Journal of the Korean Data & Information Science Society
2016, **27**(4), 1059–1066

http://dx.doi.org/10.7465/jkdi.2016.27.4.1059
한국데이터정보과학회지

Robust varying coefficient model using L1 regularization†

Changha Hwang[1] · Jongsik Bae[2] · Jooyong Shim[3]

[1]Department of Applied Statistics, Dankook University
[2]Department of Mathematics, Sungkyunkwan University
[3]Institute of Statistical Information, Department of Statistics, Inje University

Received 24 May 2016, revised 11 June 2016, accepted 23 June 2016

Abstract

In this paper we propose a robust version of varying coefficient models, which is based on the regularized regression with L1 regularization. We use the iteratively reweighted least squares procedure to solve L1 regularized objective function of varying coefficient model in locally weighted regression form. It provides the efficient computation of coefficient function estimates and the variable selection for given value of smoothing variable. We present the generalized cross validation function and Akaike information type criterion for the model selection. Applications of the proposed model are illustrated through the artificial examples and the real example of predicting the effect of the input variables and the smoothing variable on the output.

Keywords: Akaike's information criterion, generalized cross validation function, iteratively reweighted least squares procedure, L1-regularization, locally weighted regression, smoothing variable, variable selection, varying coefficient model.

1. Introduction

The varying coefficient (regression) model (VCM) is flexible and powerful for modeling the dynamic changes of regression coefficients, which was firstly introduced by Hastie and Tibshirani (1993). VCM is known to be a useful extension of the linear regression model, in which the regression coefficients are not fixed as constants but are allowed to change with the values of certain variables (smoothing variables, environmental variables). Since VCM inherits the simplicity and the easy interpretation of the linear regression models, it therefore gains the popularity in statistical literature in recent years. The introductions and current research areas of VCM are found in Hastie and Tibshirani (1993), Hoover *et al.* (1998), Fan and Zhang (2008) and Shim and Hwang (2015). The problems of estimating coefficient functions and analyzing them appropriately have been studied in many areas of

† This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2015R1D1A1A01056582). This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046715).

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 16890, Korea.
[2] Professor, Department of Mathematics, Sungkyunkwan University, Suwon 16419, Korea.
[3] Corresponding author: Adjunct professor, Institute of Statistical Information, Department of Statistics, Inje University, Kimhae 50834, Korea. E-mail: ds1631@hanmail.net

applied Statistics. Most of this attention of estimating coefficient functions has been focused on using kernel smoothing techniques. Fan and Zhang (2008) gave good reviews of VCM and discussed three approaches in estimating the coefficient functions: kernel smoothing, polynomial splines and smoothing splines. Wang and Xia (2009) proposed to combine the local polynomial smoothing (Fan and Gijbels, 1996) and the LASSO (least absolute shrinkage and selection operator; Tibshirani, 1996) for efficient estimation of VCM. There are some possibilities in constructing VCM. One is to set all regression coefficients be functions of a single input variable or multiple input variables. Another is to set each regression coefficient be function of different variables. There are various extensions of VCM (Park *et al.*, 2015).

We denote the smoothing variable by u_i and an input vector by $\boldsymbol{x}_i \in R^{d_x}$ and an output corresponding to $\boldsymbol{x}_i$ and u_i by y_i, where $i = 1, 2, \cdots, n$. Let $f(u_i, \boldsymbol{x}_i)$ be the regression function of the output given u_i and $\boldsymbol{x}_i$, which can be expressed as the form of VCM as follows:

$$f(u_i, \boldsymbol{x}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}(u_i), \; i = 1, 2, \cdots, n, \tag{1.1}$$

where $\boldsymbol{X}_i' = (1, \boldsymbol{x}_i')$ and $\boldsymbol{\beta}(u_i) = (\beta_0(u_i), \beta_1(u_i), \cdots, \beta_{d_x}(u_i))'$ is a $(d_x + 1) \times 1$ coefficient function (smooth function) vector given u_i. Coefficient functions are usually estimated by the locally weighted regression (Cleveland and Susan, 1988). Generally, all the input variables may not much affect the outputs so that some $\beta(u_i)$'s may be 0's in true regression model. Lots of variable selection techniques such as the best-subset selection, stepwise selection, and Bootstrap procedures (Sauerbrei and Schumacher, 1992) for linear regression models have been proposed. Recently the LASSO has been proposed. By shrinking some regression coefficients to 0, this method provides the selection of important variables and the estimation of regression coefficients simultaneously. Huang *et al.* (2005) proposed the regularization and the variable selection approach using the LASSO (Tibshirani, 1996). Hu and Rao (2010) proposed a weighted least squares method with sparse regularization to fit censored regression models with high-dimensional input variables.

