5-2-2013

Statistical Methods and Computing for Semiparametric Accelerated Failure Time Model with Induced Smoothing

Sy Han Chiou
Department of Statistics, University of Connecticut, steven.chiou@uconn.edu

Statistical Methods and Computing for Semiparametric Accelerated Failure Time Model with Induced Smoothing

Sy Han (Steven) Chiou, PhD University of Connecticut, 2013

In survival analysis, semiparametric accelerated failure time (AFT) models directly relate the predicted failure times to covariates and are a useful alternative to relative risk models. Recent developments in rank-based estimation and least squares estimation provide promising tools to make the AFT models more attractive in practice. In this dissertation, we propose fast and accurate inferences for AFT models with applications under various sampling schemes.

The challenge in computing the rank-based estimator comes from solving nonsmooth estimating equations. This difficulty can be overcome with an induced smoothing approach. We generalize the induced smoothing approach to incorporate weights with missing data arising from case-cohort study and stratified sampling design. Parameters are estimated with smoothed estimating equations. Variance estimators are obtained through efficient resampling methods that avoid full blown bootstrap. The estimator from the smooth weighted estimating equations are shown to be consistent and have the same asymptotic distribution as that from the nonsmooth version. An univariate failure time data from a tumor study and a clustered data from a dental study are analyzed.

Sy Han (Steven) Chiou, PhD University of Connecticut, 2013

The induced smoothing approach for rank-based AFT models is natural with Gehan's weight. Using the estimator from induced smoothing with Gehan's weight as an initial value, we propose an iterative procedure that works for any weight of general form. The resulting estimator has the same asymptotic properties as the nonsmooth rank-based estimator with the same weight. Real data from an adolescent stress duration study and a case-cohort study for Wilm's tumor illustrate the methods.

As for the least square estimation, we propose a generalized estimating equations (GEE) approach. The consistency of the regression coefficient estimator is robust to misspecification of working covariance, and the efficiency is higher when the working covariance structure is closer to the truth. The marginal error distributions and regression coefficients are allowed be unique for each margin or partially shared across margins as needed. The resulting estimator is consistent and asymptotically normal, with variance estimated through a multiplier resampling method. Bivariate failure times data from a diabetic retinopathy study is analyzed.

All the aforementioned methods for AFT models are implemented in an R package **aftgee** (http://cran.r-project.org/web/packages/aftgee/index.html).

Statistical Methods and Computing for Semiparametric Accelerated Failure Time Model with Induced Smoothing

Sy Han (Steven) Chiou

B.S., Statistics, University of Connecticut, Storrs, CT, May 2008

B.S., Applied Mathematics, University of Connecticut, Storrs, CT, May 2008

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2013

APPROVAL PAGE

Doctor of Philosophy Dissertation

Statistical Methods and Computing for Semiparametric Accelerated Failure Time Model with Induced Smoothing

Presented by

Sy Han (Steven) Chiou, B.S. Statistics, B.S. Applied Mathematics.

Major Advisor _____

Dr. Jun Yan

Major Advisor _____

Dr. Sangwook Kang

Associate Advisor _____

Dr. Dipak Dey

University of Connecticut

2013

Acknowledgments

First and foremost, I would like to express my special appreciation and thanks to my major advisors, Dr. Jun Yan and Dr. Sangwook Kang. Working with them has been a real pleasure for me. Throughout my dissertation research, they have guided and supported me with promptness and care, and have always been patient and encouraging in times of new ideas and difficulties. Their approaches in research problems, scientific standards and hard work set an example for me. I appreciate all their contributions of time, ideas, and funding that made my Ph.D. experience productive and stimulating.

I would like to extend my thanks to my committee member, Dr. Dipak Dey, for constructive comments and valuable suggestions. Moreover, I will forever be thankful to my former advisors, Dr. Nalini Ravishanker and Dr. Rick Vitale. They have been helpful in providing advice on numerous occasions during my undergraduate and graduate career. They were and are my best role models as a mentor and teacher. My thanks also go to Dr. Rob Aseltine and Dr. Beth Schillings from the Institute for Public Health Research, University of Connecticut Health Center. This group has been a source of friendships as well as collaboration opportunities. I would also like to acknowledge the UConn's Quantitative Learning Center where I started my first formal teaching training.

My time at UConn was enjoyable in large part due to the many friends that became parts of my life. I am also indebted to my students who have been an invaluable support.

I truly enjoyed the time spent with students and friends. In addition, I owe special thanks to Andrew Duxbury, Robert Mancini, Waseem Mehar and Valerie Tellez for proof-reading the thesis and thereby significantly improving my English.

On a personal note, I would like to thank Dr. Xiaodi Wang, who planted a seed of thought in my head that Ph.D. was a dream which was worthwhile to pursue. Finally, I would like to thank my family for their infinite support throughout everything. Words can not express how grateful I am for all the sacrifices that they have made on my behalf. This thesis would not have happened without their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis is a statistical method for data analysis in which the outcome variable of interest is the time to the occurrence of an event subject to censoring. The semiparametric accelerated failure time (AFT) model is not as widely used as the proportional hazards model (Cox, 1972) which assumes the underlying hazard rate is a function of the independent covariates. However, the semiparametric AFT model provides an attractive alternative to the proportional hazards model because it directly relates the effect of explanatory variables on the survival time instead of the hazard as in the Cox model. This characteristic allows an easier interpretation of the results. Nevertheless, the semiparametric AFT model has not been as widely used as it should be due to lack of efficient and reliable computing algorithm to obtain both parameter estimates and their standard errors. This dissertation aims to develop a computationally more efficient approach for the semiparametric AFT model in both univariate and multivariate cases with various sampling schemes that arise quite frequently in real world problems.

There are two classes of estimators. The first one is the rank-based estimator motivated by inverting the weighted log-rank test (Prentice, 1978). Its asymptotic properties

have been rigorously studied by Tsiatis (1990) and Ying (1993). Due to a lack of an efficient and reliable computing algorithms, the rank-based estimator had not been widely used in practice until recently, with numerical strategies for drawing inference developed by Huang (2002) and Strawderman (2005). In addition to the theoretical advances, some efforts have been made to provide ways to solve for rank-based estimators, particularly with Gehan (Gehan, 1965) weight. For example, Fygenson and Ritov (1994) show that Gehan rank-based estimating equation is a gradient of an convex objective function. Taking advantage of this, Jin et al. (2003) obtained the Gehan estimator by minimizing the convex objective function through a standard linear programming technique, thereby saving the computation time. This method is eventually extended to multivariate failure time by Jin et al. (2006a). However, such linear programming is computationally demanding, especially with larger sample sizes.

The rank-based estimator is difficult to compute due to the fact that rank-based estimating equations are not smooth. The induced smoothing procedure of Brown and Wang (2005) is a more efficient approach in computation. This method relies on taking expectations of the estimating equation with respect to added to the model parameter. The resulting estimating equation is continuously differentiable with respect to the parameter and can be solved with standard numerical algorithms. The asymptotic properties of this induced smoothing procedure in the context of rank-based approach with Gehan's weight have been established. For example, Brown and Wang (2007) applied induced smoothing to Gehan rank-based estimating equation for univariate failure

times. Johnson and Strawderman (2009) and Wang and Fu (2011) applied the induced smoothing approach for clustered failure times.

We adopted the induced smoothing procedure to a setting with an extremely high censoring rate. In practice, high censoring rates are often caused by rare diseases or when the main risk factor is expensive to measure. The case-cohort design generally appears in these situations. In the case-cohort design, covariate information is only collected from cases and a representative sample of censored observations rather than collected from the entire cohort. At the first glance, we assume that the sub-cohort is sampled by simple random sampling without replacement from the full cohort. We thereby constructed a weight adjusted induced smoothing rank-based estimating equation. A class of efficient variance estimators and its asymptotic properties are presented in Chapter 2. This problem sets the tone for the next proposed method in terms of performing inference under the more complicated case-cohort setup.

Motivated by a retrospective cohort dental study (Caplan et al., 2005), we considered a generalization of the case-cohort design to the stratified case-cohort design which is a special form of stratified random sampling. In the retrospective cohort dental study, times to extraction of teeth were compared within each patient. We extended our model to multivariate failure time data. In Chapter 3, we show that the estimator from the induced smoothing weighted estimating equations are consistent and have the same asymptotic distribution as that from the nonsmooth version. As in the case-cohort design, the variance of the estimator is estimated by computationally efficient sandwich

estimators aided by a multiplier bootstrap.

The aforementioned induced smoothing rank-based approach is natural with Gehan weight. When other weights are used, the induced smoothing approach does not in general provide smoothing estimating equations that are easy to evaluate. An asymptotic equivalent smoothed estimating equation is proposed for point estimation with general weight, then an induced smoothing iterative procedure is presented in Chapter 4. The resulting estimator has the same asymptotic properties as the nonsmooth rank-based estimator with general weights.

The second class in solving semiparametric AFT model is the Buckley–James (BJ) estimator which extends the least squares principle to accommodate censoring through an expectation–maximization (EM) algorithm which iterates between imputing the censored failure times and least squares estimation (Buckley and James, 1979). Despite the nice asymptotic properties (Lai and Ying, 1991; Ritov, 1990), the BJ estimator may be hard to compute as the EM algorithm may not converge. Jin et al. (2006b) suggested an iterative least-squares procedure that starts from a consistent and asymptotically normal initial estimator, such as the one obtained from the rank-based method of Jin et al. (2003). Jin et al. (2006b) also considered their least squares method with marginal models for multivariate failure times. However, these approaches used an independent working model and left the within-cluster dependence structure unspecified. In Chapter 5, we propose an iterative GEE procedure for marginal semiparametric multivariate AFT models that generalizes the recent development of least squares approach by Jin

et al. (2006b). This method has the same spirit as GEE for complete data in that misspecification of the working covariance matrix does not affect the consistency of the parameter estimator in the marginal AFT models. When the working covariance is closer to the unknown truth, the estimator has a higher efficiency than that from working independence as used in Jin et al. (2006b). In addition, we also considered cases where all marginal distributions are identical and for cases where at least two margins are different. Our model also allow some covariates to share the same coefficients.

For each of these proposed methods, large scale of simulation studies are performed to test the adequacy of the proposed model. All the methods are implemented and made publicly available in an open source R package `aftgee` (Chiou et al., 2012a).

The rest of the dissertation is arranged as follows. Chapter 2 details the point estimation procedures based on the induced smoothing procedure with Gehan's weight for case-cohort data when the sub-cohort is a simple random sample from the a full cohort. Four variance estimation procedures, one based on full multiplier bootstrap and three based on a possibly multiplier bootstrap-aided sandwich variance estimator along with a large scale simulation and a tumor study are also included in Chapter 2. Chapter 3 extends the procedures proposed in Chapter 2 to cluster data and generalizes the case-cohort design to stratified sampling as in a retrospective dental study. Asymptotic properties are proved and verified with large scale simulation studies in Chapter 3. Chapter 4 extends the induced smoothing estimating equations to adopt general weight functions. Least squares approach with an iterative GEE procedure is proposed in

Chapter 5. A collection of these proposed methods are implemented in package `aftgee` as presented in Chapter 6. A discussion on future research concludes in Chapter 7.

Chapter 2

Efficient Rank-Based Approach from Case-Cohort Data

2.1 Introduction

A case-cohort design (Prentice, 1986) is an effective and economical design which reduces the effort and cost of a full-scale cohort study. Such design originated to allow efficient analysis of studies where it is too expensive and time consuming to collect and analyze data on all subjects. Cases and controls refer to subjects who have and have not, respectively, developed the disease of interest by the end of the study period. A case-cohort design is typically composed of two steps. First, a subset called sub-cohort is randomly selected from the whole cohort regardless of their disease status. Second, the remaining cases in the cohort are added to the sub-cohort. Cases and controls refer to subjects who have and have not, respectively, developed the disease of interest by the end of the study period. Measurement on the main risk factors are taken only on subjects in the sub-cohort and the remaining cases outside of the sub-cohort. This leads

to substantial reduction in the effort and cost of conducting large scale cohort studies, especially when the disease of interest is rare or the main risk factors are expensive to measure.

For failure time data from case-cohort studies, most statistical methods have focused on semiparametric models that work on either the hazard function (Barlow, 1994; Kang and Cai, 2009a; Kulich and Lin, 2000; Lin and Ying, 1993; Prentice, 1986; Self and Prentice, 1988; Sun et al., 2004; Therneau and Li, 1999), or the survival function (Chen, 2001a,b; Kong et al., 2004; Lu and Tsiatis, 2006). Parametric AFT models were considered by Kalbfleisch and Lawless (1988). Inferences about semiparametric AFT models for case-cohort data are much less developed, with only a few recent works (Kong and Cai, 2009; Nan et al., 2006; Yu, 2011; Yu et al., 2007).

Inferences for semiparametric AFT models have been difficult for not only case-cohort data but also for complete data. The most important estimator is the rank-based estimator motivated from inverting the weighted log-rank test (Prentice, 1978), with asymptotic properties rigorously studied (Tsiatis, 1990; Ying, 1993). Nevertheless, the estimator has not been as widely used as it should be due to lack of efficient and reliable computing algorithm to obtain both parameter estimates and their standard errors.

The parameter estimates are hard to compute because the most widely used rank-based estimating equations are not smooth. Recent works shed light on bringing AFT models into routine data analysis practice, including case-cohort studies. Jin et al. (2003) exploited that the rank-based estimating equation with Gehan's weight is the gradient

of an objective function and obtained estimates by solving it with linear programming. This approach was adapted to case-cohort data by Kong and Cai (2009). Nevertheless, the optimization with linear programming is still computationally very demanding, especially for larger sample sizes. A more computing efficient approach for rank-based inference is the induced smoothing procedure of Brown and Wang (2007). This approach is an application of the general induced smoothing method of Brown and Wang (2005), where the discontinuous estimating equations are replaced with a smoothed version, whose solutions are asymptotically equivalent to those of the former. The smoothed estimating equations are differentiable, thus facilitates rapid numerical solution.

Direct estimation of the variance is difficult because it involves nonparametric estimation of the unspecified error distribution. Most existing methods rely on bootstrap which is very computing intensive. Jin et al. (2003) estimated the variance through a multiplier resampling method, which requires a large bootstrapping sample in order to obtain a reliable variance estimate. For case-cohort data, Kong and Cai (2009) adopted a specially designed bootstrap procedure (Wacholder et al., 1989). The demanding computing task in linear programming is amplified because it requires solving estimating equations for each bootstrap sample. Huang (2002) proposed an easy-to-compute variance estimator based on the asymptotic linearity property of the estimating equations. A decomposition matrix of the variance matrix is estimated by solving estimating equations, but the number of the estimating equations to solve is much smaller; it is just the dimension of the parameters. For general nonsmooth estimating functions, Zeng

and Lin (2008) proposed a resampling strategy that does not require solving estimating equations or minimizing objective functions. Instead, it only involves evaluations of estimating functions and simple linear regression in estimating the slope matrix. The resulting variance estimators are computationally more efficient and stable than those from existing resampling methods.

In this article, we propose a fast rank-based inference procedure for semiparametric AFT models in the context of case-cohort studies. The parameters are estimated with an induced smoothing approach. Variance estimators are obtained through an efficient resampling methods for nonsmooth estimating functions that avoids full blown bootstrap. Of course, the methods also apply to full cohort data.

The rest of this article is organized as follows. Point estimation procedures based on smoothed estimating equations for case-cohort data when the sub-cohort is a simple random sample from the full cohort are proposed in Section 2.2. Four variance estimation procedures, one based on full multiplier bootstrap and three based on possibly multiplier bootstrap-aided sandwich variance estimator, are proposed in Section 2.3. A large scale simulation study is reported in Section 2.4, comparing the performances of the variance estimator and their timings. The methods are applied to a tumor study with both case-cohort data and full cohort data in Section 2.5. A discussion concludes in Section 2.6.

2.2 Point Estimation

Let $\{T_i, C_i, X_i\}$, $i = 1, \ldots, n$, be n independent copies of $\{T, C, X\}$, where T_i and C_i are log-transformed failure time and log-transformed censoring time, X_i is a $p \times 1$ covariate vector, and given X, C and T are assumed to be independent. A semiparametric AFT model has the form

$$T_i = X_i^\top \beta + \epsilon_i, \qquad i = 1, \ldots, n,$$

where β is an unknown $p \times 1$ vector of regression parameters, ϵ_i's are independent and identically distributed random variables with an unspecified distribution. It is also assumed that ϵ_i's are independent of X_i.

In a full cohort study, due to censoring, the observed data are (Y_i, Δ_i, X_i), $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$, $\Delta_i = I[T_i < C_i]$, and $I[\cdot]$ is the indicator function. A rank based estimating equation with Gehan's weight is

$$U_n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (X_i - X_j) I[e_j(\beta) \geq e_i(\beta)] = 0, \tag{2.1}$$

where $e_i(\beta) = Y_i - X_i^\top \beta$. The root of (2.1) is consistent to the true parameter β_0, and is asymptotically normal (Tsiatis, 1990). Despite these nice properties, even for the most promising method to date that solves it via linear programming (Jin et al., 2003), the computing burden increases drastically when bootstrapping is used to estimate the variance of the estimator.

For a case-cohort study, the covariate vector X_i's are not completely available for each individual. Measurement of some covariates is taken only on the subjects in the sub-cohort and cases outside the sub-cohort, and thus estimating function (2.1) cannot be evaluated. Using the observed data naively in (2.1) would lead to misleading results because the case-cohort sample is biased — it includes all cases but only a fraction of controls. It is possible, however, to adjust the biases by incorporating a weight that depends on the selection scheme of case-cohort samples. Suppose we select a sub-cohort of size $\tilde{n}$ by simple random sampling without replacement from the whole cohort. Let ξ_i be the sub-cohort indicator; $\xi_i = 1$ if the ith observation is in the sub-cohort and $\xi_i = 0$ otherwise. Let $p = \lim_{n\to\infty} p_n$, where $p_n = \tilde{n}/n$ is the sub-cohort inclusion probability. Under these assumptions, the desired case-cohort weight is $h_i = \Delta_i + (1-\Delta_i)\xi_i/p_n$. The weight-adjusted estimating equation (2.1) becomes

$$U_n^c(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} h_j \Delta_i (X_i - X_j) I[e_j(\beta) \geq e_i(\beta)] = 0. \tag{2.2}$$

The solution to (2.2), $\hat{\beta}_n$, remains to be consistent and asymptotically normal (Kong and Cai, 2009).

For full cohort data, a computationally more efficient approach for rank-based inference with Gehan's weight is the induced smoothing procedure of Brown and Wang (2007). Such smoothing method leads to continuously differentiable estimating equations that can be solved with standard numerical methods. Let Z be a p-dimensional

standard normal random vector. The estimating function $U_n(\beta)$ in (2.1) is replaced with $E[U_n(\beta + n^{-1/2}Z)]$, where the expectation is taken with respect to Z. This lead to

$$\tilde{U}_{n,G}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i (X_i - X_j)\Phi\left[\frac{e_j(\beta) - e_i(\beta)}{r_{ij}^2}\right] = 0, \tag{2.3}$$

where $r_{ij}^2 = n^{-1}(X_i - X_j)^\top (X_i - X_j)$ and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The solution to (2.3) is consistent to β_0 and has the same asymptotic distribution as the solution to (2.1) (Johnson and Strawderman, 2009).

For case-cohort data, we propose a smoothed version of (2.2) by adapting the idea of Brown and Wang (2007). Specifically, we replace $U_n^c(\beta)$ with $E[U_n^c(\beta + n^{-1/2}Z)]$ to obtain the induced smooth version of (2.2),

$$\tilde{U}_n^c(\beta) = E[U_n^c(\beta + n^{-1/2}Z)] = \sum_{i=1}^{n}\sum_{j=1}^{n} h_j \Delta_i (X_i - X_j)\Phi\left[\frac{e_j(\beta) - e_i(\beta)}{r_{ij}^2}\right]. \tag{2.4}$$

The solution $\tilde{\beta}_n$ to (2.4) is a consistent estimator to β_0 and is asymptotically normal. Furthermore, the asymptotic distribution of $\tilde{\beta}_n$ is also the same as that of $\hat{\beta}_n$. These arguments can be justified similarly as those in Johnson and Strawderman (2009).

2.3 Variance Estimation

The asymptotic variance of $\tilde{\beta}_n$ is even harder to estimate for case-cohort data than for full cohort data because of the extra complexity caused by the data structure. The terms

in the summation in $\tilde{U}_{n,G}(\beta)$ are not independent since the sub-cohort is drawn from the full cohort without replacement. We propose four variance estimators; one is fully resampling based while the other three use resampling to a component of the sandwich variance estimator.

2.3.1 Multiplier Bootstrap

The multiplier bootstrap estimator of Jin et al. (2003) is adapted to case-cohort data by inserting proper case-cohort weights, h_i's, in the multiplier bootstrap estimating equations. Let η_i, $i = 1, \ldots, n$, be independent and identically distributed positive random variables with $\mathrm{E}(\eta_i) = \mathrm{Var}(\eta_i) = 1$. Define

$$\tilde{U}_n^{c*}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_i \eta_j h_j \Delta_i (X_i - X_j) \Phi \left[\frac{e_j(\beta) - e_i(\beta)}{r_{ij}^2} \right]. \tag{2.5}$$

For a realization of $(\eta_1, \ldots, \eta_n)$, the solution to (2.5) provides one draw of $\tilde{\beta}_n$ from its asymptotic distribution. By repeating this process a large number B times, the variance matrix of $\tilde{\beta}_n$ can be estimated directly by the sampling variance matrix of the bootstrap sample of $\tilde{\beta}_n$.

Since the asymptotic variance of $\hat{\beta}_n$ is the same as that of $\tilde{\beta}_n$, the covariance matrix of $\tilde{\beta}_n$ can also be estimated by (2.2) through multiplier bootstrap. This is, however, not recommended because it would need to solve a large number B nonsmooth estimating equations. As will be seen in our simulation study, even with the computationally more

efficient smoothing estimating equations, the multiplier bootstrap approach can still be very time consuming, especially for larger sample sizes or more covariates.

2.3.2 Sandwich Estimator

To improve the computational efficiency, we consider alternative variance estimation procedures based on the sandwich form that avoid solving estimating equations repetitively. The asymptotic variances of $\hat{\beta}_n$ and $\tilde{\beta}_n$ are the same, both having a sandwich form. Under some regularity conditions (Zeng and Lin, 2008), uniformly in a neighborhood of β_0, equation (2.2) can be expressed as

$$n^{-1/2}U_n^c(\beta) = n^{-1/2}\sum_{i=1}^{n} h_i S_i(\beta_0) + An^{1/2}(\beta - \beta_0) + o_p(1 + n^{1/2}\|\beta - \beta_0\|),$$

where $S_i(\beta_0)$ is a zero-mean random vector, and A is asymptotic slope matrix of $n^{-1/2}\tilde{U}_n^c(\beta_0)$. The analytical details of $S_i(\beta_0)$ for case-cohort data is presented in the Appendix A.1. The asymptotic variance matrix of $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ is $n\Sigma = nA^{-1}V(A^{-1})^\top$, where V is the variance of $n^{-1/2}\sum_{i=1}^{n} h_i S_i(\beta_0)$. Estimation of Σ involves estimating V and A by estimator V_n and A_n, respectively. The variance estimator then has the sandwich form $\hat{\Sigma}_n = A_n^{-1}V_n(A_n^{-1})^\top$.

Estimation of V

Matrix V can be estimated either through a closed-form estimator or through bootstrapping the estimating equations. For case-cohort data, due to the correlated feature of ξ_i's in h_i's, V is different from its full cohort counterpart. There are two sources of variations contributing to V: variation due to the sampling of a full cohort (V_1) and variation due to the sampling of a sub-cohort within the full cohort (V_2). In particular, we have

$$V = V_1 + \frac{1-p}{p} V_2 = E\left[S_i(\beta_0) S_i(\beta_0)^\top\right] + \frac{1-p}{p} \mathrm{Var}\left[(1-\Delta_i) S_i(\beta_0)\right],$$

where V_2 vanishes if full cohort data are available.

Closed-form With explicit expressions for $S_i(\beta)$'s in the Appendix, a closed-form estimator of V is

$$V_n = V_{1n} + \frac{1-p_n}{p_n} V_{2n}$$

where

$$V_{1n} = n^{-1} \sum_{i=1}^{n} h_i \hat{S}_i(\hat{\beta}_n) \hat{S}_i^\top(\hat{\beta}_n),$$

and

$$V_{2n} = n^{-1} \sum_{i=1}^{n} h_i (1-\Delta_i) \hat{S}_i(\hat{\beta}_n) \hat{S}_i^\top(\hat{\beta}_n) - \left\{ n^{-1} \sum_{i=1}^{n} h_i (1-\Delta_i) \hat{S}_i(\hat{\beta}_n) \right\} \left\{ n^{-1} \sum_{i=1}^{n} h_i (1-\Delta_i) \hat{S}_i(\hat{\beta}_n) \right\}^\top,$$

and $\hat{S}_i(\hat{\beta}_n)$ is obtained by replacing unknown quantities in $S_i(\beta)$ with their sample counterparts.

Multiplier Bootstrap When $\hat{S}_i(\hat{\beta}_n)$ have complicated expressions, it is more convenient and perhaps more accurate to estimate V via bootstrap (Zeng and Lin, 2008). Because U_n^c and $\tilde{U}_n^c$ have the same asymptotic distribution, we apply the multiplier bootstrap approach to $\tilde{U}_n^c$. Evaluation of (2.5) at $\hat{\beta}_n$ with each realization of $(\eta_1, \ldots, \eta_n)$ provides one bootstrap replicate of $\tilde{U}_n^{c*}(\hat{\beta}_n)$. With B replicates, we estimate V by the sample variance of the bootstrap sample of $\tilde{U}_n^{c*}(\hat{\beta}_n)$. The bootstrap here is much less demanding than the full multiplier bootstrap above, because it only involves evaluations of estimating equations instead of solving them to obtain each bootstrap replicate.

Estimation of A

With V estimated by V_n, we next propose three approaches to estimate the slope matrix A. Depending whether V_n is based on closed-form or multiplier bootstrap, we will have two versions of estimator of Σ for each approach of slope matrix estimation.

Induced Smoothing With $\tilde{U}_n^c$, the smoothed version of U_n^c, the slope matrix A can be estimated directly by

$$A_n = \frac{1}{n} \frac{\partial}{\partial \beta^\top} \tilde{U}_n^c(\hat{\beta}_n).$$

The close-form expression of A_n can be evaluated easily. The variance estimator then has the sandwich form $\hat{\Sigma}_n = A_n^{-1} V_n (A_n^{-1})^\top$.

Smoothed Huang's (2002) Approach Huang (2002) avoided the difficulty in estimating the slope matrix of nonsmooth estimating equations by exploiting the asymptotic linearity of the estimating equations. Nevertheless, this approach still requires solving p nonsmooth estimating equations, whose convergence may be a problem. We adapt Huang's approach by replacing the p nonsmooth estimating equations with their smoothed versions. Let $V_n = L_n^\top L_n$ be the Cholesky decomposition of V_n. Let q_{nj} be the solution to the following estimating equations for γ, $j = 1, \ldots, p$,

$$n^{-1} \tilde{U}_n^c(\gamma) = n^{-1/2} l_j,$$

where l_j is the jth column of L_n. The solutions can be obtained with from general purpose nonlinear equation solvers; in our implementation we used R packages nleqslv (Hasselman, 2012) and BB (Varadhan and Gilbert, 2009). Let Q_n be the matrix whose jth column is $q_{nj} - \hat{\beta}_n$. Then $Q_n^\top Q_n$ is an estimate of Σ.

With the adaptation to smooth estimating equations, this approach has an advantage compared to the induced smoothing approach in that the closed-form derivative matrix is not required, and, hence, can be applied to more general nonsmooth estimating equations.

Zeng and Lin's (2008) Approach Zeng and Lin (2008) proposed to estimate the slope matrix by regressing the perturbed estimating functions on the perturbations. Let Z_b, $b = 1, \ldots, B$, be B realizations of a p-dimensional standard normal random vector. For case-cohort data, let U_{nj}^c be the jth component of U_n^c. We estimate the jth row of A, $j = 1, \ldots, p$, by A_{nj}, the least squares estimate of the regression coefficients when regressing $n^{-1/2}U_{nj}(\hat{\beta}_n + n^{-1/2}Z_b)$ on Z_b, $i = 1, \ldots, n$. The variance estimator also has the sandwich form $\hat{\Sigma}_n = A_n^{-1}V_n(A_n^{-1})^\top$.

This approach differs from the induced smoothing approach in that the slope matrix A is estimated via a resampling procedure that involves p least squares regressions, instead of taking the derivatives of a smooth function. It can be viewed as an empirical version of the induced smoothing approach.

2.4 Simulation

We conducted an extensive simulation study to assess the performance of the our point and variance estimators. Failure time T was generated from AFT model

$$\log(T) = 2 + X_1 + X_2 + X_3 + \epsilon,$$

where X_1 was Bernoulli with rate 0.5, X_2 and X_3 were uncorrelated standard normal variables. Censoring time C was generated from unif$(0, \tau)$ where τ was tuned to achieve desired censoring rate C_p. The distribution of ϵ had three types: standard normal,

standard logistic, or standard Gumbel, abbreviated by N, L, and G, respectively. The censoring rate C_p had two levels, 90% and 97%, representing a mildly rare disease and a very rare disease, respectively. For the mildly rare disease, the full cohort size was set to be 1500 and the case-cohort size was set to $\bar{m} = 300$ on average. For the very rare disease, the full cohort sizes were set to be 1500 and 3000, each with case-cohort sizes averaged at $\bar{m} \in \{150, 300\}$. The sub-cohort sampling proportion p_n was set to yield the desired average case-cohort size given censoring rate and full cohort size. For each viable combination, we generated 1000 datasets.

