Published in final edited form as:

Comput Stat Data Anal. 2011 March 1; 55(3): 1479–1487. doi:10.1016/j.csda.2010.10.012.

Semiparametric Accelerated Failure Time Partial Linear Model and Its Application to Breast Cancer

Yubo Zou^a, Jiajia Zhang^a, and Guoyou Qin^{b,c}

Yubo Zou: zou@mailbox.sc.edu; Jiajia Zhang: jzhang@mailbox.sc.edu; Guoyou Qin: gyqin@fudan.edu.cn

^aDepartment of Epidemiology and Biostatistics, University of South Carolina Columbia, SC 29208, USA

^bDepartment of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, P.R. China

^cKey Laboratory of Public Health Safety, Ministry of Education (Fudan University), Shanghai 200032, P.R. China

Abstract

Breast cancer is the most common non-skin cancer in women and the second most common cause of cancer-related death in U.S. women. It is well known that the breast cancer survival varies by age at diagnosis. For most cancers, the relative survival decreases with age but breast cancer may have the unusual age pattern. In order to reveal the stage risk and age effects pattern, we propose the semiparametric accelerated failure time partial linear model and develop its estimation method based on the P-spline and the rank estimation approach. The simulation studies demonstrate that the proposed method is comparable to the parametric approach when data is not contaminated, and more stable than the parametric methods when data is contaminated. By applying the proposed model and method to the breast cancer data set of Atlantic county, New Jersey from SEER program, we successfully reveal the significant effects of stage, and show that women diagnosed around 38s have consistently higher survival rates than either younger or older women.

Keywords

Accelerated failure time model; Partial linear model; Penalized spline; Rank estimation; Robustness

1. Introduction

The proportional hazards (PH) model and the accelerated failure time (AFT) model are the most popular models in survival analysis. The PH model assumes that there is the linear relationship between the logarithm of hazard function and covariates of interest. One advantage of the PH model is its partial likelihood estimator which does not require additional parametric assumption for the hazard distribution. The existing softwares, such as R and SAS, can provide the partial likelihood estimator directly, so the PH model is used

© 2010 Elsevier B.V. All rights reserved.

Correspondence to: Jiajia Zhang, jzhang@mailbox.sc.edu; Guoyou Qin, gyqin@fudan.edu.cn.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

extensively in many areas, such as epidemiological cancer studies. However, the proportional hazard assumption may be avoided or hard to verify in real data analysis.

The AFT model, which considers the linear relationship between the logarithm of survival time and covariates of interests, is a useful alternative to the PH model when the proportional hazard assumption is not satisfied. Comparison with the PH model, estimated parameters in the AFT model can be easily interpreted in practice. There are many discussions on parametric estimation methods (Lawless, 2003; Kalbfleisch and Prentice, 2002) and semiparametric estimation methods (Tsiatis, 1990; Ritov, 1990; Jin et al., 2003, 2006; Zeng and Lin, 2007) for the AFT model. Tsiatis (1990) proposed the rank estimation method; Ritov (1990) considered the general linear square estimation method; and Jin et al. (2003, 2006) developed the rank estimation method and the least square estimation method. Most of the existing softwares can provide estimates of the parametric AFT model under the lognormal, loglogistic and Weibull distribution. Huang and Jin (2007) implemented the Gehan-rank estimation method and least square method in R by “lss” package.

For more flexible situation, a nonlinear structure of covariate is employed in the PH model, such as Gray (1992); Hastie and Tibshirani (1993). Huang (1999) investigated the efficient estimation of the partly linear additive Cox model and Heller (2001) developed the asymptotic distribution theory for the maximum profile partial likelihood estimate. Lu et al. (2001) studied a semiparametric survival model through a generalized profile likelihood method. Ma and Kosorok (2005) proposed the penalized log-likelihood estimation for partial linear transformation models with current status data. Cai et al. (2007a, b) studied the partial linear hazard regression for multivariate survival data.

Similarly, the nonlinear structure can also be incorporated into the AFT model, which is called as the accelerated failure time partial linear model (AFT-PLM). Because of recent advances in the semiparametric AFT model, such as Jin et al. (2003, 2006); Zeng and Lin (2007), it is possible to develop the estimation method in the semiparametric AFT-PLM. Orbe et al. (2003) firstly discussed the AFT-PLM based on the weighted least square method under certain conditions. As they pointed out, asymptotic properties are needed for their estimates. Chen et al. (2005) proposed an estimation method for the semiparametric AFT-PLM by suitably stratifying a Gehan-type extension of the Wilcoxon-Mann-Whitney estimating function. Due to the stratifying technique, the nonlinear structure can not be estimated directly. In this paper, we consider the semiparametric AFT-PLM and develop a new estimation method based on the rank estimation method and penalized spline approach and establish its asymptotic properties. The algorithm is easy to implement and the nonlinear structure can be estimated accurately.

