AN ALTERNATIVE APPROACH TO SOLVE THE LAD-LASSO PROBLEM

Esra EMIROGLU Kadri Ulas AKAY

PRESENTATION PLAN

1) Informations

2) Our Approach To Solve the LAD-LASSO

3) Example

4) Discussion

1.INTRODUCTION

★ Regression analysis is a statistical process for estimating the relationships among variables.

★ The general objectives of regression analysis are that

- Description of the change in the dependent variable
- Finding the corresponding average value of any observation
- Fitting the best curve to points.

★ The concept of the regression was first used in 1897 by Galton to display some of the relationships and correlations in studies in relation to the theory of genetics [1,2]. Today, regression theory is used widely and computational difficulties have been eliminated with pre-prepared programs (SAS, Minitab, STATGRAPHICS, S-PLUS…).

★ Let us consider the linear regression model which is described as follows

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of the observations, $\mathbf{X}$ is an $n \times p$ matrix of the levels of the regressor variables, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_{p-1}\right)'$ is a $p \times 1$ vector of the unknown coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the random errors satisfying $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

★ In regression analysis, the most important aim is the estimation of unknown parameters. The most popular method is the Least Squares (LS) method. The LS estimator is a solution to the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p-1} x_{ij} \beta_j \right)^2 . \qquad (1.2)$$

★ According to the Gauss-Markov theorem, the LS estimator is the best linear unbiased estimators of $\boldsymbol{\beta}$ when the errors ε_i are normally distributed. On the other hand, when the distribution of the errors is nonnormal and the data has outliers or multicollinearity the LS estimator is known to be very sensitive (Montgomery *et al.*, 2001).

- ★ One of the important problems of regression analysis is multicollinearity. When there are near-linear dependencies among the regressors, the problem of multicollinearity occurs. For multicollinearity, several alternative estimation techniques are proposed, but Ridge regression estimator, proposed by Hoerl and Kennard (1970), is one of the most widely used estimators.

★ Ridge regression estimator $\hat{\boldsymbol{\beta}}_R$ is a solution to the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2 \quad . \qquad (1.3)$$
$$\sum_{j=0}^{p-1} \left(\beta_j \right)^2 \leq s$$

.

★ The ridge regression solutions are easily seen to be

$$\hat{\boldsymbol{\beta}}_R = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \qquad k \geq 0 \tag{1.4}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. Note that when $k = 0$, the ridge estimator is the LS estimator (Montgomery *et al.*, 2001).

★ When there are outliers, robust regression methods are more powerful than the LS method. (Huber, 1981). One of these robust estimation methods is the Least Absolute Deviation (LAD) method. The LAD estimator is a solution to the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij} \beta_j \right| \quad . \qquad (1.5)$$

★ In applications, one can frequently face with *x*-space and/or *y*-space outliers in the data sets. It is known that the LS estimator is unsuccessful in producing a reliable result under these circumstances, and the LAD estimator is better in the case of *y*-space outliers (Arslan, 2011). However, there are some computational difficulties as the number of regressor increases.

★ Variable selection is another important subject in regression analysis. A large number of regressors decrease possible modelling biases. However, including unnecessary regressors yields less accurate predictions. On the other hand, omitting important regressors may produce biased parameter estimates and prediction results. Therefore, selecting the significant regressors is an important task of regression analysis.

★ The problem of selecting a model under suitable conditions for the remainder is studied extensively in the literature. Some of the recommended and best applied methods are the Akaike Information Criterion (AIC) (Akaike 1973), the Bayes Information Criterion (BIC) (Schwarz 1978), and the Mollows-Cp statistic. Theoretically speaking there is no confirmed knowledge as to which criterion will be better (Shi and Tsai, 2002).

★ In order to eliminate this insufficiency, Tibshirani (1996) proposed the following the Least Absolute Shrinkage and Selection Operator (LASSO) which is minimized the penalized LS regression as follows

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p-1}x_{ij}\beta_j\right)^2+n\lambda\sum_{j=0}^{p-1}\left|\beta_j\right| \qquad (1.6)$$

where $\lambda>0$ is tuning parameter.

★ Minimizing criterion in (1.6) is equal to

$$\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2$$

$$\textit{subject to} \quad \sum_{j=0}^{p-1}\left|\beta_j\right| \le s \qquad (1.7)$$

where $s \ge 0$ is tuning parameter selected by the analyst.

★ The finite-dimensional performance of the LASSO estimator under standard errors was shown by Tibshirani (1996) and its statistical properties were studied by Knight and Fu (2000), Fan and Li (2001), Rosset and Zhu (2004) and Zhau and Yu (2006) .

