

# A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause

Bin Nan<sup>1</sup>, Xihong Lin<sup>2</sup>, Lynda D. Lisabeth<sup>3</sup>, and Siobán D. Harlow<sup>4</sup>

<sup>1,2</sup> Department of Biostatistics

<sup>3</sup> Department of Neurology

<sup>4</sup> Department of Epidemiology

University of Michigan, Ann Arbor, MI 48109

<sup>1</sup> *email*: bnan@umich.edu

June 1, 2004

**SUMMARY.** It is of recent interest in reproductive health research to investigate the validity of a marker event for the onset of menopausal transition and to estimate age at menopause using age at the marker event. We propose a varying coefficient Cox model to investigate the association between age at a marker event, defined as a specific bleeding pattern change, and age at menopause, where both events are subject to censoring and their association varies with age at the marker event. Estimation proceeds using the regression spline method. The proposed method is applied to the Tremin Trust Data to evaluate the association between age at onset of the 60-day menstrual cycle and age at menopause. The performance of the proposed method is evaluated using a simulation study.

**KEY WORDS:** *B*-splines; Cox regression; Generalized cross validation; Marker events; Non-parametric regression; Survival Analysis; Time-dependent covariates.

# 1 Introduction

It is of recent interest in female reproductive aging research to identify marker events for the onset of the menopausal transition, and to investigate their use for estimating age at menopause. Menopause is defined as the final menstrual period (FMP), with the FMP confirmed after at least 12 months of amenorrhea. Although several marker events based on menstrual bleeding criteria have been proposed (Mitchell, et al., 2000; Soules, et al., 2001; Taffe and Dennerstein, 2001), there is a lack of appropriate statistical models to formally evaluate their validity due to the complex nature of the data.

This paper is motivated by the analysis of the Tremin Trust data. This data set provides a unique opportunity to evaluate the association between age at menopause and ages at onset of the marker events proposed by reproductive health experts based on bleeding criteria (Treloar, et al., 1967). The study enrolled 1997 white college students at the University of Minnesota between 1935 and 1939 and followed them up to 40 years through their reproductive life. The study participants were asked to use menstrual diary cards to record the days when bleeding was experienced. Only limited covariate information was available in the data.

Lisabeth et al. (2004) analyzed a subset of 562 women from the original Tremin Trust cohort, who were age 25 or less at enrollment and still participating in the study at age 35. A total of 282 women experienced the 60-day cycle marker event. The median age at the 60-day cycle marker was 48.7 years. A total of 193 women experienced natural menopause. The median age at menopause was 51.7 years. There were 9 women who experienced menopause without having the 60-day cycle marker, and 271 women who were censored for both 60-day cycle and menopause events. Note that these 271 women who were censored for the marker event were part of the 369 women who were censored for menopause and their censoring times for those two events were the same. The median age at enrollment was 19 years,

the median age at menarche was 12 years and ranged from 9 to 18 years, and the length of follow-up ranged from 9 to 39 years with median of 27 years. The descriptive analysis results of Lisabeth et al. (2004) suggest that the 60-day cycle might be a useful marker for predicting age at menopause.

To explore the relationship between age at the 60-day cycle marker and age at menopause, we first restricted ourselves to the 282 women who had an observed marker event and classified them into several groups based on their ages at onset of the marker event as  $[35, 40)$ ,  $[40, 43)$ , and so on, which is similar to what Lisabeth et al. (2004) did. For each marker age group, we calculated the quartiles of age at menopause using the Kaplan-Meier method and displayed these estimated quartiles using a boxplot. These boxplots are given in Figure 1. The number of women in each marker age group is given above the corresponding boxplot. Figure 1 shows that the relationship between age at the 60-day cycle marker and age at menopause is complicated and varies with age at the 60-day cycle marker. This relationship, however, is only explorative and may not be able to reflect the truth quantitatively since women who were censored for the marker event were excluded from the analysis. In other words, the complete case analysis makes a strong assumption that the marker event is missing completely at random (see e.g. Little and Rubin (2002)). An appropriate statistical model is proposed in this article which can easily handle censored markers without making this strong assumption. For more discussions, see the Discussion Section.

The first scientific interest is to quantify the association between age at the 60-day marker and age at menopause using a statistical model. The second scientific interest, especially for clinicians and women themselves, is to estimate the distribution of age at menopause given age at onset of the 60-day cycle marker. For example, if a woman first experiences a 60-day cycle at age 40, she would like to know from her physician her expected median age of menopause. From a clinical point of view, this would be a very useful piece of information for helping determine a woman's need for continued contraception and the likelihood of

initiating interventions such as bone density screening.

Several approaches have been proposed for modeling intermediate marker events. Crowley and Hu (1977) analyzed the Stanford heart transplant data using the Cox partial likelihood method by treating the transplant status, an intermediate marker event, as a time dependent covariate. Lefkopoulou and Zelen (1995) and Nam and Zelen (2001) studied the same model from a different angle, which leads to a contingency table interpretation. For an overview of the existing methods handling intermediate marker events, see Kalbfleisch and Prentice (2002, Section 6.4). All of these authors assumed a constant regression coefficient for modeling the effect of the intermediate marker event (See Figure 2(a)). The results in Figure 1 however suggest that this assumption is not appropriate for the Tremin Trust data. We need to allow the regression coefficient of age at onset of the 60-day marker event to vary with age at the marker event. We hence consider a varying coefficient model.

Hastie and Tibshirani (1993) proposed general varying-coefficient models. In the Cox model setting, it is commonly assumed in such models that the regression coefficient  $\beta(\cdot)$  of the time-dependent covariate is a function of the follow-up time, e.g., see Murphy and Sen (1991), Marzec and Marzec (1997), among others. Their model can be illustrated by replacing the dot line (and dash line) in Figure 2(a) by an arbitrary curve that may not be parallel to the log baseline hazard function. The interest of the Tremin Trust data, however, lies in evaluating the effect of age at the 60-day cycle marker event on age at menopause as a function of age of the marker event, as demonstrated in Figure 1. Hence it is natural and biologically more desirable to assume that the regression coefficient  $\beta(\cdot)$  of the time dependent covariate, which indicates the onset of the marker event, be a function of age at the marker event, instead of a function of the follow-up time. The resulting regression coefficients are biologically more interpretable to address the scientific interest of the Tremin Trust data. The model is illustrated in Figure 2(b). Both of these two situations are special cases of the general framework of varying-coefficient models for survival data of Hastie and

Tibshirani (1993). The first situation has been investigated in detail, while limited attention has been paid to the second situation, which is of the major interest in the Tremin Trust data. Parameters in both models can be estimated by using the (penalized) partial likelihood method.

The model discussed in this article can also be regarded to fall into the general framework of the illness-death model of Joly et al. (2002). However, the estimation procedure in Joly et al. (2002) is only for the model where the effect of the marker event varies with the follow-up time, not the time at onset of the marker event which is our major interest.

