Communications of the Korean Statistical Society Vol. 15, No. 2, 2008, pp. 205–211

Kernel Ridge Regression with Randomly Right Censored Data

Jooyong Shim[1], Kyung Ha Seok[2]

Abstract

This paper deals with the estimations of kernel ridge regression when the responses are subject to randomly right censoring. The iterative reweighted least squares(IRWLS) procedure is employed to treat censored observations. The hyperparameters of model which affect the performance of the proposed procedure are selected by a generalized cross validation(GCV) function. Experimental results are then presented which indicate the performance of the proposed procedure.

Keywords: Generalized cross validation function; kernel ridge regression; randomly right censoring; iterative reweighted least squares procedure; Kaplan-Meier estimator.

1. Introduction

The accelerated failure time model and the least squares method to accommodate the censored data seem appealing since they are familiar and well understood. Miller (1976) proposed a simple estimation in censored regression model by applying the weighted least square method. Koul *et al.* (1981) proposed a censored regression model using the weighted observations. Zhou (1992) proposed the M-estimators of regression parameter with the weights suggested by Koul *et al.* (1981). Yang (1999) proposed a censored median regression model as an alternative to the mean regression model for examining the input vector effect with the data subject to randomly right censoring and showed that the estimators are consistent and asymptotically distributed. Heuchenne and Van Keilegom (2005) proposed an estimation procedure which extends the least squares procedures for nonlinear regression with censored data.

Ridge regression (Hoerl and Kennard, 1970) is the classical statistical technique which implements a regularized form of the least squares regression. Kernel ridge regression (Saunders et al., 1998), which is a nonlinear form of ridge regression, is developed by introducing kernel functions satisfying Mercer conditions (Mercer, 1909). The least squares support vector machine, a formulation of kernel ridge regression including a bias term

1) Adjunct professor, Department of Applied Statistics, Catholic University of Daegu, Kyungbuk 712-702, Korea.

2) Professor, Institute of Statistical Information, Department of Data Science, Inje University, Kyungnam 621-749, Korea. Correspondence: statskh@paran.com

has been proposed for classification and regression by Suykens and Vanderwalle (1999). In kernel ridge regression the solution is given by a linear system instead of a quadratic program problem. The fact that kernel ridge regression has an explicit formulations has a number of advantages such as simple and fast computation.

In this paper we estimate the regression function by kernel ridge regression, where the weighted squared error loss function are included in the optimization problem. It is well known that the estimation performance of kernel ridge regression is affected by the hyper-parameters. We apply the generalized cross-validation method to kernel ridge regression with censored data. The rest of this paper is organized as follows. In Section 2 we give a brief review of kernel ridge regression. In Section 3 we present the estimation procedure of kernel ridge regression with the censored data. In Section 4 GCV function for selecting hyper-parameters is introduced. In Section 5 we perform the numerical studies through examples. In Section 6 we give the conclusions.

2. Kernel Ridge Regression

Let the data set be denoted by $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with each input vector $\boldsymbol{x}_i \in R^d$ including a constant 1 and the response $y_i \in R$ which is the output corresponding to $\boldsymbol{x}_i$. For kernel ridge regression, we can assume the functional form of unknown regression function f for given input vector $\boldsymbol{x}_0$ by $f(\boldsymbol{x}_0) = \boldsymbol{\omega}'\phi(\boldsymbol{x}_0)$ where $\boldsymbol{\omega}$ is an appropriate weight vector. Here the feature mapping function $\phi(\cdot) : R^d \to R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. The optimization problem is defined with a regularization parameter C which controls the trade-off between the goodness-of-fit on the data and $\boldsymbol{\omega}'\boldsymbol{\omega}$ as follows,

$$\min \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{C}{2}\sum_{i=1}^{n} e_i^2 \tag{2.1}$$

over $\{\boldsymbol{\omega}, \boldsymbol{e}\}$ subject to equality constraints,

$$y_i = \boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + e_i, \ i = 1, \ldots, n. \tag{2.2}$$

The Lagrangian function can be constructed as

$$L(\boldsymbol{\omega}, \boldsymbol{e}; \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{C}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i(\boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + e_i - y_i), \tag{2.3}$$

where α_i's are the Lagrange multipliers. The Karush-Kuhn-Tucker (Smola and Schölkopf, 1998) conditions for optimality are given by

