EFFICIENT MODEL SELECTION THROUGH STANDARD OPERATING PROCEDURE USING HYBRID OF SPARSE AND ROBUST ESTIMATORS

ANAM JAVAID

UNIVERSITI SAINS MALAYSIA

2020

EFFICIENT MODEL SELECTIONN THROUGH STANDARD OPERATING PROCEDURE USING HYBRID OF SPARSE AND ROBUST ESTIMATORS

by

ANAM JAVAID

Thesis submitted in fulfillment of the requirements
For the degree of
Doctor of Philosophy

November 2020

ACKNOWLEDGEMENT

I am grateful to ALLAH Almighty for giving me this opportunity to study abroad at a higher level of Ph.D. and for guiding me on this path to make my journey smoother, and I would also like to express my sincere gratitude to my supervisor, Associate professor Dr. Mohd Tahir Ismail and Co-Supervisor, Dr. Majid Khan bin Majahar Ali for all their help, suggestions, comments and guidance on this full journey to completion of the thesis as without their guidance I will not be able to accomplish this task. I am also grateful to the Women University, Multan, Pakistan for providing me with financial support for studying abroad and the University Sains Malaysia for providing me an opportunity to study in this reputable university. Special thanks to my colleagues and friends for their assisting along the journey.

Lastly, this thesis is dedicated to my husband, my daughter and members of my family for being always with me and for their patience and support.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LI ST OF SYMBOLS

n	Number of sample size
p	Number of Variables
k	Number of variables left in the model
R^2	Coefficient of determination
L1	Manhattan Distance or Taxicab norm (sum of the magnitudes of the vectors in a space)
L2	Euclidean norm(Calculated as Euclidean distance from the origin)
Lp	Provides alternatives to least squares for estimating the coefficients of a linear regression model.

LIST OF ABBREVIATIONS

SOP	Standard operating Procedure
SSE	Sum of Square of Error
MAPE	Mean absolute Percentage Error
IoT	Internet of things
AIT	Agriculture Information Technology
CGMs	Crop Growth Models
OLS	Ordinary Least Square
ANOVA	Analysis of Variance
LASSO	Least Absolute Shrinkage and Selection Operator
8SC	Eight Selection Criteria
9SC	Nine Selection Criteria
RMSE	Root Mean Square Error
QTL	Quantitative Trait Loci
DMR	Delete or merge regressor
P-PLS	*PRESS* statistics of PLS regression
CAS-ANOVA	For Collapsing and Shrinkage in ANOVA
DASSO	Connection between Dantzing selector and Lasso
NIR	Near Infrared Instruments
RMSEP	Root mean square error of prediction
FLARE	Family of LASSO regression
LAD LASSO	Least absolute deviation LASSO
SQRT LASSO	Square root LASSO
ZINB	Zero Inflated Negative binomial
EM-Algorithm	Expectation- Maximization algorithm
BIC	Bayes Information Criteria
RF	Random Forest
ARIMA	Autoregressive integrated moving average
MAE	Mean Absolute Error
RE	Reduction error
SMAPE	Symmetric Mean Absolute Percentage Error
SVR	Support Vector Regression

GAM	Generalized Additive Model
CSDH	Chronic subdural hematoma
GWL	Geographically Weighted LASSO
GWR	Geographically Weighted Regression
EMD	Empirical mode decomposition
VIF	Variance Inflation Factor
GM	Generalized M
TRME	Robust two parameter ridge M-estimator
CN	Condition Number
VDP	Variance Decomposition Proportion
LTS	Least Trimmed Squares
LAD	Least Absolute Deviation
StARS	Stability Approach to Regularization Selection
CAIC	Consistent AIC
AIC	Akaike Information criteria
MM	Majorize-Minimization
S estimator	Estimator of Scale
Hqreg	Huber loss penalized by lasso or elastic net
E.Net	Elastic net
E.netLTS	Elastic net with LTS estimator
Pense	Elastic net with S estimator
MPense	M step Elastic net with S estimator
SVD	Singular value decomposition
MLE	Maximum Likelihood Estimators
IRLS	Iteratively Reweighted Least Squares
LMS	Least Median of Squares
AIC$_c$	Corrected AIC
PRESS	Predicted Residual Sum of Squares
FPE	Final Prediction Error
HQ	Hanan and Quinn Information Criteria
LTS-ELA	Trimmed square elastic net
LTS-LASSO	Trimmed square LASSO
SCAD	Smoothly Clipped absolute Deviation

