Journal of the Korean Data & Information Science Society
2012, **23**(2), 385–392

한국데이터정보과학회지

Semiparametric kernel logistic regression with longitudinal data[†]

Jooyong Shim[1] · Kyung Ha Seok[2]

[12]Department of Data Science, Inje University

Received 25 January 2012, revised 20 February 2012, accepted 25 February 2012

Abstract

Logistic regression is a well known binary classification method in the field of statistical learning. Mixed-effect regression models are widely used for the analysis of correlated data such as those found in longitudinal studies. We consider kernel extensions with semiparametric fixed effects and parametric random effects for the logistic regression. The estimation is performed through the penalized likelihood method based on kernel trick, and our focus is on the efficient computation and the effective hyperparameter selection. For the selection of optimal hyperparameters, cross-validation techniques are employed. Numerical results are then presented to indicate the performance of the proposed procedure.

Keywords: Generalized cross-validation function, kernel trick, logistic regression, longitudinal data, mixed-effects model, penalized likelihood.

1. Introduction

Logistic regression (Amemiya, 1985; Agresti, 2002) is a popular method for binary classification problems. The output of a logistic regression model can be interpreted as a posterior estimate of the probability that an observation belongs to each of two disjoint classes. The probabilistic nature of the logistic regression model affords many practical advantages, such as the ability to accommodate unequal relative class frequencies in the training set or to apply an appropriate loss matrix in making predictions that minimize the expected risk. As a result, this model has been adopted in a diverse range of applications, including cancer classification and analysis of DNA binding sites. For data that are clustered and/or longitudinal, mixed-effect regression models are becoming increasingly popular (Hedeker and Gibbons, 2006; Wu and Zhang, 2006). Mixed-effects models constitute both fixed and random effects. In clustered data, subjects are clustered within an organization such as a hospital, school, clinic or firm. In longitudinal data where individuals are repeatedly assessed, measurements

[†] This work was supported by Inje research and scholarship foundation in 2010.
[1] Adjunct professor, Department of Data Science, Institute of Statistical Information, Inje University, Obang-Dong, Kimhae 621-749, Korea.
[2] Corresponding author: Professor, Department of Data Science, Institute of Statistical Information, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr

are clustered within individuals. For clustered data the random effects represent cluster effects, while for longitudinal data the random effects represent subject effects. There has been much work done on mixed-effect models for continuous responses, and non-continuous responses data (Winkelmann, 2003; Shim and Seok, 2008).

In this paper we propose a semiparametric kernel logistic regression model (SKLR) for the binary classification of longitudinal data. The proposed model is derived by employing the penalized likelihood method based on kernel tricks in Vapnik (1995), Smola and Schölkopf (1998). For the easy selection of optimal hyperparameters to achieve high generalization performance, we propose a generalized cross validation (GCV), which uses quadratic loss function instead of the idea of exponential family of Xiang and Wahba (1996).

The rest of this paper is organized as follows. In Section 2 we review the kernel logistic regression briefly. In Section 3 we propose SKLR with longitudinal data using mixed-effect kernel extensions with semiparametric fixed effects and parametric random effects. In Section 4 we propose GCV function for the model selection. In Section 5 we perform the numerical studies through examples. In Section 6 we give the conclusions.

2. Kernel logistic regression

A nonlinear form of logistic regression, known as kernel logistic regression, can be obtained via kernel tricks (Pi *et al.*, 2011), whereby a conventional logistic regression model is constructed in a high dimensional nonlinear feature space induced by a Mercer's kernel (1909).