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{\omega}} = \mathbf{0} \to \boldsymbol{\omega} = \sum_{i=1}^{n} \alpha_i \phi(\boldsymbol{x}_i), \\
&\frac{\partial L}{\partial e_i} = 0 \to e_i = \frac{1}{C}\alpha_i, \quad i = 1, \ldots, n, \\
&\frac{\partial L}{\partial \alpha_i} = 0 \to 0 = y_i - \boldsymbol{\omega}'\phi(\boldsymbol{x}_i) - e_i, \quad i = 1, \ldots, n,
\end{aligned}$$

leading to the solution

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + I/C)^{-1}\boldsymbol{y}, \tag{2.4}$$

where $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$ is a kernel matrix with $(k,l)^{th}$ element $K_{kl} = \phi(\boldsymbol{x}_k)'\phi(\boldsymbol{x}_l)$, $k, l = 1, \ldots, n$, which are obtained from the application of Mercer's conditions (1909), and I is an identity matrix. From the equation (2.4) the fitted regression function on $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2', ..., \boldsymbol{x}_n')'$ is obtained as

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{K}\hat{\boldsymbol{\alpha}} = \boldsymbol{K}(\boldsymbol{K} + I/C)^{-1}\boldsymbol{y}. \tag{2.5}$$

3. Kernel Ridge Regression with Censored Data

Consider the censored linear regression model for the response variables t_i's,

$$t_i = \boldsymbol{\beta}'\boldsymbol{x}_i + e_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}$ is the regression parameter vector of the model and e_i's are unobservable errors assumed to be independent with zero means and bounded variances. Let c_i's be the censoring variables assumed to be independent and identically distributed. The parameter vector of interest is $\boldsymbol{\beta}$, and t_i is not observed directly but we have

$$\delta_i = I_{(t_i < c_i)} \quad \text{and} \quad y_i = \min(t_i, c_i),$$

where $I_{(\cdot)}$ denotes the indicator function. The problem considered here is the estimation of $\boldsymbol{\beta}$ based on $(\delta_1, y_1, \boldsymbol{x}_1)$,...,$(\delta_n, y_n, \boldsymbol{x}_n)$. Miller (1976) proposed the least squares method for the estimation of regression parameters by minimizing the objective function as follows,

$$\int e^2 d\hat{F}(e) = \sum_{i=1}^{n} wt_i(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2,$$

where $\hat{F}$ is Kaplan-Meir estimator (1958) based on $\{e_i, \delta_i\}$, $e_i = y_i - \boldsymbol{\beta}'\boldsymbol{x}_i$, and the weight wt_i is the i^{th} jump of Kaplan-Meir estimator.

We consider the kernel ridge regression for censored nonlinear case, replacing the optimal problem (2.1) by

$$\min \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{C}{2}\sum_{i=1}^{n} wt_i e_i^2, \tag{3.1}$$

which is equivalent to

$$\min \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{K}\boldsymbol{\alpha} + \frac{C}{2}\sum_{i=1}^{n} wt_i(y_i - \boldsymbol{K}_i\boldsymbol{\alpha})^2, \tag{3.2}$$

where $\boldsymbol{K}_i$ is the i^{th} row of $\boldsymbol{K}$. From the equation (3.2) we obtain the Lagrange multipliers $\hat{\boldsymbol{\alpha}}$, but not in a explicit form since wt_i depends on $\hat{\boldsymbol{\alpha}}$, which leads to use the iterative reweighted least squares(IRWLS) procedure as follows:

1) Obtain the initial value of $\hat{f}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}^u)\hat{\boldsymbol{\alpha}}^u$ using $\hat{\boldsymbol{\alpha}}^u = (\boldsymbol{K}(\boldsymbol{x}^u, \boldsymbol{x}^u) + I/C)^{-1}\boldsymbol{y}^u$, where $\{\boldsymbol{x}^u, \boldsymbol{y}^u\}$ are uncensored data and $\boldsymbol{K}(\boldsymbol{a}, \boldsymbol{b})$ is a kernel matrix with data $\boldsymbol{a}$ and $\boldsymbol{b}$.

2) Obtain $W = \text{diag}\{wt_i\}$, where wt_i is the i^{th} normalized jump of Kaplan-Meier estimator from $\{e_i^{(l)}, \delta_i\}_{i=1}^n$ with $e_i^{(l)} = y_i - \hat{f}(\boldsymbol{x}_i)$.

3) Obtain $\hat{\boldsymbol{\alpha}}^{(l+1)} = (\boldsymbol{K}W\boldsymbol{K} + \boldsymbol{K}/C)^{-1}\boldsymbol{K}W\boldsymbol{y}$ and $\hat{f}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\hat{\boldsymbol{\alpha}}^{(l+1)}$.

4) Iterate steps 2–3 until convergence.

The fitted regression function is obtained as

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{K}\hat{\boldsymbol{\alpha}} = \boldsymbol{K}(\boldsymbol{K}W\boldsymbol{K} + \boldsymbol{K}/C)^{-1}\boldsymbol{K}W\boldsymbol{y}, \tag{3.3}$$

where $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$ is a kernel matrix.

