Vary-Coefficient Cox Models and Cross-Ratio Models for Evaluating the Usefulness of a Marker Event For the Age-at-Onset of Menopause

Xihong Lin

Department of Biostatistics

Harvard School of Public Health

October 2005

Joint work with Bin Nan, Sioban Harlow, Lynda Lisabeth

Outline

- Motivation: Markers for Menopause
- The Tremin Trust Data
- Varying Coefficient Cox Model [1]
- Bivariate Survival Model with Piecewise Constant Cross-Ratios [2]
- Analysis of the Tremin Trust Data
- Simulation Study
- Conclusions

[1] *Biometrics*, 2004

[2] *JASA*, December, 2005

Motivation

- An important component of female reproductive aging research is to establish a staging system, i.e., stages a woman experiences in her reproductive life span.
- The Stages of reproductive Aging Workshop (2001): Interested in identifying early and late stage bleeding pattern based markers of menopause transition.
- The experts proposed 9 such markers. Examples include age at onset of experiencing a menstrual cycle at least 45 days; 60 days, 90 days.

Motivation

- Goal: Evaluate the usefulness of a bleeding pattern based age at onset marker for age at onset of menopause.
-
- Why?
  - Determine a woman's need of contraception
  - Initiate interventions such as hormone replacement
  - Assess the approach of menopause
  - Predict age of onset of menopause using age of onset of a marker.

The Tremin Trust Data

- The largest cohort study of the whole female reproductive life span in the US history with detailed menstrual cycle data.
- Enrolled 1997 white college students in the University of Minnesota in 1950's and followed them up to 40 years.
- Each woman was asked to use menstrual diary cards to record dates of menstrual cycles.
- Menopause is defined as the final menstrual period, confirmed at least 12 months of amenorrhea.
- Covariate information is limited, e.g., age at menarche.
- It provides a unique opportunity to evaluate how good a proposed marker is for age at onset of menopause.

First Look at the Tremin Trust Data: Lisabeth et al (2003)

- A subset of 562 women: age 25+ at enrollment, had age at menarche, still in study at age 35.
- Some descriptive statistics

Type of Event	# of Events	% Censoring	Median Age
Menopause	193	68%	51.7
45-day cycle marker	357	36%	42.7
60-day cycle marker	282	50%	48.7

- Challenges: Both ages at onset of the marker event and menopause are subject to censoring.
- Q: How to evaluate the usefulness of a censored survival marker for a censored survival endpoint of interest?

Connection with joint modeling longitudinal and survival outcomes

- There exists a considerable recent literature on joint modeling a longitidudinal marker and a survival endpoint (See Jeremy's talk).

  Common approach using the mixed-effect and Cox model:

$$\begin{aligned} \lambda_i(t) &= \lambda_0(t)e^{(b_{0i}+b_{1i}t)\beta} \\ W_i(t_{ij}) &= b_{0i} + b_{1i}t_{ij} + e_{ij} \end{aligned}$$

  where $(b_{0i}, b_{1i}) \sim N(\mu, \Sigma)$ and $e_{ij} \sim N(0, \sigma)$.

- How to jointly model a survival marker and a survival endpoint, both subject to censoring?

Paired Marker-Time Specific Box-plots

- Divide age into several intervals: [35,40), [40-43), etc.
- In each age interval, compare the distributions of age to menopause between two groups:

  A: women who have experienced the marker event in the age interval.

  B: women who have not experienced the marker event by the end of the age interval.
- The distribution of age to menopause in each group can be estimated by the KM method.
- Display the KM curves by side-by-side boxplots
- If a good marker, the distributions of age to menopause between the two groups would be different.

