MASTER'S THESIS

Interuniversity Master in Statistics and Operations Research UPC-UB

Title: Modeling pseudo-observations with covariate dependent censoring: robustness of the method against misspecified censoring models.

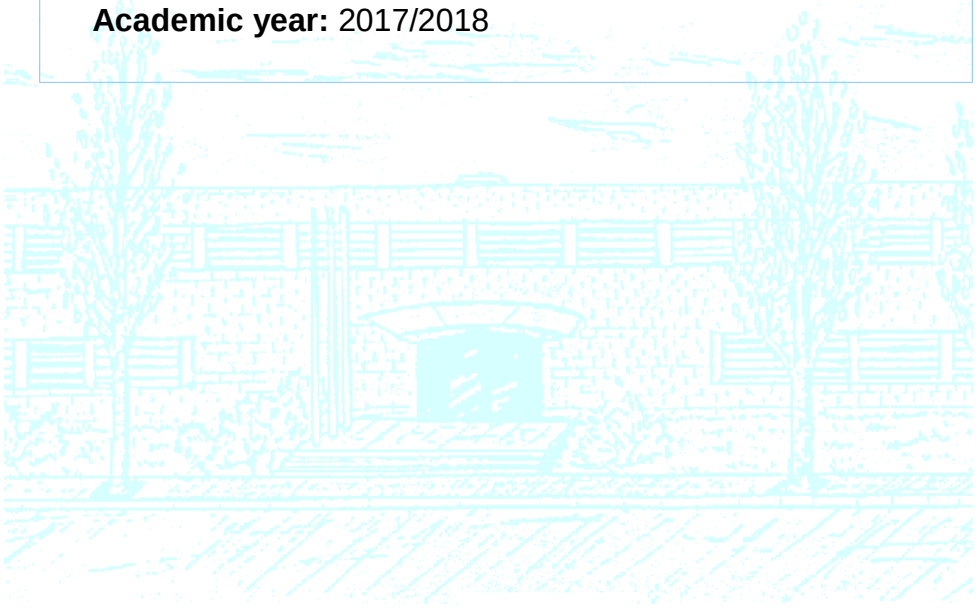
Author: Andreu Schoenenberger López

Advisors: Klaus Langohr / Per Kragh Andersen

Department: Statistics and Operations Research (UPC) / Department of Biostatistics (UCPH - KU)

University: Universitat Politècnica de Catalunya (UPC) / Københavns Universitet (KU)

Academic year: 2017/2018



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA
DE CATALUNYA

UNIVERSITY OF
COPENHAGEN

Universitat Politècnica de Catalunya (UPC)
Facultat de Matemàtiques i Estadística (FME)

University of Copenhagen (UCPH)
Faculty of Health Sciences
Institute of Public Health
Department Of Biostatistics

Master's Thesis

Modeling pseudo-observations with covariate dependent censoring: robustness of the method against misspecified censoring models

Andreu Schoenenberger López

Directed by (UCPH):
Per Kragh Andersen

Supervised by (UPC):
Klaus Langohr

January 9, 2018

Dedicated to all people
I met during my stay in Copenhagen
and special thanks to Per for his time,
help and teaching.

"The journey is more important than the end or the start"

Abstract

The so called pseudo-observations in survival analysis were introduced by recent studies that reviewed this method when estimating different parameters using regressions models (Andersen and Perme, Stat. Meth. Med. Res., 2010) with the condition that the censoring distribution is independent from covariates. If censoring depends on covariates, the method based on pseudo-observations requires modeling of the censoring distribution, which leads to the construction of alternative estimators based on censoring probability weighting. This master thesis will present the proposal of Andersen and Perme and – by means of Monte Carlo simulation – will also study its robustness if the model for the censoring distribution is misspecified. Two alternative estimators will be explained and used for the study of robustness of the method: the Cumulative Incidence Function and the Restricted Mean Lifetime.

Keywords: Survival analysis, Cox Model, Dependent censoring, Pseudo-values, Monte Carlo Simulation, Cumulative Incidence Function, Restricted Mean Lifetime

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context of this study

This study is founded on the context of survival analysis (with competing risks) as well as the modeling techniques available for this time-to-event data with the important (and widely spread) remark of right-censored data, that is, we only know *partially* this time-to-event data.

Censored data is usually always a problem for three main reasons: the first and most obvious one is that we partially lose information about (usually) several time-to-event points, which is lowering our statistical power when computing point or interval estimates with different techniques or models. Secondly is that the most generalized statistical tools for survival analysis have the assumption that censoring is independent from the covariates (Kaplan-Meier, Aalen-Johansen, ...). A violation of this assumption can introduce a bias in our point and interval estimates and ultimately lead to wrong results or conclusions about a certain clinical trial, study, etc. This is particularly important when we have a high load of censored time points in our data. Ultimately, censoring is also a problem when our aim is to do a linear regression model using an outcome which can be the survival time itself or a function of that survival time $T_i \rightarrow f(T_i)$.

Thus, (P.K. Andersen et al., 2003; P.K. Andersen & Pohar Perme M., 2010) proposed *pseudo-observations* as a way of using same survival analysis tools or methods when right censoring is present. These pseudo-observations can be used to solve the generalized estimating equations in a regression model. N. Binder et al. (2014) expressed the ways of defining pseudo-values in situations where censoring depends on covariates because if we don't have independent censoring, Kaplan-Meier is not consistent, that was the motive for finding alternative estimators. However, our aim is to study if the robustness still remains when the model for censoring is misspecified based on different scenarios and using the Cumulative Incidence Function and the Restricted Mean Lifetime.

The main advantage of this study is that results from Monte Carlo simulations can be easily compared between them, giving us a clear view of the possible robustness of the method, as well as the situations where we might have biased results. Nevertheless, the simulation is not straight forward and it requires high knowledge of *R* (https://cran.r-project.org/). Also, is very time-consuming even with parallelized code and using 48 CPUs computational servers.

1.2 Structure of this report

This study is structured in 5 main chapters, the first 4 of them contain different sections.

Chapter 2: aims to explain the theoretical background used in this study which is split in 2 basic sections:

- Inverse Probability of Censoring Weighted Estimators: will present and explain the alternative estimators that will take into account the covariate-dependent censoring in the data. These are the censoring weighted cumulative incidence function and restricted mean lifetime, which, as a point in common, are weighted by the inverse of the probability of not being lost to follow-up before a given time t and for a given value of the covariates (conditional censoring).
- Pseudo-observations: Explanation of this method, the assumptions and the advantages of using it.

Chapter 3: first of the two practical and Monte Carlo simulation-based parts of this study. Having the theoretical background, the estimating equations for the cumulative incidence using pseudo-observations will be used and then applied to the simulation where the study of robustness of the method respect to the specification of the censoring model will be done.

Chapter 4: this second Monte Carlo simulation-based part will have a similar structure as Chapter 3 but in this case we will look at the behavior of the restricted mean lifetime when the censoring model is misspecified. The main difference with Chapter 3 is that in this case we need to calculate the link function between the restricted mean lifetime and the linear predictor, as well as using the appropriate algorithm to fit a GEE model specifying a user-defined link function, inverse link function, first derivative of the inverse link function and the variance function.

Chapter 5: will explain and show the results from the two previous simulations.

Chapter 6: will explain the conclusions and open the discussion based on the previous shown results.

Chapter 2

Theoretical Background

2.1 Inverse Probability of Censoring Weighted Estimators (IPCW Estimators)

Standard non-parametric estimator (Kaplan-Meier, Aalen-Johansen, ...) can not only handle censoring but they also have (under random censoring) an exact representation as inverse probability of censoring weighted estimators.

In a competing risks setup with i number of individuals ($i = 1, ..., n$), j number of endpoints ($j = 1, 2, 3, ..., k$) and T_i as the *true* survival times ($T_i = T_1, ..., T_n$, which are defined as the *true* time from 0 to the first possible event j), using the notation used by N. Binder et al. (2014), let's define the following (where I is the *indicator function*):

- Right censoring times U_i, observed times $\tilde{T}_i$ and censoring indicator Δ_i for subject i:

  $$U_i = U_1, ..., U_n$$

  The observed times $\tilde{T}_i$ then are defined as,

  $$\tilde{T}_i = T_i \text{ if } T_i \leq U_i \text{ and } \tilde{T}_i = U_i \text{ if } U_i \leq T_i$$

  Then the censoring indicator (0 for censored observation and 1 otherwise),

  $$\Delta_i = I(T_i \leq U_i)$$

- Number of observed cause j events for a given subject i between 0 and t:

  $$N_{ij}(t) = I(\tilde{T}_i \leq t; D_i = j; \Delta_i = 1)$$

- Number of subjects still at risk at time t (or count number of observed times $\tilde{T}_i$ greater or equal than t):

  $$Y(t) = \sum_{i=1}^{n} I(\tilde{T}_i \geq t)$$

2.1.1 Cumulative Incidence Function (CIF)

Using the previous notation, Cumulative Incidence Function (CIF) can be defined as:

$$F_j(t) = P(T \leq t, D = j) = E(I(T \leq t, D = j))$$

As an example of this Cumulative Incidence Function the following figure is shown with 2 possible endpoints or events from the competing risks setup:

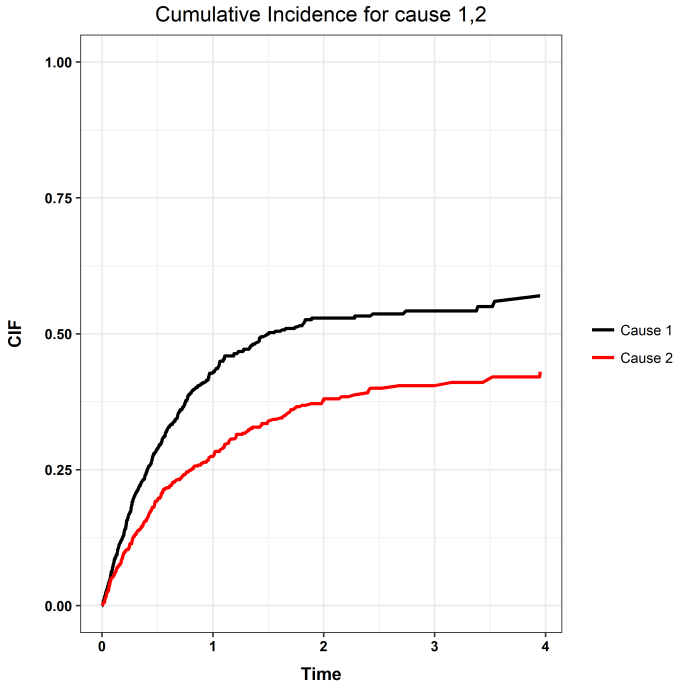


Figure 2.1: Example of Cumulative Incidence Function (step function) for causes $j = 1, 2$.

Competing risks data generated with the algorithm described in section 3.3

CIF is giving us the probability that an event of type j cause has occurred before or at time t. We start by defining the Kaplan-Meier estimator (Kaplan & Meier, 1958):

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{N_i}{Y_i}\right)$$

Then, in a non-parametrical way using the Aalen-Johansen estimator for a given cause j:

$$\hat{A}_j(t) = \int_0^t \frac{\sum_{i=1}^n dN_{ij}(u)}{Y(u)}$$

The CIF is defined as (Aalen, 1978):

$$\hat{F}_j(t) = \int_0^t \hat{S}(u-) d\hat{A}_j(u) = \int_0^t \hat{S}(u-) \frac{\sum_{i=1}^n dN_{ij}(u)}{Y(u)} \tag{1}$$

Where $\hat{S}(u-)$ is the probability of having an event just before time u (thus notation $u-$ is used). We can also use Gill's relation (Gill, 1980) between the number of subjects still at risk at time t, the total number of subjects, the survival probability and the estimated censoring probability which is defined as the product limit estimator for $C_0(t)$: $\hat{C}_0(t)$. This estimator is giving us the (marginal or unconditional) probability of not being lost to follow-up (being uncensored) just before time t. If we define $N^c = I(\tilde{T}_i \leq t, \Delta_i = 0)$ (analogously to $N_{ij}(t)$), then:

$$\frac{Y(t)}{n} = \hat{S}(t) \prod_{s \leq t} \left(1 - \frac{N^c(ds)}{Y(s) - N(ds)}\right) = \hat{S}(t)\hat{C}_0(t)$$

$$Y(t) = n\hat{S}(t)\hat{C}_0(t) \tag{2}$$

Following the work done by (N.Binder et al, 2014), the Aalen-Johansen estimator in this case can be represented, using (1) and (2):

$$\hat{F}_j(t) = \sum_{i=1}^{n} \int_0^t \hat{S}(u-) \frac{N_{ij}(du)}{n\hat{S}(u-)\hat{C}_0(u-)} = \frac{1}{n} \sum_{i=1}^{n} \frac{N_{ij}(t)}{\hat{C}_0(\tilde{T}_i-)} \tag{3}$$

The previous estimator is consistent when censoring is independent from covariates (Aalen, 1978), but for the case of covariate-dependent censoring, Aalen-Johansen estimator is biased, giving us a biased estimate for $\hat{F}_j(t)$ (N.Binder et al., 2014), also shown by Andersen and Pohar Perme (2010) for the marginal survival function. The proposed solution to obtain an un-biased estimator for $\hat{F}_j(t)$ when the censoring is depending on covariates is to estimate $\hat{C}(t|Z)$ which is the censoring distribution conditional on covariates. Then, if we are able to do a consistent estimation for $C(t|Z)$ (which requires a correct specification of the model for censoring), we should also obtain a consistent estimation for $\hat{F}_j(t)$.