More formally, given a training data, $D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, $\boldsymbol{x}_i \in X \subset R^d$, $y_i \in R$, a feature space $F; (\phi : X \to F)$, is defined by a kernel function, $K : X \times X \to R$, that evaluates the inner product between the images of input vectors in the feature space, i. e. $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \phi(\boldsymbol{x}_k)'\phi(\boldsymbol{x}_l)$. The most popular kernel function used for nonlinear case is the Gaussian kernel,

$$K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{x}_k - \boldsymbol{x}_l\|^2\right), \tag{2.1}$$

where σ^2 is the kernel parameter. The negative log-likelihood function of the kernel logistic regression model constructed in the feature space is given as follows:

$$\ell(\boldsymbol{\eta}) = -\sum_{i=1}^{n} y_i\eta_i + \sum_{i=1}^{n} \log\left(1 + \exp(\eta_i)\right), \tag{2.2}$$

where $\eta_i = \boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + b$. Often, model penalization improves generalization performance and so we employ the following penalty term $\|\boldsymbol{\omega}\|^2$ to the negative log-likelihood function during model fitting:

$$\ell(\boldsymbol{\omega}, b) = -C_1 \sum_{i=1}^{n} (y_i(\boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + b)) + C_1 \sum_{i=1}^{n} \log\left(1 + \exp(\boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + b)\right) + \frac{1}{2}\|\boldsymbol{\omega}\|^2, \tag{2.3}$$

where $C_1 > 0$ is a penalty parameter which controls the trade-off between the goodness-of-fit on the data and the smoothness.

From now we denote by $K$the $n \times n$ matrix consisting of $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then the representation theorem (Kimeldorf and Wahba, 1971) guarantees that the minimizer of the penalized negative log-likelihood (2.3) to be $\eta_i = \sum_{j=1}^{n} K_{ij}\alpha_j + b = K_{1i}\boldsymbol{\alpha}$, where $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, K_{1i} is the ith row of the kernel matrix $K_1 = (K, \mathbf{1})$ and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n, b)'$. Now the problem becomes obtaining $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n, b)'$ to minimize

$$\ell(\boldsymbol{\alpha}) = -C_1 \sum_{i=1}^{n} y_i(K_{1i}\boldsymbol{\alpha}) + C_1 \sum_{i=1}^{n} \log\left(1 + \exp\left(K_{1i}\boldsymbol{\alpha}\right)\right) + \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{K}_0\boldsymbol{\alpha} \tag{2.4}$$

where $K_0 = \begin{pmatrix} K & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}$. Note that $p_i = \exp(\eta_i)/(1+\exp(\eta_i)) = \exp\left(K_{1i}\boldsymbol{\alpha}\right)/(1+\exp\left(K_{1i}\boldsymbol{\alpha}\right))$.

Using Newton's method, the estimate of $\boldsymbol{\alpha}$ can be obtained iteratively as follows:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha} - \left(K_1' W K_1 + \frac{1}{C_1}K_0\right)^{-1}\left(K_1'(\boldsymbol{p} - \boldsymbol{y}) + \frac{1}{C_1}K_0\boldsymbol{\alpha}\right), \tag{2.5}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)'$ $\boldsymbol{p} = (p_1, \cdots, p_n)'$ and W is a diagonal matrix with the ith diagonal element of $p_i(1-p_i)$. The estimate of $\boldsymbol{\alpha}$ can be obtained iteratively reweighted least squares (IRWLS) algorithm from (2.5) as follows:

$$\boldsymbol{\alpha} = \left(K_1' W K_1 + \frac{1}{C_1}K_0\right)^{-1} K_1' W \boldsymbol{z}, \tag{2.6}$$

where $\boldsymbol{z} = (z_1, \cdots, z_n)'$ is the working response vector such that $z_i = \eta_i + (y_i - p_i)/(p_i(1-p_i))$.

