2023-05-01

Comparison Of Different Robust Methods In Linear Regression And Applications In Cardiovascular Data

Jagannath Das
University of Texas at El Paso

COMPARISON OF DIFFERENT ROBUST METHODS IN LINEAR REGRESSION AND APPLICATIONS IN CARDIOVASCULAR DATA

JAGANNATH DAS

Master's Program in Mathematical Sciences

APPROVED:

Abhijit Mandal, Ph.D., Chair

Suneel Babu Chatla, Ph.D.

Mohammad Iqbal H. Bhuiyan, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

To my

FATHER & MOTHER

with love

COMPARISON OF DIFFERENT ROBUST METHODS IN LINEAR REGRESSION AND APPLICATIONS IN CARDIOVASCULAR DATA

by

JAGANNATH DAS

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

May 2023

Acknowledgement

I would especially like to thank Lord Krsna, the supreme personality of godhead, for making this voyage successful. Additionally, I would like to express my sincere gratitude to my thesis advisor, Dr. Abhijit Mandal of the Department of Mathematical Sciences at The University of Texas at El Paso for his resolute support and suggestions. I also want to express my profound appreciation to Dr. Suneel Baabu Chatla from the Department of Mathematical Sciences and Dr. Mohammad Iqbal H. Bhuiyan from the Department of Pharmaceutical Sciences (PS) at The University of Texas at El Paso for their constant endeavors.

Abstract

Due to advanced technology and wide source of data collection, high-dimensional data is available in several fields, including healthcare, bioinformatics, medicine, epidemiology, economics, finance, sociology, and climatology. In those datasets, outliers are generally encountered due to technical errors, heterogeneous sources, or the effect of some confounding variables. As outliers are often difficult to detect in high-dimensional data, the standard approaches may fail to model such data and produce misleading information. In this thesis, we studied Huber and Tukey's M-estimators for linear regression that automatically down-weight outliers and provide a good fit. We also investigated two variable selection methods – LASSO and LAD-LASSO. In addition, we performed a simulation study to compare different estimators in pure and contaminated data. Finally, we analyzed cardiovascular data to model systolic and diastolic blood pressure. The results show that Huber and Tukey's M-estimators perform better for this dataset.

Keywords: Regression models; variable selection; robust estimator.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the linear regression model, we estimate the relationship between a dependent variable with one or more independent variables. In practice, the dataset may contain unexpected data points or outliers. The outliers are often difficult to detect during data cleaning, especially in high-dimensional data. Moreover, if we delete some good data points by mistake, we may lose important information. On the other hand, if we keep outliers, the classical estimate for the linear regression model is unlikely to provide the actual result. Thus, we need some other robust methods that automatically down-weight outlying observations. For this purpose, we use different robust estimators, including Huber M-estimation, Tukey MM-estimator, and Least Absolute Deviation (LAD) estimation, to better understand the dataset.

Another critical issue with regression analysis is variable selection for high-dimensional datasets. In order to mitigate potential modeling biases, a significant number of regressors are typically introduced at the beginning of the regression model. However, adding extra predictors can reduce the effectiveness of the resulting estimation process and produce less accurate predictions. On the other hand, leaving out a crucial explanatory variable could lead to inaccurate parameter values and incorrect predictions. To choose a model and perform model estimation in regression models, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO). The parameter estimation and variable (model) selection is done concurrently using the LASSO method. In a single minimization problem based on a penalized technique, these two methods have been combined. The LASSO, however, is affected by heavy-tailed errors or outliers in the data since it is a special case of penalized least squares regression. The LAD-LASSO is a robust version of

the LASSO technique, making it resistant to outliers and heavy-tailed error distributions (Hesterberg et al. 2008). Combining the LAD and LASSO methods, the LAD-LASSO regression recently became popular for performing simultaneous robust parameter estimation and variable selection.

Chapter 2

Ordinary Least-Squares (OLS) estimator

Assume there are n observations $\{x_i, y_i\}_{i=1}^n$ in a dataset. A scalar response y_i and column vector x_i with p regressors are included in each observation i. It can expressed as $x_i = [x_{i1}, \cdots, x_{ip}]^T$. The target (response) variable y_i, in a model of linear regression is a linear function of the regressors (dependent variables): $y_i = \beta_1 x_{i1}, \cdots, \beta_p x_{ip} + u_i$, in vector notation it can be expressed as follows:

$$y_i = x_i^T \beta + u_i$$

where β is $p \times 1$ vector of unknown parameters, x_i are the explanatory variables of the column vector of i-th observation. This model can alternatively be expressed in matrix notation as

$$Y = X\beta + u$$

where y and u are the $n \times 1$ vectors of the dependent (response) variable and the noise of the n observations respectively, and $n \times p$ be the dimension of the design matrix X. A systematic component ($X\beta$) and a stochastic component (u) make up the model. Our objective is to estimate the population parameters included in the vector.

The primary goal of Least Square method is to reduce the sum of the squares of noise

components of the estimated parameter β, i.e.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |Y_i - \sum_{j=1}^{p} X_{ij}\beta_j|^2 = \arg\min_{\beta} \|Y - X\beta\|^2$$

The vector of residuals is represented by

$$u = Y - X\hat{\beta}$$

It should be clear that the total of squared residuals can be expressed as:

$$u'u = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

$$\implies u'u = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$\implies u'u = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Take the derivative of the above equation with respect to $\hat{\beta}$ to obtain the $\hat{\beta}$ that minimizes the sum of squared residuals. This gives,

$$\frac{\delta u'u}{\delta \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

To check the minimum of the above equation, we went for the second derivative, the result gave us $2X'X$. Thus it is easy to take a decision that X has a full rank and positive definite matrix. Therefore, the Ordinary Least Squares (OLS) estimate is given by

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

Assumptions of the multiple linear model:

a. The population mean of the error term is zero.

b. The error terms are not correlated with any independent variables.

c. There is no correlation between the observations of the error term.

d. There is no heteroscedasticity in the error term; it has a constant variance.

e. The noise term is normally distributed.

2.1 Asymptotic Distribution of OLS

We know from the above discussion $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$. We now assume that random errors term (u_i) are i.i.d with $E(u_i) = 0$ and $var(u_i) = \sigma^2$. If X is fixed, then $E(\hat{\beta}_{LS}) = \beta$ and $var(\hat{\beta}_{LS}) = \sigma^2(X'X)^{-1}$. Thus if the X is full rank and u_i are normal, then $\hat{\beta}_{LS}$ is a multivariate normal. That is,

$$\hat{\beta}_{LS} \sim N_p(\beta, \sigma^2(X'X)^{-1}).$$

Chapter 3

M-Estimation

Let $x_1, \cdots x_n$ be a random sample from a distribution $f(x_i, \mu)$, where μ is the parameter of interest. The maximum likelihood estimate (MLE) of μ is defined as:

$$\hat{\mu} = \arg\max_{\mu} \left(\prod_{i=1}^{n} f(x_i, \mu) \right), \quad or \quad \hat{\mu} = \arg\min_{\mu} (-\log f(x_i, \mu)). \tag{3.0.1}$$

Under some regularity conditions, the MLE is the most efficient estimator. However, it is well known that the MLE breaks down in the presence of outliers. In 1964, Peter J. Huber proposed generalizing maximum likelihood estimation to the minimization of

$$\sum_{i=1}^{n} \rho(x_i, \mu), \tag{3.0.2}$$

where the ρ function can be chosen in such a way to provide the estimator desirable properties in terms of bias and efficiency. The minimizer of the above equation is called an M-estimator. The aim of an M-estimator is to down-weight the effect of outliers and make the estimator robust. In case of the multiple linear regression model, an M-estimator is obtained as follows:

$$\hat{\beta}_M = min \sum_{i=1}^{n} \rho \left(\frac{(y_i - \sum_{j=0}^{k} x'_{ij} \beta_j)}{\hat{\sigma}_{MAD}} \right), \tag{3.0.3}$$

where

$$\hat{\sigma}_{MAD} = \frac{median|u_i - median(u_i)|}{0.6745}. \tag{3.0.4}$$

Here, median absolute deviation ($\hat{\sigma}_{MAD}$) is an estimate of scale frequently created by combining the residuals in a linear fashion. The constant 0.6745 helps sample standard deviation S to become an unbiased estimate of population standard deviation σ.

