Bootstrapping methods for assessing causality in survival analysis

A case study on major adverse cardiovascular events

Paulina Benthem Ciano

Master's thesis
2023:E42

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

The combination of graphical models with Aalen's additive hazards model, resulting in a model known as dynamical path analysis, permits assessing the effects of different variables on times until an event and decomposing these total effects into direct and indirect effects. This thesis proposes novel bootstrapping methods designed for left-truncated right-censored data, conditional on covariates within the framework of Aalen's additive hazards model, in order to obtain confidence intervals for the estimates.

To illustrate the practical application of the bootstrapping methods, we conduct a case study utilising data from the Malmö diet and cancer study. The data set consists of left-truncated right-censored data. Our analysis aims to examine causality and estimate the direct effects of various covariates on the incidence of major adverse cardiovascular events and indirect effects between covariates. We compute confidence intervals for these effects with the proposed bootstrapping methods.

Acknowledgements

I would like to thank my supervisor, Dragi Anevski, for his guidance and support throughout this thesis. Additionally, I am grateful to Olle Melander for providing the data for the case study and to my examiner, Magnus Wiktorsson, for his valuable feedback.

Thank you to my parents for giving me the opportunity to study abroad; and to Mikkel and my friends for accompanying me and making sure I had fun along the way.

Contents

1 Introduction

Even though some events happen without a reason, many others can be attributed to underlying factors. However, studying causality in statistics requires caution due to the difficulty in differentiating statistical dependence from causal dependence. The study of causal effects in times to an event is possible with graphical models and regression. However, there are no asymptotic results for these effects. Bootstrapping, a non-parametric method, can be applied in situations where no asymptotic results are known. This thesis proposes bootstrapping methods for causality effects in the context of survival analysis.

Wright (1921) pioneered path analysis, a methodology used to examine causal relationships between variables [17]. It utilises graphical models to visualise these dependencies. However, path analysis fails to consider time in the model and causality relations vary with time. Fosen et al. (2005) introduced dynamic path analysis as a solution to this limitation [7]. This method combines graphical models with continuous time development to study the temporal dependencies among variables. Dynamic path analysis can investigate causal dependencies in survival analysis problems. Different regression models can be used to estimate the effects of covariates on the time until an event using dynamic path analysis. Aalen et al. proposed combining dynamic path analysis with Aalen's additive hazards model [1]. This approach enables the estimation of effects and the decomposition into direct and indirect components. The regression functions of Aalen's additive hazards model determine the direct effect of the covariates on the event times. The indirect effects between the covariates are estimated by regressing each covariate on its parents by (multiple) linear regression. This thesis investigate different bootstrapping techniques to compute confidence intervals for the estimates due to the lack of asymptotic results.

Bootstrapping is a non-parametric method introduced by Efron (1979) that can assess the uncertainty of statistics without making assumptions about the underlying distribution of the data [5]. Given a data set with sample size n, bootstrapping consists of sampling with replacement from the data set to obtain bootstrapped samples of size n. By repeatedly computing the statistics for different bootstrapped samples, one can analyse the distribution of the statistics. Efron (1981) developed two bootstrapping methods, known as the 'simple' and the 'obvious' method, developed to handle right-censored data, and they

are, in fact, equivalent [6]. Wang (1991) proposed a generalisation of Efron's 'obvious' bootstrapping method to sample from left-truncated right-censored (l.t.r.c.) data under some assumptions [16]. Gross and Lai (1996) discussed the generalisation of Efron's 'simple' bootstrapping method for l.t.r.c. data, which does not rely on Wang's assumptions. The 'obvious' and the 'simple' method for l.t.r.c. data are not equivalent, c.f. [16]. Burr (1994) studied bootstrapping considering covariates and suggested two methods for bootstrapping right-censored data within the Cox regression model framework [3]. We modify the methods presented by [3] and [16] for bootstrapping right-censored and l.t.r.c. data conditional on the covariates within Aalen's additive hazards model framework.

This thesis considers a method presented for l.t.r.c. data under Aalen's model. In order to showcase the feasibility of the method, it is applied on survival data. The data used comes from the Malmö diet and cancer study, having participants entering the study in their middle age, the data is left-truncated. The data is also right-censored since individuals are followed until they experienced a major adverse cardiovascular event (MACE), or until death, emigration or the end of the study. Thus, this data is a suitable basis for illustrating the applicability of the method.

The thesis is organised as follows:

In Section 2, we introduce the mathematical foundations required for our study. In particular, we explain the Doob-Meyer decomposition, which allows us to decompose counting processes into martingales and predictable processes.

Section 3 focuses on survival analysis with right-censored data, commonly encountered in longitudinal studies and clinical trials. Right-censored data refers to situations where individuals may not experience the event of interest within the study period. To handle right-censored data, we give an overview of several methods. From a counting-process approach, we discuss estimators of the survival function and cumulative hazard rates, specifically the Kaplan-Meier and Nelson-Aalen estimators. We then examine two regression models, the Cox proportional hazards model and Aalen's additive hazards model, to analyse the influence of covariates on survival. Furthermore, we investigate the study of causality in survival analysis by combining dynamic path analysis with Aalen's additive hazards model. This approach allows us to examine temporal dependencies among variables and estimate direct, indirect, and total effects. To finish the section, we examine bootstrapping methods for sampling from

right-censored data. We provide an overview of the methods proposed by Burr in [3] to sample from right-censored data considering covariates, within the framework of Cox proportional hazards model. We present a modification of these methods to sample from right-censored data within the framework of Aalen's additive hazards model.

Section 4 introduces modifications to the methods presented in Section 3 to handle l.t.r.c. data. This type of data occurs when individuals are not observed from the origin but conditionally on having survived until a specific point. We give an overview of the literature for bootstrapping l.t.r.c. data and propose bootstrapping methods to sample from this data within the framework of Aalen's additive hazards model.

In Section 5, we present data from the Malmö diet and cancer study and hypothesise causal relationships between the covariates and the outcome of interest, MACE. We apply one of the proposed bootstrapping methods for l.t.r.c. data within Aalen's additive hazards model framework (Method 2) to the data set. We present and interpret the resulting bootstrapped confidence intervals for the causality effects. Lastly, we compare these results with results obtained using sampling with replacement from the data. This naive approach is Efron's 'simple' method for l.t.r.c. data (Method 1).

Finally, in Section 6, we present a conclusion and some open problems.

2 Mathematical background

2.1 Processes, filtrations and stopping times

To model the occurrence in time of random events, begin by fixing a continuous and finite time interval $\mathcal{T} = [0, \tau]$. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ is defined as an increasing right-continuous family of sub-σ-algebras of $\mathcal{F}$. The σ-algebra $\mathcal{F}_t$ contains all events up to time t. The σ-algebra $\mathcal{F}_{t-}$ is the smallest σ-algebra containing all $\mathcal{F}_s$ such that $s < t$, i.e. it contains all events that happen strictly before time t.

A stochastic process $X = \{X(t)\}_{t\in\mathcal{T}}$ is said to be adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ if $X(t)$ is $\mathcal{F}_t$-measurable for all t, which means that at time t, the value of $X(s)$ is known for all $s \leq t$.

A realisation of the process X can be seen as a function of t, and this function is referred to as a sample path. The process X is called càdlàg (continue à droite, limité à gauche) if its sample paths are right-continuous with left hand limits.

A random variable C taking values in $\mathcal{T}$ is defined as a stopping time if $\{C \leq t\}$ is $\mathcal{F}_t$-measurable for all t, i.e. at time t it is known whether $C \leq t$ or $C > t$.

Given the process X and a stopping time C, the stopped process X^C is defined as

$$X^C(t) = X(t \wedge C),$$

where $t \wedge C$ denotes the minimum of t and C. If X is càdlàg and adapted and C is a stopping time, then X^C is càdlàg and adapted, cf. [2].

2.2 Martingales

A stochastic process $M = \{M(t); t \in \mathcal{T}\}$ is a martingale relative to the filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ if it is adapted to the filtration, is integrable and satisfies the

martingale property given by

$$\mathrm{E}(M(t)|\mathcal{F}_s) = M(s), \text{ for all } s \leq t.$$

If the process satisfies the inequality

$$\mathrm{E}(M(t)|\mathcal{F}_s) \geq M(s), \text{ for all } s \leq t,$$

then it is a sub-martingale. Any non-decreasing process, like a counting process, is a sub-martingale.

A stochastic process H is called predictable if, loosely speaking, its value is known just before time t. Sufficient conditions for H to be predictable are that H is adapted to the filtration $\mathcal{F}_t$ and that all its sample paths are left-continuous.

2.3 Doob-Meyer decomposition

The Doob-Meyer decomposition states that any sub-martingale can be decomposed uniquely into the sum of a martingale and a predictable process. Specifically, let $X = \{X(t); t \geq 0\}$ be a sub-martingale relative to a filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$, then there exists a unique decomposition of X of the form

$$X(t) = \tilde{X}(t) + M(t),$$

where $M = \{M(t); t \in \mathcal{T}\}$ is a mean-zero martingale, and $\tilde{X} = \{\tilde{X}(t); t \in \mathcal{T}\}$ is a non-decreasing predictable process, often denoted as the compensator of X. So,

$$d\tilde{X}(t) = \mathrm{E}(dX(t)|\mathcal{F}_{t-}), \tag{2.1}$$

and

$$dM(t) = dX(t) - \mathrm{E}(dX(t)|\mathcal{F}_{t-}). \tag{2.2}$$

Therefore, the process X can be decomposed into a predictable part $\tilde{X}$, the sum of the conditional expectations (2.1); and an innovation part M, the sum of the increments minus the conditional expectations (2.2).

2.4 Predictable and optional variation processes

The predictable variation process $\langle M \rangle$ and the optional variation process $[M]$ are defined as the following limits in probability

$$\langle M \rangle(t) = \underset{n\to\infty}{\text{P-lim}} \sum_{k=1}^{n} \text{Var}(\Delta M_k | \mathcal{F}_{(k-1)t/n})$$

and

$$[M](t) = \underset{n\to\infty}{\text{P-lim}} \sum_{k=1}^{n} (\Delta M_k)^2,$$

where the interval $[0, t]$ is partitioned into n sub-intervals of equal length, and $\Delta M_k = M_{kt/n} - M_{(k-1)t/n}$ is the increment of the martingale over the kth sub-interval. It can be shown that $M^2 - \langle M \rangle$ and $M^2 - [M]$ are mean-zero martingales [1]. Thus, since $M(t)$ is a mean-zero martingale,

$$\text{Var}(M(t)) = \text{E}(M(t)^2) = \text{E}(\langle M \rangle) = \text{E}([M]).$$

The predictable and the optional variation processes are unbiased estimators of the variance of $M(t)$.

2.5 Counting processes

The counting process $N_i = \{N_i(t); t \in \mathcal{T}\}$, $i = 1, ..., n$, are n adapted càdlàg process with piece-wise constant and non-decreasing paths, with jumps of size 1 at event times. Assume that none of the n processes jumps simultaneously. Then, the aggregated process, given by

$$N(t) = \sum_{i=1}^{n} N_i(t),$$

is also a counting process.

Consider a small interval $[t, t + dt)$ and assume that, at most, one event can happen during the time interval. The intensity process $\lambda_i(t)$ of a counting process $N_i(t)$ with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ is the probability of the

event occurring given the information prior to time t divided by the length of the interval

$$\lambda_i(t)dt = P(dN_i(t) = 1|\mathcal{F}_{t-}),$$

where $dN_i(t)$ equals the number of jumps of process i in $[t, t+dt)$. Since $dN_i(t)$ is Bernoulli distributed, $\lambda_i(t)dt$ also equals

$$\lambda_i(t)dt = \mathrm{E}(dN_i(t)|\mathcal{F}_{t-}).$$

Let $\Lambda_i = \int_0^t \lambda_i(u)du$ be the cumulative intensity of the counting process N_i. Then, $\Lambda(t) = \sum_{i=1}^n \Lambda_i(t)$ is the compensator of the aggregated counting process $N(t)$. Since the process $N(t)$ is non-decreasing, it is a sub-martingale. By the Doob-Meyer decomposition, $N(t)$ can be decomposed into its compensator and a zero-mean martingale

$$N(t) = \Lambda(t) + M(t).$$

Thus, the increments of the counting process can be written as

$$dN(t) = \lambda(t)dt + dM(t). \tag{2.3}$$

3 Survival analysis for right-censored data

Given data times to an event, one might be interested in calculating the probability that an individual survives past a specific time. Let the data times to an event, $t_1, ..., t_n$, be observations of the independent identically distributed positive random variables $T_1, ..., T_n$ with distribution function F.

The survival function S specifies the unconditional probability that the event of interest has not happened by time t, and is given by

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= P(T > t). \end{aligned}$$

A goal of survival analysis is to estimate S or, equivalently, estimate F.

If the data times $t_1, ..., t_n$ are observed, then the empirical distribution function F_n is the optimal estimator of F. The empirical distribution function is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{t_i \leq t\}, \tag{3.1}$$

where n is the total number of observations.

Survival analysis deals with situations where not all event times are directly observed. Instead, the data comes as a mixture of complete and incomplete observations. For instance, in observational studies, individuals are monitored until the event of interest occurs or until a predetermined endpoint, resulting in right-censored observation times. In the case of left-truncation, the observation is not recorded from the beginning but is instead conditioned on surviving until a specific starting point.

However, it is important to note that there are other types of survival data not considered in this thesis, namely interval-censored data and left-censored data. Interval-censored data refers to situations where the exact timing of an event is unknown but falls within a certain time interval. This type of data typically arises when information is gathered during follow-up visits, with no additional data available between visits. Furthermore, left-censored data occurs when the event process is not observed from the beginning, but it is known whether the event has occurred prior to entering the study. This type of data occurs in, for example, registers, such as population cancer registers.

3.1 Right-censored data

Survival analysis often deals with incomplete data as not all individuals in the study might experience the event of interest. As a result, not all of the samples $t_1, ..., t_n$ from the underlying random variables $T_1, ..., T_n$ are observed. Instead, the observed data consists of

$$\tau_i = \min(T_i, C_i),$$

for $i = 1, ..., n$, where C_i is the stopping or right-censoring random variable for individual i and C_i is assumed to be independent of T_i. For each individual, it is known whether the event time was observed, referred to as an exact observation, $T_i \leq C_i$, or was censored, $T_i > C_i$. The indicator δ_i, defined by

$$\delta_i = \mathbb{1}\{T_i \leq C_i\},$$

denotes whether the event time was observed or censored.

The right-censored processes N_i, $i = 1, ..., n$, are defined as

$$N_i(t) = \mathbb{1}\{\tau_i \leq t, \delta_i = 1\},$$

so that the counting process jumps to one at the time of an exact observation.

