UNIVERSITY OF CALGARY

Minimum Hellinger Distance Estimation of Cox Proportional Hazard Model with Right Censored Data

by

Pengkun Liang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN STATISTICS

CALGARY, ALBERTA

AUGUST, 2021

Abstract

Cox Proportional Hazard (Cox PH) models are simple but frequently used in survival analysis. The unknown coefficient parameters in a Cox PH model are usually estimated by the partial maximum likelihood estimation (PMLE) introduced by Cox (1975). Nevertheless, maximum likelihood estimation (MLE) is generally non-robust against model misspecification and outlying observations. When data is contaminated, PMLE produces inaccurate estimates with large bias. In this thesis, we proposed instead a robust distance-based method, specifically, minimum Hellinger distance estimation (MHDE) for Cox model with right censored data. I considered both discrete and continuous covariates and accommodated the difference in the type of covariates by introducing different versions of MHDE. Through extensive simulation studies, we examined the performance of the proposed MHDEs and compare them with the PMLE. Our numerical results showed that the proposed MHDEs are competitive with the PMLE under true model, while they outperforms the PMLE when outliers are present which testified the excellent robustness property of the proposed MHDEs. We also demonstrated the applications of the proposed MHDE to two real data analysis.

Acknowledgements

I would first like to thank my supervisor, Professor Jingjing Wu, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Secondly, I would like to express my gratitude to Dr. Chel Hee Lee and Dr. Xuwen Lu for their time spent on reading my thesis and their insightful comments on this thesis. I am also indebted to my course instructors, Dr. Xuwen Lu and Dr. Rob Deardon, for their valuable guidance throughout my studies. You helped me cultivate the sense of statistics and provided me with the tools that I needed to choose the right direction and successfully complete my thesis.

I would also like to express my sincere appreciation to the faculty and staff members of the Department of Mathematics and Statistics at University of Calgary for creating a supportive environment for my study.

Finally, I could not have completed this thesis without the support of my beloved parents and girlfriend, who kept me company and provided happy distractions to rest my mind outside of my research. Thank you all for supporting me in all ways.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
Cox PH	Cox Proportional Hazard model
OLS	ordinary least squares
OLM	ordinary linear regression model
GLM	generalized linear regression model
MLE	maximum likelihood estimation/estimator/estimate
PMLE	partial maximum likelihood estimation/estimator/estimate
MHDE	minimum Hellinger distance estimator
MPHDE	minimum profile Hellinger distance estimation/estimator/estimate
NA-estimator	Nelson-Aalen estimator
MSE	mean squared error
p.d.f.	probability density function
c.d.f.	cumulative distribution function
i.i.d.	independent and identically distributed
r.v.	random variable
i.i.d.	independent and identically distributed
$\|\|\cdot\|\|$	L_2-norm
T	event time
C	censoring time
X	observed time; $\min\{T, C\}$
δ	indicator of censoring; 1 if $T \leq C$ and 0 if otherwise
$f(t)$	probability density function of event time
$S(t)$	survival function
$\lambda(t)$	hazard rate function
$\lambda_0(t)$	baseline hazard rate function
$H(t)$	cumulative hazard rate function
z	vector of covariates; explanatory factors

Chapter 1

INTRODUCTION

In this chapter, we first review in Section 1.1 the Cox Proportional Hazard (PH) model, the model of our focus in this thesis. Then in Section 1.2 we present the most commonly used estimation, partial maximum likelihood estimator (PMLE), of the unknown coefficients in the Cox PH model.

1.1 Cox PH Model

Since first introduced by Cox (1972), Cox PH model has been widely used to study associations, such as in clinical trials with two arms (control and experimental) or in studies examining relationship between patient survival time and one or more predictor variables, i.e. covariates such as age and gender. Cox PH model examines the conditional hazard rate function of an event time given specified covariates as a product of an undetermined baseline hazard rate function and an exponential function of a linear combination of covariates with unknown coefficients, assuming proportionality across individuals with different values of covariates.

Let T be a continuous random variable (r.v.) that represents the time to event of interest and z be a p-dimensional vector of covariates. In the PH model, the hazard rate function of T given z is represented as a multiple linear regression of the logarithm of the hazard on the covariates z, with the baseline hazard that changes with time. In another word, the PH model is given by

$$\lambda(t|z) = \lambda_0(t)\exp(\beta^\top z), \tag{1.1}$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$ is the vector of coefficients and λ_0 is the baseline hazard function giving the hazard rate when covariates $z = 0$. The associated conditional survival function of T is

$$S(t|z) = \exp\left\{-H_0(t)\exp(\beta^\top z)\right\}, \tag{1.2}$$

where $H_0(t) = \int_0^t \lambda_0(u)du$ is the baseline cumulative hazard rate function. Cox model allows us to examine the impact of specified factor(s) on the rate of event of interest happening at time t with the coefficients β and consider the odds ratio of two individuals while treating baseline hazard rate function λ_0 as a nuisance parameter. Basu et al. (2004) observed that when the distribution of event time is skewed to the right, Cox PH model outperforms ordinary linear regression model (OLM) and generalized linear model (GLM).

1.2 Partial Maximum Likelihood Estimation (PMLE)

For Cox model (1.1) with right censored data, the commonly used estimation method for the unknown coefficients is the PMLE introduced by Cox (1975) which treats the baseline hazard function λ_0 as a nuisance parameter. For the i^{th} individual, let T_i and C_i denote the event time and censoring time respectively, $i = 1, \ldots, n$. Denote $X_i = \min\{T_i, C_i\}$ as the observed time and $\delta_i = I_{\{X_i \le C_i\}}$ as the observed censorship indicator, where I_A is the indicator function of event A which takes value 1 when A is true and takes value 0 otherwise. For the i^{th} individual, let z_i denote the covariates or factors that might influence the event time. Note that the triples (X_i, δ_i, z_i), $i = 1, \ldots, n$, are the actually observed data.

Following Cox (1975), let $t_1 < \cdots < t_D$ represent the D ordered distinct event times among a sample of n possibly censored observed times ($D \le n$). Assuming no tied event times, let $z_{(i)}$ be the covariates associated with the individual whose event time is t_i. Denote the risk set $R(t_i)$ as the collection of individuals who are free of the event of interest and uncensored just prior to t_i, i.e. those with observed time x such that $x \ge t_i$. Then the conditional probability that, given the occurrence of the event at time t_i, it occurred on the individual with covariates $z_{(i)}$ is given by

$$\frac{\lambda(t_i|z_{(i)})}{\sum\limits_{l\in R(t_i)} \lambda(t_i|z_l)} = \frac{\lambda_0(t_i)\exp(\beta^\top z_{(i)})}{\sum\limits_{l\in R(t_i)} \lambda_0(t_i)\exp(\beta^\top z_l)} = \frac{\exp(\beta^\top z_{(i)})}{\sum\limits_{l\in R(t_i)} \exp(\beta^\top z_l)}. \tag{1.3}$$

Multiplying these probabilities over all the observed k distinct event times gives the partial likelihood of β as

$$L_0(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^\top z_{(i)})}{\sum_{l\in R(t_i)} \exp(\beta^\top z_l)}. \tag{1.4}$$

Taking tied events (multiple events at the same time t_i) into account, the partial likelihood (1.4) can be generalized as

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^\top S_i)}{\left[\sum_{l\in R(t_i)} \exp\left(\beta^\top z_l\right)\right]^{d_i}}, \tag{1.5}$$

where d_i is the number of individuals with events occurred at time t_i and S_i is the sum of the covariate vectors of these d_i individuals.

Following the idea of maximum likelihood estimation (MLE), the parameter vector β can be estimated by finding the vector $\hat{\beta}$ that maximizes the partial likelihood $L(\beta)$ given in (1.5), or equivalently maximizes the partial log-likelihood function

$$\log L(\beta) = \sum_{i=1}^{D} \beta^\top S_i - d_i \log\left[\sum_{l\in R(t_i)} \exp(\beta^\top z_l)\right] \tag{1.6}$$

Setting the first derivatives of the partial log-likelihood function (1.6) with respect to β to zero yields the PMLE of β, denoted as $\hat{\beta}_{PML}$. Without explicit expression of solution, the PMLE is usually calculated through Newton-Raphson iteration or other numerical approach.

The PMLE has been proved to have good properties by Wong (1986) and many others, such as consistency and asymptotic normality. The PMLE method can be implemented and calculated in statistical programming language **R** using the package 'coxph'. Breslow (1974) and Kalbfleisch and Prentice (1973) obtained statistical inferences for baseline hazard function and cumulative hazard function, respectively.

Breslow (1972) showed that the PMLE is equivalent to estimations obtained by simultaneously maximising the full likelihood function with respect to regression coefficients and the baseline hazard function, with the assumption that the baseline hazard function is a step function with a constant hazard rate for all uncensored observations and all censored observations are censored at the preceding uncensored time. However, this equivalence is inefficient in terms of PMLE calculation, especially for big data sets. Clayton and Cuzick (1985) proposed an alternate estimation method, in which the baseline cumulative hazard rate function is iteratively estimated using Brewslow estimator and the regression coefficients are estimated by maximizing the full log-likelihood function of Cox model. However this approach does not guarantee convergence, while when it does the estimated coefficients converge to the PMLE.

In this thesis, we introduce a new estimation method based on minimum distance technique. More specifically, we propose minimum Hellinger distance estimation (MHDE) of the Cox PH model. Generally speaking, an MHDE is the estimated parameter value which minimizes the Hellinger distance between the assumed model and its nonparametric estimate. With the proposed methods, we hoped the proposed MHDEs achieve relatively high efficiency when model assumption is valid and at the same time retain good robustness properties when assumed model is not strictly correct or when data is contaminated. These attractive properties will be demonstrated for the proposed MHDEs through numerical studies in the thesis.

The remainder of the thesis is organized as follows. In Chapter 2, we introduce MHDE and construct different versions of MHDE for both quantitative and categorical covariates. We also present in Chapter 2 the kernel estimations of both conditional and non-conditional hazard functions. Chapter 3 presents an extensive simulation study which examines the finite-sample performance of the proposed MHDEs and compare them with with the PMLE. In Chapter 4, we demonstrate the application and implementation of the proposed MHDE via two real data analysis. Finally, some concluding remarks are given in Chapter 5.

Chapter 2

MINIMUM HELLINGER DISTANCE ESTIMATOR (MHDE)

In this chapter, we first review in Section 2.1 the core notion of MHDE in both parametric and semiparametric models and their applications in survival analysis. In Section 2.2, we propose and construct our MHDE for Cox PH model and accommodate different type of covariates by introducing different version of MHDE. Section 2.3 presents the kernel function estimation of hazard functions that we employ in the proposed MHDEs.

2.1 Review of MHDE

2.1.1 MHDE for parametric models

Classical estimation methods, such as MLE, are known to be efficient but sensitive to the presence of outliers. Comparatively, MHDE can provide a reliable alternative with good robustness properties while maintaining desirable properties such as consistency and asymptotic normality. As a result, MHDE has gained popularity and has been largely explored in the past decades.

Various estimators have been examined since Wolfowitz (1957) which initially established the class of minimal distance estimators. Among many other minimum distance estimators, MHDE was first proposed by Beran (1977) for univariate parametric density functions. Beran (1977) demonstrated that MHDE is stable in the presence of outliers and model perturbations within a small neighbourhood measured by Hellinger metric of the assumed model. Beran (1977) also proved that the use of the Hellinger distance for parametric models leads to an estimation asymptotically equivalent to MLE and thus achieves fully efficiency under the true parametric model assumption.

Suppose $X_1, X_2, \ldots, X_n$ are n independent and identically distributed (i.i.d.) random samples/variables (r.v.) following the probability density function (p.d.f.) f. For parametric modelling, f is assumed belonging to the parametric family $\{f_\theta : \theta \in \Theta\}$, where the finite-dimensional parameter space $\Theta \in \mathbb{R}^p$. Frequently Ω is assumed compact in current literatures. Suppose the true parameter value is θ, i.e. $f = f_\theta$. The Hellinger Distance between two functions f and g is defined as $\|f^{1/2} - g^{1/2}\|$, where $\|\cdot\|$ denotes the L_2-norm. Following Beran (1977), the MHD functional T_0 evaluated at a p.d.f. g is defined as the parameter value t which minimizes the Hellinger distance between g and the parametric model f_t, i.e.

$$T_0(g) = \underset{t\in\Theta}{\arg\min}\, \|f_t^{1/2} - g^{1/2}\|. \tag{2.1}$$

Suppose $\hat{f}_n$ is a suitable nonparametric density estimator of f (such as kernel estimator used in this thesis) based on the sample X_i's. Now the proposed MHDE $\hat{\theta}$ is the parameter value t in the parameter space Θ which minimizes the Hellinger distance between f_t and $\hat{f}_n$, i.e.

$$\hat{\theta} = T_0(\hat{f}_n) = \underset{t\in\Theta}{\arg\min}\, \|f_t^{1/2} - \hat{f}_n^{1/2}\|. \tag{2.2}$$

Donoho and Liu (1988) demonstrated that all minimum distance functionals are "automatically" robust over contamination neighbourhoods and proved that the functional associated with the MHDE has the lowest sensitivity to contamination among Fisher consistent functionals. Tamura and Boos (1986) focused on MHDE for parametric families within the class of elliptically symmetric distributions and particularly for multivariate normal distributions. They examined the MHDEs of the location and covariance parameters for multivariate cases and proved their affine invariance, consistency, asymptotically normality and robustness measured by breakdown point. MHDE was also proved to be efficient in analysis of count data (Simpson (1987)), branching processes (Sriram and Vidyashankar (2000)), and parametric finite mixture of Poisson regression models (Karlis and Xekalaki, 1998; Lu et al., 2003).

2.1.2 MHDE for semiparametric models

Suppose $X_1, X_2, \ldots, X_n$ are n i.i.d. r.v.s following the probability density function (p.d.f.) f. For semiparametric modelling, f is assumed belonging to the semiparametric family $\{f_{\theta,\eta} : \theta \in \Theta, \eta \in \Gamma\}$, where the parameter spaces $\Theta \in \mathbb{R}^p$ and Γ is a subset of some Banach space with norm $\|\cdot\|$. Assume the true parameter value is (θ, η), i.e. $f = f_{\theta,\eta}$. Suppose θ is the parameter of our interest, treating η as a nuisance parameter. Wu and Karunamuni (2012) proposed a plug-in type MHDE of θ. The idea is that if we could find a suitable estimator of the nuisance parameter η, say $\hat{\eta}$, then we can plug it in to obtain an estimated parametric model $f_{\theta,\hat{\eta}}$. Now the problem is reduced to estimating θ in the parametric model $f_{\theta,\hat{\eta}}$ and then the MHDE of θ can be defined as

$$\hat{\theta} = T_1(\hat{f}) = \underset{\theta \in \Theta}{\arg\min} \|f_{\theta,\hat{\eta}}^{1/2} - \hat{f}^{1/2}\|, \tag{2.3}$$

where again $\hat{f}$ is a suitable nonparametric (say kernel) density estimator of f based on X_i's. Wu and Karunamuni (2012) proved the consistency, asymptotic normality and robustness properties of the MHDE given by (2.3). The plug-in procedure is also used and studied in various models including two-sample semiparametric model (Wu et al., 2010), mixture model (Karunamuni and Wu, 2009) and others.

However, in some cases, a suitable estimator of the nuisance parameter η is hard to find in explicit form, or hard to be estimated at the parametric rate $O(n^{-1/2})$ when it is infinite-dimensional. This is the main restriction of plug-in type MHDE. Wu and Karunamuni (2015) proposed estimating θ through a profile approach which introduced a new estimator called minimum profile Hellinger distance estimator (MPHDE). The idea is that we first profile η out and then minimize the profiled Hellinger distance over the parameter space Θ to find the estimate of θ. Specifically, for any fixed $t \in \Theta$, let η_t denote the $h \in \Gamma$ that minimizes the Hellinger distance between $f_{t,h}$ and $\hat{f}$, i.e.

$$\eta_t = \eta_t(\hat{f}) = \underset{h \in \Gamma}{\arg\min} \|f_{t,h}^{1/2} - \hat{f}^{1/2}\|. \tag{2.4}$$

With this profiled η_t, the profiled Hellinger distance is $\|f_{t,\eta_t}^{1/2} - \hat{f}^{1/2}\|$. Now the MPHDE $\hat{\theta}$ is defined as the value $t \in \Theta$ which minimizes the profiled Hellinger distance, i.e.

$$\hat{\theta} = \arg\min_{t \in \Theta} \|f_{t,\eta_t}^{1/2} - \hat{f}^{1/2}\|. \tag{2.5}$$

It is usually not easy to derive the explicit expression of the profiled η_t, though it is feasible for some simpler models such as location model. As a result, one may give a numerical approximation of η_t and may update it iteratively. Suppose $\hat{\theta}^{(k)}$ and $\hat{\eta}^{(k)}$ are the estimates of θ and η respectively in the k^{th} iteration. Then similar to (2.4) one can find the updated estimate of η as

$$\hat{\eta}^{(k+1)} = \arg\min_{h \in \Gamma} \|f_{\hat{\theta}^{(k)},h}^{1/2} - \hat{f}^{1/2}\|, \tag{2.6}$$

and further the updated estimate of θ as

$$\hat{\theta}^{(k+1)} = \arg\min_{t \in \Theta} \|f_{t,\hat{\eta}^{(k+1)}}^{1/2} - \hat{f}^{1/2}\|. \tag{2.7}$$

One can update the estimates of θ and η iteratively until convergence. MPHDE has not been studied extensively except for location models (Wu and Karunamuni, 2015) and two-component mixture models (Xiang et al., 2014; Wu et al., 2017).

2.1.3 MHDE for survival models

In survival models, observations are possibly censored and mechanism needs to be found to deal with censorship. Yang (1991) discussed the MHDE of parametric density family when data are censored and investigated the tail-behaviour of product-limit process. The authors also proved the asymptotic efficiency of MHDE in the random censorship models. Zhu et al. (2013) considered MHDE for a two-sample semiparametric cure rate model with censored survival data. Note that above two papers constructed MHDE based on Hellinger distance between c.d.f.s or p.d.f.s, while in survival analysis survival function and hazard function are even more relevant than distribution function. Following this direction, Ying (1992) proposed a MHDE for parametric survival models based on Hellinger distance between hazard functions. Ying (1992) also proposed a weight function to tackle tail-behaviour and obtained the asymptotic normality of MHDE under considerably weaker conditions than normal assumptions. Nevertheless, all these research focused on one or two populations' lifetime without relating it to potential factors or covariates which will be focus of this thesis.

Since later our proposed method borrows some idea from Ying (1992), here we review the MHDE in Ying (1992). Suppose for n individuals, their event times $T_1,\ldots,T_n$ are i.i.d. r.v.s following parametric p.d.f. f_θ with corresponding c.d.f. F_θ, where the unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$. These events could possibly be right censored with censoring times $C_1,\ldots,C_n$. These C_i's are assumed mutually independent and independent of T_i's. Note that only pairs of (X_i,δ_i), $i=1,\ldots,n$, are observed, where $X_i=\min\{T_i,C_i\}$ and $\delta_i = I_{\{X_i\leq C_i\}}$ with I_A the indicator function of event A. Let $S_\theta(t) = 1-F_\theta(t)$ denote the survival function and $\lambda_\theta(t) = f_\theta(t)/[1-F_\theta(t)]$ be the hazard rate function. Define counting processes

$$\begin{aligned} N_n(x) &= \sum_{i=1}^n I_{\{X_i\leq x,\delta_i=1\}}, \quad \bar{N}_n(x) = N_n(x)/n, \\ Y_n(x) &= \sum_{i=1}^n I_{\{X_i\geq x\}}, \qquad\quad \bar{Y}_n(x) = Y_n(x)/n. \end{aligned} \tag{2.8}$$

These are commonly used empirical processes in survival analysis and proved to have asymptotic properties. Following Ying (1992), one can construct the kernel estimator of the underlying hazard rate function λ_θ as

$$\hat{\lambda}_n = \frac{1}{b_n}\int_{-\infty}^{\infty} K(\frac{x-u}{b_n})d\hat{\Lambda}_n(u) \ \text{ with } \ \hat{\Lambda}_n(x) = \int_{-\infty}^{x} \frac{dN_n(u)}{Y_n(u)}, \tag{2.9}$$

where K is a kernel function (a non-negative p.d.f.) and b_n is the bandwidth such that $b_n \to 0$ and

$nb_n \to \infty$ as $n \to \infty$. Then Ying (1992) proposed a MHDE, restricted to interval $[a,b]$, defined as

$$\hat{\theta}_n(a,b) = \underset{\theta\in\Theta}{\arg\min} \int_a^b \left[\lambda_\theta^{1/2}(x) - \hat{\lambda}_n^{1/2}(x)\right]^2 \bar{Y}_n(x)dx. \tag{2.10}$$

Here $\bar{Y}_n(x)$ can be viewed as a weight function to ensure consistency. Ying (1992) proved that $\hat{\theta}_n$ is asymptotically equivalent to the MLE if the parametric model assumption is valid.

2.2 Proposed MHDE for Cox PH Model

Frequently in survival analysis we are interested in studying the association of lifetime with potential factors. This thesis extends the MHDE of survival model for single population (without covariates) to models quantifying associations (with covariates). To start with, we will consider Cox PH model in this work. Suppose for the i^{th} individual, besides (X_i, δ_i) we also observe a covariate vector z_i, and thus the observed data are the triples (X_i, δ_i, z_i), $i = 1, \ldots, n$.

In the Cox PH model (1.1), assume the unknown parameter $\beta \in \mathbb{B}$ with $\mathbb{B}$ a compact subset of $\mathbb{R}^p$. In this section we construct MHDEs for PH model (1.1) with Hellinger distance based on hazard functions. Note that the Hellinger distance could be based on either baseline hazard functions or conditional hazard functions. Our numerical results show that the former performs slightly better and thus becomes our choice. We consider both the case of known baseline hazard function λ_0 and the case of unknown λ_0.

2.2.1 MHDE for categorical covariate

Cox PH model is widely used in clinical trials to test or measure the effect of certain medications, technologies or treatments. The main covariate of relevance in a Cox PH model of such situations would be a categorical factor indicating which treatment a patient is receiving. In other words, these observations may be categorized into groups based on their covariate, which can only take categorical values such as 0,1 and 2. Note that in categorical circumstances, the group with $z = 0$

is also known as the reference group.

Assume λ_0 is known. Taking the grouping effect into consideration, we may make the following changes to our notations and methodology. Suppose we observe n triples of time X, indicator of censorship δ and covariate z, and these n individuals are grouped into k groups ($k << n$) corresponding to k different covariate values $z_1, \ldots, z_k$. Therefore we use $(X_{ij}, \delta_{ij}, z_i)$ to denote the observation of the j^{th} individual in the i^{th} group with covariate z_i, $i = 1, 2, ..., k$ and $j = 1, 2, ..., n_i$, where n_i is the number of individuals with covariate z_i so that $\sum_{i=1}^{k} n_i = n$. By the fact that $\lambda_0(t) = \lambda(t|z)/\exp(\beta^\top z)$, we define

$$v_{ij}(b) = \frac{\hat{\lambda}(X_{ij}|z_i)}{\exp(b^\top z_i)}, \quad i = 1, \ldots, n, \tag{2.11}$$

where $\hat{\lambda}(t|z)$ is an appropriate nonparametric estimator of the conditional hazard function $\lambda(t|z)$. Such an estimator is given with details in Section 2.3. With the grouping information, the estimates $\hat{\lambda}(X_{ij}|z_i)$, $j = 1, \ldots, n_i$, can be obtained nonparametrically based on data in the i^{th} group only, which is more cost-effective and has better asymptotic properties.

When $b = \beta$, $v_{ij}(\beta)$ can be treated as an estimate of $\lambda_0(X_{ij})$. Now we need to estimate the function $\lambda_0(t)$ given the pairs $(X_{ij}, v_{ij}(\beta))$. One way is to use kernel regression (see details in Appendix) to provide a smooth estimate of λ_0, denoted as $\hat{\lambda}_{0,\beta}$. As a result, $\hat{\lambda}_{0,b}$ and λ_0 should be very close to each other when we use $b = \beta$ or b very close to β. Following this idea, the estimate of β can be defined as the value $b \in \mathbb{B}$ which minimizes the Hellinger distance between $\hat{\lambda}_{0,b}$ and λ_0, i.e.

$$\hat{\beta} = \arg\min_{b \in \mathbb{B}} \left\| \lambda_0^{1/2}(t) - \hat{\lambda}_{0,b}^{1/2}(t) \right\|. \tag{2.12}$$

In order to make the Hellinger distance meaningful, the Hellinger distance is calculated over the range formed by the smallest and largest observed time. The calculation of the MHDE defined in such a way is straight forward, nevertheless it is computationally heavier compared to the method below.

In order to ease the calculation of the MHDE, we use a discretized version of the definition given in (2.12). Empirically, we can minimize the sum of point-wise Hellinger distance between λ_0 and its estimator at each observed time. In another word, we define the MHDE of β as

$$\hat{\beta}_{MHD1} = \arg\min_{b\in\mathbb{B}} \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left\{ \lambda_0^{1/2}(X_{ij}) - v_{ij}^{1/2}(b) \right\}^2. \tag{2.13}$$

Assume λ_0 is unknown. Recall that in Cox PH model, the baseline hazard function λ_0 can be viewed as the hazard function when the covariate $z = 0$. In case we have enough data in the group with $z = 0$, we can estimate λ_0 using the data in such group only. In case we don't have enough data in the group with $z = 0$, we can use the nonparametric (e.g. kernel) estimator $\hat{\lambda}(t|z)$ of the hazard function $\lambda(t|z)$ to be introduced in Section 2.3 and set $z = 0$ to get an estimator of λ_0. In either case, we denote the estimator of λ_0 as $\hat{\lambda}(t|0)$. Now the MHDE can be defined as

$$\hat{\beta}_{MHD2} = \arg\min_{b\in\mathbb{B}} \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left\{ \hat{\lambda}^{1/2}(X_{ij}|0) - v_{ij}^{1/2}(b) \right\}^2. \tag{2.14}$$

2.2.2 MHDE for quantitative covariate

Now we consider the case that the covariate is quantitative. We first assume λ_0 is known. Suppose we have a nonparametric (e.g. kernel) estimator of the hazard function $\lambda(t|z)$, again denoted as $\hat{\lambda}(t|z)$. Since there is no grouping effect in quantitative covariate case, instead of $v_{ij}(b)$ given in (2.11) for categorical covariate, we define

$$v_i(b) = \frac{\hat{\lambda}(X_i|z_i)}{\exp(b^\top z_i)}, \quad i = 1,\ldots,n. \tag{2.15}$$

Accordingly, the MHDE of β with known λ_0 is defined as

$$\hat{\beta}_{MHD3} = \arg\min_{b\in\mathbb{B}} \sum_{i=1}^{n} \left\{ \lambda_0^{1/2}(X_i) - v_i^{1/2}(b) \right\}^2. \tag{2.16}$$

Assume λ_0 is unknown, then similarly we can set $z = 0$ in the same kernel estimator $\hat{\lambda}(t|z)$, i.e.

$\hat{\lambda}(t|0)$ to derive a nonparametric estimator of λ_0 as discussed in Section 2.2.1. Now the MHDE of β can be defined as

$$\hat{\beta}_{MHD4} = \arg\min_{b \in \mathbb{B}} \sum_{i=1}^{n} \left\{ \hat{\lambda}^{1/2}(X_i|0) - v_i^{1/2}(b) \right\}^2, \tag{2.17}$$

where $v_i(b)$ is given by (2.15).

2.2.3 MHDE for mixed types of covariates

In practice, the covariates of interest may be mixed-type consisting of both quantitative and categorical factors. In such cases, there is no grouping effect as for the case of quantitative only covariates and thus we can follow the same steps and methods as in Section 2.2.2 to construct $v_i(b)$ and a nonparametric estimator $\hat{\lambda}(t|z)$ via kernel method. When λ_0 is known, the MHDE is the same as (2.16). When λ_0 is unknown, with λ_0 replaced by its estimator $\hat{\lambda}(\cdot|0)$, the MHDE is given as in (2.17).

2.3 Kernel Estimation of Hazard Function

As shown in Section 2.2, in the definition of the proposed MHDEs, we need to estimate the conditional hazard rate for each observation when constructing $v_i(b)$ or $v_{ij}(b)$. In this section we discuss two commonly used approaches to estimate hazard function and one of which is used in our simulation study.

2.3.1 Estimation of conditional hazard function

In survival analysis, it is important and useful to estimate the hazard rate at any time t, i.e. the hazard rate function $\lambda(t)$. In practice, the hazard rate may vary from one individual to the other and the hazard function for a specific individual may be affected by its characteristics such as age and gender. Therefore, it is even crucial to estimate the conditional hazard function $\lambda(t|z)$ on a set of specified covariates z in survival models such as Cox PH model in this thesis and accelerated failure time (AFT) model.

Spierdijk (2008) proposed an estimation of conditional hazard function written as the ratio of local linear estimators of conditional density and survival function respectively, which is straightforward by definition. Given covariate z, it is assumed that the event time T and the censoring time C are independent with conditional survival functions $S(\cdot|z)$ and $G(\cdot|z)$ respectively and conditional p.d.f.s $f(\cdot|z)$ and $g(\cdot|z)$ respectively. Let $L(\cdot|z)$ denote the conditional survival function of the observed time $X := \min\{T, C\}$ given covariate z, and let $l(\cdot,\cdot|z)$ denote the conditional density of (X, δ) given z. Specially, we use $r(\cdot|z) := l(\cdot, 1|z)$ to denote the conditional density of X given z for an uncensored case. Followed by the independence of T and C given z, we have

$$L(\cdot|z) = S(\cdot|z)G(\cdot|z). \tag{2.18}$$

Consequently,

$$r(\cdot|z) = f(\cdot|z)G(\cdot|z). \tag{2.19}$$

Recall that the conditional hazard function $\lambda(t|z)$ is defined as

$$\lambda(t|z) = \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t | T > t, z)}{\Delta t} = \frac{f(t|z)}{S(t|z)}. \tag{2.20}$$

From (2.18)-(2.20) we have

$$\lambda(t|z) = \frac{f(t|z)}{S(t|z)} = \frac{r(t|z)}{L(t|z)}. \tag{2.21}$$

Spierdijk (2008) considered the local linear estimators for $r(t|z)$ and $L(t|z)$, where the local linear estimation is a special case of local polynomial regression and has better bias properties than Linton and Nielsen (1995)'s smoothing estimator.

Following the idea of (2.21), Selingerová et al. (2014) provided another method to estimate $r(t|z)$ and $L(t|z)$. For the i^{th} individual with covariate z_i, the Nadaraya-Watson weight is defined as

$$w_i(z) = \frac{K(\frac{z-z_i}{h_1})}{\sum_{j=1}^{n} K(\frac{z-z_i}{h_1})}, \tag{2.22}$$

where K is a kernel function (i.e. a non-negative p.d.f.) and h_1 is the bandwidth such that $h_1 \to 0$ as $n \to \infty$. Note that the choice of h_1 depends on sample size n but we remove this dependence in notation for notation simplicity. Since the choice of kernel function in principle doesn't influence the asymptotic properties of kernel estimation, in all our numerical studies, we always use the same Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ for simplicity. For the bandwidth, we always use $n^{-1/5}$ in our simulation study. This choice of bandwidth ensures the optimal convergence rate of the integrated mean square error (IMSE) (Rao, 2014). The Nadaraya-Watson weights are commonly used in kernel estimation of regression function. With this weight function, the conditional density $r(t|z)$ and the conditional survival function $L(t|z)$ at time t can be respectively estimated as locally weighted averages of corresponding functions at observed times within a neighbourhood of t. More specifically, we can estimate $r(t|z)$ using a weighted kernel density estimator given by

$$\hat{r}(t|z) = \frac{1}{h_2} \sum_{i=1}^{n} \delta_i w_i(z) K\left(\frac{t - X_i}{h_2}\right), \tag{2.23}$$

where h_2 is a bandwidth such that $h_2 \to 0$ as $n \to \infty$. Note that one may choose different kernel function K in (2.22) and (2.23). Since the conditional p.d.f. of X given z, $l(x|z) = -\frac{\partial}{\partial t} L(x|z)$, can be estimated by a kernel estimation $\hat{l}(x|z) = \frac{1}{h_3} \sum_{i=1}^{n} w_i(z) K\left(\frac{x-X_i}{h_3}\right)$, by the fact that $L(t|z) = \int_t^\infty l(x|z)dx$, the conditional survival function $L(t|z)$ can be estimated by

$$\hat{L}(t|z) = \sum_{i=1}^{n} w_i(z) W\left(\frac{t - X_i}{h_3}\right), \tag{2.24}$$

where $W(x) = \int_x^\infty K(u)du$ is the cumulative function of kernel K and h_3 is the bandwidth such that $h_3 \to 0$ as $n \to \infty$. Now by (2.21), (2.23) and (2.24), the conditional hazard function $\lambda(t|z)$ can be estimated by

$$\hat{\lambda}(t|z) = \frac{\hat{r}(t|z)}{\hat{L}(t|z)} = \frac{\frac{1}{h_2} \sum_{i=1}^{n} \delta_i w_i(z) K(\frac{t-X_i}{h_2})}{\sum_{i=1}^{n} w_i(z) W(\frac{t-X_i}{h_3})}. \tag{2.25}$$

In our simulation study, we write a self-defined function to calculate the kernel estimation of hazard function given by (2.25), with which we can estimate $\lambda(X_i|z_i)$ and construct v_i or v_{ij} in the following steps.

2.3.2 Estimation of non-conditional hazard function

In our proposed MHDE (2.13) for categorical covariates, within the group with covariate z, the estimation of the conditional hazard function $\lambda(\cdot|z)$ could be an estimation of a non-conditional hazard function based on the data in this group only. In this subsection, we introduce rank kernel estimation of, particularly, the baseline hazard function that is used in our simulation study.

Recall the cumulative hazard function $H(t)$ of T is given by

$$H(t) = \int_0^t \lambda(u)du = -\log S(t).$$

One of most popular estimators of cumulative hazard function H is the Nelson-Aalen (N-A) estimator which considers the cumulative counts at observed time. Assume that both T and C are continuous r.v.s so that there is a zero chance that we observe two identical X_i's. Thus without loss of generality, we assume all the X_i's are distinct. Denote R_i as the rank of X_i and $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the ordered observations. For notation convention, let $X_{(0)} = 0$ and $X_{(n+1)} = +\infty$. Then the N-A estimator is given by

$$\hat{H}(t) = \sum_{i=1}^{k} \frac{\delta_i}{n - R_i + 1} \text{ when } X_{(k)} \leq t < X_{(k+1)}, \ \ k = 0, 1, \ldots, n.$$

which is a step function that is flat between the k^{th} and $(k+1)^{th}$ ordered observed time $X_{(k)}$ and $X_{(k+1)}$. Ramlau-Hansen (1983) proposed a kernel estimator of the hazard function given by

$$\hat{\lambda}(t) = \frac{1}{h_4} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_4}\right) \frac{\delta_i}{n - R_i + 1}, \tag{2.26}$$

where h_4 is the bandwidth such that $h_4 \to 0$ as $n \to \infty$. The $\hat{\lambda}$ in (2.26) can be treated as a convolution of the N-A estimator with a kernel weighting function. While h_4 in (2.26) is a global bandwidth depending only on sample size n, Hess et al. (1999) made an extension allowing the bandwidth to depend on t and thus proposed an adaptive hazard function estimator given by

$$\hat{\lambda}(t) = \frac{1}{h_t} \sum_{i=1}^{n} K\left(\frac{t - X_{(i)}}{h_t}\right) \frac{\delta_i}{n - R_i + 1}, \tag{2.27}$$

where h_t is the locally optimal bandwidth h which minimizes the local mean square error (MSE)

$$MSE(t,h) = Var(t,h) + Bias^2(t,h), \tag{2.28}$$

i.e. $h_t = \min_h MSE(t,h)$. Here $Bias(t,h)$ and $Var(t,h)$ denote the bias and variance of the estimator $\hat{\lambda}(t)$ in (2.27) when the bandwidth takes value h. In fact, one can choose other objective functions instead of (2.28) to obtain optimal bandwidth in other sense. The benefit of using (2.28) is that it produces a hazard function estimator with the smallest MSE while MSE is a commonly used criteria to assess the performance of an estimator.

Based on Hess et al. (1999)'s work, an **R** package called 'muhaz' was developed and widely used for estimating non-conditional hazard rate function. In our simulation study, we adopt the estimator (2.27) and use this package to estimate the conditional hazard $\lambda(X_{ij}|z_i)$ in a non-conditional way based on observations in the group with covariate z_i only. When the baseline hazard function $\lambda_0(t) = \lambda(t|0)$ is unknown, we also use 'muhaz' to calculate its estimator $\hat{\lambda}_0$ using the data in the group with $z = 0$ only.

Chapter 3

SIMULATION STUDY

In this chapter, we conduct simulation studies using **R** to examine the finite sample performance of our proposed MHDEs and compare them with some existing methods in literature. We first present in Section 3.1 some model settings for our numerical studies. In Sections 3.2 and 3.3 we present our simulation results for case of single categorial covariate and case of single quantitative covariate respectively, while Section 3.4 gives the results for case of two dimensional covariates of three various types. When investigating the finite sample performance of the proposed MHDEs, we examine not only the efficiency but also the robustness properties of the proposed MHDEs.

3.1 Simulation Settings

In all our simulation studies, we arbitrarily choose linear baseline hazard function $\lambda_0(t) = 0.1 + 0.2t$. Then given the covariate z, the hazard function, cumulative hazard functioin, survival function and c.d.f. are given respectively by

$$\begin{aligned}
\lambda(t|z) &= (0.1+0.2t)\exp(\beta^\top z), \\
H(t|z) &= (0.1t+0.1t^2)\exp(\beta^\top z), \\
S(t|z) &= \exp[-H(t|z)] = \exp\left[-(0.1t+0.1t^2)\exp(\beta^\top z)\right], \\
F(t|z) &= 1-S(t|z).
\end{aligned}$$

We consider varying sample sizes $n = 50, 100, 150, 300, 900$ and always take $N = 100$ repetitions in all our calculations. We investigate the efficiency of an estimator by examining its bias and MSE given by

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta), \quad MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)^2,$$

where $\hat{\beta}_i$ is the estimate of β in the i^{th} repetition, $i = 1, \ldots, N$.

3.2 Single Categorical Covariate

Consider the Cox PH model with a single categorial covariate. Thus β is one-dimensional. We take $\beta = 1$ as the true value and assume the covariate z takes three different values 0, 1 and 2. We consider sample sizes $n = 150, 300, 900$ and consider both balanced sampling and unbalanced sampling. For the former, equal number of observations are drawn from each of the three populations corresponding to $z = 0, 1, 2$ respectively. In the latter, the sizes of the three populations are at the ratio of either $1:2:2$ or $3:1:1$.

For the group with $z = 0$, the survival function and the p.d.f. are given respectively by

$$\begin{aligned} S(t|z=0) &= \exp[-(0.1t + 0.1t^2)], \\ f(t|z=0) &= -\frac{dS(t|Z=0)}{dt} = (0.1 + 0.2t)\exp\left[-(0.1t + 0.1t^2)\right]. \end{aligned}$$

Given these functions, the survival time can be generated via the inversion sampler (Bender et al., 2005). We similarly generate the event times in groups with $z = 1$ and $z = 2$. For individuals in the group with $z = i$, $i = 0, 1, 2$, the censoring time C_{ij} is randomly selected from the uniform distribution over interval I_i, where I_i's vary from group to group and they are chosen to control the overall censoring rate to be around either 20% or 40%. As a result, we use $I_0 = (0, 12)$, $I_1 = (0, 9)$ and $I_2 = (0, 6.5)$ for censoring rate 20% and $I_0 = (0, 5.9)$, $I_1 = (0, 4.4)$ and $I_2 = (0, 3.3)$ for censoring rate 40%.

We examine the performance of the MHDEs given in (2.13) when λ_0 is known and in (2.14) when λ_0 is unknown, and compare their performance not only with the PMLE presented in Section 1.2 based on the partial log-likelihood given in (1.6), but also with the MHDE (2.10) proposed by Ying (1992). Note that the MPHDE in Ying (1992) is for parametric models and it doesn't incorporate covariates, thus for comparison purpose we modify Ying (1992)'s method in such a way that we minimize the sum, over $z = 0, 1, 2$, of the weighted Hellinger distance between the assumed parametric conditional hazard function $\lambda(t|z) = \lambda_0(t)\exp(\beta^\top z)$ and its kernel estimator $\hat{\lambda}(t|z)$, i.e.

$$\hat{\beta}_{Ying1} = \arg\min_{\theta\in\Theta} \sum_{z=0}^{2} \int \left[\left(\lambda_0(x)e^{\theta z}\right)^{1/2} - \hat{\lambda}^{1/2}(x|z) \right]^2 \bar{Y}_n(x)dx, \tag{3.1}$$

where the weight function $\bar{Y}_n(x) = \sum_{i=1}^{n} I_{\{X_i \geq x\}}/n$ is given in (2.8) as proposed by Ying (1992).

The MHDEs given in (2.13) and (2.14) are calculated following the estimation procedures described in Chapter 2 with self-written $\mathbb{R}$ code. We also write our own code to calculate the Ying (1992)'s MHDE $\hat{\beta}_{Ying1}$ given in (3.1). To calculate the PMLE with partial log-likelihood given in (1.6), we use the well-developed 'coxph' function in $\mathbb{R}$ package 'survival'. The numerical results for the cases when λ_0 is known and unknown are given in Subsections 3.2.1 and 3.2.2.

3.2.1 λ_0 known

When λ_0 is known, the difference between our proposed MHDE $\hat{\beta}_{MHD1}$ given in (2.13) and the $\hat{\beta}_{Ying1}$ in (3.1) by Ying (1992) is that our Hellinger distance is based on baseline hazard and is an empirical version while Ying (1992)'s is based on conditional hazard with attached weight. Table 3.1 displays the results of the proposed $\hat{\beta}_{MHD1}$, Ying (1992)'s $\hat{\beta}_{Ying1}$ and the PMLE $\hat{\beta}_{PML}$ for varying sample sizes $n = 150, 300, 900$ and censoring rates $0, 20\%, 40\%$.

Table 3.1: Bias (MSE) of $\hat{\beta}_{MHD1}$, $\hat{\beta}_{Ying1}$ and $\hat{\beta}_{PML}$ for single categorical covariate and known λ_0

Censoring Rate	n	$\hat{\beta}_{MHD1}$	$\hat{\beta}_{Ying1}$	$\hat{\beta}_{PML}$
0	150	.023 (.002)	.076 (.008)	.015 (.011)
	300	.016 (.002)	.058 (.025)	.001 (.006)
	900	.011 (.001)	.046 (.001)	.003 (.002)
20%	150	.008 (.005)	.083 (.015)	.003 (.020)
	300	.008 (.003)	.060 (.008)	.013 (.008)
	900	.001 (.001)	.050 (.004)	.003 (.003)
40%	150	.009 (.007)	.097 (.019)	.032 (.029)
	300	.001 (.004)	.088 (.014)	.004 (.012)
	900	.005 (.001)	.052 (.005)	.004 (.004)

From Table 3.1 we observe that in general $\hat{\beta}_{MHD1}$ has the best performance, followed by $\hat{\beta}_{PML}$

and then $\hat{\beta}_{Ying1}$. Note that both $\hat{\beta}_{MHD1}$ and $\hat{\beta}_{Ying1}$ make use of the known λ_0 while $\hat{\beta}_{PML}$ doesn't, which explains why $\hat{\beta}_{PML}$ performs a bit worse than $\hat{\beta}_{MHD1}$. More specifically, when bias is concerned, $\hat{\beta}_{Ying1}$ always performs far worst, and $\hat{\beta}_{MHD1}$ outperforms $\hat{\beta}_{PML}$ when censoring is present while it performs worse than $\hat{\beta}_{PML}$ when no censoring. When MSE is concerned, $\hat{\beta}_{MHD1}$ always has much better performance than $\hat{\beta}_{Ying1}$ and $\hat{\beta}_{MPL}$ which two are competitive. Intuitively the performance of all the three estimators improve when sample size increases. Surprisingly, the bias and MSE of $\hat{\beta}_{MHD1}$ seems very stable when censoring rate increases while obvious increase is observed in the bias and MSE of $\hat{\beta}_{PML}$ and $\hat{\beta}_{Ying1}$. In terms of computing time, when sample sizes grow, $\hat{\beta}_{Ying1}$ becomes more time-consuming than the other two methods.

3.2.2 λ_0 unknown

When λ_0 is unknown, our proposed MHDE is $\hat{\beta}_{MHD2}$ given in (2.14). In $\hat{\beta}_{MHD2}$, we use the method described in Subsection 2.3.2 (i.e. (2.27)) to give the baseline hazard estimator $\hat{\lambda}(t|0)$ based on the data with $z=0$ only. When λ_0 is unknown, Ying (1992)'s method is not applicable as it needs to make use of λ_0. Intuitively, we can use the same estimator $\hat{\lambda}(t|0)$ as in $\hat{\beta}_{MHD2}$ to replace the unknown λ_0. Thus Ying (1992)'s MHDE becomes

$$\hat{\beta}_{Ying2} = \arg\min_{\theta\in\Theta} \sum_{z=0}^{2} \int \left[\left(\hat{\lambda}(x|0) e^{\theta z} \right)^{1/2} - \hat{\lambda}^{1/2}(x|z) \right]^2 \bar{Y}_n(x) dx, \tag{3.2}$$

where the weight function $\bar{Y}_n$ is the same as the one used in (3.1).

Since our baseline hazard estimator $\hat{\lambda}(t|0)$ is based on the observations with $z=0$ only, the size of such a group influences its performance. In order to examine this influence, we consider both the balanced sampling, with the sample size ratio of the three groups corresponding to $z=0,1,2$ being 1 : 1 : 1, and unbalanced sampling with the sample size ratio of both 1 : 2 : 2 and 3 : 1 : 1. The simulation results for $\hat{\beta}_{MHD2}$, $\hat{\beta}_{Ying2}$ and $\hat{\beta}_{PML}$ are presented in Table 3.2 for balanced sampling, Tables 3.3 and 3.4 for sample size ratio of 1 : 2 : 2 and 3 : 1 : 1 respectively.

From Table 3.2 we observe that in general $\hat{\beta}_{PML}$ has the best performance, followed by $\hat{\beta}_{MHD2}$

Table 3.2: Bias (MSE) of $\hat{\beta}_{MHD2}$, $\hat{\beta}_{Ying2}$ and $\hat{\beta}_{PML}$ for single categorical covariate and unknown λ_0 with balanced sampling (group size ratio 1 : 1 : 1)

Censoring Rate	n	$\hat{\beta}_{MHD2}$	$\hat{\beta}_{Ying2}$	$\hat{\beta}_{PML}$
0	150	.056 (.006)	.066 (.007)	.006 (.005)
	300	.042 (.003)	.049 (.003)	.006 (.002)
	900	.030 (.002)	.039 (.003)	.001 (.001)
20%	150	.047 (.026)	.094 (.029)	.017 (.022)
	300	.060 (.012)	.079 (.015)	.005 (.009)
	900	.039 (.005)	.043 (.005)	.008 (.002)
40%	150	.045 (.024)	.095 (.030)	.047 (.029)
	300	.062 (.020)	.094 (.023)	.011 (.019)
	900	.043 (.006)	.054 (.006)	.005 (.004)

Table 3.3: Bias (MSE) of $\hat{\beta}_{MHD2}$, $\hat{\beta}_{Ying2}$ and $\hat{\beta}_{PML}$ for single categorical covariate and unknown λ_0 with group size ratio 1 : 2 : 2

Censoring Rate	n	$\hat{\beta}_{MHD2}$	$\hat{\beta}_{Ying2}$	$\hat{\beta}_{PML}$
0	150	-.103 (.021)	-.120 (.026)	.011 (.018)
	300	-.083 (.012)	-.086 (.012)	.016 (.011)
	900	-.054 (.005)	-.052 (.005)	.001 (.003)
20%	150	-.093 (.030)	-.120 (.032)	-.008 (.024)
	300	-.074 (.020)	-.085 (.019)	.005 (.010)
	900	-.047 (.006)	-.044 (.005)	.005 (.003)
40%	150	-.036 (.045)	-.083 (.040)	.047 (.038)
	300	-.059 (.027)	-.080 (.026)	.004 (.018)
	900	-.052 (.009)	-.057 (.009)	.009 (.004)

and then $\hat{\beta}_{Ying2}$. More specifically, when bias is concerned, $\hat{\beta}_{PML}$ always performs much better than $\hat{\beta}_{MHD2}$ which in turn performs much better than $\hat{\beta}_{Ying2}$. When MSE is concerned, the three methods yield very similar MSEs even though $\hat{\beta}_{PML}$ produces the smallest MSE while $\hat{\beta}_{Ying2}$ produces the largest in most cases. Intuitively the performance of all the three estimators improve when sample size increases and deteriorate when censoring rate increases. In Tables 3.3 and 3.4 for unbalanced samplings, the findings are are very similar to those for balanced sampling.

Since we observe from Tables 3.1-3.4 that Ying (1992)'s method always performs worse than

Table 3.4: Bias (MSE) of $\hat{\beta}_{MHD2}$, $\hat{\beta}_{Ying2}$ and $\hat{\beta}_{PML}$ for single categorical covariate and unknown λ_0 with group size ratio 3 : 1 : 1

Censoring Rate	n	$\hat{\beta}_{MHD2}$	$\hat{\beta}_{Ying2}$	$\hat{\beta}_{PML}$
	150	-.058 (.007)	-.096 (.015)	.024 (.012)
0	300	-.070 (.009)	-.097 (.015)	-.009 (.009)
	900	-.037 (.002)	-.054 (.005)	.004 (.002)
	150	-.072 (.023)	-.132 (.032)	-.002 (.018)
20%	300	-.042 (.009)	-.081 (.015)	.015 (.008)
	900	-.037 (.004)	-.054 (.007)	-.001 (.003)
	150	-.058 (.031)	-.134 (.042)	.004 (.030)
40%	300	-.047 (.012)	-.125 (.027)	-.017 (.012)
	900	-.036 (.005)	-.066 (.005)	.001 (.004)

our proposed MHDE method, in the following studies we remove Ying (1992)'s method and only focus on the comparison with PMLE.

3.3 Single Quantitative Covariate

Consider the Cox PH model with a single quantitative covariate. We take $\beta = 0.5$ as the true value and generate z_i randomly from uniform distribution $U(0.1, 2)$. We set the lower bound as 0.1 to avoid the random chaos when z is too close to zero. For each z_i, we generate the event time T_i following the assumed model. The censoring time C_i is randomly selected from an interval chosen such that it ensures the overall censoring rate is closed to the designed 20% or 40%. As a result, we use the intervals $(0, 9)$ and $(0, 4.8)$ for censoring rates 20% and 40% respectively.

For a single quantitative covariate, the proposed MHDE is $\hat{\beta}_{MHD3}$ given in (2.16) when λ_0 is known and is $\hat{\beta}_{MHD4}$ given in (2.17) when λ_0 is unknown. Since now z is quantitative, it may possibly take outlying values in practice that don't follow an assumed distribution. Thus in such quantitative covariate cases, we not only examine the efficiency of the proposed MHDEs when model assumption is valid, but also scrutinize their robustness when outlying observations are present.

3.3.1 Efficiency study

When data is generated following exactly the assumed model, the simulation results for $\hat{\beta}_{MHD3}$ with known λ_0, $\hat{\beta}_{MHD4}$ with unknown λ_0 and $\hat{\beta}_{PML}$ are given in Table 3.5.

Table 3.5: Bias (MSE) of $\hat{\beta}_{MHD3}$ with known λ_0, $\hat{\beta}_{MHD4}$ with unknown λ_0 and $\hat{\beta}_{PML}$ for single quantitative covariate

Censoring Rate	n	$\hat{\beta}_{MHD3}$	$\hat{\beta}_{MHD4}$	$\hat{\beta}_{PML}$
0	50	.009 (.015)	-.095 (.065)	.017 (.081)
	100	.005 (.009)	-.056 (.035)	.029 (.028)
	300	-.006 (.004)	-.005 (.025)	-.026 (.010)
20%	50	.005(.022)	.004(.100)	.021 (.101)
	100	.046 (.013)	-.004 (.058)	.025 (.054)
	300	.005 (.004)	-.010 (.021)	.003 (.013)
40%	50	.029 (.027)	.046 (.097)	.043 (.092)
	100	.048 (.014)	.039 (.087)	.061 (.070)
	300	.002 (.004)	.010 (.026)	.004 (.016)

From Table 3.5 we observe that when λ_0 is known, $\hat{\beta}_{MHD3}$ has much smaller MSE in all cases and much smaller bias in most cases than $\hat{\beta}_{PML}$. Again the higher efficiency in $\hat{\beta}_{MHD3}$ than $\hat{\beta}_{PML}$ could be explained by the fact that the former makes use of the known λ_0 while the latter doesn't. When λ_0 is unknown, $\hat{\beta}_{PML}$ outperforms $\hat{\beta}_{MHD4}$ in the sense that their bias are competitive while $\hat{\beta}_{PML}$ generally has smaller MSE than $\hat{\beta}_{MHD4}$.

3.3.2 Robustness study

In this subsection, we examine the robustness properties of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ when data is contaminated with outlying observations.

Study A: Influence of an outlier

We firstly examine how the estimates change when a single outlying observation is added. To add an outlier after simulating the data, we select the individual with smallest uncensored event time and then replace its observed event time and covariate with outlying values. Specifically, since the covariates are generated from $U(0.1,2)$ and the observed times corresponding to $z=4$ are typically in the range $(0,2)$, we replace the covariate of the selected individual with value 4 and replace the observed event time with value 4. By the ways we choose it, the outlier is always one of the largest observed time in the sample and thus appears in most of the risk sets. The simulation results for $\hat{\beta}_{MHD3}$ with known λ_0, $\hat{\beta}_{MHD4}$ with unknown λ_0 and $\hat{\beta}_{PML}$ are presented in Table 3.6, in which the results for both before and after the outlier being added are given and compared.

Table 3.6: Bias (MSE) of $\hat{\beta}_{MHD3}$ with known λ_0, $\hat{\beta}_{MHD4}$ with unknown λ_0 and $\hat{\beta}_{PML}$ for single quantitative covariate, before and after single outlier added

Censoring		$\hat{\beta}_{MHD3}$		$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
Rate	n	before	after	before	after	before	after
0	50	.017 (.016)	-.042 (.017)	-.056 (.081)	-.089 (.085)	.043 (.099)	-.501 (.261)
	100	.009 (.009)	-.024 (.009)	-.069 (.041)	-.091 (.046)	-.016 (.038)	-.386 (.156)
	300	.007 (.003)	-.004 (.003)	-.034 (.020)	-.044 (.020)	-.001 (.010)	-.221 (.052)
20%	50	-.026 (.019)	-.056 (.019)	-.027 (.098)	.060 (.099)	.086 (.104)	-.506 (.272)
	100	.029 (.012)	.033 (.012)	-.024 (.049)	-.066 (.052)	-.027 (.037)	-.414 (.178)
	300	-.008 (.003)	.020 (.004)	-.055 (.031)	-.068 (.032)	-.009 (.016)	-.246 (.066)
40%	50	-.020 (.006)	-.068 (.010)	-.055 (.037)	-.098 (.054)	.026 (.022)	-.253 (.072)
	100	.023 (.010)	-.019 (.009)	.052 (.078)	.055 (.086)	.035 (.041)	-.433 (.196)
	300	.004 (.006)	-.009 (.006)	-.006 (.024)	-.029 (.025)	.014 (.018)	-.269 (.078)

From Table 3.6 we observe that when the outlying observation is added, $\hat{\beta}_{PML}$ is significantly influenced in the sense that its bias increases approximately tenfold and its MSE also increases sharply in most cases. Comparatively, when λ_0 is known, $\hat{\beta}_{MHD3}$ is influenced very little by the outlier and its bias and MSE have small changes before and after the outlier is added. When λ_0 is unknown, $\hat{\beta}_{MHD4}$ performs worse than $\hat{\beta}_{PML}$ in most cases, but after the outlier is added $\hat{\beta}_{MHD4}$ performs much better than $\hat{\beta}_{PML}$. When the outlier is present, $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ always performs much better than $\hat{\beta}_{PML}$ in terms of bias and MSE. Thus we can conclude that $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are much more robust against outliers than $\hat{\beta}_{PML}$.

Study B: Influence of different values of an outlier

Now we allow the outlying value varies and investigate how an estimator changes when the outlying value changes. Under the model of our consideration, we can either let the event time vary or let the covariate vary. In this thesis we let the covariate vary, but one can easily let event time vary and do similar analysis as in this thesis which is omitted to save space and avoid redundance.

In order to let the covariate in the outlying observation vary, we first select the individual with smallest observed event time, and then let its covariate z take integer numbers from the range $[-1, 10]$ and change its observed event time to the fixed value 4, similar to what we did in Study A. The rest of data remains unchanged, and we would like to see how the $\hat{\beta}_{MHD}$ and $\hat{\beta}_{PML}$ change when the covariate z of the outlying observation varies from -1 to 10. In order to quantify the influence of an outlier on an estimator $\hat{\beta}$, we adopt the α-influence function (IF) (Beran, 1977; Lu et al., 2003) which is defined as the change in the estimate before and after the outlier is added, divided by the contamination rate $1/n$, i.e.

$$IF(z) = \frac{\hat{\beta}^z - \hat{\beta}}{1/n}, \tag{3.3}$$

where $\hat{\beta}$ is the estimate based on the original uncontaminated data while $\hat{\beta}^z$ is the estimate based on the contaminated data with the covariate of the outlying observation taking value z. In Studies B and C, we always use sample size $n = 100$ and the results for other sample sizes are similar and thus omitted to save space. The α-IF defined in (3.3) are calculated and displayed in Figure 3.1 for both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ when either no censoring or censoring rate is 40%. The α-IFs for censoring rate 20% is very similar and between those with no censoring and censoring rate 40%, and thus they are omitted to make the figure clearer. In addition, since we observe from Table 3.6 that both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ behave very similarly when an outlier is added, we expect that the α-IF of $\hat{\beta}_{MHD4}$ will be similar to that of $\hat{\beta}_{MHD3}$ and thus is omitted to save space.

From Figure 3.1 we observe that $\hat{\beta}_{MHD3}$ always has much smaller α-IFs (absolute value) than $\hat{\beta}_{PML}$ for varying outlying covariate values and different censoring rate. More specifically, when the covariate value z falls within the normal range [0.1,2] (thus z is not quite an outlier), both methods yield fair estimates and their α-IFs are quite close to 0. However, when the outlying

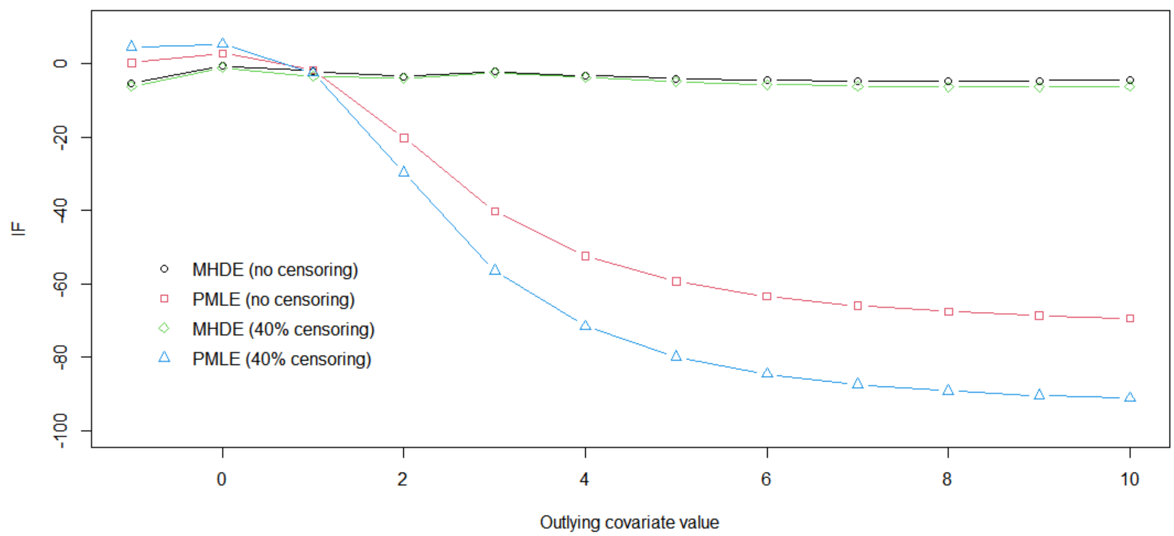

Figure 3.1: α-IF for single outlier

covariate value z increases from 2, $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ perform very differently: the α-IF of $\hat{\beta}_{MHD3}$ is very stable and always stay close to 0, while that of $\hat{\beta}_{PML}$ increases dramatically and reaches absolute values as high as over 90. This indicates that $\hat{\beta}_{MHD3}$ is very robust against outliers while $\hat{\beta}_{PML}$ is very sensitive to outliers. In terms of the impact of censoring rate, both estimators perform better when censoring rate is lower than higher. Surprisingly, the censoring rate has very little impact on the α-IF of $\hat{\beta}_{MHD3}$ while it has much higher impact on that of $\hat{\beta}_{PML}$.

The robustness of $\hat{\beta}_{MHD3}$ is partially due to the use of kernel estimation. When an outlier is far away from the rest of data, the mass of the estimated kernel function at or around the outlier is very close to 0 since the rest of data is too far to contribute much to the kernel estimation at or around the outlier. Comparatively, when the outlying covariate z moves towards 10, $\hat{\beta}_{PML}$ tries to fit this outlier so that the outlier drags the estimate away from the true value $\beta = 0.5$.

Study C: Influence of number of outliers

In Studies A and B we only consider a single outlier. Now we consider the influence of multiple outliers on the proposed MHDE method. In order to add multiple outliers, we order the simulated data in ascending event times. We first replace the individual with the smallest event time by an outlier, then replace the individual with the second smallest event time by another outlier, and this process continues in order to add desired number of outliers. For simplicity, all these added outliers are identical with covariate value $z = 4$ and observed event time $T = 4$. When k outliers are added, $k = 1, 2, \ldots, 30$, the α-IFs of both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ are calculated by (3.3) but with $1/n$ replaced by contamination rate k/n. The α-IFs of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ for no censoring and 40% censoring rate are displayed in Figure 3.2.

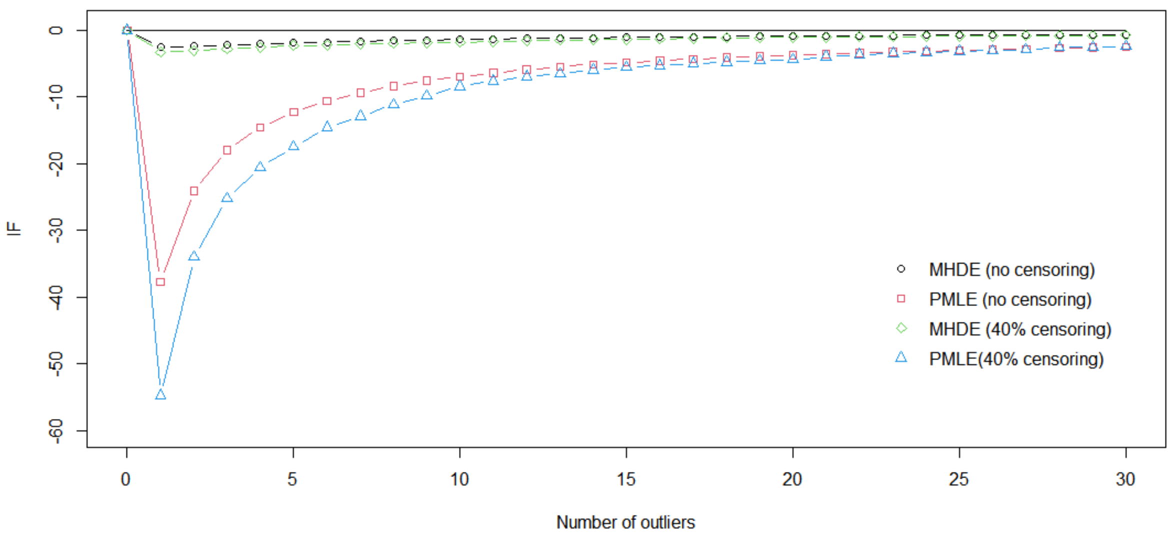

Figure 3.2: α-IF for multiple outliers

From Figure 3.2 we observe that $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ behave very differently. The α-IF of $\hat{\beta}_{MHD3}$ is very stable and always close to 0, and it seems converging to 0 as the number of outliers increases. Comparatively, the absolute value of α-IF of $\hat{\beta}_{PML}$ jumps to a very large value when one outlier is added and then decreases as the number of outliers increases. It seems that the α-IF of $\hat{\beta}_{PML}$ converges to a negative constant when the number of outliers approaches to 30. The α-IFs for no censoring and 40% censoring rate are very similar for each of the two estimators, though censoring rate has a much higher impact on the α-IF of $\hat{\beta}_{PML}$ than on that of $\hat{\beta}_{MHD3}$. These observations indicate that $\hat{\beta}_{MHD3}$ is much robust against multiple outliers than $\hat{\beta}_{PML}$.

Here we want to mention that the α-IF of $\hat{\beta}_{PML}$ may look different if different multiple outliers are added than the ones used for Figure 3.2. For example, one may use different and more deviated values as outliers so that the α-IF of $\hat{\beta}_{PML}$ may converge to a different non-zero value or even keep increasing in absolute value. With the identical outliers used for Figure 3.2, PMLE produces estimates close to 0.3 when more and more outliers are added, and as a result the α-IF is roughly $100(0.3-0.5)/k$ when k approaches 30.

3.4 Multiple Covariates

In previous Sections 3.2 and 3.3, we examined the performance of the proposed MHDE methods for single covariate, either categorical or quantitative. We will show in this section that the methods can be easily extended to multiple covariates. Specifically, we conduct numerical studies for cases of two categorical covariates, two quantitative covariates and one categorical and one quantitative covariates. For more than two covariates, the extension will be very similar.

Consider Cox model (1.1) with $z = (z_1, z_2)^\top$ and $\beta = (\beta_1, \beta_2)^\top$. For our numerical studies in this section, we always use sample sizes $n = 50, 100, 300$. We consider possibly different types of z_1 and z_2. For categorical covariate, data are generated with balanced sampling.

3.4.1 Two categorical covariates

Suppose z_1 and z_2 are both categorical covariates which take either value 0 or 1. Suppose the true parameter values are $(\beta_1, \beta_2) = (0.5, 0.5)$. With these categorical covariates, individuals can be categorized into four groups and then it is very similar to the case of single categorical covariate so that the MHDE is $\hat{\beta}_{MHD1}$ given in (2.13) when λ_0 is known and $\hat{\beta}_{MHD2}$ given in (2.14) when λ_0 is unknown except that now the optimization in (2.13) and (2.14) is over two-dimensional parameter space instead of one-dimensional.

Since our results in Section 3.2 demonstrated that $\hat{\beta}_{MHD1}$ and $\hat{\beta}_{MHD2}$ behave similarly for different censoring rates, as examples we only report our numerical results of $\hat{\beta}_{MHD1}$ with known λ_0 and no censoring in Table 3.7 and $\hat{\beta}_{MHD2}$ with unknown λ_0 and 20% censoring rate in Table

3.8. The results for other censoring rates are similar and thus omitted to avoid repetition. For the purpose of comparison, the results of $\hat{\beta}_{PML}$ are also included in Tables 3.7 and 3.8. In Table 3.8, for individuals in each of the four groups with $z = (0,0), (0,1), (1,0)$ or $(1,1)$, the censoring times are randomly selected from an uniform distribution over interval I_z, where I_z's vary from group to group and they are chosen to control the overall censoring rate to be around 20% . As a result, we use $I_{(0,0)} = (0,12)$, $I_{(0,1)} = I_{(1,0)} = (0,7)$ and $I_{(1,1)} = (0,3.5)$.

Table 3.7: Bias (MSE) of $\hat{\beta}_{MHD1}$ and $\hat{\beta}_{PML}$ for two categorical covariates, known λ_0 and no censoring

	$\hat{\beta}_{MHD1}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	.003 (.019)	.019 (.019)	.016 (.027)	.026 (.041)
100	-.001 (.007)	-.001 (.008)	.013 (.011)	.007 (.019)
300	.006 (.003)	-.010 (.002)	.007 (.002)	.003 (.005)

Table 3.8: Bias (MSE) of $\hat{\beta}_{MHD2}$ and $\hat{\beta}_{PML}$ for two categorical covariates, unknown λ_0 and censoring rate 20%

	$\hat{\beta}_{MHD2}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	-.030 (.047)	-.028 (.047)	.015 (.035)	.029 (.029)
100	.032 (.030)	.010 (.027)	.004 (.019)	.002 (.014)
300	.016 (.011)	.007 (.010)	-.008 (.005)	.006 (.004)

From Table 3.7 we observe similar phenomenon as in Table 3.1 for single categorical covariate, i.e. $\hat{\beta}_{MHD1}$ generally outperforms $\hat{\beta}_{PML}$ in terms of bias and MSE. More specifically, $\hat{\beta}_{MHD1}$ always has better performance than $\hat{\beta}_{PML}$ when sample size is relatively small ($n = 50, 100$) while it is very competitive when sample size is relatively large ($n = 300$). From Table 3.8 we observe similar phenomenon as in Table 3.2 for single categorical covariate, i.e. $\hat{\beta}_{PML}$ generally performs better than $\hat{\beta}_{MHD2}$ in terms of bias and MSE.

3.4.2 One categorical and one quantitative

Suppose z_1 is categorical covariate which takes either value 0 or 1, while z_2 is quantitative which follows uniform distribution $U(0.1,2)$. Suppose the true parameter values are $(\beta_1,\beta_2)=(0.5,0.5)$. With a quantitative covariate involved, individuals are not able to be split into groups from a unique population (i.e. with exactly same (z_1,z_2) values). As a result, the case of mixed types of covariates (e.g. one categorical and one quantitative) is very similar to the case of single quantitative covariate, and the MHDEs are $\hat{\beta}_{MHD3}$ given in (2.16) when λ_0 is known and $\hat{\beta}_{MHD4}$ given in (2.17) when λ_0 is unknown except that now the optimization in (2.16) and (2.17) is over two-dimensional parameter space instead of one-dimensional.

Since our results in Section 3.3 demonstrated that $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ behave similarly for different censoring rates, as examples we only report our numerical results of $\hat{\beta}_{MHD3}$ with known λ_0 and no censoring in Table 3.9 and $\hat{\beta}_{MHD4}$ with unknown λ_0 and 20% censoring in Table 3.10. The results for other censoring rates are similar and thus omitted to avoid repetition. For Table 3.10, censoring times are randomly selected from uniform distributions over interval $(0,10)$ for individuals with $z_1=0$ and $(0,5)$ for individuals with $z_1=1$ so that the overall censoring rate is around 20%.

Table 3.9: Bias (MSE) of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ for mixed types of covariates, known λ_0 and no censoring

	$\hat{\beta}_{MHD3}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	.030 (.026)	-.067 (.007)	.035 (.053)	.002 (.035)
100	.005 (.016)	-.048 (.004)	-.015 (.022)	-.010 (.015)
300	.013 (.004)	-.038 (.001)	.003 (.007)	.002 (.007)

From Tables 3.9 and 3.10 we observe similar phenomenon as in Table 3.5 for single quantitative covariate. More specifically, both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ always have much smaller MSE than $\hat{\beta}_{PML}$ while their biases are competitive with $\hat{\beta}_{PML}$.

With a quantitative covariate, outlying observations are likely to occur due to measurement

Table 3.10: Bias (MSE) of $\hat{\beta}_{MHD4}$ and $\hat{\beta}_{PML}$ for mixed types of covariates, unknown λ_0 and censoring rate 20%

	$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	-.050 (.037)	-.070 (.015)	.107 (.125)	-.037 (.101)
100	-.051 (.036)	.023 (.010)	-.016 (.067)	-.006 (.049)
300	-.036 (.014)	.028 (.004)	-.053 (.017)	.020 (.015)

error or poorly-designed experiment. Thus we look at the impact of outliers on both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ and compare their robustness. Similar to Study A in Subsection 3.3.2, we select the individual with smallest uncensored event time and replace it by an abnormal individual from another population. To be more specific, we replace its quantitative covariate z_2 with value -1 while its normal range is $(0.1, 2)$, and also replace its observed event time with outlying value 10 while its normal range is $(0, 6)$. The results before and after the outlier being added are presented in Table 3.11 for $\hat{\beta}_{MHD3}$ with known λ_0 and no censoring and in Table 3.12 for $\hat{\beta}_{MHD4}$ with unknown λ_0 and censoring rate 20%.

Table 3.11: Bias (MSE) of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ for mixed types of covariates, known λ_0 and no censoring, before and after single outlier added

		$\hat{\beta}_{MHD3}$		$\hat{\beta}_{PML}$	
	n	β_1	β_2	β_1	β_2
Before	50	.040 (.021)	-.058 (.007)	.042 (.049)	.017 (.038)
	100	.041 (.012)	-.008 (.012)	-.056 (.005)	.026 (.025)
	300	.018 (.004)	-.037 (.002)	.017 (.005)	.005 (.005)
After	50	.010 (.020)	-.021 (.007)	.068 (.049)	.109 (.029)
	100	.011 (.025)	.003 (.005)	-.018 (.013)	.036 (.011)
	300	.003 (.005)	-.018 (.002)	.023 (.005)	.028 (.005)

From Tables 3.11 and 3.12 we observe that both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are much less influenced by outliers than $\hat{\beta}_{PML}$. Specifically, when sample size is relatively small ($n = 50$), the bias of the PMLE of β_2 grows dramatically whereas the bias of the PMLE of β_1 also increases but in much smaller scale. When sample size increases, the influence of the single outlier on $\hat{\beta}_{PML}$ is reduced,

Table 3.12: Bias (MSE) of $\hat{\beta}_{MHD4}$ and $\hat{\beta}_{PML}$ for mixed types of covariates, unknown λ_0 and censoring rate 20%, before and after single outlier added

		$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
	n	β_1	β_2	β_1	β_2
Before	50	-.049 (.054)	-.011 (.017)	.052 (.160)	-.004 (.121)
	100	-.049 (.029)	-.010 (.010)	-.020 (.063)	.019(.046)
	300	-.045 (.013)	.009 (.004)	-.044 (.019)	.025 (.016)
After	50	-.055 (.037)	-.025 (.012)	.091 (.137)	.205 (.114)
	100	-.032 (.022)	-.023 (.009)	-.043 (.062)	.135 (.051)
	300	-.022 (.010)	-.012 (.004)	-.052 (.020)	.073 (.019)

which is expected due to the decreased contamination rate $(1/n)$ when sample size increases. It is not surprising that the PMLE of β_2 is influenced more than the PMLE of β_1 by the added outlier, since the outlier is added on the quantitative covariate z_2. Comparatively, both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ have very little change in both their bias and MSc before and after the data is contaminated, and in some cases their bias and MSE are even improved after the outlier added. Thus, we can conclude that both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are much more robust against outliers than $\hat{\beta}_{PML}$.

3.4.3 Two quantitative covariates

Suppose both z_1 and z_2 are quantitative covariates both of which follow uniform distribution $U(0.1,2)$. Suppose the true parameter values are $(\beta_1,\beta_2) = (0.5,0.5)$. The MHDEs of $(\beta_1,\beta_2)^\top$ are $\hat{\beta}_{MHD3}$ given in (2.16) when λ_0 is known and $\hat{\beta}_{MHD4}$ given in (2.17) when λ_0 is unknown, except that now the optimization in (2.16) and (2.17) is over two-dimensional parameter space instead of one-dimensional. With the same argument, we only report our numerical results of $\hat{\beta}_{MHD3}$ with known λ_0 and no censoring in Table 3.13 and $\hat{\beta}_{MHD4}$ with unknown λ_0 and censoring rate 20% in Table 3.14. For Table 3.14, the censoring times are randomly selected from the uniform distribution over interval $(0,7)$ so that the overall censoring rate is around 20%.

From Tables 3.13 and 3.14 we observe similar phenomenon as in Table 3.5 for single quantitative covariate. More specifically, both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ always have much smaller MSE than

Table 3.13: Bias (MSE) of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ for two quantitative covariates, known λ_0 and no censoring

	$\hat{\beta}_{MHD3}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	-.010 (.024)	-.024 (.036)	.069 (.355)	-.034 (.304)
100	.004 (.019)	-.027 (.018)	.083 (.149)	.018 (.157)
300	-.011 (.006)	-.024 (.009)	.031 (.045)	-.013 (.037)

Table 3.14: Bias (MSE) of $\hat{\beta}_{MHD4}$ and $\hat{\beta}_{PML}$ for two quantitative covariates, unknown λ_0 and censoring rate 20%

	$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
n	β_1	β_2	β_1	β_2
50	-.015 (.012)	-.019 (.014)	.002 (.075)	-.001 (.093)
100	.002 (.009)	-.015 (.009)	.018 (.033)	.005 (.032)
300	-.014 (.006)	.025 (.005)	-.017 (.011)	-.012 (.017)

$\hat{\beta}_{PML}$ and very competitive bias with $\hat{\beta}_{PML}$.

To investigate the robustness of the MHDE when there are two quantitative covariates, as in previous sections we select the individual with smallest uncensored event time and replace its covariates with $(z_1, z_2) = (4, 4)$ and replace its observed event time with value 4. The results before and after the outlier being added are presented in Table 3.15 for $\hat{\beta}_{MHD3}$ with known λ_0 and no censoring and in Table 3.16 for $\hat{\beta}_{MHD4}$ with unknown λ_0 and censoring rate 20%. .

Table 3.15: Bias (MSE) of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{PML}$ for two quantitative covariates, known λ_0 and no censoring, before and after single outlier added

		$\hat{\beta}_{MHD3}$		$\hat{\beta}_{PML}$	
	n	β_1	β_2	β_1	β_2
Before	50	-.018 (.027)	-.016 (.021)	-.002 (.242)	-.027 (.405)
	100	-.030 (.015)	.001 (.018)	-.047 (.142)	.052 (.166)
	300	-.030 (.008)	-.020 (.009)	-.004 (.043)	-.008 (.043)
After	50	-.041 (.017)	-.046 (.013)	.277 (.173)	.239 (.272)
	100	-.069 (.011)	-.037 (.012)	.117 (.106)	.224 (.128)
	300	-.060 (.007)	-.051 (.008)	.071 (.038)	.065 (.038)

Table 3.16: Bias (MSE) of $\hat{\beta}_{MHD4}$ and $\hat{\beta}_{PML}$ for two quantitative covariates, unknown λ_0 and censoring rate 20%, before and after single outlier added

		$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
	n	β_1	β_2	β_1	β_2
Before	50	-.030(.018)	-.025(.019)	-.014(.097)	.055(.110)
	100	-.023(.010)	-.029(.011)	.006(.043)	.020(.046)
	300	-.015(.005)	-.025(.006)	.007(.012)	.027(.024)
After	50	-.075(.023)	-.066(.022)	-.604(.398)	-.529(.314)
	100	-.045(.012)	-.049(.013)	-.499(.271)	-.483(.252)
	300	-.022(.005)	-.031(.006)	-.367(.141)	-.356(.134)

From Tables 3.15 and 3.16 we observe similar phenomenon as in Table 3.6 for single quantitative covariate. More specifically, after the outlier is added, $\hat{\beta}_{PML}$ is significantly influenced in the sense that its bias increases dramatically and when censoring is present the MSE greatly increases as well. Comparatively, both $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are much less influenced by the outlier. We observe that the MSEs of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are always much smaller than that of $\hat{\beta}_{PML}$ no matter before or after adding the outlier. The biases of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are very competitive with $\hat{\beta}_{PML}$ before adding the outlier, whereas after adding the outlier, the biases of $\hat{\beta}_{MHD3}$ and $\hat{\beta}_{MHD4}$ are always much smaller than $\hat{\beta}_{PML}$.

In summary, we show in this section that the proposed MHDEs for single covariate can be easily extended to cases with multiple covariates of different type and the resulting estimations retain good efficiency and robustness properties as in single covariate case.

Chapter 4

REAL DATA ANALYSIS

In this chapter we use two real datasets to demonstrate the implementation of the proposed MHDE. Specifically we analyze a bone marrow transplant dataset and a male laryngeal cancer dataset in Sections 4.1 and 4.2 respectively and compare our MHDE estimates with the PMLE estimates. Both datasets are available on R package 'KMsurv'.

4.1 Bone Marrow Transplants for Hodgkin's and Non-Hodgkin's Lymphoma

The dataset *hodg* was collected on 43 bone marrow transplant patients at The Ohio State University Bone Marrow Transplant Unit (`https://www.rdocumentation.org/packages/KMsurv/versions/0.1-5/topics/hodg`). All the 43 patients had either Hodgkin's disease (HOD) or non-Hodgkin's lymphoma (NHL) and were given either an allogeneic (Allo) transplant from a Human Leukocyte Antigens (HLA) match sibling donor or an autogenic (Auto) transplant; i.e., their own marrow was cleansed and returned to them after a high dose of chemotherapy. These information are recorded by indicator *gtype* which takes 0 if graft type is allogenic and takes 1 if graft type is autologous, and indicator *dtype* which takes 0 if disease type is non-Hodgkin lymphoma and takes 1 if disease type is Hodgkins disease. Also included in this dataset is time to death or relapse (in days) and the event indicator which takes 1 if dead or relapse and takes 0 otherwise. Note that the time to death is count in days with the longest time being 2144 days and the shortest being only 2 days. We rescale the time in months by dividing 30.

We model this data using Cox PH model with the two categorical covariates *gtype* and *dtype*, and thus the hazard function is given by

$$\lambda(t|gtype, dtype) = \lambda_0(t)\exp(\beta_1 \cdot gtype + \beta_2 \cdot dtype).$$

According to different combinations of the *gtype* and *dtype* values, the 43 patients can be divided into four groups. However the sample size is so small in each group that it is not possible to obtain any sound and reliable estimates of the conditional hazard functions. Therefore we use $\hat{\beta}_{MHD4}$ with kernel estimators of conditional hazard function and baseline hazard function to estimate the coefficients β_1 and β_2. The calculated coefficient estimates are $\hat{\beta}_{MHD4,1} = -0.060$ and $\hat{\beta}_{MHD4,2} = 0.128$. The PMLEs of β_1 and β_2 are $\hat{\beta}_{PML,1} = -0.260$ and $\hat{\beta}_{PML,2} = 0.274$.

4.2 Death Times of Male Laryngeal Cancer Patients

The dataset *larynx* was collected on 90 males diagnosed with cancer of the larynx during the period 1970-1978 at a Dutch hospital (`https://www.rdocumentation.org/packages/KMsurv/versions/0.1-5/topics/larynx`) This dataset consists of time (in years). between first treatment and either death or the end of the study (January 1, 1983), indicator of death, patient's age (*age*) at the time of diagnosis, and the stage of patient's cancer (*stage*). According to the T.N.M. (primary tumor (T), nodal involvement (N) and distant metastasis (M) grading) classification used by the American Joint Committee for Cancer Staging (1972), the disease in the study were classified into four stages denoted as stage I to stage IV, ordered from least serious to most serious. Among these 90 patients, 33 patients were in Stage I, 17 in Stage II, 27 in Stage III and all other 13 in Stage IV. In this dataset, the stages were recorded and rescaled as categorical values 0, 1, 2 and 3. The recorded ages vary from 41 to 86, and in our analysis we rescale them by subtracting the minimum age 41 so that the new range is from 0 to 45. We are interested in the quantified relationship between the hazard rate of a patient at time t given his/her specific age and stage of disease.

For this data, we model the survival time using Cox PH model with one categorical covariate *stage* and one quantitative covariate *age*, and thus the hazard function is given by

$$\lambda(t|stage, age) = \lambda_0(t)\exp(\beta_1 \cdot stage + \beta_2 \cdot age).$$

With unknown λ_0, we use $\hat{\beta}_{MHD4}$ with kernel estimators of conditional hazard function and baseline hazard function to estimate the coefficients β_1 and β_2. The calculated coefficient estimates are $\hat{\beta}_{MHD4,1} = 0.817$ and $\hat{\beta}_{MHD4,2} = -0.045$. On the other hand, $\hat{\beta}_{PML}$ gives estimates as $\hat{\beta}_{PML,1} = 0.501$ and $\hat{\beta}_{PML,2} = 0.023$.

We further use this dataset to demonstrate the robustness of the MHDE against outliers. In this data, the patients' ages vary from 0 to 45 (after shifting) and their observed times are within the range $[0.1, 10.7]$. We also notice that the No. 83 patient has a reference age 0 and a relatively short survival time 1.0. We deliberately replace the age of this patient by 24, which is the median age of patients in this dataset. We also replace the survival time of this patient by 15, which is larger than the sample maximum value 10.7. Consequently No. 83 patient can be regarded as an outlying observation. Now we recalculate all the estimates based on this modified data and the results are presented in Table 1.4.

Table 4.1: Estimates of $\hat{\beta}_{MHD4}$ and $\hat{\beta}$ before and after the larynx data is contaminated

	$\hat{\beta}_{MHD4}$		$\hat{\beta}_{PML}$	
	β_1	β_2	β_1	β_2
original data	.817	-.045	.501	.023
contaminated data	.830	-.045	.398	.028
difference	-.013	.000	.103	-.005

From Table 4.1 we observe that $\hat{\beta}_{PML}$ is affected much more than $\hat{\beta}_{MHD4}$. More specifically, the MHDE of β_2 has almost no change and that of β_1 has very little change before and after the data is contaminated. Comparatively, the PMLE of β_1 decreases about 20% before and after the data is contaminated, though the PMLE of β_2 doesn't change much. This testifies again that the proposed MHDE has much better robustness against outliers than PMLE.

Chapter 5

CONCLUDING REMARKS

In this thesis, we proposed MHDEs for estimating the coefficients in Cox PH model. We not only considered the case of a single categorical covariate and the case a single quantitative covariate, but also the case of mixed types of covariates. For all these different scenarios, we constructed different versions of MHDE. In order to examine the finite-sample performance of the proposed estimators, we conducted extensive simulation studies and two real data analysis.

Our numerical results demonstrated that when the model assumption is valid without outliers, the proposed MHDEs are very competitive with the PMLE (Cox, 1975) in terms of bias and MSE, regardless of covariate type. In order to examine how the MHDEs react to outliers, we considered not only adding a single outlier but also contaminating data with multiple outliers. In all these situations considered, the MHDEs are much more robust against outliers than the PMLE in the sense that the MHDEs don't change much before and after the data contamination.

When one is confident that the model assumption is correct and the data has no outliers, the PMLE might be a better choice over MHDE since the computing packages for PMLE in statistical softwares are well developed and ready available. However, in cases when outliers are present, which is very common in big data that are everywhere nowadays, the proposed MHDEs significantly outperform PMLE and thus are highly preferred.

This thesis focuses more on numerical studies of the proposed MHDEs. Thus theoretical examination of the MHDEs need to be done in our future work. One can study their consistency and asymptotic normality which will contribute to statistical inferences based on MHDE. In a different direction, we may construct MHDE for more general or more complicated models other than Cox model, e.g. accelerated failure time (AFT) model and cure rate model.

Appendix A

Kernel Regression

Assume i.i.d. sample pairs (x_i, y_i), $i = 1, \ldots, n$ are from the regression model

$$y_i = r(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{A.1}$$

where ε follows some unknown distribution. Define

$$\hat{r}(x) = \sum_{i=1}^{n} w(x, x_i) \cdot y_i, \tag{A.2}$$

$$w(x, x_i) = \frac{K(\frac{x_i - x}{h})}{\sum_{j=1}^{n} K(\frac{x_j - x}{h})}, \tag{A.3}$$

where K is some kernel function satisfying $\int K(x)dx = 1$ and $K(x) = K(-x)$. Then the kernel regression estimator of r is given by

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} K(\frac{x_i - x}{h}) \cdot y_i}{\sum_{j=1}^{n} K(\frac{x_j - x}{h})}. \tag{A.4}$$

Bibliography

A. Basu, W. G. Manning, and J. Mullahy. Comparing alternative models: log vs cox proportional hazard? *Health economics*, 13(8):749–765, 2004.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

R. Beran. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

N. E. Breslow. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34: 216–217, 1972.

N. E. Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

D. Clayton and J. Cuzick. The em algorithm for cox's regression model using glim. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):148–156, 1985.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D. L. Donoho and R. C. Liu. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

K. R. Hess, D. M. Serachitopol, and B. W. Brown. Hazard function estimators: a simulation study. *Statistics in medicine*, 18(22):3075–3088, 1999.

J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

D. Karlis and E. Xekalaki. Minimum hellinger distance estimation for poisson mixtures. *Computational statistics & data analysis*, 29(1):81–103, 1998.

R. Karunamuni and J. Wu. Minimum hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 139(3):1118–1133, 2009.

O. Linton and J. P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100, 1995.

Z. Lu, Y. V. Hui, and A. H. Lee. Minimum hellinger distance estimation for finite mixtures of poisson regression models and its applications. *Biometrics*, 59(4):1016–1026, 2003.

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466, 1983.

B. P. Rao. *Nonparametric functional estimation*. Academic press, 2014.

I. Selingerová, I. Horová, and J. Zelinka. Kernel estimation of conditional hazard function for cancer data. 2014.

D. G. Simpson. Minimum hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association*, 82(399):802–807, 1987.

L. Spierdijk. Nonparametric conditional hazard rate estimation: a local linear approach. *Computational statistics & data analysis*, 52(5):2419–2434, 2008.

T. Sriram and A. Vidyashankar. Minimum hellinger distance estimation for supercritical galton–watson processes. *Statistics & probability letters*, 50(4):331–342, 2000.

R. N. Tamura and D. D. Boos. Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

W. H. Wong. Theory of partial likelihood. *The Annals of statistics*, 14(1):88–123, 1986.

J. Wu and R. J. Karunamuni. Efficient hellinger distance estimates for semiparametric models. *Journal of Multivariate Analysis*, 107:1–23, 2012.

J. Wu and R. J. Karunamuni. Profile hellinger distance estimation. *Statistics*, 49(4):711–740, 2015.

J. Wu, R. Karunamuni, and B. Zhang. Minimum hellinger distance estimation in a two-sample semiparametric model. *Journal of Multivariate Analysis*, 101(5):1102–1122, 2010.

J. Wu, W. Yao, and S. Xiang. Computation of an efficient and robust estimator in a semiparametric mixture model. *Journal of Statistical Computation and Simulation*, 87(11):2128–2137, 2017.

S. Xiang, W. Yao, and J. Wu. Minimum profile hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics*, 42(2):246–267, 2014.

S. Yang. Minimum hellinger distance estimation of parameter in the random censorship model. *The Annals of Statistics*, 19(2):579–602, 1991.

Z. Ying. Minimum hellinger-type distance estimation for censored data. *The Annals of Statistics*, pages 1361–1390, 1992.

Y. Zhu, J. Wu, and X. Lu. Minimum hellinger distance estimation for a two-sample semiparametric cure rate model with censored survival data. *Computational Statistics*, 28(6):2495–2518, 2013.