Given a dataset, point estimates of regression coefficients were obtained from both nonsmooth and smoothed estimating equations. The estimator from the nonsmooth version was obtained using linear programming (Jin et al., 2003), denoted by LP. The estimator from the induced smoothing approach with estimating equations (2.4) was obtained using R package nleqslv (Hasselman, 2012), denoted by IS. The two estimators are expected to be asymptotically the same, but with the IS estimator obtained much faster. Eight variance estimates were computed for the point estimate. The first two were full multiplier bootstrap estimates, denoted by MB, one based on the LP approach and the other based on the IS approach. The rest six were sandwich estimates constructed by combinations of three approaches to estimate A and two approaches to estimate V. We use abbreviations IS, sH, and ZL to denote the induced smoothing, smoothed Huang's, and Zeng and Lin's approach for A, respectively. We use abbreviations CF and MB to denote the closed-form estimate approach and the multiplier bootstrap approach for V,

respectively.

Results for the mildly rare disease case with censoring percentage $C_p = 90\%$, full cohort size 1500, and average case-cohort size $\bar{m} = 300$ are summarized in Table 1. Both the LP and the IS estimators appear to be virtually unbiased. In fact, they agreed with each other closely on a 45 degree line (not shown). Consequently, their empirical standard errors agreed with each other, and their bootstrap based standard errors agreed with each other. The bootstrap standard errors and the empirical standard errors match closely, suggesting that the bootstrap variance estimators provide good estimation of the empirical variantion. The other six standard errors based on sandwich variance estimators agreed quite well with the empirical standard errors too. The associated 95% confidence intervals based on all eight standard errors had empirical coverage percentages reasonably close to the nominal level. These observations were invariant to the error distributions.

Table 2 summarizes the results for the very rare disease case with censoring rate 97% and full cohort size 3000. The results for full cohort size 1500 were similar and not reported. The two point estimates, their empirical standard errors, and their average bootstrap standard errors still agree with each other. The bootstrap standard errors for case-cohort size 150, however, are underestimating the true variation, and as a result, the 95% confidence intervals had coverage percentage smaller than the nominal level. Not surprisingly, the six sandwich variance estimators performed no better than the two multiplier bootstrap variance estimators. When the case-cohort size was increased to

Table 1: Summary of simulation results based on 1000 replications for full cohort size 1500 and censoring rate 90%. The bootstrapping size is 500 for each replication. PE is average of point estimates; ESE is the empirical standard deviation of the parameter estimates; ASE is the average of the standard error of the estimator; CP is the coverage percentage of 95% confidence interval.

Error	β	PE		ESE		ASE								CP(%)							
		LP	IS	LP	IS	MB		IS		sH		ZL		MB		IS		sH		ZL	
						LP	IS	CF	MB	CF	MB	CF	MB	LP	IS	CF	MB	CF	MB	CF	MB
N	β_1	0.997	1.000	0.170	0.170	0.161	0.161	0.161	0.163	0.155	0.157	0.161	0.163	93.4	93.8	94.0	94.2	93.0	93.5	94.0	94.5
	β_2	1.004	1.009	0.090	0.090	0.089	0.089	0.089	0.090	0.086	0.087	0.089	0.090	93.9	93.8	94.0	94.3	93.0	93.5	94.0	94.1
	β_3	1.000	1.004	0.093	0.093	0.089	0.088	0.089	0.090	0.102	0.103	0.089	0.090	93.8	94.0	94.1	94.3	96.6	96.7	94.0	94.6
L	β_1	0.998	1.000	0.284	0.284	0.274	0.274	0.273	0.275	0.269	0.271	0.273	0.275	94.4	94.5	94.6	94.9	93.9	94.2	94.4	94.7
	β_2	1.008	1.011	0.150	0.150	0.149	0.149	0.148	0.149	0.148	0.149	0.148	0.149	94.9	94.9	94.6	94.6	94.1	94.8	94.7	94.9
	β_3	1.012	1.015	0.151	0.151	0.149	0.149	0.149	0.150	0.159	0.160	0.149	0.150	94.2	94.4	94.0	94.0	95.9	96.1	93.8	94.6
G	β_1	0.999	1.003	0.148	0.148	0.142	0.143	0.143	0.145	0.137	0.139	0.143	0.145	94.7	94.9	94.4	94.5	93.1	93.2	94.4	94.7
	β_2	0.999	1.004	0.082	0.082	0.079	0.079	0.079	0.080	0.075	0.077	0.079	0.080	93.2	93.6	93.6	94.1	91.9	92.3	93.6	94.2
	β_3	1.001	1.006	0.082	0.082	0.079	0.079	0.079	0.081	0.093	0.094	0.079	0.081	93.3	92.7	93.3	93.6	97.3	97.2	93.2	94.2

Table 2: Summary of simulation results based on 1000 replications for full cohort size 3000 and and censoring rate 97%. The bootstrapping size is 500 for each replication. PE is average of point estimates; ESE is the empirical standard deviation of the parameter estimates; ASE is the average of the standard error of the estimator; CP is the coverage percentage of 95% confidence interval.

Error	β	PE		ESE		ASE								CP(%)							
		LP	IS	LP	IS	MB		IS		sH		ZL		MB		IS		sH		ZL	
						LP	IS	CF	MB	CF	MB	CF	MB	LP	IS	CF	MB	CF	MB	CF	MB
case cohort size = 150:																					
N	β_1	1.009	1.021	0.315	0.317	0.281	0.281	0.280	0.304	0.268	0.293	0.280	0.304	91.5	91.6	90.5	93.6	89.3	92.4	90.6	92.8
	β_2	1.014	1.029	0.172	0.172	0.151	0.150	0.149	0.160	0.145	0.157	0.149	0.161	91.9	92.1	90.8	93.5	89.3	91.0	90.8	93.1
	β_3	1.015	1.030	0.180	0.180	0.151	0.149	0.148	0.160	0.178	0.189	0.148	0.160	88.4	88.5	88.8	91.1	92.6	94.8	88.9	91.1
L	β_1	1.031	1.036	0.466	0.467	0.472	0.473	0.463	0.479	0.452	0.469	0.463	0.479	94.7	95.0	94.1	95.0	93.9	93.9	94.1	94.9
	β_2	1.032	1.039	0.254	0.254	0.256	0.256	0.250	0.257	0.252	0.261	0.250	0.258	95.0	95.0	94.5	95.1	95.1	95.5	94.4	95.3
	β_3	1.032	1.039	0.271	0.271	0.255	0.255	0.250	0.258	0.276	0.284	0.250	0.258	94.0	94.2	93.0	94.1	95.6	95.9	92.7	94.3
G	β_1	0.995	1.007	0.291	0.293	0.234	0.235	0.238	0.262	0.229	0.253	0.238	0.262	86.6	87.0	85.6	88.4	83.8	87.9	85.4	88.6
	β_2	1.015	1.030	0.152	0.153	0.129	0.128	0.128	0.139	0.123	0.134	0.128	0.139	90.0	90.5	88.4	91.4	85.9	89.0	88.5	91.8
	β_3	1.013	1.028	0.149	0.149	0.129	0.128	0.128	0.139	0.152	0.162	0.128	0.139	92.2	91.5	89.7	92.8	93.7	96.1	89.7	92.9
case cohort size = 300:																					
N	β_1	1.002	1.008	0.228	0.229	0.214	0.215	0.215	0.221	0.205	0.210	0.215	0.221	92.9	93.8	92.8	93.8	91.3	92.1	92.8	93.9
	β_2	1.005	1.012	0.130	0.130	0.118	0.118	0.119	0.122	0.111	0.114	0.119	0.122	92.8	93.6	93.3	94.6	91.2	92.7	93.4	94.0
	β_3	0.998	1.005	0.129	0.129	0.118	0.118	0.118	0.122	0.144	0.148	0.118	0.122	92.1	92.6	92.6	93.3	97.0	97.0	93.0	92.9
L	β_1	1.048	1.050	0.394	0.395	0.373	0.373	0.368	0.373	0.362	0.366	0.369	0.373	93.7	93.8	93.2	93.5	92.7	92.9	93.2	93.9
	β_2	1.031	1.035	0.221	0.222	0.205	0.205	0.204	0.207	0.204	0.207	0.204	0.207	92.1	92.9	92.6	93.0	92.9	92.9	92.6	92.8
	β_3	1.026	1.029	0.210	0.210	0.202	0.202	0.202	0.204	0.219	0.221	0.202	0.204	93.8	93.6	93.2	93.6	94.5	95.1	93.1	93.3
G	β_1	1.010	1.016	0.188	0.189	0.179	0.179	0.181	0.186	0.171	0.177	0.181	0.187	93.2	92.9	93.4	94.4	91.2	92.7	93.2	94.3
	β_2	1.007	1.015	0.107	0.108	0.099	0.098	0.099	0.102	0.092	0.095	0.099	0.102	91.9	92.6	92.4	93.3	90.1	91.4	92.3	93.0
	β_3	1.009	1.017	0.109	0.109	0.100	0.099	0.100	0.103	0.124	0.127	0.100	0.103	92.4	92.0	92.4	93.4	96.6	97.1	92.2	93.4

Table 3: Summary of timing results in seconds with both point estimation and variance estimation from the simulation study.

C_p	$\bar{m}$	Error	PE		Variance							
			LP	IS	MB		IS		sH		ZL	
					LP	IS	CF	MB	CF	MB	CF	MB
Full cohort size = 1500												
90%	300	N	7.2	1.6	2007.5	561.3	2.9	10.9	2.9	11.4	2.1	11.6
		L	6.5	1.5	1708.0	499.9	2.8	10.2	2.9	10.8	2.2	10.9
		G	6.9	1.6	1899.7	544.6	2.8	10.4	2.8	10.9	2.0	11.1
Full cohort size = 3000												
97%	150	N	0.8	0.6	183.4	150.2	0.4	2.8	0.5	3.0	0.6	3.0
		L	0.7	0.4	143.7	118.2	0.3	2.5	0.5	2.6	0.6	2.7
		G	1.1	0.7	262.3	191.6	0.5	3.5	0.7	3.7	0.7	3.7
	300	N	3.6	0.9	629.6	301.7	2.1	5.5	2.2	5.8	1.3	5.8
		L	3.3	0.7	544.9	237.1	2.0	4.8	2.1	5.1	1.3	5.2
		G	4.1	1.2	816.8	367.8	2.2	6.5	2.3	6.8	1.4	6.9

300, all variance estimators performed reasonably well in estimating the true variation and the coverage percentage was reasonably close to the nominal level. Among all the sandwich variance estimators, the IS-MB and ZL-MB approaches seem to provide confidence intervals with the best coverage percentage. The sH-MB approach is slight inferior, which might be explained by the fact that this approach has two layers of approximation — one from asymptotic linear approximation and the other from induced smoothing.

Of more interest is Table 11, which summarizes the timing results in seconds averaged from 1000 replicates for both point estimation and variance estimation on a 2GHz linux machine. For point estimation with full cohort size 1500 and censoring percentage $C_p = 0.90\%$, the IS approach was up to 4.5 times as fast as the LP approach (with normal error distribution). The multipler bootstrap variance estimation with the IS approach was up to 3.6 times as fast as the LP approach (again with normal error). Nevertheless, the multiplier bootstrap IS approach still needed about 9 minutes on average to obtain a variance estimator. All sandwich variance estimators are strikingly much faster, especially with the closed-form approach: ZL-CF approach took about 2 seconds on average; the IS-CF and sH-CF approaches took about 3 seconds on average. For each sandwich variance estimator, the version with CF estimation of V is over 5 times faster than the version with MB estimation of V. Using the LP approach as benchmark, the IS-CF and sH-CF estimators is 695 times faster and the ZL-CF estimators 1003 times faster. Since the performance of all variance estimators are similar for this setting, the IS-CF, sH-CF

and ZL-CF approaches are obviously preferred for this setup with a mildly rare event.

The timing results for full cohort size 3000 and censoring percentage $C_p = 0.97\%$ follow a similar pattern. Compare to case $C_p = 0.90\%$, time for point estimation is shorter because the number of cases decreases in the case $C_p = 0.97\%$ even when the average case-cohort size were both at 300. Sandwich variance estimators with CF estimation of V is up to 8 times faster than with those with MB estimation of V, at the expense of slightly worse performance in coverage percentage. The IS-MB and ZL-MB approaches yield the most reliable variance estimates but IS-MB is slightly faster than faster than ZL-MB. As the average case-cohort size doubles, the computing time of the sandwich variance estimates with MB estimation for V appear to double accordingly, in contrast to those with CF estimation for V, which do not necessarily double linearly. In summary, based on the performance and speed, our recommended variance estimator is the IS-MB estimator.

2.5 National Wilm's Tumor Study

We demonstrate the performance of our proposed methods with an application to the cohort study conducted by the National Wilm's Tumor Study Group (NWTSG) (D'Angio et al., 1989; Green et al., 1998). Wilm's tumor is a rare kidney cancer in young children. The interest of the study was to assess the relationship between the tumor histology and the outcome, time to tumor relapse. Tumor histology can be classified into two

categories, favorable or unfavorable, depending on the cell type. The central histological diagnosis was made by an individual pathologist at the central pathology center, which was believed to be more accurate than a local diagnosis yet more expensive to measure and required more efforts to obtain. Although the central histology measurements were available for all the cohort members, we assume, in this example, that this measurement was taken only for the subjects in the case-cohort sample. Other covariates that were available for all cohort members were patient age, disease stage and study group. According to the staging system employed by NWTSG, four stages (I – IV) of Wilms' tumors, with Stage IV as the latest stage, indicated the spread of the tumor. Each subject came from one of the two study groups, NWTSG-3 and NWTSG-4. The case-cohort version of the data was analyzed with Cox models (Breslow et al., 2009a; Kulich and Lin, 2004) and additive hazards models (Kulich and Lin, 2000), respectively.

There were a total of 4028 subjects in the full cohort. Among them, 571 were cases who experienced the relapse of tumor — a censoring rate of about 86%. We considered an AFT model for the time to relapse with the following covariates: central histology measurement (1 = favorable, 0 = unfavorable), age (measure in year) at diagnosis, three tumor stages indicators (Stage I as reference) and a study group indicator (NWTSG-3 as reference). The case-cohort version of the data had 668 patients selected as sub-cohort sample and the total case-cohort sample size was 1154. To take advantage of availability of full cohort data, we drew 1000 new sub-cohort samples with size 668 and formed a case-cohort by including the remaining cases for each replicate. We then averaged these

Table 4: National Wilm's tumor study and timing result in seconds.

	PE	SE						
Effects	IS	MB	IS		sH		ZL	
		IS	CF	MB	CF	MB	CF	MB
Case-Cohort Analysis:								
(time)	(8.5)	(3682.2)	(7.7)	(13.9)	(9.5)	(15.7)	(9.4)	(15.0)
histol	−3.428	0.465	0.409	0.458	0.372	0.423	0.410	0.458
age	−0.190	0.079	0.074	0.080	0.221	0.243	0.074	0.080
stage2	−1.283	0.613	0.590	0.621	0.516	0.544	0.590	0.622
stage3	−1.401	0.612	0.579	0.616	0.572	0.602	0.580	0.616
stage4	−2.092	0.717	0.665	0.712	0.717	0.763	0.666	0.712
study	−0.128	0.475	0.455	0.484	0.451	0.482	0.455	0.484
Full-Cohort Analysis:								
(time)	(266.0)	(126927.7)	(309.9)	(453.0)	(341.3)	(486.1)	(321.1)	(494.0)
histol	−2.749	0.202	0.148	0.213	0.138	0.214	0.148	0.196
age	−0.127	0.037	0.029	0.039	0.081	0.092	0.029	0.039
stage2	−1.335	0.280	0.233	0.285	0.200	0.280	0.234	0.271
stage3	−1.341	0.286	0.239	0.297	0.211	0.288	0.240	0.299
stage4	−2.203	0.319	0.245	0.321	0.219	0.300	0.247	0.334
study	−0.106	0.226	0.175	0.229	0.162	0.224	0.176	0.219

estimates and estimated standard errors from the 1000 replicates of the case-cohort analysis.

The results of the average from 1000 replicates of case-cohort analyses are summarized in Table 4. Due to its poor timing performance, the LP approach was not considered. Since the MB standard error is considered to reflect the true variation quite well from the simulation study, we are interested in how close the various sandwich standard errors to the MB standard error. For all three sandwich estimators, the CF versions systematically underestimate noticeably, although the underestimation is less

severe in the IS and ZL approach than in the sH approach. The MB versions of the sandwich estimates appear to agree with the MB standard error closely, and again the agreement appears to be better for the IS and ZL approach than for the sH approach. In particular, the sH standard error for the age effect is about three times as much as that from other approaches. The standard errors from IS-MB and ZL-MB are almost identical, both very close to the time consuming MB standard error. Based on the IS-MB standard errors, the coefficients of central histological diagnosis, age, and all three stage indicators were found to be significantly different from zero with p-values 0.000, 0.009, 0.019, 0.011 and 0.002, respectively. No significant difference was found between the two study groups. In terms of timing, the MB-IS standard error took over a hour whereas the MB-based sandwich estimates only took 14–15 seconds on average.

For comparison purpose, we also analyzed the full cohort data with the same approaches and reported the results in Table 4. Point estimates are close to these in case-cohort analysis, with their standard errors taken into consideration. All the standard errors decrease compared to the case-cohort analyses, which is expected as full information became available for all covariates. The best sandwich variance estimators are still IS-MB and ZL-MB, both closely approximates the full blown MB standard error. With full cohort size 4028 and censoring rate 86%, the IS point estimates took 4 and a half minutes, the MB variance estimation took 35.36 hours, while IS-MB only took 7 and a half minutes.

2.6 Discussion

In AFT modeling of case-cohort data, both point estimation and variance estimation are challenging with the nonsmooth estimating equations. Resampling methods are commonly used to estimate the variance, which are time consuming even with a computationally efficient point estimator such as our induced smoothing approach with rank-based estimating equations. We have proposed six sandwich variance estimators and compared their performances with the bootstrap variance estimator in numerical studies. The IS-MB and ZL-MB approaches were found to provide good approximation to the true variation and are computationally very efficient. All the methods are implemented in an R package aftgee (Chiou et al., 2012a). The package had the potential to bring AFT modeling of case-cohort data into routine analysis.

The IS approach was built on Gehan's weight for rank-based estimating equations, in which case closed-form expectations of the perturbed estimating equations are available. Alternative weights such as the logrank weight are possible, though the computation is less straightforward than that for Gehan's weight. Incorporating a general, possibly optimized weight in the IS approach merits further investigation for both full cohort and case-cohort data. The estimates from Gehan's weight always serve as a good initial value in numerical equations solving.

It would be worthwhile to consider extensions of the proposed methods in several directions. It is often the case that some auxiliary covariates are available for the entire

cohort. Then, the selection of the subcohort members can depend on the strata constructed by using the information available for the whole cohort members. Resulting estimators from this stratified case-cohort design were shown to be more efficient than their traditional case-cohort counterpart for the Cox model (Kulich and Lin, 2004) and the additive hazards model (Kulich and Lin, 2000). An extension of the proposed methods with the AFT model to a stratified case-cohort design is straightforward and we also expect an improvement on the efficiency.

Another extension to consider is to accommodate multivariate failure time data from a case-cohort study. Unlike the univariate failure time data which assumes independence among failure times, a possible dependence among failure times within the same cluster needs to be taken into account. For the Cox model, Kang and Cai (2009a) used a marginal model approach and a similar approach can be considered for the AFT model. We will explore these possibilities as the next step to the present work.

Chapter 3

Efficient Rank-Based Approach from Clustered Failure Times with Stratified Sampling

3.1 Introduction

Clustered failure times from stratified sampling designs present methodological challenges in statistical analysis to correct the sampling bias and account for within cluster dependence. One such example arises from a retrospective cohort dental study where times to extraction of teeth were compared within each patient (Caplan et al., 2005). A tooth with pulpal involvement due to deep caries or restoration requires extraction or root canal therapy (RCT). While RCT is expected to extend tooth life, it may not last a lifetime. Using the RCT status as an indicator of pulpal involvement, the survival time of a root canal filled (RCF) tooth was compared to a tooth with no pulpal involvement from the same patient. A stratified random sampling design was used where two strata

were defined using the RCF tooth status within each patient. The first stratum consists of all the patients whose RCF tooth had been extracted and the second stratum consists of all the patients whose RCF tooth had not been extracted. Patients were randomly sampled within each stratum but with different selection probabilities: 85% for the first stratum and 11% for the second stratum. After the sampling of two strata, a matched non-RCF tooth was selected for each patient for comparison. The observed data has two complications. First, because of the stratified sampling design, the sample did not constitute a random sample but instead a biased sample. Second, because the teeth being compared were clustered within a patient, their survival times might not be independent. A proper statistical inference procedure needs to take these two statistical issues into account: biased nature of the study sample and correlated feature of the failure times.

Stratification is a sampling technique in which the population is divided into different mutually exclusive subgroups, or strata, and then randomly selecting subjects from each strata (Cochran, 1977; Smith, 2001). Stratified random sampling is often utilized instead of simple random sampling to ensure inclusion of subjects from underrepresented subgroups. For failure time outcomes, an example of the stratified random sampling design is the stratified case-cohort design, a stratified variant of the classical case-cohort design (Prentice, 1986). Under a stratified case-cohort design, strata are usually constructed based on the failure status of the subjects. Cases and controls refer to those who have experienced the event of interest by the end of the study and those

who has not, respectively. The cases form a stratum themselves, whereas the controls can be further classified into multiple strata based on some covariates that are available for all cohort members. This design is frequently employed in epidemiological studies. It is especially useful when the outcome is rare and the main covariates are expensive to measure. The design of the aforementioned dental study is a example where case patients and control patients formed two strata. Analysis of stratified case-cohort study data requires special attention since the observed sample is biased. Such biases can be adjusted by incorporating a weight that is the inverse of the sampling probabilities or their estimates (Borgan et al., 2000; Kulich and Lin, 2004)

Most of the existing works on stratified case-cohort design have focused on Cox proportional hazards models (Cox, 1972). For univariate failure times, an exposure-stratified case-cohort design and related estimation procedures were proposed by Borgan et al. (2000) and extended to more efficient estimation procedures by Kulich and Lin (2004), Breslow et al. (2009a) and Breslow et al. (2009b) rigorously studied both the theoretical properties and practical implementation of efficient estimators in stratified case-cohort samples under the framework of two-stage sampling. Kim and Gruttola (1999) and Samuelsen et al. (2007) considered various stratification for general cohort sampling designs which includes stratified case-cohort design as a special case. For multivariate failure times, the literature is much less developed. Lu and Shih (2006) and Zhang et al. (2011) considered an extension of the classical case-cohort study design to clustered failure times and proposed related inference procedures. Kang and Cai (2009b)

proposed estimation procedures for data from case-control studies with clustered failure times which is a special case of the stratified case-cohort design.

A semiparametric accelerated failure time (AFT) model is an alternative to the popular Cox model in analyzing failure time data. It directly links the expected failure time to covariates through a log-linear model without specifying the error distribution. A popular estimation procedure for semiparametric AFT models is the rank-based estimating equations approach motivated from inverting the weighted log-rank test (Prentice, 1978). The asymptotic properties of the resulting estimator has been rigorously studied by Tsiatis (1990) and Ying (1993). Rank-based estimators have also been considered for case-cohort data (Nan et al., 2006; Yu, 2011; Yu et al., 2007) and stratified case-cohort data (Kong and Cai, 2009). While this class of estimators have nice theoretical properties, it has not been as popular as the Cox model in practice due to lack of efficient and reliable computing algorithm in obtaining both parameter estimates and their standard errors. The main difficulty in rank-based estimation for AFT models comes from the nonsmoothness in the rank-based estimating equations. A linear programming approach to solving nonsmooth estimating equations and a resampling approach for estimating the variance of the estimator was proposed by Jin et al. (2003). The method has been extended to case-cohort designs (Kong and Cai, 2009) and to clustered failure times (Jin et al., 2006c). Because of the linear programming component, this method can be computationally very demanding even for moderate sample sizes with a moderate number of covariates. The induced smoothing approach proposed by Brown and Wang

(2005) is computationally efficient because it replaces the nonsmooth equations with an asymptotically equivalent version. This approach has been adapted to case-cohort data (Chiou et al., 2013a) and to clustered failure times (Johnson and Strawderman, 2009; Wang and Fu, 2011). It opens a route to inferences for semiparametric AFT models with clustered failure times from stratified case-cohort designs.

In this article, we propose weighted rank-based estimating equations for fitting semiparametric AFT with clustered failure times from stratified random sampling with the induced smoothing approach. The asymptotic properties of the estimator from the nonsmooth weighted rank-based estimating equations in this setting, which have not been studied in the literature, are established. The estimator from the smoothed version of the estimating equations are shown to be consistent and have the same asymptotic distribution as that from the nonsmooth version. The variance of the estimator is estimated by computationally efficient sandwich estimators aided by a multiplier bootstrap. The whole methods are available in an R package aftgee (Chiou et al., 2012a, 2013b).

The rest of this article is organized as follows. The weighted rank-based estimating equations for AFT models with clustered data under stratified sampling are presented, and the asymptotic properties of the resulting estimator are established in Section 3.2. In Section 3.3, an estimation procedure based on induced smoothing is developed and the asymptotic properties of the estimator is established. Variance estimation procedures are also provided. The performance of the estimators are investigated via large scale simulation studies in Section 3.4. The method is applied to the dental study data in

Section 3.5. A discussion concludes Section 3.6.

3.2 AFT Model and Stratified Sampling

Consider a random sample of n independent clusters with K members in each cluster. Let T_{ik} be the log-transformed failure time, C_{ik} be a corresponding log-transformed censoring time and X_{ik} be the $p \times 1$ covariate vector for the kth member in the ith cluster, $k = 1, \ldots, K, i = 1, \ldots, n$. In our dental study example, each patient i has $K = 2$ teeth. In particular, T_{i1} and T_{i2} are the log-transformed failure times of the RCF tooth and non-RCF tooth for the ith patient, respectively. Suppose T_{ik} is conditionally independent of C_{ik} given X_{ik} then the multivariate accelerated failure time model is

$$T_{ik} = X_{ik}^{\top}\beta + \epsilon_{ik}, \qquad i = 1, \ldots, n, k = 1, \ldots, K,$$

where β is an unknown $p \times 1$ vector of regression parameters and the error terms, $\epsilon_i = \{\epsilon_{i1}, \ldots, \epsilon_{ik}\}$, are independent and identically distributed throughout clusters. We also assume that $\epsilon_{i1}, \ldots, \epsilon_{iK}$ have identical marginal distribution for clustered data. Define $Y_{ik} = \min(T_{ik}, C_{ik})$, $\Delta_{ik} = I[T_{ik} < C_{ik}]$, where $I[\cdot]$ is the indicator function. The full cohort data are independent and identically distributed copies of $\{Y_{ik}, \Delta_{ik}, X_{ik}\}$, $k = 1, \ldots, K$, $i = 1, \ldots, n$. Suppose the full cohort are divided into S mutually exclusive strata. A study sample is assumed to be taken via a stratified simple random sampling at the cluster level with possibly different inclusion probabilities for different strata,

$p_{n,s} = \tilde{n}_s/n_s$, where n_s and $\tilde{n}_s$ are the numbers of clusters and sampled clusters in the sth stratum, respectively, $s = 1, \ldots, S$.

If the full cohort data are available, β can be estimated as the root of a set of rank-based estimating equations. Let $e_{ik}(\beta) = Y_{ik} - X_{ik}^\top \beta$ and $N_{ik}(\beta; t) = \Delta_{ik} I(e_{ik}(\beta) \leq t)$. For a vector a, define $a^{\otimes 0} = 1$ and $a^{\otimes 1} = a$. The rank-based estimating equation with Gehan's weight (e.g., Jin et al., 2006c) is

$$U_{n,G}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_{-\infty}^{\infty} W^{(0)}(\beta; t) \left\{ X_{ik} - \bar{X}(\beta; t) \right\} \mathrm{d}N_{ik}(\beta; t) = 0$$

where $\bar{X}(\beta; t) = W^{(1)}(\beta; t)/W^{(0)}(\beta; t)$, and $W^{(d)}(\beta; t) = n^{-1} \sum_{j=1}^{n} \sum_{l=1}^{K} X_{jl}^{\otimes d} I\{e_{jl}(\beta) \geq t\}$, $d = 0, 1$. A frequently used equivalent form is

$$U_{n,G}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} \Delta_{ik} (X_{ik} - X_{jl}) I\{e_{jl}(\beta) \geq e_{ik}(\beta)\} = 0. \tag{3.1}$$

When the full cohort data are not available, (3.1) cannot be evaluated and using the observed data naively would lead to misleading results because the observed data are a biased sample. To accommodate the stratified random sampling scheme, equation (3.1) needs to be modified with proper weight. Let ψ_{is} be the strata indicator; $\psi_{is} = 1$ if the ith cluster is in the sth stratum and $\psi_{is} = 0$ otherwise. Let ξ_i be the sampling indicator; $\xi_i = 1$ if the ith cluster is sampled and $\xi_i = 0$ otherwise. Then, the weighted version

of (3.1) is

$$U_n^c(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{K} h_i h_j \Delta_{ik}(X_{ik} - X_{jl}) I\{e_{jl}(\beta) \geq e_{ik}(\beta)\} = 0 \qquad (3.2)$$

where $h_i = \sum_{s=1}^{S} \xi_i \psi_{is}/p_{n,s}$ is the inverse of the inclusion probability for the ith cluster. Notice that both the strata indicator and the stratified weight are at the cluster level that is shared by all K members in the ith cluster. If we sample all the clusters within each strata, i.e., $p_{n,s} = \xi_i = 1$ for $i = 1, \ldots, n$ and $s = 1, \ldots, S$, then we have the full cohort data and (3.2) reduces to (3.1). The aforementioned dental study is a special case with $S = 2$ where $\psi_{i1} = 1$ if $\Delta_{i1} = 1$, and $\psi_{i2} = 1$ if $\Delta_{i1} = 0$.

Let $\hat{\beta}_n^c$ be the solution to (3.2). Under regularity conditions, $\hat{\beta}_n^c$ is consistent to the true regression coefficients β_0 and asymptotically normal. The following theorem, whose proof is sketched in the Appendix A.2, summarizes the asymptotic property of $\hat{\beta}_n^c$ and lays out the structure of the asymptotic variance matrix.

