Rein, Carina:

# Identification of Mediators in High Dimensional Survival Data in the Presence of Confounding

**Masterarbeit, Sommersemester 2017**

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.41010

Master's thesis

# Identification of Mediators in High Dimensional Survival Data in the Presence of Confounding


Department of Statistics
Ludwig-Maximilians-Universität Munich

| | |
|---|---|
| Supervising Professor | Prof. Dr. Anne-Laure Boulesteix |
| Supervisor | Dr. Roman Hornung |
| Author | Carina Rein |
| Matriculation number | 10084774 |
| Submission date | 07. June 2017 |

# Contents

# List of Figures

# List of Tables

# 1 Abstract

The aim of this thesis is to develop a multivariate method for identifying mediators in high dimensional survival data and to compare this approach with a univariate mediation analysis method, which is proposed by Lange and Hansen (2011). Zhang et al. (2016) published a paper addressing a three step multivariate high dimensional mediation analysis for continuous response variables which is the basis for this thesis. However, the main difference between the method of Zhang et al. (2016) and this work is that the developed method is embedded in a survival data setting, in contrast to Zhang et al. (2016), who are working with a continuous outcome.
Building a new package for R containing both, the univariate and the multivariate approach is also a part of this thesis.
The comparison of the multivariate and the univariate method is performed with a simulation which consists of three different dependency settings with 500 simulated data sets each. After checking the simulation design, every setting was analyzed using the univariate and the developed multivariate method.

Ziel dieser Arbeit ist es, eine multivariate Methode zur Identifizierung von Mediatoren im Falle hochdimensionaler survival Daten zu entwickeln und diesen Ansatz mit einer univariaten Mediatorenanalyse, welche von Lange and Hansen (2011) entwickelt wurde, zu vergleichen.
Die Veröffentlichung von Zhang et al. (2016), welches die Basis für diese Arbeit darstellt, beschäftigt sich mit einer multivariaten hochdimensionalen Mediatorenanalyse, die aus drei Schritten besteht. Die Methode von Zhang et al. (2016) und die Methode, welche im Rahmen dieser Arbeit entwickelt wird, unterscheiden sich in erster Linie darin, dass in dieser Arbeit mit survival Daten gearbeitet wird, im Gegensatz zu Zhang et al. (2016), welche von einem stetigen outcome ausgehen.
Ein weiteres Ziel dieser Thesis ist es, ein neues Packet für R zu entwickeln, welches sowohl den univariaten als auch den multivariaten Ansatz beinhaltet.
Um die multivariate und die univariate Methode miteinander vergleichen zu können wird eine Simulation durchgeführt, welche aus drei verschiedenen Abhängigkeitsstrukturen (3 Settings) mit jeweils 500 simulierten Datensätzen besteht. Nach einer Prüfung des Simulationsdesigns wird jedes Setting sowohl mit der univariaten als auch der entwickelten multivariaten Methode analysiert.

# 2 Introduction

The aim of this thesis is to develop a multivariate method for identifying mediators in high dimensional survival data and to compare this multivariate approach with a univariate mediation analysis method. Zhang et al. (2016) published a paper addressing the multivariate mediation analysis of high dimensional survival data for continuous response variables, which is the basis of this thesis.

Before explaining some basic terms of this work and how it is structured, the main terms and definitions used all throughout this thesis are explained. The mediating variables are denoted as $M$, $X$ is called exposure and $Y$ represents the outcome variable. In this work the exposure $X$ is binary, which means taking the value 1 if the exposure is present and 0 otherwise. The mediators $M$ are continuous and the response $Y$ is considered to be survival times. A variable is called a mediating variable $M$ if it "is intermediate in the causal chain relating $X$ and $Y$" (MacKinnon, 2008, p.8). That means the exposure $X$ does have an effect on the outcome $Y$, but there also exists an effect from $X$ to the mediator $M$, and an effect from $M$ to $Y$. The structure and a detailed description of mediators is included in section 3.1.

Mediation analysis can be performed with a univariate or multivariate approach. One possibility for a univariate analysis method by Lange and Hansen (2011) is proposed in section 4.1, which analyzes each mediator separately. The approach in Zhang et al. (2016) as well as the method developed in this work is a multivariate analysis, which means that the mediation effects are not analyzed separately, but including the consideration of other variables. One advantage of multivariate regression is that it can adjust confounding variables.

A major topic of this thesis is the concept of high dimensional data. Hastie et al. (2015) put the matter of high dimensional data in a nutshell: "There is a crucial need to sort through this mass of information, and pare it down to its bare essentials. For this

process to be successful, we need to hope that the world is not as complex as it might be. For example, we hope that not all of the 30,000 or so genes in the human body are directly involved in the process that leads to the development of cancer" (Hastie et al., 2015, p.1). Data are considered as high dimensional if the amount of observed units $n$ is smaller than the number of potential influences $p$, thus $n < p$. Standard regression methods usually cannot handle such type of data and the interpretation in case of a large number of predictors would not be meaningful. Such problems can be solved by using a regularization method like Lasso or Ridge.

The main difference between the work of Zhang et al. (2016) and this thesis is that the mediator analysis developed in this work is embedded in a survival data setting, in contrast to a continuous outcome in Zhang et al. (2016). Survival data own special characteristics and are usually collected during survival studies in which individuals, who for example experienced a disease, are observed until an event of interest occurs, such as death or recurrence of the disease. Individuals "are followed from the time they experience a particular event such as the diagnosis of disease, and the time to recurrence of the disease or death is recorded" (Kirkwood and Sterne, 2003, p.225). The characteristics of survival data and possible methods for an analysis are further described in section 3.2.

Before taking a look at the ideas and methods of high dimensional mediation analysis used in this work, chapter 3 clarifies some basic concepts and essential methods needed. These concepts include an explanation of different kinds of two- and three-variable effects, including mediators in section 3.1, as well as survival data and methods for survival analysis (Cox Regression) in section 3.2. The idea of high dimensional data and penalized regression is described in section 3.3.
Since multivariate analysis adjusts for confounding variables the method developed in this thesis contain a multivariate analysis combined with the MCP technique for variable selection, which is described in section 3.4. Chapter 4 contains an approach for univariate mediator analysis by Lange and Hansen (2011) in section 4.1, a general description of multivariate mediator analysis in section 4.2, as well as the detailed description of the multivariate method developed by Zhang et al. (2016) in section 4.3. Zhang et al. (2016) use a three step analysis with a pre-selection of variables using sure independence screening (SIS), estimating and selecting variables with a minimax concave penalty regression and a joint significance test to finally identify

the mediators, The multivariate mediation analysis for high dimensional survival data developed in this thesis is described in section 4.4 and contains three steps as well. However, compared to Zhang et al. (2016) the pre-selection is not performed using SIS, some minor changes were made and the method was adapted to survival data.

Within the framework of this thesis a package for R, names *himasurv*, containing both, the univariate and the multivariate approach was built. The implementation of the univariate and multivariate methods, named *metest()* and *himasurv()* respectively, is described in section 4.5. To compare *metest()* and *himasurv()*, a simulation is performed. Section 4.6 displays the simulation design and results.

The analysis and simulation in this work were performed with R-3.4.0 for (Mac) OS X/R-Studio Version 1.0.143. The thesis itself was created with TeXstudio 2.12.4.

# 3 Essentials and methods

## 3.1 Two- and three-variable effects

Basically the main interest in analyzing data is to identify relations and how strong influences are. The simplest relation between variables is the one shown in Figures 3.1 and 3.2. The relation between $X$ and $Y$ can be asymmetric, where one variable $X$ causes another variable $Y$ (see Figure 3.1), or the relation can be symmetric, where $X$ and $Y$ cause each other (see Figure 3.2). If a third variable $Z$ is added to the system, an interpretation of possible relations between the three variables becomes more complex. Some concepts of relationships among three variables are those of confounder, covariate, mediator and moderator variables, which will be explained in the following sections.

(MacKinnon, 2008, p.6)

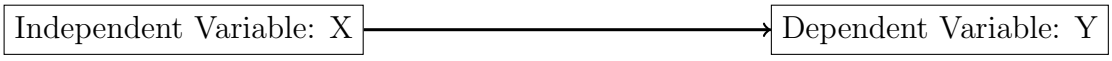


Figure 3.1: Two-variable effect: asymmetric
(based on MacKinnon (2008))

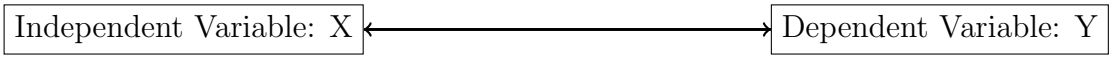


Figure 3.2: Two-variable effect: symmetric
(based on MacKinnon (2008))

The first concept is $Z$ being considered a confounding variable. This relationship is shown in Figure 3.3. Accounting for the confounder leads to a meaningfully different interpretation of the relationship between the independent variable $X$ and the dependent variable $Y$ (Aparasu and Bentley, 2015, p.180). If $Z$ is not included in the analysis this relation may wrongly be considered as a causal relationship between $X$ and $Y$. If one is interested in the effect of $X$ on $Y$ and is not considering the confounding variable $Z$, the result may be biased.

(Kirkwood and Sterne, 2003, p.179; MacKinnon, 2008, p.7)

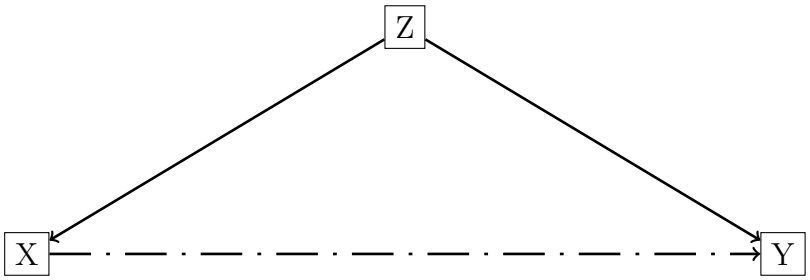


Figure 3.3: Three-variable effect: Confounder
(based on Kirkwood and Sterne (2003))

Another concept is that the third variable $Z$ being considered as a covariate, which is shown in Figure 3.4. In case of $Z$ being a covariate, the prediction of $Y$ will be more accurate considering $Z$, as it explains variability within $Y$. It is possible that covariates are related to the dependent $Y$ and independent variable $X$. However, the difference between a confounder and a covariate is that considering the "confounder leads to a meaningfully different interpretation of the relationship between the independent variable $X$ and the dependent variable $Y$" (Aparasu and Bentley, 2015, p.180) and considering a covariate does not.

(MacKinnon, 2008, p.7)

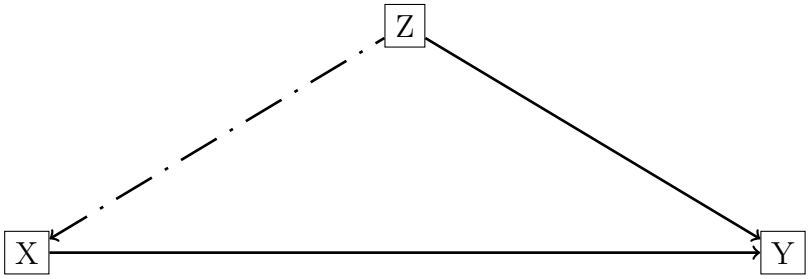


Figure 3.4: Three-variable effect: Covariate
(based on MacKinnon (2008))

The next concept within three variable effects is a moderator variable. Figure 3.5 shows such a relationship between $X$, $Z$ and $Y$. In this case $Z$ changes the sign or strength of the effect of n $X$ and $Y$ because it moderates the relation between them. Therefore the relation between $X$ and $Y$ changes depending on the level of the moderator variable. In literature (cf. Jaccard and Turrisi (2003)) the moderator effect is often referred to as the interaction effect.

(MacKinnon, 2008, p.11)

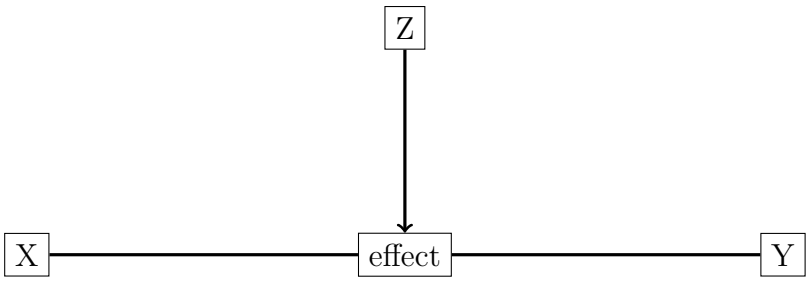


Figure 3.5: Three-variable effect: Moderator
(based on Küster-Rohde (2010))

The main interest of this thesis focuses on mediating variables. In case of mediation the third variable $Z$ "is intermediate in the causal chain relating $X$ and $Y$" (MacKinnon, 2008, p.8). That means the exposure $X$ does have a direct effect on the outcome $Y$, but there also exists an effect from $X$ to the mediator $M$, and an effect from $M$ to $Y$. This relationship is shown in Figure 3.6. Chapter 4 contains a detailed description about each path and different ways of mediator analysis.

(MacKinnon, 2008, p.8)

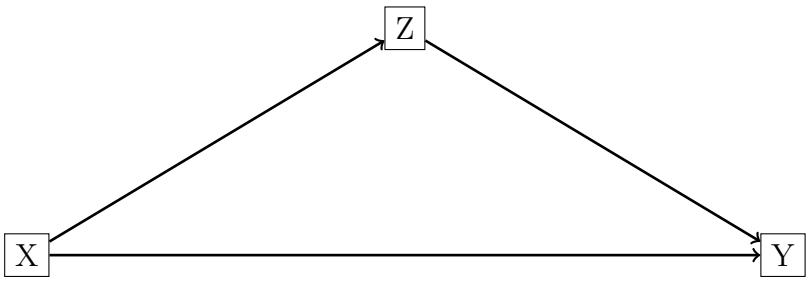


Figure 3.6: Three-variable effect: Mediator
(based on Baron and Kenny (1986))

## 3.2 Survival data and survival analysis

The first section of this chapter (3.2.1) introduces the concept of survival data, what such data looks like, what characteristics they have and when they are needed. Since survival data need particular methods for analysis, caused by their special structure, section 3.2.2 and 3.2.3 address some ideas on how to analyze survival data.

### 3.2.1 Survival data

Survival data are usually collected during survival studies in which individuals are for example "followed from the time they experience a particular event such as the diagnosis of a disease, and the time to recurrence of the disease or death is recorded" (Kirkwood and Sterne, 2003, p.225). An observation of survival time consists of a starting point, denoted as $t = 0$ and an endpoint which is reached when the event of interest occurs. The survival time is the distance between the starting point ($t = 0$) and the time a subject reaches the event of interest, for example dies from the observed disease, or the time a subject experiences "some other non-fatal, well-defined, condition such as meeting clinical criteria for remission" (Hosmer et al., 2008, p.3) of a disease. Incomplete observations due to censoring and truncation are one characteristic of survival data. An observation is considered as censored if it is incomplete due to random factors and a truncated observation is incomplete due to a selection process. Censoring can occur in different ways. The most common censored data is right censored. This type of observed survival time starts at the beginning point $t = 0$ and ends before the event of interest occurs. If the event of interest has already occurred when the observation begins, the survival time data is considered left censored. Interval censoring occurs if the particular event times are unknown, but it is known between what points of time they are located. Details about different types of censored and truncated data are explained in detail by Liu (2012). Left censoring, interval censoring and truncation are mentioned for the sake of completeness, the most common type of censoring is right censoring which can be included in the estimation of a survival model.

(Hosmer et al., 2008, p.3ff.)

### 3.2.2 Survival analysis

A descriptive data analysis should be the beginning of every statistical analysis. As survival data includes censored or truncated observations, common methods for the estimation of mean, variance etc. cannot be applied. In this case an estimated cumulative distribution provides parameter estimates.

Let $T$ be the random variable (the survival time), then the cumulative function of $T$ ($F(t)$) is the probability that a subject selected at random will have a survival time less than or equal to time $t$ and is denoted as

$$F(t) = P(T \leq t)$$

The survival function or survival curve $S(t)$ is the probability of observing a survival time greater than a time $t$:

$$S(t) = P(T > t) = 1 - F(t)$$

(Hosmer et al., 2008, p.16)

The survival curve can be estimated with the Kaplan-Meier method, which is calculated considering all risk sets of the individuals that are still in the study at each time at which an event occurs ($t$). The calculation of the survival probability at time $t$ ($s_t$), with $n_t$ being the number of individuals in the risk set and $d_t$ being the number of events that occur at exactly that time $t$, can be performed with

$$s_t = 1 - r_t = \frac{n_t - d_t}{n_t}$$

Thus the risk $r_t$ of time $t$ is equal to $\frac{d_t}{n_t}$. Based on this it is possible to estimate the survivor function via the Kaplan-Meier estimator with

$$S(t_j) = S(t_{j-1}) \times s_{t_j} = s_{t_1} \times s_{t_2} \times ... \times s_{t_j}$$

(Kirkwood and Sterne, 2003, p.277)

### 3.2.3 Regression analysis for survival data

The most common regression analysis for survival data is the Cox method, or proportional hazards regression.
Let $y_i$ be the observed survival time with $i = 1, ..., n$. The vector of predictors is denoted as $\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})$ and $\delta_i$ as the censoring indicator, with $\delta_i = 1$ if the survival time $y_i$ is the time of failure or $\delta_i = 0$ if $y_i$ is right censored. Then the Cox proportional hazards model can be written as

$$log(h(t)) = log(h_0(t)) + \beta_1 \boldsymbol{x_{i1}} + \beta_2 \boldsymbol{x_{i2}} + ... + \beta_p \boldsymbol{x_{ip}}$$

with $h(t)$ being the hazard at time $t$.

The shared baseline hazard is denoted as $h_0(t)$ and $h_i(t)$ is the hazard for patient $i$ at time $t$. $\boldsymbol{\beta}$ is a vector of the predictive effects of length $p$. Then the hazard for patient $i$ can be written as

$$h_i(t) = h_0(t) e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}$$

The assumption used by the Cox regression is called proportional hazard assumption, which means that the ratio of the hazards comparing different exposure groups remains constant over time. Considering $x_{i1}$ being a binary variable with $x_{i1} = 1$ if unit $i$ is exposed and $x_{i1} = 0$ if it is unexposed. Then the hazard ratio $HR(t)$ compares individuals who are exposed to them who are not exposed at time $t$ and can be written as

$$HR(t) = \frac{h_0(t) e^{(\beta_1)}}{h_0(t)} = e^{(\beta_1)}$$

Suppose $R_i$ being the set of indices $j$ with $y_i \geq t_i$, which includes those at risk at time $t_i$. Therefore the indices included in $R_i$ did not experience the event of interest and are uncensored right before $t_i$. The inference of the Cox model can be calculated with the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[ \frac{e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x_j}^T \boldsymbol{\beta}}} \right]^{\delta_i}$$

By maximizing this partial likelihood it is possible to estimate the parameter vector $\boldsymbol{\beta}$. This is equivalent to maximizing a log partial likelihood of the form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[ \boldsymbol{x_i}^T \boldsymbol{\beta} - log(\sum_{j \in R_i} e^{\boldsymbol{x_j}^T \boldsymbol{\beta}}) \right]$$

(Simon et al., 2011, p.1ff.; Verweij and Van Houwelingen, 1994, p.2428; Kirkwood and Sterne, 2003, p.287f.)

Further reading about survival data and survival analysis can be found in Kirkwood and Sterne (2003) and Hosmer et al. (2008).

## 3.3 High-dimensional data and penalized regression

In the age of digitalization and scientific progress, the volume of data grows every day and high-dimensional data is more common in a multitude of research areas than ever before, one example being the medical field which is increasingly delving into the world of genetics. For this reason, scientists have to handle a large number of variables and quite often with a comparably small number of observed units (e.g. observed patients). Hastie et al. (2015) refer to the assumption of simplicity, or more precisely of sparsity: "Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters (or predictors) play an important role" (Hastie et al., 2015, p.1).

Before introducing penalized regression, a way to analyze high-dimensional data, it is necessary to start explaining how estimation with linear regression works. Linear Regression is a popular and easy way to analyze data as it can be used to form a model for prediction as well as for measuring the predictor's importance. For the following section suppose $y_i$ is the outcome of unit $i$, with $i = 1, ..., n$, $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})^T$ are the $p$ predictor variables, $\beta_0$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are unknown parameters and $\epsilon_i$ is an error term. Then the linear model can be written as

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta_j} + \epsilon_i$$

The unknown parameters $\beta_0$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ can be estimated with the least squares estimator $RSS(\beta)$:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

which can then be obtained by minimizing the least squares objective function

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

The least squares estimator is a common estimator and easy to use, but it has some drawbacks. The first is the prediction accuracy because "the least squares estimates often have low bias but large variance" (Hastie et al., 2009, p.57). Using shrinkage or variable selection methods may improve the prediction accuracy by increasing the bias a bit to gain a lower variance of the predicted values.

The second drawback to consider is the case of a large $p$. It is difficult to provide a reasonable interpretation in case of a large number of predictors. Therefore it is desirable to select a smaller subset of variables with the strongest effects, as an interpretation will be easier.

At last in case of $p > n$ there will be an infinite set of solutions for the estimated parameters, which set the objective function equal to zero. Therefore, if $p > n$ the least squares estimates for $\boldsymbol{\beta}$ are not unique and will overfit the data (Bühlmann and van de Geer, 2011, p.9). Such problems can be solved by using a regularization method. In general, a regularized (or penalized) regression problem can be written as

$$\hat{\boldsymbol{\beta}} = \min_{\beta_0,\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \sum_{j=1}^{p} p(\beta_j)$$

where $p(\beta_j)$ is called penalty term and $\hat{\boldsymbol{\beta}}$ is the estimated parameter vector.

(Hastie et al., 2009, p.44; Zhang et al., 2016, p.3151; Hastie et al., 2015, p.1f.)

The most popular regularization methods are the least absolute shrinkage and selection operator (Lasso) and Ridge. In the Lasso regression the parameters are estimated by solving the following problem

$$\hat{\boldsymbol{\beta}}^{lasso} = \min_{\beta_0,\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

with the penalization parameter $\lambda$, which controls the shrinkage of the parameter $\beta_j$. The Ridge regression uses a different penalization term and minimizes the following penalized regression problem

$$\hat{\boldsymbol{\beta}}^{ridge} = \min_{\beta_0,\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso and Ridge show that regularization processes differ depending on the penalization term $p(\beta_j)$ they use. The regularization process used in this thesis is called minimax concave penalty and will be explained in chapter 3.4.

(Hastie et al., 2015, p.1f.; Tibshirani, 1996, p.268)

## 3.4 Minimax concave penalty techniques (MCP)

Some of the problems of estimation in case of high dimensional data were addressed in section 3.3. As described, these problems can be solved by using a regularization method like Lasso or Ridge. Nevertheless both methods do have drawbacks. Performing Ridge regression will not result in a variable selection, as this method shrinks the coefficients towards 0 but does not set any of them to 0 (Tibshirani, 1996, p.267). Lasso solves that problem because it performs a variable selection by setting coefficients equal to 0. However, in case of $p > n$, which means that the analysis includes more covariates than observed units, Lasso selects a maximum of $n$ variables for the model. Besides that Zhang (2010) pointed out that it is biased. Therefore a penalty is used, containing a second threshold level, like the minimax concave penalty (MCP), as proposed by Fan and Li (2001).

(Zhang et al., 2016, p.3151; Breheny and Huang, 2011, p.235f.)

The MCP term, defined on $[0, \infty)$, with the regularization parameters $\lambda > 0$ and $\gamma > 0$ which determines the concavity of MCP, is written as follows

$$p_{\lambda,\gamma}(\beta_k) = \begin{cases} \lambda\beta_k - \frac{\beta_k^2}{2\gamma} & , \text{ if } 0 \leq \beta_k < \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & , \text{ if } \beta_k \geq \gamma\lambda \end{cases}$$

The estimation of the predictors covariance matrix in section 4.4 requires the derivative of that penalty term, which is

$$p'_{\lambda,\gamma}(\beta_k) = \begin{cases} \lambda - \frac{\beta_k}{\gamma} & , \text{ if } 0 \leq \beta_k < \gamma\lambda \\ 0 & , \text{ if } \beta_k \geq \gamma\lambda \end{cases}$$

Looking at the penalty's derivative $p'_{\lambda,\gamma}(\beta_k)$ helps to understand the concept of MCP. Figure 3.7 shows the penalty terms for Lasso and MCP and Figure 3.8 displays the derivative of those penalty terms. The MCP starts with the same rate as Lasso and continuously eases the penalization until the rate drops to 0 when $\beta_k > \gamma\lambda$.

(Zhang et al., 2016, p.3151; Breheny and Huang, 2011, p.235f.)