★ However, when errors in (1.1) are distributed in a heavy-tailed manner, the performance of the LASSO becomes weaker due to LS estimator’s sensitivity to the heavy-tailed error distributions and outliers. Due to this sensitivity, the LAD regression which is resistant to outliers and heavy-tailed errors is combined with the LASSO .

★ The obtained LAD-LASSO is successful in simultaneously estimating robust regression and selecting variables. When the LAD and the LAD-LASSO are compared, the LAD-LASSO is seen to be able to perform parameter estimation while at the same time for selecting the model. Also the LAD-LASSO is resistant to heavy-tailed distributions and outliers than the LASSO. The aim of this presentation is to reformulate LAD-LASSO and solve the reformulated LAD-LASSO with the Simplex algorithm .

2.LAD-LASSO

★ The LAD-LASSO is obtained by minimizing the penalized LAD regression criterion as follows

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right| + n\lambda \sum_{j=0}^{p-1} \left|\beta_j\right| \qquad (2.1)$$

where $\lambda > 0$ is tuning parameter.

★ In studies of Wang, Li and Jiang (2007), the parametres are estimated by minimizing the following objective function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij} \beta_j \right| + n \sum_{j=0}^{p-1} \lambda_j \left| \beta_j \right| \qquad (2.2)$$

by using the different tuning parameters for different regression coefficients.

★ They considered an augmented dataset $\{(y_i^*, \mathbf{x}_i^*)\}$ with $i = 1, 2, \ldots, n + p - 1$, where $(y_i^*, \mathbf{x}_i^*) = (y_i, \mathbf{x}_i)$ for $1 \le i \le n$, $(y_{n+j}^*, \mathbf{x}_{n+j}^*) = (0, n\lambda_j \mathbf{e}_j)$ for $1 \le j \le p-1$, and $\mathbf{e}_j$ is a $(p-1)-$dimensional vector with the *j*th component equal to 1 and all others equal to 0. They obtained

$$\text{LAD-LASSO} = \sum_{i=1}^{n+p-1} \left| y_i^* - \mathbf{x}_i^{*\prime} \boldsymbol{\beta} \right|. \qquad (2.3)$$

This is just a traditional LAD criterion. Consequently, any standard unpenalized LAD program (rq in the QUANTREG package of R) can be used to find the LAD-LASSO estimator.

★ In our study, we find that the LAD-LASSO estimator of $\boldsymbol{\beta}$ is obtained by

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n} \left| d_i \right| \\ &\textit{subject to} \quad \sum_{j=0}^{p-1} \left| \beta_j \right| \le t \\ &\mathbf{d}, \boldsymbol{\beta} \ \textit{unrestricted in sign} \end{aligned} \tag{2.4}$$

where $t \ge 0$ is tuning parameter and d_i is defined as $d_i = y_i - \sum_{j=0}^{p-1} x_{ij} \beta_j$.

★ Minimizing (2.4) is equal to

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right| + \lambda \sum_{j=0}^{p-1} \left| \beta_j \right| \qquad (2.5)$$

★ For estimation of β_j parameter in problem (2.5), LAD-LASSO is reformulated as follows

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n} \left| d_i \right| + \lambda \sum_{j=0}^{p-1} \left| \beta_j \right| \\ &\textit{subject to} \;\; \mathbf{X}\boldsymbol{\beta} + \mathbf{d} = \mathbf{y} \\ &\mathbf{d}, \boldsymbol{\beta} \;\; \textit{unrestricted in sign} \end{aligned} \qquad (2.6)$$

★ Also minimizing (2.6) is equal to minimizing

$$
\begin{aligned}
&\min_{\boldsymbol{\beta}} \qquad \sum_{i=1}^{n} \left| d_i \right| \\
&\textit{subject to } \mathbf{X}\boldsymbol{\beta} + \mathbf{d} = \mathbf{y} \\
&\qquad\qquad \sum_{j=0}^{p-1} \left| \beta_j \right| \le t \\
&\mathbf{d}, \boldsymbol{\beta} \;\; \textit{unrestricted in sign}
\end{aligned}
\tag{2.7}
$$

★ Note that $|d_i| = d_{1i} + d_{2i}$ and $d_i = d_{1i} - d_{2i}$ where d_{1i} and d_{2i} are nonnegative and $|\beta_j| = \beta_{1j} + \beta_{2j}$ and $\beta_j = \beta_{1j} - \beta_{2j}$ where β_{1j} and β_{2j} are nonnegative. We can reformulate the problem as

$$\begin{aligned} &\min_{\boldsymbol{\beta}} && \sum_{i=1}^{n} d_{1i} + \sum_{i=1}^{n} d_{2i} \\ &\textit{subject to} && \mathbf{X\beta_1} - \mathbf{X\beta_2} + \mathbf{d_1} - \mathbf{d_2} = \mathbf{y} \\ & && \sum_{j=0}^{p-1} \beta_{1j} + \sum_{j=0}^{p-1} \beta_{2j} \le t \\ & && \mathbf{d_1}, \mathbf{d_2}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2} \ge \mathbf{0} \end{aligned} \qquad (2.8)$$