LIST OF APPENDICES

PEMILIHAN MODEL CEKAP MELALUI PROSEDUR OPERASI PIAWAI MENGGUNAKAN PENGANGGAR HIBRID TIPIS DAN TEGUH

ABSTRAK

Internet untuk segalanya (IoT) menjadi lebih kritikal seiring dengan peredaran masa. Penggunaan produk yang berkaitan dengan IoT membantu mengurangkan usaha manusia dan dapat memberikan kualiti setinggi mungkin pada waktu minimum. Pengering suria adalah salah satu kegunaan IoT dalam sektor pertanian untuk pengeringan barang. Kajian ini memfokuskan pada pengenalpastian faktor-faktor yang mempengaruhi kecekapan pengering solar pemungut dan penyingkiran nisbah kelembapan rumpai laut. Prosedur Operasi Piawai (SOP) disediakan berdasarkan empat Fasa untuk tujuan ini. Model hibrid berdasarkan analisis regresi yang tipis dan teguh digunakan untuk tujuan ini. Enam jenis penganggar hibrid dibangunkan dengan menggunakan penganggar tipis dan teguh serta kombinasi terbaik dipilih untuk set data sederhana dan besar. Kesan hubungan dalam semua model yang mungkin dipertimbangkan terutamanya dalam kajian ini. Sembilan kriteria pemilihan (9SC) telah digunakan untuk pemilihan model yang cekap. Keseluruhan prosedur dijalankan dalam empat Fasa. Semua model yang mungkin akan dijalankan dalam Fasa 1; model terbaik akan dipilih dalam Fasa 2, Fasa 3 akan merangkumi pemilihan model yang cekap, dan kecekapan ramalan akan diuji pada Fasa 4. Penyusutan mutlak terkecil dan operator pemilihan (LASSO) dan jaringan Elastik (E.Net) digunakan sebagai penganggar tipis, sementara Huber M, Hample M dan Bisquare M digunakan untuk analisis yang teguh. Dalam set data medium, hibrid jaringan elastik dengan fungsi berwajaran penganggar Bisquare M didapati berguna. Sebaliknya, jaringan elastik dengan hibrid dari penganggar Hampel M memberikan hasil terbaik untuk analisis data yang besar. Perbandingan dibuat berdasarkan nilai ralat min kuasa dua (MSE) dan nilai ralat peratusan mutlak (MAPE). MAPE untuk model yang cekap dalam analisis data medium didapati 24.24. Sementara MAPE untuk model yang efisien didapati 9.216 dalam analisis data besar. Dapatan menunjukkan bahawa SOP yang dibangunkan lebih baik daripada kaedah lain yang ada dari segi kecekapan. SOP yang dikembangkan dapat digunakan dalam apa jua bidang dan juga untuk set data dimensi tinggi. Plot reja terpiawai juga diperhatikan untuk bukti sokongan.

EFFICIENT MODEL SELECTION THROUGH STANDARD OPERATING PROCEDURE USING HYBRID OF SPARSE AND ROBUST ESTIMATORS

ABSTRACT

The Internet of Things (IoT) is becoming more critical as time passes by. The use of IoT-related products helps to reduce human effort and can provide the highest possible quality at a minimum of time. Solar dryer is one of the uses of IoT in the agricultural sector for the drying of goods. This study focuses on the identification of factors affecting the collector's solar dryer efficiency and the removal of seaweed moisture ratio. The Standard Operational Procedure (SOP) is provided on the basis of four Phases for this purpose. A hybrid model based on a sparse and robust regression analysis is intended for this purpose. Six types of hybrid estimators are developed using sparse and robust estimators and the best combination is selected for the medium and large data set. Interaction effects in all possible models are primarily addressed in this study. Nine model selection criteria (9SC) have been used for the efficient selection of models. The whole procedure is carried out in four Phases. All possible models will be run in Phase 1; the best model will be selected in Phase 2, Phase 3 will include efficient model selection, and forecasting efficiency will be tested in Phase 4. The least absolute shrinkage and selection operator (LASSO) and Elastic net (E.Net) are used as a sparse estimator, while Huber M, Hample M and Bisquare M are used for robust analysis. In the medium data set, hybrid of elastic net with the weighted function of the Bisquare M estimator is found to be useful. On the other hand, the elastic net with a hybrid of the Hampel M estimator provided the best result for the large data analysis. The comparison is made on the basis of the mean square error (MSE) and the mean absolute percentage error (MAPE) values. The MAPE for an efficient model in the medium data analysis was found to be 24.24. While the MAPE for an efficient model is found to be 9.216 in the large data analysis. The results show that the developed SOP is better than the other existing methods in terms of efficiency. The developed SOP can be used in any field as well as for the high-dimensional data. Standardized residual plots are also observed for the supporting evidence.