4. Model Selection

The functional structure of kernel ridge regression with censored data is characterized by the hyper-parameters(regularization parameter and the kernel parameter). To determine these hyper-parameters, we define the cross-validation(CV) function with uncensored data $\{\boldsymbol{x}^u, \boldsymbol{y}^u\}$ as follows,

$$\text{CV}(\lambda) = \frac{1}{n_u}\sum_{i=1}^{n_u}\{y_i^u - \hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda)\}^2, \tag{4.1}$$

where λ is a set of hyper-parameters, $\hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda)$ is the regression function estimated without i^{th} observation of uncensored data. Since for each candidates of hyper-parameters, $\hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda), i = 1, \ldots, n_u$, should be evaluated, selecting parameters using CV function is computationally formidable. By the leave-out-one lemma (Craven and Wahba, 1979), we have

$$\begin{aligned}&\{y_i^u - \hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda)\} - \{y_i^u - \hat{f}(\boldsymbol{x}_i^u|\lambda)\}\\ &= \hat{f}(\boldsymbol{x}_i^u|\lambda) - \hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda) \approx \frac{\partial\hat{f}(\boldsymbol{x}_i^u|\lambda)}{\partial y_i^u}\{y_i^u - \hat{f}^{(-i)}(\boldsymbol{x}_i^u)\},\end{aligned}$$

we have $\hat{f}(\boldsymbol{x}_i^u|\lambda) - \hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda) \approx s_{ii}(y_i^u - \hat{f}^{(-i)}(\boldsymbol{x}_i^u|\lambda))$, where $s_{jk} = \partial\hat{f}(\boldsymbol{x}_j^u)/\partial y_k^u$ is the $(j,k)^{th}$ element of $S = \boldsymbol{K}(\boldsymbol{x}^u, \boldsymbol{x}^u)(\boldsymbol{K}(\boldsymbol{x}^u, \boldsymbol{x}^u) + I/C)^{-1}$ which is the hat matrix such that $\hat{f}(\boldsymbol{x}^u) = S\boldsymbol{y}^u$. Then $\hat{f}^{(-)}(\boldsymbol{x}^u|\lambda) = (\hat{f}^{(-1)}(\boldsymbol{x}_1^u|\lambda), \ldots, \hat{f}^{(-n_u)}(\boldsymbol{x}_{n_u}^u|\lambda))'$ can be expressed in terms of S as $\hat{f}^{(-)}(\boldsymbol{x}^u|\lambda) \approx H\boldsymbol{y}^u$, where $H = (I - D(S))^{-1}(S - D(S))$ and $D(S)$ the diagonal matrix of diagonals of S. Thus the ordinary cross validation(OCV) function can be obtained as

$$\text{OCV}(\lambda) = \frac{1}{n_u}\boldsymbol{y}^{u'}(I - H)'(I - H)\boldsymbol{y}^u. \tag{4.2}$$

Replacing s_{ii} by their average, the generalized cross validation(GCV) function can be obtained as

$$\mathrm{GCV}(\lambda) = n_u(n_u - tr(S))^{-2}\boldsymbol{y}^{u'}(I-S)'(I-S)\boldsymbol{y}^u. \tag{4.3}$$

5. Numerical Studies

We illustrate the performance of the censored regression estimation using kernel ridge regression through the simulated example on the nonlinear regression case and real example on the linear case. We ran MATLAB 7.0 over Pentium IV at 2.0GHz for the numerical studies. For the nonlinear censored regression case, 100 of x's are generated from a uniform distribution, $U(0,1)$, and (t,c)'s are generated as follows,

$$t_i = 1 + \sin(0.75\pi x_i) + \epsilon_{t_i},\ c_i = \sin(0.75\pi x_i) + \epsilon_{c_i},\ i = 1,\ldots,100,$$

where ϵ_{t_i}'s and ϵ_{c_i}'s are generated from normal distributions, $N(0,0.1)$ and $N(cc,0.1)$, respectively and cc is chosen for 25% censoring proportion. We repeated the above procedure 100 times to obtain 100 data sets. Then the regression function of t given x can be modelled as $f(x) = 1 + \sin(0.75\pi x_i)$. The Gaussian kernel is utilized in this example, which is,

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

The regularization parameter C and the kernel parameter σ^2 are obtained by GCV function (4.3). We used grid search method to find minimizers of equation (4.3). From IRWLS procedure in Section 3, $\boldsymbol{\alpha}$ is obtained. Then by the equation (3.3), we have the fitted regression function. Figure 5.1 (Left) shows true regression function (dotted line) and fitted regression function (solid line) for one of 100 data sets. Uncensored data points are denoted by '·' and those by 'o' are censored. In the figure we can see that the fitted regression function behaves similarly as the true regression function does. The average of 100 mean squared errors are obtained as 0.0044 and their standard deviation as 0.0038.

