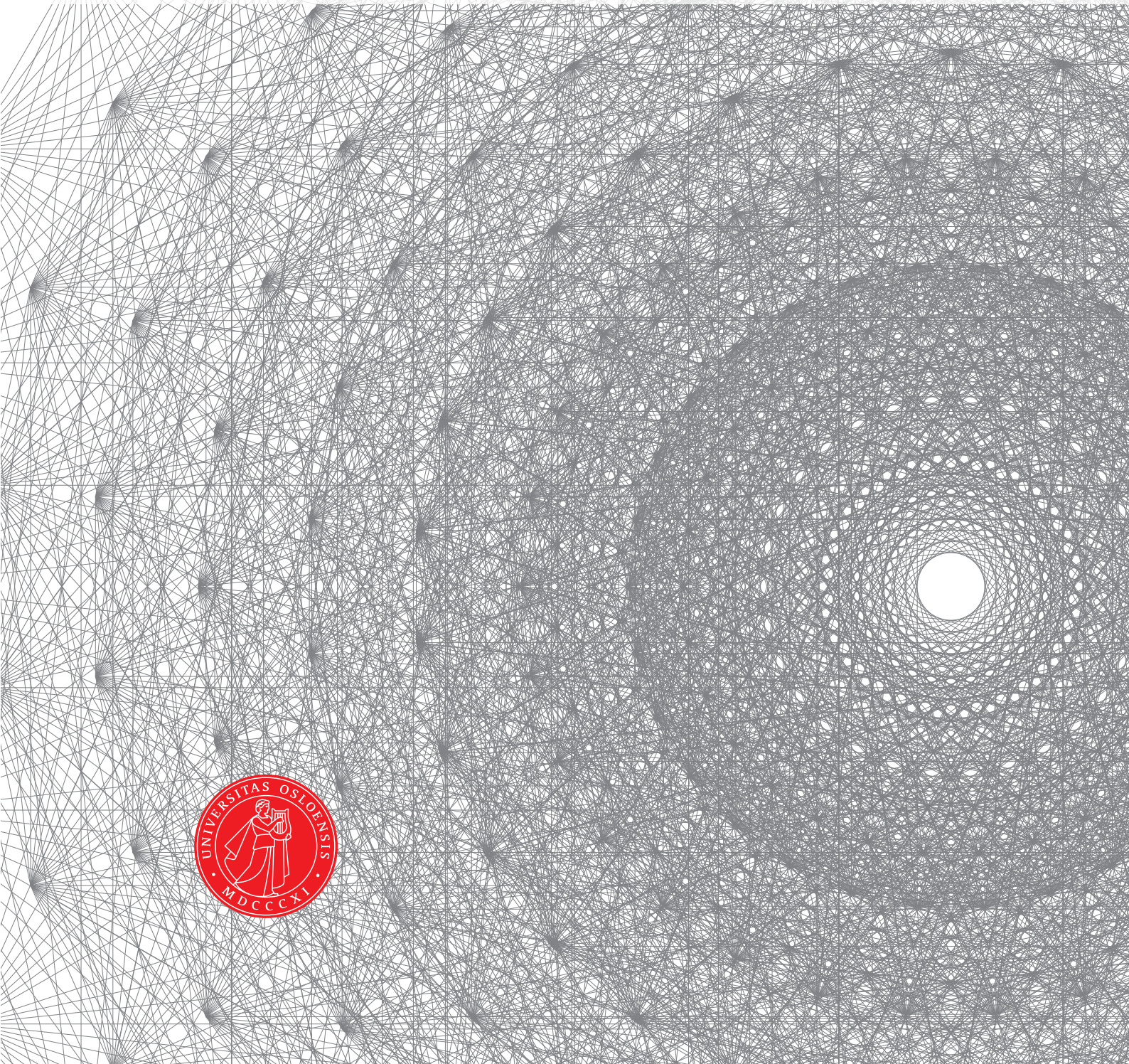
UiO : Department of Mathematics
University of Oslo

Two-stage predictor substitution for time-to-event data

Simon Lergenmuller

Master's Thesis, Spring 2017



This master's thesis is submitted under the master's programme *Modelling and Data Analysis*, with programme option *Statistics and Data Analysis*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

This thesis is devoted to presenting and illustrating a novel estimation method offering a way to reduce confounding bias in time-to-event situations. In order to reduce this bias, without having to observe all of the confounding, one can use *instrumental variables* (IV). These are observed variables independent of the unknown confounder, correlated with the exposure of interest and affecting the response only through the exposure. Under these assumptions, one can perform a *two-stage predictor substitution*, that allows for consistent estimation of the hazard difference for Aalen's additive hazard model, and under the additional assumption of rare outcome, allows for approximate consistent estimation of the hazard ratio for Cox's proportional hazards model.

For the aforementioned models, we then illustrate the consistency of this estimate for varying IV strength, and investigate the meaning of "rare outcome", by illustrating how inconsistent the estimator can become as the number of outcomes increases.

We also found, somewhat surprisingly, that bootstrapping might not necessarily be the best choice for estimating the variance of the IV estimate, as both the aalen function and the coxph function in R already seem to report the correct variance. As for small samples, bootstrapping an estimate of the variance for Cox's proportional hazard might be cumbersome.

Under misspecification of the first stage, only the two-stage predictor substitution for Aalen's additive hazard model seems to still yield consistent estimates but the resulting variances are very high in comparison to a correctly specified first stage. We thus advise against misspecification. However, for testing the null hypothesis of no causal effect, both Aalen's additive hazard model and Cox's proportional hazards model seem to perform well under misspecification of the first stage and with an arbitrary number of outcome.

We then perform an analysis of the effect of mothers body mass index (BMI) on the pregnancy duration using data from the Norwegian Mother and Child cohort. Even though previous studies showed a clear effect of the mother's BMI on premature birth, this analysis shows only a very small overall effect of the mother's BMI. We hope, however, that this analysis can provide the reader with an illustration on how to apply the method to a real data set.

Acknowledgments

First and foremost, I would like to thank my supervisor S. O. Samuelsen for introducing me to a method that I was completely unfamiliar with prior to beginning this thesis. Thank you for your support, your optimism, and thank you for the time you took to guide me through this Master.

I would also like to thank the amazing people I was lucky to share the reading room with, and the equally amazing people I was lucky enough to spend two years of coffee and laughs with. Thank you Molly Wood for giving me access to the Norwegian Mother and Child cohort, making it possible to illustrate the methods presented in this thesis on real data. I am also very grateful to my family and friends for always supporting me.

A special thanks goes to Mona for being there when I needed it the most.

Simon Lergenmuller
Oslo, May 2017

Contents

1

Introduction

In clinical, epidemiological and economical studies, it is very common that there are unobserved confounders. In the case of non-randomized studies, one can seldom get rid of the confounding bias. Because they often are not measured, and because unknown confounders can either influence the outcome or the exposure or both, fitting a model ignoring these will typically lead to inconsistent estimation of the measure of association of interest (Didelez et al., 2010; Li et al., 2014).

One way to reduce this bias, without having to observe all of the confounding, is to use *instrumental variables* (IV). These are observed variables satisfying some core assumptions. In short, they are independent of the unknown confounder, correlated with the exposure of interest and affecting the response only through the exposure. Under these assumptions, an IV can offer a way to obtain consistent estimates of the measure of association of interest. Instrumental variable estimation has been mainly used in economical studies, and more recently in epidemiological studies. In economics, instrumental variables are often seen in structural equation models (Angrist et al., 1996). In epidemiological studies, instrumental variables are often used in the context of Mendelian randomization (Didelez and Sheehan, 2007; Didelez et al., 2010), in the form or genes closely associated with the exposure of interest.

The first time instrumental variable appears in the literature is perhaps in the appendix of a book written by Wright (1928), but it is the Norwegian statistician Olav Reiersøl, who in 1945 used a similar approach to Philip G. Wright, and coined the term "instrumental variable" (Reiersøl, 1945). Instrumental variable estimation methods have since then been well studied in the context of linear regression (Leigh and Schembri, 2004) and for certain generalized linear models (Didelez et al., 2010). The traditional way to perform an instrumental variable estimation is a method called *two-stage least squares regression*, which, as the name suggests, entails performing two ordinary least squares regressions after one another. Recent developments (Li et al., 2014; Tchetgen et al., 2015), some of which will be presented here, ad-

dress the problem of bias reduction in the context of time-to-event data. A two stage procedure is also used to achieve this, but in order to differentiate it from the one mentioned earlier and in order to avoid confusion as to which method is used at each stage, we will simply refer to it as *two-stage predictor substitution*.

This thesis will present these latest results, and attempt to derive some of them in a somewhat different fashion. Instrumental variable estimation in epidemiological studies often requires prior knowledge of causal inference calculus (Pearl, 2009), and, this being reputed difficult literature, we will try to minimize it here.

Perhaps because IV estimation for time-to-event data is a fairly novel method, very few simulation studies have been done, except in Li et al. (2014), where the authors have illustrated one very specific situation, that of the additive hazard model, with constant parameters, also called the Lin-Ying model (Lin and Ying, 1995). Therefore, and because one may argue that in statistics it is often important and useful to present illustrations and examples, a large part of this thesis will be focused on simulations, from the simplest situation to some more complicated ones.

Finally, and for completeness, there will be an attempt to apply these methods to real data, namely to the Norwegian Mother and Child cohort (MoBa), where we will look at the event "time until birth", or "time until premature birth". One of the challenges will be to find appropriate instrumental variables for the exposures we wish to study.

2

Methods

Before we get into the core of the subject, we need to define the models that will be studied in the thesis, as well as the ones needed to illustrate the simplest situations. We will start with the latter so that we can introduce the two-stage predictor substitution. Only the necessary material needed to do the subsequent demonstrations will be presented here.

In what follows, unless otherwise specified, we will use uppercase letters such as X, Y, or L, when referring to the general aspect of a random variable. We will use lowercase letters for observed values, thus when referring to i'th observed value of X, we will write $\boldsymbol{x}_i$ if it is a vector, and x_i (or x_{ij} for the j'th value of $\boldsymbol{x}_i$) if it is a scalar. Furthermore, matrices of observed values will be denoted by uppercase bold letters such as $\boldsymbol{X}$.

2.1 Generalized linear models, an overview

Generalized linear models (GLM) are a class of models used to obtain information about the relationship between a response variable Y and explanatory variables. To asses this relationship, it is important to know the probabilistic behavior of Y. The distribution of the response variable is specific to the situation of interest. For example, if we consider the response to be the height of an individual (in cm), one can assume that Y follows a normal distribution. Another example could be if we consider the response to be counting the number of car crashes every year, then one may assume that Y follows a Poisson distribution.

In this section, we will present these two types of response, that is, when Y is continuous and normally distributed, and when Y is a natural number and Poisson distributed.

Assume $Y_1, \dots Y_n$ are independent and identically normally distributed with mean μ and variance σ. We write

$$Y_i \overset{iid}{\sim} N(\mu, \sigma^2) \quad i = 1, \dots, n,$$

and we can write the joint probability density function of $(Y_1, \ldots, Y_n)$ as

$$f(y_1, \ldots, y_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mu)^2). \tag{2.1}$$

Assume $Y_1, \ldots Y_n$ are independent and identically Poisson distributed with mean μ (and therefore variance μ). We write

$$Y_i \overset{iid}{\sim} Po(\mu) \quad i = 1, \ldots, n,$$

and we can write the joint probability density function of $(Y_1, \ldots, Y_n)$ as

$$f(y_1, \ldots, y_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{\mu^{y_i}}{y_i!} \exp(-\mu). \tag{2.2}$$

There exist a lot of literature on this subject, but we will not dwell on the details here. The two main aspects of GLMs are that the response is chosen from the exponential family of distributions, and that an adequate transformation (via a link function) of the expected value of the response given the covariates is expressing a linear relationship with the explanatory covariates.
Assume that we have $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ with joint distribution F_θ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is an unknown parameter. We say that a family of probability distribution functions is an *exponential family* if the density function of Y_i is of the form

$$f(y_i|\theta_i) = \exp(\theta_i y_i - c(\theta_i) + h(y_i)), \tag{2.3}$$

for $y_i \in A$, where A is some set that does not depend on θ_i and where $h(y_i)$ is a real-valued function also independent of θ_i. The function $c(\theta_i)$ is a real-valued function depending only on θ_i. We can *enrich* this frequency by an additional scale parameter $\phi \geq 0$, and write

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{\theta_i y_i - c(\theta_i)}{\phi} + h(y_i, \phi)\right). \tag{2.4}$$

When ϕ is known, this family of density functions is in the exponential family, when ϕ is unknown, it may or may not be. A common thing to do is to choose $\theta_i = g(\mu_i) = \boldsymbol{\beta'} \boldsymbol{x_i}$, with $\boldsymbol{\beta'} = (\beta_1, \ldots, \beta_p)$, and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})'$, in which case the family of distributions of $\boldsymbol{Y}$ are called a $(p+1)$-parameter exponential family. The link $g(\cdot)$ is then called the canonical link function.

As it turns out, both the normal distribution and the Poisson distribution belong to the exponential family (Piet de Jong, 2008, chapter 3)(Knight, 1999).

2.1.1 Normally distributed response – identity link

This model is maybe the simplest model of all, and there is a lot of literature available on the subject, from the most detailed and easiest to read (Jay L. Devore, 2011, p. 613-716), to some more advanced material (Friedman et al., 2009, p.43-56). We will assume that one is familiar with simple linear regression, and give a quick outline of the multivariate regression model.

Assume we have n observations, each consisting of a pair $(\boldsymbol{x_i}', y_i)$, where $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ with $p > 0$ and $y_i \in \mathbb{R}$, and that the response comes from a normal distribution with mean μ_i and variance σ^2. We can then express μ_i with explanatory variables through a link function $g(\cdot)$. We write (Piet de Jong, 2008, chapter 5)

$$Y_i \sim N(\mu_i, \sigma^2) \quad i = 1, \ldots, n,$$

where $g(\mu_i) = \boldsymbol{\beta'} \boldsymbol{x}_i \quad \forall \quad i = 1, \ldots, n$, with $\boldsymbol{\beta'} = (\beta_0, \beta_1, \ldots, \beta_p)$. With the identity link function $g(\mu_i) = \mu_i$ we can express the mean as

$$\mu_i = \boldsymbol{\beta'} \boldsymbol{x}_i, \tag{2.5}$$

and we can assume the following model for y_i:

$$y_i = \boldsymbol{\beta'} \boldsymbol{x_i} + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \quad i = 1, \ldots, n. \tag{2.6}$$

We want to estimate the $p \times 1$ vector $\boldsymbol{\beta}$ using the principle of least squares (Jay L. Devore, 2011, p.626), that is, we want to minimize

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2, \tag{2.7}$$

with respect to $\boldsymbol{\beta}$, where $\boldsymbol{X'} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Assuming that $\boldsymbol{X'X}$ has an inverse, by taking the partial derivative with respect to $\boldsymbol{\beta}$ and solving for $\boldsymbol{\beta}$, we obtain:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1} \boldsymbol{X'y}. \tag{2.8}$$

Alternatively, the solution of (2.7) can be found via maximum likelihood estimation, which we briefly describe in Section 2.1.3

2.1.2 Poisson distributed response – log-link

Assume now that we have n observations, each consisting of a couple $(\boldsymbol{x_i}', y_i)$, where $\boldsymbol{x_i}' \in \mathbb{R}^p$ with $p > 0$ and $y_i \in \mathbb{N}$, and that the response comes from a Poisson distribution with mean μ_i for individual i. We can then express μ_i with explanatory variables through a link function $g(\cdot)$. We write (Piet de Jong, 2008, chapter 6)

$$Y_i \sim Po(\mu_i), \tag{2.9}$$

where $g(\mu_i) = \boldsymbol{\beta'} \boldsymbol{x}_i \quad \forall \quad i = 1, \ldots, n$, with $\boldsymbol{\beta'} = (\beta_0, \beta_1, \ldots, \beta_p)$. With the log-link function $g(\mu_i) = \log(\mu_i)$, we can express the mean as

$$\mu_i = e^{\boldsymbol{\beta'} \boldsymbol{x}_i}. \tag{2.10}$$

One can then estimate the parameter $\boldsymbol{\beta'}$ using maximum likelihood estimation described briefly in the next section.

2.1.3 Maximum likelihood estimation

Assume $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are independent random variables with joint probability density function in the exponential family of distributions. If Y_i has the enriched distribution 2.4, we have the relations (Knight, 1999)

$$\mu_i = E(Y_i) = c'(\theta_i),$$

and

$$\mathrm{Var}(Y_i) = \phi c''(\theta_i),$$

where c' and c'' are the first and second derivatives of c. If c' is a one-to-one function, it follows that

$$\mathrm{Var}(Y_i) = \phi V(\mu_i),$$

where V is the variance function, indicating the relation between the mean and variance of the response. Assume we have a strictly increasing canonical link function $g(\cdot)$ such as

$$g(\mu_i) = g(c'(\theta_i)) = \boldsymbol{\beta'} \boldsymbol{x}_i = \theta_i, \tag{2.11}$$

where $\mu_i = \mathrm{E}(Y_i)$. Given $\boldsymbol{Y} = \boldsymbol{y}$, the log-likelihood function (which for continuous probability distributions is equal to the logarithm of the joint probability distribution) is given by

$$l(\boldsymbol{\beta}, \phi) = \log \mathcal{L}(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left[\frac{y_i \boldsymbol{\beta'} \boldsymbol{x}_i - c(\boldsymbol{\beta'} \boldsymbol{x}_i)}{\phi} + c(y_i, \phi) \right] \tag{2.12}$$

Thus, the maximum likelihood estimator for $\boldsymbol{\beta}$ can be found by taking the first derivative of (2.12) and solving the equations:

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{\phi V(\mu_i(\boldsymbol{\beta}))} = 0, \tag{2.13}$$

where $\mu_i(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{\beta'} \boldsymbol{x}_i)$.

2.2 Regression models for survival data

Before moving into the presentation of hazard models, one needs to introduce the survival function and the hazard rate, the two important concepts for analyzing survival data, or in other words, time-to-event data (Aalen et al., 2008).

Definition 1 (Survival function)**.** *The survival function, denoted $S(t)$, gives the probability for which an event has not yet occurred by time t. Assuming that T is a random variable denoting the survival time, we write (Aalen et al., 2008, p.5),*

$$S(t) = P(T > t). \tag{2.14}$$

Assuming T is continuous, we consider the probability for the event to happen in a small time interval $[t, t + dt]$, given that is has not happened yet at time t. We denote this "probability" by $h(t)dt$. It can be interpreted as an infinitesimal, and we can define it more specifically as

Definition 2 (Hazard rate)**.** *The hazard rate $h(t)$ is defined as (Aalen et al., 2008, p.6),*

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t). \tag{2.15}$$

The hazard rate can be any non-negative function, whereas the survival function is a positive decreasing function that starts at 1.

We define the cumulative hazard as

$$H(t) = \int_0^t h(s)ds. \tag{2.16}$$

Note that the hazard rate can also be written as

$$h(t) = -\frac{S'(t)}{S(t)}, \quad \text{or} \quad h(t) = \frac{f(t)}{S(t)}, \tag{2.17}$$

where $f(t)$ denotes the density of T, and we have the following relation between the survival function and the hazard rate:

$$S(t) = e^{-H(t)}. \tag{2.18}$$

Assume now that we have a counting process $N_i(t)$ recording the number of events that have occurred up to, and including time t for an individual i. Let λ_i be the intensity of the counting process $N_i(t)$. Using martingale theory without dwelling on the details, we have the relation

$$N_i(t) = \int_0^t \lambda_i(s)ds + M_i(t),$$

where $M_i(t)$ is a zero-mean martingale (see Aalen et al., 2008, chapter 2), that is, if $\mathcal{F}_t$ is the history at time t of the counting process, we have

$$\mathrm{E}(M_i(t)|\mathcal{F}_s) = M_i(s) \quad \text{for all} \quad t > s,$$

and

$$\mathrm{E}[M_i(t)] = 0.$$

In particular, we can write:

$$dN_i(t) = \lambda_i(t)dt + dM_i(t), \tag{2.19}$$

where $dN_i(t)$ denotes the number of jumps of the process in the small interval $[t, t + dt]$, and we have the following relation between the intensity and the hazard:

$$\lambda_i(t) = Y_i(t)h(t),$$

where $Y_i(t) = I(T_i \geq t)$.

The above derivations can be extended by the inclusion of covariates for a number n of individuals, so that $h(t)$ depends on covariates. Assume that at time t, for individual i, we have at our disposal a covariate vector $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{in})'$. Assume that the counting process $N_i(t)$ has intensity $\lambda_i(t) = Y_i(t)h(t|\boldsymbol{x_i})$, where $Y_i(t) = I(T_i \geq t)$ takes the value 1 if individual i is at risk just before time t and 0 otherwise (T_i here denotes the event time for individual i). We will consider two different hazard rates for individual i, namely the proportional hazards and the additive hazard.

2.2.1 The proportional hazards model

The semi-parametric proportional hazards model is due to Cox (1972). The proportional hazards model assumes that the hazard rate of an individual i with covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})'$ is defined as

$$h(t|\boldsymbol{x_i}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x_i}), \tag{2.20}$$

with $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ and where $h_0(t)$ is called the baseline hazard, which characterizes the shape of the hazard rate as a function of time. The second part of (2.20), $\exp(\boldsymbol{\beta}'\boldsymbol{x_i})$ is called the *hazard ratio*, and can be interpreted as how much the covariates influence the size of the hazard rate.

Estimation of $\boldsymbol{\beta}$ is often done using the partial likelihood. It is the product over the individuals of all the conditional probabilities that an event for individual i is observed at time t given the past, and given that an event is observed at that time. These conditional probabilities can be written (Aalen et al., 2008, chapter 4)

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda_\cdot(t)} = \frac{Y_i(t)\exp(\boldsymbol{\beta}, \boldsymbol{x}_i(t))}{\sum_{k=1}^{n} Y_k(t)\exp(\boldsymbol{\beta}, \boldsymbol{x}_k(t))}. \tag{2.21}$$

Assuming there are no ties, the event times T_i are such that $T_1 < \cdots < T_n$, and the partial likelihood takes the form

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j|T_j) = \prod_{T_j} \frac{Y_{i_j}(T_j)\exp(\boldsymbol{\beta}, \boldsymbol{x}_{i_j})}{\sum_{l=1}^{n} Y_l(T_j)\exp(\boldsymbol{\beta}, \boldsymbol{x}_l(T_j))}, \tag{2.22}$$

where i_j is the index of the individual experiencing the event at time T_j. Note that since the baseline hazard $h_0(t)$ is not depending on $\boldsymbol{\beta}$, it cancels out in (2.21) and is thus not needed when estimating $\boldsymbol{\beta}$, hence the term *partial likelihood*. Estimates of the parameters are then found by maximizing (2.22), that is, by solving

$$\frac{\partial \log(\mathcal{L}(\boldsymbol{\beta}))}{\partial \beta_j} = 0 \quad \forall \quad j = 1, \ldots, p. \tag{2.23}$$

The estimate $\hat{\boldsymbol{\beta}}$ can be shown to enjoy similar large sample properties as ordinary maximum likelihood estimation (Aalen et al., 2008, chapter 4).

2.2.2 The additive hazard model

The non-parametric additive hazard model is due to Aalen (1980, 1989). Suppose we have n individuals with p covariates each, the additive hazard model assumes that the hazard rate of an individual i with covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})'$ is defined as

$$h(t|\boldsymbol{x_i}) = h_0(t) + \boldsymbol{\beta}(t)'\boldsymbol{x_i}, \tag{2.24}$$

with $\boldsymbol{\beta}(t)' = (\beta_1(t), \beta_2(t), \ldots, \beta_p(t))$ and where $h_0(t)$ is called the *baseline hazard*, which characterizes the shape of the hazard rate as a function of time. The regressions functions $\beta_1(t), \ldots, \beta_p(t)$ can be interpreted as how much the covariates affect the hazard at time t. A special case of (2.24) is the so-called Lin-Ying model (Lin and Ying, 1995), defined as

$$h(t|\boldsymbol{x_i}) = h_0(t) + \boldsymbol{\beta}'\boldsymbol{x_i}, \tag{2.25}$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is now a time invariant parameter, thus (2.25) is a semi-parametric version of (2.24), and special estimation techniques are

needed to estimate $\boldsymbol{\beta}$, but we will not discuss these here.

Note that using (2.24) in (2.19) we can write:

$$dN_i(t) = Y_i(t)\left(h_0(t)dt + \sum_{j=1}^{p} x_{ij}\beta_j(t)dt\right) + dM_i(t). \tag{2.26}$$

which we can rewrite using vector and matrix notation, so that

$$d\boldsymbol{N}(t) = \boldsymbol{X}(t)d\boldsymbol{B}(t) + d\boldsymbol{M}(t), \tag{2.27}$$

where $\boldsymbol{M}(t) = (M_1(t), \ldots, M_n(t))'$, and $\boldsymbol{B}(t) = (H_0(t), B_1(t), \ldots, B_p(t))'$ with $B_j(t) = \int_0^t \beta_j(s)ds$, and where

$$\boldsymbol{X}(t)' = \begin{pmatrix} Y_1(t) & \ldots & Y_n(t) \\ Y_1(t)x_{11} & \ldots & Y_n(t)x_{n1} \\ & \vdots & \\ Y_1(t)x_{1p} & \ldots & Y_n(t)x_{np} \end{pmatrix}. \tag{2.28}$$

Estimation in the non-parametric additive hazard model focuses on estimating the cumulative regression parameters $\boldsymbol{B}(t) = \int_0^t \boldsymbol{\beta}(s)ds$. From expression (2.27) we can find an estimate of $d\boldsymbol{B}(t)$ the same way as in ordinary least squares and obtain

$$d\hat{\boldsymbol{B}}(t) = (\boldsymbol{X}(t)'\boldsymbol{X}(t))^{-1}\boldsymbol{X}(t)'d\boldsymbol{N}(t). \tag{2.29}$$

Since there is very little information contained in the small increments considered here, estimation of $d\boldsymbol{B}(t)$ by (2.29) is in fact done very poorly, but by aggregating the estimated increments over time, i.e. $\hat{\boldsymbol{B}}(t) = \int_0^t d\hat{\boldsymbol{B}}(s)$, we can achieve stability of the estimation of $\boldsymbol{B}(t)$.

For the Lin-Ying model (2.25) another estimating method is used, that allows estimating $\boldsymbol{\beta}$ and the cumulative baseline hazard $H_0(t)$ (Lin and Ying, 1995). The timereg package (Scheike and Martinussen, 2006) in R can handle both these models, in addition to the semi-parametric model (McKeague and Sasieni, 1994), whose hazard for individual i defined as

$$h(t|\boldsymbol{x_i}, \boldsymbol{w_i}) = h_0(t) + \boldsymbol{\beta}(t)'\boldsymbol{x_i} + \boldsymbol{\phi}'\boldsymbol{w_i}, \tag{2.30}$$

with $\boldsymbol{\beta}(t)' = (\beta_1(t), \ldots, \beta_p(t))$, $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_q)$, and $\boldsymbol{w_i} = (w_{i1}, \ldots, w_{iq})'$.

2.3 Collapsibility and non-collapsibility

Because of the problem of dependence between the unknown confounders the response and the exposure, using a model ignoring the confounding will typically result in biased estimates (Didelez et al., 2010). However, as we will

see in this section, some models posses the property of *collapsibility*. In short, it means that if we fit a model for a given response with two independent explanatory covariates, one can fit a model ignoring one of these covariates without affecting the bias in the estimate of the effect of the other covariate on the response. When combined with the two-stage procedure that we will present in Chapter 3, this property can allow us, under certain assumptions, to ignore the confounding and yet still find a consistent estimate of the parameter of interest.

Definition 3 (Collapsibility)**.** *(Clogg et al., 1992; Pearl et al., 1999) Consider a generalized linear model of Y on independent covariates X and L, such that:*

$$g(\mathrm{E}[Y|X,L]) = \beta_0 + \beta_x X + \beta_l L, \tag{2.31}$$

with $g(\cdot)$ an appropriate link function. Consider the following marginal model (omitting L) and assume that for the same $g(\cdot)$,

$$g(\mathrm{E}[Y|X]) = \tilde{\beta}_0 + \tilde{\beta}_x X. \tag{2.32}$$

The model (2.31) *is said to be* collapsible *for β_x over L if $\beta_x = \tilde{\beta}_x$, and is* non-collapsible *if $\beta_x \neq \tilde{\beta}_x$.*

Example 1. *The linear regression model is collapsible:*

Assume that the response Y is normally distributed, and let X and L be two independent random variables so that, with $g(\cdot)$ the identity link,

$$g(\mathrm{E}[Y|X,L]) = \mathrm{E}[Y|X,L] = \beta_0 + \beta_x X + \beta_l L,$$

is the conditional mean model.

Using the law of total expectation, we have,

$$\begin{aligned} \mathrm{E}[Y|X] &= \mathrm{E}[\mathrm{E}[Y|X,L]|X] \\ &= \beta_0 + \beta_x X + \mathrm{E}[\beta_l L|X] \\ &= \tilde{\beta}_0 + \beta_x X, \end{aligned}$$

where $\tilde{\beta}_0 = \beta_0 + \mathrm{E}[\beta_l|X]$, which shows that the linear regression model is collapsible. □

Note here that we have not used the normality assumption for the response in order to show collapsibility, only that the model has an identity link.

2.3.1 Collapsibility of the log-linear model

Let X and L be two independent random variables, and assume that Y is Poisson distributed with mean μ, and that $g(\mu) = \log(\mu)$, so that

$$g(\mathrm{E}[Y|X,L]) = \log(\mathrm{E}[Y|X,L]) = \beta_0 + \beta_x X + \beta_l L,$$

is the predicted model.

Again, using the law of total expectation, we have,

$$\begin{aligned} \mathrm{E}[Y|X] &= \mathrm{E}[\mathrm{E}[Y|X,L]|X]] \\ &= \mathrm{E}[e^{\beta_0+\beta_x X+\beta_l L}|X] \\ &= e^{\beta_0+\beta_x X}\mathrm{E}[e^{\beta_l L}|X]. \end{aligned}$$

Hence,

$$\begin{aligned} g(\mathrm{E}[Y|X]) &= \log(\mathrm{E}[Y|X]) \\ &= \tilde{\beta}_0 + \beta_x X, \end{aligned}$$

where $\tilde{\beta}_0 = \beta_0 + \log(\mathrm{E}[e^{\beta_l L}|X])$, which shows that the log-linear regression model is also collapsible. □

We can note once more here that we have not used the Poisson assumption for the response in order to show collapsibility, only that the model has log-link.