CHAPTER 1

INTRODUCTION

1.1 Overview

The Internet of Things (IoT) includes physical device networks such as software, sensors, electronics and home appliances (Zhao et al., 2010). These types of items enable the data set to be collected, connected and exchanged. This type of process offers opportunities for more direct integration of the physical world into a computer-based system. The use of such technologies results in economic benefits, improved efficiency and a reduction in human effort (Malavade and Akulwar 2017). These technologies have a number of applications in different fields, such as health care, agriculture, transport, retail, supply chain management, environmental, infrastructure monitoring (Patil et al., 2012). The use of agricultural information technology (AIT) is one of the most efficient and effective tools used to improve agricultural productivity (Yan, 2011). Another example of cloud computing and IoT in agriculture and forestry was analysed by Bo and Wang, (2011). The incorporation of crop growth models (CGMs) into the IoT application system was proposed to make the agriculture system more intelligent and adaptive (Hu and Qian, 2011). Because by the use of IoT, data can be collected from a sensor for determining factors such as temperature, humidity, airspeed, water irrigation (Gondchawar and Kawitkar, 2016).

In agriculture, food insecurity is considered to be a major problem, with approximately 797 million people facing food insecurity problems (Ahmed et al., 2017). It is therefore obligatory to increase food production due to an increase in population and food insecurity problems (Rockstrom et al., 2009). There are many stages of crop management in the seeding process, such as nutrient supply, water, the crop production environment (Yan, 2011). The drying process is one of the important steps. Air drying is the most frequently used method in agriculture (Ali et al., 2014).

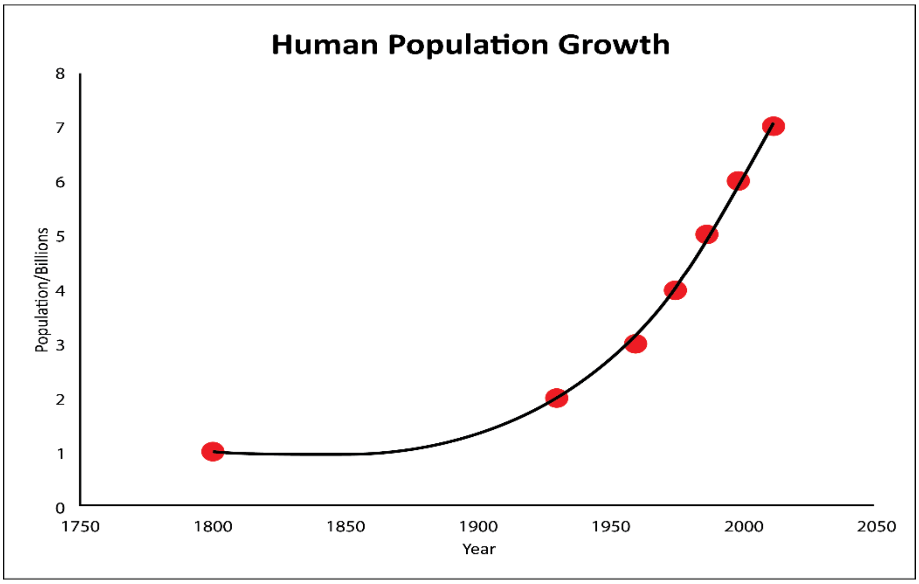

Figure 1.1 World Population: 1750-2050 (Source: United Nations Population Fund (UNFPA); Deborah Byrd in the Human world, 12 October 2019 update

Figure 1.1 shows an increase in population over the years, making it mandatory to take steps to reduce food insecurity problems in the coming years (Ahmad et al., 2017). One of the most important products used in agriculture or aquaculture is seaweed as shown in Figure 1.2 due to its use in food, fertilizers, cosmetics and industrial gums and chemical extraction. Ali et al. (2017) further investigated that carrageenan is also used as an item in human food and non-food, cosmetics, pet food, meat binders, etc.

Figure 1.2 Seaweed (Ali et al., 2014)

1.2 Problem Background

The issue of food insecurity is a problem of the study. The drying of seaweed is used for this purpose in the analysis. The main focus is to deal with the interaction effects of variables in the medium and large data analysis. In the real-life data set, when there is a need to deal with a lot of variables, the problem of multicollinearity and outliers may arise. In this research, the issue of multicollinearity and outliers are being addressed at the same time. Previous research has shown that no such model can be trusted for better forecasting, including interaction effects.

1.3 Research questions

Research questions for the current study can be defined as

Question 1: The model is efficient among all possible models with interaction effects.

Question 2: The efficient model can address multicollinearity issues, including interaction effects.

Question 3: In the case of outliers, including interaction effects, the efficiency of the model will not be affected.

Question 4: The developed hybrid model through SOP is better than any other existing hybrid models.

1.4 Objectives

In order to answer all the research questions, the objectives of this research are

- **a.** To analyze all the possible models including interaction effects.
- **b.** To remove the problem of multicollinearity in all possible models.
- **c.** To compare the efficiency of the models based on the percentage of outliers observed in each model.
- **d.** To develop an SOP based on four Phases for finding the best-performed hybrid model.