Theorem 1. *Under the conditions A1–A7 in the Appendix, $\hat{\beta}_n^c$ is strongly consistent for β_0 and $n^{1/2}(\hat{\beta}_n^c - \beta_0)$ is asymptotically normally distributed with mean zero and variance matrix $\Sigma(\beta_0) = A(\beta_0)^{-1}V(\beta_0)A(\beta_0)^{-1}$ whose explicit forms are given in the Appendix.*

3.3 Estimation Procedures

3.3.1 Smoothed Estimating Equations

Estimating β by solving estimating equation (3.2) is challenging because $U_n^c(\beta)$ is not a smooth function of β. Although the problem can be solved by a linear programming approach (Jin et al., 2003), the computation is very intensive even for moderate sample sizes, and to a even greater degree when the variance of the estimator needs to be estimated via bootstrapping. A computationally more efficient approach is induced smoothing introduced by Brown and Wang (2005), which has been applied to AFT modeling for univariate failure time (Brown and Wang, 2007), clustered failure times (Johnson and Strawderman, 2009), and univariate failure times from case-cohort studies (Chiou et al., 2013a). We extend the induced smoothing approach to accommodate clustered failure times from stratified sampling.

The essence of the induced smoothing approach is to smooth the step estimating functions in a way that the solution of the smoothed estimating equations have the same asymptotic properties as that of equation (3.2), but are much easier to obtain. In particular, let Z be a p-dimensional standard normal random vector that is independent of the data. A smoothed version of (3.2) is constructed by replacing $U_n^c(\beta)$ with $E[U_n^c(\beta+ n^{-1/2}Z)]$, where the expectation is taken with respect to Z and the scale is $n^{-1/2}$ because

$\hat{\beta}_n^c$ is $\sqrt{n}-$consistent. This lead to

$$\tilde{U}_n^c(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{K} h_i h_j \Delta_{ik}(X_{ik} - X_{jl})\Phi\left\{\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}}\right\} = 0, \qquad (3.3)$$

where $\Phi(\cdot)$ is the distribution function of a standard normal variate and $r_{ikjl}^2 = n^{-1}(X_{ik} - X_{jl})^\top(X_{ik} - X_{jl})$. The smoothed estimating equations (3.3) are now continuously differentiable with respect to β, and hence can be solved with standard numerical algorithms. For example, we used a nonlinear equation solver in R package `BB` (Varadhan and Gilbert, 2009).

Let $\tilde{\beta}_n^c$ be the solution to equation (3.3). Under regularity conditions, $\tilde{\beta}_n^c$ and $\hat{\beta}_n^c$ have the same limiting distribution as $n \to \infty$. This result is summarized by the following theorem with proof sketched in the Appendix A.2.

Theorem 2. *Under the conditions A1–A7 in the Appendix, $\tilde{\beta}_n^c$ is strongly consistent for β_0 and $n^{1/2}(\tilde{\beta}_n^c - \beta_0)$ converges to a normal distribution with mean zero and variance matrix $\Sigma(\beta_0)$.*

3.3.2 Variance Estimation

To estimate the asymptotic variance of $\tilde{\beta}_n^c$, we propose strategies similar to those in Chiou et al. (2013a). Two classes of estimators are considered: full multiplier bootstrap approach and sandwich estimators.

In the full multiplier bootstrap approach, we first generate independent and identically distributed positive multipliers η_i, $i = 1, \ldots, n$, that is independent of the observed data, with $\mathrm{E}(\eta_i) = \mathrm{Var}(\eta_i) = 1$. Then, given a realization of $(\eta_1, \ldots, \eta_n)$, a bootstrap replicate $\tilde{\beta}_n^*$ may be obtained by solving the following perturbed estimating equation

$$\tilde{U}_n^{c*}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{K} h_i\eta_i h_j\eta_j\Delta_{ik}(X_{ik} - X_{jl})\Phi\left\{\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}^2}\right\} = 0. \qquad (3.4)$$

The variance of $\tilde{\beta}_n^c$ can be approximated by the sample variance of the B bootstrap replicates, $(\tilde{\beta}_n^{c*(1)}, \ldots, \tilde{\beta}_n^{c*(B)})$, obtained from repeatedly generating multipliers $(\eta_1, \ldots, \eta_n)$ and solving equation (3.4). This procedure is much faster than the unsmooth version but still very computing intensive because (3.4) needs to be solved B times.

For sandwich estimators, we need to estimate $A(\beta_0)$ and $V(\beta_0)$ in the sandwich variance form of $\Sigma(\beta_0)$. These estimators avoid solving estimating equations repetitively, thus are expected to be computationally more efficient than the full multiplier bootstrap method. To estimate $V(\beta_0)$, we can use either a closed-form direct approximation and multiplier bootstrap procedure. The direct approximation of $V(\beta_0)$ uses its closed form expression and replaces all unknown quantities in $V(\beta_0)$ with their sample version evaluated at $\tilde{\beta}_n^c$. The explicit form of $V(\beta_0)$ is presented in the Appendix. Nevertheless, estimating these unknown quantities, such as the baseline hazard function of the error term, requires nonparametric estimation that can sometimes be complicated. To avoid this complexity while increasing the accuracy (Chiou et al., 2013a; Zeng and Lin, 2008),

we estimate $V(\beta_0)$ using multiplier bootstrap procedure. Given a realization of multipliers $(\eta_1, \ldots, \eta_n)$ with unit mean and unit variance, a bootstrap replicate of $\tilde{U}_n^{c*}(\beta)$ evaluated at $\tilde{\beta}_n$ is obtained. The bootstrap estimate of $V(\beta_0)$ is the sample variance of B bootstrap replicates of $\tilde{U}_n^{c*}(\tilde{\beta}_n)$. Unlike the full multiplier bootstrap procedure, this procedure only involves evaluations of estimating equations (3.4) instead of solving it. Thus, it is much less computationally demanding.

To estimate the slope matrix $A(\beta_0)$, we propose three estimators. The first estimator is the derivative of the smoothed estimating function $\tilde{U}_n^h(\beta)$,

$$\hat{A}_n(\beta_0) = \frac{1}{n}\frac{\partial}{\partial \beta^\top}\tilde{U}_n^h(\tilde{\beta}_n) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{K} h_i h_j \Delta_{ik}\frac{(X_{ik}-X_{jl})^2}{r_{ikjl}}\phi\left\{\frac{e_{jl}(\tilde{\beta}_n)-e_{ik}(\tilde{\beta}_n)}{r_{ikjl}}\right\},$$

evaluated at $\tilde{\beta}_n^c$ where $\phi(\cdot)$ is the density function of a standard normal variate. We name this method the induced smoothing approach. The second estimator of $A(\beta_0)$ is adapted from the estimation procedure in Zeng and Lin (2008) for nonsmooth estimating equations. Let Z_b be the bth realization of the p-dimensional standard normal random vector, $b = 1, \ldots, B$. The jth row of $A(\beta_0)$ is estimated by the least squares estimate of the regression coefficients when regressing $n^{-1/2}U_{nj}^c(\hat{\beta}_n + n^{-1/2}Z_b)$ on Z_b where $U_{nj}^c(\cdot)$ denote the jth row of $U_n^c(\cdot)$. Lastly, we adapt estimation procedure in Huang (2002) to the induced smoothing estimating equations. Given an estimator $V_n(\beta_0)$ for $V(\beta_0)$, let $V_n(\tilde{\beta}_n^c) = L_n^\top L_n$ be the its Cholesky decomposition. Let q_{nj} be the solution to the following p estimating equations for γ, $n^{-1}\tilde{U}_n^c(\gamma) = n^{-1/2}l_j$ $(j = 1, \cdots, p)$, where l_j is

the jth column of L_n. Then $\Sigma_n(\tilde{\beta}_n) = Q_n^\top Q_n$ where Q_n is a matrix with its jth column being $q_{nj} - \tilde{\beta}_n^c$. This this approach involves solving only p estimating equations.

3.4 Simulation

Two simulation studies were conducted to evaluate the performance of the proposed procedures. The first study was designed to evaluate the finite sample performance under general stratified sampling. For cluster i in a full cohort, we generated bivariate failure times, $T_i = (T_{i1}, T_{i2})$, from

$$\log T_{ik} = 0.5X_{1ik} + 0.5X_{2ik} + \epsilon_{ik}, \quad k = 1, 2,$$

where X_{1ik} was Bernoulli with rate 0.5 and X_{2ik} depended on X_{1ik}; X_{2ik} was $N(0, 1)$ if $X_{1ik} = 0$ and $N(-0.5, 1)$ if $X_{1ik} = 1$. The failure times were subject to independent censoring by a uniform distribution, from 0 to c, with c calibrated to achieve a desired censoring rate in the full cohort. The full cohort size was set to be $n = 1000$. Two levels of censoring rates were considered: 0.90 and 0.97. The error term, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})$, was a bivariate random vector specified by identical marginal distributions and a Clayton copula. Three marginal error distributions were considered: standard normal, standard logistic, or standard Gumbel (abbreviated by N, L and G). The Clayton copula was specified to give three levels of dependence measured by Kendall's tau (τ): 0, 0.3 and 0.6. After a full-cohort dataset was generated, it was further divided into three strata.

The first stratum consisted of all the clusters with $\Delta_{i1} = 1$ or $\Delta_{i2} = 1$; that is, at least one of the two members were observed with an event. For $\Delta_{i1} = 0$ and $\Delta_{i2} = 0$, the second stratum contained all clusters with $X_{2i1} + \gamma_i \leq 0$ and the third stratum contained the remaining clusters, where γ_i is an additional, independent $N(0, 1)$ variable. This means that stratum 2 and 3 were determined by the second covariate from the first member and an independent variable γ_i. A stratified sample of size 300 was drawn from the full-cohort. In the stratified sample, all clusters from stratum 1 were included; strata 2 and 3 were then sampled with equal quota to fulfill the designated case-cohort size. When censoring rate is 0.90, the corresponding weights were 1, 6.7 and 5.6 for strata 1, 2 and 3, respectively. On the other hand, under censoring rate of 0.97, the corresponding weights were 1, 4.3 and 3.3 for strata 1, 2 and 3, respectively.

The results of study 1 are summarized in Table 5 and Table 6 for censoring rates of 0.90 and 0.97, respectively. We reported the average point estimates (PE), the empirical standard error (ESE), the average of standard error (ASE), and the empirical coverage percentage (CP) of the 95% confidence intervals for the two regression coefficients. Three types of sandwich variance estimator were computed. They are named after how the slope matrix $A(\beta)$ was estimated: induced smoothing (IS) approach, Zeng and Lin's (ZL) approach, and the smoothed Huang's (SH) approach. Depending whether $V_n(\beta)$ is based on closed-form (CF) or multiplier bootstrap (MB), we will have two versions of $\hat{\Sigma}_n(\tilde{\beta}_n^c)$ for each type of sandwich estimator. This gives a total of six sandwich estimators. For all scenarios, our point estimates appear to be virtually unbiased. The average

standard errors from all six variance estimators and the empirical standard errors agree closely, suggesting that the variance estimators provide good estimation of the empirical variation. The empirical coverage percentages are reasonably close to 95%. The standard errors increase as censoring rate increase form 90% to 97% but does not seem to vary with the level of dependence. These observations were invariant to the error distributions.

The goal of the second simulation is to mimic the stratified case-cohort design of the aforementioned dental study. For cluster i, bivariate failure times were generated from

$$\log T_{ik} = 2 + X_{1ik} + X_{2ik} + \epsilon_{ik}, \quad k = 1, 2,$$

where $X_{1i1} = 1$, $X_{1i2} = 0$, X_{2ik} was $N(0, 1)$, and the error distribution had the same settings as in the first study. For each cluster, the first member mimics the RCF tooth and the second member mimics the non-RCF tooth. The clusters were categorized into cases and controls depending on the status of first failure time: a cluster is a case cluster if $\Delta_{i1} = 1$ and control cluster otherwise. The cases and controls naturally formed two strata. The full-cohort size was set to be 1000, with censoring rate 90%. A stratified sample was obtained by simple random sampling 90% of the clusters from the first stratum and then sample the same number of clusters from the second stratum. Thus, the average sub-cohort size is 180. Under this design, the inclusion probability of cluster i only depends on the failure status of its first member, and the weight is at the cluster level. The average corresponding weight for cluster 1 and 2 are 1.11 and 10, respectively.

We generated 1000 replications for each scenario.

The results of study 2 are summarized in Table 7. All estimates appear to be virtually unbiased and their standard errors are generally close to the empirical standard deviation of the estimates. The coverage probabilities are also reasonably close to the nominal level. Compared to the variance estimates obtained with closed-form estimate for $V(\beta)$, those obtained with multiplier bootstrap for $V(\beta)$ seem to provide confidence intervals with better coverage percentage. These estimators were also recommended by Chiou et al. (2013a) for univariate case-cohort data.

3.5 Retrospective Dental Study

We applied our proposed methods to the retrospective dental study analyzed in Caplan et al. (2005) and Kang and Cai (2009b). As described earlier, it was of interest to investigate the effect of pulpal involvement on time to tooth extraction. RCF teeth were used as an indicator of pulpal involvement. The sampled subjects were current or retired employees (or their dependents) of companies with dental insurance through the Kaiser Permanente Dental Care Program (KPDCP), a dental maintenance organization located in Portland, Oregon, United States, between 1987 to 1994. A total of 1795 patients who had at least one RCF tooth were in the full cohort, out of which, 272 were cases and 1523 were controls. Cases were those who lost the RCF tooth during the study period and controls were those who did not lose the RCF tooth during the same period.

Using a simple random sampling without replacement, 232 cases and 174 controls were selected; that is, the inclusion probabilities were 82.29% and 11.42%, respectively. After the sample was selected, within each patient, a contralateral non-RCF tooth was chosen to compare with the RCF tooth. A total of 202 subjects were further selected. If the contralateral tooth was missing or already had RCF on the RCF tooth's access date (index date), the tooth of the same type (anterior, premolar, or molar) adjacent to the contralateral tooth was selected. Of the 406 sampled patients, only those who satisfied the study eligibility criteria (Caplan et al., 2005) were included, which resulted in 202 patients (111 case patients and 91 control patients).

We considered an AFT model for the teeth survival time with RCF status (ROOT = 1) as the main risk factor. Since we were also interested in the effect of RCT by tooth type, we also considered an interaction term (MOLAR:ROOT) between the RCF status and molar tooth type indicator (MOLAR) as another risk factor in our model. Covariates included in the model are proximal contacts (PC) and the number of pockets larger than five millimeter (POCKET). PC is a categorical variable measuring the contact between the distal surface of one tooth and the mesial surface of an adjacent tooth. Four mutually exclusive categories for PC were created: nonbridge abutment with zero PC (PC0), nonbridge abutment with one PC (PC1), nonbridge abutment with two PCs (PC2), and bridge abutment (PCABUT), PC1 and PC2 made up the majority of the PC (combined to 90%). A periodontal pocket refers to an unusually deep space between the teeth and gums. When the measure exceeds three millimeters, regular brushing cannot

effectively remove debris from the area. For a given tooth, six sites of pocket depths were recorded. Numbers of measurements greater than five millimeters were counted. In the dataset, about 30% of the teeth have periodontal pockets. The same main risk factor and covariates were considered in Kang and Cai (2009b)

The results of the data analysis are summarized in Table 8. In addition to the six sandwich variance estimators used in simulation studies, we also included the variance estimate obtained from a full MB procedure for comparison purpose. The bootstrap sample size was set to be 500. All the sandwich variance estimators seem to provide similar results which are also close to those from using the full MB procedure the MB estimator with an exception of the SH approach. In particular, the SH standard error for the POCKET effect is about three times as much as that from other approaches, though this does not change the conclusion of insignificance at 5% level. For speed comparison purpose, we also included the timing result. As expected, the MB estimator took the longest with 1564.8 seconds and our sandwich estimators are up to 381 times faster.

The MOLAR effect was found to be significantly positive, which is not surprising because molars tend to last longer. The interaction between MOLAR and ROOT was found to be significant. The RCF effect among molars can be computed by adding the coefficient estimates of two covariates, ROOT and MOLAR:ROOT, yielding a point estimation of -2.657 and a p-value of 0.0004 under ISMB. This suggests that, after adjustment for proximal contacts and the number of pockets larger than five millimeters, a molar with RCT has shorter life span than one without RCT. For a non-molar tooth,

the effect of RCT was not statistically significant (p-value = 0.501 with the ISMB variance estimator). These findings were similar to those in Kang and Cai (2009b) who reported, for a molar tooth, a significantly increased risk of extraction with RCT after adjusting for the same covariates.

3.6 Discussion

Application of semiparametric AFT models in routine survival analysis has been held back by lack of reliable and efficient computing method. Taking advantage of recent advances, Chiou et al. (2013a) proposed fast and accurate estimators for parameters in semiparametric AFT models in case-cohort studies with univariate failure time data. We generalized the sampling weights to a stratified sampling design and extended their estimators to accommodate clustered failure times. The asymptotic properties of the estimators from nonsmooth weighted rank-based estimation equations are established first. The estimators from our induced smoothing procedure were shown to be consistent and have the same asymptotic distribution as the estimators from the nonsmooth weighted rank-based estimating equations. We proposed various ways to compute sandwich variance estimator to avoid a full bootstrap procedure which can be very expensive to compute. Our sandwich variance estimators provide good approximations of the true variation and are very computationally efficient.

The induced smoothing approach was based Gehan type of weight in the rank-based

estimating equations (Gehan, 1965). Other weights such as log-rank (Prentice, 1978), Prentice-Wilcoxon (Prentice, 1978), or the general G^p class (Harrington and Fleming, 1982) can also be considered. The smoothed estimating equations with these alternative weights, however, are in general much harder to obtain. Moreover, due to the complicated nature of the resulting estimating equations, extra efforts will be required for solving them. One possible option is to approximate the estimating equations by an asymptotically equivalent version which are simpler to work with. This could be a possible extension of our proposed methodology.

Another extension to consider is to accommodate within cluster correlations to improve efficiency. For full cohort data, Wang and Fu (2011) decomposed the estimating equations into between cluster estimating equations and within cluster estimating equations and combined them in an optimal way to obtain a more efficient estimator. In the spirit of generalized estimating equations, Chiou et al. (2012b) proposed a least squares approach to account for within cluster correlation through a working correlation structure and demonstrated substantial efficiency gain in parameter estimation relative to that under working independence when the within cluster dependence is strong. It would be worthwhile to consider incorporating weights in these approaches to model clustered failure times under stratified sampling designs for higher efficiency.

Table 5: Summary of simulation study 1 with three strata. Results are based on 1000 replications for censoring rate 90% with $n = 1000$. The bootstrapping size is 500 for each replication. PE is the average of point estimates; ESE is the empirical standard deviation of the parameter estimates; ASE is the average of the standard error of the estimator; CP is the coverage percentage of 95% confidence interval.

					ASE						CP(%)					
					IS		ZL		SH		IS		ZL		SH	
τ	dist	β	PE	ESE	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF
0	N	β_1	0.500	0.123	0.120	0.119	0.120	0.119	0.120	0.119	94.0	94.1	94.2	93.9	94.5	94.1
		β_2	0.503	0.061	0.062	0.061	0.062	0.061	0.063	0.062	96.0	95.6	96.1	95.5	95.9	95.8
	L	β_1	0.501	0.095	0.096	0.095	0.096	0.095	0.096	0.095	94.8	94.7	94.9	94.8	94.7	94.8
		β_2	0.506	0.052	0.051	0.050	0.051	0.050	0.051	0.050	94.4	93.6	94.3	94.0	94.5	94.1
	G	β_1	0.508	0.220	0.220	0.218	0.220	0.218	0.217	0.215	95.1	95.1	95.5	95.3	95.4	94.9
		β_2	0.505	0.111	0.111	0.109	0.112	0.109	0.117	0.115	94.5	94.1	94.5	94.2	95.6	95.5
0.3	N	β_1	0.503	0.122	0.119	0.118	0.119	0.118	0.119	0.118	94.9	95.0	95.0	95.0	94.3	95.1
		β_2	0.504	0.060	0.061	0.060	0.061	0.060	0.062	0.061	96.2	96.0	96.1	96.0	96.4	96.1
	L	β_1	0.505	0.097	0.095	0.094	0.095	0.094	0.095	0.094	94.9	94.3	94.7	94.2	94.3	94.4
		β_2	0.508	0.050	0.050	0.049	0.050	0.049	0.050	0.049	94.6	94.2	94.8	94.6	94.5	94.4
	G	β_1	0.511	0.216	0.213	0.211	0.212	0.211	0.210	0.209	95.7	95.8	95.8	95.7	95.2	95.3
		β_2	0.509	0.107	0.108	0.106	0.108	0.106	0.113	0.111	95.1	94.9	95.6	94.7	96.2	95.7
0.6	N	β_1	0.508	0.121	0.118	0.117	0.118	0.117	0.118	0.117	95.7	95.5	95.4	95.4	95.5	95.5
		β_2	0.507	0.058	0.061	0.060	0.061	0.060	0.061	0.061	95.2	95.1	95.7	95.1	95.6	95.5
	L	β_1	0.508	0.097	0.094	0.093	0.094	0.093	0.094	0.093	95.0	94.7	94.4	94.6	95.1	94.6
		β_2	0.505	0.049	0.049	0.049	0.050	0.049	0.050	0.049	95.2	94.8	95.3	94.8	95.0	95.0
	G	β_1	0.502	0.202	0.210	0.209	0.210	0.209	0.207	0.206	95.8	95.4	95.6	95.2	94.5	95.0
		β_2	0.509	0.100	0.106	0.104	0.106	0.104	0.112	0.110	96.0	95.5	95.6	95.6	97.0	96.8

Table 6: Summary of simulation study 1 with three strata. Results are based on 1000 replications for censoring rate 97% with $n = 1000$. The bootstrapping size is 500 for each replication. PE is the average of point estimates; ESE is the empirical standard deviation of the parameter estimates; ASE is the average of the standard error of the estimator; CP is the coverage percentage of 95% confidence interval.

					ASE						CP(%)					
					IS		ZL		SH		IS		ZL		SH	
τ	dist	β	PE	ESE	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF
0	N	β_1	0.510	0.196	0.177	0.176	0.178	0.176	0.177	0.176	92.4	92.6	93.1	92.7	92.6	92.5
		β_2	0.517	0.099	0.090	0.089	0.091	0.090	0.090	0.089	92.6	91.8	92.4	92.4	92.4	91.9
	L	β_1	0.513	0.142	0.131	0.129	0.131	0.130	0.131	0.129	93.6	93.1	93.8	93.5	93.1	93.0
		β_2	0.513	0.071	0.066	0.065	0.067	0.067	0.067	0.066	92.6	92.3	93.0	93.1	93.2	92.5
	G	β_1	0.488	0.332	0.342	0.340	0.343	0.340	0.337	0.336	96.3	96.4	96.5	96.3	95.7	95.8
		β_2	0.520	0.182	0.173	0.172	0.174	0.172	0.182	0.181	93.8	93.7	94.0	93.7	95.8	95.7
0.3	N	β_1	0.504	0.186	0.173	0.172	0.173	0.172	0.174	0.173	94.1	93.8	93.2	93.6	93.8	93.8
		β_2	0.512	0.093	0.089	0.087	0.089	0.089	0.088	0.087	93.0	93.0	93.9	93.3	92.8	92.6
	L	β_1	0.509	0.135	0.128	0.127	0.129	0.128	0.128	0.127	94.3	94.0	94.3	94.4	94.0	94.0
		β_2	0.510	0.069	0.065	0.064	0.066	0.065	0.065	0.065	93.5	93.6	94.0	94.3	94.3	93.8
	G	β_1	0.517	0.352	0.339	0.338	0.340	0.339	0.337	0.335	95.0	95.2	95.3	94.9	94.7	95.0
		β_2	0.508	0.168	0.170	0.168	0.170	0.169	0.177	0.176	94.3	94.1	94.6	94.0	95.2	95.2
0.6	N	β_1	0.514	0.177	0.172	0.171	0.173	0.172	0.174	0.173	94.6	94.7	94.0	94.6	94.9	94.8
		β_2	0.509	0.095	0.088	0.087	0.088	0.088	0.086	0.086	92.2	92.2	92.6	92.6	91.6	92.0
	L	β_1	0.514	0.134	0.128	0.126	0.128	0.127	0.128	0.126	94.6	94.8	95.1	94.9	95.3	94.8
		β_2	0.507	0.068	0.065	0.064	0.066	0.065	0.065	0.064	92.9	92.1	93.0	93.0	92.7	92.7
	G	β_1	0.532	0.351	0.341	0.340	0.342	0.340	0.339	0.337	95.5	95.6	95.5	95.7	95.0	95.2
		β_2	0.499	0.175	0.171	0.170	0.171	0.170	0.179	0.177	94.1	94.1	93.9	93.9	95.3	95.0

Table 7: Summary of simulation 2 with $n = 1000$ and two strata. Results are based on 1000 replications for censoring rate 90%. The bootstrapping size is 500 for each replication. PE is the average of point estimates; ESE is the empirical standard deviation of the parameter estimates; ASE is the average of the standard error of the estimator; CP is the coverage percentage of 95% confidence interval.

					ASE						CP(%)					
					IS		ZL		SH		IS		ZL		SH	
τ	dist	β	PE	ESE	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF	MB	CF
0	N	β_1	0.989	0.190	0.194	0.182	0.195	0.183	0.189	0.177	95.1	92.9	94.2	92.8	94.4	92.5
		β_2	1.012	0.136	0.127	0.124	0.129	0.126	0.135	0.132	93.5	92.6	93.2	92.7	94.0	93.8
	L	β_1	0.987	0.149	0.151	0.142	0.151	0.142	0.146	0.137	94.4	93.3	94.8	93.5	93.5	92.1
		β_2	1.005	0.103	0.101	0.099	0.102	0.100	0.108	0.106	93.6	92.9	93.3	93.2	94.0	93.8
	G	β_1	0.997	0.203	0.219	0.205	0.219	0.205	0.218	0.203	96.3	95.6	96.5	95.5	96.4	95.4
		β_2	1.009	0.133	0.128	0.126	0.128	0.127	0.131	0.130	92.4	92.6	92.5	92.5	93.5	93.2
0.3	N	β_1	0.986	0.161	0.161	0.152	0.161	0.153	0.156	0.147	93.7	91.8	93.4	92.4	92.2	90.8
		β_2	1.002	0.115	0.112	0.110	0.114	0.111	0.119	0.117	92.2	91.6	92.6	91.9	93.1	92.7
	L	β_1	0.995	0.127	0.133	0.126	0.133	0.126	0.128	0.121	95.3	94.1	95.2	94.2	94.1	92.5
		β_2	1.012	0.098	0.094	0.093	0.095	0.094	0.102	0.100	93.7	93.2	94.0	93.8	95.6	95.1
	G	β_1	1.000	0.149	0.164	0.156	0.164	0.156	0.162	0.154	95.7	94.9	95.6	94.7	95.3	94.5
		β_2	1.006	0.110	0.113	0.112	0.113	0.112	0.115	0.115	94.9	95.2	95.6	95.1	95.8	95.3
0.6	N	β_1	0.993	0.143	0.145	0.138	0.144	0.138	0.139	0.133	93.1	91.8	93.1	91.9	92.0	91.4
		β_2	1.009	0.115	0.108	0.106	0.109	0.107	0.114	0.112	92.9	92.4	93.1	92.7	93.9	93.7
	L	β_1	0.989	0.123	0.121	0.116	0.122	0.117	0.117	0.111	93.9	92.7	94.4	93.1	92.6	91.5
		β_2	1.006	0.091	0.091	0.089	0.092	0.090	0.097	0.096	94.5	94.4	95.1	94.5	96.8	96.4
	G	β_1	0.993	0.122	0.123	0.118	0.123	0.118	0.121	0.116	93.9	93.3	93.9	93.5	94.0	92.9
		β_2	1.003	0.105	0.104	0.103	0.104	0.103	0.106	0.106	95.0	94.5	94.5	94.5	94.9	94.8

Table 8: Summary of regression coefficient estimates and standard errors for the RCT study. In addition to the proposed sandwich variance estimators, we also included the multiplier bootstrap approach (MB) for comparison purpose. Timing result (in seconds) are presented in parentheses.

	PE	SE						
		MB	IS		ZL		SH	
			MB	CF	MB	CF	MB	CF
MOLAR	2.265	1.053	1.283	1.245	1.169	1.254	1.093	1.273
ROOT	−0.426	0.647	0.633	0.606	0.620	0.579	0.772	0.713
PC1	1.775	0.903	0.905	0.879	0.917	0.888	0.598	0.561
PC2	2.816	0.745	0.723	0.711	0.674	0.673	0.539	0.535
PCABUT	1.355	1.105	1.001	1.017	1.023	1.004	1.112	1.129
POCKET	−0.251	0.207	0.188	0.181	0.174	0.174	0.449	0.503
MOLAR:ROOT	−2.231	1.023	1.028	1.164	1.072	1.067	1.178	1.155
(Timing)		(1564.8)	(9.6)	(4.1)	(4.9)	(4.7)	(27.0)	(26.6)

Chapter 4

Efficient Rank-Based Approach with General Weight Functions

4.1 Introduction

In a study involving the observance of stressful event duration by Harnish et al. (2000),young adults in the greater Boston metropolitan area were interviewed and screened for major stressor. Treating the stressor duration as responses, researchers were interested in how coping strategies on stress affect the responses. Harnish et al. (2000) found that avoidance and active cognitive coping were linked with longer duration while active behavioral coping was linked with shorter duration.

Although the Gehan estimator is a preferred estimator for AFT model, it is desirable to explore the rank-based estimator with general weights for several reasons. First, selecting a proper weight can improve efficiency depending on the censoring distribution (Tsiatis, 1990). For example, Prentice (1978) shows log-rank is most efficient when censoring rate is low. Furthermore, the Prentice-Wilcoxon weight is more efficient when

censoring distribution is logistic (Gill, 1980). Second, the variance of the rank-based estimator is minimized if the limit of the weight is proportional to $\dot{\lambda}(t)/\lambda(t)$ where $\lambda(\cdot)$ is the common hazard function of the error terms and $\dot{\lambda}(t) = \mathrm{d}\lambda(t)/\mathrm{d}t$ (Tsiatis, 1990). Despite these observations, the rank-based estimator with general weights has not been widely used in practice mainly because of the lack of efficient and reliable computational methods. Recently, Jin et al. (2003) proposed an iterative method to approximate a class of general weight estimating equation around the true value of the regression parameters using the Gehan estimator as initial value. Within each iteration, an estimate is obtained by minimizing an convex objective function with a fixed weight. These minimisers of the objective function are obtained by the linear programming algorithm aforementioned, thus the computation time multiplies.