If we take first partial derivative to find out $\hat{\beta}_M$, then

$$\sum_{i=1}^{n} x_{ij} \Psi \left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}} \right) = 0, \quad j = 0, \cdots, k, \tag{3.0.5}$$

where first derivative of ρ is Ψ, that is $\Psi = \rho'$, x_{ij} is i-th observation on the j-th independent variable.

Let us define the weight function as

$$w_i = \frac{\Psi \left(\frac{y_i - \sum_{j=0}^{k} x'_{ij}\beta_j}{\hat{\sigma}_{MAD}} \right)}{\frac{\left(y_i - \sum_{j=0}^{k} x'_{ij}\beta_j \right)}{\hat{\sigma}_{MAD}}}. \tag{3.0.6}$$

Then, the M-estimator is derived from

$$\hat{\beta}_M = (X^T W X)^{-1} (X^T W Y), \tag{3.0.7}$$

where W is the diagonal matrix with the diagonal element as w_i, $i = 1, 2, \cdots, n$. We cannot to directly calculate the weight function in equation (3.0.6), as the weights depend on the unknown parameter β and σ. However, the weighted-means representation of M-estimators yields a straightforward iterative approach for computing the M-estimator.

(i) We take median as the initial estimate of β and then calculate sample standard deviation s,

(ii) Compute new estimate for β using equation (3.0.7),

(iii) Until the algorithm converges, repeat step (i) & (ii).

3.1 Huber M Estimation

The ρ function for Huber M-estimator is

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| < c, \\ c|u| - \frac{1}{2}c^2 & \text{if } |u| \geq c, \end{cases} \tag{3.1.1}$$

where $u = y_i - \hat{y}_i$. The Least-squares (LS) and the least absolute deviation (LAD) loss functions are combined to create Huber's loss, which uses the LS-loss function for generally minor mistakes and the LAD loss function for generally big errors. This has a double exponential distribution at the tails and a Gaussian distribution in the centre. Moreover, if $c \to \infty$, Huber loss function reduces to least square loss (LS Loss) $\rho_c(u) = \frac{u^2}{2}$. The derivative of Huber loss function, which is convex and differentiable, produces the following weight function:

$$w(u) = \begin{cases} 1 & \text{if } |u| \leq c, \\ c/|u| & \text{if } |u| > c, \end{cases} \tag{3.1.2}$$

where $w(u) = \Psi(u)/u$.

For the Huber M-estimator weighted function we take $c = 1.345$, where c is a cutoff point configured by users that affects the level of robustness. To achieve a user defined asymptotic relative efficiency with respect to least square estimate under normal (Gaussian) errors, the cutoff point c is often selected to minimise the regression problem (σ being fixed). The cutoff point are chosen $c_{0.95} = 1.345$ and $c_{0.85} = 0.7317$ correspondingly for 95% and 85% asymptotic relative efficiency.

3.2 Tukey MM Estimation

Tukey MM method is a combination of high breakdown point estimation and efficient estimation. The ρ function for Tukey Bisquare Estimator is given below

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{6}\{1 - [1 - (\frac{u_i}{c})^2]^3\} & \text{if } |u_i| \leq c, \\ \frac{u_i^2}{6} & \text{if } |u_i| > c. \end{cases} \tag{3.2.1}$$

The weight function is given by

$$w_i = \begin{cases} [1 - (\frac{u_i}{c})^2]^2\} & \text{if } |u_i| \leq c, \\ 0 & \text{if } |u_i| > c. \end{cases} \tag{3.2.2}$$

The popular choice of $c = 4.685$ for the Tukey's bisquare weighted function.

3.3 The Distribution of M-estimates

Let us consider a general linear model

$$Y = X\beta + u,$$

where X are the vectors with dimension $n \times p$, and u is the identically independently distributed random variables of a distribution F with components $(u_1, \cdots, u_n)$. Y is consists of the $(y_1, \cdots, y_n)$ observations. Assume that

$$E(u_i) = 0 \quad and \quad E(u_i^2) < \infty. \tag{3.3.1}$$

An M-estimator of β is defined in equation (3.0.3) or equivalently, the solution of the estimating equation given in equation (3.0.5). Then, the following results from Yohai & Maronna (1979) give the consistency and asymptotic distribution of the M-estimator.

Assumptions: Let us assume

(A1) Ψ is nondecreasing

(A2) there exist positive numbers b, c and d such that

$$D(u,z) \geq d \text{ if } |u| \leq c \text{ and } |z| \leq b, \tag{3.3.2}$$

where

$$D(u,z) = \frac{\phi(u+z) - \phi(u)}{z}, \tag{3.3.3}$$

and c satisfies

$$q = F(c) - F(-c) > 0.$$

(A3) $E_F(\Psi^2(u)) = v < \infty$

(A4) $E_F(\Psi(u)) = 0$

(A5) For all $n > n_0$, $X'X$ is a nonsingular matrix.

Consistency: Under assumptions (A1)–(A5), the M-estimator $\hat{\beta}$ is a consistent estimator of β if $\lambda_1(X'X) \to \infty$, where $\lambda_1(X'X)$ is the smallest eigen value of $X'X$.

Let us define

$$\beta^* = M\beta, \quad \hat{\beta}^* = M\hat{\beta}, \quad z_i = (M')^{-1}x_i, \tag{3.3.4}$$

where M is a $p \times p$ matrix so that $M'M = X'X$. Therefore, we have

$$\sum_{i=1}^{n} z_i z_i' = I \quad and \quad \sum_{i=1}^{n} |z_i^2| = p$$

and $\hat{\beta}^*$ is a solution of

$$\sum_{i=1}^{n} \Psi(y_i - z_i'\hat{\beta}^*)z_i = 0.$$

Then, the following theorem gives the asymptotic distribution of $\hat{\beta}^*$.