Right-censoring may alter the intensities of the events of interest. If the censoring preserves the intensity processes of the counting processes, then it is said to be independent. Formally, the censoring is independent if it holds that

$$P(\tau_i \in [t, t + dt), \delta_i = 1 | \tau_i \geq t, \mathcal{F}_{t-}) = P(T_i \in [t, t + dt) | T_i \geq t),$$

see [1].

The intensity process $\lambda_i(t)$ for the process $N_i(t)$ is given by

$$\begin{aligned} \lambda_i(t)dt &= P(dN_i(t) = 1 | \mathcal{F}_{t-}) \\ &= P(\tau_i \in [t, t + dt), \delta_i = 1 | \tau_i \geq t, \mathcal{F}_{t-}), \end{aligned}$$

which, for independent right-censoring, can be expressed as

$$\lambda_i(t)dt = \begin{cases} P(T_i \in [t, t + dt) | T_i \geq t), & \tau_i \geq t \\ 0, & \tau_i < t \end{cases}$$

$$= Y_i(t)h_i(t)dt,$$

where $Y_i(t) = \mathbb{1}\{\tau_i \geq t\}$ denotes whether individual i is at risk at time t, that is, it has not experienced the event, nor censoring, by time t. The hazard rate $h_i(t)$ of the process $N_i(t)$ is defined as the instantaneous rate of experiencing the event at time t, given that the individual has not experienced the event of interest by time t. Thus, $h_i(t)$ is given by

$$h_i(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_i \in [t, t + \Delta t) | T_i \geq t).$$

Consider the aggregated counting process $N(t) = \sum_{i=1}^{n} N_i(t)$. When the hazard rate of all individuals is equal, i.e. $h_i(t) = h(t)$ for all i, the aggregated process $N(t)$ has an intensity process of the form

$$\lambda(t) = Y(t)h(t),$$

where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ is the number of individuals at risk just before time t. When an intensity process is of this form, the process $N(t)$ is said to fulfil the multiplicative intensity model, and by the Doob-Meyer decomposition, the increments of the counting process can be written as

$$dN(t) = Y(t)h(t)dt + dM(t),$$

as seen in (2.3).

3.2 Kaplan-Meier estimator

The Kaplan-Meier estimator provides an estimator of the survival function in situations where censoring occurs.

Let $0 = u_0 < u_1 < \cdots < u_k = t$ be a partition of the interval $[0, t]$. The survival function S at t can be rewritten as

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t | T > u_{k-1}) P(T > u_{k-1}) \\ &= P(T > t | T > u_{k-1}) \cdots P(T > u_1 | T > u_0) \\ &= \prod_{i=1}^{k} P(T > u_i | T > u_{i-1}), \end{aligned} \tag{3.2}$$

by the multiplication rule for conditional probabilities.

Let $\tau_{(1)}, \tau_{(2)}, ...$ be the observed times ordered increasingly and let $\delta_{(i)}$ equal the indicator for the occurrence at time $\tau_{(i)}$. Then, $0 = \tau_{(0)} < \tau_{(1)} < \cdots < \tau_{(k)} = t$ is a partition of the interval $[0, t]$ and the survival function S at t equals

$$S(t) = \prod_{i=1}^{k} P(T > \tau_{(i)} | T > \tau_{(i-1)}),$$

by (3.2).

Let the risk set at time t be the set of individuals for whom the event has not happened before time t and who have not been censored before time t. Let $Y_i(t)$ indicate whether individual i is at risk at time t. Then, $Y(t) = \sum_{i=1}^{n} Y_i(t)$ determines the cardinality of the risk set at time t.

To estimate the probability

$$p_{(i)} = P(T > \tau_{(i)} | T > \tau_{(i-1)}),$$

consider the case that the ordered ith observation was an exact observation, $\delta_{(i)} = 1$, meaning that one individual experienced the event at time $\tau_{(i)}$. Since the number of individuals at risk right before $\tau_{(i)}$ is $Y\left(\tau_{(i)}\right)$ and $Y\left(\tau_{(i)}\right) - 1$ did not experience the event at time $\tau_{(i)}$, the estimate of the probability is given by $\frac{Y(\tau_{(i)})-1}{Y(\tau_{(i)})}$. Consider the case where the ordered ith observation was a censored time, $\delta_{(i)} = 0$. Since $Y(\tau_{(i)})$ individuals did not experience the event at time $\tau_{(i)}$, the estimate of the probability is given by $\frac{Y(\tau_{(i)})}{Y(\tau_{(i)})}$.

Therefore, the probability $p_{(i)}$ can be estimated by

$$\hat{p}_{(i)} = \begin{cases} \frac{Y(\tau_{(i)})-1}{Y(\tau_{(i)})} = 1 - \frac{\delta_{(i)}}{Y(\tau_{(i)})} & \text{if } \delta_{(i)} = 1, \\ \frac{Y(\tau_{(i)})}{Y(\tau_{(i)})} = 1 - \frac{\delta_{(i)}}{Y(\tau_{(i)})} & \text{if } \delta_{(i)} = 0. \end{cases}$$

Thus, the plug-in estimator of S is given by

$$\begin{aligned} \hat{S}(t) &= \prod_{i, \tau_{(i)} \leq t} \left(1 - \frac{\delta_{(i)}}{Y\left(\tau_{(i)}\right)}\right) \\ &= \prod_{t_i \leq t} \left(1 - \frac{1}{Y\left(t_i\right)}\right), \end{aligned}$$

where the equality holds since censoring times do not contribute to the product. This estimator of the survival function results in a non-increasing right-continuous step function with jumps at exact observations, where the size of the jump is affected by the censoring observations. The estimator is known as the Kaplan-Meier estimator and was introduced in [9].

3.3 Nelson-Aalen estimator

The Nelson-Aalen estimator provides an estimator of the cumulative hazard function H. The cumulative hazard function and the survival function are related, and the Kaplan-Meier estimate of the survival function can be obtained from the Nelson-Aalen estimator by the plug-in approach.

The cumulative hazard function $H(t)$ is given by

$$H(t) = \int_0^t h(t)dt,$$

where $h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T \in [t, t + \Delta t) | T \geq t)$.

In the case where the survival function S is absolutely continuous, the cumulative hazard function can be related to the survival function by the following equation

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T \in [t, t + \Delta t) | T_i \geq t) \\ &= \frac{f(t)}{S(t)} \\ &= \frac{-S'(t)}{S(t)}, \end{aligned} \tag{3.3}$$

and the survival function can be written in terms of the cumulative hazard function by

$$S(t) = e^{-H(t)}.$$

If S is not absolutely continuous, it is càdlàg. Let $dS(t)$ be the increments of S over the time interval $[t, t + dt)$

$$dS(t) = P(T \in [t, t + dt)),$$

and let $S(t-)$ denote the left hand limit of $S(t)$. Then, using the definition of h, it follows that

$$h(t)dt = P(T \in [t, t+dt) | T \geq t) = \frac{-dS(t)}{S(t-)}.$$

Integrating both sides yields an expression for the cumulative hazard function

$$H(t) = -\int_0^t \frac{dS(u)}{S(u-)}, \tag{3.4}$$

which is a Riemann-Stieltjes integral.

If $S(t)$ is absolutely continuous, the hazard rate h from equation (3.4) equals (3.3). In the case where S is discrete, then the cumulative hazard equals

$$H(t) = \sum_{u \leq t} h_u,$$

where

$$\begin{aligned} h_u &= -\frac{S(u) - S(u-)}{S(u-)} \\ &= P(T = u | T \geq u). \end{aligned}$$

As seen in (3.2), S can be written as

$$S(t) = \prod_{i=1}^{k} P(T > u_i | T > u_{i-1}),$$

where $0 = u_0 < u_1 < \cdots < u_k = t$ is a partition of the interval $[0, t]$.

From (3.4), it follows that

$$dS(t) = -S(t-)dH(t),$$

which can be approximated by

$$S(u_i) - S(u_{i-1}) \approx -S(u_{i-1})(H(u_i) - H(u_{i-1})),$$

dividing both sides by $S(u_{i-1})$ yields

$$S(u_i | u_{i-1}) \approx 1 - (H(u_i) - H(u_{i-1})).$$

By inserting this expression in (3.2), the following expression is obtained for S

$$S(t) \approx \prod_{i=1}^{k} \left(1 - \left(H(u_i) - H(u_{i-1})\right)\right).$$

By increasing the number of sub-intervals of the partition, the approximation improves. The right hand side will approach a limit, denoted the product-integral, which is defined as

$$\underset{0\leq u\leq t}{\pi}(1 + dB(u)) = \lim_{\max_i |u_i - u_{i-1}|\to 0} \prod_{i=1}^{k} \left(1 + \left(B(u_i) - B(u_{i-1})\right)\right)$$

for an arbitrary càdlàg function $B(t)$.

The survival function can be written as

$$S(t) = \underset{0\leq u\leq t}{\pi}(1 - dH(t)),$$

where π corresponds to the product-integral [1].

For a continuous distribution,

$$\begin{aligned} S(t) &= \underset{0\leq u\leq t}{\pi}(1 - dH(t)) \\ &= \underset{0\leq u\leq t}{\pi}(1 - h(u)du) \\ &= \exp\left\{-\int_0^t h(u)du\right\} \\ &= e^{-H(t)}, \end{aligned}$$

since $e^{-x} \approx (1 - x)$.

For a discrete distribution,

$$S(t) = \prod_{u\leq t}(1 - h_u).$$

In general, the cumulative hazard can be decomposed into a continuous H_c and a discrete part H_d in the following way

$$H(t) = H_c(t) + H_d(t).$$

Then, the survival function is given by

$$S(t) = e^{-H_c(t)} \prod_{t_i \leq t} (1 - \Delta H_d(t_i)).$$

The probability $h_{\tau_{(i)}} = P(T = \tau_{(i)} | T \geq \tau_{(i)})$ can be estimated by

$$\hat{h}_{\tau_{(i)}} = \frac{\delta_{(i)}}{Y(\tau_{(i)})}.$$

The cumulative hazard function can be estimated using the plug-in approach

$$\begin{aligned} \hat{H}(t) &= \sum_{\tau_{(i)} \leq t} \hat{h}_{\tau_{(i)}} \\ &= \sum_{i, \tau_{(i)} \leq t} \frac{\delta_{(i)}}{Y(\tau_{(i)})} \\ &= \sum_{t_i \leq t} \frac{1}{Y(t_i)}, \end{aligned}$$

where the last equality holds since censoring times do not contribute to the sum. This estimate of the cumulative hazard function is known as the Nelson-Aalen estimator [1]. The estimator is a non-decreasing and right-continuous step-function with jumps at the observed events.

An estimator of the survival function $\hat{S}$ can be obtained using the plug-in approach in the following way

$$\begin{aligned} \hat{S}(t) &= \prod_{0 \leq u \leq t} (1 - d\hat{H}(t)) \\ &= \prod_{t_i \leq t} \left(1 - \frac{1}{Y(t_i)}\right). \end{aligned}$$

This estimator is equal to the Kaplan-Meier estimator.

3.3.1 Handling tied data

Tied event times refer to situations where two or more exact observations occur at the same time. The estimators described beforehand assume that the event times are time-continuous. However, in practice, tied event times are often present. To address this issue, [1] suggests two approaches.

(i) One option is to assume that the events happen in continuous time so that the event times do not coincide. However, the recorded event times are equal due to rounding.

(ii) Another option is to explicitly model as time-discrete processes. Thus, tied event times are events that have happened simultaneously and not due to rounding.

Let d_i represent the number of individuals experiencing the event at time τ_i. If the observed time is a censoring time, d_i is zero. If the tied event times are due to rounding, it might be reasonable to assume that the true event times would have been slightly different had it not been for measurement error. The Nelson-Aalen estimator for approach **(i)** can be used to estimate the cumulative hazard function at time t. The estimator is given by

$$\hat{H}(t) = \sum_{i,t_i \leq t} \sum_{l=0}^{d_i - 1} \frac{1}{Y(t_i) - l}.$$

If $d_i = 1$, the Nelson-Aalen estimator reduces to the untied data case. If $d_i > 1$, it is assumed that the individuals experience the event one at a time, and the risk set decreases by one for each event.

For approach **(ii)**, the events are assumed to be discrete. Therefore, it is reasonable to estimate the cumulative hazard function by

$$\hat{H}(t) = \sum_{i,t_i \leq t} \frac{d_i}{Y(t_i)}.$$

Using the Nelson-Aalen estimator, the survival function can be estimated by plugging in the corresponding estimator for the cumulative hazard function. For both approaches, the resulting estimator equals

$$\hat{S}(t) = \prod_{i,t_i \leq t} \left(1 - \frac{d_i}{Y(t_i)}\right). \tag{3.5}$$

For approach **(ii)**, expression (3.5) is obtained directly. For approach **(i)**, the survival function is estimated by

$$\hat{S}(t) = \prod_{i,t_i \leq t} \prod_{l=0}^{d_i - 1} \left(1 - \frac{1}{Y(t_i) - l}\right)$$

$$
\begin{aligned}
&= \prod_{i,t_i \le t} \left(1 - \frac{1}{Y(t_i)}\right)\left(1 - \frac{1}{Y(t_i)-1}\right)\cdots\left(1 - \frac{1}{Y(t_i)-d_i+1}\right) \\
&= \prod_{i,t_i \le t} \frac{(Y(t_i)-1)\,(Y(t_i)-2)\cdots(Y(t_i)-d_i+1-1)}{Y(t_i)\,(Y(t_i)-1)\cdots(Y(t_i)-d_i+1)} \\
&= \prod_{i,t_i \le t} \frac{Y(t_i)-d_i}{Y(t_i)} \\
&= \prod_{i,t_i \le t} \left(1 - \frac{d_i}{Y(t_i)}\right).
\end{aligned}
$$

3.4 Regression models

Neither the Nelson-Aalen nor the Kaplan-Meier estimates include covariates. Nevertheless, there are several regression models to be able to assess the effect of covariates on survival.

3.4.1 Cox Proportional Hazards Model

The Cox proportional hazards model is widely used in survival analysis and has become a standard method for analysing time-to-event data. The Cox model is a type of relative risk regression model that is semi-parametric.

The hazard function for the relative risk model has the form

$$h(t|\boldsymbol{z}_i) = h_0(t)c(\boldsymbol{\theta}, \boldsymbol{z}_i(t)), \tag{3.6}$$

where $h_0(t)$ is the baseline hazard, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ are unknown parameters, $\boldsymbol{z}_i(t) = (z_{1,i}(t), ..., z_{p,i}(t))^T$ are the covariates for individual i and $c(\boldsymbol{\theta}, \boldsymbol{z}_i(t))$ is a known function.

