USB Pen Drive**	1	550/-
		Total	**1, 29, 949/-**

B. Books:

Sr. No.	Book	Author	Price (in Rs.)
1	Robust Regression	D.L. Kenneth, L.A. Jeffrey.	7, 343/-
2	Negative Binomial Regression	J. Hilbe.	2, 702/-
3	Applied Regression Analysis and Generalized Linear Models	J. Fox.	4, 241/-
4	Regression Analysis by Example	S. Chatterjee, A.S. Hadi.	5, 715/-

5	Applied Regression Analysis: A Research Tool	J.O. Rawlings, S.G. Pantula, D.A. Dickey.	8, 677/-
6	The Nature of Statistical Learning Theory	V.N.Vapnik.	
7	Robust Diagnostic Regression Analysis	A. C. Atkinson, M. Riani.	15, 184/-
8	Regression Modeling Strategies: With Application to Linear Models, Logistic Regression and Survival Analysis	F. E. Harrell.	
9	Log-linear Models and Logistic Regression	R. Christensen.	
10	Linear Models and Meneralizations	C.R. Rao, H. Toutenburg.	15, 496/-
11	Modern Regression Techniques Using R: A Practical Guide	D. B Wright, K. London.	
12	Introductory Statistics with R	P. Dalgaard.	
13	Logistic Regression: Self Learning Text	D. G. Kleinbaum, M. Klein.	
14	Robust Estimation and Hypothesis Testing	M. L. Tiku.	337/-
15	Quantile Regression	R. Koenker.	4, 598/-
16	Least Squares Support Vector Machines	J.A.K.Suykens, T.V.Gestel, J.D.Brabanter.	
17	Semiparametric Regression	D. Ruppert, M. P. Wand, R. J. Carroll.	11, 150/-
18	The Elements of Statistical Learning-Data Mining, Inference, and Prediction	T. Hastie, R. Tibshirani, J. Friedman.	
19	Regression Analysis of Count Data	A.C.Cameron, P.K.Trivedi.	
20	Logistic Regression: From Introductory to Advanced Concepts and Applications	S. Menard.	
21	Reliability , Life Testing and Prediction of Service Lives	S.C. Sauders.	18, 223/-
22	Functional Estimation for Density, Regression Models and Processes	O. Pons.	
23	Spatial Statistics and Spatio-Temporal Data	M. Sherman.	
24	Large Sample Techniques for Statistics	J. Jiang.	
25	Testing Statistical Hypotheses	E.L. Lehmann, J.P. Romano.	6, 334/-
26	40 Puzzles and Problems in Probability and Mathematical Statistics	W. Schwarz.	
27	Statistical Methods for the Analysis of Repeated Measurements	C.S. Davis.	

28	Numerical Optimization	J. Nocedal, S.J. Wright.	
29	Analyzing Categorical Data	J.S. Simonoff.	
30	Bayesian Core: A Practical Approach to Computational Bayesian Statistics	J.M. Marin, C.P. Robert.	
31	A First Course in Real Analysis	S.K. Berberian.	
32	Probability through Problems	M.Capinski T. Zastawniak	
33	Reliability Theory	I.Gertsbakh	
34	An Introduction to Bayesian Analysis: Theory and Methods	J.K.Ghosh, M.Delampady, T. Samanta.	
35	Time Series Analysis: with Applications in R	J. D. Cryer Kung-Sik Chan	
36	Logistic Regression: A Self Learning Text	D.G. Kleinbaum, M. Klein.	
37	Matrix: Algebra, Calculus and Generalized Inverse Part I and Part II	A.K. Hazra.	
Total			**1, 00, 000/-**

0.5 Appointment of Project Fellow(s):

Mr. Deepak M. Sakate joined as a Project Fellow (P.F.) on 21st July, 2009 and resigned on 19th December 2009. Mr. Nileshkumar H. Jadhav joined as a P.F. on 13th January, 2010 and worked up to 31st January 2012, the last date of the project. The work carried out by both of them is satisfactory. They developed computer programs to obtain the new estimation procedures and carried out extensive simulation studies to study the performance of the proposed estimators. Also, they have developed source code for new subset selection methods.