The remainder of the paper is organized as follows. We introduce in Section 2 a varying coefficient Cox model for age at menopause, where age at onset of the 60-day cycle marker is a time dependent binary covariate and its coefficient is assumed to be a smooth function of the marker event age. We discuss in Section 3 an estimation procedure using regression splines. We analyze in Section 4 the Tremin Trust data, and conduct a simulation study in Section 5 to evaluate the performance of the proposed method, followed by concluding remarks in Section 6.

## 2 The Varying-Coefficient Model

Suppose the data consist of  $n$  subjects. Let  $Y_i$  be the observed time to the event of interest, which is defined as the minimum of the survival time  $T_i$ , e.g., age at menopause, and the censoring time  $C_i$  for the  $i$ th subject ( $i = 1, \dots, n$ ). We assume independent censoring. Let  $\Delta_i$  be a censoring indicator, which takes value 1 if a failure is observed and 0 otherwise. Let  $Z_i(t)$  be a time dependent covariate.

Assuming  $\lambda_0(t)$  is the baseline hazard and  $\lambda_i\{t|Z_i(t)\}$  is the hazard rate of the survival time to the endpoint event at  $t$  given  $Z_i(t)$ . A standard Cox model with a time dependent covariate has the following form:

$$\lambda_i\{t|Z_i(t)\} = \lambda_0(t)\exp\{\beta Z_i(t)\} . \quad (1)$$

It is common to use (1) to model the effect of an intermediate marker event (Crowley and Hu, 1977; Kalbfleisch and Prentice, 2002).

In the Tremin Trust data,  $t$  is age, time to the endpoint event is age at menopause, and time to the marker event is age at the first occurrence of the 60-day cycle marker event. Model (1) assumes the log relative risk comparing subjects who have experienced the marker event with subjects who never experienced the marker event is constant  $\beta$ , and is irrelevant to the age they have experienced the marker event. However, discussions in Section 1 suggest that the association between age at menopause and age at the marker event varies with age at the marker event in the Tremin Trust data.

Let  $S_i$  be the age at the 60-day marker event for woman  $i$ . Define

$$Z_i(t) = \begin{cases} 1 & \text{if } t \geq S_i \\ 0 & \text{if } t < S_i. \end{cases} \quad (2)$$

Equivalently,  $Z_i(t) = I[t \geq S_i]$ , where  $I(\cdot)$  is an indicator function. We extend model (1) to allow the association between age at menopause and age at the marker event to depend on age at the marker event  $S_i = s$  as

$$\lambda\{t|Z_i(t)\} = \lambda_0(t)\exp\{\beta(s)Z_i(t)\} = \begin{cases} \lambda_0(t)\exp\{\beta(s)\} & \text{if } t \geq s \\ \lambda_0(t) & \text{if } t < s, \end{cases} \quad (3)$$

where  $\beta(s)$  is an unknown smooth function.

The interpretations and difference between model (1) and model (3) are clearly illustrated using Figure 2(a) and 2(b) on the log hazard scale by contrasting two subjects who have experienced the marker event at time 1 and time 2 respectively. Under the constant coefficient Cox model (1), the first subject's hazard is the baseline hazard before time 1 ( $s_1$ ) and changes by an amount of  $\beta$  since time 1, while the second subject's hazard is the baseline hazard before time 2 ( $s_2$ ) and changes from the baseline hazard since time 2 by the same amount  $\beta$ . Under the varying-coefficient Cox model (3), both women's hazards also change at time 1 and time 2 respectively, but by different constants  $\beta(s_1)$  and  $\beta(s_2)$  respectively. Note that

the lines are all parallel and reflect the proportional hazards assumption. It should be noted that  $Z_i(t)$  is always observable in the analysis at any  $t$  whenever subject  $i$  appears in the risk set at time  $t$  even if  $S_i$  is not observed. Thus  $\beta(s)$  is estimable.

If baseline covariates  $\mathbf{X}_i$  are available, model (3) can be easily extended to incorporate baseline covariates  $\mathbf{X}_i$  as

$$\lambda_i\{t|Z_i(t), \mathbf{X}_i\} = \lambda_0(t)\exp\{\beta(s)Z_i(t) + \boldsymbol{\gamma}'\mathbf{X}_i\}, \quad (4)$$

where  $\mathbf{X}_i$  is age at menarche in the Tremin Trust data. Since model (3) is a special case of model (4), we shall focus on model (4) in this paper.

### 3 The Estimation Procedure

#### 3.1 Estimation Using B-splines

We consider estimation of the nonparametric function  $\beta(s)$  using the regression spline method by approximating  $\beta(s)$  using the natural cubic  $B$ -spline basis. Let  $K$  be the number of interior knots. Knot locations are usually chosen such that there are roughly equal numbers of observed data points between any two adjacent knots. This can be done by placing the  $K$  knots using the  $100j/(K+1)$  ( $j = 1, \dots, K$ ) percentiles of the observed marker event times. We discuss in Section 3.2 estimation of the number of knots  $K$  using GCV.

Since a natural spline is constrained to be linear beyond the two boundary knots, the function  $\beta(s)$  can be parameterized using  $K+2$  natural cubic  $B$ -spline basis functions  $B_k(s)$  ( $k = 1, \dots, K+2$ ) as

$$\beta(s) = \sum_{k=1}^{K+2} \theta_k B_k(s). \quad (5)$$

Replacing  $\beta(s)$  by its  $B$ -spline approximation in equation (4), model (4) can be written as

$$\lambda\{t|Z_i(t), \mathbf{X}_i\} = \lambda_0(t)\exp\{\boldsymbol{\theta}'\tilde{\mathbf{Z}}_i(t) + \boldsymbol{\gamma}'\mathbf{X}_i\}, \quad (6)$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{K+2})'$  and  $\tilde{\mathbf{Z}}_i(t) = \{B_1(s)Z_i(t), \dots, B_{K+2}(s)Z_i(t)\}'$ . Note that  $\tilde{\mathbf{Z}}_i(t)$  is always observable during follow-up because  $Z_i(t)$  is fully observed during follow-up. Specifi-

cally, if the marker event is observed at  $S_i$  for the  $i$ th woman during follow-up, then  $\tilde{\mathbf{Z}}_i(t) = \mathbf{0}$  if  $t < S_i$  and  $\tilde{\mathbf{Z}}_i(t) = \{B_1(S_i), \dots, B_{K+2}(S_i)\}'$  if  $t \geq S_i$ . If the marker event is not observed, i.e.,  $S_i$  is censored, then  $\tilde{\mathbf{Z}}_i(t) = \mathbf{0}$  at any observed follow-up time  $t$ .