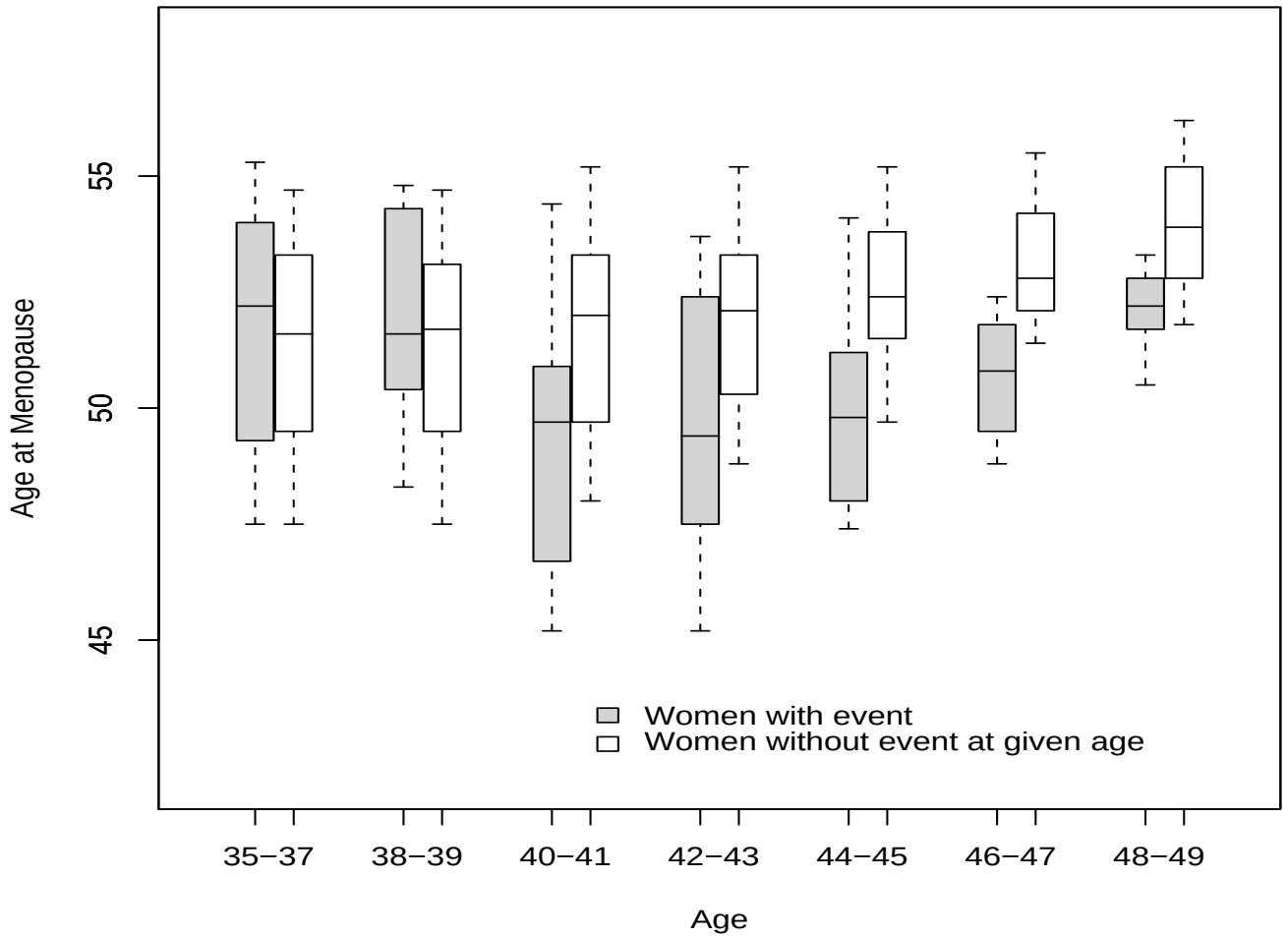


Figure 1: Age specific paired boxplots for the KM curves of age at menopause for women with/without a 45 day cycle.

Observations from of the Paired Boxplots

- The usefulness of the 45-day cycle marker for age at onset of menopause varies with age experiencing the marker event.
- The survival distributions of menopause are similar for women with/without the marker event before age 40, but are different after age 40.
- The 45-day cycle marker does not seem to be useful before age 40 but seems to be a good marker for menopause after age 40.
- How can we develop a formal statistical model to quantify these observations and make statistical inference?

First Try: Varying-Coefficient Cox Model

- Setup:

  Y_i: censored age at menopause
  S_i: true age at marker event (not observed if censored)
  $\boldsymbol{X}_i$: covariates, e.g., age at menarche
  $Z_i(t)$: time-dependent marker indicator (0 if $t \leq S_i$) and 1 if $t \geq S_i$)

- Model

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

  where $\beta(s)$ is a function of the marker event time s and is an unknown smooth function.

- Model (1) allows the relative risk of menopause at time t to depend on the marker event time s.

First Try: Varying-Coefficient Cox Model

- Comparison of the PH model with the varying coefficient model:
  - Consider two subjects: subject 1 experiences the marker event at time 1 and subject 2 experiences the marker event at time 2.
  - Figure (a): PH model $(\beta(s) = \beta)$
    Figure (b): varying coefficient model $\beta(s)$.
- Model (1) is estimable even if S_i is censored, since age at menopause is either censored at this time or observed before this time, i.e., $Z_i(t) = 0$ during the followup.

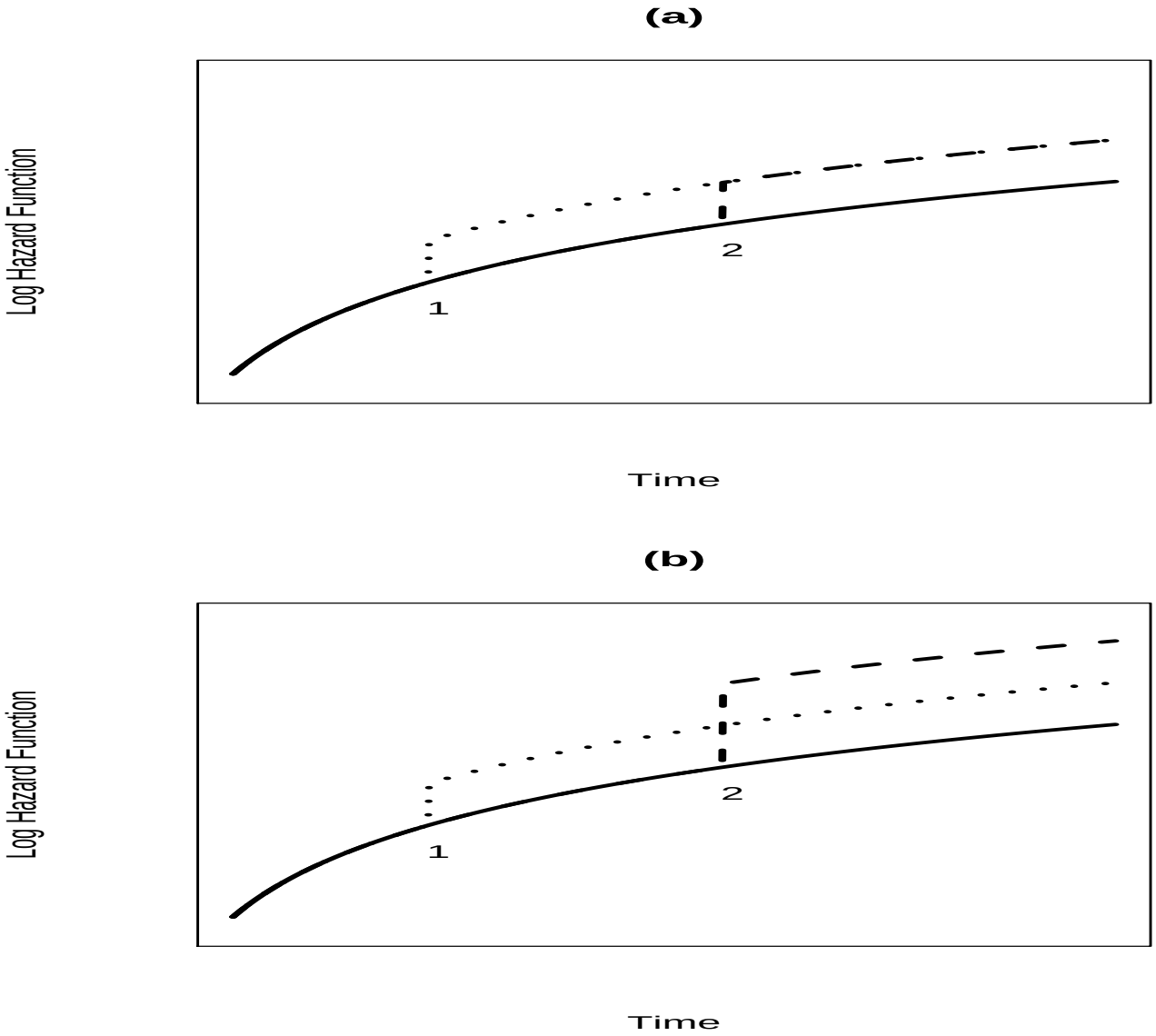


Figure 2: Figure (a): PH model; Figure (b): Varying coeffient model

Estimation in the varying coefficient Cox model

- The coefficient function $\beta(s)$ is estimated nonparametrically using a B-spline, see plot.
- Set $\beta(s) = \sum_{k=1}^{K} \alpha_k B_k(s)$, where the $B_k(s)$ are B-spline bases.
- Fix the Cox model

$$\lambda(t|Z_i(t), X_i) = \lambda_0(t)\exp\{\alpha^T \widetilde{Z}_i(t) + \gamma^T X_i\}$$

where $\widetilde{Z}_i(t) = \{B_1(s)Z_i(t), \cdots, B_K(s)Z_i(t)\}$.

