STATISTICAL METHODS FOR TESTING TREATMENT-COVARIATE INTERACTIONS IN CANCER CLINICAL TRIALS

by

Shifang Liu

A thesis submitted to

the Department of Mathematics and Statistics

In conformity with the requirements for

the degree of Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

September, 2011

Abstract

Treatment–covariate interaction is often used in clinical trials to assess the homogeneity of treatment effects over those subgroups defined by a baseline covariate, which is frequently conducted after primary analysis including all patients is completed. When the endpoint is the time to an event, as in the cancer clinical trials, the Cox proportional hazard model with an interaction term has been used exclusively to test the significance of treatment-covariate interaction in oncology literature. But the proportional hazards assumption may not be satisfied by the data from clinical trials. Although there are several procedures proposed in statistical literature to assess the interaction based on a nonparametric measure of interaction or nonparametric models, some of these procedures do not take into the account of the nature of the data well, while some are very complicated which may have limited their applications in practice. In this thesis, a non-parametric procedure based on the smoothed estimate of Patel–Hoel measure is first derived to test the interaction between the treatment and a binary covariate with censored data. The theoretical distribution of the test statistic of the proposed procedure is derived. The proposed procedure is also evaluated through Monte-Carlo simulations and applications to data from a cancer clinical trial. Jackknifed versions of two test statistics based on nonparametric models are then derived by simplifying these test statistics and applying the jackknife method to estimate their variances. These jackknifed tests are also compared with the smoothed test and other related tests.

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor Dr. Dongsheng Tu for his guidance and encouragement during my study. His broad and profound knowledge has been of great value for me. Grateful appreciation is also extended to members of my supervisory committee, Drs. Glen Takahara and Miu Lam, for their feedback and guidance. I would like also to thank the NCIC Clinical Trials Group (NCIC-CTG) for allowing me to use their data for my thesis work. Finally, I greatly appreciate the patience and encouragement of my wife Qimin and sons Jason and Timothy during the completion of this thesis.

Statement of Originality

I hereby certify that all of the work described within this thesis is the original work of the author. Any published (or unpublished) ideas and/or techniques from the work of others are fully acknowledged in accordance with the standard referencing practices.

Shifang Liu

September, 2011

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Assessment of Treatment-Covariate Interaction in Cancer Clinical Trials

Although the major interest in clinical trials is to assess the treatment effect over all subjects included in the trials, it is also often of interest to assess the heterogeneity of the treatment effects across subgroups defined by baseline characteristics of subjects to identify whether treatment effects in certain subgroup is larger than that in other subgroups. The standard approach used for this assessment is the test of interaction between the treatments and covariate. The main objective of my thesis research is to study some statistical procedures for the test of treatment-covariate interactions. In this section, I will use the correlative or translational study in cancer clinical trials as an exclusive example for the background and motivation for my thesis project.

A correlative or translational study in a cancer clinical trial is a major step in cancer research to translate the knowledge obtained on the "bench" (cancer research laboratory) to the "bed" (clinics), that is, the knowledge on the new tumor biomarkers

discovered in the basic and laboratory cancer research to the routine clinical practice. Tumor biomarkers are molecules occurring in blood or tissue produced by a tumor or released by the host body in response to a tumor, which can be cellular characteristics such as cytogenetic markers, oncogenes and their protein products and abnormally expressed cell surface molecules. A biomarker is called predictive if it allows the prospective identification of individuals who will or will not benefit from the use of a particular therapy. A decision on how the patients are treated can be made based on the information from the biomarkers.

The major statistical method used to determine whether a biomarker is predictive of treatment is the test of the interaction between the treatment and the expression level of the biomarker, which can be continuous or categorical (such as presence or absence of the biomarker) (Pritchard at al., 2003). I will consider a binary biomarker only for simplicity. When the clinical outcome of a tumor correlative study is time to an event, for example, survival time, the test for the significance of the interaction term in a Cox proportional hazard model with interaction term has been used exclusively in the clinical literature to identify the predictive biomarkers. In the following, I first give a brief introduction for this method.

Denote that X_1 and X_2 are the two variables representing respectively the treatment and biomarker, where

$$X_1 = \begin{cases} 1 & \text{if the patient received treatment A} \\ 0 & \text{if the patient received treatment B} \end{cases}$$

and

$$X_2 = \begin{cases} 1 & \text{if the biomarker is present in the tumour sample} \\ 0 & \text{if the biomarker is absent from the tumour sample.} \end{cases}$$

Then the Cox proportional hazard model that is used to assess the predictive ability of this biomarker can be defined as

$$h(t;\ x_1,\ x_2) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2),$$

where $h(t;\ x_1,\ x_2)$ is the hazard rate of a patient with given values x_1 and x_2 for the above two variables, $h_0(t)$ the unknown baseline rate, β_1, β_2 and β_{12} unknown parameters. With this model, the assessment of whether a biomarker is predictive of treatment can be formulated statistically as the testing of the following hypotheses:

$$(1) \qquad H_0: \beta_{12}=0 \quad \text{versus} \quad H_1: \beta_{12}\neq 0.$$

Let $h_{ij}(t)$ be the hazard rate of a patient who received the i-th treatment (i=A, B) and was at the j-th level of the biomarker (j=present or absent). Define the hazard ratio of treatment A to B for the patient with the j-th level of the biomarker as

$$\rho_j(t) = \frac{h_{1j}(t)}{h_{2j}(t)}.$$

Then it can be shown that the above hypotheses for the coefficient of the interaction term of the Cox model are equivalent to the following hypotheses:

(2) H_0: $\rho_1(t)=\rho_2(t)$ for all t versus H_1: $\rho_1(t)\neq\rho_2(t)$ for at least one t.

The hypotheses in (1) are based on an implied assumption from the Cox model that $\rho_1(t)$ and $\rho_2(t)$ are constants (the famous proportional hazards assumption), while such assumption is not required in (2).

The proportional hazards assumption may not be satisfied by the data from clinical trials. Tu and Pater (1999) provided some examples from the clinical trials conducted by the NCIC Clinical trials Group. Although the test on this assumption is usually performed, the power of the test procedures used may be very low. An alternative method is needed, especially if the statistical test or practical background indicates possible violation of this assumption.

In the literature, some procedures have been proposed to assess the interaction based on a nonparametric measure of interaction or nonparametric model. From the review of the literature on these procedures, I found that some of these procedures did not take into the account of the data well. For example, Schemper (1988) proposed a

non-parametric method by extending Patel and Hoel (1973) procedure to the censored data. This method is based on comparison of every pairs of observed survival times in two treatment groups. If observed survival times in a given pair are both censored, the comparison cannot be made because their true survival times are unknown. These pairs have to be discarded in the calculation of the test statistic, which may cause problem when censoring rate is moderate or high. Some other procedures are based on some nonparametric statistical models and associated test statistics are very complicated because they involve complicate functions of counting processes, which may have limited their applications in practice. In this thesis, I propose to overcome these problems by introducing a smoothed estimate for a nonparametric measure of interaction based on all the data and applying resampling methods to estimate complicated variances associated with the complicated procedures proposed in the literature.

1.2 Objectives of This Thesis

As more and more personalized medicine advance, there is an increasing need of a general procedure for testing the interaction between the treatment and biomarkers. As mentioned above, currently, the interaction test in the presence of censoring data are

most commonly performed using the Cox's proportional hazards model, although there were several nonparametric procedures proposed in the literature. It is known that proportional hazard assumption may not be held by the data from clinical trials and nonparametric methods are preferred. Some of the reasons that nonparametric procedures have not been used in statistical practice may be due to the lower power of proposed procedures or the complicated calculation of test statistics and its covariance matrix. The major objective of this research is to explore the improvements to these nonparametric statistical procedures for testing the interaction by introducing a smoothed estimate for a nonparametric measure of interaction based on the data and applying re-sampling methods to estimate complicated variances associated with some of these procedures. Specifically, the following are the objectives of this thesis:

1. Develop a nonparametric test based on the smoothed estimate of Patel-Hoel measure and censored data and a jackknife approximation (Shao and Tu, 1995) to the variance of the proposed test statistics since it is anticipated to be difficult to calculate because of censoring.

2. Develop also a jackknife approximation to the variance of the test statistics proposed by respectively Akritas and Bruner (1997) and Akritas and LaValley (1996).

3. Investigate the behavior of the proposed and modified procedures and compare the proposed procedures with those derived under proportional hazard assumptions in term of actual type I error when proportional hazard assumption holds or not holds through Monte-Carlo simulations and applications to real data from cancer clinical trials.

1.3 Organization of This Thesis

In Chapter 2, we give a detail review of the statistical methods for testing the interaction between two factors based on the non-parametric measures of interaction. A detail review of statistical method for testing interaction based on the non-parametric models will be presented in Chapter 3. Chapter 4 provides a brief overview of the kernel estimation of density and jackknife and bootstrap estimations of the variance, which are main statistical tools for the research presented in this thesis. In Chapter 5, a new nonparametric test of the interaction between the treatment and covariate based on the smoothed estimate of Patel-Hoel measure and censored data will be developed. The asymptotic distribution of the test statistics is derived. Monte-Carlo simulations are employed to investigate the performance of proposed test procedure. An application to a real clinical trial is also presented. In Chapter 6, several

general interaction test statistics based on nonparametric models are first simplified for the cases where both treatment and covariate have two levels and the jackknife method is then applied to estimate variances for these test statistics. The results of simulation studies which evaluate the methods proposed in this thesis and their applications to data from clinical trials are presented in Chapter 7. The final chapter presents some summary and discussions of results in this thesis and also some further research questions. The R programs which could be used to implement the procedures developed in this thesis are given as an appendix to this thesis.

Chapter 2

Interactions Tests Based on a Non-parametric Measures of Interaction

For a two by two factorial design, Patel and Hoel (1973) proposed a nonparametric measure of interaction and constructed a nonparametric test of interaction based on this measure. Brunner et al (1995) and Schemer (1988) extended this procedure to general factorial designs with respectively completed or censored data. I review the original idea of Patel and Hoel and the extensions made by Brunner et al. and Schemer in details in this chapter.

2.1 Methods for Two by Two Factorial Design

Consider a two-factor experiment with each factor having two possible levels. Let X_{ijk} designate the response of the k-th replicate receiving the i-th level of factor A and the j-th level of factor B. Assume that the response can be expressed linearly as

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad 1 \le i \le 2, 1 \le j \le 2, 1 \le k \le n_{ij}, \qquad (2.1.1)$$

where μ_{ij} is the effect of the i-th level of factor A and the j-th level of factor B. It is also assumed that random errors $\{\varepsilon_{ijk}\}$ are independent and identically distributed each with the continuous distribution function G. Then under the assumption of this linear model, the quantity

$$\mu^* = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$$

can be used as a measure of interaction between the factor A and B and there is usually of interest in testing whether or not $\mu^* = 0$. If the distribution *G* is normal, the common *F* test may be used to test the hypothesis of no interaction. However, when *G* is not normal, it may be more appropriate to apply non-parametric test procedures.

2.1.1 A Nonparametric Measures of Interaction

Patel and Hoel (1973) proposed a nonparametric measure of interaction for a 2x2 factorial design, which is defined as

$$\mu = P(X_{12} \leq X_{11}) - P(X_{22} \leq X_{21}), \qquad (2.1.2)$$

where the replicate subscript k of X_{ijk} has been suppressed. This measure can be easily interpreted and does not require the linear model assumption. In fact, μ is invariant under continuous strictly increasing transformations of the X_{ij}.

Let X be governed by a distribution function F. Denoted by C the class of distribution function F such that the distribution function of $X_1 - X_2$ is continuous and strictly increasing on the real line, where X_1 and X_2 are independent copies of X. Note that the class C includes many common continuous distribution functions, for instance, the normal and gamma distribution. If $G \in C$, then it can be shown that $\mu = 0$ if and only if $\mu^* = 0$. Thus we can consider a test of $\mu = 0$ to be essentially the test of no interaction ($\mu^* = 0$) in the linear model.

2.1.2 An Nonparametric Estimator of μ

To find a consistent, unbiased, and asymptotically normal estimator of μ given by (2.1.2), define

$$\psi(X_{ijk}, X_{i'j'k'}) = \begin{cases} 0 & if \quad X_{ijk} < X_{i'j'k'} \\ 1/2 & = \\ 1 & > \end{cases} \qquad (2.1.3)$$

and introduce the Wilcoxon-Mann-Whitney statistics

$$V_i = \sum_{j=1}^{n_{i1}} \sum_{k=1}^{n_{i2}} \psi(X_{i1j}, X_{i2k}) / (n_{i1} n_{i2}) . \qquad i = 1, 2 . \qquad (2.1.4)$$

From Puri and Sen (1970), we know that V_i is a minimum variance unbiased, consistent, and asymptotically normal estimator of

$$P_i = \int F_{i2} dF_{i1} = P\{X_{i2} \le X_{i1}\}, \qquad \text{i=1, 2} \qquad (2.1.5)$$

where F_{ij} is the distribution of X_{ij}. Thus, since $\mu = P_1 - P_2$, a natural estimator of μ would be

$$T = V_1 - V_2 \qquad (2.1.6)$$

If we denote by R_{ijk} the rank of X_{ijk} in the ordered array of $\{X_{ijk},\ 1 \le k \le n_{ij},\ j = 1, 2\}$, the statistic T then reduces to

$$T = \frac{1}{n_{11}n_{12}}\left(\sum_{k=1}^{n11} R_{11k} - \frac{n_{11}(n_{11}+1)}{2}\right) - \frac{1}{n_{21}n_{22}}\left(\sum_{k=1}^{n21} R_{21k} - \frac{n_{21}(n_{21}+1)}{2}\right)$$

In the case where sample size in all groups are equal (e.g., $n_{ij} = n$), it would be further simplified to

$$T = \left(\overline{R}_{11} - \overline{R}_{21}\right)/n,$$

where

$$\overline{R}_{i1} = \sum_{k=1}^{n} R_{i1k}/n, \qquad \text{i=1, 2.}$$

Denote

$$\upsilon = \min\{n_{11}, n_{12}, n_{21}, n_{22}\}. \qquad (2.1.7)$$

It can be verified that the variance of $\upsilon^{\frac{1}{2}}V_i$ is

$$\sigma_i^2 = \frac{\upsilon}{n_{i1}n_{i2}}\left\{\int F_{i2}dF_{i1} + (n_{i1}-1)\int (1-F_{i1})^2 dF_{i2} + (n_{i2}-1)\int F_{i2}^2 dF_{i1} - (n_{i1}+n_{i2}-1)\left(\int F_{i2}dF_{i1}\right)^2\right\}$$

where i=1, 2. (2.1.8)

Define

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \qquad (2.1.9)$$

It can be shown that, for $\sigma^2 \neq 0$, we have

$$\lim_{\upsilon\to\infty} P\left\{\upsilon^{\frac{1}{2}}(T-\mu)/\sigma \leq z\right\} = \Phi(z)$$

where μ and σ^2 are given by (2.1.2) and (2.1.9), respectively, and $\Phi(\cdot)$ is the standard normal distribution function.

2.1.3 An Approximate Nonparametric Test for Interaction

The existence of interaction between two factors in the 2x2 factorial design can be assessed by testing the following null hypotheses:

$$H_0 : \mu = 0$$

against the alternative hypothesis,

$$H_1 : \mu \neq 0 .$$

Assume σ^2 can be estimated by a consistent estimate $\hat{\sigma}^2$. Based on the asymptotic results in the last section, one can obtain an approximate α level test which rejects H_0 if

$$\left| \frac{\upsilon^{\frac{1}{2}} T}{\hat{\sigma}} \right| \geq z_{1-\alpha/2}, \qquad (2.1.10)$$

where

$$\Phi(z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

A consistent and distribution-free estimate of σ^2 can be obtained by replacing the various distribution functions in (2.1.8) with their corresponding empirical distribution functions and substituting them into (2.1.9).

2.2 Methods for n x 2 Factorial Design

Motivated by some problems arising from the analysis of data from multicenter clinic trails, Brunner et al (1995) extended the procedure developed by Patel and Hoel (1973) given in Section 2.1 to an n x 2 factorial design.

Let $\{X_{ijk}, k = 1, ..., m_{ij}; i=1, \ldots, n; j = 1, 2\}$ be a set of $M = \sum_{i=1}^{n} \sum_{j=1}^{2} m_{ij}$ independent real valued random variables and F_{ij} the distribution function of X_{ij} (i.e.,

$F_{ij}(x) = P(X_{ijk} \le x)$). Assume the random variables X_{ijk} satisfy the following linear model:

$$X_{ijk} = \xi_i + \mu_j + \theta_{ij} + \varepsilon_{ijk}, \qquad k = 1, \ldots, m_{ij}; \qquad i=1, \ldots, n; \quad j = 1, 2 \qquad (2.2.1)$$

where $\sum_{i=1}^{n} \xi_i = \sum_{j=1}^{2} \mu_j = \sum_{i=1}^{n} \theta_{ij} = \sum_{j=1}^{2} \theta_{ij} = 0$, ξ_i (i = 1, . . . , n) and μ_j (j = 1, 2) are two factors of fixed effect; θ_{ij}, i = 1, ..., n; j = 1, 2 is the (fixed) interaction effect between levels of two factors; and ε_{ijk}, k= 1, ... ,m_{ij}; i=1, …, n; j = 1, 2 are random errors that are independent and identically distributed with a common distribution function $F(x)$. Under this linear model and if the distribution F is normal, an F test may be used to test the hypothesis of no interaction (i.e., $\theta_{ij} = 0$).

Let

$$p_i = P(X_{i11} \le X_{i21}).$$

Brunner et al (1995) defined

$$\rho_n = \sum_{i=1}^{n} (p_i - \tau_n)^2$$

as a generalized measure of overall interaction, where

$$\tau_n = (1/n) \sum_{i=1}^{n} p_i .$$

If model (2.2.1) is satisfied and $F(x)$ is strictly increasing, it can be shown that:

$$\rho_n = 0 \quad \text{if and only if} \quad \sum_{i=1}^{n} \theta_i^2 = 0 .$$

Therefore, we can see that ρ_n is a nonparametric generalization of the measurement for the interaction effect under a linear model.