The ratio of the hazard rates of individuals i and j is equal to

$$
\begin{aligned}
\frac{h(t|\boldsymbol{z}_i)}{h(t|\boldsymbol{z}_j)} &= \frac{h_0(t)c(\boldsymbol{\theta}, \boldsymbol{z}_i(t))}{h_0(t)c(\boldsymbol{\theta}, \boldsymbol{z}_j(t))} \\
&= \frac{c(\boldsymbol{\theta}, \boldsymbol{z}_i(t))}{c(\boldsymbol{\theta}, \boldsymbol{z}_j(t))},
\end{aligned}
$$

which is constant over time if the covariates are fixed. In that case, the model (3.6) is referred to as the proportional hazards model.

A common choice for the function $c(\boldsymbol{\theta}, \boldsymbol{z}_i(t))$ is

$$c(\boldsymbol{\theta}, \boldsymbol{z}_i(t)) = e^{\theta_1 z_{1,i} + \cdots + \theta_p z_{p,i}},$$

yielding the Cox proportional hazards model for the hazard rate

$$h(t|\boldsymbol{z}_i) = h_0(t) e^{\theta_1 z_{1,i} + \cdots + \theta_p z_{p,i}}.$$

The model is semi-parametric since the baseline hazard is an arbitrary function. Thus, the parameter estimates of $\boldsymbol{\theta}$ are obtained by maximising the Cox partial likelihood, which is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{\substack{i=1,\\ \delta_i=1}}^{n} \frac{h_0(\tau_i) \exp\{\theta_1 z_{1,i} + \cdots + \theta_p z_{p,i}\}}{\sum_{j \in R_i} h_0(\tau_i) \exp\{\theta_1 z_{1,j} + \cdots + \theta_p z_{p,j}\}} \\ &= \prod_{\substack{i=1,\\ \delta_i=1}}^{n} \frac{\exp\{\theta_1 z_{1,i} + \cdots + \theta_p z_{p,i}\}}{\sum_{j \in R_i} \exp\{\theta_1 z_{1,j} + \cdots + \theta_p z_{p,j}\}}, \end{aligned}$$

where R_i is the at-risk index set just before time t_i, i.e. it consists of the indices of the individuals that have not yet experienced the event, nor have been censored, just before time t_i.

The estimate of θ is obtained by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta).$$

The assumption of proportional hazards might not hold in some situations, such as when the hazard rate varies with time or when covariates are time-varying. When the assuption does not hold, Aalen's additive hazard model may be an alternative.

3.4.2 Aalen's Additive Hazards Model

Aalen's additive hazards model is a non-parametric regression model that estimates the effect of covariates on the hazard function without assuming proportionality. Instead, it models the hazard rate as the sum of unknown functions

of covariates, which allows for a more flexible and interpretable modelling of the effect of covariates on survival.

Let the intensity process be equal to

$$\begin{aligned}\lambda_i(t) &= Y_i(t)h_i(t) \\ &= Y_i(t)h(t|\boldsymbol{z}_i).\end{aligned}$$

The hazard rate for Aalen's model has the form

$$h(t|\boldsymbol{z}_i) = \beta_0(t) + \beta_1(t)z_{1,i}(t) + \cdots + \beta_p(t)z_{p,i}(t),$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_p(t))^T$ is a vector of unknown parameter functions and $\boldsymbol{z}_i(t) = (z_{1,i}(t), ..., z_{p,i}(t))^T$ are the covariates for individual i.

Then, by the Doob-Meyer decomposition (2.3), the increments of the process can be expressed as

$$dN_i(t) = Y_i(t)\left(\beta_0(t) + \beta_1(t)z_{1,i}(t) + \cdots + \beta_p(t)z_{p,i}(t)\right)dt + dM_i(t). \qquad (3.7)$$

Let $B_j(t) = \int_0^t \beta_j(u)du$, $j = 0, 1, ..., p$, be the cumulative regression functions and introduce the matrix notation

$$\begin{aligned}\boldsymbol{N}(t) &= \left(N_1(t), ..., N_n(t)\right)^T, \\ \boldsymbol{B}(t) &= \left(B_0(t), ..., B_p(t)\right)^T, \\ \boldsymbol{X}_i(t) &= Y_i(t)\left(1, z_{1,i}(t), ..., z_{p_i}(t)\right), \\ \boldsymbol{M}(t) &= \left(M_1(t), ..., M_n(t)\right)^T,\end{aligned}$$

and the $n \times (p+1)$ matrix $\boldsymbol{X}(t)$ with ith row given by $\boldsymbol{X}_i(t)$. Then (3.7) can be rewritten as

$$d\boldsymbol{N}(t) = \boldsymbol{X}(t)d\boldsymbol{B}(t) + d\boldsymbol{M}(t).$$

Suppose $\boldsymbol{X}(t)$ has the generalised inverse

$$\boldsymbol{X}^-(t) = \left(\boldsymbol{X}^T(t)\boldsymbol{X}(t)\right)^{-1}\boldsymbol{X}^T(t).$$

If $\boldsymbol{X}(t)$ has full rank, the increments $d\boldsymbol{B}(t)$ can be estimated by

$$d\hat{\boldsymbol{B}}(t) = \boldsymbol{X}^-(t)d\boldsymbol{N}(t).$$

For a general $\boldsymbol{X}(t)$, introduce the indicator $J(t)$ of $\boldsymbol{X}(t)$ having full rank at time t, $J(t) = \mathbb{1}\{\text{rank}(\boldsymbol{X}(t)) = p+1\}$. The estimator of cumulative regression functions $\boldsymbol{B}(t)$ is obtained by accumulating the increments $d\hat{\boldsymbol{B}}(t)$ over the times when an event occurs and $\boldsymbol{X}(t)$ has full rank,

$$\begin{aligned}\hat{\boldsymbol{B}}(t) &= \int_0^t J(u)\boldsymbol{X}^-(t)d\boldsymbol{N}(u) \\ &= \sum_{t_i \le t} J(t_i)\boldsymbol{X}^-(t_i)\Delta\boldsymbol{N}(t_i),\end{aligned}$$

where $\Delta\boldsymbol{N}(t_i) = \boldsymbol{N}(t_i) - \boldsymbol{N}(t_{i-1})$ for $t \ge 1$ and $\boldsymbol{N}(t_0) = \boldsymbol{0}$, cf. [1].

3.5 Causality

Path analysis is a technique used to study causality. Introduced in [17], path analysis allows visualising the dependencies among variables and decomposing total effects into direct and indirect effects. However, a limitation of path analysis is that it does not consider time, which is often a critical aspect of causality. To address this, dynamic path analysis combines graphical models with continuous time development to study the temporal dependencies among variables [7].

3.5.1 Path analysis

Path analysis investigates the relationships between variables by examining their direct and indirect effects. Path analysis uses directed acyclic graphs (DAGs) to provide a graphical representation of the causal relationships between variables. A hypothesised causal model in the form of a DAG is first specified, where each vertex represents a variable and each directed edge represents a hypothesised causal relationship between two variables. Statistical methods such as regression can be used to estimate path coefficients.

A directed graph G is determined by (V, E), where V is a set of vertices and E is a set of directed edges. Each edge in E is represented by an ordered pair of vertices from V and is drawn as an arrow.

A directed path in G is a sequence of edges such that the ending vertex of one edge is the starting vertex of the next edge. If the starting and ending vertices of a directed path are the same, the path is called a directed cycle. Directed acyclic graphs (DAGs) are directed graphs that do not contain any directed cycles.

For a vertex $v_1 \in V$, the set of children of v_1 is defined as all vertices $v_i \in V$ such that $(v_1, v_i) \in E$, i.e. there is a directed edge from v_1 to v_i. The set of children of v_1 is denoted by $\text{ch}(v_1)$. Similarly, the set of parents of v_1 is defined as all vertices $v_j \in V$ such that $(v_j, v_1) \in E$, i.e. there is a directed edge from v_j to v_1. The set of parents of v_1 is denoted by $\text{pa}(v_1)$. More information on DAGs, and other graphical models, can be found in [4].

DAGs are used to study the direct and indirect effects of the random variables $X_1, ..., X_p$ on a outcome Y. The vertex set V can be partitioned into the set of covariates $V_c = \{X_1, ..., X_p\}$ and the outcome variable Y. The set of hypothesised edges E respects the DAG assumption. The outcome Y is the ending vertex for all paths in the graph and the set of children of Y, $\text{ch}(Y)$, is empty. Each edge (X_i, X_j) is associated with a path coefficient denoted $\theta_{i,j}$ and each edge (X_i, Y), with a path coefficient denoted β_i.

Let X_i be a variable such that the vertex $(X_i, Y) \in E$, i.e. there exists a path of length one between X_i and Y. Then the coefficient β_i represents the direct effect of the variable X_i on the output Y, denoted by

$$\text{dir}(X_i, Y) = \beta_i.$$

Indirect paths from X_i to Y, i.e. paths of length greater than one, are defined by

$$P_j = \{(X_{j_1}, X_{j_2}), (X_{j_2}, X_{j_3}), ..., (X_{j_{k_j}}, Y)\},$$

where $X_{j_1} = X_i$ for all j, and k_j is the length of the path P_j. Suppose that there are r indirect paths from X_i to Y and denote them by $P_1, ..., P_r$. The indirect effect of X_i to Y is then given by

$$\text{ind}(X_i, Y) = \sum_{j=1}^{r} \left(\prod_{l=1}^{k_j - 1} \theta_{j_l, j_{l+1}} \right) \beta_{j_{k_j}}.$$

The path coefficients θs and βs can be estimated using layered (multiple) linear regression. The output is regressed on its parents to obtain β estimates, and then each of the variables on its parents to get θ estimates.

3.5.2 Dynamic path analysis

A dynamic path is a set of directed acyclic graphs $G(t) = (V(t), E(t))$ indexed by time t, where $V(t)$ denotes the vertices and $E(t)$ the edges at time t. Let the vertex set $V(t)$ be partitioned into a set of covariates $V_c(t) = \{X_1(t), ..., X_p(t)\}$ and an outcome $Y(t)$. The partition of vertices is time-invariant, but the edges may vary with time. For all t, the set $E(t)$ respects the DAG assumption, and $Y(t)$ is not the starting vertex for any edge, i.e. the set of children of $Y(t)$, $\text{ch}(Y)$, is empty.

In the counting process framework, the output is the aggregated counting process $N(t)$ for n individuals, which counts the number of individuals who have experienced the event of interest during the time interval $[0, t]$.

Dynamic path analysis considers collections of DAGs indexed by time t, where the vertices remain constant, and the edges may vary with time. Additionally, the coefficients corresponding to the edges can vary. When the outcome is a counting process, a DAG is defined for each jump in the counting process.

It is necessary to perform a regression analysis on the occurrence of a single jump in the counting process to estimate direct, indirect and total effects. A linear model must be used for regression analysis to separate total effects into direct and indirect effects. The study of direct and indirect effects is impossible with non-linear models like the Cox model. Therefore, Aalen's additive hazards model is used in combination with linear regression between covariates.

Example 3.1. Consider the dynamic path model diagram shown in Figure 1, which illustrates a single jump in the counting process. The diagram includes a fixed covariate X_1 (e.g., a treatment indicator) and a possibly time-varying covariate $X_2(t)$. The coefficient $\theta_{1,2}(t)$ is the standard coefficient in the ordinary linear regression of $X_2(t)$ on X_1, where the analysis is performed on the relevant risk set at time t. Although X_1 and X_2 may not change over time, the coefficient $\theta_{1,2}(t)$ may vary since the risk set changes over time. The direct effect of X_1 and $X_2(t)$ on the number of events $dN(t)$ is given by $\beta_1(t)dt$ and $\beta_2(t)dt$, where $\beta_1(t)$ and $\beta_2(t)$ correspond to the regression functions in Aalen's additive model.

The equations corresponding to the path diagram are given by

$$dN(t) = (\beta_0(t) + \beta_1(t)X_1 + \beta_2(t)X_2)dt + dM(t), \tag{3.8}$$

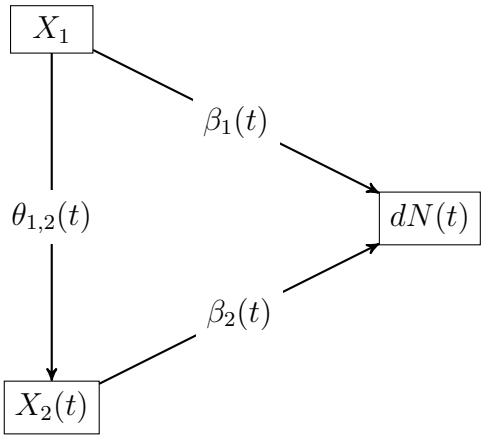


Figure 1: Dynamic path model diagram for a single jump in the counting process

$$X_2(t) = \theta_0(t) + \theta_{1,2}(t)X_1 + \varepsilon(t), \tag{3.9}$$

where equation (3.8) is the Doob-Meyer decomposition for counting processes so that the intensity of $N(t)$ is given by $h(t) = \beta_0(t) + \beta_1(t)X_1 + \beta_2(t)X_2$. In equation (3.9), $\varepsilon(t)$ is independent of X_1 and $M(t)$.

Inserting (3.9) into (3.8) leads to

$$dN(t) = [\beta_0(t)+\beta_2(t)\theta_0(t)+(\beta_1(t)+\beta_2(t)\theta_{1,2}(t))X_1+\beta_2(t)\varepsilon(t)]dt+dM(t).$$

The total effect of X_1 on $dN(t)$, referred to as treatment effect, is given by

$$\text{tot}(X_1, dN(t)) = (\beta_1(t) + \beta_2(t)\theta_{1,2}(t))dt,$$

which can be decomposed into the direct and indirect effects of X_1

$$\begin{aligned}\text{dir}(X_1, dN(t)) &= \beta_1(t)dt,\\ \text{ind}(X_1, dN(t)) &= \beta_2(t)\theta_{1,2}(t)dt.\end{aligned}$$

□

Estimation of direct and indirect effects

Combining path analysis with Aalen's additive hazards model makes it possible to estimate the total effects of a covariate on the counting process by adding the estimates of direct and indirect effects.