2.3.2 Collapsibility of Aalen's additive hazard model

The following definition of collapsibility is derived from Martinussen and Vansteelandt (2013), where the authors also show both the collapsibility of the additive hazard model, and the non collapsibility of the proportional hazards model. They do not, however, display any calculations, and only the results are presented. Furthermore, the authors also introduce new notations and use some causal inference theory that we will avoid using here. This section and the next will thus present these derivations without introducing any new notation. Note, however, that a similar proof of collapsibility for the additive hazard model appears in Aalen (1989) in the appendix without the author mentioning the term collapsibility.

Definition 4 (Collapsibility)**.** *(Martinussen and Vansteelandt, 2013) Assume that we have two independent random variables X and L and a response Y, and are interested in the measure of association between X and the response*

Y in a full model containing both X and L. This measure of association is said to be collapsible *over the variable L if it can be obtained from the model omitting L (the marginal model).*

We defined the additive hazard model in Section 2.2.2. Now let X and L be two independent random variables, and let

$$h(t|X, L) = h_0(t) + \beta_x(t)X + \beta_l(t)L. \tag{2.33}$$

For this additive hazard model, the measure of association between X and the response $h(t|X, L)$, is typically the hazard difference $\beta_x(t)$, which can be found in the full model by:

$$\beta_x(t) = h(t|X = x + 1, L) - h(t|X = x, L).$$

We now need to find an expression for the marginal model. Recall that by the law of total probability,

$$\begin{aligned} S(t|X) &= P(T > t|X) \\ &= \mathrm{E}_L[P(T > t|X, L)] \\ &= \mathrm{E}_L[S(t|X, L)]. \end{aligned}$$

Note that $\int_0^t (h_0(s) + \beta_x(s)X + \beta_l(s)L)ds = H_0(t) + B_x(t)X + B_l(t)L$. From (2.17) we have

$$h(t|X) = -\frac{S'(t|X)}{S(t|X)},$$

assuming we can interchange derivation and expectations, we have,

$$\begin{aligned}
h(t|X) &= -\frac{\frac{\partial}{\partial t}\,\mathrm{E}_L[S(t|X,L)]}{\mathrm{E}_L[S(t|X,L)]} \\
&= -\frac{\frac{\partial}{\partial t}\,\mathrm{E}_L[e^{-\int_0^t (h_0(s)+\beta_x(s)X+\beta_l(s)L)ds}]}{\mathrm{E}_L[e^{-\int_0^t (h_0(s)+\beta_x(s)X+\beta_l(s)L)ds}]} \\
&= -\frac{\frac{\partial}{\partial t}\; e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_L[e^{-B_l(t)L}]}{e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_L[e^{-B_l(t)L}]} \\
&= \frac{(h_0(t)+\beta_x(t)X)e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_Z[e^{-B_z(t)Z}]}{e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_L[e^{-B_l(t)L}]} \\
&\quad + \frac{e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_L[\beta_l(t)Le^{-B_l(t)L}]}{e^{-H_0(t)-B_x(t)X}\;\mathrm{E}_L[e^{-B_l(t)L}]} \\
&= h_0(t)+\beta_x(t)X+\frac{\mathrm{E}_L[\beta_l(t)Le^{-B_l(t)L}]}{\mathrm{E}_L[e^{-B_l(t)L}]}.
\end{aligned}$$

We thus showed that

$$h(t|X) = \tilde{h}_0(t) + \beta_x(t)X, \tag{2.34}$$

where

$$\tilde{h}_0(t) = h_0(t) + \frac{\mathrm{E}_L[\beta_l(t)Le^{-B_l(t)L}]}{\mathrm{E}_L[e^{-B_l(t)L}]}, \tag{2.35}$$

so that

$$h(t|X=x+1) - h(t|X=x) = \beta_x(t).$$

We have shown that $\beta_x(t)$ can be obtained from the marginal model, thus the hazard model is collapsible. □

2.3.3 Non-collapsibility Cox's proportional hazards model

We defined the proportional hazards model in Section 2.2.1. Now let X and L be two independent random variables, and let

$$h(t|X,L) = h_0(t)\exp(\beta_x X + \beta_l L). \tag{2.36}$$

For this proportional hazards model, the measure of association between X and the response $h(t|X, L)$ typically is the log hazard ratio β_x, which can be found in the full model by

$$\beta_x = \log(\frac{h(t|X = x + 1, L)}{h(t|X = x, L)}).$$

Now, let $\int_0^t h_0(s)ds = H_0(t)$. Assuming we can interchange derivation and expectations, we have

$$\begin{aligned}
h(t|X) &= -\frac{S'(t|X)}{S(t|X)} \\
&= -\frac{\frac{\partial}{\partial t}\ \mathrm{E}_L[S(t|X, L)]}{\mathrm{E}_L[S(t|X, L)]} \\
&= -\frac{\frac{\partial}{\partial t}\ \mathrm{E}_L[e^{-\int_0^t h_0(s)\exp(\beta_x X+\beta_l L)ds}]}{\mathrm{E}_L[e^{-\int_0^t h_0(s)\exp(\beta_x X+\beta_l L)ds}]} \\
&= -\frac{\mathrm{E}_L[\frac{\partial}{\partial t}\ e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]}{\mathrm{E}_L[e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]} \\
&= -\frac{\mathrm{E}_L[-h_0(t)e^{\beta_x X+\beta_l L}e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]}{\mathrm{E}_L[e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]} \\
&= h_0(t)e^{\beta_x X}\frac{\mathrm{E}_L[e^{-\beta_l L\ \exp(\beta_x X+\beta_l L)H_0(t)}]}{\mathrm{E}_L[e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]}.
\end{aligned}$$

We thus showed that

$$h(t|X) = h_0(t)e^{\beta_x X}\eta\{X, \beta_x, \beta_l, h_0(t)\}, \tag{2.37}$$

with

$$\eta\{X, \beta_x, \beta_l, h_0(t)\} = \frac{\mathrm{E}_L[e^{-\beta_l L\ \exp(\beta_x X+\beta_l L)H_0(t)}]}{\mathrm{E}_L[e^{-\exp(\beta_x X+\beta_l L)H_0(t)}]}.$$

Hence

$$\begin{aligned}\log\left(\frac{h(t|X=x+1)}{h(t|X=x)}\right) &= \log\left(\frac{h_0(t)e^{\beta_x(x+1)}\ \eta\{x+1,\beta_x,\beta_l,h_0(t)\}}{h_0(t)e^{\beta_x x}\ \eta\{x,\beta_x,\beta_l,h_0(t)\}}\right)\\ &= \log\left(e^{\beta_x}\frac{\eta\{x+1,\beta_x,\beta_l,h_0(t)\}}{\eta\{x,\beta_x,\beta_l,h_0(t)\}}\right)\\ &= \beta_x+\log\eta\{x+1,\beta_x,\beta_l,h_0(t)\}-\log\eta\{x,\beta_x,\beta_l,h_0(t)\},\end{aligned}$$

and

$$\eta\{X=x+1,\beta_x,\beta_l,h_0(t)\}=\frac{\mathrm{E}_L[e^{-\beta_l L\ \exp(\beta_x X+\beta_l L)H_0(t)\exp(\beta_x)}]}{\mathrm{E}_L[e^{-\exp(\beta_x X+\beta_l L)H_0(t)\exp(\beta_x)}]},$$

which is equal to $\eta\{X=x,\beta_x,\beta_l,h_0(t)\}$ if and only if $e^{\beta_x}=1$, i.e. $\beta_x=0$, which in turn means that $\log\eta\{x+1,\beta_x,\beta_l,h_0(t)\}-\log\eta\{x,\beta_x,\beta_l,h_0(t)\}=0$ if and only if $\beta_x=0$. This shows that the proportional hazards model is non-collapsible. □

2.4 Simulation material

In this section we will derive some important results and tools that will be used in Chapter 4.

2.4.1 Bootstrapping

Bootstrapping has been very well studied in the literature. It provides a robust an easy-to-implement method that can be used to assess the accuracy of a statistical quantity (Friedman et al., 2009, p.249-250). It consists as follows:

Definition 5 (Bootstrapping)**.** *Suppose we have a quantity $S(\mathcal{D})$ computed from data $\mathcal{D}$. From bootstrap sampling we can estimate aspects of the distribution of $S(\mathcal{D})$, in particular its variance. In order to do this, we randomly draw samples from the original dataset, with replacement, each of the same size than the original. We do that B times, and compute the quantity $S(\mathcal{D})$ in each of the new datasets, and obtain B bootstrap estimates. We then can estimate the variance of the distribution of $S(\mathcal{D})$, by :*

$$\hat{Var}[S(\mathcal{D})]=\frac{1}{B-1}\sum_{b=1}^{B}(S(\mathcal{D}^{\star b})-\bar{S^{\star}})^2, \tag{2.38}$$

where $\bar{S^{\star}}=\frac{1}{B}\sum_{b=1}^{B}S(\mathcal{D}^{\star b})$ and $\mathcal{D}^{\star b}$ is the b'th bootstrap sample of the data.

We may also compute bootstrap confidence intervals, which can simply be obtained by sorting the B bootstrap estimates and taking the lower and upper percentile (at a beforehand chosen significance level α). Alternatively, we can use bootstrapping to estimate the variance of the distribution of $S(\mathcal{D})$ as in (2.38) and use this to construct confidence intervals the usual way as if the variance would have been directly estimated.

2.4.2 Inverse transform sampling

In order to generate survival times directly, we can apply the Inverse probability sampling method:

Definition 6 (Inverse Transform Sampling)**.** *Let X be a continuous random variable with cumulative distribution function F_X. Then, the random variable $Y = F_X(X)$ has the uniform distribution on $[0, 1]$. The random variable $F_X^{-1}(Y)$ is called the* inverse probability integral transform *and has the same distribution as X.*

This method can be used with the additive hazard when everything is constant, or when the regression parameters are first degree polynomials, and with the proportional hazards model with continuous baseline hazard (Bender et al., 2005). We will show how to generate survival times directly for these models.

2.4.2.1 Generating survival times for Cox's proportional hazards model

We will only show the result for a Weibull baseline hazard, i.e.

$$h_0(t) = \lambda\nu t^{\nu-1}.$$

For the proportional hazards model, assume we have the following survival function

$$S(t|x) = \exp[-H_0(t)\exp(\beta_x' x)], \tag{2.39}$$

with

$$\begin{aligned} H_0(t) &= \int_0^t h_0(s)ds \\ &= \lambda t^\nu. \end{aligned}$$

Let V be a uniformly distributed random variable on $[0, 1]$ (i.e. $V \sim U(0, 1)$). For a Weibull baseline hazard, one can show (see Appendix A.1 on page 84) that we can generate survival times directly using

$$T = \left(\frac{-\log(V)}{\lambda \exp(\beta_x' x)} \right)^{\frac{1}{\nu}}. \tag{2.40}$$

Assume that we have generated, say $n = 1000$ covariates x_i. It suffices now to draw $n = 1000$ values v_i coming from a uniform distribution $U(0,1)$, and compute

$$t_i = \left(\frac{-\log(v_i)}{\lambda \exp(\beta_x' x_i)} \right)^{\frac{1}{\nu}} \quad \text{for} \quad i = 1, \ldots, n,$$

in order to obtain survival times for n individuals.

2.4.2.2 Generating survival times for Aalen's additive hazard model

For the Aalen additive hazard model, assume we have the following survival function

$$\begin{aligned} S(t|x) &= \exp[-H(t|x)] \\ &= \exp\left[-\int_0^t (h_0(s) + \beta_x(s)x)\, ds \right]. \end{aligned}$$

Assume the following expressions for the baseline hazard and the regression parameter

$$h_0(t) = h_0 + h_1 t,$$

and

$$\beta_x(t) = \chi_0 + \chi_1 t.$$

Let $V \sim U(0,1)$. One can then show (see Appendix A.2 on page 85) that we may generate survival times using

$$\begin{aligned} T &= H^{-1}[-\log(V)|X] \\ &= \frac{-(h_0 + \chi_0 X) + \sqrt{(h_0 + \chi_0 X)^2 - 2(h_1 + \chi_1 X)\log(V)}}{h_1 + \chi_1 X}. \end{aligned}$$

Assume that we have generated, say $n = 1000$ covariates x_i. It suffices now to draw, say $n = 1000$ values v_i coming from an uniform distribution $U(0,1)$, and compute

$$t_i = \frac{-(h_0 + \chi_0 x_i) + \sqrt{(h_0 + \chi_0 x_i)^2 - 2(h_1 + \chi_1 x_i)\log(v_i)}}{h_1 + \chi_1 x_i} \quad \text{for} \quad i = 1, \ldots, n,$$

in order to obtain survival times for n individuals.

An alternative way to generate survival data from Aalen's additive hazard model is to use the R function multiroot. This allows for more general forms for $\beta_x(t)$ and $h_0(t)$, but is computationally inefficient, and was therefore not the preferred method.

3

The two-stage predictor substitution

The method described in this chapter is derived from instrumental variables estimation in a linear regression context, often taught in undergraduate courses in econometrics. A lot of handbooks in econometrics mention IV estimation (Wooldridge, 2002; Russell Davidson, 2009), and few have a more statistical approach to IV estimation. One can go back as far as 1984 in order to find more theoretical literature on the subject (Bowden and Turkington, 1984).

We will begin by giving a formal definition of an instrumental variable, and introduce the two-stage predictor substitution approach for the linear and log-linear models. We then show, similarly to Tchetgen et al. (2015) how it can be applied to Aalen's additive hazard model. For Cox's proportional hazards model, we justify the two-stage procedure in a somewhat different fashion than in the aforementioned article. We also provide arguments and discuss the consistency of such estimators

In order to make this chapter more readable, we will, unless otherwise specified, only consider random variables, that we will denote by capital letters without indexes, such as Z, A or U.

3.1 Definitions and assumptions

Whether we are in the context of linear regression, generalized linear models, or survival analysis, an easy way to understand what is happening is to draw a directed acyclic graph (DAG) of the situation of interest. A directed acyclic graph is a special type of mathematical graph.

Definition 7 (Mathematical graph)**.** *(Pearl et al., 2016) A mathematical graph is a collection of nodes an edges, where the nodes are connected by the edges. A graph can be directed or undirected. It is directed if the edges are*

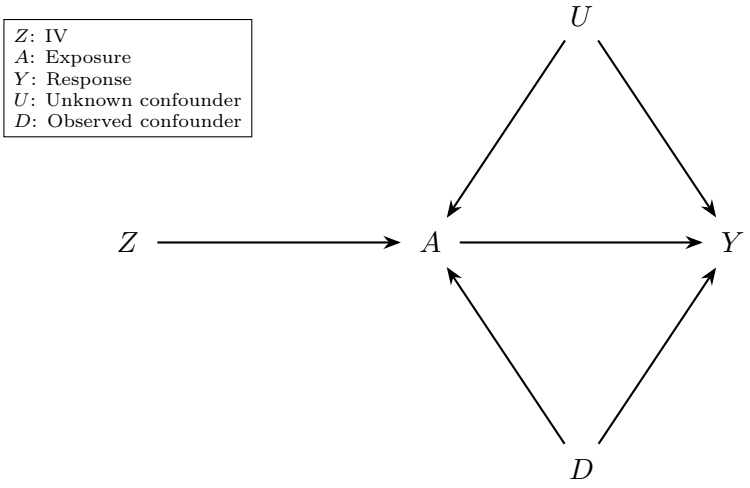


Figure 3.1: A directed, acyclic graph of the setup.

arrows pointing from one node to another, and undirected if the nodes are just connected to each other by a line.

Definition 8 (Directed acyclic graph)**.** *(Pearl et al., 2016) A directed, acyclic graph is a directed graph where no directed path exists from a node to itself.*

Such a directed acyclic graph is displayed in Figure 3.1. In this DAG, the nodes are the variables (Z, A, Y, U, D), and the edges are the arrows connecting the nodes. Directed acyclic graphs like this one are used a lot in epidemiology (Didelez and Sheehan, 2007) and in causal inference (Pearl, 2009), in order to visually represent a situation of interest, in which we would like to obtain information about dependencies between the different variables. A special type of calculus has been developed in order to make causal effects apparent from these type of graphs, and primarily in order to handle graphs much more complicated than the ones considered in this thesis.

The important variable in Figure 3.1 if variable A. The goal of this section is to present a method that can help us see how much this variable is affecting the response Y if we would force A to take a particular value. In causal inference this is often referred to the $do()$ operation. We will not dwell on the details here, but in short, when in the presence of a graph, this operator can help us isolate the important measure of association, and this measure of association is then called a causal effect. The method presented here is essentially a way to simulate the theoretical action of forcing a variable to take a specific value with the help of an instrumental variable.

Some additional clarifications need to be done regarding Figure 3.1. In particular, the arrows should be interpreted as "is affecting". Thus, from the DAG we clearly see that the exposure of interest A is affecting the response Y,

and the instrumental variable Z is affecting Y only through A. In addition, U is affecting both A and Y, but not Z. U is called an *unknown confounder*. There is also an additional observed confounder D, which is affecting both A and Y, but not Z. It is clear from this DAG that real life situations could be represented that way (without Z), as there typically will be some external unobserved variables affecting both covariates and outcome. Since U is unknown, it is often included implicitly in the error term of the explanatory equation, rather than as a variable of its own (Bowden and Turkington, 1984). Furthermore, A is typically a vector of exposure variables, which can be binary or continuous, while Z has to be at least of the size of A, and U is unspecified as to its dimension and type.

Typically, the instrumental variable need to fulfill three main assumptions (Didelez et al., 2010)

(i) The instrumental variable Z has to be unconditionally independent of the unknown confounder U. We write:

$$Z \perp\!\!\!\perp U$$

(ii) The instrumental variable Z needs to be dependent with the endogenous explanatory variable of interest A. We write:

$$Z \not\perp\!\!\!\perp A$$

(iii) The instrumental variable Z, conditionally on U, D and A needs to be independent of Y. We write:

$$Z \perp\!\!\!\perp Y | (A, D, U)$$

Note that D appears only in assumption (iii), because by being observed and not the variable of main interest, it can be included in our model in the usual way as a covariate. In addition to these three assumptions, the models we are considering in this thesis need to be *collapsible*. Collapsibility is a rather important concept in our setting, and it was defined formally in Section 2.3. Before presenting any method, we need to discuss briefly how these assumptions can be verified or tested:

(i) This assumption is unverifiable, but some authors suggest falsifying it (Jackson and Swanson, 2015). Essentially, they compare the potential bias of an ordinary multivariate regression with the potential bias of an instrumental variable estimation. But the test relies on the assumption that observed confounding gives information about the unobservable, and is mainly developed for binary exposure.

(ii) This assumption can be trivially tested by traditional methods such as F-statistics, likelihood ratio tests, or adjusted $R-$squared.

(iii) This assumption is in practice unverifiable, but knowledge about how Z and A are associated can help validate or invalidate it. There exist tests to falsify this assumption, but are mostly used in Mendelian Randomization (Glymour et al., 2012), and as discussed by the authors, these tests can fail under a wide range of circumstances, and epidemiological knowledge about associations can sometimes be a better way to invalidate the assumption.

Now suppose the above assumptions are fulfilled and that after thorough investigation of the data, we have specified a model of interest, that we will call the *original* model. Using the notations of Figure 3.1, the two-stage predictor substitution simply goes as follows:

Stage 1 : We use the instrumental variable Z and any observed confounder D to predict the exposure A via ordinary least squares.

Stage 2 : We use the fitted value of A as a plug-in variable instead of A in the original model.

Suppose our *original* model is:

$$Y = \beta_0 + \beta_a A + \beta_d D + \beta_u(U) + \epsilon, \tag{3.1}$$

where U and A are dependent, $\beta_u(U)$ can be any unrestricted function of the unknown confounders, and ϵ is Gaussian with mean zero and variance σ^2_ϵ. Let Z be a variable fulfilling assumptions (i)-(iii), and such that:

$$A = c_0 + c_z Z + c_d D + \delta, \tag{3.2}$$

where δ is of mean zero and such that $\text{corr}(U, \delta) \neq 0$. Let M be the conditional expectation of (3.2), meaning,

$$M = \text{E}[A|Z, D] = c_0 + c_z Z + c_d D.$$

Thus M is the predicted mean value of A as a function of the instrumental variable Z, and the estimate of M, that we call $\hat{M}$, gives us the first stage of the two-stage procedure. The predicted mean value of Y is then

$$\begin{aligned} \text{E}[Y|A, U, Z, D] &= \beta_0 + \beta_a A + \beta_d D + \beta_u(U) && (3.3) \\ &= \beta_0 + \beta_a M + \beta_a \delta + \beta_d D + \beta_u(U), && (3.4) \end{aligned}$$

by noting that $A = M + \delta$.

Marginalizing with respect to (A, U), using the law of total expectation, we obtain (remembering that M is independent of U and δ):

$$\begin{aligned} \mathrm{E}[Y|Z,D] &= \mathrm{E}[\mathrm{E}[Y|A,U,Z,D]|Z,D] && (3.5)\\ &= \beta_0 + \beta_a M + \beta_d D + \mathrm{E}[\beta_a\delta + \beta_u(U)|Z,D] && (3.6)\\ &= \tilde{\beta}_0 + \beta_a M + \beta_d D, && (3.7) \end{aligned}$$

where $\tilde{\beta}_0 = \beta_0 + \mathrm{E}[\beta_a\delta + \beta_u(U)|Z,D]$.

In (3.7), M is assumed known, but must be estimated in practice by $\hat{M}$, by which we obtain the second stage of the two-stage procedure. The resulting estimate of β_a obtained via ordinary least squares on M and D, is then called an instrumental variable estimate of β_a

Because of the linearity of the model a two-stage predictor substitution is possible to do, and thus we can retrieve an estimate of β_a without having to obtain information about U. Such a feature was referred to as *collapsibility* (see Section 2.3). We can thus exploit the collapsibility of a model to perform a two-stage regression with our instrument in order to retrieve the causal effect of the endogenous variable A. The resulting estimate can be shown to be consistent (Bowden and Turkington, 1984).

3.2 Exploiting collapsibility

The main problem with the presence of unknown confounders is that estimation based on a marginal model ignoring these will typically yield inconsistent and biased estimates because of the dependence between the unknown confounder U, the explanatory variable A and the response Y (Martinussen and Vansteelandt, 2013). The objective is now to use the properties of collapsibility defined in Section 2.3 to perform an IV estimation via a two-stage regression on the additive hazard model in order to retrieve consistent estimates of the causal effect. For the proportional hazards model, because of non-collapsibility, we need to make additional assumptions for the two-stage procedure to work. Furthermore, it is straightforward that the collapsible models studied in the previous chapter will still be collapsible by the inclusion of observed confounders independent of U.

We have already showed as an example how collapsibility is exploited in the two-stage least squares regression in Section 3.1. We will now briefly show how an IV estimator in a additive hazard model can in fact be obtained via a two-stage predictor substitution, and how we solve the non-collapsibility problem for the proportional hazards model. Later on, we will also argue that in fact the resulting IV estimators are expected to be consistent.

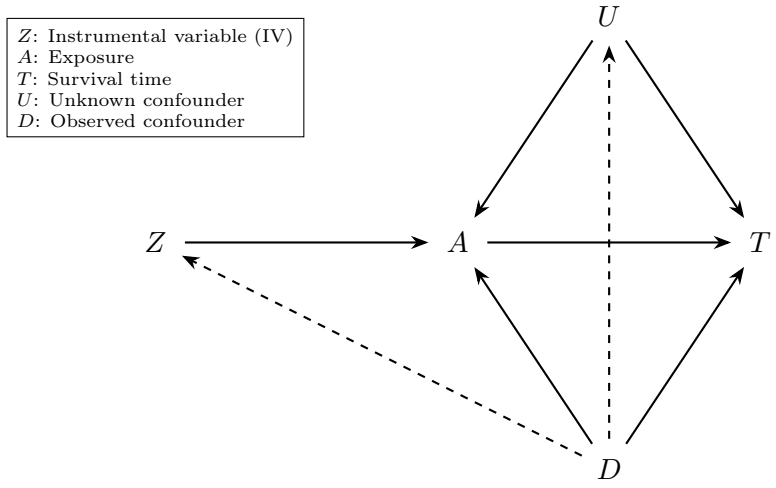


Figure 3.2: A directed, acyclic graph of the setup

So far, we have not explicitly stated the dimensions of the different elements in our IV estimation. Assume n individuals, the instrumental variable can then either be a $n \times 1$ vector or a $n \times p$ matrix (in the case of Mendelian randomization, the instrument is typically a matrix containing p gene expressions), the exposure is a $n \times 1$ vector, and the observed confounder can be either a $n \times 1$ vector or a $n \times q$ matrix. Furthermore, unless otherwise specified, we will assume that the exposure is continuous.
In the time-to-event situations studied here, we do not observe only the event times, but also censoring times. An additional assumption needs to be made for these methods to work, that the censoring time is independent of the event times and the exposure given the instrumental variable. In other words, the censoring needs to be independent of U.

Assume now the DAG displayed in Figure 3.2, it is essentially the same DAG as given in Section 3.1, but now the response is the event time $\tilde{T}$ subject to censoring C, so that we in fact observe $T = \min(\tilde{T}, C)$.

The instrumental variable assumptions described in Section 3.1 are the same here, except for the fact that the response now is called T, so that we write assumption (iii) as

(iii)
$$Z \perp\!\!\!\perp T | (A, D, U)$$

We also add an assumption on the censoring scheme(Li et al., 2014)(Tchetgen et al., 2015):

(iv) The censoring time C needs to be independent of $\tilde{T}$ and A given Z and D. We write:

$$C \perp\!\!\!\perp (\tilde{T}, A) | (Z, D)$$

Note that the DAG presented in Figure 3.2 has also been extended by additional (dotted) lines going from the observed confounder to the unknown confounder and to the instrumental variable. This should not violate any of the assumptions. This is because unlike the unknown confounder, D is observed, and thus can be included in the first and second stage. If D affects Z, because both Z and D are observed, we can include them both in the first stage and the second stage as usual. If D affects U which in turn affects A, it is not a problem either because we are not interested in the effect of D on T, but of A on T.

3.2.1 Additive hazard model – IV estimation

Assume that the survival data is generated by an Aalen additive hazard model:

$$h(t|A, U, Z, D) = h_0(t) + \beta_a(t)A + \beta_d(t)D + \beta_u(U, t), \tag{3.8}$$

where $\beta_u(U, t)$ can be any unrestricted function of the unknown confounders and time t. Note that the right hand side of (3.8) does not depend on Z because of instrumental variable assumption (iii), i.e. that given A, U, and D, the response $\tilde{T}$ is independent of the instrumental variable Z. Suppose we have a linear model for the exposure A such that

$$A = c_0 + c_z Z + c_d D + \delta, \tag{3.9}$$

where δ is a mean zero residual error independent of Z and D, and that given Z and D we have $\text{corr}(\delta, U|Z, D) \neq 0$, representing the confounding induced by U. The conditional mean model of A under (3.9) is thus:

$$\text{E}[A|Z, D] = c_0 + c_z Z + c_d D \tag{3.10}$$

so that $A = \text{E}[A|Z, D] + \delta$. We can thus write model (3.8) as

$$\begin{aligned} h(t|A, U, Z, D) &= h_0(t) + \beta_a(t)(\text{E}[A|Z, D] + \delta) + \beta_d(t)D + \beta_u(U, t) \\ &= h_0(t) + \beta_a(t)\text{E}[A|Z, D] + \beta_a(t)\delta + \beta_d(t)D + \beta_u(U, t) \\ &= h_0(t) + \beta_a(t)\text{E}[A|Z, D] + \beta_d(t)D + \tilde{\beta}_u(U, t), \end{aligned}$$

where $\tilde{\beta}_u(U, t) = \beta_a(t)\delta + \beta_u(U, t)$ is independent of Z and D. Using the fact that the additive hazard model is collapsible, we can marginalize the hazard with respect to (A, U), and we obtain:

$$h(t|Z,D) = \tilde{h}_0(t) + \beta_a(t)\mathrm{E}[A|Z,D] + \beta_d(t)D, \tag{3.11}$$

where $\tilde{h}_0(t)$ is a rather complicated baseline hazard of similar form to the one in (2.35). We can thus use the following two-stage procedure:

Stage 1 : By writing $M = \mathrm{E}[A|Z,D]$, we estimate M by $\hat{M} = \hat{c}_0 + \hat{c}_z Z + \hat{c}_d D$, the fitted value of the OLS regression of A on (Z,D).

Stage 2 : We plug-in $\hat{M}$ instead of $\mathrm{E}[A|Z,D]$ in model (3.37) to obtain Aalen's least squares estimator $\hat{\tilde{\boldsymbol{B}}}(t)$ of $\tilde{\boldsymbol{B}}(t) = (\tilde{H}_0(t), B_a(t), B_d(t))'$.

3.2.2 Proportional hazard model – IV estimation

We showed earlier (see Section 2.3.3) that the proportional hazards model is non-collapsible. Performing a similar two-stage predictor substitution as in the previous section is thus not appropriate. However, it can be shown that in the presence of rare outcomes, the model is approximatively collapsible, and a two-stage procedure can then be performed (Tchetgen et al., 2015).