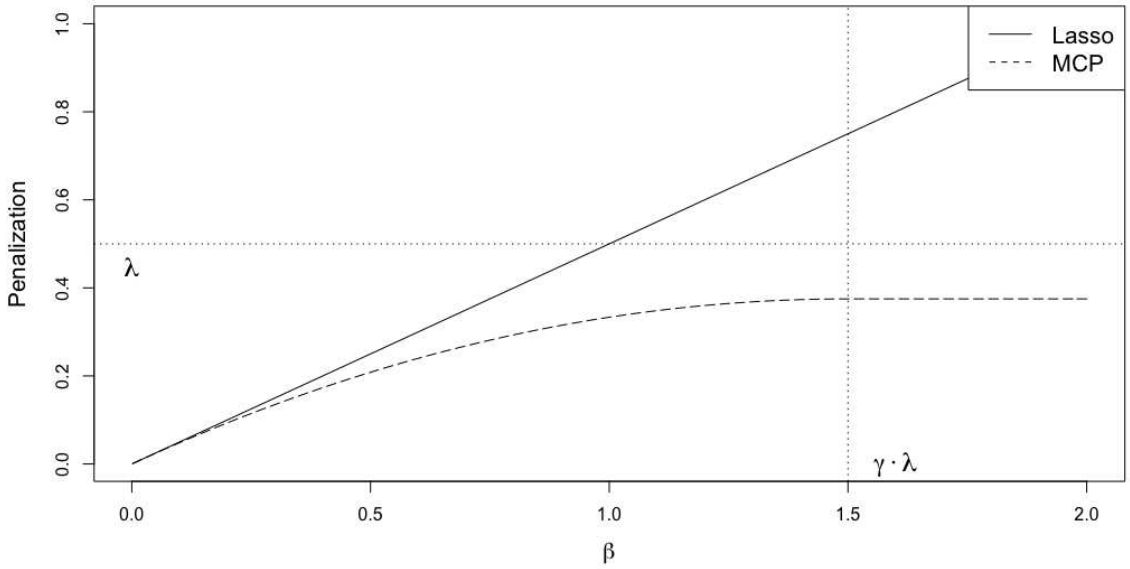


Figure 3.7: Penalization terms with $\lambda = 0.5$ and $\gamma = 3$
(based on Breheny and Huang (2011))

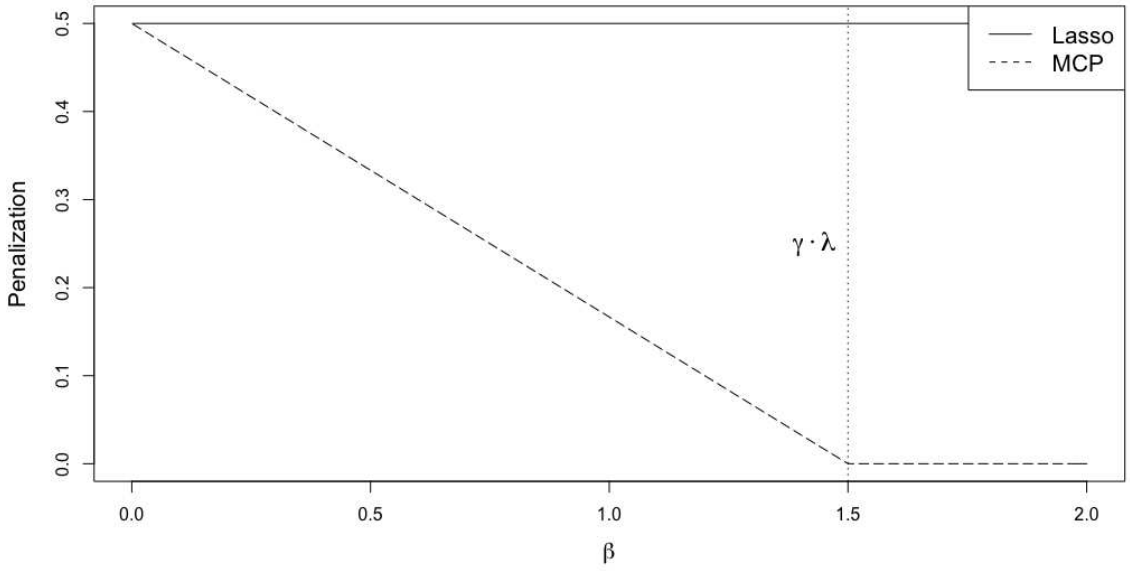


Figure 3.8: Derivative of penalization terms with $\lambda = 0.5$ and $\gamma = 3$
(based on Breheny and Huang (2011))

Since the data analyzed in this work are survival data (cf. section 3.2), the MCP method has to be adapted to the Cox proportional hazards model. As mentioned in section 3.2.3 the parameter $\boldsymbol{\beta}$ can be estimated by maximizing the log partial likelihood of the Cox proportional hazards model. Therefore the penalized log partial likelihood, with $p_{\lambda,\delta}(\beta_k)$ being the penalty term of the MCP and $l(\beta)$ being the partial likelihood, reads as follows

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - p_{\lambda,\delta}(\boldsymbol{\beta})$$

Considering the Lagrangian formulation, the vector $\boldsymbol{\beta}$ can be estimated with

$$\hat{\boldsymbol{\beta}} = argmax_{\boldsymbol{\beta}} \left[ \delta_i [\sum_{i=1}^{n} x_i^T \boldsymbol{\beta} - log(\sum_{j \in R_i} e^{x_j^T \boldsymbol{\beta}}) - p_{\lambda,\delta}(\beta)] \right]$$

The paper of Breheny and Huang (2011) covers the minimax concave penalty in detail. The function *ncvsurv()* in the package *ncvreg*, which is used in this work, uses a coordinate descent algorithm for the estimation. An explanation of that algorithm can be found in Simon et al. (2011).

(Simon et al., 2011, p.3; Verweij and Van Houwelingen, 1994, p.2428; Breheny, 2017)

### 3.4.1 The choice of $\lambda$

The recently introduced minimax concave penalty technique (MCP) (cf. section 3.4) needs a regularization parameter $\lambda$ and a tuning parameter $\gamma$ to perform a variable selection and estimation of the parameters. The package *ncvreg* contains functions performing an MCP regression for continuous, binary and survival variables. The methods for the selection of $\lambda$ offered in those functions are the information criteria Akaike's Informations Criterion ($AIC$) and the Bayesian Information Criterion ($BIC$), as well as cross-validation ($CV$). Those are common criteria for model comparison and selection. The values of the chosen criterion are calculated for each model and compared with each other. The model corresponding to the lowest value of $AIC$, $BIC$ or $CV$ is considered to be the best. Considering the MCP regression, the compared models differ in their value of the regularization parameter $\lambda$. The model with the smallest value for the chosen criterion yields the ideal value for $\lambda$, which will then be used for further analysis. The following passage addresses the calculation of $AIC$, $BIC$ and $CV$.

(Breheny, 2017)

Let $\hat{\boldsymbol{\beta}}$ be the vector of $p$ estimated parameters, $l(\hat{\boldsymbol{\beta}})$ is the corresponding log-likelihood and $n$ is the total number of observed units. Then the value for the $AIC$ can be calculated with

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2p$$

The $BIC$ can be obtained by using

$$BIC = -2l(\hat{\boldsymbol{\beta}}) + log(n)p$$

(Fahrmeir et al., 2009, p.488)

The third measurement $CV$ used for model comparison is generated using cross-validation. Verweij and Van Houwelingen (1993) displayed a cross-validation method for survival analysis and denoted the resulting value $cvl$. The $cvl$ is used as a measure of prediction accuracy because is represents how well the prediction for unit $k$ is when using the remaining observations. Before displaying how $cvl$ is calculated some notations and terms are explained first.

Suppose there are $n$ observed units, the log-likelihood is denoted by $l(\boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is the coefficient vector. With $l_{(-k)}(\boldsymbol{\beta})$ being the log-likelihood when the observed unit $k$ is left out then $l_k(\boldsymbol{\beta})$ is defined as the contribution of unit $k$ to the log-likelihood $l(\boldsymbol{\beta})$ and can be written as follows

$$l_k(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - l_{(-k)}(\boldsymbol{\beta})$$

The adaption of the method to the Cox model starts with the partial likelihood, already introduced in section 3.2.3

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[ \frac{e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x_j}^T \boldsymbol{\beta}}} \right]^{\delta_i}$$

Next, it is necessary to derive the partial likelihood $L_{(k)}(\boldsymbol{\beta})$ to calculate the partial log-likelihood $l_k(\boldsymbol{\beta})$. This can be achieved by using

$$L_k(\boldsymbol{\beta}) = \frac{L(\boldsymbol{\beta})}{L_{(-k)}(\boldsymbol{\beta})}$$

Let $w_i = e^{\boldsymbol{x_i}^T\boldsymbol{\beta}}$, then $L_{(-k)}(\boldsymbol{\beta})$ can be derived by leaving out unit $k$, which means that unit $k$ is removed from all risk sets before time $t_k$. Therefore, with $t_i < t_k$ for $i < k$, it holds that

$$L_{(-k)}(\boldsymbol{\beta}) = \prod_{i<k}(\frac{w_i}{\sum\limits_{k\in R_i} w_j - w_k})^{\delta_i} \prod_{i>k}(\frac{w_i}{\sum\limits_{j\in R_i} w_j})^{\delta_i}$$

Now it is possible to get $L_k(\boldsymbol{\beta})$, which can be interpreted as the conditional probability that unit $k$ survives until time $t_{k-1}$ or in case of $\delta_k = 1$, experiences the event of interest at time $t_k$. With $p_{ki} = \dfrac{w_k}{\sum\limits_{j\in R_i} w_j}$ being the probability that individual $k$ dies at time $t_i$, $L_k(\boldsymbol{\beta})$, can be written as

$$L_k(\boldsymbol{\beta}) = \prod_{i<k}(1 - p_{ki})^{\delta_i} p_{kk}^{\delta_i}$$

Therefore the log-likelihood is

$$l_k(\boldsymbol{\beta}) = \sum_{i<k} \delta_i log(1 - p_{ki}) + \delta_i log(p_{kk})$$

Finally the cross-validated log-likelihood $cvl$ is defined by

$$cvl = \sum_{k=1}^{n} l_k(\hat{\boldsymbol{\beta}}_{(-k)})$$

with the parameter $\hat{\boldsymbol{\beta}}_{(-k)}$ representing the value of $\boldsymbol{\beta}$ that maximizes $l_{(-k)}(\boldsymbol{\beta})$ and is called leave-one-out regression coefficient. "The determination of these coefficients involves the fitting of $n$ Cox models, each with $n - 1$ observations" (Verweij and Van Houwelingen, 1993, p.2307). Four ways for the approximation of $\hat{\boldsymbol{\beta}}_{(-k)}$ are described in Verweij and Van Houwelingen (1993).

Analogously to the criteria $AIC$ and $BIC$, the model containing the value of $\lambda$ with the smallest value of $CV$ is considered the best. Thus the $\lambda$ of this specific model will be used for further analysis.

(Verweij and Van Houwelingen, 1993, p.2306f.)

### 3.4.2 The choice of $\gamma$

However, an estimation using MCP models not only depends on the choice of $\lambda$ but also on the choice of $\gamma$. Breheny and Huang (2011) suggest an approach using a combination of AIC or BIC, cross-validation and convexity diagnostics to determine $\lambda$ as well as $\gamma$ containing the following steps:

1. Start by using a given $\gamma$ (e.g. the default 3) and select $\lambda$ using $AIC$ or $BIC$ with that given $\gamma$.

2. Look at the coefficient paths produced using MCP regression (in case of survival data with *ncvsurv()*) with the $\lambda$ which was selected in the first step. Figure 3.9 shows an exemplary coefficient path, which is created by applying *ncvsurv()* on a simulated data set, which is described in section 4.6 (setting 1 data set 1). Each colored line represents the path of one penalized coefficient $\hat{\beta}_j$ and how it changes with different values of $\lambda$. "The shaded region is the region in which the objective function is not locally convex" (Breheny and Huang, 2011, p.244). The chosen value for $\gamma$ is the one producing a coefficient path where the recently chosen $\lambda$ lies outside the shaded region (ideally near to the edge) and therefore produces a balance of sparsity and convexity.

3. Finally, use the $\gamma$ which was selected in the last step to choose the final $\lambda$ using cross-validation.

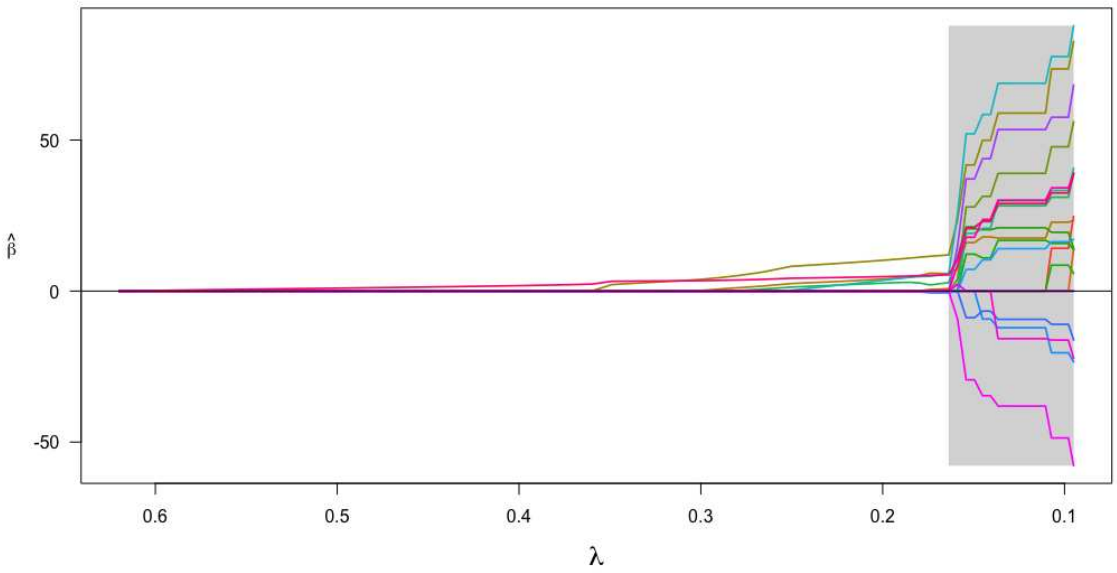


Figure 3.9: Coefficient path using *ncvsurv()* with $\gamma = 3$
(based on Breheny and Huang (2011))

The selection of $\gamma$ is just displayed for the sake of completeness as further analysis in this work use the default value $\gamma = 3$ in *ncvsurv()*. A more detailed explanation of convexity diagnostics and analysis can be found in Breheny and Huang (2011).

(Breheny and Huang, 2011, p.243ff.)

# 4 Mediator analysis

The main interest of this work is mediation analysis. "Mediation analysis plays an important role in biomedical, behavioral and psycho-social research studies, typically to understand the mechanism whereby change in one variable causes change in another" (Zhang et al., 2016, p.3150).

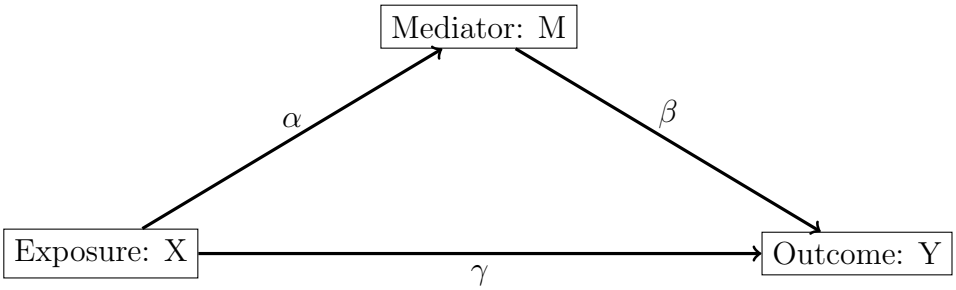


Figure 4.1: Detailed paths of mediation
(based on Baron and Kenny (1986))

A lot of different methods for mediation analysis have been developed and published and some of them are listed in Zhang et al. (2016). Mediator analysis can be explained using Figure 4.1, which shows the simplest scenario containing just one mediator (univariate). This mechanism contains different paths: $\gamma$ displays the direct impact of the independent variable (the exposure $X$) on the dependent variable (the outcome $Y$), the impact of the mediator $M$ on the outcome is $\beta$ and $\alpha$ shows the path from the exposure to $M$.
One way of univariate mediation analysis will be explained in section 4.1.

(Zhang et al., 2016, p.3151)

## 4.1 Univariate mediator analysis

Lange and Hansen (2011) proposed a univariate method for mediation analysis in a survival context. The model framework contains $T$, the survival time, either the time of interest or the time of censoring and a binary exposure $X$, equal to 1 if the exposure is present and 0 otherwise. The potential mediators are represented by $M$ and $Z$ represents other baseline covariates.

Before introducing the proposed measure or mediation it is necessary to introduce the model framework, some definitions, terms and assumptions.

"Recall that the rate at time $t$ measures the probability of experiencing an event within the next unit of time, given that a person has not experienced an event before time $t$" (Lange and Hansen, 2011, p.576). Using Cox regression it is possible to estimate how many times greater the rate is, in case the exposure $X$ is present relative to the reference $X = 0$ (hazard ratio). However, Lange and Hansen (2011) suggest using an Aalen additive hazard model for estimating the rate as the ratio modeled by Cox cannot be related to an absolute number of events. The Aalen additive hazard model though "yields an estimate of the absolute change in the rate when comparing a given" (Lange and Hansen, 2011, p.576) exposure group to the reference group. It is not assuming the hazard to be proportional and it can include time-varying covariate effects. Those are huge advantages compared to the Cox model. With $\lambda_j(t)$ being potentially time-dependent coefficient functions the Aalen additive hazard model can be written as

$$\lambda_0(t) + \lambda_1(t)x + \lambda_2(t)z + \lambda_3(t)m$$

Next, assume that the mediator $M$ is normally distributed and can be modeled by a linear regression, with $e$ being the normally distributed error with variance $\sigma^2$ and mean zero, $x$ represents the exposure and $z$ another baseline covariate. The linear model for the mediator can be written as

$$M = \alpha_0 + \alpha_1 x + \alpha_2 z + e$$

(Lange and Hansen, 2011, p.576, Abadi et al., 2011, p.3113f.)

Suppose the observations are independent, then standard techniques (like the package *timereg*) can be used to estimate the parameters $\alpha_0, \alpha_1, \alpha_2, \sigma^2$ and the functions $\lambda_0(t), ..., \lambda_3(t)$.

Lange and Hansen (2011) model causal effects and therefore define the following variables describing what would have happened if the exposure and the mediator were set to specific values. Those variables are called counterfactual variables:

- $T^{x,m}$: time to event with exposure set to $x$ and the mediator set to $m$
- $M^x$: value of mediator when exposure is set to $x$
- $T^{x,M^{x^*}}$: "event time when the exposure is set to $x$, bu the mediator is set to the value it would have had if the exposure had been set to $x^*$" (Lange and Hansen, 2011, p.576)
- $\gamma(t; x, m)$: counterfactual rate for the event in case the exposure is set to $x$ and the mediator takes the value $m$, which is the rate for the counterfactual variable $T^{x,m}$

Besides the already introduces notations and definitions, some assumptions are necessary when drawing a causal conclusion. The assumptions are defined as follows:

- A1: There are no unmeasured confounders for the exposure-outcome relationship.
- A2: There are no unmeasured confounders for the mediator-outcome relationship.
- A3: There are no unmeasured confounders for the exposure-mediator relationship.
- A4: $M^{x^*} \perp T^{x,m}|Z$: The identifiability condition, which ensures that the effect of the exposure has its effect through a distinct and a non-intertwined causal pathway. In case a variable is affected by the exposure and affects the mediator and the outcome, this assumption is violated.
- A5: The consistency assumption as shown in VanderWeele and Vansteelandt (2009), which ensures that the outcome is not affected if the exposure and mediator are set to the values they would naturally take. According to this assumption, the observed outcome $Y$ is equal to the potential outcome $Y(a)$ for subjects with an observed exposure level equal to $a$. $Y(a)$ is defined as the counterfactual outcome which would be the observed outcome if the exposure is set to $a$, for example through manipulation.

For more details about causality and causal and counterfactual effects refer to Pearl (2009) who published a book called "Causality".

(Lange and Hansen, 2011, p.576f.; VanderWeele and Vansteelandt, 2009, p.457f.)

The proposed measure of mediation from Lange and Hansen (2011) takes into account "how much of the effect of the exposure is mediated through the mediator" and "how these proportions change over time" (Lange and Hansen, 2011, p.577). This is why they suggest the counterfactual rate difference being calculated as the effect measure of the exposure change from $x$ to $x^*$.

**Theorem 1.** Given the assumptions A1-A5, "it holds that the total causal effect of changing the exposure from $x^*$ to $x$, measured on the rate difference scale at time $t$ can be expressed as" (Lange and Hansen, 2011, p.577):

$$\begin{aligned}\underbrace{\gamma(t;x,M^x)-\gamma(t;x^*,M^{x^*})}_{TE(t)} &= \gamma(t;x,M^x)-\gamma(t;x^*,M^x)+\gamma(t;x^*,M^x)-\gamma(t;x^*,M^{x^*}) \\ &= \underbrace{\lambda_1(t)(x-x^*)}_{DE(t)}+\underbrace{\lambda_3(t)\alpha_1(x-x^*)}_{IE(t)}\end{aligned}$$

with TE(t) being the total effect, DE(t) the natural direct effect and IE(t) the natural indirect effect. Thus the indirect effect corresponds to the number of deaths due to the mediator. The direct effect is the number of events caused by the direct path and the total effect represents the number of deaths, which are caused by changing the exposure $X$ and is equal to the sum of the direct effect and the indirect effect.

In case the effects in the Aalen model are not time-dependent, which means that $\lambda_1(t)$ and $\lambda_3(t)$) are both constant, Theorem 1 simplifies to

$$\underbrace{\gamma(t;x,M^x)-\gamma(t;x^*,M^{x^*})}_{TE(t)/\text{total effect}} = \underbrace{\lambda_1(x-x^*)}_{DE(t)/\text{natural direct effect}} + \underbrace{\lambda_3\alpha_1(x-x^*)}_{IE(t)/\text{natural indirect effect}}$$

The computation of the total, direct and indirect effects can be performed with standard statistical software which calculates the estimates $\hat{\lambda_1}$, $\hat{\lambda_3}$ and $\hat{\alpha_1}$. Under mild conditions, "it holds that the 3 estimators are asymptotically normally distributed and that $(\hat{\lambda_1}, \hat{\lambda_3})$ is uncorrelated with $\hat{\alpha_1}$" (Lange and Hansen, 2011, p.577). The output of the used software provides the covariance matrices for $\hat{\lambda_1}$, $\hat{\lambda_3}$ and $\hat{\alpha_1}$. Confidence intervals, and therefore p-values and tests, can be computed using the delta rule or using a simulation.

(Lange and Hansen, 2011, p.577f.)

A method for the calculation via simulation is provided by Lange and Hansen (2011) and the delta rule was implemented by Dr. Roman Hornung. The implementation in R and how the univariate mediation analysis is included in the package *himasurv* is explained in section 4.5.

A univariate perspective of mediator analysis is easy to use and the interpretation is straight forward, but if considering the different relations three variables can have (cf. section 3.1) it may not be the best way to identify and interpret mediating relationships (Wakkee et al., 2014, p.1). Therefore the following section introduces multivariate mediation analysis.

## 4.2 Multivariate mediation analysis

After displaying a way for a univariate mediation analysis, the following section will introduce the basic idea for a multivariate analysis of mediators. This theory is especially useful considering possible confounding. Figure 4.2 shows a multivariate view of mediators.

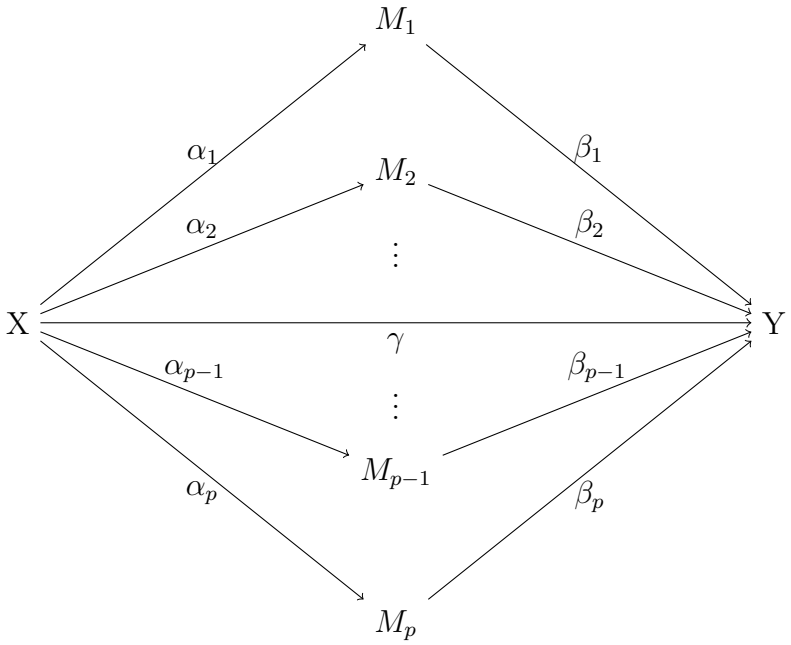


Figure 4.2: High dimensional mediators
(based on Zhang et al. (2016))

Zhang et al. (2016) consider three equations shown in 4.1, 4.2 and 4.3 for identifying mediators (see Figure 4.2). The equations include the following notations:

- $M_k$ with $k = 1, ..., p$: potential mediators
- $\gamma^*$: total effect of the independent variable $X$ on the dependent variable $Y$
- $\gamma$: parameter relating $X$ and $Y$ via the direct effect, after adjusting for all mediators of interest
- $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_p)^T$: parameter vector relating the independent variable to the mediating variables
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$: parameter vector relating the mediators to the dependent variable adjusting for the effect of the independent variable

- $c^*, c, c_k$ with $k = 1, ..., p$: intercept terms
- $\epsilon_1, \epsilon_2, \epsilon_k$ with $k = 1, ..., p$: residuals

To access the total effect $\gamma^*$ of $X$ on $Y$ regress the dependent variable on the independent variable

$$Y = c^* + \gamma^* X + \epsilon_1^* \tag{4.1}$$

Afterwards regress the mediator on the independent variable

$$M_k = c_k + \alpha_k X + \epsilon_k \tag{4.2}$$

Performing this regression for all mediators results in $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T$, the parameter vector relating $X$ to the mediators $M_k$.