We consider the regularized regression with L1 norm, which is known to have the sparsity on estimation of regression coefficients (Williams, 1995). We use the iteratively reweighted least squares (IRWLS) procedure to solve the regularized log -likelihood function with L1 norm of regression. It provides the efficient computation including variable selection.

The remainder of the paper is organized as follows. In Section 2 we briefly review VCM. In Section 3 we propose IRWLS procedure to regularized estimation for VCM with L1 regularization. In Section 4 we illustrate the applications of the proposed model to artificial data sets and a real data set. Section 5 gives the concluding remarks.

2. Varying coefficient model

For given data set $(y_i, u_i, \boldsymbol{x}_i)_{i=1}^n$, where $y_i \in R$ is an output, $u_i \in R$ is a smoothing variable and $\boldsymbol{x}_i \in R^{d_x}$ is an input vector, the output y_i is assumed to be related to u_i and $\boldsymbol{x}_i$ in the varying coefficient model as follows:

$$y_i = f(u_i, \boldsymbol{x}_i) + e_i = \boldsymbol{X}_i'\boldsymbol{\beta}(u_i) + e_i,$$

where $\boldsymbol{X}_i' = (1, \boldsymbol{x}_i')$ and e_i's are independent error terms with mean 0 and a bounded variance.

For a given value of the smoothing variable u_t, the coefficient functions in VCM are usually estimated locally by the formulating the weighted least squares problem as follows:

$$\min \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)\left(y_i - \sum_{k=0}^{d_x} X_{ik}\beta_k(u_i)\right)^2, \tag{2.1}$$

where $W_i(u_t)$ is a kernel function constructed by u_i and u_t. One of possible kernel functions is the radial basis kernel function such that $W_i(u_t) = \exp(-\frac{1}{\sigma^2}(u_i - u_t)^2)$ with a bandwidth parameter $\sigma > 0$.

Solving the weighted least squares problem leads to the estimate of $\beta_k(u_t)$ for $k = 0, \cdots, d_x$, which can be expressed as vector-matrix notation as follows:

$$\widehat{\boldsymbol{\beta}}(u_t) = (\boldsymbol{X}'W\boldsymbol{X})^{-1}\boldsymbol{X}'W\boldsymbol{y}, \tag{2.2}$$

where $\boldsymbol{X}$ is a $n \times (d_x + 1)$ input matrix and W is a diagonal matrix of $W_i(u_t)$'s.

3. L1 regularized varying coefficient model

From (2.1) we assume that $e_i(u_t) = W_i(y_i - f(u_t, \boldsymbol{x}_i))$ follows a probability distribution such as $p(e) \propto \exp(-0.5e^2)$. The negative log-likelihood function can be expressed as,

$$L_0(f|u_t, \boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)(y_i - f(u_t, \boldsymbol{x}_i))^2. \tag{3.1}$$

The regression function is estimated by a linear regression model such that $f(u_t, \boldsymbol{x}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}(u_t)$. Then the maximum likelihood estimates of $\boldsymbol{\beta}(u_t)$ are obtained by minimizing the negative log-likelihood function,

$$L_0(\boldsymbol{\beta}|\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2. \tag{3.2}$$

Generally, the maximum likelihood estimate of $\boldsymbol{\beta}(u_t)$ leads severe overfitting, so we are encouraged to use a prior over $\boldsymbol{\beta}(u_t)$. Then the regularized maximum likelihood estimate (the maximum a posteriori estimate) of $\boldsymbol{\beta}(u_t)$ can be obtained by minimizing the objective function as follows:

$$L(\boldsymbol{\beta}|\boldsymbol{x}) = L_0(\boldsymbol{\beta}|\boldsymbol{x}) + \log p(\boldsymbol{\beta}) \tag{3.3}$$

where $p(\boldsymbol{\beta})$ is some prior over $\boldsymbol{\beta}$.