Now model (6) becomes a standard Cox proportional hazards model with the time dependent covariate vector  $\tilde{\mathbf{Z}}(t)$  and the baseline covariate vector  $\mathbf{X}_i$ . Thus the estimation of parameters  $(\boldsymbol{\theta}, \boldsymbol{\gamma})$  can be obtained using partial likelihood method. Denote the maximum partial likelihood estimators of  $(\boldsymbol{\theta}, \boldsymbol{\gamma})$  by  $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$  and their covariance estimators by  $\text{cov}(\hat{\boldsymbol{\theta}})$  and  $\text{cov}(\hat{\boldsymbol{\gamma}})$ . The nonparametric function  $\beta(s)$  can then be estimated by

$$\hat{\beta}(s) = \sum_{k=1}^{K+2} \hat{\theta}_k B_k(s). \quad (7)$$

The pointwise confidence interval for  $\hat{\beta}(s)$  can be estimated using its variance estimator  $\text{var}\{\hat{\beta}(s)\} = \mathbf{B}(s)' \text{cov}(\hat{\boldsymbol{\theta}}) \mathbf{B}(s)$ , where  $\mathbf{B}(s) = \{B_1(s), \dots, B_{K+2}(s)\}'$ .

As discussed in the Introduction Section, it is of both clinical interest and a woman's own interest to estimate age at menopause if a woman has experienced the 60-day marker event at a certain age. We first estimate the baseline cumulative hazard function  $\Lambda_0(t)$  using Breslow estimator,

$$\hat{\Lambda}_0(t) = \int_0^t \left[ \sum_{i=1}^n I(Y_i \geq u) \exp\{\hat{\beta}(S_i) Z_i(u) + \hat{\boldsymbol{\gamma}}' \mathbf{X}_i\} \right]^{-1} \left\{ \sum_{i=1}^n dN_i(u) \right\}. \quad (8)$$

Then the survival function for menopause given the age at the marker event  $S = s$  and the covariates  $\mathbf{X} = \mathbf{x}$  can be estimated by

$$\hat{F}(t|s, \mathbf{x}) = \exp \left\{ - \int_0^t \exp\{\hat{\beta}(s) z(u) + \hat{\boldsymbol{\gamma}}' \mathbf{x}\} d\hat{\Lambda}_0(u) \right\}, \quad (9)$$

where  $z(u) = I(u \geq s)$ .

### 3.2 Estimation of the Number of Knots

An advantage of the use of a regression spline for estimating the nonparametric function  $\beta(s)$  is its computational simplicity. However this method requires estimation of the number of



knots. For uncensored data, cross-validation (CV) and generalized cross validation (GCV) are commonly used, see, e.g. Hastie and Tibshirani (1990). For survival data, O'Sullivan (1988) proposed CV and GCV for choosing the smoothing parameter for the smoothing spline estimator assuming that the baseline cumulative hazard function  $\Lambda_0(t) = \int_0^t \lambda_0(u)du$  is known. We extend O'Sullivan's method to choose the number of knots in the regression spline setting and account for the fact that  $\Lambda_0(t)$  is unknown and is estimated.

We first consider the case when  $\Lambda_0(t)$  is known. Following O'Sullivan (1988), under model (6), for a given number of knots  $K$ , if  $\Lambda_0(t)$  is a known function, the likelihood function of  $(\boldsymbol{\theta}, \boldsymbol{\gamma})$  is available and can be maximized using an iterated reweighted least square algorithm. If the estimators of  $(\boldsymbol{\theta}, \boldsymbol{\gamma})$  at the  $l$ th iteration are  $(\hat{\boldsymbol{\theta}}_{(l)}, \hat{\boldsymbol{\gamma}}_{(l)})$ , the working weight  $w_i$  and the working dependent variable  $y_i$  for subject  $i$  can be written as

$$\begin{aligned} w_i &= \frac{1}{2} \Lambda_0(Y_i) \exp\{\hat{\boldsymbol{\theta}}'_{(l)} \tilde{\mathbf{Z}}_i(Y_i) + \hat{\boldsymbol{\gamma}}'_{(l)} \mathbf{X}_i\} , \\ y_i &= \hat{\boldsymbol{\theta}}'_{(l)} \tilde{\mathbf{Z}}_i(Y_i) + \hat{\boldsymbol{\gamma}}'_{(l)} \mathbf{X}_i + \Delta_i / (2w_i) - 1 . \end{aligned}$$

One calculates  $(\hat{\boldsymbol{\theta}}_{(l+1)}, \hat{\boldsymbol{\gamma}}_{(l+1)})$  by minimizing  $\sum_{i=1}^n w_i \{y_i - \boldsymbol{\theta}' \tilde{\mathbf{Z}}_i(Y_i) - \boldsymbol{\gamma}' \mathbf{X}_i\}^2$ . Let  $\tilde{\mathbf{X}}_i = \{\tilde{\mathbf{Z}}_i(Y_i), \mathbf{X}_i'\}'$  and  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}'_1, \dots, \tilde{\mathbf{X}}'_n)'$ . Denote the working dependent variable, the working weight matrix, and the predicted value vector at convergence by  $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)'$ ,  $\hat{\mathbf{W}} = \text{diag}(\hat{w}_1, \dots, \hat{w}_n)$  and  $\hat{\mathbf{f}} = (\hat{f}_1, \dots, \hat{f}_n)'$ . Then  $\hat{\mathbf{f}}$  can be calculated as  $\hat{\mathbf{f}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}' \hat{\mathbf{W}} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \hat{\mathbf{W}} \hat{\mathbf{y}} = \hat{\mathbf{H}} \hat{\mathbf{y}}$ , where  $\hat{\mathbf{H}}$  is the linearized hat matrix. The generalized cross-validation (GCV), which is a function of the number of knots  $K$ , is given by

$$GCV(K) = \frac{\sum_{i=1}^n w_i (\hat{y}_i - \hat{f}_i)^2}{(1 - \bar{h})^2} , \quad (10)$$

where  $\bar{h}$  is the average of the diagonal elements of  $\hat{\mathbf{H}}$ , the so-called mean leverage.

We now consider the case when the baseline hazard  $\Lambda_0(t)$  is unknown and is estimated by the Breslow estimator (8). O'Sullivan (1988) suggested to calculate the Breslow estimator of  $\Lambda_0(\cdot)$  for each  $K$  and plug it into (10) as if it were known. However, this plug-in procedure

ignores the fact that different choices of  $K$  give different baseline hazard estimators of  $\Lambda_0(t)$ , but the above procedure assumes that the same true baseline hazard is used for different  $K$ . We hence propose a modified procedure to account for this.