Asymptotic Normality: Let us assume that Ψ and F satisfy (A1)–(A5) and the following conditions

$$\int_{-\infty}^{\infty} [\Psi(x+h) - \Psi(x-h)]^2 \, dF(x) \; = 0 \quad as \quad h \to 0$$

and

$$sup_{|q|\leq\varepsilon, |h|\leq\varepsilon} \{|h|^{-1} \int_{-\infty}^{\infty} [\Psi(x+q+h) - \Psi(x+q)] \, dF(x) \; \} < \infty \quad for \; some \; \varepsilon > 0$$

There exists $A(\Psi, F)$ such as

$$\int_{-\infty}^{\infty} [\Psi(x+h) - \Psi(x-h)] \, dF(x) \; = hA(\Psi, F) + o(h)$$

Also consider

$$\lim_{n\to\infty} \max_{i\leq n} |z_i|^2 = 0$$

Then, the distribution of $(\hat{\beta}^* - \beta^*)$ follows a multivariate normal with mean 0 and covariance $\tau^2 I$, where

$$\tau^2 = E_F \Psi^2 / A(\Psi, F)^2.$$

Chapter 4

Variable Selection

4.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Consider the multiple linear regression setup, where we have data $(x^i, y_i), i = 1, 2, \cdots, n$, where $x^i = (x_{i1}, \cdots, x_{ip})^T$ and y_i is the regressor and response for the i-th observation. The residual squared error is minimized to get the ordinary least squares (OLS) estimations. In high-dimensional data, the analyst is frequently unsatisfied with the OLS estimates for two reasons. The first is the success rate of predictions: OLS estimates frequently have low bias but high variance; sometimes reducing or lowering some coefficients to 0 might increase prediction accuracy. In exchange for reducing the variance of the prdicted values, we make a little bias sacrifice, which may increase the prediction accuracy as a whole. Interpretation is the secondary factor. We frequently want to identify the smaller selection of predictors that has the largest influence when there are many predictors. Subset selection and ridge regression, the two commonly used methods for enhancing the OLS estimates, both have shortcomings. Because subset selection is a discrete process in which regressors are either kept in the model or removed, it can produce highly variable models that are still interpretable. Its prediction accuracy may decrease as a result of significantly diverse models being chosen in response to small changes in the data. Ridge regression produces a more stable model since it continuously reduces coefficients; but, because no coefficients are set to 0, it does not produce a model that is simple to understand. We provide a novel method for least absolute shrinkage and selection operator called the LASSO. It attempts

to keep the beneficial aspects of both subset selection and Ridge regression by reducing some coefficients and setting others to 0.

4.2 Background

Let's say we have data $(x^i, y_i), i = 1, 2, \cdots, n$, where $x^i = (x_{i1}, \cdots, x_{ip})^T$ are the predicting factors and y_i is the responses. We may assume that the observations are independent, as is the case in the standard regression setup, or that the $y_i s$ are if certain conditions are met given the $x_{ij} s$. We assume that x_{ij} are standardized so that $\sum_i x_{ij}/n = 0$, $\sum_i x_{ij}^2/n = 1$.

Letting $\widehat{\beta} = (\widehat{\beta_1}, \cdots, \widehat{\beta_p})^T$, the LASSO estimate $(\widehat{\alpha}, \widehat{\beta})$ is defined by,

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right\}, \tag{4.2.1}$$

subject to $\mid \beta_j \mid \leq t$. Here $t \geq 0$ is a tuning parameter. Now for all t the solution for α is $\widehat{\alpha} = \bar{y}$. We can assume without sacrificing generality that $\bar{y} = 0$ and hence omit α.

The parameter $t \geq 0$ regulates how much shrinkage is applied to the estimates. Let $\hat{\beta}_j^0$ be the full least squares estimates and let $t_0 = \sum \mid \beta_j^0 \mid$. Values of $t < t_0$ will result in the solutions shrinking towards 0, and some coefficients might be exactly 0. For example, if $t = t_0/2$ the effect will be basically equivalent to determining the best size subset $p/2$. Also, keep in mind that the design matrix does not have to be full rank.

The idea for the LASSO was inspired by an intriguing proposition of Breiman (1993). Breiman's non-negative garotte minimizes

$$\sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^{\,o} x_{ij})^2, \tag{4.2.2}$$

subject to $c_j \geq 0, \sum c_j \leq t$. Starting with the OLS estimates, the garotte reduces them by non-negative components whose sum is restricted. Breiman shown in extensive simulation experiments that, with the exception of situations when the correct model has numer-

ous small non-zero coefficients, Garotte typically outperforms subset selection in terms of prediction error and competes favorably with ridge regression.

The garotte has the issue that the direction and size of the OLS estimations affect its solution. The garotte may suffer when the OLS estimates behave unfavorably due to overfitting or excessive correlation. When compared to the LASSO, OLS estimates are not explicitly used. Tibshirani (1996) proposed the following LASSO criterion:

$$Q(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + n\lambda \sum_{j=1}^{p} \mid \beta_j \mid,$$

where $\lambda > 0$ is the tuning parameter. The generated estimators may have a significant bias because LASSO utilizes the same tuning settings for all regression coefficients.

4.3 Orthonormal Design

The shrinkage's nature is shown by the orthonormal design scenario. Let X be the $n \times p$ design matrix with ijth entry x_{ij}, and suppose that $X^T X = I$, the identity matrix.

It is simple to demonstrate that the solution of equation (4.2) is given by

$$\hat{\beta}_j = sign(\hat{\beta}_j^0)(\mid \hat{\beta}_j^0 \mid -\gamma)^+, \tag{4.3.1}$$

where γ is decided by the condition $\sum \mid \hat{\beta}_j^0 \mid = t$. This follows the soft shrinkage suggestions exactly in terms of structure is interestingly same Donoho & Johnstone (1994) and Donoho & Johnstone (1995) made in reference to the function estimate of wavelet coefficients. In the context of signal or picture recovery, Donoho et al. (1992) have seen a connection between soft shrinkage and a low L_1 norm penalty for non-negative parameters.

In the orthonormal design scenario, selecting the optimal subset of size k reduces to selecting the k coefficients with the biggest absolute values and setting the other coefficients to 0. For some choice of λ this is equivalent to setting $\hat{\beta}_j = \hat{\beta}_j^0$ if $\mid \hat{\beta}_j^0 \mid > \lambda$ and to 0 otherwise.

On the other hand, the ridge regression minimizes

$$\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2 \tag{4.3.2}$$

or equivalently, minimizes

$$\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2 \quad subject \quad to \sum \beta_j^2 \leq t. \tag{4.3.3}$$

The solutions at the ridge include

$$\frac{1}{1+\gamma}\hat{\beta}_j^0, \tag{4.3.4}$$

where γ depends on λ or t. The garotte estimate is

$$(1 - \frac{\gamma}{\hat{\beta}_j^{o2}})^+ \hat{\beta}_j^o. \tag{4.3.5}$$

The garotte function is similar to the LASSO in that it shrinks less with increasing coefficient size.

4.4 Geometry of LASSO

For the case $p = 2$, the criterion $\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2$ equals the quadratic function

$$(\beta - \hat{\beta}^o)^T X^T X (\beta - \hat{\beta}^o).$$

The constraint region is the rotated square, and the elliptical outlines of this function are centered at the OLS estimates. There are no corners for the contours to contact, therefore zero solutions will seldom occur. The LASSO solution is the first place the contours touch the square, and this will occasionally happen at a corner, resulting to a zero coefficient.

The parameters have been standardized, when $p = 2$ the main axis that make up the outlines are at $\pm 45^o$ to the co-ordinate axes, and We can demonstrate that the contours

must come into contact with the square in the same quadrant as the one containing $\hat{\beta}^o$. However, when $p > 2$ and the data show at least a considerable degree of correlation, while this need not be the case. However, the garotte keeps the mark of each $\hat{\beta}_j^o$, LASSO is capable of changing signs.The existence of the OLS estimates in the garotte can cause it to behave differently even when the LASSO estimate and the garotte have the same sign vector. The model $\sum c_j \hat{\beta}^o x_{ij}$ with constraint $\sum c_j \leq t$ can be written as $\sum \beta_j x_{ij}$ with constraint $\sum \frac{\beta_j}{\hat{\beta}^o} \leq t$. If for example $p = 2$ and $\hat{\beta}_1^o > \hat{\beta}_2^o > 0$ then the outcome would be a horizontal stretching of the square. The garotte will therefore prefer bigger values of $\hat{\beta}_1$ and smaller values of β_2.