Thus, the modified and censoring weighted estimator that accounts for covariate-dependent censoring is (N.Binder et al., 2014):

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{N_{ij}(t)}{\hat{C}_i(\tilde{T}_i - |Z_i)} \tag{4}$$

2.1.2 Restricted Mean Lifetime (RML)

The restricted mean lifetime or restricted mean survival time can be defined as the τ-year life expectancy: if we set a horizon time $\tau > 0$, RML is the expected survival time between 0 and the horizon time τ. Mathematically:

$$RML = \mu_\tau = E[min(T, \tau)] = \int_0^\tau S(t)dt \tag{5}$$

An example of the Restricted Mean Lifetime (depending on the survival function as said before) is shown below:

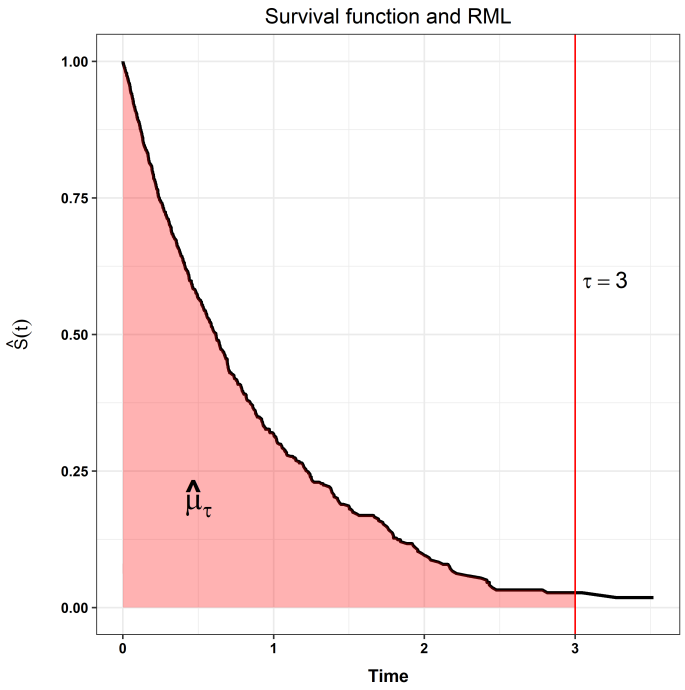


Figure 2.2:
Example of Survival function (step function) as a black line and Restricted Mean Lifetime (red colored area). Data generated with the algorithm described in section 3.3 and using the *prodlim* package for *R*

Following with the same context of a competing risk set-up as before, we now need to calculate the alternative estimator for the survival function $\hat{S}(t)$ in order to integrate (or sum in a discrete case) and compute the RML. Following this thought we can express the survival function, in a competing risk setup, as the complementary of the sum of the probabilities of having an event (cumulative incidence) at time t from all causes of failure ($j = 1 to j = k$):

$$\hat{S}(t) = 1 - \sum_{j=1}^{k} \hat{F}_j(t)$$

And taking the alternative estimator for $\hat{F}_j(t)$ in equation (4):

$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{N_{ij}(t)}{\hat{C}_i(\tilde{T}_i - |Z_i)} \tag{6}$$

In equation (6) is important to notice how the competing risks setup is not so relevant anymore, since we are summing over all types of failures and then the operation $\sum_{j=1}^{k} N_{ij}(t)$ will be equal to 1 for any type of event (from cause $j = 1$ to $j = k$ between 0 and t) and 0 if individual i is censored. It's also the natural way of looking at the survival function since if we compute $\hat{S}_{j=1}(t)$ we are obtaining the probability of either being alive at time t or having a cause $j = 2, ..., k$ event, so we are not really clarifying the probability of having an event for the cause of interest.

Because of this a competing risks setup doesn't make much sense in this case, so to simplify the formulation and the future simulation, we will join all the j causes. Thus, the resulting alternative estimator for the survival function is:

$$\hat{S}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{N_i(t)}{\hat{C}_i(\tilde{T}_i - |Z_i)} \tag{7}$$

Using equations (5) and (7):

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt \tag{8}$$

Where $\hat{S}(t)$ is the alternative estimator (using the inverse probability of not being lost to follow-up before a given time t as weights) for the survival function (equation 7).

2.2 Pseudo-observations

Pseudo-observations or *Pseudo-values* are based on a resampling technique called Jackknife (Efron & Stein, 1981). Jackknife allows us to estimate the bias of an estimator over the entire sample, hence we can recalculate the estimator corrected for bias:

- Calculate the all-sample estimator $\hat{\theta}$
- Calculate all the leave-one-out estimators $\hat{\theta}^{(-i)}$ for $i = 1, ..., n$ where n is the number of subjects/observations/rows of the data, then compute the average:

$$\hat{\theta}^{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^{(-i)}$$

- The bias-corrected Jackknifed estimator $\hat{\theta}^{(corr)}$ will be:

$$\hat{\theta}^{(corr)} = n\hat{\theta} - (n-1)\hat{\theta}^{(.)}$$

Now if we are interested in only a single correction, one for each subject/observation/row, then instead of doing the average over all the leave-one-out estimates, we can subtract the i-th leave-one-out estimate to the all-sample estimate to obtain the bias-corrected estimate for subject/observation/row i:

$$\hat{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)} \tag{9}$$

Supposing that we define some function of the survival time $T \rightarrow f(T)$, it's important to notice that if we don't have censoring in our data the expectation $\theta = E(f(T))$ can be estimated using simply the average $\frac{1}{n}\sum_{i=1}^{n} f(T_i)$ and the i-th pseudo-observation will be $f(T_i)$ (N. Binder et al., 2014).

Case for the Cumulative Incidence: Pseudo-values are (Klein & Andersen, 2005):

$$\hat{F}_{ij}(t) = n\hat{F}_j(t) - (n-1)\hat{F}_j^{(-i)}(t) \tag{10}$$

Then the generalized estimating equations (GEE) take the form:

$$U_n(\boldsymbol{\beta}\,') = \sum_{i=1}^{n} V(t)[\hat{F}_{ij}(t) - g(t, \boldsymbol{\beta}\,', Z_i)] = 0 \tag{11}$$

Following this path, (Graw et al., 2009) showed the following property which is necessary for the consistency for the solution of the GEE:

$$E(\hat{F}_{ij}(t)|Z_i) = E(I(T_i \leq t, D_i = j)|Z_i) + O_p(1)$$

However, it relies on the assumption that censoring is independent from covariates:

$$P(U > t|Z = z) = P(U > t) = C_0(t) > 0$$

That's why previously we focused on the alternatives estimators for the CIF and RML, so that pseudo-values are a “correction" for the original values when we have dependent censoring. However, for the pseudo-values to work in this case and as we will see later, the censoring model needs to be correctly specified.

Chapter 3

Monte Carlo Simulation for the Cumulative Incidence Function (CIF)

The following indexes will be used throughout this chapter:

- $i = 1, ..., n$ as an indicator for the subject.
- $j = 1, 2, 3, ...$ refers to the cause. In this case the cause of interest will be $j = 1$.
- $k = 1, ..., n-1$ is the leave-one-out index, meaning that it won't have the index i or it has to be different than index i.
- $m = 1, 2, 3, ...$ refers to the time-points where an event of cause $j = 1$ has happened.
- To simplify notation, if the cumulative incidence function is specified as $F_1(t)$, it is assumed that the proper time-points for cause 1 events are used.
- Definitions from *Theoretical Background* still apply.

3.1 Estimating Equations for the CIF

Following the *Theoretical Background* chapter and the work from N.Binder et al (2014), we can now calculate the pseudo-observations in order to calculate later the link between the pseudo-values for the CIF and our linear predictor. Taking equation (4) as the alternative estimator and a Cox model for the censoring distribution:

$$C_i(t|Z_i) = e^{-B(t)e^{\boldsymbol{\gamma}'Z_i}} = \exp(-B(t)\exp(\boldsymbol{\gamma}'Z_i))$$

We obtain:

$$\hat{F}_j(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{N_{ij}(t)}{exp(-\hat{B}(T_i)exp(\hat{\boldsymbol{\gamma}}'Z_i))} \tag{12}$$

Therein $B(t)$ is the cumulative baseline censoring hazard and $\boldsymbol{\gamma}'$ is the vector of coefficients for each covariate included. In this case $B(t)$ and $\boldsymbol{\gamma}'$ would be estimated with the whole sample. But calculating the value $\hat{F}_j^{(-i)}(t)$ is more complicated since we have different alternatives for estimating (or not estimating) the cumulative baseline censoring hazard $\hat{B}(t)$ and the covariates coefficients $\hat{\boldsymbol{\gamma}}'$. To summarize, our options are:

- Re-estimate the values for $\hat{B}(t)$ and $\hat{\boldsymbol{\gamma}}'$ each time we remove the subject i from the sample, that is, re-fitting the Cox model for the censoring distribution from $i = 1$ to n.
- Keep the estimates from the whole sample for $\hat{\boldsymbol{\gamma}}'$ and re-estimate again $\hat{B}(t)$ each time we remove an individual.
- Keep always the model with the whole sample and then use the previously estimated $\hat{B}(t)$ and $\hat{\boldsymbol{\gamma}}'$ for each i.

Since we can have an explicit formula for the cumulative baseline censoring hazard – when $\boldsymbol{\gamma}'$ coefficients are known (or estimated) – called the Breslow estimator, then keeping the estimated $\hat{\boldsymbol{\gamma}}'$ effects from the whole sample and re-estimating $\hat{B}(t)$ each time is the most effective solution in terms of accuracy and computational time or effort. Analysis of computational time and bias was also done by (N.Binder et al, 2014) and showed that using the Breslow estimator was the best choice, which is defined as:

$$\hat{B}^{(-i)}(t) = \int_0^t \frac{\sum_{k \neq i} dN^c(u)}{\sum_{k \neq i} Y_k(u) exp(\hat{\boldsymbol{\gamma}}' Z_k)}$$

Then the leave-one-out estimator for calculating the pseudo-observations can be computed as following:

$$\hat{F}_j^{(-i)}(t) = \frac{1}{n-1} \sum_{k \neq i} \frac{N_{kj}(t)}{exp(-\hat{B}^{(-i)}(\tilde{T}_k -) exp(\hat{\boldsymbol{\gamma}}' Z_k))} \tag{13}$$

Combining equations (12) and (13) we can compute the pseudo-observations in the usual way using formula (10).

Now let's suppose a competing risks scenario with at least 2 causes of failure and with right-censored data. Our cause of interest will be cause number 1. Then we want to be able to generate data that follows the Fine and Gray model (Fine & Gray, 1999) for the cause of interest which has the following cumulative incidence function:

$$F_1(t|Z) = 1 - exp(-\Lambda_0(t) exp(\boldsymbol{\beta}' Z)) \tag{14}$$

Where $\Lambda_0(t)$ is the cumulative baseline subdistribtuion hazard for cause 1 defined as:

$$\Lambda_0(t) = \int_0^t \lambda_0(u) du$$

$\lambda_0(t)$ is the subdistribtuion hazard function for cause 1 that is:

$$\lambda_0(t) = \frac{p \cdot exp(-t)}{1 - p(1 - exp(-t))}$$

Where p is the marginal probability (not individual) of having a cause 1 event.

(Klein & Andersen, 2005) proposed the following for modeling the CIF in a competing risks setup when using pseudo-observations, for cause 1:

$$\phi(F_{i,j=1}(t_m)) = \alpha_m + \boldsymbol{\beta}' Z_i$$

Where ϕ is a proper link function that can be either a logit or a complementary log-log function. In our case, since we are interested in applying a Fine and Gray model and therefor having an assumption of proportional subdistribution hazards, we will apply the complementary log-log function. On the other hand, a logit link function would lead to a proportional odds model. Operating equation (14) we can find an explicit form for the complementary log-log link function:

$$1 - F_1(t|Z) = \exp(-\Lambda_0(t)\exp(\boldsymbol{\beta}'Z))$$

Applying logarithms for the first time:

$$-\log(1 - F_1(t|Z)) = \Lambda_0(t)\exp(\boldsymbol{\beta}'Z)$$

And for the second time:

$$\log(-\log(1 - F_1(t|Z))) = \log(\Lambda_0(t)) + \boldsymbol{\beta}'Z \tag{15}$$

Which is consistent with the model $\phi(F_{i,j=1}(t_m)) = \alpha_m + \boldsymbol{\beta}'Z_i$ proposed.

Using this link function we can use equation (11) to solve the generalized estimating equations (using *R* with a proper package such as *geepack* (https://cran.r-project.org/web/packages/geepack)).

3.2 Scenarios and Misspecifications

We are mainly interested in seeing what happens to point estimates and the standard deviations from the GEE model for the CIF when we are in two basic situations: the model is misspecified by missing a covariate or the model is misspecified when the functional form is not the correct one. Furthermore, we know that dependency between covariates can be an important factor when estimating from the GEE model, that's why we duplicated the scenarios so we can compare the robustness of the pseudo-observations method when we have correlation between covariates.

In that path, the following table presents the 10 scenarios proposed for this study, where the second column will be the original model (simulated) and the third column specifies the covariates and the functional form in the model fitted to obtain the censoring probabilities. Two covariates are generated: a binary one (Z_1) with a certain probability $p = 0.25$ and a continuous one (Z_2) with a uniform distribution. The 10 scenarios can be extended in a large number of ways, but in order to keep the simulation as simple as possible we decided to only generate two with different distributions and possible dependency between them to compare the robustness of the method.