Stanford heart transplant data set (Miller and Halpern, 1982) consists of 152 patients with complete record who survived at least 10 days, there were 55 censored observations. In various analysis of this data set, the quadratic age model is considered where the age refers to the age at the first transplant time. Let $f(x_i)$ be the base 10 logarithm of the survival time of the i^{th} patient. To examine the age effect, we use the regression model as follows,

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

where x_i is the age at the first transplant time for $i = 1,\ldots,152$. The linear kernel is utilized in this example, which is,

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{x}_2,$$

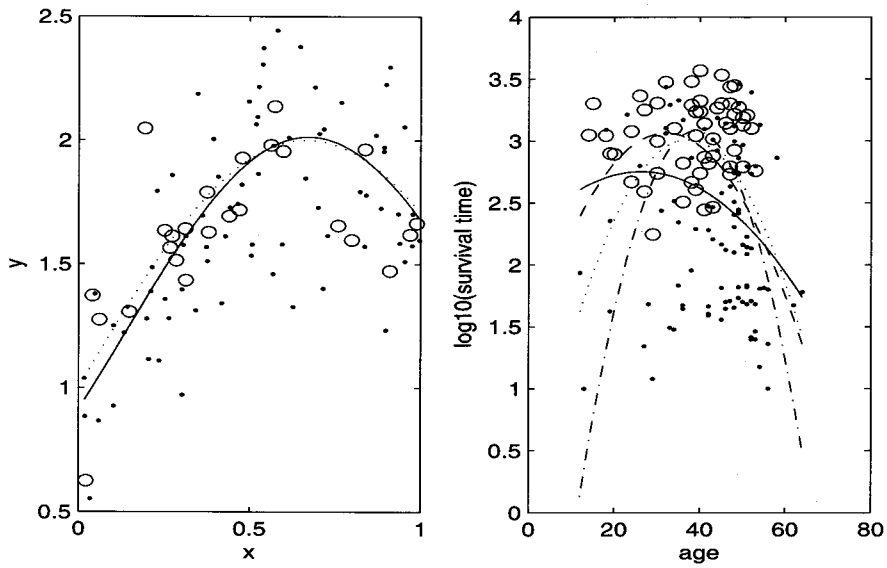

Figure 5.1: The true and the fitted regression functions for simulated data (Left). The fitted regression functions superimposed on the scatter plot of log_{10} of survival times versus age for Stanford heart transplant data (Right).

Table 5.1: The estimates of regression parameters for $\log_{10}$ of survival times on age and age squared.

	intercept	age	age squared
proposed	2.2609	0.0373	−0.0007
Zhou	−0.0208	0.1633	−0.0022
Koul *et al.*	−3.1028	0.3185	−0.0041
Buckley-James	1.3530	0.1067	−0.0017

where $\boldsymbol{x}_i = (1, x_i, x_i^2)$. The regularization parameter is obtained as 10 by GCV function (4.3). Figure 5.1 (Right) shows the fitted regression functions, where patients denoted by '·' are uncensored and those by 'o' are censored. Solid line is the proposed estimator, dotted line is Zhou estimator (1992), dash-dotted line is Koul *et al.* (1981), and dashed line is the Buckley-James estimator (1979), respectively. The plots of fitted values of regression functions show the proposed estimator agrees with Buckley-James estimator which was known well fitted on this data set (Miller and Halpern, 1982) on the fact that the younger patients had better survivals after transplant than older patients. Table 5.1 shows the estimates of regression parameters for log_{10} of survival times on age and age squared. The results of Buckley-James estimator are from Zhou (1992, Table 5.1).

6. Conclusions

In this paper, we dealt with estimating the regression function using kernel ridge regression when the responses are subject to randomly right censoring. We obtained GCV function for the proposed procedure. By using GCV function the model selection becomes easier and faster than that by a leave-one-out cross validation. Through the examples we showed that the proposed procedure derives the satisfying results and is attractive approaches to modelling of the censored data.

References

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.

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. *Numerische Mathematik*, **31**, 377–403.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Heuchenne, C and Van Keilegom, I. (2005). Nonlinear Regression with Censored Data. *Technical Report 520*, Universite catholique de Louvain.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Koul, H., Susarla, V. and Van Ryzin J. (1981). Regression analysis with randomly right censored data. *The Annals of Statistics*, **9**, 1276–1288.

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

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**, 449–464.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, 521–531.

Saunders, C., Gammerman, A. and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning.*, Madison, WI, 515–521.

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

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293–300.

Yang, S. (1999). Censored median regression using weighted empiricals survival and hazard functions. *Journal of the American Statistical Association*, **94**, 137–145.

Zhou, M. (1992). M-estimation in censored linear models. *Biometrika*, **79**, 837–841.

[Received October 2007, Accepted January 2008]