Journal of Data Science and Modeling, Vol. 1, No. 1, 21-32, December 2022

Research Manuscript

Differenced-Based Double Shrinking in Partial Linear Models

Mina Norouzirad[1]*, Mohammad Arashi[2], Mahdi Roozbeh[3]

1. PhD graduate in statistics, Shahrood University of Technology, Shahrrod, Iran.
2. Associate professor of statistics, Shahrood University of Technology, Shahrrod, Iran.
3. Associate professor of statistics, School of Sciences, Semnan University, Semnan, Iran.

Recieved: 20/3/2017 Accepted: 1/8/2017

Abstract:

The partial linear model is very flexible when the relationship between the covariates and responses, is either parametric or nonparametric. However, the estimation of the regression coefficients is challenging since one must also estimate the nonparametric component simultaneously. As a remedy, the differencing approach, to eliminate the nonparametric component and estimate the regression coefficients, can be used. Here, suppose the regression vector-parameter is subjected to lie in a sub-space hypothesis. In situations where the use of difference-based least absolute and shrinkage selection operator (D-LASSO) is desirable, we propose a restricted D-LASSO estimator. To improve its performance, LASSO-type shrinkage estimators are also developed. The relative dominance picture of suggested estimators is investigated. In particular, the suitability of estimating the nonparametric component based on the Speckman approach is explored. A real data example is given to compare the proposed estimators. From the numerical analysis, it is obtained that the partial difference-based shrinkage estimators perform better than the difference-based regression model in average prediction error sense.

Keywords: Double shrinking, Partial linear model, Preliminary test LASSO, Restricted LASSO, Stein-type shrinkage LASSO.

Mathematics Subject Classification (2010): 62G08, 62F15.

*Corresponding author: mina.norouzirad@gmail.com

1. Introduction

Partial linear regression models are popular semi-parametric modeling techniques that assume the response to be linearly dependent on some predictors, whereas its relation to other additional variables is nonparametric functions. In these models, some of the relations are believed to be of certain parametric form while others are not easily parameterized.

A partial linear model (PLM) has the following general form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}(\mathbf{t}) + \boldsymbol{\epsilon} \tag{1.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is the ith observed vector of explanatory variables including p-dimensional vector of regression coefficients, $\mathbf{g}(\mathbf{t}) = (g(t_1), \ldots, g(t_n))^\top$, t_i's are values of an extra univariate variable satisfying $0 \leq t_1 \leq \ldots \leq t_n \leq 1$, $g(t_i)$ is an unknown bounded real-valued function defined on $[0, 1]$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-vector unknown parameters.

Generally, assume that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a vector of unobservable random errors distributed with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2\mathbf{I}_n$ where $\mathbf{I}_n$ is an identity matrix of order n.

The PLM generalizes both parametric and nonparametric regression models which correspond to the cases $\mathbf{g}(\mathbf{t}) = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{0}$, respectively. The key idea is to estimate the parameter vector $\boldsymbol{\beta}$, the function $\mathbf{g}(\mathbf{t})$.

PLMs have been received considerable attention in statistics and econometrics. These models were originally studied by Engle et al. (1986) to determine the effect of weather on the electricity sales. Some earlier surveys of the estimation and application of model (1.1) can be found in the monograph of Hardle, Liang and Gao (2000). In the last decade, several authors have investigated the PLM, including Bunea (2004), Liang (2006), Sun, Kopciuk and Lu (2008), and Aydin (2014), among others.

Now, suppose that we are provided with some prior information about the whole or subset of covariates. This prior information can be utilized to improve the overall estimation of the regression coefficients using shrinkage estimation (Ahmad and Raheem , 2012). Many notable studies are incorporating prior information, in the form of restrictions, to improve estimation in the sense that the restricted and shrinkage estimators have lesser risk and prediction error values.