Finally use the following regression for $Y$ with both, the independent variable and the mediators to receive the parameter vector containing the effects of the mediators to the dependent variable $Y$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$

$$Y = c + \gamma X + \beta_1 M_1 + ... + \beta_p M_p + \epsilon_2 \tag{4.3}$$

"These three regression equations provide the tests of the linkages of the mediation model. To establish mediation, the following conditions must hold" (Baron and Kenny, 1986, p.1177): the independent variable must affect the dependent variable (cf. equation 4.1), the independent variable must affect the mediator (cf. equation 4.2) and the mediator must have an effect on the dependent variable in a multivariate setting (cf. equation 4.3).

(Baron and Kenny, 1986, p.1176f.; Judd and Kenny, 1981; Zhang et al., 2016, p.3151)

## 4.3 Paper Zhang et al. (2016)

A lot of different methods for mediation analysis have been developed and published. Some of them are listed in Zhang et al. (2016). However none of these methods deal with high dimensional mediation in case of survival data. The idea of high-dimensional data is addressed in section 3.3, where the number of predictors, or in case of mediation analysis potential mediators, is larger than the number of observed units ($p > n$).
Zhang et al. (2016) developed a multivariate method to estimate high-dimensional mediation effects for continuous outcomes and built an R package named *hima*. Based on the three regression models described in section 4.2 (cf. equations 4.1 - 4.3), they identify mediators in three steps: Screening, MCP estimation and a joint significance test. Those steps are described in the following sections 4.3.1, 4.3.2 and 4.3.3.

### 4.3.1 Step 1: Screening

For the first step, the pre-selection of potential mediators, Zhang et al. (2016) use the sure independence screening (SIS) method based on Fan and Lv (2008). Suppose the data contains a continuous outcome $y$, an exposure $X$ and $p$ potential mediators $M_j$, with $j = 1, ..., m$. The SIS identifies the following subset:

$$\mathcal{I} = \{1 \leq s \leq p : M_s \text{ is among the top } d \text{ largest effects for the response } Y\}.$$

The value of $d$ symbolizes the amount of the top $n$ variables, and will further also be denoted as *topn*. With the outcome $y$ being continuous $d$ is equal to $[2n/log(n)]$, Zhang et al. (2016) fit one linear model for each potential mediator $M_j$ of the form

$$Y = c + \gamma_{M_j} X + \beta_{M_j} M_+ \epsilon_{M_j}$$

Afterwards the *topn* potential mediators, with the lowest p-values are selected. and are further analyzed in step 2.

(Zhang et al., 2016, p.3151; Zheng et al., 2017)

### 4.3.2 Step 2: MCP estimate

The *topn* remaining potential mediators, which were selected in step 1, and the exposure $X$ are used for step 2, the MCP estimate. This step selects and estimates the parameters $\boldsymbol{\beta_s} = (\beta_1, ...., \beta_{topn})^T$, with $s = 1, ..., topn$, corresponding to the regression:

$$Y = c + \gamma X + \beta_1 M_1 + ... + \beta_{topn} M_{topn} + \epsilon_2$$

The minimax concave penalization (MCP) method is described in chapter 3.4. The R-package *ncvreg* (Breheny, 2017) contains a function for the MCP which Zhang et al. (2016) use to compute $\{\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}, s \in \mathcal{I}\}$. However it is important to note that they defined a penalization term which excludes the exposure $X$ from the penalization.

In *hima* the regularization parameter $\lambda$ is selected via BIC (cf. section 3.4.1) and the tuning parameter $\gamma$ is set to 3, which is the default value of the function *nvcreg()*. The MCP performs a variable selection of the potential mediators as well as an estimation of the remaining effects, similar to the least absolute shrinkage and selection operator (Lasso) (cf. Tibshirani (1996) and Hastie et al. (2009)). Using the MCP procedure results in a set of selected potential mediators $M_k$ with $k = 1, ..., p$ and the corresponding estimated effects $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$.

(Zhang et al., 2016, p.3151; Zheng et al., 2017)

### 4.3.3 Step 3: Joint significance test

The last step of the mediation analysis is the joint significance test. A variable $M_k$ is considered a mediator if the effects $\alpha_k$ and $\beta_k$ are significant. The purpose of the joint significance test is to identify those mediators. The MCP has the model selection consistency "which ensures the validity of the joint significance test procedure" (Zhang et al., 2016, p.3152).

Let $S = \{k : \hat{\beta}_k \neq 0\}$ be the set of remaining potential mediators based on the MCP estimation in step 2. To assess whether the effect $\beta_k$ is significant, the hypothesis $H_0 : \beta_k = 0$ is tested, for every coefficient selected by the MCP. Suppose $k \in S$, $\Phi(\cdot)$ is the cumulative distribution of $N(0, 1)$ and the estimated standard error for $\hat{\beta}_k$ is $\hat{\sigma_{1k}}$, then the raw p-value for this test is

$$P_{raw,1k} = 2\{1 - \Phi(\frac{|\hat{\beta}_k|}{\hat{\sigma_{1k}}})\}$$

In case of testing for only one hypothesis, the p-value is considered as significant if it is lower than $\alpha = 0.05$, with $\alpha$ being the probability of rejecting the null hypothesis (Type 1 error), although it is true. Therefore the hypothesis cannot be rejected if $p < 0.05$. However, when testing for multiple hypotheses, in case all are true, the probability of at least one being wrongly rejected is higher than the desired value of $\alpha = 0.05$. Therefore, since multiple parameters are tested, it is reasonable to adjust for multiple testing. This can be done by controlling the family wise error rate ($FWER$), which is the probability of rejecting at least one true hypothesis. This can be done by using the Bonferroni method, which assures that in case of multiple testing $n$ hypotheses, the $FWER$ is not higher than 0.05 if each hypothesis is tested at the level $\frac{0.05}{n}$.

Thus the corrected p-value, with $k \in S$ and the cardinality $|S|$ (number of elements in set $S$) is

$$P_{corr,1k} = min(P_{raw,1k} \cdot |S|, 1)$$

The estimates $\hat{\beta}_k$ are already available from step 2. According to Zhang et al. (2016) $\hat{\sigma_{1k}}$, the estimated standard errors for $\beta_k$ are estimated using the oracle property of the MCP technique (cf. Fan and Li (2001) and Zhang (2010)). However, looking into function *hima()* (in package *hima*), those estimates $\hat{\sigma_{1k}}$ were calculated using a linear model of the form

$$Y = c + \gamma X + \beta_1 M_1 + ... + \beta_k M_k + \epsilon_2$$

Accordingly, the hypothesis for $\alpha_k$ is $H_0 : \alpha_k = 0$. With $k \in S$ and the estimated standard error $\hat{\sigma_{2k}}$ for $\hat{\alpha_k}$, the raw p-value is

$$P_{raw,2k} = 2\{1 - \Phi(\frac{|\hat{\alpha_k}|}{\hat{\sigma_{2k}}})\}$$

and, analogous to the Bonferroni corrected p-value for $\beta_k$, the corrected p-value for $\alpha_k$ is

$$P_{corr,2k} = min(P_{raw,2k} \cdot |S|, 1)$$

Zhang et al. (2016) estimate the values of $\alpha_k$ and the corresponding raw p-values using a linear model of the form

$$M_k = c_k + \alpha_k X + \epsilon_k$$

The corrected p-value for the joint significance test is the maximum of $P_{corr,1k}$ and $P_{corr,2k}$:

$$P_{corr,k} = max(P_{corr,1k}, P_{corr,2k})$$

If $P_{corr,k} < 0.05$ one can conclude that $M_k$ is a mediator.

(Zhang et al., 2016, p.3152; Zhang, 2010; Darlington and Hayes, 2016, p.315ff.; Zheng et al., 2017)

## 4.4 Mediator analysis in high dimensional survival data

The mediation analysis for survival data developed in this thesis is based on the work of Zhang et al. (2016). Some changes were made and the method was adapted to analyze survival data. The following sections describe the procedure in detail. The terms and notations used in this section are:

- $M_j$ with $j = 1, ..., m$: potential mediators
- $X$: binary exposure
- $\gamma^*$: total effect of the independent variable $X$ on the dependent variable $Y$
- $\gamma$: parameter relating $X$ and $Y$ via the direct effect, after adjusting for all mediators of interest
- $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_m)^T$: parameter vector relating the independent variable to the mediating variables
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T$: parameter vector relating the mediators to the dependent variable, adjusting for the effect of the independent variable

### 4.4.1 Step 1: Pre-selection

The first step of the mediator analysis is the pre-selection of the potential mediators $M_j$ with $j = 1, ..., m$. Zhang et al. (2016) use the SIS, a supervised method, for a pre-selection of those variables. However, the method developed in this thesis does not include a supervised selection method in step 1, because step 2 already includes a supervised selection process. Therefore, the pre-selection chooses the $topn = [2n/log(n)]$ potential mediators based on their variance, so the $topn$ variables with the highest empirical variances are chosen for the next step. All variables are selected, if the number of potential mediators $m$ is smaller than $topn$ which means$topn = m$. The result are the pre-selected mediators $M_s$ with $s = 1, ..., topn$.

### 4.4.2 Step 2: MCP-penalized estimate

Even though a possible pre-selection of variables was performed in step 1 of the analysis, step 2 will perform another variable selection to receive a reasonable and interpretable model. Because $p > n$ it may be considered using a Lasso regression. However, step 2 uses the MCP as proposed by Zhang et al. (2016) to perform a penalized regression and variable selection. That is not only because of the bias Zhang (2010) points out, but especially because of the drawback described in section 3.4, as the selection should not be restricted to $n$ variables when using Lasso regression. Therefore in step 2 the MCP method is used to estimate $\hat{\beta}_s$ and to perform another variable selection. The method of the minimax concave penalization (MCP) method is described in chapter 3.4.

The R-package *ncvreg* (Breheny and Huang, 2011) contains the function *ncvsurv()* for the MCP procedure using survival data which is used to compute $\hat{\beta}_s$ by minimizing the MCP criterion, analogous to Zhang et al. (2016). The regularization parameter $\lambda$ can be selected via cross-validation ($CV$), Akaike information criterion ($AIC$) or Bayesian information criterion ($BIC$). The tuning parameter $\gamma$ takes the value 3 per default.
The result of step 2 is a set of selected potential mediators $M_k$ and the corresponding estimated effects $\hat{\boldsymbol{\beta}}_k$, with $k = 1, ..., p$. However, *ncvsurv()* does not give the parameters' $\hat{\beta}_k$ variances or p-values. Fan and Li (2001) display a way to estimate a covariance matrix using the oracle property of the MCP technique. They assume that all included covariates are penalized, that is why in contrast to Zhang et al. (2016) the developed method does not exclude the exposure $X$ from the penalization in step 2.

The covariance of the estimates $\hat{\boldsymbol{\beta}}_k$ can be estimated using the sandwich formula

$$\hat{cov}(\hat{\boldsymbol{\beta}}_k) = \{\nabla^2 l(\hat{\boldsymbol{\beta}}_k) + n\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}_k)\}^{-1} \hat{cov}\{\nabla l(\hat{\boldsymbol{\beta}}_k)\} \times \{\nabla^2 l(\hat{\boldsymbol{\beta}}_k) + n\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}_k)\}^{-1}$$

with $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)^T \neq \mathbf{0}$ , being a vector of the remaining estimated effects calculated in step 2 (effects not set to zero).

For an easier display of the calculation, let further $\hat{\boldsymbol{\beta}}_k$ be $\boldsymbol{\beta}$.

$\nabla^2 l(\boldsymbol{\beta})$ is the second derivative of the Cox regression partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[ \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{\sum\limits_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}} \right]^{\delta_i}$$

The result of $l(\boldsymbol{\beta})$ is a matrix with inputs corresponding to the entries of $\boldsymbol{\beta}$ used for the derivation and is displayed as

$$\nabla^2 l(\boldsymbol{\beta}) = \frac{\delta l(\boldsymbol{\beta})}{\delta^2 \boldsymbol{\beta}} = \begin{pmatrix} \frac{\delta l(\boldsymbol{\beta})}{\delta \beta_1 \delta \beta_1} & \cdots & \frac{\delta l(\boldsymbol{\beta})}{\delta \beta_1 \delta \beta_m} \\ \vdots & \ddots & \vdots \\ \frac{\delta l(\boldsymbol{\beta})}{\delta \beta_m \delta \beta_1} & \cdots & \frac{\delta l(\boldsymbol{\beta})}{\delta \beta_m \delta \beta_m} \end{pmatrix}$$

The general representation, depending on the combination of $k$ and $n$, of the second derivative is

$$\frac{\delta l(\boldsymbol{\beta})}{\delta \beta_k \delta \beta_n} = \sum_{i=1}^{n} \delta_i \left[ - \left[ \frac{[\sum_{j \in R_i} exp(\boldsymbol{x}_j^T \boldsymbol{\beta})][\sum_{j \in R_i} x_{j,k} x_{j,n} exp(\boldsymbol{x}_j^T \boldsymbol{\beta})] - [\sum_{j \in R_i} x_{j,k} exp(\boldsymbol{x}_j^T \boldsymbol{\beta})][\sum_{j \in R_i} x_{j,n} exp(\boldsymbol{x}_j^T \boldsymbol{\beta})]}{[\sum_{j \in R_i} exp(\boldsymbol{x}_j^T \boldsymbol{\beta})]^2} \right] \right]$$

$$\cdots$$

The exact development of this second derivative (and how it is implemented in R) is displayed in the file **derivative_covariance.pdf** in the digital appendix.

The estimation of $\hat{cov}(\boldsymbol{\beta})$ also includes the term $\hat{cov}\{\nabla l(\boldsymbol{\beta})\}$, which can be written as the second derivative

$$\nabla^2 l(\boldsymbol{\beta}) = \frac{\delta l(\boldsymbol{\beta})}{\delta^2 \boldsymbol{\beta}} = \hat{cov}\{\nabla l(\boldsymbol{\beta})\}$$

The last term $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta})$ needed to estimate the covariance matrix $\hat{cov}(\boldsymbol{\beta}_k)$ is

$$\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}) = diag \left\{ \frac{p'_\lambda(|\beta_1|)}{|\beta_1|}, ..., \frac{p'_\lambda(|\beta_p|)}{|\beta_p|} \right\}$$

$$= \begin{pmatrix} \frac{p'_\lambda(|\beta_1|)}{|\beta_1|} & 0 & ... & 0 & 0 \\ 0 & \frac{p'_\lambda(|\beta_2|)}{|\beta_2|} & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & \frac{p'_\lambda(|\beta_{p-1}|)}{|\beta_{p-1}|} & 0 \\ 0 & 0 & ... & 0 & \frac{p'_\lambda(|\beta_p|)}{|\beta_p|} \end{pmatrix}$$

This procedure is implemented in the R package *himasurv* (inner function *covest()*).

(Fan and Li, 2001, p.1354)

### 4.4.3 Step 3: Joint significance test

The last step of the mediation analysis for survival data is the joint significance test, which is performed similarly to Zhang et al. (2016).

The hypothesis for $\alpha_k$ is $H_0 : \alpha_k = 0$. The values of $\alpha_k$ and the corresponding raw p-values ($P_{raw,2k}$) are calculated using a linear model for each mediator $M_k$ with $k = 1, ..., p$ of the form

$$M_k = c_k + \alpha_k X + \epsilon_k$$

The Bonferroni corrected p-values for testing $\alpha_k$ are defined as $P_{corr,2k} = min(P_{raw,2k} \cdot |S|, 1)$ (cf. Zhang et al. (2016) in section 4.3.3).

Analogously to Zhang et al. (2016) the raw p-values of $\beta_k$ are calculated using

$$P_{raw,1k} = 2\{1 - \Phi(\frac{|\hat{\beta}_k|}{\hat{\sigma_{1k}}})\}$$

with $k \in S$, $\Phi(\cdot)$ being the cumulative distribution of $N(0, 1)$ and the estimated standard error for $\hat{\beta}_k$ is $\hat{\sigma_{1k}}$, obtained by the estimate covariance matrix using the oracle property of the MCP technique (cf. Fan and Li (2001)), as displayed in section 4.4.2.

The corrected p-values for testing $\beta_k$ are $P_{corr,1k} = min(P_{raw,1k} \cdot |S|, 1)$.

Accordingly, as shown in section 4.3.3 the p-value for the joint significance test is defined as

$$P_{corr,k} = max(P_{corr,1k}, P_{corr,2k})$$

Analogous to Zhang et al. (2016), if $P_{corr,k} < 0.05$ one can conclude that $M_k$ is a mediator.

(Zhang et al., 2016, p.3152)

## 4.5 Implementation in R

Building a new package for R is part of this work. It contains the univariate mediation analysis for survival data, described in section 4.1 and the multivariate analysis for high dimensional mediators in survival data which was developed in the context of this work (cf. section 4.4).

The name *himasurv* is chosen based on the high dimensional mediation analysis for survival data. It is built using the packages *devtools* and *roxygen2* by Hadley Wickham (cf. Wickham and Chang (2017) and Wickham et al. (2017)).

Section 4.6 displays a simulation which was performed not only to verify the developed method for high dimensional mediation analysis in survival data (cf. section 4.4) but also to compare it to the univariate mediation analysis explained in section 4.1. The implementation of both methods for the R package *himasurv* is explained in the following sections.

### 4.5.1 Univariate mediator analysis

First, the implementation of the univariate mediation analysis (cf. section 4.1) will be described. This method is implemented in *metest()* by Dr. Roman Hornung using R code provided in Lange and Hansen (2011) and is included in *himasurv*. The function *metest()* provides a univariate test for mediation in a survival data setting. It is possible to analyze one or more mediators at a time and the results are p-values, one for each mediator.

One important parameter included in this function is $y$, describing the survival times and the corresponding censoring indicator which takes value 1 if the survival time is a time of failure and 0 if it is a censoring time. It is either a survival object or a matrix with two columns, with the first column representing the survival times and the second one consisting of the censoring indicator. The exposure is denoted as $X$, which is a binary vector (0/1) and $M$ is a data frame or matrix of high-dimensional continuous mediators. Furthermore, it is important to choose the right entry for $riskincr$, which is set to $TRUE$ if it is assumed that the presence of the exposure leads to an increase in risk, and $FALSE$ otherwise. The parameters $exposinflc$ and $minflc$ indicate whether the exposure or the mediator should be modeled as constant over time ($TRUE$) or time-variying ($FALSE$). If the estimation of the p-values is calculated with a simulation,

the parameter *simul* is set to $TRUE$ and using $FALSE$ leads to an estimation with the delta rule. More parameters can be looked up in the description of package *himasurv*.

The calculations performed in *metest()* are based on the method developed by Lange and Hansen (2011), which is explained in section 4.1. First, depending on the selection of $exposinflc$ and $minflc$ an Aalen additive hazard model (*aalen()*, from package *timereg*) of the form $Surv(y, delta) \sim X + m$ is used to fit $\lambda_1$, $\lambda_3$ and the corresponding covariance for each mediator separately. $Surv(y, delta)$ represents the survival object of the survival times $y$ and the censoring indicator *delta*, created with the function *Surv()* in the package *survival*. $X$ is the vector of exposure and $m$ represents one single mediator. Afterwards a linear model (*lm()*) with the formula $m \sim X$, leads to an estimation of $\alpha$ and the corresponding variance. Based on those solutions it is possible to calculate the requested p-values via a simulation approach or the delta rule.

### 4.5.2 Multivariate mediator analysis

This section describes the implementation of the function *himasurv()*, based on the three steps in section 4.4, which are used for the multivariate mediation analysis in the case of survival data. Besides the parameters $X$, $y$ and $M$, which have already been described in section 4.5.1, it is important to choose an option for the parameter *method.lambda*, which describes how the regularization parameter $\lambda$ is selected. One of the following three possibilities can be chosen: cross-validation ($CV$), Akaike information criterion ($AIC$) or Bayesian information criterion ($BIC$) (cf. section 3.4.1). *himasurv()* imports the functions *ncvsurv()* and *cv.ncvsurv()* from the package *ncvreg* to perform the MCP regression. In case of cross-validation *cv.ncvsurv()* in *ncvsurv* detects the best value of $\lambda$. Here the cross-validation error is based on the work of Verweij and Van Houwelingen (1993), which is described in section 3.4.1. The function *cv.ncvsurv()* calls *ncvsurv()* $nfolds$ times (the default of the parameter $nfolds$ is 10) and leaves out $\frac{1}{nfolds}$ of the data each time. The default for the tuning parameter $\gamma$ is set to value 3 and the maximum number of *topn* variables which are selected in step 1 is $[2n/log(n)]$ per default (cf. section 4.4.1). The joint significance test can be performed using the raw (set parameter *test="raw"*) or the Bonferroni corrected (set parameter *test= "adjusted"*) p-values (cf. section 4.4.3). The result of *himasurv()* is a table of the remaining potential mediators which were selected by MCP, containing the effects and the corresponding p-values. The variable can be considered a mediator, if the p-value is lower than 0.05.

## 4.6 Simulation

To test and compare the methods covered in this thesis, various simulations are performed which make it possible to compare the univariate mediator analysis (described in section 4.1) with the multivariate mediator analysis (described in section 4.4). The simulation consists of three different settings with 500 simulated data sets each. Setting 1 includes data with no dependency structure between the variables (potential mediators). The data in setting 2 contain a dependency between the variables that affect the survival and in setting 3 the dependency exists between all potential mediators. Each data set consists of $n = 50$ observed units, 100 potential continuous mediators and an exposure $X$, which takes value 1 if the observed unit is exposed and 0 otherwise. 25 observations of $X$ are 1, the other 25 are equal to 0 and the assignment is random. Furthermore, the following notations and terms are applied for the structure of the simulation:

- $X$: binary exposure, with $\beta_X$ being the direct effect of $X$ on survival $y$
- $M_{V1}$: 10 mediators affected by the exposure and with effect on the survival
- $M_{V2}$: 10 variables affected by exposure, but with no effect on the survival
- $M_{V3}$: 10 variables with effect on the survival but they are not affected by the exposure
- $M_{noise}$: 70 noise variables, which are not affected by the exposure and have no effect on the survival

The allocation of the indices is performed in a straightforward manner, as $M_{V1}$ represents the first ten variables $(1 - 10)$, $M_{V2}$ the next ten $(11 - 20)$, $M_{V3}$ the variables $21 - 30$ and $M_{noise}$ the remaining seventy $(31 - 100)$. The variables $M$ originate from a multivariate normal distribution using the function *mvrnorm()* from the package *MASS*. If a variable $M$ is affected by the exposure $X$, a constant value $v$ is added to that variable if $X = 1$. This is the case for the variables with index $V1$ and $V2$. The effects of the variables on survival are $\boldsymbol{\beta}_{V1}$, $\boldsymbol{\beta}_{V2}$, $\boldsymbol{\beta}_{V3}$ and $\boldsymbol{\beta}_{noise}$, with $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3} \neq 0$ and $\boldsymbol{\beta}_{V2} = \boldsymbol{\beta}_{noise} = 0$. The value of $v$ represents the strength of the correlation depending on the simulation setting. The variance of all potential mediators is constant $\frac{1}{100}$. The sections 4.6.4, 4.6.5 and 4.6.6 describe the characteristics and choices of the parameters $\boldsymbol{\beta}_{V1}$ and $\boldsymbol{\beta}_{V3}$, $v$ and $c$ (constant covariance between variables) for each setting.

### 4.6.1 Simulation of survival time

The simulation of the survival time $\boldsymbol{y}$ consists of two parts: the simulation of the time of failure $T$ and the simulation of the censoring times $C$.
The times of failure $T$ are simulated based on Bender et al. (2005) using

$$T = \frac{-log(u)}{\lambda exp(\boldsymbol{x}^T \boldsymbol{\beta})}$$

with $u \in$ uniform distribution on $[0, 1]$ (*runif()*).

The continuous, non-negative random variable $X$ is exponentially distributed with $\lambda > 0$. This characteristic is denoted as $X \sim Expo(\lambda)$ with the exponential distribution denoted as

$$f(x) = \lambda exp(-\lambda x)$$

(Fahrmeir et al., 2009, p.460)

The censoring times $C$ originate from an exponential distribution and can be simulated using *rexp(n,rate)*, with length $n$ and rate $\lambda_C$. Before explaining how $\lambda_C$ is selected, some characteristics of the censoring time $C$ need to be clarified. If the censoring time $C$ is lower than the time of failure $T$, which means the observed unit was censored before the event of interest occurred, the survival time $y$ takes this specific value of $C$. In this case, the censoring indicator $\delta$ is equal to 1. If $C \geq T$, then $y = T$ and $\delta = 0$. In this simulation the rate of $\delta = 1$ should be approximately 20%, which means that approximately 20% of the observations are censored. To fulfill this condition, it is important to choose an appropriate rate $\lambda_C$ for the exponential distribution ($rexp()$) when simulating the survival times $y$.

The selection of $\lambda_C$ is made by looking at the mean rate of censored observations in the survival time $y$ of 1000 data sets with a specific value for $\lambda_C$. For this, $T$ and $C$ are simulated 1000 times and comparisons drawn on how often $C < T$. If the mean rate of $C < T$ and therefore $\delta = 1$ for the corresponding cases is approximately equal to 20%, the used value of $\lambda_C$ used for simulating $C$.

### 4.6.2 Methods used for the simulation design

Before describing how the parameters of the simulations are correctly selected, some methods need to be introduced. Those methods are needed for the checks that are performed to control each simulation design after choosing the parameters. One of those measurements is called the concordance index (C index), which is a way to estimate the prediction error, in this case of a random survival forest. The C index is described in section 4.6.2.2, but first the idea of random forests and random survival forests is explained in the following section 4.6.2.1.