Let

$$P = (p_1, \cdots, p_n)' .$$

It is easy to see that it can be estimated by a nonparametric estimate

$$\hat{P} = (\hat{p}_1, \cdots, \hat{p}_n)' ,$$

where $\hat{p}_i, i = 1, \cdots, n$ are given by

$$\hat{p}_i = \frac{1}{m_{i1}} \left(\overline{R}_{i2\cdot} - \frac{m_{i2} + 1}{2} \right) \qquad (2.2.2)$$

with

$$\overline{R}_{ij\cdot} = m_{ij}^{-1} \sum_{k=1}^{m_{ij}} R_{ijk} ,$$

$$R_{ijk} = \sum_{s=1}^{2} \sum_{t=1}^{m_{is}} c(X_{ijk} - X_{ist}) ,$$

and $c(u) = 0$ or 1 according to u <0 or u≥ 0.

It was shown that under $H_0 : \rho_n = 0$, the statistics

$$Q_M = \sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2} \left(\hat{p}_i - \frac{1}{\sum_{j=1}^{n} 1/\hat{\sigma}_j^2} \sum_{j=1}^{n} \frac{\hat{p}_j}{\hat{\sigma}_j^2} \right)$$

has a central χ^2_{n-1} distribution, where

$$\hat{\sigma}_i^2 = \frac{1}{m_{i1} m_{i2}} \left(\sum_{j=1}^{2} \frac{M_i - m_{ij} - 1}{m_{ij} (M_i - m_{ij})^2} \sum_{k=1}^{mij} \left[R_{ijk} - R_{ijk}^{(j)} - \overline{R}_{ij\cdot} + \frac{m_{ij} + 1}{2} \right]^2 + \frac{1}{4} \right), \quad (2.2.3)$$

$M_i = m_{i1} + m_{i2}$, and $R_{ijk}^{(j)}$ is the rank of X_{ijk} within level i of first factor and level j of the second factor.

Based on the above result, one can obtain an approximate α level test which rejects H_0 if

$$Q_M \geq \chi^2_{n-1,1-\alpha}, \quad (2.2.4)$$

where $\chi^2_{n-1,1-\alpha}$ is the 1-α percentile of a chi-square distribution with n-1 degrees of freedom.

2.3 Nonparametric Test of Interaction with Censored Data

Schemper (1988) extended procedures of Patel and Hoel (1973) and Brunner et al. (1995) reviewed respectively in Sections 2.1 and 2.2 to the censored data. We first consider the case where there are two factors and each has two levels.

Let T_{ik} be the true survival time for a patient in the level i of factor 1 (i=A, B) and level k of factor 2 (k=1, 2). From Patel and Hoel (1973), we can define a nonparametric measure of interaction between these two factors as:

$$\mu=P(T_{A1}< T_{B1})- P(T_{A2}< T_{B2}).$$

Assume that X_{ikm} is the observed survival time for the m-th subject in the level i of factor 1 and level k of factor 2. Define the Wilcoxon-Mann-Whitney scores as

$$u_{k,mn} = \begin{cases} 1 & \text{when } X_{Akm} - X_{Bkn} < 0 \text{ and } X_{Akm} \text{ uncensored} \\ -1 & \text{when } X_{Akm} - X_{Bkn} > 0 \text{ and } X_{Bkn} \text{ uncensored} \\ 0 & \text{otherwise} \end{cases},$$

the number of patients with positive $u_{k,mn}$ over all m and n as PO_k ($PO_k > 0$), and number of patients with negative $u_{k,mn}$ over all m and n as NE_k ($NE_k > 0$). Then we can estimate $P(T_{Ak}< T_{Bk})$ by

$$\hat{p}_k = \frac{PO_k}{PO_k + NE_k}$$

The nonparametric measure of the interaction can thus be estimated by:

$$\hat{\mu} = \hat{P}_1 - \hat{P}_2 .$$

A nonparametric test of interaction would be constructed based on this estimate and an estimate for its standard error.

Schemper (1988) proposed a test for a general 2xI factorial design based on the transformation of $\hat{p}_k$ and its jackknife variance estimate. Specifically, let $\tilde{p}_k = \sin^{-1}\sqrt{\hat{p}_k}$ and calculate the jackknife variance of $\hat{p}_k$ based on the following formula:

$$S_k^{\,2} = (n_k - 1)\left[\sum_{q=1}^{n_k} \tilde{p}_{kq}^2 / n_k - (\sum_{q=1}^{n_k} \tilde{p}_{kg} / n_k)^2\right],$$

where $\tilde{p}_{kq}$ is the transformation of $\hat{p}_k$ based on the dataset when the observation for the q-th patient in the k-th level of second factor is deleted out and n_k is the total number of patients in the k-th level of the second factor. Define

$$\overline{P} = (\sum_{k=1}^{I} \tilde{p}_k S_k^{-2}) / \sum_{k=1}^{2} S_k^{-2}$$

.

Then the Schemper test statistic is defined as

$$H = \sum_{k=1}^{I} (\tilde{p}_k - \overline{P})^2 / S_k^{-2}$$

The null hypothesis of no interaction between two factors would be rejected at α level if H > $\chi^2_{1-\alpha,I-1}$.

Tu and Gou (2004) have compared Schemper's procedure with that based on the Cox model by Monte-Carlo simulations. They found that Cox model is liberal,

which, in the context of biomarker study, may result in more non-predictive biomarkers to be identified as predictive. The significance level of Schemper's test is less than the nominal level in most cases, but in some cases, it went too far below. This implies that Schemper's procedure is more likely to lead to a situation where, to use a biomarker study as an example again, a biomarker that is truly predictive would be declared as non-predictive. This may be due to the conservative approach that defines the Wilcoxon-Mann-Whitney scores as zero, when, for fixed j, both X_l and $X_{l'}$ are censored. Therefore, Schemper's test may be improved by a more smoothed estimate of μ, which will be explored in Chapter 5.

Chapter 3

Interaction Tests Based on Nonparametric Models

Akritas and Arnold (1994) introduced nonparametric version of some hypotheses usually tested in analysis of variance and repeated measures models, such as the hypotheses of no interaction effects, and constructed nonparametric tests for these hypotheses. Akritas, Arnold and Brunner (1997) generalized their ideas to general factorial designs. Akritas and Bruner (1997) extended further these procedures to the situations where some data may be censored. In the other direction, Akritas and LaValley (1996) considered linear models for the potentially censored data from a factorial design without imposing parametric assumptions for the error terms and developed a test of interaction based on an extended Hodges-Lehmann estimator. In this chapter, the basic ideas of these procedures are reviewed.

3.1 Test of Nonparametric Hypotheses in a Factorial Design

Consider a $a \times b$ factorial experiment (factor A with a levels and factor B with b levels) with several observations per factor level combination. The k-th observation in cell (i, j) is denoted by Y_{ijk}. Assume Y_{ijk} are independent and $F_{ij}(u)$ is the distribution function of Y_{ijk}. Following the idea of Akritas and Arnold (1994), Akritas, Arnold and Brunner (1997) give an ANOVA (analysis of variance) type decomposition of $F_{ij}(u)$ as:

$$F_{ij}(u) = M(u) + A_i(u) + B_j(u) + C_{ij}(u),$$

where $\sum_i A_i(u) = 0$, $\sum_j B_j(u) = 0$, $\sum_i C_{ij}(u) = 0$, and $\sum_j C_{ij}(u) = 0$. Specifically, define

$$F_{i\cdot}(u) = \sum_{j=1}^{b} F_{ij}(u), \qquad F_{\cdot j}(u) = \sum_{i=1}^{a} F_{ij}(u), \text{ and } F_{\cdot\cdot}(u) = \sum_{i=1}^{a}\sum_{j=1}^{b} F_{ij}(u).$$

Then one has:

$$M = F_{\cdot\cdot}(u),\ A_i(u) = F_{i\cdot}(u) - F_{\cdot\cdot}(u),\ B_j(u) = F_{\cdot j}(u) - F_{\cdot\cdot}(u), \text{ and}$$

$$C_{ij}(u) = F_{ij}(u) - F_{i\cdot}(u) - F_{\cdot j}(u) + F_{\cdot\cdot}(u).$$

$C_{ij}(u)$ can be interpreted as the nonparametric interaction effects. Therefore, a test of interaction between two factors is equivalent to testing $H_0 : C_{ij}(u) = 0$ for i=1, 2, …, a and j=1, 2, …, b, which can also be written as H_0: C**F** = 0 for a suitably defined matrix C, where F=(F_{11}, …, F_{1b}, …, F_{a1}, …, F_{ab})' denote the *abx1* column vector consisting of the distribution functions F_{ij}.

Akritas, Arnold and Brunner (1997) proposed a nonparametric estimate of C**F** based on the rank of Y_{ijk} among all $N = \sum_i \sum_j n_{ij}$ observations, which is defined as

$$R_{ijk} = \frac{1}{2} + N\hat{H}(Y_{ijk}),$$

where

$$\hat{H}(x) = \sum_i \sum_j \lambda_{ij} \hat{F}_{ij}(x)$$

with $\lambda_{ij} = n_{ij} / N$ and $\hat{F}_{ij}(x)$ the empirical distribution of all observations in cell (i, j). With these ranks, C**F** can be consistently estimated by:

$$\hat{T}_C = N^{-1} C(R_{11.}, ..., R_{ab.})'.$$

Based on this estimate, Akritas, Arnold and Brunner (1997) defined a test statistic for the null hypothesis of no interaction between two factors as

$$Q_C = N\hat{T}_C{}' (C\hat{V}C')^{-1} \hat{T}_C,$$

where

$$\hat{V} = diag(\lambda_{ij}^{-1}\hat{\sigma}_{ij}^{2})$$

with

$$\hat{\sigma}_{ij}^{2} = N^{-2}(n_{ij} - 1)^{-1}\sum_{k=1}^{n_{ij}}(R_{ijk} - R_{ij.})^{2}.$$

It is shown that the asymptotic distribution of Q_c is a chi-square with *(a-1)x(b-1)* degrees of freedom. Therefore, one can reject the null hypothesis of no interaction at α level if $Q_c > \chi^2_{1-\alpha,(a-1)(b-1)}$.

3.2 Test of Nonparametric Hypotheses in a Factorial Design with Censored Data

Akritas and Bruner (1997) extended the method presented in Section 3.1 to the factorial designs with censored data by replacing the empirical distribution function with the Kaplan-Meier estimator.

Consider again a $a \times b$ factorial experiment (factor A with a levels and factor B with b levels) with several observations per factor level combination. From each cell (i, j), observed data are $(X_{ijk}, \Delta_{ijk}), k = 1,..., n_{ab}$, where X_{ijk} is the minimum between the failure time of interest T_{ijk} and the censoring variable C_{ijk} and $\Delta_{ijk} = I(X_{ijk} = T_{ijk})$,

where $I(E)$ denotes the indicator of the event E. It is assumed that T_{ijk} are independent and identically distributed according to F_{ij}, C_{ijk} are independent and identically according to G_{ij}, observations in different cells are independent, and censoring variables are independent from the failure time variables. The survival function of the T_{ijk} is denoted by S_{ij}, i.e., $S_{ij} = 1 - F_{ij}$. As discussed in Section 3.1, the nonparametric hypotheses of no interaction can be represented as

$$H_0: \quad \mathbf{CF} = 0 \qquad (3.2.1)$$

with a suitably defined matrix C, where $\mathbf{F} = (F_{11},...,F_{1b},...,F_{ab})'$, and this null hypothesis would be tested through weighted combinations of the empirical distribution function $\hat{F}_{ij}$ evaluated based on data from each cell. Specifically, the test statistic for H_0 is defined as

$$C\int_0^T \hat{S}_H \, d\hat{F} \qquad (3.2.2)$$

where

$\hat{S}_H = 0.5(Y_{\cdots}(s) + Y_{\cdots}(s-))$ with $Y_{ijk}(t) = I(X_{ijk} \geq t)$ and $Y_{\cdots}(t) = \sum_i \sum_j \sum_k I(X_{ijk} \geq t)$,

$\hat{F} = (\hat{F}_{11},...,\hat{F}_{1b},...,\hat{F}_{ab})'$ is the vector of Kaplan-Meier estimators of distribution function from each factor-level combination, and T is the smallest of largest observed survival times in all cells.

Under the basic assumption:

$$\int_0^{\tau_{ij}} \frac{dF_{ij}(s)}{1-G_{ij}(s-)} < \infty, \qquad \forall (i, j) \tag{3.2.3}$$

where

$$\tau_{ij} = \sup\{t : (1-F_{ij}(t-))(1-G_{ij}(t-)) > 0\},$$

Akritas and Brunner (1997) proved that under H_0: **CF** = 0, one has

$$Q(C) = n\left(C\int_0^T \hat{S}_H \, d\hat{F}\right)' (C\hat{V}C')\left(C\int_0^T \hat{S}_H \, d\hat{F}\right) \longrightarrow \chi^2_{(a-1)(b-1)}, \tag{3.2.4}$$

where $\hat{V}$ is the $ab \times ab$ diagonal matrix with diagonal elements $(N/n_{ij})\hat{\sigma}^2_{ij}(T)$ with

$$\hat{\sigma}^2_{ij}(t) = \int_0^t \hat{h}_{ij}(s)\left(1-\frac{\Delta N_{ij\cdot}(s)-1}{Y_{ij\cdot}(s)-1}\right)\frac{dN_{ij\cdot}(s)}{Y_{ij\cdot}(s)}, \tag{3.2.5}$$

$$\hat{h}_{ij}(s) = \hat{S}^2_{ij}(s-)\left[\hat{S}_H(s) - \frac{1}{\hat{S}_{ij}(s)}\int_{(s,T]} \hat{S}_H \, d\hat{F}_{ij}\right]^2 \frac{I(s \le T)}{n_{ij}^{-1} Y_{ij\cdot}(s)}, \tag{3.2.6}$$

$N_{ijk}(t) = I(X_{ijk} \le t, \Delta_{ijk} = 1), N_{ij\cdot}(t) = \sum_{k=1}^{n_{rc}} N_{ijk}(t)$, and $Y_{ij\cdot}(t) = \sum_{k=1}^{n_{rc}} Y_{ijk}(t)$.

Therefore, one can reject the null hypothesis of no interaction at α level if

$$Q(C) > \chi^2_{1-\alpha,(a-1)(b-1)}.$$

The calculation of test statistics and its variance matrix are complicate since they involve complicate function of counting process. This may have limited the potential application of the proposed test in practice. In Chapter 6, I will derive a simplified form of the test statistics under a 2x2 factorial design and apply the jackknife method to estimate the variance matrix.

3.3 Test of Interaction Based on Hodges-Lehmann Estimator

Akritas and LaValley (1996) proposed a method for testing the hypotheses of no main effects and no interaction in factorial designs based on a linear model for a transformation of the survival time. The method used the fact that these hypotheses could be expressed in terms of a vector of contrasts which could be estimated via a generalization of the Hodges-Lehmann estimator (Hodges and Lehmann, 1963).

Assume that the data consist of (Y_{ijk}, δ_{ijk}), $i = 1,...,a$, $j = 1,...,b$, $k = 1,...,n_{ij}$, with $Y_{ijk} = Y^t_{ijk} \wedge Y^c_{ijk}$, $\delta_{ijk} = I(Y_{ijk} = Y^t_{ijk})$, where Y^t_{ijk} is the true value of the time to event variable of interest and Y^c_{ijk} is the censoring variable which is assumed to be

independent of Y_{ijk}^{t}. Denote the total sample size as $N = \sum_i \sum_j n_{ij}$. Assume the following linear model to the time to event variable:

$$Y_{ijk}^{t} = \mu_{ij} + \varepsilon_{ijk} \quad \text{with} \quad \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad , \qquad (3.4.5)$$

where the error ε_{ijk} are assumed to be independent and identically distributed and the parameters in the model to satisfy the following constrains:

$$\sum_i \alpha_i = 0 \,,\ \sum_j \beta_j = 0 \,,\ \sum_i \gamma_{ij} = 0 \,,\ \sum_j \gamma_{ij} = 0.$$

Then the hypotheses of no interaction can be written as

$$H_I :\ \gamma_{ij} = 0 \text{ , for all } i \text{ and } j.$$

Let μ denote the $ab \times 1$ column vector consisting of the location parameter μ_{ij} defined as

$$\mu = (\mu_{11}, ..., \mu_{1b}, ..., \mu_{a1}, ..., \mu_{ab})' .$$

Then one can find a suitable matrix A_I (see Akritas and LaValley, 1996, for details) such that $\gamma_{ij} = 0$ is equivalent to $A_I \mu = 0$.

The basic idea behind the method of Akritas and LaValley is to write down $A_I \mu$ as a function of mean differences $\Delta_{ij,i'j'} = \mu_{ij} - \mu_{i'j'}$ and then replace $\Delta_{ij,i'j'}$ by its Hodges-Lehmann estimator $\hat{\Delta}_{ij,i'j'}$. For this purpose, they first introduced a new

extension of Hodges-Lehmann location difference estimator to the censored data. Let $\widehat{A_I\mu}$ denote the estimates of the contrast $A_I\mu$ based on the extended Hodges-Lehmann location estimators, it is shown that:

$$N^{\frac{1}{2}}\left(\widehat{A_I\mu} - A_I\mu\right) \rightarrow N\left(0, \sum_I\right).$$

The estimate of variance – covariance matrix $\sum_I$ is given by Akritas and LaValley (1996) but it is also very complicated since they also involve complicate functions of counting process. With this estimate of variance – covariance matrix, a test for the interaction between two factors can be obtained.

Akritas and LaValley's method is based on a linear model for a suitable transformation. This approach, like the approach based on the Cox proportional hazard models, is semi-parametric in nature since no parametric form needs to be assumed for the residuals in the linear model for the transformed data. One problem associated with this method is the definition of interaction may depend on the transformation used, another is that the estimation of the variance – covariance matrix may be very complicate. In Chapter 6, I will also derive a jackknife estimate for the variance – covariance matrix in their method.

Chapter 4

Kernel Density and Jackknife and Bootstrap Variance Estimates

In this chapter I will give an overview of the kernel smooth method for estimation of a nonparametric density function, the jackknife and bootstrap methods for the variance estimation, which are two major statistical tools used in the subsequent chapters to derive the new approaches for the assessment of interaction between two binary variables. The kernel density estimation with the right censored data will be first introduced, followed by an overview of the bandwidth selection for the kernel smoothing. The jackknife and bootstrap methods for variance estimation are introduced afterwards.

4.1 Kernel Density Estimation

Kernel-type density estimators are widely used in applications due to their simple form. We first consider the case where complete independent data are observed.