1.5 Scope

For forecasting, model selection has its importance in an era of life. This thesis focus on the model selection issue in the agriculture field. In previous research, ordinary least squares (OLS) are used by different researchers for the efficient forecasting (Zuur et al., 2009). There is also another kind of regression methods are used for forecasting. Multicollinearity is considered a big issue in regression analysis. But not every model can deal with this issue (Gujrati, 2004). Similarly, the presence of outliers is considered as a big issue in data analysis. So, robust regression methods are available to handle these kinds of problems (Huber, 1973). Some researchers have already been working on interaction effects (Ali et al., 2017). But through SOP, they developed an efficient model using an OLS. The current study focused on developing a hybrid model using a sparse and robust estimator. In this study, all possible models are involved. Due to the inclusion of all possible models, only six important variables are used in the analysis of the large data set. In literature, only a few people are working with the interaction effects (Abdullah et al., 2008). According to their study, it is noted that the interaction factors cannot be ignored in any study because sometimes the interaction effect can be significant even if the simple effect of two variables is non-significant.Thus in this study the interaction importance is also highlighted. By including the interaction effect, the study now consisted of a total of 63 variables. The developed SOP shows that an efficient model can deal with the issue of multicollinearity and outliers. In the context of the high dimensionality issue, the developed SOP can also provide efficient forecasting results.

1.6 Significance of the Study

The research project proposed an efficient model selection using hybrid approach of sparse and robust regression analysis. There is no such research has been done for interaction factors effect using a hybrid kind of approach. People are using SOP but with the simple OLS method (Abdullah et al., 2008). That method is used in this study for comparison purpose. The more efficient results are noted for the purposed technique as compared to the existing ones (Abdullah et al., 2008). Robust regression has no ability to deal with the high dimensional dataset, while the sparse regression can deal with it. So, as a result, the proposed hybrid model can deal with the high dimensional dataset in case of multicollinearity and outlier’s issues. Thus the main focus of the research is the hybrid efficient model selection through a proper SOP, that has not done yet in literature. The proposed procedure is used to to identify the key factors related to the dryness of seaweed. The factors affecting the removal of the moisture ratio and the factors affecting the collector efficiency are observed. The key significance of the research is the inclusion of the interaction factors in the analysis. It can also highlight the importance of dealing with all possible models. The four Phases provided for this project are in a position to obtain an efficient chosen model. The efficient model is now ready to predict factors related to the collector efficiency and the moisture ratio removal of seaweed using the solar dryer. The developed SOP that has been proposed can be used in any field of life. The issue of food insecurity can be addressed through the development of SOPs by considering the important factors in the field of agriculture.

1.7 Outlines

The thesis is organised as follows. Chapter 1 covers the background to the problem and the research gap. Chapter 2 discusses the types of regression analysis that researchers have already used. The theory regarding the methods applied in this research is discussed in Chapter 3. Chapter 4 makes provision for the application of the proposed medium data set procedure. Comparisons are observed with other existing estimators, and results are noted. Chapter 5 applies the proposed SOP to the large data set. The results are observed, and the comparison with other existing estimators is also made. Chapter 6 comprises the summary and conclusions based on the results of Chapter 4 and Chapter 5. It also discusses the future directions and limitations of the research project.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter contains the types of regression analysis that various researchers had critically discussed in the literature. The use of linear regression analysis in various fields has been discussed in Section 2.2. Section 2.3 reviews some literature concerning the analysis of logistic regression. Section 2.4 includes a discussion on using an analysis of sparse regression. The use of robust regression analysis with issue of multicollinearity has been discussed in Section 2.5. Section 2.6 describes various other forms of regression analyses used by previous researchers. The critical review in Section 2.7 and the summary in Section 2.8 are discussed. For the sample size issue, the sample size less than 50 is considered to be a small sample size (Maman et al. ,2017), whereas the sample size between 50 and 100 is considered to be a medium sample size (Hawa Yahay et al. ,2012) and the sample size greater than 100 is considered to be a large sample size (Hoffmann et al., 2017). p denotes the number of variables and n denotes the number of observations in which $p > n$ shows a high dimensional study and $p < n$ shows a low dimensional study (Maj-Kanska et al.,2015)

2.2 Review about linear regression analysis

The simplest form of regression analysis is considered in linear regression analysis. Although this type of regressions suffers from some disadvantages if some assumptions are not fulfilled Gujarati (2004) but these regression analysis has been used in many social sciences issues due to the simplicity of models and reduction of computational procedure. The current study is concerned with the agricultural and aquacultural field. So, the literature related to both fields is also observed in the application of linear regression analysis. Work of the researchers on low dimensional and higher dimensional issues is also observed in Table 2.1. In which the field of study can also be observed.

