ESTIMATION OF REGRESSION COEFFICIENTS USING GEOMETRIC MEAN OF SQUARED ERROR FOR SINGLE INDEX LINEAR REGRESSION MODEL

Prasanna Mayilvahanan

Research Assistant, Analytics Vidhya
B.Tech, Department of Civil Engineering, Indian Institute of Technology – Guwahati

ABSTRACT

Regression models and their statistical analyses is one of the most important tool used by scientists and practitioners. The aim of a regression model is to fit parametric functions to data. It is known that the true regression is unknown and specific methods are created and used strictly pertaining to the problem. For the pioneering work to develop procedures for fitting functions, we refer to the work on the methods of least absolute deviations, least squares deviations and minimax absolute deviations. Today's widely celebrated procedure of the method of least squares for function fitting is credited to the published works of Legendre and Gauss. However, the least squares based models in practice may fail to provide optimal results in non-Gaussian situations especially when the errors follow distributions with the fat tails. In this paper an unorthodox method of estimating linear regression coefficients by minimising GMSE(geometric mean of squared errors) is explored. Though GMSE(geometric mean of squared errors) is used to compare models it is rarely used to obtain the coefficients. Such a method is tedious to handle due to the large number of roots obtained by minimisation of the loss function. This paper offers a way to tackle that problem. Application is illustrated with the 'Advertising' dataset from ISLR and the obtained results are compared with the results of the method of least squares for single index linear regression model.

KEYWORDS

Linear Regression, Geometric Mean of Squared Errors, Geometric regression

1. INTRODUCTION

The supervised learning in the form of classification and regression is an significant constituent of statistics and machine learning. The most basic supervised learning techniques is Linear regression which is the basic building block of all Machine Learning models. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. The linear regression model assumes a linear relationship between the response and the predictor variables as show.

$$Y_i^* = X_i^T\beta + \varepsilon_i,^* \qquad i = 1,\ldots\ldots,n \qquad (1)$$

Where Y_i^* is the response variable(dependent variable), X_i^T is the set of predictor variables(set of independent variables), β is the set of coefficients we estimate and $\varepsilon_i^*$is the unobservable error term independent of X_i. Various procedures have been proposed to estimate β [5, 6]. The estimation of β is entirely dependent on how we define 'error'. In all known procedures of estimation, a standard error term is established first and the error is minimised to obtain β. The most popular techniques were the *L1* and the *L2* norms [1, 2]. The L1 is otherwise known as Least absolute deviations method and L2 is known as the Least squares method. These methods are widely used due to their intuitive and mathematical simplicity. In the L1 norm the error term

DOI: 10.5121/ijaia.2016.7606

is established as the arithmetic mean of absolute differences and the in L2 norm the error term is established as the arithmetic mean of squared differences [3]. Latest techniques include estimation of coefficients using minimax regret and algorithmic methods [14, 15].

In this paper, we establish the error term as the geometric mean of squared differences/errors(GMSE). The robustness to large values of the Geometric Mean in comparison to the Arithmetic Mean calls for such a such a method. The loss function is established as the geometric mean of squared errors(GMSE), which is to be minimised to estimate the coefficients. Unlike the popular norms, minimising this function provides a large number of real roots(estimates of β). In other words minimising the function returns a large number of models as the loss function has a substantial number of minima. To tackle this problem we use a non-linear optimizing technique with a threshold to filter and retrieve a reasonable number of models similar to some methods used in [15]. The best fit is selected by choosing the model with the least geometric mean of squared differences(GMSE) of the retrieved models. The loss function has a large number of minima due to zero inputs in the geometric mean of squared errors(GMSE). Various techniques have been proposed to calculate geometric mean when there are zero values [16]. Although such techniques are intriguing the scope of this work does not include them. The proposed method is also dependent on the optimizing technique. Although there are a number of robust optimizing techniques like the stochastic gradient descent,In this paper we use fmincon owing to its simplicity and usability [17, 18].