The organization of this study is given as follows: the full model estimators are given in section 2, the preliminary test, shrinkage estimators are also presented in section 3. Section 4 consists of a real data example that illustrates the usefulness of

the suggested estimators. Finally, the conclusion and remarks are given in section 5.

2. Full Model Estimation

In this situation, difference-based technique has been used to remove the nonparametric component in the PLM. Let $\mathbf{d} = (d_0, d_1, \ldots, d_m)$ be a $(m+1)$-vector, where m is the order of differencing weights minimizing the variance of linear estimators satisfying the condition

$$\sum_{j=0}^{m} d_j = 0 \qquad \sum_{j=0}^{m} d_j^2 = 1 \tag{2.2}$$

Define the $(n-m) \times m$ differencing matrix $\mathbf{D}$ whose elements satisfy (2.2) as

$$\mathbf{D} = \begin{bmatrix} d_0 & d_1 & \cdots & d_m & 0 & 0 & \ldots & 0 \\ 0 & d_0 & d_1 & \cdots & d_m & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & d_0 & d_1 & \cdots & d_m \end{bmatrix} \tag{2.3}$$

The optimal values for elements of this matrix, is given for example in Yatchew (2003).

Applying the difference matrix to model (1.1) permits direct estimation of the parametric effect. In particular, take

$$\mathbf{Dy} = \mathbf{DX}\boldsymbol{\beta} + \mathbf{Dg}(\mathbf{t}) + \mathbf{D}\boldsymbol{\epsilon}. \tag{2.4}$$

Since the data have been ordered so that the $\mathbf{X}$'s are close, the application of the differencing matrix $\mathbf{D}$ in model (1.1) removes the nonparametric effect in large samples (Yatchew , 2000).

If $g(\cdot)$ is an unknown function that is the inferential object has a bounded first derivative, then $\mathbf{Dg}(\mathbf{t})$ is close to zero, so that applying the differencing matrix ignores the presence of $\mathbf{Dg}(\mathbf{t})$. Thus, we may rewrite (1.1) as

$$\mathbf{Dy} = \mathbf{DX}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\epsilon},$$

or

$$\mathbf{y}_{\mathrm{D}} = \mathbf{X}_{\mathrm{D}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{D}} \tag{2.5}$$

where $\mathbf{y}_{\mathrm{D}} = \mathbf{Dy}$, $\mathbf{X}_{\mathrm{D}} = \mathbf{DX}$ and $\boldsymbol{\epsilon}_{\mathrm{D}} = \mathbf{D}\boldsymbol{\epsilon}$. So that $\boldsymbol{\epsilon}_{\mathrm{D}}$ is an $(n-m)$-vector of disturbances distributed with $\mathrm{E}[\boldsymbol{\epsilon}_{\mathrm{D}}] = \mathbf{0}$ and $\mathrm{E}[\boldsymbol{\epsilon}_{\mathrm{D}}\boldsymbol{\epsilon}_{\mathrm{D}}^{\top}] = \sigma^2\mathbf{D}\mathbf{D}^{\top} \neq \mathbf{I}_{n-m}$.

For arbitrary differencing coefficients satisfying (2.2), Yatchew (1997) defines a simple differencing (D) estimator of $\boldsymbol{\beta}$ in model (1.1) as

$$\hat{\boldsymbol{\beta}}^{\mathrm{D}} = \left(\mathbf{X}_{\mathrm{D}}^{\top}\mathbf{X}_{\mathrm{D}}\right)^{\top}\mathbf{X}_{\mathrm{D}}^{\top}\mathbf{y}_{\mathrm{D}} \tag{2.6}$$

Thus, differencing allows one to perform inferences on $\boldsymbol{\beta}$ as if there were no nonparametric component $\mathbf{g}(\cdot)$ in the model (1.1) (Yatchew , 2003). Once $\boldsymbol{\beta}$ is estimated, a variety of nonparametric techniques could be applied to estimate $\mathbf{g}(\cdot)$ as if $\boldsymbol{\beta}$ were known.