Table 2.1 Previous studies related to linear regression

Author(s) (year)	**Sample size**			**field of data set**		**Simulations**		**Data Dimension**		**SOP**
	small	**medium**	**large**	**Agriculture**	**Non agriculture**	**Monte Carlo**	**Other**	***p*<*n***	***p* > *n***	
Angelini et al. (2003)			✓	✓			✓	✓		
Budin et al., (2008)	✓				✓			✓		
Abdullah et al. (2008)			✓	✓				✓		✓
Ho et al. (2009)			✓				✓	✓	✓	
Abdullah et al., (2011)	✓			✓				✓		
Humann et al. (2012)			✓	✓				✓		
Williams et al.(2012)			✓	✓				✓		
Hawa Yahay et al., (2012)		✓			✓			✓		
Payus et al.,(2013)		✓			✓			✓		
Jilcott Pitts et al. (2013)			✓	✓				✓		
Rodríguez-Galino et al. (2014)			✓		✓	✓	✓	✓		
Efron(2014)	✓				✓			✓		
Vu et al., (2015)	✓				✓			✓		
Roozbeh et al. (2016)			✓		✓	✓		✓		
Maj-Kanska et al. (2015)	✓	✓	✓	✓			✓	✓		
Rischbeck et al.(2016)			✓	✓				✓		
Abdullah et al., (2016)	✓				✓			✓		✓

de Paula et al.(2017)		✓					✓	✓		
Martínez et al. (2017)	✓	✓	✓	✓	✓	✓		✓	✓	
Hoffmann et al. (2017)	✓			✓				✓	✓	

Table 2.1 provided the use of linear regression analysis in different fields of life. Angelini et al. (2003) proposed the best linear unbiased estimator using an orthogonal wavelet base later on Budin et al. (2008) used multiple linear regression analysis and found an efficient model using eight selection criteria (8SC) from all possible 32 models. After that, 8SC was used by Abdullah et al. (2008) to study the volumetric stem biomass of the tropical tree species. For this purpose, they used 80 of all possible models and compared Huber's and Newton's multiple regression analysis. Later on Ho et al. (2009) used the quantile regression technique to identify genes. They have changed the expression or variability, depending on the age pattern. Another example of all possible models was found in Abdullah et al. (2011) study, which used a polynomial regression analysis for 32 possible models. In their study, 8SC was used for efficient model selection.

Humann et al. (2012) used multiple linear regression to identify significant factors affecting hearing loss among farmers in the agricultural field. The analysis of variance was used by Williams et al. (2012) for the analysis of yields from different varieties of maize. Factors affecting yield production can be found in East Timor. From October 2008 to March 2009 in Manukan Island, Hawa Yahay et al. (2012) made the best selection of models for electrical conductivity levels for 59 samples. The use of linear, quadratic, cubic, quartic, quintic and sextic regression models for the observation of cholesterol-related factors can be observed in the Efron (2014) study. Subsequently, Vu et al. (2015) used a multiple regression method with a backward elimination procedure to forecast the monthly demand for electricity.

The Delete or Merge Regressor (DMR) algorithm was proposed by Maj-Kanska et al. (2015) for linear model selection. When both types of categorical and continuous variables are present in a data set. Abdullah et al . (2016) used all possible models for the study of log production. The data set has been taken from the Department of Statistics in Malaysia and the World Data Bank in Indonesia and Malaysia. Partial least square method was found to be better than multiple linear regression models in the yield prediction study of the barley data set (Rischbeck et al.,2016). De Paula et al. (2017) proposed a parallel regression analysis to minimise the multicollinearity problem. The data set was taken from a grain research laboratory in the near-infrared whole wheat sample for 250 variables. Predicted residual square error (PRESS) statistics for PLS regression (P-PLS) were proposed by Martínez et al. (2017) for best predictive model selection. Partial least square (PLS) regression was found to be good for the optimisation of soybean treatments in the Hoffmann et al. (2017) study.

2.3 Review about logistic regression analysis

Logistic regression analysis is another type of regression analysis used by different researchers. It uses a logit function in its basic form for modeling of binomial or multinomial response variables. Some highlights based on the use of logistic regression analysis are noted in this study**.** Table 2.2 shows some example of literatures using logistic regression analysis.