In order to show this, it is useful to introduce the Laplace transform of a variable W, defined by (Aalen et al., 2008, chapter 6)

$$\mathscr{L}_W(c) = \mathrm{E}(e^{-cW}).$$

Now, let the data be generated by the proportional hazards model

$$h(t|A,U,Z,D) = h_0(t)e^{\beta_a A + \beta_d D + \beta_u(U)}, \tag{3.12}$$

where $\beta_u(U)$ can be any unrestricted function of the unknown confounders, and so that

$$S(t|A,U,Z,D) = e^{-H(t|A,U,Z,D)},$$

is our survival function, with

$$H(t|A,U,Z,D) \;=\; \int_0^t h_0(s)e^{\beta_a A + \beta_d D + \beta_u(U)}ds,$$

Suppose that we have the conditional mean model (3.10) for A, so that $A = \mathrm{E}[A|Z,D] + \delta$. Then,

$$\begin{aligned} H(t|A,U,Z,D) &= \int_0^t h_0(s)e^{\beta_a(\mathrm{E}[A|Z,D]+\delta)+\beta_d D+\beta_u(U)}ds \\ &= e^{\beta_a\delta+\beta_u(U)}\int_0^t h_0(s)e^{\beta_a\mathrm{E}[A|Z,D]+\beta_d D}ds \\ &= L\bar{H}(t|Z,D), \end{aligned}$$

where $L = e^{\beta_a\delta+\beta_u(U)}$ and $\bar{H}(t|Z,D) = \int_0^t h_0(s)e^{\beta_a\mathrm{E}[A|Z,D]+\beta_d D}ds$, so that we can write

$$S(t|L,Z,D) = e^{-L\bar{H}(t|Z,D)}. \tag{3.13}$$

We should note here that L can be considered a frailty, that is to say, it is an unobservable random variable specifying a level of frailty (Aalen et al., 2008, chapter 6). By integrating out L, in other words, by taking the expectation of (3.13) with respect to its distribution, we observe that

$$S(t|Z,D) = \mathrm{E}[e^{-L\bar{H}(t|Z,D)}] = \mathscr{L}_L(\bar{H}(t|Z,D)). \tag{3.14}$$

where the survival function is depending on U through L. Furthermore, note that under rare disease assumption, we have that

$$S(t|A,U,Z,D) = P(T \geq t|A,U,Z,D) \approx 1, \tag{3.15}$$

hence

$$S(t|Z,D) = P(T \geq t|Z,D) = \mathscr{L}_L(\bar{H}(t|Z,D)) \approx 1, \tag{3.16}$$

Now, recall the relation between the population hazard and the individual hazard rate (Aalen et al., 2008, p.235),

$$\mu(t) = -h(t)\frac{\mathscr{L}'_L(H(t))}{\mathscr{L}_L(H(t))},$$

where $\mathscr{L}'_L$ is the first derivative of $\mathscr{L}_L$ with respect to the hazard H, so that under the rare disease assumption,

$$\mu(t|Z,D) = -\bar{h}(t|Z,D)\frac{\mathscr{L}'_L(\bar{H}(t|Z,D))}{\mathscr{L}_L(\bar{H}(t|Z,D))} \approx -\bar{h}(t|Z,D)\mathscr{L}'_L(\bar{H}(t|Z,D)). \tag{3.17}$$

We thus only need to find an expression for $\mathscr{L}'_L(\bar{H}(t|Z,D))$, i.e.

$$\begin{aligned} \mathscr{L}'_L(\bar{H}(t|Z,D)) &= \frac{\partial}{\partial\bar{H}}\mathrm{E}[e^{-L\bar{H}(t|Z,D)}] \\ &= \mathrm{E}[-Le^{-L\bar{H}(t|Z,D)}] \\ &= \mathrm{E}[-e^{\beta_u+\beta_a\delta}S(t|A,U,Z,D)] \\ &\approx \mathrm{E}[-e^{\beta_u+\beta_a\delta}], \end{aligned}$$

hence, under the rare disease assumption,

$$\begin{aligned}\mu(t|Z,D) &\approx \bar{h}(t|Z,D)\mathrm{E}[e^{\beta_u+\beta_a\delta}] \\ &= -h_0(t)e^{\beta_a\mathrm{E}[A|Z,D]+\beta_d D}\mathrm{E}[e^{\beta_u+\beta_a\delta}] \\ &= \tilde{h}_0(t)e^{\beta_a\mathrm{E}[A|Z,D]+\beta_d D}.\end{aligned}$$

We can then write

$$h(t|Z,D) \approx \tilde{h}_0(t)e^{\beta_a\mathrm{E}[A|Z,D]+\beta_d D}. \tag{3.18}$$

Which shows that under the rare disease assumption, the model is approximately collapsible, one can thus use the following two-stage procedure:

Stage 1 : As before, by writing $M = \mathrm{E}[A|Z,D]$, we estimate M by $\hat{M} = \hat{c}_0 + \hat{c}_z Z + \hat{c}_d D$, the fitted value of the OLS regression of A on (Z,D).

Stage 2 : We plug-in $\hat{M}$ instead of $\mathrm{E}[A|Z,D]$ in model (3.18) and can then estimate β_a by standard maximum partial likelihood.

An interesting thing to note here is that in the process of showing non-collapsibility in Section 2.3.3, we also showed is that if the true parameter is in fact zero (i.e., no causal effect), the marginal model will also retrieve the zero-parameter. This could be used in order to test the null hypothesis:

$$H_0 : \quad \beta_a = 0$$

This exact feature has in fact been noted in a recent paper (Burgess, 2015), and will be illustrated in Chapter 4.

3.3 Consistency of the IV estimator

We have shown how collapsibility can be exploited in order to retrieve an estimate of the target parameter of interest. But we yet have to argue that the estimate remains consistent. In Tchetgen et al. (2015), the authors show consistency of the IV estimator resulting from the two-stage predictor substitution with the Aalen additive hazard model analytically, and overlapping authors (Vansteelandt and Didelez, 2015) showed that this can also be done via M-estimation for a wide range of models. They show this in a more general setup, but it still holds under the three core assumptions outlined earlier and with the conditional mean model for the exposure (3.10).

M-estimator are estimators obtained by solving unbiased estimating equations. The linear or generalized linear regression is closely related to M-estimation. In fact, the parameter estimates resulting from a OLS estimation

or maximum likelihood estimation are M-estimators, and therefore enjoy the properties of M-estimators, that is to say, they are consistent and asymptotically normal (Stefanski and Boos, 2002). More generally, M- estimators are the solutions $\hat{\boldsymbol{\theta}}$ of the system of equations:

$$\sum_{i=1}^{n} \psi(Y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{3.19}$$

where $(Y_1 \ldots, Y_n)$ are independent random response variables, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are explanatory variables, $\boldsymbol{\theta}$ is a p-dimensional parameter, and ψ is a known $p \times 1$ function independent of i and n. In addition, if the true parameter value $\boldsymbol{\theta}_0$ is defined uniquely by

$$\mathrm{E}[\psi(Y_1, \boldsymbol{x}_1, \boldsymbol{\theta}_0)] = \mathbf{0}, \tag{3.20}$$

then under certain regularity conditions, there exist a sequence $\hat{\boldsymbol{\theta}}$ of M-estimators satisfying

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0. \tag{3.21}$$

In words, $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ in probability, which we recognize as being the definition of consistency, and under some more regularity conditions one can show that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\left(0, \frac{V(\boldsymbol{\theta}_0)}{n}\right)$$

where $V(\boldsymbol{\theta}_0) = A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) \{A(\boldsymbol{\theta}_0)^{-1}\}^T$ is called the *sandwich matrix* and can be estimated from the data. We refer to Stefanski and Boos (2002) for explicit expressions for $A(\boldsymbol{\theta}_0)$ and $B(\boldsymbol{\theta}_0)$.

3.3.1 Consistency of the two-stage procedure for collapsible generalized linear models

Assume that we have at our disposition n independent random variables $(Y_1, \ldots, Y_n)$ whose joint density function is in the exponential family of distributions, and assume that the Y_i's are related to their mean μ_i via a link function $g(\cdot)$, and that we have a parameter $\boldsymbol{\beta}$ and a vector of covariates $\boldsymbol{x}_i$ such that $g(\mu_i(\boldsymbol{\beta})) = \boldsymbol{x_i}'\boldsymbol{\beta}$. For these generalized linear models, estimating the parameter $\boldsymbol{\beta}$ amounts to solving the score equations (see Section 2.1.3)

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{\phi V(\mu_i(\boldsymbol{\beta}))} = 0, \tag{3.22}$$

where $\mathrm{Var}(Y_i) = \phi \mathrm{V}(\mu_i(\boldsymbol{\beta}))$. And $\mu_i(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x_i}'\boldsymbol{\beta})$ for link function $g(\cdot)$ and covariate vector $\boldsymbol{x}_i$. Then $\hat{\boldsymbol{\beta}}$ is in fact an M-estimator found by solving (3.19) with

$$\psi(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{\phi V(\mu_i(\boldsymbol{\beta}))}. \tag{3.23}$$

Note that the estimates remain the same whether or not the parameter ϕ is known, in that case, we are then solving (3.22) with

$$\psi(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i(\boldsymbol{\beta})}{V(\mu_i(\boldsymbol{\beta}))}. \tag{3.24}$$

of which the resulting estimates can be shown to be consistent (Knight, 1999).

Assume now that we have n observations, and the following collapsible conditional mean model for the response

$$g(\mathrm{E}[Y_i|A_i, U_i, Z_i, D_i]) = \beta_0 + \beta_a A_i + \beta_d' D_i + \beta_u(U_i)$$

for some link function $g(\cdot)$, and where β_a is the parameter of interest, A_i is an exposure variable, β_d is a $k \times 1$ parameter, $D_i = (d_{i1}, \ldots, d_{ik})$ is an additional observed confounder and where $\beta_u(U)$ represent an unspecified function of unknown confounders. We will write $\boldsymbol{\beta} = (\beta_0, \beta_a, \beta_d', \beta_u(.))$. Assume we have parameters $\boldsymbol{c} = (c_0, c_z', c_d')$ and the following conditional mean model (correctly specified) for the exposure A,

$$\mathrm{E}[A_i|Z_i, D_i] = c_0 + c_z' Z_i + c_d' D_i \tag{3.25}$$

where c_z is a $(q-1) \times 1$ parameter, c_d is a $k \times 1$ parameter and $Z_i = (z_{i1}, \ldots, z_{iq-1})$ is an instrumental variable satisfying the three IV assumptions. For the exposure model, at the first stage, we are in fact solving:

$$\sum_{i=1}^{n} \psi_1(A_i, Z_i, \boldsymbol{c}) = \mathbf{0} \tag{3.26}$$

with $\psi_1(A_i, Z_i, \boldsymbol{c}) = (c_0, c_z', c_d')'(A_i - \mathrm{E}[A_i|Z_i, D_i])$.

For the second stage, recall that for collapsible models, we can write, $g(\mathrm{E}[Y_i|Z_i, D_i]) = \tilde{\beta}_0 + \beta_a \mathrm{E}[A_i|Z_i, D_i] + \beta_d' D_i = \tilde{\beta}_0 + \beta_a(c_0 + c_z' Z_i + c_d' D_i) + \beta_d' D_i$. At the second stage, using expression (3.24), we are thus solving:

$$\sum_{i=1}^{n} \psi_2(Y_i, Z_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{c}) = \mathbf{0}, \tag{3.27}$$

with

$$\psi_2(Y_i, Z_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{c}) = \left(\frac{d\mu_i(\tilde{\boldsymbol{\beta}})}{d\tilde{\beta}_0}, \frac{d\mu_i(\tilde{\boldsymbol{\beta}})}{d\beta_a}, \frac{d\mu_i(\tilde{\boldsymbol{\beta}})}{d\beta_d} \right)' \frac{Y_i - \mu_i(\tilde{\boldsymbol{\beta}})}{\mathrm{V}(\mu_i(\tilde{\boldsymbol{\beta}}))}$$

and where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \beta_a, \beta_d)$ and $\mu_i(\tilde{\boldsymbol{\beta}}) = \mathrm{E}[Y_i|Z_i, D_i]$.

In fact, from the general theory of M-estimation (Stefanski and Boos, 2002) it turns out that first estimating (c_0, c_z, c_d) in (3.26) and then plugging in the

estimates $(\hat{c}_0, \hat{c}_z, \hat{c}_d)$ in (3.27), and solving for $(\tilde{\beta}_0, \beta_a, \beta_d)$ yields consistent estimators $(\hat{\tilde{\beta}}_0, \hat{\beta}_a, \hat{\beta}_d)$.

We can see why, by noting that solving the above is equivalent to considering a new ψ function:

$$\psi(Y_i, Z_i, D_i, A_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{c}) = \begin{pmatrix} \psi_1(A_i, Z_i, D_i, \boldsymbol{c}) \\ \psi_2(Y_i, Z_i, D_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{c}) \end{pmatrix}, \tag{3.28}$$

and solving

$$\sum_{i=1}^{n} \psi(Y_i, Z_i, D_i, A_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{c}) = \mathbf{0}. \tag{3.29}$$

The resulting estimates are thus M-estimators and they all enjoy the properties of M-estimators, that is to say, they are asymptotically normal distributed around the true value of the parameters. This justifies the two-stage procedure for the linear and log-linear models discussed earlier, and provides us with an argument for the consistency of the estimate.

3.3.2 Consistency of the two-stage procedure for Aalen's additive hazard model

Earlier authors have shown almost surely convergence of the two-stage Aalen additive hazard estimator for constant parameters (Li et al., 2014). And recent authors have elicited a proof for consistency of the two-stage Aalen additive hazard estimator (Tchetgen et al., 2015) for time dependent parameters, but the proof is complicated and contains many typos. Chapter 4 will therefore be very important in order to illustrate this consistency.

We could however attempt to argue why the estimator resulting from the two-stage procedure with Aalen's additive hazard at the second stage is consistent. Assume the counting process $N(t)$ adapted to the history

$$\mathcal{G}_t = \{N_i(s), Y_i(s), A_i, U_i, Z_i\}$$

of an individual i with exposure covariates A_i unknown confounder U_i and instrumental variable Z_i is related to its intensity by:

$$\begin{aligned} \mathrm{E}[dN_i(t)|\mathcal{G}_{t-}] &= \lambda_i^{\mathcal{G}}(t)dt \\ &= Y_i(t)(\boldsymbol{\beta}_0(t) + \boldsymbol{\beta}_u(t)U_i + \boldsymbol{\beta}_a(t)A_i)dt. \end{aligned}$$

We first need to show that the structure

$$dN(t) = \lambda(t)dt + dM(t),$$

is preserved under the two-stage predictor substitution setup. In order to continue further, we need to introduce the *Innovation Theorem.*

Theorem 1. *(Innovation Theorem (Aalen et al., 2008)) Assume that two histories* $\{\mathcal{F}_t\}$ *and* $\{\mathcal{G}_t\}$ *are nested, that is,* $\mathcal{F}_t \subseteq \mathcal{G}_t$ *for all* t*. The intensity process of* n *is related to two the histories as follows:*

$$\lambda^{\mathcal{F}}(t) = \mathrm{E}[\lambda^{\mathcal{G}}(t)|\mathcal{F}_{t-}].$$

Let $\{\mathcal{F}_s\} = \{N_i(s), Y_i(s), Z_i\}$ so that we have $\mathcal{F}_s \subseteq \mathcal{G}_s$ for all s. Using the Innovation Theorem, we have:

$$\begin{aligned}
\lambda_i^{\mathcal{F}}(t) &= \mathrm{E}[\lambda^{\mathcal{G}}(t)|\mathcal{F}_{t-}] \\
&= \mathrm{E}\left[Y_i(t)\left(\boldsymbol{\beta}_0(t) + \boldsymbol{\beta}_u(t)U_i + \boldsymbol{\beta}_a(t)A_i\right)|\mathcal{F}_{t-}\right] \\
&= Y_i(t)\left(\boldsymbol{\beta}_0(t) + \boldsymbol{\beta}_u(t)\mathrm{E}[U_i|\mathcal{F}_{t-}] + \boldsymbol{\beta}_a(t)\mathrm{E}[A_i|\mathcal{F}_{t-}]\right) \\
&= Y_i(t)\left(\tilde{\boldsymbol{\beta}}_0(t) + \boldsymbol{\beta}_a(t)\mathrm{E}[A_i|\mathcal{F}_{t-}]\right),
\end{aligned}$$

with $\tilde{\boldsymbol{\beta}}_0(t) = \boldsymbol{\beta}_0(t) + \boldsymbol{\beta}_u(t)\mathrm{E}[U_i|\mathcal{F}_{t-}]$. Note that this in fact illustrates the collapsibility property of the additive hazard model discussed in Section 2.3.2, and by giving an alternative definition of collapsibility, we could have shown it formally with the innovation theorem as we essentially did here. We can now write

$$dN_i(t) = dM_i^{\mathcal{F}}(t) + \lambda_i^{\mathcal{F}}(t)dt,$$

where $dM_i^{\mathcal{F}}(t)$ is a zero-mean martingale with respect to the smaller history $\mathcal{F}_t$. Hence, when $\mathrm{E}[A_i|\mathcal{F}_{t-}]$ is known, we can estimate $d\boldsymbol{B}(t)$ the usual way by:

$$d\hat{\boldsymbol{B}}(t) = (X(t)'X(t))^{-1}X(t)'dN(t),$$

where $d\hat{\boldsymbol{B}}(t) = (d\hat{\tilde{B}}_0(t), d\hat{B}_a(t))$ and where

$$X(t)' = \begin{pmatrix} Y_1(t) & \dots & Y_n(t) \\ Y_1(t)\mathrm{E}[A_1|\mathcal{F}_{t-}] & \dots & Y_n(t)\mathrm{E}[A_n|\mathcal{F}_{t-}]. \end{pmatrix} \tag{3.30}$$

We however do not observe $\mathrm{E}[A_i|\mathcal{F}_{t-}]$, but estimate it consistently by $\hat{M}$, so that $\hat{M} \xrightarrow{P} \mathrm{E}[A_i|\mathcal{F}_{t-}]$, so that, with

$$\hat{X}(t)' = \begin{pmatrix} Y_1(t) & \dots & Y_n(t) \\ Y_1(t)\hat{M}_1 & \dots & Y_n(t)\hat{M}_n \end{pmatrix}, \tag{3.31}$$

and we have

$$\left|\frac{1}{n}(\hat{X}(t)'\hat{X}(t) - X(t)'X(t))\right| \xrightarrow{P} 0,$$

by the law of large numbers. What we actually obtain from the two-stage estimation is

$$\tilde{\boldsymbol{B}}(t) \quad = \quad \int_0^t (\hat{X}(t)'\hat{X}(t))^{-1}\hat{X}(t)'dN(t).$$

We have

$$\tilde{\boldsymbol{B}}(t) - \hat{\boldsymbol{B}}(t) = \int_0^t \left[(\hat{X}(t)'\hat{X}(t))^{-1}\hat{X}(t)' - (X(t)'X(t))^{-1}X(t)'\right] dN(t), \quad (3.32)$$

and we can write

$$(\hat{X}(t)'\hat{X}(t))^{-1} = (X(t)'X(t))^{-1} + (\hat{X}(t)'\hat{X}(t))^{-1} - (X(t)'X(t))^{-1},$$

so that

$$\tilde{\boldsymbol{B}}(t)-\hat{\boldsymbol{B}}(t) = \int_0^t \left[(X(t)'X(t))^{-1}(\hat{X}(t)' - X(t)') + \left((\hat{X}(t)'\hat{X}(t))^{-1} - (X(t)'X(t))^{-1}\right)\hat{X}(t)'\right] dN(t),$$

which can be written

$$\begin{aligned}\tilde{\boldsymbol{B}}(t) - \hat{\boldsymbol{B}}(t) = \quad & \int_0^t (X(t)'X(t))^{-1}(\hat{X}(t)' - X(t)')dN(t) \\ & + \int_0^t \left((\hat{X}(t)'\hat{X}(t))^{-1} - (X(t)'X(t))^{-1}\right)\hat{X}(t)'dN(t),\end{aligned}$$

and by an extension of the law of large numbers, it should be possible to show that the second term yields

$$\left|\int_0^t \left((\frac{1}{n}\hat{X}(t)'\hat{X}(t))^{-1} - (\frac{1}{n}X(t)'X(t))^{-1}\right)\frac{1}{n}\hat{X}(t)'dN(t)\right| \xrightarrow{P} 0,$$

provided that for fixed t, the term $\frac{1}{n}\hat{X}(t)'$ converges to a vector of finite constants, and provided that the inverse of $\frac{1}{n}\hat{X}(t)'\hat{X}(t)$ and $\frac{1}{n}X(t)'X(t)$ exist for all t. Now, for fixed t, one can hope that under some additional regularity conditions,

$$|\hat{X}(t)' - X(t)'|dN(t) \xrightarrow{P} 0,$$

so that in the end, we should obtain

$$|\tilde{\boldsymbol{B}}(t) - \hat{\boldsymbol{B}}(t)| \xrightarrow{P} 0,$$

and because

$$|\hat{\boldsymbol{B}}(t) - \boldsymbol{B}(t)| \xrightarrow{P} 0,$$

we will have

$$\tilde{\boldsymbol{B}}(t) \xrightarrow{P} \boldsymbol{B}(t), \tag{3.33}$$

so that $\tilde{\boldsymbol{B}}(t)$ is consistent for $\boldsymbol{B}(t)$. Note that this will also lead expressions for the asymptotic variance of the estimate, something that has been done in Tchetgen et al. (2015), but as noted in the beginning of this section, it is unclear as to how usable such an expression is.

3.3.3 Approximate consistency of the estimate for Cox's proportional hazards model

In general, one can fit a Cox proportional hazards model by a Poisson regression with log-link function (Aalen et al., 2008), however, as discussed earlier, the model is not collapsible, so it would be a mistake trying to fit the marginal model without additional assumptions. Under the rare disease assumption, however, the proportional hazards model is approximately collapsible, which means that the two-stage procedure should be applicable. Authors have argued that under this assumption, the two-stage procedure works equivalently to the two-stage procedure for log-linear models, and that this ables us to say that the resulting estimates are approximately consistent (Tchetgen et al., 2015). They then refer to other papers where the two-stage procedure has been shown to work for log-linear models (Didelez et al., 2010). It is however unclear if this is a fact or not, and it is probably highly depending on the rarity of the outcome. The results from Section 3.3.1 were meant to be used in order to argue for the consistency of the two-stage procedure with Cox's proportional hazards under the additional assumption of rare outcome, but more research needs to be done here.

3.3.4 Estimating the variance

One problem with two-stage procedures, is the variance of the resulting estimator. In fact, two-stage predictor substitution entails estimating the exposure at the first stage, so the second stage estimate will also inherit the variance obtained from the first stage estimation. Therefore, the real variance of the IV estimate will typically be higher than the one obtained from a model using the observed A (rather that the predicted A). Because of the difficulty to obtain a closed expression for an estimate of the variance, a way to remedy this could be bootstrapping (Tchetgen et al., 2015), which we will investigate further in Chapter 4.

We briefly describe how to apply bootstrapping to our situation. For both the Lin-Ying model and the proportional hazards model, we are interested in the variance of the IV estimate of β_a coming from the two-stage predictor

substitution, (i.e. $\hat{\beta}_{a_{IV}}$). We generate for example $B = 500$ bootstrap samples of $\hat{\beta}_{a_{IV}}$, and then estimate the variance of $\hat{\beta}_{a_{IV}}$ by

$$\hat{\text{Var}}[\beta_{a_{IV}}] = \frac{1}{B-1}\sum_{b=1}^{B}(\beta_{a_{IV}}^{\star b} - \bar{\beta_{a_{IV}}^{\star}})^2,$$

where

$$\bar{\beta_{a_{IV}}^{\star}} = \frac{1}{B}\sum_{b=1}^{B}\beta_{a_{IV}}^{\star b},$$

When looking at the additive hazard model with time varying parameters, we estimate $B_a(t)$, the cumulative parameter, and by fixing t to a value of interest, bootstrapping can be done as usual.

A problem we might encounter when in the presence of few events, is that some bootstrap samples might contain no events at all, we thus should keep this in mind when applying the bootstrap to real data.

3.4 Alternative models for the first stage

3.4.1 Non-linear model for the exposure

In Section 3.2, we used a linear model for the exposure. In general, however, one needs to investigate the relationship between the instrumental variable and the exposure. In fact, after further investigation, one might find that for example, the square of the instrument should be included in the model. It is quite straightforward to show that because of collapsibility, one can still retrieve the correct measure of association when the exposure comes for example from the model

$$A = c_0 + c_{z1}Z + c_{z2}Z^2 + c_dD + \delta. \tag{3.34}$$

One could even extend (3.34) to the more general

$$A = g(Z, \boldsymbol{c}) + c_dD + \delta, \tag{3.35}$$

where $g(Z, \boldsymbol{c})$ is a function of the instrumental variable Z and some parameter $\boldsymbol{c}$. Provided that (3.35) holds, and that A can be estimated consistently that way, the two-stage procedure should still work. In fact, if we assume model (3.35), the conditional mean model of A is thus:

$$\text{E}[A|Z, D] = g(Z, \boldsymbol{c}) + c_dD. \tag{3.36}$$

In 3.3.1 we showed that the two-stage procedure for collapsible generalized linear models under a linear first stage should lead consistent estimators of

the parameter of interest. One can see how this is not restricted to the linear first stage, as any consistent estimation of the first stage should lead consistent estimates on the second stage using the same argument. Even under misspecification of the first stage, provided that misspecification still yield consistent first stage estimates, the resulting second stage estimates should be consistent. This feature is often referred to in the literature as *robustness*.

For the additive hazard model (3.8), we can write:

$$h(t|A,U,Z,D) \quad = \quad \beta_0(t) + \beta_a(t)\mathrm{E}[A|Z,D] + \beta_d(t)D + \tilde{\beta}_u(U,t),$$

where $\tilde{\beta}_u(U,t) = \beta_a(t)\delta + \beta_u(U,t)$ is independent of Z and D. Using the fact that the additive hazard model is collapsible, we can marginalize the hazard with respect to (A,U), and we obtain:

$$h(t|Z,D) = \tilde{\beta}_0(t) + \beta_a(t)\mathrm{E}[A|Z,D] + \beta_d(t)D, \tag{3.37}$$

where $\tilde{\beta}_0(t)$ is a rather complicated baseline hazard. We can thus use the following two-stage procedure:

Stage 1 : By writing $M = \mathrm{E}[A|Z,D]$, we estimate M by $\hat{M} = g(Z,\hat{\boldsymbol{c}}) + \hat{c}_d D$, the fitted value of the OLS regression of A on (Z,D).

Stage 2 : We plug-in $\hat{M}$ instead of $\mathrm{E}[A|Z,D]$ in model (3.37) to obtain Aalen's least squares estimator $\hat{\boldsymbol{\beta}}(t)$ of $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_a(t), \beta_d(t))'$.

However, consistency of the instrumental variable estimate and more specifically the speed of convergence should be investigated. In Section 4.4 we attempt to show the behavior of such a model via simulations.

For the proportional hazards model (3.12), and under the rare disease assumption, a similar argument could be used to justify the two-stage procedure, but because it is subject to the additional assumption of rare outcome, one needs to investigate this further, something that will be done in Chapter 4

3.4.2 Misspecification of the exposure model

In the case of collapsible generalized linear models, even under misspecification of the first stage, two-stage regression will yield consistent estimators (Wooldridge, 2002; Vansteelandt and Didelez, 2015). For categorical endogenous covariates, one might be tempted to use a different first stage than OLS, for example using probit-link. This is commonly know in the econometric literature as the *forbidden regression* (Wooldridge, 2010). What it means is that

the subsequent IV estimates obtained at the second stage will be inconsistent, even though we go through an efficient (and consistent) first stage. But because of the robustness of the two-stage procedure, one could still use a linear model on the first stage and it will still lead to consistency of the resulting estimator but result in a loss of efficiency. However, when the second stage is an additive hazard regression model or a proportional hazards model, it has been argued that misspecification of the first stage leads to inconsistency, and a robust alternative was suggested in Tchetgen et al. (2015). We will nonetheless investigate this through simulations in Section 4.5.1.

4

Simulations

As noted in the Introduction, very little simulation studies have been done regarding instrumental variable estimation with survival models. The only study that could be found at the start of this thesis was for Aalen's additive hazard model (Li et al., 2014), for constant regression parameters and arbitrary baseline hazard (i.e. the Lin-Ying model). This model is however included in the larger class of Aalen additive hazard models with time varying parameters. An attempt in illustrating these situations will be made here, especially looking at varying IV strengths, starting with the Lin-Ying model, followed by the more general additive hazard model with time varying parameters. We will then attempt to exhibit what is meant by "rare outcome" for the Cox proportional hazards model, and after a suitable choice for the rarity of the outcome, look at varying IV strengths. We will also look at IV estimation with a polynomial model for the exposure, and see what happens under misspecification of the first stage. Lastly, we will conclude this chapter with some remarks on testing the null hypothesis of no causal effect.
The censoring scheme used in this whole section is satisfying the assumption (iv) described in Section 3.2.1, i.e. independent censoring.

In this chapter and the next we will use the R programming language in which we imported various packages necessary for the simulations and the analysis. In order to fit Lin-Ying models and Aalen's additive hazard model, we used the aalen function in the package timereg (Scheike and Martinussen, 2006). For fitting the proportional hazards model we used the function coxph in the package survival (Therneau, 2015), which also includes other functions used in Chapter 5. In order to perform bootstrapping we used the function boot in the package boot (Davison and Hinkley, 1997). We used the package MASS (Venables and Ripley, 2002) in order to generate most of the datasets. We refer to Appendix C for the source codes used in Section 4.1, 4.2 and 4.3, and to Appendix D for the simulation codes used for these sections. The R codes for Section 4.4 to 4.7 are available upon demand, as they are quite similar to the ones used for Section 4.1 to 4.3.

4.1 Additive hazard with strong IV

Using the method described in the previous sections, survival data was generated, and the two-stage predictor substitution was used to find an estimate of the parameter of interest. An additional observed confounder D was also included in the models. Data was thus simulated from the following additive hazard model

$$h(t|A,U,Z,D) = h_0(t) + \beta_a(t)A + \beta_d(t)D + \beta_u(t)U, \tag{4.1}$$

where $h(t|A,U,Z,D)$ is the hazard function of T evaluated at t, conditional on A, U, D and Z, and the functions $h_0(t)$, $\beta_a(t)$, $\beta_d(t)$ and $\beta_u(t)$ are time dependent parameters.

Throughout this section, the exposure A is supposed continuous, and generated by,

$$A = c_0 + c_z Z + c_d D + \delta, \tag{4.2}$$

where Z was chosen to be a categorical variable taking values $(1,2,3,4)$, with equal probabilities, D is standard normal distributed, and δ is a zero-mean residual error independent of Z and such that $\text{cov}(\delta, U|Z,D) \neq 0$. In Section 4.4.1, we consider the case where A is generated from

$$A = c_0 + c_{z1} Z + c_{z2} Z^2 + c_d D + \delta, \tag{4.3}$$

Both δ and U where simulated simultaneously using a bivariate normal distribution, i.e.

$$\begin{pmatrix} \delta \\ U \end{pmatrix} \sim BVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} b & a \\ a & b \end{pmatrix} \right] \tag{4.4}$$

The parameter c_z in (4.2) can be varied in order to vary the correlation between A and Z. The parameters a and b in (4.4) decide the amount of unknown confounding, and in order to hold this to a fixed amount, these parameters will be found adequately for each value of c_z, by simply generating A for a fixed value c_z first, and then finding a and b so that the correlation between A and U is of a desired amount.

4.1.1 Time constant parameters – Lin-Ying model

In this illustration, we will restrict ourselves to constant parameters and exposure model (4.2), with $(c_0, c_z, c_d) = (3.5, 1.24, 1)$ so that the correlation between A and Z is of about 0.70. The parameters a and b were chosen in order to have a correlation between the unknown confounder U and the exposure A of about 0.40. We have the following Lin-Ying model:

$$h(t|A,U,Z,D) = h_0(t) + \beta_a A + \beta_d D + \beta_u U,$$

which gives the cumulative hazard:

$$\begin{aligned} H(t|A,U,Z,D) &= \int_0^t h(s|A,U,Z,D)\,\mathrm{d}s \\ &= H_0(t) + (\beta_a A + \beta_d D + \beta_u U)t, \end{aligned}$$

from which we obtain the survival function:

$$S(t|A,U,Z,D) = \exp\{-H_0(t) - (\beta_a A + \beta_d D + \beta_u U)t\},$$

or

$$-\log\{S(t|A,U,Z,D)\} = H_0(t) + (\beta_a A + \beta_d D + \beta_u U)t.$$

Using the method described in Section 2.4.2, $n = 1000$ values v_i coming from an Uniform distribution $U(0,1)$ where picked, and the latter equation was solved for t. If we fix $h_0(t) = h_0$ a constant, is straightforward to show that we have,

$$t_i = \frac{-\log(v_i)}{h_0 + \beta_a A_i + \beta_d D_i + \beta_u U_i}, \quad \text{for} \quad i = 1, \ldots n, \tag{4.5}$$

so that t_i can be computed directly, and where (A_i, D_i, U_i) are the values of (A, D, U) for individual i. A uniform censoring scheme was then used, to obtain about 25% of events.

