The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.403
25*(3), 403–413*

Further Results on Piecewise Constant Hazard Functions in Aalen's Additive Risk Model

Daiho Uhm[1] · Sunghae Jun[2]

[1]Department of Statistics, Oklahoma State University; [2]Department of Statistics, Cheongju University

(Received February 4, 2012; Revised March 6, 2012; Accepted June 4, 2012)

Abstract

The modifications suggested in Uhm *et al.* (2011) are studied using a partly parametric version of Aalen's additive risk model. A follow-up time period is partitioned into intervals, and hazard functions are estimated as a piecewise constant in each interval. A maximum likelihood estimator by iteratively reweighted least squares and variance estimates are suggested based on the model as well as evaluated by simulations using mean square error and a coverage probability, respectively. In conclusion the modifications are needed when there are a small number of uncensored deaths in an interval to estimate the piecewise constant hazard function.

Keywords: Aalen's additive risk model, piecewise constant, hazard function, weighted least square, survival analysis.

1. Introduction

In a survival analysis Aalen's additive risk model (1980) is a useful alternative regression model to Cox's proportional hazard model (1972). In the Aalen's model the effects of covariates might vary on time. Aalen (1989) suggested a non-parametric estimation for the additive risk model, and Huffer and McKeague (1991) introduced a weighted least squares (WLS) estimator for the cumulative hazard function with constant covariates. They also suggested WLS estimators for a grouped data case and variance estimates. For the grouped data case a follow-up time period $[0, T]$ is partitioned into d intervals, and the hazard functions are estimated as a piecewise constant in each interval. The WLS estimator was modified by smoothing ordinary least squares (OLS) estimates over the past using a kernel function since weights might blow up or be negative with too samll or negative estimates of the hazard function. McKeague and Sasieni (1994) studied a partly parametric version of Aalen's risk model, which was defined as

$$h(t|x, z) = \alpha(t)'x + \beta' z, \tag{1.1}$$

where x and z are p- and q-dimensional covariate vectors, $\alpha(t)$ is the time-varying hazard function, and β is the unknown constant hazard. The first component of x might be 1 for a baseline hazard.

[2]Corresponding author: Associate Professor, Department of Statistics, Cheongju University, Chungbuk 360-764, Korea. E-mail: shjun@cju.ac.kr

They suggested semiparametric estimators for the cumulative hazard function, $A(\cdot) = \int_0^{\cdot} \alpha(s)ds$ and β. If β is known, the least squares estimator for the cumulative hazard function could be given. Also $\alpha(\cdot)$ is given, estimate of β is obtained by maximum likelihood. To avoid unstable weights, they also used a predictable kernel smoother and modified the hazard function. Lin and Ying (1994) studied Aalen's model, and suggested a simple procedure to estimate the regression coefficients with high efficiency. Recently, Uhm *et al.* (2011) studied the WLS estimator for Aalen's model which allowed a very flexible handling of covariates, and they extended the grouped data version of Huffer and McKeague's (1991) estimator. They obtained maximum likelihood estimators (MLE) for the cumulative hazard function by iterative reweighted least squares (IRLS), and they modified the hazard functions over neighboring intervals.

The previous studies mentioned that the estimates were not stable when there were a small number of uncensored deaths in each interval. With negative or very small hazard estimates the weights should be negative and blow-up, and the IRLS estimates might not converge. In this article the modifications of smoothing and updating in Uhm *et al.* (2011) are studied based on the model (1.1). The modifications of smoothing and updating are compared based on various conditions, and the effects of the modifications are discussed. The smoothing window size and its relationship with updating modification are also studied depending on sample size. Guidelines in modifications are suggested for application using the additive risk model.

In Section 2 the WLS estimator is suggested. In addition, it discusses the weight modifications and variance estimates. The modifications of weights and the variance estimations are evaluated using a mean square error (MSE) and a coverage probability by simulations in Section 3. An application with a lung cancer data is reported in Section 4. Results and implications are concluded and summarized in Section 5.

2. Piecewise Constant Hazard Function

Uhm *et al.* (2011) extended the grouped data case of Huffer and McKeague's model (1991), and studied the WLS estimators for the Aalen's additive risk model. In this section their estimation and modifications are introduced based on the model in (1.1). The follow-up time period $[0, 1]$ is partitioned into d intervals.