- Advantages:
  - Easy to Fit using the standard Cox model e.g., by SAS PROC PHREG.
  - Convenient for predicting age at menopause using age at the marker event.

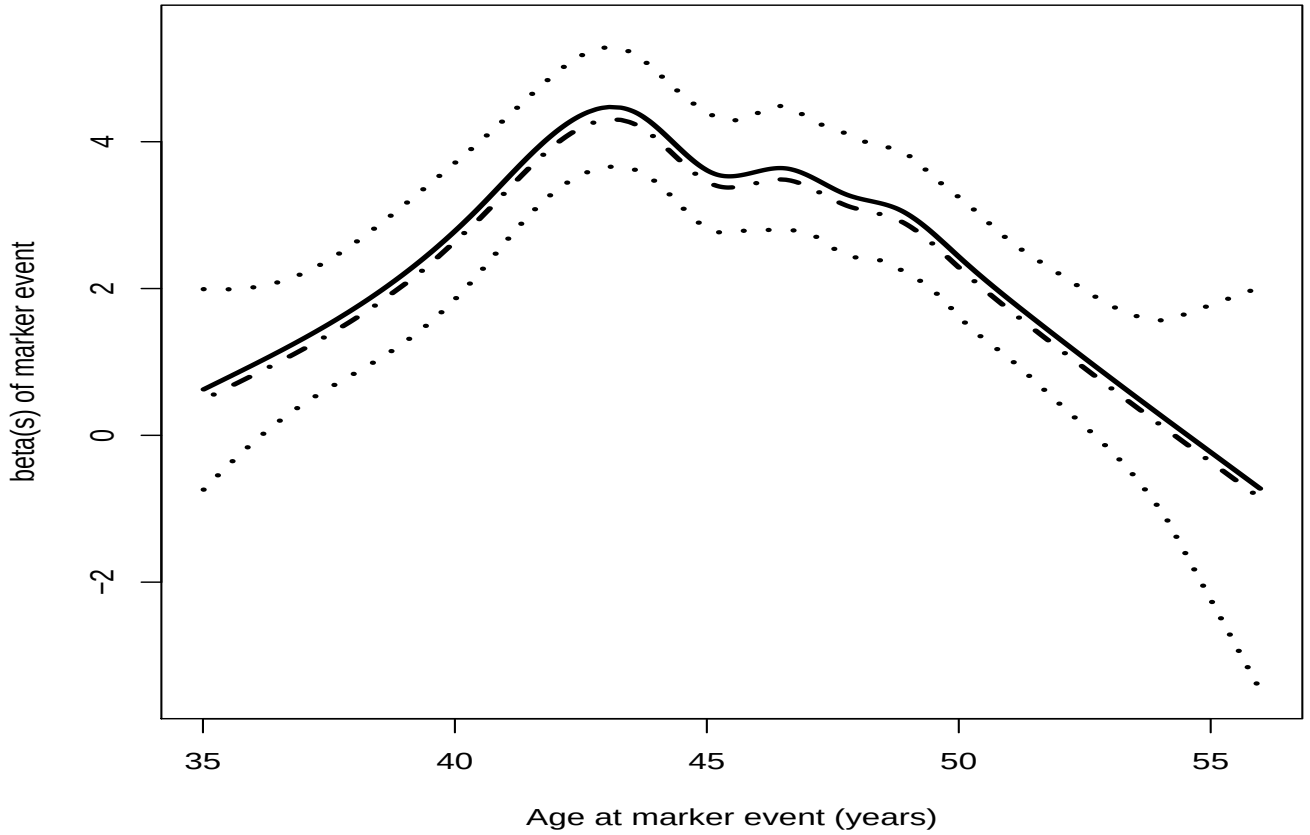


Figure 3: Estimates of $\beta(s)$ using the B-spline in the varying-coefficient model using the 60-day cycle marker: —— B-spline estimate with the covariate; · · · 95% CI; – – – B-estimate without the covariate.

Issues of the Varying-Coefficient Cox Model

- The interpretation of the coefficient function $\beta(s)$ is not intuitive and not particularly of practical interest.
- It is the relative risk comparing those who experience the marker event at age s with those who have never experienced the marker event in her life.
- This is not what the paired Box-plots want to capture: we are interested in the relative risk comparing those who experience the marker event at age s with those who have not experienced the marker event by age s.

Cross-Ratio in Bivariate Survival Data

- Revisit of the paired boxplots: Interested in modeling the relative risk of menopause comparing two groups

  Group A: Women who experience the marker event at age t_1

  Group B: women who have not had the marker event by age t_1

  as a function of marker event age t_1.

- This is exactly what the cross ratio for bivariate survival data measures!
- Consider bivariate survival data: T_1 is time to marker event and T_2 is time to menopause, both subject to censoring.

$$\text{Cross Ratio}: \quad \theta(t_1, t_2) = \frac{\lambda_2(t_2|T_1 = t_1)}{\lambda_2(t_2|T_1 > t_1)}$$