A common problem that one may encounter with the additive hazard model is the risk of generating negative hazards. Choosing c_0 and $h_0(t)$ sufficiently big takes care of this, we therefore choose $(h_0(t), \beta_a, \beta_d, \beta_u) = (4, 0.5, 0.5, 0.5)$. The R-code used for this section can be found in Appendix D.1.

We are interested in estimating β_a by using only A, D and Z. In order to do this, a two-stage procedure described in Section 3.1 was used, first regressing D and Z on A to obtain a fitted value of A that we can call $\hat{M}$, and then doing an Aalen regression of $\hat{M}$ and D on T. In particular, we fitted the following three different Lin-Ying models,

(1) $h(t|A,U,Z,D) = h_0(t) + \beta_a A + \beta_d D + \beta_u U$

(2) $h(t|A,Z,D) = h_0(t) + \beta_a A + \beta_d D$

(3) $h(t|Z,D) = \tilde{h}_0(t) + \beta_a \hat{M} + \beta_d D$

We generated data 5000 times, and the three different models were fitted each time. Table 4.1 displays the resulting estimates together with their empirical variance and mean squared error (MSE). Note that model (1) cannot be fitted in practice, as we do not have access to U in reality.

Table 4.1: Result of 5000 estimations for additive hazard model with constant parameters (the Lin-Ying model) and with corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.40. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 1000$

	model (1)	model (2)	model (3)
estimate	0.499	0.636	0.498
empirical variance	0.091	0.070	0.109
MSE	0.091	0.088	0.109

Both the true model (model (1)) and the IV model (model (3)) seem to estimate the true β_a correctly, with very small bias. The naive model (model (2)) however seems to overestimate the true β_a. As for the empirical variance of the estimates, it comes as no surprise that the variance obtained from model (1) is higher than the variance of model (2), as model (1) contains one more covariate (the unknown confounder) as model (2). The variance obtained from model (3) is higher than the other two, because of the additional variance gained from estimating the first stage. An interesting thing to note here is that even-though model (2) overestimates the true parameter, the simulations report the smallest mean squared error for this model. Selecting a model based on MSE would thus result in choosing the wrong model here.

We will now estimate the variance of the estimated parameter (i.e. of $\hat{\beta}_{a_{IV}}$) using the bootstrap method introduced in Section 2.4.1, and described in more details in Section 3.3.4. A new dataset was generated so that corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.40, and we then repeated the bootstrap procedure 5000 times with $B = 500$ bootstrap samples each time, for both model (1), model (2) and model (3). We then computed the average of the estimated bootstrap variances for the three models (bootstrap variance), together with the average of the variances reported by the aalen function (aalen variance), and the empirical variance. The results are reported in Table 4.2.

Table 4.2: Result of 500 bootstrap variances for 5000 IV estimations for additive hazard model with constant parameters and with corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.40. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 1000$.

	model (1)	model (2)	model (3)
bootstrap variance	0.0941	0.0735	0.1116
aalen variance	0.0927	0.0726	0.1105
empirical variance	0.0928	0.0721	0.1112

The bootstrap variance for model (3) seems to slightly overestimate the

empirical variance. Surprisingly, the variance computed from the aalen (from the timereg package) function seems to report the correct variance already, however slightly underestimating it. We also presented the results for model (1) and model (2), and as expected, the variance reported by the aalen function this time agrees with the empirical variance. The bootstrap variances for model (1) and model (2) are slightly off, due to perhaps the low amount of bootstrap samples. When comparing the naive model (model (2)) and the IV model (model (3)), an important thing to note is that in addition to have biased estimates, model (2) has very low variance compared to model (3). Therefore, not taking in account unknown confounding could result in constructing confidence intervals that are too narrow and maybe not even including the true parameter value. In practice, one should of course both report the aalen variance and compute a bootstrap estimate of the variance in order to compare the two. Typically, one should expect the bootstrap variance to be closer to the truth, and higher than the aalen variance. Hence, computing confidence intervals based upon the bootstrap variance will give us more conservative intervals, which can be appealing when the amount of confounding is thought to be high. However, when we performed the same bootstrap method with $n = 100$ (see Table 4.3), we found that the bootstrap variance overestimates the variance of all three models, and even more surprisingly, the aalen function seems to estimate the variance correctly.

Table 4.3: Result of 500 bootstrap variances for 5000 IV estimations for additive hazard model with constant parameters and with corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.40. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 100$.

	model (1)	model (2)	model (3)
bootstrap variance	1.2151	0.9127	1.4024
aalen variance	1.0503	0.8069	1.2259
empirical variance	1.0434	0.8119	1.2355

In this section we have generated data with constant parameters. Therefore, when fitting the different models, we "told" the aalen function that it should estimate constant parameters, with the exception of the baseline hazard $h_0(t)$. In practice, however, we do not know whether the parameters are constant or not, it is therefore useful to fit a model that estimates time varying parameters, then plot the cumulative parameters as a function of time, and see whether the estimated cumulative functions are following a straight line or not. For illustration, Figure 4.1 displays the estimated cumulative baseline hazard $\hat{H}_0(t)$ (left) and the estimated cumulative parameter $\hat{B}_a(t)$ (right) as a function of time resulting from fitting a model estimating time varying cumulative parameters using one of the generated datasets.

The two lines plotted in Figure 4.1 show that the estimated cumulative baseline hazard and cumulative parameter are not time dependent as the two curves are almost straight lines. However, if one takes a closer look at (2.35) in

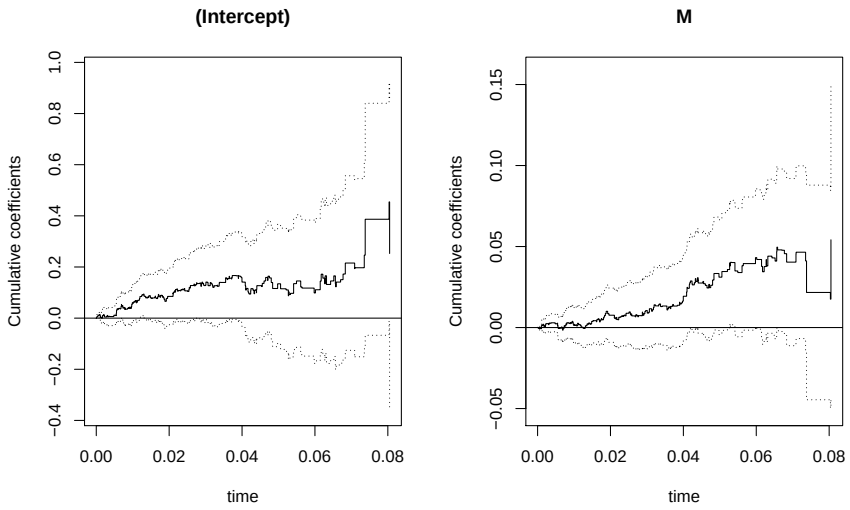


Figure 4.1: Estimated cumulative baseline hazard $\hat{\tilde{H}}_0(t)$ (left) and estimated cumulative $\hat{B}_a(t)$ (right) as a function of time. The data was generated using a constant parameter ($\beta_a = 0.5$), and such that corr(A,Z)$\approx$ 0.70 and corr(A,U)$\approx$ 0.40. The number of observations in each simulation was $n = 1000$.

Section 2.3.2,one can see that it is not guarantied that $\tilde{h}_0(t)$ is time constant even though $h_0(t)$ and the other parameters are. To see why, recall that (using the notations in 2.3.2),

$$\tilde{h}_0(t) = h_0(t) + \frac{\mathrm{E}_L[\beta_l(t) L e^{-B_l(t)L}]}{\mathrm{E}_L[e^{-B_l(t)L}]},$$

where this time L contains both δ from the exposure model and U from the response. With $h_0(t) = h_0$, and if all the parameters are constant, $\beta_l(t)$ will also be constant (i.e. $\beta_l(t) = \beta_l$), and we can then take it out of the expectation so that

$$\tilde{h}_0(t) = h_0 + \beta_l \frac{\mathrm{E}_L[L e^{-B_l(t)L}]}{\mathrm{E}_L[e^{-B_l(t)L}]}, \tag{4.6}$$

we see that the second term in (4.6) is time dependent, as the ratio of the two expectations contains the cumulative regression parameter $B_l(t)$ which is time dependent. However, from the left hand plot in Figure 4.1, and at least for the parameters and variables used in these simulations, this effect seems to be negligible as the cumulative baseline hazard follows an almost straight line.

4.1.2 Lin-Ying model with varying IV strengths

In this section we will try to illustrate the behavior of the estimated parameter when the strength of the instrumental variable is varied. As noted in the

beginning of Section 4.1, the strength of the IV can be varied by varying the correlation between Z an the exposure variable A. Ideally we would like to see how well the two-stage procedure performs when the strength of the IV varies while all the other dependence structures remain fixed. That is to say, in order to simulate a realistic enough situation, while the correlation between Z and A varies, the correlation between A and U, between A and T, as well as the number of events must be held fixed. If not, the number of events might vary quite erratically, and there is no guaranty that the illustrated consistency is the real thing. This was done relatively well by the R-code given in Appendix D.1. We generated $n = 1000$ observations where Z and D were as in 4.1, and where corr(A,U)$\approx$ 0.40. The true parameter values were $(h_0(t), \beta_a, \beta_d, \beta_u) = (4, 0.5, 0.5, 0.5)$ as in 4.1.1, and the number of events was fixed to approximately 25%.

Once all these restrictions had been taken care of, we generated data 5000 times for each IV strength, estimation of β_a was performed with models (1)-(3) and the resulting plots are presented in Figure 4.2.

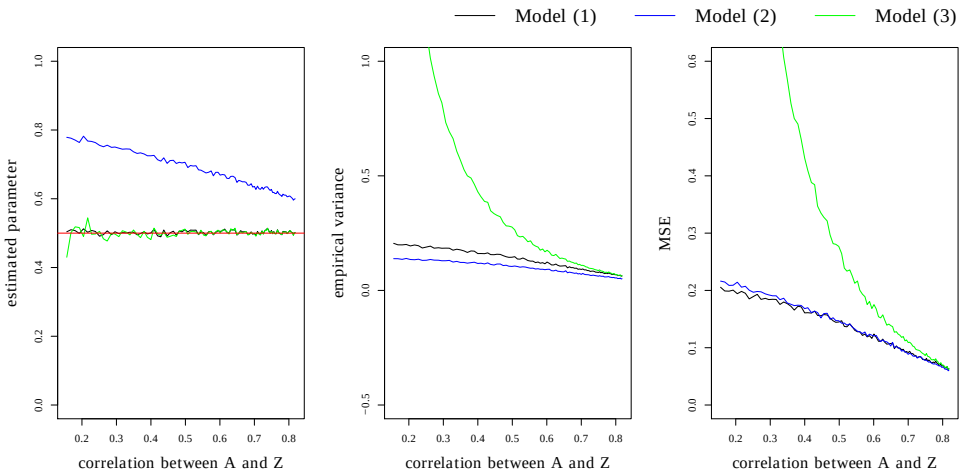


Figure 4.2: Estimated parameters (left), empirical variance (middle) and mean squared error (right) as a function of the correlation between the IV Z and the covariate A. We fixed corr(A,U)$\approx$ 0.40. We display the results from true model (model (1)), from the naive model (model (2)), and from the IV model (model (3)). The true parameter value is $\beta_a = 0.50$ and is represented by a red line in the left hand figure. The number of observations in each simulation was $n = 1000$

We can see from Figure 4.2 that the IV model (model (3)) seems to perform uniformly better than the naive model (model (2)) throughout the whole range of correlations. We can see the appeal of such a method quite clearly. While the naive model (model (2)) tends to overestimate the effect of A, the two-stage procedure (model (3)) seems to estimate the effect of A consistently. However, as the instrument becomes weaker, the variance of the IV estimate becomes exponentially higher compared to variance of the other two estimates.

It can be interesting to calculate averaged 95% confidence interval around the estimates, and the resulting plot is displayed in Figure 4.3.

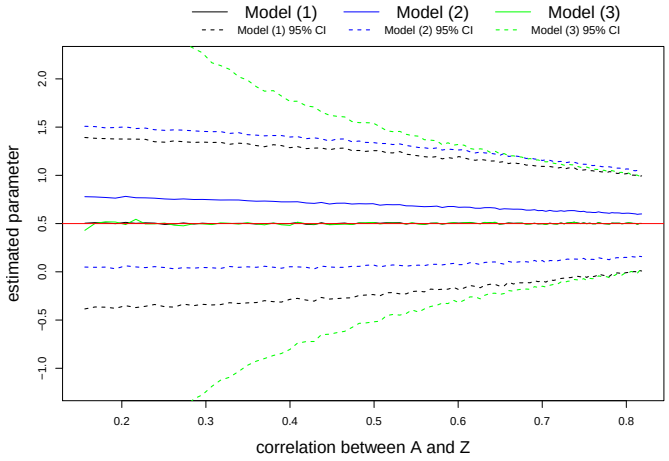


Figure 4.3: Estimated parameters and their confidence intervals as a function of the correlation between the IV Z and the covariate A, for the true model (model (1)), the naive model (model (2)), and for the IV model (model (3)). Thick lines represent the estimated parameter while the dotted lines represent the bounds of the 95 % confidence intervals. We fixed corr(A,U)$\approx$ 0.20. The true parameter value is $\beta_a = 0.50$ and is represented by a red line. The number of observations in each simulation was $n = 1000$

The confidence bound around the IV estimate (green) can be seen as much more conservative, and it converges to the true confidence interval (the one of the true model) as the instrument becomes stronger. The naive confidence interval (blue) can bee seen to be biased upwards, and one can imagine that there exist situations where the confidence interval around the naive estimate does not include the true parameter.

For comparison, two other simulations have been performed, one with corr(A,U)$\approx$ 0.20, and one with corr(A,U)$\approx$ 0.60. The results are presented side by side in Figure 4.4.

It is apparent from Figure 4.4 that more confounding translates into more bias in the naive estimate. Indeed, the curve representing the estimated parameter for the naive model (in blue) misses the true value by much more in when the amount of confounding is high (corr(A,U)$\approx$ 0.60) than when it is low (corr(A,U)$\approx$ 0.20). The variance of all the IV estimate seems to have the same behavior regardless of the amount of confounding, and the peculiar feature of the MSE for the naive model seen earlier, does not seem to repeat itself when the amount of confounding is high.

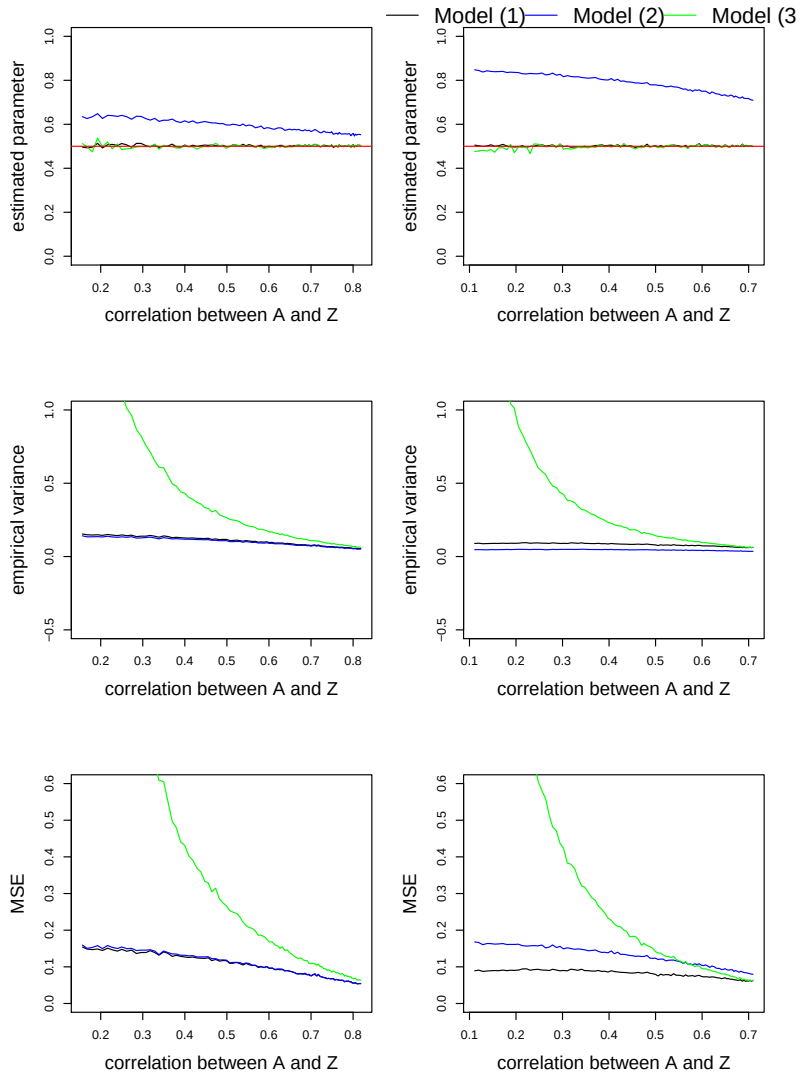


Figure 4.4: Comparison of the estimated parameters (top), empirical variance (middle) and mean squared error (bottom) as a function of the correlation between the IV Z and the covariate A for corr(A,U)$\approx$ 0.20 (left), and corr(A,U)$\approx$ 0.60 (right) . We display the results from true model (model (1)), the naive model (model (2)), and from the IV model (model (3)). The true parameter value is $\beta_a = 0.50$ and is represented by a red line in the top plots. The number of observations in each simulation was $n = 1000$

4.1.3 Time varying parameters

In this section, we will look at the natural extension to the Lin-Ying model (4.1) where $h_0(t)$, $\beta_a(t)$, $\beta_d(t)$ and $\beta_u(t)$ are continuous and time dependent functions. For illustration, they were chosen to be of the same form as in Section 2.4.2.2, i.e. of the form:

$$\begin{cases} h_0(t) = h_0 + h_1 t, \\ \beta_a(t) = a_0 + a_1 t, \\ \beta_d(t) = d_0 + d_1 t, \\ \beta_u(t) = u_0 + u_1 t. \end{cases}$$

with $(h_0, a_0, d_0, u_0) = (1, 1.5, 0.5, 0.6)$ and $(h_1, a_1, d_1, u_1) = (10, 1.5, 0.5, 0.6)$, where h_1 was chosen to be big in order to avoid generating negative hazards.

Because the parameters are now continuous functions, we need to specify values t of T, for which we are interested in seeing the effect of $\beta_a(t)$ on the hazard difference. For the purpose of illustration, we settled with the value at time $t = 0.1$ and $t = 0.2$. The method used to simulate the event times directly is described in Section 2.4.2.2, i.e., event times were generated by first picking $n = 1000$ values v_i from the uniform distribution on $[0, 1]$, and then generating:

$$t_i = \frac{-\tilde{a}_i + \sqrt{\tilde{a}_i^2 - 2\tilde{b}_i \log(v_i)}}{\tilde{b}_i}, \quad \text{for} \quad i = 1, \ldots, n, \tag{4.7}$$

where

$$\begin{aligned} \tilde{a}_i &= h_0 + a_0 A_i + d_0 D_i + u_0 U_i, \\ \tilde{b}_i &= h_1 + a_1 A_i + d_1 D_i + u_1 U_i. \end{aligned}$$

and where (A_i, D_i, U_i) are the values of (A, D, U) for individual i. Again, A was generated as in (4.2), with $c_0 = 3.5$ and $c_z = 1.24$ giving an average correlation between A and Z of about 0.70. Furthermore, δ and U where simulated as in (4.4), so the correlation between A and U is about 0.40. The R-code used to perform the simulations can be found in Appendix D.2. The true value of $B_a(t)$ at $t = 0.1$ and $t = 0.2$ are

$$\begin{aligned} B_a(0.1) &= 0.1575, \\ B_a(0.2) &= 0.33, \end{aligned}$$

and the results of the simulations are shown in Table 4.4a for the true model (which cannot be fitted in practice as we do not have access to U in reality), i.e. the model

(4) $h(t|A, U, Z, D) = h_0(t) + \beta_a(t)A + \beta_d(t)D + \beta_u(t)U,$

in Table 4.4b for the naive model, i.e. the model

(5) $h(t|A, Z, D) = h_0(t) + \beta_a(t)A + \beta_d(t)D,$

and in Table 4.4c for the IV estimation obtained through the two-stage predictor substitution method, i.e., the model

(6) $h(t|Z, D) = \tilde{h}_0(t) + \beta_a(t)\hat{M} + \beta_d(t)D.$

Table 4.4: Results of 5000 simulations and fitting of the true model (model (4)), the naive model (model (5)), and the IV model (model (6)) with corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.40. The true parameter values are $B_a(0.1) = 0.1575$, and $B_a(0.2) = 0.33$. The number of observations in each simulation was $n = 1000$.

(a) True model - model (4)

	t=0.1	t=0.2
estimate of $B_a(t)$	0.1574	0.3302
empirical variance	0.00127	0.01518
bias	-0.0001	0.0002
MSE	0.00127	0.01517

(b) Naive model - model (5)

	t=0.1	t=0.2
estimate of $B_a(t)$	0.1749	0.3671
empirical variance	0.00101	0.01182
bias	0.01699	0.03707
MSE	0.0013	0.01319

(c) IV model - model (6)

	t=0.1	t=0.2
estimate of $B_a(t)$	0.1573	0.3308
empirical variance	0.00152	0.01812
bias	-0.0002	0.0008
MSE	0.00152	0.01811

Once more, the IV estimation seems to succeeds in estimating the cumulative parameter $B_a(0.1)$ and $B_a(0.2)$. Even though the naive estimates have very small variance, they overshoot the true values of the parameter. While the IV estimate has no bias (asymptotically), it seems to have a higher spread than the estimate resulting from a model also including U.

In an applied situation, one might wish to estimate the variance of the estimator via bootstrapping. However, here we need to do so for each event time point. Let T_j be an event time at which one wishes to estimate the variance of the estimator at that point. If n is small, and if there are few individuals experiencing an event, there is no guaranty that the individual experiencing an event at time T_j will be picked in the bootstrap sample. We thus choose the nearest event time from T_j that is smaller than T_j. This allows for estimation of the variance in applied situations. But because of the high computational time we do not present bootstrapping results for these models.

In this section, when fitting the different models, we "told" the aalen function that it should estimate time varying parameters. As said in the previous section, we do not know in practice whether the parameters are constant or not, it is therefore useful to fit a model that estimates time varying parameters, then plot the cumulative parameters as a function of time, and see whether the estimated cumulative functions are following a straight line or not. For illustration, Figure 4.5 displays the estimated cumulative baseline hazard $\hat{H}_0(t)$ (left) and the estimated cumulative parameter $\hat{B}_a(t)$ (right) as a function of time resulting from fitting a model estimating time varying cumulative parameters using one of the generated datasets.

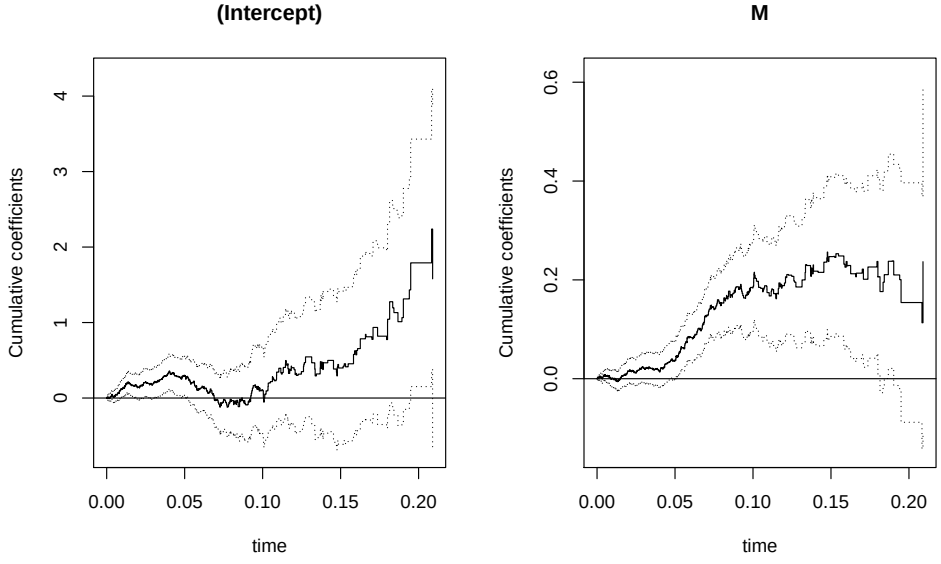


Figure 4.5: Estimated cumulative baseline hazard $\hat{\hat{H}}_0(t)$ (left) and estimated cumulative $\hat{B}_a(t)$ (right) as a function of time. The data was generated using the time varying parameters, and such that corr(A,Z)$\approx$ 0.70 and corr(A,U)$\approx$ 0.40. The number of observations in each simulation was $n = 1000$.

The two lines plotted in Figure 4.5 are seemingly not following a straight

line, thus fitting models with time varying parameters seems appropriate here. However, the non-linearity of the two lines is perhaps due to chance, as the effect on the hazard, even though time dependent, is relatively small in our simulations.

4.2 Proportional hazards model – varying the outcome

This time we generated data coming from the proportional hazards model (2.20), i.e.

$$h(t|A,U,Z,D) = h_0(t)e^{\beta_a A+\beta_d D+\beta_u U}. \tag{4.8}$$

where $h_0(t) = \lambda t^\nu$ with $\lambda = 0.1$ and $\nu = 1$, and with $(\beta_a, \beta_d, \beta_u) = (0.5, 0.5, 0.5)$. A noted in Section 3.2.2, due to the non-collapsibility of the proportional hazards, we have to restrict our simulations to situations where the outcome is rare. Using the inverse transform sampling method described in Section 2.4.2, event times were generated by first picking $n = 10000$ values v_i from the uniform distribution on $[0, 1]$, and then generating

$$t_i = \left(\frac{-\log(v_i)}{\lambda \exp(\beta_a A_i + \beta_d D_i + \beta_u U_i)}\right)^{\frac{1}{\nu}}, \quad \text{for} \quad i = 1, \ldots, n, \tag{4.9}$$

where (A_i, D_i, U_i) are the values of (A, D, U) for individual i. The R-code used for this section can be found in Appendix D.3. The instrument Z and the observed confounder D were generated as in Section 4.1, and the exposure A was generated as in (4.2), with $(c_0, c_z, c_d) = (0.01, 1.122, 0.8)$ so that the correlation between A and Z was controlled at around 0.70, the correlation between A and U at around 0.20, and the total number of events was varied from 0.5% to 5% . A two-stage procedure described in Section 3.1 was then performed, first regressing D and Z on A to obtain a fitted value of A that we can call $\hat{M}$, and then doing an Cox regression of $\hat{M}$ and D on T. In particular, we fitted the following three different models,

(7) $h(t|A,U,Z,D) = h_0(t)e^{\beta_a A+\beta_d D+\beta_u U}$

(8) $h(t|A,Z,D) = h_0(t)e^{\beta_a A+\beta_d D}$

(9) $h(t|Z,D) = \tilde{h}_0(t)e^{\beta_0+\beta_a \hat{M}+\beta_d D}$

Note here again that model (7) cannot be fitted in practice as we do not have acces to U in reality. Data was generated 5000 times, and the three different models fitted each time. Table 4.5 displays the resulting estimates of the true model (model (7)), the naive model (model (8)) and the IV model (model (9)), for different percentages of outcome.

Table 4.5: Results of 5000 estimations for the true model (model (7)), the naive model (model (8)) and the IV model (model (9)) with corr(A,Z)= 0.70, and corr(A,U)= 0.20. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$.

rarity of event (in %)	model (7)	model (8)	model (9)
0.6089	0.5043	0.5841	0.5035
1.0495	0.5029	0.5806	0.4990
1.4770	0.5011	0.5773	0.4936
2.0279	0.5018	0.5763	0.4908
2.5661	0.5005	0.5735	0.4870
3.0826	0.5014	0.5728	0.4860
3.5829	0.5008	0.5712	0.4824
4.0766	0.5004	0.5693	0.4798
4.5557	0.5000	0.5676	0.4778
5.0243	0.5004	0.5673	0.4772

From the results presented in Table 4.5 it seems like the two-stage estimation for anything under 2% of outcome estimates β_a consistently, and anything over 2% underestimates the true parameter. We first did these simulations with $n = 1000$, but the results were unstable for low percentages of events (see Table B.1 in Appendix B). We thus presented the results when $n = 10000$. It seems that for further simulations, controlling the number of events to 2% or less is a suitable enough choice. However, depending on both the correlation between A and U and the correlation between A and Z, one might obtain a different optimal number of events. Figure 4.6 displays six different plots in each of which the strength of the IV is varied, while the correlation between A and U is held fixed to approximately 0.20 (see Figure B.1 in Appendix B for the results when $n = 1000$).

From Figure 4.6 one sees that the assumption for rare outcome does not seem to be depending on the strength of the IV. The curves in the first three plots (on top in Figure 4.6) have a slightly more erratic behavior, due to the combination between having a weak instrument and a relatively small amount of events. In fact, Table 4.6 reports the empirical variance for the three estimates obtained from fitting models (7), (8) and (9) for about 1% of events and varying IV strengths.

The results shown in Table 4.6 confirm what was said about the first three plots in Figure 4.6 and what was seen for the additive hazard model, that as the instrument becomes weaker, the resulting IV estimate is subject to much more variation.