Now, suppose that we are provided with some prior information about the whole or subset of covariates. This prior information can be utilized to improve the overall estimation of the regression coefficients using shrinkage estimation (Ahmad and Raheem , 2012)

Many notable studies are incorporating prior information, in the form of restrictions, to improve estimation in the sense that the restricted and shrinkage estimators have lesser risk and prediction error values. Saleh (2006) gives extensive overviews on a preliminary test (PT), and shrinkage estimators using the ordinary least square (OLS), ridge and maximum likelihood (ML) estimators as starting points. Hossain and Ahmed (2014) start by maximum partial likelihood estimator and propose shrinkage and positive shrinkage estimators, while Roozbeh (2015, 2016) develops shrinkage estimators in ridge regression.

However, in this study, we have different concerns. As a prelude, Tibshirani (1996) proposed a new method for variable selection that produces an accurate, stable, and parsimonious model, called least absolute shrinkage and selection operator (LASSO). We define differenced-based LASSO (D-LASSO) estimator, obtained by

$$\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \left\|\mathbf{y}_{\mathrm{D}} - \mathbf{X}_{\mathrm{D}}\boldsymbol{\beta}\right\|_2^2 + \lambda_n \left\|\boldsymbol{\beta}\right\|_1 \right\}, \qquad \lambda_n \geq 0, \tag{2.7}$$

where for an arbitrary vector $\mathbf{v} = (v_1, \ldots, v_k)^{\top}$, $\left\|\mathbf{v}\right\|_p = \left(\sum_{i=1}^{k} v_j^p\right)^{\frac{1}{p}}$, and λ_n is the tuning parameter, controlling the level of sparsity in $\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}}$.

Now, the questions are as follows:

(1) How can we build the theory if we start with the D-LASSO instead of using a differencing estimator of $\boldsymbol{\beta}$?

(2) What will the form of shrinkage estimators be under restriction, when D-LASSO is used as the starting point?

In this paper, we cover the above issues. In the following section, the restricted D-LASSO estimator is defined for inference under restriction, and the concept of double shrinking is introduced.

3. Double Shrinking Notion

The differenced-based LASSO estimator has been denoted as $\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}}$ and termed as unrestricted D-LASSO estimator. Now, suppose that some non-sample information (a priori restriction on the parameters) about the whole covariates is available. A set of q linear restrictions on the vector $\boldsymbol{\beta}$ can be written as $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$. Or, we can suppose that our model is subjected to lie in the linear sub-space restriction,

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{h}, \tag{3.8}$$

where $\mathbf{H}$ is a $q \times p$ $(q \leq p)$ matrix of known elements, and $\mathbf{h}$ is a q vector of known components. The rank of $\mathbf{H}$ is q, which implies that the restrictions are linearly independent.

The restriction (3.8) may be (i) a fact known from theoretical or experimental considerations, (ii) a hypothesis that may have to be tested or (iii) an artificially imposed condition to reduce or eliminate redundancy in the description of model (Sengupta and Jammalamadaka , 2003).

Our proposal is to consider the following estimator as the restricted differenced-based LASSO (RD-LASSO) estimator ,

$$\hat{\boldsymbol{\beta}}^{\mathrm{RD-LASSO}} = \hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} - \boldsymbol{\Sigma}_{\mathrm{D}}^{-1}\mathbf{H}^{\top}(\mathbf{H}\boldsymbol{\Sigma}_{\mathrm{D}}^{-1}\mathbf{H}^{\top})^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} - \mathbf{h}), \tag{3.9}$$

where $\boldsymbol{\Sigma}_{\mathrm{D}} = \mathbf{X}_{\mathrm{D}}^{\top}\mathbf{X}_{\mathrm{D}}$. The above closed form RD-LASSO estimator cannot be achieved via routine optimization techniques. Indeed, we proposed it by the analogy of differenced-based estimator of $\boldsymbol{\beta}$ subject to the restriction $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ (Roozbeh et al. , 2010).