Scenario	Model for censoring (Cox model)	Fitted model for cens.	Description
0	Independent censoring	$Z_1 + Z_2$	Control Scenario
1	$0.5Z_1 + 0.5Z_2$ (Z_1, Z_2 indep.)	$Z_1 + Z_2$	Good model fitted
2	$0.5Z_1 + 0.5Z_2$ (Z_1, Z_2 indep.)	Z_1	Missing covariate when Z_1, Z_2 indep.
3	$0.5Z_1 + 0.5Z_2^*$ (Z_1, Z_2^* dep.)	$Z_1 + Z_2^*$	Good model fitted
4	$0.5Z_1 + 0.5Z_2^*$ (Z_1, Z_2^* dep.)	Z_1	Missing covariate when Z_1, Z_2^* dep.
5	$0.5Z_1 + 0.5Z_2^2$ (Z_1, Z_2 indep.)	$Z_1 + Z_2^2$	Good model fitted
6	$0.5Z_1 + 0.5Z_2^2$ (Z_1, Z_2 indep.)	$Z_1 + Z_2$	Missing functional form when Z_1, Z_2 dep.
7	$0.5Z_1 + 0.5Z_2^2$ (Z_1, Z_2 indep.)	Z_1	Missing covariate when Z_1, Z_2 dep.
8	$0.5Z_1 + 0.5(Z_2^*)^2$ (Z_1, Z_2^* dep.)	$Z_1 + (Z_2^*)^2$	Good model fitted
9	$0.5Z_1 + 0.5(Z_2^*)^2$ (Z_1, Z_2^* dep.)	$Z_1 + Z_2^*$	Missing functional form when Z_1, Z_2^* dep.
10	$0.5Z_1 + 0.5(Z_2^*)^2$ (Z_1, Z_2^* dep.)	Z_1	Missing covariate when Z_1, Z_2^* dep.

Table 3.1: 10 Scenarios for the Monte Carlo simulations for CIF.

These scenarios will be repeated in different cases that are enumerated here:

- **Two different sample sizes**, one with 500 individuals in total (counting censoring, cause 1 events and cause 2 events) and the other one with 200 individuals in total.
- We will do the simulations with two levels of censoring, the first one will be **moderate censoring** where approximately 30-33% of the data will be right-censored. The second one will be **heavy censoring** and will contain approximately 65-67% of censoring.

This will allow us to also see the robustness of the method when we misspecify the censoring model and we also have a reduced sample size or a fairly amount of censored data.

It's important to notice that scenario 0 is used as a control since censoring is independent from covariates Z_1 and Z_2 (or Z_2^*). If the method is consistent we should get the same results using the censoring weighted estimator and the Aalen-Johansen estimator. In other words, random censoring should not affect the modified or alternative estimator if we want a robust method.

3.3 Simulation Algorithm

The Monte Carlo simulation for the CIF is based, in a general basis, in a competing risks setup with $j = 2$ and right censored data with $j = 1$ as the cause of interest. This simulation can be split in 4 main parts:

1. **Generate the data:** this includes time to either cause 1, cause 2 or being censored; a status indicator (0 for censoring, 1 for cause 1 and 2 for cause 2) and the two (possibly dependent) covariates Z_1 and Z_2 or Z_2^*.
2. **Compute pseudo-observations:** based on the method explained in *Theoretical introduction*, we generate either the censoring-adjusted pseudo-values using the alternative estimator for the CIF or the not-censoring-adjusted pseudo-values using the Aalen-Johansen estimator.

3. **Fit of the GEE model:** with the linear predictor and the pseudo-values as the outcome, we fit a linear regression and obtain the point estimates and robust standard errors (Huber-White standard error estimates).

4. **Get the estimates:** since we are doing a Monte Carlo based simulation we have to repeat the previous steps n times (in our case $n = 500$) and computing the average we get the final point and robust-interval estimates.

3.3.1 Data generation

Generating covariates

The two covariates are generated using the following distributions and conditions:

$$Z_1 \sim Binomial(n = 1, p = 0.25)$$

$$\text{If } Z_2 \text{ dependent of } Z_1 \rightarrow Z_2 \sim Unif[-1, 1]$$

$$\text{If } Z_2 \text{ dependent of } Z_1 \rightarrow Z_2^* \sim \begin{cases} Unif[-1, 0] & \text{If } Z_1 = 0 \\ Unif[0, 1] & \text{If } Z_1 = 1 \end{cases}$$

Generating censoring times

Next step is to simulate the censoring times based on a Cox model with the covariates we just generated:

$$C(t|Z) = \exp(-B(t)\exp(\gamma Z))$$

With the corresponding survival function:

$$S_c(t|Z) = \exp(-H_0(t)\exp(\boldsymbol{\gamma}\,'Z))$$

where:

$$H_0(t) = \int_0^t B(u)du$$

So, it is direct to see that the cumulative distribution function (cdf) denoted as G for the Cox model is:

$$G(t|Z) = 1 - \exp(-H_0(t)\exp(\boldsymbol{\gamma}\,'Z))$$

Let T be the randomly distributed censoring times for the Cox model, then from the previous equation we have (keeping in mind that $G(t|Z)$ is a cdf and if $U \sim Unif[0, 1]$ then
$1 - U \sim Unif[0, 1]$):

$$U = \exp(-H_0(T)\exp(\boldsymbol{\gamma}\,'Z)) \sim Unif[0, 1]$$

Furthermore, if $H_0(T) > 0$ for all positive $T \neq 0$, the previous expression can be inverted:

$$T = H_0^{-1}(T)(-\log(U)\exp(-\boldsymbol{\gamma}\,'Z))$$

So, considering an exponential distribution for the censoring times, we can have a constant baseline hazard function which will allow us to simplify our coding and computational times. Then knowing that in this case $H_0(t) = B \cdot t$, the inverse of this relation is direct: $H_0(t)^{-1} = B^{-1} \cdot t$. Finally, times T will be:

$$T = B^{-1}(-\log(U)\exp(-\boldsymbol{\gamma}\,'Z)) = -\frac{\log(U)}{B \cdot \exp(\boldsymbol{\gamma}\,'Z)}$$

Which will follow an exponential distribution with scale parameter $\lambda = B \cdot \exp(\boldsymbol{\gamma}\,'Z)$. Keep in mind that the censoring times will be generated differently in 5 cases (scenarios 0, 1-2, 3-4, 5-7, 8-10) since the true model is also different. The coefficients will be invariant during the whole simulation for the CIF and the censoring baseline hazard will change our censored sample size due to the selection process between censored observation, cause 1 or cause 2 observation (that will be seen later on this chapter), which by definition will also change our distribution: if we want longer censored times we have to increase the value of the constant censoring baseline hazard.

Generating cause1 & cause2 survival times

A binomial experiment is used to choose whether a single individual will have an event of type 1 or type 2. But this decision is also constraint by the fact that the probability of having cause 1 or cause 2 of failure has to be consistent with the Fine and Gray model (Fine $ Gray, 1999). In other words, since cumulative incidence can also be defined as the probability of having an event before a given time t, we can define the probability of having an event of type 1 as the cumulative incidence directly. However, we have to make the distinction between the marginal probability of having that type 1 event and the individual probability of having the type 1 event given the covariates that the individual has. Following that thought, marginal probability of having a cause 1 failure time can be seen as the cumulative incidence without conditioning on covariates and at time $t \to \infty$. Furthermore, since it doesn't make sense that the individual probability depends on time, hence has to be seen as a function of the marginal probability ($t \to \infty$) and the value of the covariates that the individual has, combined with the fact that the Fine and Gray model has to apply.

- Marginal probability of having a cause 1 failure time:

  $$p = 1 - \exp(-\Lambda_0(\infty))$$

  This probability can also be seen as an overall frequency of cause 1 events, meaning that we can define that value arbitrarily depending on how many cause 1 events we want to have.

- Individual probability of having a cause 1 failure time, from equation (15), where the effect of the marginal probability is seen as $\Lambda_0(\infty)$:

  $$cloglog(1 - p_i) = \log(-\log(1 - p_i)) = \log(\Lambda_0(\infty)) + \boldsymbol{\beta}\,'Z_i$$

  Applying exponentials:

  $$-\log(1 - p_i) = \exp(\boldsymbol{\beta}\,'Z_i) \cdot \Lambda_0(\infty)$$

$$\log(1-p_i) = \exp(\boldsymbol{\beta}\,'Z_i) \cdot -\Lambda_0(\infty)$$

Taking logarithm to an exponential:

$$\log(1-p_i) = \exp(\boldsymbol{\beta}\,'Z_i) \cdot \log(\exp(-\Lambda_0(\infty)))$$

Summing $(1-1)$ inside the logarithm we obtain the numerical expression for the marginal probability p and then isolating p_i as a function of p:

$$\log(1-p_i) = \exp(\boldsymbol{\beta}\,'Z_i) \cdot \log(1-1+\exp(-\Lambda_0(\infty))) = \exp(\boldsymbol{\beta}\,'Z_i) \cdot \log(1-p)$$

$$1-p_i = (1-p)^{\exp(\boldsymbol{\beta}\,'Z_i)} \rightarrow p_i = 1-(1-p)^{\exp(\boldsymbol{\beta}\,'Z_i)}$$

And that gives us the probability of having a cause 1 event for each subject. Hence we can calculate the probabilities for each subject and using a Binomial experiment, classify them either as status 1 (which is a subject that will have a type 1 event) or status 2 (subject that will have a type 2 event). Once we have the cause of failure which we can define as D_i, we can compute the time t_i (every subject will have a time to failure given the cause D_i, that's why index i for the time is used) to event from cause D_i from the conditional distribution of t_i given D_i. This is a proper distribution with the following distribution functions (Fine & Gray, 1999).
For cause 1:

$$F_1(t|Z) = \frac{1-[1-p(1-\exp(-t))]^{\exp(\boldsymbol{\beta}\,'Z)}}{p_i}$$

For cause 2:

$$F_2(t|Z) = \frac{(1-p)^{\exp(\boldsymbol{\beta}\,'Z)} \cdot (1-\exp(-t \cdot \exp(\boldsymbol{\beta}\,'Z)))}{1-p_i}$$

Select minimum time

To summarize, what we have now is 2 times per subject. The first one corresponding to the censoring time according to the Cox model and the second one is either a type 1 or type 2 event time according to the Fine and Gray model. So the time that happens first is the one that we would "observe" in a real data set. Thus, we select the minimum between the two calculated times.

Defining values

For the simulation we have some freedom when defining certain parameters that are:

- Covariate coefficients for censoring ($\boldsymbol{\gamma}\,'$): will change the effect of covariates to the censoring times, and consequently the amount of "observed" final censored subjects.
- Covariate coefficients for CIF ($\boldsymbol{\beta}\,'$): will change the effect of covariates to our outcome (the CIF) as well as the type 1 and type 2 failure times.
- Cumulative censoring baseline hazard (B): as covariate effects, will also change censoring times and censored sample size.
- Marginal probability of being a type 1 event (p): will change the amount of subjects that have a cause 1 event.

The effects of the covariates and the intercept for the censoring model are defined as:

$$\boldsymbol{\gamma}' = \begin{pmatrix} \alpha_c = 0.2 \\ \gamma_1 = 0.5 \\ \gamma_2 = 0.5 \end{pmatrix}$$

The effects of the covariates and the intercept for the CIF (Fine and Gray) model are defined as:

$$\boldsymbol{\beta}' = \begin{pmatrix} \alpha_f = 0.2 \\ \beta_1 = 0.75 \\ \beta_2 = 0.5 \end{pmatrix}$$

For our simulation we also found that good values for B given the coefficients and covariates defined before are:

- $B = 0.5$ for the case of moderate censoring (around 33% of data).
- $B = 1.6$ for the case of heavy censoring (around 66% of data).

Finally, since we aim to have a 50/50 ratio between cause 1 and cause 2 failures (e.g for moderate censoring we would have around 33% of censored data, around 33% of cause 1 failure data and around 33% of cause 2 failure data), the optimal value for p is $p = 0.45$. An example of this generated data corresponding to scenario number 1 (independent Z_1, Z_2) is shown below:

id	time	status	z1	z2
1	0.36	1.00	0.00	-0.35
2	0.04	0.00	1.00	-0.30
3	0.17	0.00	1.00	-0.07
4	0.30	0.00	0.00	-0.63
5	0.25	2.00	0.00	0.03
6	0.29	0.00	0.00	0.46
7	0.31	0.00	0.00	0.42
8	1.10	1.00	0.00	-0.52
9	1.43	2.00	0.00	-0.18
10	0.67	0.00	0.00	0.01
11	2.38	1.00	0.00	0.31
12	0.24	0.00	1.00	-0.41
13	1.32	0.00	0.00	0.10
14	0.53	0.00	0.00	0.92
15	0.05	1.00	0.00	0.13
16	1.70	0.00	0.00	-0.26
17	0.27	2.00	0.00	-0.80
18	0.89	1.00	0.00	0.69
19	0.98	2.00	0.00	-0.76
20	0.09	1.00	0.00	0.34

Table 3.2: Example of generated data with scenario 1.
id is the subject number (one for each subject), *time* is the time to failure, *status* indicates whether the subject is censored (status = 0), a cause 1 failure (status = 1) or a cause 2 failure (status = 2), (*z1*, *z2*) are the two covariates (generated for each subject) without dependency between Z_1 and Z_2.

3.3.2 Computation of pseudo-observations

Once we generate the data we have defined two options of estimators to compute the pseudo-observations. The first one uses the Aalen-Johansen estimator already seen in *Theoretical Background*, the other option and the one that makes most interest is the estimator weighted by the inverse censoring probability (IPCW estimators) also seen in the *Theoretical Background* section. Making both of them will allow us to see – once we fit the model and get the coefficient estimates – how the IPCW estimators perform in comparison with the Aalen-Johansen estimator as well as the different scenarios already mentioned. Then, we can assess more precisely the robustness of the method when misspecifying the model for censoring.