In Figure 4.7 we display similar plots to Figure 4.6, but this time the correlation between the IV Z and the exposure A was held fixed to 0.70, while the correlation between A and U was controlled to 0.10, 0.20, 0.30 and 0.40

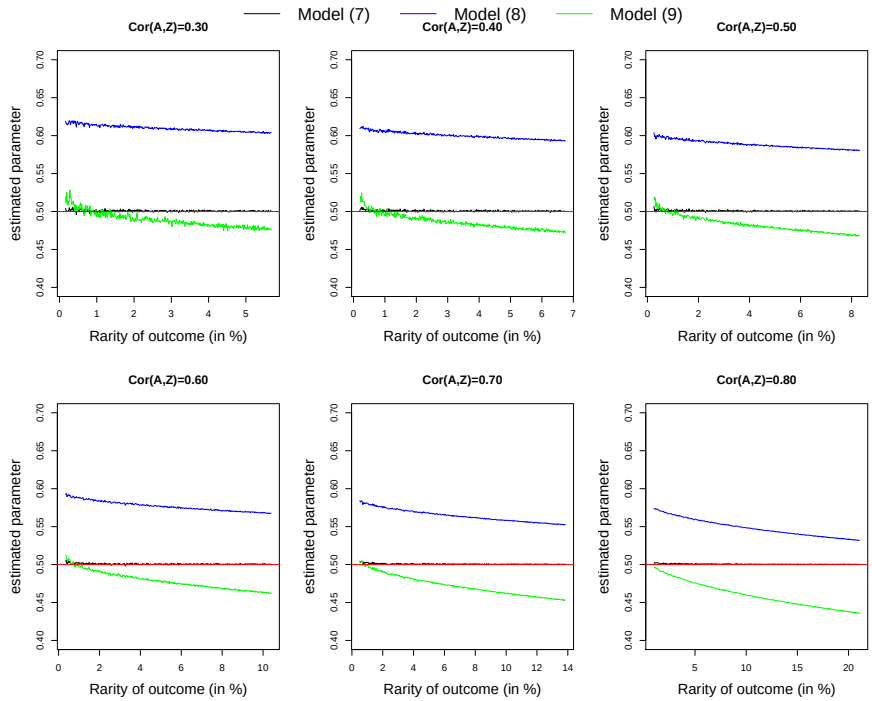


Figure 4.6: Results of 5000 estimations from the naive model (model (7)), the IV model (model (8)) and the true model (model (9)), for six different IV strengths, with corr(A,U)$\approx$ 0.20, as a function of the percentage of events. The number of observations in each simulation was $n = 10000$

Table 4.6: Estimated parameter and empirical variance, as a results of 5000 estimations for the true model (model (7)), the naive model (model (8)) and the IV model (model (9)) with varying IV strengths and corr(A,U)= 0.20, and about 1% of events. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$.

	Estimated parameter			*Empirical variance*		
corr(A,Z)	model (7)	model (8)	model (9)	model (7)	model (8)	model (9)
0.30	0.5014	0.6146	0.5011	0.0106	0.0100	0.0738
0.40	0.5009	0.6067	0.4946	0.0091	0.0086	0.0387
0.50	0.4996	0.5947	0.4999	0.0082	0.0080	0.0225
0.60	0.5024	0.5884	0.4992	0.0067	0.0065	0.0139
0.70	0.5019	0.5803	0.4989	0.0054	0.0055	0.0092
0.80	0.5014	0.5732	0.4977	0.0038	0.0041	0.0054

It is apparent from Figure 4.7 that the rare disease assumption is only very slightly depending on how much confounding there is, but it seems that above 2% there are too many events to consider them rare. What we also see in this figure is how the effectiveness of the IV estimation compared to a naive model depends on the amount of confounding. In particular the first and last plot in Figure 4.7 show us how much the naive model (model (8))

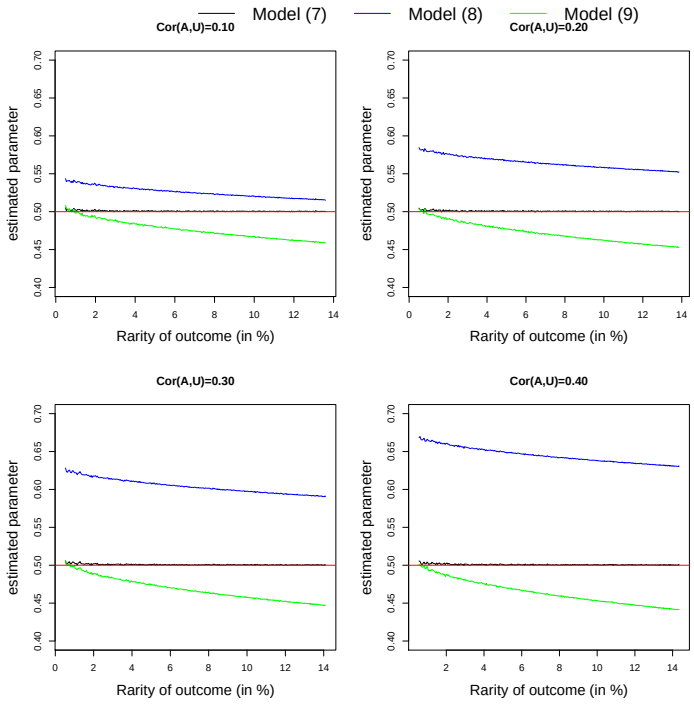


Figure 4.7: Results of 5000 estimations from the true model (model (7)), the naive model (model (8)), and from the IV model (model (9)), for six different confounding strength, with corr(A,Z)$\approx$ 0.70, as a function of the percentage of events. The true value of the parameter is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$

fails to estimate the true value of the parameter. Indeed, for low confounding (top-left plot), the naive estimate has a relatively small bias, while for high amount of confounding (bottom-right plot), the naive estimate is very highly biased. However, other simulations - in particular one not using an observed confounder - showed that the rare disease assumption can be slightly relaxed, and using 5% then seemed a judicious choice. Results from these simulations are presented in Table B.2 in Appendix B.2.

As for the variance of the estimates, under the rare disease assumption, if n is too small ($n = 1000$ is already too small), bootstrap samples might not contain any event at all, and bootstrapping will not work as expected. When n is big, however, bootstrapping 500 times inside 5000 simulations with $n = 10000$ will be computationally inefficient. Instead, Table 4.7 displays the average variance reported by the coxph function (both the robust variance and the so-called naive variance), together with the empirical variance for the three models. As surprisingly as in Section 4.1.1, the coxph function seems to already report the correct variance (both the robust and the naive variance). Even though it could be due to the fact that these are very small to begin with, this needs further investigating.

Table 4.7: Result of 5000 simulations for the proportional hazards model with corr(A,Z)≈ 0.70, and corr(A,U)≈ 0.40. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$.

	model (1)	model (2)	model (3)
coxph naive variance	0.0066	0.0056	0.0092
coxph robust variance	0.0066	0.0060	0.0091
empirical variance	0.0066	0.0060	0.0092

4.3 Proportional hazard – rare outcome

In this section, we will control the number of events to an approximately fixed amount so that we obtain about 1% of outcome. Once more, Z and D were as in 4.1, the correlation between the unknown confounder U and A was held fixed at about 0.40, and the correlation between the instrumental variable Z and the exposure A was varied. Once all these restrictions had been taken care of, a dataset with 10000 observations was generated 5000 times for varying IV strengths, estimation of β_a was performed with models (7)-(9), and the resulting plots are displayed in Figure 4.8. The R-code used for this section can be found in Appendix D.4.

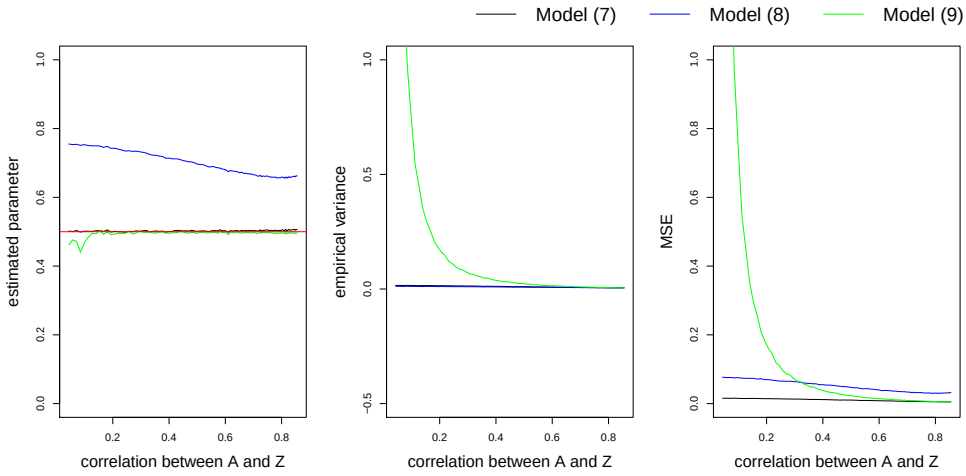


Figure 4.8: Estimated parameters (left), empirical variance (middle) and mean squared error (right) as a function of the correlation between the IV Z and the covariate A. We display the results from true model (model (7)), from the naive model (model (8), and from the IV model (model (9)). The number of events was held fixed to about 1%, with corr(A,U)≈ 0.40. The true value of the parameter is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$

The two-stage procedure (model (9)) seems to perform better than naive model (model (8)). Once more, the IV estimate seems to be consistent, but with high variance, especially when the instrument is weak. Figure 4.9 displays the plot of the estimated parameters as a function of the correlation between A and Z and their 95% confidence intervals.

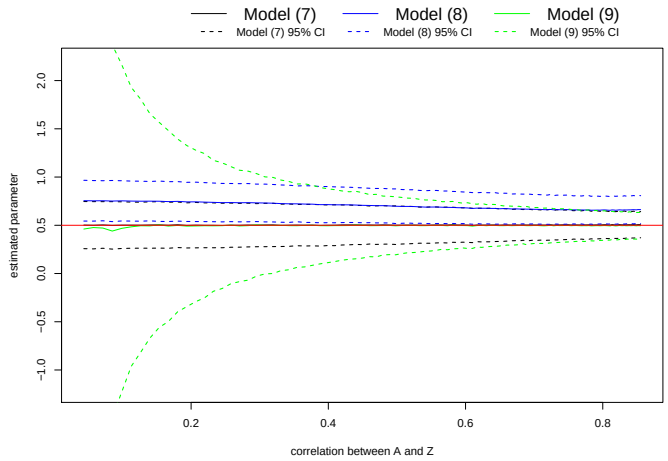


Figure 4.9: Estimated parameters and their confidence intervals as a function of the correlation between the IV Z and the covariate A, from model (7), model (8), and model (9). Thick lines represent the estimated parameter while the dotted lines represent the bounds of the 95 % confidence intervals. The number of events was held fixed to about 1%, with corr(A,U)$\approx$ 0.40. The true parameter value is $\beta_a = 0.50$ and is represented by a red line. The number of observations in each simulation was $n = 10000$.

The plot presented in Figure 4.9 illustrates once more the consistency of the IV estimate, and shows us that the resulting confidence intervals (seen in green) are much more conservative than the ones obtained from the naive estimates (seen in blue). More importantly, note that for these simulations, the confidence bounds around the naive estimate do not include the true parameter at all.

4.4 IV estimation under a polynomial exposure model

In this section, we will look at an alternative model for the exposure, of the form (3.35), i.e.

$$A = g(Z, \boldsymbol{c}) + c_d D + \delta,$$

where we will look specifically at

$$g(Z, \boldsymbol{c}) = c_0 + c_{z1} Z + c_{z2} Z^2$$

for both the additive hazard model with constant parameters, and proportional hazards model, and where $\boldsymbol{c} = (c_0, c_{z1}, c_{z2})$. This time, however, both the instrumental variable Z and the observed confounder D were generated from a normal distribution with mean 1 and variance 1. The reason for this is purely for the sake of illustration, as it is easier to see if a second order

polynomial is adequate when investigating the relationship between A and Z when both variables are continuous.

4.4.1 Polynomial first stage for the Lin-Ying model

We generate the exposure from the model

$$A = c_0 + c_{z1}Z + c_{z2}Z^2 + c_dD + \delta, \tag{4.10}$$

with $c_0 = 0.01$, $c_d = 0.8$, and c_{z1} and c_{z2} where varied adequately in order to obtain varying IV strengths. We then generate $n = 1000$ event times from (4.5). The correlation between A and U has been held constant to about 0.40 and the number of events to about 20%. We then generate survival times and fit models (1), (2) as in Section 4.1 and model (3) is fitted using a second order polynomial to predict M. Here we know how the exposure is depending on the instrumental variable. In reality, this might not be the case, and we thus need to investigate this relationship. Figure 4.10 shows a simple plot of A as a function of Z and we clearly see that a second degree polynomial seems appropriate when fitting a model for the exposure. Alternatively, one could perform a series of likelihood ratio tests by sequentially adding higher order terms in the model, and see which terms have a significance.

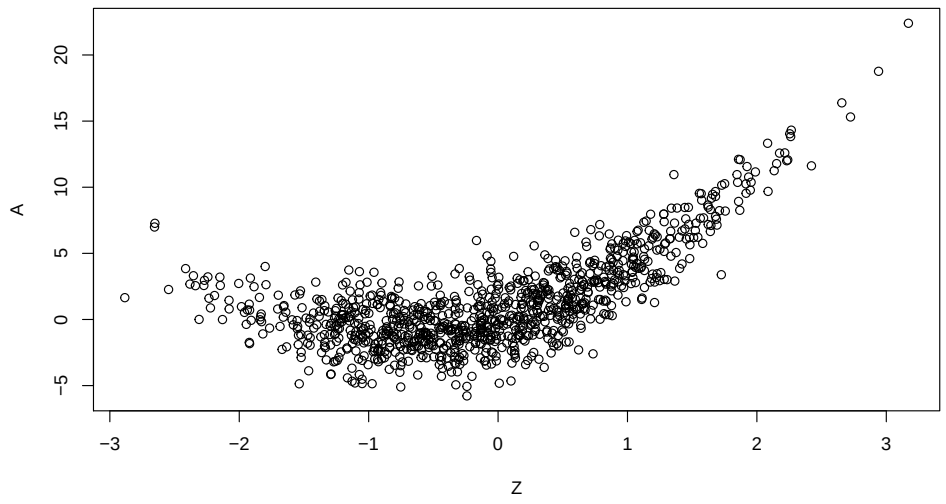


Figure 4.10: Scatterplot of A versus Z.

The results of 5000 simulations are presented in Figure 4.11. The two-stage procedure seems to perform quite well asymptotically under this alternative first stage model, even for weaker IV's. The reported empirical variance is again quite high compared to model (2) and model (3).

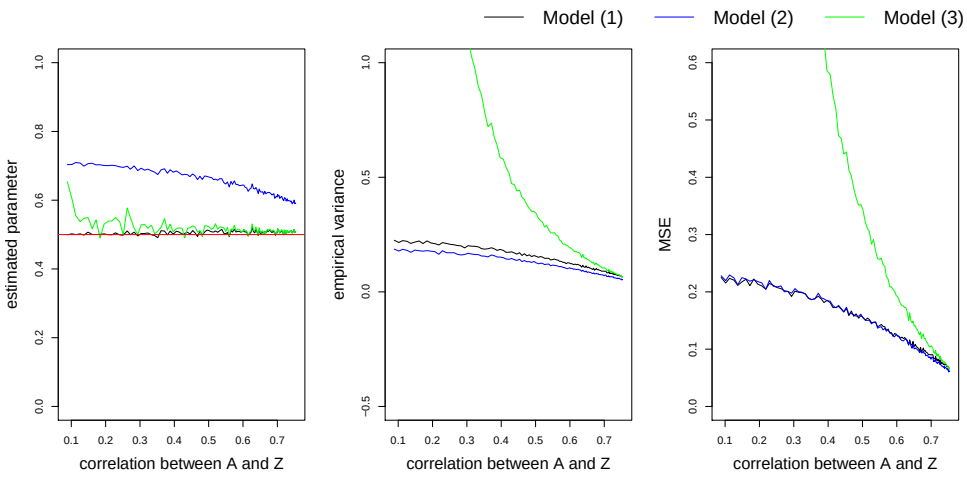


Figure 4.11: Estimated parameters (left), empirical variance (middle) and mean squared error (right) as a function of the correlation between the IV Z and the covariate A. We fixed corr(A,U)$\approx$ 0.40, and displayed he results from true model (model (1)), from the naive model (model (2)), and from the IV model under the alternative first stage (model (3)). The true parameter value is $\beta_a = 0.50$ and is represented by a red line in the left hand figure. The number of observations in each simulation was $n = 1000$.

4.4.2 Polynomial first stage for Aalen's additive hazard model

We generate the exposure from the model (4.10), and generate $n = 1000$ event times from (4.7). The correlation between A and Z has been held constant to about 0.70, and the correlation between A and U to about 0.20. We fit models (4)-(5) by a similar procedure to the one described in 4.1.3, and model (6) is fitted using a second order polynomial to predict M. We then specify values of t for which we would like to see the effect of $\beta(t)$ on the hazard difference. We choose the values at time $t = 0.1$ and $t = 0.14$ so that the true value of $B_a(t)$ at these times are:

$$\begin{aligned} B_a(0.1) &= 0.1575, \\ B_a(0.14) &= 0.2247, \end{aligned}$$

We present the results of 5000 simulations in Table 4.8a for model (4), in Table 4.8b for model (5), and in Table 4.8c for model (6).

The results displayed in the tables illustrate that under a polynomial first stage, the two-stage procedure still recovers a consistent estimate of the cumulative parameter $B_a(t)$, while the estimates coming from the naive model are clearly biased.

Table 4.8: Results of 5000 simulations and fitting of the true model (model (4)), the naive model (model (5)), and the IV model (model (6)) under a polynomial first stage, with corr(A,Z)$\approx$ 0.70, and corr(A,U)$\approx$ 0.20. The true parameter values are $B_a(0.1) = 0.1575$, and $B_a(0.14) = 0.2247$. The number of observations in each simulation was $n = 1000$.

(a) True model - model (4)

	t=0.1	t=0.14
estimate of $B_a(t)$	0.15751	0.22461
empirical variance	0.00148	0.00872
MSE	0.00148	0.00872

(b) Naive model - model (5)

	t=0.1	t=0.14
estimate of $B_a(t)$	0.17447	0.25006
empirical variance	0.00107	0.00587
MSE	0.00136	0.00651

(c) IV model - model (6)

	t=0.1	t=0.14
estimate of $B_a(t)$	0.15805	0.22429
empirical variance	0.00186	0.01161
MSE	0.00186	0.01161

4.4.3 Polynomial first stage for Cox's Proportional hazards model

We generate the exposure from the model (4.10), with $c_0 = 0.01$, $c_d = 0.8$, and c_{z1} and c_{z2} where varied adequately in order to obtain varying IV strength. The correlation between A and U was held fixed to about 0.40. We then generate survival times and fit models (7), (8) as in Section 4.2 and model (9) is fitted using a second degree polynomial to predict M. Here we know how the exposure is depending on the instrumental variable. In reality, this might not be the case, and we thus need to investigate the dependence structure. When making a plot of A versus Z (as a result of one simulation), we obtain a similar figure to Figure 4.10.

Figure 4.12 displays the results of 5000 simulations for six different percentage of events (from 0.5% to 3% of events), and varying IV strength for the three different models. It seems that the IV estimation is biased downwards. It performs better than the naive estimate for medium IV strengths, but decreases in performance when the IV becomes stronger. This is quite peculiar,

but the bias for weak IV's is perhaps due to the combination between rare outcome, polynomial exposure and weak IV.

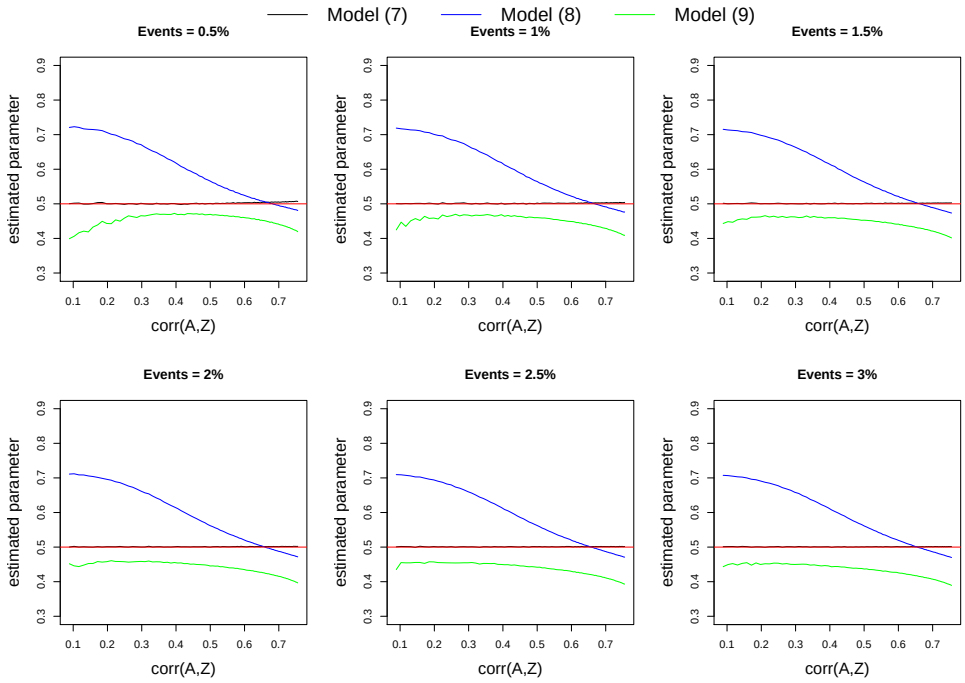


Figure 4.12: Results of 5000 estimations from true model (model (7)), from the naive model (model (8)), and from the IV model under the alternative first stage (model (9)), for different amount of outcome, as a function of the correlation between the IV Z and the covariate A. We fixed corr(A,U)≈ 0.40. The true parameter value is $\beta_a = 0.50$ and is represented by a red line. The number of observations in each simulation was $n = 1000$.

4.5 IV estimation under misspecified first stage

In this section, an attempt will be made to illustrate what happens if the exposure is categorical (or binary), but is still estimated by ordinary least squares (i.e. the first stage is misspecified). Because the exposure is now binary, it was slightly harder to fix the amount of confounding and the strength of the IV to fixed values. For both the additive and proportional hazards model, we will therefore restrict ourselves to a correlation between the instrumental variable Z and the expose A of about 0.60, and a correlation between A and the unknown confounder of about 0.10.

4.5.1 Misspecified first stage for the Lin-Ying model

We generated binary exposure A by first generating a continuous variable $A^\star$ as in (4.2), and then passing it through a inverse logit-function. In other words

we first generate:

$$A^\star = c_0 + c_z Z + c_d D + \delta, \tag{4.11}$$

and then generating probabilities by

$$P(A = 1|Z, D) = \frac{1}{1 + e^{-A^\star}} \tag{4.12}$$

which we then use to generate $n = 1000$ binomial observations with probabilities $P(A = 1|Z, D)$. The variable Z and D were this time both normally distributed, and the coefficients (c_0, c_z, c_d) chosen so that the IV is relatively strong (corr(A,Z)$\approx$ 0.64), δ was chosen so that there is some non-negligible unknown confounding (corr(A,U)$\approx$ 0.10), and the censoring was done so that we obtain about 25% of events. We then did a two-stage procedure where the first stage is a linear regression as before, instead of a logistic regression (i.e. misspecified first stage), and the second stage with constant parameters also as usual (i.e. model (3)). We also fitted model (1) and (2) for comparison. The results of 5000 simulations are presented in Table 4.9.

Table 4.9: Result of 5000 estimations for additive hazard model with constant parameters, and misspecified first stage. The true parameter value is $\beta_a = 0.5$, and the number of observations in each simulation was $n = 1000$.

	model (1)	model (2)	model (3)
estimate	0.493	0.625	0.515
emp. variance	0.468	0.458	1.218
MSE	0.467	0.473	1.218

The results presented in Table 4.9 seem to indicate that misspecification of the first stage still yields approximately consistent estimates, but result in a much higher variance. Misspecification of the first stage should thus be avoided until further investigations or simulations have been done. One should expect similar results, or worse, when misspecifying the first stage with time-varying parameters at the second stage.

4.5.2 Misspecified first stage for the proportional hazards model

Here we generated the exposure the same way as in Section 4.5.1, but with $n = 10000$ because of the rare disease assumption, and used corr(A,Z)$\approx$ 0.64 and corr(A,U)$\approx$ 0.10. The number of events was varied between about 1% to 3.5%, and the two-stage procedure was performed by ordinary least squares at the first stage, and fitting the proportional hazards model at the second stage (model (9)). We also fitted the true model (model (7)) and the naive model (model (8)) for comparison. The results of 5000 simulations are displayed in Table 4.10.

Table 4.10: Results of 5000 simulations and fitting of the true model (model (7)), the naive model (model (8)), and the IV model (model (9) with misspecified first stage), with corr(A,Z)$\approx$ 0.64 and corr(A,U)$\approx$ 0.10. The value of the true parameter is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$.

	Estimated parameter			*Empirical variance*		
% of events	model (7)	model (8)	model (9)	model (7)	model (8)	model (9)
1%	0.5586	0.6894	0.3861	0.1470	0.1460	0.1656
1.5%	0.5354	0.6668	0.3856	0.0909	0.0905	0.1126
2%	0.5277	0.6570	0.3918	0.0632	0.0627	0.0799
2.5%	0.5184	0.6482	0.3866	0.0497	0.0494	0.0645
3%	0.5166	0.6456	0.3875	0.0413	0.0407	0.0533
3.5%	0.5161	0.6445	0.3892	0.0344	0.0343	0.0474

The results reported in Table 4.10 show that the two-stage procedure using the proportional hazards at the second stage is not robust against misspecification of the first stage. Indeed, only the true model (model (7)) seems to estimate the parameter β_a correctly and only once we have enough outcomes, i.e. beyond 3.5% of events. This shows once more the importance of investigating the relationship between the instrument and the exposure before doing anything else.

4.6 Additional remarks on the proportional hazards model

One obvious problem that we will encounter when applying the IV estimation using Cox proportional hazards to real data, will be the construction of confidence intervals around the estimated effect. As noted earlier, because of the rare disease assumption, we typically will have very few events. Hence, if one wishes to perform a non-parametric bootstrap to obtain an estimate of the variance of the estimated effect, especially for small samples, there is no guaranty that bootstrap samples will also satisfy the rare disease assumption. In fact, some bootstrap samples might contain no events at all. Therefore, until an expression for the asymptotic variance is derived, one should be careful when using Cox proportional hazards for IV estimation.

What might be possible to do, however, is to test the null hypothesis of no causal effect. Indeed, as noted in Section 3.2.2, the non-collapsibility of the proportional hazards model might not be a burden if the true causal effect is zero, in which case a confidence interval can be constructed by bootstrapping, and the hypothesis can be accepted if the interval contains zero. This exact feature has in fact been noted in a recent commentary (Burgess, 2015) to Tchetgen et al. (2015).

We can illustrate this feature by generating a new dataset from

$$h(t|A,U,Z,D) = h_0(t)e^{\beta_a A+\beta_d D+\beta_u U}, \tag{4.13}$$

where $\beta_a = 0$, $\beta_d = 0.50$ and $\beta_u = 0.50$, and $h_0(t)$ is as in 4.2. We saw previously that misspecification of the first stage will lead inconsistent results, we could thus see if it is still the case if the true parameter is zero. We therefore generate A as in 4.5, (i.e. binary A), and so that corr(A,Z)$\approx$ 0.64, and corr(A,U)$\approx$ 0.10. The number of events here is arbitrary, but we made sure we obtain at least 25% of events. We then generate $n = 1000$ observations and fit models (7), (8) and (9). The results of 5000 simulations are presented in Table 4.11.

Table 4.11: Result of 5000 estimations for the proportional hazards model, and misspecified first stage. The true value of the parameter is $\beta_a = 0$, and corr(A,Z)$\approx$ 0.64, and corr(A,U)$\approx$ 0.10. The number of observations in each simulation was $n = 1000$.

	model (1)	model (2)	model (3)
estimate	0.0030	0.1345	0.0006
emp. variance	0.0166	0.0165	0.0370
MSE	0.0165	0.0346	0.0370

What we suspected in Section 3.2.2 seems to be confirmed by the results in Table 4.11. It shows that both the rare disease assumption and correct specification of the first stage are not necessary when testing the null hypothesis of no causal effect. Indeed, the two-stage procedure seems to have estimated correctly the true parameter as zero even though the rare disease assumption is not satisfied and the first stage misspecified.

4.7 Additional remarks on the additive hazard model

For the additive hazard model (but in fact also for Cox's proportional hazards), we are restricted to the linear first stage, i.e. continuous exposure. We showed that misspecification of the first stage will yield very high variance of the estimate, and thus advised against this. What should be possible to do, however, is to test the null hypothesis of no causal effect as in the previous section (Tchetgen et al., 2015).

We can illustrate this feature by generating a new dataset from the Lin-Ying model:

$$h(t|A,U,Z,D) = h_0(t) + \beta_a A + \beta_d D + \beta_u U \tag{4.14}$$

where $\beta_a = 0$, $\beta_d = 0.50$ and $\beta_u = 0.50$, and $h_0(t)$ is as in 4.1.1. We generate A as in 4.5, (i.e. binary A), and so that corr(A,Z)$\approx$ 0.64, and corr(A,U)$\approx$ 0.10. The number of events here is arbitrary, but we made sure we obtain at least

25% of events. We then generate $n = 1000$ observations and fit models (1), (2) and (3). The results of 5000 simulations are displayed in Table 4.12.

Table 4.12: Result of 5000 estimations for the additive hazard model, and misspecified first stage. The true value of the parameter is $\beta_a = 0$, and corr(A,Z)$\approx$ 0.64, and corr(A,U)$\approx$ 0.10. The number of observations in each simulation was $n = 1000$.

	model (1)	model (2)	model (3)
estimate	-0.0085	0.0459	-0.0004
emp. variance	0.0010	0.0009	0.0032
MSE	0.0010	0.0030	0.0032

Table 4.12 shows that correct specification of the first stage are not necessary when testing the null hypothesis of no causal effect. Indeed, the two-stage procedure seems to have estimated correctly the true parameter as zero even though the first stage was misspecified. Note here that the naive model (model (2)) does not "miss" the true value of the parameter by much, as there is very low amount of unknown confounding. By generating data in a different fashion, one could add confounding, and the naive estimate can be expected to be much more biased.

5

Application - MoBa cohort

In this chapter, we will try to apply two-stage predictor substitution to the Norwegian Mother and Child Cohort Study (MoBa). This cohort study has been assembled from 1999 and some parts of the cohort are still being assembled now. In 2008, more than 100 000 women and 70 000 men had participated to the study. The primary reason for assembling such a cohort, is that there is not enough information available about the causes for serious diseases or illnesses in children. Women participating in this study have answered numerous questionnaires throughout their pregnancy and in the eight years following birth. They answered thousand of questions, and due to the size of this cohort, we will have to restrict ourselves to one exposure, a couple of confounders, and one IV. In Chapter 4, we have discussed and illustrated possible pitfalls of using Cox's proportional hazards models for IV estimation. For the application, we will thus only use Aalen's additive hazard model. The outcome will have to be of the time-to-event type, as for example the time until the child walks, or the time from conception until birth.

Every child in the dataset has a pregnancy ID. There are so far nine questionnaires for the mothers. In addition, we have data from the MFR, containing additional (and sensitive) information about the mothers and the children. These contain variables like mothers age, pregnancy duration, what medication they used during pregnancy, and much more. Originally, we wished to look at the event time-to diagnose of ADHD but we did not get access to this part of the dataset. We will therefore consider the event time from conception until child birth, or more specifically, gestational age. We will refer to Appendix E for the R-codes used to perform the analysis.