Let $X_1, X_2, \cdots X_n$ be a random sample from an unknown distribution with a density f. Estimation of f based on observed data is important in many statistical applications. One of the popular approaches is to use the kernel method, which uses

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (4.1)$$

as an estimate for f, where $K(y)$ is known as the kernel function and h the bandwidth which controls the smoothness of the estimated curve. The estimate defined in equation (4.1) is called as the kernel density estimate, which was first introduced by Rosenblatt (1956) and Parzen (1962). A comprehensive overview on the properties and application of the kernel estimate can be found in, for example, Silverman (1986), Simonoff (1996) and Hardle et. al. (2004).

To extend the kernel method to estimate the density function based on potentially censored survival data, we can rewrite $\hat{f}_h(x)$ in the following form:

$$\hat{f}_h(x) = \sum_{i=1}^{n} w(X_i) W(x, X_i), \qquad (4.2)$$

where

$$W(x, y) = \frac{1}{h} K\left(\frac{y - x}{h}\right) \text{ and } \quad w(x) = \frac{1}{n}.$$

With a general weight function $W(\cdot)$, which is usually taken to be a symmetric density function, and a re-weighting function $w(\cdot)$, which can be adjusted to control the roles of different data points in the sample, (4.2) defines a general class of density estimators, which is called weighted kernel estimators (Klein and Moeschberger, 2003). For the estimation of the density function based on the censored survival data, Marron and Padgett (1987) suggested to use the same kernel weight function for $W(\cdot)$ as for the complete data but a different weighting function *w(x)*, defined to be the jump sizes at *x* of the product-limit (PL) estimator by Kaplan and Meier (1958), to correct the random right-censoring bias. This estimate will be used in Chapter 5 to derive the smoothed nonparametric interaction test based on censored data.

In the application of kernel smoothing method to data analysis, an important issue is to determine the bandwidth *h*. In the following, I review some methods which can be used for bandwidth selection based on Wang and Wang (2007).

4.1.1 Bandwidth Selection Based on Mean Integrated Squared Error (MISE) Criterion

The Mean Integrated Squared error (MISE) criterion for bandwidth selection was first proposed by Rosenblatt (1956) and has been widely used for bandwidth selection.

Specifically, the MISE for an estimate $\hat{f}$ of an unknown density function f is defined as

$$MISE(\hat{f}) = \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx$$

and can be decomposed into a sum of an integrated square bias term and an integrated variance term as below:

$$MISE(\hat{f}) = \int \left\{ E\hat{f}(x) - f(x) \right\}^2 dx + \int \text{var } \hat{f}(x) dx. \qquad (4.4)$$

For the complete data with $w(\cdot) = 1/n$ in (4.2), it can be shown that the optimal bandwidth minimizing the MISE (see, for example, Silverman ,1986) can be written as:

$$h_{opt} = k_2^{-2/5} \left\{ \int K(t)^2 dt \right\}^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5} \qquad (4.5)$$

where $k_2 = \int t^2 K(t)dt$. This formula cannot, however, be used to the kernel estimate based on censored data with a data-dependent weight function.

Wang and Wang (2007) adjusted formula (4.5) and derived the following formula to calculate the bandwidth for the kernel density estimate based on the view that the censored survival data *{(X_i, Δ_i): i=1, 2, …, n}* could be considered as a data set consists of observed events and latent true survival times, with the same size and all data points having equal weights, and using the exponential distribution as the reference density in the estimation of $\int f^{"2}$:

$$h_e = CBn^{-1/5},$$

where

$$\mathrm{C}= 2^{1/5}\left\{\int t^2 K(t)dt\right\}^{-2/5}\left\{\int K(t)^2\,dt\right\}^{1/5},$$

$$B = \min(\hat{\lambda}, IQR_w/1.34),$$

$$\hat{\lambda} = \sum X_i \Big/ \sum \Delta_i,$$

and IQR_w is the sample inter-quartile range.

4.1.2 Bandwidth Selection Based on Asymptotic Mean Absolute Error (AMAE) Criterion

The bandwidth introduced above is a global bandwidth in the sense that it is the same for all *x*. In the uncensored case, a global bandwidth was found to perform acceptably for all values of *x* in the domain of the distribution. However, when estimating densities based on censored data, it makes sense to use locally optimized bandwidth for each value of *x* since proportion of censoring may be different at different *x*. Kuhn and Padgett (1997) proposed a method to select local bandwidth for right-censored data by minimizing the asymptotic mean absolute error defined as

$$AMAE(\hat{f}) = \lim_{n\to\infty} E \mid \hat{f} - f \mid .$$

They showed the optimal bandwidth can be written as:

$$h_{AMAE}(x) = \left[\frac{4\alpha^2 f(x) \int K^2(t)dt}{n\left(\int t^2 K(t)dt\right)^2 \left(f''(x)\right)^2 H^*(x)} \right]^{\frac{1}{5}},$$

where α was found to be 0.4809489 by Scott and Wand (1991). Kuhn and Padgett (1997) proposed that the censoring survival function, $H^*(x)$, can estimated by the Product-Limit (PL) estimator, $\hat{H}^*(x) = 1 - \hat{H}(x)$, where

$$\hat{H}(x)=\begin{cases}1, & 0\le x\le X_{(1)}\\ \sum_{i=1}^{k-1}\left(\dfrac{n-i}{n-i+1}\right)^{1-\Delta_i}, & X_{(k-1)}<x\le X_{(k)},k=2,\cdots,n\\ 0, & x>X_{(n)}\end{cases}$$

Here $X_{(1)},\cdots,X_{(n)}$ are order statistics of $X_1,\cdots,X_n$. For the unknown density $f(x)$ and its second derivatives $f''(x)$, they proposed to use the exponential distribution as a "reference distribution" in place of the normal reference distribution which is often used with the uncensored data situation. Therefore, there is only one parameter that needs to be estimated for the exponential "reference" density $f_R(x)=1/\lambda\exp(-x/\lambda)$ and its second derivative, and for the right-censored data, λ is estimated by the maximum likelihood estimator

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}X_i}{\sum_{i=1}^{n}\Delta_i}.$$

Therefore, we have the following formula for the local bandwidth:

$$h_{kp}=0.7644174\hat{\lambda}\hat{H}^{*}(x)^{-1/5}e^{x/(5\hat{\lambda})}n^{-1/5}.$$

Both global and local bandwidths can be used for the smoothed nonparametric interaction test defined in Chapter 5.

4.2 Jackknife and Bootstrap Variance Estimation

4.2.1 Jackknife Variance Estimation

The jackknife is an older, more specialized re-sampling procedure than the bootstrap, which was originally introduced by Quenouille (1956) to remove the bias and was extended by Tukey (1958) to the estimation of variance. Several other authors, including Miller (1964, 1974), Brillinger (1964, 1977) and Reeds (1978) have found conditions under which the jackknife variance estimate is consistent. There are two versions of jackknife re-sampling. First is based on the deleting single one case from the original sample (called as delete-one jackknife), and second is based on the deleting multiple case from the original sample (delete-k or block Jackknife) sequentially [Efron and Gong (1983); Wu (1986); Shao and Tu (1995)]. In the following, I will introduce the details of these two approaches.

Delete-one Jackknife

Let $\hat{\theta}_n(X) = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ be an estimator of the parameter θ based on the samples $X = (X_1, X_2, \cdots, X_n)$. The i^{th} *jackknife pseudo value* of $\hat{\theta}_n(X)$ is defined as

$$\hat{\varphi}_i(x) = n\hat{\theta}_n(X_1, X_2, \cdots X_n) - (n-1)\hat{\theta}_{n-1}\left((X_1, X_2, \cdots X_n)_{[-i]}\right) \qquad (4.2.1)$$

In (4.2.1), $X_{[i]} = (X_1, X_2, ..., X_n)_{[i]}$ denotes the sample $X = (X_1, X_2, \cdots, X_n)$ with i^{th} value X_i deleted, which is of a size $n-1$. Note

$$\hat{\varphi}_i(X) = \hat{\theta}_n(X) + (n-1)(\hat{\theta}_n(X) - \hat{\theta}_{n-1}(X_{[i]}))$$

so that $\hat{\varphi}_i(X)$ can be viewed as a biased-corrected version of $\hat{\theta}_n(X)$ determined by the trend in the estimation $\hat{\theta}_n(X)$ from $\hat{\theta}_{n-1}(X_{[i]})$ to $\hat{\theta}_n(X)$.

The basic idea of the jackknife is to treat the pseudo-values $\hat{\varphi}_i(X)$ as if they were independent random variable with mean θ. One can then use the sample variance of the jackknife pseudo-values to estimate the variance of $\hat{\theta}_n(X)$. Specifically the jackknife variance estimate can be defined as:

$$\hat{V}_\theta(X) = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\varphi}_i(X) - \hat{\varphi}(X))^2 , \qquad (4.2.2)$$

where

$$\hat{\varphi}(X) = \frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}_i(X) . \qquad (4.2.3)$$

With this variance estimate, we can obtain the jackknife confidence interval and carry out statistical tests using the Central Limit Theorem. Specifically, the jackknife 95% confidence interval for θ is defined as:

$$\left(\hat{\theta}(X) - 1.96\sqrt{\frac{1}{n}\hat{V}_\theta(X)}, \qquad \hat{\theta}(X) + 1.96\sqrt{\frac{1}{n}\hat{V}_\theta(X)} \right). \quad (4.2.4)$$

Similarly, one can get the test statistics for the hypothesis $H_0 : \theta = \theta_0$

$$Z = \frac{\hat{\theta}(X) - \theta_0}{\sqrt{\frac{1}{n}\hat{V}_\theta(X)}}, \qquad (4.2.5)$$

which has an approximate standard normal distribution under $H_0 : \theta = \theta_0$..

Delete k or Block Jackknife

If *n* is large, the pseudo-values $\hat{\varphi}_i(X)$ in (4.2.1) may be too close together, and the variance estimate $\hat{V}_\theta(X)$ may be mostly sampling error. In that case, one can define a block jackknife instead of the delete–one jackknife. Assume $n = gh$, where *h* will be the block size and *g* is the number of blocks. Define

$$\hat{\varphi}_i(X) = g\hat{\theta}_n(X_1, X_2, \cdots X_n) - (g-1)\hat{\theta}_{n-k}((X_1, \cdots, \cdots X_n))_{[i]} \qquad (4.2.6)$$

where $1 \le i \le g$ and $X_{[i]}$ means the sample $X = (X_1, X_2, \cdots, X_n)$ with i^{th} block of *h* values was removed. In the same way, $\hat{\varphi}_i(X)$ in (4.2.6) could be treated as approximately independent and identically distributed random variables. Therefore,

one can then estimate sample mean and variance and obtain the confidence interval and carry out statistical testing as that in the delete – one jackknife.

4.2.2 Bootstrap Variance Estimation

Bootstrap, which is similar to jackknife, is another resampling method which can be used in estimating the variance of a population parameter. This method was originally proposed by Efron (1979). Algorithm for bootstrap variance estimation can be discribed as follows:

Step 1. Calculate the estimator $\hat{\theta}_n(X) = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ of the parameter θ based on the samples $X = (X_1, X_2, \cdots, X_n)$.

Step 2 . Draw a bootstrap sample $X_b^* = (X_1^*, X_2^*, \ldots, X_n^*)$, with replacement, from the empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}$ and calculate the bootstrapped estimator $\hat{\theta}_b^* = \hat{\theta}_n(X_1^*, X_2^*, \ldots, X_n^*)$.

Step 3. Repeat the Step 2 *B* times, getting $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$.

Then the variance estimate based on the B bootstrap samples can be written as:

$$\hat{V}_\theta(X) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^*(X) - \hat{\theta}_\cdot^*(X)\right)^2 \qquad (4.2.7)$$

where

$$\hat{\theta}_{\cdot}^{*} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_{b}^{*}(X) \qquad (4.2.8)$$

As pointed out by Shao and Tu (1995), the bootstrap can be applied to both variance and distribution estimation problems. However, the bootstrap variance estimator is not as good as the jackknife in terms of the empirical results. Furthermore, the bootstrap variance estimator usually requires more computations than the jackknife. Thus, the bootstrap method is mainly recommended for distribution estimation. Therefore, in this thesis, only the delete-one jackknife method will be used to estimate the variance in simulation study in Chapter 5 and 7.

Chapter 5

A Smoothed Test for Treatment-Covariate Interactions with Censored Data

5.1 Introduction

As described in the first chapter of this thesis, analysis by subgroups defined based on baseline covariates of patients is often interest in clinical trials to assess the treatment-covariate interaction, which measures the heterogeneity of treatment effects across the subgroups. For example, in NCIC CTG MA.17 trial, it was shown treatment with letrozole is very effective in reducing the chance of disease recurrence for women with early breast cancer who had received five years of treatment with tamoxifen (Goss et al., 2003). Muss et al. (2008) assessed whether letrozole is equally effective for all women randomized in this trial, regardless of their age. Based on the insignificance in the test of the interaction between letrozole and age, they suggested

that both young and old patients should be considered treatment with letrozole. On the other hand, in the NCIC CTG CO.17 trial, it was shown that cetuximab would improve the survival rate for patients with end-stage colorectal cancer but the difference between two treatment groups, cetuximab and best supportive care, was small. It would be interesting to identify a group of patients, based on their baseline characteristics, who would have larger benefit from cetuximab treatment. Karapetis et al. (2008) tested the interaction between the cetuximab treatment and K-ras mutation status. The results showed that, for patients with mutated K-ras, the difference in median survival between cetuximab and best supportive care treatments was very small, whereas the difference was very large for patients with wild-type K-ras. Since the p-value of the interaction test was highly significant, they recommended that cetuximab should only be given to patients with wild-type K-ras.

The traditional approach to assess the treatment-covariate interaction is to use the Cox proportional hazard model and to include an interaction term as one of the covariates in clinical trials with time to an event endpoint. The key assumption behind the Cox model, the proportional hazard assumption, however, may not be satisfied by the data from clinical trials. As reviewed in Chapter 2 of this thesis, Schemper (1988) proposed a non-parametric test for the treatment-covariate interaction based on a non-

parametric definition of the covariate-treatment interaction proposed by Patel and Hoel (1973). This test is based on the sum of U-scores generated from the Wilcoxon statistics under a random censoring mechanism. As pointed out by Koziol and Jia (2009), the expected value of this type of test statistics depends on the underlying censoring distributions, and therefore, may be far moved from the true value. They suggested to construct a test statistic based on Kaplan-Meier estimates but, as pointed out by Bassiakos et al. (1991), this type of statistics may not be consistent beyond the supports of the censored distribution. Koziol and Jia (2009) also pointed out any other estimate of survival function would be used to construct the test statistic. In this chapter, we consider using the kernel smoothing method to estimate the survival function and derive a test for interaction based on this type of kernel estimates.

This chapter organized as follows: I introduce a smoothed estimate for a nonparametric measure of interaction in Section 5.2 and present the major theoretical result which establishes the asymptotic distribution of this estimate in Section 5.3. The new smoothed test of interaction based on this estimate is defined in Section 5.4. The setup and results of the simulation studies assessing the type I errors of the proposed procedure are presented in Section 5.5. The proposed procedures are illustrated with real data from a clinical trial in Section 5.6. Section 5.7 concludes this chapter with

some discussions. As an appendix, a detail proof of the major results presented in Section 5.3 is given at the end of this chapter.

5.2 The Smoothed Estimate for the Nonparametric Measurement of Interaction

Let $F_j(x)$ and $G_j(x)$ be the distribution function of T_{Aj} and T_{Bj} respectively, where *A* and *B* represent two treatments; *j*=1, 2 denotes the binary level of a baseline covariate, such as a biomarker. Then Patel-Hoel measure of interaction between the treatment and baseline covariate can be written as

$$\mu = P_1 - P_2,$$

where

$$P_j = P\left\{T_{Aj} < T_{Bj}\right\} = \int_0^\infty F_j(x) dG_j(x).$$

Since data from two groups defined by the baseline covariate are independent, in the next, I only consider the estimate for each P_j.

The Wilcoxon-Mann-Whitney statistic is an estimate of P_j by replacing $F_j(x)$ and $G_j(x)$ respectively with their empirical estimates. Empirical estimates are

discrete, which may lead to an inaccurate estimate for P_j when both $F_j(x)$ and $G_j(x)$ are smooth.

Assume the $F_j(x)$ and $G_j(x)$ are smooth function. This is in general true for survival data in medical studies. Let $g_j(x)$ be the density function of $G_j(x)$. Then p_j can be rewritten as

$$P_j = \int_0^\infty F_j(x) g_j(x) dx.$$

A smoothed estimate for P_j can be obtained by estimating $F_j(x)$ and $g_j(x)$ using the kernel method. Assume $F_{jn}(x)$ and $g_{jm}(x)$ are respectively the kernel estimate of $F_j(x)$ and $g_j(x)$ with respectively kernel functions $L(\cdot)$ and $K(\cdot)$ and bandwidths h_{Aj} and h_{Bj} (for example, see Mielniczuk, 1986). Then P_j can be estimated by

$$\hat{P}_j = \int_0^\infty F_{jn}(x) g_{jm}(x) dx.$$

A test statistic for interaction could then be constructed based on the difference of $\hat{P}_1$ and $\hat{P}_2$, and its variance estimate under the null hypothesis. For this purpose, in the next section, I first present the asymptotic distributions $\hat{P}_j$.

5.3 The Asymptotic Distribution of $\hat{P}_j$

In this section, I will state the main theoretical result of this chapter. First I introduce some notations. Let $T_{11},...,T_{1n}$ and $T_{21},...,T_{2m}$ be independent failure times from subjects in two groups, which have respectively absolutely continuous distribution functions $F_1(t)$, $F_2(t)$ and density functions $f_1(t)$, $f_2(t)$ $(t>0)$. They are independently censored from the right by respectively random sequences $C_{11},...,C_{1n}$ and $C_{21},...,C_{2m}$, which also have absolutely continuous distribution functions $G_1(t)$ and $G_2(t)$. Further assume that all T_{1i}, C_{1i} and T_{2j}, C_{2j} are mutually independent. Define the observed data from these two groups as:

$$X_{1i} = \min(T_{1i}, C_{1i}), \qquad X_{2j} = \min(T_{2j}, C_{2j}),$$

$$\delta_{1i} = I_{(T_{1i}<C_{1i})}, \quad i = 1,2,...,n, \qquad \delta_{2j} = I_{(T_{2j}<C_{2j}),} \quad j = 1,2,...,m$$

where I_A is the indicator function of set *A*. Then the distribution functions H_1 and H_2 of X_{1i} and X_{2j} satisfy

$$1 - H_k(t) = (1 - F_k(t))(1 - G_k(t)), \qquad (k = 1, 2)$$

Let $\overline{F}_k(t) = 1 - F_k(t)$, $\overline{G}_k(t) = 1 - G_k(t)$ and $\overline{H}_k(t) = 1 - H_k(t)$ be the corresponding survival function of F_k, G_k and H_k (k = 1,2). Define

$$T_{1F} = \sup\{t \mid \overline{F}_1(t) > 0\},\ T_{2F} = \sup\{t \mid \overline{F}_2(t) > 0\} \text{ and } T = \min(T_{1F}, T_{2F}).$$

Because of difficulty of estimating survival distribution beyond T, P has to be modified as:

$$P_M = P(M > T_1 > T_2) = \int_0^M F_2(x) dF_1(x),$$

where M is a given number which is smaller than T.

