Fitting Semiparametric Additive Hazards Models using Standard Statistical Software

Douglas E. Schaubel[*] and **Guanghui Wei**

Department of Biostatistics, University of Michigan, M4039 SPH II, 1420 Washington Heights, Ann Arbor, MI, 48109-2029, USA

Received 2 August 2006, revised 23 December 2006, accepted 1 March 2007

Summary

The Cox proportional hazards model has become the standard in biomedical studies, particularly for settings in which the estimation covariate effects (as opposed to prediction) is the primary objective. In spite of the obvious flexibility of this approach and its wide applicability, the model is not usually chosen for its fit to the data, but by convention and for reasons of convenience. It is quite possible that the covariates add to, rather than multiply the baseline hazard, making an additive hazards model a more suitable choice. Typically, proportionality is assumed, with the potential for additive covariate effects not evaluated or even seriously considered. Contributing to this phenomenon is the fact that many popular software packages (e.g., SAS, S-PLUS/R) have standard procedures to fit the Cox model (e.g., proc phreg, coxph), but as of yet no analogous procedures to fit its additive analog, the Lin and Ying (1994) semiparametric additive hazards model. In this article, we establish the connections between the Lin and Ying (1994) model and both Cox and least squares regression. We demonstrate how SAS's phreg and reg procedures may be used to fit the additive hazards model, after some straightforward data manipulations. We then apply the additive hazards model to examine the relationship between Model for End-stage Liver Disease (MELD) score and mortality among patients wait-listed for liver transplantation.

Key words: Additive risk model; Least squares regression; PROC PHREG; PROC REG; Schoenfeld residuals; Semiparametric estimation.

Supporting information is available from the author or on the WWW under http://www.wiley-vch.de/contents/jc_2221/2007/200610349_s.pdf

1 Introduction

Currently, the most popular regression method for survival analysis in biomedical studies is the Cox (1972) proportional hazards model, wherein the hazard at time t for subject i is given by:

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}_0^T \boldsymbol{Z}_i(t)\}, \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\beta}_0$ is a $p \times 1$ regression parameter and $\boldsymbol{Z}_i(t)$ is a $p \times 1$ vector of possibly time-varying covariates. The Cox model is especially popular in settings where the estimation of covariate effects is of chief interest, in which case $\lambda_0(t)$ is treated like a nuisance parameter. Estimation of $\boldsymbol{\beta}_0$ proceeds through partial likelihood (Cox, 1975) such that $\lambda_0(t)$ is not involved in the estimation of $\boldsymbol{\beta}_0$. Cox regression is the predominant model in biomedical studies and the original paper proposing model (1) is one of the most cited papers in science, let alone statistics.

[*] Corresponding author: e-mail: deschau@umich.edu, Phone: +01 734 615 9825, Fax: +01 734 763 2215

Under model (1), covariates are assumed to have multiplicative effects on the baseline hazard. While this is no doubt a plausible model, there is no guarantee of its appropriateness for a particular application. Models with additive covariate effects are well-accepted in other types of regression, most notably the linear regression model. Additive risk models have been considered by several authors (e.g., Aalen (1980); Breslow and Day (1980); Buckley (1984); Cox and Oakes (1984); Thomas (1986); Breslow and Day (1987); Aalen (1989); Huffer and McKeague (1991); Andersen et al. (1993)). The model which can be considered to be most closely connected to the Cox model was proposed by Lin and Ying (1994):

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{\theta}_0^T \boldsymbol{Z}_i(t)\,. \tag{2}$$

It is quite possible that an additive model may be more appropriate for a given application; particularly for continuous covariates since, under model (1), the hazard is assumed to increase exponentially per unit increase a given covariate element (i.e., with all other covariates held constant). An exponential increase may be too extreme in many practical applications. Due to its close connection with the Cox model, model (2) would appear to be a natural choice if the fit of the Cox model and/or the appropriateness of multiplicative covariate effects was in question. However, despite the potential liabilities of model (1), it is quite difficult to find real-data examples of hazard regression models with additive effects in the applied literature.

Since, in most applications, neither the investigators nor the statistical analysts have any reason a priori to believe the appropriateness of the Cox model, it would appear that model (1) is often applied by convention and out of convenience. That the Cox model is the default can hardly be questioned. Due to its semiparametric nature, the model is extremely flexible. However, the role of convenience cannot be understated and herein lies perhaps the biggest practical advantage of the Cox model; it can be fitted using standard implementations of widely available software, such as PROC PHREG in SAS and the *coxph* function in R/S-PLUS. It is very likely that use of additive hazard regression models would greatly increase such models could be fitted more easily.

In this article, we draw connections between estimation of $\boldsymbol{\theta}_0$ in the Lin and Ying (1994) additive hazard model and (i) Cox regression and (ii) least squares estimation. Specifically, we demonstrate that, following some straightforward data manipulations, $\hat{\boldsymbol{\theta}}$ can be computed by combining PROC PHREG and PROC REG, which are standard SAS procedures for Cox and linear regression, respectively.