To attain the robustness to the estimation of regression function, we use a Laplacian prior (Williams, 1995) to have the sparsity of estimates of $\boldsymbol{\beta}(u_t)$,

$$p(\boldsymbol{\beta}) \propto \exp(-\lambda||\boldsymbol{\beta}||_1),$$

where $||\boldsymbol{\beta}||_1 = \sum_{k=0}^{d_x} |\beta_k|$ denotes L1 norm and λ is a positive regularization parameter.

The objective function can be reexpressed as follows:

$$L(\boldsymbol{\beta}|\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||_1. \tag{3.4}$$

Here λ controls the tradeoff between the goodness-of-fit on $||\boldsymbol{\beta}||_1$ and the data. The objective function $L(\boldsymbol{\beta}|\boldsymbol{x})$ in (3.4) is not differentiable with respect to $\boldsymbol{\beta}$ at 0, we modify $L(\boldsymbol{\beta}|\boldsymbol{x})$ for easy estimation by using IRWLS procedure.

We consider the objective function given $(\boldsymbol{\beta}^*, \boldsymbol{x})$ as follows:

$$L(\boldsymbol{\beta}|\boldsymbol{\beta}^*, \boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{k=0}^{d_x}\left(\frac{\beta_k^2}{|\beta_k^*|} + |\beta_k^*|\right), \tag{3.5}$$

then $L(\boldsymbol{\beta}|\boldsymbol{\beta}^*, \boldsymbol{x}) \geq L(\boldsymbol{\beta}|\boldsymbol{x})$ with equality if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ (Krishnapuram *et al.*, 2005) and $L(\boldsymbol{\beta}|\boldsymbol{\beta}^*, \boldsymbol{x})$ is differentiable with respect to $\boldsymbol{\beta}$. At ℓth iteration of IRWLS procedure, we have

$$L(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(\ell)}, \boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{n} W_i(u_t)(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{k=0}^{d_x}\left(\frac{\beta_k^2}{|\widehat{\beta}_k^{(\ell)}|} + |\widehat{\beta}_k^{(\ell)}|\right). \tag{3.6}$$

Then $\widehat{\boldsymbol{\beta}}^{(\ell+1)}$ is obtained by minimizing $L(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(\ell)}, \boldsymbol{x})$ with respect to $\boldsymbol{\beta}$ as follows:

$$\widehat{\boldsymbol{\beta}}^{(\ell+1)} = (\boldsymbol{X}'W(u_t)\boldsymbol{X} + \lambda V(\widehat{\boldsymbol{\beta}}^{(\ell)}))^{-1}\boldsymbol{X}'W(u_t)\boldsymbol{y}, \tag{3.7}$$

where $W(u_t)$ and $V(\widehat{\boldsymbol{\beta}}^{(\ell)})$ are diagonal matrices with $W_i(u_t)$'s and $(1/|\widehat{\beta}_k^{(\ell)}|)$'s, respectively.

During iterations, we find that some $hatbeta_k$'s tend to be 0 keeping the value of objective function $L(\boldsymbol{\beta}|\boldsymbol{x})$ decreasing, which motivates that we can find sparse estimates of $\boldsymbol{\beta}(u_t)$ which provides decreasing value of the objective function $L(\boldsymbol{\beta}|\boldsymbol{x})$ simultaneously.

Algorithm of L1 regularized locally weighted regression using IRWLS Procedure is given as follows:

i) Set $v = (0 : d_x)'$ and $\widehat{\boldsymbol{\beta}}(v)^{(0)} = (\boldsymbol{X}'W(u_t)\boldsymbol{X})^{-1}\boldsymbol{X}'W(u_t)\boldsymbol{y}$.
ii) Find $\widehat{\boldsymbol{\beta}}(v)^{(\ell+1)}$ from (3.7).
iii) Set $\beta_k = 0$ when $\widehat{\beta}_k^{(\ell+1)} = 0$ is sufficiently close to zero. Find $v = \{k|\beta_k \neq 0\}$.
iv) Iterate ii) and iii) until convergence.