The coefficients $\beta_i(t)dt$ represent the direct effects of variable $X_i(t)$ on the single jumps of the counting process,

$$\text{dir}(X_i(t), dN(t)) = dB_i(t),$$

and the cumulative direct effect is given by

$$\text{cdir}(X_i(t), N(t)) = B_i(t).$$

Assume that there are r paths from $X_i(t)$ to $dN(t)$. Then, indirect effect of X_i on $dN(t)$ is given by

$$\text{ind}(X_i(t), dN(t)) = \sum_{j=1}^{r} \prod_{l=1}^{k_j-1} \theta_{j_l, j_{l+1}}(t) dB_{j_{k_j}}(t),$$

and the cumulative direct effect is given by

$$\text{cind}(X_i(t), N(t)) = \int_0^t \sum_{j=1}^{r} \prod_{l=1}^{k_j-1} \theta_{j_l, j_{l+1}}(s) dB_{j_{k_j}}(s).$$

The total effect and the cumulative total effects are given by

$$\begin{aligned}\text{tot}(X_i(t), dN(t)) &= \text{dir}(X_i(t), dN(t)) + \text{ind}(X_i(t), dN(t)),\\ \text{ctot}(X_i(t), N(t)) &= \text{cdir}(X_i(t), N(t)) + \text{cind}(X_i(t), N(t)).\end{aligned}$$

As seen in Section 3.3, the cumulative regression functions can be estimated by

$$\hat{\boldsymbol{B}}(t) = \sum_{t_m \leq t} J(t_i) \boldsymbol{X}^{-}(t_m) \Delta \boldsymbol{N}(t_m),$$

where t_m are the observed time events. The estimator can be rewritten as

$$\hat{\boldsymbol{B}}(t) = \sum_{t_m \leq t} \Delta \hat{\boldsymbol{B}}(t_m),$$

where $\Delta \hat{\boldsymbol{B}}(t) = J(t) \boldsymbol{X}^{-}(t) \Delta \boldsymbol{N}(t)$.

Then, the cumulative direct and indirect effects can be estimated using the plug-in approach, resulting in

$$\widehat{\text{cdir}}(X_i(t), N(t)) = \sum_{t_m \leq t} \Delta \hat{B}_i(t_m),$$

$$\widehat{\text{cind}}(X_i(t), N(t)) = \sum_{t_m \leq t} \sum_{j=1}^{r} \prod_{l=1}^{k_j - 1} \hat{\theta}_{j_l, j_{l+1}}(t_m) \Delta \hat{B}_{j_{k_j}}(t_m).$$

The cumulative total effect of the variable X_i is then estimated by

$$\widehat{\text{ctot}}(X_i(t), N(t)) = \widehat{\text{cdir}}(X_i(t), N(t)) + \widehat{\text{cind}}(X_i(t), N(t)).$$

3.6 Bootstrap for survival data

Bootstrapping is a non-parametric method introduced by Efron (1979) that can assess the uncertainty of statistics without making assumptions about the underlying distribution of the data [5]. Given a data set with sample size n, bootstrapping consists of sampling with replacement from the data set to obtain bootstrapped samples of the size n. By repeatedly computing the statistics for different bootstrapped samples, one can analyse the distribution of the statistics.

To introduce the general bootstrap, let $x_1, ..., x_n$ be independent identically distributed observations from the distribution F that is known except for the parameter θ. The empirical distribution function F_n of F is given by (3.1). Given that the unknown parameter θ is a functional of the distribution function

$$\theta = T(F),$$

the plug-in estimator of θ is given by

$$\hat{\theta}_n = T(F_n).$$

By Donsker's theorem [15], it holds that

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} B \circ F,$$

on the space $D[-\infty, \infty]$ of càdlàg functions, where B is an standard Brownian bridge. If the functional T is Hadamard differentiable and the derivative T'_f is linear, then it holds that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} T'_F(B \circ F),$$

by [15]. However, since F is unknown, the asymptotic distribution $T'_F(B \circ F)$ is also unknown.

3.6.1 The bootstrap idea

Given the observational data $z_1, ..., z_n$, a bootstrap sample $z_1^*, ..., z_n^*$ is a sample with replacement from the data. For each bootstrapped sample, one can calculate the bootstrapped empirical distribution function F_n^*. The following result holds

$$\sqrt{n}(F_n^* - F_n) \xrightarrow{d^*} B \circ F,$$

by [15], and with the exact meaning of $\xrightarrow{d^*}$ described in [15, p. 332-333]. If T is Hadamard differentiable, then

$$\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) \xrightarrow{d^*} T_F'\left(B \circ F\right),$$

where $\hat{\theta}_n^* = T(F_n^*)$. Thus $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ and $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ have the same asymptotic distribution. The asymptotic distribution for $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$ only depends on the sample and can be approximated by resampling the original data $x_1, ..., x_n$ many times and by providing the empirical densities based on the resampled samples. The bootstrap method was introduced by Efron in [5].

3.6.2 Bootstrap for right-censored data

Efron [6] introduces two methods for bootstrapping right-censored data: the 'simple' and the 'obvious' method. The 'simple' method consists of sampling n times with replacement from the observed data $\{(\tau_i, \delta_i)\}_{i=1}^n$. On the other hand, the 'obvious' method involves constructing a new data set by drawing bootstrapped event times t_i^* and censoring times c_i^*.

In the 'obvious' method, event times t_i^* are drawn from the Kaplan-Meier estimate of the survival function $S(t) = P(T > t)$. Similarly, censoring times c_i^* are drawn from the Kaplan-Meier estimate of the cumulative censoring function $G(t) = P(C > t)$, which can be computed as

$$\hat{G}(t) = \prod_{i,\tau_i \leq t} \left(1 - \frac{\tilde{\delta}_i}{Y\left(\tau_i\right)}\right),$$

where $\tilde{\delta}_i = 1 - \delta_i$ indicates whether a censoring time has been observed. Once $t_i^* \sim \hat{S}$ and $c_i^* \sim \hat{G}$ are generated, the bootstrapped sample is given by (τ_i^*, δ_i^*), where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$.

Both the 'simple' and 'obvious' methods yield equivalent results, because of certain properties of the Kaplan-Meier estimates $\hat{S}$ and $\hat{G}$, c.f. [6].

3.6.3 Bootstrap for Cox multiplicative hazards model

Burr introduces in [3] three methods for bootstrapping the Cox model. The first method is non-parametric and does not rely on any assumptions about the model. The second method draws samples of the time-to-events from the survival function estimated from the fitted Cox model. The censoring times are drawn from a Kaplan-Meier estimate of the survival function of the censoring variable. The third method utilises the fact that the censoring pattern is an ancillary statistic and uses the available information, specifically that C_i is either observed or $C_i \geq T_i$.

Method 1: Resample with replacement the triplets $(\tau_i, \delta_i, \boldsymbol{z}_i^T)$, where $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ are the covariates for individual i. Equivalent to Efron's 'simple' method, but including covariates. □

Method 2: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and assume the time to the events $t_1, ..., t_n$ are completely observed. The fitted model is obtained by maximising the Cox partial likelihood. The Breslow estimator $\hat{H}_0(t)$ estimates the cumulative baseline hazard. The fitted survival curve for individual i is given by

$$\begin{aligned}\hat{S}_i(t|\boldsymbol{z}_i) &= \prod_{0\leq u\leq t}(1 - d\hat{H}(t)) \\ &= \prod_{t_i\leq t}(1 - \Delta\hat{H}(t_i)) \\ &= \prod_{t_i\leq t}\left(1 - \Delta\hat{H}_0(t_i)e^{\hat{\theta}_1 z_{1,i}+\cdots+\hat{\theta}_p z_{p,i}}\right).\end{aligned}$$

To obtain an observation time t_i^*, simulate from $\hat{S}_i$.

Let $\tilde{\delta}_i = 1-\delta_i$ indicate whether a censoring time has been observed. Since the censoring is independent of the covariates, the survival function for the random variable C can be estimated by the Kaplan-Meier estimator of $G(t) = P(C > t)$. The censoring times c_i^* can be simulated from $\hat{G}(t)$.

After generating t_i^* and c_i^*, the bootstrapped sample is constructed as $(\tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

Method 3: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and generate t_i^* as in Method 2. The censoring pattern is an ancillary statistic since the distribution of C_i, by assumption, does not depend on the parameters of the Cox model. If $\delta_i = 0$, i.e. an observation from C_i is observed, let c_i^* be equal to τ_i. If $\delta_i = 1$, draw the censoring time c_i^* from the Kaplan-Meier estimate of the survival function G given $C_i \geq T_i$, namely

$$\begin{aligned}\hat{G}(t|C_i \geq t_i) &= \hat{P}(C_i > t|C_i \geq t_i) \\ &= \frac{\hat{P}(C_i > t)}{\hat{P}(C_i \geq t_i)} \\ &= \frac{\hat{G}(t)}{\hat{G}(t_i)}.\end{aligned}$$

After generating t_i^* and c_i^*, the bootstrapped sample is constructed as $(\tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

3.6.4 Bootstrap for Aalen's additive hazards model

Method 2 and 3 can be modified to draw bootstrap samples from Aalen's model.

Method 2: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and assume the time to the events $t_1, ..., t_n$ are completely observed. Given the estimates of the cumulative regression functions $\boldsymbol{B}(t)$ for Aalen's model, to obtain the fitted survival curve for individual i compute

$$\begin{aligned}\tilde{S}_i(t|\boldsymbol{z}_i) &= \prod_{0 \leq u \leq t}(1 - d\hat{H}(t|\boldsymbol{z}_i)) \\ &= \prod_{t_k \leq t}(1 - \Delta\hat{H}(t_k|\boldsymbol{z}_i)) \\ &= \prod_{t_k \leq t}\left(1 - \left(\Delta\hat{B}_0(t_k) + \Delta\hat{B}_1(t_k)z_{1,i} + \cdots + \Delta\hat{B}_p(t_k)z_{p,i}\right)\right).\end{aligned}$$

Since there is no condition on $\Delta\hat{H}(t_k|\boldsymbol{z}_i)$ being positive, the resulting function $\tilde{S}_i(t|\boldsymbol{z}_i)$ is not monotone. To obtain a monotone estimate of

the survival curve, perform antitonic regression on $\tilde{S}_i(t|\boldsymbol{z}_i)$. Antitonic regression is equivalent to performing isotonic regression on $\tilde{F}_i(t|\boldsymbol{z}_i) = 1 - \tilde{S}_i(t|\boldsymbol{z}_i)$ to obtain $\hat{F}_i$ and the survival curve $\hat{S}_i(t|\boldsymbol{z}_i) = 1 - \hat{F}_i(t|\boldsymbol{z}_i)$.

Isotonic regression is a technique used when the underlying regression function is assumed to have specific order restrictions. It aims to estimate a non-decreasing function that fits the data by optimising a criterion function under the monotonicity constraint. Consider the data set $(t_i, \tilde{F}(t_i))$, $i = 1, ..., k$, where k is the number of distinct times at which an individual has experienced the event. Define $\mathscr{F}$, the set of increasing sequences, as follows

$$\mathscr{F} = \{x = (x_1, ..., x_k) \in \mathbb{R}^k; x_1 \leq \cdots \leq x_k\}.$$

The isotonic regressor of the data is defined by

$$\hat{F} = \underset{x \in \mathscr{F}}{\operatorname{argmin}} \sum_{i=1}^{k} (x_i - \tilde{F}(t_i))^2.$$

The isotonic regressor $\tilde{F}$ exists and is characterised as the slopes of the greatest convex minorant of the partial sum process, defined by

$$\left(t_i, \sum_{j=1}^{i} \tilde{F}(t_j) \right),$$

for $i = 1, ..., k$. The greatest convex minorant is the supremum of all convex functions which lie entirely below the partial sum process. The pool-adjacent-violators algorithm (PAVA), implemented by the `isoreg` function in the statistical package R, is a well-known algorithm to compute the isotonic regression. A comprehensive study of order restricted inference can be found in [12].

Note that the isotonic regression of $\tilde{F}$ does not give a function that ends in 1, i.e. that is necessarily a distribution function. However, this is a general problem for the plug-in estimator of F in the Cox model, Aalen's model and, in fact, even for the Kaplan-Meier estimator since all of these give possibly defective distributions.

To obtain an observation time t_i^*, simulate from $\hat{S}_i$. The censoring times c_i^* can be simulated from the Kaplan-Meier estimator of $G(t) = P(C > t)$, i.e. $\hat{G}$.

After generating t_i^* and c_i^*, the bootstrapped sample is constructed as $(\tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

Method 3: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and generate t_i^* as in Method 2. The censoring pattern is an ancillary statistic since the distribution of C_i, by assumption, does not depend on the parameters of Aalen's model. If an observation from C_i is observed, i.e. $\delta_i = 0$, let c_i^* be equal to τ_i. Otherwise, if $\delta_i = 1$, draw the censoring time c_i^* from the Kaplan-Meier estimate of the survival function G given $C_i \geq t_i$.

After generating t_i^* and c_i^*, the bootstrapped sample is constructed as $(\tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

4 Survival analysis for left-truncated right-censored data

Section 3 addresses right-censored data. This section presents the modifications necessary for dealing with left-truncated right-censored (l.t.r.c.) data. As seen previously, right-censored data occurs when the event of interest does not happen for all individuals within the study period. In contrast, the data is left-truncated if individuals are not observed from the origin but conditional on having survived until a baseline.

The method presented here is adapted from [16]. However, the application to Aalen's model and to graphical models is, to the best of our knowledge, novel.

4.1 Left-truncated right-censored data

Let $e_1, ..., e_n$ be the entry points in the study, generated from the underlying truncation random variables $E_1, ..., E_n$ with distribution function F_E. The observed data consists of (e_i, τ_i, δ_i), where τ_i and δ_i are given by

$$\tau_i = \min(T_i, C_i)$$

and

$$\delta_i = \mathbb{1}\{T_i \leq C_i\}.$$

The entry point and the censoring random variables (E_i, C_i) are assumed to be independent of the time until the event T_i. Also, (τ_i, δ_i) is only observed if $\tau_i \geq E_i$.

4.2 Bootstrap for left-truncated right-censored data

Wang [16] generalises Efron's 'obvious' bootstrap method to sample from l.t.r.c. data under the assumptions

$$E_i \leq C_i \text{ and } C_i - E_i \text{ is independent of } E_i, \tag{4.1}$$

$$S \text{ is continuous, } a < \inf\{t|F(t) > 0\} \text{ and } b < \sup\{t|S(t) > 0\}, \tag{4.2}$$

where $a = \inf\{e|F_E(e) > 0\}$ and $b = \sup\{e|1 - F_E(e) > 0\}$. Assumption (4.1) ensures the independence of T_i, E_i and $C_i - E_i$. Assumption (4.2) ensures that the minimum truncation time is less than the minimum life-time and the maximum truncation time is less than the maximum life-time. This technical assumption avoids non-identifiability problems regarding estimating S and F_E. If this assumption is not satisfied in practice, it suffices to trim part of the observations so that it is.

Let the risk set at time t be the set of individuals who have entered the study by time t, i.e. the set of indices i for which $E_i \leq t$; for whom the event has not happened before time t and who have not been censored before time t. Let $Y_i(t) = \mathbb{1}\{e_i \leq t \leq \tau_i\}$ indicate whether individual i is at risk at time t. Then, $Y(t) = \sum_{i=1}^{n} Y_i(t)$ equals the cardinality of the risk set at time t.