Table 2.2 Literature about the use of logistic regression

Author(s) **(Year)**	**Sample sizes**			**Field of study**		**simulation**		**Data dimension**		**SOP**
	small	**medium**	**large**	**Agriculture**	**Non agriculture**	**Monte Carlo**	**Other**	***p*<*n***	***p* > *n***	
Bergtold et al. (2011)					✓		✓	✓		
Guns and Vanacker (2013)			✓		✓			✓		
Yahaya et al. (2013)			✓		✓			✓		
Huda et al. (2017)			✓		✓			✓		
Ranganathan, P. et al., (2017)			✓		✓			✓		
Chen, W. et al., (2019)			✓		✓			✓		
Jie, M.A. et al., (2019)			✓		✓			✓		

Bergtold et al. (2011) investigated the impact of sample size on average bias and the estimation of parameter efficiency and later on, Guns and Vanacker (2013) found that rare event logistic regression was a good choice in land-covering changes in landslides controlling factors. Yahaya et al. (2013) used logistic regression analysis for all possible models using 8SC in their study. Multilinear regression has been used by Huda et al. (2017) for malware selection problems. Ranganathan et al. (2017) performed a multivariate logistic regression analysis to predict different factors for gestational hypertension. Subsequently, they highlighted the limitations of the technique applied by Chen et al. (2019) in the investigation of landslide susceptibility using a kernel-based logistic regression analysis. They found the method with a high predictive capability. Jie et al. (2019) found that there was no benefit in the study of the clinical risk prediction model from machine learning over logistic regression analysis.

2.4 Review related to sparse regression analysis

Sparse regression analysis has many advantages over the issue of variable selection. These types of regressions have the ability to automatically select variables by removing insignificant factors from the model (Tibshirani, 1996). There are different kinds of sparse methods that researchers have developed. Sparse regression was mostly used by medical researchers due to the increased number of factors in the analysis (Jiang et al., 2017). It has the ability to deal with the high dimensional issues Zhang et al. (2016). Some of them are examined here for the purpose of observing the application of such techniques in different fields of life. Table 2.3 provide some of the sparse methods used by researchers in previous studies.

Table 2.3 Literature about the sparse regression

Author(s) (Year)	**Regression analysis**		**Sample Size**			**Simulation**		**Data Dimensions**		**SOP**
	Sparse	**Others**	**Small**	**Med**	**Large**	**Monte Carlo**	**Others**	***p<n***	***p>n***	
Wang et al. (2007)	✓	✓		✓	✓		✓		✓	
Park and Hastie (2007)	✓	✓			✓			✓		
Li and Yin (2008)	✓	✓	✓	✓			✓	✓	✓	
Bondell and Reich (2009)	✓			✓				✓		
Witten and Tibshirani (2009)	✓	✓	✓				✓	✓		
James et al. (2009)	✓	✓		✓		✓		✓		
Xu and Ying (2010)	✓	✓	✓	✓	✓		✓			
Filzmoser et al. (2012)	✓	✓		✓			✓	✓		
Lockhart et al. (2014)	✓	✓		✓				✓		
Olson Hunt et al.(2014)	✓	✓			✓		✓	✓		
Gusnanto and Pawitan (2015)	✓	✓	✓		✓			✓	✓	
Li et al. (2015)		✓		✓			✓		✓	
Zhang et al.(2016)			✓				✓		✓	
Chiquet et al. (2016)	✓	✓			✓		✓	✓		
Mallick & Tiwari(2016)				✓	✓		✓	✓		
Jang & Anderson-Cook (2017)	✓	✓			✓			✓		
Liu et al (2017)	✓	✓			✓			✓		
Guo et al. (2017)	✓	✓			✓				✓	
Guo et al. (2017)	✓			✓				✓		

Pripp and Stanisic (2017)	✓		✓					✓		
Maman et al.(2017)	✓		✓					✓		
Setiyorini et al. (2017)		✓		✓				✓		
Sakurama (2017)	✓		✓					✓		
Hirose et al. (2017)	✓	✓			✓	✓	✓	✓		
Jiang et al. (2017)	✓		✓	✓	✓		✓	✓		
Ramsay et al. (2018)	✓	✓						✓		
Masselot et al., (2018)	✓	✓		✓				✓		
Gillard and Zhigljavsky (2018)	✓	✓		✓			✓		✓	

Wang et al. (2007) proposed LASSO with autoregressive error named as REGAR. After that the regularization path of L1 was proposed by Park and Hastie, 2007. Li and Yin (2008) took advantage of the L1 penalty by proposing sliced inverse regression analysis, a combination of L1 and L2 regularizations (squared penalty) that works well in case of predictors *p* exceeds the sample size *n*. CAS-ANOVA (for collapsing and shrinkage in ANOVA) is one example of the method proposed by Bondell and Reich, (2009). Later on, Witten and Tibshirani (2009) worked on the covariance-regularised regression that belonged to the family of high-dimensional prediction methods, while James et al. (2009) worked on a new algorithm called DASSO, which was a link between the Dantzing selector and LASSO. It consisted of the computational cost as the least angle of regression. Another median regression analysis with the LASSO type penalty was suggested by Xu and Ying (2010). Filzmoser et al. (2012) have chosen an adaptive degree of penalty and proposed a two-stage method for simultaneous estimation and variable selection after a sparse regression analysis has been used. Lockhart et al . (2014) used a 10-fold LASSO cross-validation to identify and predict schizophrenia-related factors. Olson Hunt et al . (2014) proposed all possible sparse partial least square for all tuning parameters through a set grid.