3. SKLR with longitudinal data

Let y_{ij} be the indicator function such that 1 if the jth observation in the ith subject belongs to a certain class, 0 otherwise. Then we can consider a logistic regression model of the form,

$$\log \frac{p_{ij}}{1 - p_{ij}} = b_0 + \eta_1(\boldsymbol{z}_{ij}) + \eta_2(\boldsymbol{x}_{ij}) + b_i \text{ for } \; i = 1, \cdots, N, \; j = 1, \cdots, n_i,$$

where $(\boldsymbol{z}_{ij}, \boldsymbol{x}_{ij})$ are covariates, the random effect b_i assumed to follow iid $N(0, \sigma_B^2)$ can be interpreted as the subject effect and p_{ij} is the probability of the jth observation in the ith subject,

$$p_{ij} = P(y_{ij} = 1 | \boldsymbol{z}_{ij}, \boldsymbol{x}_{ij}) = \frac{\exp(b_0 + \eta_{1ij} + \eta_{2ij} + b_i)}{1 + \exp(b_0 + \eta_{1ij} + \eta_{2ij} + b_i)}.$$

The negative log-likelihood can be written as

$$L(\boldsymbol{\eta}, \boldsymbol{b}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \{-y_{ij}(b_0 + \eta_{1ij} + \eta_{2ij} + b_i) + \log(1 + \exp(b_0 + \eta_{1ij} + \eta_{2ij} + b_i))\}.$$

The weights η_{1ij} and η_{2ij} are related to covariates $\boldsymbol{z}_{ij}$ and $\boldsymbol{x}_{ij}$ such as $\eta_{1ij} = \boldsymbol{\beta}'\boldsymbol{z}_{ij}$, and $\eta_{2ij} = \boldsymbol{\omega}'\phi(\boldsymbol{x}_{ij})$, respectively. Known that $\phi(u)'\phi(v) = K(u, v)$ which are obtained from the

application of Mercer's conditions (1909). Then the estimates of $(b_0, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{b})$ are obtained by minimizing the penalized negative log likelihood:

$$\begin{aligned} L(b_0, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{b}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \{ & -y_{ij}(b_0 + \boldsymbol{\beta}' \boldsymbol{z}_{ij} + \boldsymbol{\omega}' \phi(\boldsymbol{x}_{ij}) + b_i) \\ & + \log(1 + \exp(b_0 + \boldsymbol{\beta}' \boldsymbol{z}_{ij} + \boldsymbol{\omega}' \phi(\boldsymbol{x}_{ij}) + b_i))\} \\ & + \frac{\lambda_1}{2} ||\boldsymbol{\omega}||^2 + \frac{\lambda_2}{2} ||\boldsymbol{b}||^2, \end{aligned} \tag{3.1}$$

where λ_1 and λ_2 are nonnegative constants which control the tradeoff between the goodness-of-fit on the data and $||\boldsymbol{\omega}||^2$ and $||\boldsymbol{b}||^2$.

The representation theorem (Kimeldorf and Wahba, 1971) guarantees that the minimizer of the penalized negative log likelihood to be $\eta_{1ij} = K^{ij}\boldsymbol{\alpha}$, where K^{ij} is the row of K^* corresponding to $\boldsymbol{x}_{ij}$, K^* is the $\sum_i n_i \times \sum_i n_i$ kernel matrix generated from $\{\boldsymbol{x}_{ij}\}$, and $\boldsymbol{\alpha}$ is a $\sum_i n_i \times 1$ parameter vector to be estimated.

Now the problem (3.1) becomes obtaining $b_0, \boldsymbol{\beta}, \boldsymbol{\alpha}$ and $\boldsymbol{b}$ ($N \times 1$ vector) to minimize

$$\begin{aligned} L(b_0, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \{ & -y_{ij}(b_0 + \boldsymbol{\beta}' \boldsymbol{z}_{ij} + K^{ij}\boldsymbol{\alpha} + b_i) \\ & + \log(1 + \exp(b_0 + \boldsymbol{\beta}' \boldsymbol{z}_{ij} + K^{ij}\boldsymbol{\alpha} + b_i)) \\ & + \frac{\lambda_1}{2} \boldsymbol{\alpha}' K^* \boldsymbol{\alpha} + \frac{\lambda_2}{2} ||\boldsymbol{b}||^2. \end{aligned} \tag{3.2}$$