Under these assumptions, the survival function S can be consistently estimated by the Kaplan-Meier estimate of S given by

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{1}{Y(t_i)}\right).$$

The distribution function of the truncation variables E is obtained by the non-parametric maximum likelihood estimator (NPMLE) of the "working" data, (e_i, t_i, c_i). The NPMLE of F_E is given by

$$\hat{F}_E\left(e, \hat{S}\right) = \frac{1}{\sum_{j=1}^{n} \frac{1}{\hat{S}(e_j)}} \sum_{e_i \leq e} \frac{1}{\hat{S}(e_i)}.$$

The distribution Q of $D_i = C_i - E_i$ is given by

$$\hat{Q}(d) = \begin{cases} 1 & \text{if } d \geq \max_i\{\tau_i - E_i\}, \\ 1 - \prod_{d_i \leq d}\left(1 - \frac{1}{Y'(d_i)}\right) & \text{if } d < \max_i\{\tau_i - E_i\}, \end{cases}$$

where d_i are the observed differences $D_i = C_i - E_i$, and where the cardinality of the risk set at time d is given by $Y'(d) = \sum_{i=1}^{n} \mathbb{1}\{\tau_i - e_i \geq d\}$. Note in particular that $\prod_{d_i \leq d}\left(1 - \frac{1}{Y'(d_i)}\right)$ is a Kaplan-Meier type estimator, but with censored differences $d_i = c_i - e_i$ being event times.

Bootstrap samples from E, D and T are generated independently. The truncation times e_i^* are sampled from $\hat{F}_E$, the differences d_i^* are sampled from $\hat{Q}$

and, finally, the event times t_i, from $\hat{S}$. Let $c_i^* = e_i^* + d_i^*$, $\tau_i = \min\{t_i^*, c_i^*\}$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. If $t_i \leq e_i$, the observation is truncated and discarded. Generate samples until n observations have been accepted.

This generalisation of Efron's 'obvious' method for l.t.r.c. data is not equivalent to the 'simple' method. The 'simple' method for l.t.r.c. data is discussed in [8] and does not require the assumtions (4.1) and (4.2).

4.3 Bootstrap for Aalen's additive hazards model with left-truncated right-censored data

We next propose three methods to sample from l.t.r.c. data with covariates. The first is a trivial generalisation of Efron's 'simple' method. The second and the third are more substantial adaptions of Wang's [16] method to regression models for survival data. Also, we note the need for order-based inference methods, i.e. isotonic regression. The method we propose is, to the best of our knowledge, not previously studied.

Method 1: Perform the generalisation of Efron's 'simple' method for l.t.r.c. data [8]. Resample the quadruplets $\{(e_i, \tau_i, \delta_i, \boldsymbol{z}_i^T)\}_{i=1}^n$ to obtain a bootstrapped data set of size n. □

Method 2: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and assume the time to the events $t_1, ..., t_n$ are completely observed. Given the estimates of the cumulative regression functions $\boldsymbol{B}(t)$ for Aalen's model, the fitted survival curve for individual i is given by

$$\begin{aligned}\tilde{S}_i(t|\boldsymbol{z}_i) &= \underset{0\leq u\leq t}{\pi}(1 - d\hat{H}(t|\boldsymbol{z}_i)) \\ &= \prod_{t_k \leq t}(1 - \Delta\hat{H}(t_k|\boldsymbol{z}_i)) \\ &= \prod_{t_k \leq t}\left(1 - \left(\Delta\hat{B}_0(t_k) + \Delta\hat{B}_1(t_k)z_{1,i} + \cdots + \Delta\hat{B}_p(t_k)z_{p,i}\right)\right).\end{aligned}$$

Since there is no condition on $\Delta\hat{H}(t_k|\boldsymbol{z}_i)$ being positive, the resulting function $\tilde{S}_i(t|\boldsymbol{z}_i)$ is not monotone. To obtain a monotone estimate of the survival curve, perform isotonic regression on $\tilde{F}_i(t|\boldsymbol{z}_i) = 1 - \tilde{S}_i(t|\boldsymbol{z}_i)$ to obtain $\hat{F}_i$ and the survival curve $\hat{S}_i(t|\boldsymbol{z}_i) = 1 - \hat{F}_i(t|\boldsymbol{z}_i)$.

The truncation variables e_i^* can be sampled from the NPMLE

$$\hat{F}_E\left(e, \hat{S}\right) = \frac{1}{\sum_{j=1}^{n} \frac{1}{\hat{S}_j(e_j|\boldsymbol{z}_j)}} \sum_{e_i \leq e} \frac{1}{\hat{S}_i(e_i|\boldsymbol{z}_i)}.$$

Obtain an observation time t_i^* and a truncation time e_i^* by sampling from $\hat{S}_i$ and $\hat{F}_E\left(e, \hat{S}\right)$, respectively. If $t_i^* \geq e_i^*$ keep the samples. Repeat until samples t_i^* and e_i^* are obtained for each i.

The differences d_i^* can be obtained by sampling from the distribution Q of $D_i = C_i - E_i$ which is estimated by

$$\hat{Q}(d) = \begin{cases} 1 & \text{if } d \geq \max_i\{\tau_i - E_i\}, \\ 1 - \prod_{d_i \leq d}\left(1 - \frac{1}{Y'(d_i)}\right) & \text{if } d < \max_i\{\tau_i - E_i\}, \end{cases}$$

where d_i are the observed differences $D_i = C_i - E_i$, the cardinality of the risk set at time d is given by $Y'(d) = \sum_{i=1}^{n} \mathbb{1}\{\tau_i - e_i \geq d\}$. The censoring times are given by $c_i^* = e_i^* + d_i^*$.

After generating e_i^*, t_i^* and c_i^*, the bootstrapped sample is constructed as $(e_i^*, \tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

Method 3: Let $\boldsymbol{z}_i = (z_{1,i}, ..., z_{p,i})^T$ be the fixed covariates for individual i and generate t_i^* and e_i^* as in Method 2. The differences pattern is an ancillary statistic since the distribution of D_i does not depend on the parameters of Aalen's model. If an observation from D_i is observed, i.e. $\delta_i = 0$, let d_i^* be equal to $c_i - e_i$. Otherwise, if $\delta_i = 1$, draw the difference time d_i^* from the distribution function estimate $\hat{Q}$ of D given $D_i \geq t_i - e_i$. The censoring times are given by $c_i^* = e_i^* + d_i^*$.

After generating e_i^*, t_i^* and c_i^*, the bootstrapped sample is constructed as $(e_i^*, \tau_i^*, \delta_i^*, \boldsymbol{z}_i^T)$, where $\tau_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = \mathbb{1}\{t_i^* \leq c_i^*\}$. □

5 Implementation

5.1 Data

This project employs data from the Malmö diet and cancer study, which is a population-based prospective cohort study that recruited middle-aged men and women residing in Malmö during the early 1990s. The baseline screening of the study was conducted between 1991 and 1996, and the participants were monitored until 31st December 2019.

The data set consists of baseline measurements, an indicator of major adverse cardiovascular event (MACE), and the duration of the period from baseline until the first MACE, death, emigration or last follow-up. The data set components considered are listed in Table 1.

The response variable to study is the age at which individuals experience the first MACE. In this study, individuals begin to be followed at a specific age and are monitored from that point, which leads to the data being left-truncated. Thus, the age at entry into the study is used as a left-truncation time instead of a covariate. The variable of interest is the sum of fumc and age, which represents the age at which an individual experiences the first MACE, death, emigration, or the end of the follow-up period.

The data set contains missing values, and individuals with incomplete information are excluded from the analysis. The original data set included information about 30,447 individuals, but after removing those with missing data, the remaining sample size is 27,679.

The goal of the analysis is to examine the risk factors associated with the incidence of MACE during the study period, identify potential causal relationships between covariates, and calculate the direct effects between the variables and MACEs, utilising bootstrapping to compute confidence intervals for these effects. The statistical analysis was conducted in R, using the `survival` package [14].

Table 1: Data set variables

Covariate	Information
sex	Sex (0 = male, 1 = female)
age	Age at baseline (years)
systolic	Systolic blood pressure at baseline (mmHg)
diastolic	Diastolic blood pressure at baseline (mmHg)
pr-dm	History of prevalent diabetes event (0 = no, 1 = yes)
BMI	Body mass index (BMI) at baseline (kg/m2)
ApoA-I	Apolipoprotein A-I (ApoA-I) level at baseline (mg/dl)
ApoB	Apolipoprotein B (ApoB) level at baseline (mg/dl)
current-smoker	Smoking status at baseline (0 = no, 1 = yes)
lipid-low	Low lipid levels at baseline (0 = no, 1 = yes)
AHT	Anti-hypertensive therapy (AHT) (0 = no, 1 = yes)
hypertension	Hypertension at baseline (0 = no, 1 = yes)
inc-mc	History of MACE incidence until last follow-up (0 = no, 1 = yes)
fumc	Follow-up period from baseline until first MACE, death, emigration, or last follow-up (years)

5.2 Causality

This section examines the causal relationships between variables and the incidence of MACEs. Specifically, the focus is on apolipoproteins, including apolipoprotein A-I (ApoA-I) and apolipoprotein B (ApoB), which are accurate markers of cardiovascular risk. Apart from their direct effects on MACE incidence, these apolipoproteins also influence other variables. Additionally, smoking has been linked to decreased ApoA-I levels and elevated ApoB levels, further affecting cardiovascular health.

Apolipoproteins are proteins that bind to lipids to form lipoproteins, which are complex particles that transport lipids throughout the body. There are several classes of apolipoproteins, including apolipoprotein A-I (ApoA-I) and apolipoprotein B (ApoB). ApoB is a component of low-density lipoprotein (LDL). LDL cholesterol is colloquially referred to as 'bad cholesterol' because high levels of LDL and ApoB are associated with an increased risk of atherosclerosis, the underlying cause of heart attack and stroke [11]. Conversely, ApoA-I is the major structural protein component of high-density lipoprotein (HDL). HDL cholesterol is colloquially referred to as 'good cholesterol' because it absorbs excess cholesterol in the blood and carries it back to the liver for excretion, thus preventing the accumulation of cholesterol in the arterial wall. High levels of HDL and ApoA-1 are protective against atherosclerosis [11].

The levels of apolipoproteins in the body can affect the risk of cardiovascular problems by influencing lipid metabolism and blood clotting. In particular, high ApoB levels are associated with an increase in cardiovascular risk, whereas high ApoA-I levels are associated with a reduction in cardiovascular risk.

Besides the direct effect of ApoA-I and ApoB on the incidence of MACEs, ApoA-I and ApoB levels also influence other variables. Studies on mice have shown that ApoA-I has anti-obesity properties [13], implying that ApoA-I affects BMI, which is also associated with coronary heart disease. ApoA-I, ApoB and high blood pressure are also related since high cholesterol levels can cause plaque build-up in the arteries. Plaque build-up can restrict blood flow and increase blood pressure. A variable that influences apolipoprotein levels is smoking. Smoking has been associated with decreased levels of ApoA-I and with elevated levels of ApoB [10].

Taking into account the previously known risk factors and relations between these, we have in Figure 2 hypothesised causal relations between variables and the outcome. All the direct effects between the variables and the MACE are considered. The only direct effects between the variables that are examined are linear, since these can be combined with Aalen's additive hazards model to study causality.

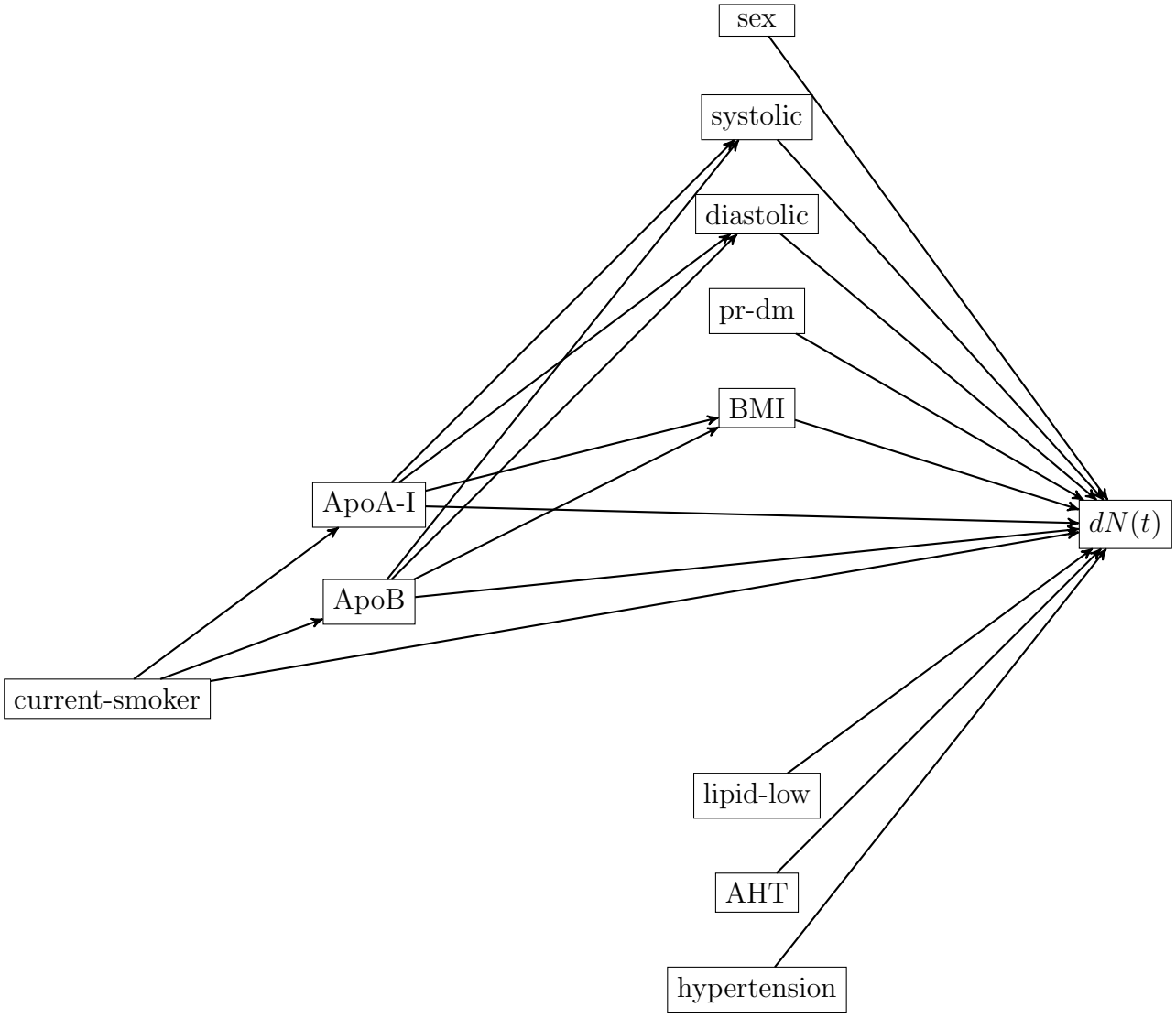


Figure 2: Hypothesised causal path between variables and the aggregated counting process $dN(t)$ of the number of MACE occurred by time t

5.3 Direct effects of covariates on MACEs

The cumulative direct effects of the variables on the number of MACEs are given by the cumulative regression functions from Aalen's additive hazards model. These are computed using the function `aareg` from the `survival` package.