When (3.8) is satisfied, the RD-LASSO estimator has a smaller asymptotic risk than the D-LASSO estimator. However, for $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{h}$, the RD-LASSO estimator may be biased and inconsistent in many cases. Now, how can we decide on D-LASSO (as an unrestricted) or RD-LASSO (as a restricted) estimator, since we do not know whether the restriction holds? To solve this, it is plausible to follow Fisher's recipe and define the preliminary test differenced-based LASSO (PTD-LASSO) estimator by taking D-LASSO or RD-LASSO estimator according to the acceptance or rejection of the null hypothesis, $\mathcal{H}_o : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$.

This estimator will have the form

$$\hat{\boldsymbol{\beta}}^{\text{PTD}-\text{LASSO}} = \hat{\boldsymbol{\beta}}^{\text{D}-\text{LASSO}} - (\hat{\boldsymbol{\beta}}^{\text{D}-\text{LASSO}} - \hat{\boldsymbol{\beta}}^{\text{RD}-\text{LASSO}})I(\mathcal{L}_n \leq \mathcal{L}_{n,\alpha}), \quad (3.10)$$

where $\mathcal{L}_{n,\alpha}$ is the upper α-level critical value of the exact distribution of the test statistic $\mathcal{L}_n$ under $\mathcal{H}_o$. In order to define test statistics, we need the following theorem.

Theorem 3.1. *(Yatchew , 1997) Under the assumed regularity conditions as $n \to \infty$,*

$$(n-m)^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}^{\text{D}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, (1+2\delta)\sigma^2\boldsymbol{\Sigma}^{-1}) \quad (3.11)$$

where $\delta = \sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right)^2$, and

$$\begin{aligned} s_{\text{D}}^2 &= \frac{1}{n-m}(\mathbf{y}_{\text{D}} - \mathbf{X}_{\text{D}}\hat{\boldsymbol{\beta}}^{\text{D}})^\top(\mathbf{y}_{\text{D}} - \mathbf{X}_{\text{D}}\hat{\boldsymbol{\beta}}^{\text{D}})^\top \xrightarrow{\mathcal{P}} \sigma^2 \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n-m}(\mathbf{X}_{\text{D}}^\top\mathbf{X}_{\text{D}}) \xrightarrow{\mathcal{P}} \boldsymbol{\Sigma} \end{aligned}$$

where $\xrightarrow{\mathcal{D}}$ and $\xrightarrow{\mathcal{P}}$ denote convergence in distribution and probability, respectively.

Using Theorem 3.1 and following Saleh (2006), we can define the test statistics as

$$\mathcal{L}_n = \frac{(\mathbf{H}\hat{\boldsymbol{\beta}}^{\text{D}} - \mathbf{h})^\top(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^\top)^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}^{\text{D}} - \mathbf{h})}{(1+2\delta)s_{\text{D}}^2}. \quad (3.12)$$

Indeed $\mathcal{L}_n$ has the non-central chi-square distribution with q degrees of freedom (d.f.)and the non-central parameter Δ^2, given by

$$\Delta^2 = \frac{(\mathbf{H}\boldsymbol{\beta} - \mathbf{h})^\top(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^\top)^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h})}{(1+2\delta)\sigma^2}. \quad (3.13)$$

Under the null hypothesis, $\mathcal{H}_o$, the $\mathcal{L}_n$ is distributed as χ_q^2, the central chi-square with q d.f.

Here, we use the test as in (3.12), since we can build the asymptotic theory. We think if one uses a test based on the D-LASSO, it makes the analytical computation much easier; here, our aim is only the application aspect.