The remainder of this article is organized as follows. In Section 2, we set up the requisite notation and review the estimation procedures proposed by Lin and Ying (1994) to fit the semiparametric additive hazard model. In Section 3, we establish the connection between parameter estimation in the additive hazard model and both Cox and least squares regression. We then demonstrate how to fit the additive hazard model in SAS using PROC PHREG and REG. In Section 4, we describe how to implement the proposed procedures. We then apply the Lin and Ying (1994) model to data on liver failure patients. Finally, discussion of our results and some concluding remarks are given in Section 5. SAS code to implement the proposed techniques is provided in the Appendix.

2 Additive Hazards Model: Parameter Estimation

In this section, we set up the requisite notation and list the formulae for estimating the regression parameter in the additive hazard model proposed by Lin and Ying (1994). Consistent with usual set-up for univariate survival data, the study population consists of n independent vectors, $(X_i, \Delta_i, \boldsymbol{Z}_i)$, where, for subject i ($i = 1, \ldots, n$), T_i is the failure time, C_i is the censoring time, $X_i = \min\{T_i, C_i\}$ is the observation time, $\Delta_i = I(T_i < C_i)$ is the observed-event indicator and $\boldsymbol{Z}_i$ is a $p \times 1$ vector of covariates. It is also useful to set up the commonly employed counting process notation, with the at-risk and observed-event counting processes defined as $Y_i(t) = I(X_i \geq t)$ and $N_i(t) = I(T_i \leq t, \Delta_i = 1) = \int_0^t \mathrm{d}N_i(s)$, respectively. The key quantities with respect to inference on $\hat{\boldsymbol{\theta}}$

are given by:

$$\boldsymbol{U} = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)] \, \mathrm{d}N_i(t) \tag{3}$$

$$\boldsymbol{A} = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)]^{\otimes 2} \, Y_i(t) \, \mathrm{d}t \tag{4}$$

$$\boldsymbol{B} = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)]^{\otimes 2} \, \mathrm{d}N_i(t) \,, \tag{5}$$

where, for any vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$; τ is a pre-specified time point usually set to $\max\{X_1, \ldots, X_n\}$ such that all observed failures are included in the analysis, and

$$\bar{\boldsymbol{Z}}(t) = \sum_{i=1}^{n} Y_i(t) \, Z_i(t) \sum_{i=1}^{n} Y_i(t) \tag{6}$$

is the at-risk weighted covariate mean at time t. Using counting process analogs of generalized estimating equation (GEE; Liang and Zeger (1986)) methods, Lin and Ying (1994) propose to estimate $\boldsymbol{\theta}_0$ by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}\boldsymbol{U} \,, \tag{7}$$

while the estimated variance of $\hat{\boldsymbol{\theta}}$ was derived to be:

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} \,. \tag{8}$$

In the next section, we propose techniques for exploiting standard software to implement the above-described estimation methods.

3 Fitting the Additive Hazard Model using Software for Cox and least Squares Regression

In this section we describe the connection between methods described in Section 2 and Cox and least squares regression. Through these connections, we propose techniques for fitting the Lin and Ying (1994) model using standard software designed for Cox and linear regression. Since the code in the Appendix is written in SAS, we describe the proposed techniques in the context of SAS's Cox (PROC PHREG) and linear regression (PROC REG) procedures. A SAS program is provided in the Appendix, and various code segments are provided up front in this section for continuity. Basically, PROC PHREG and PROC REG are combined to compute $\hat{\boldsymbol{\theta}}$, $\boldsymbol{A}$ and $\boldsymbol{B}$, with $\hat{V}(\hat{\boldsymbol{\theta}})$ then computed using only a few lines of basic PROC IML code. For notational convenience and clarity of illustration, we consider time-independent covariates. The extension of the proposed computational techniques to the case where $\boldsymbol{Z}_i(t) \neq \boldsymbol{Z}_i$ is straightforward and will be presented in Section 4. Supporting Information for this article is available from the first author or on the WWW under http://www.wiley-vch.de/contents/jc_2221/2007/200610xxx_s.pdf

To begin, we review estimation procedures for the proportional hazards (PH) model. Inference is carried out through partial likelihood (Cox, 1975) such that the quantity assumed to be of primary interest, $\boldsymbol{\beta}_0$, can be estimated without estimating the baseline hazard, $\lambda_0(t)$. The regression parameter vector is estimated by $\hat{\boldsymbol{\beta}}$, which is the solution to the score equation, $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}_p$, where

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t;\boldsymbol{\beta})] \, \mathrm{d}N_i(t) \,, \tag{9}$$

with the risk-weighted covariate mean given by

$$\bar{\boldsymbol{Z}}(t;\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} Y_i(t) \, \boldsymbol{Z}_i \exp\{\boldsymbol{\beta}^T \boldsymbol{Z}_i\}}{\sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{Z}_i\}} \,, \tag{10}$$

and where $\boldsymbol{a}_p$ is a $p \times 1$ vector with all elements equal to a. It is convenient to recycle the notation from Section 2. First, the formulae in (9) and (10) are quite standard. Second, $\bar{\boldsymbol{Z}}(t)$ in (6) in Section 2 equals $\bar{\boldsymbol{Z}}(t;0)$ in (10); similarly, $U = U(0)$. Such connections between the procedures of Lin and Ying (1994) and Cox (1972) are key to the methods we propose for fitting the additive hazards model.