The remaining paper is organized as follows. Section 2 describes the motivated data set, Section 3 presents the AFT partial linear model. A semiparametric estimation method for the proposed model is discussed in Section 4. Section 5 reports a simulation study to investigate the performance of the proposed model and method. Section 6 applies the proposed model to the breast cancer data set of Atlantic, New Jersey from SEER program. Finally, conclusions and some discussions are given in Section 7. The asymptotic properties is established in Appendix.

2. Motivating Data and Modeling Issues

Breast cancer is the most common non-skin cancer in women and the second most common cause of cancer-related death in U.S. women. A breast cancer data 2000–2004 of Atlantic County, New Jersey from the SEER program motivates this study. Observations with missing values on race, age, stage and marital status at diagnosis are excluded in this analysis. The total subjects in this study is 1584, among which the ratio of black women is

just 3.4%. Therefore, we exclude the black women in this studies. For each subject level, we will consider age, marital status (Single, Married and Other), and stage (Local, Regional and Distant), which are the most important effects of breast cancer survival. Local denotes an invasive neoplasm confined entirely to the organ of origin. Regional presents a neoplasm that has extended, and distant means that a neoplasm has spread to parts of the body remote from the primary tumor. The age ranges from 24 to 97 with the median age 61.

Commonly, the nonparametric Kaplan-Meier (KM) survival curve is used as a rule of thumb to evaluate the proportional hazards assumption (Klein and Moeschberger, 1997). If the proportional hazards assumption holds, the logarithm of the cumulative hazard function are expected to be parallel. We plot the logarithm of the cumulative hazard function from the Kaplan-Meier estimators (Figure 1) with respect to the marital status, which is done by “survfit” in R.

From visual inspection, we find there exists the crossover between the curve of single and the curve of married. It is also hard to see the parallel relationship between the curve of single and the curve of other (Figure 1). Therefore, the PH model will be questionable in this study. Furthermore, we perform the Schoenfeld residual test to check the PH assumption (Therneau and Grambsch, 2000). If the P-value from the Schoenfeld residual test is less than significant level (such as, 0.05), it indicates that the PH assumption is not satisfied. The Schoenfeld residual test can be conducted by “cox.zph” in R. The global p-value of this data set is 0.00149, which provides adequate evidence that the PH assumption does not hold. The time dependent PH model may be considered in this case, but the interpretation of estimates from the time dependent PH model is complex. Therefore, we prefer the AFT model in this study.

Since the AFT model is the regression model on the logarithm of survival time, then, the second question arising is whether we can model all risk covariate linearly? We plot the logarithm of the survival time with the age to check their linear relationship (Figure 2). The smooth curves is fitted by “lowess” function in R, which performs the computations for the LOWESS smoother using locally-weighted polynomial regression.

From Figure 2, we can see that the survival time tends to increase then decrease according to the increasing of age. It is apparent that the linear assumption may not be appropriate in this data set. Furthermore, we check the linear relationship between the logarithm of uncensored survival time and age by several correlation test, such as “Pearson test”, “Kendall τ test”, and “Spearman test”, and the P-values are 0.4564, 0.5247, and 0.5099. The P-values suggests that we do not have enough evidence to reject the null hypothesis and then we believe there is no linear relationship between the logarithm of survival time and age. Therefore, we consider to fit the data set by the AFT model considering the nonlinear effects.

3. Accelerated Failure Time Partial Linear Model

The accelerated failure time partial linear model (AFT-PLM) can be described as:

$$\log T = X^T \beta + f(u) + \varepsilon, \quad (1)$$

where β is the p -dimensional vector of regression coefficient, X is the p -dimensional vector of covariates and u is a 1-dimensional covariate. The AFT-PLM assumes that the covariate u is related with $\log T$ with a function $f(\cdot)$ and ε 's are independent error terms with a common distribution.

There are three situations we may consider in the AFT-PLM (1).

Case I: The function $f(\cdot)$ is fully specified. If the distribution of ε is also fully specified, we have fully parametric structure of the AFT-PLM and it can be fitted by the maximum likelihood method, and the existing software, such as “survreg” in R, can realize it. If the error distribution of ε is not specified, we have semiparametric structure of the AFT model, which can be fitted by the least square or rank estimation method (Jin et al., 2003, 2006). The “lss” package in R can provide the Gehan-rank estimates or least square estimates.