The estimated regression function given $(u_t, \boldsymbol{x}_j)$ is obtained as follows:

$$\widehat{f}(u_t, \boldsymbol{x}_j) = \boldsymbol{X}_j'\widehat{\boldsymbol{\beta}}(u_t).$$

The performance of L1 regularized regression is affected by the hyperparameters, which are the regularization parameter λ and the bandwidth parameter. To choose the optimal values of hyperparameters, we consider the cross validation (CV) function for the model selection criterion as follows:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} W_i(u_t)\left(y_i - \widehat{f}_\lambda^{(-i)}(u_t, \boldsymbol{x}_i)\right)^2. \tag{3.8}$$

Here $\widehat{f}_{\lambda}^{(-i)}(u_t, \boldsymbol{x}_i)$ is the regression function estimated without the ith observation. Since for each candidate set of hyperparameters, $\widehat{f}_{\lambda}^{(-i)}(u_t, \boldsymbol{x}_i)$ for $i = 1, \cdots, n$, should be computed, choosing the optimal values of hyperparameters using CV function (3.8) is computationally burdensome. We consider GCV function as follows:

$$GCV(\lambda) = \frac{n \sum_{i=1}^{n} W_i(u_t)(y_i - \widehat{f}_{\lambda}(u_t, \boldsymbol{x}_i))^2}{(n - tr(H))^2}, \tag{3.9}$$

where $H = (\boldsymbol{X}(:,v)'W\boldsymbol{X}(:,v)+\lambda V(v,v))^{-1}\boldsymbol{X}(:,v)'W$ is the hat matrix such that $\widehat{f}_{\lambda}(u_t, \boldsymbol{x}) = H\boldsymbol{y}$ with the (i,j)th element $H_{ij} = \partial \widehat{f}(u_t, \boldsymbol{x}_i)/\partial y_j$ and $\boldsymbol{X}(:,v)$ consists of X_{ik} for $i = 1, \cdots, n$ and $k \in v$. Details of derivation of GCV function can be found in Cho *et al*. (2010) and Shim *et al*. (2015).

Akaike (1974) defined Akaike's Information Criterion (AIC) for the model selection criterion as follows:

$$AIC = 2L_0(\boldsymbol{\beta}|\boldsymbol{x}) + 2\mathrm{P}, \tag{3.10}$$

where P is the number of estimable parameters in the given model and $L_0(\boldsymbol{\beta}|\boldsymbol{x})$ is the negative log-likelihood function. We use an AIC-type criterion (Hwang *et al*., 2011) to incorporate the simplicity of the model into the model selection criterion as follows:

$$GCV(\lambda)_{AIC} = \log(GCV(\lambda)) + \mathrm{P}, \tag{3.11}$$

where P is the average number of nonzero coefficient functions.

4. Numerical studies

Through the artificial data sets and the real data set, we illustrate the estimation performance of the proposed model. For each data set, the proposed model and the varying coefficient model (VCM) with the least squares-support vector regression (VCM_LSSVR, Shim and Hwang, 2015) are applied with the optimal values of the hyperparameters chosen from AIC-type criterion (3.11) and GCV function, respectively. For VCM with the locally weighted regression (VCM_LWR) in (2.1) the leave-one-out CV function is used. The radial basis kernel function is utilized for all examples in numerical studies.

Example 4.1 We generate N data sets by similar manner to Wu *et al*. (2015). We compare the performance of estimations of regression functions and nonzero coefficient functions of the proposed model with VCM_LWR and VCM_LSSVR. For each $i = 1, \cdots, n$, $x_{i1}, \cdots, x_{i15}$ and u_i are generated from a uniform distribution, $U(0,1)$, respectively, and y_i 's are generated as follows:

$$y_i = f(u_i, \boldsymbol{x}_i) + e_i = \beta_0(u_i) + \boldsymbol{x}_i'\boldsymbol{\beta}(u_i) + e_i,\ i = 1, \cdots, n,$$

where e_i's are generated from the standard normal distribution, $N(0,1)$, Student t-distribution with 3 degrees of freedom, $t(3)$ and Laplace distribution, $L(0,2)$, respectively. Coefficient functions are set as $\beta_0(u) = 0$, $\beta_1(u) = 2sin(2\pi u)$, $\beta_2(u) = 2exp(2u-1)$, $\beta_3(u) = 6u(1-u)$, $\beta_4(u) = -4u^3$, $\beta_5(u) = \beta_6(u) = \cdots = \beta_{15}(u) = 0$. From each data set we obtain the mean squared error (MSE) to measure the estimation performance.