Choosing times

Revisiting the equations for the CIF:

$$\hat{F}_j(t) = \int_0^t \hat{S}(u-) d\hat{A}_j(u) = \int_0^t \hat{S}(u-) \frac{\sum_{i=1}^n dN_{ij}(u)}{Y(u)} \text{ (using Aalen-Johansen estimator)}$$

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^n \frac{N_{ij}(t)}{\exp(-\hat{B}(T_i) \exp(\hat{\gamma}' Z_i))} \text{ (using IPCW)}$$

$$\hat{F}_j^{(-i)}(t) = \frac{1}{n-1} \sum_{k \neq i} \frac{N_{kj}(t)}{\exp(-\hat{B}^{(-i)}(\tilde{T}_k-) \exp(\hat{\gamma}' Z_k))} \text{ (using IPCW)}$$

$$\hat{F}_{ij}(t) = n\hat{F}_j(t) - (n-1)\hat{F}_j^{(-i)}(t) \text{ (Pseudo-observations)}$$

We have to keep in mind that when computing pseudo-observations, whether using Aalen-Johansen estimator or IPCW estimator, we will have a curve (or more precisely a step function) for the cumulative incidence function defined at all t and for each subject i (given a cause j). That can be seen in the pseudo-observations formula where:

- $\hat{F}_j(t)$ will be a step function given all subjects ($i = 1, ..., n$), given a cause of failure j and for all t. An example of the CIF computed as $\hat{F}_j(t)$ for $j = 1$ is shown below:

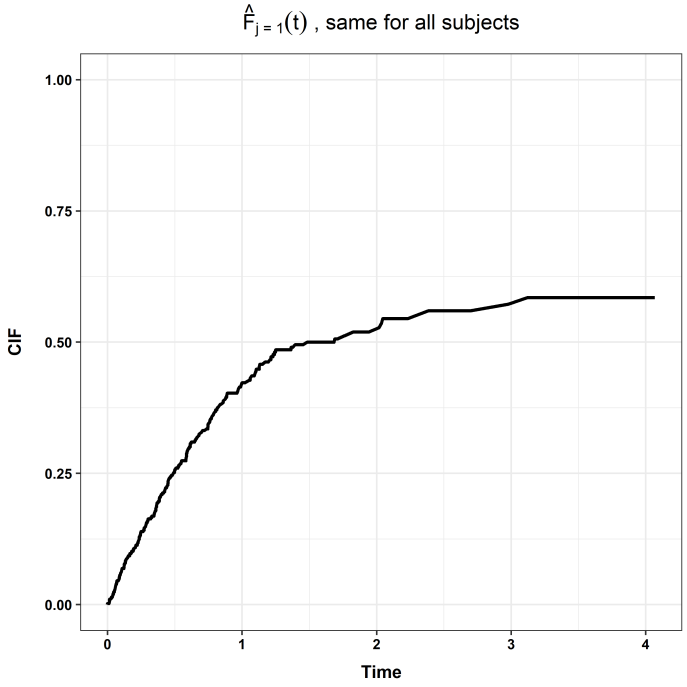


Figure 3.1: Example of Cumulative Incidence Function $\hat{F}_j(t)$ for cause $j = 1$. Computed with same previously generated data (table 3.2)

- $\hat{F}_j^{(-i)}(t)$ will also be a step function with the difference that now we have a unique curve for each subject $(i = 1, ..., n)$, for all t and for a given cause j. However, since the estimation for the CIF won't change much when we remove one subject from the whole sample, the curves will be very similar. This can be seen in the figure below where –as an example– we've plotted the leave-one-out estimator of the cumulative incidence $(\hat{F}_j^{(-i)}(t))$ for two different values of i:

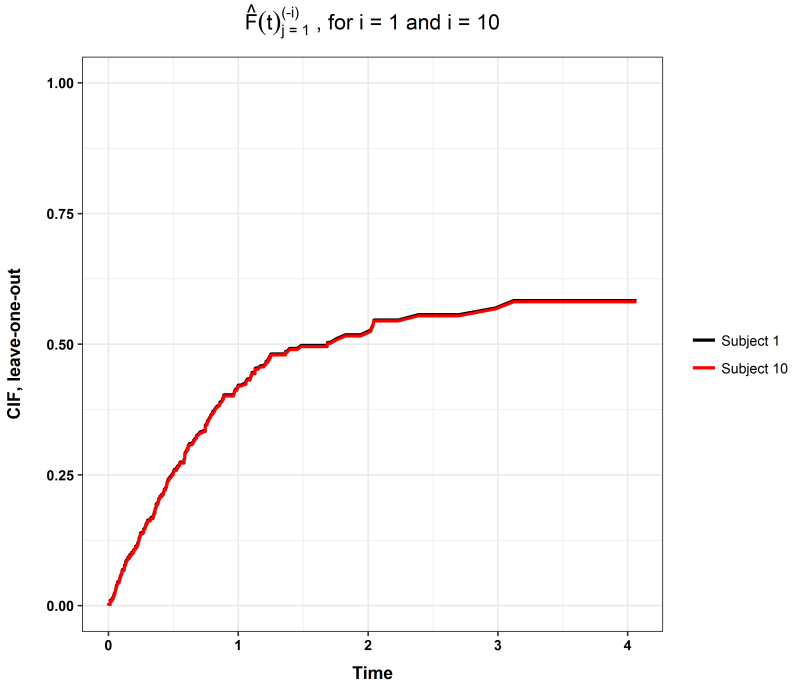


Figure 3.2: Example of Cumulative Incidence Function using leave-one-out estimator $\hat{F}_j^{(-i)}(t)$ for cause $j = 1$. Computed with same previously generated data (table 3.2)

That would lead us to a matrix n x n of pseudo-observations, where n is the sample size of our data, that would take the form:

	t_1	t_2	...	t_n
1	$F_{1,j=1}(t_1)$	$F_{1,j=1}(t_2)$	...	$F_{1,j=1}(t_n)$
2	$F_{2,j=1}(t_1)$	$F_{2,j=1}(t_2)$	...	$F_{2,j=1}(t_n)$
3	$F_{3,j=1}(t_1)$	$F_{3,j=1}(t_2)$	...	$F_{3,j=1}(t_n)$
⋮	⋮	⋮	⋮	⋮
n	$F_{n,j=1}(t_1)$	$F_{n,j=1}(t_2)$	...	$F_{n,j=1}(t_n)$

Table 3.3: Example of pseudo-observations matrix with n times chosen and $j = 1$

This has a serious problem when computing because is very memory-consuming and then the computation slows down drastically. To solve this we choose the times where we want the CIF estimated when using pseudo-values. We will loose some precision when fitting the final model and therefore we will most likely have larger standard errors, but it's necessary step in this case. Our simulation will compute the pseudo-values in 10

different (and evenly spaced or quantiles) time-points.

Computing pseudo-values

As said before, we have to make a distinction depending on if we use the Aalen-Johansen estimator or the IPCW estimator. In the first case the computation is fairly fast and direct using the *prodlim* package from Thomas A. Gerds (https://cran.r-project.org/web/packages/prodlim) which gives us the function *prodlim* to compute the cumulative incidence using the Aalen-Johansen estimator and then the function *jackknife* to compute the pseudo-values for a given cause and given time-points. Otherwise, if we want to use the IPCW estimator we have to follow the steps defined previously in this chapter, section 3.1 *Estimating Equations for CIF*, that is, estimating the censoring coefficients $\boldsymbol{\gamma}'$ with the whole sample and re-fitting the cumulative censoring baseline hazard for each leave-one-out case $\hat{B}^{(-i)}(t)$:

1. Fit the censoring model and keep the coefficients $\hat{\boldsymbol{\gamma}}'$ and the cumulative censoring baseline hazard $\hat{B}(t)$.
2. Using the previously fitted censoring model compute $\hat{F}_j(t)$ and evaluate at the chosen time-points.
3. Compute the Breslow estimator for the leave-one-out for all t:
$$\hat{B}^{(-i)}(t) = \int_0^t \frac{\sum_{k\neq i} dN^c(u)}{\sum_{k\neq i} Y_k(u)\exp(\hat{\boldsymbol{\gamma}}' Z_k)}$$
4. Compute the leave-one-out CIF $\hat{F}_j^{(-i)}(t)$ using $\hat{B}^{(-i)}(t)$ and the same $\hat{\boldsymbol{\gamma}}'$ from step 1, evaluate the function at the chosen time-points.
5. Compute the pseudo-observations for subject i and cause of interest j (in our case, as said before, $j = 1$) using $\hat{F}_j(t)$ from step 2 (already evaluated at chosen time-points):
$$\hat{F}_{ij}(t) = n\hat{F}_j(t) - (n-1)\hat{F}_j^{(-i)}(t)$$
6. Repeat steps 3 to 6 for $i = 1, ..., n$.

The final matrix contains n rows corresponding to our n subjects and 10 columns corresponding to the chosen time-points, a sample can be seen below:

t.0.0582	t.0.1116	t.0.1812	t.0.2551	t.0.3444	t.0.44016	t.0.5622	t.0.7071	t.0.9718	t.1.3173
-0.00058	-0.00251	-0.00547	-0.01143	-0.01814	1.14937	1.13626	1.12117	1.09858	1.08315
0.01342	0.05791	0.09854	0.14762	0.18499	0.25391	0.30942	0.37327	0.46891	0.53427
-0.00109	-0.00468	0.00141	0.05901	0.10286	0.18375	0.24889	0.32382	0.43605	0.51277
-0.00049	-0.00210	-0.00457	-0.00956	0.00165	0.11346	0.20351	0.30709	0.46224	0.56829
-0.00074	-0.00320	-0.00698	-0.01459	-0.02148	-0.03418	-0.04441	-0.05618	-0.07381	-0.08585
-0.00098	-0.00421	-0.00917	-0.01918	0.00384	0.10277	0.18245	0.27410	0.41137	0.50521
-0.00095	-0.00409	-0.00892	-0.01865	-0.01258	0.08895	0.17072	0.26478	0.40566	0.50196
-0.00052	-0.00226	-0.00491	-0.01027	-0.01629	-0.03192	-0.04924	-0.08566	-0.16907	1.37041
-0.00065	-0.00280	-0.00610	-0.01276	-0.02025	-0.03968	-0.06120	-0.10648	-0.21023	-0.38669
-0.00073	-0.00315	-0.00687	-0.01436	-0.02279	-0.04466	-0.06889	-0.01897	0.25922	0.44945
-0.00089	-0.00382	-0.00832	-0.01739	-0.02759	-0.05407	-0.08341	-0.14517	-0.28675	-0.52809
-0.00088	-0.00377	-0.00822	-0.00695	0.04334	0.13612	0.21084	0.29679	0.42554	0.51354
-0.00078	-0.00334	-0.00728	-0.01523	-0.02416	-0.04735	-0.07304	-0.12710	-0.25100	-0.46199
-0.00131	-0.00564	-0.01229	-0.02569	-0.04077	-0.07991	-0.08781	0.03409	0.21669	0.34152
1.02595	1.02479	1.02372	1.02244	1.02147	1.01967	1.01822	1.01655	1.01406	1.01235
-0.00062	-0.00267	-0.00581	-0.01214	-0.01926	-0.03775	-0.05822	-0.10130	-0.19999	-0.36779
-0.00044	-0.00189	-0.00411	-0.00860	-0.01307	-0.02132	-0.02796	-0.03560	-0.04705	-0.05487
-0.00113	-0.00488	-0.01064	-0.02225	-0.03530	-0.06918	-0.10674	-0.18584	1.91601	1.76952
-0.00045	-0.00193	-0.00421	-0.00880	-0.01397	-0.02736	-0.04221	-0.07341	-0.14488	-0.21547
-0.00091	1.06620	1.06334	1.05989	1.05727	1.05242	1.04852	1.04403	1.03731	1.03271

Table 3.4: Example of pseudo-observations matrix with 10 chosen time-points from our simulated data.

The matrix is then modified to include the original times to failure, status and the two covariates.

3.3.3 Fit of the GEE model

What we do in this step is re-arranging the previous matrix (particularly “melting") so we can obtain a new matrix that will take the following form:

pseudo	tpseudo	z1	z2	id
-0.00	1	0.00	-0.35	1
-0.00	2	0.00	-0.35	1
-0.01	3	0.00	-0.35	1
-0.01	4	0.00	-0.35	1
-0.02	5	0.00	-0.35	1
1.23	6	0.00	-0.35	1
1.22	7	0.00	-0.35	1
1.20	8	0.00	-0.35	1
1.17	9	0.00	-0.35	1
1.15	10	0.00	-0.35	1
0.01	1	1.00	-0.30	2
0.05	2	1.00	-0.30	2
0.08	3	1.00	-0.30	2
0.12	4	1.00	-0.30	2
0.15	5	1.00	-0.30	2
0.21	6	1.00	-0.30	2
0.26	7	1.00	-0.30	2
0.32	8	1.00	-0.30	2
0.41	9	1.00	-0.30	2
0.48	10	1.00	-0.30	2

Table 3.5: Example of "melted" pseudo-observations matrix from our simulated data. *id* is the subject number (in this case $1, ..., 500$), *pseudo* are the calculated pseudo-values, *tpseudo* indicated the chosen time-point where the *pseudo* is computed and (*z1*, *z2*) are the two covariates (generated for each subject).

What we do at this point is using the *R* package *geepack* (https://cran.r-project.org/web/packages/geepack) and the function *geese* from that same package to fit the model with the *cloglog* link function that we explained in section 3.1, *Estimating Equations for CIF*. This is a known and well-defined link function, so the *R* function knows how to handle it. Something worth mentioning is that the GEE fit will also give us the robust standard errors.