A more computing efficient approach is the induced smoothing procedure of Brown and Wang (2005). This method relies on taking expectations of the estimating equation with respect to a continuous noise. The resulting estimating equation is continuously differentiable with respect to the parameter and can be solved with standard numerical algorithms. Recently, some progress has been made on the asymptotic properties of this induced smoothing procedure in the context of rank-based approach with Gehan weight. For example, Brown and Wang (2007) applied induced smoothing to Gehan rank-based estimating equation for univariate failure times. Johnson and Strawderman (2009) and Wang and Fu (2011) applied the induced smoothing approach for clustered failure times.

Often all survival outcomes and covariates are observed on all the subjects. However,

in some special cases the exposures of interest are observed based only on a sample of the cohort. This is done because when a big percentage of outcomes are censored due to rare disease. Chiou et al. (2013a) extended the induced smoothing rank-based estimation to case-cohort study and proposed a class of fast and accurate variance estimators. Their method was further extended to general stratified sampling by Chiou et al. (2013c).

We investigate the induced smoothing method for the rank-based estimating equation with general weights. The original rank-based estimating equations for AFT models are presented in Section 4.2. Section 4.3 provides methods for solving induced smoothing estimating equations. The asymptotic variance estimations are proposed in Section 4.4. Extension to case-cohort studies is discussed in Section 4.5. A large scale simulation study on performance of the estimators is reported in Section 4.6. The methods are applied to the Stressful Experiences Study and a National Wilm's Tumor Study in Section 4.7. A discussion concludes in Section 4.8.

4.2 Rank-Based Estimation with Gehan Weight

For $i = 1, \ldots, n$, let T_i, C_i and X_i denote the log-transformed failure time, log-transformed censoring time and $p \times 1$ covariate vector respectively. It is assumed that for subject i, T_i and C_i are independent conditional on X_i. A semiparametric AFT model has the form

$$T_i = X_i^\top \beta + \epsilon_i, i = 1, \ldots, n,$$

where β is a $p \times 1$ regression parameters and ϵ_i's are independent and identically distributed but unspecified random variables. The error terms, ϵ_i, are assumed to be independent of X_i. In the presence of censoring, the observed data consist of $\{Y_i, \Delta_i, X_i\}$ for $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$, $\Delta_i = I[T_i < C_i]$, and $I[\cdot]$ is the indicator function.

Under the regularity conditions of Ying (1993), the regression parameters can be estimated from rank-based weighted estimating equation

$$U_{n,\varphi}(\beta) = \sum_{i=1}^{n} \Delta_i \varphi_i(\beta) \left[X_i - \frac{\sum_{j=1}^{n} X_j I[e_j(\beta) \geq e_i(\beta)]}{\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]} \right], \tag{4.1}$$

where $e_i(\beta) = Y_i - X_i^\top \beta$ and $\varphi_i(\beta)$ is a possibly data-dependent nonnegative weight function with values between 0 and 1. Let $\hat{F}_{e_i(\beta)}$ be an estimated cumulative distribution function at $e_i(\beta)$ given Δ_i. For a fixed $\rho \geq 0$, the most commonly choices of $\varphi_i(\beta)$ including 1, $\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]$, $1 - \hat{F}_{e_i(\beta)}$ and $(1 - \hat{F}_{e_i(\beta)})^\rho$ correspond to the log-rank (Prentice, 1978), Gehan (Gehan, 1965), Prentice-Wilcoxon (Prentice, 1978) and the more general G^ρ class (Harrington and Fleming, 1982), respectively. For a given function, $\varphi_i(\beta)$, solving $U_{n,\varphi}(\beta) = 0$ for β yields a consistent estimator, $\hat{\beta}_{n,\varphi}$, and the random vector $n^{1/2}(\hat{\beta}_{n,\varphi} - \beta_0)$ is asymptotically normal for any positive monotone function $\varphi_i(\beta)$ (Jin et al., 2003; Tsiatis, 1990; Ying, 1993). Despite the theoretical advances, the solution to $U_{n,\varphi}(\beta) = 0$ often consists multiple points due to the fact that $U_{n,\varphi}(\beta)$ is a step function. Although the problem might be solved by a linear programming approach (Jin et al., 2003), the computation is intensive for large sample size. Furthermore the

covariance matrices of the estimators can be problematic. These difficulties can be reduced with an efficient induced smoothing method.

Brown and Wang (2005) proposed an induced smoothing method that has a strong computational feature. Let Z be a $p \times 1$ standard normal random vector and Γ_n be some $p \times p$ matrix such that $\Gamma_n^2 = \Sigma_n$ for a symmetric positive definite matrix Σ_n which will eventually converge to Σ, the asymptotic covariance matrix of $n^{1/2}(\hat{\beta}_{n,\varphi} - \beta_0)$. Applying the induced smoothing method, the induced smoothing estimating equations are obtained by replacing $U_{n,\varphi}(\beta)$ with $E_Z[U_{n,\varphi}(\beta + n^{-1/2}\Gamma_n Z)]$, where the expectation is taken with respect to Z. Some common initial choices for Σ_n are I_p and $XX^\top/n$. The resulting estimating equations are continuously differentiable with respect to β and can be solved with standard numerical methods.

For simplicity purpose, the most common choice of $\varphi_i(\beta)$ in (4.1) is the Gehan's weight function which equals the numbers of subjects at risk, or $\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]$. In this case, the relevant induced smoothing expression estimating equation is

$$\tilde{U}_{n,G}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \left(X_i - X_j\right) \Phi\left[\frac{e_j(\beta) - e_i(\beta)}{r_{ij}}\right], \tag{4.2}$$

where $r_{ij}^2 = (X_i - X_j)^\top \Sigma_n (X_i - X_j)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The solution to (4.2), denoted by $\tilde{\beta}_{n,G}$, is consistent to β_0 and has the same asymptotic distribution as the solution to the nonsmooth version (Brown and

Wang, 2007; Johnson and Strawderman, 2009). Johnson and Strawderman (2009) obtained $\tilde{\beta}_{n,G}$ by optimizing the corresponding objective function. Alternatively, since equation (4.2) is continuously differentiable with respect to β, it can be directly solved with standard numerical algorithms. For example, we used a nonlinear equation solver in R package `BB` (Varadhan and Gilbert, 2009). Although the Gehan estimator is easy to obtain, it has frequently been criticized for its inefficiency. To counteract this drawback, we propose rank estimating equations with various weights in the following sections.

4.3 Induced Smoothing Method with General Weights

The log-rank estimating equation can be obtained by setting $\varphi_i(\beta) = 1$ in (4.1) with the form

$$U_{n,L}(\beta) = \sum_{i=1}^{n} \Delta_i \left[X_i - \frac{\sum_{j=1}^{n} X_j I[e_j(\beta) \geq e_i(\beta)]}{\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]} \right].$$

Deriving the smoothing log-rank estimating equation is challenging because it involves solving expectation of ratio of the summations. We propose to obtain the log-rank estimator by solving an asymptotically equivalence estimating equation:

$$\tilde{U}_{n,L}(\beta) = \sum_{i=1}^{n} \Delta_i \left[X_i - \frac{\sum_{j=1}^{n} X_j \Phi\left(\kappa_{ij}(\beta)\right)}{\sum_{j=1}^{n} \Phi\left(\kappa_{ij}(\beta)\right)} \right], \tag{4.3}$$

where $\kappa_{ij}(\beta) = [e_j(\beta) - e_i(\beta)]/r_{ij}$. The smoothed log-rank estimator, $\tilde{\beta}_{n,L}$, is found by solving $\tilde{U}_{n,L}(\beta) = 0$. The asymptotically equivalence between equation (4.3) and

$E[U_{n,L}(\beta + n^{-1/2}\Gamma_n Z)]$ is verified with Theorem 3.

Theorem 3. *Under the conditions B1–B2 in the Appendix, $\tilde{U}_{n,L}(\beta)$ is asymptotically equivalent to $E[U_{n,L}(\beta + n^{-1/2}\Gamma_n Z)]$.*

Appendix A.3 gives a sketch of the proof.

When the weight function, $\varphi_i(\beta)$ depends on β, the induced smoothing equation is much more difficult to obtain. We exploit an iteratively reweighed strategy by Jin et al. (2003) to obtain an estimator for β_0. Given an initial estimator b_n of β, define

$$\tilde{U}_{n,\varphi}(b,\beta) = \sum_{i=1}^{n} \Delta_i \varphi_i(b) \left[X_i - \frac{\sum_{j=1}^{n} X_j \Phi\left(\kappa_{ij}(\beta)\right)}{\sum_{j=1}^{n} \Phi\left(\kappa_{ij}(\beta)\right)} \right]. \tag{4.4}$$

The proposed procedure consists of the following steps

1. Obtain an initial estimate $\tilde{\beta}_{n,\varphi}^{(0)} = b_n$ of β and initialize with $m = 1$.

2. Update $\tilde{\beta}_{n,\varphi}^{(m)}$ by solving $\tilde{U}_{n,\varphi}(\tilde{\beta}_{n,\varphi}^{(m-1)}, \tilde{\beta}_{n,\varphi}^{(m)}) = 0$.

3. Increase m by one and repeat 2. until $|\tilde{\beta}_{n,\varphi}^{(m)} - \tilde{\beta}_{n,\varphi}^{(m-1)}| < \tau$, for a fixed tolerance, τ.

One possible initial estimator, b_n, is the easy-to-compute Gehan's estimator, $\tilde{\beta}_{n,G}$. Once converge, the estimator $\tilde{\beta}_{n,\varphi}$ is consistent. Moreover, the large-sample distribution of $\tilde{\beta}_{n,\varphi}$ can be approximated by a multivariate normal distribution (Jin et al., 2003). In all the simulation and real data we have tested, $\tilde{\beta}_{n,\varphi}^{(m)}$ always converged within 10 steps.

4.4 Sandwich Variance Estimation

To estimate the variance matrix, one could use a novel multiplier bootstrap approach proposed by Jin et al. (2003). Let $\eta_i, i = 1, \ldots, n$, be independent and identically distributed positive random variables with $\mathrm{E}(\eta_i) = \mathrm{Var}(\eta_i) = 1$. Define a perturbed estimating equation

$$\tilde{U}^*_{n,\varphi}(b, \beta) = \sum_{i=1}^{n} \eta_i \Delta_i \varphi_i(b) \left[X_i - \frac{\sum_{j=1}^{n} \eta_j X_j \Phi\left(\kappa_{ij}(\beta)\right)}{\sum_{j=1}^{n} \eta_j \Phi\left(\kappa_{ij}(\beta)\right)} \right]. \tag{4.5}$$

For a realization of $(\eta_1, \ldots, \eta_n)$, and an initial value, b, the solution to $\tilde{U}^*_{n,\varphi}(b, \beta) = 0$ provides one draw of $\tilde{\beta}_{n,\varphi}$ from its asymptotic distribution. By repeating this process a large number B times, the variance matrix of $\tilde{\beta}_{n,\varphi}$ can be estimated directly by the sampling variance matrix of the bootstrap sample of $\tilde{\beta}_{n,\varphi}$. This method requires to solve (4.5) a large number times, thus being very time consuming.

An efficient variance estimation is to decompose the covariance matrix into sandwich form and estimate the necessary parts separately. Base on the performance and speed, Chiou et al. (2013a) recommended two efficient sandwich variance estimators, induced smoothing approach and Zeng and Lin (2008) approach. We extend these two methods to adopt general weight function.

Under some regularity conditions (Zeng and Lin, 2008), uniformly in a neighborhood

of β_0, equation (4.1) can be expressed as

$$n^{-1/2}U_{n,\varphi}(\beta) = n^{-1/2}\sum_{i=1}^{n} S_i(\beta_0) + An^{1/2}(\beta - \beta_0) + o_p(1 + n^{1/2}\|\beta - \beta_0\|),$$

where $S_i(\beta_0)$ is a zero-mean random vector, and A is asymptotic slope matrix of $n^{-1/2}\tilde{U}_{n,\varphi}(\beta_0)$. The asymptotic variance matrix of $n^{1/2}(\tilde{\beta}_{n,\varphi} - \beta_0)$ is $n\Sigma = nA^{-1}V(A^{-1})^\top$, where V is the variance of $n^{-1/2}\sum_{i=1}^{n} S_i(\beta_0)$. Estimation of Σ involves estimating V and A.

To estimate V, we apply an empirical approach with aids from multiplier bootstrap. For a realization of $(\eta_1, \ldots, \eta_n)$, evaluation of $\tilde{U}^*_{n,\varphi}(\tilde{\beta}_{n,\varphi}, \tilde{\beta}_{n,\varphi})$ provides a bootstrap replicate of $\tilde{U}^*_{n,\varphi}(\beta)$. With large sample replicates, we estimate V by the sample variance of the bootstrap sample of $\tilde{U}^*_{n,\varphi}(\beta)$. The bootstrap here is much less demanding than the aforementioned full multiplier bootstrap, because it only involves evaluations of estimating equations instead of solving them to obtain each bootstrap replicate.

The two versions of the sandwich variance estimator depending on how A is estimated. The first method is a direct extension from the induced smoothing approach. With $\tilde{U}_{n,\varphi}(\beta)$, the slope matrix A can be estimated directly by

$$A_n = \frac{1}{n}\frac{\mathrm{d}}{\mathrm{d}\beta}\tilde{U}_{n,\varphi}(\beta).$$

For the log-rank estimating equation, the closed form of A_n exists and can be obtained

through fundamental calculus. The slope matrix for the general weight estimating equation can be derived similarly but with great complication.

The second method is motivated from Zeng and Lin (2008). This method serves as an empirical version of the induced smoothing approach because it estimated A via a resampling procedure instead of taking the derivatives of a smooth function. For $r = 1, \ldots, R$, let Z_r be a p-dimensional standard normal random vector. Denote A_{nj} to be the jth row of A_n and $\tilde{U}_{nj,\varphi}(\cdot)$ to be the jth component of $\tilde{U}_{n,\varphi}(\cdot)$ for $j = 1, \ldots, p$. We obtain A_{nj} by the least squares estimate of the regression coefficients when regressing $n^{-1/2}\tilde{U}_{nj,\varphi}(\tilde{\beta}_{n,\varphi} + n^{-1/2}Z_b)$ on Z_b, $i = 1, \ldots, n$. The variance estimator also has the sandwich form $\hat{\Sigma}_n = A_n^{-1}V_n(A_n^{-1})^\top$.

4.5 Incorporating Sampling Weight

We consider the case where the data $\{Y_i, \Delta_i, X_i\}$ can be missing by design as in case-cohort studies (Prentice, 1986). A case-cohort design is known to be cost-effective when a large percentage of outcomes are censored possibly because the events of interest occur rarely, or covariates are expensive to measure. In a case-cohort design, only members in the randomly selected case-cohort sample is included. The case-cohort sample is assumed to be taken via a stratified simple random sampling from S mutually exclusive strata form by the original full cohort. Suppose $\tilde{n}_s$ is the numbers of subjects sampled from the sth stratum consists n_s subjects. Since the study sample is not complete, statistical

methods which do not account for this absence in covariates could result in biased estimates. One typical method employed to adjust for biases is weighting a complete observation by the inverse of the inclusion probability. Let ψ_{is} be the strata indicator indicator ($\psi_{is} = 1$ if the ith subject is in the sth stratum and $\psi_{is} = 0$ otherwise) and ξ_i be the sampling indicator ($\xi_i = 1$ if the ith subject is sampled and $\xi_i = 0$ otherwise). Then the case-cohort estimator can be estimated from the weighted version of (4.4)

$$\tilde{U}^c_{n,\varphi}(b,\beta) = \sum_{i=1}^{n} h_i \Delta_i \varphi_i(b) \left[X_i - \frac{\sum_{j=1}^{n} h_j X_j \Phi\left(\kappa_{ij}(\beta)\right)}{\sum_{j=1}^{n} h_j \Phi\left(\kappa_{ij}(\beta)\right)} \right] \tag{4.6}$$

where $h_i = \sum_{s=1}^{S} \xi_i \psi_{is} / p_{n,s}$ and $p_{n,s} = \tilde{n}_s / n_s$ is the inclusion probability for the sth stratum. If we sample all the subjects, i.e., $\xi_i = \psi_i = p_{n,s} = 1$ for $i = 1, \cdots, n$ and $s = 1, \cdots, S$, then we have the full cohort data and (4.6) reduces to (4.4).

The variance of the case-cohort estimator can be estimated from the aforementioned bootstrap approach with a weighted adjusted perturbed estimating equation

$$\tilde{U}^{c*}_{n,\varphi}(b,\beta) = \sum_{i=1}^{n} \eta_i h_i \Delta_i \varphi_i(b) \left[X_i - \frac{\sum_{j=1}^{n} \eta_j h_j X_j \Phi\left(\kappa_{ij}(\beta)\right)}{\sum_{j=1}^{n} \eta_j h_j \Phi\left(\kappa_{ij}(\beta)\right)} \right].$$

Alternatively, to apply Zeng and Lin's approach, one should obtain A_{nj} by the least squares estimate of the regression coefficients when regressing $n^{-1/2}\tilde{U}^c_{nj,\varphi}(\tilde{\beta}_{n,\varphi} + n^{-1/2} Z_b)$ on Z_b, $i = 1, \cdots, n$, where $\tilde{U}^c_{nj,\varphi}(\cdot)$ is the jth component of $\tilde{U}^c_{n,\varphi}(\cdot)$. The sandwich variance estimator then has the form $\hat{\Sigma}_n = A_n^{-1} V_n (A_n^{-1})^\top$ with V_n obtained by the

sample variance of a large realization of $\tilde{U}^{c*}_{n,\varphi}(\tilde{\beta}_{n,\varphi}, \tilde{\beta}_{n,\varphi})$.

4.6 Simulation

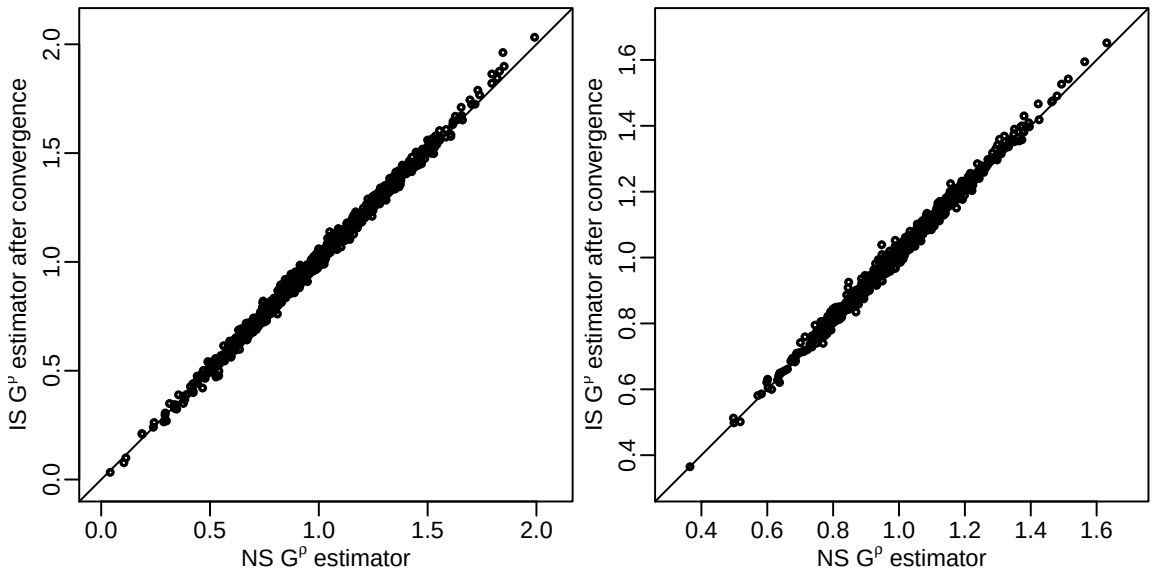
Extensive simulation studies were conducted to assess the large-sample properties of the proposed methods. Failure time T was generated from AFT model

$$\log(T) = 2 + X_1 + X_2 + X_3 + \epsilon,$$

where X_1 was Bernoulli with rate 0.5, X_2 and X_3 were uncorrelated standard normal variables. The error term, ϵ, follows standard normal, standard gumbel or standard logistic, abbreviated by N, G and L, respectively. The censoring times were generated uniform distributions $[0, c]$ where c was adjusted to yield the desirable censoring rates, C_p. Ordinary censoring rates, $C_p = 0.25$ and $C_p = 0.50$, are considered using full cohort analysis. Rank based estimating equation based on general weight functions, log-rank (LR), Prentice-Wilcoxon (PW) and the general G^ρ class (G^ρ) developed in Section 4.3, were used to fit the model. When the general G^ρ class was used, ρ was chosen to be 1/3; one over the number of the covariates. For each weight, two point estimation were considered; with non-smoothed estimating equation that solves (4.1) directly (NS) and with the proposed iterative induced smoothing method (IS). The variance is then estimated by Zeng and Lin's sandwich estimator with empirical approach. Each entry in the tables is based on 1000 replicates

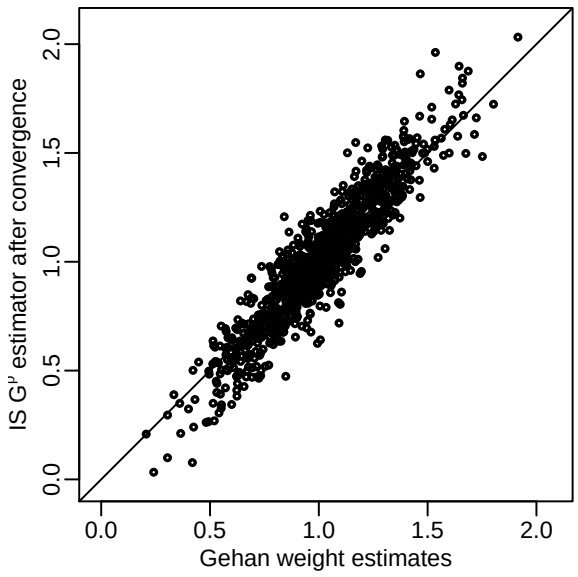
A comparison between NS estimator and IS estimator is summarized in table 9 for $C_p = 0.50$, $n = 100$ and Gumbel error distribution. Note that the intercept term is not considered, because rank is invariant to location shift and the intercept term cannot be estimated. Both the NS and IS estimators appear to be virtually unbiased. Furthermore, they agreed with each other closely on a 45 degree line as shown in figure 1(a) and figure 1(b). Their empirical standard errors and estimated standard errors also agreed with each other. The associated 95% confidence intervals based on the estimated standard errors had empirical coverage percentages reasonably close to the nominal level. The performance of the iterative induced smoothing method for full scale simulation are presented in table 10. The same observation remains and were invariant to the error distributions. Among all weights, PW estimator seems to provide the smallest standard errors and the best coverage percentage for all scenarios. For normal margin and logistic margin, standard errors estimated from the three weights are close. Under the Gumbel margin, the standard errors estimated from G^{ρ} estimator are consistently between that of LR estimator and PW estimator. Figure 1(c) and figure 1(d) displays the corresponding plots for the G^{ρ} estimator versus the Gehan estimator which was used as the initial estimator in the iterative procedure. The two estimates are noticeably different.

It is important to note the timing results summarized in table 11. Each entry represents the timing results in seconds averaged from 1000 replicates on a 2GHz linux machine. The proposed IS estimator provided faster estimation for all cases. In particular, the IS estimator is up to 23 times as fast as the NS estimator for PW estimator (with

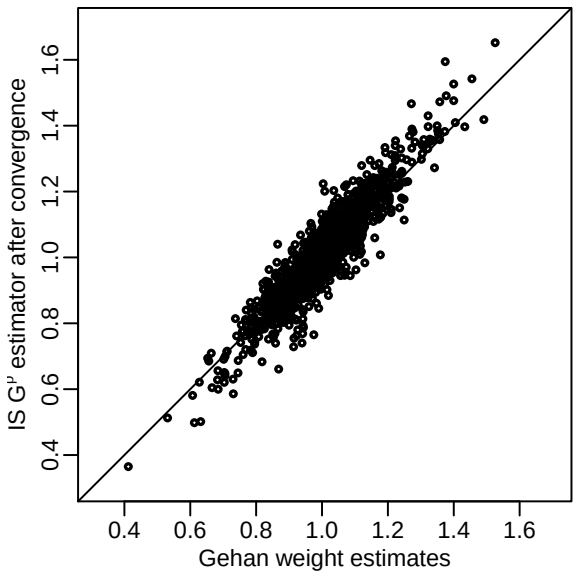


(a) Comparisons between β_1.

(b) Comparisons between β_2.



(c) Comparisons between β_1.



(d) Comparisons between β_2.

Figure 1: Comparisons of different estimates of β_1 and β_2 under 50% censoring rate, $n = 100$ and Gumbel error distribution. (a), (b): Nonsmooth estimator versus smoothed estimator for G^ρ estimator; (c), (d): Gehan estimator versus G^ρ estimator after convergence.

$n = 400$, logis margin) and is up to 69 times as fast for the G^{ρ} estimator (with $n = 100$, logis margin). Since the performance of both point estimator and variance estimator are the same, the IS approaches are obviously preferred over the NS approaches.

In the second simulation we consider $C_p = 0.9$ representing a rare disease. The failure time was generated from the same model but the full cohort size was set to be 1500. The subjects were categorized into two strata; stratum 1 formed by cases ($\Delta_i = 1$) and stratum 2 formed by controls ($\Delta_i = 0$). A stratified sample was obtained by simple random sampling all the subjects from the cases and then sample the number of subjects from the controls so that the average case-cohort size was 300. The average case-cohort weight for stratum 1 and 2 are 1 and 1.11, respectively. Induced smoothing rank-based estimator with case-cohort weight, h_i, constructed by inversing the inclusion probability was used for point estimation and Zeng and Lin's approach was used for variance estimation. Table 12 summarizes the results for simulation 2. All point estimates appear to be virtually unbiased. The empirical standard errors and the average standard error are closely agree with each other. The coverage percentages are reasonably close to the nominal level.

4.7 Application

4.7.1 Stressful Experiences Study

The Stressful Experiences Study was first studied by Harnish et al. (2000) based on data obtained from a longitudinal study of the stress process during the adolescent and early adult periods. Interviews were conducted at the location of the respondent's preference (primarily in the home) by professional staff from the Center for Survey Research, University of Massachusetts-Boston. Interviews consisted six waves from 1988 to 1999 with an initial sample of 1208 high school students in grades 9, 10, and 11 residing in three communities in the greater Boston metropolitan area. The response rate in the initial wave of the study was 77% and retention rates exceeded 85% in each successive interview. Because data on depression event duration and coping were not collected in the first four waves of data collection, data used for the current study were drawn from Wave 5. In Wave 5, interviews were conducted with 829 members of the sample, constituting 69% of those in the original high school sample. The Wave 5 sample contained more women ($\hat{60}$) than men and was predominantly Caucasian (94%). These demographics were consistent with the original sample.

All respondents were questioned about the most difficult stressor event in the past year. Out of the 829 members interviewed, 451 (44%) members still had an on going stressful event at the time of last interview and thus were censored. Due to missing data on relationship responds, 22 respondents were dropped from the analysis, resulting

in an effective sample size of 807. Preliminary exploratory analyses indicated that the distribution of the depression duration contained 68 (8%) outliers. As Harnish et al. (2000) suggested, durations was truncated to 27 months in order to reduce the degree of influence due to very large duration values. After truncating outliers, the censoring rate increased to 58%. To examine whether duration varies by event type, controlling for gender (1 = female, 0 = male) and race (1 = Caucasian, 0 = others) of the respondent, event severity on the respondent's actions was also taken into consideration.

The measure of the severity was determined on the basis of descriptive, contextual information about the stressors that respondents detail in their own words when asked in an open-ended fashion about major stresses or difficulties (Brown and Harris, 1978). The severity level was then quantified into a four point ordinal scale with 4 being most severe. A similar four point ordinal scale (0 = not at all, 1 = a little, 2 = some and 3 = a lot) was used to measure the frequency each of the six types of coping strategies was used during the course of the depression experience. The six types of coping strategies include avoidance (avoid), positive reappraisal (reapp), religion (relig), active cognitive (actcog), active behavioral (actbhvr), and social support (supp).

We first fit the data with smoothed estimating equation with Gehan's weight (GE). Using GE estimator as the initial value, we then fit with the general G^ρ weight with $\rho = 0, 1/9$ and 1 denoted by LR, G^ρ and PW respectively. These estimator are solved with our iterative procedure. Once the point estimator is obtained, Zeng and Lin's sandwich estimator with empirical approach was applied. The results are summarized in 13. All

the point estimators seem to agree closely. The coefficient for the avoidance coping and higher severity level are associated with longer stress duration for all estimators. However, only the GE estimator and the PW estimator suggested that female tend to have longer stress duration whereas the LR estimator and G^{ρ} estimator suggest gender to be insignificant. These results are consistent with that of Harnish et al. (2000).

4.7.2 National Wilm's Tumor Study

We demonstrate the performance of our proposed method with case-cohort extension to a National Wilm's Tumor Study (NWTS) (D'Angio et al., 1989; Green et al., 1998). The NWTS was first created in 1969 to address the need to study and compare treatments in different period of time. The study subjects in this example are from one of the two study groups, NWTSG-3 and NWTSG-4, where the period of study is between May 1979 to August 1995. The interest of the study was to assess the relationship between the tumor histology and time to tumor relapse. Tumor histology is classified depending on either a given histological diagnosis was favored. The central histological diagnosis was made by an individual pathologist at the central pathology center, which was believed to be more accurate than a local diagnosis yet more expensive to measure and required more efforts to obtain. A staging system employed by the NWTS Group was used as indicator of the tumor spread. This staging system consists of four stages, Stage I, Stage II, Stage III and Stage IV, with Stage IV being the latest and severest stage. Although all the measurements were available for full cohort, we assume, in this example, that

these measurements were taken only for the subjects in the case-cohort sample. The case-cohort version of the data was analyzed with Cox models (Kulich and Lin, 2004) and additive hazards models (Kulich and Lin, 2000), respectively.