Without loss of generality, let's assume that $p = 2$, and assume without loss of generality that both of the least squares estimates $\hat{\beta}_j^o$ are positive. Next, we can demonstrate that the LASSO estimates are

$$\hat{\beta} = (\hat{\beta}_j^o - \gamma)^+, \tag{4.4.1}$$

where γ is chosen so that $\hat{\beta}_1 + \hat{\beta}_2 = t$. This formula holds for $t \leq \hat{\beta}_1^o + \hat{\beta}_2^o$ is accurate despite the predictors' correlation. Solving for γ produces

$$\hat{\beta}_1 = \left(\frac{t}{2} + \frac{\hat{\beta}_1^o + \hat{\beta}_2^o}{2} \right)^+, \quad \hat{\beta}_2 = \left(\frac{t}{2} - \frac{\hat{\beta}_1^o - \hat{\beta}_2^o}{2} \right)^+. \tag{4.4.2}$$

All values of ρ are covered by the LASSO estimations along the entire curve. The estimations of the ridges that are broken depend on ρ. When $\rho = 0$, the slope of the ridge regression shrinks correspondingly. However, because the bound is compressed for larger values of ρ, the ridge estimates are reduced differently and may even increase slightly. As Jerome Friedman pointed out, this is a consequence of ridge regression's propensity to try to equilibrate the coefficients in order to lessen their squared norm.

4.5 Standard Errors

Since the LASSO estimate is a non-linear and non-differentiable function of the response values, it is difficult to calculate the standard error of the estimate even for a fixed value of t. The bootstrap method is one technique; for each bootstrap sample, t can either be fixed or optimized over. In order to fix t, it is similar to selecting the best subset and then using its least squares standard error.

One can derive an approximate closed form approximation by writing the penalty $\sum \mid \beta_j \mid$ as $\sum \frac{\beta_j^2}{|\beta_j|}$. Hence, at the LASSO estimate $\hat{\beta}$, we may the solution by a ridge regression of the norm $\beta^* = (X^T X + \lambda W^-)^{-1} X^T y$ where W is a diagonal matrix with diagonal elements $\mid \tilde{\beta}_j |$, W^- denotes the generalized inverse of W and λ is chosen so that $\sum \mid \beta_j^* \mid = t$. The covariance matrix of the estimates can be roughly calculated using

$$(X^T X + \lambda W^-)^{-1} X^T X (X^T X + \lambda X^-)^{-1} \hat{\sigma}^2, \tag{4.5.1}$$

where $\hat{\sigma}^2$ is an estimate of the error variance. This formula has the drawback of providing an estimated variance of 0 for predictors with $\hat{\beta}_j = 0$.

4.6 LAD-LASSO

Consider a linear regression model

$$y_i = x_i' \beta + \epsilon_i, \quad i = 1, \cdots, n, \tag{4.6.1}$$

where $x_i = (x_{i1}, \cdots, x_{ip})'$ is the p dimensional regression covariate, $\beta = (\beta_1, \cdots, \beta_p)'$ are the associated regression coeffficients, and ϵ_i are iid random errors with median 0. Moreover, assume that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 \geq 0$. Thus the correct p_0 significant and $(p - p_0)$ not significant regression variables. Typically, the OLS criterion can be minimized in order to estimate the model's unknown parameters, $\sum_{i=1}^n (y_i - x_i' \beta)^2$.

To reduce extraneous coefficients as well to 0 (Fan & Li 2001). Consequently, we also take into account the following modified LASSO criterion:

$$LASSO^* = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + n\sum_{j=1}^{p}\lambda_j|\beta_j|,$$

which enables the use of various tuning parameters for various coefficients. As a result, $LASSO^*$ is better than LASSO at producing sparse solutions. The OLS criterion employed in $LASSO^*$ is well known to be extremely sensitive to outliers, though. We further change the $LASSO^*$ objective function into the following LAD-LASSO criterion to produce a robust LASSO-type estimator:

$$Q(\beta) = \sum_{i=1}^{n} \mid y_i - x_i'\beta \mid + n\sum_{j=1}^{p}\lambda_j|\beta_j|.$$

As is obvious, the LAD-LASSO criteria combines the LAD criterion and the LASSO penalty; as a result, the resulting estimator is expected to be robust to outliers and to enjoy a sparse representation.

Chapter 5

Simulation Results

We also carried out a simulation study of the relative effectiveness of the Root Mean Prediction Error (RMPE). The simulation results are shown in tables (5.1)–(5.9). Simulation findings based on the characteristics of several robust estimating techniques for sample size (n) (training data), contamination proportion (p), trimming proportion for RMPE, size of test data, number of replication (R), sigma: error standard deviation, location of outliers. We considered the errors are normally distributed. The detailed information of the simulation setup is given in R code in the Appendix.

Table 5.1: Root mean prediction errors over 50 samples for 0% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.8660351	0.8738636	0.8766966	0.8862412	1.2236169	1.4025990
In 0 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.8660351	0.8738636	0.8766966	0.8862412	1.2236169	1.4025990
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.101808	1.109186	1.111754	1.113878	1.299827	1.666857

We observed that for the case of 5% outliers and for the test data, as the sample size increases mean prediction error decreases for Tukey and Huber, and for the 10% outliers we experienced the same pattern for the Tukey and Huber. Other methods provided RMPE which are larger than the Tukey and Huber. Although Tukey MM and Huber M-estimators RMPE value is very close, specifically, Tukey MM-estimator provided us the lower RMPE

Table 5.2: Root mean prediction errors over 50 samples for 5% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
3.607753	4.027983	4.124016	3.800170	4.089226	4.189711
In 5 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.8094452	0.8336395	0.8143050	1.7366329	1.1579878	1.4257752
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.390870	1.129554	1.115302	2.235038	1.344893	1.842929

Table 5.3: Root mean prediction errors over 50 samples for 10% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
5.448293	6.174247	6.385792	5.918860	6.204027	6.300127
In 10 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
3.0012419	0.9519808	0.8741759	2.9544069	1.3369843	1.7078085
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
3.729105	1.211684	1.106938	3.388778	1.496747	2.122962

than Huber M-estimator. However, when we considered 0% outliers then OLS provided us the better result, although the results of Huber and Tukey are very close to OLS.