The root of (9) is computed through Newton-Raphson iteration, with the $(j+1)$'th estimate given by

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} + \boldsymbol{I}(\hat{\boldsymbol{\beta}}^{(j)})^{-1}\, \boldsymbol{U}(\hat{\boldsymbol{\beta}}^{(j)})\,, \tag{11}$$

where $\boldsymbol{I}(\boldsymbol{\beta}) = -\partial \boldsymbol{U}/\partial \boldsymbol{\beta}^T$.

By default, PROC PHREG sets $\hat{\boldsymbol{\beta}}^{(0)} = 0$ which is a convenient choice for our purposes due to the fact that $\bar{\boldsymbol{Z}}(t;0) = \bar{\boldsymbol{Z}}(t)$ and $U(0) = U$. By setting MAXITER $= 0$ in the call to PHREG, we could extract U. We bypass this step, though, since we are able to compute $\hat{\boldsymbol{\theta}}$ directly, as we show later. We need the 1-step call to PHREG to compute the Schoenfeld residuals Schoenfeld (1982),

$$\begin{aligned} \boldsymbol{S}_i(\hat{\boldsymbol{\beta}}) &= \textstyle\int_0^{\tau} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t;\hat{\boldsymbol{\beta}})]\, \mathrm{d}N_i(t) \\ &= [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(X_i;\hat{\boldsymbol{\beta}})]\, \Delta_i\,, \end{aligned} \tag{12}$$

which are typically used to assess the proportionality assumption after fitting a Cox model. These can be obtained in PHREG through the RESSCH option, which creates a data set with all Schoenfeld residuals for subjects with $\Delta_i = 1$. For the previously prescribed call to PHREG with MAXITER $= 0$, we would obtain

$$\begin{aligned} \boldsymbol{S}_i = \boldsymbol{S}_i(0) &= \textstyle\int_0^{\tau} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]\, \mathrm{d}N_i(t) \\ &= [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(X_i)]\, \Delta_i\,. \end{aligned} \tag{13}$$

We combine the null Schoenfeld residuals into a matrix,

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1^T \\ \boldsymbol{S}_2^T \\ \vdots \\ \boldsymbol{S}_d^T \end{bmatrix}, \tag{14}$$

where the number of observed failures is denoted by $d = \sum_{i=1}^n \Delta_i$.

Next, we outline how to compute the matrix, $\boldsymbol{B}$. We can re-express the Schoenfeld residuals as follows:

$$\begin{aligned} \boldsymbol{B} &= \sum_{i=1}^n \textstyle\int_0^{\tau} [\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)]^{\otimes 2}\, \mathrm{d}N_i(t) \\ &= \sum_{i=1}^n [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(X_i)]^{\otimes 2}\, \Delta_i = \boldsymbol{S}^T\boldsymbol{S}\,. \end{aligned} \tag{15}$$

We now shift attention to ordinary least squares, with response vector $\boldsymbol{Y}$ and design matrix $\boldsymbol{X}$. Under OLS, the regression parameter is computed as $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{X}^T\boldsymbol{Y}$, while the sum-of-squares-and-cross-products (SSCP) matrix is given by $\boldsymbol{X}^T\boldsymbol{X}$. By setting $\boldsymbol{Y} = \boldsymbol{1}_d$ and $\boldsymbol{X} = \boldsymbol{S}$, we can compute $\boldsymbol{B}$ by fitting a linear regression (through-the-origin) model using PROC REG (via the NOINT option to remove the intercept), outputting the SSCP matrix. Note that the choice of $\boldsymbol{1}_d$ to serve as the response vector in the call to REG is arbitrary; any non-zero vector would suffice, since we do not aim to estimate the OLS parameter; but only wish to force SAS into computing an inner product which happens to equal $\boldsymbol{B}$.

It is now necessary to introduce some additional notation. First, we order the observation times: $X_{(1)} < X_{(2)} < \ldots < X_{(n)} = \tau$ and set $X_{(0)} = 0$. We then denote the gaps between successive observa-

tion times by $\mathrm{d}t_j = X_{(j)} - X_{(j-1)}$ for $j = 1, \ldots, n$. The number of observation gap times for which the i-th subject is under observation denoted by $K_i = \sum_{j=1}^{n} Y_i(X_{(j)})$, with $K = \sum_{i=1}^{n} K_i$. As alluded to previously, the Schoenfeld residuals and closely related quantities are essential to our proposed techniques. As such, we set up further notation along these lines, related to the observation gap times. Specifically, set $\boldsymbol{R}_{ij} = \boldsymbol{Z}_i - \boldsymbol{Z}(X_{(j)})$ for $j = 1, \ldots, K_i$ and set