Case II: The function $f(\cdot)$ is unknown and the distribution of ε is fully specified. We should consider the nonparametric smoothing methods for function $f(\cdot)$. Various nonparametric smoothing methods can be applied to estimate the nonparametric function $f(\cdot)$, such as kernel and spline methods. Then the estimator can be obtained by the penalized likelihood estimation equation.

However, several issues may exist in Case I or II, which will lead to the incorrect inference of the covariate effects in practice, such as how to verify the parametric assumption of error distribution, and how to specify the closed form of $f(\cdot)$ based on the original data.

Case III: The function $f(\cdot)$ and the distribution of ε are fully unknown. This situation is more flexible than Case I or II, but there is few discussion on its estimation method.

Orbe et al. (2003) discussed the AFT-PLM based on the weighted least square method using the natural spline, and the GCV is used to choose the smoothing parameters. In this paper we propose the alternative estimation method based on the penalized spline and the rank estimation method and establish its asymptotic properties (see Appendix). It is worthwhile pointing out that the assumption of Orbe’s method, which is the expectation of the error term is zero, is not required in the proposed model and method.

4. Estimation Procedure

Let $(T_i, \Delta_i, X_i, u_i)$ denote the observed data for the i th individual $i = 1, \dots, n$, where T_i is the observed survival time for the i th patient, Δ_i is an indicator of censoring with $\Delta_i = 1$ for the uncensored time and $\Delta_i = 0$ for the censored time. It is common to assume that the censoring is independent and noninformative about the parameters of interest.

4.1. Smoothing method for $f(\cdot)$

There exist many smoothing techniques in the literature. For example, Lin and Carroll (2001) discussed the partial linear model by using the kernel method to approximate the nonparametric function; He et al. (2002) and Huang et al. (2007) adopted B-spline method to estimate the nonparametric function; Yu and Ruppert (2002) studied the estimation of partially linear single-index model by penalized splines (P-spline) to estimate the nonparametric function. We apply the P-spline in this paper and other splines can also be considered in the similar way.

The P-spline is an extension of smoothing splines and allows a more flexible choice of knots and penalty. Similar to the definition in Yu and Ruppert (2002), $f(\cdot)$ is a r th degree spline function under the working assumption. Then we have

$$f(u) = \pi^T(u) \alpha$$

where $\pi(u) = (B_1(u), \dots, B_{N_K}(u))^T$ is a vector of r degree B-spline basis functions with K internal knots where $N_K = K + r + 1$ and $\alpha \in R^{N_K}$ is the spline coefficient vector. Usually, K

$= 5 - 10$ knots seems quite adequate for smooth of either monotonic or unimodal regression, as suggested by Yu and Ruppert (2002). The equally spaced sample quantiles of $\{u_i, i = 1, \dots, n\}$ is used as knots and 10 knots are chosen in the simulation study.

4.2. Rank-like estimation method

Replacing the nonparametric function $f(u)$ by $\pi^T(u)\alpha$, the AFT-PLM (1) can be rewritten as

$$\log T_i = X_i^T \beta + \pi^T(u_i) \alpha + \varepsilon_i = D_i^T \theta + \varepsilon_i \quad (2)$$

where $D_i = (X_i^T, \pi^T(u_i))^T$ and $\theta = (\beta^T, \alpha^T)^T$. Then we can take (2) as the general AFT model and the semiparametric estimation method will be applied. Among the estimation methods in the semiparametric AFT model, we choose the Gehan-rank estimation method in this study due to its easy implementation. As mentioned in Jin et al. (2003), other weight functions can be applied into rank estimation methods. Rank estimation method is widely used in the literature, such as Cai et al. (2009); Johnson (2008, 2009); Xu et al. (2010).

The Gehan-rank estimating equation proposed by Jin et al. (2003) can be expressed as:

$$U_G(\theta) = n^{-1} \sum_{i=1}^n \Delta_i S^{(0)}(\theta; e_i(\theta)) [D_i - \bar{D}(\theta; e_i(\theta))],$$

where

$e_i(\theta) = \log T_i - D_i^T \theta$, $S^{(0)}(\theta; t) = n^{-1} \sum_{i=1}^n Y_i(\theta; t)$, and $\bar{D}(\theta; t) = S^{(1)}(\theta; t) / S^{(0)}(\theta; t)$ with $S^{(1)}(\theta; t) = n^{-1} \sum_{i=1}^n Y_i(\theta; t) D_i$. $Y_i(\theta; t) = 1_{\{e_i(\theta) \geq t\}}$ is the risk set on the time scale of residuals. The estimating equation is the gradient of the convex function

$$L_G(\theta) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i \{e_i(\theta) - e_j(\theta)\}^-, \quad (3)$$

where $a^- = |a| 1_{\{a < 0\}}$. The minimization of $L_G(\theta)$ with respect to θ can be carried out by the linear programming method.