2.1. Estimation

For individual i, the piecewise constant hazard function in interval r $(= 1, 2, \ldots, d)$ is defined as

$$h_{ir} = \alpha_r' x_i + \beta' z_i, \tag{2.1}$$

where α_r is a p-dimensional hazard vector for the time-varying effect in interval r, and β is a q-dimensional constant hazard. Let θ be a parameter vector, which is defined as

$$\theta = \text{vec}(\alpha_1, \alpha_2, \ldots, \alpha_d, \beta).$$

The log-likelihood function for the right-censored data is given by

$$l(\theta) = \sum_{i=1}^{n} \sum_{r=1}^{d} \delta_{ir} \log h_{ir} - \sum_{i=1}^{n} \sum_{r=1}^{d} T_{ir} h_{ir},$$

where δ_{ir} is the indicator that individual i undergoes uncensored death in interval r, and T_{ir} is the total time that individual i is at risk in interval r. Differentiating the log-likelihood function with respect to θ,

$$\begin{aligned}\ell(\theta) &= \frac{\partial l}{\partial \theta} \\ &= \sum_{i=1}^{n}\sum_{r=1}^{d} \delta_{ir}\frac{1}{h_{ir}}\frac{\partial h_{ir}}{\partial \theta} - \sum_{i=1}^{n}\sum_{r=1}^{d} T_{ir}\frac{\partial h_{ir}}{\partial \theta}. \end{aligned} \tag{2.2}$$

Let $\Psi_{ir} = \partial h_{ir}/\partial\theta$, and define a weight as $w_{ir} = 1/h_{ir}$. Then the Equation (2.2) is

$$\ell(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{d} \delta_{ir} w_{ir}\Psi_{ir} - \sum_{i=1}^{n}\sum_{r=1}^{d} T_{ir}\Psi_{ir}. \tag{2.3}$$

Since $\Psi_{ir} = \text{vec}(\mathbf{0}_1, \ldots, \mathbf{0}_{r-1}, x_i, \mathbf{0}_{r+1}, \ldots, \mathbf{0}_d, z_i)$, where $\mathbf{0}$ is a p-dimensional zero-vector, the piecewise hazard function in (2.1) could be rewrite as $h_{ir} = \Psi'_{ir}\theta$. By the definition of weight, $w_{ir}h_{ir} = w_{ir}\Psi'_{ir}\theta = 1$. Now the Equation (2.3) is

$$\ell(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{d} \delta_{ir} w_{ir}\Psi_{ir} - \sum_{i=1}^{n}\sum_{r=1}^{d} T_{ir}\Psi_{ir}(w_{ir}\Psi'_{ir}\theta)\ . \tag{2.4}$$

Setting the Equation (2.4) be equal to zero, a WLS estimator is obtained as $\tilde{\theta} = D^{-1}C$, where

$$C = C(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{d} w_{ir}\delta_{ir}\Psi_{ir} \quad \text{and} \quad D = D(\theta) = \sum_{i=1}^{n}\sum_{r=1}^{d} w_{ir}T_{ir}\Psi_{ir}\Psi'_{ir}\ .$$

The MLE is obtained by iteratively reweighted least squares (IRLS).

2.2. New weights

In the IRLS, the weights are modified to obtain better estimates. When there are a small number of uncensored deaths in an interval, the WLS estimate might be unstable. The estimates of weights could blow-up with a very small estimate of h_{ir}, and the estimate of h_{ir} might also be negative. Two independent procedures to modify the piecewise constant hazard are suggested in each interval. One is the averaging of $\tilde{h}_{ir} = \Psi'_{ir}\tilde{\theta}$ over neighboring intervals depending on the minimum number c of uncensored deaths. Let $S_r = \sum_{i=1}^{n}\delta_{ir}$ be the number of uncensored deaths in interval r. The smoothing window size $s = s(r)$ is given by the smallest s (≥ 0) such that $\sum_{j=r-s}^{r+s} S_j \geq c$, where $S_j = 0, j \notin \{1, 2, \ldots, d\}$. The smoothing hazard function is defined as

$$\tilde{h}_{ir}^{sm} = \frac{\sum_{j=r-s}^{r+s} \tilde{h}_{ij}}{\sum_{j=r-s}^{r+s} I_{(1\leq j\leq d)}}, \tag{2.5}$$

where $\tilde{h}_{ij} = 0, j \notin \{1, 2, \ldots, d\}$. The smoothing hazard function is also modified to avoid negative and too small hazard by

$$\hat{h}_{ir} = \max\left\{\tilde{h}_{ir}^{sm}, \varepsilon\overline{\tilde{h}_{r}^{sm}}\right\}, \tag{2.6}$$

where $\overline{\tilde{h}_{r}^{sm}}$ is the mean of $\tilde{h}_{ir}^{sm}$ for all individuals at risk in interval r, and ε is a constant to define updating rate on the mean, $\overline{\tilde{h}_{r}^{sm}}$.