The penalized negative log-likelihood (3.2) can be rewritten as

$$L(\boldsymbol{\beta}) = -\boldsymbol{y} V \widetilde{\boldsymbol{\alpha}} + \log(\mathbf{1} + \exp(V\widetilde{\boldsymbol{\alpha}})) + \widetilde{\boldsymbol{\alpha}}' U \widetilde{\boldsymbol{\alpha}}, \tag{3.3}$$

where

$$\tilde{\alpha} = \begin{pmatrix} b_0 \\ \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \boldsymbol{b} \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_{\cdot 1} \\ \vdots \\ \boldsymbol{y}_{\cdot N} \end{pmatrix}, \ V = (\mathbf{1}, Z, K^*, L), \ U = \begin{pmatrix} \mathbf{0} & \mathbf{0}_1 & \mathbf{0}_2 \\ \mathbf{0}_1' & \frac{\lambda_1}{2} K^* & \mathbf{0}_3 \\ \mathbf{0}_2' & \mathbf{0}_3' & \frac{\lambda_2}{2} I_N \end{pmatrix},$$

$\mathbf{0} = (p+1) \times (p+1)$ zero matrix, $\mathbf{0}_1 = (p+1) \times \sum_{i=1}^{N} n_i$ zero vector, $\mathbf{0}_2 = (p+1) \times N$ zero matrix, $\mathbf{0}_3 = \sum_{i=1}^{N} n_i \times N$ zero matrix, $\boldsymbol{y}_{\cdot 1}' = \left(\boldsymbol{y}_{11}, \cdots, \boldsymbol{y}_{n_1 1}\right)'$, $L = \sum_{i=1}^{N} n_i \times N$ matrix with ith column $L_{\cdot i}' = \left(\mathbf{0}_4, \mathbf{1}_{n_i \times 1}, \mathbf{0}_5\right)'$, $\mathbf{0}_4 = \sum_{l=1}^{i-1} n_l \times 1$ zero vector, and $\mathbf{0}_5 = \sum_{l=i+1}^{N} n_l \times 1$ zero vector.

By minimizing the penalized negative log-likelihood (3.3) we obtain the estimator of parameter vector $\widetilde{\boldsymbol{\alpha}}' = \left(b_0, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}\right)'$, but not in a explicit form, which leads to use the iterative reweighted least squares (IRWLS) procedure. At $(t+1)$st iteration, the parameter vector is estimated as follows,

$$\widetilde{\boldsymbol{\alpha}}^{(t+1)} = (V'WV + U)^{-1} V'W\boldsymbol{y}^*, \tag{3.4}$$

where W is a diagonal matrix of $\boldsymbol{p}^{(t)}(1 - \boldsymbol{p}^{(t)})$ and $\boldsymbol{y}^* = W^{-1}(\boldsymbol{y} - \boldsymbol{p}^{(t)}) + V\widetilde{\boldsymbol{\alpha}}^{(t)}$ is the working response vector.

4. Model selection using GCV function

The functional structures of the kernel logistic regression for longitudinal data is characterized by hyper-parameters, the regularization parameters (λ_1, λ_2) and the kernel parameters.

For the model selection of the kernel logistic regression, we define the leave-one-out cross validation (CV) function for a set of hyperparameters, θ, as follows:

$$CV(\theta) = \frac{1}{N_n}\sum_{i}^{N}\sum_{j=1}^{n_i}(y_{ij} - \widehat{p_\theta}^{(-ij)}(\boldsymbol{x}_{ij}^*))^2,$$

which can be rewritten as follows by denoting the kth observation as $\{y_k, \boldsymbol{x}_k^*\} = \{y_k, (\boldsymbol{z}_k, \boldsymbol{x}_k)\}$,

$$CV(\theta) = \frac{1}{N_n}\sum_{k=1}^{N_n}(y_k - \widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*))^2,$$