5.1 Time-to-childbirth

We will consider pregnancy duration (or gestational age), as the time-to-event response (in days). One factor that could have an effect on the pregnancy time is the mother's weight, or more specifically, the mother's Body Mass

Index (BMI) before the pregnancy. The BMI of an individual is simply the weight (in kilograms) divided by the square of the height (in meters). It is thus reported in kg/m^2.

The reason for this choice of outcome and exposure is based on the fact that previous studies have investigated the relationship between maternal underweight and preterm birth (Han et al., 2010), and the relationship between maternal overweight and infant death (Johansson et al., 2014). The results presented in both these articles seem to indicate that maternal underweight or overweight has an effect on birth duration or birth outcome. We can thus infer that the BMI in general has an effect on the pregnancy duration.

The difficult task is finding an adequate IV for the BMI of the mother prior to pregnancy. Recall the three main assumptions for a valid IV:

(i) $Z \perp\!\!\!\perp U$

(ii) $Z \not\perp\!\!\!\perp A$

(iii) $Z \perp\!\!\!\perp T | (A, D, U)$

With the additional assumption of independent censoring. Unfortunately, as discussed in Section 3.1, assumptions (i) and (iii) are not verifiable. We will thus rely on common sense and previous studies to argue for the validity of these IV assumptions. Assumption (ii) will be verified by performing likelihood ratio tests, and by looking at the adjusted R-squared of a linear regression between the exposure and the potential instrument.

In relatively recent studies trying to relate obesity to medical costs (Cawley and Meyerhoefer, 2012) and obesity to wages (Cawley, 2004), authors argued that a valid instrumental variable for the BMI of the mother can be the BMI of an older child. In fact, they argue that the BMI squared and BMI cubed can also be included as instruments. They give some justifications for the validity of assumption (i) obtained from twin studies, however, they use a different response. In our situation, we could argue that the BMI of an older sibling should be independent of other factors influencing the pregnancy duration for a younger sibling. Furthermore it seems quite reasonable to assume that the BMI of an older child affects the pregnancy duration of a younger child only through the BMI of the mother, which informally justifies assumption (iii). Assumption (ii) will be tested later on.

The Mother and Child cohort has been assembled over a period of many years, and some women have participated in the study twice. Therefore, it should be possible to extract information about the women that had at least two children in the study, and use the reported weight and height of the oldest of the two to make an instrument for the mothers BMI. The weight and height of the children have been registered at birth, 6 months after birth, 18 month after birth, 36 month after birth, when the child is 5 years old, 7 years old

and 8 years old. Because height and weight is self reported, it is subject to measurement error, but also not every woman participating at least twice in the study has reported both age and height of the children in all the questionnaires. Furthermore, to obtain a stronger IV, we should use the BMI of the oldest possible child, which in this dataset is 8 years old. This, of course, will reduce the size of the dataset considerably, and one should keep this in mind when interpreting the results. Ideally, it would have been better to have at our disposition height and weight of the older sibling much later in life, as the calculated BMI would be more stable. This is however not possible here, we will therefore use the available data.

After extraction of the necessary information, we are left with 2324 individuals. Since all of the children come to term, there is no censored observations, the independent censoring assumption is thus irrelevant. In order to see what should be included as instruments, and check the validity of assumption (ii), we first checked the correlation between the oldest child's BMI and the mother BMI before pregnancy of the second child, and the instrument was found to be relatively weak, with a correlation of about 0.30. We then perform a simple linear regression on the mother's BMI using the oldest child's BMI as a covariate, and by adding the BMI squared and the BMI cubed sequentially and preforming a likelihood ratio test each time comparing models one by one, and conclude that only the BMI and the BMI squared should be included in the model. Table 5.1 shows the results of such tests.

Table 5.1: Results of the likelihood ratio test perfomed using anova testing the sequential addition of higher order BMI.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Z	1	2677.29	2677.29	176.99	0.0000
I(Z^2)	1	220.54	220.54	14.58	0.0001
I(Z^3)	1	19.35	19.35	1.28	0.2582
Residuals	2022	30586.50	15.13		

As seen in Table 5.1, only the BMI and BMI squared seems to have a significant effect on the response, a feature confirmed when performing a summary of the model. However, the said summary reports an adjusted-R^2 of about 0.086, which means that the model only explains about 9% of the variation in the response, and so the instrument will be rather weak. The summary of the model containing only BMI and BMI squared is reported in Table 5.2.

Using only the BMI and BMI squared of the first participating child at 8 years as an IV for the mother's BMI, we can illustrate the assumptions by DAG (see Figure 5.1).

Table 5.2: Summary of model checking if BMI and BMI squared have a significant effect on the response.

	Dependent variable:
	A
Z	-1.038^{**}
	(0.424)
I(Z^2)	0.047^{***}
	(0.012)
Constant	28.308^{***}
	(3.646)
Observations	2,026
R^2	0.086
Adjusted R^2	0.086
Residual Std. Error	3.890 (df = 2023)
F Statistic	95.771^{***} (df = 2; 2023)
Note:	*p<0.1; **p<0.05; ***p<0.01

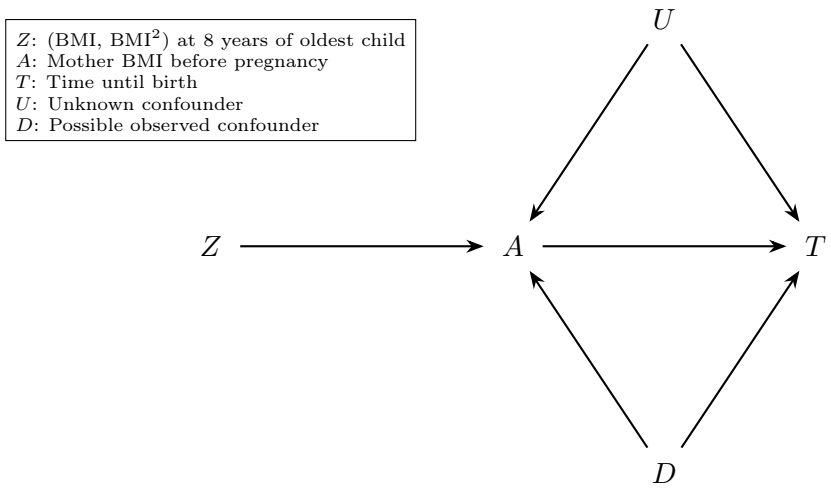


Figure 5.1: A directed, acyclic graph.

We then need to extract some observed confounders, and we gather them into a data-frame together with the IV and the exposure. We end up with a data-frame containing:

- A: BMI of mother right before pregnancy
- Z: BMI of first child at 8 years

- D1 : Mother age when pregnant with second child
- D2 : Region (South / East=1, West=2, Center=3, North=4)
  Reference level: 1
- D3 : Gender of first child (Boy=0, Girl=1)
  Reference level: 0
- D4 : Marital status (Married=1, Single=2, Living with partner=3)
  Reference level: 1
- D5 : Education - with levels:

  (1) 9 years basic school

  (2) 1-2 years of high school

  (3) professional high school

  (4) 3 years regular high school

  (5) university/private school up to 4 years

  (6) university/private school more than 4 years

  Reference level: 5

After specifying the reference levels for these confounders, we perform a similar analysis as the one we did for deciding which BMI to include in the model to see which confounders can be included. We thus fitted a first model with A as the response, Z and Z^2 as the instrument and D1 to D5 as observed confounders. Table 5.3 displays the summary of such a model, and we note that only confounders D1, D2, D4 and D5 seem to be significant enough to keep in our model. This was confirmed by performing likelihood ratio tests adding the confounders sequentiality (output not displayed).

5.1.1 First stage

The first stage is fairly simple. We predict A using the following model:

$$\mathrm{E}[A|Z, \boldsymbol{D}] = c_0 + c_{z1}Z + c_{z2}Z^2 + c_d'\boldsymbol{D}, \tag{5.1}$$

where $c_d' = (c_{d1}, c_{d2}, c_{d4}, c_{d5})$, and where $\boldsymbol{D} = (D1, D2, D4, D5)$. We call the predicted value of A obtained from this model $\hat{M}$, and include it in our data frame. The summary of this model is displayed in Table 5.4.

Table 5.3: Summary of the model including all the confounders.

	Exposure:	
	A	(Std. error)
Z	−1.059**	(0.421)
I(Z^2)	0.046***	(0.012)
D1	0.051**	(0.024)
D22	−0.337*	(0.195)
D23	0.631**	(0.267)
D24	0.014	(0.379)
D31	0.036	(0.172)
D42	4.106**	(1.725)
D43	0.226	(0.181)
D51	2.941***	(1.001)
D52	1.355***	(0.516)
D53	0.941***	(0.314)
D54	0.831***	(0.287)
D56	−0.706***	(0.208)
Constant	27.027***	(3.686)
Observations:	2,012	
R^2:	0.120	
Adjusted R^2:	0.114	
Residual Std. Error:	3.837 (df = 1997)	
F Statistic:	19.421*** (df = 14; 1997)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

5.1.2 Second stage

Before performing the second stage, it could be interesting to look at a few things. We will start with a simple Kaplan-Meier plot estimating the survival curve of the time until birth. The Kaplan-Meier estimator, when there is no ties is defined as

$$\hat{S}(t) = \prod_{T_i \leq t} \{1 - \frac{1}{Y(T_i)}\}, \tag{5.2}$$

where T_i is the event time for individual i, and $Y(T_i)$ is the number of individuals at risk “just before" time T_i. Because of the nature of our data, we do have ties, but no censored data, the Kaplan-Meier estimator can then simply

Table 5.4: Summary of the first stage model.

	Exposure: A	(Std. error)
Z	-1.064^{**}	(0.420)
I(Z^2)	0.046^{***}	(0.012)
D1	0.051^{**}	(0.024)
D22	-0.337^{*}	(0.195)
D23	0.632^{**}	(0.267)
D24	0.018	(0.379)
D42	4.102^{**}	(1.724)
D43	0.227	(0.181)
D51	2.940^{***}	(1.001)
D52	1.354^{***}	(0.516)
D53	0.939^{***}	(0.314)
D54	0.830^{***}	(0.287)
D56	-0.705^{***}	(0.208)
Constant	27.091^{***}	(3.673)
Observations:	2,012	
R^2:	0.120	
Adjusted R^2:	0.114	
Residual Std. Error:	3.836 (df = 1998)	
F Statistic:	20.921^{***} (df = 13; 1998)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

be rewritten as:

$$\hat{S}(t) = \frac{\#\{T_i > t\}}{n}, \tag{5.3}$$

where $\#\{T_i > t\}$ denotes the number of individuals remaining in the study at time t, and n denotes the total number of individuals in the study. We present such a plot in Figure 5.2

Figure 5.2 shows us that there is a very small interval of days where almost all the births happen. Alternatively, we can recode our `status` vector according to a cutoff point at day 260, which, according to specialists, births occurring before day 260 (week 37) can be considered premature. After recoding, we call this new status vector `status2`, add it to our data frame and compute a new Kaplan Meier estimate that we display in Figure 5.3.

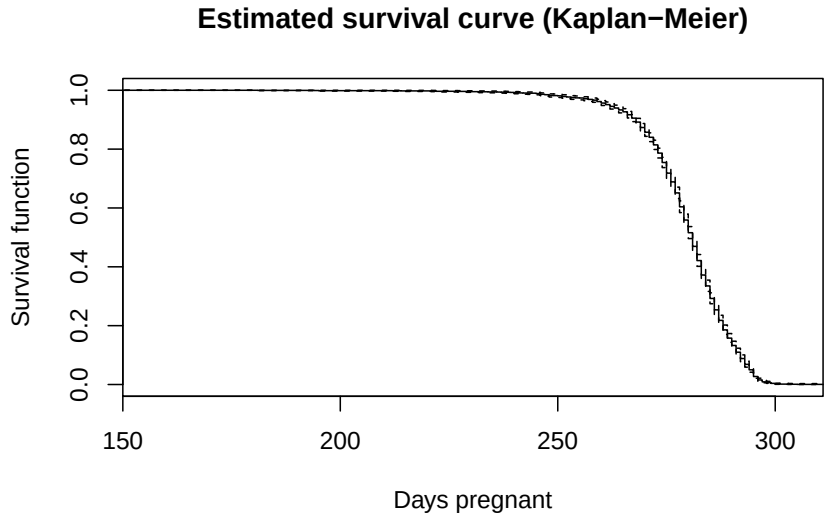


Figure 5.2: Estimated survival curve for the time until birth.

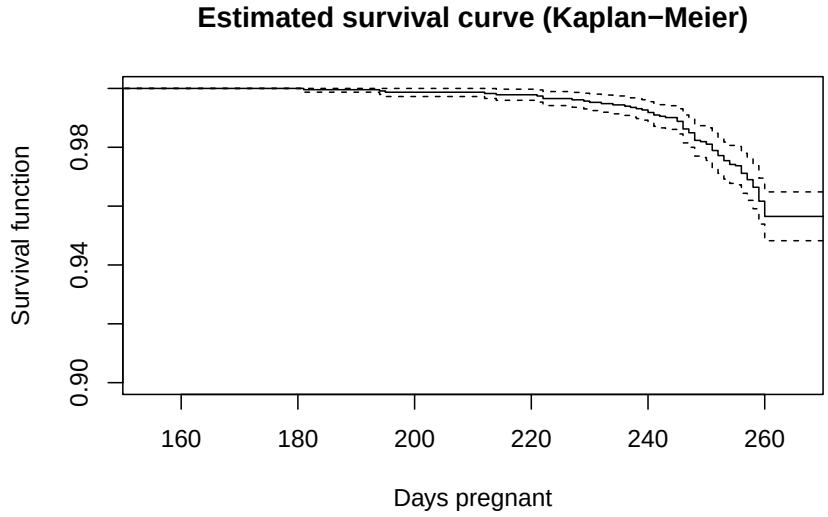


Figure 5.3: Estimated survival curve for the time until premature birth.

Figure 5.3 tells us that there are very few recorded premature birth, and we see that the estimated survival probability is nearly one across the whole pregnancy time. However, as noted If there was not so many pitfalls, this would be a perfect situation for using Cox's proportional hazards model, but we would almost certainly not know if the results are significant or not. Therefore, we will restrict ourselves to Aalen's additive hazard model for time until

birth.

We now fit a model with all the observed variables, i.e.

$$h(t|A, Z, \boldsymbol{D}) = h_0(t) + \beta_a(t)A + \beta_d'(t)\boldsymbol{D}, \tag{5.4}$$

where $\beta_d'(t) = (\beta_{d1}(t), \beta_{d2}(t), \beta_{d4}(t), \beta_{d5}(t))$ and $\boldsymbol{D} = (D1, D2, D4, D5)$. We can look at the behavior of the estimated cumulative baseline hazard $\hat{H}_0(t)$ as well as of the estimated cumulative parameter for A, i.e. $\hat{B}_a(t)$. These estimates are displayed in Figure 5.4 together with generated 95% confidence intervals.

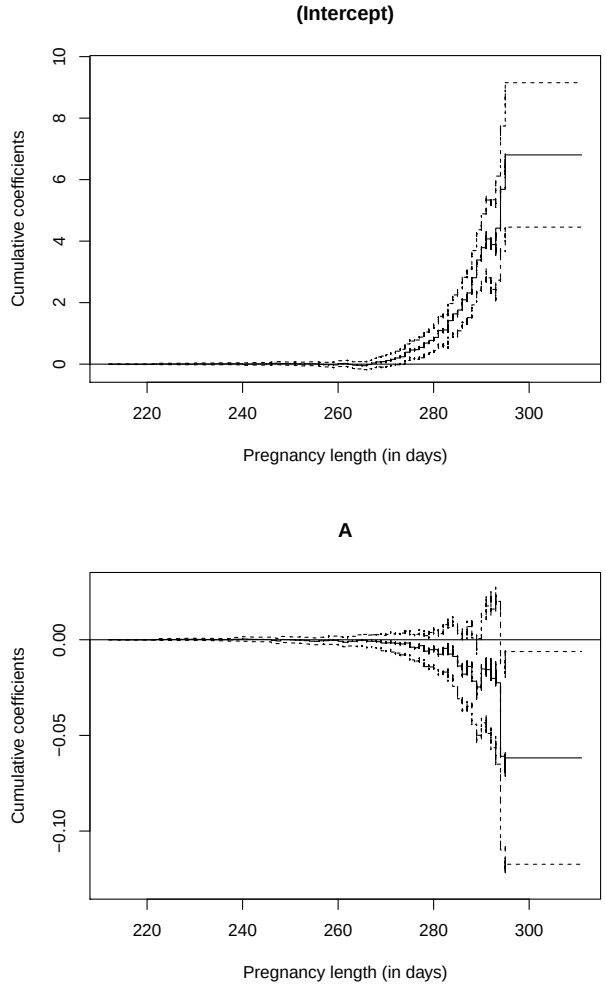


Figure 5.4: Estimated cumulative parameters for the observational model.

There are two important things to note when looking at Figure 5.4. Firstly, from the left-hand plot, there seems to be a significant baseline effect on the pregnancy duration. Secondly, from the right-hand plot, there does not seem to be a very significant effect of the mother's BMI on the pregnancy duration, except perhaps around 290 days, where an increase in BMI of one seems to decrease the hazard difference very so slightly. Doing a quick observational analysis as this one before the two-stage predictor substitution can be useful, as it enables us to have a prior idea of what is happening. It can be informative to present both results, and very big discrepancies between the two should be reported and further investigated.

We can now perform the second stage, by fitting the model

$$h(t|Z, \boldsymbol{D}) = \tilde{h}_0(t) + \beta_a(t)\hat{M} + \beta_d'(t)\boldsymbol{D}, \tag{5.5}$$

where $\beta_d'(t)$ and $\boldsymbol{D}$ are the same as in (5.4), and where $\hat{M}$ is the predicted value of A predicted using model (5.1). The resulting estimate of the cumulative regression parameter are shown in Figure 5.5, together with the bootstrapped confidence intervals (in green), and the ones reported by the aalen function (in blue).

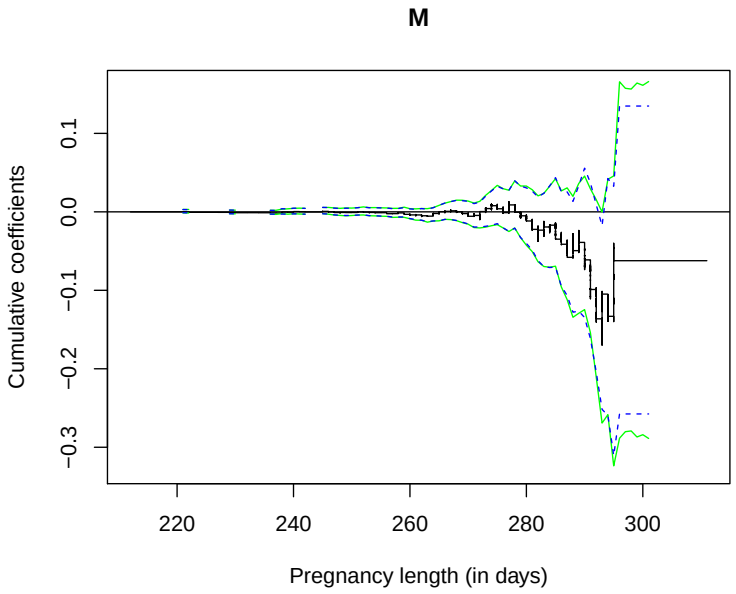


Figure 5.5: Estimated cumulative regression parameter of the BMI of the mother before pregnancy. Confidence intervals obtained from the aalen function are displayed in blue, and the bootstrapped confidence intervals are displayed in green.

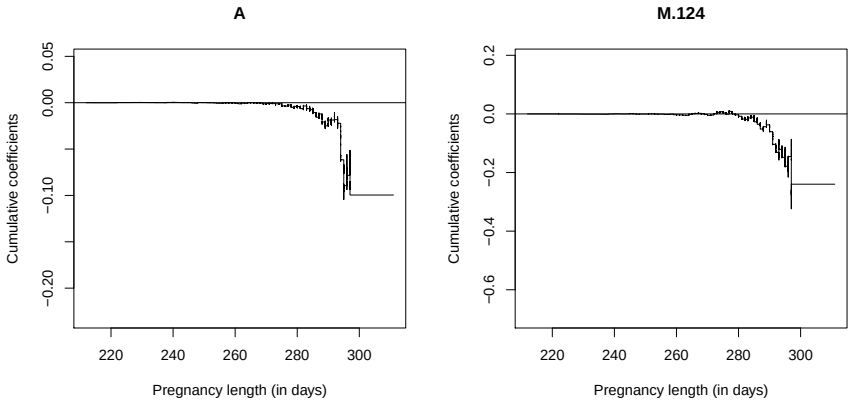
At first glance, we can be reassured that the estimated cumulative parameter in Figure 5.5 has a similar behavior as the one obtained from a model directly making use of A. The confidence intervals computed around the estimate are quite wide, and the only place where the confidence interval does not include zero is at day 293. There seem to be a small effect of BMI on the pregnancy length. It is a negative effect, thus reducing the risk giving birth. In other words, before pregnancy, a one unit increase in the mother's BMI seems to prolong pregnancy given that the said mother has not given birth yet by day 293. Furthermore, we can note that there is very little difference between the bootstrapped confidence interval and the one reported by the aalen function.

An additional and perhaps important thing to note is that we in fact performed three more IV analyses, each of which contained a different combination of confounder than the other. For each of these other IV analyses we also fitted their observational counterpart (the models with the same amount of confounding and using A as exposure). We then plotted each of these side by side with the IV models, and the results are presented in Figure 5.6.

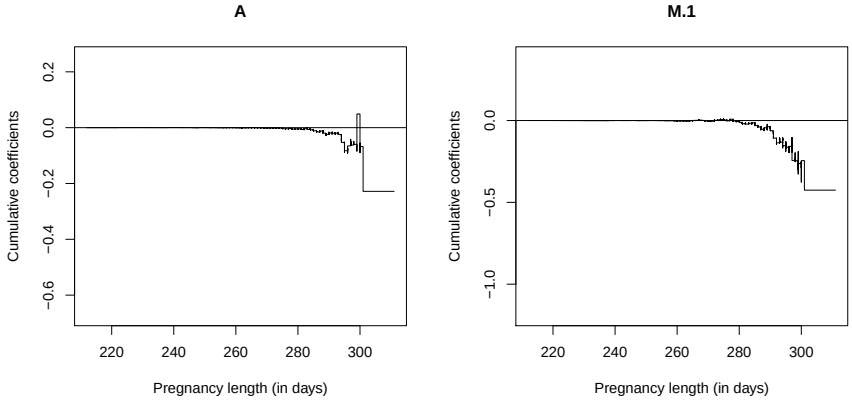
The plots that are on the left hand side in Figure 5.6 are the plots resulting from the estimation using the original exposure A. We can see that the effect of A is quite different depending on which observed confounder included in the model. However, the IV counterpart of these models shows an estimate that is much more stable and has almost the same behavior regardless of which confounder is used in the model. This does not mean nor prove that the IV assumptions hold, but assuming that they do hold, and provided that the IV is strong enough, any additional confounding added in the model might be washed out by the first stage estimation.

Of course, we have to keep in mind how we built our instrument. By choosing this specific group of mothers, we reduced considerably the sample size and we most likely lost a lot of information in the process. In addition, as noted before, it would have been ideal to obtain the BMI of the older child at a later stage in life. Other studies suggest that the age of the first menstruation of the mother is correlated with the mother's BMI (Pierce and Leon, 2005). This could be an interesting instrumental variable to use. We do however not have this information at our disposition.

(a) Models using D1, D2, D4



(b) Models using D1



(c) Models using D2

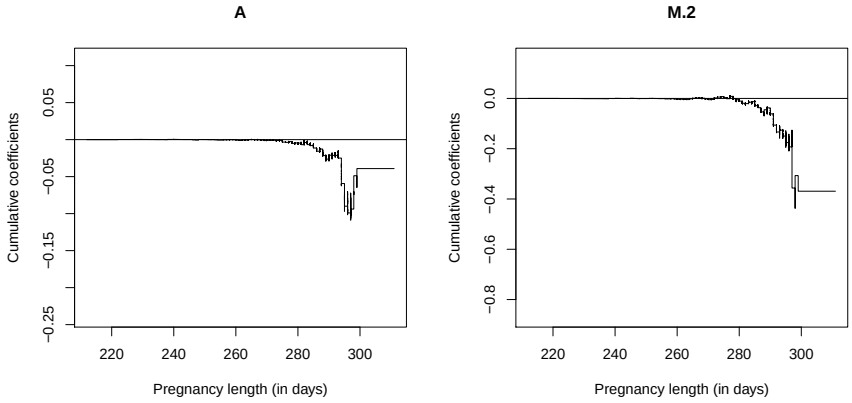


Figure 5.6: Comparison of three different IV models (right) and their observational counterparts (left). The plots displayed in 5.6a used a model with confounders D1, D2, D4, the plots in 5.6b used only D1, and the two plots in 5.6c used only D2.

6

Discussion and concluding remarks

In this thesis, we have attempted to illustrate and give an overview of a relatively novel method enabling us to use instrumental variables for estimating a measure of association between an exposure and a time-to-event response via two-stage predictor substitution. We have presented the concept of collapsibility, and how it relates to the two-stage procedure, and have also noted that this assumption might not be absolutely necessary when testing the null hypothesis of no causal effect. As for the consistency of the IV estimator, we have argued that at least for collapsible generalized linear models under both linear and non-linear models for the exposure, the estimator resulting from the two-stage procedure should be consistent. For the additive hazard, the consistency of the estimator has been shown (Tchetgen et al., 2015), but unfortunately, the appendix in which it is published has several obvious typos (see `http://links.lww.com/EDE/A893` for details). As for the consistency of the estimator under a proportional hazard method at the second stage, the same authors have deemed it sufficient to say that such a model (under the rare disease assumption) is consistent since it is essentially the same as performing a log-linear regression at each time point for the individuals in the risk set at that point.

In Section 2.3.3, when we showed non-collapsibility of the proportional hazards model, we saw that if the true parameter is in fact zero, the measure of association obtained from the marginal model is also zero. One could thus suggest that the collapsibility assumption can be relaxed if we are interested in testing the null hypothesis of no causal effect. This feature had been noted by various authors (Burgess, 2015; Tchetgen et al., 2015), and has then been confirmed by the simulations done in Section 4.6, where we also illustrated, simultaneously, that misspecification of the first stage still seems to lead to consistency in the second stage. This is a very important feature, as we often are not interested in obtaining a number, but rather a yes/no answer to the

question: "does A cause T?"

A note on bootstrapping should also be made here, as in general, unless recent development suggest a closed form for the variance of the IV estimate, one will typically have to rely on bootstrapping in order to compute confidence intervals. Bootstrapping is expected to work. However, for Cox's proportional hazards, the additional assumption of rare outcome can seem a little bit cumbersome since it will allow estimation of the parameter of interest, but performing bootstrapping when n is small, can lead to some bootstrap samples not containing any events. In addition, as discovered in the simulations, both the aalen function and the coxph function seemed to give the correct estimate of the variance, at least for the sample sizes and parameter values considered here. This is a very peculiar thing, as the aalen function does not know how much extra variation is gained when estimating the first stage, and should therefore be further investigated.

To the best of our knowledge, something that was not discussed in the literature when this thesis was started (in the context of time-to-event data), is the use of an alternative first stage. It is of course a natural extension to the simple linear first stage, but it was necessary to discuss it before applying the method to a real dataset. We illustrated the fact that consistency is still preserved under a polynomial first stage for Aalen's additive hazard model (see Section 4.4.1 and 4.4.2), however, this does not seem to be the case with Cox's proportional hazards on the second stage, which makes the two-stage predictor substitution even less attractive for this model (see Section 4.4.3).

Finally, even though the resulting analysis on the effect of a mother's BMI on the pregnancy duration did not yield very strong nor clear results, we hope that it provided the reader with a simple illustration on how to proceed.

6.1 Concluding remarks

An important point that has not been discussed in this thesis, are the assumptions regarding the censoring scheme. There is of course a broader class of censoring schemes that are not independent. In a recent research paper (Martinussen et al., 2016), the authors have also taken into account different censoring schemes including some forms of dependent censoring. Furthermore, the methods presented in that research paper can handle arbitrary exposures, and are thus not limited to the continuous exposure that we had to assume in order for the two-stage procedure to work. In addition, these authors also consider in particular the extended DAG drawn in Figure 6.1.

A note should be made on the assumptions for a valid instrument. The main problem with instrumental variables are the unverifiable assumptions (i) and (iii). As we discussed in Chapter 3, some authors suggested falsifying these assumptions, but it is a difficult task and it is sometimes better to rely on common sense or discuss it with specialists. Further work thus needs to be

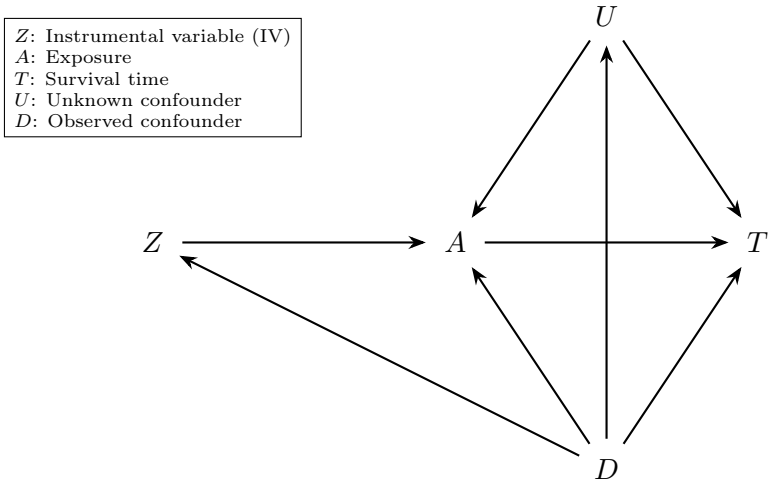


Figure 6.1: A directed, acyclic graph of the setup

made on how to test or falsify these assumptions. Alternatively, twin studies, where we gather information about twins that were separated at birth, could give us additional information on the type of unknown confounding.

The main focus of this thesis was the two-stage procedure. In fact, from a computational point of view, at least in linear regression models, one can always view an IV estimation as a two-stage least square regression (Bowden and Turkington, 1984, p.34-5,65), but intuitively, an IV estimate could be any estimator of a causal effect obtained through the use of an instrumental variable. In Bowden and Turkington (1984) and Wooldridge (2002) for example, a three stage procedure is used to find consistent estimates of the target causal effect. In Tchetgen et al. (2015), they also introduce a control function approach instead of the two-stage procedure. In a recent research paper (Vansteelandt and Didelez, 2015), the authors suggest robust IV estimation methods diverging from traditional two-stage estimators. Overlapping authors have very recently submitted a research paper on arXiv (Martinussen et al., 2016) where they make use of causal inference calculus to help justify a recursive estimate of the cumulative regression parameter, and show that this estimate is in fact very robust. Even more recently, developments have been made on the semi-parametric version of Aalen's additive hazard model in a competing risk setup (Zheng et al., 2017). Thus different IV estimators may have different asymptotic properties, and be more efficient or better at reducing the bias than others. This makes instrumental variable estimation a very broad and interesting field of study, and knowledge of causal inference seems more and more necessary in order to contribute to this field of research. We can thus recommend reading *Causal Inference in Statistics, a Primer* (Pearl et al., 2016), which is a more accessible version of *Causality* (Pearl, 2009).