#### 4.6.2.1 Random survival forest

Random forests can be used to construct a prediction rule for a supervised learning problem and it ranks the predicting variables with respect to their importance in predicting the response. This ranking of predictors is performed with respect to the variable importance measure (VIM), which is calculated for each predicting variable. It is a classification and regression method, aggregating a large number of decision trees which are built from a training data set and validated internally. A well known kind of random forest was submitted by Breiman (2001), which builds the decision trees based on bootstrap samples of the whole data. Afterwards randomly selected covariates are chosen as candidate variables for splitting at each node for each tree. Finally the predictions of all trees are aggregated. Different kinds of random forests exist which vary in the way the trees are constructed, the method used to build the data sets on which each tree is constructed and in the way the predictions are aggregated.

Decision trees are built based on the idea of recursive partitioning, which means, that the observed units at a node are divided recursively into two daughter nodes, each containing the observations with most similar responses. The different kinds of decision trees mostly differ in their splitting criterion. The most common kind of a decision tree is the classification and regression tree (CART) proposed by Breiman et al. (1984). The CART uses the Decrease of Giny Impurity (DGI) as a splitting criterion, that means that the splitting is guided by the impurity of the nodes. This means that the splitting causes the daughter nodes to be divided more precisely with respect to their classification of the response than their parent node.

More about different kinds of decision trees, splitting criteria and more detailed information about random forests can, for example, be found in Breiman et al. (1984), Breiman (2001) and Tutz (2012).

Random forests do have a special characteristic, the out-of-bag (OOB) error. An observation is called out-of-bag observation for a tree if it is not used to build that tree. Thus those observations can be used as a validation data set for those trees. The OOB error describes the average error when predicting the OOB observations with the corresponding tree.

(Tutz, 2012, p.317ff.; Boulesteix et al., 2012, p.494)

Ishawaran et al. (2008) extended the classical random forest idea by Breiman (2001) to survival data. This method is implemented in the R package *randomForestSRC* (function *rfsrc()*), which is used in this thesis. The decision tree used for the random survival forest (RSF) is called binary survival tree, which is similar to CART. The splitting at each node is performed by maximizing the survival difference between the resulting daughter nodes, which means that the cases have a similar survival to the other cases in their group. With an increasing number of nodes, those become more homogeneous and the cases each node contains have a similar survival. The splitting criterion used by *rfsrc()* per default is the log-rank splitting (Ishwaran and Kogalur, 2017). LeBlanc and Crowley (1993) describe this criterion as a weighted difference between the estimated hazard functions.

A decision tree is called saturated if no new daughter nodes can be formed "because of the criterion that each node must contain a minimum of $d_0 > 0$ unique deaths" (Ishawaran et al., 2008, p.844). Those nodes are called terminal nodes.
The prediction error calculated for the random survival forest by Ishawaran et al. (2008) is based on the ensemble cumulative hazard function (CHF), which is computed as an average of the CHF of the $B$ survival trees. The CHF of each tree is calculated based on the Nelson-Aalen estimator and with respect to the terminal nodes.
As the random survival forest being just a small part of this work it will not be covered in detail. For the sake of completeness however, the formulas needed for the calculation of the CHF and ensemble CHF are displayed in the appendix section 7.1. Further reading and details can be found in Ishawaran et al. (2008).

Nevertheless the theory of a random survival forest is easier to understand considering the algorithm, used to built a RSF. The algorithm proceeds as follows:

1. Draw $B$ bootstrap samples from data. Each sample excludes on average 38% of the original data. This is called out-of-bag data.

2. Grow a survival tree for each bootstrap sample, which randomly chooses $p$ variables as candidates for splitting at each node. The splitting is performed depending on which candidate variable maximizes the survival difference between the daughter nodes.

3. "Grow the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths" (Ishawaran et al., 2008, p.843).

4. For each tree the cumulative hazard function (CHF) is calculated and the average over all CHFs (ensemble CHF) is obtained by using their average.

5. The prediction error for the ensemble CHF is calculated using OOB data.

The last step of the algorithm is talking about calculating the prediction error. In case of the RSF this is done with the so called Concordance index (C index) based on Harrell et al. (1982). The theory of the C index which is used in this analysis is explained in the following section 4.6.2.2.

(Ishawaran et al., 2008, p.841ff.)

#### 4.6.2.2 Concordance index

The concordance index, or C index, is a common measure of predictive discrimination. It "is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant" (Harrell Jr. et al., 1996, p.370). Pencina and D'Agostino (2004) generalized the method of the Receiver operating characteristic (ROC) curve area, a measure of discrimination for logistic regression, for survival data. They start by looking at a logistic regression setting, observing two classes of individuals, one developing the event of interest and one which does not.

Assume $Y$ being a random variable describing the predicted probabilities of having the event for those observed units who actually had an event of interest and $V$ being a random variable describing the predicted probabilities of having the event for

those who did not. In case $Y$ and $V$ are continuous the area under the ROC curve, denoted by $C$, is $P(Y \geq V)$. If $Y$ and $V$ are discrete, the C value is calculated as $C = P(Y > V) + 0.5 \cdot P(Y = V)$. The C Index "can be interpreted as the probability that a subject from the event group has a higher predicted probability of having an event than a subject from the non-event group" (Pencina and D'Agostino, 2004, p.2110).

Hanley and McNeil (1982) linked the ROC curve (C index) to the Mann-Whitney statistic $W_{VY}$ which is defined as follows: With $Y_1, Y_2, ..., Y_k$ being the predicted probabilities of having an event in the event group and $V_1, V_2, ..., V_n$ being the predicted probabilities of having an event in the non event group. Each pair of subjects $(i, j)$, with the first subject not having an event and the second belonging to the event group, can be associated to a number which takes value 1 if $Y_i > V_i$, 0.5 if $Y_i = V_i$ and is equal to 0 otherwise. Summing up the results of all possible pairs leads to the Mann-Whitney statistic $W_{VY}$. Thus the C index, or the area under the ROC curve, for logistic regression can be written as

$$C = \frac{1}{k \cdot n} W_{VY}$$

(Harrell Jr. et al., 1996, p.370; Pencina and D'Agostino, 2004, p.2110f.)

Pencina and D'Agostino (2004) adapt this theory to survival data and assume that nobody leaves the study for reasons other than the event of interest. The following notations are necessary for the calculation of the C index:

- $n$: number of observed units
- $X_1, X_2, ..., X_n$: the actual survival times of the observed units
- $T_1, T_2, ..., T_n$: the corresponding predicted survival times
- $Y_1, Y_2, ..., Y_n$: predicted probabilities of survival

Two types of observed units exist at time point $T$, those who experienced the event of interest and those who did not. Looking at all possible pairs of observed units $(i, j)$, with $i < j$ one can say, that a pair is concordant if $X_i < X_j$ and $T_i < T_j$ or $X_i > X_j$ and $T_i > T_j$.

The "predicted probabilities of surviving until any fixed time point can be used instead of the predicted survival times" (Pencina and D'Agostino, 2004, p.2111).

Therefore a pair $(i, j)$ is considered concordant if $X_i < X_j$ and $Y_i < Y_j$ or $X_i > X_j$ and $Y_i > Y_j$. A pair is considered discordant if $X_i < X_j$ and $Y_i > Y_j$ or $X_i > X_j$ and $Y_i < Y_j$. Some pairs are called unusable, because they are neither concordant nor discordant.

The unconditional probability of concordance ($\pi_c$ ) of the entire population is denoted as

$$\pi_c = P(X_i < X_j \text{ and } Y_i < Y_j) + P(X_i > X_j \text{ and } Y_i > Y_j)$$

and $\pi_d$ is the corresponding probability of discordance

$$\pi_d = P(X_i < X_j \text{ and } Y_i > Y_j) + P(X_i > X_j \text{ and } Y_i < Y_j)$$

Then $\pi_t = 1 - \pi_c - \pi_d$ defines the proportion of unusable pairs.

Assuming that the distribution of $Y$ is continuous and that the pairs considered are usable, the C index is defined as

$$C = P(X_i < X_j \text{ and } Y_i < Y_j \text{ or } X_i > X_j \text{ and } Y_i > Y_j) = \frac{\pi_c}{\pi_c + \pi_d}$$

One way of estimating the C index is given by Nam and D'Agostino (2002). Assuming a sample of $n$ observed units, $X_1, X_2, ..., X_n$ being the survival times, $Y_1, Y_2, ..., Y_n$ the predicted probabilities of survival and $X_i \neq X_j$, the C index can be calculated with

$$\hat{C} = \frac{1}{Q} \sum_{(i,j) \in U} c_{ij}$$

with $U$ being a set of all usable pairs $(i, j)$ and $Q$ is the number of all comparisons made. The term $c_{ij}$ is equal to 1 if the pair is concordant, and equal to 0 if the pair is discordant. More details are given in Harrell Jr. et al. (1996), Pencina and D'Agostino (2004) and Nam and D'Agostino (2002).

(Harrell Jr. et al., 1996, p.370; Pencina and D'Agostino, 2004, p.2110f.)

### 4.6.3 Simulation design

Several checks are performed to control each simulation design after choosing the parameters $\beta_X$, $\boldsymbol{\beta}_{V1}$, $\boldsymbol{\beta}_{V3}$, $v$ and $c$ (the corresponding covariance in setting 2 and 3) for each setting separately. Table 4.1 shows the general structure of the parameters. A table like this table shown in every description of each setting to illustrate how the simulation is performed. To find a good fit the check will be performed for two different simulations in each setting.

| **Parameter** | **Description** | **Value** |
|---|---|---|
| $v$ | constant value added to $M_{V1}$ & $M_{V2}$ if $X = 1$ | numeric value |
| $c$ | constant covariance depending on the setting | numeric value |
| $\beta_X$ | effect of $X$ on survival time $y$ | numeric value |
| $\boldsymbol{\beta}_{V1}$ | effect of $M_{V1}$ on survival | constant numeric value |
| $\boldsymbol{\beta}_{V2}$ | effect of $M_{V2}$ on survival | $= 0$ |
| $\boldsymbol{\beta}_{V3}$ | effect of $M_{V3}$ on survival | constant numeric value |
| $\boldsymbol{\beta}_{noise}$ | effect of noise variables on survival | $= 0$ |

Table 4.1: Parameters of the simulation

The testing phase contains four checks of the simulated data sets. The purpose of those tests is to make sure that the chosen parameters create simulated data sets which describe realistic situations with particular characteristics. These characteristics differ depending on the setting (cf. section 4.6.4, 4.6.5 and 4.6.6).
The results of the tests are shown for each setting in each section separately.

1. Check: Calculate the concordance index (C index), based on Harrell Jr. et al. (1996) and Pencina and D'Agostino (2004) (cf. section 4.6.2.2).
    - Calculate the C index using out-of-bag prediction of a random survival forest, *rfsrc()* in package *randomForestSRC()*, which is based on the method of Ishawaran et al. (2008), described in section 4.6.2.1, including all potential mediators $M$ and the exposure $X$ using *concordance.index()* from the package *survcomp*. For more information about the random forest see section 4.6.2.1.
    - The C index should take a value around $0.7 - 0.8$.
2. Check: Take a look at the influence of variables $(V1, V3)$ on survival time.
    - For each potential mediation fit a Cox Regression for the survival time $y$ and look at the corresponding p-values (*coxph()*).

- Take a look at the box plots of those calculated p-values. One plot shows the variables with the effect on the survival time ($V1, V3$) and the other the remaining variables ($V2, noise$). The box plot corresponding to the variables with an effect on the survival time ($V1, V3$) should have appropriately smaller p-values.

3. Check: Take a look at the influence of exposure on variables ($V1, V2$).
   - Take a look at the box plots for each potential mediator with index $V1$ and $V2$ by the exposure. The box plot for $X = 1$ should show higher values for each potential mediator.
   - Perform a t-test (*t.test()*) for each mediator of $V1$ and $V2$ with the exposure and look at the p-values. If $p < 0.05$ the hypotheses $H_0 : \mu_{(X=0)} = \mu_{(X=1)}$ can be rejected and the mean of the mediators differs significantly depending on the presence of the exposure $X$.

4. Check: Take a look at the influence of the exposure on the survival times.
   - Take a look at the box plot of the survival time with the exposure.
   - Check if the corresponding survival curves (Kaplan-Meier estimator) show realistic results using *survfit()*.

### 4.6.4 Setting 1: No dependency structure between variables

In the first setting no dependency structure exists between the variables. The potential variables are simulated with *mvrnorm()* including the the vector $\boldsymbol{mean} = \mathbf{0}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}_{100\times 100} = \begin{pmatrix} \frac{1}{100} & 0 & ... & 0 \\ 0 & \frac{1}{100} & ... & 0 \\ \vdots & ... & \ddots & \vdots \\ 0 & 0 & ... & \frac{1}{100} \end{pmatrix}$$

The diagonal describes the variables' variances, which is constant and takes the value $\frac{1}{100}$. The simulation is performed as described in section 4.6.3 with two resulting different data sets simulated with two seeds, the first is 2017 (data set 1) and the second 1234 (data set 2). The chosen parameters (cf. section 4.6.3) are displayed in table 4.2. As setting 1 has no dependency structure between the variables $c$ is equal to 0. The variables $M_{V2}$ and $M_{noise}$ do not have an effect on the survival times $y$, therefore $\boldsymbol{\beta}_{V2}$ and $\boldsymbol{\beta}_{noise}$ take the value 0. $M_{V1}$ and $M_{V2}$ are affected by the exposure, hence $v = 0.01$ is added to their entries. The direct effect of $X$ on $y$ denoted by $\beta_X$ is set to 0.6. As the variables $M_{V1}$ and $M_{V3}$ have an effect on the survival times, $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3}$ take the value 3.5.

| **Parameter** | **Value** |
|---|---|
| v | 0.1 |
| c | 0 |
| $\beta_X$ | 0.6 |
| $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3}$ | 3.5 |
| $\boldsymbol{\beta}_{V2}$ | 0 |
| $\boldsymbol{\beta}_{noise}$ | 0 |

Table 4.2: Parameters of simulation setting 1

Afterwards, the checks displayed in section 4.6.3 are performed on both data sets. All corresponding Figures are attached in appendix section 7.2.1 - 7.2.3.

The C index for data set 1 is $C = 0.8081$ and $C = 0.7917$ for data set 2, therefore they fit the requirements.

Figure 7.1 and 7.2 show the influence of the variables on the survival time. The identifier takes value 1 if the variables are real mediators ($M_{V_1}$) or if they have an effect on $y$ ($M_{V_3}$) and 0 otherwise. The p-values, displayed in those box plots, are much lower for $M_{V1}$ and $M_{V3}$ compared to $M_{V2}$ and the median lies below $p = 0.05$ (red line). Therefore the second test for the simulation is successful as well.

The variables influenced by the exposure $X$ are $M_1 - M_{20}$ ($M_{V1}$ or $M_{V2}$) and the corresponding box plots in appendix section 7.2.2 for data set 1 are Figure 7.3 - 7.6 and Figure 7.7 - 7.10 for data set 2. In data set 1, most of the medians of $M_1 - M_{V20}$ are higher in case the exposure is present. Only $M_5$ shows medians of the same dimension, however the values are a bit higher if $X = 1$. The Figures for data set 2 all show higher values for $M_1 - M_{V20}$ in case the exposure is present. Nevertheless the corresponding box plot to $M_{18}$ shows, that there is not a huge difference between $X = 0$ or $X = 1$ even though the median for $X = 1$ is a bit higher.
Additional to box plots showing the influence of the exposure on the variables of interest, table 4.3 displays the corresponding t-test p-values for each variable in $M_{V1}$ or $M_{V2}$ of each data set. The p-values show a similar situation like the box plots discussed before. In data set 1 the mediators $M_5$ and $M_{15}$ do not have significantly different means for $X = 0$ and $X = 1$ and in data set 2 this is the case for $M_{18}$.

| **Variable** | **p-value data set 1** | **p-value data set 2** |
|---|---|---|
| $M_1$ | 0.00046 | 0.03165 |
| $M_2$ | 0.00208 | 0.00157 |
| $M_3$ | 0.00045 | 0.00020 |
| $M_4$ | 0.00000 | 0.00223 |
| $M_5$ | **0.13418** | 0.00001 |
| $M_6$ | 0.01309 | 0.00000 |
| $M_7$ | 0.00006 | 0.00031 |
| $M_8$ | 0.00004 | 0.00101 |
| $M_9$ | 0.00017 | 0.00527 |
| $M_{10}$ | 0.01391 | 0.00039 |
| $M_{11}$ | 0.02881 | 0.01130 |
| $M_{12}$ | 0.00375 | 0.00145 |
| $M_{13}$ | 0.00049 | 0.00268 |
| $M_{14}$ | 0.00505 | 0.00001 |
| $M_{15}$ | **0.05529** | 0.00053 |
| $M_{16}$ | 0.00155 | 0.00611 |
| $M_{17}$ | 0.00026 | 0.00024 |
| $M_{18}$ | 0.00003 | **0.35823** |
| $M_{19}$ | 0.00070 | 0.01227 |
| $M_{20}$ | 0.00010 | 0.00143 |

Table 4.3: P-values of t-test for each variable in $M_{V1}$ or $M_{V2}$, setting 1

In a final check the influence of the exposure on the survival times is displayed. The corresponding box plots are shown in section 7.2.3, Figure 7.11 and 7.12. The survival curves, estimated with Kaplan Meier are fully displayed in Figure 7.13 and 7.15, as well as in a cutout form in Figure 7.14 and 7.16. The survival times showed have some extreme values and therefore are a bit skewed. However this should not be an issue for the analysis.

### 4.6.5 Setting 2: Dependency between variables that affect survival

In setting 2 the variables affecting survival, thus those with index $V1$ and $V3$, depend on each other. Therefore the dependency structure, determined by $\boldsymbol{\Sigma}$, differs compared to setting 1. The covariance matrix for setting 2 is

$$\boldsymbol{\Sigma}_{100x100} = \begin{pmatrix} \frac{1}{100} & d_{1,2} & ... & d_{1,100} \\ d_{2,1} & \frac{1}{100} & ... & d_{2,100} \\ \vdots & ... & \ddots & \vdots \\ d_{100,1} & d_{100,2} & ... & \frac{1}{100} \end{pmatrix}$$

The simulation is performed analogously to setting 1 (cf. section 4.6.4) and table 4.4 lists the selected parameters. The parameter $d$ takes a constant value $c = \frac{0.07}{100}$ describing the dependency between two variables if both affect the survival $y$ (those with index $V1$ and $V3$), it is equal to zero otherwise. Analogous to the first setting $\boldsymbol{\beta}_{V2}$ and $\boldsymbol{\beta}_{noise}$ are set to 0. The constant value $v$ added to the variables affected by $X$, $M_{V1}$ and $M_{V2}$, is 0.11. The parameters $\beta_X$, $\boldsymbol{\beta}_{V1}$ and $\boldsymbol{\beta}_{V3}$ take the same values as in setting 1.

| **Parameter** | **Value** |
|---|---|
| v | 0.11 |
| c | $\frac{0.07}{100}$ |
| $\beta_X$ | 0.6 |
| $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3}$ | 3.5 |
| $\boldsymbol{\beta}_{V2}$ | 0 |
| $\boldsymbol{\beta}_{noise}$ | 0 |

Table 4.4: Parameters of simulation setting 2

All Figures needed for the following checks of the simulation are displayed in appendix section 7.2.4 - 7.2.6.

The C index values with $C = 0.8491$ for data set 1 and $C = 0.7183$ for data set 2 fit the requirements.

Figure 7.17 and 7.18 show the influence of the variables on the survival time. The p-values, displayed in those box plots, are much lower for $M_{V1}$ and $M_{V3}$, those variables influencing the survival time, compared to $M_{V2}$. Additionally the median is placed below $p = 0.05$ (red line). Therefore the second test for the simulation is successful as well.

The variables influenced by the exposure $X$ are $M_1 - M_{20}$. The corresponding box plots are shown in section 7.2.5, Figure 7.19 - 7.22 for data set 1, and Figure 7.23 - 7.26 for data set 2. In data set 1 and data set 2, all of the medians of $M_1 - M_{20}$ are higher in case the exposure is present. Furthermore the p-values for t-tests of each variable in $M_{V1}$ or $M_{V2}$ are displayed in table 4.5 for each data set. The variables $M_{11}$ and $M_{15}$ in data set 2, are the only ones with a p-value higher than 0.05.

| **Variable** | **p-value data set 1** | **p-value data set 2** |
|---|---|---|
| $M_1$ | 0.00196 | 0.00081 |
| $M_2$ | 0.00022 | 0.00658 |
| $M_3$ | 0.00000 | 0.00004 |
| $M_4$ | 0.00002 | 0.00456 |
| $M_5$ | 0.00691 | 0.00009 |
| $M_6$ | 0.00019 | 0.00003 |
| $M_7$ | 0.02663 | 0.00009 |
| $M_8$ | 0.00348 | 0.00000 |
| $M_9$ | 0.00000 | 0.00282 |
| $M_{10}$ | 0.00038 | 0.00964 |
| $M_{11}$ | 0.01064 | **0.30721** |
| $M_{12}$ | 0.00023 | 0.00019 |
| $M_{13}$ | 0.00233 | 0.00000 |
| $M_{14}$ | 0.00018 | 0.00001 |
| $M_{15}$ | 0.00000 | **0.06314** |
| $M_{16}$ | 0.00937 | 0.00043 |
| $M_{17}$ | 0.00029 | 0.00015 |
| $M_{18}$ | 0.00133 | 0.00004 |
| $M_{19}$ | 0.00575 | 0.00061 |
| $M_{20}$ | 0.00000 | 0.00001 |

Table 4.5: P-values of t-test for each variable in $M_{V1}$ or $M_{V2}$, setting 2

In a final check the influence of the exposure on the survival times is displayed. The corresponding box plots are shown in section 7.2.6, Figure 7.27 and 7.28. The survival curves, estimated with Kaplan Meier are fully displayed in Figure 7.29 and 7.31, as well as in a cutout form in Figure 7.30 and 7.32. The survival times showed have some extreme values and therefore are a bit skewed. However, as before in setting 1, this should not be an issue for the analysis.

### 4.6.6 Setting 3: Dependency between all variables

In setting 3 all variables depend on each other, with $c$ being a constant value. The corresponding covariance matrix is

$$\boldsymbol{\Sigma}_{100x100} = \begin{pmatrix} \frac{1}{100} & c & \dots & c \\ c & \frac{1}{100} & \dots & c \\ \vdots & \dots & \ddots & \vdots \\ c & c & \dots & \frac{1}{100} \end{pmatrix}$$

The simulation is performed analogously to setting 1 and 2 (cf. section 4.6.4 and 4.6.5). The selected parameters for setting 3 are listed in table 4.6. The constant covariance $c$ between all variables is set to $\frac{0.05}{100}$. As before $\boldsymbol{\beta}_{V2}$ and $\boldsymbol{\beta}_{noise}$ are set to 0, $\beta_X = 0.6$ and $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3} = 3.5$. The constant value added to the variables affected by $X$, $M_{V1}$ and $M_{V2}$, is $v = 0.12$.

| **Parameter** | **Value** |
|---|---|
| v | 0.12 |
| c | $\frac{0.05}{100}$ |
| $\beta_X$ | 0.6 |
| $\boldsymbol{\beta}_{V1} = \boldsymbol{\beta}_{V3}$ | 3.5 |
| $\boldsymbol{\beta}_{V2}$ | 0 |
| $\boldsymbol{\beta}_{noise}$ | 0 |

Table 4.6: Parameters of simulation setting 3

All Figures needed for the checks are displayed in appendix section 7.2.7-7.2.9 .

With $C = 0.8064$ and $C = 0.8166$, the C index values for data set 1 and data set 2, the simulations pass the first check.

The influence of $M_{V1}$ and $M_{V3}$, the variables affecting $y$, is shown in Figure 7.33 and 7.34. The box plots displaying the corresponding p-values, reveal lower p-values for $M_{V1}$ and $M_{V3}$ compared to $M_{V2}$. The medians are below $p = 0.05$ as well. Figure 7.35 - 7.38 and Figure 7.39 - 7.42 are the box plots for the variables, which are influenced by $X$ $(M_1 - M_{20})$. Most of the medians are higher in case of $X = 1$. However, for data set 1 the difference between $X = 0$ and $X = 1$ is not that obvious for variable $M_5$. This is the case for variable $M_{19}$ in data set 2 as well. Table 4.7 shows the corresponding p-values resulting from t-tests for each variable in $M_{V1}$ or $M_{V2}$. The variable $M_{18}$ in data set 2, is the only one with a p value higher than 0.05.

| **Variable** | **p-value data set 1** | **p-value data set 2** |
|---|---|---|
| $M_1$ | 0.00004 | 0.00566 |
| $M_2$ | 0.00028 | 0.00013 |
| $M_3$ | 0.00004 | 0.00001 |
| $M_4$ | 0.00000 | 0.00021 |
| $M_5$ | 0.03560 | 0.00000 |
| $M_6$ | 0.00225 | 0.00000 |
| $M_7$ | 0.00001 | 0.00003 |
| $M_8$ | 0.00000 | 0.00012 |
| $M_9$ | 0.00001 | 0.00067 |
| $M_{10}$ | 0.00197 | 0.00004 |
| $M_{11}$ | 0.00781 | 0.00129 |
| $M_{12}$ | 0.00052 | 0.00015 |
| $M_{13}$ | 0.00052 | 0.00046 |
| $M_{14}$ | 0.00066 | 0.00000 |
| $M_{15}$ | 0.01065 | 0.00047 |
| $M_{16}$ | 0.00016 | 0.00087 |
| $M_{17}$ | 0.00023 | 0.00002 |
| $M_{18}$ | 0.00000 | **0.11177** |
| $M_{19}$ | 0.00006 | 0.00103 |
| $M_{20}$ | 0.00001 | 0.00018 |

Table 4.7: P-values of t-test for each variable in $M_{V1}$ or $M_{V2}$, setting 3

Finally the box plots in section 7.2.9, Figure 7.43 and 7.44, display the influence of the exposure on the survival times. The survival curves, estimated with Kaplan Meier are fully displayed in Figure 7.45 and 7.47, as well as in a cutout form in Figure 7.46 and 7.48. Similar to the first two settings, the survival times have some extreme values and therefore are a bit skewed.