Let $\hat{F}_{KMk}$ be the Kaplan-Meier estimator of F_k (k=1, 2), then the smoothed Kaplan-Meier estimators are defined as:

$$\hat{F}_{nk}(x) = \int K((x - y / h_n) d\hat{F}_{KMk}(y)$$

$$= \int h_n^{-1} k((x - y / h_n) \hat{F}_{KMk}(y) dy, \qquad (k=1, 2)$$

where $K(x) = \int_{-\infty}^{x} k(u) du$ and $k(\cdot)$ is a kernel function. With these estimates, a smoothed estimate of $P_M = P(M > T_1 > T_2) = \int_0^M F_2(x) dF_1(x)$ could be defined as:

$$\hat{P}_{nM} = \int_0^M \hat{F}_{n2}(x) d\hat{F}_{n1}(x).$$

Assume the following regularity conditions:

(A1): Kernel function k belongs to a class of kernels, k_m, of all Borel measurable bounded function on the real line with following properties

(i) $\int k(x)dx = 1$

(ii) $\int x^i k(x)dx = 0$ for $i = 1,...,m$, and

(iii) $M_{m+1} = \int |x|^{m+1} |k(x)| dx < \infty$

(A2): The distribution function F is continuous and $G(T_F^-) < 1$.

(A3): $f^{(m)}$ is integrable for $m \geq 1$ and $\sqrt{n}h_n^m \to 0$ as $n \to \infty$.

We can have the following theorem for the asymptotic distribution of $\hat{P}_{nM}$:

Theorem 1: If kernel function $K(x)$ satisfies condition (A1); $F_1(t)$, $F_2(t)$ and $G_1(t)$, $G_2(t)$ satisfy condition (A2); $f_1(t)$, $f_2(t)$ and h_n satisfy condition (A3), then we have:

$$\sqrt{n}(\hat{P}_{nM} - P_M) \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 is a positive constant which depends on unknown distribution functions $F_1(t), F_2(t), G_1(t), and\ G_2(t)$.

The proof of this theorem and the definition of σ^2 can be found in the Appendix. In the next, I will introduce a smoothed test of interaction based on the results in this theorem.

5.4 The Smoothed Test for the Interaction

Based on the results in the last section, we can define a smoothed test for the interaction between the treatment and a baseline covariate. It is known that the existence of interaction between treatment and covariate can be assessed by testing the following null hypothesis:

$$H_0: \mu = 0$$

against the alternative hypothesis:

$$H_1: \mu \neq 0.$$

Define

$$\hat{\mu} = \hat{P}_1 - \hat{P}_2.$$

To construct a test statistic based on $\hat{\mu}$, we need to have an estimate for its variance. Although in theory, we may estimate this variance based on the asymptotic variance defined in the appendix but in practice, a direct estimate may be too difficult since the

asymptotic variance of $\hat{\mu}$ is too complicated. We will use the delete-one jackknife to estimate this variance. Let $\hat{\mu}_{[-k]}$ be the estimate based on the dataset when the observation for the k-th patients is delete out, where *k=1,2…n* and n is the total number of patients.

Let

$$\bar{\mu} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}_{[-k]}.$$

Then the Jackknife variance of $\hat{\mu}$ can be calculated based on the following formula:

$$\hat{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{\mu}_{[-k]} - \bar{\mu}\right)^2.$$

Therefore, the test statistics for the interaction can be written as

$$H = \frac{\hat{\mu}^2}{\hat{S}^2/n}.$$

Based on results in Theorem 1, the null hypothesis of no interaction between treatment and baseline covariate can be rejected at α level if $H > \chi^2_{1,1-\alpha}$. This test is called as the smoothed test for interaction.

Remark: In application of this test to the analysis of real data, one may select *M* as the maximum of the survival times in the dataset. Some more conditions are needed to generalize Theorem 1 to this case when *M* is not a constant.

5.5 Simulation Studies

Monte-Carlo simulations were performed to assess the actual type I error of the smoothed test proposed in the last section under the nominal level α=0.05. In each simulation, 1000 random samples were generated. All the simulations were designed to resemble to real clinical trials. It was assumed that there are two treatment arms and one covariate with two levels (i. e. there are a total of two strata). Let A and B represent two treatment arms, and 1 and 2 two levels of the covariate. Various sample sizes in each group ($n_{A1} = n_{A2} = n_{B1} = n_{B2} = 50, 75, 100$) were considered in the simulations.

Two settings have been investigated in the simulation studies: the case when the hazard rates in the 4 groups (A1, A2, B1, B2) are proportional and the case when they are not proportional. The true survival times for 4 groups are all from exponential distribution with scale parameter λ=0.1 in the settings where hazard rates are

proportional and log normal distribution with location parameter μ=1.95 and scale parameter σ=1 for the settings where hazard rates are not non-proportional.

The following clinical trial process was used to generate censored survival data: Let T_a and T_f denote the time lengths of the accrual phase and follow-up phase of a cancer clinical trial, respectively. Suppose the time unit is one month. This means patients are accrued uniformly until T_a months after the trial has opened and then are followed up for another T_f months. Assume that a patient entered the study at u months after the study opens. Then this patient would stay in the trials for the maximum $(T_a - u) + T_f$ months in total. This implies that a patient would experience the event of interest at time t after entering the study if $u + t < T_a + T_f$. Otherwise the patient is censored if the patient remains event free at the end of study. It is assumed that no patient leaves study early or is lost to the follow up.

In the simulations, location parameter μ and scale parameter σ of the log normal distribution were chosen to have approximately same censoring rate as the exponential distribution with scale parameter λ=0.1 for given time lengths of the accrual phase and follow-up phase of a clinical trial. For each scenario, kernel function $k(x) = \frac{15}{16}(1 - x^2)^2$ is used and two bandwidths, h_e proposed by Wang and Wang

(2007) and h_{kp} proposed by Kuhn and Padgett (1997), are used to assess the effect of the bandwidth on the type I error. T_a was fixed at 6 month for all simulations and T_f was varied from 6, 12 and 18, which represent, approximately 41%, 23% and 12% when $\lambda = 0.1$ for exponential distribution and $\mu = 1.95, \sigma^2 = 1$ for log normal distribution, respectively.

The results of the simulation studies are presented in Table 5-1 for the cases where hazard rates were proportional and in Table 5-2 for the cases where hazard rates were non-proportional, respectively. From these tables, we may observe that in most cases, the simulated type I errors of the proposed tests are slightly less than, but in most cases, quite close to the nominal level. The simulated type I errors in the cases where the sample size was 75 and hazard rates were proportional are little bit far from the nominal level. For all scenarios, both bandwidths had similar performance in term of type I error. The type I error in the case where the censoring was light seems closer to nominal level than in the cases where censoring was moderate or heave.

Table 5-1: Simulated Type I Error (Proportional Hazard Case)

Sample Size ($n_{A1}=n_{A2}=n_{B1}=n_{B2}$)	Bandwidth Type	Censoring Proportions		
		12%	23%	41%
50	h_e	0.052	0.046	0.038
	h_{kp}	0.042	0.044	0.047
75	h_e	0.028	0.022	0.034
	h_{kp}	0.030	0.027	0.032
100	h_e	0.048	0.044	0.048
	h_{kp}	0.046	0.043	0.047

Table 5-2: Simulated Type I Error (Non-Proportional Hazard Case)

Sample Size ($n_{A1}=n_{A2}=n_{B1}=n_{B2}$)	Bandwidth Type	Censoring Proportions		
		12%	23%	41%
50	h_e	0.044	0.042	0.040
	h_{kp}	0.046	0.048	0.046
75	h_e	0.052	0.044	0.042
	h_{kp}	0.050	0.044	0.042
100	h_e	0.046	0.030	0.046
	h_{kp}	0.048	0.034	0.058

5.6 An Application to Data from a Clinical Trial

In this section, we apply the proposed methods to a data set from a randomized clinical trial conducted by NCIC Clinical Trial Group. This trial was designed to compare two chemotherapy regimens, Cyclophosphamide-Methotrexate-Fluorouracil (CMF) versus Cyclophosphamide-Epirubicin-Fluorouracil (CEF), in women with early stage breast cancer. Covariates of interest are the expression levels of two genes (HER2 and TOP2A). 494 patients with expression data for both genes were included in the analysis with average censoring rate of 61.3% on overall survival.

The result based on the univariate Cox models show that the hazard ratio (CEF over CMF) in the HER2 over-expressed group was 0.654 with a 95% confidence interval from 0.396 to 1.078, while in the HER2 normal group was 1.184 with a 95% confidence interval from 0.838 to 1.672 and a p-value of 0.0554 from the treatment and HER2 interaction test. For gene TOP2A, the hazard ratio was 0.423 with a 95% confidence interval from 0.218 to 0.821 in patients with TOP2A alteration, and 1.245 with a 95% confidence interval from 0.880 to 1.761 in TOP2A normal patients. Interaction test between the treatment and gene TOP2A, on the other hand, was highly significant with a p-value 0.0047 from the analysis based on the Cox model.

Figures 5-1 to 5-4 present the Kaplan–Meier plots for different subgroups. From Figure 5-2 and 5-4, we can see that the proportional hazard assumption may not be satisfied in these subgroups since the survival curves for the CEF and CMF groups crossed.

Figure 5-1: Kaplan-Meier Curves for HER2 Over-expressed Patients

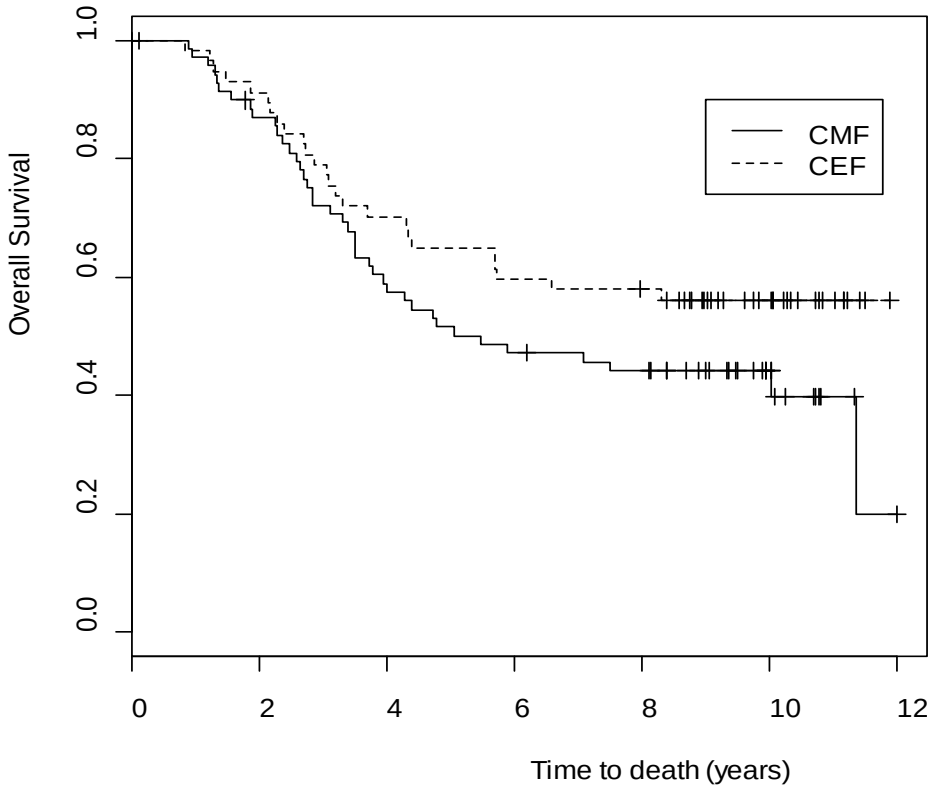

Figure 5-2: Kaplan-Meier Curves in HER2 Normal Patients

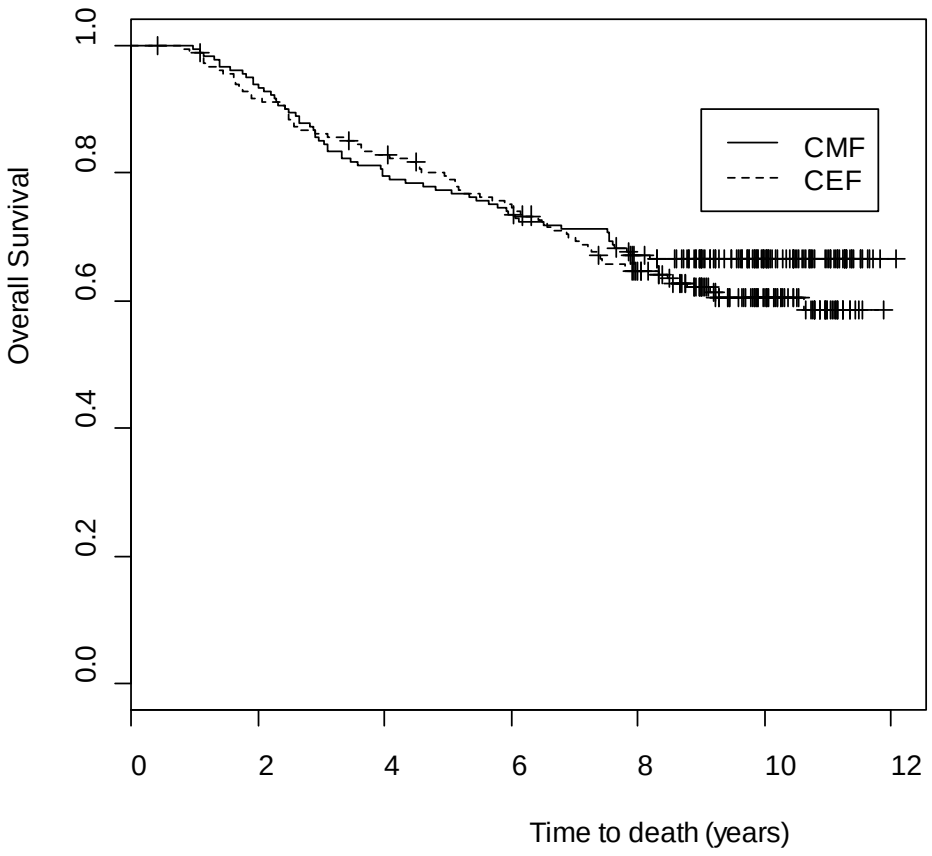

Figure 5-3: Kaplan-Meier Curves in TOP2A Alteration Patients

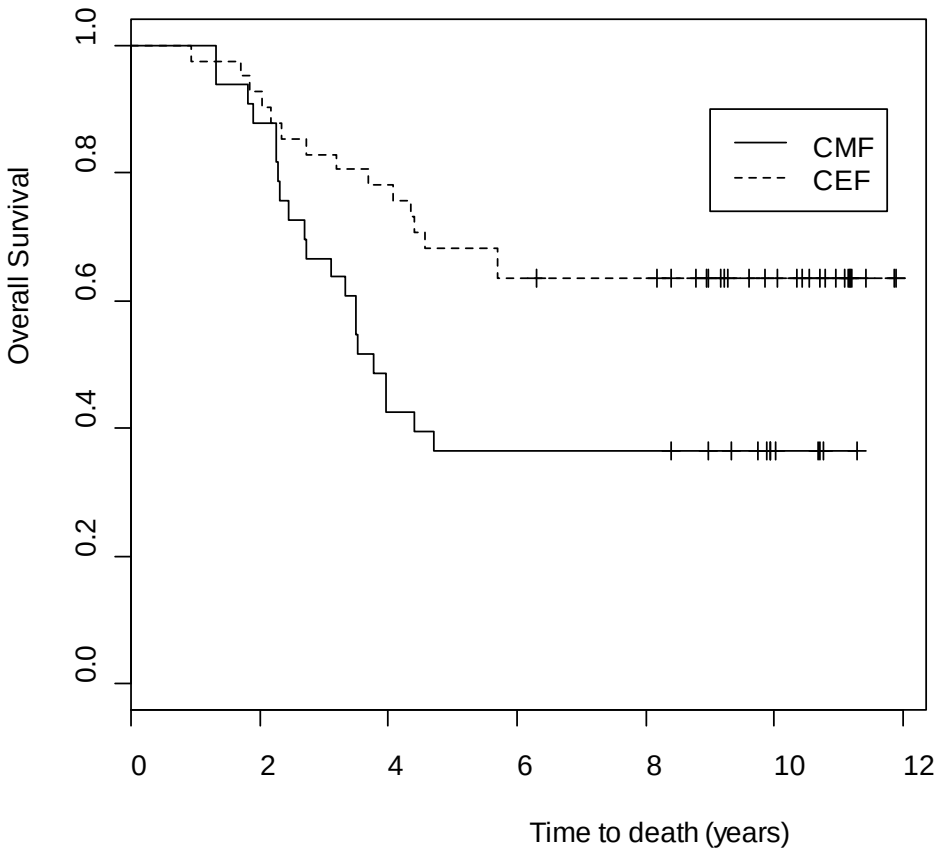

Figure 5-4: Kaplan-Meier Curves in TOP2A Normal Patients

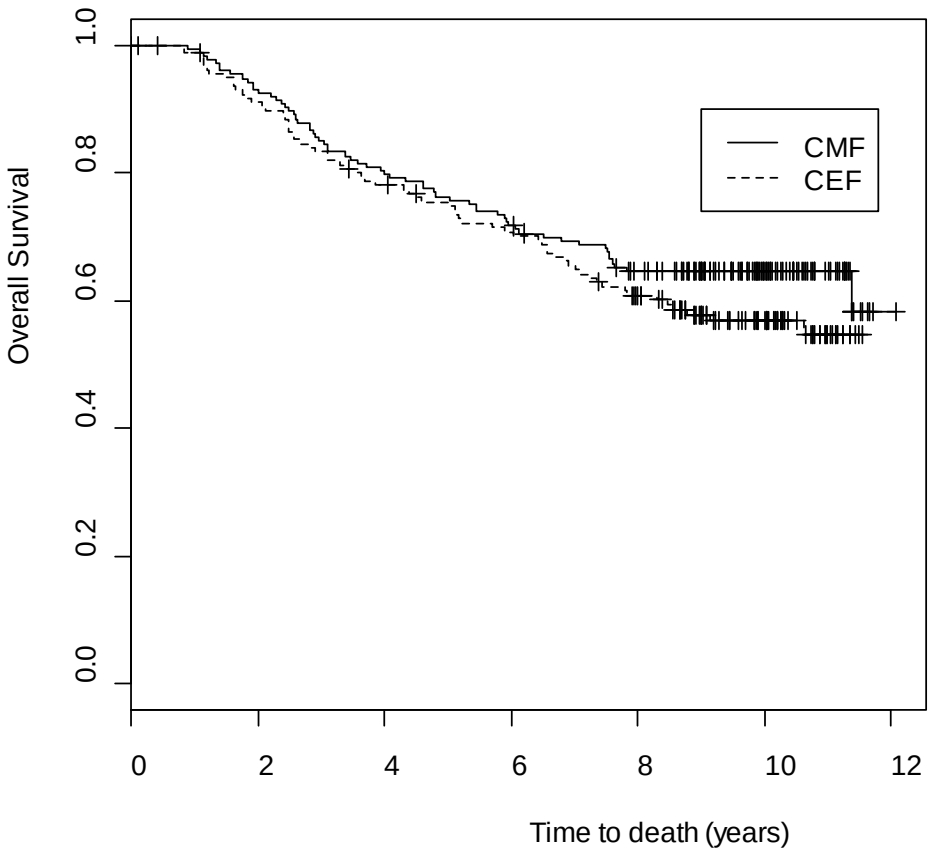

Table 5-3: P-values from Smoothed method

	Treatment-HER2 interaction	Treatment-TOP2A interaction
Smoothed test with bandwidth h_e	0.05167	0.325
Smoothed test with bandwidth h_{kp}	0.05924	0.420

Table 5-3 presents p-values from the proposed method for the testing of the interaction between the treatment and HER2 and between the treatment and TOP2A respectively. From this table, results from two bandwidths were very similar, which indicated that the interaction between treatment and HER2 was marginally significant, while the interaction between treatment and TOP2A did not reach significance.