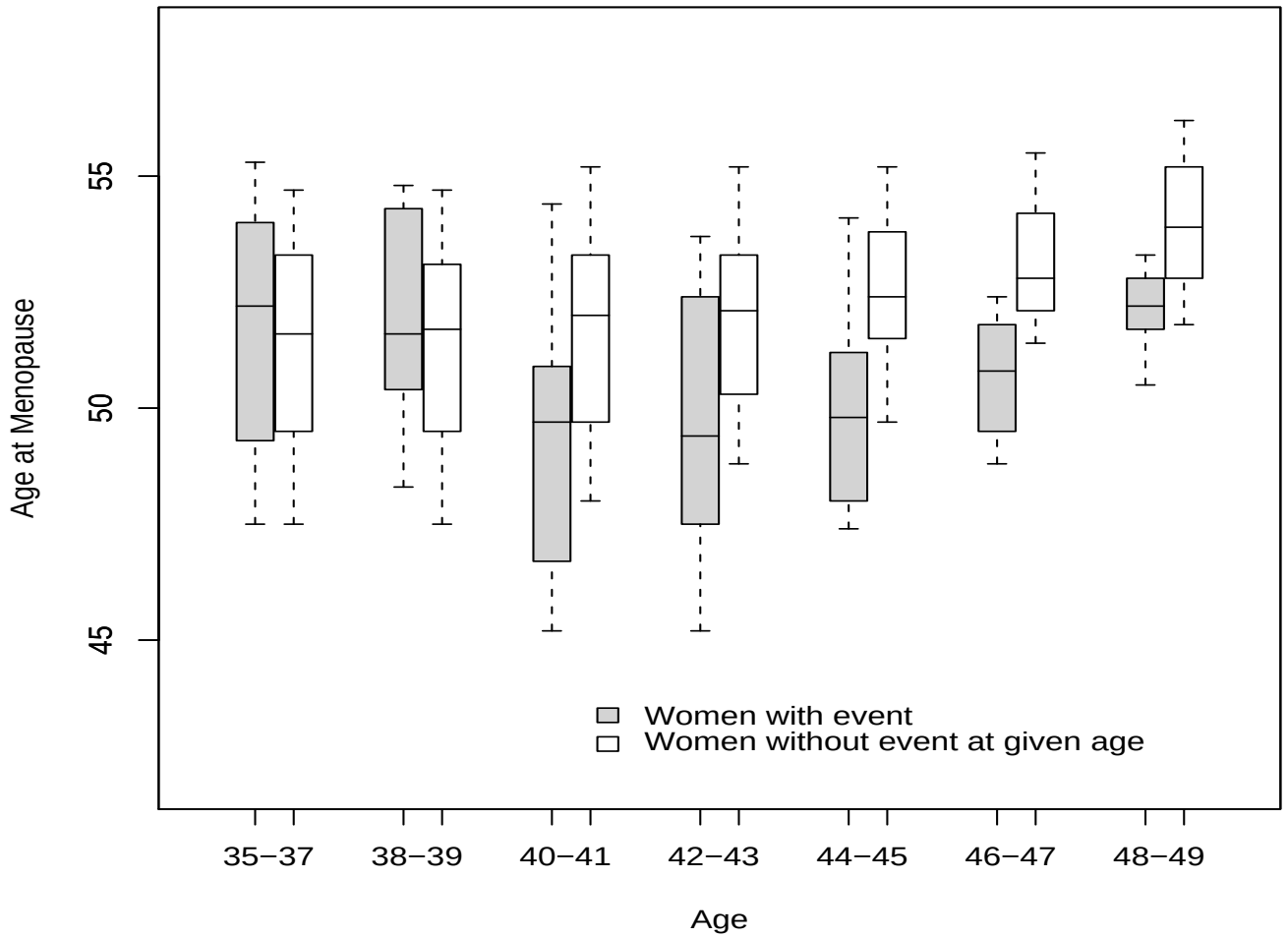


Figure 4: Age specific paired boxplots for the KM curves of age at menopause for women with/without a 45 day cycle.

Piece-wise Constant Cross Ratio Model

- Model: Assume the CR as a nonparametric function of the marker event time t_1 and the Cox model for each marginal distribution.

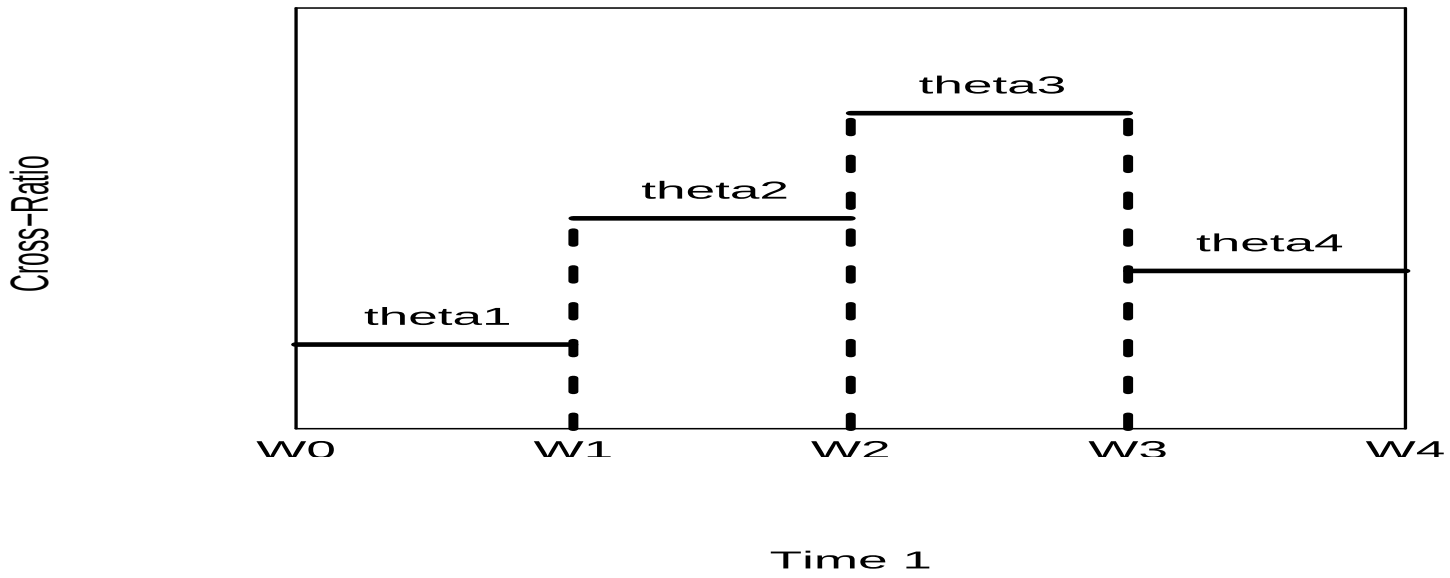
$$\begin{aligned} \theta(t_1, t_2) &= \theta(t_1) \\ \lambda_j(t_j) &= \lambda_{j0}(t_j) e^{\boldsymbol{Z}'_j \boldsymbol{\beta}_j} \quad (j = 1, 2) \end{aligned}$$