Following the same notation as the previous table (table 3.5), the formula that we will use as input is:

$$pseudo \sim factor(tpseudo) + factor(z1) + z2$$

In this case we are putting the CIF pseudo-values as outcome or response but also specifying at which time-point that pseudo-value corresponds (hence the $factor(tpseudo)$), as well as using the covariates Z_1 and (in this case) Z_2 as explanatory variables. Each subject (*id*) is treated as a different cluster.

The outcome of this function will be saved and then the coefficients and the robust

standard errors are extracted for the next step.

3.3.4 Get estimates from Monte Carlo simulation

To find the asymptotic estimates (point and interval estimates) for a given sample size the Monte Carlo simulation is used. This method is characterized for repeating the same estimation process a number B of times and that is what we have done, from the data generation till the fit of the GEE model. For reproducibility we always generate the same data using the *seed* function from *R*. In our case and following the work done by (N.Binder et al, 2014), we decided that $B = 500$ was a number large enough to get those asymptotic estimates.

The whole algorithm is parallelized using the packages *foreach* (https://cran.r-project.org/web/packages/foreach) and *doParallel* (https://cran.r-project.org/web/packages/doParallel) so we can save the GEE output each of the 500 times and then calculate the mean point estimates and robust standard errors. The mean square error (MSE) is also computed once we have the 500 point estimates.

It's also worth mentioning that this is a very time consuming process even with the code parallelized. Using 48 CPUs with 800MHz the process can take up to 30 hours for the computation of the 10 scenarios approximately. Taking 3 hours for each scenario.

Chapter 4

Monte Carlo Simulation for the Restricted Mean Lifetime (RML)

Throughout this chapter, the following notation will be used:

- $i = 1, ..., n$ as an indicator for the subject.
- Definitions from *Theoretical Background* still apply.

4.1 Estimating Equations for RML

Revisiting the alternative estimator for RML from *Theoretical Background*:

$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{N_i(t)}{\hat{C}_i(\tilde{T}_i - |Z_i)}$$

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$$

With the same defined Cox model as the previous chapter:

$$C_i(t|Z_i) = e^{-B(t)e^{\boldsymbol{\gamma}\,'Z_i}} = \exp(-B(t)\exp(\boldsymbol{\gamma}\,'Z_i))$$

We can now compute the pseudo-observations for the RML:

$$\hat{\mu}_\tau^{(i)} = n \cdot \hat{\mu}_\tau - (n-1) \cdot \hat{\mu}_\tau^{(-i)}$$

In this case, for the leave-one-out estimator, the censoring Cox model is refitted every time (different from the CIF). It has been done this way because from a coding point of view it was much easier and the computing time is fairly short.

We now aim to find a link function between the RML and the linear predictor. Let's define a randomly distributed time to failure t that follows a uniform distribution from 0 to the linear predictor $LP = \alpha + \boldsymbol{\beta}\,'Z$:

$$t \sim Unif[0, LP]$$

Then the probability density function $f(t)$ is, by definition ($Pr[0 \leq t \leq LP]$):

$$f(t) = \frac{1}{LP} \cdot I(0 \leq t \leq LP) = \begin{cases} \frac{1}{LP} & \text{If } t \in [0, LP] \\ 0 & \text{Otherwise} \end{cases}$$

Hence the cumulative distribution function $F(t)$:

$$F(t) = \begin{cases} \frac{t}{LP} & \text{If } t \in [0, LP) \\ 1 & \text{If } t \geq LP \end{cases}$$

Then the survival function, defined as $S(t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function:

$$S(t) = \begin{cases} 1 - \frac{t}{LP} & \text{If } t \in [0, LP) \\ 0 & \text{If } t \geq LP \end{cases}$$

And following the definition of RML introduced in *Theoretical Background*:

$$\mu_\tau = \int_0^\tau S(t)dt = \begin{cases} \tau - \frac{\tau^2}{2(LP)} & \text{If } \tau < LP \\ LP - \frac{1}{2} \cdot LP & \text{If } \tau = LP \\ \frac{1}{2} \cdot LP & \text{If } \tau > LP \end{cases}$$

This particular distribution of t allows us to find an explicit link function between the RML and the linear predictor, which in general can be fairly challenging. But we also realize that this is not a very realistic distribution of failure times, however, this can be interpreted as the foothold on how to operate with pseudo-values and the RML once you have a well-defined link function, inverse link function, variance function and first derivative of the inverse link function. We also have to choose a value for τ and we decided to operate with $\tau < LP$, which leads us to the following:

Link function:

$$LP(\mu_\tau) = \frac{\tau^2}{2(\tau - \mu_\tau)}$$

Inverse link function:

$$\mu_\tau(LP) = \tau - \frac{\tau^2}{2 \cdot LP}$$

First derivative inverse link:

$$\frac{d\mu_\tau}{d(LP)} = \frac{\tau^2}{2 \cdot LP^2}$$

Recovering equation (11):

$$U_n(\boldsymbol{\beta}\,') = \sum_{i=1}^{n} V_i(t)[\hat{\mu}_\tau^{(i)} - g(t, \alpha, \boldsymbol{\beta}\,', Z_i)] = 0$$

We can see that if we want a solution for this equation is enough to find that $\hat{\mu}_\tau^{(i)} = g(t, \alpha, \boldsymbol{\beta}\,', Z_i)$, the variance function is not needed explicitly. That's why we don't have a formula for the variance and we can simply put a constant number. However, we also realize that this can introduce a larger standard error to the estimates.

To solve the estimating equations using R we need to use a special package called *geeM* (https://cran.r-project.org/web/packages/geeM) and its function *geem* allows us to use the previous equations as a custom or user-defined link.

4.2 Scenarios and Misspecifications

The objective is still the same as chapter 3 but using RML instead of the CIF as pseudo-values. We have also simplified the generation of covariates to one covariate (Z) and we simulated 4 different scenarios in this case where one of them is the control. Results from the IPCW estimator and the K-M estimator will be compared to assess the robustness of the method:

Scenario	Model for censoring (Cox model)	Fitted model for cens.	Description
0	Independent censoring	Z	Control Scenario
1	$1 \cdot Z$	Z	Good model fitted
2	$1 \cdot Z^2$	Z^2	Good model fitted
3	$1 \cdot Z^2$	Z	Missing functional form

Table 4.1: 10 Scenarios for the Monte Carlo simulations for RML.

This will give us a very important insight: if something (use pseudo-values with alternative estimators) is better than nothing. We still want to see the robustness when the model for censoring is misspecified but not with the same intensity as for the CIF case. We will also use the moderate and heavy censoring as defined in Chapter 3.

4.3 Simulation Algorithm

The Monte Carlo simulation for the RML is based, in a general basis, in a single time to failure setup and right - censored data. This simulation, as before, can be split in 4 main parts:

1. **Generate the data:** this includes time to failure or being censored; a status indicator (0 for censoring, 1 for failure) and the covariate Z.
2. **Compute pseudo-observations:** we generate either the censoring-adjusted pseudo-values using the alternative estimator for the RML or the not-censoring-adjusted pseudo-values using the K-M estimator.
3. **Fit of the GEE model:** using same strategy as for the CIF but now the difference is the user-supplied link function.
4. **Get the estimates** from Monte Carlo simulation.

4.3.1 Data generation

Generating covariates

For this section with the RML we chose a continuous covariate with a uniform distribution:

$$Z \sim Unif[-1, 1]$$

Generating censoring times

The process is exactly the same as for the CIF case when generating the times to being censored.

Generating times to failure

As said in the previous section, our condition is that t follows a uniform distribution between 0 and the linear predictor $\alpha + \boldsymbol{\beta}\,'Z$, so for each subject we pick a random number from the uniform distribution of times conditional on covariate Z_i. Once we have the time to being censored and the time to failure, we select the **minimum time** for each subject i and that will be the "observed" times.

Defining values

Following again the case for the CIF:

- Covariate coefficients for censoring ($\boldsymbol{\gamma}\,'$).
- Covariate coefficients for RML ($\boldsymbol{\beta}\,'$): will change the effect of covariates to our outcome as well as the failure times.
- Cumulative censoring baseline hazard (B).

The effects of the covariates and the intercept for the censoring model are defined as:

$$\boldsymbol{\gamma}\,' = \begin{pmatrix} \alpha_c = 4 \\ \gamma_1 = 1 \end{pmatrix}$$

The effects of the covariates and the intercept for the RML model are defined as:

$$\boldsymbol{\beta}\,' = \begin{pmatrix} \alpha_r = 4 \\ \beta_1 = 1 \end{pmatrix}$$

For our simulation we also found that good values for B given the coefficients and covariates defined before are:

- $B = 0.0035$ for the case of moderate censoring (around 33% of data).
- $B = 0.011$ for the case of heavy censoring (around 66% of data).

Example of generated data:

id	time	status	z
1	0.33	1.00	-0.86
2	1.35	1.00	0.64
3	1.61	1.00	0.89
4	0.58	1.00	-0.46
5	1.74	1.00	-0.66
6	0.07	1.00	-0.93
7	3.04	1.00	-0.64
8	3.21	1.00	0.28
9	1.34	0.00	-0.95
10	2.96	1.00	-0.98
11	2.25	1.00	-0.21
12	1.93	0.00	0.63
13	1.05	1.00	-0.25
14	2.48	1.00	-0.24
15	1.22	1.00	-0.47
16	0.67	1.00	-0.12
17	0.53	1.00	-0.08
18	0.34	1.00	0.08
19	0.32	0.00	0.33
20	0.47	1.00	-0.77

Table 4.2: Example of generated data with scenario 1 (RML algorithm). *id* is the subject number (one for each subject), *time* is the time to failure, *status* indicates whether the subject is censored (status = 0) or not (status = 1), z is the generated covariate for each subject.

4.3.2 Computation of pseudo-observations

We have two options when computing pseudo-observations. The first one using the K-M estimator without taking into account the covariate dependent censoring and the second one using the IPCW estimator which should correct the bias from the covariate dependent censoring.

Now we are in a much simpler case because the RML is integrated for all times, meaning that we will have a single value for $\hat{\mu}_\tau^{(i)}$ for each subject, which will lead us to a matrix of $n \times 1$ pseudo-values instead of the $n \times n$ that we had for the CIF case. Thus, we won't have to choose times. This $n \times 1$ matrix will have the following structure:

	$\hat{\mu}_\tau$
1	$\hat{\mu}_\tau^{(1)}$
2	$\hat{\mu}_\tau^{(2)}$
3	$\hat{\mu}_\tau^{(3)}$
⋮	⋮
n	$\hat{\mu}_\tau^{(n)}$

Table 4.3: Example of pseudo-observations for RML

When using the K-M estimator the calculation of the pseudo-values is straightforward if we use the *R* package *pseudo* (https://cran.r-project.org/web/packages/pseudo) with its function *pseudomean*. When using the IPCW estimators the implementation in *R* has to be done and the algorithm is the following:

1. Fit the censoring model with the whole sample to get the censoring effects $\hat{\boldsymbol{\gamma}}'$ and the cumulative baseline censoring hazard $\hat{B}(t)$.
2. Using the previous censoring fit compute $\hat{\mu}_\tau$ for the whole sample. Important to notice that working in a discrete case, the integral of the estimated survival function $\hat{S}(t)$ has to be computed as a sum.
3. Fit the censoring model with the leave-one-out data and using the same structure as step 2 obtain the leave-one-out RML $\hat{\mu}_\tau^{(-i)}$.
4. Compute the pseudo-observation for subject i using $\hat{\mu}_\tau^{(i)} = n \cdot \hat{\mu}_\tau - (n-1) \cdot \hat{\mu}_\tau^{(-i)}$.
5. Repeat steps 3 and 4 for all subjects $i = 1, ..., n$.

The final matrix will contain n rows and 1 column, as already stated. An example of this below:

id	time	status	z	pv
1	0.00	0.00	-0.61	1.80
2	0.00	1.00	-0.45	-0.01
3	0.01	1.00	-0.73	0.03
4	0.02	1.00	0.52	0.01
5	0.02	0.00	-0.57	1.82
6	0.03	1.00	-0.63	0.03
7	0.04	1.00	-0.88	0.08
8	0.04	0.00	0.72	1.81
9	0.05	1.00	-0.36	0.02
10	0.05	0.00	0.78	1.80
11	0.05	1.00	-0.61	0.04
12	0.06	0.00	0.72	1.81
13	0.06	1.00	0.30	0.02
14	0.07	0.00	-0.64	1.83
15	0.07	1.00	-0.93	0.11
16	0.07	0.00	0.90	1.80
17	0.08	1.00	0.17	0.02
18	0.09	1.00	0.55	0.06
19	0.09	1.00	0.54	0.06
20	0.09	1.00	0.54	0.06

Table 4.4:
Example of RML pseudo-values using alternative estimator with the previously generated data (scenario 1). *id* is the subject number (one for each subject), *time* is the time to failure, *status* indicates whether the subject is censored (status = 0) or not (status = 1), z is the generated covariate for each subject, pv is the computed pseudo-value for RML

4.3.3 Fit of the GEE model

The model that will be fitted using the previous table 4.4 is:

$$pv \sim z$$

Keeping in mind that the proper link, inverse link and first derivative of the inverse link functions have to be inputted to the *geem* function.

4.3.4 Get estimates from Monte Carlo simulation

Similarly to what we did with the CIF, this will also be a Monte Carlo simulation with $B = 500$. For each iteration we will get the point estimates and the robust standard errors, once we have all of them we do the average. At the end we will also have all the point estimates for $\boldsymbol{\beta}'$ and we will also compute the MSE. Compared to the previous chapter computational time in this case is not an issue.