To achieve a smooth fit, we incorporate a penalty term into (3). The penalized loss function is defined as:

$$\begin{aligned} PL_G(\theta) &= \text{Loss} + \text{Penalty} \\ &= n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i \{e_i(\theta) - e_j(\theta)\}^- + \frac{1}{2} \lambda \theta^T \Psi \theta, \end{aligned} \quad (4)$$

where λ is the smoothing parameter, and Ψ is a $(K + r + 1 + p) \times (K + r + 1 + p)$ matrix. All elements of Ψ are zeros except the right below matrix Ψ_1 with the order of $(K + r + 1) \times (K + r + 1)$. Therefore Ψ is a belt-shaped matrix

$$\text{with } \Psi = \begin{bmatrix} 0 & 0 \\ \Psi_1 & 0 \end{bmatrix} \text{ and}$$

$$\Psi_1 = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

More details can be found in Eilers and Marx (1996).

The estimator $\hat{\theta}$ of θ_0 can be obtained by minimizing $PL_G(\theta)$. We utilize the Nelder-Mead algorithm in obtaining the estimator, which is an option in “optim” function in R. The initial value is specified by the linear regression with respect to $D_i = (X_i^T, \pi^T(u_i))^T$. Then, $\hat{f}(u_0)$ can

be estimated by $\pi^T(u_0)\hat{a}$. Assume $\lambda = o_p(\frac{1}{\sqrt{n}})$, according to the general asymptotic theory for the rank estimator, the random vector $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically distributed as a zero-mean normal. Details can be found in appendix.

4.3. Variance estimation

Since there is no closed form to estimate the covariance matrix analytically, Jin et al. (2003) employed the resampling strategy to approximate the variance-covariance matrix of estimators in the semiparametric AFT model. Similarly, we approximate the variance-covariance matrix of $\hat{\theta}$ in the AFT-PLM by considering

$$PL_G^*(\theta) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i \{e_i(\theta) - e_j(\theta)\}^- Z_i + \frac{1}{2} \lambda \theta^T \Psi \theta, \quad (5)$$

where Z_i s are independent positive random variables with $E(Z_i) = \text{var } Z_i = 1$, and are independent of the data $(T_i, \Delta_i, X_i, u_i)$. For example, Z_i can be generated from the exponential distribution with mean one. Let $\hat{\theta}^*$ denote the minimizer of PL_G^* , that is equivalent to the root of the estimating equation

$$PU_G^*(\theta) = U_G^*(\theta) + \lambda \Psi \theta, \quad (6)$$

where

$$U_G^*(\theta) = n^{-1} \sum_{i=1}^n \Delta_i S^{(0)}\{\theta; e_i(\theta)\} [D_i - \bar{D}\{\theta; e_i(\theta)\}] Z_i.$$

In Appendix, we show that the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ can be approximated by the conditional distribution of $\sqrt{n}(\hat{\theta}^* - \theta)$ given the data $(T_i, \Delta_i, X_i, u_i)$.

Conditional on the data $(T_i, \Delta_i, X_i, u_i)$, the only random elements in $PL_G^*(\theta)$ are the Z_i s. To approximate the distribution of $\hat{\theta}$, we produce a large number of realizations of $\hat{\theta}^*$ by repeatedly generating the random sample $(Z_1, \dots, Z_n)$ while holding the data $(T_i, \Delta_i, X_i, u_i)$ at their observed values. The covariance matrix of $\hat{\theta}$ can be approximated by the empirical covariance matrix of $\hat{\theta}^*$. Confidence intervals for individual components of θ can be obtained from the percentiles of the empirical distribution of $\hat{\theta}^*$ or by the Wald method.

4.4. Choice of smoothing parameters

Selecting a suitable value of smoothing parameter λ is crucial to good curve fitting. We define the generalized cross-validation (GCV) score (Qu and Li, 2006; Johnson, 2008) as,

$$\text{GCV}(\lambda) = \frac{L_G}{(1 - \frac{1}{n}\text{df})^2},$$

where $\text{df} = \text{trace}\{H(\lambda)\}$ is the effective degree of freedom, and $H(\lambda) = \left(\frac{\partial^2 L_G}{\partial \theta \partial \theta^T} + \lambda \Psi \right)^{-1} \frac{\partial^2 L_G}{\partial \theta \partial \theta^T}$. The best smoothing parameter λ will be the minimizer of the GCV score, which is

$$\hat{\lambda} = \arg\min_{\lambda} \text{GCV}(\lambda).$$

It is worthwhile pointing out that the GCV criterion was analogous to AIC in Xu et al. (2010). In practice, the minimization can be carried out by grid search over a sequence of possible λ values. Yu and Ruppert (2002) suggested selecting λ over 30 grid points where the values of $\log_{10}(\lambda)$ are equally spaced between -6 and 7 . Through our simulation experience, this rule also works in the AFT-PLM.