Table 5.4: Root mean prediction errors over 100 samples for 0% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.9492279	0.9527806	0.9532036	0.9585786	1.2191313	1.4114292
In 0 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.9492279	0.9527806	0.9532036	0.9585786	1.2191313	1.4114292
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.104925	1.109122	1.109058	1.117279	1.426157	1.636696

Table 5.5: Root mean prediction errors over 100 samples for 5% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
4.223040	4.503123	4.564310	4.290617	4.540071	4.601477
In 5 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.8484025	0.9395949	0.9265473	1.8226538	1.2106157	1.5203182
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.014507	1.106811	1.103090	1.950204	1.434388	1.791936

Table 5.6: Root mean prediction errors over 100 samples for 10% outliers

In full training data					
5.756346	6.255145	6.382624	5.907336	6.278497	6.325418
In 10 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.7037218	0.9706079	0.9268265	2.7174216	1.2495160	1.7293914
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.893549	1.130862	1.102836	2.777624	1.440567	1.984243

Table 5.7: Root mean prediction errors over 150 samples for 0% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.9617213	0.9639335	0.9640969	0.9677256	1.2271929	1.4772331
In 0 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.9617213	0.9639335	0.9640969	0.9677256	1.2271929	1.4772331
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
0.9850437	0.9878325	0.9885020	0.9888645	1.2466415	1.5521162

Table 5.8: Root mean prediction errors over 150 samples for 5% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
4.455625	4.677796	4.726180	4.506301	4.717703	4.801181
In 5 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.7781387	0.9701431	0.9590385	1.7574830	1.2565367	1.6325668
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
1.8854363	1.0002758	0.9869823	1.8944440	1.2867973	1.7429515

Table 5.9: Root mean prediction errors over 150 samples for 10% outliers

In full training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
5.870861	6.295382	6.401048	5.943603	6.299566	6.347056
In 10 % trimmed training data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.5922784	1.0017509	0.9656727	2.5559324	1.2871031	1.7747781
In test data					
OLS	Huber	Tukey	LASSO	LAD	LAD-LASSO
2.7836584	1.0426653	0.9968226	2.7803825	1.3515616	1.9378616

Chapter 6

Real Data Analysis

A real dataset is used to assess how well the robust model strategy performs. We studied a set of collected data at the moment of medical examination for the patients to be healthy or suffering from cardiovascular disease (Source: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset). This dataset comprises 70,000 patient records with information on 12 characteristics, including age, gender, systolic and diastolic blood pressure, among others.

Description of the dataset: (i) Age: integer (in days) (ii) Height: integer (in cm) (iii) Weight: (in kilogram) (iv) Sex: 1 = Male; 2 = Female (v) Blood pressure level: Systolic: integer. (vi) Blood pressure level: Diastolic: integer (vii) Cholesterol level: 1: normal, 2: above normal, 3: well above normal (viii) Glucose level: 1: normal, 2: above normal, 3: well above normal (ix) Smoking status: binary: 1: smoker, 0: nonsmoker. (x) Alcohol intake: 1: alcoholic 0: nonalcoholic (xi) Physical activity: 1: active 0: inactive (xii) Presence or absence of cardiovascular disease: 1: present, 0: absent.

At the time of the physical examination, all dataset values were gathered. After cleaning the dataset, we took 68,986 observations with 12 important features. Compared to more conventional, smaller datasets like the Cleveland Dataset and the Hungarian Heart Disease Dataset (200–1000 variables), the used dataset is considerably larger (68,986 values). This aids in the development of more accurate and effective models. All variables were continuous, binary, or categorical, and there were no missing values in it. In the case of systolic blood pressure, we experienced anomalies as well as diastolic blood pressure too. These implausible data points were down-weighted by the robust models so that we produce more realistic predictions. We considered the response variable as a continuous variable, which

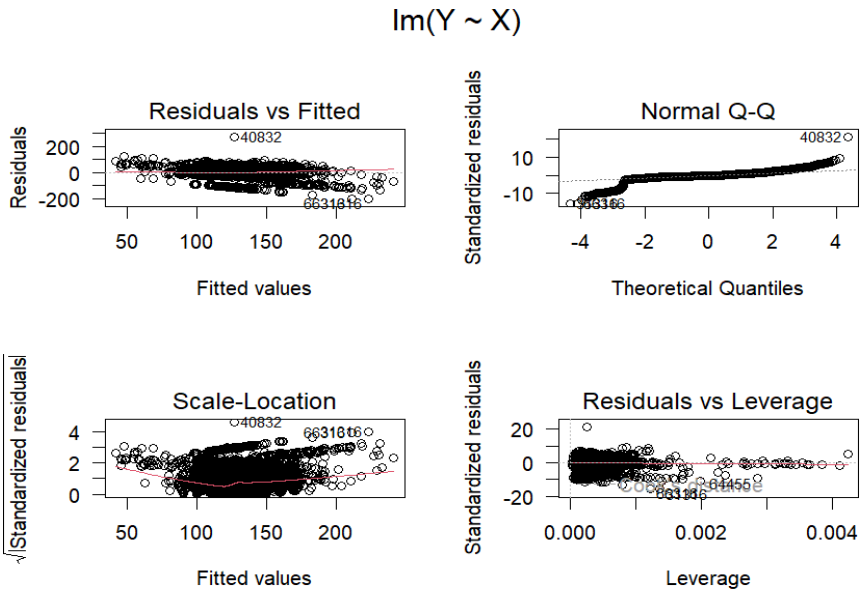

Figure 6.1: OLS residuals plot of the cardiovascular disease dataset: response variable is systolic BP.

is systolic blood pressure; in another case, we considered diastolic blood pressure.

6.1 Data Summary

From the residual plots (Fig. 6.1), density plot (Fig. 6.4), boxplot (Fig. 6.7), and QQ plot (Fig. 6.6), we have seen, in case of systolic blood pressure, there are many values lies above the third quartile and below the first quartile. The data is positively skewed (0.1414235). The distribution curve's outliers are more extreme to the right. Kurtosis of systolic blood pressure is 8.261528, which means the data is leptokurtic and produces more outliers compared to normal distribution. The maximum value and minimum value are recorded as 401 and 7, respectively, though the systolic blood pressure range is generally (100 - 130) (source: https://www.cdc.gov/bloodpressure/about.htm). The average systolic blood pressure is 126.5. Similarly, we observed that diastolic blood pressure also has many extreme values, for the diastolic blood pressure outliers lie above the third quartile

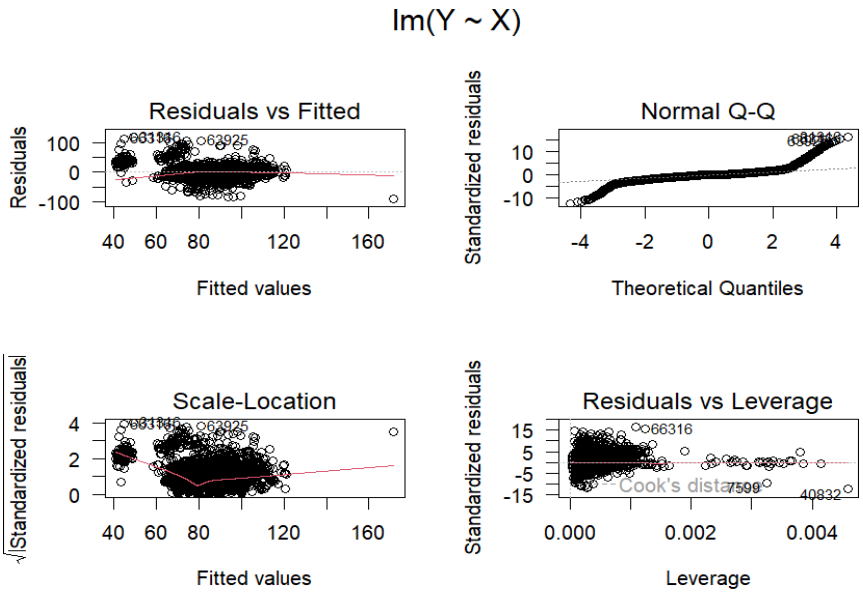

Figure 6.2: OLS residuals plot of the cardiovascular disease dataset: response variable is diastolic BP.