★ Any $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{d}_1,\mathbf{d}_2)$ satisfying $\mathbf{X}\boldsymbol{\beta}_\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_\mathbf{2}+\mathbf{d}_\mathbf{1}-\mathbf{d}_\mathbf{2}=\mathbf{y}$ is called a solution to (2.4).Let $\begin{pmatrix} \mathbf{X}_{n\times p} & -\mathbf{X}_{n\times p} & \mathbf{I}_{n\times n} & -\mathbf{I}_{n\times n} \\ \mathbf{1}'_{1\times \mathrm{p}} & \mathbf{1}'_{1\times \mathrm{p}} & \mathbf{0}_{1\times n} & \mathbf{0}_{1\times n} \end{pmatrix}$ be denoted by the matrix $\mathbf{A}$ of order $n+1\times 2p+2n$, $(\boldsymbol{\beta}_\mathbf{1},\boldsymbol{\beta}_\mathbf{2},\mathbf{d}_\mathbf{1},\mathbf{d}_\mathbf{2})$ be denoted by the vector $\mathbf{W}'$ of order $1\times 2p+2n$ and $\begin{pmatrix} \mathbf{y}_{n\times 1} \\ t_{1\times 1} \end{pmatrix}$ be denoted by the vector $\mathbf{P}$ of order $n+1\times 1$. Any $\mathbf{W}$ satisfying $\mathbf{AW}\begin{Bmatrix} = \\ \leq \end{Bmatrix}\mathbf{P}$ is a solution to (2.4).

★ Let $\mathbf{c}'$ be the vector $\left(\mathbf{0}_{1\times p},\mathbf{0}_{1\times p},\mathbf{1}'_{1\times n},\mathbf{1}'_{1\times n}\right)$ where $\mathbf{0}=\left(0,0,\ldots 0\right)$ and $\mathbf{1}'=(1,1,\ldots 1)$. Then $\mathbf{c}'\mathbf{w}$ is called the objective function of problem (2.4). Any solution $\mathbf{w}$ to (2.4), if it further satisfies $w_j \geq 0$, $j=1,2,\ldots,2p+2n$, we call it a feasible solution to problem. Thus LAD-LASSO is translated into a mathematical programming problem and can be solved with Simplex Algorithm.

3.EXAMPLE

★ To illustrate parameter estimation by using LAD-LASSO, we consider Hald data, which is used widely in literature. Hald (1952) present data concerning the heat evolved in calories in calories per gram of cement $(\mathbf{y})$ as a function of the amount of each of four ingredient in the mix: tricalcium aluminate $(\mathbf{x}_1)$, tricalcium silicate $(\mathbf{x}_2)$, tetracalcium alumino ferrite $(\mathbf{x}_3)$, and dicalcium silicate $(\mathbf{x}_4)$. The data is shown in Table 3.1.

Table 3.1 Hald Cement Data

Observation i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

★ Simple correlations are shown in Table 3.2. Note that the pairs of regressor variables $(\mathbf{x}_1, \mathbf{x}_3)$ and $(\mathbf{x}_2, \mathbf{x}_4)$ are higly correlated since $r_{13} = -0.824$ and $r_{24} = -0.973$.

Table 3.2 Simple Correlations

	x_1	x_2	x_3	x_4	y
x_1	1				
x_2	0.229	1			
x_3	-0.824	-0.139	1		
x_4	-0.245	-0.973	0.030	1	
y	0.731	0.816	-0.535	-0.821	1

★ On the other hand, statistics for detecting outliers for the Hald cement data set is given in Table 3.3. Based on the result of Table 3.3, e_6, e_8 and e_{13} residual seem suspiciously different . Therefore, we can say that Hald data has *y*-direction outliers. On the other hand, according to the leverage $\left(h_{ii}\right)$, Cook’s distance and DFITS values, it seems that there is no *x*-direction outliers in Table 3.3. In this situation, LAD regression is much more powerful estimation method than LS regression.