In Section 2 we look at a brief overview of the popular methods of estimation. In Section 3 we frame our primary method(Minimisation of Geometric Mean Squared Error) of estimation and state the challenges faced with ways to tackle them in this method of estimation. We also develop an Algorithm to carry out the proposed method in Section 3. Section 4 includes the working of the Algorithm on real-time data with the interpretation of results. In the following sections we conclude the work with the provision of appropriate references.

2. Brief Overview Of Standard Error Measures And Estimation Techniques

2.1 Introduction

Most of the existing estimation techniques involve minimisation of the loss function. Thus depending on how error is defined. Methods have been proposed to deal with high dimensional regression problems and also a number of sparse and bayesian estimation methods are popular too [19 - 22]. This paper does not deal with such methods. In the following subsections we look at the classic error measures which are easy and intuitive to understand and we also discuss how they are minimised to estimate the regression coefficients. The error measures we look at are Mean squared Error(MSE), Mean of Absolute Errors(MAE) and the Least Absolute Relative Error(LARE).

2.2 The Least Squares Method

The mean square error is defined as follows(MSE):

$$\sum_{i=1}^{n} (Y_i^* - X_i^T \beta)^2 \qquad (2)$$

Where Y_i* is the set of values of the response variable (dependent variable), X_i^T is the corresponding values of the set of predictor variables (set of independent variables), n is the number of data points and β is the set of coefficients we want to estimate.

The MSE, as its name implies, provides for a quadratic loss function as it squares and subsequently averages the various errors. Such squaring gives considerably more weight to large errors than smaller ones (e.g., the square error of 100 is 10000 while that of 50 and 50 is only 2500 + 2500 = 5000, that is half). MSE is, therefore, useful when we are concerned about large errors whose negative consequences are proportionately much bigger than equivalent smaller ones (e.g., a large error of 100 vs two smaller ones of 50 each) [7].

The two biggest advantages of MSE or RMSE are that they provide a quadratic loss function and that they are also measures of the uncertainty in forecasting. Their two biggest disadvantages are that they are absolute measures that make comparisons across forecasting horizons and methods highly problematic as they are influenced a great deal by extreme values [7].

To estimate β the loss function, that is the MSE is minimised to give a set of linear equations. A comprehensive look at the procedure can be found in [8].

2.3 The Least Absolute Deviation Method

The mean absolute error or mean absolute deviation is defined as follows(MAE):

$$\sum_{i=1}^{n} |Y_i^* - X_i^T \beta| \qquad (3)$$

Where Y_i* is the set of values of the response variable (dependent variable), X_i^T is the corresponding values of the set of predictor variables(set of independent variables), *n* is the number of data points and β is the set of coefficients we want to estimate.

The MAE is also an absolute measure like the MSE and this is its biggest disadvantage. However, since it is not of quadratic nature, like the MSE, it is influenced less by outliers. Furthermore, because it is a linear measure its meaning is more intuitive; it tells us about the average size of forecasting errors when negative signs are ignored. The biggest advantage of MAE is that it can be used as a substitute for MSE for determining optimal inventory levels. The MAE is not used much by either practitioners or academicians [7].

Here, the MAE is minimised to obtain β**.** A detailed procedure can be found in [9].

2.4 The Least Absolute Relative Error Estimation Method

The Least Absolute relative error is defined as follows(LARE):

$$\sum_{i=1}^{n} |\{Y_i - X_i^T \beta)\} / Y_i| \qquad (4)$$

Where Y_i* is the set of values of the response variable(dependent variable), X_i^T is the corresponding values of the set of predictor variables(set of independent variables), *n* is the number of data points and β is the set of coefficients we want to estimate.

The Least Absolute relative error(LARE) is a relative measure which expresses errors as a percentage of the actual data. This is its biggest advantage as it provides an easy and intuitive way of judging the extent, or importance of errors. Least Absolute relative error(LARE) is used a great deal by both academicians and practitioners and it is the only measure appropriate for evaluating budget forecasts and similar variables whose outcome depends upon the proportional size of errors relative to the actual data (e.g., we read or hear that the sales of company X increased by

3% over the same quarter a year ago, or that actual earnings per share were 10% below expectations) [7].