and below the first quartile. The average diastolic blood pressure is 81.35. Diastolic Blood pressure data points also fluctuate as the systolic data points. The highest and lowest value recorded as 190 and 01, respectively. From the figure (6.2, 6.3, 6.5 & 6.7), we have an idea that diastolic blood pressure data also has some disturbing values. Diastolic blood pressure data were positively skewed (0.4960359) and leptokurtic (8.130113). Here we consider systolic blood pressure as our target variable, and we took diastolic blood pressure as a response variable again. However, we have taken robust methods for the down-weighted outliers and compared those robust methods based on their RMPE. We are also cautious about the dataset so that these extreme values do not hinder the result. From the descriptive statistics, we observed that male patients' cardiac risk is much more than female patients before the average age (<53). Due to systolic blood pressure above the normal range (greater than 130), males are affected by around 52.47%, and females are affected by approximately 30.95%. Diastolic blood pressure is responsible for around 19.48% and 8.92% for male and female patients, respectively. Cholesterol level (well above normal) having an

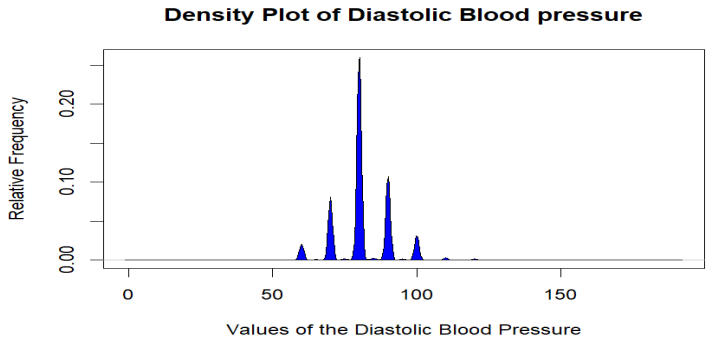

Figure 6.3: The Density Plot of Diastolic Blood Pressure in the cardiovascular disease dataset.

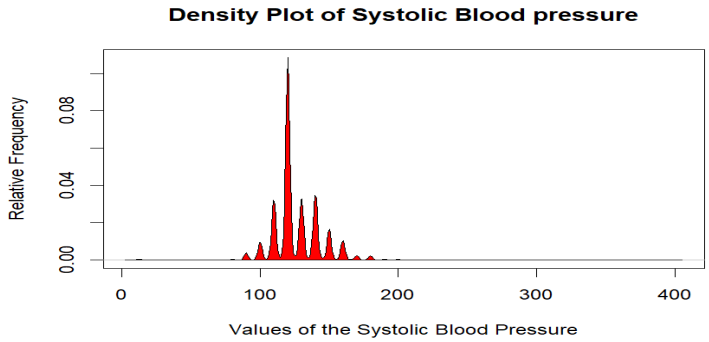

Figure 6.4: The Density Plot of Systolic Blood Pressure in the cardiovascular disease dataset.

obligation of about 35.35% & 23.25% proportionately for men and women. Glucose level (well above normal) control over around 43.30% and 18.50% for male and female patients, respectively. Surprisingly, we have got two different characteristics, one is smoking, and another one is alcohol intake, where males are less sufferer than females. For these two attributes, males clearly detected more cardiac disease symptoms than females, physically active and not active. According to the CDC (https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html), Body Mass Index (BMI) is interpreted into four categories. These are: below 18.5, 18.5 – 24.9, 25.0 – 29.9, 30.0, and Above defines as

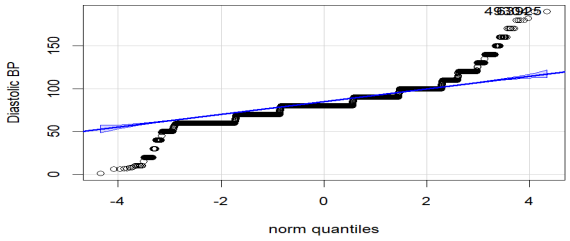

Figure 6.5: The QQ Plot of Diastolic Blood Pressure

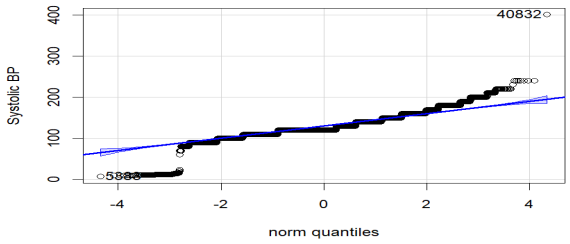

Figure 6.6: The QQ Plot of Systolic Blood Pressure

Table 6.1: Percentage of males and females having Cardiovascular Disease (CVD) for different features.

Features	Male	Female
Age < 53 (average age)	23.79%	4.07%
Systolic Blood Pressure (>130)	52.47%	30.95%
Diastolic Blood Pressure (<80)	19.48%	8.92%
Cholesterol (well above normal)	35.35%	23.25%
Glucose (well above normal)	43.30%	18.50%
Smoker	6.10%	40.66%
Alcohol intake	15.21%	32.56%
Physical activity	31.49%	17.07%
Physically not active	8.41%	4.63%
BMI (18.5 < BMI < 24.9)	11.93%	7.3%
BMI (>25)	36.68%	18.93%
BMI (<18.5)	17.88%	9.6%

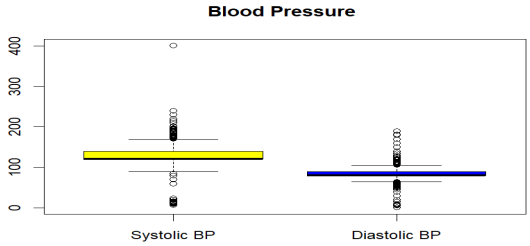

Figure 6.7: The Box Plot of Systolic and Diastolic Blood Pressure in the cardiovascular disease dataset.

Underweight, Healthy Weight, Overweight, and Obesity, respectively. The data shows a much more heart disease risk with a BMI greater than 25. Here, we have seen that systolic blood pressure is a critical factor in causing cardiovascular disease. Indeed, it is a growing concern for medical science.

6.2 Variable Selection

Table 6.2: Coefficients estimates of Cardiovascular Disease dataset using different methods ("0" indicates that the corresponding variable is not selected.) Response variable is Systolic Blood Pressure.

Variable	OLS	Huber	Tukey	LAD	LASSO	LAD-LASSO
Age	0.0005	0.0433	0.0324	0.0006	0.0001	0
Gender	0.0002	0.0141	0.0093	0.0001	0	0
Height	-0.0003	-0.0238	-0.0194	-0.0027	0	0
Weight	0.0082	0.0613	0.0505	0.0008	0	0
Diastolic	0.0553	0.6347	0.6649	0.0552	0.0642	0.0545
Cholesterol	0.0038	0.0294	0.0289	0.0036	0.0059	0
Glucose	-0.0001	0.0012	0.0008	0	0	0
Smoke	0.0012	0.0088	0.0076	0.0011	0.0001	0
Alcohol Intake	-0.0016	0.0033	0.0034	0	0	0
Activity	0.009	0.0072	0.0069	0.0007	0	0
Cardio	0.0181	0.1373	0.1081	0.1802	0.1474	0.0282

Table 6.3: Coefficients estimates of Cardiovascular Disease dataset using different methods ("0" indicates that the corresponding variable is not selected.) Response variable is Diastolic Blood Pressure.