4.5. Algorithm

We describes the algorithm process as follows:

Step 1: Give initial value of θ from the linear regression model. “glm” in R can be applied here.

Step 2: For each grid points $\lambda_1, \dots, \lambda_{30}$, minimize (5) to obtain the estimates of $\theta_1, \dots, \theta_{30}$. Calculate GCV score based on each θ_i , $i = 1, \dots, 30$. Then $\hat{\lambda} = \arg\min_{\lambda} \text{GCV}(\lambda)$ and the corresponding $\hat{\theta}$ is the estimate of θ .

Step 3: Resampling Z 's from the exponential distribution with mean one M times and minimizing (6) to obtain M estimates of θ^* . The empirical variance of θ^* is the approximation of variance of $\hat{\theta}$.

5. Simulation Study

In order to check the performance of the proposed model and method, we conduct simulations under several settings. The model we consider is

$$\log T_i = x_i \beta + \sin(\pi u_i) + \varepsilon_i, i = 1, \dots, n,$$

where x_i are drawn independently from the normal distribution with mean zero and standard deviation one ($N(0, 1)$) and u_i are independently generated from the uniform distribution on $(0, 1)$. ε_i 's follow the normal distribution $N(0, 0.5)$ and mixture normal distribution $0.5N(0, .$

$5) + 0.5N(0, 5)$. The former one is referred as the non-contaminated case and the latter one is the contaminated case. The purpose of these two distributions is to check the robustness of the proposed method. We take $\beta = 1$. The censoring time is generated from the exponential distribution to achieve 15% (light censoring) and 30% (moderate censoring) censoring rate. Both sample size $n = 200$ and $n = 400$ are considered in each setting with simulation time $N = 500$.

Firstly, we fit the data set by the proposed method, which is referred as “Proposed” in the table. In the proposed method, 200 resampling is used in the variance estimation. For the purpose of comparisons, we then fit the data set by the parametric approach with unknown $f(\cdot)$, which is denoted as “Parametric” in the table and the method from Orbe et al. (2003), which is denoted by “Orbe”. In the parametric approach, we assume that the error term comes from the normal distribution and $f(\cdot)$ is modeled by P-spline. In Orbe’s method, 200 resampling is used to estimated the variance. Similar penalty term and selection criterion are used in both approaches.

The results are reported in tables. For $\hat{\beta}$, we record its bias (Bias), empirical standard deviation (EMPSD) (average of 500 simulations), estimated standard deviation from resampling (ESTSD) and the coverage probability (CP). For the nonparametric function $f(u)$, we report the estimated integrated mean square error (IMSE), where

$$IMSE = \frac{1}{n} \sum_{i=1}^n (\hat{f}(u_i) - f(u_i))^2.$$

The results are reported in Table 1 for 200 sample size and Table 2 for 400 sample size.

From Table 1, we can see that the bias, standard deviation and IMSE are comparable from the parametric approach, proposed method and Orbe’s method under the non-contaminated case; while the bias from the proposed method is smallest among three methods under the contaminated case and the coverage probability is most stable from the proposed method. It is worthwhile pointing out that the empirical standard deviation from the parametric is not comparable to the estimated standard deviation from the model under the non-contaminated case since the normal assumption is applied in the contaminated case, while it is comparable in the non-contaminated case since the underlying distribution assumption is correct. The EMPSD and ESTSD are comparable from the proposed method and most of cases in Orbe’s method, which indicates that the estimated standard deviation from the resampling method works well. With the increase of the censoring rate, the bias and standard deviation will increase. Same tendency can be found in the IMSE.

Similar performance can be found from Table 2. Comparing the results from both tables, we find that the bias, the estimated standard deviation and the IMSE will decrease with the increasing of the sample size.

We also plot estimated nonparametric function $\hat{f}(u)$ along with their 95% confidence interval. The estimated confidence interval is obtained from the normal approximation using the empirical standard error of $\hat{f}(u_i^*)$ and normal approximation using the estimated standard error from resampling. For illustration purpose, we only illustrate the curve from the normal distribution with sample size 400 and 15% censoring.