Tables 4.1, 4.2 and 4.3 are the results from example 4.1, which show averages of MSE's for regression functions and nonzero coefficient functions, respectively. The boldfaced figure in each column signifies the smallest MSE. For $n = 100$, the proposed model shows the best performance. For $n = 200$ with $t(3)$ distribution and $L(0, 2)$ distribution, the proposed model shows the best performance. For $n = 500$ with $L(0, 2)$ distribution, the proposed model shows the best performance. Thus, we can see that the proposed model shows the better performance for the smaller sample sizes and the thicker tailed distributions, which indicates that the proposed model provides the robust estimations of regression functions and nonzero coefficient functions.

Table 4.1 Results of example 4.1: Averages of MSE's of $\Delta f = f(u, \boldsymbol{x}) - \widehat{f}(u, \boldsymbol{x})$ and $\Delta\beta_k = \beta_k(u) - \widehat{\beta}_k(u)$ (standard error in parenthesis) for $n = 100$ and $N = 1000$.

		proposed	VCM_LSSVR	VCM_LWR
$N(0,1)$	Δf	**0.3741** (0.0021)	0.4518 (0.0032)	0.5718 (0.0051)
	$\Delta\beta_1$	**1.0197** (0.0112)	1.2338 (0.0190)	1.4407 (0.0199)
	$\Delta\beta_2$	0.6999 (0.0151)	**0.6055** (0.0172)	0.9484 (0.0420)
	$\Delta\beta_3$	**0.3818** (0.0091)	0.6181 (0.0176)	0.7362 (0.0189)
	$\Delta\beta_4$	**0.6427** (0.0135)	0.6624 (0.0187)	0.8026 (0.0197)
$t(3)$	Δf	**0.7993** (0.0811)	1.4397 (0.2930)	1.9578 (0.3064)
	$\Delta\beta_1$	**1.3706** (0.0511)	2.6598 (0.3925)	2.1923 (0.0567)
	$\Delta\beta_2$	**1.1548** (0.0465)	1.5294 (0.0901)	2.4728 (0.1094)
	$\Delta\beta_3$	**0.8319** (0.1716)	2.3507 (0.8951)	1.3993 (0.0182)
	$\Delta\beta_4$	**1.0446** (0.0686)	1.5665 (0.0988)	1.6431 (0.0492)
$L(0,2)$	Δf	**1.6941** (0.0175)	2.7964 (0.0295)	5.1772 (0.0917)
	$\Delta\beta_1$	**2.0350** (0.0503)	3.9624 (0.1209)	3.5339 (0.0989)
	$\Delta\beta_2$	**2.0982** (0.0605)	3.3975 (0.1127)	4.7891 (0.0955)
	$\Delta\beta_3$	**1.5504** (0.0492)	3.3692 (0.1276)	2.8735 (0.0904)
	$\Delta\beta_4$	**1.8407** (0.0517)	3.4107 (0.1114)	3.3649 (0.0841)

Table 4.2 Results of example 4.1: Averages of MSE's of $\Delta f = f(u, \boldsymbol{x}) - \widehat{f}(u, \boldsymbol{x})$ and $\Delta\beta_k = \beta_k(u) - \widehat{\beta}_k(u)$ (standard error in parenthesis) for $n = 200$ and $N = 1000$.

		proposed	VCM_LSSVR	VCM_LWR
$N(0,1)$	Δf	0.3201 (0.0013)	**0.2688** (0.0016)	0.4396 (0.0019)
	$\Delta\beta_1$	0.7814 (0.0066)	**0.6901** (0.0094)	1.0328 (0.0076)
	$\Delta\beta_2$	0.4132 (0.0076)	**0.2675** (0.0068)	0.3172 (0.0098)
	$\Delta\beta_3$	**0.2299** (0.0050)	0.2691 (0.0075)	0.3671 (0.0083)
	$\Delta\beta_4$	0.4025 (0.0072)	**0.2795** (0.0069)	0.3672 (0.0075)
$t(3)$	Δf	**0.5510** (0.0177)	0.7567 (0.0381)	1.1013 (0.1606)
	$\Delta\beta_1$	**0.9675** (0.0185)	1.4049 (0.1168)	1.4310 (0.0623)
	$\Delta\beta_2$	**0.6226** (0.0168)	0.7579 (0.0707)	0.6674 (0.0317)
	$\Delta\beta_3$	**0.4264** (0.0165)	0.8096 (0.1462)	0.7095 (0.0391)
	$\Delta\beta_4$	**0.6226** (0.0211)	0.7258 (0.0372)	0.8129 (0.1029)
$L(0,2)$	Δf	**1.0850** (0.0083)	1.6031 (0.0130)	1.6299 (0.0198)
	$\Delta\beta_1$	**1.3954** (0.0265)	2.1173 (0.0433)	2.1301 (0.0444)
	$\Delta\beta_2$	**1.1709** (0.0324)	1.5311 (0.0471)	1.4638 (0.0482)
	$\Delta\beta_3$	**0.8895** (0.0263)	1.5073 (0.0455)	1.4613 (0.0423)
	$\Delta\beta_4$	**1.0968** (0.0290)	1.4879 (0.0443)	1.4436 (0.0436)

Table 4.3 Results of example 4.1: Averages of MSE's of $\Delta f = f(u, \boldsymbol{x}) - \widehat{f}(u, \boldsymbol{x})$ and $\Delta\beta_k = \beta_k(u) - \widehat{\beta}_k(u)$ (standard error in parenthesis) for $n = 500$ and $N = 1000$.

		proposed	VCM_LSSVR	VCM_LWR
$N(0,1)$	Δf	0.2823 (0.0008)	**0.1219** (0.0007)	0.3327 (0.0019)
	$\Delta\beta_1$	0.6351 (0.0034)	**0.2387** (0.0048)	0.6646 (0.0089)
	$\Delta\beta_2$	0.2770 (0.0037)	**0.0980** (0.0022)	0.2326 (0.0059)
	$\Delta\beta_3$	0.1390 (0.0021)	**0.0922** (0.0022)	0.2637 (0.0052)
	$\Delta\beta_4$	0.2755 (0.0034)	**0.1008** (0.0022)	0.2672 (0.0047)
$t(3)$	Δf	0.3725 (0.0026)	**0.3277** (0.0050)	0.5076 (0.0031)
	$\Delta\beta_1$	0.7061 (0.0059)	**0.6249** (0.0103)	0.9754 (0.0059)
	$\Delta\beta_2$	0.3604 (0.0074)	**0.2508** (0.0086)	0.2587 (0.0067)
	$\Delta\beta_3$	**0.2174** (0.0049)	0.2427 (0.0071)	0.3091 (0.0065)
	$\Delta\beta_4$	0.3478 (0.0067)	**0.2542** (0.0075)	0.3135 (0.0064)
$L(0,2)$	Δf	**0.6232** (0.0038)	0.7741 (0.0054)	0.8440 (0.0046)
	$\Delta\beta_1$	**0.9041** (0.0116)	1.1849 (0.0158)	1.2704 (0.0145)
	$\Delta\beta_2$	0.5763 (0.0135)	0.5829 (0.0158)	**0.5217** (0.0144)
	$\Delta\beta_3$	**0.4222** (0.0104)	0.5965 (0.0160)	0.5828 (0.0143)
	$\Delta\beta_4$	**0.5520** (0.0126)	0.6118 (0.0165)	0.5918 (0.0156)

Example 4.2 We consider a subset of the wage data set studied in Wooldridge (2012), collected on each of 526 working individuals for the year 1976. The variables we use in this example are the wage (in dollars per hour, output), years of education (smoothing variable) and marital status (the person is married or not, input variable). The input variable is a binary (1,0) in nature and serves to indicate qualitative features of the each individual. Taking these variables into account, we consider the varying coefficient model as follows:

$$\log(y_i) = \beta_0(u_i) + \beta_1(u_i)x_i + e_i,\ i = 1, \cdots, 526,$$

where y, u and x denote the wage, years of education and marital status, respectively.