We created $n_B = 100$ bootstrapped samples from the l.t.r.c. data using Method 2, provided in Section 4.3. We obtained the survival curve estimates $\hat{S}_i$ using the function `isoreg` to perform isotonic regression. To sample the times t_i^* from the survival curve, we performed inverse transform sampling.

Inverse transform sampling is a method to sample from any distribution or, equivalently, survival function S. We will apply the method to sample from the survival function estimate $\hat{S}$. The method consists of drawing a value u from a uniform distribution $U \in \text{Un}(0, 1)$. Since the survival curve estimates are possibly defective if the value $u < \hat{S}(t_{(\max_k)})$ is less than the minimum value of $\hat{S}$, the sampled value t_i^* equals the time $t_{(\max_k)}$. If u is greater or equal than the minimum value of $\hat{S}$, the sampled value t_i^* is equal to the time $t_{(k)}$ such that $\hat{S}(t_{(k)}) \leq u < \hat{S}(t_{(k-1)})$, since $\hat{S}(t)$ is a right-continuous step-function. These two cases are illustrated in Figure 3.

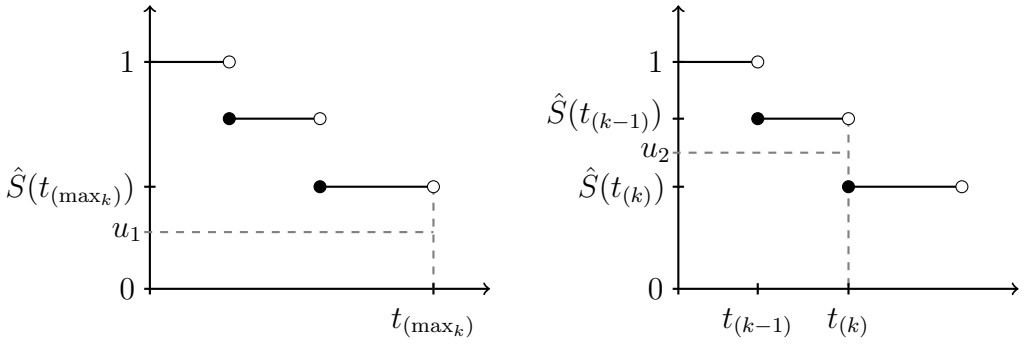


Figure 3: Inverse transform sampling from a possibly defective survival curve estimate

For each bootstrapped sample $\{(e_i^{(b)}, \tau_i^{(b)}, \delta_i^{(b)}, \boldsymbol{z}_i^T)\}_{i=1}^n$, $b = 1, ..., n_B$, we computed the cumulative regression functions for each of the variables. At each time point, we compute confidence intervals at 95% confidence level. The resulting estimates and confidence intervals are presented in Figure 4 and 5.

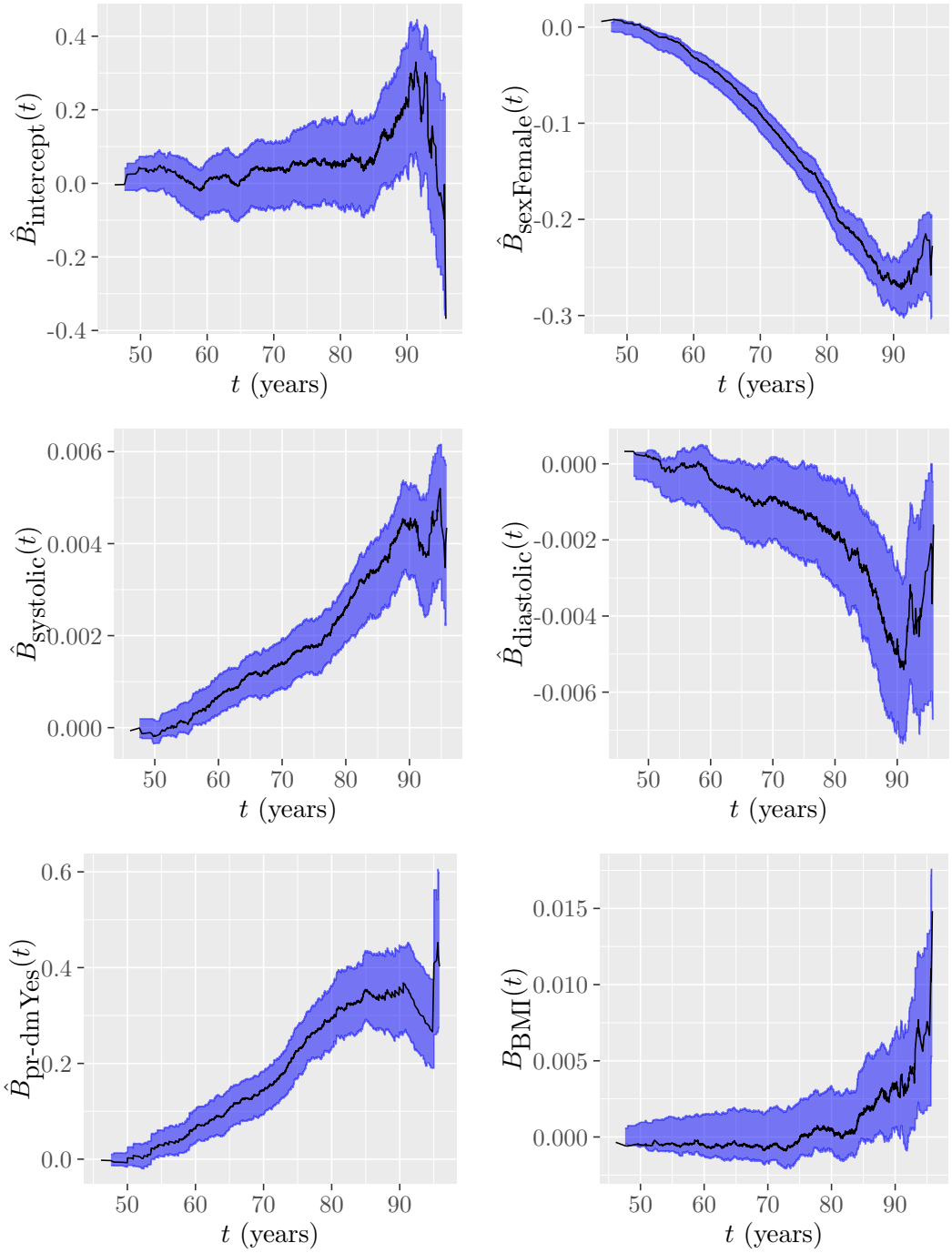


Figure 4: Estimates of the cumulative direct effects of the variables sex, systolic, diastolic, pr-dm and BMI on the incidence of MACE (black) and 95% point-wise bootstrapped confidence intervals (blue)

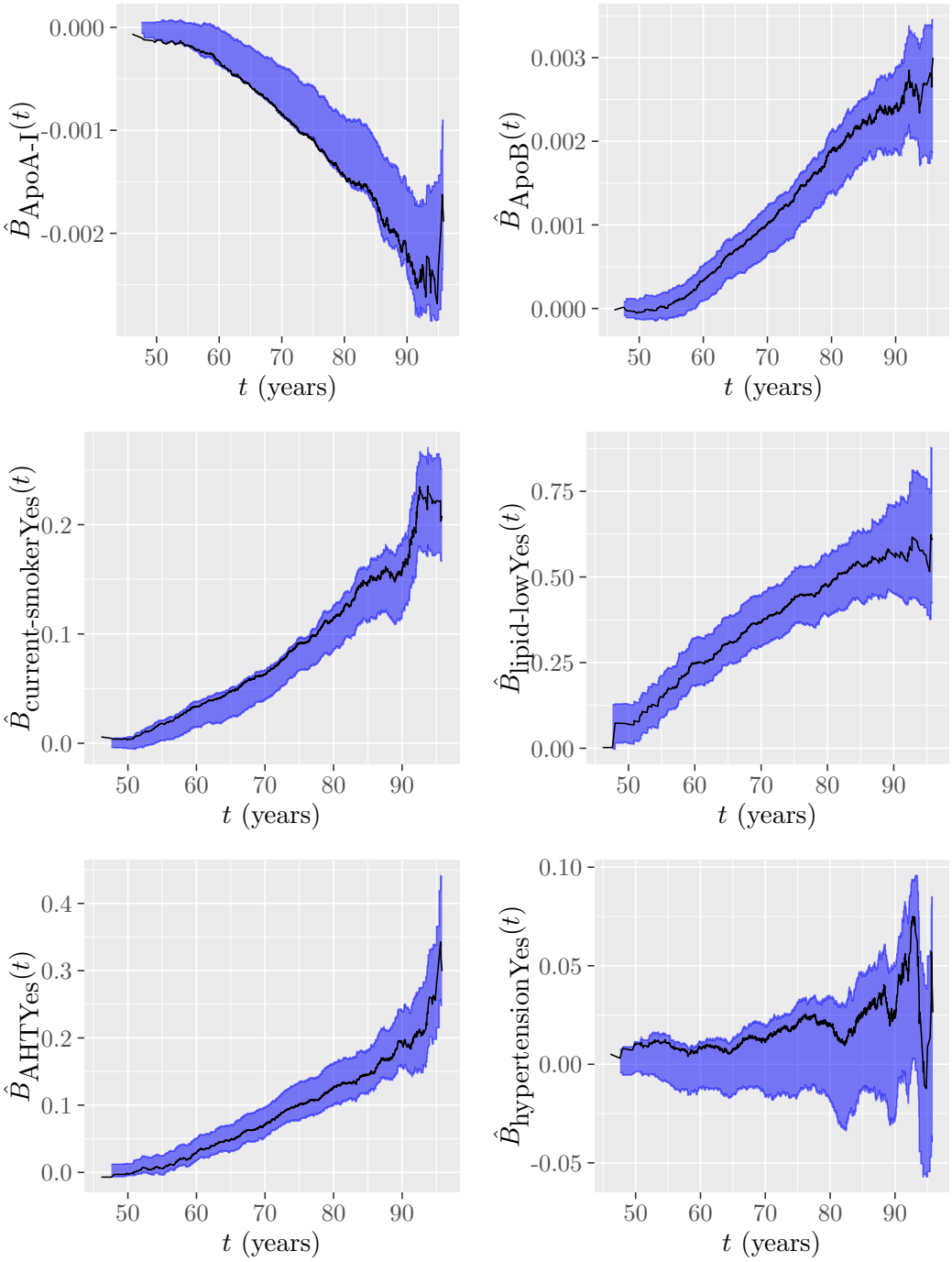
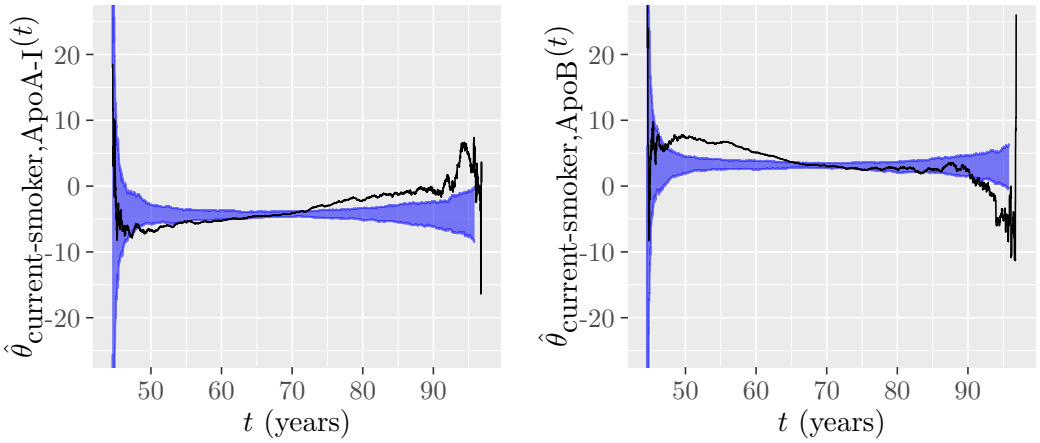


Figure 5: Estimates of the cumulative direct effects of the variables ApoA-I, ApoB, current-smoker, lipid-low, AHT and hypertension on the incidence of MACE (black) and 95% point-wise bootstrapped confidence intervals (blue)

5.4 Direct effects between covariates

We computed the direct effects between the variables by regressing each variable on its hypothesised parents, as presented in Figure 2. The coefficients are calculated point-wise on the risk set at time t using the function `lm`.

For each bootstrapped sample $\{(e_i^{(b)}, \tau_i^{(b)}, \delta_i^{(b)}, \boldsymbol{z}_i^T)\}_{i=1}^n$, $b = 1, ..., n_B$, we computed the coefficients of each variable on its parents on the risk set at time t. At each time point, we calculated confidence intervals at 95% confidence level. The resulting estimates and confidence intervals are presented as follows. Figure 6a and 6b present the results from regressing the variables ApoA-I and ApoB on current-smoker, respectively. Figure 7 presents the results from regressing the variable systolic on ApoA-I and ApoB. Figure 8 presents the results from regressing the variable diastolic on ApoA-I and ApoB. Finally, Figure 9 presents the results from regressing the variable BMI on ApoA-I and ApoB.