5.7 Conclusions and Discussion

In this chapter, a non-parametric procedure based on the smoothed estimate of Patel–Hoel measure and censored data is derived to test the interaction between the treatment and a binary covariate with censored data. The proposed procedure is evaluated by the Monte-Carlo simulations and applied to a real data set from a cancer

clinical trial. The results of the simulations showed that actual significance levels are well controlled and are very close to the nominal significance level.

These are also several other non-parametric procedures for testing the interaction between the treatment and the covariate. In Chapter 7, we will present the results from a systemic Monte-Carlo simulation study which compared the procedure proposed in this chapter to some other procedures.

5.8 Appendix: The Proof of Theorem 1

Before I present the proof of Theorem 1, I first introduce some results for the strong approximation and weak convergence of kernel smoothed Kaplan-Meier estimator, as well as continuous mapping, as lemmas.

Lemma 1: [Folder, A., Rejto, L. and Winter B.B. (1981)] Suppose that Condition (A2) hold for *F* and *G*, and let $\hat{F}_n$ be smoothed Kaplan-Meier estimator defined as in the Section 5.3. Then

$$\sup_{x \le M} | \hat{F}_n(x) - F(x) | \to 0 \qquad a.s.$$

Lemma 2: [Ghorai and Susarla (1990)] Suppose the kernel function satisfies Condition (A1), and Condition (A3) hold for density function f and h_n. Then for $\hat{F}_n$, the smoothed Kaplan-Meier estimator defined as in the Section 5.3, we have

$$\{\sqrt{n}(\hat{F}_n(x) - F(x)) : x \le M\} \to \{W(x) : x \le M\} \text{ in D[0, M],}$$

where $W(\cdot)$ is a zero mean Gaussian process with covariance function

$$\sigma(x, y) = \overline{F}(x)\overline{F}(y)\int_0^{x \wedge y} (\overline{H}(u))^{-2}\overline{G}(u)dF(u)$$

and D[0, M] is a function space containing all right continuous functions possessing left limit.

Lemma 3: [Durbin (1973)] The mapping

$$Z(t) \to \int_0^M Z(t)dt$$

from $D[0, M](M < \infty)$ to R^1 is a continuous mapping.

Lemma 4: [Billingsley (1968)] Suppose h is a continuous mapping from a function space D to another one D'. If a sequence of random elements X_m of D converges weakly to X according to the topology on D, then $h(X_m)$ will converge weakly to $h(X)$ according to topology on D'.

Proof of Theorem 1: One can see that:

$$\sqrt{n}(\hat{P}_{nM} - P_M) = \sqrt{n}\left(\int_0^M \hat{F}_{n2}(x)d\hat{F}_{n1}(x) - \int_0^M F_2(x)dF_1(x) \right)$$

$$= \sqrt{n}\int_0^M \left(\hat{F}_{n2}(x) - F_2(x)\right)d\left(\hat{F}_{n1}(x) - F_1(x)\right)$$

$$+ \int_0^M F_2(x)d\left(\sqrt{n}\left(\hat{F}_{n1}(x) - F_1(x)\right)\right) + \int_0^M \sqrt{n}\left(\hat{F}_{n2}(x) - F_2(x)\right)dF_1(x)$$

$$= I_{n1} + I_{n2} + I_{n3}.$$

I first prove that $I_{n1} \xrightarrow{P} 0.$

Given $\varepsilon > 0$, let $\varepsilon_1 > 0$ such that $\sqrt{\frac{2}{\pi}}\frac{2\varepsilon_1\sigma_M}{\varepsilon}e^{-\frac{1}{2\sigma_M^2}\left(\frac{\varepsilon}{2\varepsilon_1}\right)^2} < \frac{\varepsilon}{3}.$ Since by Lemma 2, one has

$$\sqrt{n}\int_0^M d\left(\hat{F}_{n1}(x) - F_1(x)\right) = \sqrt{n}\left(\hat{F}_{n1}(M) - F_1(M)\right) \xrightarrow{d} N(0, \sigma_M^2),$$

one can select n large enough such that

$$P\left(\left|\sqrt{n}\left(\hat{F}_{n1}(M) - F_1(M)\right) - N(0, \sigma_M^2)\right| > \varepsilon/(2\varepsilon_1)\right) < \varepsilon/3.$$

Also, by Lemma 1, one can select n large enough such that

$$P\left(\sup_{x \le M} |\hat{F}_{n2}(x) - F_2(x)| > \varepsilon_1\right) < \frac{\varepsilon}{3}.$$

Therefore, I have

$$P(|I_{n1}|>\varepsilon)$$

$$\leq P\left[\left(|I_{n1}|>\varepsilon\right)\cap\left(\sup_{x\leq M}|\hat{F}_{n2}(x)-F_2(x)|>\varepsilon_1\right)\right]$$

$$+P\left[\left(|I_{n1}|>\varepsilon\right)\cap\left(\sup_{x\leq M}|\hat{F}_{n2}(x)-F_2(x)|\leq\varepsilon_1\right)\right]$$

$$\leq P\left(sup_{x\leq M}\left|\hat{F}_{n2}(x)-F_2(x)\right|>\varepsilon_1\right)+P\left(\varepsilon_1\left|\sqrt{n}\int_0^M d\left(\hat{F}_{n1}(x)-F_1(x)\right)\right|>\varepsilon\right)$$

$$\leq\frac{\varepsilon}{3}+P\left(\varepsilon_1\left|\sqrt{n}\int_0^M d\left[\left(\hat{F}_{n1}(x)-F_1(x)\right)\right]-N(0,\sigma_M^2)\right|>\varepsilon/2\right)+P(|N(0,\sigma_M^2)|$$

$$>\frac{\varepsilon}{2\varepsilon_1})$$

$$\leq\frac{\varepsilon}{3}+P\left(\left|\sqrt{n}\left(\hat{F}_{n1}(M)-F_1(M)\right)-N(0,\sigma_M^2)\right|>\varepsilon/(2\varepsilon_1)\right)+P(|N(0,\sigma_M^2)|>\frac{\varepsilon}{2\varepsilon_1})$$

$$<\frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}$$

$$=\varepsilon.$$

The probability inequality $1-\Phi(x)<\frac{\phi(x)}{x}$, where $\Phi(x)$ and $\phi(x)$ are respectively the distribution and density functions of a standard normal distributions, are used in the last inequality above. $I_{n1} \xrightarrow{P} 0$ is now proved.

Using the same technique that was used in the proof of Theorem 3.5 in Rao and Tu (1990), now I prove that $I_{n2} \xrightarrow{d} N(0, \sigma_1^2)$. Since

$$I_{n2} = \int_0^M F_2(x) d\left(\sqrt{n}\left(\hat{F}_{n1}(x) - F_1(x)\right)\right)$$

$$= \sqrt{n}\left[F_2(M)\left(\hat{F}_{n1}(M) - F_1(M)\right) - \int_0^M \left(\hat{F}_{n1}(x) - F_1(x)\right) dF_2(x)\right],$$

for a function h from $D[0, M]$ to R^1 defined as

$$h\big(f(x)\big) = F_2(M)f(M) - \int_0^M f(x)dF_2(x),$$

for any $f(x) \in D[0, M]$, I_{n2} can be written as

$$Q_n = h\left(\sqrt{n}\big(\hat{F}_{n1}(x) - F_1(x)\big)\right)$$

From Lemma 3, one knows that h is a continuous mapping. Therefore by Lemma 4 and Lemma 2, one has

$$Q_n \xrightarrow{w} Q = h(W(t)) .$$

Since

$$Q = W(M)F_2(M) - \int_0^M W(x)dF_2\,(x)\,,$$

$W(x)$ is a mean zero Gaussian process, one can obtain that Q has a normal distribution with mean zero and variance

$\sigma_1^2 = E(Q^2) =$

$\delta_1(M,M)F_2^2(M) - 2F_2(M)\int_0^M \delta_1\,(x,M)dF_2(x) + \int_0^M \int_0^M \delta_1(x,y)dF_2\,(x)dF_2(y),$

where $\delta_1\,(x,y)$ is the covariance function which is defined from Lemma 2 as

$$\delta_1(x,y) = \bar{F}_1(x)\bar{F}_1(y)\int_0^{x\wedge y}\left(\bar{H}_1(u)\right)^{-2}\bar{G}_1(u)dF_1(u)\,,$$

with $\bar{F}_1(x), \bar{G}_1(x) and\ \bar{H}_1(x)$ defined in Section 5.3. This completes the proof $I_{n2} \xrightarrow{d} N(0, \sigma_1^2)$.

Similarly, I can use the same technique to prove $I_{n3} \xrightarrow{d} N(0, \sigma_2^2)$. Specifically, since $I_{n3} = \int_0^M \left(\hat{F}_{n2}(x) - F_2(x)\right) dF_1(x)$, with g a function from $D[0,M]$ to R^1 defined as

$$g\big(f(x)\big) = \int_0^M f(x)dF_1\,(x),$$

for any $f(x) \in D[0,M]$, it can be written as

$$H_n = g\left(\sqrt{n}\left(\hat{F}_{n2}(x) - F_2(x)\right)\right).$$

From Lemma 3, g is also a continuous mapping. Therefore by Lemma 4 and Lemma 2, one has

$$H_n \xrightarrow{w} H = g(W(t)) .$$

Since

$$H = \int_0^M W(x) dF_1(x) ,$$

$W(x)$ is a mean zero Gaussian process, one can obtain that H has a normal distribution with mean zero and variance

$$\sigma_2^2 = E(H^2) = \int_0^M \int_0^M \delta_2(x,y) dF_1(x) dF_1(y),$$

where $\delta_2(x,y)$ is the covariance function which is defined from Lemma 2 as

$$\delta_2(x,y) = \bar{F}_2(x)\bar{F}_2(y) \int_0^{x \wedge y} \left(\bar{H}_2(u)\right)^{-2} \bar{G}_2(u) dF_2(u)$$

with $\bar{F}_2(x), \bar{G}_2(x) and\ \bar{H}_2(x)$ defined in Section 5.3. This completes the proof of $I_{n3} \xrightarrow{d} N(0, \sigma_2^2)$.

Since I_{n2} and I_{n3} are independent, we have

$$\sqrt{n}(\hat{P}_{nM} - P_M) \xrightarrow{d} N(0, \sigma^2) ,$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. The proof of the theorem now is completed.

Chapter 6

Jackknifed Interaction Tests Based on Nonparametric Models

6.1 Introduction

As reviewed in Chapter3, Akritas and Bruner (1997) proposed a method for testing the interaction based on the ANOVA (analysis of variance) type decomposition of distribution function, and for the censored data, they used the Kaplan-Meier estimator to replace the empirical distribution function to derive the test. Alternatively, Akritas and LaValley (1996) proposed a method for testing the hypotheses of no main effect and no interaction in factoring designs based on a linear model for a transformation of the survival time.

All of these methods were developed for a general a *x b* factorial design, which results in some complicated formulas for the test statistics. Since in most clinical applications, there are only two treatment groups and also two levels for a baseline covariate, we will first drive some simplified test statistics specifically for a 2 x 2

design for these methods. From the overview of these statistical methods, I found also that the variances of the proposed test statistics are very complicated. This may limit on the application of these statistical methods for the testing of interaction between the treatment and covariate. To overcome this limitation, I develop in this chapter the jackknife estimates for the complicated variances associated with these procedures.

6.2 Jackknifed Interaction Tests Based on Akritas and Bruner's Method

Consider specifically a 2×2 factorial experiment and denote that i=A, B the treatment group and j=1, 2 the levels of the covariate. From each cell (i, j), observed data are $(X_{ijk}, \delta_{ijk}), k = 1,..., n_{ab}$, where X_{ijk} is the minimum between the failure time of interest T_{ijk} and the censoring variable C_{ijk} and $\delta_{ijk} = I(X_{ijk} = T_{ijk})$, where $I(E)$ denotes the indicator of the event E. It is assumed that T_{ijk} are independent and identically distributed according to F_{ij}, C_{ijk} are independent and identically distributed according to G_{ij}, observations in different cells are independent, and censoring variables are independent from the failure time variables. Denote the survival function of the T_{ijk} by S_{ij}, i. e., $S_{ij} = 1 - F_{ij}$. As reviewed in Chapter 3, Akritas, Arnold and Brunner

(1997) proposed that the nonparametric hypotheses of no interaction can be represented as

$$H_0: \ \boldsymbol{CF} = 0,$$

where C is a known matrix, this null hypothesis has been tested through weighted comparisons of the empirical distribution function $\hat{F}_{ij}$ evaluated from each cell. For a 2 x2 design, matrix C can be written down as (1 -1 -1 1), where $\mathbf{F}=(F_{A1}F_{B1}F_{A2}F_{B2)})'$, and the test statistic for H_0 can be defined as

$$C\int_0^T \hat{S}_H d\hat{F},$$

where

$\hat{S}_H = 0.5(Y_{...}(s) + Y_{...}(s-))$ with $Y_{ijk}(t) = I(X_{ijk} \geq t)$, and $Y_{...}(t) = \sum_i \sum_j \sum_k I(X_{ijk} \geq t)$,

$\hat{F} = (\hat{F}_{A1}\hat{F}_{B1}\hat{F}_{A2}\hat{F}_{B2})'$ is the vector of Kaplan-Meier estimators of distribution function from each factor-level combination, and T is the smallest of largest observed survival times in all cells. It is easy to see the test statistics for H_0 can be rewritten as

$$\hat{\mu} = \int_0^T \hat{S}_H \, d\hat{F}_{A1} - \int_0^T \hat{S}_H \, d\hat{F}_{B1} - \int_0^T \hat{S}_H \, d\hat{F}_{A2} + \int_0^T \hat{S}_H \, d\hat{F}_{B2}.$$

Under the basic assumption:

$$\int_0^{\tau_{ij}} \frac{dF_{ij}(s)}{1-G_{ij}(s-)} < \infty, \qquad \forall (i,j)$$

where

$$\tau_{ij} = \sup\{t : (1-F_{ij}(t-))(1-G_{ij}(t-)) > 0\},$$

from Akritas and Brunner (1997) one has under H_0: **CF** = 0 that

$$Q(C) = n\frac{\hat{\mu}^2}{\hat{V}} \xrightarrow{d} \chi_1^2,$$

where

$$\hat{V} = n\left(\frac{\hat{\sigma}_{A1}^2(T)}{n_{A1}} + \frac{\hat{\sigma}_{B1}^2(T)}{n_{B1}} + \frac{\hat{\sigma}_{A2}^2(T)}{n_{A2}} + \frac{\hat{\sigma}_{B2}^2(T)}{n_{B2}}\right)$$

with

$$\hat{\sigma}_{ij}^2(t) = \int_0^t \hat{h}_{ij}(s)\left(1-\frac{\Delta N_{ij\cdot}(s)-1}{Y_{ij\cdot}(s)-1}\right)\frac{dN_{ij\cdot}(s)}{Y_{ij\cdot}(s)}, \qquad \text{(i=A, B and j=1,2)}$$

$$\hat{h}_{ij}(s) = \hat{S}_{ij}^2(s-)\left[\hat{S}_H(s) - \frac{1}{\hat{S}_{ij}(s)}\int_{(s,T]} \hat{S}_H\, d\hat{F}_{ij}\right]^2 \frac{I(s \le T)}{n_{ij}^{-1} Y_{ij\cdot}(s)},$$

$N_{ijk}(t) = I(X_{ijk} \le t, \Delta_{ijk} = 1), N_{ij\cdot}(t) = \sum_{k=1}^{n_{ij}} N_{ijk}(t)$, and $Y_{ij\cdot}(t) = \sum_{k=1}^{n_{ij}} Y_{ijk}(t)$.