The dataset is available in survival package (Therneau, 2012). There were a total of 4028 subjects in the full cohort. Among them, 571 belong to the cases stratum who experienced the relapse and 3457 belong to the control stratum who did not experience the relapse. The censoring rate was about 86% reflecting the fact that Wilm's tumor is a rare kidney cancer in young children. The covariates included central histology measurement (1 = favorable, 0 = unfavorable), age (measure in year) at diagnosis, three tumor stages indicators (Stage I as reference) and a study group indicator (NWTSG-3 as reference). The case-cohort version of the data had 1154 patients selected as case-cohort sample, including 571 cases and 583 controls. Thus, the case-cohort weights are 1 and 5.930 for the cases stratum and controls stratum respectively. In the analysis, we used LR estimator, PW estimator and G^{ρ} estimator with $\rho = 1/6$.

The results of the National Wilm's Tumor Study is presented in Table 14. All estimators lead to the same conclusion, the coefficients of central histological diagnosis, age and all stages are significantly different from zero. For comparison purpose, we also analyzed the full cohort data with the same covariates and reported the results in Table 14. Point estimates are close to these in case-cohort analysis, with their standard errors taken into consideration. All the standard errors decrease and timing increases compared to the case-cohort analyses, which is expected as full information became

available for all covariates.

4.8 Discussion

The rank-based approach in solving AFT model has seldom been applied with general weight functions mainly due to lack of efficient and reliable computing algorithm. With the recent developed induced smoothing approach, we have proposed an iterative procedure in estimation with general weight function. The point estimators from our procedure were shown to be consistent and the corresponding variance estimators were found to provide good approximation to the true variation and are computationally efficient. All the methods are made public available in an R package aftgee (Chiou et al., 2012a, 2013b).

It might be worthwhile to consider extensions of weights functions mentioned. For example, Buyske et al. (2000) proposed an extension to the general G^ρ weight that is shown to be more efficient when the censoring rate is high. When the censoring rate is high, the general G^ρ class of weight does not have a good range of flexibility. This is because the high censoring rate yields $(1 - \hat{F}_{e_i(\beta)})^\rho$ to be close to 1 for all $e_i(\beta)$ and $0 \leq \rho \leq 1$. Therefore, any member in the G^ρ family are essentially the same as the log-rank weight. Buyske et al. (2000) proposed to subtract $\hat{F}_{e_i(\beta)}$ by .min$(\hat{F}_{e_i(\beta)})$. Specifically, $\varphi_i(\beta) = [\hat{F}_{e_{\gamma,i}(\beta)} - \hat{F}_{e_i(\beta)}]^\rho$. Another extension is to consider optimal weight function. One possibility to construct an optimal weight is to perform a power study on

selecting ρ in the general G^{ρ} weight function.

Table 9: Summary of simulation1 on selected results to compare NS and IS under Gumbel error margin and $n = 100$. PE is the point estimator; ESE is the empirical standard error; ASE is the average of the standard errors of the estimator; CP is the coverage percentage.

		PE		ESE		ASE		CP(%)	
C_p	β	IS	NS	IS	NS	IS	NS	IS	NS
					Logrank				
0.25	β_1	0.992	0.981	0.299	0.292	0.287	0.285	93.9	94.2
	β_2	1.016	1.004	0.161	0.157	0.146	0.144	91.3	92.6
	β_3	1.015	1.002	0.159	0.157	0.147	0.144	92.2	92.4
0.50	β_1	1.023	1.007	0.335	0.324	0.301	0.294	90.9	92.2
	β_2	1.026	1.003	0.174	0.168	0.161	0.157	92.0	92.6
	β_3	1.022	1.002	0.177	0.173	0.161	0.156	91.1	90.4
					Prentice-Wilcoxon				
0.25	β_1	1.014	1.012	0.247	0.250	0.243	0.237	94.0	92.2
	β_2	1.000	0.996	0.131	0.132	0.126	0.123	93.0	91.1
	β_3	1.009	1.004	0.124	0.124	0.124	0.121	94.5	93.5
0.50	β_1	1.007	1.000	0.270	0.269	0.266	0.256	93.9	92.9
	β_2	1.009	1.000	0.144	0.144	0.142	0.137	93.2	93.3
	β_3	1.011	1.002	0.147	0.148	0.142	0.138	93.3	92.4
					General G^ρ with $\rho = 1/3$				
0.25	β_1	1.022	1.015	0.280	0.277	0.265	0.261	91.9	92.3
	β_2	1.003	0.996	0.149	0.148	0.136	0.133	91.6	91.1
	β_3	1.012	1.005	0.140	0.138	0.135	0.133	93.8	93.4
0.50	β_1	1.009	0.996	0.302	0.297	0.289	0.278	93.7	92.9
	β_2	1.015	1.001	0.160	0.158	0.153	0.148	93.2	93.5
	β_3	1.017	1.002	0.166	0.162	0.153	0.148	92.9	92.3

Table 10: Summary of simulation 1. PE is the point estimator; ESE is the empirical standard error; ASE is the average of the standard errors of the estimator; CP is the coverage percentage; MD is marginal distribution.

				PE			ESE			ASE			CP (%)		
MD	n	C_p	β	LR	PW	G^ρ	LR	PW	G^ρ	LR	PW	G^ρ	LR	PW	G
N	200	0.25	β_1	1.002	0.996	0.999	0.172	0.159	0.163	0.158	0.157	0.157	92.7	94.8	93.8
			β_2	1.006	1.003	1.004	0.091	0.083	0.086	0.083	0.083	0.082	92.0	94.1	92.9
			β_3	1.006	1.002	1.004	0.091	0.084	0.086	0.083	0.082	0.082	91.8	93.9	94.0
		0.50	β_1	1.006	1.000	1.002	0.191	0.176	0.182	0.186	0.181	0.183	93.7	94.5	95.1
			β_2	1.015	1.009	1.011	0.110	0.101	0.105	0.100	0.098	0.099	90.7	93.6	92.3
			β_3	1.021	1.012	1.016	0.104	0.098	0.100	0.101	0.098	0.099	92.4	94.7	93.5
	400	0.25	β_1	0.998	0.998	0.998	0.121	0.112	0.115	0.114	0.111	0.111	93.0	94.7	93.7
			β_2	1.003	1.001	1.002	0.061	0.056	0.058	0.059	0.058	0.058	93.1	95.5	94.3
			β_3	1.004	1.001	1.002	0.064	0.060	0.061	0.059	0.057	0.058	92.2	93.2	93.2
		0.50	β_1	1.015	1.011	1.013	0.140	0.130	0.134	0.132	0.127	0.128	93.6	94.6	94.3
			β_2	1.007	1.002	1.004	0.076	0.071	0.073	0.071	0.069	0.070	93.2	92.9	94.1
			β_3	1.011	1.006	1.008	0.076	0.072	0.073	0.072	0.069	0.070	92.0	93.8	93.6
G	200	0.25	β_1	0.996	0.996	0.996	0.215	0.168	0.191	0.205	0.169	0.187	93.3	95.0	94.5
			β_2	1.007	1.001	1.004	0.108	0.085	0.096	0.104	0.087	0.095	93.3	94.9	94.2
			β_3	1.002	0.999	1.000	0.111	0.086	0.098	0.104	0.087	0.095	93.5	95.6	95.2
		0.50	β_1	1.005	1.001	1.002	0.227	0.188	0.207	0.216	0.186	0.202	93.0	94.0	93.8
			β_2	1.009	1.002	1.005	0.122	0.101	0.112	0.115	0.099	0.107	91.9	93.0	92.2
			β_3	1.016	1.008	1.011	0.120	0.098	0.109	0.114	0.099	0.107	92.7	94.9	94.0
	400	0.25	β_1	1.006	1.004	1.005	0.155	0.125	0.139	0.146	0.120	0.133	93.5	94.1	94.2
			β_2	1.006	1.003	1.004	0.076	0.061	0.068	0.074	0.062	0.068	93.5	95.9	94.4
			β_3	1.004	1.002	1.003	0.079	0.063	0.071	0.074	0.062	0.068	93.0	94.6	94.5
		0.50	β_1	1.002	0.999	1.000	0.154	0.127	0.140	0.153	0.130	0.142	94.4	94.5	94.7
			β_2	1.004	1.001	1.003	0.085	0.072	0.078	0.081	0.070	0.076	93.4	94.8	94.5
			β_3	1.006	1.002	1.004	0.085	0.070	0.077	0.081	0.070	0.076	93.1	94.6	94.1
L	200	0.25	β_1	1.010	1.003	1.007	0.278	0.255	0.262	0.272	0.258	0.262	93.7	94.5	94.5
			β_2	1.012	1.008	1.009	0.145	0.133	0.137	0.139	0.133	0.135	91.8	94.3	93.1
			β_3	1.009	1.007	1.008	0.142	0.133	0.135	0.140	0.134	0.135	93.5	94.0	94.4
		0.50	β_1	1.004	1.004	1.003	0.316	0.301	0.306	0.301	0.290	0.293	93.9	93.6	93.7
			β_2	1.002	0.993	0.997	0.166	0.157	0.160	0.159	0.155	0.156	93.7	94.8	93.6
			β_3	1.017	1.009	1.013	0.172	0.159	0.164	0.159	0.155	0.155	91.8	93.7	93.1
	400	0.25	β_1	1.003	1.004	1.003	0.205	0.186	0.192	0.193	0.181	0.184	92.2	93.4	93.0
			β_2	1.002	0.999	1.000	0.107	0.097	0.100	0.099	0.093	0.095	92.4	93.8	93.4
			β_3	1.002	1.001	1.001	0.102	0.094	0.096	0.099	0.093	0.095	93.8	94.7	94.0
		0.50	β_1	0.987	0.987	0.987	0.215	0.204	0.207	0.213	0.206	0.208	94.5	95.0	94.8
			β_2	1.006	1.002	1.004	0.114	0.108	0.110	0.113	0.110	0.111	94.0	94.8	95.0
			β_3	1.005	1.004	1.004	0.119	0.112	0.114	0.114	0.110	0.111	93.4	94.5	93.9

Table 11: Timing results in seconds for point estimation.

		$n = 100$		$n = 200$		$n = 400$	
MD	C_p	IS	NS	IS	NS	IS	NS
		Log-rank					
N	25%	0.5	1.3	2.0	2.9	7.4	9.2
	50%	0.4	1.1	1.3	2.1	5.0	6.0
G	25%	0.5	1.2	1.9	2.7	7.1	8.8
	50%	0.4	1.1	1.3	2.0	5.1	5.9
L	25%	0.5	1.3	1.8	2.5	6.5	8.6
	50%	0.3	1.1	1.2	1.8	4.4	6.1
		Prentice-Wilcoxon					
N	25%	18.4	103.4	36.6	178.4	34.1	403.0
	50%	11.7	101.6	15.4	174.2	18.5	389.6
G	25%	17.3	103.5	33.6	175.4	33.6	407.8
	50%	10.7	101.5	12.7	169.5	17.3	390.9
L	25%	18.5	99.5	39.8	170.8	46.4	405.4
	50%	9.4	100.0	11.6	166.6	17.7	385.0
		General G^ρ weight					
N	25%	2.3	104.9	5.6	180.1	18.2	390.2
	50%	1.8	103.3	4.1	173.6	12.9	384.4
G	25%	1.9	103.7	5.7	174.0	18.5	394.6
	50%	1.5	103.5	3.9	169.6	12.6	379.5
L	25%	2.1	103.0	6.2	171.8	18.2	395.7
	50%	1.6	101.5	3.7	166.4	12.1	373.8

Table 12: Summary of simulation 2. Full cohort size = 1500, average case-cohort size = 300; $C_p = 0.9$; PE is the point estimator; ESE is the empirical standard error; ASE is the average of the standard errors of the estimator; CP is the coverage percentage.

		PE			ESE			ASE			CP (%)		
ME	β	LR	PW	G^ρ	LR	PW	G^ρ	LR	PW	G^ρ	LR	PW	G^ρ
N	β_1	0.998	1.002	1.000	0.220	0.196	0.209	0.195	0.182	0.190	90.6	93.3	92.1
	β_2	1.007	1.007	1.007	0.117	0.105	0.111	0.104	0.098	0.101	90.9	93.1	91.9
	β_3	1.009	1.010	1.009	0.109	0.099	0.104	0.104	0.098	0.101	93.5	94.6	93.9
G	β_1	1.007	1.006	1.006	0.225	0.188	0.209	0.203	0.178	0.192	91.7	93.6	93.0
	β_2	1.014	1.010	1.012	0.119	0.099	0.110	0.110	0.097	0.104	91.5	93.8	92.2
	β_3	1.010	1.007	1.008	0.120	0.100	0.111	0.110	0.097	0.104	92.0	93.6	93.0
L	β_1	1.019	1.022	1.020	0.302	0.283	0.294	0.284	0.276	0.280	93.1	95.0	94.2
	β_2	1.011	1.009	1.009	0.161	0.152	0.157	0.151	0.148	0.149	92.8	93.6	93.5
	β_3	1.012	1.012	1.011	0.163	0.155	0.159	0.151	0.148	0.150	92.2	92.6	92.0

Table 13: Stressful Experiences Study

	GE		LR		PW		G^ρ	
	PE	EST	PE	EST	PE	EST	PE	EST
gender	0.396	0.180	0.394	0.210	0.393	0.175	0.396	0.235
race	0.059	0.466	−0.079	0.554	0.002	0.468	−0.070	0.473
avoid	0.109	0.035	0.105	0.053	0.108	0.041	0.106	0.046
reapp	0.008	0.035	−0.005	0.047	0.003	0.038	−0.003	0.046
relig	0.022	0.023	0.015	0.031	0.020	0.026	0.016	0.028
actcog	0.015	0.050	0.034	0.060	0.023	0.049	0.032	0.054
actbhor	−0.027	0.042	−0.074	0.062	−0.045	0.051	−0.070	0.059
supp	−0.024	0.032	−0.039	0.052	−0.030	0.037	−0.038	0.038
severity	0.216	0.091	0.207	0.099	0.214	0.090	0.207	0.097

Table 14: National Wilm's Tumor Study

Effects	GE		LR		PW		G^ρ	
	PE	SE	PE	SE	PE	SE	PE	SE
Case-Cohort Analysis:								
(time)	(66.8)		(58.1)		(163.8)		(167.3)	
histol	−2.743	0.213	−3.705	0.238	−3.521	0.280	−3.672	0.269
age	−0.127	0.038	−0.143	0.054	−0.143	0.046	−0.144	0.051
stage2	−1.334	0.264	−1.578	0.344	−1.513	0.327	−1.565	0.306
stage3	−1.340	0.312	−1.405	0.350	−1.389	0.340	−1.403	0.403
stage4	−2.201	0.324	−3.070	0.416	−2.775	0.389	−3.011	0.376
study	−0.145	0.227	−0.269	0.322	−0.244	0.296	−0.265	0.319
Full-Cohort Analysis:								
(time)	(565.9)		(210.4)		(627.1)		(552.6)	
histol	−2.749	0.202	−3.758	0.162	−3.614	0.143	−3.731	0.160
age	−0.127	0.037	−0.177	0.039	−0.172	0.029	−0.176	0.038
stage2	−1.335	0.280	−1.466	0.233	−1.414	0.200	−1.458	0.238
stage3	−1.341	0.286	−1.808	0.251	−1.694	0.195	−1.789	0.253
stage4	−2.203	0.319	−2.627	0.294	−2.404	0.239	−2.584	0.281
study	−0.106	0.226	−0.361	0.197	−0.304	0.191	−0.350	0.214

Chapter 5

Multivariate Analysis with Generalized Estimating Equations

5.1 Introduction

Multivariate failure times arise frequently in biomedical research. For example, a diabetic retinopathy study assessed the efficacy of a laser treatment on decelerating vision loss, measured by time to blindness in the left eye and in the right eye from the same patient with diabetes (Diabetic Retinopathy Study Research Group, 1976). The failure times from the same patient are associated. The primary interest most often lies in the marginal covariate effects on the failure times, and exploiting the within-cluster dependence may lead to more efficient statistical inferences. For non-censored multivariate data, the generalized estimating equations (GEE) approach (Liang and Zeger, 1986) has become an important piece in statisticians' toolbox for marginal regression. For censored multivariate failure times, the marginal accelerated failure time (AFT) model is a counterpart of the marginal model. This paper aims to develop a GEE approach for

marginal semiparametric AFT models with censored data, taking advantage of recent developments on AFT models with least squares and induced smoothing.

A semiparametric AFT model is a linear model for the logarithm of the failure times with error distribution unspecified. A nice interpretation is that the effect of a covariate is to multiply the predicted failure time by some constant. It provides an attractive alternative to the popular Cox relative risk model (Cox, 1972). Three main classes of estimator exist for univariate AFT models. The first class is the aforementioned rank-based estimator which is motivated by inverting the weighted log-rank test (Prentice, 1978). The second class is the Buckley–James (BJ) estimator which extends the least squares principle to accommodate censoring through an expectation–maximization (EM) algorithm which iterates between imputing the censored failure times and least squares estimation (Buckley and James, 1979). Despite the nice asymptotic properties (Lai and Ying, 1991; Ritov, 1990), the BJ estimator may be hard to get as the EM algorithm may not converge. Further, the limiting covariance matrix is difficult to estimate because it involves the unknown hazard function of the error term. The third class is obtained by minimizing an inverse probability of censoring weighed (IPCW) loss function (Robins and Rotnitzky, 1992). The IPCW estimator is easy to compute, consistent and asymptotically normal (Stute, 1993, 1996; Zhou, 1992), but it requires correct specification of the conditional censoring distribution and overlapping of the supports of the censoring time and the failure time.

More recent works have led to a promising perspective on bringing AFT models

into routine data analysis practice. For rank-based inference, Jin et al. (2003) proposed a linear programming approach, exploiting the fact that the weighted rank estimating equation is the gradient of an objective function which can be readily solved by linear programming. Variances of the estimators were obtained from a resampling method. A computationally more efficient approach for rank-based inference with Gehan's weight (Gehan, 1965) is the induced smoothing procedure of Brown and Wang (2007). This approach is an application of the general induced smoothing method of Brown and Wang (2005), where the discontinuous estimating equations are replaced with a smoothed version, whose solutions are asymptotically equivalent to those of the former. The smoothed estimating equations are differentiable, which facilitates rapid numerical solution and sandwich variance estimator. Jin et al. (2006b) suggested an iterative least-squares procedure that starts from a consistent and asymptotically normal initial estimator, such as the one obtained from the rank-based method of Jin et al. (2003). The resulting estimator is consistent and asymptotically normal, with variance estimated from a multiplier resampling approach.

For multivariate AFT models, Jin et al. (2006c) developed rank-based estimating equations that are solved via linear programming for marginal regression parameters. Johnson and Strawderman (2009) extended the induced smoothing approach for a rank-based estimator with Gehan's weight to the case of clustered failure times and showed that the smoothed estimates perform as well as those from the best competing methods at a fraction of the computational cost. Jin et al. (2006b) considered their least squares

method with marginal models for multivariate failure times. All these approaches used independent working model and left the within-cluster dependence structure unspecified. Li and Yin (2009) developed a generalized method of moments approach for rank-based estimator using the quadratic inference function approach (Qu et al., 2000) to incorporate within-cluster dependence. Wang and Fu (2011) incorporated within-cluster ranks for the Gehan type estimator with the aid of induced smoothing. To the best of our knowledge, almost no work has been done to extend the GEE approach to the setting of multivariate AFT models. In an unpublished technical report, Hornsteiner and Hamerle (1996) attempted to combine the BJ estimator with GEE. Nevertheless, having no access to recent advances on AFT models, they did not solve the convergence problem of the EM algorithm. More importantly, their sandwich variance estimator overestimates the empirical variation by about 40–60% in most of the simulation scenarios, including one with a large sample size 2000. The problem might be caused by the fact that the variance estimator depends on the derivatives of imputed failure times with respect to regression parameters, which could not be reliably computed.

We propose an iterative GEE procedure for marginal semiparametric multivariate AFT models that generalizes the recent development of least squares approach by Jin et al. (2006b). This method has the same spirit as GEE for complete data in that misspecification of the working covariance matrix does not affect the consistency of the parameter estimator in the marginal AFT models. When the working covariance is closer to the unknown truth, the estimator has higher efficiency than that from working

independence as used in Jin et al. (2006b). Our initial estimator is the computationally efficient, rank-based estimator from Johnson and Strawderman (2009), whose consistency and asymptotic normality is inherited by the resulting GEE estimator. Unlike Hornsteiner and Hamerle (1996), we do not have convergence issues because our estimator is not BJ estimator, and our variance estimator estimates the true variation well. Further, we use a general model formulation that allows the marginal error distributions and regression coefficients to be unique for each margin or partially shared across margins as needed. All the methods are implemented in an open source R package `aftgee` (Chiou et al., 2012a).

The rest of the article is organized as follows. The semiparametric multivariate accelerated failure time model and the notation are introduced in Section 5.2. In Section 5.3, we propose the GEE approach with computational details and asymptotic properties; the sketch of proofs is relegated to appendix A.4. A large scale simulation study is reported in Section 5.4 to assess the properties of the proposed estimator. The proposed methods are applied to the aforementioned diabetic retinopathy study data in Section 5.5. A discussion concludes in Section 5.6.

5.2 Multivariate AFT Model

Under the assumptions made in Chapter 3, the multivariate AFT model can be express in a vector form. Let $\mathbf{Y}_i$, $\mathbf{T}_i$, $\mathbf{C}_i$ and $\mathbf{\Delta}_i$ be $K \times 1$ vector formed by staking Y_{ik}, T_{ik},

C_{ik}, and Δ_{ik}, respectively. Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iK})^\top$ be a $K \times p$ covariate matrix, with the kth row denoted by X_{ik} for Y_{ik}. The observed data are independent and identically distributed copies of $\{\mathbf{Y}, \mathbf{\Delta}, \mathbf{X}\}$: $\{(\mathbf{Y}_i, \mathbf{\Delta}_i, \mathbf{X}_i) : i = 1, \ldots, n\}$. We assume that $\mathbf{T}_i$ and $\mathbf{C}_i$ are conditionally independent given $\mathbf{X}_i$.

The multivariate AFT model is

$$\mathbf{T}_i = \mathbf{X}_i \beta + \boldsymbol{\epsilon}_i, \tag{5.1}$$

where β is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}, \ldots, \boldsymbol{\epsilon}_{iK})^\top$ is a random error vector with an unspecified multivariate distribution. The error vectors $\boldsymbol{\epsilon}_i$'s, $i = 1, \ldots, n$, are independent and identically distributed. Depending on the set up of the design matrix $\mathbf{X}_i$, this formulation accommodates margin-specific regression coefficients, identical regression coefficients across margins, and any compromise in between. For instance, in a model with margin-specific regression coefficients, $\mathbf{X}_i$ is a block diagonal matrix with one block for each margin and β is a stack of all marginal coefficients.

With right censoring, Buckley and James (1979) replaced each response T_{ik} in the least squares normal equations with its conditional expectation $\hat{Y}_{ik}(\beta) = E_\beta(T_{ik}|Y_{ik}, \Delta_{ik}, X_{ik})$, where the expectation is evaluated at regression coefficients β. Let $\hat{\mathbf{Y}}_i(\beta) = \big(\hat{Y}_{i1}(\beta), \ldots, \hat{Y}_{iK}(\beta)\big)^\top$. To avoid numerical problems in obtaining the BJ estimator, Jin et al. (2006b) defined

estimating equation

$$U_n^L(\beta, b) = \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{X}\right)^\top \left(\hat{\mathbf{Y}}_i(b) - \mathbf{X}_i\beta\right) = 0, \tag{5.2}$$

where $\bar{X} = \sum_{i=1}^{n} \mathbf{X}_i/n$, and b is an initial estimator of β. The BJ estimator is the solution to $U_n^L(\beta, \beta) = 0$, which is hard to obtain because $U_n^L(\beta, \beta)$ is neither continuous nor monotone in β. Let the $L_n(b)$ be the solution to (5.2) given b. Then $L_n(b)$ has a closed-form

$$L_n(b) = \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^\top (\mathbf{X}_i - \bar{X})\right]^{-1} \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^\top \left(\hat{\mathbf{Y}}_i(b) - \bar{Y}(b)\right)\right], \tag{5.3}$$

where $\bar{Y}(b) = \sum_{i=1}^{n} \hat{\mathbf{Y}}_i(b)/n$. Equation (5.3) leads to an iterative algorithm: $\hat{\beta}_n^{(m)} = L_n(\hat{\beta}_n^{(m-1)})$, $m \geq 1$. If the initial estimator b is consistent and asymptotically normal, $\hat{\beta}_n^{(m)}$ is consistent and asymptotically normal for every m. As pointed out by Jin et al. (2006b), each $\hat{\beta}^{(m)}$ is asymptotically a linear combination of the initial estimator and the BJ estimator.

Although this estimator is consistent, its efficiency might be improved because it ignores the within-cluster dependence. As will be shown, it is a special case of our GEE estimator with a working independence covariance structure.

5.3 Inference with GEE

For noncensored data, the GEE approach increases the efficiency of the marginal regression coefficient estimator by incorporating an inverse working covariance matrix as weight into the estimating equations (Liang and Zeger, 1986). A working covariance matrix does not need to be correctly specified, though the closer to the truth, the higher the efficiency. It may involve additional working parameters, whose estimation does not affect the consistency of the regression coefficient estimator. For censored data, we add an inverse working covariance weight matrix to the estimating equations of Jin et al. (2006b).

Suppose an working covariance matrix has parameter vector α. For a given initial estimator b of β, let $\alpha(b)$ be an estimator of α; more details will be given later. Our GEE for β given b is

$$U_n(\beta, b, \alpha) = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^\top \Omega_i^{-1}\big(\alpha(b)\big) \left(\hat{\mathbf{Y}}_i(b) - \mathbf{X}_i\beta\right) = 0, \tag{5.4}$$

where $\bar{X} = \sum_{i=1}^{n} \mathbf{X}_i/n$, and $\Omega_i\big(\alpha(b)\big)$ is a $K \times K$ nonsingular working covariance matrix. For given b and $\alpha(b)$, the solution to the GEE (5.4) has a closed-form

$$L_n(b, \alpha) = \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^\top \Omega_i^{-1}\big(\alpha(b)\big)(\mathbf{X}_i - \bar{X})\right]^{-1} \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^\top \Omega_i^{-1}\big(\alpha(b)\big) \left(\hat{\mathbf{Y}}_i(b) - \bar{Y}(b)\right)\right]. \tag{5.5}$$

This process can be carried out iteratively, summarized as follows.

1. Obtain an initial estimate $\hat{\beta}_n^{(0)} = b_n$ of β and initialize with $m = 1$.

2. Obtain an estimate of α given $\hat{\beta}_n^{(m-1)}$, $\hat{\alpha}_n(\hat{\beta}_n^{m-1})$.

3. Update with $\hat{\beta}_n^{(m)} = L_n\big(\hat{\beta}_n^{(m-1)}, \hat{\alpha}_n(\hat{\beta}^{(m-1)})\big)$.

4. Increase m by one and repeat 2 and 3 until convergence.

As in Jin et al. (2006b), a consistent and asymptotically normal initial estimator is important for avoiding convergence problems. We propose to use the rank-based estimator with Gehan's weight from the induced smoothing approach of Johnson and Strawderman (2009). This estimator has the same asymptotic property as the non-smooth version in Jin et al. (2003), but can be obtained with computational ease; its finite sample performance was also reported to be competitive with the best competing methods (Johnson and Strawderman, 2009).

The matrix Ω_i^{-1} is a weight matrix which does not affect the consistency of the GEE estimator, but higher efficiency can be achieved if Ω_i is chosen closer to the covariance matrix of $\hat{\mathbf{Y}}_i(b)$. When Ω_i's are the identity matrix (working independence with all marginal variances the same), our estimator reduces to the least squares estimator of Jin et al. (2006b). Since ϵ_i's are independent and identically distributed, $\Omega_i = \Omega$ for $i = 1, \ldots, n$. For convenience, we assume from now on that $E(\epsilon_{ik}) = 0$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. This can be achieved by incorporating appropriate columns of ones in X_i.