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1^T \\ \vdots \\ \boldsymbol{R}_n^T \end{bmatrix}, \qquad \boldsymbol{R}_i = \begin{bmatrix} \boldsymbol{R}_{i1}^T \\ \vdots \\ \boldsymbol{R}_{iK_i}^T \end{bmatrix}. \tag{16}$$

Note that $\boldsymbol{S}_i = \boldsymbol{R}_{iK_i}\Delta_i$. Correspondingly, we set up observation block diagonal gap time vectors as follows,

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & & \\ & \vdots & \\ & & \boldsymbol{D}_n \end{bmatrix}, \qquad \boldsymbol{D}_i = \begin{bmatrix} \mathrm{d}t_{i1} & & \\ & \vdots & \\ & & \mathrm{d}t_{iK_i} \end{bmatrix}. \tag{17}$$

In addition, we define corresponding observed-event indicator vectors; set $\Delta_{ij} = \Delta_i I\{X_i = X_{(j)}\}$ and let

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}_1 \\ \vdots \\ \boldsymbol{\Delta}_n \end{bmatrix}, \qquad \boldsymbol{\Delta}_i = \begin{bmatrix} \boldsymbol{\Delta}_{i1} \\ \vdots \\ \boldsymbol{\Delta}_{iK_i} \end{bmatrix}.$$

Next, we depart briefly from notational considerations to describe a data set that we hereafter refer to as the "expanded" data set which has one record per subject per observation gap time at risk. That is, while the original data set contained n rows, 1 row per subject, the expanded data set will contain K rows. For example, consider a study with $n = 3$ and observed data: $(X_i, \Delta_i) = (10, 1), (5, 0), (7, 1)$ for $i = 1, 2, 3$, respectively. The original data would be given by:

idnum	obs_time	dead
1	10	1
2	5	0
3	7	1

while the expanded data set would look like:

idnum	t1	t2	dead	dt	dead1
1	0	5	0	5	1
1	5	7	0	2	1
1	7	10	1	3	1
2	0	5	0	5	1
3	0	5	0	5	1
3	5	7	1	2	1

Creating the expanded data set is straightforward, and SAS code is provided in the Appendix. The only provision is that the subject-identification numbers be sequenced $1, \ldots, n$ in the original data set. By setting the event indicator to 1 for all records (through the *dead1* variable listed above), we can obtain the $\boldsymbol{R}_i$ vectors by fitting a Cox model to the expanded data set and again getting PHREG to output the Schoenfeld-type residuals (to a data set we subsequently refer to as the expanded-Schoenfeld data set). As a quick look ahead, we would then fit a weighted least squares model to the Schoenfeld-type residuals; the output of which will be used to estimate the additive hazard model regression parameter and its variance.

We now describe algebraically the connection between estimators for the additive hazard model and the quantities we have just defined which relate to the expanded data set. The matrix $\boldsymbol{A}$ can be re-expressed as follows:

$$\begin{aligned}\boldsymbol{A} &= \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]^{\otimes 2}\, Y_i(t)\, \mathrm{d}t \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{X_{(j-1)}}^{X_{(j)}} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]^{\otimes 2}\, Y_i(t)\, \mathrm{d}t \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]^{\otimes 2}\, Y_i(X_{(j)}) \{X_{(j)} - X_{(j-1)}\} \\ &= \sum_{i=1}^{n} \boldsymbol{R}_i \boldsymbol{D}_i \boldsymbol{R}_i^T = \boldsymbol{R}^T \boldsymbol{D} \boldsymbol{R}\,. \end{aligned} \tag{18}$$

Similarly, we can re-write $\boldsymbol{U}$ as:

$$\begin{aligned}\boldsymbol{U} &= \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]\, \mathrm{d}N_i(t) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{X_{(j-1)}}^{X_{(j)}} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)]\, \mathrm{d}N_i(t) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} [\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(X_{(j)})]\, \Delta_{ij} \\ &= \sum_{i=1}^{n} \boldsymbol{R}_i^T \boldsymbol{\Delta}_i = \boldsymbol{R}^T \boldsymbol{\Delta}\,. \end{aligned} \tag{19}$$

For weighted least squares (WLS), the regression parameter and its estimated variance are given by $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}$, where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are (as in our description of OLS) the response vector and design matrix, respectively, while $\boldsymbol{W}$ is a matrix of weights. Exploiting the close relationship between the WLS set-up and (18) and (19), we can now estimate compute $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{A}$ directly using a program that can perform weighted least squares (e.g., through the WEIGHT command in PROC REG) by setting $\boldsymbol{X} = \boldsymbol{R}$, $\boldsymbol{Y} = \boldsymbol{D}^{-1}\boldsymbol{\Delta}$ and $\boldsymbol{W} = \boldsymbol{D}$. The regression parameter from this fit to the expanded-Schoenfeld data set (again using the NOINT option to remove the model intercept) would equal $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = \hat{\boldsymbol{\theta}}$. The matrix $\boldsymbol{A}$, can be obtained by outputting the weighted SSCP matrix, $(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})$, from PROC REG. The estimated variance for the regression parameter additive hazard model, $\hat{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$ can now be computed through a very brief and basic call to PROC IML, as listed in the Appendix.