Since the estimation of variance $\hat{V}$ is too complicated directly from the above formulas, I will use the delete-one jackknife to estimate the variance. Specifically, let

$\hat{\mu}_{[-k]}$ be the estimate based on the dataset when the observation for the k-th patient is delete out, where *k=1,2…n* and n is the total number of patients. Then the jackknife variance estimate can be calculated based on the following formula:

$$\hat{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{\mu}_{[-k]} - \frac{1}{n}\sum_{k=1}^{n}\hat{\mu}_{[-k]}\right)^2$$

With this estimate, a test statistic for interaction can be written as

$$H = n\frac{\hat{\mu}^2}{\hat{S}^2}$$

and the null hypothesis of no interaction between treatment and covariate can be rejected at α level if $H > \chi^2_{1,1-\alpha}$. This test is called as the jackknifed Akritas-Bruner test.

6.3 Jackknifed Interaction Tests Based on Akritas and LaValley's Method

Consider again a 2x2 factor design with censored data. In this case, the observed data consist of (Y_{ijk}, δ_{ijk}), $i = A, B$ the treatment groups, the two level of a covariate $k = 1,..., n_{ij}$, where n_{ij} is the sample size of the group in which patients receive treatment *i* and have level *j* of the covariates. Let $Y_{ijk} = Y^t_{ijk} \wedge Y^c_{ijk}$, $\delta_{ijk} = I(Y_{ijk} = Y^t_{ijk})$,

where Y_{ijk}^{t} is the true value of the time-to-event variable of interest and Y_{ijk}^{c} is the censoring variable assumed to be independent of Y_{ijk}^{t}. Denote the total sample size as $N = \sum_i \sum_j n_{ij}$. Assume a linear model for the time-to-event variable of interest:

$$Y_{ijk}^{t} = \mu_{ij} + \varepsilon_{ijk} \quad \text{with} \quad \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \qquad (3.4.5)$$

where the error ε_{ijk} are assumed to be independent and identically distributed and parameters in the model satisfy the following constraints:

$$\sum_i \alpha_i = 0, \ \sum_j \beta_j = 0, \ \sum_i \gamma_{ij} = 0, \ \sum_j \gamma_{ij} = 0.$$

Then the hypotheses of no interaction can be written as

$$H_I: \ \gamma_{ij} = 0, \text{ for } i=A, B \text{ and } j=1,2.$$

Let μ denote the 4×1 column vector consisting of the location parameter μ_{ij} defined as

$$\mu = (\mu_{A1}, \mu_{B1}, \mu_{A2}, \mu_{B2})'.$$

Then it is easy to see that $\gamma_{ij} = 0$, for $i=A, B$ and $j=1, 2$, is equivalent to $A_I \mu = 0$ for a matrix A_I =(1,-1,-1,1), which can also be written as:

$$(\mu_{A1} - \mu_{B1}) - (\mu_{A2} - \mu_{B2}) = 0.$$

From the basic idea of Akritas and LaValley as summarized in Chapter 3, we can define $\Delta_{ij,i'j'} = \mu_{ij} - \mu_{i'j'}$ and then estimate them by extended Hodges-Lehmann estimator $\hat{\Delta}_{ij,i'j'}$ based on the censored data. Therefore, the interaction between the treatment and covariate could be estimated by

$$\hat{\mu} = \hat{\Delta}_{A1B1} - \hat{\Delta}_{A2B2}.$$

A test statistics for interaction could then be constructed based $\hat{\mu}$.

From Akritas and LaValley (1996), the extended Hodges-Lehmann estimator $\hat{\Delta}_{AjBj}$(*j=1, 2*) can be obtained by finding b_j which reduce the following to zero:

$$T(b_j) = \sum_{k=1}^{n_{Aj}} \sum_{l=1}^{n_{Bj}} \left(\delta_k I\{Y_{Ajk} - b_j < Y_{Bjl}\} - \delta_l I\{Y_{Bjl} < Y_{Ajk} - b_j\}\right) \doteq 0.$$

The symbol $\doteq$ means that the estimator $\hat{\Delta}_{AjBj}$is defined as the value of b_j setting $T(b_j)$ to zero.

To obtain the a test statistic of interaction based on $\hat{\mu}$, we also need to know the asymptotic distribution of these extended Hodges-Lehmann estimator. Define $D_{Ajk} = (Y_{Ajk}, \delta_{Ajk})$ and $D_{Bjl} = (Y_{Bjl}, \delta_{Bjl})$, where *k=1,…,*$n_{Aj}$ and *l=1, …,*n_{Bj} and let $h(D_{Ajk}, D_{Bjl}, b_j)$ denote the summand on the right side of above formula. The the estimating equation can be written as the generalized *U*–statistic forms (Lee, 1990):

$$U(b_j) = \frac{1}{n_{Aj}n_{Bj}} T(b_j) = \frac{1}{n_{Aj}n_{Bj}} \sum_{k=1}^{n_{Aj}} \sum_{l=1}^{n_{Bj}} h(D_{Ajk}, D_{Bjl}, b_j) \doteq 0 \ (j = 1, 2).$$

Under assumptions that as $n_j \to \infty (n_j = n_{Aj} + n_{Bj})$, ${}^{n_{Aj}}/_{n_j} \to p$ where $0 < p < 1$, it can be shown based on the theory of the generalized U-statistics that

$$n_j^{1/2}(\hat{\Delta}_{AjBj} - \Delta_{AjBj}) \xrightarrow{D} N\left(0, \frac{\xi(0)}{[\xi'(0)]^2}\right).$$

$\xi(0)$ can be consistently estimated by $\hat{\xi}(0) = p^{-1}\hat{\xi}_{1,0} + (1-p)^{-1}\hat{\xi}_{0,1}$ and for any $q \in (0,0.5), n_j^{0.5-q} U(n_j^{-0.5+q} + \hat{\Delta}_{AjBj})$ is a consistent estimator of $\xi'(0)$, where

$$\hat{\xi}_{1,0} = \frac{1}{n_{Aj} - 1} \sum_{k=1}^{n_{Aj}} \left(\frac{1}{n_{Bj}} \sum_{l=1}^{n_{Bj}} h(D_{Ajk}, D_{Bjk}; \hat{\Delta}_{AjBj}) \right)^2,$$

$$\hat{\xi}_{0,1} = \frac{1}{n_{Bj} - 1} \sum_{l=1}^{n_{Bj}} \left(\frac{1}{n_{Aj}} \sum_{k=1}^{n_{Aj}} h(D_{Ajk}, D_{Bjk}; \hat{\Delta}_{AjBj}) \right)^2.$$

These formulas are quite involved and we can also use the delete-one jackknife to estimate directly the variance of $\hat{\mu}$.

Let $\hat{\mu}_{[-k]}$ be the estimate based on the dataset when the observation for the k-th patient is delete out, where *k=1,2…n* and n is the total number of patients. Then the Jackknife variance of $\hat{\mu}$ can be calculated based on the following formula:

$$\hat{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{\mu}_{[-k]} - \frac{1}{n}\sum_{k=1}^{n}\hat{\mu}_{[-k]}\right)^2.$$

With this estimate, the test statistics can be written as

$$H = n\frac{\hat{\mu}^2}{\hat{S}^2}.$$

The null hypothesis of no interaction between treatment and biomarker can be rejected at α level if $H > \chi^2_{1,1-\alpha}$, which will be denoted as the jackknifed Akritas-LaValley test.

In the next chapter, I will present the results of simulation studies which compared the type one errors of these tests with that of the smoothed nonparametric tests proposed in Chapter 5.

Chapter 7

Comparisons and Applications of the Proposed Methods

7.1 Introduction

For the clinical trials with survival endpoints, in the context that there are two treatment groups and a binary baseline covariate, I have first developed in Chapter 5 a non-parametric method based on the smoothed estimate of Patel-Hoel measure to test the interaction between the treatment and a binary baseline covariate. I also proposed to use the jackknife method to estimate the variance for the test statistics derived by Akritas and Bruner (1997) and Akritas and LaValley (1996) to test the interaction based on respectively the ANOVA (analysis of variance) type decomposition of distribution function and a linear model for a transformation of the survival time. In this chapter, I will present results of a systematic comparison between these proposed methods through Monte-Carlo simulations and application to the data from a cancer

clinical trial conducted by NCIC Clinical Trials Group. I also included two traditional methods, the method based on the Cox proportional hazards models with an interaction term and the method of Schemper (1988) reviewed in Chapter 2, in the comparison.

The rest of this chapter is organized as follows: The setup and results of the simulation studies assessing the type I error of tests compared are presented in Section 7.2. Section 7.3 applies these methods to a real data set from the clinical trial. Some conclusions and discussions are presented in the Section 7.4.

7.2 Simulation Studies

In this section, I first summarize the designs of the simulations studies and then the results of these studies.

7.2.1 Design of Simulation Studies

Since the research presented in this thesis was motivated by the biomarker studies using the data from cancer clinical trials, which were described in Chapter 1, the parameters used in the simulations were selected to resemble to these studies. Specifically, I considered the situation where there are two treatment groups (denoted

as A and B) and one biomarker variable with two levels (denoted as 1 and 2). This divides all the patients in the following four groups denoted as respectively A1, A2, B1, and B2, where the first letter denotes the treatment group a patient was allocated and the second number the level of the biomarker for this patient.

As in the simulation studies presented in Chapter 5, two settings have been investigated: the case where the hazard rates in the 4 groups (A1, A2, B1, B2) are proportional and the case where they are not proportional. In this former case, the distributions of the survival times in the two biomarker groups were assumed to be exponential with parameters respectively $\lambda_{A1} = \lambda_{A2} = 0.1$ for patients in treatment group A and with parameters respectively $\lambda_{B1} = \lambda_{B2} = 0.1\gamma$ (where γ varies from 1 to 2) for patients in treatment group B. In the late case, the distributions of the survival times in the two biomarker groups were assumed to be log-normal with parameters respectively $\mu_{A1} = \mu_{A2} = 1.95$ (all the shape parameters from log normal distributions are assumed as 1) for patients in treatment group A and with parameters respectively $\mu_{B1} = \mu_{B2} = 1.95 - \beta$ (where β varies from 0 to 0.8 and shape parameters are also assumed as 1) for patients in treatment group B.

For the censoring, as in Chapter 5, I also assumed the unit for the survival times is months and generated the censored survival times based on the following

clinical trial process where patients were recruited with a uniform distribution over T_a months and followed for at least T_f months. This is equivalent to assume that the censoring distribution for each patient is uniform over [T_a, T_a +T_f]. Three different uniform censoring distribution were simulated for each scenario: U[6, 12], U[12, 18] and U[18, 24]. The censoring rates with various parameters that are used in simulation studies are displayed in Table 7-1. In this table, $EXP(\lambda)$ represents the exponential distribution with parameter λ, and $LN(\mu, \sigma^2)$ represents log-normal distribution with location parameter μ and shape parameter σ^2.

Table 7-1: Censoring rate

Exponential distribution			
	EXP (0.1)	*EXP (0.15)*	*EXP (0.2)*
U [6, 6]	41%	27%	18%
U [6, 12]	23%	11%	5%
U [6, 18]	12%	4%	4%
Log Normal distribution			
	LN (1.95,1)	LN (1.55,1)	LN (1.15,1)
U [6, 6]	41%	27%	16%
U [6, 12]	23%	13%	6%
U [6, 18]	14%	7%	3%

In each simulation, it was assumed that the total sample sizes in the two treatment groups were the same. The proportions of patients with two levels of the biomarker were also the same in the two treatment groups. This implies that the treatment was balanced by the biomarker, which is the case in a randomized clinical trial. Three different total sample sizes (n=100, 150, and 200) per treatment group and two different proportions of biomarker levels were considered for all the scenarios described above.

In the simulations, the random samples were generated using the R functions *rexp* for the exponential distribution or *rlnorm* for the log normal distribution. For the proposed method based on the smoothed estimate, kernel function $k(x) = \frac{15}{16}(1-x^2)^2$ was used to define the kernel estimate. Since simulation results presented in Chapter 5 showed that bandwidth h_e proposed by Wang and Wang (2007) or h_{kp} proposed by Kuhn and Padgett (1997) generated similar results, only bandwidth h_e proposed by Wang and Wang (2007) was used in the simulation studies presented in this Chapter.

Under each scenario of the simulations, a set of survival data with fixed parameters and sample size was first generated. The p-value from each test included

in the comparison was then calculated. The p-value was compared with nominal level $\alpha = 0.05$ to decide the whether non-hypothesis was rejected or not. This was repeated for 1,000 times. The type I error of a test was estimated from the proportion of hypotheses being rejected in these1,000 trials.

7.2.2 Results of Simulation Studies

The simulation results are presented in Tables 7.2 – 7.5. In the tables, CEN, Smoothed, Schemper, JAL, JAB and Cox, represent respectively, uniform censoring distribution, the results from respectively, smoothed test, Schemper test, jackknifed Akritas-LaValley test, jackknifed Akritas-Bruner test and the test based on the Cox proportion hazard model with an interaction term. Other symbols used in the tables are p, γ_1, γ_2, β_1 and β_2. "p" represents the percentage of patients in the first level of the biomarker. $\gamma_1$and $\gamma_2$are the ratios of the hazard rates between treatment B and treatment A in respectively biomarker levels 1 and 2 for the exponential distributions. β_1 and β_2 are the difference in mean between arm B and arm A in biomarker levels 1 and 2 for the log normal distributions.

From the results summarized in these tables, we may first observe that all the methods except the jackknifed Akritas-LaValley test produce the comparatively satisfying results. The type I errors from the jackknifed Akritas-LaValley test are

much lower than the nominal level 0.05 regardless of whether the proportional hazard assumption is satisfied or not and the 2 levels of biomarker were balanced or not. Type I errors from all other tests are all close to the nominal level. When the proportional hazard assumption is satisfied and 2 levels of biomarker are balanced (Table 7-2), the type I errors from the smoothed test are below nominal level in almost cases, while the type I errors from the Schemper test, jackknifed Akrita-Bruner test and the test based on the Cox model are larger than the nominal level in most cases when the treatment difference is large and sample size is high. But when the proportional assumption is satisfied and 2 levels of biomarker were imbalanced (Table 7-3), the type I errors from the smoothed test become larger and in some cases are higher than the nominal level, especially when the treatment difference and sample size are large. When the proportional assumption is not satisfied (Table 7-4 and Table 7-5), the type I error from the smoothed test are very stable and, regardless of the treatment difference and sample size, they are very close to nominal level. The Schemper test, jackknifed Akritas-Bruner test and the test based on the Cox model became, however, a little bit more conservative compared to the corresponding cases when the proportional assumption is satisfied. In most cases, their type I errors were below the nominal level.

Table 7-2: Type I Errors of Proposed Tests for Treatment-Covariate Interaction Proportional Hazard Rates (λ_{A1}=0.1, λ_{A2}=0.1, p=0.5)

p	n_A	n_B	CEN	γ_1	γ_2	Smoothed	Schemper	JAL	JAB	Cox
0.5	100	100	U [6,6]	1	1	0.038	0.050	0.022	0.044	0.040
			U [6,18]	1	1	0.046	0.036	0.008	0.048	0.036
			U [6,24]	1	1	0.052	0.034	0.016	0.036	0.040
0.5	150	150	U [6,6]	1	1	0.034	0.040	0.006	0.038	0.048
			U [6,18]	1	1	0.022	0.042	0.016	0.054	0.046
			U [6,24]	1	1	0.028	0.038	0.008	0.040	0.046
0.5	200	200	U [6,6]	1	1	0.047	0.062	0.016	0.058	0.054
			U [6,18]	1	1	0.044	0.060	0.014	0.062	0.054
			U [6,24]	1	1	0.042	0.059	0.014	0.063	0.053
0.5	100	100	U [6,6]	1.5	1.5	0.064	0.032	0.016	0.044	0.044
			U [6,18]	1.5	1.5	0.024	0.038	0.014	0.044	0.042
			U [6,24]	1.5	1.5	0.034	0.040	0.020	0.038	0.048
0.5	150	150	U [6,6]	1.5	1.5	0.030	0.042	0.010	0.054	0.056
			U [6,18]	1.5	1.5	0.038	0.052	0.024	0.042	0.048
			U [6,24]	1.5	1.5	0.026	0.044	0.022	0.044	0.048
0.5	200	200	U [6,6]	1.5	1.5	0.04	0.058	0.022	0.063	0.064
			U [6,18]	1.5	1.5	0.04	0.061	0.022	0.060	0.056
			U [6,24]	1.5	1.5	0.043	0.061	0.019	0.059	0.054
0.5	100	100	U [6,6]	2	2	0.040	0.038	0.014	0.036	0.036
			U [6,18]	2	2	0.032	0.034	0.030	0.042	0.046
			U [6,24]	2	2	0.030	0.038	0.030	0.044	0.050
0.5	150	150	U [6,6]	2	2	0.022	0.052	0.026	0.056	0.054
			U [6,18]	2	2	0.042	0.046	0.024	0.046	0.058
			U [6,24]	2	2	0.036	0.048	0.032	0.044	0.056
0.5	200	200	U [6,6]	2	2	0.037	0.060	0.02	0.064	0.060
			U [6,18]	2	2	0.041	0.062	0.03	0.064	0.055
			U [6,24]	2	2	0.039	0.063	0.027	0.060	0.053

Table 7-3: Type I Errors of Proposed Tests for Treatment-Covariate Interaction Proportional Hazard Rates (λ_{A1}=0.1, λ_{A2}=0.1, p=0.3)

p	n_A	n_B	CEN	γ_1	γ_2	Smoothed	Schemper	JAL	JAB	Cox
0.3	100	100	U [6,6]	1	1	0.044	0.042	0.014	0.038	0.048
			U [6,18]	1	1	0.052	0.040	0.012	0.052	0.050
			U [6,24]	1	1	0.044	0.036	0.004	0.036	0.040
0.3	150	150	U [6,6]	1	1	0.048	0.050	0.018	0.048	0.048
			U [6,18]	1	1	0.064	0.052	0.018	0.050	0.046
			U [6,24]	1	1	0.066	0.052	0.026	0.050	0.048
0.3	200	200	U [6,6]	1	1	0.058	0.076	0.022	0.072	0.066
			U [6,18]	1	1	0.054	0.074	0.026	0.076	0.068
			U [6,24]	1	1	0.050	0.068	0.018	0.070	0.062
0.3	100	100	U [6,6]	1.5	1.5	0.068	0.046	0.014	0.052	0.042
			U [6,18]	1.5	1.5	0.046	0.040	0.018	0.034	0.040
			U [6,24]	1.5	1.5	0.030	0.040	0.030	0.034	0.046
0.3	150	150	U [6,6]	1.5	1.5	0.070	0.048	0.010	0.052	0.050
			U [6,18]	1.5	1.5	0.048	0.052	0.026	0.056	0.048
			U [6,24]	1.5	1.5	0.052	0.050	0.034	0.048	0.054
0.3	200	200	U [6,6]	1.5	1.5	0.052	0.068	0.026	0.076	0.064
			U [6,18]	1.5	1.5	0.072	0.058	0.014	0.062	0.060
			U [6,24]	1.5	1.5	0.056	0.058	0.032	0.060	0.052
0.3	100	100	U [6,6]	2	2	0.060	0.046	0.022	0.050	0.044
			U [6,18]	2	2	0.036	0.034	0.026	0.036	0.048
			U [6,24]	2	2	0.042	0.036	0.046	0.034	0.044
0.3	150	150	U [6,6]	2	2	0.054	0.052	0.020	0.050	0.046
			U [6,18]	2	2	0.056	0.058	0.038	0.052	0.046
			U [6,24]	2	2	0.048	0.056	0.052	0.052	0.054
0.3	200	200	U [6,6]	2	2	0.066	0.060	0.022	0.064	0.052
			U [6,18]	2	2	0.058	0.054	0.026	0.058	0.056
			U [6,24]	2	2	0.058	0.058	0.038	0.062	0.058

Table 7-4: Type I Errors of Proposed Tests for Treatment-Covariate Interaction Non-Proportional Hazard Rates (μ_{A1}=1.95, μ_{A2}=1.95, p=0.5)

p	n_A	n_B	CEN	β_1	β_2	Smoothed	Schemper	JAL	JAB	Cox
0.5	100	100	U [6,6]	0	0	0.040	0.054	0.020	0.060	0.040
			U [6,18]	0	0	0.042	0.046	0.020	0.054	0.050
			U [6,24]	0	0	0.044	0.038	0.014	0.040	0.046
0.5	150	150	U [6,6]	0	0	0.042	0.062	0.020	0.062	0.060
			U [6,18]	0	0	0.044	0.060	0.016	0.062	0.066
			U [6,24]	0	0	0.052	0.060	0.028	0.056	0.058
0.5	200	200	U [6,6]	0	0	0.046	0.054	0.024	0.066	0.060
			U [6,18]	0	0	0.030	0.052	0.010	0.056	0.044
			U [6,24]	0	0	0.046	0.056	0.012	0.050	0.052
0.5	100	100	U [6,6]	0.4	0.4	0.042	0.054	0.026	0.058	0.044
			U [6,18]	0.4	0.4	0.036	0.042	0.022	0.044	0.050
			U [6,24]	0.4	0.4	0.038	0.042	0.022	0.046	0.048
0.5	150	150	U [6,6]	0.4	0.4	0.052	0.064	0.028	0.076	0.068
			U [6,18]	0.4	0.4	0.036	0.064	0.026	0.058	0.060
			U [6,24]	0.4	0.4	0.040	0.058	0.022	0.062	0.056
0.5	200	200	U [6,6]	0.4	0.4	0.038	0.050	0.010	0.054	0.050
			U [6,18]	0.4	0.4	0.040	0.054	0.016	0.056	0.052
			U [6,24]	0.4	0.4	0.048	0.050	0.022	0.058	0.042
0.5	100	100	U [6,6]	0.8	0.8	0.034	0.044	0.024	0.062	0.046
			U [6,18]	0.8	0.8	0.036	0.040	0.042	0.040	0.048
			U [6,24]	0.8	0.8	0.034	0.038	0.044	0.038	0.056
0.5	150	150	U [6,6]	0.8	0.8	0.040	0.058	0.022	0.062	0.056
			U [6,18]	0.8	0.8	0.040	0.052	0.024	0.058	0.060
			U [6,24]	0.8	0.8	0.048	0.052	0.034	0.052	0.058
0.5	200	200	U [6,6]	0.8	0.8	0.040	0.050	0.012	0.056	0.040
			U [6,18]	0.8	0.8	0.042	0.050	0.032	0.054	0.040
			U [6,24]	0.8	0.8	0.048	0.050	0.028	0.050	0.046

Table 7-5: Type I Errors of Proposed Tests for Treatment-Covariate Interaction Non-Proportional Hazard Rates (μ_{A1}=1.95, μ_{A2}=1.95, p=0.3)

p	n_A	n_B	CEN	β_1	β_2	Smoothed	Schemper	JAL	JAB	Cox
0.3	100	100	U [6,6]	0	0	0.070	0.052	0.014	0.064	0.050
			U [6,18]	0	0	0.052	0.038	0.018	0.042	0.044
			U [6,24]	0	0	0.058	0.038	0.020	0.038	0.052
0.3	150	150	U [6,6]	0	0	0.048	0.052	0.024	0.056	0.050
			U [6,18]	0	0	0.044	0.048	0.012	0.048	0.056
			U [6,24]	0	0	0.062	0.052	0.020	0.052	0.042
0.3	200	200	U [6,6]	0	0	0.058	0.042	0.014	0.044	0.050
			U [6,18]	0	0	0.044	0.038	0.008	0.038	0.036
			U [6,24]	0	0	0.022	0.040	0.008	0.038	0.036
0.3	100	100	U [6,6]	0.4	0.4	0.050	0.048	0.026	0.036	0.054
			U [6,18]	0.4	0.4	0.054	0.040	0.012	0.044	0.050
			U [6,24]	0.4	0.4	0.048	0.042	0.020	0.042	0.050
0.3	150	150	U [6,6]	0.4	0.4	0.038	0.052	0.024	0.060	0.052
			U [6,18]	0.4	0.4	0.048	0.042	0.018	0.048	0.044
			U [6,24]	0.4	0.4	0.044	0.044	0.030	0.048	0.046
0.3	200	200	U [6,6]	0.4	0.4	0.048	0.052	0.014	0.040	0.048
			U [6,18]	0.4	0.4	0.036	0.042	0.012	0.040	0.038
			U [6,24]	0.4	0.4	0.024	0.042	0.028	0.042	0.042
0.3	100	100	U [6,6]	0.8	0.8	0.058	0.054	0.026	0.052	0.056
			U [6,18]	0.8	0.8	0.048	0.042	0.042	0.048	0.048
			U [6,24]	0.8	0.8	0.048	0.042	0.046	0.046	0.060
0.3	150	150	U [6,6]	0.8	0.8	0.052	0.052	0.024	0.058	0.048
			U [6,18]	0.8	0.8	0.044	0.048	0.024	0.046	0.048
			U [6,24]	0.8	0.8	0.040	0.050	0.022	0.046	0.054
0.3	200	200	U [6,6]	0.8	0.8	0.048	0.044	0.018	0.050	0.056
			U [6,18]	0.8	0.8	0.032	0.042	0.032	0.038	0.040
			U [6,24]	0.8	0.8	0.028	0.046	0.046	0.038	0.046

7.3 Application to Data from a Clinical Trial

In this section, we make a direct comparison of methods proposed in this thesis by analyzing the biomarker studies from a randomized clinical trial conducted by NCIC Clinical Trial Group (CTG). The original trial (Levine et al., 1998 and 2005) was designed to compare two chemotherapy regimens, Cyclophosphamide-Methotrexate-Fluorouracil (CMF) versus Cyclophosphamide-Epirubicin-Fluorouracil (CEF), in women with early stage breast cancer and showed the CEF is superior to CMF in disease-free and overall survival. Because of toxicity associated with CEF, it is desirable to find a biomarker which would predict the benefits of the CEF.

Two biomarkers have been studied recently by NCIC CTG investigators: HER2 (Pritchard et al., 2006) and TOP2A (O'Malley, 2009). In this analysis, I included 494 patients with expression data for both biomarkers and considered only the overall survival endpoint. The average censoring rate was 61.3% for overall survival. Patient's distribution among the combination of the treatment group and the level of each biomarker are presented in Table 7-6.

Results based on the univariate Cox models were summarized in details in Section 5.6. Table 7-7 presents the p-values for the testing the interaction between

treatment and biomarker HER2 and TOP2A respectively from all the tests studied in the simulation studies presented in the last section. We may see from Table 7-7, no significant interaction between the treatment and HER2 at 0.05 level was found by any of the tests. Interaction p-values from the smoothed test, the jackknifed Akristas-LaValley test and the test based on the Cox model were close to 0.05, while the p-values from Schempter test and the jackknifed Akristmas-Bruner test were close to each other and both were much bigger than 0.05. Significant interaction between the treatment and TOP2A was found by all methods except the smoothed test which had a very high p-value. This may be because the smoothed test is too conservative and thus, has very low power in this situation.

Table 7-6: Distributions of Treatment and Biomarkers

Genes		Treatment Group	
		CEF	CMF
HER2	Over-expressed	58 (24.0%)	70 (27.7%)
	Normal	182 (75.5%)	181 (71.5%)
	Missing	1 (0.5%)	2 (0.8%)
TOP2A	Alteration	41 (17%)	33 (13.0 %)
	Normal	157 (65.1%)	175 (69.2%)
	Missing	43 (17.0%)	45 (17.8%)

Table 7-7: Interaction p-values from Different Tests

	Treatment-HER2 interaction	Treatment-TOP2A interaction
Smoothed Test	0.05167	0.325
Schemper Test	0.1487	0.0068
Jackknifed Akritas-LaValley Test	0.072	0.0188
Jackknifed Akritas-Bruner Test	0.1404	0.007
Test based on the Cox model	0.0553	0.0048

7.4 Conclusion and Discussions

In this chapter, the smoothed test proposed in Chapter 5 and the jackknifed tests based on nonparametric models derived in Chapter 6 for the testing of the interaction between the treatment and biomarker are evaluated and compared through Monte-Carlo simulations and applications to a real data set from a cancer clinical trial.

From the simulation results, we observed that the jackknifed Akritas-LaValley method was conservative in almost all the scenarios that were studied. The reason may be because of the linear model assumption behind this method. The smoothed test performed better in controlling the type I error compared with all other methods, such as, the Schemper test, jackknifed Akritas-Bruner test and the test based on the Cox model. The type I error of the smoothed test was below but very close the nominal level 0.05, while the type I errors from the Schemper test, the jackknifed Akritas-Brunner test and the test based on the Cox model were higher than 0.05 in most scenarios when the treatment difference and sample size are larger. Since controlling type I error is important in clinical practice to avoid that a useless biomarker is identified as useful, the smoothed test would be the test I recommend to use in practice.

In the simulation studies presented in this chapter, I only considered comparisons of the tests studied based on their true type I errors. It is expected that a conservative test would have lower power based on the relationship between type I and II errors but a fair comparison between these tests based on their power is difficult. This is because the definitions of interaction are different among these tests and, therefore, it is difficult to determine the alternatives which have the same degrees of departure from the null hypothesis of no interaction for all of these tests to calculate the powers of these tests.

Chapter 8

Summary and Discussions

In this thesis, the current literatures on the testing of the interaction between treatment and covariates were first reviewed. I then developed a new nonparametric procedure based on the smoothed estimate of the Patel-Hoel measure for the interaction between treatment and a covariate in a 2 x 2 design with censored data. I also simplified the test statistics for some interaction tests proposed in the literature based on nonparametric models in the context of a 2x2 design and apply the jackknife method to estimate the variance of these test statistics to make these tests easily applicable to analysis of real data. All the test procedures proposed/derived in this thesis were directly compared in term of the type I error under various scenarios by using the Monte-Carlo simulations and through applications to the analysis of real data from cancer clinical trials. The following are the summary of the major results in this thesis:

In Chapter 5, a non-parametric procedure based on the smoothed estimate of the Patel–Hoel measure of interaction was derived to test the interaction between the

treatment and a binary covariate with censored data. The theoretical distribution of the test statistic of the proposed procedure was derived. The proposed procedure was evaluated by Monte-Carlo simulations and applied to a real data set from a cancer clinical trial. The results of the simulations showed that actual type I errors of proposed tests are well controlled and are very close to the nominal significance level.

In Chapter 6, the test statistics proposed to test the treatment-covariate interaction by respectively Akritas and Bruner (2007) based on ANOVA type decomposition of distribution function and Akritas and LaValley (2006) based on linear model for the survival times were first simplified for a study with two treatment groups and a covariate with two levels. Since the variances of these test statistics are still very complicated for the application in analysis of data from clinical trials, the jackknife method was applied to estimate these variances. Two versions of jackknifed tests for treatment-covariate interaction were defined in this chapter.

In Chapter 7, the true type I errors of the interaction tests defined in Chapters 5 and 6 were compared through Monte-Carlo simulations and applications to a real data set from a cancer clinical trial. From the simulation results, it was observed that jackknifed Akritas-LaValley test was too conservative in almost all the scenarios that were conducted. The smoothed test performed better than other methods, such as the

Schemper test, the jackknifed Akritas-Bruner and the test based on the Cox model, in controlling the type I error and, thus, was recommended in practical applications.

In this thesis, I only considered the situation where the baseline covariate has only two levels. Although this is the situation encountered in most of the applications, there are also examples where the baseline covariate has more than two levels. For example, there were three categories in the original definition of TOP2A gene: amplified, normal, deleted. The procedures proposed in this thesis can be generalized to this situation with little more complicated algebra.

For the smoothed test, I considered two methods for the selection of bandwidth: one based on Mean Integrated Squared Error (MISE) criteria and another Mean Absolute Error (AMAE) criteria. Those bandwidths were optimal for estimation of density function but it is not known whether they are also optimal for hypothesis testing. Determination of the bandwidth based on criteria which minimize the type I error or maximize the power of the smoothed test would be an interesting problem for further research but it would be mathematically very challenging since asymptotic expansion for the type I error of a test is mathematically very difficult.

Bibliography

1. Akritas, M. G. (2004). Nonparametric Survival Analyses, *Statistical Sciences*, 19, 615-623.

2. Akritas, M.G. and Arnold, S.F. (1994), Fully Nonparametric Hypotheses for Factorial Designs I: Multivariate Repeated Measures Designs, *Journal of the American Statistical Association,* 89, 336-343.

3. Akritas, M. G., Arnold, S. F. and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of American Statistical Association,* 92, 258-265.

4. Akritas, M. G. and Brunner, E. (1997). Nonparametric methods for factorial designs with censored data, *Journal of American Statistical Association*, 92, 568-576.

5. Akritas, M. G. and LaValley, M. P. (1996). Nonparametric inference in factorial designs with censored data, *Biometrics,* 52, 913-924.

6. Bassiakos, Y., Meng, X., and Lo, S. (1991). A general estimator of the treatment effect when the data are heavily censored, *Biometrika*, 78, 741-748.

7. Billingsley, P. (1968). *Weak Convergence of Probability Measures*, Wiley, New York.

8. Breslow, N. E. (1970). A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship, *Biometrika*, 57, 579-594.

9. Brillinger, D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likehood estimate. *Rev. Inst. Internat. Statist.* 32, 202-206.

10. Brillinger, D. R. (1977). Approximate estimation of the standard errors of complex statistics based on sample surveys. *New Zealand Statistician* 11, 35-41.

11. Brunner, E., Puri , M.L. and Sun, S. (1995). Nonparametric methods for stratified two-sample sesigns with application to multi clinic trials. *Journal of American Statistical Association*. 90, 1004-1014.

12. Brown, B. W., Hollander, M., and Korwar, R. M. (1974). Nonparametric tests of independence for censored data with applications to heart transplant studies. *In Reliability and Biometry: Statistical Analysis of Lifelength,* 327-354, SIAM Philadelphia.

13. Cox, D. R. (1984). Interaction, *International Statistical Review*, 52, 1-31.

14. Durbin, J. (1973). *Distribution Theory for Tests Based on the sample Distribution Function.* SIAM Philadephia.

15. Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Statist.,* 7, 1-26.

16. Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *Amer. Statist.,* 37, 36-48.

17. Fleiss, J. E. (1986). Analysis of data from multiclinic trials, *Controlled Clinical Trials*, 7, 267-275

18. Folder, A., Rejto, L. and Winter B.B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, II: estimation of density and failure rate. *Periodica Mathematica Hungarica*, 12, 15-29.

19. Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41, 361-372.

20. Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored sample, *Biometrika*, 52, 203-223

21. Ghorai, J. K. and Susarla, V. (1990). Kernel estimation of a smooth distribution function based on censored data, *Metrika*, 37, 71-86.

22. Goss, P.E., Ingle, J. N., Robert, N. J., Muss, H. B., Piccart, M. J., Castiglione, M., Tu, D., Shepherd, L. E., Pritchard, K. L., Livingston, R. B., Davidson, N. E., Norton, L., Perez, E. A., Abrams, J. S., Therasse, P., Palmer, M. J., and Pater, J. L. (2003). A randomized trial of letrozole in postmenopausal women after five years of tomoxifen therapy for early-stage breast cancer, *N. Engl. J. Med.*, 349, 1793-1802.

23. Hardle, W., Muller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametrics and semiparameteric models*, Berlin/Heidelberg, Springer.

24. Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank test, *Annals of Mathematical Statistics*, 34, 598-611.

25. Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observation, *J. Amer. Statist. Assoc.* 53, 457-481.

26. Karapetis, C. S., Khambata-Ford, S. , Jonker, D. J. , O'Callaghan, C. J. , Tu, D. , Tebbutt, N. C. , Simes, R. J. , Chalchal, H. , Shapiro, J. , Robitaille, S. , Price, T. , Shepherd, L. , Au, H-J. , Langer, C. , Moore, M. J. , and Zalcberg, J. R. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*, 359, 1757-1765.

27. Klein, J. P. and Moeschberger, M. L. (2003). *Survivals analysis: Techniques for censored and truncated data, second edition.* New York, Springer.

28. Koziol, J. A. and Jia, Z., (2009). The concordance index C and Mann-Whitney parameter Pr (X>Y) with randomly censored data. *Biometrical Journal*, 51, 467-474.

29. Kuhn, J. W. and Padgett, W. J. (1997). Local bandwidth selection for kernel density estimation from right-censored data based on asymptotic mean absolute error, *Nonlinear analysis, Theory, Methods & Applications.* 30, 4375-4384.

30. Lee, A. J. (1990). *U-Statistics: Theory and Practice*, New York: Marcel Dekker.

31. Levine, M. N., Bramwell, V. H. C., Pritchard, K. I., Norris, B. D., Shepherd, L. E., Abu-Zahra, H., Findlay, B., Warr, D., Bowman, D., Myles, J., Arnold, A., Vandenberg, T., MacKenzie, R., Robert, J., Ottaway, J., Burnell, M., Williams, C. K. O., and Tu, D. (1998). A Randomized Trial of Cyclophosphamide, Epirubicin, Fluorouracil Chemotherapy Compared with Cyclophosphamide, Methotrexate, Fluorouracil in Premenopausal Women with Node Positive Breast Cancer, *Journal of Clinical Oncology*, 16, 2651-2658.