2.3. Estimates of variance

Uhm *et al.* (2011) suggested three estimations of variance which were adopted from Huffer and McKeague (1991). In Equation (2.4) we obtained

$$\begin{aligned}\ell(\theta) &= C - D\theta,\\ \ell\big(\tilde{\theta}\big) &= C - D\tilde{\theta} = 0.\end{aligned}$$

It is rewritten that

$$\begin{aligned}\ell(\theta) = \ell(\theta) - \ell\big(\tilde{\theta}\big) &= D\big(\tilde{\theta} - \theta\big),\\ \mathrm{Var}(\ell(\theta)) &= D\mathrm{Var}\big(\tilde{\theta}\big)D',\end{aligned}$$

where D is regarded as a constant. Approximately the variance of $\hat{\theta}$ is defined as

$$\mathrm{Var}\big(\hat{\theta}\big) \approx D^{-1}\mathrm{Var}(\ell(\theta))D^{-1},$$

where D is a symmetric matrix. By martingale properties (see Appendix in Uhm *et al.* (2011)) the variance of $\ell(\theta)$ is

$$\begin{aligned}\mathrm{Var}(\ell(\theta)) &= \mathrm{Var}\left(\sum_{i=1}^{n}\sum_{r=1}^{d} w_{ir}\Psi_{ir}\left(\delta_{ir} - T_{ir}\Psi_{ir}'\theta\right)\right)\\ &= \sum_{i=1}^{n}\sum_{r=1}^{d} w_{ir}^2\Psi_{ir}\Psi_{ir}'\mathrm{Var}\left(\delta_{ir} - T_{ir}\Psi_{ir}'\theta\right)\\ &= \sum_{i=1}^{n}\sum_{r=1}^{d} w_{ir}^2\Psi_{ir}\Psi_{ir}'\mathrm{E}\left(T_{ir}\Psi_{ir}'\theta\right). \qquad (2.7)\end{aligned}$$

They suggested three estimates of $\mathrm{Var}(\ell(\theta))$ by

$$\begin{aligned}\hat{D} &= \sum_{i=1}^{n}\sum_{r=1}^{d} \hat{w}_{ir}T_{ir}\Psi_{ir}\Psi_{ir}',\\ \hat{E} &= \sum_{i=1}^{n}\sum_{r=1}^{d} \hat{w}_{ir}^2\delta_{ir}\Psi_{ir}\Psi_{ir}',\\ \hat{F} &= \sum_{i=1}^{n}\sum_{r=1}^{d} \hat{w}_{ir}^2T_{ir}\left(\Psi_{ir}'\hat{\theta}\right)\Psi_{ir}\Psi_{ir}'.\end{aligned}$$

Since the piecewise hazard function was defined as $h_{ir} = \Psi_{ir}'\theta$, the expectation in (2.7) could be estimated by $T_{ir}\hat{h}_{ir} = T_{ir}/\hat{w}_{ir}$ after estimating the weights by $\hat{w}_{ir} = 1/\hat{h}_{ir}$. Then the estimate of $\hat{D}$ is suggested. Uhm *et al.* (2011) showed that

$$\mathrm{E}\left(\delta_{ir} - T_{ir}\Psi_{ir}'\theta\right) = 0$$

in their Appendix, now we know that

$$\mathrm{E}\left(\delta_{ir}\right) = \mathrm{E}\left(T_{ir}\Psi_{ir}'\theta\right). \qquad (2.8)$$

Therefore δ_{ir} is replaced with $\mathrm{E}(T_{ir}\Psi_{ir}'\theta)$ in (2.7) as an unbiased estimator for $\mathrm{E}(T_{ir}\Psi_{ir}'\theta)$, and the estimate of $\hat{E}$ is obtained . Finally, $\mathrm{E}(T_{ir}\Psi_{ir}'\theta)$ is estimated by $T_{ir}\Psi_{ir}'\hat{\theta}$ leading to $\hat{F}$.

Table 3.1. Comparison of MSE's with $n = 1{,}000$ and $\varepsilon = 0.15$ depending on c

c	Intervals	Only Smoothing				Smoothing+Updating			
		0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0	0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0
	Baseline	36.72	178.66	434.38	626.61	2e-04	0.0018	0.0060	0.0161
100	$A(t)$	114.01	327.20	710.00	675.80	8e-04	0.0057	0.0186	0.0563
	$B(t)$	21.81	348.98	1068.75	2181.13	4e-04	0.0060	0.0183	0.0374
	Baseline	0.3367	0.3677	0.4143	0.4270	2e-04	0.0018	0.0060	0.0160
200	$A(t)$	1.1823	1.4405	1.5150	6.0675	8e-04	0.0057	0.0185	0.0560
	$B(t)$	0.0174	0.2782	0.8519	1.7386	4e-04	0.0060	0.0182	0.0372
	Baseline	0.0874	0.0038	0.0126	0.0231	2e-04	0.0018	0.0060	0.0160
300	$A(t)$	0.1503	0.0144	0.0404	0.0779	8e-04	0.0057	0.0185	0.0560
	$B(t)$	0.0007	0.0110	0.0336	0.0686	4e-04	0.0060	0.0183	0.0373
	Baseline	2e-04	0.0018	0.0060	0.0161	2e-04	0.0018	0.0060	0.0161
400	$A(t)$	8e-04	0.0057	0.0185	0.0561	8e-04	0.0057	0.0185	0.0561
	$B(t)$	4e-04	0.0060	0.0184	0.0375	4e-04	0.0060	0.0184	0.0375
	Baseline	2e-04	0.0018	0.0061	0.0162	2e-04	0.0018	0.0061	0.0162
500	$A(t)$	8e-04	0.0057	0.0185	0.0562	8e-04	0.0057	0.0185	0.0562
	$B(t)$	4e-04	0.0061	0.0185	0.0378	4e-04	0.0061	0.0185	0.0378