Table 3.3 Statistics for detecting outliers for Hald Cement Data

ID	y	$\hat{y}$	e_i	h_{ii}	Cook's Distance	DFITS
1	78.5	78.495	0.005	0.473	0	0.006
2	74.3	72.789	1.511	0.256	0.057	0.755
3	104.3	105.971	-1.671	0.500	0.301	-2.279
4	87.6	89.327	-1.727	0.218	0.059	-0.724
5	95.9	95.649	0.251	0.281	0.002	0.140
6	109.2	105.275	3.925	0.047	0.083	0.556
7	102.7	104.149	-1.449	0.290	0.064	-0.840
8	72.5	75.675	-3.175	0.332	0.394	-2.193
9	93.1	91.722	1.378	0.217	0.038	0.575
10	115.9	115.619	0.282	0.623	0.021	0.658
11	83.8	81.809	1.991	0.349	0.171	1.475
12	113.3	112.327	0.973	0.186	0.015	0.347
13	109.4	111.694	-2.294	0.227	0.110	-1

★ Because of this results, we can say that this data has outliers and serious multicollinearity. If we want to estimate parameters and select significant regressors simultaneously under these circumstances. We will use LAD-LASSO estimator. Finally in Table 3.4, the parameter estimates based on reformulated LAD-LASSO are given with various *t* values.

Table 3.4 Estimates of Reformulated LAD-LASSO

t	0	0.001	1.61	2.16	3.058	4.690	4.691	10	18.579	64.424
β_0	0	0	0	0	0	0	-0.0003	-5.098	-13.337	-13.337
β_1	0	0	0	0.008	1.008	2.213	2.213	2.267	2.354	2.354
β_2	0	0.001	1.609	1.491	1.437	1.145	1.145	1.196	1.280	1.280
β_3	0	0	0	0	0.0003	0.865	0.865	0.920	1.007	1.007
β_4	0	0	0.001	0.661	0.612	0.468	0.468	0.518	0.601	0.601
MS_{Res}	9314.	10080.	638.75	126.54	33.63	6.68	7.51	7.51	7.64	7.64

★ According to the result which are obtained in Table 3.4, variable selection is done between 0 and 4.691. In this range, a model which has less parameter, is obtained for a suitable *t* value. Therefore the obtained model is less affected from multicollinearity and outliers.

★ After this point which variables selection stops. Notice that if t is chosen larger than $\sum_{j=0}^{p-1}\left|\hat{\beta}_j^{LAD}\right|$, the LAD-LASSO estimates are equal to $\hat{\boldsymbol{\beta}}^{LAD}$. On the other hand, until 4.691, while t increases, MS_{Res} decreases. Therefore the best point t is previous point from 4.691.

4.DISCUSSION

★ In this study, the tuning parameter is in augmented observations vector in our approach but in study of Wang, Li and Jiang (2007) the different tuning parameters are in augmented regressor variables matrix for different regressor coefficients. Therefore the dimension of matrix is larger and using Simplex Algoritm is more difficult than ours. The other difference is the range of tunig parameter is known in our approach.

★ Finally based on the analysis result of Hald Data, by using the reformulated LAD LASSO, it is shown that a regression model, which is less affected from multicollinearity and outliers, can be obtained for suitable *t* value.

5.REFERENCES

[1] AKAIKE, H. (1973) Information Theory and an Estimation of the Maximum Likelihood Principle. In 2nd International Symposium on Information Theory, eds. B. N. Petrov and F.Csaki, Budapest:Akademia Kiado, pp. 267-281.

[2] ARSLAN, O. (2011) Weighted LAD-LASSO Method for Robust Parameter Estimation and Variable Selection in Regression. Computational Statistics and Data Analysis 56, 1952-1965.

[3] ARTHANARI, T. S., DODGE, Y., (1993) Mathematical Programming in Statistics, John Wiley&Sons Inc., New York, USA.

[4] FAN and LI, (2001), Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties, Journal of the American Statistical Association, 96, 1348-1360.

[5] FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2001) The elements of statistical learning. New York: Springer Series in Statistics, 2001.

[6] HOERL, A. E. and KENNARD, R. V. (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

5.REFERENCES

[7] MONTGOMERY, D. C., PECK, E. A., VINIG, G. G. (2001) Introduction to Linear Regression Analysis, 3th. Ed. John Wiley &Sons Inc., USA.

[8] ROBERT TIBSHIRANI, (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B Vol. 58, No. 1, pp 267-288.

[9] ROSSET, S., ZHU, J. (2004), Least Angle Regression: discussion. The Annals of Statistics 32, 469-475.

[10] SCHWARZ, G. (1978), Estimating the Dimension of a Model. The Annals of Statistics, 6, 461-464.

[11] SHI, P. and TSAI, C. L. (2002), Regression Model Selection a Residual Likelihood Approach. Journal of the Royal Statistical Society: Series B 64.2: 237-252.

[12] TIBSHIRANI, 1996, Regression Shrinkage and Selection via the LASSO, Journal of the Royal Statistical Society, Ser.B, 58, 923-941.

[13] WANG, H., LI, G. and JIANG, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through The LAD_LASSO. Journal of Business & Economic Statistics 25, 347-355.

Mehmet Çetin
Bozkurt.B