The PTD-LASSO estimator is highly dependent on the level of significance α and has discrete nature, which is simplified to one of the extremes D-LASSO or RD-LASSO estimator according to the output of the test. In this respect, making use of a continuous and α-free estimator may make more sense. Now, we propose a double shrinking idea which reflects a relevant estimator. It is well-known that the LASSO estimator shrinks coefficients toward the origin, however, when the

restriction $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ is subjected to the model, it is of major importance that the estimator is shrunk toward the restricted one as well. Hence, there must be shrinking toward two directions or double shrinking concept, say. Consequently, we combine the idea of James and Stein (1961) shrinkage and DLASSO to propose the following Stein-type shrinkage differenced-based LASSO (SD-LASSO) estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{SD-LASSO}} = \hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} - (p-2)(\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} - \hat{\boldsymbol{\beta}}^{\mathrm{RD-LASSO}})\mathcal{L}_n^{-1}, \tag{3.14}$$

where $(p-2)$ is the shrinkage constant.

The SD-LASSO may go past the RD-LASSO. So, we define the positive-rule Stein-type shrinkage differenced-based LASSO (PRD-LASSO) estimator (PRSSLE) given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\mathrm{PRD-LASSO}} &= \hat{\boldsymbol{\beta}}^{\mathrm{RD-LASSO}} + (1-(p-2)\mathcal{L}^{-1})I(\mathcal{L}_n > (p-2)),\\ &\quad \times(\hat{\boldsymbol{\beta}}_n^{\mathrm{D-LASSO}} - \hat{\boldsymbol{\beta}}^{\mathrm{RD-LASSO}}),\\ &= \hat{\boldsymbol{\beta}}^{\mathrm{SD-LASSO}} - (1-(p-2)\mathcal{L}_n^{-1})I(\mathcal{L}_n \leq (p-2)),\\ &\quad \times(\hat{\boldsymbol{\beta}}^{\mathrm{D-LASSO}} - \hat{\boldsymbol{\beta}}^{\mathrm{RD-LASSO}}).\end{aligned} \tag{3.15}$$

We note that, as the test based on $\mathcal{L}_n$ is consistent against fixed $\boldsymbol{\beta}$ such that $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{h}$, the PTD-LASSO, SD-LASSO and PRD-LASSO are asymptotically equivalent to the D-LASSO for fixed alternative. Hence, we will investigate the asymptotic risks under local alternatives and compare the performance of the estimators.

4. Application

Mroz (1987) used a sample of 1975 Panel Study on Income Dynamics (PSID) labor supply data to systematically study several theoretical and statistical assumptions used in many empirical models of female labor supply. PSID data is freely available from https://ideas.repec.org/p/boc/bocins/mroz.html.

The female labor, were collected from married white women between ages 30 and 60 in 1975, supply data consists of 753 observations on 19 variables: INLF (= 1 if the labor force in 1975), **hours** (Hours worked in 1975), K5 (kids less than 6 years), K618 (kids 6-18 years), AGE (Woman's age in years), EDUC (Years of schooling), WAGE (Estimated hourly wage from earnings), REPWAGE (Reported wage ate interview in 1976), HUSHRS (Hours worked by husband in 1975), HUSAGE (Husband's age), HUSEDUC (HUsband's years of schooling), HUSWAGE (Husband's hourly wage in 1975), FAMINC (Family income in 1975), MTR (Federal marginal tax rate facing woman), MOTHEDUC (Mother's years of schooling), FATHEDUC (Father's years of schooling), UNEM (Unemployment rate in country of residence),

CITY (= 1 if living in SMSA), EXPER (Actual labor market experience), NWIFEINC ((FAMINC - WAGE × HOURS)/1000).

Similar to Mroz (1987), we consider HOURS-woman's hours of work in 1975, as our response variable. Following Raheem et al. (2012), because of the nature of our response variable, we only used the portion of the data when the women were in labor force. Thus, we had 428 cases (rows) in our working data. Our candidate full model consists of AGE, NWIFEINC, K5, K618, WC (= 1 if EDUC$\geq$ 12), (textschc) (= 1 if HUSEDUC$\geq$ 12), UNEM, EXPER, andMTR. With the inclusion of a nonparametric part, they suggested the following model:

$$\text{HOURS} = \text{WC} + \text{g(NWIFEINC)} + \text{MTR} + \text{EXPER} + \text{UNEM} + \text{K5} + \text{AGE} + \text{K618} + \text{HC}$$

Here NWIFEINC is the nonparametric component.