Variable	OLS	Huber	Tukey	LAD	LASSO	LAD-LASSO
Age	0.0011	0.0039	0.0027	0.0011	0	0
Gender	0.0022	0.0143	0.0123	0.0018	0	0
Height	-0.0005	0.0064	0.0098	-0.0002	0	0
Weight	0.0076	0.0.0430	0.0304	0.0075	0	0
Systolic	0.0593	0.07351	0.7949	0.0592	0.0724	0.0606
Cholesterol	0.0017	0.0062	0.0003	0.0017	0	0
Glucose	0	-0.0038	-0.0037	0	0	0
Smoke	-0.0008	-0.0076	-0.0075	-0.0006	0	0
Alcohol Intake	0.0019	0.0070	0.0047	0.0017	0	0
Activity	0.0002	0.0004	-0.0010	0	0	0
Cardio	0.0072	0.0251	0.0118	0.0072	0	0

From the coefficients estimation table, we have considered the full data set. For these, we carried out our study considering systolic blood pressure, and then we considered diastolic blood pressure as a response variable. Among the test procedure, systolic blood pressure is the most common one that was selected by all methods; the second most selected characteristic was diastolic blood pressure. In the coefficients estimation tables, we saw, LAD-LASSO criteria do not select most variables but systolic and diastolic variables. However, Huber and Tukey's method takes all the variables, though the weight differed for all variables. Hence, the importance of variable selection criteria converged with our descriptive statistics. In the descriptive statistics, we have seen that systolic blood pressure is mostly responsible for cardiac risk. All methods chose the systolic blood pressure variable in our variable selection criteria. For further evaluation, we checked the RMPE of these methods. We predict that systolic and diastolic blood pressure are the two most significant variables in our study.

6.3 Prediction

We tried to determine which model provided us with less RMPE compared to our models' findings. In addition, we thought carefully about mean dimension reduction typically before making a decision. In our study, we have seen that the LAD-LASSO method reduced the data dimension for systolic BP more than the LASSO method. A similar criterion we have seen for diastolic BP. From the RMPE table, we saw Huber and Tukey's methods provided very close results and less RMPE value than other models. A similar study has been done by Qin et al. (2017), Song & Liang (2015). From the β coefficients results, we have also seen that evaluating of β coefficient of this data, systolic blood pressure, and diastolic blood pressure were the top selection category of all the models.

Table 6.4: The RMPE of different methods when all observations are used for training as well as test data.

RMPE						
Response Variable	OLS	Huber	Tukey	LAD	LASSO	LAD-LASSO
Systolic BP	0.524	0.461	0.469	0.471	0.491	0.495
Diastolic BP	0.525	0.502	0.501	0.525	0.524	0.525
Dimension Reduction						
Response variable	LASSO	LAD-LASSO				
Systolic BP	9.09	80.01				
Diastolic BP	9.1	90				

Table 6.5: The RMPE of different methods when 80% of observations are used for training, and remaining 20% are test data.

RMPE						
Response Variable	OLS	Huber	Tukey	LAD	LASSO	LAD-LASSO
Systolic BP	1.191	0.461	0.469	1.249	1.643	1.554
Diastolic BP	1.525	0.402	0.401	1.575	1.554	1.763

We further split the dataset into two categories. We randomly chose 80% of observations

as training and the remaining 20% as test data. The RMPE of the test data set showed that Huber and Tukey provided a lower RMPE than the other methods. For the systolic BP, Huber provided the lowest value; for the diastolic BP, Tukey provided the lowest RMPE.

Conclusion

We compared different robust methods in linear regression, including Huber and Tukey's M-estimators. We also studied the two variable selection methods – LASSO and LAD-LASSO. We reviewed their theoretical properties and performed a simulation study. Finally, we evaluated how well robust models can forecast systolic and diastolic blood pressure based on different features. To begin with, we explore the data patterns and the critical features of the cardiovascular disease dataset. Root Mean Prediction Error (RMPE) determines how closely predictions or estimates match actual values. To choose the best estimator from the OLS, Huber, Tukey, LAD, LASSO, and LAD-LASSO, we observed that Huber M-estimator and Tukey MM-estimator provided better results. These two estimators give low RMPE in full data and when data is divided into training and test sets. We also observed that the OLS estimate provided a large RMPE for this dataset due to the anomalies or more extreme values.

Bibliography

Breiman, L. (1993), Better subset selection using the non-negative garotte, Technical report, Technical Report. University of California, Berkeley.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *biometrika* **81**(3), 425–455.

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the american statistical association* **90**(432), 1200–1224.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. & Stern, A. S. (1992), 'Maximum entropy and the nearly black object', *Journal of the Royal Statistical Society: Series B (Methodological)* **54**(1), 41–67.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

Hesterberg, T., Choi, N. H., Meier, L. & Fraley, C. (2008), 'Least angle and l_1 penalized regression: A review'.

Qin, Y., Li, S., Li, Y. & Yu, Y. (2017), 'Penalized maximum tangent likelihood estimation and robust variable selection', *arXiv preprint arXiv:1708.05439* .

Song, Q. & Liang, F. (2015), 'High-dimensional variable selection with reciprocal l 1-regularization', *Journal of the American Statistical Association* **110**(512), 1607–1620.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Yohai, V. J. & Maronna, R. A. (1979), 'Asymptotic behavior of m-estimators for the linear model', *The Annals of Statistics* pp. 258–268.