First, a series of Cox models as in (6) are fitted for a range of the number of interior knots  $K$ . We used 1 to 20 in the analysis of Tremin Trust data. For each choice of  $K$ , the cumulative baseline hazard function estimator  $\hat{\Lambda}_0(t; K)$  and the B-spline estimator  $\hat{\beta}(s; K)$  are calculated. They are then plugged into equation (10) to calculate  $\text{GCV}(K)$ . Note that different baseline hazard estimators are used for different  $K$  at this step. We then select  $K$  that minimizes  $\text{GCV}(K)$ , call it  $K_*$  and obtain the corresponding baseline hazard estimator  $\hat{\Lambda}_0(t; K_*)$ . At the next step, we replace the true  $\Lambda_0(t)$  by this estimated  $\hat{\Lambda}_0(t; K_*)$  and treat it as fixed and known. Then recalculate the GCV statistic (10) using the above least square procedure for each of the possible values of  $K$ , 1 to 20 in our analysis, and select a new  $K$ , the updated  $K_*$ , that minimizes  $\text{GCV}(K)$ . Note here a common  $\hat{\Lambda}_0(t; K_*)$  is used to calculate GCV for different possible values of  $K$ . The procedure is repeated using a newly updated common  $\hat{\Lambda}_0(t; K_*)$  until the chosen  $K_*$  at the current step is the same as the  $K_*$  at the previous step. The cross-validation (CV) statistic can be calculated similarly.

## 4 The Analysis of the Tremin Trust Data

We applied the proposed varying coefficient Cox model to the analysis of the Tremin Trust data. The goals of our study were to investigate the relationship between age at menopause and age at the 60-day cycle marker event, and to estimate the distribution of age at menopause given any particular age at onset of the 60-day cycle marker. The data used in our analysis were the same as that used in Lisabeth, et al. (2003), which were described in detail in Section 1. We used age 35 as the time origin in our analysis.

For each woman, the data set contained the observed menopause age which is the minimum of the age at menopause and the censoring age; a censoring indicator for age at

menopause; a 60-day cycle marker event indicator; the age at the marker event if it occurred during the follow-up time; and age at menarche.

Let  $Z_i(t)$  be a time-dependent binary indicator for the onset of the 60-day cycle marker event and the baseline covariate  $X_i$  be age at menarche. We first fitted (4) by assuming simple parametric forms for  $\beta(s)$  as a cubic function. The linear, quadratic and cubic terms were found to be highly significant. This implies that a simple third order polynomial function does not seem to be adequate for describing the effect of age at marker event. We then fit the semiparametric varying coefficient Cox model (4) by estimating  $\beta(s)$  nonparametrically using the B-spline method via the Cox model (6).

The method of Therneau and Grambsch (2000) was used to expand the data set for the time dependent covariate  $Z_i(t)$ . The observed marker times were used to determine the knot allocations and generate the natural cubic  $B$ -spline basis functions  $B_k(s)$  used for estimating  $\beta(s)$ . The extreme values of the observed marker times were used as the two boundary knots. The optimal number of interior knots estimated using GCV method described in Section 3.2 is  $K_{optimal} = 8$ . The spline estimator of  $\beta(s)$  and its 95% point-wise confidence interval are plotted in Figure 3. For illustrative purpose, we also considered approximating  $\beta(s)$  using piecewise constants as  $\beta(s) = \sum_{k=1}^{K+1} \beta_j I[s_{k-1} < S_i \leq s_k]$ , where  $\{s_0, s_1, \dots, s_{K+1}\}$  is the set of knots including the boundary knots, and fit  $\lambda\{t|Z_i(t), X_i\} = \lambda_0(t) \exp\left\{\sum_{k=1}^{K+1} \beta_j I[s_{k-1} < S_i \leq s_k]Z_i(t) + \gamma X_i\right\}$ . The piecewise constant estimator of  $\beta(s)$  using the age intervals  $[35, 38)$ ,  $[38, 40)$ , etc., is superimposed in Figure 3. We can see that the B-spline estimate and the piece-wise constant estimate of  $\beta(s)$  agree well with each other.

The results in Figure 3 suggest that the 60-day cycle marker is strongly associated with age at menopause, and its effect varies with age at the 60-day cycle marker event. But when age at marker event is close to 35, the estimated  $\beta(s)$  does not significantly differ from zero which implies that having a marker around age 35 is uninformative about age at menopause. The curve is mainly positive and increases before age 44 and then starts to

decrease. This indicates that before age 44, the association between age at menopause and age at the 60-day cycle marker becomes stronger as age increases. Among women who first experience the 60-day cycle before 44, as age at onset of the 60-day cycle increases, she is likely to have menopause more quickly. For example, consider two women: the first woman experiences the 60-day cycle at age 39 and the second woman experiences the 60-day cycle at age 42. Then relative risk of menopause at any age after age 42 for the second woman is  $\exp\{\hat{\beta}(42) - \hat{\beta}(39)\} = \exp(4.1 - 2.2) = 6.7$  times higher than the first woman ( $p$ -value  $< 0.0001$ ).

The estimated  $\beta(s)$  curve starts decreasing after age 44. This indicates that after age 44, the association between age at menopause and age at the 60-day cycle marker becomes weaker as age increases. Among women who first experience the 60-day cycle after 44, as age at onset of the 60-day cycle increases, a woman is likely to have menopause at a later age. For example, consider two women: the first woman experiences the 60-day cycle at age 48 and the second woman experiences the 60-day cycle at age 51. Then relative risk of menopause at any age after age 48 for the second woman is  $\exp\{\hat{\beta}(51) - \hat{\beta}(48)\} = \exp(1.9 - 3.2) = 0.27$  times lower than the first woman ( $p$ -value  $< 0.0001$ ). In other words, the relative risk of menopause at any age after age 51 for the first woman is  $1/0.27 = 3.7$  times higher of the second woman.

The estimated log relative risk for age at menarche was  $-0.16$  ( $RR = 0.85$ ) for a one year increment ( $p$ -value  $= 0.01$ ). This means that a younger age at menarche has a significant effect on advancing the expected age at menopause. We also found that the effect of age at the 60-day cycle marker was independent of age at menarche. Particularly the estimated curves of  $\beta(s)$  were almost identical with and without adjusting for age at menarche.

The survival probabilities of age at menopause were calculated using equation (9) for several selected ages at the 60-day cycle marker event given age of menarche equaled to 12, which was the median age of menarche. The estimated survival curves are plotted in Figure 4(a) and the estimated corresponding percentiles are summarized in Table 1. These

results are consistent with the pattern of the estimated  $\beta(s)$  curve in Figure 3. For a woman who experiences the 60-day cycle marker before age 44, the later she experiences the marker event, the earlier she is likely to experience menopause. For a woman who experiences the 60-day cycle marker after age 44, the later she experiences the marker event, the later she is likely to experience menopause.

These results are biologically meaningful. Women who are observed to have a 60-day cycle before age 40 may belong to a subgroup of women who cycle infrequently, e.g. women with polycystic ovarian disease, and for whom the pattern of change in menstrual bleeding with age may differ from other women. Additional research on this subgroup of women is needed. To explore this, we conducted a subgroup analysis restricted to women with age at marker greater or equal to 40. The estimated curve (with the same knots) matched the curve in Figure 3 well except for the left end within the first 2 or 3 years after age 40, where the curve was lower but still within the pointwise confidence band of the previous curve.

Another interesting and more intuitive piece of information for both clinicians and midlife women is the number of years from the onset of marker event to menopause. The percentiles of this quantity can be easily calculated by subtracting age at marker event from the corresponding estimated percentiles for age at menopause, which are also given in Table 1. The survival curves for menopause after the onset of marker event are plotted in Figure 4(b).