The two biggest disadvantages of Least Absolute relative error(LARE) are that it lacks a statistical theory (similar to that available for the MSE) on which-to base itself and that equal errors when Yi is larger than XiTβ then LARE gives smaller relative errors than when Yi is smaller than $X_i^T\beta$ [7].

Squared relative error is also widely used. A detailed explanation of the Least Absolute relative error(LARE) coefficient estimation procedure can be found in [6], [10] & [11].

3. The Approach: The Geometric Mean Of Squared Error And The Proposed Estimation Procedure

3.1 Introduction

Geometric means average the product of square errors rather than their sums as in MSE. The geometric mean of squared error(GMSE) is therefore defined as:

$$\prod_{i=0}^{n}(Y_i^* - X_i^T\beta)^{2/n} \tag{5}$$

Where Y_i* is the set of values of the response variable(dependent variable), X_i^T is the corresponding values of the set of predictor variables(set of independent variables), *n* is the number of data points and β is the set of coefficients we want to estimate.

The biggest advantage of geometric means is that they are influenced to a much lesser extent from outliers than squared means or absolute means [7].

The coefficients of regression (β) can be estimated by minimising the loss function:

$$L = \prod_{i=0}^{n}(Y_i^* - X_i^T\beta)^{2/n} \tag{6}$$

$$\frac{\partial L}{\partial a_i} = 0 \qquad i = 1,2,\ldots p \tag{7}$$

$$Y - X^T\beta = 0 \tag{8}$$

The loss function is not as simple as the quadratic loss function of Least Squares Method [12]. The next section deals with the various challenges in obtaining the best linear fit by minimising the Loss function of the GMSE and retrieving the best linear fit for the data. Equation (8) is the final parametric equation we are trying to obtain.

3.2 Challenges In Obtaining The Best Linear Fit

3.2.1 Large Number Of Minima

As you can see in Figure 1, the Loss function of the Least squares method contains a single global minima from which the best linear fit can be obtained. Whereas the loss function of GMSE contains a large number of minima as shown Figure 2. This calls for:

a) An optimization algorithm to retrieve a finite number of minima.

b) A metric to compare the obtained minima so as to obtain the one best linear fit to the data.

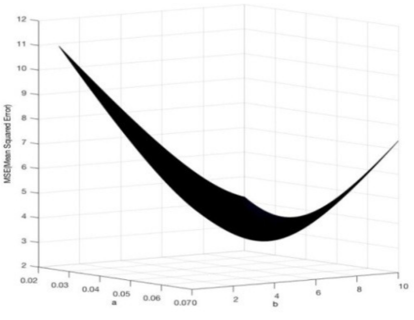


Figure 1. Mean Squared Error Vs B(A, B) (Loss Function for One Predictor Variable)

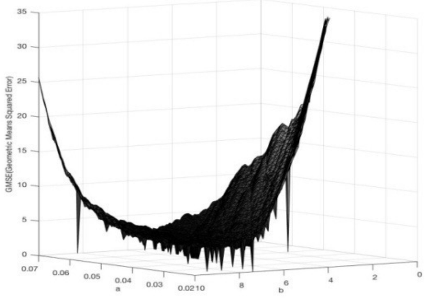


Figure 2. Geometric Mean Squared Error Vs B(A, B) (Loss Function for One Predictor Variable)

3.2.2 LARGE NUMBER OF UNWANTED B

Clearly $L = \prod_{i=0}^{n} (Y_i^* - X_i^T \beta)^{2/n}$ is a function greater than or equal to zero. So at certain minima the value of the Geometric Loss function(Equation 6) is 0 as shown in Figure 2. Say, for any fixed data :

$$\text{If } L = 0$$

$$\text{Then } \beta \text{ is such that at least one } Y_i^* - X_i^T\beta = 0$$

Hence a large number of minima will return values of β such that the linear modelY- $X^T\beta = 0$ passes through exactly one point in the data, or exactly any two points in the data, or exactly any three collinear points in the data and so on. These β do not generalise the data well and is rarely the best linear fit for the data as shown in Figure 3 and Figure 4. For example say if a dataset consist of the points (0, 0) and (1, 0) then linear models {y = x, y = 2x, y = 3x etc.} and {y = 1, y= 2x + 1, y = 3x + 1} all are going to give the least possible value for the Geometric Loss function (Equation 6) that is zero, but not all are going to generalize the overall data set.