Filzmoser et al. (2012) used the sparse regression analysis in their study. One of the works on the FLARE package in R language was done by Li et al. (2015) as FLARE was a family of new high-dimensional regression methods such as the least absolute deviation LASSO (LAD LASSO), the square root LASSO (SQRT LASSO) and the Dantzig selector. A comparison was also made with LASSO by Zhang et al. (2016). They looked at an algorithm based on adaptive lasso plus dynamic shrinkage. Later on, Chiquet et al. (2016) continued the discussion of the paper by Tutz and Gertheiss

(2016). They demonstrated the importance of the coding effect in the regularised ordinal and categorical regression. After that Mallick and Tiwari (2016) proposed flexible penalised model of regression by fitting zero-inflated negative binomial (ZINB) of expectation maximisation (EM) algorithm with an adaptive lasso. They combined it with the ZINB regression using the BIC tunning parameter criteria. Work on the LASSO influence plot was done by Jang and Anderson (2017) as they examined the influence plot for understanding the contribution of the individual observations and the robustness of the results for the estimated values of model parameters.

One of the uses of the sparse regression analysis was observed in the study of Liu et al.(2017) as they predicted on-demand ride services using the LASSO, random forest (RF) model, autoregressive integrated moving average model (ARIMA) model, and support vector regression. Guo et al. (2017) worked on penalising regression for the detection of influenza epidemics. They used data from China from the Baidu search engine. They compared the ridge, LASSO, elastic net, ensemble ridge, ensemble lasso, and ensemble elastic net. Root mean square error (RMSE), mean absolute error (MAE), reduction error (RE) and symmetric mean absolute percentage error (SMAPE) was used for the model selection. For best model selection, Guo et al.(2017) fitted on support vector regression (SVR) algorithm, step-down linear regression model, gradient boosted regression tree algorithm, negative binomial regression model, LASSO, linear regression model and generalized additive model (GAM). They calculated RMSE and R^2 for all of them and found SVR as an effective tool. Pripp and Stanisic (2017) also performed LASSO regression in the chronic subdural hematoma (CSDH) study of 93 patients in the radiological department. They used 10-fold cross-validation and calculated predictive performance per area under the curve in R.

Maman et al. (2017) made a further comparison of LASSO with ridge and elastic net. MAE was used by a sample of eight healthy participants, consisting of 3 females and 5 males aged 18 to 62 years. One of the examples of comparisons is the study of Setiyorini et al. (2017) by proposing a geographically weighted LASSO (GWL). They used data set from three different surveys and compared the performance of GWL and geographically weighted regression (GWR) and found GWL to be better. One of the contributions in the literature on LASSO was Sakurama (2017)'s work on the leader selection problem of information control with velocity assignment in the multi-agent system network of heterogeneous time delays.

There are different types of LASSO regression analysis as Hirose et al. (2017) worked on γ –lasso based on the majorize-minimization algorithm for treatment of outliers. Later on Jiang et al. (2017) examined the exponential squared loss function, a robust estimation method for varying coefficient nonlinear model by properly selecting the tunning parameter. One of the contributions was from Ramsay et al.(2018) who investigated LASSO by using 10 fold cross-validation of target cognitive training for patients. Data set from recent patients with schizophrenia consisted of 42 observations was taken for analysis. Later on, Masselot et al. (2018) examined the regression procedure for empiric mode decomposition (EMD). It consisted of an EMD algorithm in a series of data. After that, the components in LASSO were used as a regression model.

Studies show that the analysis of sparse regression has been proposed and used by different researchers in terms of model selection, prediction and model accuracy.

2.5 Overview related to robust regression

Robust regression is used for outliers because it has the ability to identify outliers and minimises the outlier impact on the regression coefficients. In robust regression, certain weights are assigned to each observation in the form of influence function (Alma, 2011). A number of researchers used robust regression in previous studies, where some of the studies had been done with a multicollinearity problem.In Table 2.4 shows the studies related to robust regression with multicollinearity problem.