(a) ApoA-I is regressed on current-smoker

(b) ApoB is regressed on current-smoker

Figure 6: Estimates of the direct effects of the variable current-smoker on ApoA-I and ApoB (black) and 95% point-wise bootstrapped confidence intervals (blue)

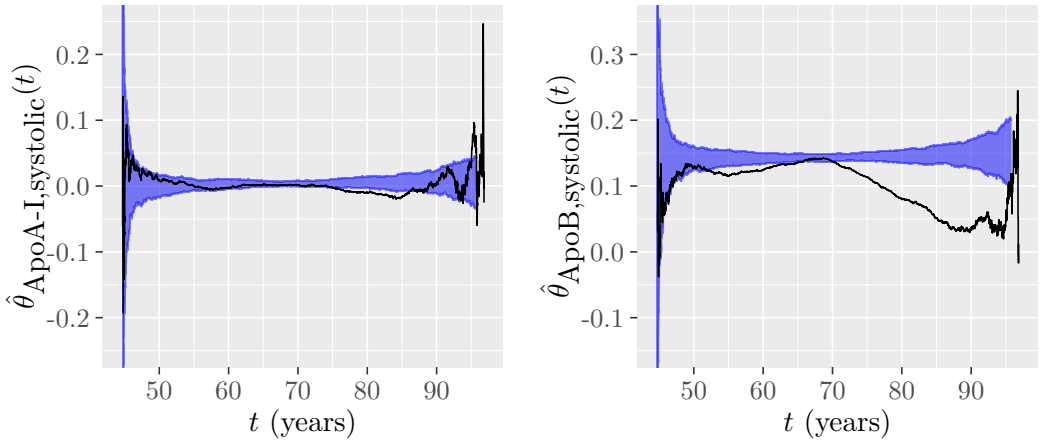


Figure 7: Estimates of the direct effects of the variables ApoA-I and ApoB on systolic (black) and 95% point-wise bootstrapped confidence intervals (blue)

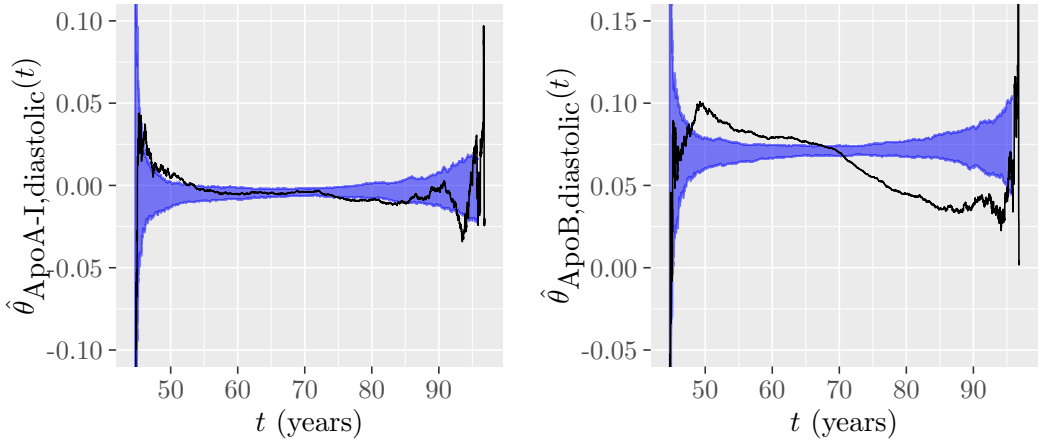


Figure 8: Estimates of the direct effects of the variables ApoA-I and ApoB on diastolic (black) and 95% point-wise bootstrapped confidence intervals (blue)

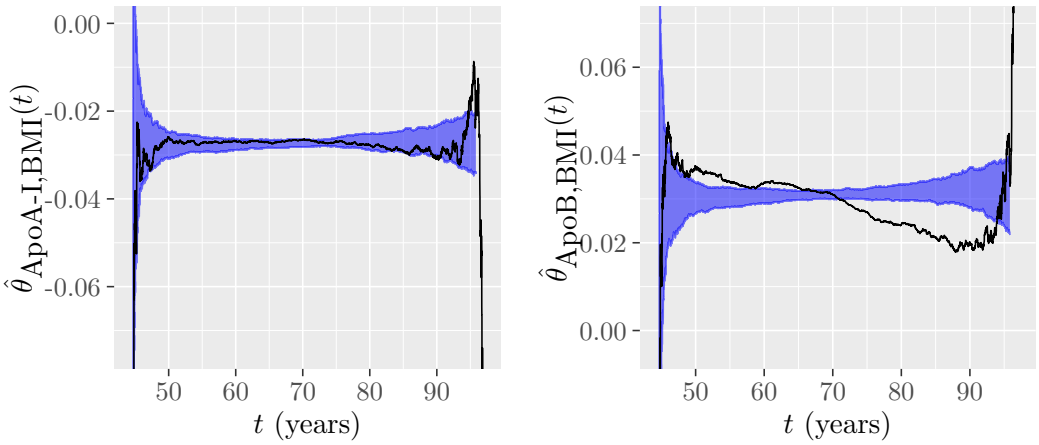


Figure 9: Estimates of the direct effects of the variables ApoA-I and ApoB on BMI (black) and 95% point-wise bootstrapped confidence intervals (blue)

5.5 Interpretation of results

This section focuses on interpreting the previous results. We begin by discussing the results in Section 5.3.

The cumulative regression functions describe how the covariates influence survival over time. If the cumulative regression function estimate is constant, the effect of the covariate does not vary with time. On the other hand, if the estimate is increasing or decreasing, there is a time-varying effect of the covariate.

For numerical variables, given that it takes positive values, estimates that are increasing indicate an increase in the hazard. Conversely, for categorical variables, the cumulative regression function compares the effect of the variable with respect to the reference category (assigned a value of 0). The cumulative regression function for the reference category of all categorical variables is given by the intercept. An increase in the cumulative regression function for the other categories indicates an increase in the hazard compared to the reference category.

Examining Figure 4, the cumulative regression function for the intercept reveals a higher hazard rate for individuals in older age groups, particularly

those nearing 90 years old, in experiencing a MACE. Furthermore, the regression function for female individuals indicates that females, compared to males, have a lower hazard rate of experiencing a MACE.

Regarding blood pressure, higher systolic blood pressure appears to slightly increase the hazard rate for MACE from ages 60 onwards, while higher diastolic blood pressure seems to slightly decrease this hazard rate from ages 70 onwards.

Individuals with prevalent diabetes before entering the study (pr-dm) seem to have a higher hazard rate of experiencing a MACE. Lastly, the cumulative regression function for the variable BMI indicates that BMI does not significantly effect MACEs, except for older individuals where the hazard increases.

Moving to Figure 5, it appears that high levels of ApoA-I lower the hazard rate of MACE. Conversely, high ApoB levels increase the hazard rate. Individuals who smoke also have a greater hazard rate compared to non-smokers. Low levels of lipids and anti-hypertensive therapy (AHT) are two factors that increase the hazard rate for MACE. Hypertension, however, does not seem to affect the hazard.

Nevertheless, some of these plots indicate that the bootstrapped confidence intervals are biased, particularly for the cumulative regression function for ApoA-I, ApoB, current-smoker and hypertension.

Additionally, the results from Section 5.4 further support that the original estimates are biased. Upon exploring Figure 6 - 9, it becomes apparent that the confidence intervals only capture the average direct effect between the variables but lacks variability in the individual bootstrapped samples.

5.6 Method comparison

The same confidence intervals calculated using Method 2 are computed using Method 1. This is done to demonstrate the possible advantages and limitations of Method 2 since Method 1 is a naive way of sampling. The results are presented in Appendix 1. This section focuses on comparing the quality of the results obtained for Method 1 and 2.

The confidence intervals for Method 1 do not present bias, whereas some of the plots for Method 2 do. Nonetheless, it is worth noting that some of the confidence intervals computed for Method 1 show higher variance compared to their counterpart for Method 2.

In total, the bootstrapped results present bias and show a somewhat odd lack of variability in the regression coefficients. We would like to emphasise that our suggested method is, to the best of our knowledge, new, has not been studied before, and there is, of course, the question of whether 'bootstrapping works'.

6 Conclusions, discussion, and open problems

The main goal of this thesis was to study bootstrapping methods for survival data, in particular, for right-censored and left-truncated right-censored (l.t.r.c.) data, and in combination with graphical models for assessing causal relations and direct/indirect effects of covariates. This thesis proposes three methods for generating bootstrapped samples for l.t.r.c. data. Method 2 and 3 are novel and designed within the framework of Aalen's additive hazards model. The first method consists of sampling with replacement from the l.t.r.c. data. This method is referred to as the generalisation of Efron's 'simple' method for l.t.r.c. data [8]. The second method draws data conditional on the covariates under Aalen's model. Method 3 is an extension of Method 2, which is suitable under the assumption that the censoring time is an ancillary statistics.

The second method is applied to the Malmö diet and cancer study data set to compute confidence intervals for the effects of the covariates on MACEs and between the variables. The method seems to generate results that capture underlying trends. However, some confidence intervals for the effects of covariates on the output seem to indicate that the original estimates are biased. Also, the method cannot effectively compute the confidence intervals for the effects between the variables. Conversely, Method 1, bootstrapping with replacement from l.t.r.c. data, provides unbiased confidence intervals.

Therefore, we encourage further research to solve the limitations of Method 2. Concerning the bias, studies with an empirical approach could perform a simulation study to better understand the factors contributing to the observed biases in the results. This can help identify whether the bias is due to implementation issues or limitations with the suggested method. Concerning the estimation technique, the survival function obtained with the plug-in of Aalen's hazard is not monotone. However, by the use of antitonic regression, we obtain a monotone survival curve. It might be fruitful studying how other ways of estimating the survival curve impact the results.

We would like to suggest three alternative ways of continuing this project. First, studying different methods for sampling and extrapolating defective survival curves. In this thesis, the survival curve estimates were found to be defective and drawing from them was done by an inverse transform sampling approach. It would be of interest to study ways to extrapolate the survival

curve to unseen times.

Second, the approach employed to investigate causality in survival only considers linear effects between the covariates. Another approach to continuing this project would be exploring methods to incorporate non-linear effects of variables in the study of causality in survival. This would enable a more comprehensive analysis of the relationships between covariates and survival outcomes.

Third, it is of interest to consider alternative approaches for studying the effects of covariates on survival that do not assume time-varying effects. While this thesis primarily focuses on investigating time-varying effects, it is important to contemplate the possibility that the resulting effects may not exhibit variations over time. In such instances, computing non-time-varying effects would be a simpler and more efficient approach for studying causality in survival.

References

[1] O. O. Aalen, Ø. Borgan, and H. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer New York, 2008.

[2] P. Andersen, Ø. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York, 1993.

[3] D. Burr. A comparison of certain bootstrap confidence intervals in the cox model. *Journal of the American Statistical Association*, 89(428):1290–1302, 1994.

[4] D. Edwards. *Introduction to Graphical Modelling*. Springer New York, NY, 2000.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] B. Efron. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319, 1981.

[7] J. Fosen, E. Ferkingstad, Ø. Borgan, and O. O. Aalen. Dynamic path analysis-a new approach to analyzing time-dependent covariates. *Lifetime Data Anal.*, 12(2):143–67, 2006.

[8] S. T. Gross and T. L. Lai. Bootstrap methods for truncated and censored data. *Statistica Sinica*, 6(3):509–530, 1996.

[9] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[10] A. R. Kauss, M. Antunes, G. de La Bourdonnaye, S. Pouly, M. Hankins, A. Heremans, and A. van der Plas. Smoking and apolipoprotein levels: A meta-analysis of published data. *Toxicology reports*, 9:1150–1171, 2022.

[11] M. F. Linton, P. G. Yancey, S. S. Davies, W. G. Jerome, E. F. Linton, W. L. Song, A. C. Doran, and K. C. Vickers. *The Role of Lipids and Lipoproteins in Atherosclerosis*. Comprehensive Endocrinology Book, 2019.

[12] T. Robertson, F. Wright, and R. Dykstra. Order restricted statistical inference. *Statistical Papers*, 30(316):1613–9798, 1989.

[13] X. Ruan, Z. Li, Y. Zhang, L. Yang, Y. Pan, Z. Wang, G. S. Feng, and Y. Chen. Apolipoprotein A-I possesses an anti-obesity effect associated with increase of energy expenditure and up-regulation of UCP1 in brown fat. *Journal of cellular and molecular medicine*, 15(4):763–772, 2011.

[14] T. M. Therneau. *A Package for Survival Analysis in R*, 2023. R package version 3.5-5.

[15] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[16] M.-C. Wang. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143, 1991.

[17] S. Wright. Correlation and causality. *Journal of the Agritultural Research*, 20:557–585, 1921.

Appendix 1

Results for Method 1: Direct effects of covariates on MACEs

The cumulative direct effects of the variables on the number of MACEs are given by the cumulative regression functions from Aalen's additive hazards model. These are computed using the function `aareg` from the `survival` package.

We created $n_B = 100$ bootstrapped samples from the l.t.r.c. data using Method 1, provided in section 4.3. Given the original data, sample with replacement to obtain a data set of size n. For each of bootstrapped data set, we computed the cumulative regression functions for each of the variables. At each time point, we compute confidence intervals at 95% confidence level. The resulting estimates and confidence intervals are presented in Figure 10 and 11.

Results for Method 1: Direct effects between covariates

We computed the direct effects between the variables by regressing each variable on its hypothesised parents, as presented in Figure 2. The coefficients are calculated point-wise on the risk set at time t using the function `lm`.

For each bootstrapped sample $\{(e_i^{(b)}, \tau_i^{(b)}, \delta_i^{(b)}, \boldsymbol{z}_i^T)\}_{i=1}^n$, $b = 1, ..., n_B$, we computed the coefficients of each variable on its parents on the risk set at time t. At each time point, we calculated confidence intervals at 95% confidence level. The resulting estimates and confidence intervals are presented as follows. Figure 12a and 12b present the results from regressing the variables ApoA-I and ApoB on current-smoker, respectively. Figure 13 presents the results from regressing the variable systolic on ApoA-I and ApoB. Figure 14 presents the results from regressing the variable diastolic on ApoA-I and ApoB. Finally, Figure 15 presents the results from regressing the variable BMI on ApoA-I and ApoB.