From Figure 3, we can see that the estimated curve $\hat{f}(u)$ is very close to the true curve $\sin(\pi u)$ and the estimated confidence interval from empirical standard error (Figure 3(a)) is very

similar to that from the estimated standard error (Figure 3(b)). Therefore, the estimated curve performs well in the proposed method.

6. SEER Breast Cancer Data

We apply the proposed model and method to the breast cancer data set mentioned in Section 2. The model we consider is

$$\log T = \beta_1 \times \text{regional} + \beta_2 \times \text{distant} + \beta_3 \times \text{married} + \beta_4 \times \text{other} + f(\text{age}) + \varepsilon.$$

The estimated variances are from bootstrap method with 500 replications.

From Table 3, we can see that both stages have significant negative impact on the survival time. If the patient's stage change from the local to regional, the survival time will reduce to $e^{-0.534} = 0.586$ of local stage. If the patient's stage change from the local to distant, the survival time will reduce to $e^{-1.304} = 0.271$ of local stage. There are no significance effects of marital status. The results are pretty consistent with breast cancer studies. The nonlinear impact of age is shown in Figure 4.

From Figure 4, the patients at age around 38 always have higher survival probability than the younger or older age. It is interesting to see this nonlinear pattern and it may indicate the this age group can be treated well.

7. Discussion and Conclusion

In this paper, we proposed the AFT partial linear model and semiparametric estimation method based on the P-spline and the rank estimation method. The maximization can be realized by the Nelder-Mead method in R. The variance of the estimated parameters can be obtained from the resampling. The simulation studies illustrate the good performance of the proposed method, which is comparable with the parametric method and Orbe's method under the non-contaminated case and better than parametric and Orbe's method under the contaminated case. Thus the proposed method is much robust in the contaminated case. Compared with Chen et al. (2005), the proposed method can provide the nonlinear pattern directly. However, the resampling method for the variance estimation is quite time consuming and the non-simulation based method for the variance estimation may be an interest topic in the future work.

As suggested by the reviewer, it is interesting to extend this work to incorporate more than one covariate. One possible way is to extend the AFT-PLM model to the additive model with censored data, which can be written as

$$\log T_i = \sum_{k=1}^d h_k(R_{ki}) + X_i \beta + \varepsilon_i,$$

where $h_k(\cdot)$, $k = 1, \dots, d$ are unknown smooth functions, β is a p -dimensional vector of regression parameters. Comparing to the univariate nonlinear component, the estimation procedure in the additive model is more complicated. Another approach is the partial linear single index AFT model, which is

$$\log T_i = \eta(X_i^T \alpha) + Z_i^T \beta + \varepsilon_i, i = 1, \dots, n,$$

where $X_i \in R^d$, $Z_i \in R^{d_z}$, the unknown single index parameter α is in R^d , the unknown linear parameter β is in R^{d_z} , and $\eta(\cdot)$ is an unknown univariate function. By reducing the dimension from multivariate predictors to a univariate index $X_i^T \alpha$, single-index models avoid the so-called “curse of dimensionality” while still capturing important features in high-dimensional data. To adapt the proposed method to the additive model and the single index model will be interesting topics for the future study.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (10801039), Youth Science Foundation of Fudan university(08FQ29) and Shanghai leading Academic discipline Project (Project Number:B118).

Appendix

The proof of the asymptotic properties of $\hat{\theta}$ and $\hat{\theta}^*$ is similar to Jin et al. (2003). The

conditions 1–4 of Ying (1993) are assumed, furthermore we assume $\lambda = o_p(\frac{1}{\sqrt{n}})$ which is also used in Yu and Ruppert (2002) to obtain the asymptotic normality of $\hat{\theta}$ and $\hat{\theta}^*$. We sketch the outline of the proof in the following.

The asymptotic consistency of $\hat{\theta}$ and $\hat{\theta}^*$ can be obtained by the convex analysis. The estimation functions $n^{-1}PL_G$ and $n^{-1}PL_G^*$ converge almost sure to the same limiting function

according to the strong law of large number and $\lambda = o_p(\frac{1}{\sqrt{n}})$. Considering $n^{-1}PL_G$ and $n^{-1}PL_G^*$ are convex functions, the limiting functions have unique minimizer at θ_0 since their first derivatives PU_G and PU_G^* are zero at θ_0 and the second derivative, denoted as A_G , is nonsingular at θ_0 . Therefore, both $\hat{\theta}$ and $\hat{\theta}^*$ converge to θ_0 by the convex analysis.