We summarize the proposed fitting procedure as follows:

1. Fit a Cox model to the original data set and output Schoenfeld residuals, $\boldsymbol{S}_i$.
2. Fit OLS regression model with $\boldsymbol{S}_i$ as the covariate vector and 1 as the response; obtain B as the SSCP matrix, $\boldsymbol{X}^T \boldsymbol{X}$.
3. Create an expanded data set which contains one record per subject per observation gap time at risk.
4. Fit a Cox model to the expanded data set with $\boldsymbol{Z}_i$ as the covariate and 1 as the event indicator. Output the Schoenfeld-type residuals, $\boldsymbol{R}_i$.
5. Fit WLS regression model with R_{ij} as the covariate, $\Delta_{ij}/\mathrm{d}t_{ij}$ as the response and $\mathrm{d}t_{ij}$ as the weight; $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = \hat{\boldsymbol{\theta}}$, while the weighted SSCP matrix, $\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ equals $\boldsymbol{A}$.
6. $\hat{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$ is computed using PROC IML.

4 Application of the Proposed Techniques to the Analysis of End-stage Liver Disease Data

In this section we describe how to apply the procedures proposed in Section 3. We begin by describing the data set of interest in Section 4.1. In Section 4.2, we detail how to implement the proposed procedures using the SAS program provided in the Appendix. The results of our analysis are presented in Section 4.3.

4.1 Description of end-stage liver disease data

The data set to be analyzed was constructed by merging records from the Pennsylvania Health Care Cost Containment Council (PHC4) and the Scientific Registry for Transplant Recipients (SRTR). The SRTR data were originally collected by the Organ Procurement and Transplant Network (OPTN). We analyzed the $n = 764$ chronic liver failure patients aged ≥ 18 who were initially wait-listed for liver transplantation in Pennsylvania between September, 2001 and December, 2002. For each patient, follow-up began ($t = 0$) at the date of initial wait listing and ended at the earliest of death, liver transplantation, loss to follow-up, or the conclusion of the observation period (December 31, 2002).

We fitted the following additive hazards model,

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{\theta}_0^T \boldsymbol{Z}_i(t),$$

where $\boldsymbol{Z}_i(t)$ contained terms for Model of End-stage Liver Disease (MELD) score, age (18–29, 30–49, 50–59, $\geq$60), gender, race (Caucasian, Minority) and diagnosis (non-cholesteric cirrhosis, cholesteric cirrhosis, acute hepatic necrosis, other). The covariate of chief interest was MELD, which is the only time-dependent covariate and is used to rank patients on the waiting list. MELD is scored as an integer between 6 and 40, with higher scores generally associated with greater degree of liver failure. The higher the MELD score, the higher the estimated wait-list mortality. Under the current allocation system, chronic liver failure patients awaiting transplantation are ranked by decreasing order of MELD. Updated MELD scores are mandated according to a schedule that depends on the current value.

4.2 Fitting the semiparametric additive hazards model

We now describe how to fit the Lin and Ying (1994) additive hazards model using SAS. The code is provided in the Appendix. As described in Section 4.1, the liver data set has one time-dependent covariate, MELD. As would be the case for a SAS or R/S-PLUS user who wanted to fit a Cox model when $\boldsymbol{Z}_i(t) \neq \boldsymbol{Z}_i$, records for the input data set (liver 1) would consist of multiple records per subject. That is, each subject would generate a separate record each time their MELD changed. For example, consider a patient (e.g., IDNUM 999; $i = 999$) who is age 56 and has the following MELD history: began follow-up ($t = 0$) with MELD = 11; increased to MELD = 14 at $t = 30$; increased to MELD = 21 at $t = 45$; died at $t = 63$. This patient would be represented in the raw input data set as follows:

IDNUM	t1	t2	MELD	AGE	DEAD
999	0	30	11	56	0
999	30	45	14	56	0
999	45	63	11	56	1

The proposed procedure assumes this basic data structure for the input data in the presence of time-dependent covariates. It also assumes that the ID numbers are sequenced $1, \ldots, n$.

The algorithm we propose in Section 3 was, for ease of illustration, presented in the context of time-independent covariates; i.e., $\boldsymbol{Z}_i(t) = \boldsymbol{Z}_i$. However, techniques carry over to the time-dependent

covariate require no new ideas and carry over with very little modification. In cases where $\boldsymbol{Z}_i(t) \neq \boldsymbol{Z}_i$, assuming that the original data contains a new start/stop (t_1/t_2) record each time a subject's covariate vector changes, the creation of the expanded data set is the same. That is, each record in the original data will be expanded into multiple records, where each record spans a $(t_1, t_2]$ time interval during which $Y(t)$ and $\bar{\boldsymbol{Z}}(t)$ are constant; i.e., the same algorithm we propose for the $\boldsymbol{Z}_i(t) = \boldsymbol{Z}_i$ case.