Since one of the biggest problems in estimation is to determine $\mathbf{H}$ and $\mathbf{h}$, we suppose that $\mathbf{H} = \mathbf{I}_7$. This choice is just for simplicity and also avoiding errors obtained by incorrect selection of parameters. In order to show the impact of correctness or incorrectness of hypothesis, let $\mathbf{h} = (0, 0, 0, 0, 0, 0, 0)^\top$. The null hypothesis changes into $\mathcal{H}_o : \boldsymbol{\beta} = \mathbf{0}$ and thus, all variables are insignificant.

Hall et al. (1990) suggested optimal differencing values by numerical analysis. Here, we consider $m = 6$. Thus, using their suggestion, let

$$\mathbf{d} = (0.9200, -0.2238, -0.1925, -0.1635, -0.1369, -0.1926, -0.0906)$$

and construct $\mathbf{D}$ matrix.

In the following, K-fold cross validation was used to obtain an estimate of the prediction errors of the model. In a K-fold cross validation, the dataset is randomly divided into K subsets of roughly equal size. One subset is left aside, $\{(\mathbf{X}_{\mathrm{D}}^{\mathrm{test}}, \mathbf{y}_{\mathrm{D}}^{\mathrm{test}})\}$, termed as test set, while the remaining $K-1$ subsets, called the training set, are used to fit model. The result estimator is called $\hat{\boldsymbol{\beta}}^{\mathrm{train}}$. The fitted model is then used to predict the responses of the test data set. Finally, prediction errors are obtained by taking the squared deviation of the observed and predicted values in the test set, i.e.

$$\mathrm{PE}^k = \left\| \mathbf{y}_{\mathrm{D}k}^{\mathrm{test}} - \hat{\mathbf{y}}_{\mathrm{D}k}^{\mathrm{test}} \right\|^2; \qquad k = 1, \ldots, K,$$

where $\hat{\mathbf{y}}_{\mathrm{D}k}^{\mathrm{test}} = \mathbf{X}_{\mathrm{D}k}^{\mathrm{test}} \hat{\boldsymbol{\beta}}_{\mathrm{D}k}^{\mathrm{train}}$. The process is repeated for all K subsets and the prediction errors are combined. To account for the random variation of the cross validation, the process is reiterated N times and is estimated the average prediction

error (APE) that is given by

$$\text{APE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K}\text{PE}_i^k\right),$$

where PE_i^k is the prediction error of considering kth test set in ith iteration.

The performance of the arbitrary estimator, $\hat{\boldsymbol{\beta}}^{\star}$, with respect to the full model estimator, $\hat{\boldsymbol{\beta}}^{\text{D-LASSO}}$, is obtained by the below Efficiency (Eff) formula, that is defined as

$$\text{Eff}(\hat{\boldsymbol{\beta}}^{\star};\hat{\boldsymbol{\beta}}^{\text{D-LASSO}}) = \frac{\text{APE}(\hat{\boldsymbol{\beta}}^{\text{D-LASSO}})}{\text{APE}(\hat{\boldsymbol{\beta}}^{\star})}.$$

If the value of Eff is greater than 1, then $\hat{\boldsymbol{\beta}}^{\star}$ performs better than $\hat{\boldsymbol{\beta}}^{\text{D-LASSO}}$.

Table 1 shows the estimator and their Eff in case of $N = 5000$.