- Open question: How to construct a bivariate survial function with two marginals and an arbitrary cross ratio model?
- Estimate $\theta(t_1)$ by a piecewise constant function of the marker event time t_1 by cutting the plane into K strips A_k's using cutoff points $(\omega_1, \cdots, \omega_K)$ of t_1 and setting

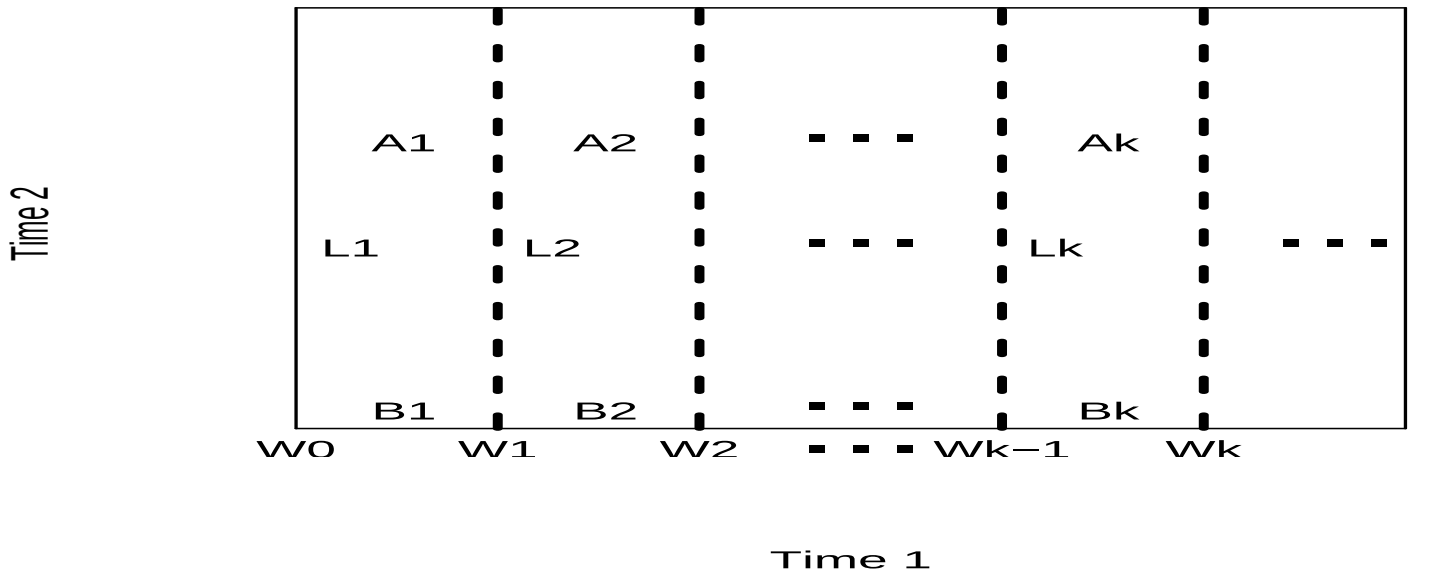
$$\theta(t_1) = \theta_k$$

for $(t_1, t_2) \in A_k$, i.e., $t_1 \in [\omega_{k-1}, \omega_k)$.

(a)


Cross-Ratio
theta1
theta2
theta3
theta4
W0
W1
W2
W3
W4
Time 1

(b)


Time 2
A1
A2
Ak
L1
L2
Lk
B1
B2
Bk
W0
W1
W2
Wk−1
Wk
Time 1

Features of the Model

- The two marginal distributions and the piecewise constant cross ratios fully determine the bivariate survival function in the whole plane and can be constructed sequentially.
- For each strip A_k, given the constant CR θ_k and the left margin (L_k) and the bottom margin (B_k), the joint survival function in A_k is fully specified through the generalized Clayton model:

$$\bar{F}_{A_k}(t_1,t_2) = \Big[\{\bar{F}(t_1,0|\boldsymbol{z})\}^{-(\theta_k-1)} + \{\bar{F}(w_{k-1},t_2|\boldsymbol{z})\}^{-(\theta_k-1)} - \{\bar{F}(w_{k-1},0|\boldsymbol{z})\}^{-(\theta_k-1)}\Big]^{-1/(\theta_k-1)}.$$

where left margin L_k : $\bar{F}(w_{k-1},t_2)$ and right margin B_k : $\bar{F}(t_1,0)$ and $\bar{F}(\cdot,\cdot)$ denotes a survival function.

Features of the Model

- To construct $\bar{F}(t_1, t_2)$, start from the very left strip A_1 to the right:

$$L_1, B_1, \theta_1 \Rightarrow F_{A_1}(t_1, t_2) \Rightarrow L_2$$
$$L_2, B_2, \theta_2 \Rightarrow F_{A_2}(t_1, t_2) \Rightarrow L_3$$
$$\ldots$$

- Check for piece-wise constant CR assumption:

$$\log(-\log\{\bar{F}(t_2|T_1 = t_1)\})$$
$$= \log(-\log(\{\bar{F}(t_2|T_1 > t_1)\}) + log(\theta_k)$$

i.e.,

Plot the log-log survival curves corresponding to the paired boxplots to see whether they are parallel for each strip.

Features of the model

- The left-truncated failure times ($\tilde{t}_1 = t_1 - w_{k-1}, \tilde{t}_2 = t_2$) follow the standard clayton model in each strip.

$$\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2|\boldsymbol{z}) = \Pr(T_1 > w_{k-1} + \tilde{t}_1, T_2 > \tilde{t}_2 | T_1 > w_{k-1}, T_2 > 0, \boldsymbol{z})$$

$$= \left[\{\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0|z)\}^{-(\theta_k - 1)} + \{\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2|z)\}^{-(\theta_k - 1)} - 1\right]^{-1/(\theta_k - 1)}$$

- Implications: For each strip A_k, if the two marginal survival functions on L_k and B_k are known, the existing methods for fitting the Clayton model can be used to estimate the CR θ_k.