Chapter 5

Results

Numerical and graphical results are shown in the next pages, both the Monte Carlo simulation for the Cumulative Incidence Function regression and the Monte Carlo Simulation for the Restricted Mean Lifetime regression. All numerical results contain the point estimates for the coefficients, their robust standard errors (Huber-White standard error estimates) and the mean square error defined in general, given a vector of point estimates for $\hat{\boldsymbol{\beta}}'$:

$$MSE(\hat{\boldsymbol{\beta}}') = \text{Var}(\hat{\boldsymbol{\beta}}') + \text{Bias}(\hat{\boldsymbol{\beta}}', \boldsymbol{\beta}')^2$$

5.1 Results of the Monte Carlo simulation for the CIF regression

Three tables of numerical results are presented:

- Table 5.1: shows the results with a sample size of $n = 500$ and 500 replications with moderate censoring (approximately 33% of data is censored). It also shows the censoring estimates from the fitted Cox model for the 10 different scenarios presented in Chapter 3.
- Table 5.2: same structure, sample size and replications as table 5.1 but in this case we have heavy censoring (approximately 66% of data is censored).
- Table 5.3: shows the results with a sample size of $n = 200$ and 500 replications only for moderate censoring.

What we can see in the three tables at first glance is that the bias for the point estimates, the robust standard errors and the MSE (both using the IPCW estimator for the CIF and the Aalen-Johansen estiamtor (AJ)) are accentuated. This is mainly due to the change of the proportion of censored/cause 1/cause 2 number of subjects: if we increase the number of subjects that are censored we will decrease (with a fixed total sample size of $n = 500$) the number of subjects that have either a cause 1 failure time or a cause 2 failure time. Thus, a reduced sample size for the cause of interest (cause 1) leads to an increased uncertainty in the estimates when doing the regression.

What we also see immediately is that the algorithm is working well and the pseudo-values regression method is consistent because in the control scenario (scenario 0) we get almost the same point estimates and robust standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ in both

moderate censoring and heavy censoring. This also shows that even we don't have a covariate dependent censoring scenario, using pseudo-values won't affect the accuracy of the estimation.

Next, the following remarks are made when looking at the results:

Scenarios with independent covariates (1, 2, 5, 6, 7):

- The point estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$ using IPCW estimators are better than the same point estimates using the AJ estimator. In fact, the pseudo-values regression when using the alternative estimator are always accurate when we have moderate censoring even if the model for censoring is misspecified. When we have heavy censoring the misspecification of the censoring model can introduce bias to the point estimates, but even if we misspecify the model for censoring we get a much smaller bias than doing the pseudo-values regression using the AJ estimator, this leads to the next remark.
- The estimates are slightly better when the model for censoring is correctly specified. But we also have to notice that if our censoring is moderate, the misspecification of the censoring Cox model won't affect much the results.
- When we have moderate censoring, the AJ standard errors are always slightly smaller or equal to the standard errors from the IPCW estimator. This effect is accentuated when we have heavy censoring.
- In general, using pseudo-values is much better than doing nothing or not adjusting for censoring when we have independent covariates or low correlation between covariates. Thus, the use of pseudo-values is necessary if you don?t want to underestimate.

Scenarios with dependent covariates (3, 4, 8, 9, 10):

- The two covariates Z_1 and Z_2^* are positively correlated, so we should get negatively correlated estimates for the parameters $\hat{\beta}_1$ and $\hat{\beta}_2$. That is what we can see in all scenarios where we have a covariate dependency, so the simulations for the different scenarios are consistent.
- When the model for censoring has a linear dependency (scenarios 3 and 4), estimates with moderate censoring are fairly correct and much better than using the Aalen-Johansen estimator (for the same scenarios) which is underestimating again the coefficients. Worth notice that is slightly better to misspecify the model for censoring if we look at the MSE.
- When the model for censoring has a non-linear dependency (scenarios 8 to 10), estimates, whether using IPCW or Aalen-Johansen, are always biased unless the model for censoring is correctly specified. So, in case of dependency between covariates and non-linear dependency with the censoring model, we have to keep in mind that Aalen-Johansen estimator could be a better choice.

Regardless of the fitted model, when covariates are dependent we get worse results than if we have independent covariates, also with much larger standard errors (approximately twice than the independent covariates scenarios). That also depends on the functional dependency for the censoring model, better case scenario where dependency

with censoring is linear and we have moderate censoring, worst case scenario being dependency between covariates and censoring dependency with non-linear form, having heavy censoring as well as not fitting the right model for censoring.

For the sake of completeness, results with $n = 200$ show the same tendency as results with $n = 500$, however in this case since we have a smaller (and more realistic) sample size for the cause of interest, our standard errors will be larger.

5.2 Results of the Monte Carlo simulation for the RML regression

Results for this simulation are shown in Tables 5.4 and 5.5 where, with a sample size of $n = 500$ and 500 replications, estimates, robust standard errors and MSE are presented. Table 5.4 is for moderate censoring (again approximately 33% of data) and Table 5.5 is for heavy censoring (approximately 66% of data). Graphical results are also shown in figures 5.3, 5.4, 5.5 and 5.6. These show the distribution of the estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}'$ from the 500 replications. That way it is easier to compare the bias and the variance when the pseudo-observations regression is made with the K-M estimator or the IPCW (alternative) estimator.

Like for the CIF case, the control scenario shows us how the simulation and estimation algorithm is consistent because we obtain fairly the same point and interval estimates using both the K-M and the IPCW estimator for the pseudo-observations.

Since we only have one covariate (Z) in this case, comparison between scenarios is much easier. With moderate censoring we can actually see how the point estimates using the alternative estimator IPCW are not biased and the point estimates using the K-M estimator for the pseudo-observations computation present bias in all scenarios except in 0 (control). However, in scenario number 1 (linear censoring simulated) the standard errors for the IPCW estimator are a little bit larger, although the difference is so small that cannot be seen in figures 5.1 and 5.2.

With heavy censoring, we still see good accuracy when using the IPCW estimator (except scenario 3 where the estimate for $\hat{\boldsymbol{\beta}}'$ is a bit lower than the true value). When using the K-M estimator however, the bias in the point estimates is much larger than before due to the reduced sample size and the use of this estimator when the censoring depends on covariates. This can also be seen in figures 5.3 and 5.4 where some peaks of the empirical distribution for $\hat{\alpha}_r$ and $\hat{\boldsymbol{\beta}}'$ are displaced respect to the horizontal line marking the true values of the coefficients. However, we still can see larger standard errors and MSE specially in scenarios 1 and 2, this can also be seen in figures 5.3 and 5.4 with wider distributions for the estimates when using the IPCW estimator.

As a general point we can say that using pseudo-observations with the alternative estimator for the RML is always better in terms of bias than using the K-M estimator.

CRR model with 500 replications; n = 500 subjects; 10 event-scale equally spaced timepoints. MODERATE CENSORING													Censoring estimates	
SCENARIO	$\hat{\beta}_1$	$\hat{\beta}_2$	**SE-**$\hat{\beta}_1$	**SE-**$\hat{\beta}_2$	**MSE-**$\hat{\beta}_1$	**MSE-**$\hat{\beta}_2$	**AJ-**$\hat{\beta}_1$	**AJ-**$\hat{\beta}_2$	**AJ-SE-**$\hat{\beta}_1$	**AJ-SE-**$\hat{\beta}_2$	**AJ-MSE-**$\hat{\beta}_1$	**AJ-MSE-**$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
0	0.7529	0.5159	0.1640	0.1396	0.0289	0.0185	0.7515	0.5154	0.1638	0.1395	0.0292	0.0183	0.0142	0.0050
1	0.7510	0.5137	0.1764	0.1475	0.0328	0.0216	0.6922	0.4877	0.1669	0.1434	0.0336	0.0196	0.5154	0.5024
2	0.7370	0.4962	0.1729	0.1458	0.0322	0.0201	0.6922	0.4877	0.1669	0.1434	0.0336	0.0196	0.5011	-
3	0.7337	0.5192	0.3395	0.3066	0.1220	0.0934	0.6365	0.4637	0.3262	0.2878	0.1266	0.0841	0.5085	0.5073
4	0.7545	0.4922	0.3474	0.3045	0.1271	0.0915	0.6365	0.4637	0.3262	0.2878	0.1266	0.0841	1.0085	-
5	0.7544	0.5133	0.1791	0.1537	0.0329	0.0231	0.7020	0.4917	0.1712	0.1448	0.0328	0.0205	0.5141	0.4983
6	0.7540	0.4954	0.1786	0.1478	0.0326	0.0213	0.7020	0.4917	0.1712	0.1448	0.0328	0.0205	0.5061	-0.0179
7	0.7547	0.4948	0.1786	0.1476	0.0328	0.0213	0.7020	0.4917	0.1712	0.1448	0.0328	0.0205	0.5050	-
8	0.7328	0.5225	0.3471	0.3170	0.1302	0.0996	0.7323	0.4322	0.3457	0.3012	0.1304	0.0977	0.5094	0.5033
9	0.8417	0.3997	0.3613	0.3114	0.1520	0.1087	0.7323	0.4322	0.3457	0.3012	0.1304	0.0977	0.7630	-0.2777
10	0.8263	0.4140	0.3571	0.3102	0.1465	0.1063	0.7323	0.4322	0.3457	0.3012	0.1304	0.0977	0.4933	-

Table 5.1:
Results for the CIF Monte Carlo simulation for n = 500 and 500 replications, moderate censoring. From left to right we have: scenario, estimates for Z_1 and Z_2 (which can be Z_2^*) using IPCW, standard errors when using IPCW and MSE when using IPCW estimator. Then the estimates, standard errors and MSE when using the Aalen-Johansen estimator (denoted as AJ) for the pseudo-values. Last two columns are the point estimates for the fitted Cox Model for censoring.

CRR model with 500 replications; n = 500 subjects; 10 event-scale equally spaced timepoints. HEAVY CENSORING													Censoring estimates	
SCENARIO	$\hat{\beta}_1$	$\hat{\beta}_2$	**SE-**$\hat{\beta}_1$	**SE-**$\hat{\beta}_2$	**MSE-**$\hat{\beta}_1$	**MSE-**$\hat{\beta}_2$	**AJ-**$\hat{\beta}_1$	**AJ-**$\hat{\beta}_2$	**AJ-SE-**$\hat{\beta}_1$	**AJ-SE-**$\hat{\beta}_2$	**AJ-MSE-**$\hat{\beta}_1$	**AJ-MSE-**$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
0	0.7528	0.5146	0.2147	0.1866	0.0470	0.0359	0.7507	0.5139	0.2140	0.1861	0.0473	0.0354	0.0104	0.0057
1	0.7537	0.5235	0.2685	0.2253	0.0747	0.0550	0.6308	0.4546	0.2183	0.1927	0.0631	0.0383	0.5115	0.5096
2	0.7132	0.4662	0.2386	0.2065	0.0613	0.0421	0.6308	0.4546	0.2183	0.1927	0.0631	0.0383	0.4885	-
3	0.7082	0.5518	0.5230	0.5081	0.3089	0.2766	0.5315	0.4346	0.4301	0.3775	0.2757	0.1608	0.5235	0.4880
4	0.7859	0.4636	0.5743	0.4916	0.3648	0.2604	0.5315	0.4346	0.4301	0.3775	0.2757	0.1608	1.0002	-
5	0.7572	0.5193	0.2639	0.2360	0.0724	0.0598	0.6554	0.4690	0.2313	0.1959	0.0638	0.0412	0.5130	0.5017
6	0.7563	0.4818	0.2584	0.2119	0.0692	0.0485	0.6554	0.4690	0.2313	0.1959	0.0638	0.0412	0.5041	-0.0036
7	0.7562	0.4807	0.2579	0.2114	0.0693	0.0484	0.6554	0.4690	0.2313	0.1959	0.0638	0.0412	0.5032	-
8	0.7102	0.5655	0.5133	0.5030	0.3073	0.2733	0.7239	0.3785	0.4843	0.4160	0.2827	0.2058	0.5080	0.5073
9	0.9645	0.2798	0.5683	0.4754	0.4289	0.2963	0.7239	0.3785	0.4843	0.4160	0.2827	0.2058	0.7609	-0.2777
10	0.9156	0.3249	0.5422	0.4668	0.3848	0.2772	0.7239	0.3785	0.4843	0.4160	0.2827	0.2058	0.4956	-

Table 5.2:
Results for the CIF Monte Carlo simulation for n = 500 and 500 replications, heavy censoring. From left to right we have: scenario, estimates for Z_1 and Z_2 (which can be Z_2^*) using IPCW, standard errors when using IPCW and MSE when using IPCW estimator. Then the estimates, standard errors and MSE when using the Aalen-Johansen estimator (denoted as AJ) for the pseudo-values. Last two columns are the point estimates for the fitted Cox Model for censoring.

CRR model with 500 replications; n = 200 subjects; 10 event-scale equally spaced timepoints. MODERATE CENSORING													Censoring estimates	
SCENARIO	$\hat{\beta}_1$	$\hat{\beta}_2$	**SE-**$\hat{\beta}_1$	**SE-**$\hat{\beta}_2$	**MSE-**$\hat{\beta}_1$	**MSE-**$\hat{\beta}_2$	**AJ-**$\hat{\beta}_1$	**AJ-**$\hat{\beta}_2$	**AJ-SE-**$\hat{\beta}_1$	**AJ-SE-**$\hat{\beta}_2$	**AJ-MSE-**$\hat{\beta}_1$	**AJ-MSE-**$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
0	0.7575	0.5228	0.2629	0.2243	0.0717	0.0520	0.7569	0.5228	0.2627	0.2239	0.0728	0.0520	0.0142	0.0050
1	0.7536	0.5209	0.2842	0.2375	0.0836	0.0611	0.6952	0.4950	0.2677	0.2299	0.0795	0.0542	0.5008	0.4964
2	0.7406	0.5016	0.2780	0.2341	0.0817	0.0568	0.6952	0.4950	0.2677	0.2299	0.0795	0.0542	0.4880	-
3	0.7408	0.5262	0.5479	0.4950	0.2885	0.2328	0.6411	0.4687	0.5222	0.4619	0.2829	0.2095	0.5246	0.4831
4	0.7581	0.5029	0.5580	0.4903	0.2967	0.2303	0.6411	0.4687	0.5222	0.4619	0.2829	0.2095	0.9993	-
5	0.7616	0.5211	0.2892	0.2477	0.0895	0.0657	0.7087	0.4991	0.2753	0.2327	0.0823	0.0568	0.5048	0.5101
6	0.7602	0.5015	0.2881	0.2383	0.0876	0.0598	0.7087	0.4991	0.2753	0.2327	0.0823	0.0568	0.4993	-0.0355
7	0.7621	0.5015	0.2880	0.2377	0.0884	0.0600	0.7087	0.4991	0.2753	0.2327	0.0823	0.0568	0.4966	-
8	0.7369	0.5288	0.5578	0.5113	0.2858	0.2418	0.7320	0.4383	0.5540	0.4837	0.2945	0.2287	0.5122	0.5142
9	0.8443	0.4063	0.5831	0.5026	0.3349	0.2462	0.7320	0.4383	0.5540	0.4837	0.2945	0.2287	0.7815	-0.2966
10	0.8295	0.4206	0.5741	0.5741	0.3202	0.3202	0.7320	0.4383	0.5540	0.4837	0.2945	0.2287	0.4940	-

Table 5.3:
Results for the CIF Monte Carlo simulation for n = 200 and 500 replications, moderate censoring. From left to right we have: scenario, estimates for Z_1 and Z_2 (which can be Z_2^*) using IPCW, standard errors when using IPCW and MSE when using IPCW estimator. Then the estimates, standard errors and MSE when using the Aalen-Johansen estimator (denoted as AJ) for the pseudo-values. Last two columns are the point estimates for the fitted Cox Model for censoring.

RML regression model with 500 replications; n = 500 subjects; Moderate censoring											
Sim. indep. cens.			**Linear cens. simulated**			**Sim. squared cens (Good fit)**			**Sim. squared cens (Linear fit)**			
$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	
4.0283	0.1832	0.0338	3.9434	0.1727	0.0333	4.0213	0.1885	0.0345	4.0213	0.1885	0.0345	K-M Estimator
1.0113	0.2858	0.0842	0.9852	0.2677	0.0872	0.9118	0.2910	0.0932	0.9118	0.2910	0.0932	
4.0283	0.1828	0.0338	4.0297	0.1914	0.0364	4.0293	0.1912	0.0362	4.0408	0.1925	0.0377	IPCW Estimator
1.0107	0.2846	0.0845	1.0149	0.2979	0.0936	1.0076	0.3098	0.0992	0.9993	0.2911	0.0972	

Table 5.4:
Results for the RML Monte Carlo simulation for n = 500 and 500 replications, moderate censoring. From left to right we have: scenario 0 as control, scenario 1 with simulated linear censoring and a good fit, scenario 2 with simulated squared censoring and a good fit and scenario 3 with simulated squared censoring and misspecified censoring model. All scenarios show the point estimates ($\hat{\alpha}_r$ & $\hat{\beta}$), robust standard error (SE) and mean squared error (MSE). First two rows are results using the K-M estimator and last two rows, results using the alternative estimator.

RML regression model with 500 replications; n = 500 subjects; Heavy censoring											
Sim. indep. cens.			**Linear cens. simulated**			**Sim. squared cens (Good fit)**			**Sim. squared cens (Linear fit)**			
$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	$\hat{\alpha}_r$ & $\hat{\beta}$	**SE**	**MSE**	
4.0375	0.2271	0.0520	3.7859	0.1821	0.0840	4.0255	0.2534	0.0626	4.0255	0.2534	0.0626	K-M Estimator
1.0049	0.3553	0.1301	0.8540	0.2899	0.1464	0.7486	0.3665	0.1951	0.7486	0.3665	0.1951	
4.0378	0.2260	0.0528	4.0403	0.2871	0.0879	4.0685	0.2922	0.0984	4.0656	0.2644	0.0743	IPCW Estimator
1.0029	0.3511	0.1304	1.0175	0.4589	0.2350	1.0187	0.5558	0.3926	0.9581	0.3712	0.2079	

Table 5.5:
Results for the RML Monte Carlo simulation for n = 500 and 500 replications, heavy censoring. From left to right we have: scenario 0 as control, scenario 1 with simulated linear censoring and a good fit, scenario 2 with simulated squared censoring and a good fit and scenario 3 with simulated squared censoring and misspecified censoring model. All scenarios show the point estimates ($\hat{\alpha}_r$ & $\hat{\beta}$), robust standard error (SE) and mean squared error (MSE). First two rows are results using the K-M estimator and last two rows, results using the alternative estimator.

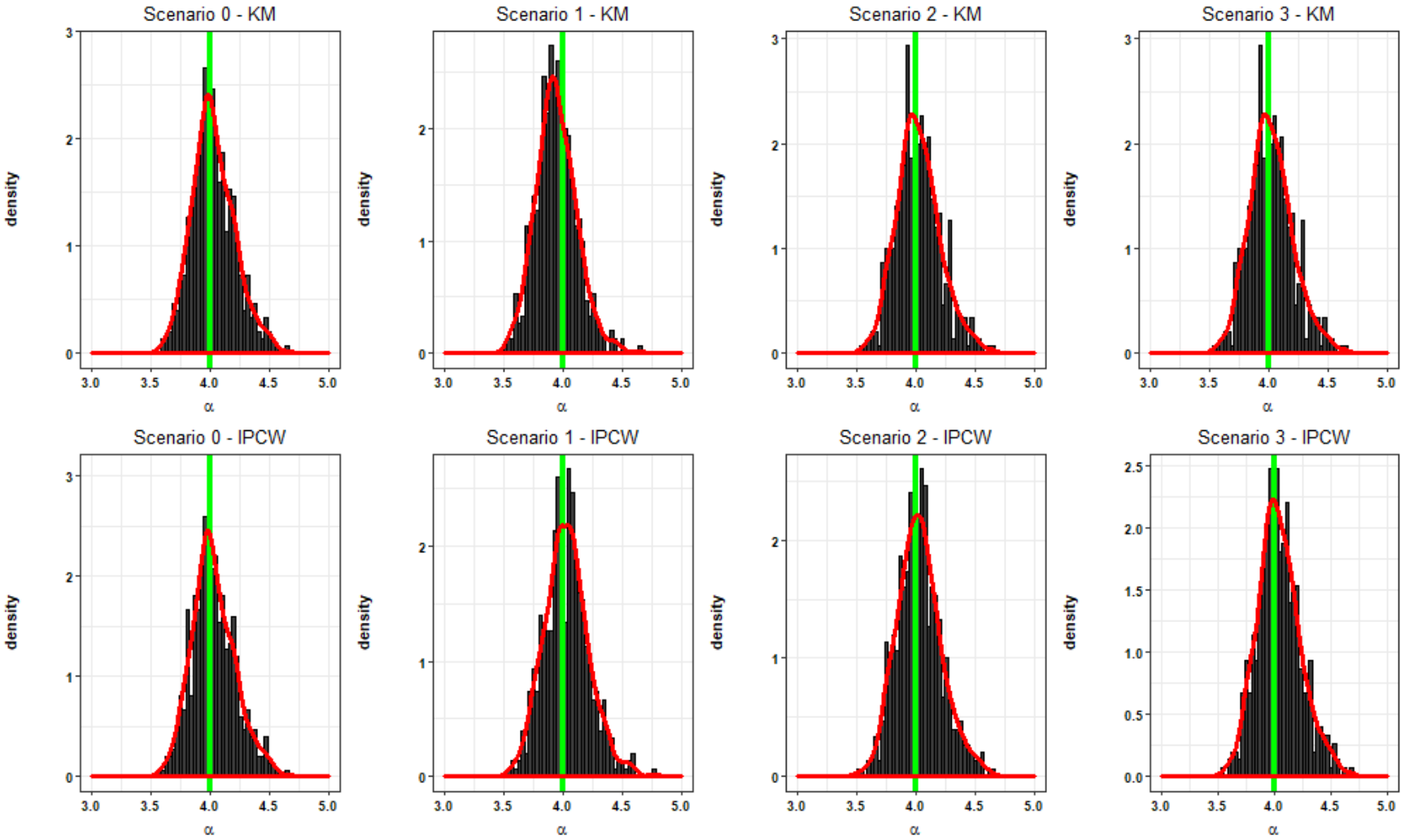


Figure 5.1: Results of Monte Carlo Simulation for RML regression: distribution of estimated $\hat{\alpha}_r$ for K-M and IPCW estimators. Moderate censoring. The red line shows the estimated density and the green line the true value for α_r.

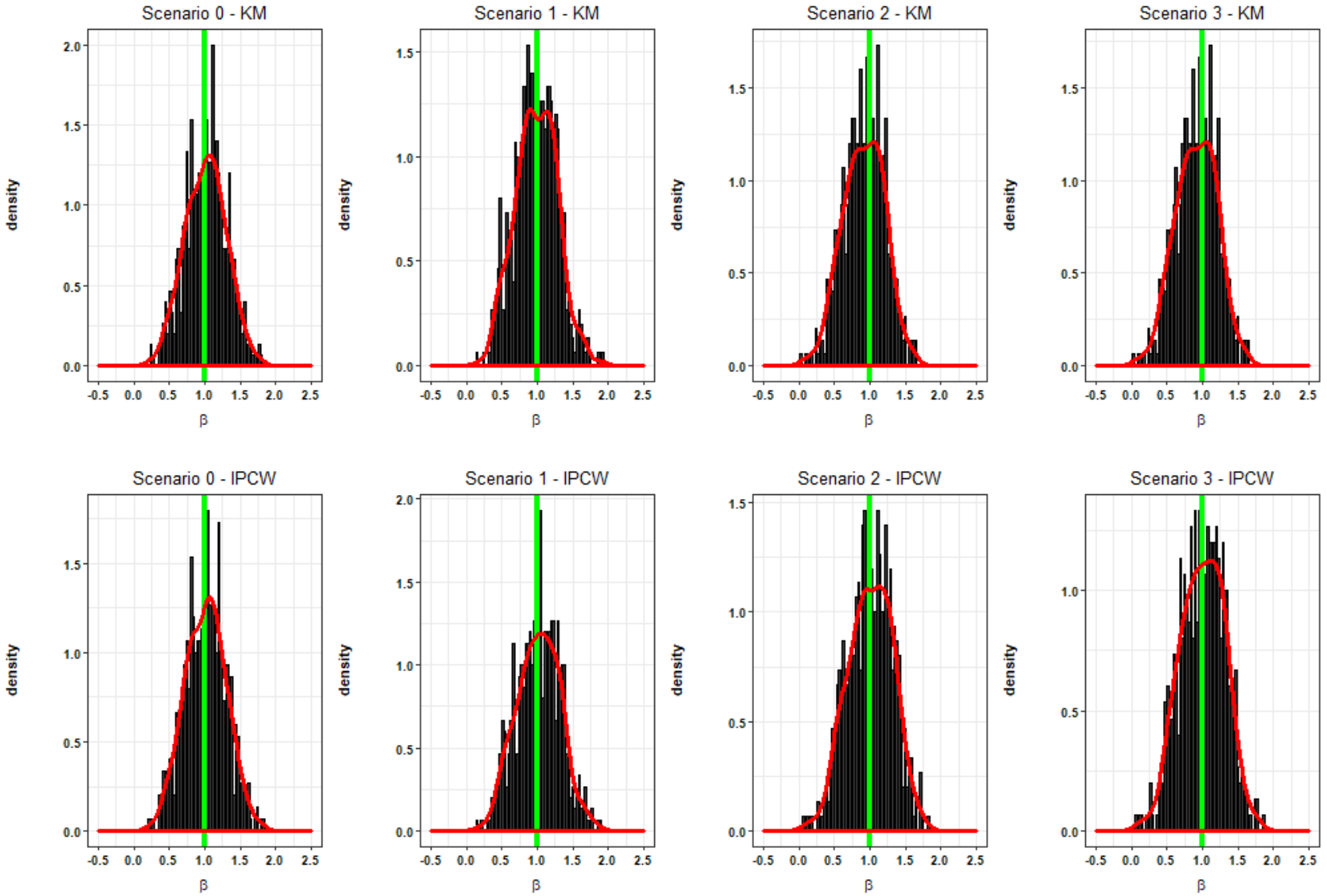


Figure 5.2: Results of Monte Carlo Simulation for RML regression: distribution of estimated $\hat{\beta}$ for K-M and IPCW estimators. Moderate censoring. The red line shows the estimated density and the green line the true value for β.

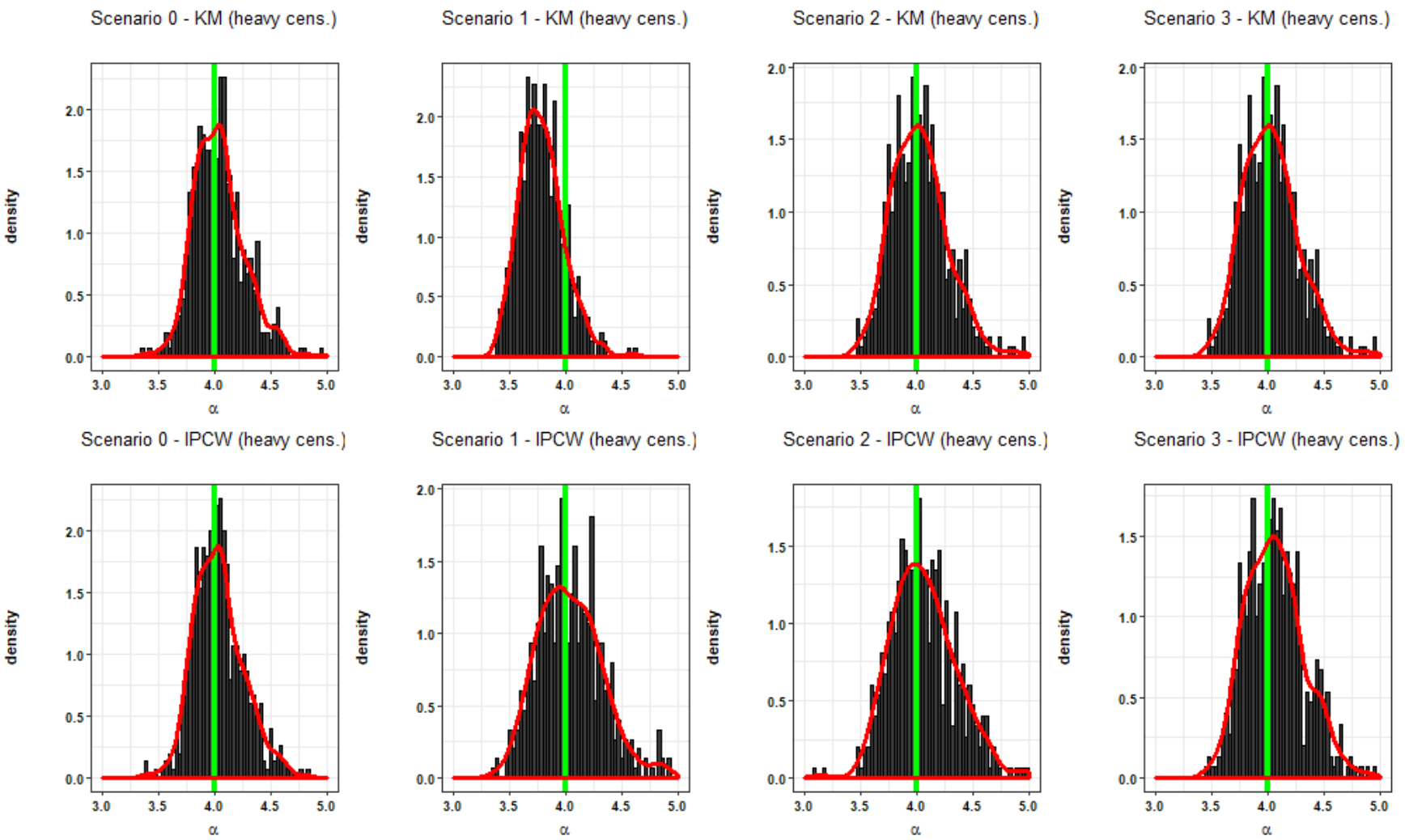


Figure 5.3: Results of Monte Carlo Simulation for RML regression: distribution of estimated $\hat{\alpha}_r$ for K-M and IPCW estimators. Heavy censoring. The red line shows the estimated density and the green line the true value for α_r.

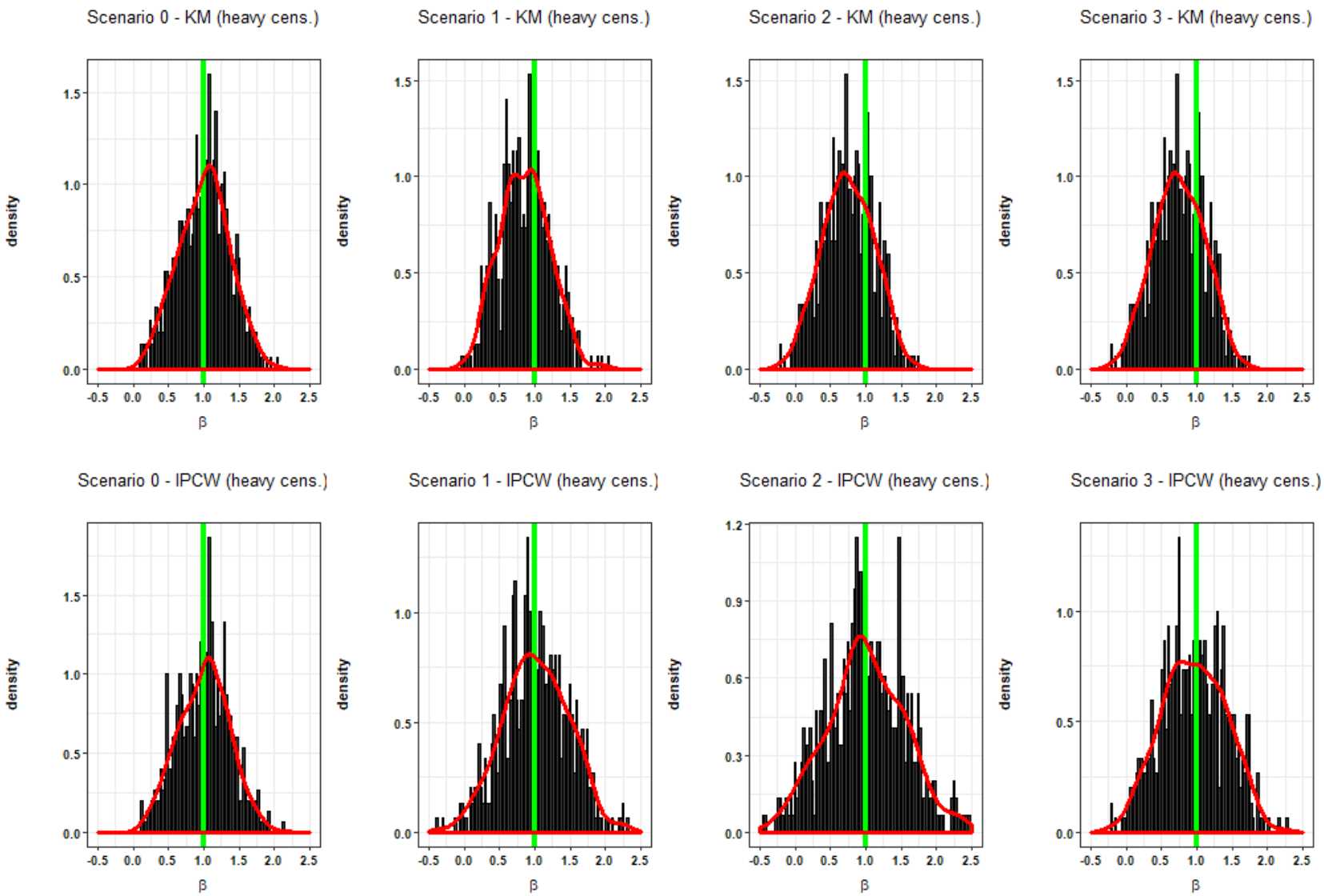


Figure 5.4: Results of Monte Carlo Simulation for RML regression: distribution of estimated $\hat{\beta}$ for K-M and IPCW estimators. Heavy censoring. The red line shows the estimated density and the green line the true value for β.

Chapter 6

Discussion

In this master thesis, we have presented two alternative estimators (for the Cumulative Incidence Function and the Restricted Mean Lifetime) that - by modeling the censoring distribution - allow us to adjust for censoring depending on covariates. Then, the pseudo-values are computed in both cases and a regression model for the pseudo-observations has been presented. Using Monte Carlo Simulations we have studied the robustness of this method when the model for the censoring distribution is misspecified.

Given the results shown before it is accurate to say that, in a general basis, the use of pseudo-values using alternative estimators is robust when we misspecify the censoring model. Another good result is that if censoring is independent from covariates we will get the same results as using a standard estimator like Aalen-Johansen or Kaplan-Meier when computing pseudo-observations. However, some distinctions have to made regarding the robustness: when we have censoring depending in a non-linear way with the covariates and these covariates have some type of dependency or correlation between them, we should be careful when specifying the model for censoring because if we misspecify that model, it can lead to a larger bias in the estimates than using for instance Aalen-Johansen or Kaplen-Meier.

It is also fair to say that when using pseudo-observations with alternative estimators we should expect an increase of the robust standard errors respect to the usual estimators when we have high correlation between covariates and in particular when our sample has a high load of censored times and low sample size. In this study we checked if the estimated coefficients when doing the Monte Carlo simulation had low probability of not being lost to follow-up before a given time t and conditional on covariates ($C_i(t|Z)$), which would lead to very high values of the estimators and eventually outliers present in our vector of estimates from the simulation, however, that was not the case. Further checks need to be done regarding this issue to pinpoint exactly the cause of this increase in the standard errors but the previous results show that it can be directly related with the sample size and the level of censoring we are working with, both for the RML and the CIF.

For the CIF case, we realize that the results we get are particular to these 10 scenarios, so a generalization of the findings should not be made freely. However, we also think that, in model fitting, missing a covariate or missing the functional form for the model are common problems. Thus, we think that studying the robustness with those misspecifications gave us a lot of insight on what can happen to a particular problem or data. Using the same reasoning, that's why we also wanted to include dependent and

independent covariates, since correlation between covariates is also a very usual problem.

Following a similar thought for the RML case, the distribution of failure times as uniform is not a very realistic case, but it was a fast solution that we found in order to find an explicit link and inverse link functions between the Restricted Mean Lifetime and the linear predictor. That allowed us to see the robustness of pseudo-observations method using alternative estimators for the RML and also using a particular (and very useful) function of R. Thus, we think we are in a more restrictive scenario(s) than for the CIF case, but this also opens a path to be followed or an algorithm that can be adapted to other explicit and more realistic link functions, although from what we have seen, finding a link, inverse link and first derivative of the inverse link with different (and more realistic) distributions for the time to failure is a fairly big challenge. Another comment regarding this method for the RML is that improvement can be done by specifying explicitly the variance function which, if correctly specified, will improve the results.

Bibliography

[1] Andersen P.K., Pohar Perme M. (2010) *Pseudo-observations in survival analysis.* Statistical Methods in Medical Research, 2010; 19(1):71-99. [https://doi.org/10.1177/0962280209105020]

[2] Andersen P.K., Klein J.P. and Rosthøj S. (2003) *Generalised linear models for correlated pseudo-observations, with applications to multi-state models.* Biometrika. 2003; 90(1):15-27. [https://doi.org/10.1093/biomet/90.1.15]

[3] Binder N., Gerds T.A., Andersen P.K. (2014). *Pseudo-observations for competing risks with covariate dependent censoring.* Lifetime Data Anal. 2014; 20:303-315. [https://doi.org/10.1007/s10985-013-9247-7]

[4] Kaplan E.L., Meier P. *Nonparametric estimation from incomplete observations.* Journal of the American Statistical Association. 1958; 53:457-481. [http://www.jstor.org/stable/2281868?origin=JSTOR-pdf]

[5] Aalen O. (1978) *Nonparametric estimation of partial transition probabilities in multiple decrement models.* Ann Stat. 1978; 6(3):534-545. [https://doi.org/10.1214/aos/1176344198]

[6] Gill, R.D. (1980) PhD thesis. Math. Centre Tracts 124. Mathematical Centre; Amsterdam: 1980. *Censoring and stochastic integrals.*

[7] Thomas A. Gerds (2017). prodlim: *Product-Limit Estimation for Censored Event History Analysis.* R package version 1.6.1. https://CRAN.R-project.org/package=prodlim

[8] Efron B., Stein C. (1981) *The Jackknife Estimate of Variance.* Ann Stat. 1981; 9(3):586-596. [https://doi.org/10.1214/aos/1176345462]

[9] Klein J.P., Andersen P.K. (2005). *Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function.* Biometrics. 2005; 61(1):223-229. [https://doi.org/10.1111/j.0006-341X.2005.031209.x]

[10] Graw F., Gerds T.A. and Schumacher M. (2009) *On pseudo-values for regression analysis in competing risks models.* Lifetime Data Anal. 2009; 15(2):241-255. [https://doi.org/10.1007/s10985-008-9107-z]

[11] Fine J.P., Gray R.J. (1999) *A proportional hazards model for the subdistribution of a competing risk.* Journal of the American Statistical Association. 1999; 94(446):496-509. [https://doi.org/10.2307/2670170]

[12] Høsgaard S., Halekoh U. & Yan J. (2006) *The R Package geepack for Generalized Estimating Equations.* Journal of Statistical Software. 2006. 15(2):1-11. [https://doi.org/10.1002/sim.1650]

[13] Yan J. & Fine J.P. (2004) *Estimating Equations for Association Structures.* Statistics in Medicine. 2004. 23:859-880. [https://doi.org/10.1002/sim.1650]

[14] Yan J (2002) geepack: *Yet Another Package for Generalized Estimating Equations.* R-News, 2/3, pp12-14.

[15] Microsoft and Steve Weston (2017) foreach: *Provides Foreach Looping Construct for R.* R package version 1.4.4. https://CRAN.R-project.org/package=foreach

[16] Microsoft Corporation and Steve Weston (2017) doParallel: *Foreach Parallel Adaptor for the 'parallel' Package.* R package version 1.0.11. https://CRAN.R-project.org/package=doParallel

[17] McDaniel L.S., Henderson N.C. & Rathouz P.J. (2013) *Fast pure R implementation of GEE: application of the Matrix package.* The R Journal, 5/1:181-187.

[18] Pohar Perme M. & Gerster M. (2017) pseudo: *Computes Pseudo-Observations for Modeling.* R package version 1.4.3. https://CRAN.R-project.org/package=pseudo