The following gives the proof of the asymptotic normality. If $\lambda = o_p(\frac{1}{\sqrt{n}})$, applying Theorem 2 of Ying (1993), we have

$$PU_G(\hat{\theta}) = PU_G(\theta_0) + A_G(\hat{\theta} - \theta_0) + o(n^{-1/2} + \|\hat{\theta} - \theta_0\|) \quad (7)$$

and

$$PU_G^*(\hat{\theta}^*) = PU_G^*(\theta_0) + A_G(\hat{\theta}^* - \theta_0) + o(n^{-1/2} + \|\hat{\theta}^* - \theta_0\|) \quad (8)$$

almost surely. The functions PU_G and PU_G^* have the same asymptotic slope matrix A_G

because $E\{U_G^*(\theta) | \mathfrak{N}\} = U_G(\theta)$, $\hat{\theta}$ and $\hat{\theta}^*$ are consistent and $\lambda = o_p(\frac{1}{\sqrt{n}})$, where $\mathfrak{N}$ denotes the σ -field generated by the original data.

Then we have $PU_G^*(\hat{\theta}) = PU_G^*(\hat{\theta}) - PU_G(\hat{\theta}) + o(\frac{1}{\sqrt{n}})$. since $\hat{\theta}$ is a root of $PU_G(\theta)$. Thus,

$$\sqrt{n}PU_G^*(\widehat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_{-\infty}^{\infty} S^{(0)}(\widehat{\theta}; t) \{D_i - \overline{D}(\widehat{\theta}; t)\} dN_i(\widehat{\theta}; t) (Z_i - 1) + o_p(1). \quad (9)$$

Conditional on $\mathfrak{N}$, the right-hand side of the above equation is a normalized sum of independent zero-mean random vectors. Since its conditional covariance matrix converges to B_G , then $\sqrt{n}PU_G^*(\widehat{\theta})$ converges in distribution to $N(0, B_G)$. According to (8), the distribution of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ conditional on $\mathfrak{N}$ converges to $N(0, A_G^{-1} B_G A_G^{-1})$, which is the limiting distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$.

References

- Cai J, Fan J, Jiang J, Zhou H. Partially linear hazard regression for multivariate survival data. *J. Amer. Statist. Assoc.* 2007a; 102(478):538–551.
- Cai J, Fan J, Zhou H, Zhou Y. Hazard models with varying coefficients for multivariate failure time data. *Ann. Statist.* 2007b; 35(1):324–354.
- Cai T, Huang J, Tian L. Regularized Estimation for the Accelerated Failure Time Model. *Biometrics.* 2009; 65(2):394–404. [PubMed: 18573133]
- Chen K, Shen J, Ying Z. Rank estimation in partial linear model with censored data. *Statist. Sinica.* 2005; 15(3):767–779.
- Eilers PHC, Marx BD. Flexible smoothing with B -splines and penalties. *Statist. Sci.* 1996; 11(2):89–121. with comments and a rejoinder by the authors.
- Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association.* 1992; 87:942–951.
- Hastie T, Tibshirani R. Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B.* 1993; 55(4):757–796. with discussion and a reply by the authors.
- He X, Zhu Z-Y, Fung W-K. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika.* 2002; 89(3):579–590.
- Heller G. The Cox proportional hazards model with a partly linear relative risk function. *Lifetime Data Anal.* 2001; 7(3):255–277. [PubMed: 11677830]
- Huang J. Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* 1999; 27(5):1536–1563.
- Huang JZ, Zhang L, Zhou L. Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Statist.* 2007; 34(3):451–477.
- Huang L, Jin Z. Lss: Splines/r program for the accelerated failure time model based on least-squares principle. *Cmp. Biomd.* 2007; 86:45–50.
- Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika.* 2003; 90(2):341–353.
- Jin Z, Lin DY, Ying Z. On least-squares regression with censored data. *Biometrika.* 2006; 93(1):147–161.
- Johnson, Brent, A. Rank-based estimation in the $\uparrow_1$ -regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics.* 2009; 10(4):659–666. [PubMed: 19553356]
- Johnson BA. Variable selection in semiparametric linear regression with censored data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 2008; 70(2):351–370.
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data.* Hoboken, NJ: John Wiley & Sons; 2002.
- Klein, JP.; Moeschberger, ML. *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer-Verlag Inc; 1997.
- Lawless, JF. *Statistical Models and Methods for Lifetime Data.* 2nd Edition. Hoboken, NJ: John Wiley & Sons; 2003.

- Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* 2001; 96(455):1045–1056.
- Lu X, Singh RS, Desmond AF. A kernel smoothed semiparametric survival model. *J. Statist. Pl. Inf.* 2001; 98(1–2):119–135.
- Ma S, Kosorok MR. Penalized log-likelihood estimation for partly linear transformation models with current status data. *Ann. Statist.* 2005; 33(5):2256–2290.
- Orbe J, Ferreira E, Núñez Antón V. Censored partial regression. *Biostatistics (Oxford)*. 2003; 4(1): 109–121.
- Qu A, Li R. Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*. 2006; 62(2):379–391. [PubMed: 16918902]
- Ritov Y. Estimation in a linear regression model with censored data. *Ann. Statist.* 1990; 18(1):303–328.
- Surveillance, Epidemiology, and End Results (SEER) Program www.seer.cancer.gov Limited-Use Data, 1973–2005. National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch. released April 2008; based on the November 2007 submission
- Therneau, TM.; Grambsch, PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag Inc; 2000.
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* 1990; 18(1):354–372.
- Xu J, Leng C, Ying Z. Rank-based variable selection with censored data. *Statist. Comput.* 2010; 20(2): 165–176.
- Ying Z. A large sample study of rank estimation for censored regression data. *Ann. Statist.* 1993; 21:76–99.
- Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* 2002; 97(460):1042–1054.
- Zeng D, Lin D. Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* 2007; 102(480):1387–1396.

Figure 1.
Logarithm of the cumulative hazard function with survival time.

Figure 2. Logarithm of survival time with age. Triangle denotes the uncensored observations and circle presents the censored observations.

Figure 3.

Estimated $f(u)$ from the proposed method with sample size 400 and 15% censoring when the error term comes from normal distribution, along with 95% confidence interval.

Figure 4. Estimated $f(\text{age})$ from the proposed method along with 95% confidence interval using the estimated standard error from resampling.

Table 1

Bias, EMPSD, ESTSD and coverage probability of $\hat{\beta}$ and IMSE for $\hat{f}(u)$ from 500 simulated data set with sample size 200.

		$\hat{\beta}$			$\hat{f}(u)$	
		Bias	EMPSD	ESTSD	CP(%)	IMSE
N(0,.5)						
15%	Parametric	-.0163	.039	.041	91.8	.1872
	Proposed	.0131	.042	.042	90.4	.1887
	Orbe's	-.0144	.043	.048	95.0	.1874
30%	Parametric	-.0141	.044	.045	92.8	.1828
	Proposed	.0108	.047	.045	92.0	.1890
	Orbe's	-.0261	.052	.058	92.6	.2473
.5N(0,.5)+.5N(0,.5)						
15%	Parametric	.2357	.252	.519	80.2	2.9282
	Proposed	-.0719	.292	.277	92.0	2.8754
	Orbe's	-.1982	.392	.278	87.2	1.2935
30%	Parametric	.2358	.292	.846	84.4	5.0698
	Proposed	-.0872	.311	.284	92.0	5.4699
	Orbe's	-.2446	.415	.286	81.6	3.0820

Table 2

Bias, EMPSD, ESTSD and coverage probability of $\hat{\beta}$ and IMSE for $\hat{f}(u)$ from 500 simulated data set with sample size 400.

		$\hat{\beta}$			$\hat{f}(u)$	
		Bias	EMPSD	ESTSD	CP(%)	IMSE
N(0,.5)						
15%	Parametric	-.0105	.084	.094	90.0	.1792
	Proposed	.0056	.094	.090	93.8	.1915
	Orbe's	-.0029	.031	.035	95.2	.1491
30%	Parametric	-.0123	.094	.100	91.6	.1765
	Proposed	.0079	.105	.100	90.6	.1885
	Orbe's	-.0190	.040	.044	92.2	.1695
.5N(0,.5)+.5N(0,.5)						
15%	Parametric	.1410	.192	.330	82.4	2.2277
	Proposed	-.0273	.201	.192	93.6	2.3703
	Orbe's	-.1510	.291	.213	86.6	2.0530
30%	Parametric	.0795	.169	.314	84.4	4.6899
	Proposed	-.0486	.205	.204	91.8	4.3977
	Orbe's	.2486	.357	.222	76.0	2.8022

Table 3

Estimates (Est), standard deviation (SD) and confidence interval (CI) of estimated parameters for breast cancer data set from the AFT-PLM. The estimated variances are from bootstrap method with 500 replications.

	Est	SD	CI
reginal	-0.534	0.078	(-0.687, -0.381)
distant	-1.304	0.098	(-1.497, -1.112)
married	0.254	0.121	(0.018,0.490)
other	0.100	0.120	(-0.135,0.336)