Features of the model

- Recall: We assume the marker event time T_1 and the menopause time T_2 follow the Cox model marginally

$$\lambda_j(t_j) = \lambda_{j0}(t_j)e^{\boldsymbol{Z}'_j\boldsymbol{\beta}_j}.$$

- This implies the bottom margins of all the strip $\{B_1, \cdots, B_K\}$ follow the Cox model,

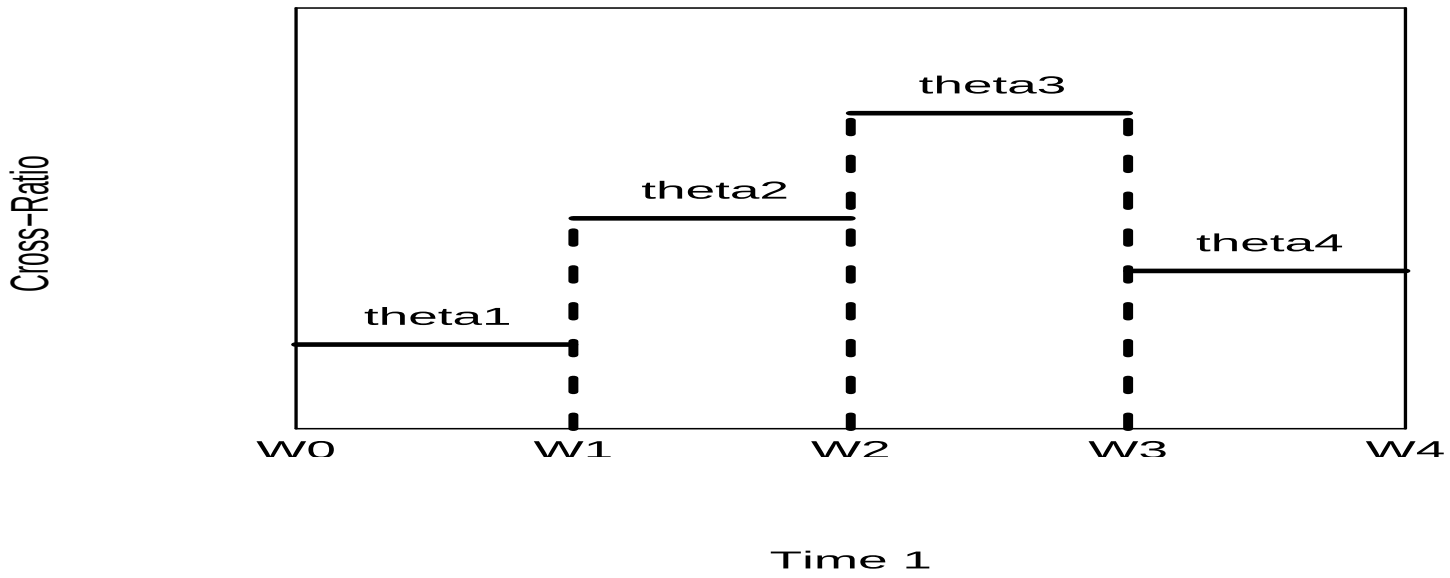
Features of the model

- The left margin of the first strip L_1 follows the Cox model. However, the left margins of the left margins on the other strips ($L_k, k > 1$) do not follow the Cox model if $\theta_k \neq 1$!

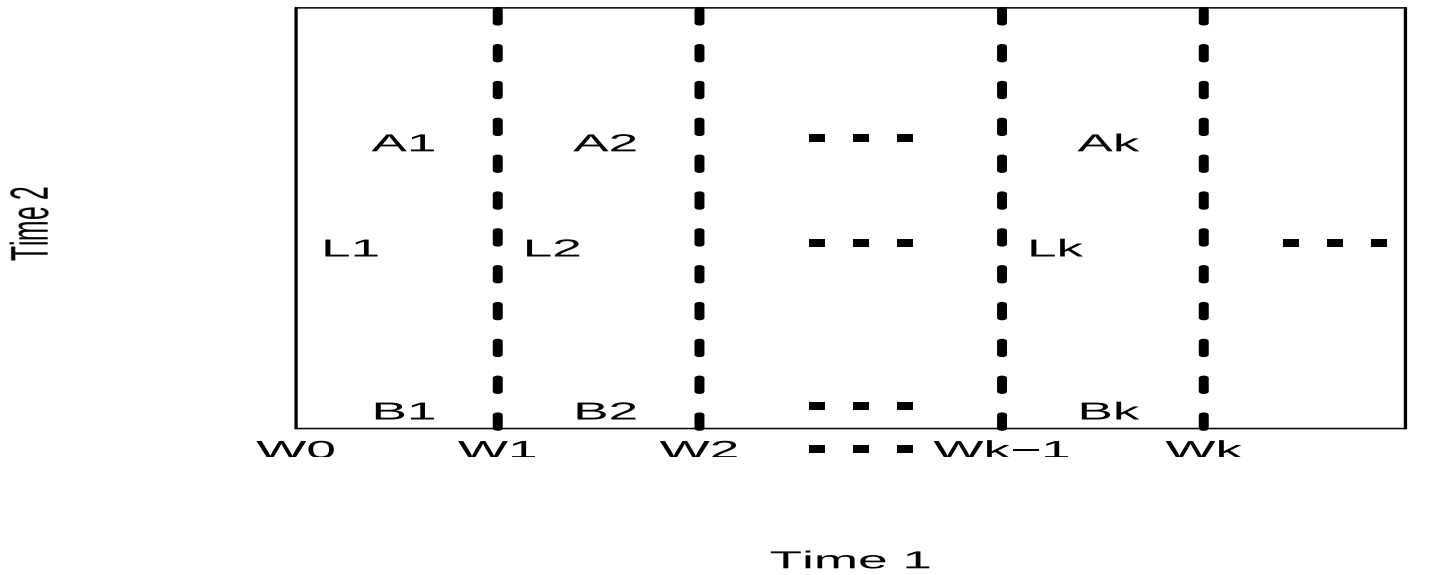
$$\begin{aligned}
\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2|z) &= \bar{F}_{A_k}(w_{k-1}, t_2|z)/\bar{F}_{A_k}(w_{k-1}, 0|z) \\
&= \Big[1 + \{\bar{F}_{A_{k-1}}(w_{k-2}, t_2|z)/\bar{F}_{A_{k-1}}(w_{k-1}, 0|z)\}^{-(\theta_{k-1}-1)} \\
&\quad -\{\bar{F}_{A_{k-1}}(w_{k-2}, 0|z)/\bar{F}_{A_{k-1}}(w_{k-1}, 0|z)\}^{-(\theta_{k-1}-1)}\Big]^{-1/(\theta_{k-1}-1)}.
\end{aligned}$$

- The distribution on the left margin L_k of the kth strip A_k depends on the marginal distributios of T_1 and T_2 $\{\lambda_{01}(t_1), \lambda_{02}(t_2), \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ and all the CRs of the previous strips θ_l ($l < k$).