To tackle this problem we introduce a threshold to which we crop the Loss function. That is instead of minimising the empirical Geometric Loss function, we minimise only the part of the Geometric Loss function greater than a given threshold(t) which is close to 0 but greater than 0.

$$L = \prod_{i=0}^{n} (Y_i^* - X_i^T \beta)^{2/n} > t \tag{9}$$

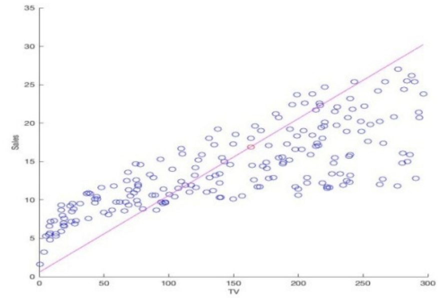


Figure 3. B(A, B) Obtained by Minimising the L = 0 Giving a Line that Passes through just One Points

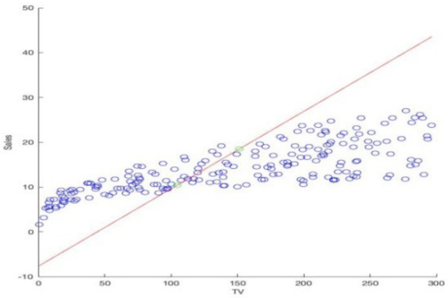


Figure 4. B(A, B) Obtained by Minimising the L = 0 Giving a Line that Passes through just Two Points

3.2.3 THE ALGORITHM

a. The Loss function (Equation 6) is coded and plotted fixing the threshold at t(some value close to zero) as concluded in section 3.2.2.

b. A random β is initialized and the Loss function(Equation 6) is minimised by an optimization algorithm (section 3.2.1a) to the threshold t and the corresponding β is obtained and noted [25].

c. Step b) is repeated for ‘i’ iterations and the corresponding β are obtained and noted.

d. Now that we have ‘i’ good β values(‘i’ linear models), we need a metric to compare them and obtain the best fit for the data. We calculate the Geometric Mean of Squared Errors(GMSE) for each model and select the model with the least Geometric Mean of Squared Errors(GMSE).

4. Calculations And Results

4.1 The Data Set

In this paper the algorithm is demonstrated from the 'advertising' data set from ISLR [4] as shown in Figure 3. The Advertising dataset consists of sales of the product in 200 different markets(200 data points), along with advertising budgets for the product in each of those markets for three different media: TV, Radio and Newspaper. In this paper we only consider Sales vs TV and other variables are ignored. All optimizations and calculations are performed in MATLAB.

Since the data set contains only one predictor variable the parametric equation(Equation 8) reduces to:

$$Y- a*X-b = 0 \qquad (10)$$

Where $\beta = (a, b)$, $X^T = (TV, 1)$ and $Y= (Sales)$

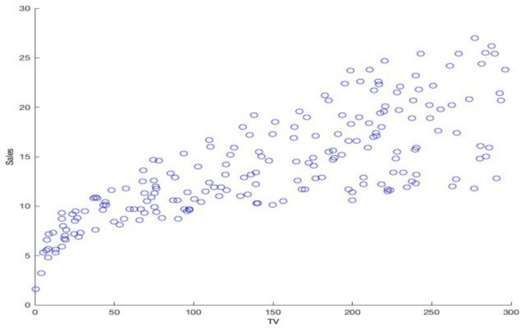


Figure 5. The Plot Displays Sales in Thousands of Units and TV Budgets in Thousands of Dollars

4.2 Performing The Algorithm

a) The Geometric Loss function L (Equation 9) is such that (Yi, Xi) are the $(Sales, TV)$ data points from the data set is input and plotted in MATLAB. Here, threshold t is taken as 0.001.

b) A random $\beta = (a, b)$ is initialized between the interval (0.01, 0.06) for *a* and (1, 10) for *b*. The Geometric Loss function(Equation 6) is minimised by a non-linear optimization solver *fmincon* [13] to the threshold t = 0.001 and the corresponding β is obtained and noted [23, 24]. See Table 1.

c) Now, Step b) is repeated for 'i' = 15 iterations and the corresponding β are obtained and noted. See Table 1.

d) Now that we have '15' good β values('15' linear models), the Geometric Mean Squared Error(GMSE) for each model is calculated as show in and select the model with the least MSE as shown in Table 1.

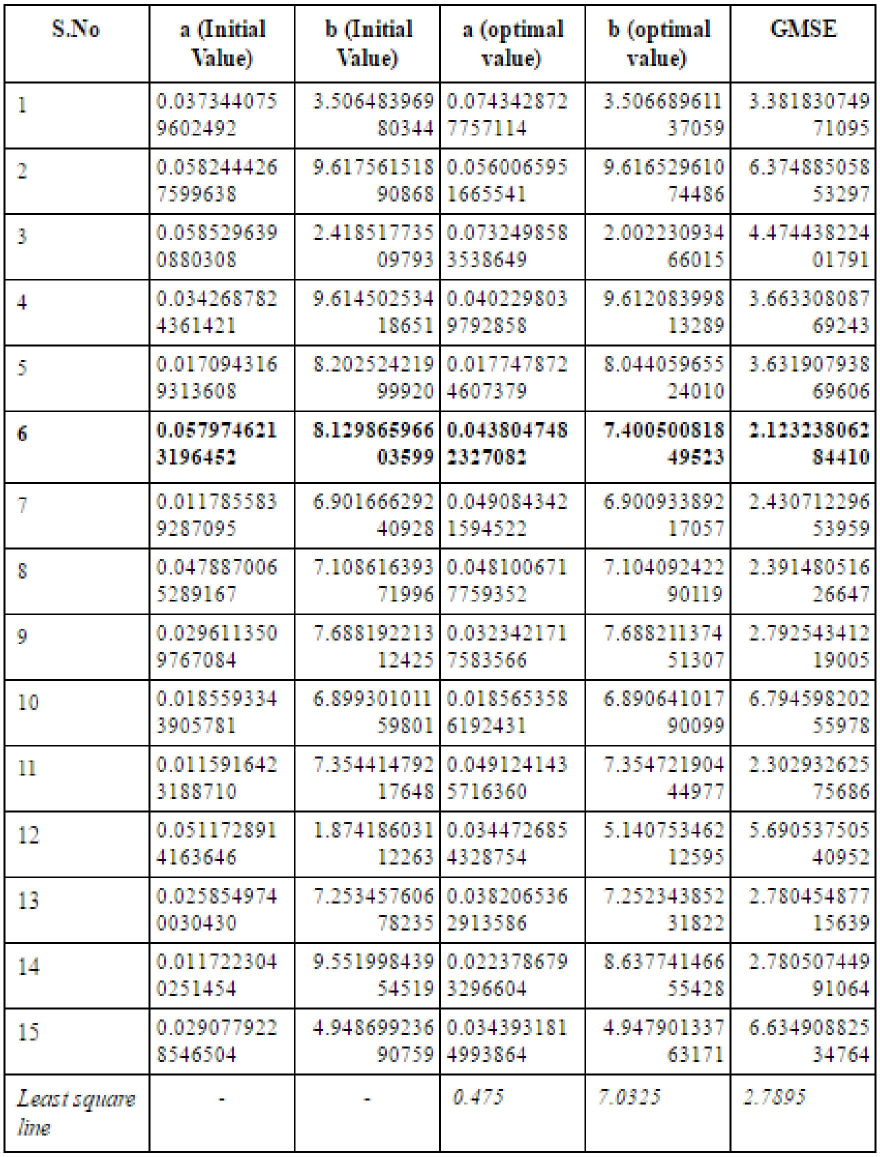
4.3 Plotting The Results

Following the algorithm we obtain the values of $\beta = (0.0438,\ 7.4005)$, thus obtaining the parametric equation:

$$Y- 0.0438*X-7.4005 = 0 \qquad (11)$$

The least square line obtained by standard procedure is [8]:

$$- 0.0475*X-7.0325 = 0 \qquad (12)$$



S.No	a (Initial Value)	b (Initial Value)	a (optimal value)	b (optimal value)	GMSE
1	0.0373440759602492	3.50648396980344	0.0743428727757114	3.50668961137059	3.38183074971095
2	0.0582444267599638	9.61756151890868	0.0560065951665541	9.61652961074486	6.37488505853297
3	0.0585296390880308	2.41851773509793	0.0732498583538649	2.00223093466015	4.47443822401791
4	0.0342687824361421	9.61450253418651	0.0402298039792858	9.61208399813289	3.66330808769243
5	0.0170943169313608	8.20252421999920	0.0177478724607379	8.04405965524010	3.63190793869606
6	**0.0579746213196452**	**8.12986596603599**	**0.0438047482327082**	**7.40050081849523**	**2.12323806284410**
7	0.0117855839287095	6.90166629240928	0.0490843421594522	6.90093389217057	2.43071229653959
8	0.0478870065289167	7.10861639371996	0.0481006717759352	7.10409242290119	2.39148051626647
9	0.0296113509767084	7.68819221312425	0.0323421717583566	7.68821137451307	2.79254341219005
10	0.0185593343905781	6.89930101159801	0.0185653586192431	6.89064101790099	6.79459820255978
11	0.0115916423188710	7.35441479217648	0.0491241435716360	7.35472190444977	2.30293262575686
12	0.0511728914163646	1.87418603112263	0.0344726854328754	5.14075346212595	5.69053750540952
13	0.0258549740030430	7.25345760678235	0.0382065362913586	7.25234385231822	2.78045487715639
14	0.0117223040251454	9.55199843954519	0.0223786793296604	8.63774146655428	2.78050744991064
15	0.0290779228546504	4.94869923690759	0.0343931814993864	4.94790133763171	6.63490882534764
Least square line	-	-	*0.475*	*7.0325*	*2.7895*

Table 1. Initial B Values, their Corresponding Final B Values and their Geometric Mean of Squared Errors(GMSE)

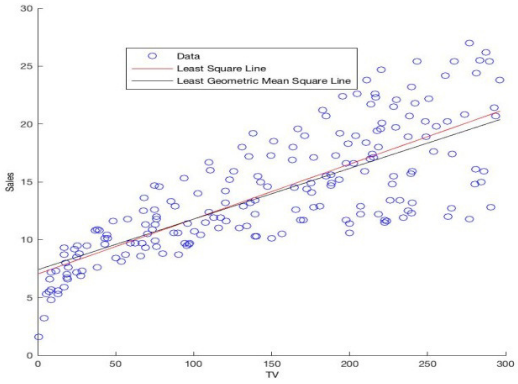


Figure 6. Scatterplot of Sales Vs TV with the Least Squares Line(Equation 11) and the Geometric Mean Square Line (Equation 10)

5. Conclusion

With this approach we have successfully obtained a line comparable to the Least squares line (11) as shown in Figure 4. The Geometric regression line (12) has a lesser GMSE(2.123238) compared to the GMSE(2.7895) of the Least square line(See Table 1). The advantages of this model is its robustness to outliers when compared other linear models like the L1 and L2 norm. Also the method is not computationally expensive. The disadvantages are that the results could vary with the change in initial β, also this method is not simple and intuitive like the L2 norm.

Further work includes usage of a better optimizer, increase in the number of iterations and extension to multivariate linear models. Also the geometric mean of squared error(GMSE) term can be improved to deal with non-positive terms [16].

References

[1] A. M. Legendre, “Nouvelles methodes pour la determination des orbites des cometes,” Mme. Courcier, Paris, 1805.

[2] C. F. Gauss, “Theoria motus corporum coellestium in sectionibus conicis solem ambientium,” Hamburg, 1809.

[3] Pranesh Kumar and Jai Narain Singh ,“Regression Model Estimation Using Least Absolute Deviations, Least Squares Deviations and Minimax Absolute Deviations Criteria”, IJCSEE Volume 3, Issue 4 (2015).

[4] Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, “Introduction to Statistical Learning”, Springer Texts in Statistics, ISBN 978-1-4614-7138-7 .

[5] P. S. Laplace, “Sur quelques points du système du monde,” Mme. De l’Academie Royale des Sciences de Paris, 1793.

[6] Kani CHEN, Shaujan GUO, “Least Absolute Relative Error Estimation,” J Am Stat Assoc. 2010; 105(491): 1104–1112.

[7] S. MAKRIDAKIS and M. HIBON, “EVALUATING ACCURACY (OR ERROR) MEASURES”,Printed at INSEAD, Fontainebleau, France; 95/18/TM.

[8] Steven J. Miller, “The Method of Least Squares”, Mathematics Department Brown University, 2006.

[9] TERRY E. DIELMAN, “Least absolute value regression: recent contributions”, Journal of Statistical Computation and Simulation Vol. 75, No. 4. April 2005. 263-286.

[10] Zhanfeng Wanga, Zimu Chena, Yaohua Wua, “A relative error estimation approach for single index model”, arXiv:1609.01553v1 [stat.ME] 6 Sep 2016.

[11] Arnaud de Myttenaere, Boris Golden , B´en´edicte Le Grand, Fabrice Rossic, “Mean Absolute Percentage Error for Regression Models”, arXiv:1605.02541v1 [stat.ML] 9 May 2016.

[12] Mark Schmidt, “Least Squares Optimization with L1-Norm Regularization”, Journal: CS542B Project Report, 2005/12.

[13] MATLAB solver, “fmincon”, https://in.mathworks.com/help/optim/ug/fmincon.html.

[14] Peter L. Bartlett, Wouter M. Koolen, “Minimax Fixed-Design Linear Regression”, JMLR: Workshop and Conference Proceedings vol 40:1–14, 2015.

[15] Dimitris Bertsimas, Angela King, “OR Forum—An Algorithmic Approach to Linear Regression ”, OPERATIONS RESEARCH Vol. 64, No. 1, January–February 2016.

[16] Elsayed A. E. Habib, “GEOMETRIC MEAN FOR NEGATIVE AND ZERO VALUES”, IJJRAS Vol 11 Issue 3 Jun 2012.[17] Jun He, “Adaptive Stochastic Gradient Descent on the Grassmannian for Robust Low-Rank Subspace Recovery and Clustering”, arXiv:1412.4044v2 [stat.ML] 18 Apr 2015.

[18] Danilo P. Mandic, “A Generalized Normalized Gradient Descent Algorithm”, IEEE SIGNAL PROCESSING LETTERS, VOL. 11, NO. 2, FEBRUARY 2004.

[19] Yan Zhang, Brian J. Reich and Howard D. Bondell, “High Dimensional Linear Regression via the R2-D2 Shrinkage Prior”, arXiv:1609.00046v1 [stat.ME] 31 Aug 2016.

[20] Chang Liu, Bo Li, “Robust High-Dimensional Linear Regression”, arXiv:1608.02257v2 [cs.LG] 9 Aug 2016.

[21] Brendan Juba, “Conditional Sparse Linear Regression”, arXiv:1608.05152v1 [cs.LG] 18 Aug 2016.

[22] Shuichi Kawano, Hironori Fujisawa, Toyoyuki Takada, “Sparse principal component regression for generalized linear models”, arXiv:1609.08886v3 [stat.ML] 13 Oct 2016.

[23] Coleman, T.F. and Y. Li, "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," SIAM Journal on Optimization, Vol. 6, pp. 418-445, 1996.

[24] Gill, P.E., W. Murray, and M.H. Wright, Practical Optimization , Academic Press, London, 1981.

[25] Powell, M.J.D., "The Convergence of Variable Metric Methods For Nonlinearly Constrained Optimization Calculations," Nonlinear Programming 3, (O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, eds.) Academic Press, 1978.