Table 2.4 Literature for robust regression analysis with multicollinearity issue

Author(s)/(Year)	**Sample Size**			**Simulations**		**Data dimensions**		**SOP**
	Small	**Medium**	**Large**	**Monte Carlo**	**Others**	***p*<*n***	***p*>*n***	
Chen (2012)			✓	✓		✓		
Mansson et al. (2012)		✓	✓	✓		✓		
Dupuis and Victoria(2013)			✓		✓	✓		
Ma and Du (2014)	✓					✓		
Ertas et al.(2015)	✓			✓		✓		
Sinan and Alkan (2015)		✓	✓		✓	✓		
Jadhav and Kashid (2016)	✓				✓	✓		
Amini and Roozbeh (2016)	✓			✓		✓		
Wang et al. (2017)		✓	✓		✓	✓		
Norouzirad et al. (2017)	✓	✓		✓		✓		
Giacalone et al., (2018)	✓	✓	✓			✓		
Shariff and Ferdaos (2017)		✓				✓		
Lukman et al. (2017)	✓					✓		
Huang et al. (2017)			✓			✓		
Zhang (2017)	✓	✓	✓			✓		

Multicollinearity is considered to be a major issue in the case of more variables in the model. To address this problem, Chen (2012) investigated a simple method for dealing with multicollinearity. It was based on four simple steps, which only required a basic statistical method with prior knowledge. This technique has the advantage of a robust technique, even in the case of incorrect prior knowledge. Later on, the generalisation of Liu (1993) was introduced (Mansson et al., 2012). They used the Logit Model Shrink Estimator, a generalisation estimator of linear regression. This could be applied to the solution of general problems due to multicollinearity. Another contribution to the literature on robust regression was made by Dupuis and Victoria, (2013). It was based on robust variance inflation factor (VIF) regression and was very fast, efficient. It has outperformed non-robust VIF in the case of outliers. Later , a new class of robust bias estimator was examined by Ma and Du (2014) as they called it generalised shrunken type generalised M (GM) estimation. They used the combination of GM estimator and bias estimators such as ridge, principal components, Liu estimates, etc. Ertas et al . (2015) investigated the robust two-parameter ridge M-estimator (TRME) as an alternative to the RME robust Liu type M estimator for the combined problem of multicollinearity and *y*-direction outliers. Subsequently, Sinan and Alkan (2015) proposed a correlation matrix robust estimation based on a minimum covariance determinant for multicollinearity identification and outliers when classical methods such as VIF, condition number (CN), variance decomposition proportion (VDP) were unable to identify multicollinearity problems. In these situations, their proposed method worked well.

In the real-life data set, outliers and multicollinearity can both occur, particularly in the large data analysis, so that Jadhav and Kashid (2016) presented a new estimator called linear ridge M estimator for the combined problem of multicollinearity and outliers. Amini and Roozbeh (2016) also contributed to the simultaneous handling of both problems by introducing robust ridge estimators from the family of regression parameters and for non-linear parts after the addition of the L2 penalty for the trimmed square estimates. It was a combination of the least trimmed squares (LTS) and the ridge estimation methods used in partial linear models. A comparison of different robust estimators was made from Lukman et al. (2017) by comparing M, MM, LTS, LAD. Next, Huang et al. (2017) investigated linear regression with both multicollinearity and heteroscedasticity problems. They discussed the properties of the method of perturbation. Zhang (2017) examined the prevalence and environment of diabetes by considering socioeconomic effect changes in China by fitting OLS, robust regression and a set of binary-choice models. They used diabetes as a dependent variable.

2.6 Performance of other types of regression analysis in different fields

A number of other types of regression analysis have been used by many researchers in previous studies. Table 2.5, summarize some of the work done by previous researchers.

Table 2.5 Study about other regression analysis in different fields

Author(s) (Year)	**Sample Size**			**Field of study**		**Simulation**		**Data dimension**		**SOP**
	small	**medium**	**large**	**Agriculture**	**Non agriculture**	**Monte Carlo**	**Other**	***p*<*n***	***p* > *n***	
Thompson and Dunlap (2008)								✓		
Liu et al. (2010)			✓		✓			✓		
Beck et al. (2010)		✓						✓		
Kisi (2010)			✓		✓			✓		
Gunes and Bondell (2012)	✓	✓	✓		✓		✓	✓		
Umali and Barrios (2014)	✓				✓		✓	✓		
Rodrigues-Galino et al. (2014)			✓		✓			✓		
Aunsmo et al. (2014)					✓			✓		
Howe and Nicolis (2015)	✓				✓			✓		
Rina et al. (2015)			✓		✓		✓	✓		
Ding et al. (2015)		✓	✓	✓			✓	✓		
Benoy et al. (2016)					✓			✓		
Ho and Ermon(2016)			✓		✓			✓		
Hsu and Huang (2017)					✓			✓		
Shim et al. (2017)			✓		✓			✓		
Schirrmann et al.(2017)		✓	✓					✓		
Takeshima (2017)			✓	✓				✓	✓	
Zelenkov et al.(2017)			✓		✓			✓		
Seo (2017)			✓		✓		✓	✓		
Weight and Harpending(201			✓		✓		✓	✓		