To allow possible shared distributions across margins, suppose that there are $\kappa \leq K$ unique marginal distributions. Let $e_{ik}(b) = Y_{ik} - X_{ik}^\top b$ be the right-censored error

evaluated at $\beta = b$. Let $m_r \in \{1, \ldots, \kappa\}$ denote the index of the rth margin among the κ unique marginal distributions. For a given b, let $\{e_{ir}(b), \Delta_{ir} : m_r = m_k\}$ be a set of pooled data from all the margins that have the same marginal distribution as the kth margin. Then, the conditional expectation $\hat{Y}_{ik}(b)$ in (5.4) is computed as

$$\hat{Y}_{ik}(b) = \Delta_{ik} Y_{ik} + (1 - \Delta_{ik}) \left[\frac{\int_{e_{ik}(b)}^{\infty} u \mathrm{d}\hat{F}_{k,b}(u)}{1 - \hat{F}_{k,b}\{e_{ik}(b)\}} + X_{ik}^{\top} b \right],$$

where $\hat{F}_{k,b}$ is the Kaplan–Meier estimator of the kth marginal distribution function based on pooled data, $\{e_{ir}(b), \Delta_{ir} : m_r = m_k\}$. In particular, $\hat{F}_{k,b}$ is

$$\hat{F}_{k,b}(t) = 1 - \prod_{1 \le i \le n, 1 \le r \le K : m_r = m_k, e_{ir} < t} \left(1 - \frac{\Delta_{ir}}{\sum_{j=1}^{n} \sum_{1 \le l \le K : m_l = m_k} I\left(e_{jl}(b) \ge e_{ir}(b)\right)} \right).$$

We now estimate the components in the working covariance matrix Ω, Ω_{kl} for $k, l \in \{1, \cdots, K\}$. For the diagonal elements Ω_{kk}, $1 \le k \le K$, we evaluate the conditional second moment of $\epsilon_{ik}(b)$ given the observed data:

$$\hat{V}_{ik}(b) = \Delta_{ik} e_{ik}^2(b) + (1 - \Delta_{ik}) \frac{\int_{e_{ik}(b)}^{\infty} u^2 \mathrm{d}\hat{F}_{k,b}(u)}{1 - \hat{F}_{k,b}\{e_{ik}(b)\}}, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K. \tag{5.6}$$

For a given b, we estimate Ω_{kk} by an unbiased estimator of $\mathrm{Var}\big(\epsilon_{ik}(b)\big)$

$$\hat{\Omega}_{kk}(b) = \frac{\sum_{i=1}^{n} \sum_{1 \le r \le K : m_r = m_k} \hat{V}_{ik}(b)}{n \sum_{1 \le r \le K} I\{m_r = m_k\}}. \tag{5.7}$$

In the extreme case where every margin has a unique distribution, estimator (5.7) reduces to

$$\hat{\Omega}_{kk}(b) = \frac{\sum_{i=1}^{n} \hat{V}_{ik}(b)}{n}.$$

For off-diagonal elements Ω_{kl}, $k \neq l$, define conditionally expected version of $\epsilon_{ik}(b)$,

$$\hat{e}_{ik}(b) = \hat{Y}_{ik}(b) - X_{ik}^{\top} b, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K. \tag{5.8}$$

For a given b, we estimate Ω_{kl}, $k \neq l$, by

$$\hat{\Omega}_{kl}(b) = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{ik}(b) \hat{e}_{il}(b). \tag{5.9}$$

Note that, given b, $\hat{e}_{ik}(b)$ and $\hat{e}_{il}(b)$, $\hat{\Omega}_{kl}(b)$ does not necessarily have expectation $\mathrm{Cov}\big(\epsilon_{ik}, \epsilon_{il}\big)$ unless $\Delta_{ik} = \Delta_{il} = 1$. This is because the information about dependence between $e_{ik}(b)$ and $e_{il}(b)$, $l \neq k$ is not used in $\hat{e}_{ik}(b)$ for $\Delta_{ik} = 0$. Consequently, $\hat{\Omega}_{kl}(b)$ may converge to a limit that is not necessarily $\mathrm{Cov}(\epsilon_{ik}, \epsilon_{il})$ for $k \neq l$ and the difference depend on censoring rate. This is in contrast to the diagonal element $\hat{\Omega}_{kk}$ which only uses marginal information and converges to $\mathrm{Var}(\epsilon_{ik})$. Fortunately, the consistency of the GEE estimator for β is not affected by the limit of $\hat{\Omega}$ and the use of working covariance matrix still improves the efficiency, as evident from our simulation study.

Parsimonious working covariance structures such as exchangeable (EX) or autoregressive with order one (AR1) can be imposed, in addition to the working independence

(IND) structure. Parameters α in the working covariance can be estimated with method of moment estimator $\hat{\alpha}_n$ by pooling $\hat{\Omega}_{kl}$'s as in the non-censored case (Liang and Zeger, 1986). Once $\hat{Y}_{ik}(b)$'s have been computed, the iteration in (5.5) can be carried out with the GEE approach of Yan and Fine (2004) which allows different variances for different margins, available in R package `geepack` (Halekoh et al., 2006).

Under certain regularity conditions, the proposed estimator is consistent to the true regression coefficients β_0 and asymptotically normal. The asymptotic results are summarized in the following theorems, whose proofs are sketched in Appendix A.4.

Theorem 4. *Under conditions B1– B9 in Appendix A.4, $\hat{\beta}_n^{(m)}$ is a consistent estimator of the true parameter β_0 for each $m \geq 1$.*

Theorem 5. *Under conditions B1 – B9 in Appendix A.4, $n^{1/2}(\hat{\beta}_n^{(m)} - \beta_0)$ converges in distribution to multivariate normal with mean zero for each $m \geq 1$.*

The resampling approach developed by Jin et al. (2006b) is adapted to estimate the covariance matrix of $\hat{\beta}_n^{(m)}$. Let Z_i, $i = 1, \cdots, n$, be independent and identically distributed positive random variables, independent of the observed data, with $E(Z_i) = \mathrm{Var}(Z_i) = 1$. Define

$$\hat{Y}_{ik}^*(b) = \Delta_{ik} Y_{ik} + (1 - \Delta_{ik}) \left[\frac{\int_{e_{ik}(b)}^{\infty} u \mathrm{d}\hat{F}_{k,b}^*(u)}{1 - \hat{F}_{k,b}^* \{e_{ik}(b)\}} + X_{ik}^\top b \right],$$

where

$$\hat{F}^*_{k,b}(t) = 1 - \prod_{1\le i\le n, 1\le r\le K: m_r = m_k, e_{ir}<t} \left(1 - \frac{Z_i \Delta_{ir}}{\sum_{j=1}^n \sum_{1\le l\le K: m_l = m_k} Z_i I\left(e_{jl}(b) \ge e_{ir}(b)\right)}\right).$$

Then the multiplier resampling version of equation (5.5) has the following form,

$$L^*_n(b,\alpha) = \left[\sum_{i=1}^n Z_i(\mathbf{X}_i - \bar{X})\Omega_i^{-1}\big(\alpha(b)\big)(\mathbf{X}_i - \bar{X})\right]^{-1} \left[\sum_{i=1}^n Z_i(\mathbf{X}_i - \bar{X})\Omega_i^{-1}\big(\alpha(b)\big)\left\{\hat{\mathbf{Y}}^*_i(b) - \bar{Y}^*(b)\right\}\right],$$

where $\bar{Y}^*(b) = \sum_{i=1}^n \hat{\mathbf{Y}}^*_i(b)/n$. For a realization of $\{Z_1, \ldots, Z_n\}$ and an initial estimator $\hat{\beta}_n^{(0)}$, a bootstrap estimator of β is obtained from iteration $\hat{\beta}_n^{(m)*} = L^*_n(\hat{\beta}_n^{(m-1)*})$. The covariance matrix of $\hat{\beta}_n^{(m)}$ can be estimated from the sample covariance matrix of a bootstrap sample of $\hat{\beta}_n^{(m)*}$. The consistency of this variance estimator can be proved following arguments similar to those in Jin et al. (2006b).

5.4 Simulation Study

We conducted two simulation studies to assess the performance of the proposed estimators with parsimonious working covariance structures and compared their efficiency with those from the least squares approach by Jin et al. (2006b). The latter estimator is also our GEE estimator with working independence covariance structure. The first study had a clustered failure time setting with identical regression coefficients across margins and identical marginal error distributions. The cluster sizes were fixed at three.

For cluster i, the multivariate failure time $T_i = (T_{i1}, T_{i2}, T_{i3})^\top$ was generated from

$$\log T_{ik} = 2 + X_{1ik} + X_{2ik} + \epsilon_{ik},$$

where X_{1ik} was Bernoulli with rate 0.5, X_{2ik} was $N(0, 0.5^2)$, and the joint distribution of $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})^\top$ was specified by identical marginal distributions and a Clayton copula. Three marginal error distributions were considered: standard normal, standard logistic, and standard Gumbel, abbreviated by N, L, and G, respectively; the tail of the three distributions gets heavier from N to L to G. Three levels of dependence measured by Kendall's tau were considered for the Clayton copula: 0, 0.3, and 0.6. Censoring times were independently generated from the uniform distributions over $(0, c)$, where c was tuned for each margin to achieve three levels of censoring percentage: 0%, 25%, and 50%. We considered sample size $n = 200$ clusters. The rank-based estimator with Gehan's weight from the induced smoothing approach of Johnson and Strawderman (2009) was used as the initial estimator in the iterative estimation procedure. Three working covariance structures were used: IND, EX and AR1. The covariance matrix of the estimator was estimated from the resampling approach with 200 bootstrap size in Section 5.3. For each configuration, we did 1000 replicates.

The results of the first study are summarized in Table 15. To save space, only results for nonzero Kendall's tau were reported. All estimators appear to be virtually unbiased. The empirical variation of the estimates and the estimated variation from the resampling

procedure agree closely for all estimators, suggesting that the resampling procedure provides valid inference. For a given censoring percentage, as the dependence level increases, the variance of the estimator changes little under the IND working covariance structure, but decreases under both the EX and AR1 structures. Further, the variances under the EX structure are in general smaller than those from the AR1 structure, which is expected because the true covariance structure is exchangeable in this simulation setting. For a fixed dependence level, the effect of censoring percentage on the variances of the estimator depends on the marginal error distributions. The variance increases clearly as censoring gets heavier when the errors are normally distributed, but this pattern is not observed when the marginal error distribution is Gumbel or logistic. The relative efficiency of the proposed GEE estimator under the EX structures in relative to that with IND structure is up to 3.369 (logistic margin at Kendall's tau 0.6 and censoring percentage 25%) — a substantial gain in efficiency.

The second simulation study had multiple event data with different regression coefficients and different marginal error distributions. The cluster sizes were still fixed at three. For cluster i, the multivariate failure times were generated from

$$\log T_{ik} = \beta_{0k} + \beta_{1k} X_{1ik} + \beta_{2k} X_{2ik} + \epsilon_{ik},$$

where $(\beta_{0k}, \beta_{1k}, \beta_{2k})^\top$, $k = 1, 2, 3$, was the regression coefficient vector for margin k,

Table 15: Summary of simulation results with identical regression coefficients and identical marginal error distributions based on 1000 replications. MD is marginal distribution; Cens. (%) is censoring percentage. Empirical SE is the standard deviation of the parameter estimates; Estimated SE is the mean of the standard error of the estimator; RE is the empirical relative efficiencies in relative to the estimator under working IND structure.

				Bias			Empirical SE			Estimated SE			RE	
MD	τ	Cens. (%)	β	IND	EX	AR1	IND	EX	AR1	IND	EX	AR1	EX	AR1
N	0.3	0%	β_1	0.003	0.000	0.000	0.087	0.073	0.076	0.081	0.068	0.071	1.419	1.306
			β_2	0.001	0.003	0.003	0.083	0.071	0.073	0.080	0.068	0.071	1.345	1.264
		25%	β_1	0.007	0.005	0.004	0.086	0.074	0.078	0.086	0.073	0.076	1.340	1.218
			β_2	−0.001	−0.002	−0.002	0.087	0.074	0.078	0.087	0.074	0.078	1.382	1.258
		50%	β_1	0.005	0.005	0.005	0.098	0.085	0.090	0.097	0.084	0.088	1.347	1.177
			β_2	−0.001	−0.002	−0.004	0.101	0.089	0.093	0.098	0.086	0.089	1.279	1.177
	0.6	0%	β_1	0.005	0.000	0.000	0.085	0.048	0.052	0.080	0.046	0.050	3.143	2.671
			β_2	0.002	0.001	0.000	0.079	0.045	0.050	0.080	0.046	0.050	3.159	2.491
		25%	β_1	0.002	0.001	0.000	0.088	0.053	0.056	0.086	0.052	0.057	2.717	2.444
			β_2	0.003	0.000	0.000	0.092	0.057	0.061	0.087	0.053	0.058	2.605	2.281
		50%	β_1	−0.001	−0.003	−0.004	0.100	0.065	0.070	0.096	0.065	0.071	2.340	2.004
			β_2	−0.002	−0.004	−0.003	0.100	0.071	0.076	0.098	0.067	0.073	2.025	1.728
L	0.3	0%	β_1	0.000	0.000	0.001	0.142	0.120	0.126	0.146	0.123	0.129	1.402	1.274
			β_2	0.002	0.002	0.002	0.150	0.127	0.134	0.144	0.122	0.128	1.390	1.241
		25%	β_1	−0.007	−0.006	−0.007	0.151	0.122	0.128	0.148	0.122	0.129	1.529	1.391
			β_2	0.000	−0.001	−0.003	0.153	0.127	0.134	0.149	0.122	0.129	1.443	1.301
		50%	β_1	0.008	0.007	0.005	0.168	0.134	0.144	0.166	0.135	0.142	1.564	1.357
			β_2	−0.004	−0.003	−0.003	0.174	0.139	0.145	0.167	0.137	0.144	1.565	1.437
	0.6	0%	β_1	0.003	0.001	0.000	0.145	0.087	0.095	0.146	0.085	0.092	2.789	2.290
			β_2	0.003	0.007	0.007	0.151	0.084	0.091	0.146	0.085	0.092	3.241	2.741
		25%	β_1	0.005	0.001	0.001	0.151	0.081	0.089	0.148	0.081	0.088	3.432	2.846
			β_2	0.000	0.001	−0.001	0.156	0.082	0.090	0.148	0.081	0.089	3.632	2.987
		50%	β_1	−0.003	−0.002	−0.003	0.169	0.093	0.104	0.165	0.091	0.100	3.327	2.679
			β_2	0.001	0.007	0.008	0.174	0.101	0.113	0.167	0.094	0.104	2.974	2.364
G	0.3	0%	β_1	0.009	0.010	0.009	0.103	0.094	0.099	0.103	0.093	0.096	1.190	1.090
			β_2	−0.001	−0.001	0.000	0.108	0.100	0.102	0.102	0.093	0.096	1.158	1.102
		25%	β_1	−0.004	−0.003	−0.004	0.099	0.086	0.089	0.097	0.087	0.090	1.319	1.224
			β_2	−0.003	−0.002	−0.001	0.100	0.089	0.091	0.098	0.087	0.090	1.258	1.204
		50%	β_1	−0.004	−0.005	−0.005	0.103	0.088	0.092	0.097	0.087	0.090	1.349	1.245
			β_2	−0.002	0.001	0.001	0.100	0.092	0.095	0.099	0.089	0.093	1.183	1.109
	0.6	0%	β_1	0.000	−0.001	0.000	0.100	0.071	0.079	0.103	0.072	0.077	1.956	1.596
			β_2	−0.001	−0.001	−0.002	0.106	0.073	0.080	0.103	0.072	0.077	2.115	1.770
		25%	β_1	−0.004	−0.003	−0.003	0.101	0.066	0.072	0.097	0.065	0.070	2.350	1.998
			β_2	−0.002	0.001	0.000	0.097	0.066	0.072	0.097	0.066	0.071	2.192	1.833
		50%	β_1	0.000	0.001	0.000	0.099	0.071	0.076	0.097	0.070	0.076	1.950	1.720
			β_2	0.002	0.000	0.000	0.100	0.075	0.081	0.100	0.074	0.080	1.768	1.510

and the joint distribution of $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})^\top$ was specified by three marginal distributions and a Clayton copula. The marginal distributions of ϵ_i were set to be standard normal, standard logistic, and standard Gumbel, respectively, for the first, second and third margin. The Clayton copula had three levels of dependence measured by Kendall's tau: 0, 0.3, and 0.6. The regression coefficients $(\beta_{0k}, \beta_{1k}, \beta_{2k})$ were set to be $(-1, 1, -1)$, $(1, -1, 1)$, and $(1, 1, 1)$, respectively for $k = 1$, 2, and 3. Other settings such as the covariates, censoring time, sample size, initial estimator, bootstrap sample size for variance estimation, replication size were all the same as in the first study. Three working covariance structures were considered: IND, EX, unstructured (UN).

The results of the second study are summarized in Table 16. Similar to the first simulation study, all estimators are virtually unbiased, and their variance estimators are generally close to the empirical variances of the replicates. The variances of the GEE estimators under EX and UN structures decrease as the dependence gets stronger at any level of censoring percentage. Holding the dependence level, as the censoring percentage increases, the variance increases at the normal margin, but the pattern is different for the other two margins. The variance has little changes at the logistic margin. At the Gumbel margin, it changes little as the censoring percentage increases from 0% to 25%, but increases notably as the censoring percentage increases from 25% to 50%. There is almost no difference between the GEE estimator under EX and that under UN, both leading to about the similar variance for all cases. With independent covariance structure as reference, the relative efficiency of GEE with EX and AR1 structures show significant

Table 16: Summary of simulation results with different regression coefficients and different marginal error distributions based on 1000 replications. Cens. (%) is censoring percentage. Empirical SE is the standard deviation of the parameter estimates; Estimated SE is the mean of the standard error of the estimator; RE is the empirical relative efficiencies in relative to the estimator under working IND structure.

			Bias			Empirical SE			Estimated SE			RE	
τ	Cens. (%)	β	IND	EX	UN	IND	EX	UN	IND	EX	UN	EX	UN
0.3	0%	β_1	0.007	0.003	0.003	0.142	0.122	0.123	0.140	0.120	0.119	1.341	1.323
		β_2	0.000	−0.003	−0.004	0.148	0.130	0.130	0.139	0.120	0.119	1.294	1.294
		β_1	−0.001	−0.003	−0.002	0.182	0.163	0.164	0.179	0.160	0.159	1.242	1.233
		β_2	−0.002	−0.005	−0.005	0.179	0.160	0.161	0.178	0.158	0.157	1.251	1.237
		β_1	0.001	−0.004	−0.003	0.257	0.219	0.219	0.253	0.217	0.217	1.372	1.371
		β_2	0.006	−0.001	−0.003	0.264	0.227	0.228	0.251	0.217	0.217	1.361	1.351
	25%	β_1	0.006	0.004	0.003	0.150	0.131	0.132	0.150	0.127	0.127	1.311	1.306
		β_2	−0.001	−0.007	−0.006	0.156	0.132	0.132	0.150	0.129	0.128	1.414	1.414
		β_1	−0.003	−0.001	0.000	0.171	0.151	0.151	0.169	0.148	0.147	1.283	1.283
		β_2	−0.007	−0.010	−0.010	0.176	0.154	0.154	0.170	0.149	0.149	1.306	1.306
		β_1	−0.003	−0.006	−0.006	0.266	0.228	0.230	0.260	0.220	0.219	1.359	1.339
		β_2	−0.009	−0.011	−0.012	0.269	0.229	0.228	0.262	0.221	0.221	1.381	1.393
	50%	β_1	0.002	0.001	0.000	0.167	0.144	0.145	0.170	0.146	0.145	1.336	1.331
		β_2	−0.008	−0.008	−0.007	0.176	0.150	0.150	0.172	0.148	0.147	1.373	1.373
		β_1	−0.001	−0.005	−0.004	0.175	0.153	0.152	0.170	0.149	0.148	1.312	1.335
		β_2	0.004	0.004	0.002	0.189	0.165	0.166	0.173	0.153	0.152	1.317	1.296
		β_1	−0.004	0.001	0.000	0.320	0.270	0.271	0.308	0.262	0.260	1.411	1.398
		β_2	0.011	0.006	0.007	0.328	0.283	0.283	0.308	0.264	0.262	1.342	1.342
0.6	0%	β_1	0.004	0.000	−0.001	0.146	0.089	0.087	0.140	0.084	0.092	2.707	2.809
		β_2	−0.015	−0.003	−0.002	0.137	0.085	0.085	0.138	0.082	0.090	2.637	2.638
		β_1	−0.009	0.000	−0.001	0.187	0.126	0.126	0.179	0.120	0.142	2.202	2.199
		β_2	0.000	0.000	0.000	0.186	0.124	0.124	0.176	0.119	0.166	2.276	2.272
		β_1	0.001	−0.004	−0.005	0.257	0.159	0.156	0.253	0.156	0.192	2.615	2.705
		β_2	0.000	−0.001	0.000	0.249	0.158	0.156	0.250	0.154	0.189	2.504	2.556
	25%	β_1	0.005	0.003	0.002	0.152	0.093	0.092	0.150	0.091	0.113	2.667	2.740
		β_2	−0.003	−0.004	−0.005	0.152	0.093	0.092	0.151	0.093	0.112	2.670	2.703
		β_1	0.002	−0.003	−0.002	0.172	0.113	0.113	0.169	0.111	0.114	2.327	2.331
		β_2	−0.007	−0.006	−0.006	0.176	0.118	0.118	0.170	0.112	0.114	2.251	2.248
		β_1	−0.004	0.000	0.000	0.271	0.160	0.160	0.261	0.155	0.175	2.861	2.864
		β_2	0.007	0.000	0.000	0.267	0.153	0.152	0.260	0.155	0.174	3.053	3.053
	50%	β_1	−0.001	0.004	0.004	0.171	0.112	0.111	0.170	0.112	0.112	2.387	2.392
		β_2	−0.003	−0.005	−0.005	0.188	0.120	0.119	0.171	0.118	0.117	2.464	2.484
		β_1	−0.002	0.002	0.003	0.184	0.120	0.120	0.168	0.119	0.120	2.332	2.329
		β_2	0.008	0.005	0.003	0.176	0.127	0.127	0.171	0.125	0.126	1.935	1.936
		β_1	−0.015	−0.006	−0.003	0.311	0.199	0.196	0.311	0.200	0.203	2.452	2.511
		β_2	0.010	0.004	0.004	0.323	0.207	0.205	0.309	0.204	0.203	2.431	2.479

increases as Kendall's tau increased from 0.3 to 0.6.

5.5 Diabetic Retinopathy Study

The diabetic retinopathy study (DRS) was started in 1971 (Diabetic Retinopathy Study Research Group, 1976) with the aim to investigate the efficacy of laser photocoagulation in delaying onset of severe vision loss. Diabetic retinopathy is the most common and serious eye complication of diabetes, which may lead to poor vision or even blindness. A subset of the DRS data for patients with "high-risk" diabetic retinopathy, categorized by risk group 6 or higher, has been analyzed by many authors (e.g., Huster et al., 1989; Lee and Wei, 1993; Liang et al., 1993; Spiekerman and Lin, 1996). Each of the 197 patients in this subset had one eye randomized to laser treatment and the other eye received no treatment. The outcomes of interest were the actual times from initiation of treatment to the time when visual acuity dropped below 5/200 at two visits in a row (defined as "blindness"). The scientific interest was the effectiveness of the laser treatment and the influence of other risk factors. In addition to the treatment indicator, three covariates are available: age at diagnosis of diabetes, type of diabetes (1 = adult, 0 = juvenile), and risk group (6 to 12, rescaled to 0.5 to 1.0). Since the interaction between treatment and diabetes type was found to be significant in Spiekerman and Lin (1996), we also included this interaction in the model.

Table 17: Summaries of results of marginal semiparametric AFT models for data from the diabetic retinopathy study.

		JS		IND		EX	
Margin	Effects	EST	SE	EST	SE	EST	SE
Identical error margins and identical regression coefficients:							
pooled	risk group	−2.659	0.739	−2.408	0.859	−2.306	0.775
	age	−0.010	0.012	−0.010	0.013	−0.010	0.014
	diabetes	−0.140	0.349	−0.065	0.440	−0.065	0.369
	treatment	0.520	0.197	0.545	0.330	0.542	0.263
	interaction	1.116	0.301	0.961	0.466	0.964	0.410
Different error margins and different regression coefficients:							
left	risk group	−2.819	1.114	−2.832	1.195	−2.654	1.242
	age	−0.042	0.016	−0.037	0.019	−0.036	0.020
	diabetes	0.825	0.463	0.706	0.554	0.702	0.544
	treatment	0.925	0.422	0.645	0.549	0.652	0.489
	interaction	1.719	0.650	1.742	0.855	1.739	0.820
right	risk group	−2.087	1.013	−1.944	1.316	−1.805	1.283
	age	0.011	0.014	0.009	0.016	0.009	0.018
	diabetes	−0.770	0.432	−0.640	0.528	−0.639	0.656
	treatment	0.383	0.326	0.481	0.381	0.477	0.446
	interaction	0.752	0.476	0.600	0.639	0.603	0.646
Identical error margins with partially common regression coefficients:							
left	age	−0.039	0.015	−0.036	0.021	−0.036	0.022
	diabetes	0.892	0.406	0.848	0.607	0.846	0.621
right	age	0.011	0.015	0.009	0.019	0.009	0.017
	diabetes	−0.870	0.435	−0.837	0.499	−0.835	0.574
common	treatment	0.630	0.227	0.606	0.250	0.607	0.267
	risk group	−2.588	0.747	−2.409	1.034	−2.264	0.938
	interaction	1.067	0.318	1.014	0.344	1.014	0.409

We first fit a bivariate AFT model with identical error margins and identical regression coefficients for both left and right eyes. The second AFT model we fit was the opposite, with different error margins and different regression coefficients for left and right eyes. For each model, we report GEE estimators with working independence and working exchangeable covariance structures; see Table 17. For comparison, the rank-based JS estimator is also reported. The GEE estimator with exchangeable working structure from the first model suggests that the treatment was significant in delaying the onset of vision loss; it had a significant higher effect for adult than for juvenile, and patients in higher risk groups tended to lose vision sooner. Note that almost all standard errors are small from the exchangeable structure than those from the independence structure. In particular, the treatment effect was not significant under independence but was significant under the exchangeable structure. The second model offered a possibility to check whether the marginal error distributions and regression coefficients should indeed be identical as assumed in the first model. Figure 2 shows the the Kaplan–Meier survival curves of the censored residuals for the left margin and right margin respectively, overlaid with the pooled estimate from the first model. All three curves appear to be mingled together tightly. A naive log-rank test to compare the two margins, ignoring that the regression coefficients were not known but estimated, yielded a p-value of 0.907, confirming the visual observation. The second model also allows hypothesis testing of equal coefficients for each covariate across the two margins with Wald-type tests. The coefficients of treatment, risk group, and treatment-diabetes interaction were found to

be not significantly different across the two margins, with p-values 0.400, 0.278, and 0.147, respectively. The coefficients of age and diabetes were found to be significantly different across the two margins, with p-values 0.036 and 0.042, respectively.

We then fit an bivariate AFT model with identical error margins, same coefficients for treatment, risk group and treatment-diabetes interaction, and different coefficients for age and diabetes. This is one of the many models with intermediate complexity between the first model and the second model. The results are summarized in the last section of Table 17. This time, the shared coefficients of treatment, risk group, and treatment-diabetes interaction remained significant as before. An interesting finding is that the difference between the coefficient of diabetes (0.846 versus -0.835) is significantly nonzero with a p-value 0.002, suggesting that patients with adult diabetes had earlier onset of vision loss in right eye than patients with juvenile diabetes, but the opposite was true for the left eye. This finding has not been reported in existing analyses and may worth closer investigation with experts in diabetics.

5.6 Discussion

The working covariance structure of the proposed GEE approach with censored data does not affect the consistency of the estimator as in the complete data case, but may improve the efficiency. At each margin, the errors are assumed to be independent and identically distributed, and hence have the same variance across clusters. The homoskedasticity

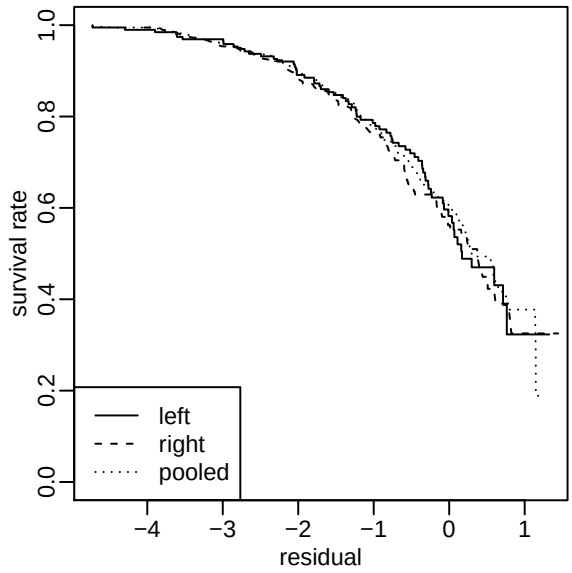


Figure 2: Kaplan–Meier survival curves for censored residuals of the DRS Study.

assumption may be relaxed by imposing a structure on the marginal variances of the errors. In particular, ϵ_{ik} in model (5.1) can be replaced with $\sigma_{ik}\nu_{ik}$, where ν_{ik}'s are independent and identically distributed across clusters $i = 1, \ldots, n$ with mean zero and variance one, and the scale parameter σ_{ik} may be described by a regression model with covariates. Such specification leads to heteroskedasticity in errors across clusters and merits further investigation.

Extension of the methods to other practical settings can be considered. When some of the covariates are missing by design as in case-cohort studies, the GEE approach can be adapted with a weight accounting for the inclusion probability to provide a possibly more efficient alternative to rank-based inferences (Chiou et al., 2013a). For clustered failure times with unequal cluster sizes, the working covariance matrix Ω_i's, which are of different dimensions, can still be constructed with exchangeable or autoregressive

structures from estimators of a small number of working parameters. In the special case of recurrent gap times, where clustered failure times occur sequentially and only the last one is censored, the GEE approach needs to be applied with caution. The second and later gap times are subject to dependent censoring, and the last censored gap time tends to be longer than uncensored gap times due to intercept sampling (Wang and Chang, 1999). Marginal AFT models can be fitted with the last censored gap times removed for clusters with size greater than one to correct the bias (Luo and Huang, 2011).

For applications like the DRS study, where identical distribution across margins may be needed and justified, a rigorous test to compare the survival curves of the residuals would be desirable. We used naive tests ignoring the fact that the residuals were calculated based on estimated regression coefficients. A rigorous test procedure should take into account of the variation caused by the estimation procedure.

Chapter 6

R Package: aftgee

6.1 Introduction

The linear regression model is the most commonly used regression model in data analysis for uncensored data. When data are right-censored, two of the most frequently used regression models are the relative risk model and the accelerate failure time (AFT) model. Recently, the AFT model with an unspecified error distribution has been studied extensively. In particular, two methods have received special attention. One is the rank-based approach motivated by inverting the weighted log-rank test discussed in Chapter 2, 3 and 4. The other method is the extension of the least squares principle from Buckley and James (1979). Thus, the convergence of the algorithm is not guaranteed. Due to lack of efficient and reliable computing algorithm, both approaches have not been widely used in practice until recently (Jin et al., 2003, 2006b,c). We present package **aftgee** aiming to provide easy access to fitting AFT models with both methods based on recent methodological development.