Table 1: The values of proposed estimators and their Efficiency

Variables	D	D-LASSO	RD-LASSO	PTD-LASSO	SD-LASS	PRD-LASSO
EXPER	30.26	23.57	0.00	23.57	18.41	18.41
MTR	−416.00	−1213.50	0.00	−1213.50	−947.88	−947.88
WC	−68.84	−45.56	0.00	−45.56	−35.59	−35.59
AGE	−8.36	−5.92	0.00	−5.92	−4.62	−4.62
UNEM	−9.85	−9.19	0.00	−9.19	−7.18	−7.18
K5	−128.1	−132.11	0.00	−132.11	−103.19	−103.19
K618	−43.06	−25.56	0.00	−25.56	−19.96	−19.96
HC	22.22	0.00	0.00	0.00	0.00	0.00
Eff	–	1.00	0.95	1.08	1.12	1.14

Although, the RD-LASSO estimator does not have less prediction error in comparison with the D-LASSO estimator, the prediction error of the difference-based preliminary test, Stein and positive rule Stein estimators are less than the LASSO estimator. As we see in Table 1, these estimators shrink full model coefficients towards a restricted estimator and improve the accuracy of the prediction error.

The differencing estimates used 6 - order ($m = 6$) differencing, we have used the kernel regression procedure with bandwidth $h_n = 1.2$ for estimation HOURS. In order to estimate non-parametric effect, first we estimated the parameter by differencing method, improved that and then , kernel approach applied to fit $\mathbf{Z}_i(\hat{\boldsymbol{\beta}}^{*}) = \text{HOURS}_i - \mathbf{x}_{\text{D}i}\hat{\boldsymbol{\beta}}^{*}$ on NWIFEINC where $\mathbf{x}_i = (\text{EXPER}, \text{MTR}, \text{WC}, \text{AGE}, \text{UNEM}, \text{K5}, \text{K618}, \text{HC})$ and $\hat{\boldsymbol{\beta}}^{*}$ is one of the differencing-based, differenced-based LASSO, and positive rule Stein-type differenced-based LASSO estimators (Figure 1).