We now outline the proposed techniques for fitting the Lin and Ying (1994) additive hazards model, referring to the sequence of steps listed at the end of Section 3 and the SAS code provided in the Appendix.

As a preliminary, the program first counts the number of subjects in the data set using PROC MEANS, then saves the result as a global (macro) variable, $\&n$.

In Step 1, we fit a Cox model to the input data, stopping the Newton-Raphson procedure at the first iteration (MAXITER = 0) in order to obtain the Schoenfeld residuals given in (13).

In Step 2, we fit a no-intercept (NOINT) ordinary least squares regression using PROC REG which has the Schoenfeld residuals as the covariate and the event indicator as the response. Since PROC PHREG only keeps non-zero Schoenfeld residuals (i.e., corresponding to events), the response vector in the OLS model is actually a vector of 1s. The SSCP matrix, $\boldsymbol{X}^T\boldsymbol{X}$, equals $\boldsymbol{S}^T\boldsymbol{S}$, where $\boldsymbol{S}$ is the matrix of Schoenfeld residuals defined in (14). The SSCP matrix equals $\boldsymbol{B}$ from (5), as indicated in (15). Note that the response vector in the linear model is arbitrary in this case, since OLS is only used to compute the SSCP matrix.

In Step 3, we create an expanded data set which splits each row of the input data into sub-intervals, $(t_1^*, t_2^*]$, within which $\boldsymbol{Z}_i(t)$, $Y(t)$ and hence $\bar{\boldsymbol{Z}}(t)$ are constant. That is, $Y(t)$ and $\bar{\boldsymbol{Z}}(t)$ will change whenever there is a death or censoring event, while $\boldsymbol{Z}_i(t)$ changes whenever one of the time-dependent covariates for subject i changes. We first create a data set which contains all censoring, event and covariate change times; this data set will merely contain the t_2 variable from the raw input data set.

In Step 4, we compute the $\boldsymbol{R_i}$ quantities related to $\boldsymbol{A}$ and given in (18). It is necessary to compute

$$\int_{t_1^*}^{t_2^*} \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}\, Y_i(t)\, \mathrm{d}t = Y_i(t_1^*)\, \{\boldsymbol{Z}_i(t_1^*) - \bar{\boldsymbol{Z}}(t_1^*)\}\, \{t_2^* - t_1^*\},$$

for each $(t_1^*, t_2^*]$ subinterval for each subject. The equality holds since the expanded data set is structured such that the risk sets and covariates are constant across each of the $(t_1^*, t_2^*]$ subintervals. We fit a Cox model to the expanded data, with the death indicator (dead1) set to 1 for all records. The desired quantities from this model are given by the Schoenfeld residuals from this model in (16).

In Step 5, we compute the regression parameter estimator, $\hat{\boldsymbol{\theta}}$ and the matrix $\boldsymbol{A}$ using weighted least squares through the origin. Specifically, the covariate equals the Schoenfeld residuals from Step 4; the response equals a death indicator, scaled by the length of the subinterval; while the weight equals the subinterval length. The WLS regression parameter equals $\hat{\boldsymbol{\theta}}$, while the weighted SSCP matrix equals $\boldsymbol{A}$.

Step 6 involves various simple matrix calculations which complete the proposed procedure.

4.3 Analysis of end-stage liver disease data

We applied the Lin and Ying (1994) additive hazards model to a data set of patients on the waiting list for liver transplantation. We began by coding MELD in categories, with the estimated covariate-adjusted risk (hazard) differences by MELD category listed Table 1. Compared to patients with MELD score between 15 and 19, patients in the 6–9 and 10–14 MELD categories have significantly reduced risk of death, while the risk of death for patients with MELD scores on 25–29 and 30–40 levels is significantly increased. As indicated in the table, $\hat{\boldsymbol{\theta}}$ has been multiplied by 10^4 and hence would be interpreted as per 10 000 patients.

Due to the monotone nature of the MELD effect, which is evident from Table 1, we sought to fit a more parsimonious model. After some experimentation, it was found that the mortality hazard increased quadratically with MELD. In Figure 1, the hazard difference associated with MELD is plotted

Table 1 Covariate-adjusted mortality hazard difference by MELD category.