32. Levine, M. N., Pritchard, K. I., Bramwell, V. H. C., Shepherd, L. E., Tu, D., Paul, N. (2005). Randomized Trial Comparing Cyclophosphamide, Epirubicin, and Fluorouracil With Cyclophosphamide, Methotrexate, and Fluorouracil in Premenopausal Women With Node-Positive Breast Cancer: Update of National Cancer Institute of Canada Clinical Trials Group Trial MA5, *Journal of Clinical Oncology*, 23, 5166–5170.

33. Marron, J. S. and Padgett, W. J. (1987). Asymptotically optimal bandwidth selection from randomly right-censored sample, *The Annals of Statistics* 15, 1520-1535.

34. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, 50, 163-170.

35. Miller, R. G. (1964). A trustworth jackknife. *Ann. Math. Statist.* 35, 1594-1605.

36. Miller, R. G. (1974). The jackknife: a review. *Biometrika* 61 1-15.

37. Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data, *The Annals of Statistics*, 14, 766-773.

38. Muss, H. , Tu, D. , Ingle, J. , Martino, S. , Robert, N. , Pater, J. , Whelan, T. , Palmer, M. J. , Piccart, M. , Shepherd, L. E. , Prichard, K. I. , He, Z. , Goss, P. E. (2008). Efficacy, toxicity and quality of life in older woman with early breast cancer treated with letrozole or placebo after five years of tamoxifen:

NCIC CTG intergroup trial MA. 17, *Journal of Clinical Oncology*, 26, 1956-1964.

39. Nenci, I. (1981). Estrogen receptor cytochemistry in human breast: status and prospects, *Cancer* 48, 2674 – 2686.

40. O'Malley, F. P., Chia, S., Tu, D., Shepherd, L. E., Levine, M. N., Bramwell, V. H., Andrulis, I. L., Pritchard, K. I. (2009). Topoisomerase II Alpha and Responsiveness of Breast Cancer to Adjuvant Chemotherapy, *Journal of National Cancer Institute*, 101, 644-650.

41. Pandha, H. S., Waxman, J., and Sikora K. (1994). Tumor Marker, *British J. of Hospl. Med,* 51, 298-303.

42. Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics,* 33, 1065-1076.

43. Patel, K. M. and Hoel, D.G. (1973). A Nonparametric Test for Interaction in Factorial Experiments, *Journal of the American Statistical Association,* 68, 615-620.

44. Pritchard, K. I., Levine, M. N. and Tu, D. (2003). Neu/erb B-2 Overexpression and Response to Hormonal Therapy in Premenopausal Women in the Adjuvant Breast Cancer Setting: Will it Play in Peoria? Part II, *Journal of Clinical Oncology*, 21, 399-400.

45. Pritchard, K. I., Shepherd, L. E., O'Malley, F. P., Andrulis, I. L., **Tu, D.,** Bramwell, V. H., Mark N. Levine, M. N. (2006). HER2/neu and Responsiveness of Breast Cancer to Adjuvant Chemotherapy, *New England Journal of Medicine,* 354, 2103-2111.

46. Puri, M. L. and Sen, P. K. (1970). Nonparametric Methods in Multivariate Analysis, *New York: John Wiley and Sons, Inc*., 1970.

47. Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* 43, 353-360.

48. Rao, C. R. and D. Tu (1990). Inference on the occurrence/exposure rate with mixed censoring models, *Calcutta Statistical Association Bulletin*, 40, 65-87.

49. Reeds, J. A. (1978). Jackknifing maximum likelihood estimates. *Ann. Statist.*, 6, 727-739.

50. Rosenblatt, M. (1956). Remark on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, 27, 832-837.

51. Savage, I. R. (1956). Contributions to the theory of rank order statistics – the two sample case, *Annals of Mathematical Statistics*, 27, 590-615.

52. Schemper, M. (1988). Non-parametric analysis of treatment-covariate interaction in the presence of censoring. *Statistics in Medicine*. 7, 1257-1266.

53. Scott, D. W. and Wand, M. P. (1991). Feasibility of multivariate density estimates, *Biometrika* 78, 197 - 205

54. Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.

55. Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, London, Chapman and Hall.

56. Simonoff, J. S. (1996). *Smoothing method in statistics*, New York, Springer.

57. Trape, J., Buxo J., De Olaguer, J. P., and Vidal, C. (2003). Tumor markers as prognostic factors in treated non-small cell lung cancer, *Anticancer Research*, 23, 2674 – 2686.

58. Tu, D. and Pater, J. (1999). Some issues related to applications of Cox's models in cancer clinical trials, *Controlled Clinical Trials*, 20, 84S.

59. Tu, D. and Gou, S. (2004). A Comparison of Statistical Tests for the Interaction of Treatment and Biomarker in Tumor Correlative Studies, *Presented at IBC 2004 (XXIInd International Biometric Conference) in parallel with ASC 2004 (2004 Australia Statistical Conference), Cairns, Queensland, Australia.*

60. Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Ann. Math. Statist.* 29, 614.

61. Wang, B. and Wang, X. F. (2010). Bandwidth Selection for Weighted Kernel Density Estimation, Technical Report, Mathematics and Statistics Department, University of South Alabama, Mobile, Alabama.

62. Whittle, P. (1958). On the smoothing of probability density functions, *J. Roy. Statist. Soc. B.*, 20, 334-343.

63. Wu, C. F. J. (1986). Jackknife, Bootstrap and other Resampling Methods in Regression Analysis, *Annals of Statistics*, 14, 1343-1350.

Appendix

R Code for Simulations

#######define some functions

R function for Smoothed Method

```
Pdiff <- function (datain) {
x1=datain$x[datain$arm=='A1']
d1=datain$event[datain$arm=='A1']
km1 <- survfit(Surv(x1, d1) ~ 1)
jump1=c(1,km1$surv[-length(km1$surv)])-km1$surv
x2=datain$x[datain$arm=='A2']
d2=datain$event[datain$arm=='A2']
km2 <- survfit(Surv(x2, d2) ~ 1)
jump2=c(1,km2$surv[-length(km2$surv)])-km2$surv
x3=datain$x[datain$arm=='B1']
d3=datain$event[datain$arm=='B1']
```

```
km3 <- survfit(Surv(x3, d3) ~ 1)
jump3=c(1,km3$surv[-length(km3$surv)])-km3$surv
x4=datain$x[datain$arm=='B2']
d4=datain$event[datain$arm=='B2']
km4 <- survfit(Surv(x4, d4) ~ 1)
jump4=c(1,km4$surv[-length(km4$surv)])-km4$surv
##### smooth CDF function for group1
B1=min(sum(x1)/sum(d1),IQR(x1)/1.34)
h1=2.3* B1 * round(sizea1 ^ (-0.2),2)
Fhat1<-function(x)
{
s=rep(0,length(x))
for (i in 1:length(x))
s[i]=sum(Kn((x[i]-km1$time)/h1)*jump1)
s
}
##### smooth density function for group3
B3=min(sum(x3)/sum(d3),IQR(x3)/1.34)
```

```
h3=2.3* B3 * round(sizeb1 ^ (-0.2),2)


fhat3<-function(x)
{
s=rep(0,length(x))
for (i in 1:length(x))
s[i]=sum(kn((x[i]-km3$time)/h3)*jump3/h3)
s
}
T1=min(max(km1$time[km1$surv >0]), max(km3$time[km3$surv >0]))
inte1<-function(z1) Fhat1(z1)*fhat3(z1)
P1=integrate(inte1,0,T1, subdivisions=10000)$value
##### smooth CDF function for group2
B2=min(sum(x2)/sum(d2),IQR(x2)/1.34)
h2=2.3* B2 * round(sizea2 ^ (-0.2),2)
Fhat2<-function(x)
{
s=rep(0,length(x))
```

```
for (i in 1:length(x))
s[i]=sum(Kn((x[i]-km2$time)/h2)*jump2)
s
}
##### smooth density function for group4
B4=min(sum(x4)/sum(d4),IQR(x4)/1.34)
h4=2.3* B4 * round(sizeb2 ^ (-0.2),2)
fhat4<-function(x)
{
s=rep(0,length(x))
for (i in 1:length(x))
s[i]=sum(kn((x[i]-km4$time)/h4)*jump4/h4)
s
}
T2=min(max(km2$time[km2$surv >0]), max(km4$time[km4$surv >0]))
inte2<-function(z2) Fhat2(z2)*fhat4(z2)
P2=integrate(inte2,0,T2, subdivisions=10000)$value
P=P1-P2
```

```
}


##### R function for Schemper Method #####
patho<- function(grp1="A1", grp2="A2", indata=X){
        suv1<-indata$x[indata$arm==grp1]
    eve1<-indata$event[indata$arm==grp1]
        suv2<-indata$x[indata$arm==grp2]
    eve2<-indata$event[indata$arm==grp2]
    yi<-matrix(suv1, nrow=length(suv1), ncol=length(suv2), byrow=F)
        ei<-matrix(eve1, nrow=length(suv1), ncol=length(suv2), byrow=F)
    yj<-matrix(suv2, nrow=length(suv1), ncol=length(suv2), byrow=T)
        ej<-matrix(eve2, nrow=length(suv1), ncol=length(suv2), byrow=T)
        ukmn<-yi-yj
        pok<- sum((ukmn<0)*ei)
        nek<- sum((ukmn>0)*ej)
        pkhat<- pok/(pok+nek)
        }
```

```
##### R function for Jackknifed Akritas - Lavalley Method #####
AkLav<- function(grp1="A1", grp2="A2", indata=X){
        suv1<-indata$x[indata$arm==grp1]
    eve1<-indata$event[indata$arm==grp1]
        suv2<-indata$x[indata$arm==grp2]
    eve2<-indata$event[indata$arm==grp2]
    yi<-matrix(suv1, nrow=length(suv1), ncol=length(suv2), byrow=F)
        ei<-matrix(eve1, nrow=length(suv1), ncol=length(suv2), byrow=F)
    yj<-matrix(suv2, nrow=length(suv1), ncol=length(suv2), byrow=T)
        ej<-matrix(eve2, nrow=length(suv1), ncol=length(suv2), byrow=T)
        a=min(suv1)-max(suv2)
        b=max(suv1)-min(suv1)
        y=1
        x1=a
        x2=b
        while (abs(y)> 1E-3 ) {
```

```
            x=mean(c(x1,x2))
            yib1<- (yi-x)<yj
            yib2<- (yi-x)>yj
            y=sum(eve1*yib1-eve2*yib2)
            if (y>0) {x2=x}
            else if (y<0) {x1=x}
            if (abs(y)<=1E-3) {bhat=x}
            else if (abs(y)>1E-3 && abs(x2-x1) <=1E-3) {bhat=mean(c(x1, x2))
                                                        y=0
                                                       }
        }
    return(bhat)
        }

##### R function for Jackknifed Akritas - Brunner Method #####
AkBru<-function(indata=X, arm='arm=A1') {
surv<-summary(survfit(formula = Surv(x, event) ~ 1, data = indata))
survarm<-summary(survfit(formula = Surv(x, event) ~ arm, data = indata))
```

```
svall=data.frame(time=surv$time, SH=surv$surv)
svgrp=data.frame(time=survarm$time, F=1-survarm$surv,arm=survarm$strata)
sv=merge(svall, svgrp, by="time")
m<-X[X$event==1,]
t<-min(by(m$x, m$arm, max))
final=sv[sv$time<=t,]
absub=final[final$arm==arm,]
leng=length(absub$time)
if (leng >1) {val=sum(absub$SH[1:(leng-1)]*(absub$F[2:leng]-absub$F[1:(leng-1)]))}
else {val=0}
}
#####Define  kernel function #####
kn <- function(x) {15/16*(1-2*x^2+x^4)*(abs(x) <= 1)}
Kn <- function(x) {
                    0*(x<0)+15/16*(x-2/3*x^3+.2*x^5+8/15)*(x<=1&x>-1)+1*(x>1)
                    }
```

####### End of definition of R functions #######

```
library(survival)
sim<-1000
Mu<-vector('numeric',sim)
sde<-vector('numeric',sim)
pathomu<-vector('numeric', sim)
pathosde<-vector('numeric', sim)
aklvmu<-vector('numeric', sim)
aklvsde<-vector('numeric', sim)
akbrmu<-vector('numeric', sim)
akbrsde<-vector('numeric', sim)
coxpv<-vector('numeric',sim)
#pvalue<-vector('numeric',sim)
#aklvpv<-vector('numeric',sim)

#####Set up the Parameters and Input Parameters Based on Simulated Scenarios###
sizea1 <-50
```

```
sizea2 <-50

sizeb1 <-50

sizeb2 <-50

size <- sizea1 + sizea2 + sizeb1 + sizeb2

lamdaa1 <- 0.1

lamdab1 <- 0.1

lamdaa2 <- 0.1

lamdab2 <- 0.1

T0 <-6

Tf <-18

##### Starts Simulation

for (i in (1:sim)) {

set.seed(372+i)

## generating random sample

x<- vector('numeric',  size)

arm <- vector('character',  size)

event <- vector('numeric',  size)

x[1:sizea1]<-rexp(sizea1, lamdaa1)
```

```
x[(sizea1+1) : (sizea1 + sizea2)]<-rexp(sizea2, lamdaa2)

x[(sizea1 + sizea2+1) : (sizea1 + sizea2 +sizeb1)]<-rexp(sizeb1, lamdab1)

x[(sizea1 + sizea2 +sizeb1+1) : size]<-rexp(sizeb2, lamdab2)

arm[1:sizea1]<-'A1'

arm[(sizea1+1) : (sizea1 + sizea2)]<-'A2'

arm[(sizea1 + sizea2+1) : (sizea1 + sizea2 +sizeb1)]<-'B1'

arm[(sizea1 + sizea2 +sizeb1+1) : (size)]<-'B2'

acr<-T0 * runif(size, 0,1)

event <- 1*(  (acr + x) <= (T0 + Tf))

x[event==0] <-(T0 +Tf) - acr[event ==0]

X <- data.frame(x, event, arm)

#######generate numeric treatment variable and baseline factor

X$trt[X$arm=='A1' | X$arm=='A2'] <-1

X$trt[X$arm=='B1' | X$arm=='B2'] <-0

X$pf[X$arm=='A1' | X$arm=='B1'] <-1

X$pf[X$arm=='A2' | X$arm=='B2'] <-0

#### run cox model to detect any interation

cox <-coxph(formula = Surv(x, event) ~ trt + pf + trt * pf, data = X)
```

```
coxpv[i]<-summary(cox)$coefficients[3,5]


############## Call R Function for Akritas and Brunner method #######
akbrmu[i]<-AkBru(indata=X, arm='arm=A1')-AkBru(indata=X, arm='arm=B1')-
AkBru(indata=X, arm='arm=A2')+AkBru(indata=X, arm='arm=B2')


############## Call R Function for Smoothed Method ###############
Mu[i]<-Pdiff(X)
############## Call R Function for Schemper  Method  ###############
pathomu[i]<-patho(grp1="A1", grp2="B1", indata=X)- patho(grp1="A2", grp2="B2",
indata=X)
############## Call R Function for Akritas-Lavalley Method ##########
aklvmu[i]<-AkLav(grp1="A1", grp2="B1", indata=X)- AkLav(grp1="A2",
grp2="B2", indata=X)


############## Jackknife for Getting Standard Errors ###############
JackMu<-vector('numeric',size)
```

```
pathojackmu<-vector('numeric',size)
aklvjackmu<-vector('numeric',size)
akbrjackmu<-vector('numeric',size)
for (j in (1:(size))) {
        X1<-X[-j,]
        JackMu[j]<-Pdiff(X1)
        pathojackmu[j]<-patho(grp1="A1", grp2="B1", indata=X1)- patho(grp1="A2",
grp2="B2", indata=X1)
        aklvjackmu[j]<-AkLav(grp1="A1",grp2="B1",indata=X1)-AkLav(grp1="A2",
grp2="B2", indata=X1)
      akbrjackmu[j]<-AkBru(indata=X1,arm='arm=A1')-AkBru(indata=X1,
arm='arm=B1')-AkBru(indata=X1,arm='arm=A2')+AkBru(indata=X1, arm='arm=B2')
}
JackMu<-size*Mu[i]-(size-1)*JackMu
pathojackmu<-size*pathomu[i]-(size-1)*pathojackmu
aklvjackmu<-size*aklvmu[i]-(size-1)*aklvjackmu
akbrjackmu<-size*akbrmu[i]-(size-1)*akbrjackmu
sde[i]<-sd(JackMu)/sqrt(size)
```

```
pathosde[i]<-sd(pathojackmu)/sqrt(size)
aklvsde[i]<-sd(aklvjackmu)/sqrt(size)
akbrsde[i]<-sd(akbrjackmu)/sqrt(size)
}
#######Getting Type I Errors ########
pvalue<-2 * pnorm(-abs(Mu/sde))
pathopv<-2*pnorm(-abs(pathomu/pathosde))
aklvpv<-2*pnorm(-abs(aklvmu/aklvsde))
akbrpv<-2*pnorm(-abs(akbrmu/akbrsde))
pwsl<-sum(pvalue <= 0.05)/sim
pwpatho<-sum(pathopv <= 0.05)/sim
pwaklv<-sum(aklvpv <= 0.05)/sim
pwakbr<-sum(akbrpv <= 0.05)/sim
pwcox<-sum(coxpv<=0.05)/sim
############ Output all parameter to file with simulation results #########
sink(file="H:/PhD project/output/typeIerrors1", append=T)
output<-c(sim, size, T0, Tf, lamdaa1, lamdab1, lamdaa2, lamdab2, sizea1, sizeb1,
sizea2, sizeb2)
```

```
names(output)<-c("sim=", "size=", "T0=", "Tf=", "lamdaa1=", "lamdab1=",
"lamdaa2=", "lamdab2=", "sizea1=", "sizeb1=", "sizea2=", "sizeb2=")


print(output)
outpw<-c(pwsl, pwpatho,  pwaklv,  pwakbr, pwcox)
names(outpw)<-c("mymethodpw=", "Pat-Ho methodpw=", "aklvmethodpw=",
"akbrmethodpw=", "coxmethodpw=")
print(outpw)
sink ()
```