where $\widehat{p}_\theta(\boldsymbol{x}_k^*)$ is the estimate of $p(\boldsymbol{x}_k^*)$ from full data and $\widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*)$ is the estimate of $p(\boldsymbol{x}_k^*)$ from data without kth observation. Since for each candidate of hyper-parameter sets, $\sum_i n_i$ of $\widehat{p_\theta}^{(-k)}(\boldsymbol{x_k^*})$'s should be computed, selecting parameters using CV function is computationally formidable. By leaving-out-one lemma (Craven and Wahba, 1979), we have

$$\begin{aligned}(y_k - \widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*)) - (y_k - \widehat{p}_\theta(\boldsymbol{x}_k^*)) &= \widehat{p}_\theta(\boldsymbol{x}_k^*) - \widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*) \\ &\approx \frac{\partial\widehat{p}_\theta(\boldsymbol{x}_k^*)}{\partial y_k}(y_i - \widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*))\end{aligned}$$

and

$$\frac{\partial\widehat{p}_\theta(\boldsymbol{x}_k^*)}{\partial y_k} = \frac{\partial\widehat{\nu}_\theta(\boldsymbol{x}_k^*)}{\partial y_k^*}\frac{\partial\widehat{p}_\theta(\boldsymbol{x}_k^*)}{\partial\widehat{\nu}_\theta(\boldsymbol{x}_k^*)}\frac{\partial y_k^*}{\partial y_k} = s_{kk}, \tag{4.1}$$

where $\widehat{\nu}_\theta(\boldsymbol{x}_k^*) = \widehat{b}_0 + \widehat{\eta}_1(\boldsymbol{z}_k) + \widehat{\eta}_2(\boldsymbol{x}_k) + \widehat{b}_k$ and $s_{jk} = \partial\widehat{\nu}_\theta(\boldsymbol{x}_j^*)/\partial y_k^*$ is the (j,k) th element of S which is the hat matrix such that $\widehat{\boldsymbol{\nu}}_\theta(\boldsymbol{x}^*) = S\boldsymbol{y}^*$, $S = V(V'WV + U)^{-1}V'W$. From (4.1) we have $\widehat{p}_\theta(\boldsymbol{x}_k^*) - \widehat{p}_\theta^{(-k)}(\boldsymbol{x}_k^*) \approx s_{kk}(y_i - \widehat{p}_\theta(\boldsymbol{x}_k^*))$.Then the ordinary cross validation (OCV) function can be obtained as

$$OCV(\theta) = \frac{1}{N_n}\sum_{k=1}^{N_n}\left(\frac{y_k - \widehat{p}_\theta(\boldsymbol{x}_k^*)}{1 - s_{kk}}\right)^2. \tag{4.2}$$

Replacing s_{kk} by their average $tr(S)/N_n$, the generalized cross validation (GCV) function can be obtained as

$$GCV(\theta) = \frac{N_n\sum_{k=1}^{N_n}(y_k - \widehat{p}_\theta(\boldsymbol{x}_k^*))^2}{(N_n - tr(S))^2}. \tag{4.3}$$

Details of derivation of GCV function can be found in Cho *et al.* (2010), Hwang (2010), and Hwang (2011).

5. Numerical studies

We illustrate the performance of the semiparametric kernel logistic regression through the simulated data and the real data on the nonlinear cases.

For the simulated data, we set $N = 20$, $n_i = 10$, and

$$\log \frac{p_{ij}}{1-p_{ij}} = -0.25 + \sin(2\pi x_{ij}) + b_i,$$

where $x_{ij} \sim U(0,1)$, $b_i \sim N(0, 0.1^2)$ for $i = 1, \cdots, 20(N)$. The radial basis kernel function is utilized in this example, which is

$$K(x_1, x_2) = \exp(-\frac{1}{\sigma^2}(x_1 - x_2)^2).$$

$(\lambda_1, \lambda_2, \sigma^2)$ is selected as (1, 1, 0.004) from the GCV function (4.3). b_0 is estimated as -0.7662.