Appendix

R CODE

```
Simulation_LM = function(n=150, p=0.1, trim=p, n_test=100, R=100,
                         sigma=1, mu_outlier = 20){
  # n: sample size (training data)
  # p: contamination proportion
  # trim: trimming proportion for RMPE
  # n_test: size of test data
  # R: number of replication
  # sigma: error sd
  # mu_outlier: location of outliers

  library(glmnet) #for lasso
  library("MTE") #for LADlasso
  library(gamsel)
  library(MASS)

  set.seed(123)

  # for gamsel
  n.basis=5

  k = 10
  # X for Training Data
  # x1 <- runif(n)
  # x2 <- runif(n)
  # x3 <- runif(n)
```

```
# X <- cbind(x1, x2, x3)
# reg <- 5*x1 - 3*x2
X = matrix(runif(n*k), nrow = n)
reg = 5 * X[,1] - 3 * X[,2] + 2 * X[,3] - 4 * X[,4] + 10 * X[,5]

# Test data
# x1 <- runif(n_test)
# x2 <- runif(n_test)
# x3 <- runif(n_test)
# X_test <- cbind(x1, x2, x3)
# reg_test <-  5*x1 - 3*x2
X_test = matrix(runif(n_test*k), nrow = n_test)
reg_test = 5 * X_test[,1] - 3 * X_test[,2] + 2 * X_test[,3] -
  4 * X_test[,4] + 10 * X_test[,5]
eps <- rnorm(n_test,0,sd=sigma)
Y_test = reg_test + eps

# prediction errors
RMPE_train_full = RMPE_test = RMPE_train_trim = matrix(NA, R, 8)
colnames(RMPE_train_full) = colnames(RMPE_train_trim) =
  colnames(RMPE_test) = c("OLS", "Huber", "Tukey", "LASSO", "LAD",
                          "LAD LASSO", "GAM", "GAMSEL")

for (i in 1:R) {

  if((i==1) | !(i%%10)) cat(sprintf("Step %d/%d\n", i, R))
  eps <- rnorm(n, 0, sd=sigma)
  outlier_index = sample(1:n, round(n*p))
  eps[outlier_index] = rnorm(length(outlier_index), mean = mu_outlier, sd=sigma)
```

```
Y <- reg + eps


###### OLS ######
lm_fit = lm(Y ~ X)
beta_lm = lm_fit$coefficients #estimator of beta
RMPE_train_full[i,"OLS"] = sqrt(mean((cbind(1, X) %*% beta_lm - Y)^2))
RMPE_train_trim[i,"OLS"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_lm - Y)^
RMPE_test[i,"OLS"] = sqrt(mean((cbind(1, X_test) %*% beta_lm - Y_test)^2))

###### Huber ######
Huber_fit = rlm(Y ~ X, maxit = 100)
beta_Huber = Huber_fit$coefficients #estimator of beta
RMPE_train_full[i,"Huber"] = sqrt(mean((cbind(1, X) %*% beta_Huber - Y)^2))
RMPE_train_trim[i,"Huber"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_Huber
RMPE_test[i,"Huber"] = sqrt(mean((cbind(1, X_test) %*% beta_Huber - Y_test)^2)

###### Tukey ######
Tukey_fit = rlm(Y ~ X, method = "MM", maxit = 100)
beta_Tukey = Tukey_fit$coefficients #estimator of beta
RMPE_train_full[i,"Tukey"] = sqrt(mean((cbind(1, X) %*% beta_Tukey - Y)^2))
RMPE_train_trim[i,"Tukey"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_Tukey
RMPE_test[i,"Tukey"] = sqrt(mean((cbind(1, X_test) %*% beta_Tukey - Y_test)^2)


###### LASSO ######
lasso_cv = cv.glmnet(x = X, y=Y, alpha = 1, nlambda = 100)
lambda_opt_lasso = lasso_cv$lambda.min #minimum MSE
lasso_fit = glmnet(x = X, y = Y, lambda = lambda_opt_lasso)
#estimators
```

```
beta_lasso = as.numeric(coef(lasso_fit))
RMPE_train_full[i,"LASSO"] = sqrt(mean((cbind(1, X) %*% beta_lasso - Y)^2))
RMPE_train_trim[i,"LASSO"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_lasso
                                                   trim=trim))
RMPE_test[i,"LASSO"] = sqrt(mean((cbind(1, X_test) %*% beta_lasso - Y_test)^2)

# ____________________________________________________________________
#                              LAD
# ____________________________________________________________________
beta_LAD = LAD(y=Y, X = X, intercept = TRUE)
RMPE_train_full[i,"LAD"] = sqrt(mean((cbind(1, X) %*% beta_LAD - Y)^2))
RMPE_train_trim[i,"LAD"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_LAD - Y)
RMPE_test[i,"LAD"] = sqrt(mean((cbind(1, X_test) %*% beta_LAD - Y_test)^2))

# ____________________________________________________________________
#                          LAD LASSO
# ____________________________________________________________________
# LADlasso_fit = LADlasso(y=Y, X = X, beta.ini = beta_LAD, intercept = TRUE)
LADlasso_fit = LADlasso(y=(Y - mean(Y)), X = X, beta.ini = beta_LAD[-1])
beta_LADlasso = c(mean(Y), LADlasso_fit$beta)
RMPE_train_full[i,"LAD LASSO"] = sqrt(mean((cbind(1, X) %*% beta_LADlasso - Y)
RMPE_train_trim[i,"LAD LASSO"] = sqrt(upper.trim.mean((cbind(1, X) %*% beta_LA
RMPE_test[i,"LAD LASSO"] = sqrt(mean((cbind(1, X_test) %*% beta_LADlasso - Y_



# MPE_LADlasso[i] = upper.trim.mean((X_test %*% beta_LADlasso + mean(Y) - Y_te
# ____________________________________________________________________

###== GAM and GAMSEL ==###
n.basis_gamsel = rep(n.basis, ncol(X))
```

```
n.x_unique = apply(X, 2, function(t) length(unique(t)))
low_degree_index = (n.x_unique <= n.basis)
n.basis_gamsel[low_degree_index] = n.x_unique[low_degree_index] - 1

bases = pseudo.bases(X, degree=n.basis_gamsel, df=4)
gamsel_cv = cv.gamsel(x = X, y = Y, family="gaussian", bases=bases)
lambda_opt_gamsel = gamsel_cv$lambda.min
# gamsel needs at least two values of lambda
temp_lambda_gamsel = c(gamsel_cv$lambda.min, 0)
gamsel_fit = gamsel(x = X, y = Y, lambda = temp_lambda_gamsel,
                    bases=bases, family="gaussian")

###### GAMSEL ======##
# for training data
Y_hat_gamsel = predict(object=gamsel_fit, X)[,1]
RMPE_train_full[i, "GAMSEL"] = sqrt(mean((Y_hat_gamsel - Y)^2))
RMPE_train_trim[i, "GAMSEL"] = sqrt(upper.trim.mean((Y_hat_gamsel - Y)^2, trim

#for test data
Y_hat_gamsel = predict(object=gamsel_fit, X_test)[,1]
RMPE_test[i, "GAMSEL"] = sqrt(mean((Y_hat_gamsel - Y_test)^2))



###### MPE for ordinary GAM ======##
# for training data
Y_hat_gam = predict(object=gamsel_fit, X)[,2]
RMPE_train_full[i,"GAM"] = sqrt(mean((Y_hat_gam - Y)^2))
RMPE_train_trim[i,"GAM"] = sqrt(upper.trim.mean((Y_hat_gam - Y)^2, trim=trim))

#for test data
```

```
    Y_hat_gam = predict(object=gamsel_fit, X_test)[,2]
    RMPE_test[i,"GAM"] = sqrt(mean((Y_hat_gam - Y_test)^2))
  }

  average_RMPE_train_full = colMeans(RMPE_train_full)
  average_RMPE_train_trim = colMeans(RMPE_train_trim)
  average_RMPE_test = colMeans(RMPE_test)

  # Print root mean prediction errors over
  cat(sprintf("Root mean prediction errors over %d samples for %g outliers:\n", n,
  cat(sprintf("In full training data:\n"))
  print(average_RMPE_train_full)
  cat(sprintf("In %g%% trimmed training data:\n", 100*trim))
  print(average_RMPE_train_trim)
  cat(sprintf("In test data:\n"))
  print(average_RMPE_test)
  browser()

}

#-------------------------------------------------------------------
#                    Upper trimmed mean
#-------------------------------------------------------------------
upper.trim.mean = function(x,trim=0.05) {
  #trim: the fraction of observations to be trimmed from the top
  if (trim==0) return(mean(x))

  x <- sort(x)
  mean(x[1:floor(length(x)*(1-trim))])
}
```

```
# ____________________________________________________________________________

Simulation_LM()
```

Curriculum Vitae

Jagannath Das is the second child of a very generous lady Pratima Bala Das and a very great gentleman Nitish Das. He was born in a beautiful village near Bay of Bengal, Chittagong Division, Bangladesh. He has done his Bachelor of Science and Master of Science in Statistics, Biostatistics & Informatics from University of Dhaka, Bangladesh. Now he is going to complete his another Master of Science in Statistics and Data Science from the University of Texas at El Paso. In future, he likes to do research on biomedical and genetics data.