Appendix A

From Chapter 2.4 - Generating survival times directly

Independently of the type of hazard $h(t|X)$, we have a survival function

$$S(t|X) = \exp[-H(t|X)], \tag{A.1}$$

where $H(t|X) = \int_0^t h(s|X)ds$, and recall that the distribution function of T can be written

$$F(t|X) = 1 - \exp[-H(t|X)]. \tag{A.2}$$

By definition, $F(t|X)$ takes values between 0 and 1, so if T is a random variable, then $U = F(T|X)$ follows a uniform distribution on the interval $[0, 1]$, i.e.

$$U \sim U[0, 1].$$

In addition, if $U \sim U[0, 1]$ then $1 - U \sim U[0, 1]$, so from (A.2), we have

$$V = F(T|X) = 1 - \exp[-H(T|X)] \sim U[0, 1],$$

so that

$$U = 1 - V = \exp[-H(T|X)]. \tag{A.3}$$

Let $G(T) = \exp[-H(T|X)]$. And it follows from Definition 6, since T is a continuous random variable with cumulative distribution function $F(t|X)$, and $U = 1 - F(T|X) = G(T) \sim U[0, 1]$, the random variable $G^{-1}(U)$ is the

inverse probability integral transform and it has the same distribution as T. It follows that the survival time T can be written as

$$T = G^{-1}(U), \tag{A.4}$$

which can be found explicitly by noting

$$\begin{aligned} & U & &= \exp[-H(t|X)] \\ \Leftrightarrow \quad & -\log(U) & &= H(t|X), \end{aligned}$$

which, provided that $H(t|X) > 0$ for all t, we can invert H, and obtain

$$T = H^{-1}[-\log(U)]. \tag{A.5}$$

A.1 The Cox proportional hazards model

For the Cox proportional hazards model, we have the following survival function

$$\begin{aligned} S(t|X) &= \exp[-H(t|X)] \\ &= \exp[-H_0(t)\exp(\beta_x' X)]. \end{aligned}$$

We will restrict ourselves to the Weibull baseline hazard $h_0(t) = \lambda v t^{v-1}$, so that

$$\begin{aligned} H_0(t) &= \int_0^t h_0(s)ds \\ &= \lambda t^v. \end{aligned}$$

We thus need to find H^{-1}, the inverse of the function $H(t|X) = \lambda t^v \exp(\beta_x' X)$, such that

$$\begin{aligned} & H(H^{-1}(t|X)|X) & &= t \\ \Leftrightarrow \quad & \lambda(H^{-1}(t|X))^v \exp(\beta_x' X) & &= t \\ \Leftrightarrow \quad & (H^{-1}(t|X))^v & &= \frac{t}{\lambda \exp(\beta_x' X)} \\ \Leftrightarrow \quad & H^{-1}(t|X) & &= \left(\frac{t}{\lambda \exp(\beta_x' X)}\right)^{\frac{1}{v}}, \end{aligned}$$

so that, for a Weibull baseline hazard, we have

$$\begin{aligned} T &= H^{-1}[-\log(U)|X] \\ &= \left(\frac{-\log(U)}{\lambda \exp(\beta_x' X)}\right)^{\frac{1}{v}}. \end{aligned}$$

It suffices now to draw, say $n = 1000$ values u_i coming from an Uniform distribution $U(0,1)$, and compute

$$t_i = \left(\frac{\log(u_i)}{\lambda \exp(\beta_x' X)}\right)^{\frac{1}{v}} \text{ for all } i = 1, \ldots, n,$$

in order to obtain survival times.

A.2 The Aalen additive hazard model

For the Aalen additive hazard model, we have the following survival function

$$\begin{aligned} S(t|X) &= \exp[-H(t|X)] \\ &= \exp[-\int_0^t (h_0(s) + \beta_x(s)X)ds]. \end{aligned}$$

Once more, we will restrict ourselves to parameters of the Weibull type, but with additional restrictions, being that the regression parameters should be first order polynomials. The reason for this choice is that if the regression parameters and/or baseline hazard have more complicated forms, it will be difficult to generate survival times directly. We will also assume only one covariate, but it can be easily extended to many. Let

$$h_0(t) = h_0 + h_1 t,$$

and

$$\beta_x(t) = \chi_0 + \chi_1 t,$$

so that

$$\begin{aligned} H(t|X) &= \int_0^t (h_0(s) + \beta_x(s)X)ds \\ &= \int_0^t (h_0 + h_1 s + (\chi_0 + \chi_1 s)X)ds \\ &= (h_0 + \chi_0 X)t + \frac{h_1 + \chi_1 X}{2} t^2. \end{aligned}$$

We thus need to find H^{-1}, the inverse of the function $H(t|X) = (h_0+\chi_0 X)t+\frac{h_1+\chi_1 X}{2}t^2$, such that

$$\begin{array}{rcl} & H(H^{-1}(t|X)|X) & = t \\ \Leftrightarrow & (h_0+\chi_0 X)H^{-1}(t|X) + \frac{h_1+\chi_1 X}{2}H^{-1}(t|X)^2 & = t. \\ \Leftrightarrow & -t + (h_0+\chi_0 X)H^{-1}(t|X) + \frac{h_1+\chi_1 X}{2}H^{-1}(t|X)^2 & = 0 \end{array}$$

After some fairly simple calculations, we obtain

$$H^{-1}(t|X) = \frac{-(h_0+\chi_0 X) \pm \sqrt{(h_0+\chi_0 X)^2 + 2(h_1+\chi_1 X)t}}{h_1+\chi_1 X}, \qquad \text{(A.6)}$$

so that we have

$$\begin{array}{rcl} T & = & H^{-1}[-\log(U)|X] \\ & = & \dfrac{-(h_0+\chi_0 X) + \sqrt{(h_0+\chi_0 X)^2 - 2(h_1+\chi_1 X)\log(U)}}{h_1+\chi_1 X}, \end{array}$$

It suffices now to draw, say $n = 1000$ values u_i coming from an Uniform distribution $U(0,1)$, and compute

$$t_i = \frac{-(h_0+\chi_0 X) + \sqrt{(h_0+\chi_0 X)^2 - 2(h_1+\chi_1 X)\log(U)}}{h_1+\chi_1 X}, \text{ for all } i = 1,\ldots,n,$$

in order to obtain survival times.

If one wishes to extend this to twoadditional covariates (for example U and D) with regression parameters say $\beta_u(t) = u_0 + u_1 t$ and $\beta_d(t) = d_0 + d_1 t$ respectively, it is straightforward to extend the above to

$$t_i = \frac{-\tilde{h} + \sqrt{\tilde{h}^2 - 2\tilde{b}\log(U)}}{\tilde{b}}, \text{ for all } i = 1,\ldots,n,$$

where

$$\begin{array}{rcl} \tilde{a} & = & h_0 + \chi_0 X + u_0 U + d_0 D, \\ \tilde{b} & = & h_1 + \chi_1 X + u_1 U + d_1 D. \end{array}$$

Appendix B

Additional Tables and Figures

Table B.1: Results of 5000 estimations for the true model (model (7)), the naive model (model (8)) and the IV model (model (9)) with corr(A,Z)= 0.70, and corr(A,U)= 0.20. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 1000$.

rarity of event (in %)	model (7)	model (8)	model (9)
1.002	0.5191	0.5945	0.5702
1.533	0.521	0.596	0.526
2.027	0.512	0.585	0.510
2.520	0.512	0.584	0.505
3.008	0.512	0.582	0.498
3.491	0.510	0.581	0.496
4.042	0.508	0.577	0.489
4.506	0.506	0.574	0.486

Table B.2: Results of 5000 estimations for the true model (model (7)), the naive model (model (8)) and the IV model (model (9)) without observed confounders, with corr(A,Z)= 0.70, and corr(A,U)= 0.20. The true parameter value is $\beta_a = 0.50$. The number of observations in each simulation was $n = 10000$.

rarity of event (in %)	model (7)	model (8)	model (9)
1.869	0.5028	0.6164	0.5055
2.778	0.5055	0.6091	0.5049
3.660	0.5044	0.6058	0.5045
5.381	0.5049	0.6053	0.502
7.034	0.5026	0.6012	0.492
10.154	0.5025	0.5989	0.4874
13.081	0.5007	0.5952	0.4826
15.783	0.5022	0.5942	0.4786

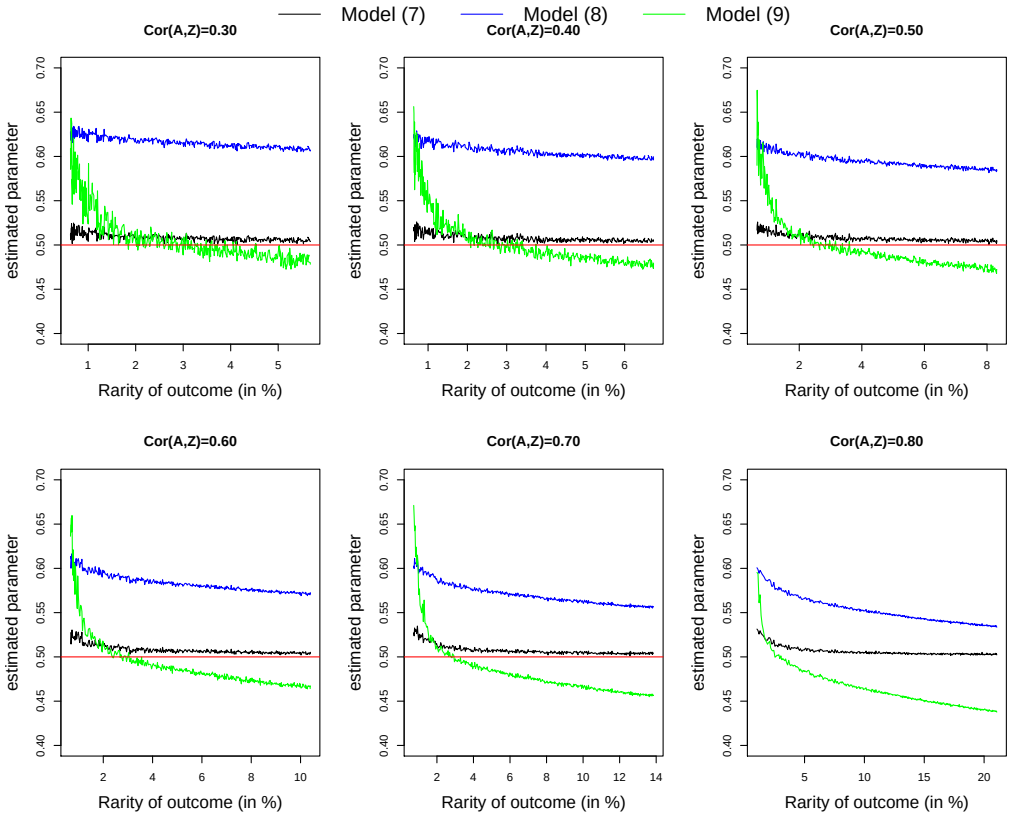


Figure B.1: Results of 5000 estimations from the true model (model (7)), the naive model (model (8)), and the IV model (model (9)), for six different IV strengths, with corr(A,U)$\approx$ 0.20, as a function of the percentage of events. The number of observations in each simulation was $n = 1000$

Appendix C

Source codes for Section 4.1, 4.2 and 4.3

C.1 Function generating the data

Here is the function generating the data as explained in the beginning of Section 4.1. It takes in:

- the number of individuals desired (N $= n$)
- the parameters for the response model (beta)
- the parameters for the exposure model (gamma)
- whether one wants to use the Lin-Ying model or the proportional hazards model (type= 1 for Lin-Ying)
- the cutoff point for the censoring scheme (eta)
- the values for the baseline hazard if one chooses Cox (lambda and rho)
- whether one wants categorical IV or continuous IV (Zcat=TRUE for categorical)
- the parameters a, b and the mean of U in (4.4)(a, b, and mu.U)

```
T.fun22 <- function(N, beta, gamma, type, eta, lambda, rho,
                    Zcat, a, b, mu.U)
{
  b0 <- beta[1]
  ba <- beta[2]
  bd <- beta[3]
  bu <- beta[4]
  c0 <- gamma[1]
  cz <- gamma[2]
  cd <- gamma[3]
```

```
  deltaU <- mvrnorm(N, c(0,mu.U), matrix(c(b,a,a,b),ncol=2))

  if (Zcat==TRUE){
    Z<- sample(x=c(1,2,3,4), size=N, replace=TRUE,
               prob=c(0.25, 0.25, 0.25, 0.25))
      } else {Z <- rnorm(N,0,1)}

  U <- deltaU[,2]
  D <- rnorm(N,0,1)
  A <- c0 + cz*Z + cd*D + deltaU[,1]

  v <- runif(n=N)
  #Using inverse transform method to get T values (Tval)
  if (type==1){
    # ADDITIVE:
    Tval <- -log(v)/(b0+ba*A+bd*D+bu*U)
  } else {
    #PROPORTIONAL:
    Tval <- (- log(v)/(lambda*exp(A*ba+bd*D+U*bu)))^(1/rho)
  }

  # censoring scheme:
  cens <-runif(n=N,min=0, max=eta)

  # times and status:
  time <- pmin(Tval, cens)
  status <- as.numeric(Tval <= cens)

  # data set:
   data.frame(id=1:N,
              time=time,
              status=status,
              A=A, U=U,Z=Z,D=D)
}
```

C.2 Function fixing cor(A,U)

This function finds approximate values of a in (4.4) for the desired correlation between A and U using the unitoot function. It takes in similar parameters as the T.fun22 function, but instead of a and b it takes in the amount of correlation desired (corAU), and an interval on which to seek the solution (inte).

```
superfun22 <- function(N, beta, gamma, type, eta, lambda, rho,
                       Zcat, corAU, inte, mu.U)
{
  find.a <- function(par)
  {
   #Solutions are  more stable by averaging, hence the loop.
    res <- rep(NA,1000)
    for (i in 1:1000)
    {
```

```
      fn<- T.fun22(N=N, beta=beta, gamma=gamma, type=type,
                   eta=eta, lambda=lambda, rho=rho, Zcat=Zcat,
                   a=par, b=inte[2], mu.U=mu.U)
      res[i] <- cor(fn$A,fn$U)
    }
    ress <- round(mean(res),4)- corAU
    ress
  }
  a.win <- uniroot(find.a, interval=inte)$root
  a.win
}
```

C.3 Function fixing the number of events

This function finds the approximate censoring cutoff point in order to obtain the desired number of events via the unitoot. It takes in similar parameters as the T.fun22 function, but instead of b and eta it takes in the percentage of events desired (per) and an interval on which to seek the solution (inte).

```
superfun33 <- function(N, beta, gamma, per, lambda, rho,
                       type, Zcat, a, inte, mu.U)
{
  find.eta <- function(par)
  {
   #Solutions are  more stable by averaging, hence the loop.
    res <- rep(NA,1000)
    for (i in 1:1000)
    {
      fn<- T.fun22(N=N, beta=beta, gamma=gamma, type=type,
                   eta=par, lambda=lambda, rho=rho, Zcat=Zcat,
                   a=a, b=inte[2], mu.U=mu.U)
      res[i] <- sum(fn$status)
    }
    round(mean(res),3)- N*per/100
  }
  eta.win <- uniroot(find.eta, interval=inte)$root
  eta.win
}
```

Appendix D

Simulation codes for Section 4.1, 4.2 and 4.3

These codes use the three source codes presented earlier, that is, T.fun22, superfun22 and superfun33.

D.1 For Section 4.1.1 and 4.1.2

```
#### ======================================================== ####
#### ============== IV - Additive hazard model ============= ####
#### ================= Constant parameters  ================ ####
#### ======================================================== ####

### Defining all the different parameters needed ================

### Default eta value (deciding the censoring cutoff point)
etadum <- 0.15
### Value for the mean of U:
mu.U <- 0
### Number of observations wanted:
N<- 10^3
### Decide if Z categorical (TRUE) of normal (FALSE)
zcat <- TRUE
### Decide if Aalen (1) or Cox (2)
type=1
### Values for the Weibull distribution
lambda<-0.1 ;rho=1
### Decide what we want for the correlation between A and U
my.corAU=0.4
### The percentage of outcome wanted
my.per <- 25
### Value on the diagonal of covariance matrix of U and delta:
bb <-1 #(when using higher corAU, raise this)
        # 0.2 and 0.4 is ok with 1, for 0.6, use 3
### Interval for the superfun22 function:
intep <- c(0.00000000001,bb)
```

```
### Parameters for the exposure model and the response model:

### Values for  b_0, b_a, b_d, b_u respectively (response)
beta<-c(4, 0.5,   0.5,  0.5)
### Values for c_z vector (deciding for the IV strength)
c_z <- seq(0.2,1.8, by=0.016)
ng <- length(c_z)
### Values for c_0,  c_z,  c_d  respectively (exposure model)
gammamat <- as.matrix(cbind(rep(3.5, ng), c_z ,rep(1, ng)))

### Getting "a"'s: ============================================

AUadd <- foreach(i=1:ng,.combine = c) %dopar% {
  superfun22(N=N, beta=beta, gamma=gammamat[i,],
             corAU=my.corAU, lambda=lambda, rho=rho, type=type,
             Zcat=zcat, eta=etadum, inte=intep, mu.U=mu.U)
}

AUadd # the resulting "a" values to be used to control
      # the correlation between A and U

## Once the "a" values are found, we need to find values "eta"
## for controlling the number of event to a certain percentage

### Getting the corresponding "eta"'s: ==========================

etaadd <- foreach(i=1:ng,.combine = c) %dopar% {
  superfun33(N=N, beta=beta, gamma= gammamat[i,], per=my.per,
             lambda=lambda, rho=rho, type=type, Zcat=zcat,
             a=AUadd[i], inte = intep, mu.U=mu.U)
}

etaadd # the resulting "eta" values to be used to obtain
        # a chosen amount of outcome

### SIMULATIONS - Additive hazard - Constant parameters =========

sim <- 5000
addrescorplot <- matrix(NA,ncol=15,nrow=ng)

for (g in 1:ng)
{
  qwe <- foreach(i=1:sim,.combine = rbind) %dopar% {

    dat <- T.fun22(N=N, beta=beta, gamma=gammamat[g,], type=type,
                   eta=etaadd[g], lambda=lambda, rho=rho, Zcat=zcat,
                   a=AUadd[g], b=bb, mu.U=mu.U)

    corAZ <- cor(dat$A,dat$Z)
    corAU <- cor(dat$A,dat$U)
    sums  <- sum(dat$status)

    #This assures that no negative hazards are generated.
```

```
    while (length(which(dat$time<0))>0)
    {
      dat <- T.fun22(N=N, beta=beta, gamma=gammamat[g,],
                     type=type, eta=etaadd[g], lambda=lambda, rho=rho,
                     Zcat=zcat, a=AUadd[g], b=bb, mu.U=mu.U)

      corAZ <- cor(dat$A,dat$Z)
      corAU <- cor(dat$A,dat$U)
      sums  <- sum(dat$status)
    }

    fit.true <- aalen(Surv(time, status)~ const(A)+ const(D)+ const(U),
                      data=dat, robust=0)
    fit.nottrue <- aalen(Surv(time, status)~ const(A)+ const(D),
                         data=dat, robust=0)

    #Now IV:
    ###
    dat$M  <- predict(lm(A ~ Z + D , data=dat))
    fit.IV <- aalen(Surv(time, status) ~ const(M) + const(D),
                    data=dat, robust=0)
    ###

    # Extracting what we need:
    betatrue  <- fit.true$gamma[1]
    betanottrue <- fit.nottrue$gamma[1]
    betaIV <- fit.IV$gamma[1]

    biastrue  <- fit.true$gamma[1] - beta[2]
    biasnottrue <- fit.nottrue$gamma[1] -beta[2]
    biasIV <- fit.IV$gamma[1] - beta[2]

    msetrue <- (fit.true$gamma[1] - beta[2])^2
    msenottrue <- (fit.nottrue$gamma[1] -beta[2])^2
    mseIV <- (fit.IV$gamma[1] - beta[2])^2

    c(corAZ, corAU, sums,
      betatrue, betanottrue, betaIV,
      biastrue, biasnottrue, biasIV,
      msetrue, msenottrue, mseIV)
  }
  addrescorplot[g,]<- c(apply(qwe,2,mean),
                        var(qwe[,4]), var(qwe[,5]), var(qwe[,6]))
}
colnames(addrescorplot) <- c("cor AZ", "cor AU", "sums",
                             "betatrue","betanottrue","betaIV",
                             "biastrue","biasnottrue","biasIV",
                             "msetrue","msenottrue","mseIV",
                             "vartrue","varnottrue","varIV")

### Bootstrapping - Comparing variances ==========================

# The "gamma", "eta", and "a" values doing :
# corAZ = 0.70
```

```
# corAU = 0.40
# 25 % of events
# are the values in position 66.

N=1000
# Now, generate new dataset, and do IV regression on it:
qwe <- foreach(b=1:5000,.combine = rbind) %dopar% {
  dat <- T.fun22(N=N, beta=beta, gamma=gammamat[66,], type=type,
                 eta=etaadd[66], lambda=lambda, rho=rho,
                 Zcat=zcat, a=AUadd[66], b=bb, mu.U=mu.U)
  while (length(which(dat$time<0))>0)
  {
    dat <- T.fun22(N=N, beta=beta, gamma=gammamat[66,], type=type,
                   eta=etaadd[66], lambda=lambda, rho=rho, Zcat=zcat,
                   a=AUadd[66], b=bb, mu.U=mu.U)
  }

  ### Now, extracting coefs and variances:

  # Variance from aalen function - IV model:
  dat$M  <- predict(lm(A ~ Z + D , data=dat))
  fit.IV <- aalen(Surv(time, status) ~ const(M) + const(D),
                  data=dat,robust=0)
  betaIV <- fit.IV$gamma[1]
  IV.var <- fit.IV$var.gamma[1,1]

  # Variance from aalen function - True model:
  fit.true <- aalen(Surv(time, status)~const(A)+const(D)+const(U),
                    data=dat, robust=0)
  betatrue <- fit.true$gamma[1]
  true.var <- fit.true$var.gamma[1,1]

  # Variance from aalen function - Naive model:
  fit.nottrue <- aalen(Surv(time, status) ~ const(A) + const(S),
                       data=dat, robust=0)
  betanottrue <- fit.nottrue$gamma[1]
  nottrue.var <- fit.nottrue$var.gamma[1,1]

  ### BOOTSTRAP FUNCTION ###
  nuke.fun <- function(data, indices)
  {
    boot.data <- data[indices, ]

    ### IV Model:  ================================================
    # First Stage:
    first <- lm(A ~ Z + D, data=boot.data)
    # Predicting A:
    boot.data$m <- predict(first , type ="response")
    # Second Stage:
    fit1 <- aalen(Surv(time, status) ~ const(m)+const(D),
                  data=boot.data, robust=0)

    ### Naive Model: ==============================================
    fit2 <- aalen(Surv(time, status)~ const(A)+const(D),
```

```
                  data=boot.data, robust=0)

    ### True model:  ===============================================
    fit3 <- aalen(Surv(time, status)~ const(A)+const(D)+const(U),
                  data=boot.data, robust=0)

    ### Returing: ==================================================
    c(fit1$gamma[1], fit2$gamma[1], fit3$gamma[1])
  }

  nuke.boot <- boot(dat, nuke.fun, R=500)

  res.var <- apply(nuke.boot$t, 2, var)
  IV.bootvar    <- res.var[1]
  naive.bootvar <- res.var[2]
  true.bootvar  <- res.var[3]

  c(IV.var, true.var, nottrue.var,
    IV.bootvar, true.bootvar, naive.bootvar,
    betaIV, betatrue, betanottrue,
    sums, corAZ, corAU)
}

my.boot.res <- t(matrix(c(mean(qwe[,1]), mean(qwe[,2]),
                          mean(qwe[,3]),
                          mean(qwe[,4]), mean(qwe[,5]),
                          mean(qwe[,6]),
                          mean(qwe[,7]), mean(qwe[,8]),
                          mean(qwe[,9]),
                          var(qwe[,7]), var(qwe[,8]),
                          var(qwe[,9]), nrow=3))

my.boot.assump <- c(mean(qwe[,10])*100/N,
                    mean(qwe[,11]),mean(qwe[,12]))

colnames(my.boot.res) <- c("IV", "True", "Naive")
rownames(my.boot.res) <- c("aalen variance","bootstrap variance",
                           "empirical variance","parameter estimates")

### PLOTTING - COR PLOT ========================================

### Example:
dev.new()
par(mfrow=c(1,3))
plot(addrescorplot[,1],addrescorplot[,4],
     xlab="correlation between A and Z",
     ylab="estimated parameter",
     ylim= c(0,1),lty=1,type="l",
     cex.lab=1.5)
lines(addrescorplot[,1],addrescorplot[,5],col="blue")
lines(addrescorplot[,1],addrescorplot[,6],col="green")
abline(h=0.5, col="red")
### (Legend code not displayed)
```

```
### PLOTTING - CI PLOT ==========================================

### Calculatinc 95 CI:
truesd.a    <- sqrt(addrescorplot[,13])
nottruesd.a <- sqrt(addrescorplot[,14])
Ivsd.a      <- sqrt(addrescorplot[,15])

truemean.a    <- addrescorplot[,4]
nottruemean.a <- addrescorplot[,5]
Ivmean.a      <- addrescorplot[,6]

trueci.L.a <-  truemean.a- 1.96*truesd.a
trueci.U.a <-  truemean.a+ 1.96*truesd.a

nottrueci.L.a <-   nottruemean.a- 1.96*nottruesd.a
nottrueci.U.a <-   nottruemean.a+ 1.96*nottruesd.a

Ivci.L.a <-  Ivmean.a- 1.96*Ivsd.a
Ivci.U.a <-  Ivmean.a+ 1.96*Ivsd.a

## Example:
par(mfrow=c(1,1))
plot(addrescorplot[,1],addrescorplot[,4],
     xlab="correlation between A and Z",
     ylab="estimated parameter",
     ylim= c(-1.2,2.2),lty=1,type="l",
     cex.lab=1.5,,lwd=1.4)#ylim= c(-1,3),lty=1)
lines(addrescorplot[,1],trueci.L.a,lty=2,lwd=1.5)
lines(addrescorplot[,1],trueci.U.a,lty=2,lwd=1.5)
lines(addrescorplot[,1],addrescorplot[,5],col="blue",lwd=1.4)
lines(addrescorplot[,1],nottrueci.L.a,col="blue",lty=2,lwd=1.5)
lines(addrescorplot[,1],nottrueci.U.a,col="blue",lty=2,lwd=1.5)
lines(addrescorplot[,1],addrescorplot[,6],col="green",lwd=1.4)
lines(addrescorplot[,1],Ivci.L.a,col="green",lty=2,lwd=1.5)
lines(addrescorplot[,1],Ivci.U.a,col="green",lty=2,lwd=1.5)
abline(h=0.5, lty=1, col="red")
### (Legend code not displayed)

### PLOTTING - B_a(t) ===========================================
N=1000
dat <- T.fun22(N=N, beta=beta, gamma=gammamat[66,], type=type,
               eta=etaadd[66], lambda=lambda, rho=rho, Zcat=zcat,
               a=AUadd[66], b=bb, mu.U=mu.U)

# Making the plot
dat$M  <- predict(lm(A ~ Z + D , data=dat))
fit.IV.t <- aalen(Surv(time, status) ~ M + const(D),
                  data=dat,robust=0)

par(mfrow=c(1,2))
plot(fit.IV.t, specific.comps = c(1,2), pointwise.ci = 3,
     xlab = "time")
```