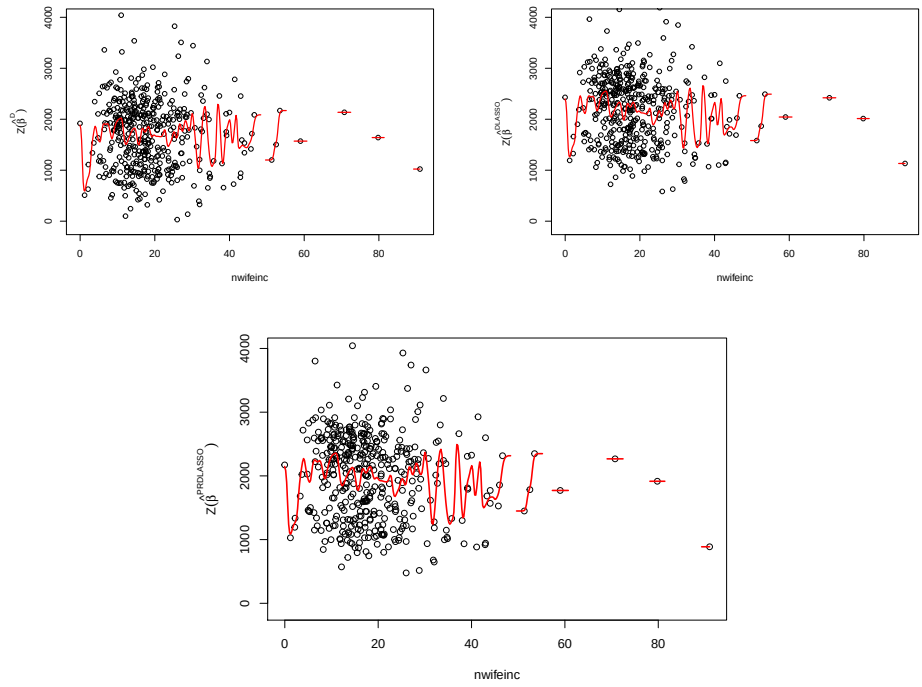

Figure 1: Graph of the estimation of nonparametric function

5. Conclusions

In this paper, we proposed improved differenced-based LASSO estimators for partially linear models by imposing a sub-space restriction to the linear part of these models. Particularly, we introduced the preliminary test, Stain and positive-rule Stein difference-based LASSO estimators. Indeed the test statistic for $\mathcal{H}_o : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ plays a determining role. The nonparametric component is estimated by using the Speckman approach based on the residual sum of squares method. As an application, a real dataset analyzed, where a 10-folded cross-validation average of the prediction errors evaluated for the differenced-based LASSO and its other four variants. The new estimators dominated the LASSO in average prediction error sense.

Acknowledgment

The authors would like to sincerely thank the Associate Editor and the reviewers for constructive and important comments that substantially improved the presentation of the paper.

References

Ahmed, S.E. and Raheem, S.M.E. (2012). Shrinkage and absolute penalty estimation in linear regression models, *Wires: Computational Statistics*, **4**(6), 541-553.

Aydın, D. (2014) Estimation of partially linear model with smoothing spline based on different selection methods: A comparative study. *Pakistan Journal of Statistics*, **30**, 35-56.

Bunea, F., (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics*, **32**, 898-927.

Engle R.F., Granger, C. W. J., Rice, C. A., Weiss, A. (1986) Semi-parametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, 310-320.

Hall, P., Kay, J. W. and Titterington, D. M., (1990). On estimation of noise variance in two-dimensional signal processing. *Advances in Applied Probability*, **23**, 476-495.

Härdle, W., Liang, H., Gao, J., (2000). *Partially Linear Models*, Physica-Verlag, Heidelber

Hossain, S. and Ahmed, S.E. (2014). Penalized and Shrinkage Estimation in the Cox Proportional Hazards Model. *Communications in Statistics-Theory and Methods*, **43**(5), 1026-1040.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. Berkeley, Calif.: University of California Press, 361-379.

Liang, H. (2006). Estimation partially linear models and numerical comparison. *Computational Statistics and Data Analysis*, **50**, 675–687.

Mroz, T. A., (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, **55**(4), 765–799.

Raheem, S. M. E., Ahmed, S. E. and Doksum, K. A. (2012). Absolute penalty and shrinkage estimation in partially linear models, *Computational Statistics and Data Analysis*, **56**, 874-891.

Roozbeh, M. (2015). Shrinkage ridge estimators in semiparametric regression models. *Journal of Multivariate Analysis*, **136**, 56 - 74.

Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models. *Journal of Multivariate Analysis*, **147**, 127–144.

Roozbeh, M., Arashi, M. and Niroumand, H. A. (2010). Semiparametric Ridge Regression approach in Partially Linear Models, *Communications in Statistics - Simulation and Computation*, **39**, 449-460.

Saleh, A.K.M.E. (2006). *Theory of preliminary test and stein-type estimation with applications*, John Wiley & Sons, New York.

Saleh, A. K. Md., Arashi, M., Norouzirad, M. and Kibria, B. M. G. (2017), On shrinkage and selection, ANOVA model, *Journal of Statistical Research*, **51**(2), 165 -191.

Sengupta, D. and Jammalamadaka, S.R. (2003). *Linear models: An integrated approach*, World Scientific Publishing Company.

Sun, J., Kopciuk, K.A., Lu, X., (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Computational Statistics and Data Analysis*, **53**, 176–188.

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

Yatchew, A., (1997). An elementary estimator of the partial linear model. *Economics Letters*, **57**, 135-143.

Yatchew, A. (2000). Scale economies in electricity distribution: A semiparametric analysis. *Journal of Applied Economics*, **57**, 187–210.

Yatchew, A., (2003). *Semiparametric regression for the applied econometrician*, Cambridge University press, Cambridge.