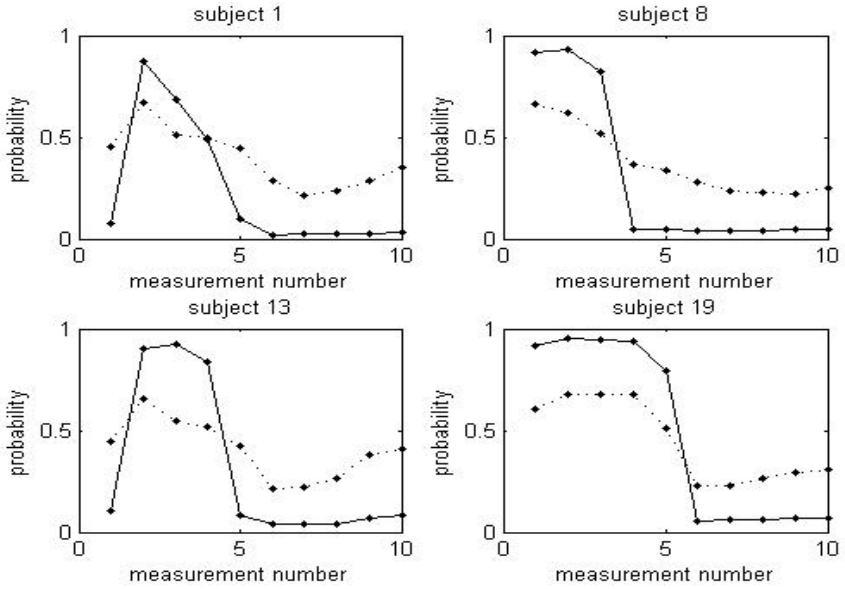

Figure 5.1 True probability (solid line), estimated probability (dotted line) of randomly selected 4 subjects

72 patients with acute spinal cord injury and bacteriuria were randomly divided into two treatment groups. 36 patients in the first group were treated for all episodes of urine tract infection and 36 patients in the second group were treated only if two specific symptoms occurred (Joe, 1997).

We set the probability of bacteriuria of patient i at j-th time,

$$P(y_{ij} = 1 | \boldsymbol{z}_{ij}, \boldsymbol{x}_{ij}) = \frac{\exp(b_0 + \beta z_{ij} + \eta_2(x_{ij}) + b_i)}{1 + \exp(b_0 + \beta z_{ij} + \eta_2(x_{ij}) + b_i)},$$

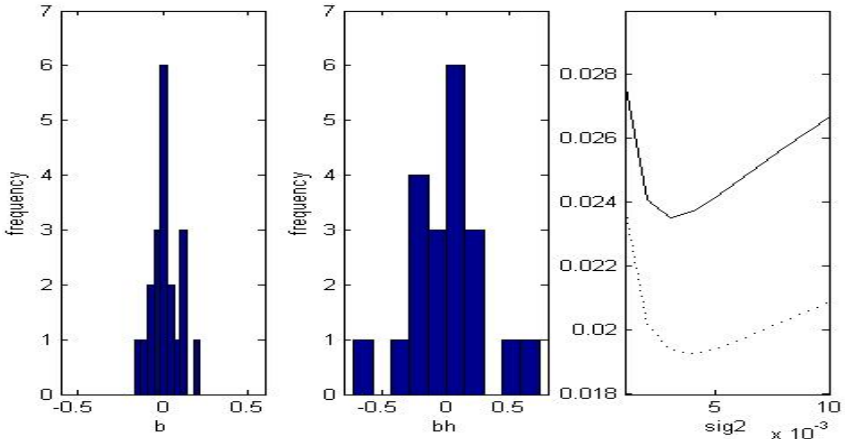

Figure 5.2 Histogram of true b's (Left) and estimated b's (Middle). CV function (solid line) and GCV function (dotted line) with $(\lambda_1, \lambda_2\ )$=(1,1) (Right)