### 4.6.7 Results of simulation

The univariate and multivariate mediation analyses are compared based on the analysis of 500 simulated data sets. The simulation for each setting is performed in R using the package *simulation* which was built only for the purpose of this analysis. The functions *sim1*, *sim2* and *sim3* create $m$ data sets of each setting using the simulation designs displayed in section 4.6.4 - 4.6.6.

The first two data sets of each setting are performed using the seed 2017 and 1234. The remaining 498 data sets are simulated based on a random seed resulting from a sample of full numbers from the interval $[100; 10000]$. The detailed simulation is performed as described in section 4.6. Each data set includes 10 mediators and 90 not mediating variables, thus $500 \times 10 = 5000$ mediators and $500 \times 90 = 45000$ not mediating variables altogether.

Each single data set is analyzed using *himasurv()* and *metest()*. When using *himasurv()* with adjusted p-values in step 3, the joint significance test does not identify a reasonable amount of mediators. Therefore it is not useful to compare the multivariate analysis with adjusted p-values to the univariate analysis.

However, it is possible to compare *himasurv()* to *metest()* when using the raw p-values in *himasurv()*. The results are displayed in table 4.8.

The identified mediators can be divided in two groups. One is the true positive group, which means that the selected variables are really mediators and therefore identified correctly. The false positive group contains the variables wrongly selected to be mediators. The rate for the true positive group is calculated as the number of detected variables ($TP$) divided by the number of all existing mediators: $\frac{TP}{5000}$. The rate of false positive ($FP$) selections are obtained analogously with $\frac{FP}{45000}$.

| | himasurv | | | | | | metest | |
|---|---|---|---|---|---|---|---|---|
| | *AIC* | | *BIC* | | *CV* | | | |
| **Setting 1** | | | | | | | | |
| *True positive* | 1273 | (25.46%) | 757 | (15.14%) | 151 | (3.02%) | 1836 | (36.72%) |
| *False positive* | 818 | (1.82%) | 355 | (0.79%) | 33 | (0.07%) | 887 | (1.97%) |
| **Setting 2** | | | | | | | | |
| *True positive* | 1815 | (36.30%) | 1508 | (30.16%) | 732 | (14.64%) | 3085 | (61.70%) |
| *False positive* | 1016 | (2.26%) | 514 | (1.14%) | 81 | (0.18%) | 1134 | (2.52%) |
| **Setting 3** | | | | | | | | |
| *True positive* | 1548 | (30.96%) | 1006 | (20.12%) | 297 | (5.94%) | 1977 | (36.54%) |
| *False positive* | 966 | (2.15%) | 453 | (1.01%) | 64 | (0.14%) | 949 | (2.11%) |

Table 4.8: Identifying mediators in 500 simulated data sets with raw p-values

The performance of *himasurv()* with the selection criterion $CV$ for $\lambda$ is the worst. Using $BIC$ gives better results in this simulation, especially the false positive identifications are low in all simulation settings for $BIC$. The rate of true positive selections are 15.14%, 30.16% and 20.12% for data set 1, data set 2 and data set 3 respectively and the false positive rates are very low.

Selecting $\lambda$ with $AIC$ gives the best results for *himasurv()*. It results in true positive rates of 25.46%, 36.30% and 30.96% and false positive rates of 1.82%, 2.26% and 2.15%.

Compared to *himasurv()* the univariate method *metest()* is a bit better at detecting true positive mediators, because the rates are higher. The false positive rates of $AIC$ and *metest()* are similar.

# 5 Conclusion

The aim of this thesis was to develop a multivariate method for identifying mediators in high dimensional survival data, based on the work of Zhang et al. (2016) and compare this method to a univariate approach proposed by Lange and Hansen (2011).

After explaining some basic ideas and concepts like different kinds of two- and three-variable effects, including mediators in section 3.1 as well as survival data and methods for survival analysis (Cox regression) in section 3.2, section 3.3 describes the concept of high dimensional data and how regularization methods like Lasso and Ridge can deal with them.

Chapter 4 addressed different methods for mediator analysis and contains an approach for univariate mediator analysis in section 4.1, a general description of multivariate mediator analysis in section 4.2, as well as the detailed description of the multivariate method developed by Zhang et al. (2016) in section 4.3.
Based on this three step analysis proposed by Zhang et al. (2016) section 4.4 showed an adjusted adaption of this theory to survival data. This multivariate analysis for high dimensional mediation analysis in case of survival data is made of three steps: a pre-selection, a MCP regression and a joint significance test. The result of the method is a list of variables identified as mediators.

Within the framework of this thesis a package for R, named *himasurv*, containing both, the univariate and the multivariate approach was built. The implementation of the univariate and multivariate methods, named *metest()* and *himasurv()* respectively, were described in section 4.5.

The comparison of both methods was performed with a simulation, consisting of three different dependency settings with 500 simulated data sets each (cf. section 4.6). After checking the simulation design, every setting was analyzed using the developed method and the univariate method.

The simulation revealed that selecting $\lambda$ using the criterion $AIC$ for the multivariate analysis approach gives the best results for *himasurv()*. Compared to *himasurv()* the univariate method *metest()* is a bit better at detecting true positive mediators, because the rates are higher. The false positive rates of $AIC$ and *metest()* are similar.

Mediation analysis is a very important field in statistics, especially in combination with high-dimensional data, as it is more common in a multitude of research areas than ever before. The more data available, the more important it gets to consider the different relationships variables can have with each other (cf. section 3.1). Therefore multivariate analysis in case of high dimensional mediators is an important topic, not only for continuous or binary outcomes, but also in case of survival data. This thesis covers a first approach for this kind of analysis.

Nevertheless, more research on this topic is recommended. For example, it is interesting how the selection of $\gamma$ in the MCP regression, as proposed by Breheny and Huang (2011) (cf. section 3.4.2), changes the results.
Furthermore, the multivariate method contains three steps and all of those steps can be changed. One possibility is to observe how the results differ when changing some components of the original analysis. For example it is possible to change the pre-selection method and use e.g. SIS like in Zhang et al. (2016). An alternative to MCP in step 2 are methods like Lasso or elastic net. Therefore, this topic offers a lot of new research approaches.

# 6 Statutory declaration

## Statutory declaration

I declare that I have authored the thesis

**Identification of Mediators in High Dimensional Survival Data in the Presence of Confounding**

independently, that I have not used any other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

## Eidesstattliche Erklärung

Ich erkläre an Eides Statt, dass ich die Masterarbeit mit dem Titel

**Identification of Mediators in High Dimensional Survival Data in the Presence of Confounding**

selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Munich, 07. June 2017 Carina Rein

# 7 Appendix

## 7.1 Random survival forest: cumulative hazard function (CHF)

Section 4.6.2.1 addresses the random survival forest, an adaption of a random forest for survival data, developed by Ishawaran et al. (2008).

The prediction error calculated for the random survival forest is based on the ensemble cumulative hazard function (CHF), which is computed as an average of the CHF of the $B$ survival trees. The CHF of each tree is calculated with respect to the terminal nodes and based on the Nelson-Aalen estimator.

The following notations are important for the explanation and calculation of the CHF estimate.

- $\mathcal{J}$: set of terminal nodes in a saturated tree
- $h \in \mathcal{J}$: a terminal node
- $(T_{1,h}, \delta_{1,h}), ..., (T_{n(h),h}, \delta_{n(h),h})$: the survival times and the censoring indicator for the cases included in node $h$, with $n(h)$ being the number of individuals in this node
- $d_{l,h}$: number of deaths at time $t_{l,h}$
- $Y_{l,h}$: number of individuals at risk at time $t_{l,h}$
- $t_{1,h} < t_{2,h} < ... < t_{N(h),h}$: distinct event times
- $\boldsymbol{x}_i$: vector of covariates corresponding to case $i$, with $x$ representing one single covariate

The Nelson-Aalen estimator is used to estimate the CHF for one tree with the terminal node $h$ and all cases $i$ within that node have the same CHF

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{7.1}$$

Let $\boldsymbol{x}_i$ be all possible covariates and $H(t|\boldsymbol{x}_i)$ be the CHF for case $i$ which is the Nelson-Aalen estimator (cf. equation 7.1) for the terminal node of $\boldsymbol{x}_i$ and defines the CHF for all cases and therefore for the tree

$$H(t|\boldsymbol{x}_i) = \hat{H}_h(t)$$

The prediction error for the random forest uses the so called ensemble CHF, which is computed as an average of the CHF of the $B$ survival trees the RSF is built of.

The package *randomForestSRC* uses the prediction error based on the C index (cf. Ishawaran et al. (2008)).

The ensemble CHF can be calculated, with $I_{i,b}$ identifying if $i$ is an OOB case for bootstrap sample $b$ (= 1 if it is, and = 0 otherwise) and $H_b^*(t|\boldsymbol{x}_i)$ being the CHF shown in equation 7.1 for sample $b$ as follows

$$H_e^{**}(t|\boldsymbol{x}_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|\boldsymbol{x}_i)}{\sum_{b=1}^{B} I_{i,b}}$$

The bootstrap ensemble CHF for $i$ is calculated as follows

$$H_e^*(t|\boldsymbol{x}_i) = \frac{1}{B} \sum_{b=1}^{B} H_b^*(t|\boldsymbol{x}_i)$$

Further reading and details can be found in Ishawaran et al. (2008).

(Ishawaran et al., 2008, p.841ff.)

## 7.2 Simulation

### 7.2.1 Setting 1: Influence of variables on survival time (check 2)

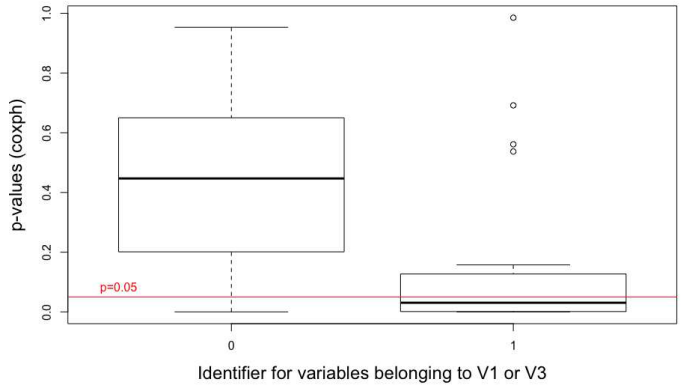


Figure 7.1: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 1.1

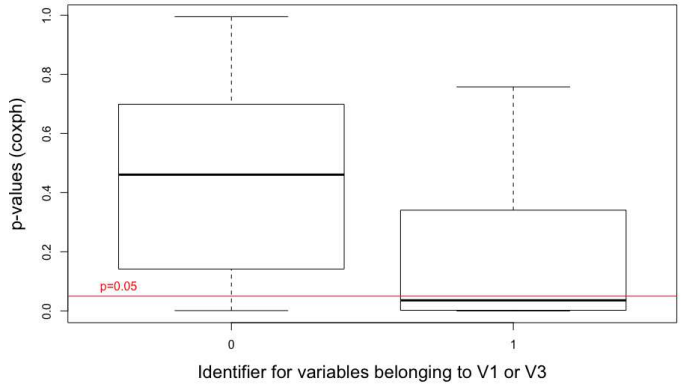


Figure 7.2: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 1.2

### 7.2.2 Setting 1: Influence of exposure on variables (check 3)

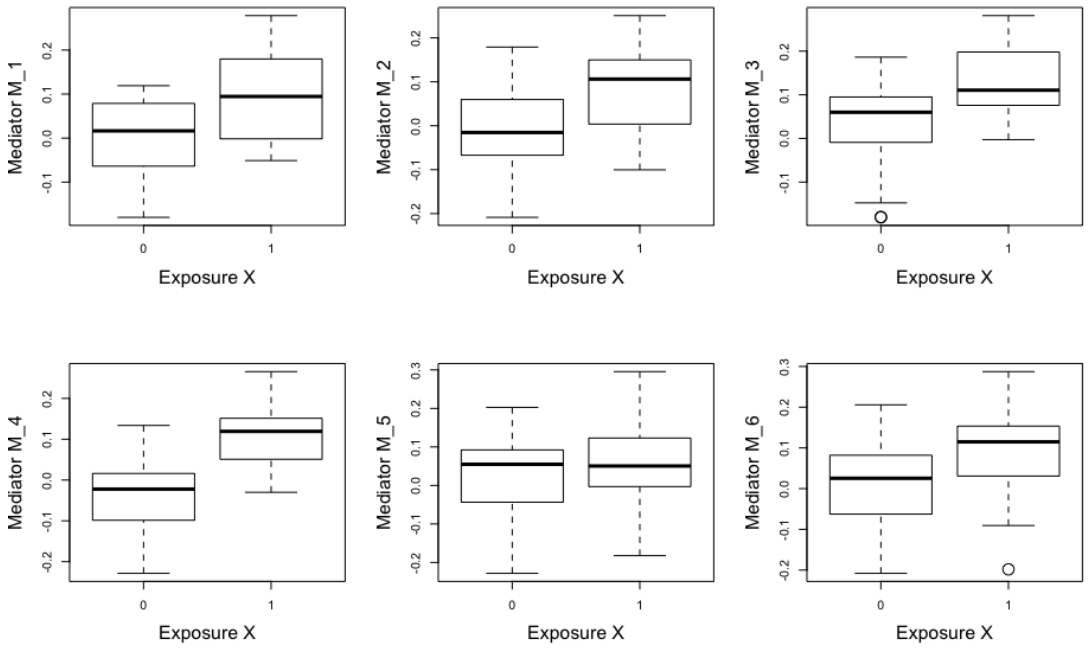


Figure 7.3: Box plot for exposure and Mediators $M_1$ - $M_6$ (V1), setting 1.1

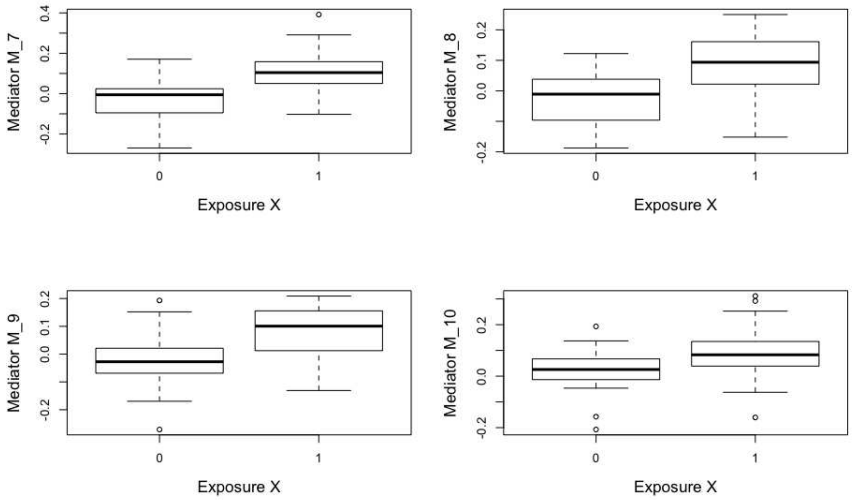


Figure 7.4: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 1.1

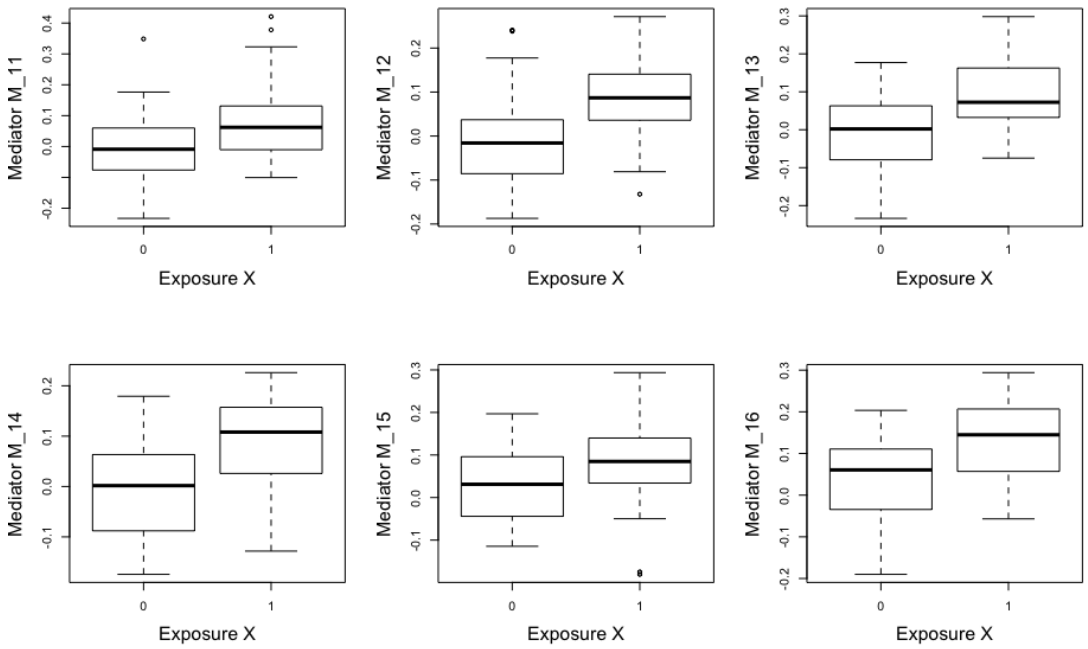


Figure 7.5: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 1.1

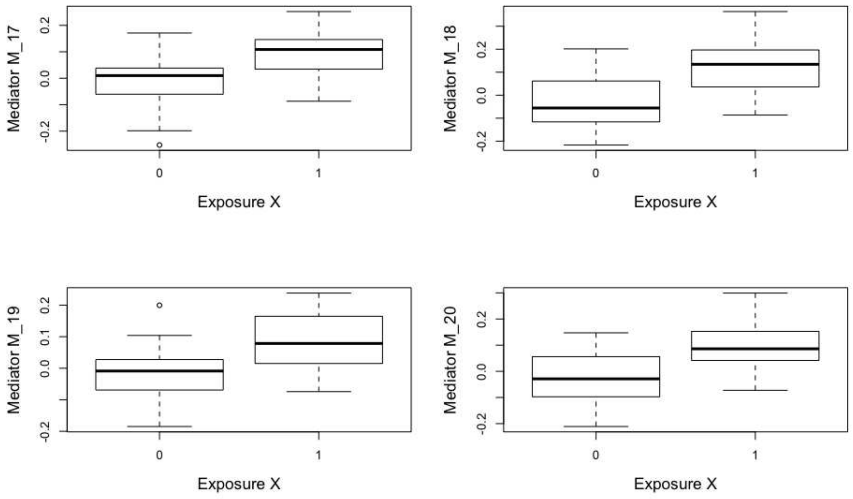


Figure 7.6: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 1.1

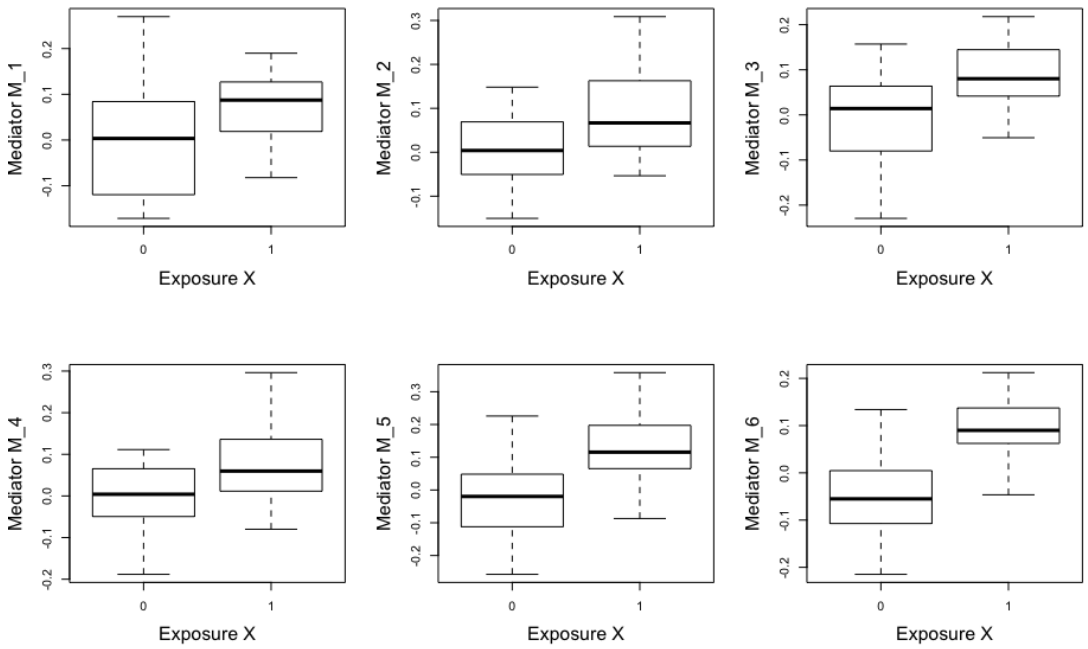


Figure 7.7: Box plots for exposure and Mediators $M_1$ - $M_6$ (V1), setting 1.2

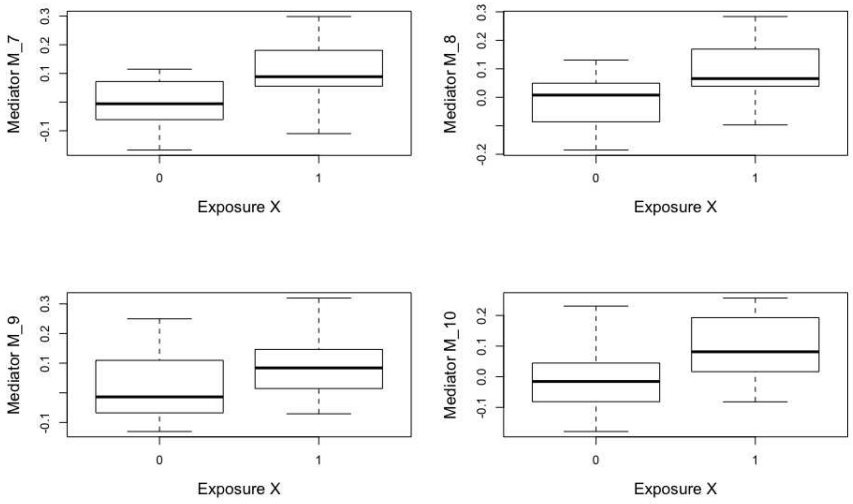


Figure 7.8: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 1.2

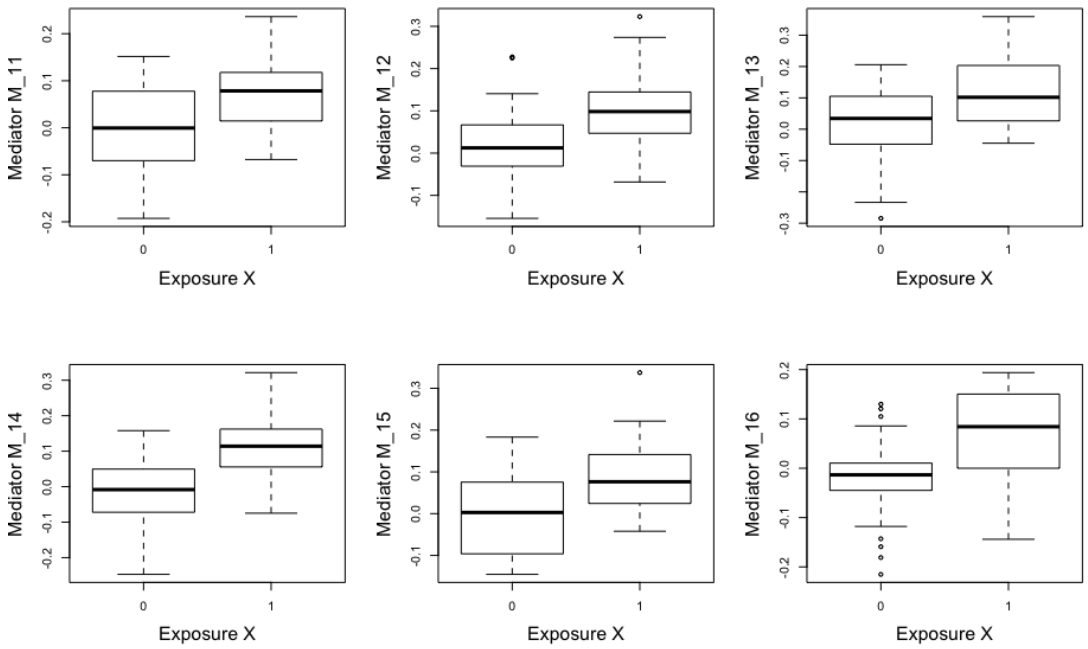


Figure 7.9: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 1.2

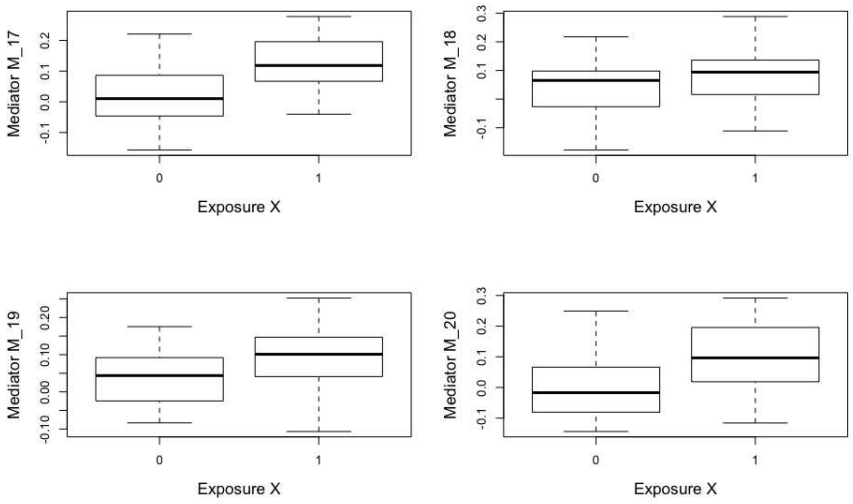


Figure 7.10: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 1.2

### 7.2.3 Setting 1: Influence of exposure on survival time (check 4)

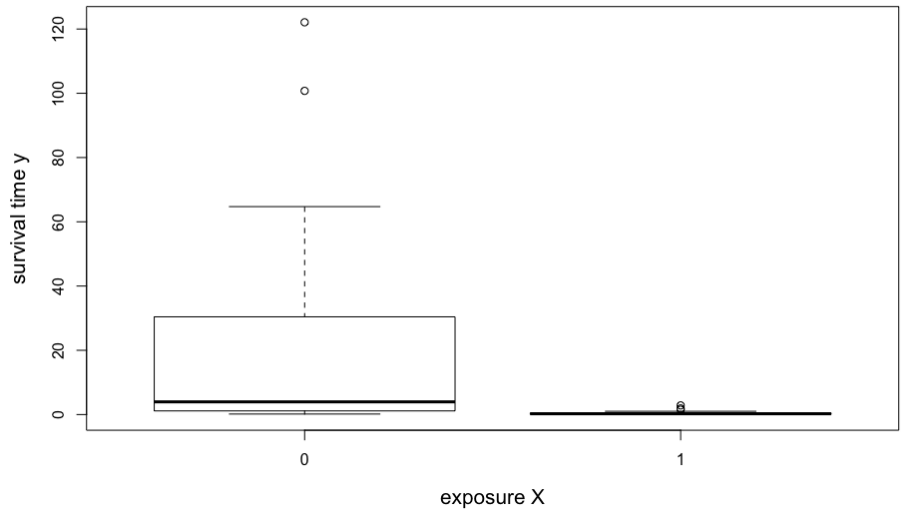


Figure 7.11: Box plot for influence of exposure $X$ on survival times $y$, setting 1.1