Several packages for fitting AFT models are available for the R environment (R Development Core Team, 2012). For example, `survreg` in package **survival** (Therneau,

2012) and `aftreg` in package **eha** (Brostrom, 2012) provide simple parametric AFT fit. The major limitation is that these functions requires a parametric specification on error distribution. The function `bj` from package **rms** (Harrell, 2012) fits semiparametric AFT models through a Buckley-James estimator (Buckley and James, 1979). Moreover, there is no reliable method to estimate the covariance matrix of the resulting estimators. To overcome these difficulties, Jin et al. (2006b) developed a linear programing approach for semiparametric AFT models. Their method yields consistent and asymptotically normal estimators. The implementation is available in package **lss** (Jin and Huang, 2007). For for multivariate failure times, package **lss** considered their least squares method with marginal models. All these approaches used independent working model and left the within-cluster dependence structure unspecified.

Our package is comprehensive in fitting semiparametric AFT models in numbers of ways. First, to improve the estimation efficiency, we apply the induced smoothing technique in Brown and Wang (2005) onto the unsmooth Gehan (Gehan, 1965) rank-based estimating equations. The procedure permits fast and accurate estimation for the unknown regression parameter. The resulting estimators are also consistent and have the same asymptotic distribution as its unsmooth version. Second, we extended this induced smoothing technique to general weight functions including log-rank (Prentice, 1978), Prentice-Wilcoxon (Prentice, 1978) and the more general G^ρ class (Harrington and Fleming, 1982). To bypass the difficulties caused by incorporating general weights,

we approximate the estimating equations by an asymptotically equivalent version. To estimate the standard errors, the full bootstrap variance estimation can be time demanding even with the fast induced smoothing techniques. Several sandwich variance estimators (Chiou et al., 2013a) are also implemented and provide valid inferences. Third, for the least squares approach, principles of generalized estimating estimator (GEE) are applied to improve efficiency by working covariance structure for multivariate dependence. Incorporating GEE, our least squares approach includes that of package **lss** as a special case. Lastly, in the case of missing observations, a weight extension on the rank-based estimation is developed. Because of these features, the **aftgee** package is appealing to analysts who would like to fit AFT models in their routine analysis of survival data.

The rest of the article is organized as follows. In Section 6.2, we summarized the usages of package functions. Two examples, one univariate and one multivariate, are in Section 6.3. Conclusion and some remarks are summarized in Section 6.4.

6.2 Package Implementation

The two major functions for package `aftgee` are `smoothrr` and `aftgee` for the rank-based approach and the least squares approach respectively. The arguments of `smoothrr` are

```
> library("aftgee")
> args(smoothrr)

function (formula, data, subset, contrasts = NULL, id, weights = NULL,
```

```
    rankweights = "gehan", binit = "lm", sigmainit = NULL, variance = "ISMB",
    B = 100, strataid = NULL, control = aftgee.control())
NULL
```

The required argument are `formula`, `data` and `id`. The argument `formula` specifies the model to be fit with the variables coming with `data`. This argument has the same format as the formula argument in the function `survreg` from **survival** with response created from `Surv`. Clusters are defined by vector `id`. The vector `subset` is a logical expression indicating elements or rows to be used in the fit. All observations are included by default when subset is unspecified. `contrast` is an optional list describing how factors should be coded. When stratification is considered, a vector specifying strata and the the observation weights (h_i in Chapter 2) can be supplied in `strataid` and `weights` respectively. The length of the arguments `id`, `weights` and `strataid` need to be the same as the number of observation. The type of weight estimating function, $\varphi_i(\beta)$, is controlled by a character string, `rankweight`. The available weights for rank-based estimating functions include log-rank, Gehan, Prentice-Wilcoxon and the more general G^{ρ} class denoted by `"logrank"`, `"gehan"`, `"PW"` and `"GP"` respectively.

The initial values for parameter estimator and variance estimator are determined by `binit` and `sigmainit` respectively. A vector consists of coefficients from simple linear regression is the default value for `binit` whereas the identity matrix is the default value for `sigmainit`.

Given the initial values, variance estimates can be obtained from several approaches. These variance estimations are specified by `variance` argument. The most straight

forward variance estimator is the multiplier bootstrap approach (`"MB"`) motivated by (Jin et al., 2003). The multiplier resampling approach is computationally inefficient because it requires to solve the estimating equation repeatedly. A more efficiency method is the class of sandwich variance estimators considered by Chiou et al. (2013a). Suppose the variance of the estimator has sandwich form, $\Sigma = A^{-1}V(A^{-1})^\top$ where V is the asymptotic variance of the estimating equation and A is the slope matrix. Chiou et al. (2013a) proposed to estimate V by either a closed-form formulation (`CF`) or through bootstrapping the estimating equations (`MB`). The bootstrapping estimates of V is much less demanding than the full multi1plier bootstrap, because it only involves evaluations of estimating equations instead of solving them. On the other hand, to estimate the slope matrix A, Chiou et al. (2013a) proposed three methods base on induced smoothing approach (`IS`), smoothed Huang's approach (`sH`) motivated by Huang (2002) or Zeng and Lin's approach (`ZL`) by Zeng and Lin (2008). Combinations between estimating V and A yield six sandwich estimators, `"ISCF"`, `"ISCF"`, `"ZLCF"`, `"ZLMB"`, `"sHCF"`, `"sHMB"` for `variance`. When bootstrap is needed, the bootstrap size is controled by `B` default at 500.

The convergence criterion for the procedure is controlled by relative tolerance. The iteration stops and the output is given when the tolerance is met or iteration reaches the pre-specified maximum iteration number. The default relative tolerance is set at 0.001 and the default max iteration step is at 30. The control argument, `aftgee.control`, has the following default set up

```
> args(aftgee.control)

function (maxiter = 30, reltol = 1e-04, abstol = 1e-04, trace = FALSE)
NULL
```

where `maxiter` controls the maximum number of iteration, `reltol` is the relative convergence tolerance and `trace` is a logical value that determine whether to print the output for each iteration.

The least squares estimator can be obtained by calling `aftgee` with the following arguments

```
> args(aftgee)

function (formula, data, subset, id, contrasts = NULL, weights = NULL,
    margin = NULL, corstr = "independence", binit = "lm", B = 100,
    control = aftgee.control())
NULL
```

Most of the arguments of `smoothrr` are shared by `aftgee` with some special arguments including `margin` and `corstr`. The `margin` is a vector with the same length as data, it is used to specify the marginal distribution within clusters. Identical marginal distribution is assumed if `margin` is not specified. A character string, `corstr` passed to `geese`, is used to specify the predefined working correlation structures. Four predefined working correlation structures are independence (`indep`), exchangeable (`ex`), autoregressive model of order one (`ar1`) and unstructured (`unstructured`). The default is `independence`. Under `aftgee`, the initial values for parameter estimator is default at `"srrgehan"`, corresponding to the induced smooth rank-based approach with Gehan's

weight. Alternatively, although not recommended, the simple linear regression (`"lm"`) which ignores censoring, can also be used for faster result.

6.3 Illustrations

6.3.1 National Wilm's Tumor Study

We first demonstrate the performance of our proposed methods with a cohort studies conducted by the National Wilm's Tumor Study Group (NWTSG) (D'Angio et al., 1989; Green et al., 1998). The dataset is available in **survival** package as `nwtco`. The interest of the study is to assess the relationship between the measurement from central histology (`histol`) and the time to tumor relapse (`edrel`). In addition to the central histology measurement (1 = favorable, 0 = unfavorable), we also include the relapse of tumor on the patient's age (`age`) in years as covariates. There are two study groups (`study`), NWTSG-3 and NWTSG-4, denoting the third and the fourth Wilms tumor studies. Patients are further categorized into four stages (`stage`) with stage 4 being the latest and most severest. The dataset consists of 4028 patients, among which, 571 patients experienced tumor relapse (`rel` = 1) and 4028 patients did not (`rel` = 0).

We first prepare the data by rescale the time to relapse and age in years.

```
> library("survival")
> data("nwtco")
> nwtco$age <- nwtco$age/12
> nwtco$edrel <- nwtco$edrel/12
```

```
> head(nwtco)

  seqno instit histol stage study rel edrel   age in.subcohort
1     1      2      2     1     3   0 506.2 2.083        FALSE
2     2      1      1     2     3   0 343.4 4.167        FALSE
3     3      2      2     1     3   0 505.8 0.750        FALSE
4     4      2      1     4     3   0 516.7 2.333         TRUE
5     5      2      2     2     3   0 103.7 4.583        FALSE
6     6      1      1     2     3   0 244.3 2.667        FALSE
```

To take advantage of availability of the full cohort data, we first fit the full-cohort data. Gehan's type rank-based approaches with standard errors estimated via multiplier bootstrap approach (`MB`) and the induced smoothing approach (`ISMB` and `ISCF`) are considered. For comparison propose, the first model is fitted with `MB` and the second model is fitted with both `ISMB` and `ISCF`.

```
> system.time(fit.MB <- smoothrr(Surv(edrel, rel) ~ histol + age - 1,
+                                id = seqno, data = nwtco,
+                                variance = "MB",
+                                rankweights = "gehan"))

   user  system elapsed
   2025.734       0.052    2032.037

> system.time(fit.IS <- smoothrr(Surv(edrel, rel) ~ histol + age - 1,
+                                id = seqno, data = nwtco,
+                                variance = c("ISCF", "ISMB"),
+                                rankweights = "gehan"))

Timing stopped at: 99.982 3.268 103.354
```

The summary gives the following information:

```
> summary(fit.MB)

Call:
smoothrr(formula = Surv(edrel, rel) ~ histol + age - 1, data = nwtco,
    id = seqno, rankweights = "gehan", variance = "MB")


Variance Estimator: MB
       Estimate  StdErr z.value p.value
histol  -3.2206  0.1422  -22.64  <2e-16 ***
age     -0.2313  0.0244   -9.48  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(fit.IS)

Call:
smoothrr(formula = Surv(edrel, rel) ~ histol + age - 1, data = nwtco,
    id = seqno, rankweights = "gehan", variance = c("ISCF", "ISMB"))

Variance Estimator: ISCF
       Estimate  StdErr z.value p.value
histol  -3.2206  0.1438  -22.40  <2e-16 ***
age     -0.2313  0.0256   -9.03  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variance Estimator: ISMB
       Estimate  StdErr z.value p.value
histol  -3.2206  0.1407  -22.89  <2e-16 ***
age     -0.2313  0.0261   -8.88  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors for all three variance estimations agree closely. The coefficients

of central histological diagnosis and age are found to be significantly different from zero. In terms of timing, the `MB` standard error took substantial amount of time; `MB` method over half an hour while combining `ISMB` and `ISCF` took 78 seconds.

With the same data set, we next demonstrate incorporating weights via case-cohort design. Define cases and controls as those who experience the event of interest by the end of the study period and who do not, respectively. The case-cohort sample is the union of all the cases and the sub-cohort sample which is selected from population via simple random sampling. The sampling indicator for sub-cohort sample is available in `nwtco` as `in.subcohort`. There are 668 patients in the sub-cohort sample and 1154 patients in the case-cohort sample. This produces a sub-cohort inclusion probability of 0.165. The biases due to sampling can be adjusted by inserting appropriate weights, h_i. In this particular example, among the case-cohort sample, $h_i = 1$ for cases and $h_i = 1/p$ for controls. On the other hand, $h_i = 0$ for non case-cohort sample. The weights, h_i can be prepared as follow.

```
> table(nwtco$in.subcohort, nwtco$rel)


           0    1
  FALSE 2874  486
  TRUE   583   85

> nwtco$in.casecohort <- (nwtco$in.subcohort | nwtco$rel == 1)
> p <- sum(nwtco$in.subcohort) / 4028
> nwtco$hi <- nwtco$in.casecohort * (nwtco$rel == 1) / p
> table(nwtco$hi)
```

```
              0 6.02994011976048
           3457              571
```

For the case-cohort design, we also demonstrate the usage of different rank weights in rank-based approach; we first fit with Gehan's type then with logrank. Standard errors are estimated with the efficient sandwich variance estimator, `ISMB`. Commands for these estimators are presented below followed by summary.

```
> system.time(fit.g <- smoothrr(Surv(edrel, rel) ~ histol + age - 1,
+                               id = seqno, weights = hi, data = nwtco,
+                               variance = "ISMB", rankweights = "gehan",
+                               subset = in.casecohort))

   user  system elapsed
  26.37    0.07   43.87

> system.time(fit.l <- smoothrr(Surv(edrel, rel) ~ histol + age - 1,
+                               id = seqno, weights = hi, data = nwtco,
+                               variance = "ISMB", rankweights = "logrank",
+                               subset = in.casecohort))

   user  system elapsed
  35.22    0.02   35.63
```

```
> summary(fit.g)

Call:
smoothrr(formula = Surv(edrel, rel) ~ histol + age - 1, data = nwtco,
    subset = in.casecohort, id = seqno, weights = hi, rankweights = "gehan",
    variance = "ISMB")


Variance Estimator: ISMB
```

```
       Estimate  StdErr z.value p.value
histol  -0.3718  0.0708   -5.25 1.5e-07 ***
age      0.0394  0.0110    3.58 0.00034 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(fit.l)

Call:
smoothrr(formula = Surv(edrel, rel) ~ histol + age - 1, data = nwtco,
    subset = in.casecohort, id = seqno, weights = hi, rankweights = "logrank",
    variance = "ISMB")


Variance Estimator: ISMB
       Estimate  StdErr z.value p.value
histol  -0.3099  0.0791   -3.92   9e-05 ***
age      0.0292  0.0107    2.74  0.0062 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two types, Gehan's type and logrank, of rank-based approach yield similar results. The points estimates from case-cohort design are close to these in full cohort analysis, with their standard errors taken into consideration. All the standard errors increased compared to the full cohort analysis, which is expected as some information became unavailable for all covariates. Moreover, the coefficients for the two covariates are found to be significantly different from zero under case-cohort design. This result is also found in the full cohort analysis.

6.3.2 Kidney Catheter Data

An bivariate example for illustration is the Kidney Catheter Data from the **survival** package as `kidney` (McGilchrist and Aisbett, 1991). The interest of the study is to examines the recurrence times to infection at point of catheter insertion for kidney patients using portable dialysis equipment. The data contains 38 patients that each has exactly two observations. The two observations correspond to the recurrence times which measures the time between catheter insertion annd infection at where the catheter is inserted. Catheter is removed when the infection occurs. After some pre-determined time, the catheter is then reinserted. In addition to infection, catheters may be removed for other reasons, in which case the time to infection is treated as censored. In such, the second recurrence time may also be censored if the follow-up period terminated prior to infection. Among the two recurrence time, patient's age (`age` in years) and gender (`sex` = 0 if male, `sex` = 1 if female) are also included in the dataset.

We first fit a bivariate AFT model with identical error margins and identical regression coefficients for the two margins. Since it is reasonable to expect some correlation between the two recurrence times for a given patient, we can model this by least squares approach with some dependent working covariance structure. However, to accounting for multivariate dependence, we will first fit the least squares approach with working independent covariance structures then with exchangeable working structure. For both least squares approaches we use the induced smoothing rank-based estimator with Gehan's weight as initial estimator. The standard errors are estimated by multiplier resampling

method with bootstrap size 1000. The models can be obtained with the following commands.

```
> data("kidney")
> kfit.ind <- aftgee(Surv(time, status) ~ age + sex, id = id,
+                     data = kidney, binit = "srrgehan")
> kfit.ex <- aftgee(Surv(time, status) ~ age + sex, id = id,
+                   data = kidney, corstr = "ex",
+                   binit = "srrgehan")
> summary(kfit.ind)

Call:
aftgee(formula = Surv(time, status) ~ age + sex, data = kidney,
    id = id, binit = "srrgehan")

AFTGEE Estimator
            Estimate   StdErr z.value p.value
(Intercept)  2.07063  0.74311    2.79 0.00533 **
age         -0.00526  0.00849   -0.62 0.53552
sex          1.37386  0.35504    3.87 0.00011 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(kfit.ex)

Call:
aftgee(formula = Surv(time, status) ~ age + sex, data = kidney,
    id = id, corstr = "ex", binit = "srrgehan")

AFTGEE Estimator
            Estimate   StdErr z.value p.value
(Intercept)  2.06989  0.60050    3.45 0.00057 ***
age         -0.00524  0.00888   -0.59 0.55502
sex          1.37382  0.31527    4.36 1.3e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient of `sex` is found to be significantly different from zero for both models. This suggests that female patients tend to have longer recurrence times to infection.

In addition to the common error margin and common coefficient assumption, we also consider a more complicate case where the marginal error distributions and regression coefficients are different. In this case, we need to specify `margin` and construct the corresponding block diagonal design matrix.

```
> kidney$margin <- as.factor(rep(1:2, 38))
> head(model.matrix( ~ age:margin + sex:margin - 1, data = kidney))

  age:margin1 age:margin2 margin1:sex margin2:sex
1          28           0           1           0
2           0          28           0           1
3          48           0           2           0
4           0          48           0           2
5          32           0           1           0
6           0          32           0           1
```

Once the block diagonal design matrix is constructed, least squares estimator with both independent covariance working structure and exchangeable working structure are fitted. For each model, we continue to use the induced smoothing rank-based estimator with Gehan's weight.

```
> kfit2.ind <- aftgee(Surv(time, status) ~ age:margin + sex:margin + margin,
+                     id = id, margin = margin,
+                     data = kidney, binit = "srrgehan")
> kfit2.ex <- aftgee(Surv(time, status) ~ age:margin + sex:margin + margin,
```

```
+                     id = id, margin = margin, data = kidney,
+                     corstr = "ex", binit = "srrgehan")
> summary(kfit2.ind)

Call:
aftgee(formula = Surv(time, status) ~ age:margin + sex:margin +
    margin, data = kidney, id = id, margin = margin, binit = "srrgehan")

AFTGEE Estimator
            Estimate   StdErr z.value p.value
(Intercept)  1.67602  0.83877    2.00   0.046 *
margin2      0.86643  1.03083    0.84   0.401
age:margin1 -0.01335  0.01190   -1.12   0.262
age:margin2  0.00526  0.01353    0.39   0.698
margin1:sex  1.74379  0.42804    4.07 4.6e-05 ***
margin2:sex  0.89353  0.46191    1.93   0.053 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(kfit2.ex)

Call:
aftgee(formula = Surv(time, status) ~ age:margin + sex:margin +
    margin, data = kidney, id = id, margin = margin, corstr = "ex",
    binit = "srrgehan")

AFTGEE Estimator
            Estimate   StdErr z.value p.value
(Intercept)  1.67216  0.86395    1.94 0.05293 .
margin2      0.87218  0.86429    1.01 0.31291
age:margin1 -0.01326  0.01109   -1.20 0.23170
age:margin2  0.00547  0.01213    0.45 0.65212
margin1:sex  1.74389  0.46272    3.77 0.00016 ***
margin2:sex  0.88730  0.51225    1.73 0.08325 .
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors for all coefficients decreased as the working structure changes form independent to exchangeable. Under exchangeable working structure, the gender effects are significant for both margins.

6.4 Conclusion

For right censored data, the relative risk model has been widely used in the analysis. The interpretation of the results from the relative risk model is done with the concept of hazard ratio. In this regard, the AFT model provides an attractive alternative by providing a more direct physical interpretation. Package `aftgee` provides easy access to fitting semiparametric AFT models via both the rank-based approach and the least squares approach. For the rank-based approach, package `aftgee` extended the induced smoothing procedure to a broad class of rank estimators. Multiple variance estimators are also included to improve computational efficiency. In addition, each rank-based implementation is made possible to handle missing data by incorporating weight. For the least squares approach, we use the smooth rank-based estimator with Gehan's weight as initial then use GEE to proceed the iteration. By combining principle of GEE, efficiency is gained when within cluster dependence is strong. With functions `smoothrr` and `aftgee`, we fulfilled the shortage of a reliable software in AFT model and aim to bring AFT model into routine survival analysis.

The current version of `aftgee` allows weights for handling missing data. Similar weight can be applied to our least squares approach. For the rank-based approach, Wang and Fu (2011) proposed a way to incorporate within cluster correlations in estimating the asymptotic variance by decomposing the estimating equations into between and within cluster estimating equations. It would be worthwhile to consider including this method in the future version of `aftgee`. Moreover, to account for possible error in covariates **simexaft** package (He et al., 2012) implemented a simulation-extrapolation approach for AFT models. Such approach can be extended into semiparametric AFT model.

Chapter 7

Future Research

In this dissertation, we have studied statistical methods raised from scenarios in both univariate and multivariate failure time data. Extensions of the research include, but are not limited to improving the proposed methods on AFT models and making them more comprehensive and accessible. There are several possible directions the work could be expanded. For example, our case-cohort weight was constructed based off of an idea commonly used in missing data analysis, the inverse of inclusion probability. It is worthwhile to consider different case-cohort weights, such as the local average type of weights and the time-dependent weighting scheme used in Barlow (1994), to improve efficiency.

For multivariate failure time data raised from case-cohort studies, the estimating equations can be decomposed into between-cluster estimating equations and within-cluster estimating equations. Combining them can substantially improve efficiency when strong multivariate dependence exists (Wang and Fu, 2011). Such reconstruction methods with induced smoothing estimating have been studied by Fu et al. (2010) but they have not yet been extended to adopt case-cohort weights. Incorporating case-cohort

weights might cause more complications in its variance estimator because possible dependence from case-cohort sampling schemes need to be considered.

The GEE approach accounts for within-cluster correlation through a working correlation structure. The success from this approach is encouraging. It is possible to incorporate case-cohort weights on the least squares approach by modifying the Buckley-James estimator according to case-cohort weights. Yu et al. (2007) and Yu and Yu (2007)have studied the Buckley-James estimator under case-cohort designs by utilizing non-parametric likelihood. Yu (2011) further extended the generalized Buckley-James estimator based on inverse weighted estimating equations. Nevertheless, little has been discussed in improving efficiency in multivariate case. Accounting for dependence with our GEE approach, the performance can be compared to the rank-based approach where the within-cluster dependence are account by decomposition. Furthermore, an equivalence relationship might be achieved.

Partial linear models have proven useful, especially when the dependence of response on one of the covariates is not certain and also not our main interest. There is little discussion on the estimation method for the semiparametric AFT partial linear model. It is possible to extend this approach to incorporate a case-cohort study from a weighted least squares aspect.

Appendix A

Appendix

A.1 Analytical Details for $S_i(\beta)$

We give the analytical form of $S_i(\beta)$'s here. Recall the general rank based weighted estimating function (Jin et al., 2003) defined in Chapter 4,

$$U_n(\beta) = \sum_{i=1}^{n} \Delta_i \varphi_i(\beta) \left[X_i - \frac{W_{n,i}^{(1)}(\beta)}{W_{n,i}^{(0)}(\beta)} \right],$$

where $\varphi_i(\beta)$ is an nonnegative weight function and

$$W_{n,i}^{(d)}(\beta) = \frac{1}{n} \sum_{j=1}^{n} X_j^k I[e_j(\beta) \geq e_i(\beta)], \qquad d = 0, 1.$$

Equation (2.1) can be obtained by setting $\varphi_i(\beta) = W_{n,i}^{(0)}(\beta)$. On the other hand, the general rank based weighted estimating function for case-cohort samples has the following form:

$$U_n^c(\beta) = \sum_{i=1}^{n} \Delta_i \varphi_i(\beta) \left[X_i - \frac{\hat{W}_{n,i}^{(1)}(\beta)}{\hat{W}_{n,i}^{(0)}(\beta)} \right],$$

where

$$\hat{W}_{n,i}^{(d)}(\beta) = \frac{1}{n}\sum_{j=1}^{n} h_j X_j^d I[e_j(\beta) \geq e_i(\beta)], \qquad d = 0, 1.$$

Similarly, equation (2.2) can be obtained by setting $\varphi_i(\beta) = \hat{W}_{n,i}^{(0)}(\beta)$.

With these settings, an explicit form of $S_i(\beta_0)$ is

$$\begin{aligned} S_i(\beta_0) =& \int_{-\infty}^{\infty} w^{(0)}(\beta_0)\left[X_i - \frac{w^{(1)}(\beta_0)}{w^{(0)}(\beta_0)}\right] \mathrm{d}M_i(t) \\ =& \Delta_i w^{(0)}(\beta_0)\left[X_i - \frac{w^{(1)}(\beta_0)}{w^{(0)}(\beta_0)}\right] - \int_{-\infty}^{e_i(\beta)} w^{(0)}(\beta_0)\left[X_i - \frac{w^{(1)}(\beta_0)}{w^{(0)}(\beta_0)}\right]\lambda(t)\,\mathrm{d}t, \end{aligned}$$

where

$$w^{(d)}(\beta) = \lim_{n\to\infty} \hat{W}_{n,i}^{(d)}(\beta), \text{ for } d = 0, 1,$$

$$M_i(t) = N_i(\beta; t) - \int_0^t I(e_i(\beta) \geq u)\lambda(u)\,\mathrm{d}u,$$

$N_i(\beta; t) = \Delta_i I(e_i(\beta) \leq t)$ and $\lambda(u)$ is the common hazard function of ϵ_i.

The unknown quantities in $S_i(\beta_0)$ include β_0, $w^{(0)}$, $w^{(1)}$ and $\lambda(t)$. With the explicit form of $S_i(\beta_0)$, $\hat{S}_i(\hat{\beta})$ is obtained by replacing these unknown quantities by their sample estimators.

A.2 Proof of Theorems 1 and 2

Here we provide a brief sketch of the proof of Theorems 1 and 2. First, we impose the following regularity conditions:

A1: The parameter space $\mathbb{B}$ containing β_0 is a compact set of $\mathbb{R}^p$.

A2: $\sum_{k=1}^{K} \|X_{ik}\| + K$ is bounded almost surely by a nonrandom constant $(i = 1, \ldots, n)$.

A3: $\mathrm{Var}(\epsilon_{11}) < \infty$.

A4: The matrix $A(\beta_0)$ is nonsingular.

A5: Let $f_0(\cdot)$ denote the marginal density associated with model error term ϵ_{11}. Then, $f_0(\cdot)$ and $f_0'(\cdot)$ are bounded functions on $\mathbb{R}$ with

$$\int_{\mathbb{R}} \left\{ \frac{f_0'(t)}{f_0(t)} \right\}^2 f_0(t) \mathrm{d}t < \infty$$

A6: The marginal distribution of C_{ik} is absolutely continuous and has a bounded density $g_{ik}(\cdot)$ on $\mathbb{R}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K)$.

A7: $\lim_{n\to\infty} n_s/n = \alpha_s$ and $\lim_{n\to\infty} p_{n,s} = p_s$ for all $s = 1, \ldots, S$, where $0 < \alpha_s, p_s < 1$.

Conditions A1, A2, A4, A5 and A6 are standard and ensure the consistency and asymptotic normal of the solution to (3.1) (Jin et al., 2006c; Tsiatis, 1990; Ying, 1993). Since $|\mathrm{Cov}(\epsilon_{ik}, \epsilon_{il})| \leq \mathrm{Var}(\epsilon_{11})$ $(i = 1, \ldots, n;\ k, l = 1, \ldots, K)$, condition A3 ensures that the covariance between all error terms within a cluster are bounded. This is required for the lemmas in Johnson and Strawderman (2009) to hold. Condition A7 is needed to ensure the desired asymptotic convergence of the stratified samples.

A.2.1 Proof of Theorem 1

First note that equations (3.1) and (3.2) are the gradient of the convex objective function, $L_n(\beta)$ and $L_n^h(\beta)$ with the following forms:

$$L_n^c(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} h_i h_j \Delta_{ik} \left\{ e_{jl}(\beta) - e_{ik}(\beta) \right\} I \left\{ e_{ik}(\beta) - e_{jl}(\beta) \leq 0 \right\},$$

$$L_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} \Delta_{ik} \left\{ e_{jl}(\beta) - e_{ik}(\beta) \right\} I \left\{ e_{ik}(\beta) - e_{jl}(\beta) \leq 0 \right\}.$$

The Lemma 1 in Johnson and Strawderman (2009) states $\sup_{\beta \in \mathbb{B}} |L_n(\beta) - L_0(\beta)| \to 0$. One can also show $\sup_{\beta \in \mathbb{B}} |L_n^c(\beta) - L_n(\beta)| \to 0$ by the strong law of large numbers for U-statistics (Serfling, 2001, Section 5.4), asymptotic convergence results on finite population sampling (Hájek, 1960), and Lemma 1 in Kong et al. (2006). Then, by the triangle inequality,

$$|L_n^c(\beta) - L_0(\beta)| \leq |L_n^c(\beta) - L_n(\beta)| + |L_n(\beta) - L_0(\beta)|,$$

$L_n^c(\beta)$ uniformly converges almost surely to $L_0(\beta)$. Note that, by condition A4, $L_0(\beta)$ is strictly convex at β_0, a unique minimizer of $L_0(\beta)$. Then, the unique minimizer $\hat{\beta}_n^c$ of $L_c^h(\beta)$ converges to β_0 almost surely (Andersen and Gill, 1982).

Under conditions A1–A5 and A7, one can show

$$n^{1/2}(\hat{\beta}_n^c - \beta_0) = -A^{-1}(\beta_0) n^{-1/2} \tilde{U}_n^c(\beta_0) + o_p(1 + n^{1/2} \|\hat{\beta}_n^c - \beta_0\|), \tag{A.2.1}$$

by using the arguments in Ying (1993, Theorem 2). In addition, from Lemma 1 in Jin et al. (2006c), it can be shown that

$$n^{-1/2}U_n^c(\beta_0) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} h_i u_{ik} + o_p(1). \tag{A.2.2}$$

Then, by combining (A.2.1) and (A.2.2), applying Lemma 3 in Kang and Cai (2009b, Supplementary Materials), the desired asymptotic normality of $n^{1/2}(\hat{\beta}_n^c - \beta_0)$ follows.