where z=indicator function of the second treatment, x=time (weeks), $i = 1, \cdots, 72(N)$, $j = 1, \cdots, n_i$ with $\sum_{i=1}^{N} n_i = 820$. $(\lambda_1, \lambda_2, \sigma^2)$ is selected as (10, 5, 2) from GCV function (4.3). b_0 is estimated as -0.4680 and β is estimated as 1.2391, which implies that the second treatment tends to have larger probabilities than the first treatment, that is, the first treatment looks superior.

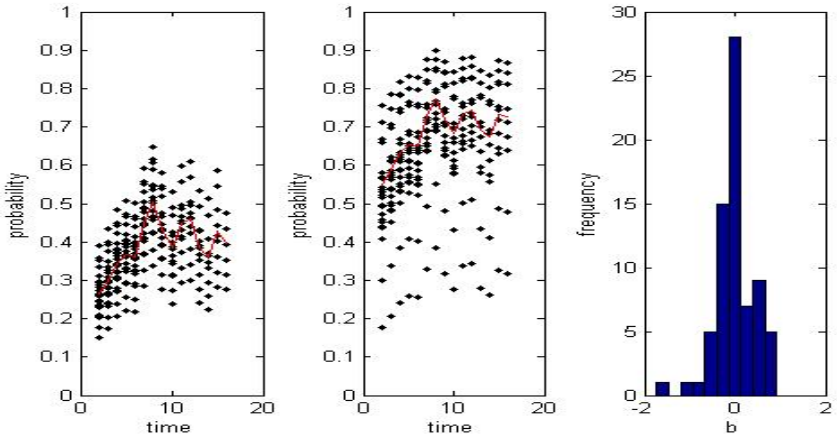

Figure 5.3 Probability of bacteriuria for the first treatment (Left) and the second treatment (Middle), solid line : average of probabilities at each time. Histogram of $\widehat{b}$ (Right)

6. Conclusions

In this paper, we dealt with estimating semiparametric kernel logistic regression of longitudinal data using IRWLS procedure and obtained GCV function. Through the examples we showed that the proposed procedure derives the satisfying results. We focused on the binary classification problem in this paper. The multi-classification problem will be studied in the future paper.

References

Agresti, A. (2002). *Categorical data analysis,* Wiley-Interscience, New York.

Amemiya, T. (1985). *Advanced econometrics,* Harvard University Press, Boston.

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. *Numerical Mathematics.,* **31**, 377-403.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis,* John Wiley and Sons.

Hwang, C. (2010). Kernel method for autoregressive data. *Journal of the Korean Data & Information Science Society*, **20**, 467-472.

Hwang, C. (2011). Asymmetric least squares regression estimation using weighted least squares support vector machine. *Journal of the Korean Data & Information Science Society*, **22**, 999-1005.

Joe, H. (1997). *Multivariate models and dependence concepts,* Chapman and Hall, London.

Kimeldorf, G. S. and Wahba, G. (1971) Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society A*, 415-446.

Pi, S. Y., Park, H. J. and Rhu, K. H. (2011). An analysis of satisfaction index on computer education. *Journal of the Korean Data & Information Science Society*, 22, 921-929.

Shim, J. and Seok, K. H. (2008). Kernel Poisson regression for longitudinal data. *Journal of the Korean Data & Information Science Society*, **19**, 1353-1360.

Smola, A. and Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator Inversion. *Algorithmica*, **22**, 211-231.

Vapnik, V. N. (1995). *The nature of statistical learning theory,* Springer, New York.

Winkelmann, R. (2003). *Econometric analysis of count data,* Springer Verlag, New York.

Wu, H. and Zhang, J. (2006). *Nonparametric regression methods for longitudinal data analysis,* Wiley, New York.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, **6**, 675-692.