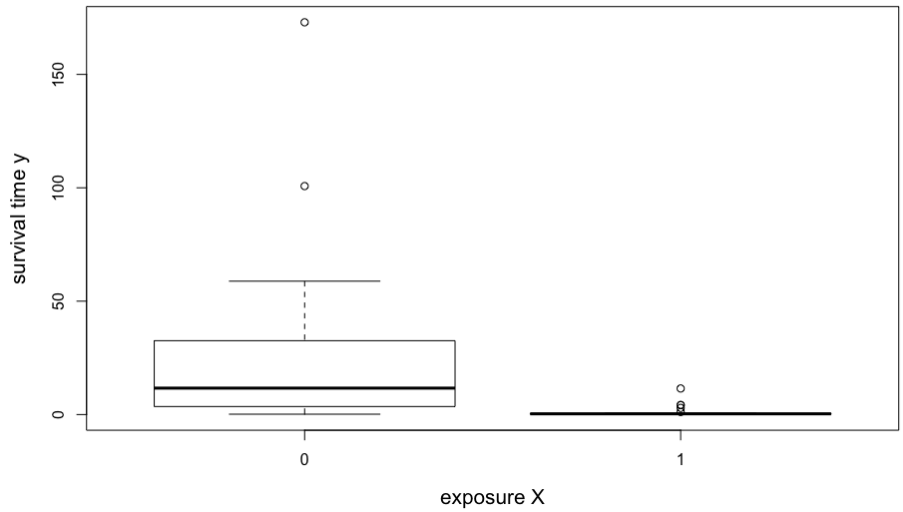


Figure 7.12: Box plot for influence of exposure $X$ on survival times $y$, setting 1.2

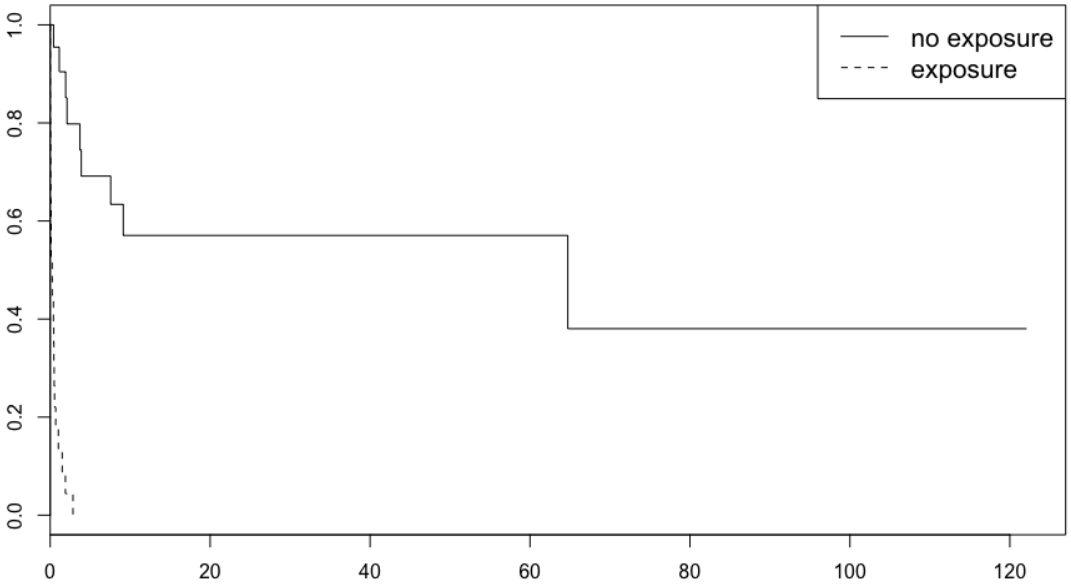


Figure 7.13: Survival curve, influence of exposure $X$ on survival times $y$, setting 1.1

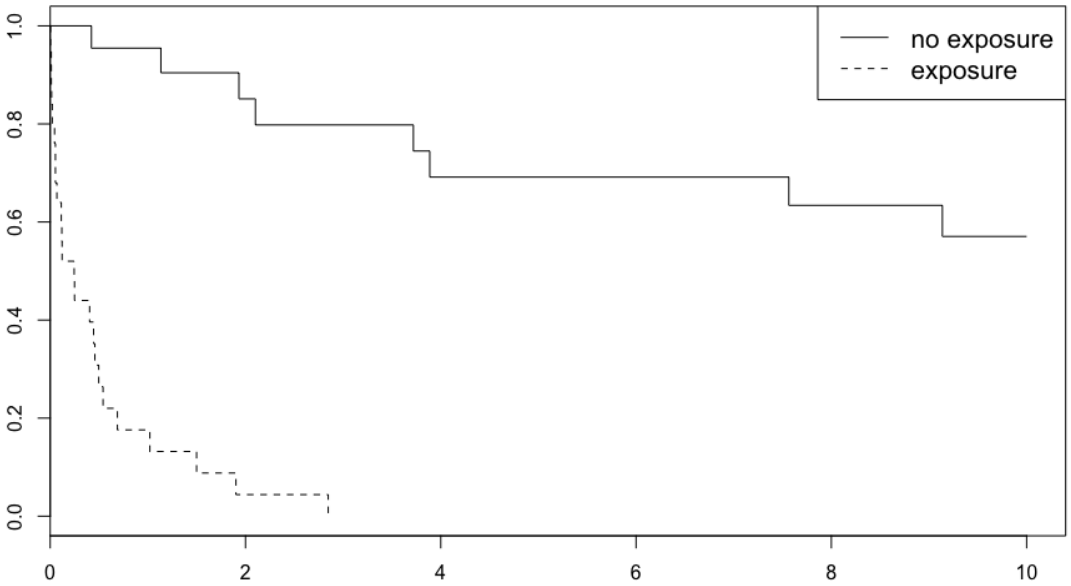


Figure 7.14: Survival curve, cut at $t = 10$, influence of exposure $X$ on survival times $y$, setting 1.1

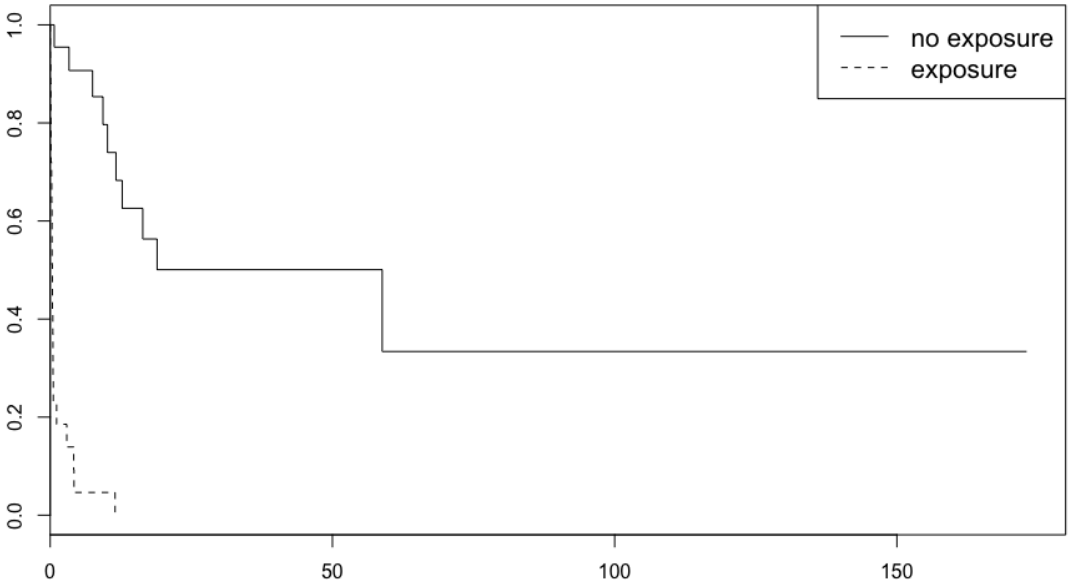


Figure 7.15: Survival curve, influence of exposure $X$ on survival times $y$, setting 1.2

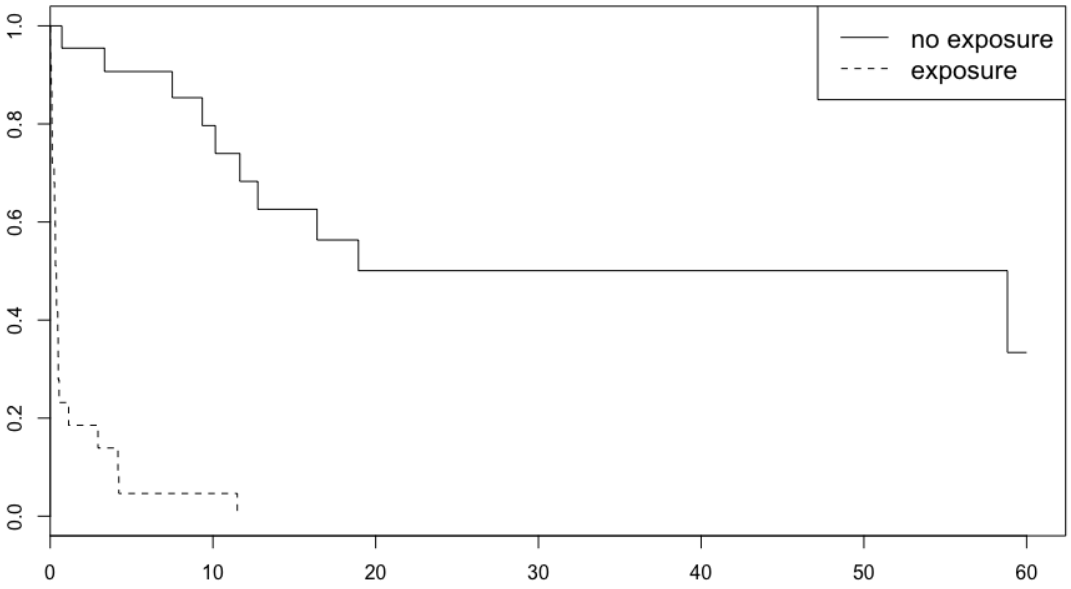


Figure 7.16: Survival curve, cut at $t = 60$, influence of exposure $X$ on survival times $y$, setting 1.2

### 7.2.4 Setting 2: Influence of variables on survival time (check 2)

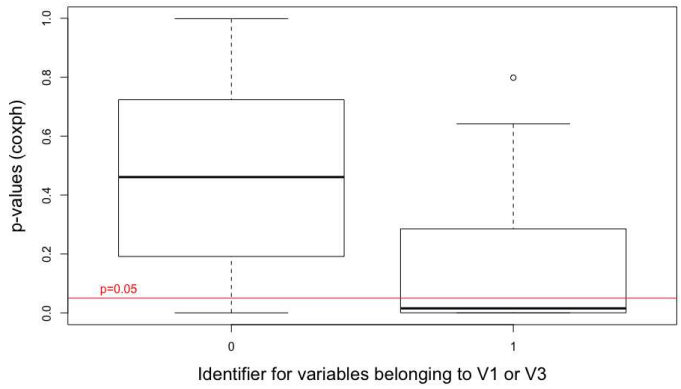


Figure 7.17: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 2.1

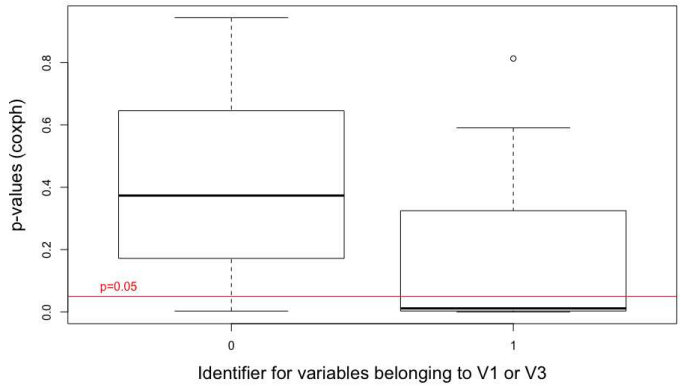


Figure 7.18: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 2.2

### 7.2.5 Setting 2: Influence of exposure on variables (check 3)

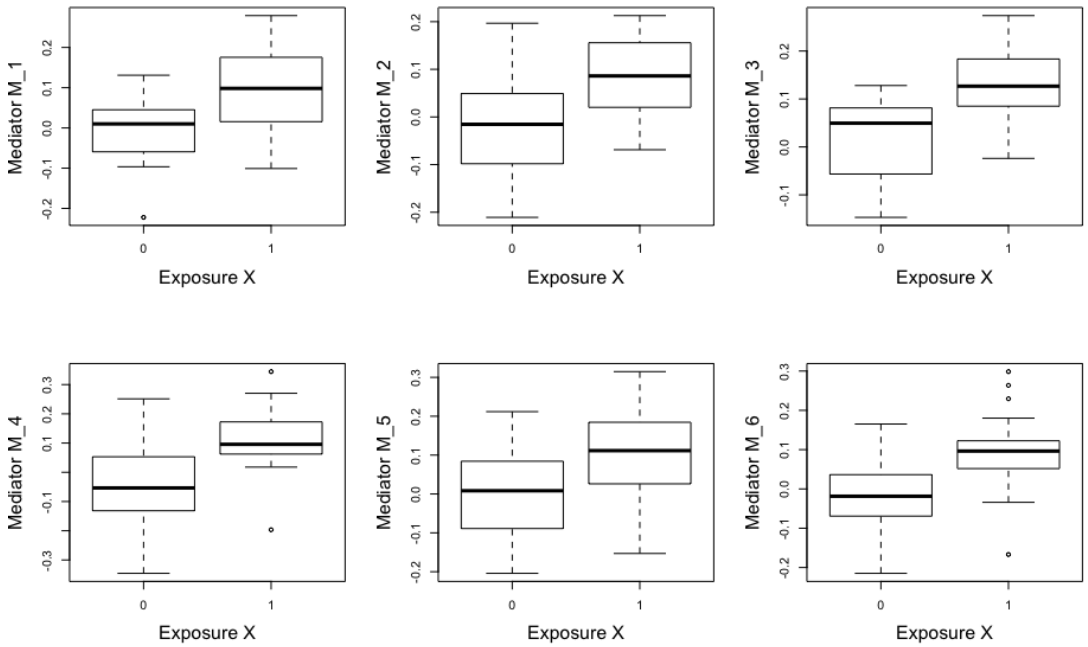


Figure 7.19: Box plots for exposure and Mediators $M_1$ - $M_6$ (V1), setting 2.1

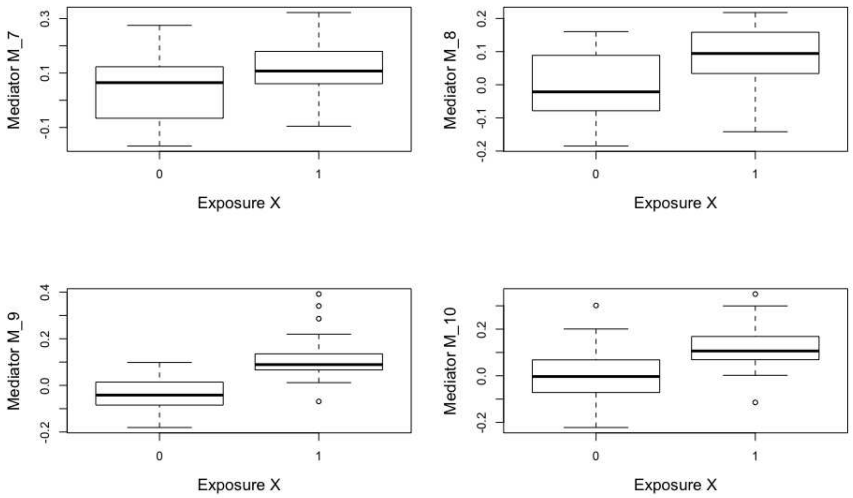


Figure 7.20: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 2.1

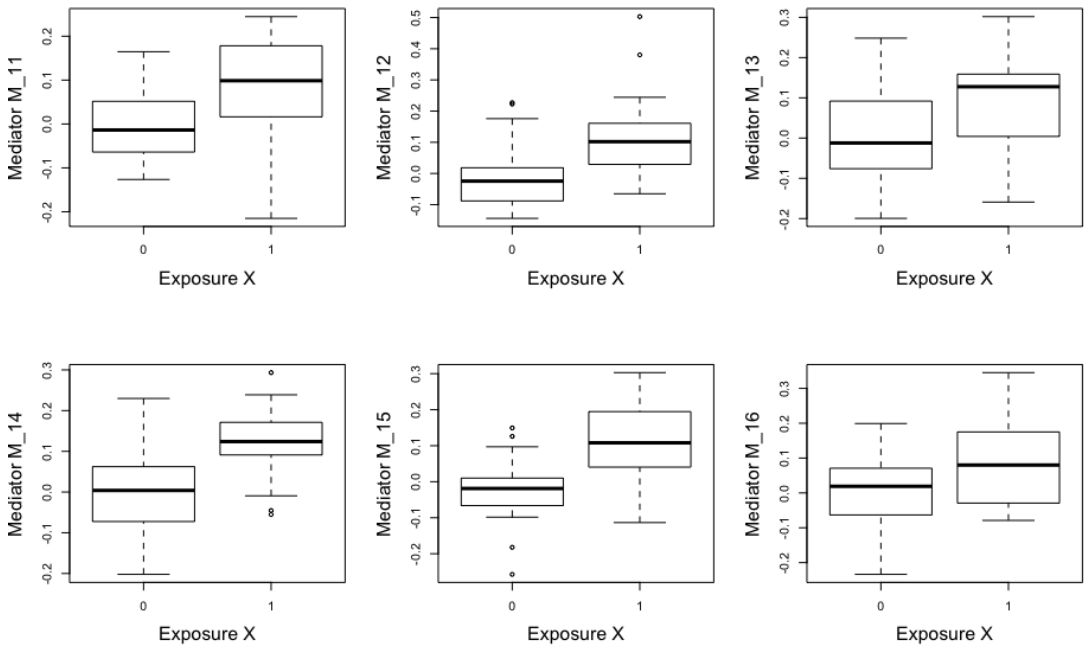


Figure 7.21: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 2.1

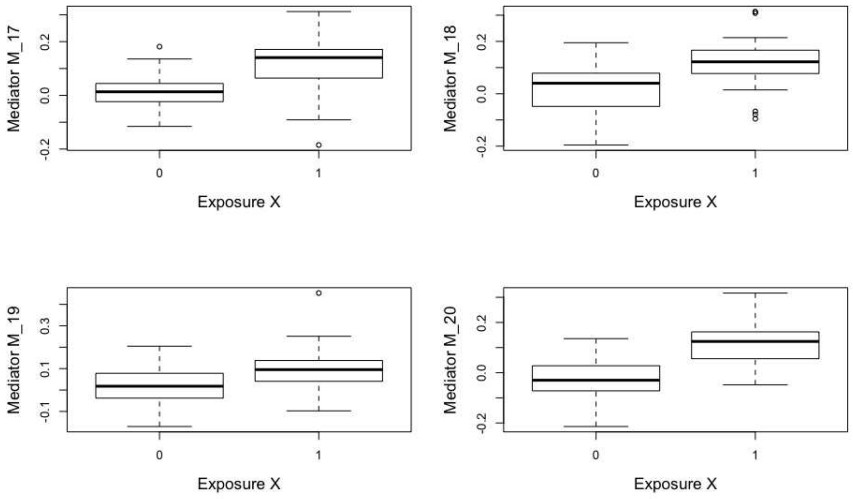


Figure 7.22: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 2.1

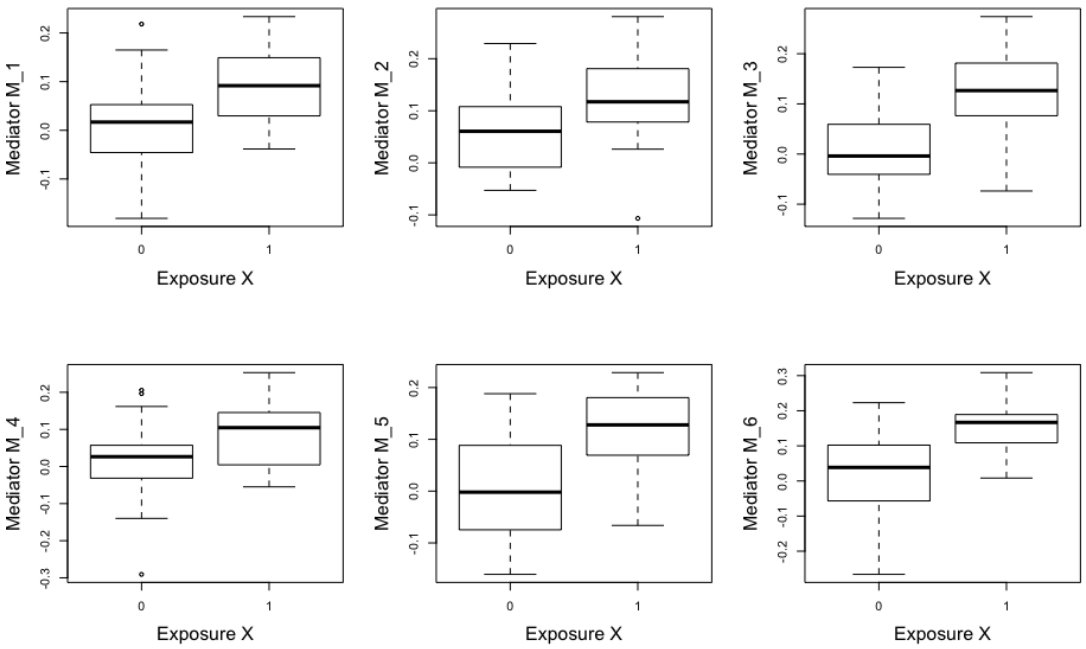


Figure 7.23: Box plots for exposure and Mediators $M_1$ - $M_6$ (V1), setting 2.2