(a)


Cross-Ratio
theta1
theta2
theta3
theta4
W0
W1
W2
W3
W4
Time 1

(b)


Time 2
A1
A2
Ak
L1
L2
Lk
B1
B2
Bk
W0
W1
W2
Wk−1
Wk
Time 1

Estimation Method I: Two-stage Direct Method

- Key idea: Treat each strip data separately and apply the estimation procedure for the Clayton model.
- Procedure:
  1. For strip A_k, left truncate the data by ω_{k-1} and right censor the data by ω_k.
  2. Stage 1: Estimate the left (L_k) and bottom (B_k) marginal survival functions by the KM.
  3. Stage 2: Estimate θ_k by MLE by plugging in the two marginal survival function estimates (Shih and Louis, 95).
- Advantage: Easy.
- Limitations:
  (1) The survival function might not be proper.
  (2) does not work with covariates, since the left margins of the strips do not follow the Cox model except for L_1.

Estimation Method: Two-stage Sequential Method

- Key idea: Sequentially estimate θ_k's from the very left strip A_1 to the right by following the way $F(t_1, t_2)$ is constructed.
- Procedure:
  1. Estimate the bottom marginal survival function using the KM or the Cox model in the presence of covariates.
  2. Create left truncated and right censored data for each strip.
  3. Stage 1: Determine the left marginal survival function L_k by the joint survival function estimated from the previous strip.
  4. Stage 2: Estimate θ_k by MLE by plugging in the two marginal survival function estimates.
- Advantages: (1) The bivariate survival function is proper; (2) Allows for covariates; (3) More efficient estimates.
- Disadvantage: Computationally more intensive.

Analysis of the Tremin Trust Data

- Marker: age at onset of a 45-day cycle.
- Fit the piece-wise cross ratio models assuming 4-8 age intervals, and assume marginally ages at onset of the 45-day cycle marker and menopause follow the Cox models with covariate age at menarche.
- Key finding: The 45-day cycle marker is non-informative for age at onset of menopause before 40 and is informative between 40-50 and less informative after age 50.
- Estimate menopause distribution given age at onset of the 45 day cycle marker.

Age at the marker event	36	39	42	45	48	51
Median Age at menopause	52.2	52.2	50.2	51.1	51.7	53.9

Cross-Ratio Estimates of the Tremin Trust Data Using Direct Method w/o Covariates

T_1	$\hat{\theta}\,(SE_{SL}, SE_{BS})$	$\hat{\theta}\,(SE_{SL}, SE_{BS})$
35-37	0.80 (0.14, 0.15)	0.80 (0.12, 0.11)
38-39	0.83 (0.20, 0.17)	
40-41	2.44 (0.50, 0.70)	2.19 (0.27, 0.33)
42-43	1.98 (0.47, 0.71)	
44-45	2.18 (0.39, 0.63)	
46-47	4.48 (1.19, 1.92)	3.64 (0.71, 0.78)
48-49	3.18 (0.81, 0.97)	
50+	1.57 (0.40, 0.55)	1.57 (0.40, 0.55)

Cross-ratio estimates of the Tremin Trust data Using the Sequential Method

	Without covariate		With covariate
T_1	$\hat{\theta}\ (SE_{BS})$	$\hat{\theta}\ (SE_{BS})$	$\hat{\theta}\ (SE_{BS})$
35-37	0.80 (0.15)	0.80 (0.10)	0.81 (0.11)
38-39	0.82 (0.18)		
40-41	2.48 (0.74)		
42-43	2.04 (0.77)	2.17 (0.35)	2.17 (0.34)
44-45	2.12 (0.70)		
46-47	5.72 (1.95)	4.11 (0.85)	4.41 (0.94)
48-49	3.28 (1.10)		
50+	1.39 (0.38)	1.37 (0.41)	1.47 (0.47)

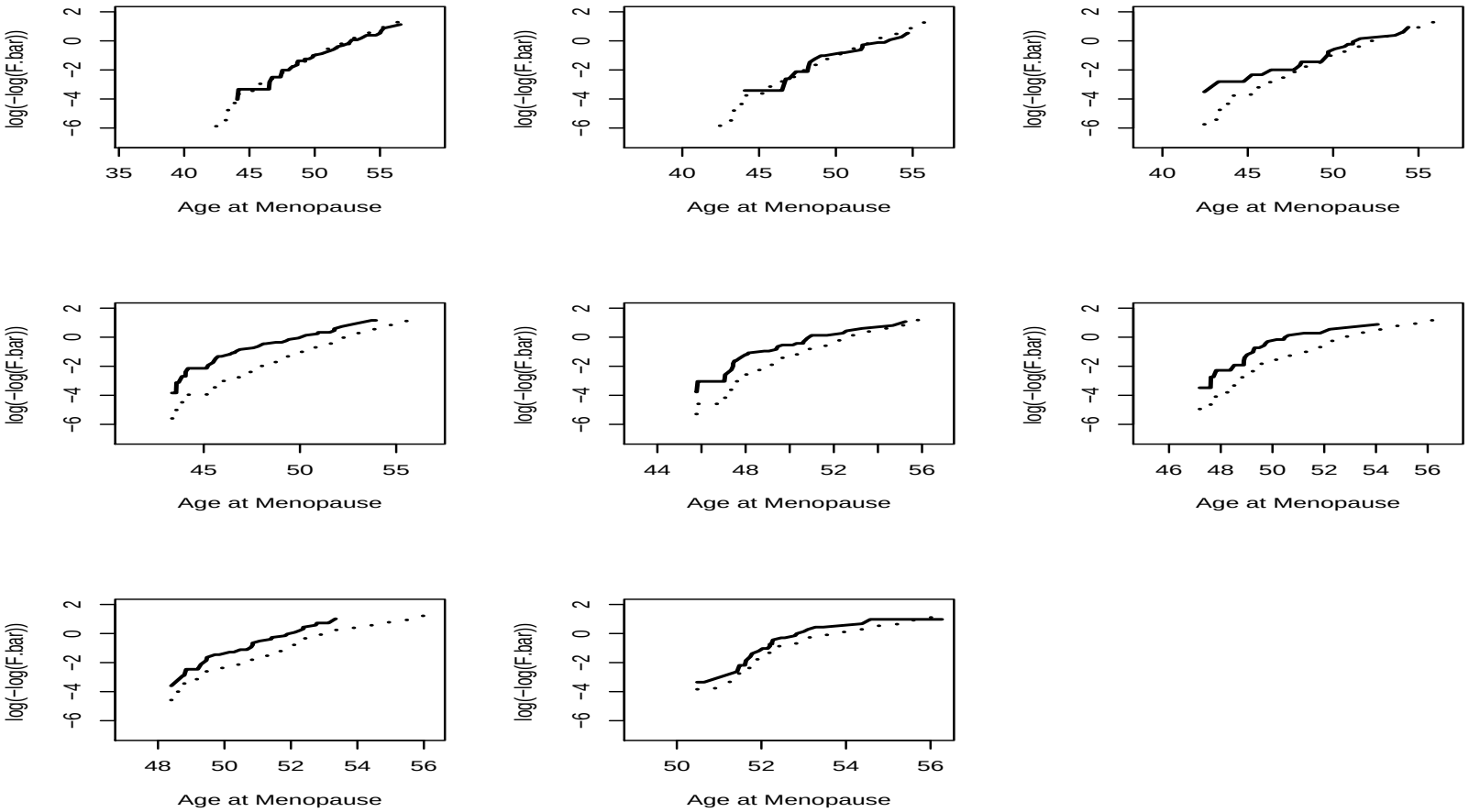


Figure 5: Checking the piece-wise constant cross ratio assumption of $\theta(t_1)$: $t_1 = 36.5, 38, 40, 42, 44, 46, 48,$ and 50.

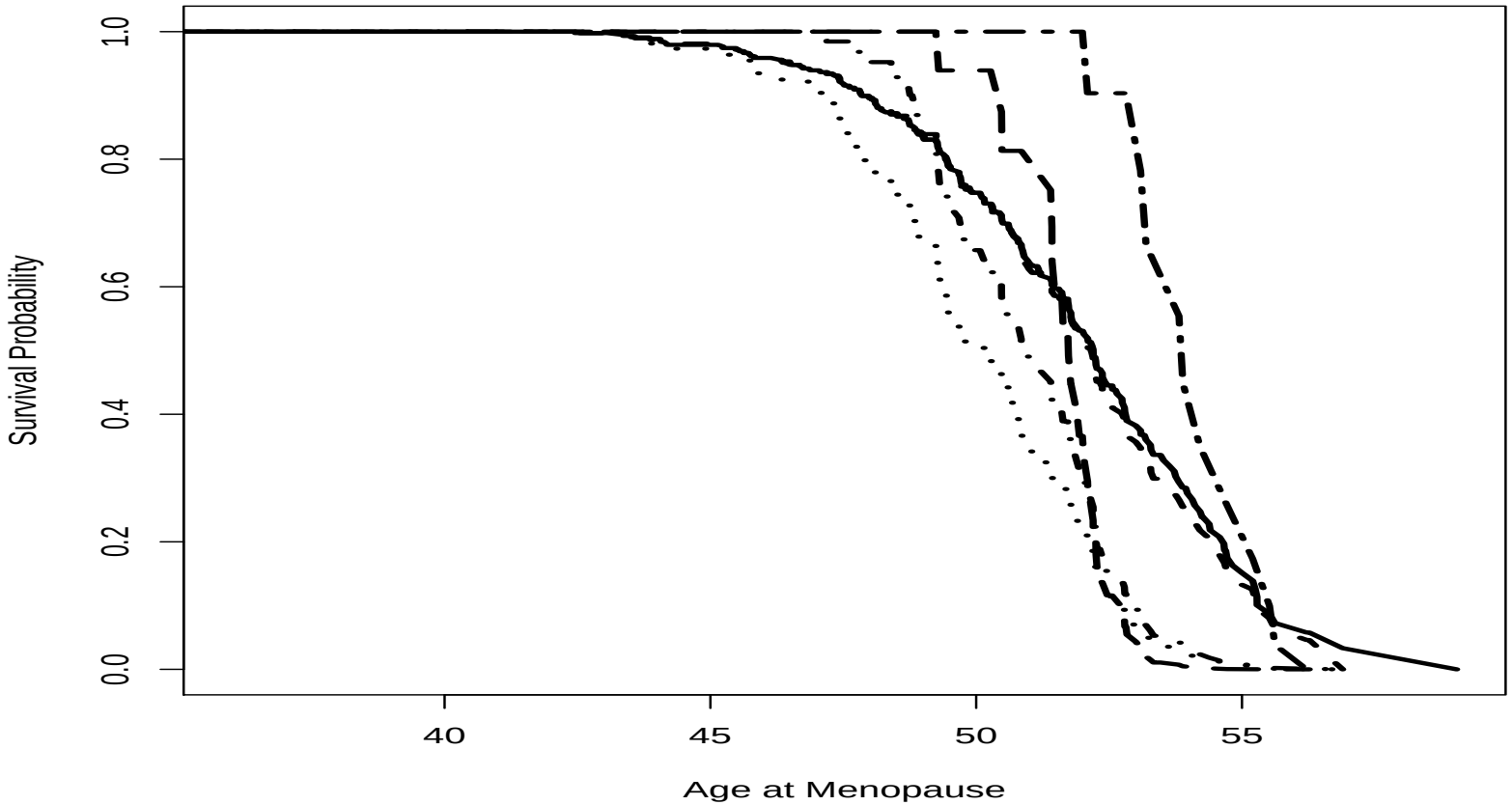


Figure 6: Estimated survival functions for time to menopause given age at the 45-day cycle marker event: —— Age 36; – – – Age 39; · · · Age 42; – · – Age 45; — — Age 48; — · — Age 51.

Simulations

Simulation results for cross-ratio estimates based on 100 replications, sample size=500, 4 intervals, with covariates in marginal Cox models.

	Direct Method		Sequential Method	
θ	$\hat{\theta}$	SE_E	$\hat{\theta}$	SE_E
0.9	0.89	0.08	0.89	0.08
2.0	2.13	0.22	2.02	0.27
4.0	2.35	0.33	3.98	0.68
1.5	1.52	0.31	1.51	0.30

Conclusions

- Two methods are provided to evaluate the informativeness of a survival marker for a survival endpoint.
- The varying coefficient model provides a convenient model to estimate age at menopause given age at a marker event.
- The piece-wise cross ratio model provides parameters with attractive and easy RR interpretations.
- The direct method is easy but does not allow for covariates and the survival function might not be proper. These difficulties are overcome by the sequential method.
- For future research it would be useful to develop a nonpametric MLE method.
- Open question: How to construct of a bivariate survival function using two marginals and an arbitrary cross-ratio regression function.