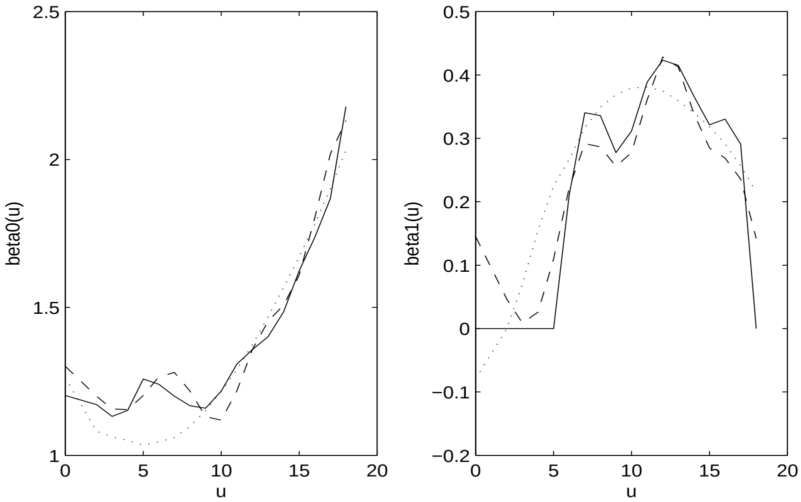

Figure 4.1 Estimated coefficient functions for Wage data set in Example 2. Solid lines: proposed estimates, dashed lines: estimates by VCM_LSSVR, dotted lines: estimates by VCM_LWR.

Figure 4.1 shows the extent to which the coefficient function varies with years of education, which shows that the smoothing variable (u, 'educ') gives the strong effect on the regression coefficient function. Three models show similar pattern on this example but estimates of coefficient function $\beta_1(u)$ by the proposed model are 0's under 6 years of education.

5. Conclusions

In this paper, we dealt with a robust version of varying coefficient model based on the regularized locally weighted regression model with L1-regularization. We use the iteratively reweighted least squares procedure to solve the L1 regularized log- likelihood function. The proposed model provides the efficient computations and the generalized cross validation function for the model selection. We showed that the proposed model derives the satisfying solutions through the examples. The proposed model is found to be simple and reliable in that both estimation of the coefficient functions and variable selection.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716-723.

Cho, D. H., Shim, J. and Seok, K. (2010). Doubly regularized kernel method for heteroscedastic autoregressive data. *Journal of the Korean Data & Information Science Society*, **21**, 155-162.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Associationl*, **83**, 596-610.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*, Chapman & Hall, Boca Raton.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**, 179-195.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, **55**, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Hu, S. and Rao, J. S. (2010). *Sparse penalization with censoring constraints for estimating high dimensional AFT models with applications to microarray data analysis*, Technical Report 07 of Division of Biostatistics, Case Western Reserve University, Ohio.

Huang, J., Ma, S. and Xie, H. (2005). *Regularized estimation in the accelerated failure time model with high dimensional covariates*, Technical Report No. 349, Department of Statistics and Actuarial Science, The University of Iowa, Iowa.

Hwang, C., Kim, M. and Shim, J. (2011). Variable selection in L1 penalized censored Regression. *Journal of the Korean Data & Information Science Society*, **22**, 951-959.

Krishnapuram, B., Carlin, L., Figueiredo, M. A. T. and Hartermink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 957-968.

Park, B. U., Mammen, E., Lee, Y. K. and Lee, E. R. (2015). Varying coefficient regression models: A review and new developments. *International Statistical Review*, **83**, 36-64.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistical Medicine*, **11**, 2093-2099.

Shim, J. and Hwang, C. (2015). Varying coefficient modeling via least squares support vector regression. *Neurocomputing*, **161**, 254-259.

Shim, J., Kim, M. and Seok, K. (2015). SVQR with asymmetric quadratic function. *Journal of the Korean Data & Information Science Society*, **26**, 1537-1545.

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

Wang, H. and Xia, Y. (2009). Shrinkage Estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**, 747-757.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, **7**, 117-143.

Wooldridge, J. M. (2012).*Introductory econometrics: A modern approach*, South-Western Cengage Learning, Mason.

Wu, C., Shi, X., Cui, Y. and Ma, S. (2015). A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, doi:10.1002/sim.6609.