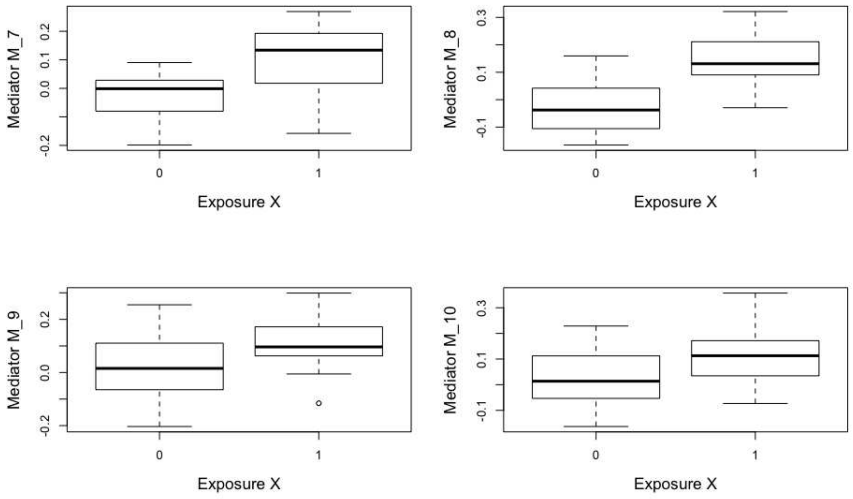


Figure 7.24: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 2.2

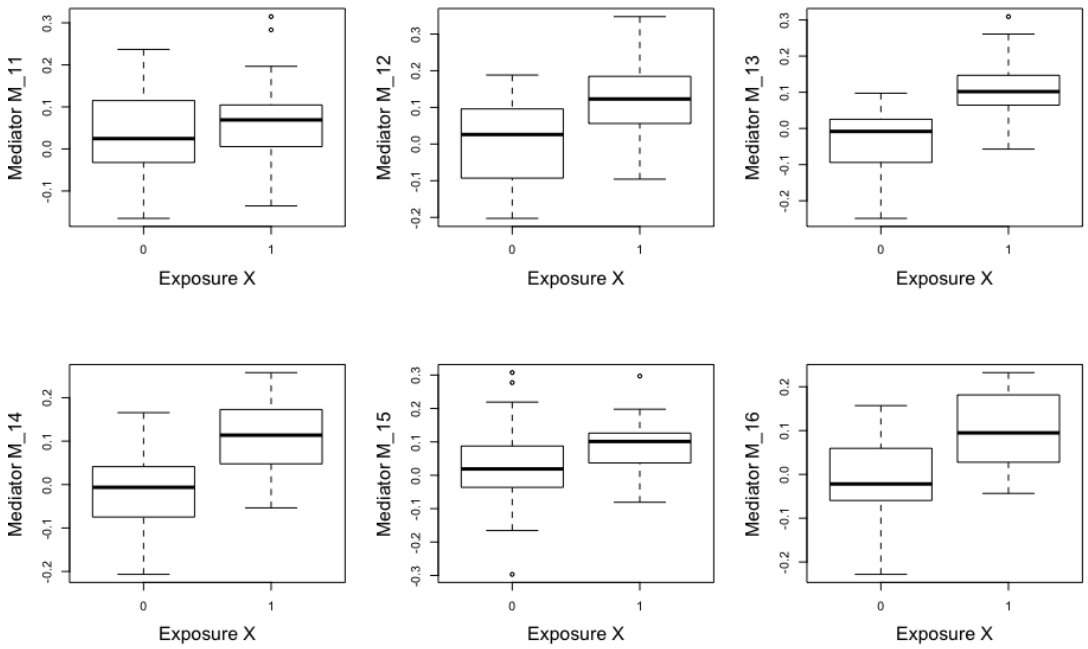


Figure 7.25: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 2.2

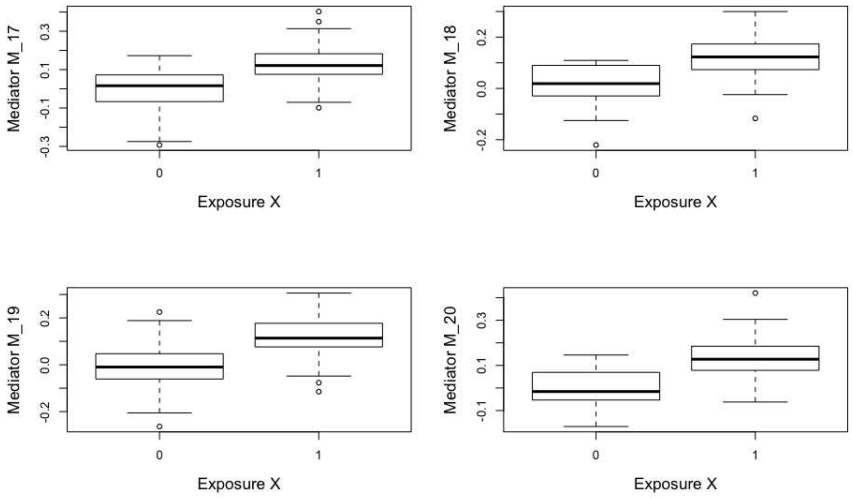


Figure 7.26: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 2.2

### 7.2.6 Setting 2: Influence of exposure on survival time (check 4)

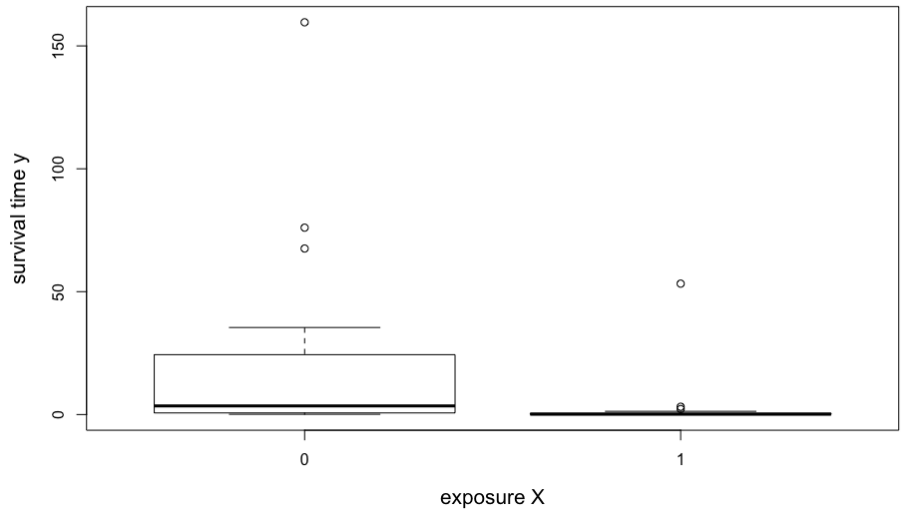


Figure 7.27: Box plot for influence of exposure $X$ on survival times $y$, setting 2.1

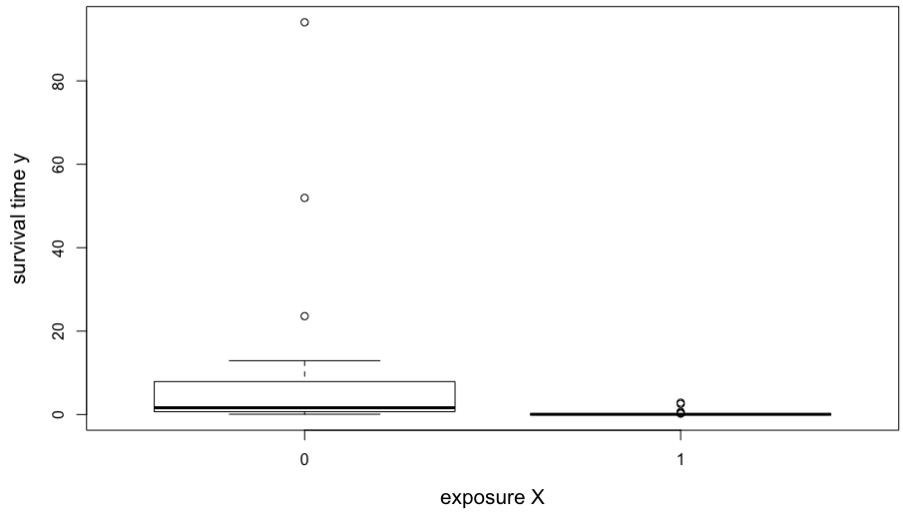


Figure 7.28: Box plot for influence of exposure $X$ on survival times $y$, setting 2.2

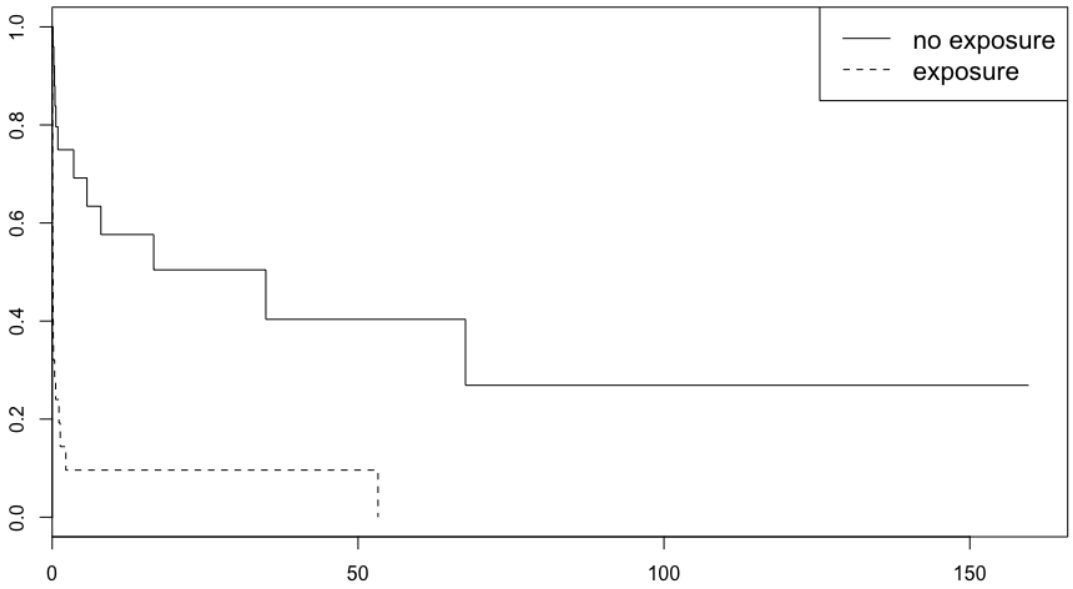


Figure 7.29: Survival curve, influence of exposure $X$ on survival times $y$, setting 2.1

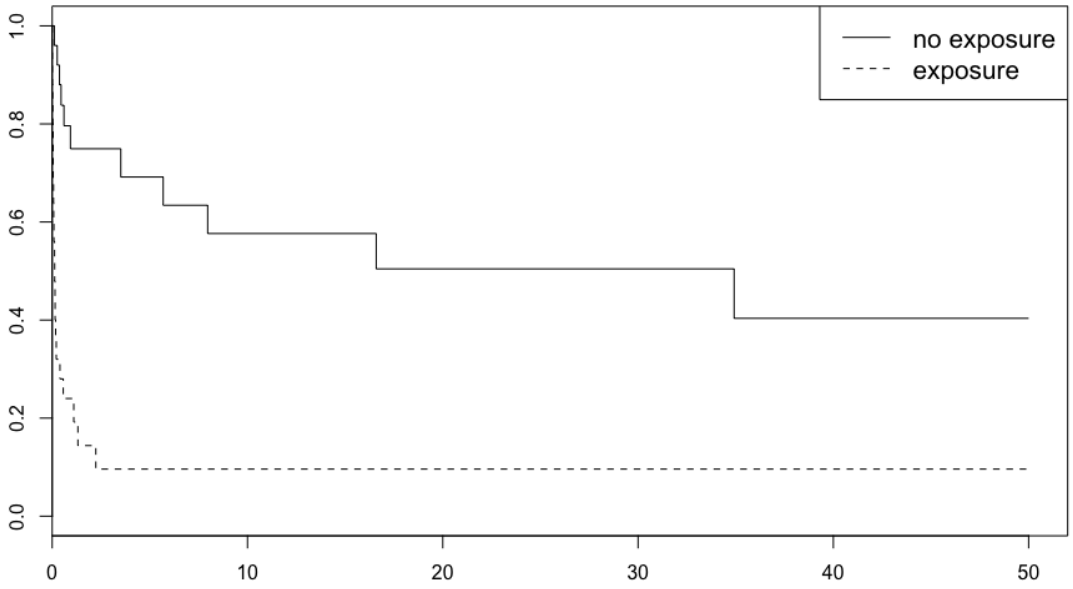


Figure 7.30: Survival curve, cut at $t = 50$, influence of exposure $X$ on survival times $y$, setting 2.1

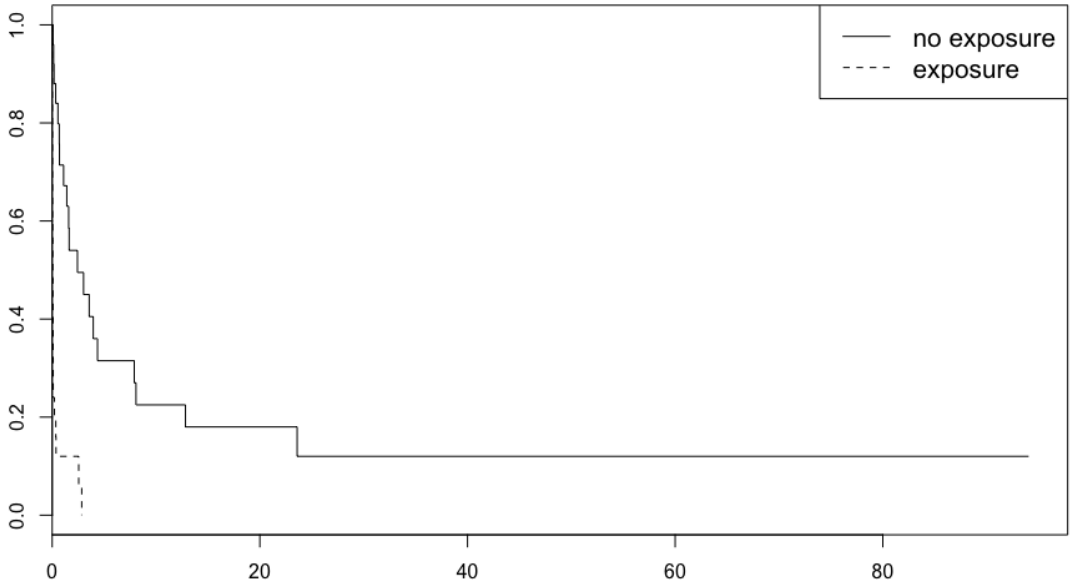


Figure 7.31: Survival curve, influence of exposure $X$ on survival times $y$, setting 2.2

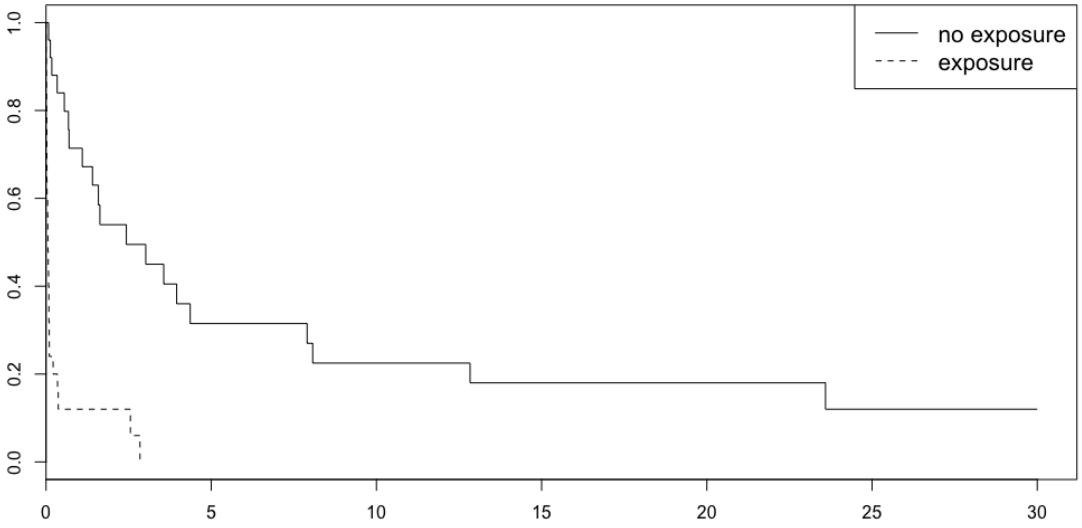


Figure 7.32: Survival curve, cut at $t = 30$, influence of exposure $X$ on survival times $y$, setting 2.2

### 7.2.7 Setting 3: Influence of variables on survival time (check 2)

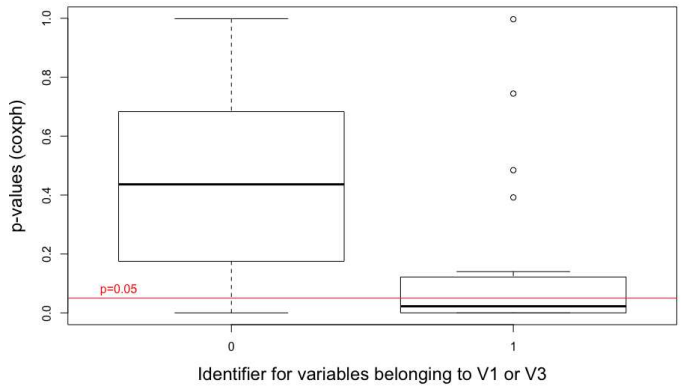


Figure 7.33: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 3.1

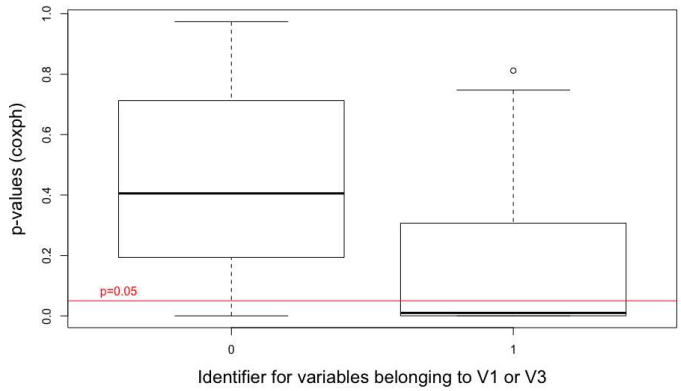


Figure 7.34: Influence on the survival time for $M_{V1}$ and $M_{V3}$ (p-values), setting 3.2

### 7.2.8 Setting 3: Influence of exposure on variables (check 3)

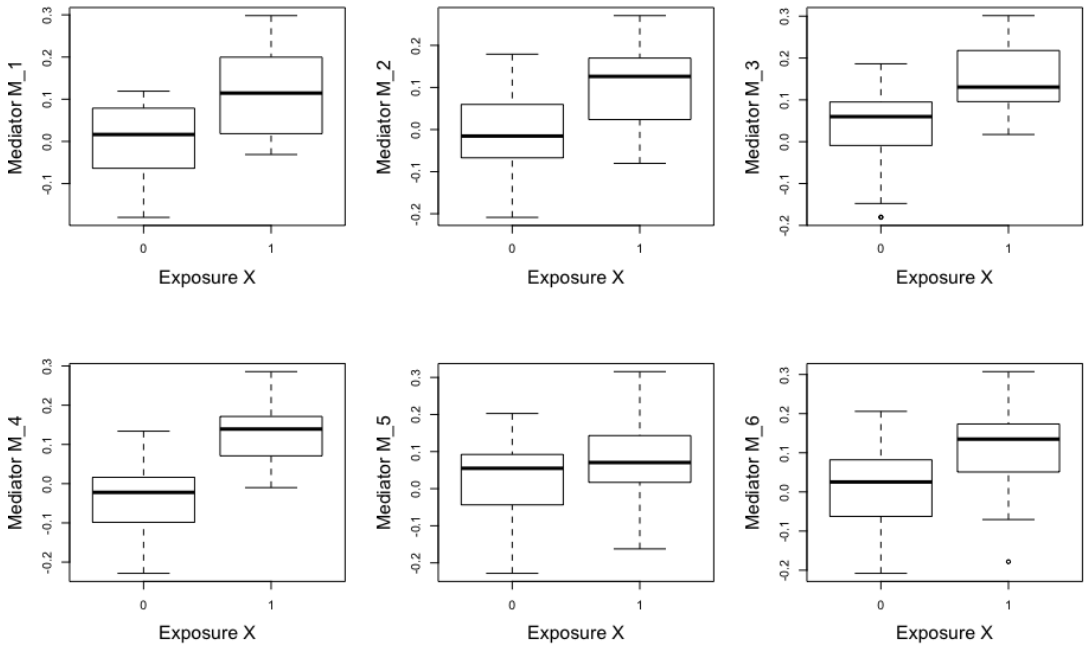


Figure 7.35: Box plots for exposure and Mediators $M_1$ - $M_6$ (V1), setting 3.1

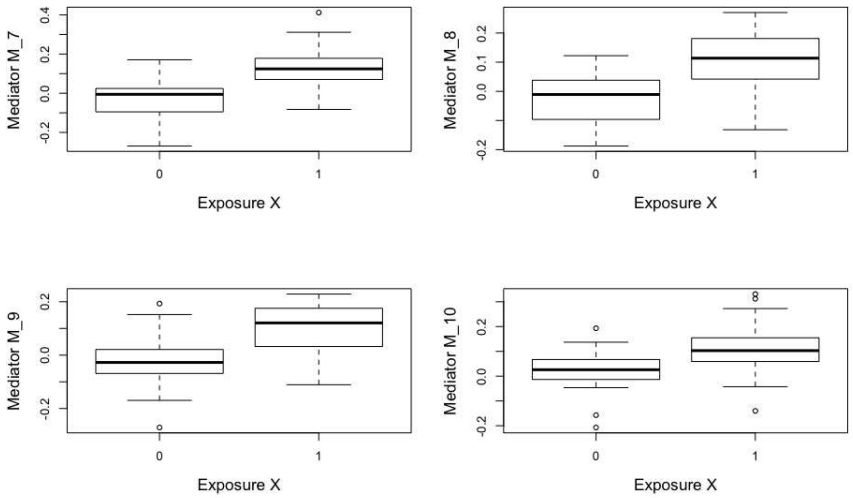


Figure 7.36: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 3.1

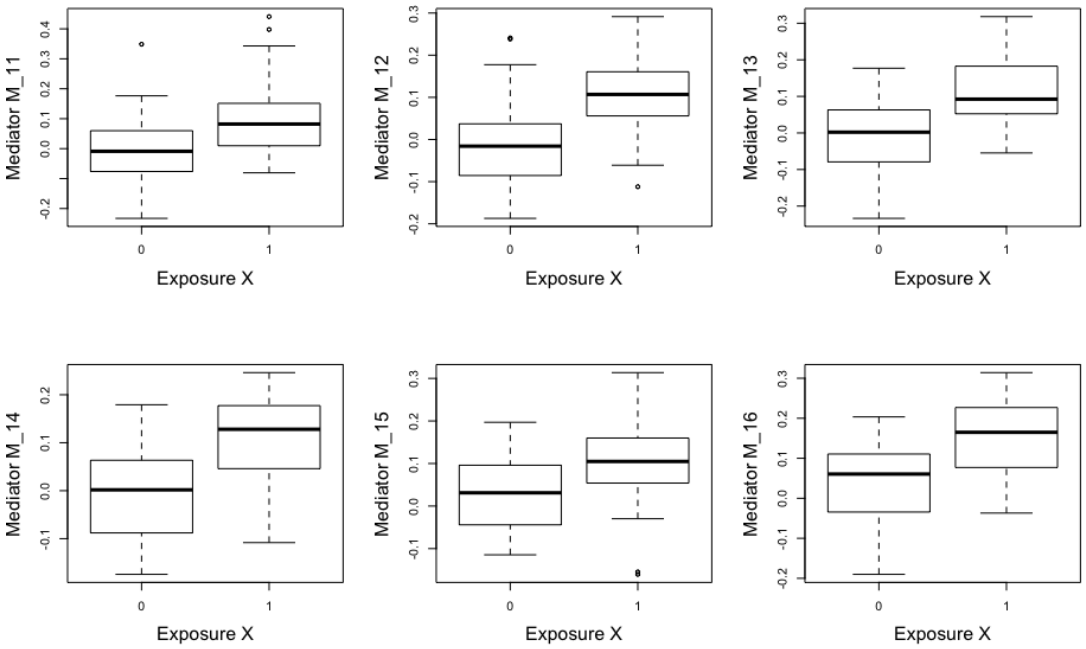


Figure 7.37: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 3.1

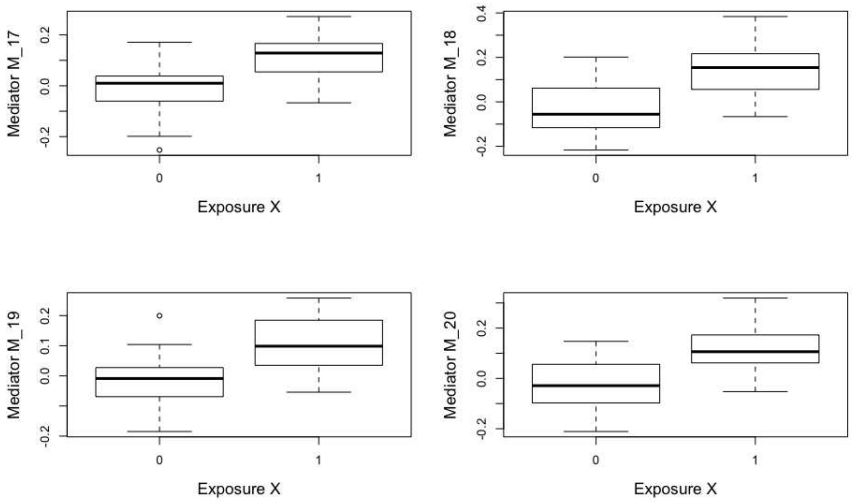


Figure 7.38: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 3.1

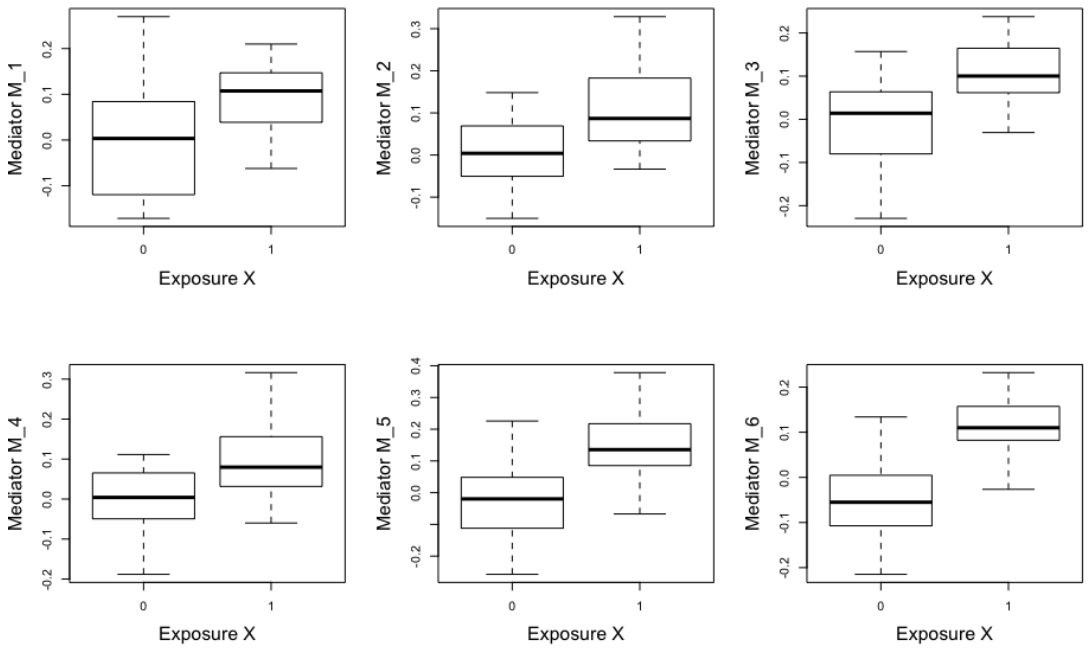


Figure 7.39: Box plots for exposure and Mediators $M_1$ - $M_6$ (V1), setting 3.2

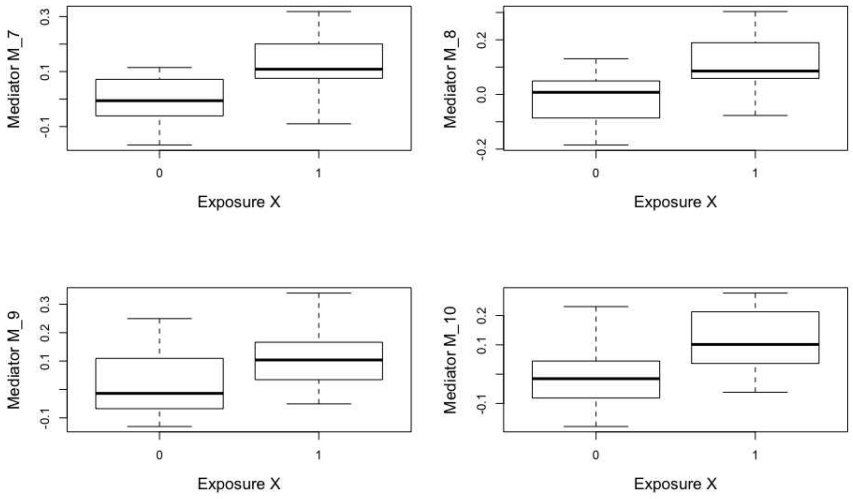


Figure 7.40: Box plots for exposure and Mediators $M_7$ - $M_{10}$ (V1), setting 3.2

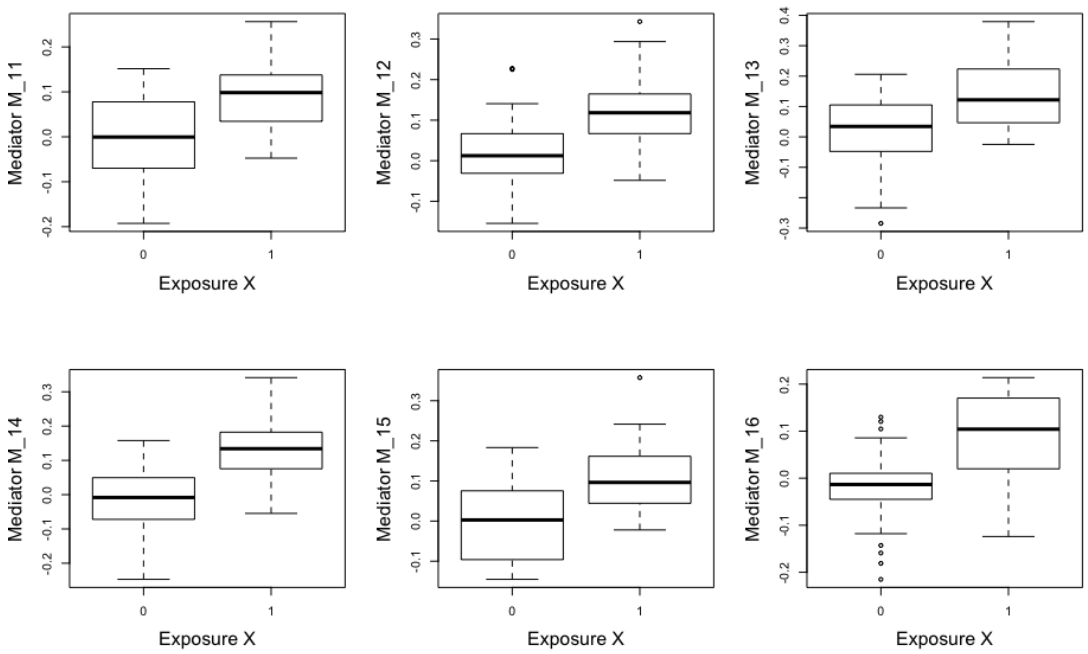


Figure 7.41: Box plots for exposure and Mediators $M_{11}$ - $M_{16}$ (V2), setting 3.2

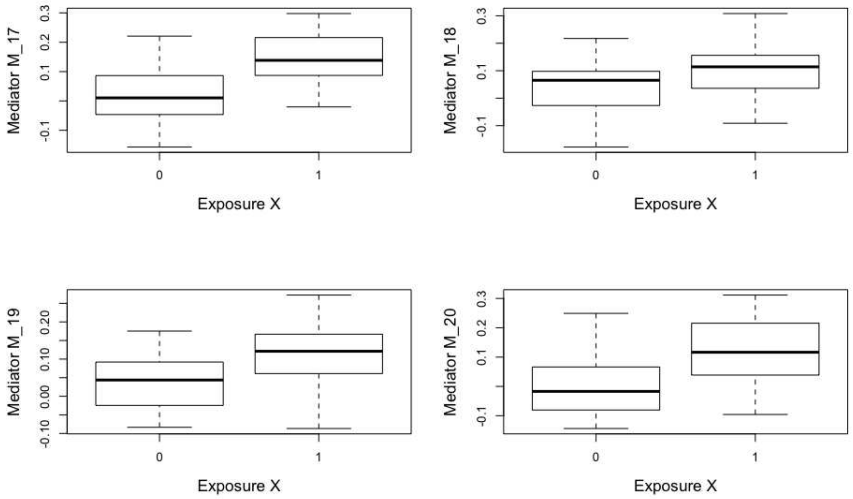


Figure 7.42: Box plots for exposure and Mediators $M_{17}$ - $M_{20}$ (V2), setting 3.2

### 7.2.9 Setting 3: Influence of exposure on survival time (check 4)

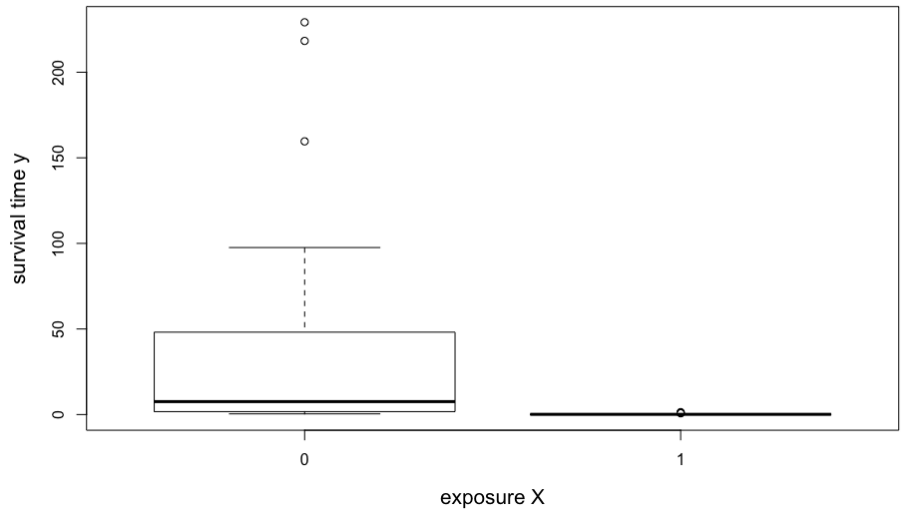


Figure 7.43: Box plot for influence of exposure $X$ on survival times $y$, setting 3.1

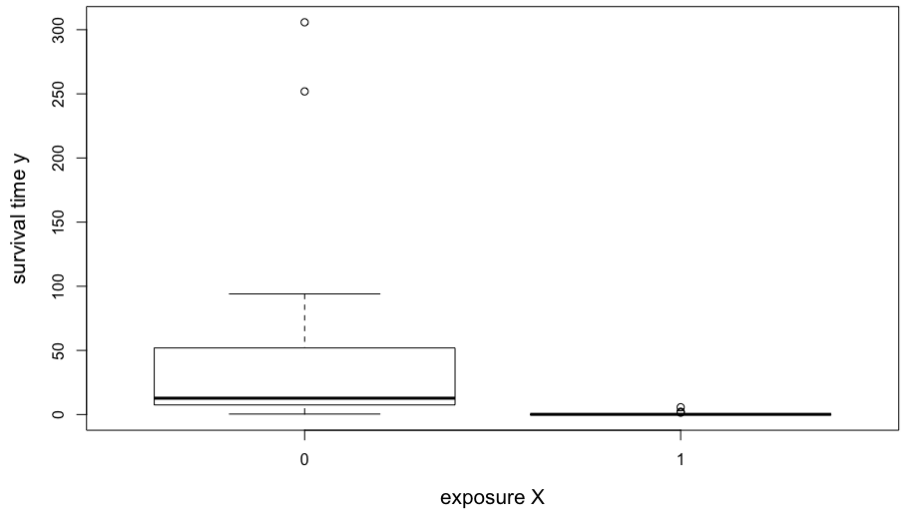


Figure 7.44: Box plot for influence of exposure $X$ on survival times $y$, setting 3.2

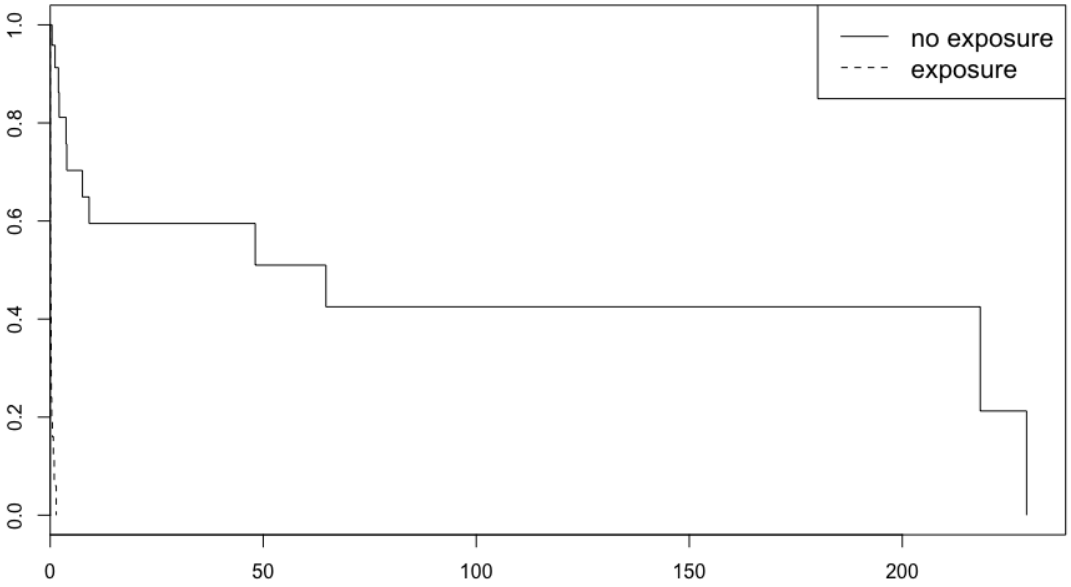


Figure 7.45: Survival curve, influence of exposure $X$ on survival times $y$, setting 3.1

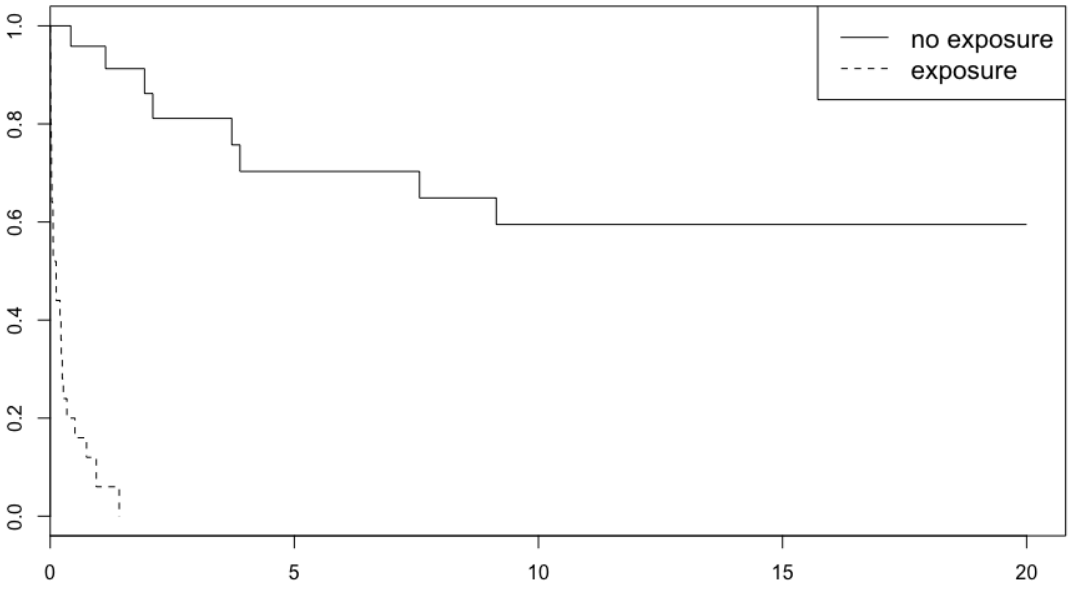


Figure 7.46: Survival curve, cut at $t = 20$, influence of exposure $X$ on survival times $y$, setting 3.1

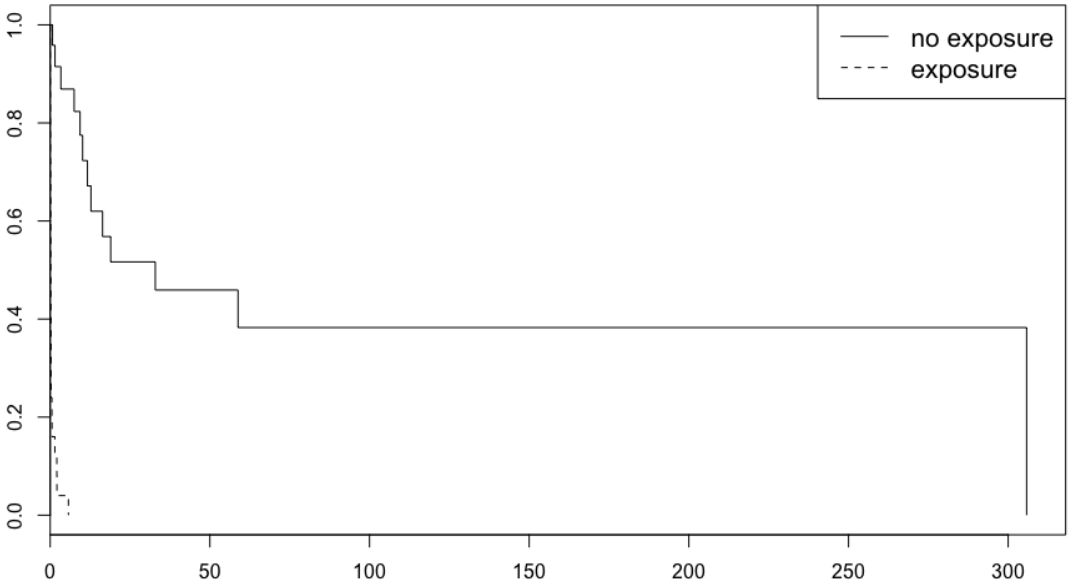


Figure 7.47: Survival curve, influence of exposure $X$ on survival times $y$, setting 3.2

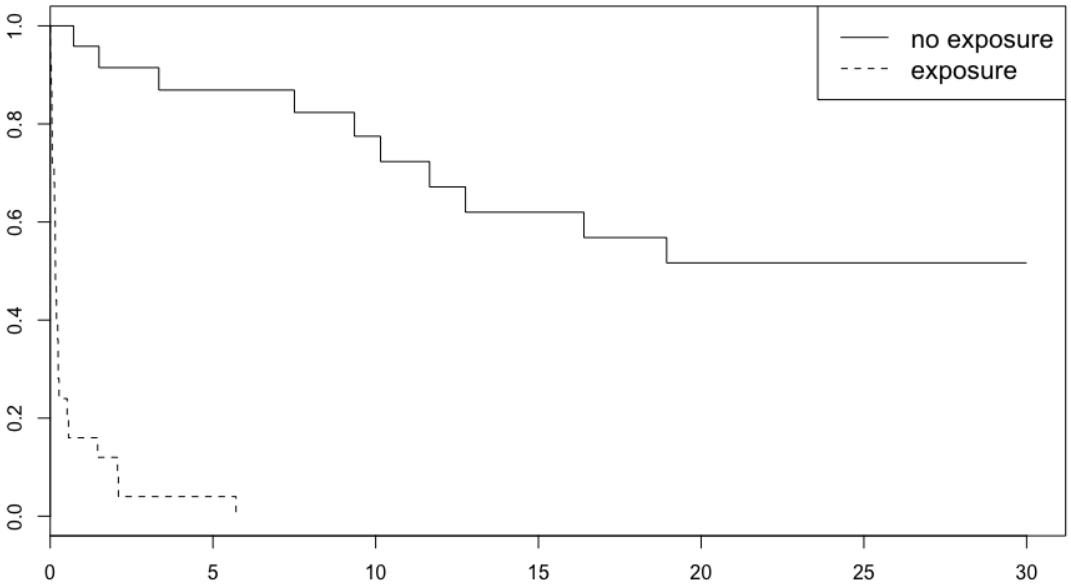


Figure 7.48: Survival curve, cut at $t = 30$, influence of exposure $X$ on survival times $y$, setting 3.2

# 8 Digital appendix

The USB stick contains the following digital appendix:

- **masters_thesis_Carina_Rein.pdf**
  A digital version of this thesis.
- **derivative_covariance.pdf**
  Contains the detailed description if the second derivative for the estimation of the covariance matrix using the oracle property of the MCP technique (cf. Fan and Li (2001)) and displays the idea of the implementation in R for function *covest()* in the package *himasurv*.
- **read_me.pdf**
  A guidance for the R codes used in this thesis.
- **install_dependencies.R**
  Install and load required packages.
- The folder `figures_thesis` includes:
  - **plots_penalization_MCP.R**
    Code to produce Figures 3.7 and 3.8 in section 3.4.
  - **plot_convexity_MCP.R**
    Code to produce Figure 3.9 in section 3.4.2.
- The folder `packages` includes:
  - Package **himasurv**
    Mediation Analysis in High-dimensional survival data
    including:
    * Function *himasurv()*: High-dimensional mediation analysis with survival data.

    - Function *metest()*: Univariate test for mediation effect with respect to one or several numerical variables in a survival setting.
  - Package **simulations**
    Simulation for 3 different dependency settings
    including:
    - Function *sim1()*, *sim2()* and *sim3()*
      Perform a simulation for high dimensional survival data with mediators for setting 1, 2 or 3.

- The folder `simulation` includes:
  - **simulation_1_1.R** & **simulation_1_2.R**
    Code to simulate data set 1 and data set 2 of setting 1 and to perform the required checks (cf. section 4.6.4).
  - **simulation_2_1.R** & **simulation_2_2.R**
    Code to simulate data set 1 and data set 2 of setting 2 and to perform the required checks (cf. section 4.6.5).
  - **simulation_3_1.R** & **simulation_3_2.R**
    Code to simulate data set 1 and data set 2 of setting 3 and to perform the required checks (cf. section 4.6.6).
  - **sim_set1_500.R** and **ws_sim_set1_500.RData**
    Code to simulate and analyze 500 data sets of setting 1 (cf. section 4.6.7). and the corresponding workingspace
  - **sim_set2_500.R** and **ws_sim_set2_500.RData**
    Code to simulate and analyze 500 data sets of setting 2 (cf. section 4.6.7). and the corresponding workingspace
  - **sim_set3_500.R** and **ws_sim_set3_500.RData**
    Code to simulate and analyze 500 data sets of setting 3 (cf. section 4.6.7). and the corresponding workingspace

# Bibliography

Alireza Abadi, Saeed Saadat, Parvin Yavari, Chris Bajdik, and Parvin Jalili. Comparison of aalen's additive and cox proportional hazards models for breast cancer survival: Analysis of population-based data from british columbia, canada. *Asian Pacific Journal of Cancer Prevention*, 12(11):3113–3116, 2011.

Rajender R. Aparasu and John P. Bentley. *Principles of Research Design and Drug Literature Evaluation*. Jones & Bartlett Learning, 2015.

Reuben M. Baron and David A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

Ralf Bender, Tomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24:1713–1723, 2005.

Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:437–507, 2012.

Patrick Breheny. *ncvreg: Regularization Paths for SCAD and MCP Penalized Regression Models*, 2017. R package version 3.9-1.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):231–253, 2011.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer-Verlag Berlin Heidelberg, 1 edition, 2011.

Richard B. Darlington and Andrew F. Hayes. *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*. Methodology in the Social Sciences. The Guilford Press, 1 edition, 2016.

L. Fahrmeir, T. Kneib, and S. Lang. *Regression*. Statistik und ihre Anwendungen. Springer Berlin Heidelberg, 2 edition, 2009.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Associations*, 96(456), 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, pages 29–36, 1982.

FE Jr Harrell, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Refression Modeling of Time-to-Event Data*. John Wiley & Sons, Inc., 2 edition, 2008.

Hermant Ishawaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

H. Ishwaran and U.B. Kogalur. *randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*, 2017.

James Jaccard and Robert Turrisi. *Interaction Effects in Multiple Regression, 2nd Ed. (Quantitative Applications in the Social Sciences)*. Quantitative Applications in the Social Sciences 72. Sage Publications, Inc, 2nd edition, 2003.

Charles M. Judd and Davis A. Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Research*, 5:602–619, 1981.

Betty R. Kirkwood and Jonathan A. C. Sterne. *Medical Statistics*. Birkwell Publishing, 2 edition, 2003.

Franziska Küster-Rohde. *Die Wirkung von Glaubwürdigkeit in der Marketingkommunikation*. Gabler, 2010.

Theis Lange and Jørgen V. Hansen. Direct and indirect effects in a survival context. *Epidemiology*, 22(4):5754581, 2011.

Michael LeBlanc and John Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993. ISSN 01621459.

Xian Liu. *Survival Analysis*. Higher Education Press, 2012.

David P. MacKinnon. *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, 2008.

Byung-Ho Nam and Ralph B. D'Agostino. Discrimination index, the are under the roc curve. In *Goodness-of-Fit Tests and Model Validity*, chapter 20. Birkhäuser, 2002.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Michael J. Pencina and Ralph B. D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58(1):267–288, 1996.

Gerhard Tutz. *Regression for Categorical Data*. Cambridge University Press, 2012.

Tyler VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009.

Pierre J. M. Verweij and Hans C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314, 1993.

Pierre J. M. Verweij and Hans C. Van Houwelingen. Penalized likelihood in cox regression. *Statistics in Medicine*, 13(23-24):2427–2436, 1994.

Marlies Wakkee, Loes M. Hollestein, and Tamar Nijsten. Multivariable analysis. *Journal of Investigative Dermatology*, 134(5):1 – 5, 2014.

Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2017. URL https://github.com/hadley/devtools. R package version 1.12.0.9000.

Hadley Wickham, Peter Danenberg, and Manuel Eugster. *roxygen2: In-Line Documentation for R*, 2017. URL https://CRAN.R-project.org/package=roxygen2. R package version 6.0.1.

C. H. Zhang. Nearly unbiased viariable selection under minimax concave penality. *The Annals of Applied Statistics*, 38:894–942, 2010.

Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, Pantel Vokonas, Lihui Zhao, Jinchi Lv, Andrea Baccarelli, Lifang Hou, and Lei Liu. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150, 2016.

Yinan Zheng, Haixiang Zhang, Zhou Zhang, Lifang Hou, and Lei Liu. *HIMA: High-Dimensional Mediation Analysis*, 2017. R package version 1.0.4.