k	$m_i(t)$	$\hat{\boldsymbol{\theta}}_{mk} \times 10^4$	(95% CI)	p
1	6–9	−4.7	(−7.9, −1.7)	<0.005
2	10–14	−4.5	(−7.6, −1.5)	<0.005
3	15–19	0	–	–
4	20–24	3.5	(−3.3, 10.3)	0.32
5	25–29	19.5	(0.1, 38.9)	0.05
6	30–40	77.1	(44.8, 109.4)	<0.005

for both the original model,

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{\theta}_1^T \boldsymbol{Z}_{i1} + \boldsymbol{\theta}_2^T \boldsymbol{Z}_{i2}(t)\,, \tag{35}$$

which assumes a step-function MELD effect (Table 1), and the revised model,

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{\theta}_1^T \boldsymbol{Z}_{i1} + \theta_3 m_i(t) + \theta_4 m_i(t)^2\,, \tag{36}$$

which assumes a quadratic MELD effect. In both models (20) and (21), $\boldsymbol{Z}_{i1}$ contains elements corresponding to the adjustment covariates (age, gender, race, diagnosis), while $m_i(t)$ is the MELD score at time t. In model (20),

$$\boldsymbol{Z}_{i2}(t) = \begin{bmatrix} I\{6 \le m_i(t) \le 9\} \\ I\{10 \le m_i(t) \le 14\} \\ I\{20 \le m_i(t) \le 24\} \\ I\{25 \le m_i(t) \le 29\} \\ I\{30 \le m_i(t) \le 40\} \end{bmatrix}, \qquad \boldsymbol{\theta}_2 = \begin{bmatrix} \theta_{m1} \\ \theta_{m2} \\ \theta_{m4} \\ \theta_{m5} \\ \theta_{m6} \end{bmatrix},$$

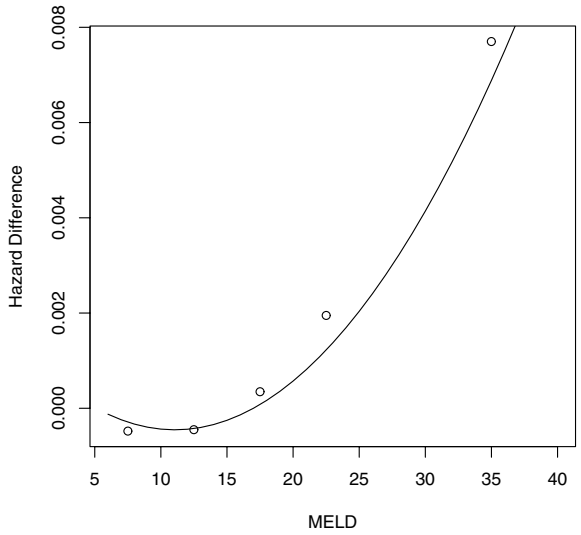

Figure 1 Covariate-adjusted hazard difference associated with MELD based on step-function ('o') and quadratic ('−') models described by Eqs. (20) and (21), respectively.

as implied by Table 1. Based on Figure 1, modelling the MELD effect through a quadratic appears quite reasonable.

5 Discussion

In this report, we develop techniques for fitting the Lin and Ying (1994) additive hazards model. We describe the connections between the additive hazards, proportional hazards and least squares regression procedures and demonstrate that, after modifying the original data, the Lin and Ying (1994) model can be fitted through basic calls to Cox and least squares regression procedures. We then apply the additive hazards model to a data set consisting of liver failure patients awaiting transplantation.

The contribution of our report to the literature is two-fold. First, we describe in detail the relationships between the Cox (1972) proportional hazards and Lin and Ying (1994) semiparametric additive hazards models. Second, we show how standard software can be used to fit the Lin and Ying (1994) model. The proposed computational techniques save the practitioner from relying on C/C++ code or more extensive IML code. In practice, these perhaps substantial savings in development time would need to be traded off with computing time due to the data management steps in the proposed procedure. This is particularly true for very large data sets or even perhaps moderate-sized data sets with time-dependent covariates since, in such cases, the 'expanded' version of the data could be quite large and hence cumbersome.

The Cox (1972) proportional hazards model is the predominant regression model for survival analysis in biomedical studies. In addition to its wide applicability, at least part of its appeal is the lack of computationally convenient alternatives; particularly if one restricts attention to semiparametric approaches. As we demonstrate, it is possible to fit this model after some straightforward data manipulation using PROC PHREG, PROC REG and some basic computations in PROC IML. Minimal knowledge of SAS-IML (and, none beyond what is listed in this report) is required. Given that the Lin and Ying (1994) model can be fitted using standard software, the model should be considered as an alternative to the Cox model by practitioners in settings where poor fit of the Cox model is uncovered or where additive covariate effects are suggested based on previous literature or preliminary descriptive analysis.

SAS (v9.1.3) was the statistical computing language used in this report. S-PLUS and R can compute Schoenfeld residuals through the *coxph* function and can carry out ordinary and weighted least squares estimation through the *lm* function. Both the *coxph* and *lm* functions can return all required elements for the $\hat{\boldsymbol{\theta}}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ vectors and matrices. As such, S-PLUS/R could be used to fit the additive hazards model using algorithms similar to those proposed in this report.