The explicit forms of $A(\beta_0)$ and $V(\beta_0)$ in $\Sigma(\beta_0)$ can be used to provide closed-form estimators for them. Define limiting quantities $w^{(d)}(t) = \lim_{n\to\infty} W^{(d)}(\beta_0; t)$, $d = 0, 1$, and $\bar{x}(t) = w^{(1)}(t)/w^{(0)}(t)$. Let $\lambda_0(\cdot)$ be the common hazard function for ϵ_{ik}'s, and $M_{ik}(\beta; t) = N_{ik}(\beta; t) - \int_{-\infty}^{t} I\{e_{ik}(t) \geq u\}\lambda_0(u)\mathrm{d}u$. Then,

$$A(\beta) = \lim_{n\to\infty}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_{-\infty}^{\infty} w_0(t)\left\{X_{ik} - \bar{x}(t)\right\}^{\otimes 2}\left\{\frac{\mathrm{d}\log\lambda_0(t)}{\mathrm{d}t}\right\}\mathrm{d}N_{ik}(\beta; t),$$

$$V(\beta) = \lim_{n\to\infty}\mathrm{Var}\left\{n^{1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} h_i u_{ik}(\beta_0)\right\} = \mathrm{E}\left[\left\{\sum_{k=1}^{K} u_{ik}(\beta)\right\}^{\otimes 2}\right] + \sum_{s=1}^{S}\alpha_s\frac{1-p_s}{p_s}\mathrm{Var}_s\left\{\sum_{k=1}^{K} u_{ik}(\beta)\right\},$$

where $\mathrm{Var}_s(\cdot)$ is the variance within the stratum s.

A.2.2 Proof of Theorem 2

Let $\tilde{L}_n^c(\beta)$ be the convex objective function corresponding to equation (3.2) where

$$\tilde{L}_n^c(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} h_i h_j \Delta_{ik} \left(\{e_{jl}(\beta) - e_{ik}(\beta)\} \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right] \right.$$
$$\left. + \frac{r_{ikjl}}{n^{1/2}} \phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right] \right).$$

Then, the consistency of $\tilde{\beta}_n^c$ follows from applying Lemma 2 in Johnson and Strawderman (2009) and using the similar arguments employed in showing the consistency of $\hat{\beta}_n^c$. In order to show that $n^{1/2}(\hat{\beta}_n^c - \beta_0)$ and $n^{1/2}(\tilde{\beta}_n^c - \beta_0)$ converge to the same asymptotic distribution, it suffices to establish the following two convergence results: As $n \to \infty$, (i) $\|\partial \tilde{U}_n^c(\beta)/\partial \beta^\top|_{\beta=\beta_0} - A(\beta_0)\| \to 0$, and (ii) $\|\sqrt{n} \left\{ \tilde{U}_n^c(\beta_0) - U_n^c(\beta_0) \right\}\| \to 0$.

By Lemma 3 in Johnson and Strawderman (2009), $\|\partial \tilde{U}_{n,G}(\beta)/\partial \beta^\top|_{\beta=\beta_0} - A(\beta_0)\| \to 0$. One can also show $\|\partial \tilde{U}_n^c(\beta)/\partial \beta^\top|_{\beta=\beta_0} - A(\beta_0)\| \to 0$ by asymptotic convergence results on finite population sampling (Hájek, 1960) and Lemma 1 in (Kong et al., 2006). Then, (i) follows from the triangular inequality.

Let $\kappa_{ijkl} = \{e_{jl}(\beta) - e_{ik}(\beta)\}/r_{ikjl}$. Then,

$$\left\| \sqrt{n} \left\{ \tilde{U}_n^c(\beta_0) - U_n^h(\beta_0) \right\} \right\|$$
$$\leq \left\| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} h_i h_j \Delta_{ik} (X_{ik} - X_{jl}) \frac{1}{\kappa_{ijkl}} \right\| \cdot \left| \sqrt{n} \kappa_{ijkl} \left\{ \Phi \left(\sqrt{n} \kappa_{ijkl} \right) - I \left(\kappa_{ijkl} \geq 0 \right) \right\} \right|$$
$$= \left\| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{l=1}^{K} h_i h_j \Delta_{ik} (X_{ik} - X_{jl}) \frac{1}{\kappa_{ijkl}} \right\| \cdot \left| \sqrt{n} \kappa_{ijkl} | \Phi \left(-\sqrt{n} | \kappa_{ijkl} | \right) \right|$$

Note that $\lim_{x\to+\infty} x\Phi(-x) = 0$ since $\Phi(-x) \leq (\sqrt{2\pi}x)^{-1}e^{-\frac{x^2}{2}}$. Then, (ii) follows from applying the strong law of large numbers for U-statistics (Serfling, 2001, Section 5.4), asymptotic convergence results on finite population sampling (Hájek, 1960), and Lemma 1 in Kong et al. (2006).

A.3 General Weight

Here we provide a brief sketch of the proof of the following theorem. For $k = 0, 1$, we first define

$$\omega_{n,i}^{(k)}(\beta) = W_{n,i}^{(k)}(\beta + n^{-1/2}\Gamma_n Z), \text{ and}$$
$$\tilde{W}_{n,i}^{(k)}(\beta) = E(\omega_{n,i}^{(k)}(\beta)).$$

We impose the following regularity conditions:

B1: $W_{n,i}^{(k)}(\beta)$ and $\omega_{n,i}^{(k)}(\beta)$ exist and are nonzero for $k = 1, 2$.

B2: $\lim_{n\to\infty} W_{n,i}^{(k)}(\beta)$ and $\lim_{n\to\infty} \omega_{n,i}^{(k)}(\beta)$ exist and are nonzero for $k = 1, 2$.

Condition B1 and B2 also implie $\tilde{W}_{n,i}^{(k)}(\beta)$ and $\lim_{n\to\infty} \tilde{W}_{n,i}^{(k)}(\beta)$ to be nonzero.

A.3.1 Proof of Theorems 3

To prove Theorem 3, one only need to show $E_Z[\omega_{n,i}^{(1)}(\beta)/\omega_{n,i}^{(0)}(\beta)]$ is asymptotically equivalent to $\tilde{W}_{n,i}^{(1)}(\beta)/\tilde{W}_{n,i}^{(0)}(\beta)$. We first obtain that

$$E_Z\left[\frac{\omega_{n,i}^{(1)}(\beta)}{\omega_{n,i}^{(0)}(\beta))}\right] = \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\frac{\omega_{n,i}^{(1)}(\beta)}{\omega_{n,i}^{(0)}(\beta))}\prod_{i=1}^{p}\varphi(z_i)dz_i$$
$$= \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\frac{\omega_{n,i}^{(1)}(\beta))}{\tilde{W}_{n,i}^{(0)}(\beta)}\prod_{i=1}^{p}\varphi(z_i)dz_i + \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\left(\frac{1}{\omega_{n,i}^{(0)}(\beta)} - \frac{1}{\tilde{W}_{n,i}^{(0)}(\beta)}\right)\omega_{n,i}^{(1)}(\beta)\prod_{i=1}^{p}\varphi(z_i)dz_i \tag{A.3.1}$$

The first term in (A.3.1) is the ratio of two expectations. In order to show that the second term in (A.3.1) is asymptotically negligible, it suffices to establish the following result: $\omega_{n,i}^{(0)}(\beta)$ and $\tilde{W}_{n,i}^{(0)}(\beta)$ converges to the same limit.

By triangular inequality,

$$\left|\omega_{n,i}^{(0)}(\beta) - \tilde{W}_{n,i}^{(0)}(\beta)\right| \leq \left|\omega_{n,i}^{(0)}(\beta) - W_{n,i}^{(0)}(\beta)\right| + \left|W_{n,i}^{(0)}(\beta) - \tilde{W}_{n,i}^{(0)}(\beta)\right|.$$

One can show $\left|\omega_{n,i}^{(0)}(\beta) - W_{n,i}^{(0)}(\beta)\right| \to 0$ as $n \to \infty$ by the uniform strong law of large numbers (Pollard, 1990, section 8). On the other hand, since $W_{n,i}^{(0)}(\beta)$ is a monotone function, $\left|W_{n,i}^{(0)}(\beta) - \tilde{W}_{n,i}^{(0)}(\beta)\right| \to 0$ as $n \to \infty$ by Brown and Wang (2005).

A.4 Proof of Theorems 4 and 5

We impose the following regularity conditions:

C1: $\|\mathbf{X}_i\|$ is bounded for all $i = 1, \cdots, n$ where $\|\cdot\|$ is matrix norm.

C2: The density function of $F_{k,\beta}$ exists such that $\int_{-\infty}^{\infty} t^2 \mathrm{d}F_{k,\beta}(t) < \infty$, for $k = 1, \cdots, K$.

C3: The distribution function $F_{k,\beta}$ is twice differentiable with density $f_{k,\beta}$ such that

$$\int_{-\infty}^{\infty} \left(\frac{f'_{k,\beta}(t)}{f_{k,\beta}(t)} \right)^2 \mathrm{d}F_{k,\beta}(t) < \infty$$

where $1 \leq k \leq K$, and both $f_{k,\beta}(t)$ and $f'_{k,\beta}(t)$ are bounded functions.

C4: $E[\exp(\theta \epsilon_{ik}^-)] + \sup_{k \in \{1,\cdots,K\}} E[\exp(\theta C_{ik}^-)] < \infty$ for some $\theta > 0$, where $a^- = |a| I_{\{a \leq 0\}}$.

C5: $\sup_{|b|<\infty; -\infty<t<\infty} \sum_{i=1}^{n} \sum_{k=1}^{K} \Pr(t \leq C_{ik} - X_{ik}^\top b \leq t + h) = O(nh)$ as $h \to 0$ and $nh \to \infty$.

C6: As $n \to \infty$, $\hat{\alpha}_n$ is bounded and is $n^{1/2}$ consistent to α_0 given β.

C7: As $n \to \infty$, initial estimator b_n is $n^{1/2}$ consistent to β_0 and $\sqrt{n}(b_n - \beta_0)$ is asymptoticly normal with zero mean.

C8: The slope matrices $n^{-1} \partial U_n / \partial \beta$ and $n^{-1} \partial U_n / \partial b$ evaluated at $(\beta_0, \beta_0, \alpha_0)$ converge to nondegenerate, finite limit A and B, respectively.

C9: The derivative $\partial \Omega_i^{-1}(\alpha) / \partial \alpha$ is finite for all $i = 1, 2, \ldots n$.

Conditions C1–C5 are standard and ensure the existence of the solution of equation (5.2) (Lai and Ying, 1991). It is natural to assume that the working covariance matrix Ω in equation (5.4) is a symmetric positive definite matrix. Then there exist a $K \times K$ nonsingular matrix, Γ, such that $\Omega(\alpha_0) = \Gamma^{1/2}\Gamma^{1/2}$. Let $\mathbb{X}_i = \Gamma^{-1/2}X_i$, $\mathbb{T}_i = \Gamma^{-1/2}Y_i$, $\mathbb{C}_i = \Gamma^{-1/2}C_i$, and $\omega_i = \Gamma^{-1/2}\epsilon_i$. Then equation (5.4) evaluated at $\alpha = \alpha_0$ can be viewed as equation (5.2) with the transformed data $\mathbb{X}_i$ and $\mathbb{Y}_i = \min(\mathbb{Y}_i, \mathbb{C}_i)$, with error ω_i, $i = 1, \ldots, n$. The existence of the solution to equation (5.4) can be verified by the same arguments as in Lai and Ying (1991), with assumptions similar to C1 to C5 on the transformed data. The consistency and asymptotic normality of the estimator given $\alpha = \alpha_0$ follow from the same arguments as in Jin et al. (2006b).

The extra complexity here comes from the fact that equation (5.4) is solved at $\alpha = \hat{\alpha}_n$, an estimator of α_0. Under condition C9, the ith term in the summation of $\partial U_n/\partial\alpha$ evaluated at $(\beta_0, \beta_0, \alpha_0)$ is a linear function of $\hat{Y}_i(\beta_0) - X_i^\top\beta_0$, $i = 1, \ldots, n$, with expectation zero. By the law of large number, $n^{-1}\partial U_n/\partial\alpha$ evaluated at $(\beta_0, \beta_0, \alpha_0)$ converges to zero in probability.

A.4.1 Proof of Theorem 4

At the solution $\hat{\beta}_n^{(1)}$ given b_n and $\hat{\alpha}_n$, we have $n^{-1}U_n(\hat{\beta}_n^{(1)}, b_n, \hat{\alpha}_n) = 0$. Taylor expansion at $(\beta_0, \beta_0, \alpha_0)$ gives

$$\begin{aligned} 0 =& \frac{1}{n}U_n(\beta_0, \beta_0, \alpha_0) + \frac{1}{n}\frac{\partial}{\partial\beta}\left[U_n(\beta_0, \beta_0, \alpha_0)\right](\hat{\beta}_n^{(1)} - \beta_0) \\ &+ \frac{1}{n}\frac{\partial}{\partial b}\left[U_n(\beta_0, \beta_0, \alpha_0)\right](b_n - \beta_0) + \frac{1}{n}\frac{\partial}{\partial\alpha}\left[U_n(\beta_0, \beta_0, \alpha_0)\right](\hat{\alpha}_n - \alpha_0) + o_p(n^{-1/2}) \\ =& \frac{1}{n}U_n(\beta_0, \beta_0, \alpha_0) + A_n(\hat{\beta}_n^{(1)} - \beta_0) + B_n(b_n - \beta_0) + C_n(\hat{\alpha}_n - \alpha_0) + o_p(n^{-1/2}). \quad \text{(A.4.1)} \end{aligned}$$

With regularity conditions A1–A5, the first term converges in probability to zero by the law of large number. The convergence of b_n and α_n in A6 and A7, combined with the limit condition in A8 and A9, then gives consistency of $\hat{\beta}_n^{(1)}$ to β_0. By induction, $\hat{\beta}_n^{(m)}$ is consistent for β_0 at every m.

A.4.2 Proof of Theorem 5

Under regularity conditions $\sqrt{n}(\hat{\beta}_n^{(1)} - \beta_0)$ can be expressed as

$$\sqrt{n}(\hat{\beta}_n^{(1)} - \beta_0) = [A_n]^{-1}\left[\frac{1}{\sqrt{n}}U_n(\beta_0, \beta_0, \alpha_0) + B_n\sqrt{n}(b_n - \beta_0) + C_n\sqrt{n}(\hat{\alpha}_n - \alpha_0)\right] + o_p(1). \quad \text{(A.4.2)}$$

With condition A9, C_n converges to zero in probability, and, hence, with $\sqrt{n}$ consistency of $\hat{\alpha}_n$, $C_n\sqrt{n}(\hat{\alpha}_n - \alpha_0) = o_p(1)$. Equation (A.4.2) is then asymptotically equivalent to

$$[A_n]^{-1}\left[\frac{1}{\sqrt{n}}U_n(\beta_0, \beta_0, \alpha_0) + B_n\sqrt{n}(b_n - \beta_0)\right].$$

With the assumption that $b_n - \beta_0$ is asymptoticly normal, there exist some nonrandom functions η_i with zero mean such that,

$$\sqrt{n}(b_n - \beta_0) = n^{-1/2}\sum_{i=1}^{n}\eta_i + o_p(\|b_n - \beta_0\|).$$

On the other hand, $U_n(\beta_0, \beta_0, \alpha_0)$ is a sum of independent and identically distributed quantities with zero mean, denoted by ϕ_i's, $i = 1, \ldots, n$. Equation (A.4.2) reduces to

$$\sqrt{n}(\hat{\beta}_n^{(1)} - \beta_0) = [A_n]^{-1}\left[n^{-1/2}\sum_{i=1}^{n}(\phi_i + B_n\eta_i)\right] + o_p(\|b_n - \beta_0\|).$$

By multivariate central limit theorem for sums of independent random vectors, the asymptotic distribution for $\hat{\beta}_n^{(1)}$ is zero mean multivariate normal as $n \to \infty$. The limit covariance matrix Σ have the form $A^{-1}\Phi A^{-1}$, where $\Phi = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\imath_i\imath_i^{\top}$ with $\imath_i = \phi_i + B\eta_i$. Induction then implies that $\hat{\beta}_n^{(m)}$ is multivariate normal for every m.

Bibliography

Andersen, P. and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Barlow, W. E. (1994), "Robust Variance Estimation for the Case-cohort Design," *Biometrics*, 50, 1064–1072.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pagoda, J. (2000), "Exposure Stratified Case-Cohort Designs," *Lifetime Data Analysis*, 6, 36–58.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009a), "Improved Horvitz-Thompson Estimation of Model parameters from Two-Phase Stratified Samples: Applications in Epidemiology," *Statistics in Biosciences*, 1, 32–49.

— (2009b), "Using the Whole Cohort in the Analysis of Case-Cohort Data," *American Journal of Epidemiology*, 169, 1398–1405.

Brostrom, G. (2012), **eha***: Event History Analysis*, R package version 2.0-7.

Brown, B. M. and Wang, Y.-G. (2005), "Standard Errors and Covariance Matrices for Smoothed Rank Estimators," *Biometrika*, 92, 149–158.

— (2007), "Induced Smoothing for Rank Regression with Censored Survival Times," *Statistics in Medicine*, 26, 828–836.

Brown, G. W. and Harris, T. . (1978), *Social Origins of Depression. A Study of Psychiatric Disorder in Women.*, Tavistock Publications, London.

Buckley, J. and James, I. (1979), "Linear Regression with Censored Data," *Biometrika*, 66, 429–436.

Buyske, S., Fagerstrom, R., and Ying, Z. (2000), "A Class of Weighted Log-rank Tests for Survival Data When the Event Is Rare," *Journal of the American Statistical Association*, 95, 249–258.

Caplan, D., Cai, J., Yin, G., and White, B. A. (2005), "Root Canal Filled Versus Non-Root Canal Filled Teeth: A Retrospective Comparison of Survival Times," *Journal of Public Health Dentistry*, 65, 90–96.

Chen, H. Y. (2001a), "Fitting Semiparametric Transformation Regression Models to Data from a Modified Case-Cohort Design," *Biometrika*, 88, 255–268.

— (2001b), "Weighted Semiparametric Likelihood Method for Fitting a Proportional Odds Regression Model to Data from the Case Cohort Design," *Journal of the American Statistical Association*, 96, 1446–1458.

Chiou, S. H., Kang, S., and Yan, J. (2012a), *aftgee: Accelerated Failure Time Model with Generalized Estimating Equations*, r package version 0.2-27.

— (2013a), "Fast Accelerated Failure Time Modeling for Case-cohort Data," *Statistics and Computing*, forthcoming.

— (2013b), "Fitting Accelerated Failure Time Model in Routine Survival Analysis with R Package Aftgee," Tech. Rep. 14, Department of Statistics, University of Connecticut.

— (2013c), "Semiparametric Accelerate Failure Time Modeling for Clustered Failure Times from Stratified Sampling," Tech. Rep. 15, Department of Statistics, University of Connecticut.

Chiou, S. H., Kim, J., and Yan, J. (2012b), "Marginal Semiparametric Multivariate Accelerated Failure Time Model with Generalized Estimating Equations," Tech. Rep. 13, Department of Statistics, University of Connecticut.

Cochran, W. G. (1977), *Sampling Techniques*, John Wiley & Sons.

Cox, D. R. (1972), "Regression Models and Life-Tables (with discussion)," *Journal of the Royal Statistical Society, Series B, Methodological*, 34, 187–220.

D'Angio, G. J., Breslow, N., Beckwith, J. B., Evans, A., Baum, E., Delorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, H. B., Tefft, M., and Thomas, P. R. M. (1989), "Treatment of Wilms' Tumor. Results of the Third National Wilms' Tumor Study," *Cancer*, 64, 349–360.

Diabetic Retinopathy Study Research Group (1976), "Preliminary Report on Effects of Photocoagulation Therapy," *American Journal of Ophthalmology*, 81, 383–396.

Fu, L., Wang, Y.-G., and Bai, Z. (2010), "Rank Regression for Analysis of Clustered Data: A Natural Induced Smoothing Approach," *Computational Statistics & Data Analysis*, 54, 1036–1050.

Fygenson, M. and Ritov, Y. (1994), "Monotone estimating equations for censored data," *The Annals of Statistics*, 22, 732–746.

Gehan, E. A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples," *Biometrika*, 52, 203–223.

Gill, R. D. (1980), *Censoring and Stochastic Integrals*, CWI, Math. Centrum [Centrum voor Wiskunde en Informatica].

Green, D., Breslow, N., Beckwith, J., Finklestein, J., Grundy, P., Thomas, P., Kim, T., Shochat, S., Haase, G., Ritchey, M., Kelalis, P., and D'Angio, G. (1998), "Comparison between Single-dose and Dvided-dose Administration of Dactinomycin and Doxorubicin for Patients with Wilms' Tumor: A Report from the National Wilms' Tumor Study Group," *Journal of Clinical Oncology*, 16, 237–245.

Hájek, J. (1960), "Limiting distributions in simple random sampling from a finite population," *Pub. Math. Inst. Hungar. Acad. Sci.*, 5, 361–374.

Halekoh, U., Højsgaard, S., and Yan, J. (2006), "The R Package geepack for Generalized Estimating Equations," *Journal of Statistical Software*, 15/2, 1–11.

Harnish, J. D., Aseltine, JR., R. H., and Gore, S. (2000), "Resolution of Stressful Experiences as an Indicator of Coping Effectiveness in Young Adults: An Event History Analysis," *Journal of Health and Social Behavior*, 41, 121–136.

Harrell, Jr., F. E. (2012), **rms***: Regression Modeling Strategies*, R package version 3.5-0.

Harrington, D. P. and Fleming, T. R. (1982), "A class of rank test procedures for censored survival data," *Biometrika*, 69, 133–143.

Hasselman, B. (2012), *nleqslv: Solve systems of non linear equations*, r package version 1.9.3.

He, W., Xiong, J., and Yi, G. Y. (2012), "**SIMEX**: R Package for Accelerated Failure Time Models with Covariate Measurement Error," *Journal of Statistical Software, Code Snippets*, 46, 1–14.

Hornsteiner, U. and Hamerle, A. (1996), "A Combined GEE/Buckley-James Method for Estimating an Accelerated Failure Time Model of Multivariate Failure Times," .

Huang, Y. (2002), "Calibration Regression of Censored Lifetime Medical Cost," *Journal of the American Statistical Association*, 97, 318–327.

Huster, W. J., Brookmeyer, R., and Self, S. G. (1989), "Modelling Paired Survival Data with Covariates," *Biometrics*, 45, 145–156.

Jin, Z. and Huang, L. (2007), "**lss**: An S-Plus/R Program for the Accelerated Failure Time Model to Right Censored Data Based on Least-Squares Principle," *Computer Methods and Programs in Biomedicine*, 86, 45–50.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003), "Rank-based Inference for the Accelerated Failure Time Model," *Biometrika*, 90, 341–353.

— (2006a), "Rank Regression Analysis of Multivariate Falure Time Data Based on Marginal Linear Models," *Scandinavian Journal of Statistics*, 33, 1–23.

Jin, Z., Lin, D. Y., and Ying, Z. (2006b), "On Least-squares Regression with Censored Data," *Biometrika*, 93, 147–161.

— (2006c), "Rank Regression Analysis of Multivariate Failure Time Data Based on Marginal Linear Models," *Scandinavian Journal of Statistics*, 33, 1–23.

Johnson, L. M. and Strawderman, R. L. (2009), "Induced Smoothing for the Semiparametric Accelerated Failure Time Model: Asymptotics and Extensions to Clustered Data," *Biometrika*, 96, 577–590.

Kalbfleisch, J. D. and Lawless, J. F. (1988), "Likelihood Analysis of Multistate Models for Disease Incidence and Mortality," *Statistics in Medicine*, 7, 149–160.

Kang, S. and Cai, J. (2009a), "Marginal Hazards Model for Case-cohort Studies with Multiple Disease Outcomes," *Biometrika*, 96, 887–901.

— (2009b), "Marginal Hazards Regression for Retrospective Studies within Cohort with Possibly Correlated Failure Time Data," *Biometrics*, 65, 405– 414.

Kim, S. and Gruttola, V. (1999), "Strategies for Cohort Sampling Under the Cox Proportional Hazards Model, Application to an AIDS Clinical Trial," *Lifetime Data Analysis*, 5, 149–172.

Kong, L. and Cai, J. (2009), "Case-Cohort Analysis with Accelerated Failure Time Model," *Biometrics*, 65, 135–142.

Kong, L., Cai, J., and Sen, P. K. (2004), "Weighted Estimating Equations for Semiparametric Transformation Models with Censroed Data from a Case-cohort Design," *Biometrika*, 91, 305–319.

— (2006), "Asymptotic Results for Fitting Semiparametric Transformation Models to Failure Time Data from Case-cohort Studies," *Statistica Sinica*, 16, 155–151.

Kulich, M. and Lin, D. (2000), "Additive Hazards Regression for Case-cohort Studies," *Biometrika*, 87, 73–87.

— (2004), "Improving the Efficiency of Relative-risk Estimation in Case-cohort Studies," *Journal of the American Statistical Association*, 99, 832–844.

Lai, T. L. and Ying, Z. (1991), "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data," *The Annals of Statistics*, 19, 1370–1402.

Lee, E. W. and Wei, L. J. aand Ying, Z. (1993), "Linear Regression Analysis for Highly Stratified Failure Time Data," *Journal of the American Statistical Association*, 88, 557–565.

Li, H. and Yin, G. (2009), "Generalized Method of Moments Estimation for Linear Regression with Clustered Failure Time Data," *Biometrika*, 96, 293–306.

Liang, K.-Y., Self, S. G., and Chang, Y.-C. (1993), "Modelling Marginal Hazards in Multivariate Failure Time Data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 55, 441–453.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, D. Y. and Ying, Z. (1993), "Cox Regression with Incomplete Covariate Measurements," *Journal of the American Statistical Association*, 88, 1341–1349.

Lu, S.-E. and Shih, J. H. (2006), "Case-cohort Designs and Analysis for Clustered Failure Time Data," *Biometrics*, 62, 1138–1148.

Lu, W. and Tsiatis, A. A. (2006), "Semiparametric Transformation Models for the Case-cohort Study," *Biometrika*, 93, 207–214.

Luo, X. and Huang, C.-Y. (2011), "Analysis of Recurrent Gap Time Data Using the Weighted Risk Set Method and the Modified Within-Cluster Resampling Method," *Statistics in Medicine*, 30, 301–311.

McGilchrist, C. A. and Aisbett, C. W. (1991), "Regression with Frailty in Survival Analysis," *Biometrics*, 47, 461–466.

Nan, B., Yu, M., and Kalbfleisch, J. D. (2006), "Censored Linear Regression for Case-cohort Studies," *Biometrika*, 93, 747–762.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Institute of Mathematical Statistics.

Prentice, R. L. (1978), "Linear Rank Tests with Right Censored Data (Corr: V70 P304)," *Biometrika*, 65, 167–180.

— (1986), "A Case-cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836.

Ritov, Y. (1990), "Estimation in a Linear Regression Model with Censored Data," *The Annals of Statistics*, 18, 303–328.

Robins, J. M. and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology — Methodological Issues*, eds. Jewell, N., Dietz, K., and Farewell, V., Boston, MA: Birkhäuser, pp. 297–331.

Samuelsen, S. O., Ånestad, H., and Skrondal, A. (2007), "Stratified Case-Cohort Analysis of General Cohort Sampling Designs," *Scandinavian Journal of Statistics*, 34, 103–119.

Self, S. G. and Prentice, R. L. (1988), "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies," *The Annals of Statistics*, 16, 64–71.

Serfling, R. J. (2001), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Smith, T. M. F. (2001), "Biometrika Centenary: Sample Surveys," *Biometrika*, 88, 167–243.

Spiekerman, C. F. and Lin, D. Y. (1996), "Checking the Marginal Cox Model for Correlated Failure Time Data," *Biometrika*, 83, 143–156.

Strawderman, R. L. (2005), "The Accelerated Gap Times Model," *Biometrika*, 92, 647–666.

Stute, W. (1993), "Consistent Estimation under Random Censorship When Covariables Are Present," *Journal of Multivariate Analysis*, 45, 89–103.

— (1996), "Distributional Convergence under Random Censorship When Covariables Are Present," *Scandinavian Journal of Statistics*, 23, 461–471.

Sun, J., Sun, L., and Flournoy, N. (2004), "Additive Hazards Model for Competing Risks Analysis of the Case-cohort Design," *Communications in Statistics: Theory and Methods*, 33, 351–366.

Therneau, T. (2012), *A Package for Survival Analysis in S*, r package version 2.36-12.

Therneau, T. M. and Li, H. (1999), "Computing the Cox model for case cohort designs," *Lifetime Data Analysis*, 5, 99–112.

Tsiatis, A. A. (1990), "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *The Annals of Statistics*, 18, 354–372.

Varadhan, R. and Gilbert, P. (2009), "BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function," *Journal of Statistical Software*, 32, 1–26.

Wacholder, S., Gail, M. H., Pee, D., and Brookmeyer, R. (1989), "Alternative variance and efficiency calculations for the case-cohort design," *Biometrika*, 76, 117–123.

Wang, M.-C. and Chang, S.-H. (1999), "Nonparametric Estimation of a Recurrent Survival Function," *Journal of the American Statistical Association*, 94, 146–153.

Wang, Y.-G. and Fu, L. (2011), "Rank Regression for Accelerated Failure Time Model with Clustered and Censored Data," *Computational Statistics and Data Analysis*, 55, 2334–2343.

Yan, J. and Fine, J. P. (2004), "Estimating Equations for Association Structures," *Statistics in Medicine*, 23(6), 859–874.

Ying, Z. (1993), "A Large Sample Study of Rank Estimation for Censored Regression Data," *The Annals of Statistics*, 21, 76–99.

Yu, M. (2011), "Buckley-James Type Estimator in Censored data with Covariates Missing by Design," *Scandinavian Journal of Statistics*, 38, 252–267.

Yu, Q., Wong, G. Y. C., and Yu, M. (2007), "Buckley-James-type of Estimators under the Classical Case Cohort Design," *Annals of the Institute of Statistical Mathematics*, 59, 675–695.

Yu, Q. and Yu, M. (2007), "Estimation with Modified Case Cohort Data under Linear Regression Models," *Journal of Applied Probability and Statistics*, 2, 49–70.

Zeng, D. and Lin, D. Y. (2008), "Efficient Resampling Methods for Nonsmooth Estimating Functions," *Biostatistics*, 9, 355–363.

Zhang, H., Schaubel, D. E., and Kalbfleisch, J. D. (2011), "Proportional Hazards Regression for the Analysis of Clustered Survival Data from Case-cohort Studies," *Biometrics*, 67, 18–28.

Zhou, M. (1992), "M-estimation in Censored Linear Models," *Biometrika*, 79, 837–841.