D.2 For Section 4.1.3

```
#### ======================================================== ####
#### ================ Additive hazard model ================ ####
#### ============== Time varying parameters  ============== ####
#### ======================================================== ####


### Defining all the different parameters needed ================
### Number of observations:
N<- 10^3
### function to be used (already the cumulative version):
### using the following function: b(t) = a + gt
b.func <- function(a,g,t){b.t <-a+g*t  ;b.t}
### parameters for exposure
c_z <- seq(0.2,1.8, by=0.016)
ng <- length(c_z)
gammamat <- as.matrix(cbind(rep(3.5, ng), c_z ,rep(1, ng)))
### parameter for censoring
eta=0.24
### Parameter for corr(A,U)
my.cor = 0.40

### Parameters for the exposure model and the response model :

### Parameters of regression function b(t)

#            b0   ba   bd   bu
alpha.s <- c( 1, 1.5, 0.5, 0.6)
gamma.s <- c(10, 1.5, 0.5, 0.6)

B.func <- function(a,g,t){B.t <-a*t+(g/2)*(t^2) ;B.t}
bet.1 <-  B.func(1.5,1.5,0.1)    #0.1575
bet.2 <-  B.func(1.5,1.5,0.2)    #0.33

### parametres  of A=c0+cz*Z +cd*D+delta
c0 <- c.vec[1]
cz <- c.vec[2]
cd <- c.vec[3]

### Getting "a"'s: ==============================================
## can use same values as for fixed parameters as
## cor(A,U) does not depend on how the response is modeled

AUadd

### Getting value of gammamat  giving cor(A,Z)=0.70: ============
## For the same reason as before, and because we used the same
## gammamat, wecan use the same value as for the Lin-Ying model,
## that is, the value in position 66.

### SIMULATIONS - Additive hazard - Time-varying parameters =====
```

```
sim <- 5000

### Using values in row 66 of gammamat
gam.vec <- gammamat[66,]
c0 <- gam.vec[1]
cz <- gam.vec[2]
cd <- gam.vec[3]

qwe <- foreach(i=1:sim,.combine = rbind) %dopar% {
    deltaU <- mvrnorm(N, c(0,0),
         matrix(c(bb,AUadd.t[66],AUadd.t[66],bb),ncol=2))
  Z<- sample(x=c(1,2,3,4), size=N, replace=TRUE,
        prob=c(0.25, 0.25, 0.25, 0.25))

  D <- rnorm(N, 0,1)
  A <- c0 + cz*Z +cd*D+ deltaU[,1]
  U <- deltaU[,2]

  v <- runif(n=N)

  num1 <- alpha.s[1] + alpha.s[2]*A + alpha.s[3]*D + alpha.s[4]*U
  num2 <- gamma.s[1] + gamma.s[2]*A + gamma.s[3]*D+ gamma.s[4]*U

  #The while loop assures that no negative hazards are generated.
  while( length(which((num1^2-2*(num2)*log(v))<0))>0)
  {
    deltaU <- mvrnorm(N, c(0,0),
        matrix(c(bb,AUadd.t[my07],AUadd.t[my07],bb),ncol=2))

    Z<- sample(x=c(1,2,3,4), size=N, replace=TRUE,
         prob=c(0.25, 0.25, 0.25, 0.25))

    D <- rnorm(N, 0,1)
    A <- c0 + cz*Z +cd*D+ deltaU[,1]
    U <- deltaU[,2]

    v <- runif(n=N)

    num1 <- alpha.s[1]+ alpha.s[2]*A+ alpha.s[3]*D+ alpha.s[4]*U
    num2 <- gamma.s[1]+ gamma.s[2]*A+ gamma.s[3]*D+ gamma.s[4]*U
  }

  num3 <- sqrt(num1^2-2*(num2)*log(v))

  ### Using formula (4.5)
  T.add  <- (-num1+num3)/num2

  corAU <- cor(A,U)
  corAZ <- cor(A,Z)
  .
  # times and status:
  cens <-runif(n=N,min=0, max=eta)
```

```
  time   <- pmin(T.add, cens)
  status <- as.numeric(T.add <= cens)
  sums <- sum(status)

  fit.true    <- aalen(Surv(time, status) ~ A+ U+ D,robust=0)
  fit.nottrue <- aalen(Surv(time, status) ~ A+ D ,robust=0)

  #Now IV:
  ###
  M <- predict(lm(A~Z+D))
  fit.IV <- aalen(Surv(time, status) ~ M+D,robust=0)
  ###

  w.n1 <- max(which(fit.true$cum[,1]<=0.1))
  w.n2 <- max(which(fit.true$cum[,1]<=0.2 ))

  betatrue1  <- fit.true$cum[w.n1,3]
  betatrue2  <- fit.true$cum[w.n2,3]

  betanottrue1 <- fit.nottrue$cum[w.n1,3]
  betanottrue2 <- fit.nottrue$cum[w.n2,3]

  betaIV1 <- fit.IV$cum[w.n1,3]
  betaIV2 <- fit.IV$cum[w.n2,3]

  msetrue1 <- (betatrue1  - bet.1)^2
  msetrue2 <- (betatrue2  - bet.2)^2

  msenottrue1 <- (betanottrue1 - bet.1)^2
  msenottrue2 <- (betanottrue2 - bet.2)^2

  mseIV1 <- (betaIV1 - bet.1)^2
  mseIV2 <- (betaIV2 - bet.2)^2

  c(corAZ,corAU, sums,
    betatrue1, betatrue2,
    betanottrue1, betanottrue2,
    betaIV1,betaIV2,
    msetrue1,msetrue2,
    msenottrue1,msenottrue2,
    mseIV1,mseIV2)
}

### Interested in values of A at t=0.1 and t=0.2
### The true values are :
bet.1  #0.1575
bet.2  #0.33

qwe.vec <-c(apply(qwe[,1:9],2,mean),
            apply(qwe[,4:9],2,var),
            apply(qwe[,10:15],2,mean))
names.vec <- c("corAZ","corAU", "sums",
               "betatrue1", "betatrue2",
               "betanottrue1", "betanottrue2",
```

```
                "betaIV1","betaIV2",
                "var betatrue1", "var betatrue2",
                "var betanottrue1", "var betanottrue2",
                "var betaIV1","var betaIV2",
                "msetrue1","msetrue2",
                "msenottrue1","msenottrue2",
                "mseIV1","mseIV2")
cbind(names.vec,round(qwe.vec,5))


### B_a(t) plot: ================================================
### (obtained similarly to the one for Lin-Ying)
```

D.3 For Section 4.2

```
#### ======================================================== ####
#### ============== Proportional hazards model ============= ####
#### ==================== Fixed cor(A,U) ================== ####
#### ======================================================== ####

### Defining all the different parameters needed ================
etadum <- 0.15
bb <-1
mu.U <- 0
N<- 10^4
zcat <- TRUE
type=2
lambda<-0.1 ;rho=1
intep <- c(0.00000000001,bb)
my.corAU=0.4
### censoring scheme:
my.rar <- seq(0.001,0.4,length=400)
mr <- length(my.rar)

### Parameters for the exposure model and the response model:

beta<-c(0.1, 0.5,   0.5,  0.5)
c_z <- seq(0.05,1.9, by=0.016)
ng <- length(c_z)
gammamat <- as.matrix(cbind(rep(0.01, ng), c_z ,rep(0.8, ng)))


### GETTING "a"s for fixed corAU ================================
### (code not displayed - similar to Lin-Ying)
AUprop

### Finding the IV strength for which to plot ===================

## Once i have found the "a" values for when corAU fixed,
## I choose an appropriate index in the gamma vector, so that
## the corresponding gamma values give a wanted correlation
## between A and Z (choose a range of correlation values)
```

```
## I then use the value of the AUprop vector at the same index,
## and make the eta vector vary in order to get
## different types of rarity:

#(code not displayed)

# Using six different values for the correlation between A and Z:
my.val <- c(20,29,40,52,68,94)
my.gam <- gammamat[my.val,]
my.a   <- AUprop[my.val]
list.length1 <- length(my.a)

#### ========================================================== ####
#### ==================== Simulations ======================= ####
#### === Fixing corAU to see if rarity depends on corAZ ==== ####
#### ========================================================== ####

new.rar <- my.rar[10:mr]
nr <- length(new.rar)
sim <- 5000
finallist1 <- list(NULL)
resrarplot <- matrix(NA,ncol=18,nrow=nr)

for (l in 1:list.length1)
{
  for (g in 1:nr)
  {
    qwe <- foreach(i=1:sim,.combine = rbind) %dopar% {
      dat <- T.fun22(N=N, beta=beta, gamma=my.gam[l,],
                     type=type, eta=new.rar[g], lambda=lambda,
                     rho=rho, Zcat = zcat, a=my.a[l], b=bb,
                     mu.U=mu.U)

      corAZ <- cor(dat$A,dat$Z)
      corAU <- cor(dat$A,dat$U)
      sums  <- sum(dat$status)

      fit.true    <- coxph(Surv(time, status)~ A+ D+ U, data=dat)
      fit.nottrue <- coxph(Surv(time, status)~ A+ D, data=dat)

      #Now IV:
      ###
      dat$M  <- predict(lm(A~Z+D , data=dat))
      fit.IV <- coxph(Surv(time, status) ~ M + D ,data=dat)
      ###

      #[...] Retrieving estimates not displayed

      c(corAZ, corAU, sums,
            betatrue, betanottrue, betaIV,
            biastrue, biasnottrue, biasIV,
            msetrue, msenottrue, mseIV)
    }
    resrarplot[g,]<- c(apply(qwe,2,mean),
```

```
                      var(qwe[,4]), var(qwe[,5]) ,var(qwe[,6]),
                      var(qwe[,1]), var(qwe[,2]), var(qwe[,3]))
  }
  finallist1[[l]] <- resrarplot
}


#### ======================================================== ####
#### ================== Simulations ======================== ####
#### === Fixing corAZ to see if rarity depends on corAU ==== ####
#### ======================================================== ####

# In order to hold corAZ fixed and vary corAU to see if more
# confounding (unknown) implies the need for rare outcome,
# need to use  use corAU=0.1,0.2,0.3.0.4

# retrieving values of "a" using superfun22 (code not displayed)

for (l in 1:list.length2)
{
  start <- proc.time()

  for (g in 1:nr)
  {
    qwe <- foreach(i=1:sim,.combine = rbind) %dopar% {

      dat <- T.fun22(N=N, beta=beta, gamma=my.gam[5,], type=type,
                     eta=new.rar[g], lambda=lambda, rho=rho,
                     Zcat=zcat, a=AUprop.vals[l], b=bb,mu.U = mu.U)

      corAZ <- cor(dat$A,dat$Z)
      corAU <- cor(dat$A,dat$U)
      sums  <- sum(dat$status)

      fit.true    <- coxph(Surv(time, status) ~ A + D + U,
                           data=dat)
      fit.nottrue <- coxph(Surv(time, status) ~ A + D,
                           data=dat)
      #Now IV:
      ###
      dat$M  <- predict(lm(A~Z+D , data=dat))
      fit.IV <- coxph(Surv(time, status) ~ M + D ,data=dat)
      ###

      #[...] Retrieving estimates not displayed

      c(corAZ, corAU, sums,
         betatrue, betanottrue, betaIV,
        biastrue, biasnottrue, biasIV,
        msetrue, msenottrue, mseIV)
    }
    resrarplot[g,]<- c(apply(qwe,2,mean),
                       var(qwe[,4]),var(qwe[,5]),var(qwe[,6]),
                       var(qwe[,1]),var(qwe[,2]),var(qwe[,3]))
```

```
  }
   finallist2[[l]] <- resrarplot
}

### We end up with two lists: ===================================
### - finallist1 (for fixed cor(A,U))
### - finallist2 (for fixed cor(A,Z))

###  (Plotting codes not displayed. )
```

D.4 For Section 4.3

```
#### ======================================================== ####
#### ============ IV - Proportional hazard model ============ ####
#### ================== Correlation plot ================== ####
#### ======================================================== ####

### Defining all the different parameters needed ================

etadum <- 0.15
bb <-1
mu.U <- 0
N<- 10^4
zcat <- TRUE
type=2
lambda<-0.1 ;rho=1
intep <- c(0.00000000001,bb)
my.corAU=0.4
my.per <- 0.9

### PARAMETERS for the exposure model and the response model
beta<-c(0.1, 0.5,   0.5,  0.5)
c_z <- seq(0.05,1.9, by=0.016)
ng <- length(c_z)
gammamat <- as.matrix(cbind(rep(0.01, ng), c_z ,rep(0.8, ng)))


### Getting "a"s ================================================
### (code not displayed - similar to Lin-Ying)
AUprop

### Getting "eta"s ==============================================
### (code not displayed - similar to Lin-Ying)
etaprop

### NOW IV SIMULATIONS ==========================================

sim <- 5000   #maybe a bit higher
rescorplot <- matrix(NA,ncol=15,nrow=ng)

for (g in 1:ng)
{
```

```
  qwe <- foreach(i=1:sim,.combine = rbind) %dopar% {

  ### (Code not displayed, essentially the same as the one in the
  ###  previous section, when we varied number of events)

    c(corAZ, corAU, sums,
      betatrue, betanottrue, betaIV,
      biastrue, biasnottrue, biasIV,
      msetrue, msenottrue, mseIV)
  }
  rescorplot[g,]<- c(apply(qwe,2,mean),var(qwe[,4]),var(qwe[,5]),var(qwe[,6]))
}


colnames(rescorplot) <- c("cor AZ", "cor AU", "sums",
                          "betatrue","betanottrue","betaIV",
                          "biastrue","biasnottrue","biasIV",
                          "msetrue","msenottrue","mseIV",
                          "vartrue","varnottrue","varIV")

### Plotting - correlation plot ========================================
# (code not displayed)

### PLOTTING - CI PLOT ==========================================

truesd.a <- sqrt(rescorplot[,13])
nottruesd.a <- sqrt(rescorplot[,14])
Ivsd.a <- sqrt(rescorplot[,15])
truemean.a <- rescorplot[,4]
nottruemean.a <- rescorplot[,5]
Ivmean.a <- rescorplot[,6]
trueci.L.a <-  truemean.a- 1.96*truesd.a
trueci.U.a <-  truemean.a+ 1.96*truesd.a
nottrueci.L.a <-  nottruemean.a- 1.96*nottruesd.a
nottrueci.U.a <-   nottruemean.a+ 1.96*nottruesd.a
Ivci.L.a <-  Ivmean.a- 1.96*Ivsd.a
Ivci.U.a <-  Ivmean.a+ 1.96*Ivsd.a

#(plotting code not displayed)
```

Appendix E

R-codes for Chapter 5

```
### IMPORTING DATA - Not shown===================================

#### ======================================================== ####
#### === Splitting data into first child and second child == ####
#### ======== And Extracting IV and exposure =============== ####
#### ======================================================== ####

### Splitting data into first child and second child ============
### Merge.mat contains the MFR data, the data from questionnaire 1
### and the data from the questionnaire when the child is 8 years.

my.names <- names(table(Merge.mat$M_ID_1992))
                  [table(Merge.mat$M_ID_1992)==2] F

my.new.mat <- matrix(NA,ncol=2,nrow=length(my.names))
for (i in 1:length(my.names))
{
  my.new.mat[i,] <- which(Merge.mat$M_ID_1992==my.names[i])
  cat("iteration:", i, " \n")
}
my.mat <- matrix(NA,ncol=2,nrow=length(my.names))

for ( i in 1:length(my.names))
{
  ind <- my.new.mat[i,]
  low <- ind[1]
  up  <- ind[2]
  year.low <- Merge.mat$FAAR[low]
  year.up  <- Merge.mat$FAAR[up]
  year.low;year.up
  if (year.low<year.up) {my.mat[i,] <- c(low,up)
  } else if (year.low>year.up) {my.mat[i,] <- c(up,low)
  } else {my.mat[i,] <- c(NA,NA)}
}

my.mat <- my.mat[-which(is.na(my.mat[,1])),]
```

```
first.ch  <- Merge.mat[my.mat[,1],]
second.ch <- Merge.mat[my.mat[,2],]

### Extracting IV: BMI 8 years 1st child=========================

first.8y.wei <- first.ch$NN25
first.8y.hei <- first.ch$NN24
first.8y.BMI <- first.8y.wei/((first.8y.hei/100)^2)

par(mfrow=c(1,2))
hist((first.8y.BMI),100)
hist(log(first.8y.BMI),100)
#slightly skewed, perhaps log transform is a good idea.

### Extracting Exposure: Mother BMI prior second preg ===========
## (We used other data to fill out the NAs.)

#1 Height
mother.wei <- second.ch$AA85

#2 Weight
mother.hei <- first.ch$AA87

#3 BMI
mother.BMI <- mother.wei/((mother.hei/100)^2)


### Finding and removing extreme values ==========================

hist(mother.BMI,100)
mother.BMI[which(mother.BMI>60)] <-NA
mother.BMI[which(mother.BMI<7)]  <-NA
length(which(is.na(mother.BMI))) #43 NAs left

### Checking correlations and IV strength/significance ==========

nas <- c(which(is.na(mother.BMI)),
         which(is.na(first.8y.BMI)))

cor(mother.BMI[-nas],first.8y.BMI[-nas])           # 0.2826
cor(mother.BMI[-nas],(first.8y.BMI^2)[-nas])       # 0.2895
cor(mother.BMI[-nas],(first.8y.BMI^3)[-nas])       # 0.2929

#### Models without confounders:

my.first.stage <- data.frame(A=mother.BMI, Z=first.8y.BMI)

mod1 <- lm(A ~ Z, data=my.first.stage)
mod2 <- lm(log(A) ~ log(Z), data=my.first.stage)
mod3 <- lm(A ~ Z + I(Z^2), data=my.first.stage)
mod4 <- lm(A ~ Z + I(Z^2) + I(Z^3), data=my.first.stage)
mod5 <- lm(log(A) ~ log(Z) + I(log(Z)^2) + I(log(Z)^3),
           data=my.first.stage)
```

```
mod6 <- lm(log(A) ~ Z + I(Z^2) + I(Z^3), data=my.first.stage)
mod7 <- lm(log(A) ~ Z, data=my.first.stage)
mod8 <- lm(log(A) ~ Z+ I(Z^2), data=my.first.stage)
mod9 <- lm(log(A) ~ Z+ I(Z^2) + I(Z^3), data=my.first.stage)

summary(mod3)$adj.r.sq # 0.0856  #highest one

mod3 <- lm(A ~ Z + I(Z^2), data=my.first.stage)
drop1(mod3,test="Chisq") # shows all good
anova(mod1, mod3)
anova(mod3, mod4)
anova(mod1, mod4)
summary(mod3)

#### ======================================================== ####
#### ========== Extracting some confounders =============== ####
#### ======================================================== ####
#e.g.

# D1 : mother age at second.child
# D2 : helseregion
# D3 : gender of first child
# D4 : martial status
# D5 : education

# Missing values were replaced using other questionnaires
# (codes not displayed)
### D1: mother age ============================================
age.mother <- second.ch$MORS_ALDER
hist(age.mother)

### D2: helseregion ===========================================
hel.mother <- second.ch$HELSEREGION
table(hel.mother)
levels(hel.mother) <- c(1,2,3,4)
# Soer/Oest ------------------------------------=1
# Vest -----------------------------------------=2
# Midt -----------------------------------------=3
# Nord -----------------------------------------=4
table(hel.mother)

### D3: gender first child ====================================
gen.first.ch <- first.ch$KJONN
table(gen.first.ch)
levels(gen.first.ch) <- c(0,1,NA)  #boy=0 girl=1

### D4: civil status ==========================================
civ.mother <- second.ch$AA1123
table(civ.mother)
civ.mother[which(civ.mother=="Mer enn 1 kryss")]=NA
levels(civ.mother) <- c(NA,1,2,3,NA,2,2)
# Gift -----------------------------------------=1
# Samboer --------------------------------------=2
# Singel ---------------------------------------=3
```

```
table(civ.mother)

### D5: education================================================
education <- first.ch$AA1124
education <- relevel(education,ref=5)
# 9-aarig grunnskole ------------------------------=1
# 1-2-aarig videregaaende -------------------------=2
# Videregaaende yrkesfaglig  ----------------------=3
# 3-aarig videregaaende allmennfaglig -------------=4
# Distriktshoeyskole, universitet inntil 4 aar ----=5  -> refer.
# Universitet, hoeyskole, mer enn 4 aar -----------=6


#### ========================================================= ####
#### ============== Trying different models ================ ####
#### ========================================================= ####

my.IV.frame <- data.frame(A  = mother.BMI,
                          Z  = first.8y.BMI,
                          D1 = age.mother,
                          D2 = hel.mother,
                          D3 = gen.first.ch,
                          D4 = civ.mother,
                          D5 = education)


### A ~ Z+ Z^2+ D1+ D2+ D5+ D3+ D4 ==============================
first.stage1 <- lm(A ~ Z+ I(Z^2)+  D5+ D2+ D4+ D1+ D3,
                   data = my.IV.frame)
first.stage1 <- lm(A ~ Z+ I(Z^2)+  D1+ D2+ D3+ D4+ D5,
                   data = my.IV.frame)
length(my.IV.frame$A)
summary(first.stage1)  #312  missing / adj.r.sq = 0.114
drop1(first.stage1,test="Chisq")
# Confounder 1, 2, 4, 5
# D3 should be dropped

### A ~ Z+ Z^2+ D1+ D2+ D4+ D5 ==================================
first.stage2 <- lm(A ~ Z+ I(Z^2)+ D1+ D2+ D4+ D5,
                   data = my.IV.frame)
summary(first.stage2)  #312 missing / adj.r.sq = 0.1141
drop1(first.stage2,test="Chisq")

AIC(first.stage2)  #lowest

### FINAL FIRST STAGE ===========================================
stage.1 <- lm(A ~ Z+ I(Z^2)+ D1+ D2+ D4+ D5,
              data = my.IV.frame,
              na.action = na.exclude)
summary(stage.1)
my.IV.frame$M     <- predict(stage.1, data= my.IV.frame)
```

```
#### ========================================================== ####
#### ============ Now, Making the second stage ============= ####
#### ========================================================== ####

#Can try 4 different things here:

# 1 - total time to birth in days
# 2 - total time to birth in weeks
# 3 - time (in days)to birth for premature children
# 4 - time (in weeks)to birth for premature children


### 1- total time to birth in days ==============================

my.IV.frame$time   <- second.ch$SVLEN_DG
time.na <- which(is.na(second.ch$SVLEN_DG))
# (used other questionnaires to fill the NAs - not displayed)
summary(my.IV.frame$time)

### Kaplan - Meier plot:
surv1<-survfit(Surv(time,status,type="right")~1,
               type="kaplan-meier",data = my.IV.frame)
par(mfrow=c(1,1))
plot(surv1, mark.time=F,conf.int=TRUE,xlim=c(150,311),
     main="Estimated survival curve (Kaplan-Meier)",
     xlab="Days pregnant",ylab="Survival function")

### IV Aalen model:
my.aalen1 <-  aalen(Surv(time, status) ~ M + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)
my.aalen1.obs <-  aalen(Surv(time, status) ~ A + D1+ D2+ D4+ D5,
                        data = my.IV.frame,
                        robust=0)

dev.new()
par(mfrow=c(2,1))
plot(my.aalen1.obs,
     specific.comps=c(1,2),
     start.time=200,
     pointwise.ci=2,
     xlab = "Pregnancy length (in days)")


### 2- total time to birth in weeks ============================

my.IV.frame$time2   <- second.ch$SVLEN_UL
# (used other questionnaires to fill the NAs - not displayed)
my.IV.frame$status2 <- rep(1,length(second.ch$SVLEN_UL))

### IV Aalen model:
my.aalen2 <-  aalen(Surv(time2, status2) ~ M + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)
```

```
my.aalen2.obs <-  aalen(Surv(time2, status2) ~ A + D1+ D2+ D4+ D5,
                        data = my.IV.frame,
                        robust=0)

### 3- time (in days) to premature birth ========================

my.IV.frame$time3   <- my.IV.frame$time
my.IV.frame$status3 <- rep(1,length(my.IV.frame$time))
cens3 <- which(my.IV.frame$time3>260)
my.IV.frame$status3[cens3] <- 0

### Kaplan - Meier plot:
surv3<-survfit(Surv(time3,status3,type="right")~1,
               type="kaplan-meier",data = my.IV.frame)
par(mfrow=c(1,1))
plot(surv3, mark.time=F,conf.int=TRUE,xlim=c(150,270),ymin=0.9,
     main="Estimated survival curve (Kaplan-Meier)",
     xlab="Days pregnant",ylab="Survival function")

### IV Aalen model:
my.aalen3 <-  aalen(Surv(time3, status3) ~ M + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)

my.aalen3.obs <-  aalen(Surv(time3, status3) ~ A + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)


### 4- time (in weeks) to premature birth ========================

my.IV.frame$time4   <- my.IV.frame$time2
my.IV.frame$status4 <- rep(1,length( my.IV.frame$time2))
cens4 <- which(my.IV.frame$time4>36)
my.IV.frame$status4[cens4] <- 0

###IV Aalen model:

my.aalen4 <-  aalen(Surv(time4, status4) ~ M + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)

my.aalen4.obs <-  aalen(Surv(time4, status4) ~ A + D1+ D2+ D4+ D5,
                    data = my.IV.frame,
                    robust=0)

### other first stages for comparison with observational model ==
stage.1.124 <- lm(A ~ Z+ I(Z^2)+ D1+ D2+ D4,
                  data = my.IV.frame,
                  na.action = na.exclude)
stage.1.2 <- lm(A ~ Z+ I(Z^2)+  D2,
                data = my.IV.frame,
                na.action = na.exclude)
```

```
stage.1.1 <- lm(A ~ Z+ I(Z^2)+ D1,
                data = my.IV.frame,
                na.action = na.exclude)

my.IV.frame$M.124 <- predict(stage.1.124, data= my.IV.frame)
my.IV.frame$M.2   <- predict(stage.1.2, data= my.IV.frame)
my.IV.frame$M.1   <- predict(stage.1.1, data= my.IV.frame)

my.aalen.124 <-  aalen(Surv(time, status) ~ M.124 + D1+ D2+ D4,
                    data = my.IV.frame,
                    robust=0)
my.aalen.124.obs <-  aalen(Surv(time, status) ~ A + D1+ D2+ D4,
                        data = my.IV.frame,
                        robust=0)
my.aalen.1 <-  aalen(Surv(time, status) ~ M.1 + D1,
                    data = my.IV.frame,
                    robust=0)
my.aalen.1.obs <-  aalen(Surv(time, status) ~ A + D1,
                        data = my.IV.frame,
                        robust=0)
my.aalen.2 <-  aalen(Surv(time, status) ~ M.2+ D2,
                    data = my.IV.frame,
                    robust=0)
my.aalen.2.obs <-  aalen(Surv(time, status) ~ A + D2,
                        data = my.IV.frame,
                        robust=0)

### Plotting the comparison models (just the first plot) =======
dev.new()
par(mfrow=c(1,2))
plot(my.aalen.124.obs,
     specific.comps=c(2),
     start.time=200,
     pointwise.ci=0,
     xlab = "Pregnancy length (in days)")
#[...]

#### ========================================================== ####
#### ================== Bootstrapping ===================== ####
#### ========================================================== ####


### Making a bootstrap function ================================

nuke.fun <- function(data, indices, ti)
{
  new.zuxy<-data[indices,]

   # First stage model:=======
  first <- lm(A ~ Z+ I(Z^2)+ D1+ D2+ D4+ D5,
              data = new.zuxy,
              na.action = na.exclude)

  # Predicting A:
```

```
  new.zuxy$m<-predict(first, type="response")

  # Second stage model========
  fit1 <-  aalen(Surv(time, status) ~ m + D1 + D2 + D4 + D5,
                 data = new.zuxy,
                 robust=0)

  #Extracting coefficients of interest (at time ti)
   fit1$cum[which.min(abs(fit1$var.cum[,1]-ti)),3]
}

### Applying Bootstrap to model of interest =====================

j<-0
my.sd <- matrix(NA,ncol=3,nrow=132)
my.sd[1,] <- 0
for(j in 1:131)
{
   jj <-188 +(j-1)
  nuke.boot <- boot(my.IV.frame,  nuke.fun,ti=jj, R = 1000)
  my.sd[j+1,] <- c(jj, apply(nuke.boot$t,2, sd))
  cat("iteration:", j, " of " , 132, "\n")
}

### Plotting the CI's  ==========================================
my.t <- c(0,seq(188,308))
who <- rep(NA, 122)
for (i in 1:122)
{
  who.is <- which(my.aalen1$cum[,1]==my.t[i])
  if (length(who.is)==0) {who[i] <- NA} else {who[i] <- who.is}
}

my.est <- my.aalen1$cum[who,3]
ci.low <- my.est-1.96*my.sd
ci.up  <- my.est+1.96*my.sd

### Plotting cumulative estimate with CI ========================
dev.new()
par(mfrow=c(1,1))
plot(my.aalen1,
     specific.comps=c(2),
     start.time=200,
     pointwise.ci=0,
     xlab = "Pregnancy length (in days)")
lines(my.t, ci.low,col="green",lty=1,type="l")
lines(my.t, ci.up,col="green",lty=1,type="l")
f.ci.low <- my.aalen1$cum[who,3]-
            1.96*sqrt(my.aalen1$var.cum[who,3])
f.ci.up  <- my.aalen1$cum[who,3]+
            1.96*sqrt(my.aalen1$var.cum[who,3])
lines(my.aalen1$cum[who,1], f.ci.low,col="blue",lty=2,type="l")
lines(my.aalen1$cum[who,1], f.ci.up,col="blue",lty=2,type="l")
```

Bibliography

Aalen, O. (1980). *A Model for Nonparametric Regression Analysis of Counting Processes*, pages 1–25. Springer New York, New York, NY.

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statist. Med.*, 8(8):907–925.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Bowden, R. J. and Turkington, D. A. (1984). *Instrumental variables*. Number 8. Cambridge University Press.

Burgess, S. (2015). Commentary: Consistency and collapsibility: Are they crucial for instrumental variable analysis with a survival outcome in mendelian randomization? *Epidemiology*, 26(3):411–413.

Cawley, J. (2004). The impact of obesity on wages. *Journal of Human Resources*, XXXIX(2):451–474.

Cawley, J. and Meyerhoefer, C. (2012). The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics*, 31(1):219–230.

Clogg, C. C., Petkova, E., and Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational and Behavioral Statistics*, 17(1):51–74.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge. ISBN 0-521-57391-2.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer series in statistics Springer, Berlin.

Glymour, M. M., Tchetgen, E. J. T., and Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339.

Han, Z., Mulla, S., Beyene, J., Liao, G., and and, S. D. M. (2010). Maternal underweight and the risk of preterm birth and low birth weight: a systematic review and meta-analyses. *International Journal of Epidemiology*, 40(1):65–101.

Jackson, J. W. and Swanson, S. A. (2015). Toward a clearer portrayal of confounding bias in instrumental variable applications. *Epidemiology*, 26(4):498–504.

Jay L. Devore, K. N. B. (2011). *Modern Mathematical Statistics with Applications*. Springer-Verlag GmbH.

Johansson, S., Villamor, E., Altman, M., Bonamy, A.-K. E., Granath, F., and Cnattingius, S. (2014). Maternal overweight and obesity in early pregnancy and risk of infant mortality: a population based cohort study in Sweden. *BMJ*, 349(dec02 6):g6572–g6572.

Knight, K. (1999). *Mathematical Statistics*. CRC PR INC.

Leigh, J. P. and Schembri, M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on sf-12. *Journal of clinical epidemiology*, 57(3):284–293.

Li, J., Fine, J., and Brookhart, A. (2014). Instrumental variable additive hazards models. *Biometrics*, 71(1):122–130.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *The Annals of Statistics*, 23(5):1712–1734.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in cox and aalen regression models. *Lifetime Data Analysis*, 19(3):279.

Martinussen, T., Vansteelandt, S., Tchetgen, E. J. T., and Zucker, D. M. (2016). Instrumental variables estimation of exposure effects on a time-to-event response using structural cumulative survival models. *arXiv*. `http://arxiv.org/pdf/1608.00818v1`.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81(3):501–514.

Pearl, J. (2009). *Causality*. Cambridge University Pr.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics*. John Wiley and Sons Ltd.

Pearl, J., Robins, J. M., and Greenland, S. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Pierce, M. B. and Leon, D. A. (2005). Age at menarche and adult bmi in the aberdeen children of the 1950s cohort study. *American Journal of Clinical Nutrition*, 82(4):733–739.

Piet de Jong, G. Z. H. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

Reiersøl, O. (1945). *Confluence Analysis by Means of Instrumental Sets of Variables*. Arkiv for matematik, astronomi och fysik. Almqvist and Wiksells Boktryckeri A B.

Russell Davidson, J. G. M. (2009). *Econometric Theory & Methods*. Oxford University Press.

Scheike, T. H. and Martinussen, T. (2006). *Dynamic Regression models for survival data*. Springer, NY.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, 56(1):29–38.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Vansteelandt, S. and Didelez, V. (2015). Robustness and efficiency of covariate adjusted linear instrumental variable estimators. *arXiv*. `http://arxiv.org/pdf/1510.01770v1`.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* Econometric Analysis of Cross Section and Panel Data. MIT Press.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data.* Econometric Analysis of Cross Section and Panel Data. MIT Press.

Wright, P. G. (1928). *The tariff on animal and vegetable oils.* New York : The Macmillan company.

Zheng, C., Dai, R., Hari, P. N., and Zhang, M.-J. (2017). Instrumental variable with competing risk model. *Statistics in Medicine*, 36(8):1240–1255.