3. Simulations

In this section the effects of modifications are evaluated and the variance estimates are compared by simulating the model (3.1). Its weights are modified in the two ways. One is the smoothing hazard function to take the average over neighboring intervals in (2.5), and the other updates the hazard functions which were very small or negative using the mean of the individual hazard rates in (2.6). The follow-up time period $[0, 1]$ is partitioned into $d = 10$ intervals that have even lengths. It is easy to evaluate the two modifications in each interval using MSE, and compare the variance estimates by coverage probabilities.

Let X and Z be the constant covariates. The simulation model is given as

$$h(t) = 2t + e^{-t}X + 2Z. \tag{3.1}$$

Setting $x = (1, X)'$ and $\beta = 2$, it is same as the semiparametric model of McKeague and Sasieni in (1.1). Generate an identically independent random sample from exponential distribution with mean $1/2$ for the covariates of X and Z, and censoring times independently from the exponential distribution with mean $10/3$. Simulate 10,000 times to estimate for the cumulative hazard functions, and calculate their MSE's and coverage probabilities from the cumulative simulation model:

$$H(t) = t^2 + \left(1 - e^{-t}\right) X + 2tZ,$$

where t^2 is a baseline hazard, and the cumulative hazard functions are $A(t) = 1 - e^{-t}$ and $B(t) = 2t$. The WLS estimates are calculated with four iterations.

3.1. Evaluation of weight modifications

In Table 3.1 the MSE's for cumulative hazard functions are suggested depending on the smoothing constant c in (2.5) at setting $\varepsilon = 0.15$ with $n = 1{,}000$. When the smoothing is applied only, the MSE's using $c = 400$ and 500 are better than smaller c's. All MSE's increase on time, except for $A(t)$ with $c = 100$, and baseline and $A(t)$ with $c = 300$. Applying smoothing and updating

Table 3.2. Comparison of MSE's with $n = 1{,}000$ and $c = 400$ depending on ε

ε	Intervals	Only Updating				Smoothing+Updating			
		0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0	0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0
0.15	Baseline	2e-04	0.0018	0.0061	0.0162	2e-04	0.0018	0.0060	0.0161
	$A(t)$	8e-04	0.0057	0.0188	0.0572	8e-04	0.0057	0.0185	0.0561
	$B(t)$	4e-04	0.0060	0.0183	0.0374	4e-04	0.0060	0.0184	0.0375
0.25	Baseline	2e-04	0.0018	0.0061	0.0162	2e-04	0.0018	0.0061	0.0161
	$A(t)$	8e-04	0.0057	0.0188	0.0571	8e-04	0.0057	0.0185	0.0561
	$B(t)$	4e-04	0.0060	0.0184	0.0376	4e-04	0.0060	0.0184	0.0376
0.35	Baseline	2e-04	0.0018	0.0061	0.0162	2e-04	0.0018	0.0061	0.0161
	$A(t)$	8e-04	0.0057	0.0188	0.0570	8e-04	0.0057	0.0185	0.0561
	$B(t)$	4e-04	0.0060	0.0185	0.0377	4e-04	0.0060	0.0185	0.0378

Table 3.3. Comparison of MSE's with $n = 100$ and $\varepsilon = 0.15$

c	Intervals	Only Smoothing				Smoothing+Updating			
		0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0	0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0
20	Baseline	2146	1124	2096	5096	0.0022	0.0202	0.0676	0.2116
	$A(t)$	16316	8801	15632	24025	0.0084	0.0643	0.2439	1.1019
	$B(t)$	1302	20832	63797	130197	0.0040	0.0648	0.1984	0.4050
30	Baseline	145	1138	1361	12440	0.0022	0.0200	0.0664	0.2077
	$A(t)$	87	264.2	355	6309	0.0086	0.0644	0.2440	1.0838
	$B(t)$	878	14049	43026	87809	0.0040	0.0643	0.1970	0.4021
40	Baseline	3.4	37.0	46.4	182.0	0.0023	0.0199	0.0666	0.2071
	$A(t)$	5.6	58.6	92.7	966.6	0.0088	0.0646	0.2459	1.0886
	$B(t)$	2.8	44.3	135.5	276.6	0.0040	0.0647	0.1982	0.4044
50	Baseline	140.2	90.1	85.0	99.7	0.0024	0.0199	0.0664	0.2068
	$A(t)$	741.0	550.1	560.7	685.0	0.0090	0.0649	0.2452	1.0904
	$B(t)$	4.0	64.4	197.2	402.5	0.0041	0.0651	0.1993	0.4068
60	Baseline	0.7	4.6	187.2	205.9	0.0025	0.0202	0.0667	0.2073
	$A(t)$	9.8	187.4	1103.7	1097.5	0.0093	0.0658	0.2444	1.0927
	$B(t)$	24.3	388.9	1190.9	2430.4	0.0041	0.0659	0.2020	0.4122

in (2.6) together, all cases suggested almost the same MSE's which are all increasing on time. Using the smaller constant $c = 100$, 200, and 300 in smoothing, the updating procedure works effectively. Using $c = 400$ and 500, the updating procedure does not provide any more improvement in MSE. The parameters could change little since they already had enough small MSE's in the smoothing procedure only. Using only smoothing, all MSE's with $c = 400$ are best, and there are all smallest MSE's with $c = 200$ using both smoothing and updating. In Table 3.2 the MSE's are compared depending on ε in (2.6) using $n = 1{,}000$ and $c = 400$. Over all there are almost no differences among the values in ε. When there are enough uncensored deaths in each interval, using an adequate constant c provides good estimates even with only smoothing estimates over neighbor intervals.

In Table 3.3 the MSE's are given with a small sample size, $n = 100$ setting $\varepsilon = 0.15$ depending constant c. The MSE's are larger and unstable with only the smoothing procedure. With $c = 30$ MSE's have a range between 87.2 and 87808.5, and they are all increasing on time. However they are not increasing always with $c = 50$ for a baseline and $A(t)$. For $A(t)$ it has the smallest MSE of 5.6 in interval between 0.0–0.1 with $c = 40$. When there are even small sample size, the MSE's are almost the same and stable with any constant c after applying the two modifications of smoothing

Table 3.4. Comparison of MSE's with $n = 100$ and $c = 60$

ε	Intervals	Only Updating				Smoothing+Updating			
		0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0	0.0–0.1	0.3–0.4	0.6–0.7	0.9–1.0
	Baseline	0.74	4.61	187.17	205.86	0.0025	0.0202	0.0667	0.2073
0.15	$A(t)$	9.77	187.43	1103.74	1097.53	0.0093	0.0658	0.2444	1.0927
	$B(t)$	24.30	388.86	1190.88	2430.37	0.0041	0.0659	0.2020	0.4122
	Baseline	0.74	4.61	187.17	205.86	0.0025	0.0203	0.0665	0.2076
0.25	$A(t)$	9.77	187.43	1103.74	1097.53	0.0092	0.0657	0.2420	1.0957
	$B(t)$	24.30	388.86	1190.88	2430.37	0.0041	0.0661	0.2023	0.4129
	Baseline	0.74	4.61	187.17	205.86	0.0025	0.0206	0.0668	0.2086
0.35	$A(t)$	9.77	187.43	1103.74	1097.53	0.0093	0.0659	0.2411	1.1052
	$B(t)$	24.30	388.86	1190.88	2430.37	0.0042	0.0665	0.2037	0.4158

Table 3.5. Range of weights

n	c	None	Only Smoothing	Smoothing+Updating
	20		(0.1635, 5.5582)	(0.1635, 4.5493)
	30		(0.1635, 1.7643)	(0.1635, 1.7643)
100	40	(−187.7944, 14.3184)	(0.1993, 1.5122)	(0.1993, 1.5122)
	50		(0.2012, 1.3742)	(0.2012, 1.3742)
	60		(0.2012, 1.3808)	(0.2012, 1.3808)
	100		(−248.5916, 195.5474)	(0.0881, 5.1871)
	200		(−462.2076, 168.3407)	(0.0890, 4.5063)
1,000	300	(−248.5916, 195.5474)	(0.0900, 5.4520)	(0.0900, 4.1878)
	400		(0.0897, 4.7216)	(0.0897, 3.9824)
	500		(0.0897, 2.4629)	(0.0897, 2.4629)

Notes: Range = (Min, Max), $\varepsilon = 0.15$, and seed = 5264 in R

and updating. When the number of uncensored deaths is not enough large in intervals, the two procedures should be applied to modify the estimates. Table 3.4 shows that the choice of ε does not effect the MSE with small sample size, $n = 100$, like as with $n = 1{,}000$.

In Table 3.5 the range of weights after four iterations are given depending on sample size and the smoothing constant c. A sample set with $n = 100$ or $n = 1{,}000$ is used setting seed be 5264 in R 2.12.1. When no modifications are applied, the weights are too big or even negative with $n = 100$ and $n = 1{,}000$. However only after the smoothing modification, the weights are more stable except with $n = 100$ at $c = 20$, and $n = 1{,}000$ at $c = 100$ and 200. Additionally updating, the ranges of weights are smaller with a bigger constant c.

3.2. Confidence intervals

A coverage probability of a 95% confidence interval for the cumulative hazard function is considered to evaluate the variance estimations. In Figure 5.1 the coverage probabilities are displayed for the three parameters in each column depending on the smoothing constant c (in each row) with $n = 1{,}000$, $\varepsilon = 0.15$ and after both modifications. The coverage probabilities in $d = 10$ intervals are connected by line segments for convenient viewing. The coverage probabilities for estimations of E and F are lower than 95% for the baseline hazard and $A(t)$, and they are getting closer to 95% with the increasing of the constant c. For $B(t)$ they are almost same (in $\pm 0.002\%$) as 95%. However the coverage probabilities of D are decreasing on time for baseline hazard and $A(t)$, and they are worse than the other two estimations for $B(t)$ except with $c = 100$.

Table 4.1. Output from R using *aareg* function for Aalen's additive regression model

	slope	coef	se(coef)	z	p
Intercept	4.93e-03	0.005690	4.71e-03	1.21	0.22700
age	9.21e-05	0.000124	6.91e-05	1.80	0.07220
sex	−3.22e-03	−0.003980	1.22e-03	−3.25	0.00114

In Figure 5.2 the coverage probabilities based on $n = 100$ are all worse than the bigger sample size in Figure 5.1. However with even a small sample size they are also getting better with bigger constant c, and for the baseline hazard and $A(t)$ the coverage probabilities of D are decreasing on time. For $B(t)$ the estimation of H is worse than the other estimations.

In Figure 5.3 and Figure 5.4 the coverage probabilities of 95% confidence intervals with only smoothing are suggested depending on the variance estimations and the constant c with $n = 100$ and $n = 1{,}000$, respectively. They were used the same data sets as that in Figure 5.1 and Figure 5.2. Based on the bigger sample size of $n = 1{,}000$, the coverage probabilities are worse than those in Figure 5.1 when a small constant $c = 100$ is applied. However with enough constant of $c = 300$ and 500, they have little difference when compared to Figure 5.1 as expected from Table 3.1. When the Figure 5.2 and Figure 5.4 with $n = 100$ are compared, they are changed little with $c = 60$. However with a smaller constant, the coverage probabilities are significantly worse with only smoothing than with both modifications.

4. Application

The studied model is applied to a data set from the North Central Cancer Treatment Group, which is a sample data set in R named *lung*. It has 63 censored data out of 228 observations with 10 variables. In Table 4.1 the lung cancer data set are fitted by Aalen's additive regression model using *aareg* function in R assumed constant effect for the covariates of AGE and SEX. The coefficients of intercept (or baseline) and AGE are not significant at the 5% significance level, and the estimated hazard effect of SEX is −3.22e-03 with p-value of 0.00114. In Figure 5.5 the estimates of cumulative hazard function by the proposed piecewise constant hazard model are plotted with 95% confidence intervals for the baseline and the two covariates. The estimates in $d = 10$ intervals are connected by line segment for convenience, and set $c = 30$ and $\varepsilon = 0.15$. Arbitrarily the variables of AGE and SEX are assigned as an effect of time-varying and constant hazard, respectively. For baseline and AGE the estimates of cumulative hazard function are positive and have an increasing trend most of the time. However the 95% confidence intervals for both contain a zero, and it means they are not significant at the 5% level. The estimate of SEX is negative and it is significant. It is the same conclusions using Aalen's additive regression model with a constant effect in Table 4.1.

5. Conclusions

The WLS estimator and its modifications on weights are studied based on the partly parametric version of Aalen's additive risk model. The follow-up time period is partitioned, and the cumulative hazard functions are estimated as a piecewise constant in each interval. In the modifications the estimates of the hazard functions are smoothing over neighboring intervals and updated with the mean of the individual hazard rates in its interval multiplying a small constant ε to avoid negative or too much weight on one observation. By simulations the constant c in smoothing and ε in updating

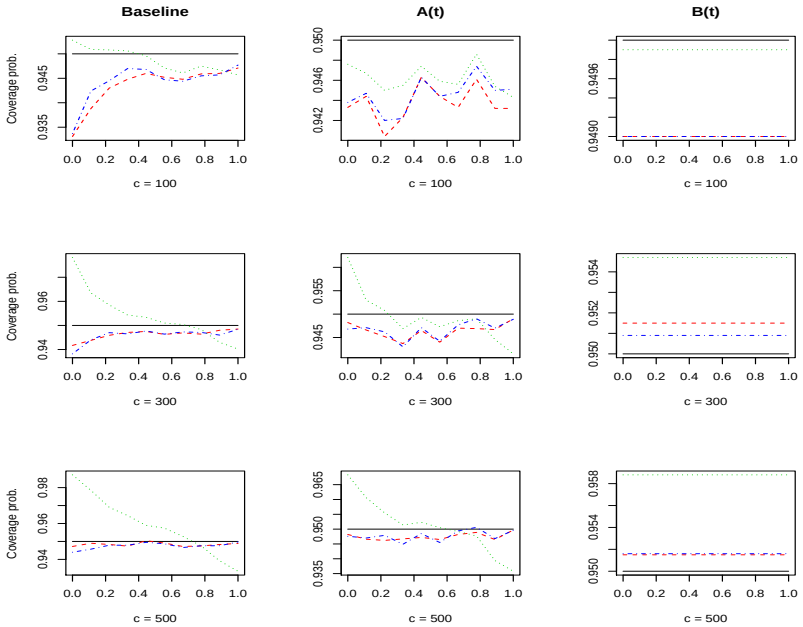

Figure 5.1. Coverage probabilities based on $n = 1{,}000$ for three estimates of variances depending on constants c. Solid lines are a reference at 95%, dotted lines are for the estimate of $\hat{D}$, dashed lines are for the estimate of $\hat{E}$, and irregular dashed lines are for the estimate of $\hat{F}$.

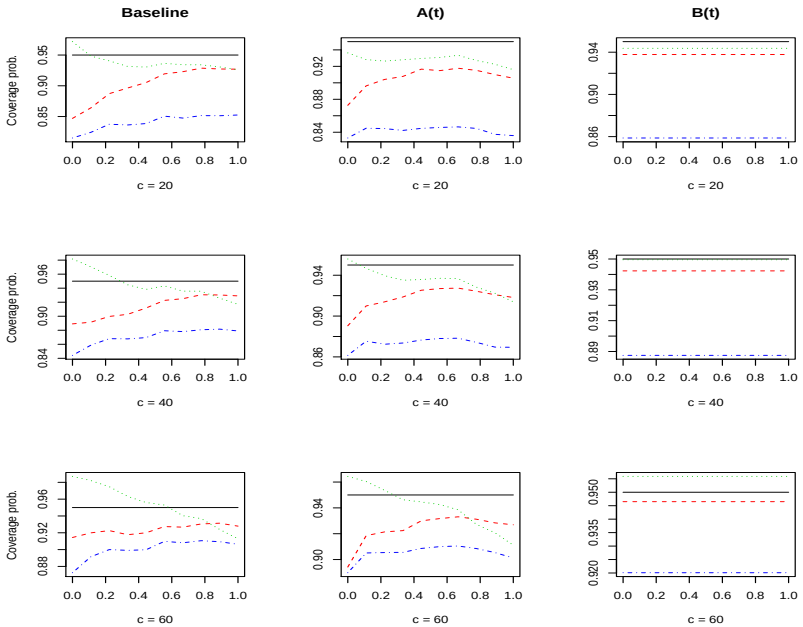

Figure 5.2. Coverage probabilities based on $n = 100$ for three estimates of variances depending on constants c. Solid lines are a reference at 95%, dotted lines are for the estimate of $\hat{D}$, dashed lines are for the estimate of $\hat{E}$, and irregular dashed lines are for the estimate of $\hat{F}$.

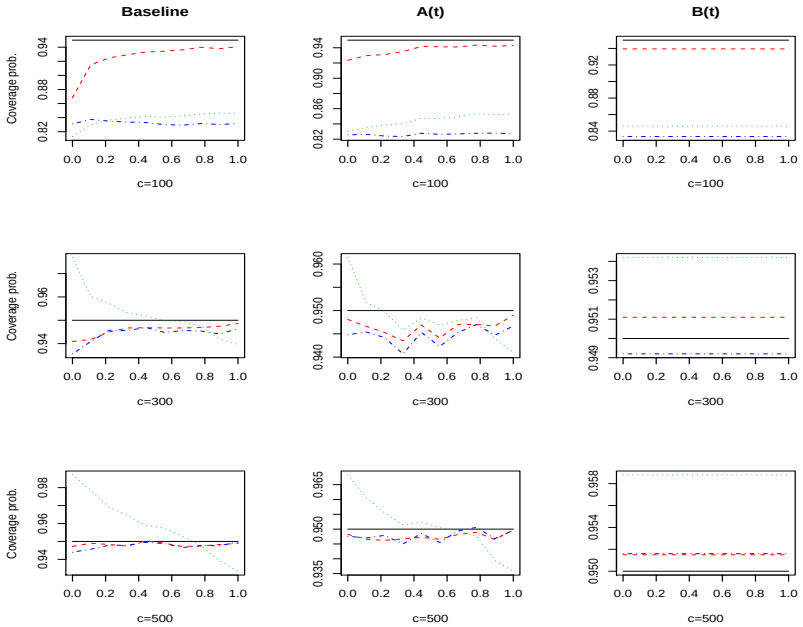

Figure 5.3. Coverage probabilities based on $n = 1{,}000$ with only smoothing for three estimates of variances depending on constants c. Solid lines are a reference at 95%, dotted lines are for the estimate of $\hat{D}$, dashed lines are for the estimate of $\hat{E}$, and irregular dashed lines are for the estimate of $\hat{F}$.

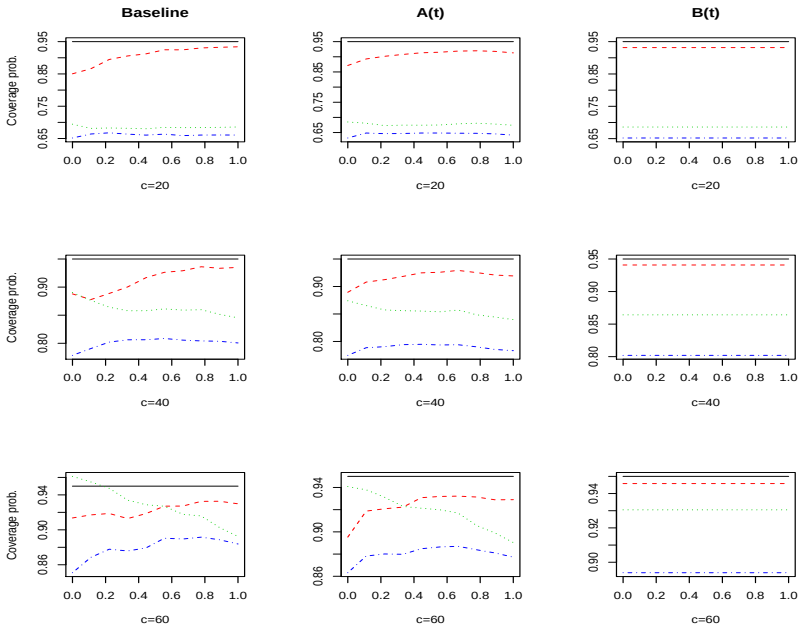

Figure 5.4. Coverage probabilities based on $n = 100$ with only smoothing for three estimates of variances depending on constants c. Solid lines are a reference at 95%, dotted lines are for the estimate of $\hat{D}$, dashed lines are for the estimate of $\hat{E}$, and irregular dashed lines are for the estimate of $\hat{F}$.

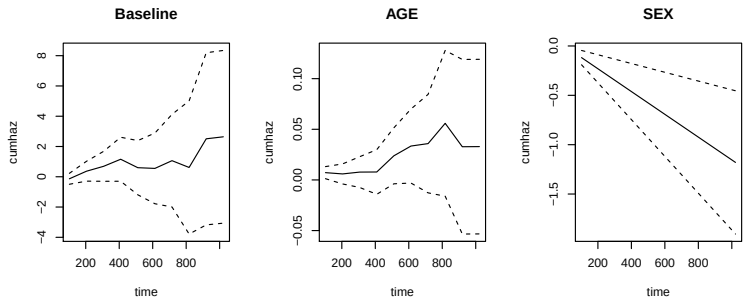

Figure 5.5. Estimates of cumulative hazard functions with 95% confidence intervals for Lung cancer data

are evaluated, and the coverage probabilities are compared among three estimations of variance with various c and sample size n.

When there are enough uncensored deaths in an interval, only smoothing with a sufficient big smoothing window size c suggests stable estimates. When we apply only smoothing, the constant c could be proposed at least $0.4n$. Adding updating in (2.6), better estimates are suggested even with a small constant c after smoothing, and there are no differences among the constant c between 0.15 and 0.35. Based on a small data set, the modifications of smoothing and updating should be applied together to provide good estimates. When we have a small data set, we should use a small number of intervals to place more uncensored deaths in each interval. Then the weights are estimated to be more stable after smoothing and updating. The follow-up time period could be partitioned into intervals according to the nature of the time-varying effects. In comparing coverage probabilities of 95% confidence intervals the estimations of E and F suggest good estimates using sufficient smoothing constant c. The estimates of E and F are getting better with bigger constant c, and they are similar as expected in (2.8). However the estimates of D become worse with a bigger c at the beginning and end of the time period.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes, In *Lecture Notes in Statistics*, Springer-Verlag, New York.

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

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

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**, 38–53.

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

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model, *Biometrika*, **81**, 501–514.

Uhm, D., Huffer, F. W. and Park, C. (2011). Additive risk model using piecewise constant hazard function, *Communications in Statistics-Simulation and Computation*, **40**, 1458–1477.