We propose techniques for fitting the semiparametric additive hazard model of Lin and Ying (1994). This report has focused on the Lin and Ying (1994) model since, among the hazard regression models assuming additive covariates effects, it would appear to be the most closely related to the Cox (1972) model. There are several different versions of the of the additive hazards model, including the non-parametric model of Aalen (1989). The Aalen (1989) nonparametric additive hazards model also features least squares type closed form estimators and, for this reason, procedures similar to those derived in this report could be applied. However, it should be noted that there are several pertinent software packages already available to fit the Aalen (1989) model. Examples include the S-PLUS *survival* package by Therneau (http://mayoresearch.mayo.edu/mayo/research/biostat/splusfuctions.cfm); the R/S-PLUS *addreg* package by Weedon-Fekjaer (http://www.med.uio.no/imb/stat/addreg/); the R *timereg* package by Scheike (http://staff.pubhealth.ku.dk/ ts/timereg.html); and a SAS macro by Howell and Klein (1997).

The numerical techniques proposed in this report illustrate computational similarities between the semiparametric proportional hazards (Cox, 1972) and additive hazards (Lin and Ying (1994) models. For the additive model, Lin and Ying proposed zero-mean estimating functions, analogous to the generalized estimating equations (GEE) approach (Liang and Zeger (1986) for longitudinal data. The score equations for the proportional hazards model, while derived through partial likelihood (Cox

(1975)), can also be expressed as the solutions to zero-mean estimating functions. That is, when $dM_i(t;\boldsymbol{\beta}) = dN_i(t) - Y_i(t)\exp\{\boldsymbol{\beta}^T\mathbf{Z}_i(t)\}\,d\Lambda_0(t)$ and $\int_0^\tau \mathbf{Z}_i\,dM_i(t;\boldsymbol{\beta})$ are used as the basis of a GEE-type procedure, the maximum partial likelihood estimators are obtained. This is evident from the literature in various places, including a related development in the recurrent event setting by Schaubel and Cai (2005). Thus, computational similarity between the semiparametric additive and multiplicative models is less surprising if one considers that the estimation procedure for either can be derived through an estimating equations approach.

In practice, investigators may have valid reasons for preferring the Cox model over the additive hazard model, or vice versa. Given that both the additive and multiplicative hazard models can be fitted easily, practitioners may prefer to fit both and base their inference on the model with the superior fit; particularly in the absence of any a priori preference for one model over the other. However, few methods are available to compare the two models in a global sense. Lin and Ying (1995) proposed a general additive-multiplicative model, which subsumes the Lin and Ying (1994) and Cox (1972) models. A related approach was developed by Martinussen and Scheike (2002). More research directed at comparing the use of multiplicative and additive hazard models is needed. Reports comparing the two approaches do exist (e.g., Klein, 2006), but are rare. Yin and Cai (2004) proposed an additive hazard model for multivariate failure time data and, in their real-data application, the estimated survival curves based on their proposed additive model were notably different than those based on the multiplicative counterpart, the Wei, Lin and Weissfeld (1989) model. It is not known whether such lack of agreement between the multiplicative and additive approaches is the rule or the exception and whether one model tends to out-perform the other with respect to goodness-of-fit and/or prediction.

Acknowledgements The authors thank the Scientific Registry of Transplant Recipients (SRTR), the Organ Procurement and Transplantation Network (OPTN), and the Pennsylvania Health Care and Cost and Containment Council (PHC4) for access to the linkage of their databases. This research was supported in part by National Institutes of Health grant R01 DK-70869 to the first author. The SRTR is funded by contract number 231-00-0116 from the Health Resources and Services Administration, U.S. Department of Health and Human Services. The views expressed herein are those of the authors and not necessarily those of the U.S. Government. This study was approved by HRSA's SRTR project officer. HRSA has determined that this study satisfies the criteria for the IRB exemption described in the "Public Benefit and Service Program" provisions of 45 CFR 46.101(b)(5) and HRSA Circular 03.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. in: N. Klonecki, A. Kosek, and J. Rosinski (eds.) *1980 Lecture Notes in Statistics*. Springer: New York.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, 1, The Design and Analysis of Case-Control Studies*. IARC, Lyon.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, 2, The design and Analysis of Cohort Studies*. IARC, Lyon.

Buckley, J. D. (1984). Additive and multiplicative models for relative survival rates. *Biometrics* **40**, 51–62.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Howell, A. and Klein, J. P. (1997). A SAS macro for the additive regression hazards model. *1997 Proceedings of the Statistical Computing Section*. American Statistical Association, 282–287.

Huffer, F. W. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association* **86**, 114–129.

Klein, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine* **25**, 1015–1034.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**, 13–22.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting process. *The Annals of Statistics* **23**, 1712–1734.

Lin, D. Y. and Ying, Z. (1997). Additive regression models for survival data. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Springer, New York. 185–198.

Martinussen, T. and Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283–298.

Schaubel, D. E. and Cai, J. (2005). Semiparametric methods for clustered recurrent event data. *Lifetime Data Analysis* **11**, 405–425.

Schoenfeld, D. (1982). Residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.

Thomas, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. *Modern Statistical Methods in Chronic Disease Epidemiology*, Wiley, New York.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91**, 801–818.