## 5 The Simulation Study

We conducted a simulation study to evaluate the performance of the natural cubic  $B$ -spline estimator for  $\beta(s)$  in model (3). The follow-up time was restricted from 0 to 1. To roughly mimic the shape of the estimated  $\hat{\beta}(s)$  for the 60-day cycle marker event in Figure 3, we assumed that true  $\beta(s) = 3\sin(\pi s)$ . The age at the marker event  $S$  was generated from a Weibull distribution with shape parameter 2 and scale parameter 1. The age at menopause  $T$  was generated from the model  $\lambda\{t|Z(t)\} = \lambda_0(t)\exp\{\beta(s)Z(t)\}$ , where  $Z(t) = I(t \geq S)$  and

the baseline hazard  $\lambda_0(t) = 0.5t^2$ , which corresponds to the hazard of a Weibull distribution with shape parameter 2 and scale parameter 4. The censoring time  $C$  was generated by  $C = U \cdot I(U \leq 1) + I(U > 1)$ , where  $U \sim \text{Uniform}(0, 2)$ . Thus the observed time  $Y = \min(T, C)$  was within the interval  $[0, 1]$ . The censoring percentage was about 70%. We assumed a sample size of  $n = 500$  in each simulated data set.

To reduce the computational burden, we chose the optimal number of interior knots in estimating  $\beta(\cdot)$  by minimizing the mean square error of  $\hat{\beta}(\cdot)$  defined as  $\text{MSE} = \sum_{j=1}^J \left\{ \hat{\beta}(t_j) - \beta(t_j) \right\}^2$ , where  $t_j, j = 1, \dots, J$ , are equally spaced grid points in  $(0, 1)$  with two boundary knots at 0 and 1. We used  $J = 1000$ . The MSE criterion suffices for our purpose.

We performed 100 simulations and analyzed each simulated data set using the varying coefficient model (3) using B-splines by fitting the Cox model (6). Each data set was analyzed using the proposed method using all the data and using only the complete cases, i.e., those women who have experienced the marker event. The estimated optimal numbers of interior knots varied from 1 to 6 with the average number of estimated knots equal to 1.6. The average of the 100 estimated  $\hat{\beta}(\cdot)$  and the true curve  $\beta(\cdot)$  are plotted in Figure 5. The 95% pointwise confidence intervals for  $\hat{\beta}(\cdot)$  using the empirical standard errors and the average of the 100 estimated standard errors are also plotted. Figure 5 suggests that the pointwise biases of the B-spline estimator  $\hat{\beta}(\cdot)$  are close to zero, and the pointwise model based SEs of  $\hat{\beta}(\cdot)$  agree well with their empirical counterparts, except for the boundary. The average of the estimated  $\hat{\beta}(\cdot)$  using complete case analysis is also plotted, and the bias is obvious.

## 6 Discussion

We have proposed in this paper a varying-coefficient Cox model to investigate the association between time to an intermediate marker event and time to a primary endpoint event, where the coefficient of the time dependent marker indicator is assumed to be a nonparametric function of time at the marker event, and baseline covariate effects are modeled paramet-

rically. We estimate the nonparametric regression function using  $B$ -splines which can be easily formulated into a standard Cox model and fitted using the standard partial likelihood method. We estimate the number of knots using a modification of O’Sullivan (1988)’s GCV method. Our simulation results suggest the proposed method works well in finite samples.

The large sample theory for the partial likelihood based regression spline estimator  $\hat{\beta}(s)$  is beyond the scope of this paper. For discussions of such spline estimators in linear regression settings, see e.g. Huang (2003). An extension to the Cox model setting requires further research. Our simulation results provide empirical evidence that similar results are likely to hold for Cox regression with varying-coefficients.

We calculated the pointwise confidence intervals for the nonparametric function  $\beta(s)$  in the analysis of the Tremin Trust data. One could also calculate the global confidence band. However, such a global confidence band is often found to be too wide to be practically useful.

We used all the data in our analysis. Similar to Figure 1, one might want to restrict analysis to the subset of women who have experienced the marker event when fitting (3). However, such analysis requires a strong assumption, i.e., the age at marker is missing completely at random, which apparently does not hold for right censored marker events. The estimators would be biased if this assumption is violated (Paik and Tsai, 1997). Our analysis however does not require this strong assumption and also does not add any major complexity in computation compared to the complete case analysis. An interesting phenomenon for the Tremin Trust data is that the complete case analysis only shows small bias comparing to the analysis using all the data. The bias is obvious in the simulation study (see e.g. Figure 5).

## Acknowledgment

The work of Nan and Lin was supported in part by U. S. National Cancer Institute Grant R01 CA76404. The work of Lisabeth and Harlow was supported in part by U. S. National Institute of Aging Grant AG021543.

## References

- Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* **72**, 27-36.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757-796.
- Huang, J. H. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics* **31**, 1600-1635.
- Joly, P., Commenges, D., Helmer, C., and Letenneur L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3**, 433-443.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, New Jersey: John Wiley & Sons.
- Lefkopoulou, M. and Zelen, M. (1995). Intermediate clinical events, surrogate markers and survival. *Lifetime Data Analysis* **1**, 73-85.
- Lisabeth, L. D., Harlow, S. D., Gillespie, B., Lin, X., and Sowers, M. F. (2004). Staging reproductive aging: a comparison of proposed bleeding criteria for the menopausal transition. *Menopause* **11**, 186-197.
- Little R. J. A. and Rubin D. B. (2002). *Statistical Analysis with Missing Data*. New Jersey: John Wiley & Sons.
- Marzec, L. and Marzec, P. (1997). On fitting Cox's regression model with time-dependent coefficients. *Biometrika* **84**, 901-908.



- Mitchell, E. S., Woods, N. F., and Mariella A. (2000). Three stages of the menopausal transition from the Seattle Midlife Women's Health Study: toward a more precise definition. *Menopause* **7**, 334-349.
- Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and Their Applications* **39**, 153-180.
- Nam, C. M. and Zelen, M. (2001). Comparing the survival of two groups with an intermediate clinical events. *Lifetime Data Analysis* **7**, 5-19.
- O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* **9**, 531-542.
- Paik, M. C. and Tsai, W. Y. On using the Cox proportional hazards model with missing covariates. *Biometrika* **84**, 579-593.
- Soules, M. R., Sherman, S., Parrott, E. P., Rebar R., Santoro, N., Utian, W., Woods, N. (2001). Executive summary: stages of reproductive aging workshop. *Fertil Steril* **76**, 874-878.
- Taffe, J. and Dennerstein, L. (2001). Menstrual patterns leading to final menstrual period. *Menopause* **9**, 32-40.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag New York, Inc.
- Treloar, A. E., Boynton, R. E., Behn, B. G., and Brown, B. W. (1967). Variation of the human menstrual cycle through reproductive life. *International Journal of Fertility* **12**, 77-126.