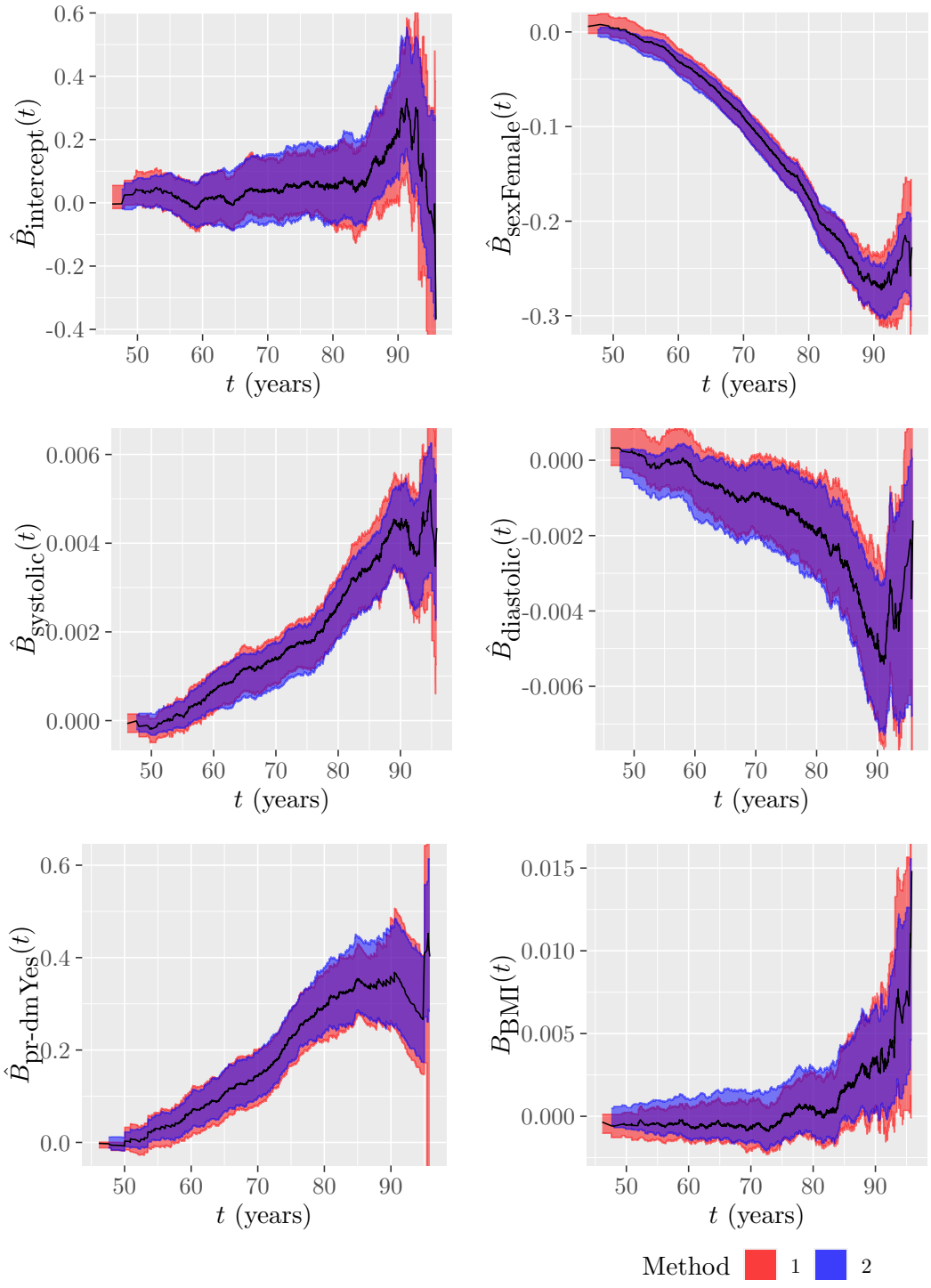


Figure 10: Estimates of the cumulative direct effects of the variables sex, systolic, diastolic, pr-dm and BMI on the incidence of MACE (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2

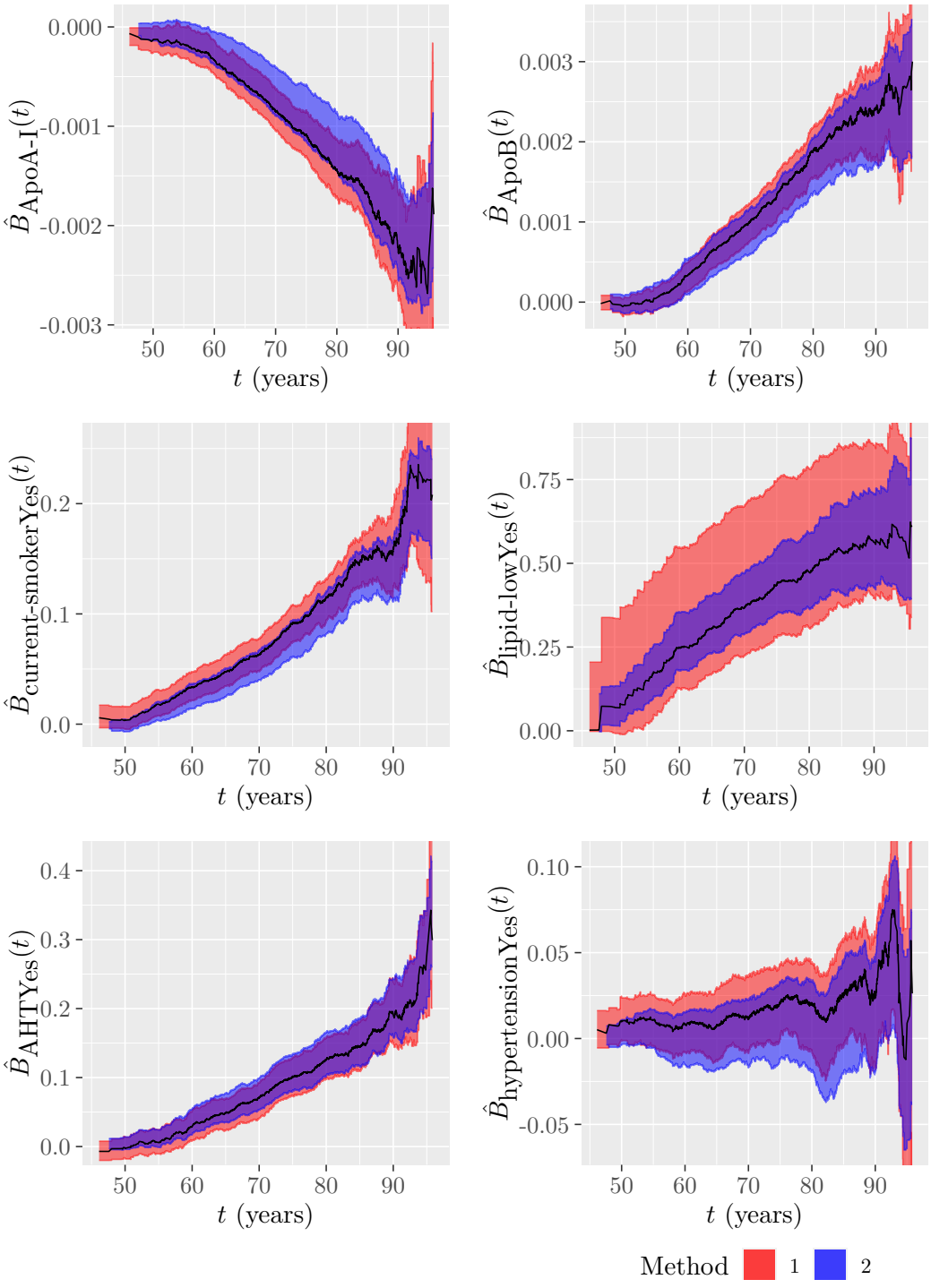
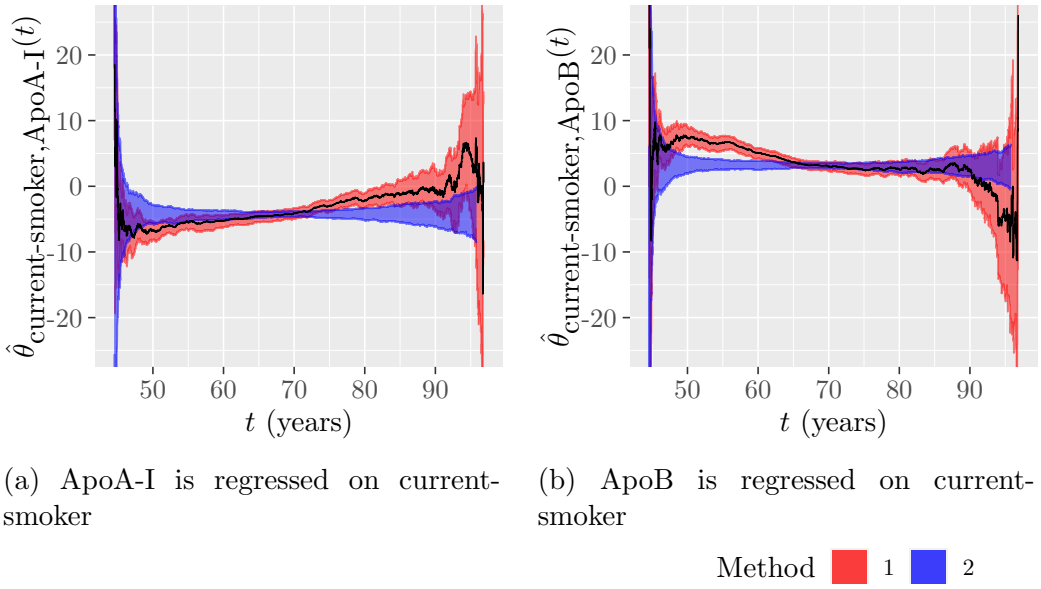


Figure 11: Estimates of the cumulative direct effects of the variables ApoA-I, ApoB, current-smoker, lipid-low, AHT and hypertension on the incidence of MACE (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2



(a) ApoA-I is regressed on current-smoker

(b) ApoB is regressed on current-smoker

Figure 12: Estimates of the direct effects of the variable current-smoker on ApoA-I and ApoB (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2

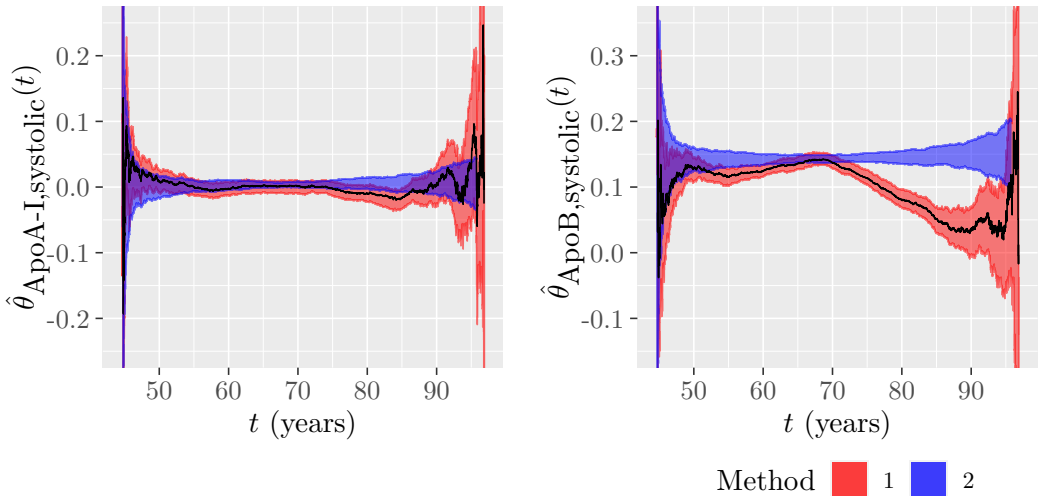


Figure 13: Estimates of the direct effects of the variables ApoA-I and ApoB on systolic (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2

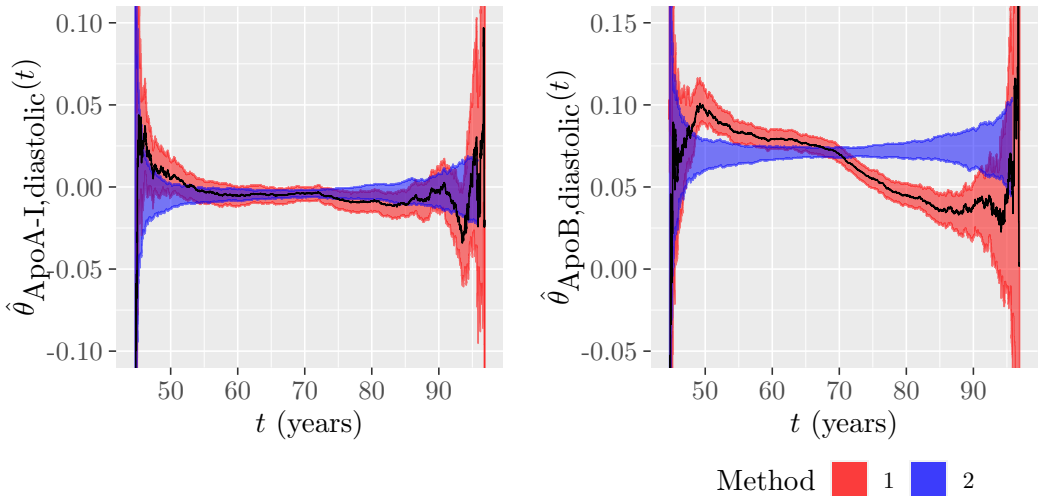


Figure 14: Estimates of the direct effects of the variables ApoA-I and ApoB on diastolic (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2

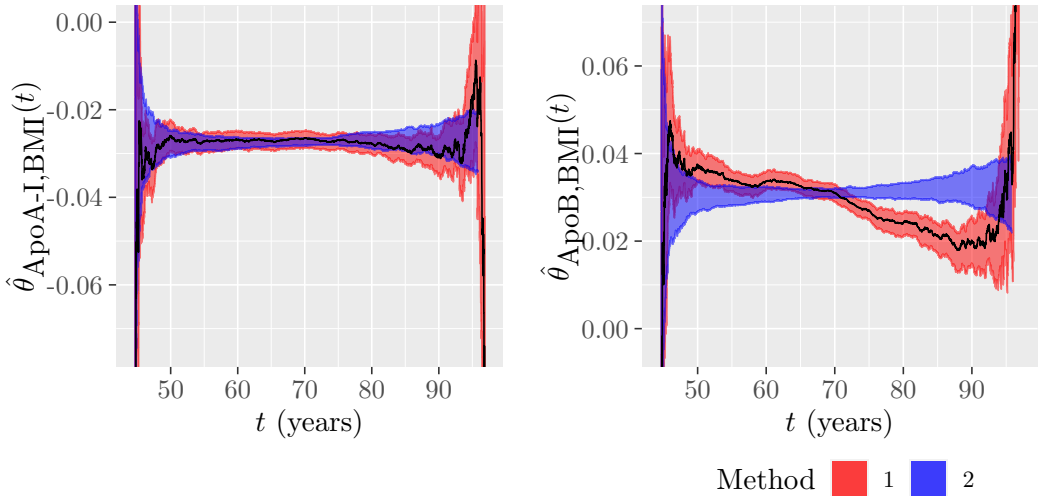


Figure 15: Estimates of the direct effects of the variables ApoA-I and ApoB on BMI (black) and 95% point-wise bootstrapped confidence intervals for Method 1 and 2

Master's Theses in Mathematical Sciences 2023:E42
ISSN 1404-6342

LUNFMS-3119-2023

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/