Table 1: Estimated percentiles for the survival probabilities of age at menopause given age at onset of the 60-day cycle marker. For each marker event age, the upper entries are the estimated percentiles of chronological age at menopause; and the lower entries are the estimated percentiles of the length in years from the onset of marker event to menopause.

Age at the marker event (in years)	Estimated percentiles				
	90%	75%	50%	25%	10%
36	50.6	52.5	54.4	55.3	56.2
	14.6	16.5	18.4	19.3	20.2
39	47.5	49.8	51.9	53.1	54.3
	8.5	10.8	12.9	14.1	15.3
42	43.3	45.2	47.5	49.2	50.3
	1.3	3.2	5.5	7.2	8.3
45	46.3	47.6	49.3	50.6	51.7
	1.3	2.6	4.3	5.6	6.7
48	48.6	49.3	50.6	51.7	52.4
	0.6	1.3	2.6	3.7	4.4
51	51.5	52.2	53.1	54.4	54.9
	0.5	1.2	2.1	3.4	3.9

## List of Illustrations

*Figure 1.* The box-plots for the estimated KM survival functions of age at menopause given different ages at the 60-day cycle marker event among the subset of women who had experienced the marker event. The number of women in each marker event age group is given above the corresponding boxplot.

*Figure 2.* An illustration of the log hazard functions at two marker event times under the constant coefficient Cox model (1) (Fig. (a)) and the varying-coefficient Cox model (3) (Fig. (b)): — Baseline hazard;  $\cdot \cdot \cdot$  Hazard if the marker event occurs at time 1;  $- - -$  Hazard if the marker event occurs at time 2.

*Figure 3.* Estimates of  $\beta(s)$  using the  $B$ -spline and the step-function for the Tremin Trust data: — estimated  $\beta(s)$  using the B-spline basis;  $\cdot \cdot \cdot$  95% CI;  $- - -$  estimated  $\beta(s)$  using piece-wise constants.

*Figure 4.* Estimated survival curves for age at menopause (Fig. (a)) and for time from onset of the 60-day cycle marker to menopause (Fig. (b)) given different ages at the 60-day cycle marker: — Age 36;  $- - -$  Age 39;  $\cdot \cdot \cdot$  Age 42;  $- \cdot -$  Age 45;  $- - -$  Age 48;  $- \cdot$  — Age 51.

*Figure 5.* Average of the estimated nonparametric functions  $\hat{\beta}(s)$  based on 100 simulations and its 95% pointwise confidence intervals: — true curve;  $- - -$  estimated curve;  $\cdot \cdot \cdot$  95% CI using the pointwise estimated SEs;  $- \cdot -$  95% CI using the pointwise empirical SEs;  $- - -$  estimated curve using complete case analysis.

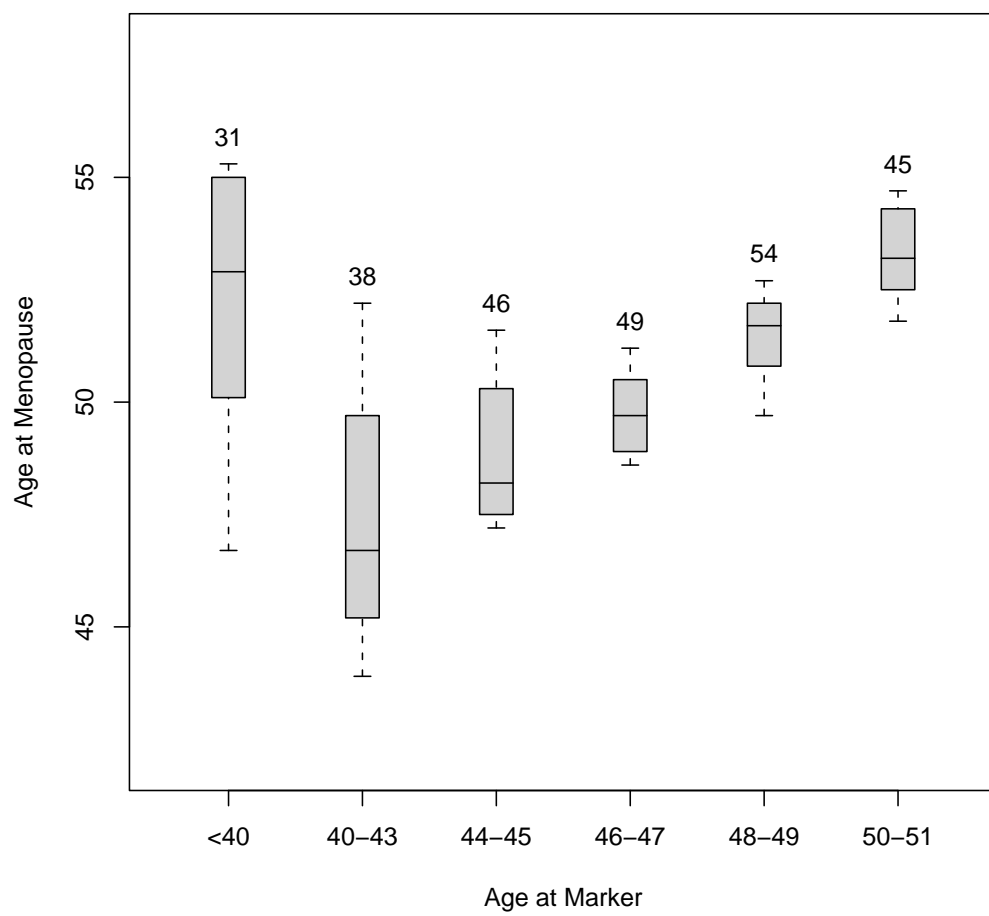


Figure 1:

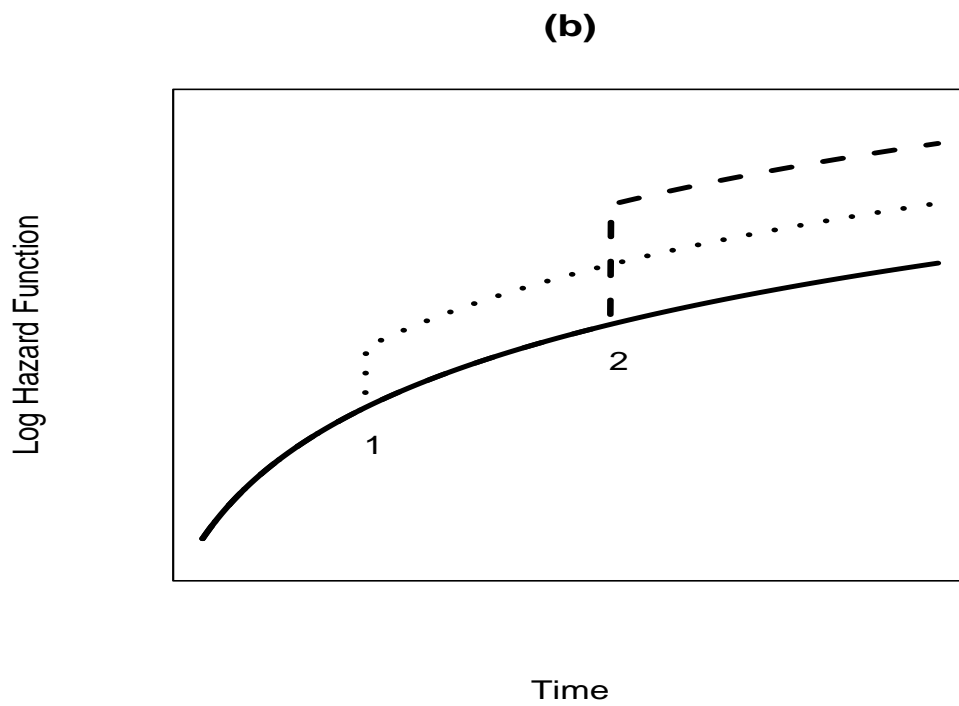
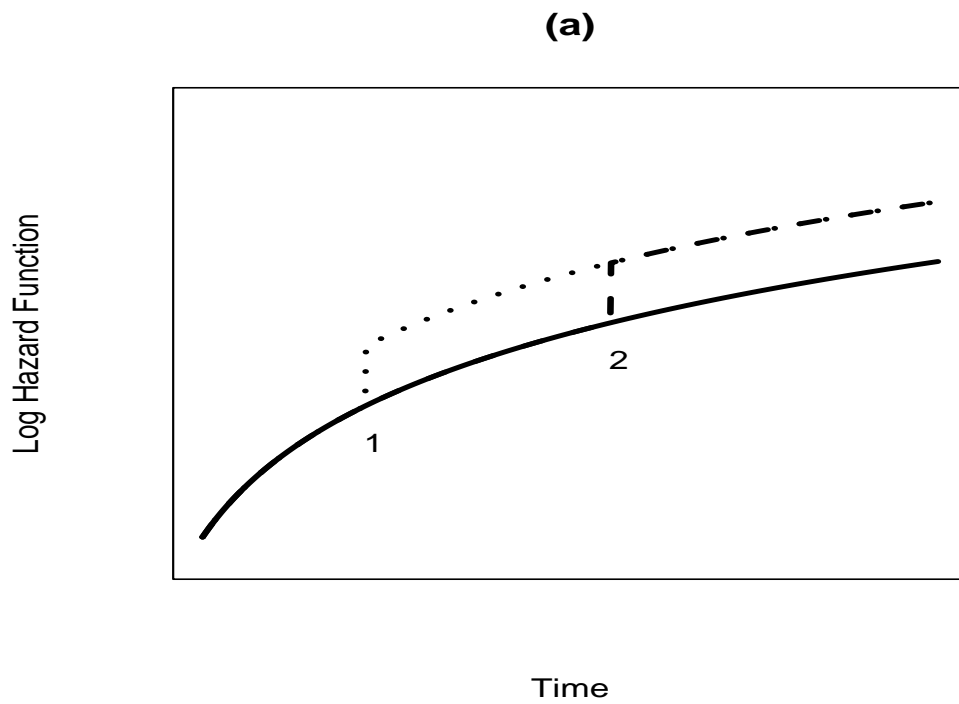


Figure 2:

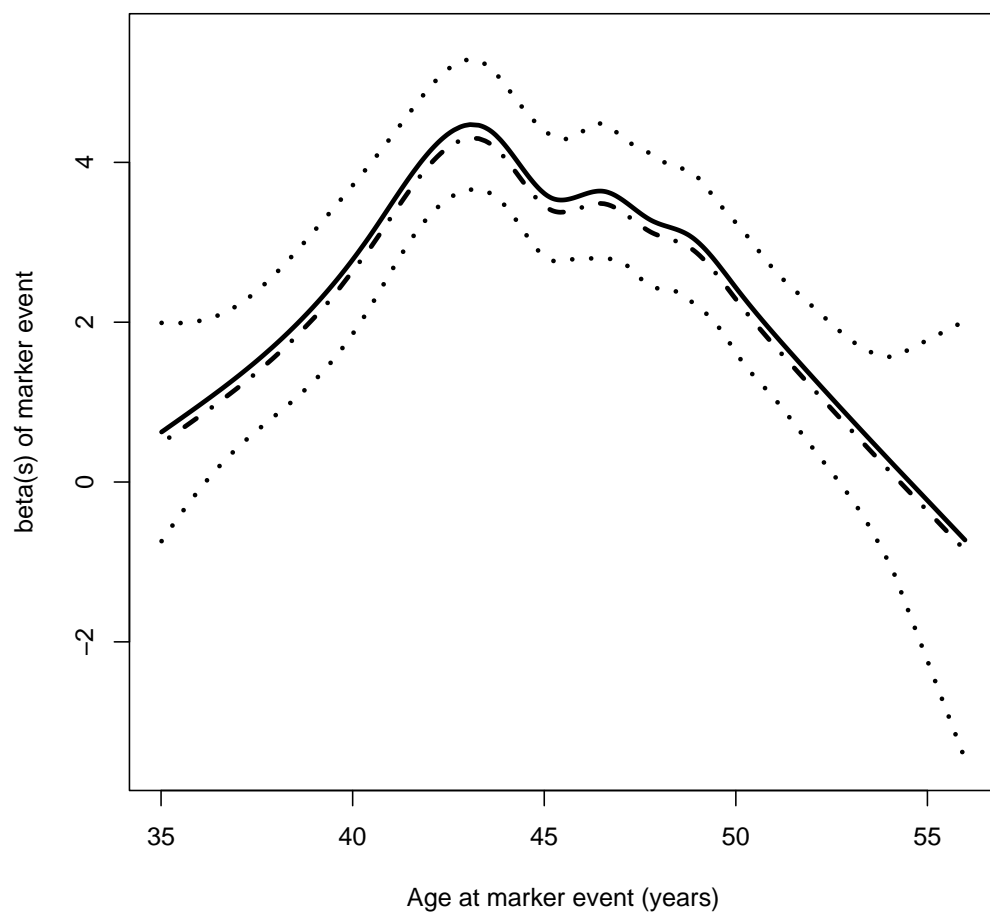


Figure 3:

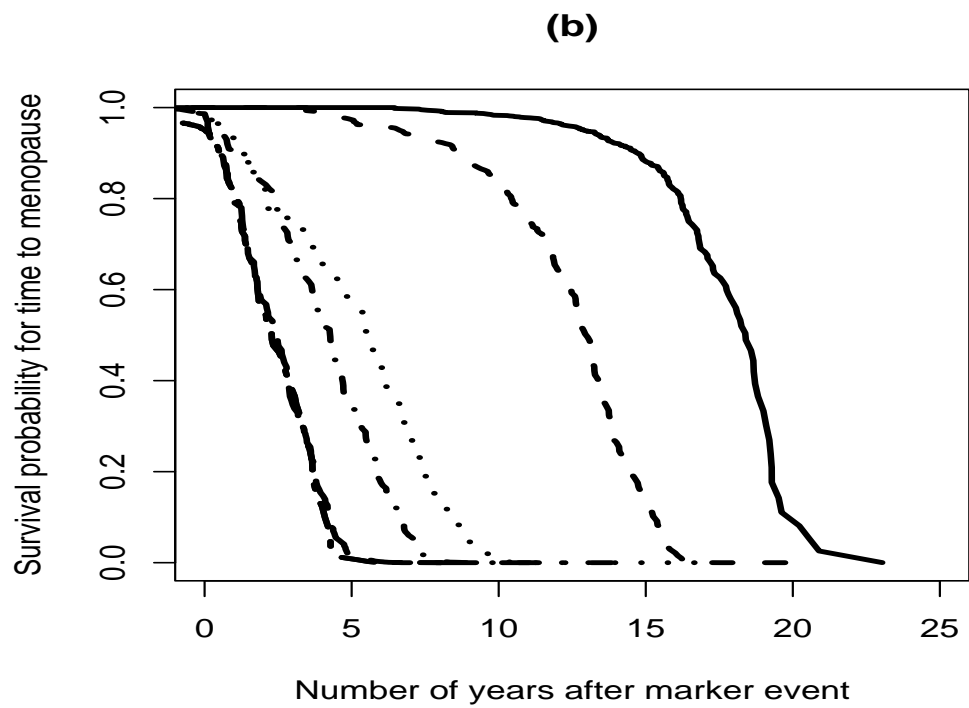
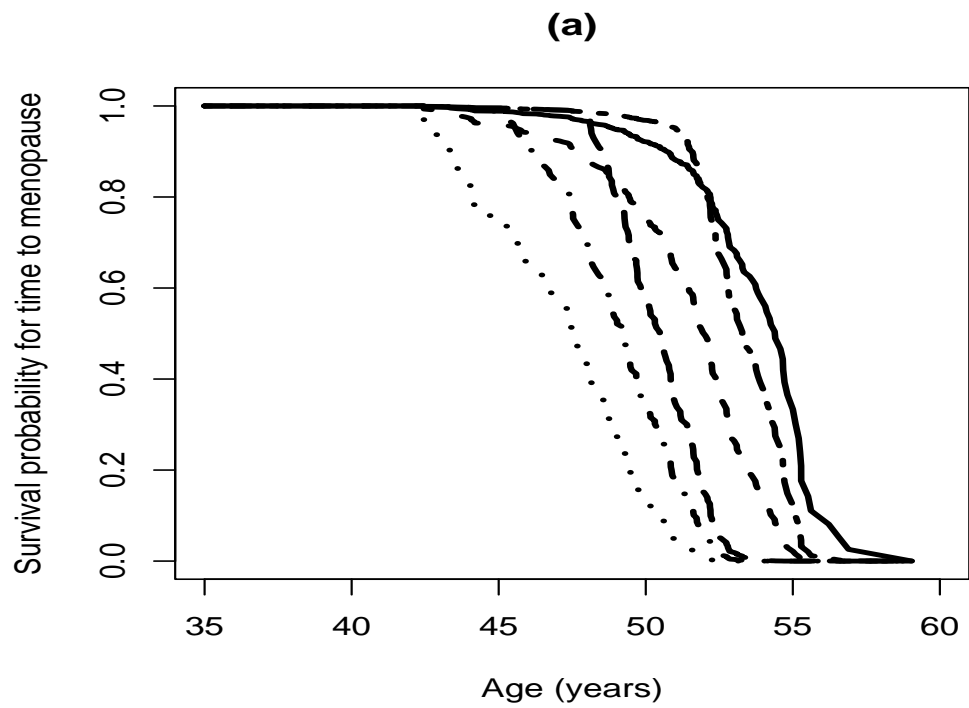


Figure 4:

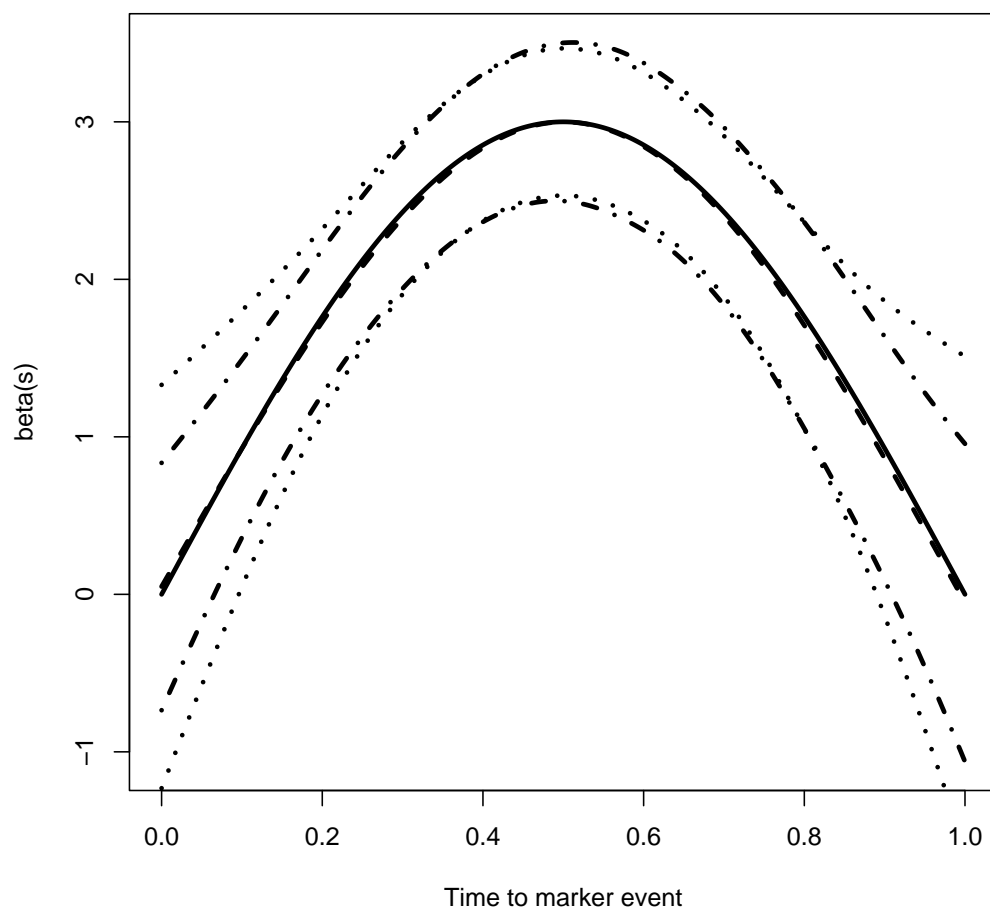


Figure 5: